gtexture: Haralick texture analysis for graphs and its application to biological networks

R Barker-Clarke[1*], D Weaver[1,2], and J G Scott[1,21]**

[1]Department of Translational Hematology & Oncology Research, Lerner Research Institute, Cleveland, OH 44195, United States

[2]School of Medicine, Case Western Reserve University, Cleveland, OH 44195, United States

[*]rowanbarkerclarke@gmail.com

[**]ScottJ10@ccf.org

ABSTRACT

The calculation and use of Haralick texture features has been traditionally limited to imaging data and gray-level co-occurrence matrices calculated from images. We generalize the calculation of texture to graphs and networks with node attributes, focusing on cancer biology contexts such as fitness landscapes and gene regulatory networks with simulated and publicly available experimental gene expression data. We demonstrate the potential to calculate texture over multiple data set types including complex cancer networks and illustrate the potential for texture to distinguish cancer types and topologies of evolutionary landscapes through the summary metrics derived.

Keywords: GLCM, image analysis, networks, texture analysis

Introduction

Topology and texture have been studied widely across biomedical research, with textural and topological analysis methods providing insight in medical imaging, in the analysis of biological signalling networks, and genotype-phenotype maps of evolution (Fig. 1). Across these fields textures and topologies have been used to identify biologically meaningful structures or patterns and have, in some applications to cancer, been associated with clinical outcomes.

Within medical disciplines, traditional image analysis heavily utilizes "texture" features, derived from a staple within imaging, the gray-level co-occurrence matrix (GLCM)(Haralick 1979). GLCMs are 2D histograms that record the frequency of neighboring pixel gray-level values in an image. The Haralick texture features summarize this distribution of value pairs and include measures that reflect heterogeneity, homogeneity and contrast within images. These are very commonly used in medical physics where texture features from CT and MRI images have been related to tumor type, severity and prognosis (Mohanty, Beberta, and Lenka 2011; Yang et al. 2012; Zulpe and Pawar 2012; Jain 2013; Torheim et al. 2014; Novitasari et al. 2019). We note that co-occurrence matrices, although most commonly used in imaging, have also been used within NLP fields (Momtazi, Khudanpur, and Klakow 2010; Benoit et al. 2018), audio processing (Terzopoulos 1985; Sayedelahl et al. 2011; Muhammad et al. 2017) and recently in pathology in a form derived by Saito et. al., describing the co-occurrence of nuclear features in physical cell neighborhoods (Saito et al. 2016).

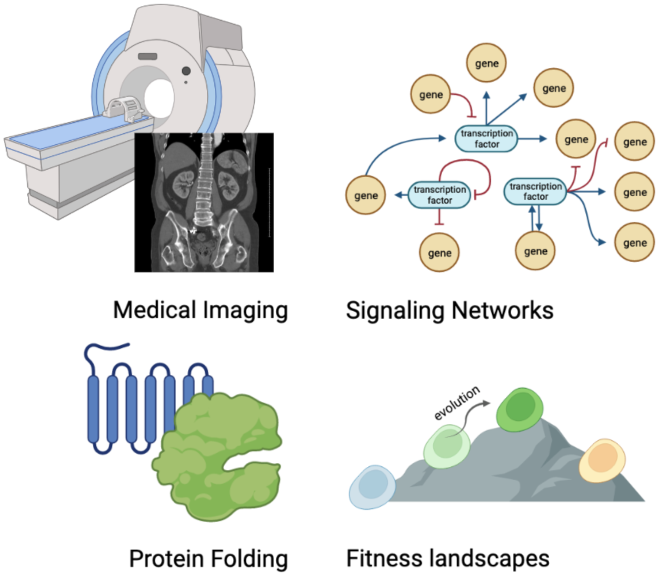


Figure 1. Illustration of areas in which topology is studied in biomedical research. Textural and topological studies are carried out in medical imaging, protein folding, signalling network and fitness landscape analysis.

Recent interdisciplinary work has successfully extended different graph-based topological analyses to image derived point clouds and more recently to images themselves, including the use of cubical complexes to derive prognostic topological features from medical images (Lawson et al. 2019; Hajij, Zamzmi, and Batayneh 2021; Somasundaram, Litzler, et al. 2021; Somasundaram, Wadhwa, et al. 2022).

Topology has also shed light on biological networks. As increasing amounts of proteomic and transcriptomic data become available, there arises a wealth of information about gene expression and protein-protein interaction networks. Within cancer, the frequent dysregulation of signalling pathways and modified interactions between mutant proteins means that holistic network analyses may have the potential to identify critical features in these data sets. Topological analysis of gene and protein networks has identified regulating gene sub-networks for potential drug targeting, improved understanding of the stability of gene signalling networks and even given prognostic indications in breast cancer (**weaver2021network**; Sardiu et al. 2019; Kumar, Blondel, and Extavour 2020; Guo and Amir 2021; Yin et al. 2021).

Another area within biology in which topology has been of interest is in the study of fitness landscapes(Lum et al. 2013), a special subclass of networks. Fitness landscapes typically encode a genotype space and their associated fitnesses. In cancer these are of particular interest as fitness landscapes encode the constraints of Darwinian evolution and are informative in the modelling of resistance and optimization of treatment (J. Scott and Marusyk 2017; Nichol, Rutter, et al. 2019; King et al. 2022). As the topology of a landscape can restrict or promote access to certain evolutionary trajectories, constraining the accessibility of local and global maxima (Levinthal 1997) measures have been developed to evaluate landscape "ruggedness" (Barnett et al.

1998). Modelling of "tunably rugged" landscapes has allowed the direct exploration of the effect of topology and texture upon evolution, demonstrating strong associations with evolutionary timescales and outcomes (Kauffman and Weinberger 1989; Barnett et al. 1998; Franke et al. 2011). As the ability to engineer and measure fitness landscapes experimentally has become easier, the nature of fitness landscapes is of growing interest; particularly in modern studies of evolutionary cancer therapies, drug resistance and biological control(Nichol, Jeavons, et al. 2015; Diaz-Uriarte 2018; Nichol, Rutter, et al. 2019; Hosseini et al. 2019; Iram et al. 2021; Hsu et al. 2022).

The aim of this work is to extend topological research by bringing the tools of image analysis to analyze network structures. In particular we believe we can gain new perspectives on networks in biological contexts. We present our method and associated package for calculating GLCM-equivalents and Haralick texture features and apply it to several network types. We developed the translation of co-occurrence matrix analysis to generic networks for the first time. We analyze networks with accompanying categorical and continuous node attributes, demonstrating this method on examples of social networks, protein-interaction networks in cancer and evolutionary fitness landscapes(see Fig. 1 for illustration).

Our R package for calculating texture of graphs, **gtexture**, is available at https://github.com/sbarkerclarke-phd/gtexture.

Methodology

We show how co-occurrence matrices and texture calculations can be generated from and applied to graph objects. Co-occurrence matrices are 2D histograms, traditionally reflecting the pairwise distribution of neighboring pixel values in images. To apply this method to graphs or networks they must have node attributes or weights. These weights can be in the form of discrete weights or ordered categorical attributes. We consider node attributes to be analogous to pixel values and a nodes' edges to be equivalent to pixel neighborhoods. Co-occurrence matrices can be described in network terms as node-weight adjacency matrices.

Network examples

To demonstrate the method, we used multiple network examples. We utilize social networks, gene expression networks and fitness landscapes. The Cross-Parker networks (Cross and Parker 2004) from the **tnet** R package (Opsahl 2009) provide an ideal example for demonstrating methods on graph structures. These networks are from a manufacturing company (77 employees) and a consulting company (56 employees). We used these structures to compare the original network structure to bootstrapped networks with randomised node attributes.

In order to analyse gene expression within graphical structures of established human protein-protein interaction networks, we used STRINGDB, the KEGG database and the KEGGGraph package(Szklarczyk et al. 2015; Zhang and Wiemann 2009; Ogata et al. 1998). These frameworks were used to obtain pathway specific subnetworks and to convert between gene and protein identifiers to assign gene expression to network nodes. To look at experimental gene expression on these networks we used the publicly available Cancer Cell Line Encyclopedia (CCLE) gene expression dataset (Barretina et al. 2012). For comparison to experimental data we also used the R package **graphsim** as a method of simulating gene expression values on

PI3-Kinase and TGF-Beta co-expression networks with varying correlation strength(Kelly and Black 2020).

Another specialized network type is the evolutionary fitness landscape. Genotypes in the fitness landscape are neighbors, connected by an edge if they are accessible through a single evolutionary timestep (eg. mutation). The underlying network structure is defined by this evolutionary access and the node weights are the fitness values. The number of experimental fitness landscapes that have been published is limited and as such we lacked graphically connected landscapes with measured fitnesses under different conditions to compare metrics across. We therefore utilized basic landscapes networks with specific fitness distributions to demonstrate our methodology. Utilizing the R package **OncoSimulR** (Diaz-Uriarte 2017) we generated three classes of basic model landscape and sets of NK landscapes and converted these into fitness landscape objects using the R package **fitscape**.

Additive model landscapes In the additive model, mutations have a specific fitness increase or decrease and multiple mutations increase or decrease fitness in a linear, additive fashion.

Eggbox model landscapes In the eggbox model there are only 2 different possible fitness values, the base fitness and base fitness + e (the “height” of the eggbox), each mutation swaps a genotype from low to high fitness, neighboring genotype fitness values are always distinct.

House of cards model landscapes The House of Cards (HOC) model is a name for a random fitness model, here the fitnesses of different genotypes are uncorrelated and not dependent on the genotype, this is an effective null/random model.

The outline of the method and approach underlying the discretization, co-occurrence and texture calculations, follow below.

Discretization

Given a number of nodes n, a network’s adjacency matrix is size $n \times n$. If the number of distinct node weights is w, the dimension of the co-occurrence matrix, C, is $w \times w$. Co-occurrence matrices summarize a network when the number of distinct node weights is less than the number of nodes, $w < n$. Although this is already the case for some networks, we provide methods to reduce the number of unique node weights, including node weight binning options for continuous node weights within the package. Continuous data can be transformed via several discretisation methods.

The following methods are useable within the package and several are demonstrated within **Fig. 2**.

Equal: We can use a breaks method to slice the node weights into n equally spaced levels containing potentially different proportions of the data.

Quantiles: In this method the values are split into n groups containing equal numbers of values.

k-means: Values are split into $n = k$ groups using 1D kmeans clustering.

Co-occurrence matrix calculation

For any graphical structure, the edges between nodes are captured in an adjacency matrix. These edges are used for the calculation of the distribution of co-occurring neighbor pairs. In an undirected network (symmetric adjacency matrix), the

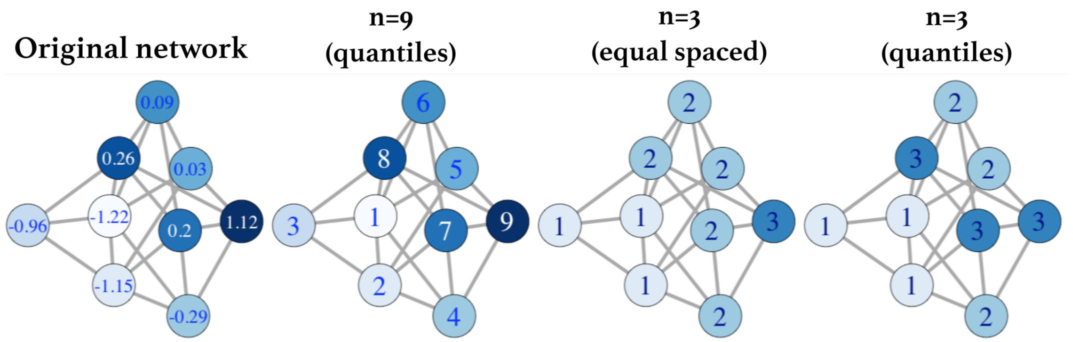
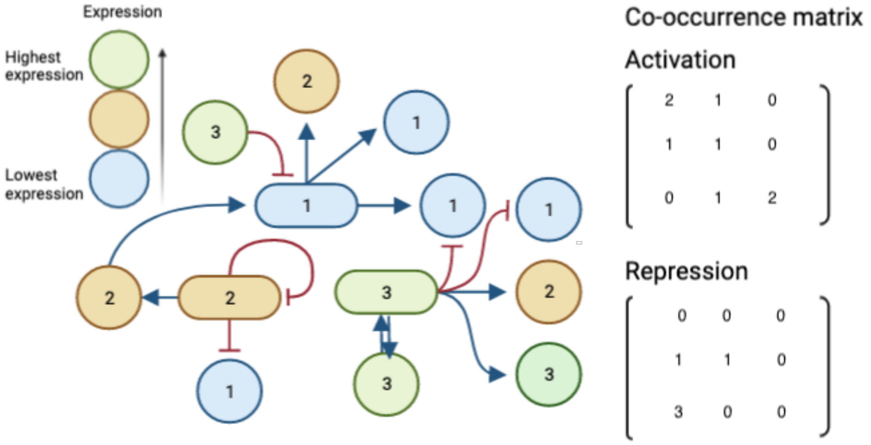


Figure 2. A demonstration of different discretization methods for continuous node values is shown. One example of a randomly generated undirected network with different random continuous expression values attributed to the nodes is shown. Discretization with 9 quantile levels matching the number of unique values and 3 levels with both equally spaced numerical bins and with 3 levels assigned to tertile groups are shown.



$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 0 & 0 \end{bmatrix}$$

Figure 3. Co-occurrence matrices calculated on a toy gene regulation network. In the case of a directed graph only directions included are counted. In directed activation and repression graphs two separate co-occurrence matrices can be calculated.

neighboring node values are summed over all edges. In a directed graph, the adjacency matrix is used directly to iterate through pairs of connected node values in a single direction. The element C_{ij} of the co-occurrence matrix is the number of times within the network a node with weight i shares an edge with a node of weight j. Examples of two separate co-occurrence matrices for a toy gene regulation network with four bins of expression values are shown in **Fig. 3.**

Haralick feature metrics and comparison

Standard image analysis practice uses the co-occurrence matrix to generate texture features for the image. Haralick defined several statistical features and these calculations on the co-occurrence matrix traditionally reflect properties of an image's texture (Haralick 1979). The definitions of eight of these key texture features calculated in this paper are shown in **Table 1**. Our package extracts these features and in order to compare these features across different categories of network, metrics are normalized across compared groups. -5mm

Feature		Term	Definition
Energy	$\sum_i \sum_j p(i,j)^2$	$p(i,j)$	Probability neighboring nodes have weights (i,j)
Contrast	$\sum_{n=0}^{N_g-1} n^2 (\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j))$ where $\lvert i-j \rvert = n$	$p_x(i,j)$	Marginal probability distribution over rows
Correlation	$\frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$	$p_y(i,j)$	Marginal probability distribution over columns
Entropy	$-\sum_i \sum_j p(i,j) \log(p(i,j))$	μ_x	mean of $p_x(i,j)$
Autocorrelation	$\sum_i \sum_j (i \cdot j) p(i,j)$	μ_y	mean of $p_y(i,j)$
Homogeneity	$\sum_i \sum_j \frac{p(i,j)}{1+(i-j)^2}$		
Cluster Shade	$\sum_{i=0}^{G} \sum_{j=0}^{G} (i+j-\mu_x-\mu_y)^3 p(i,j)$		
Cluster Prominence	$\sum_{i=0}^{G} \sum_{j=0}^{G} (i+j-\mu_x-\mu_y)^4 p(i,j)$		

Table 1. Definitions of a selection of Haralick texture features and variables needed to calculate them.

Results

In order to demonstrate both the efficacy and potential of texture analysis as applied to networks we apply our method to a selection of biological and cancer specific networks. Networks and graphs, as a general mathematical structure, have been great tools for encapsulating biological information which has underlying connected structure, we utilize several categories of graphs in order to demonstrate co-occurrence calculation and texture feature generation. Due to the intrinsic heterogeneity and complexity of biology, we include an example of organisational social network before analysing gene expression networks with simulated and publicly available data for node values and fitness landscapes, all randomly generated, with defined fitness distributions and tunable ruggedness.

Application to Organization Social Networks

Social networks have been collected and analysed across many social structures, these networks typically contain hierarchical information but options for the joint analysis of both node labels and network structure are limited. In order to demonstrate how these metrics can be applied to graphs with ordered categorical node attributes we applied our pipeline to the Cross-Parker consulting and manufacturing networks (Cross and Parker 2004). These networks consist of nodes representing personnel. Edges represent familiarity of people with each other in the network, in the original data set these edges are weighted. We removed these weights for the purpose of our analysis. These graph datasets included node attributes, reflecting organisational level, the manufacturing company dataset also included information on tenure. In these networks analyses we weighted all connections equally. We created the co-occurrence matrices for these networks and compared the original graph to 100 bootstrapped graphs with randomisation of the allocation of original node values (**Fig. 4**).

When comparing the Haralick features generated for the same network structure with different node values, we can see clearly that the original network is a signficant outlier, suggesting the neighboring node values, differences in tenure and

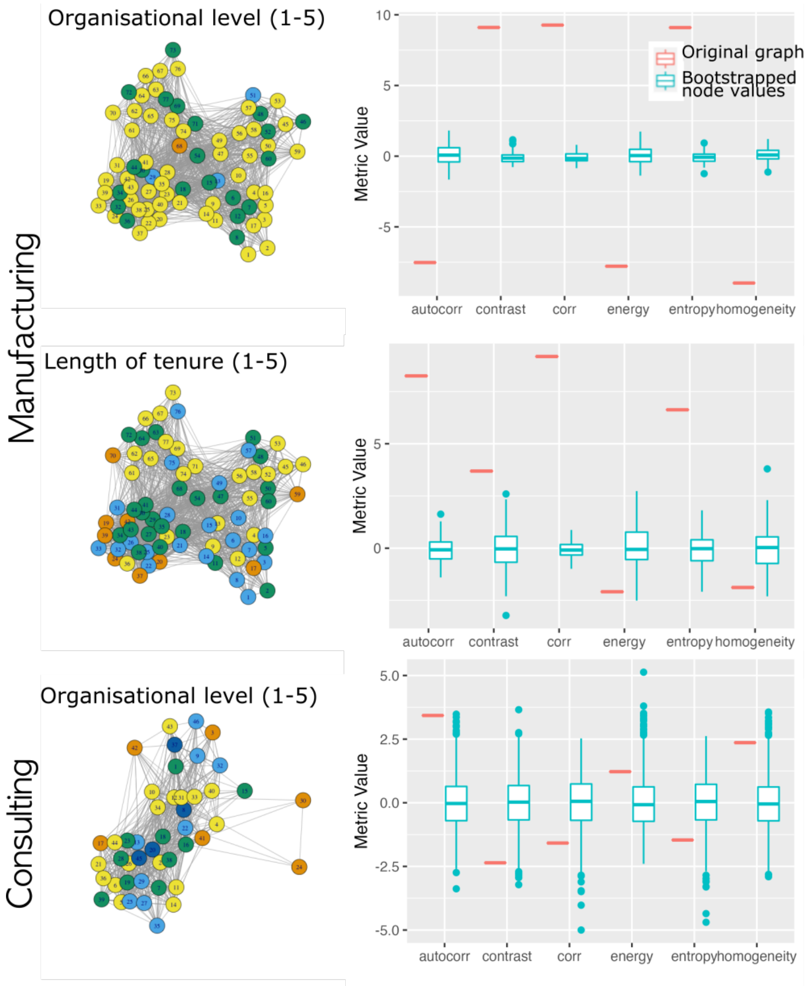


Figure 4. Haralick features in real-world networks differ greatly from random. Plots of Haralick features shown for two different social networks within a company, within manufacturing and consulting departments. Node values or attributes are organisational level and length of tenure where available. Edges reflect connected people within the network. Comparison distributions of metrics using bootstrapped node values are shown (n=1000). Clear departures from random distribution of node values between connected nodes are seen.

organizational status, in the manufacturing company lie outside the random distribution. This is reflective of a very hierarchical structure and a strong association with length of tenure demonstrates organisational and tenure-based structure.

In the case of the consulting company, autocorrelation is an outlier within the distribution, but other features are more randomly distributed, suggesting a less hierarchical organisational structure.

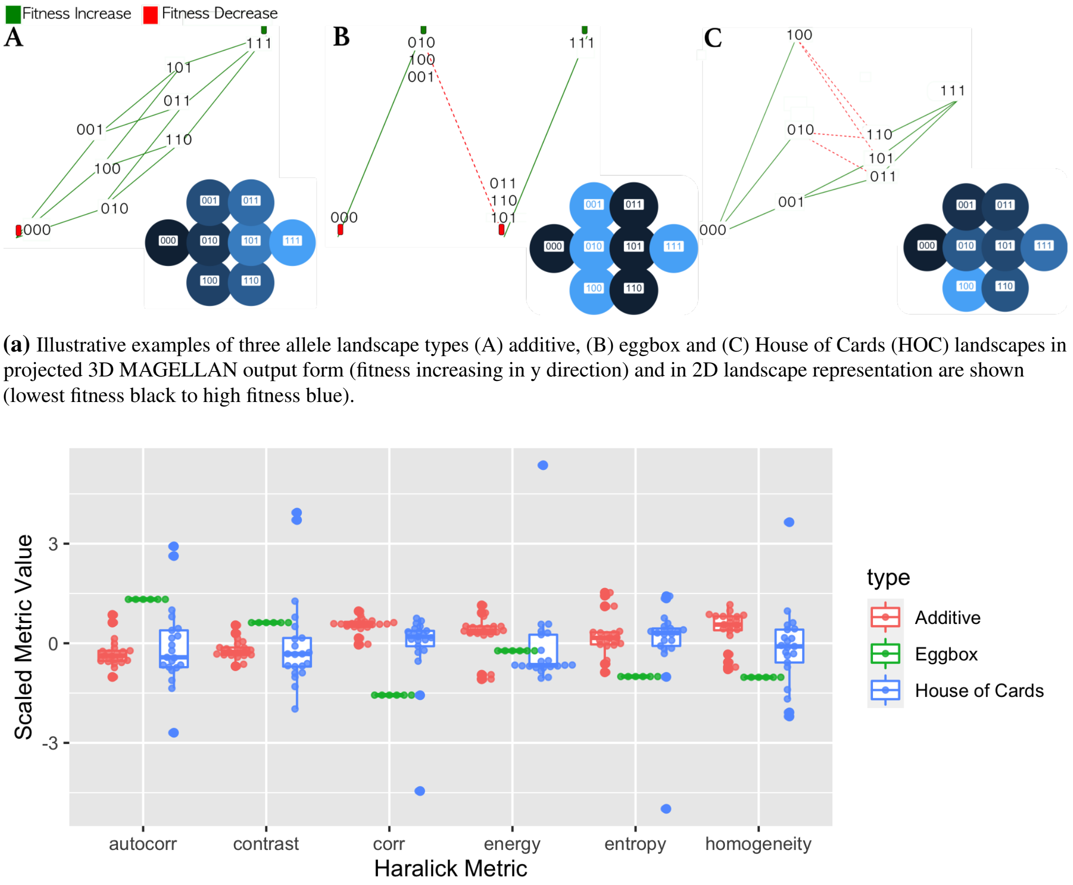
Application to fitness landscapes

Fitness landscapes encode the fitness (often considered as the growth rate) for an underlying distrbution of genotypes. As genomic processes such as mutation can allow the range of genotypes to be accessible via evolution, the different fitness values and connectivity (ie topology and texture) of a genotype landscape is associated with evolvability. Co-occurrence matrices and texture metrics may be valuable information generated from a fitness landscape, as these encode properties of the distribution of neighboring fitness values. Whilst the number of currently available experimental fitness landscapes is limited, statistically generated landscapes are available within packages such as *fitscape*. In order to assess the ability of co-occurrence matrices and

Haralick features to extract meaningful data from fitness landscapes we carried out our analysis pipeline to compare three basic landscape types and to compare tunably rugged landscapes. We modelled these landscape types with 4 alleles (16 genotypes).

Application to model landscapes

We utilized MAGELLAN, a fitness landscape analysis toolset, (Brouillet et al. 2015) to generate some standard models of fitness landscapes; additive, eggbox and house of cards. Single, illustrative examples of these landscapes are shown in **Fig. 5a**. We tested our pipeline on these models using 4 level node weight equal discretization on 4 allele (16 genotype) model landscapes.



(a) Illustrative examples of three allele landscape types (A) additive, (B) eggbox and (C) House of Cards (HOC) landscapes in projected 3D MAGELLAN output form (fitness increasing in y direction) and in 2D landscape representation are shown (lowest fitness black to high fitness blue).

(b) Scaled autocorrelation, contrast, correlation, energy, entropy and homogeneity are shown to differ in value and distribution across these types of artificial landscape.

Figure 5. Illustration of landscapes and distribution of GLCM metrics on them. **a)** Illustrative landscapes are shown for each type. **b)** "Eggbox" landscapes collapse under discretization and normalization. The eggboxes have highest contrast and lowest homogeneity as neighboring genotypes have alternating fitnesses, the additive model shows highest correlation, homogeneity and lowest contrast whereas the house of cards (HOC) model with its random fitnesses shows the largest range of values due to a wider spread of neighboring fitness pairs.

Traditional landscape models

For each basic network type we generated a set of 4 allele, 16 genotype fitness landscapes for analysis. We create sets of ten random additive, eggbox and House of Card landscapes. The Haralick texture features are calculated on these landscapes, and the normalized metrics are shown in **Fig. 5b** for comparison.

As expected the eggbox landscapes show the highest contrast and lowest homogeneity, the additive landscape shows the highest correlation and the random "House of Cards" landscape shows the largest variation in all the metrics.

Tunably rugged "NK" model landscapes

In order to demonstrate these metrics on some more realistic simulated fitness landscapes, we analysed a simulated set of tunably rugged "NK" landscapes. Our simulated landscapes had 4 alleles (16 genotypes) and we varied K from 1 to 3. For a 4 allele system, we generated 500 random "NK" landscapes for each value of K (1 to 3). We looked at the distribution of the Haralick features for these different landscape classes (Fig. S1). We compared these to some traditional measures of landscape ruggedness. As epistatic interactions increase, the contrast between neighboring fitness values decreases (therefore dissimilarity decreases). At K=0, the landscape is smooth and additive. As K is increased, the landscape becomes more rugged as epistatic interactions increase, the correlation increases with K.

Application to Gene Expression on PPI Networks

The development of biological networks has been driven by growing works in transcriptomic and proteomic studies. Protein protein interaction (PPI) networks have been built based upon experimental evidence probing the interaction of different proteins.

We hypothesized that our technique may be useful in assessing experimental data gathered in different samples for established biological networks, in particular as a way of summarizing expression patterns across different topologies of protein interaction networks.

We examined the phosphoinositide-3-kinase (Pi3K) cascade network and assiged gene expression values to nodes, using both the CCLE experimental dataset and simulated through the *graphsim* package. In order to evaluate our metrics under varying simulated gene expression, we varied the correlation parameter of the underlying expression simulation from 0.2 to 0.8. Expression levels from both the simulation and the CCLE data were discretized into 4 equal levels and these expression values used as node weights.

Fig. 6 shows the PI3K network and how the Haralick metrics vary with increasing expression correlation. In the network describing Pi3K regulation we see the expected results, that contrast decreases, correlation increases, entropy decreases and homogeneity increases as correlation in the underlying expression simulation increases. When the gene expression data from the cancer cell lines in the CCLE is compared, we see that these are significantly different (more extreme) than metrics upon the simulated expression, showing results that correspond to increased correlation strengths, lying outside the simulated distributions.

We decided to examine the distribution of metrics within the CCLE in more detail, analysing a biological sub-network with expected differences in expression patterns between cell lines. We calculated the metrics using the EGFR signalling pathway subnetwork and the metrics on this network with the expression values for some of the most common cancer subtypes within the dataset.

EGFR (epithelial growth factor) dysregulation is associated with solid tumors and we see corresponding differences in

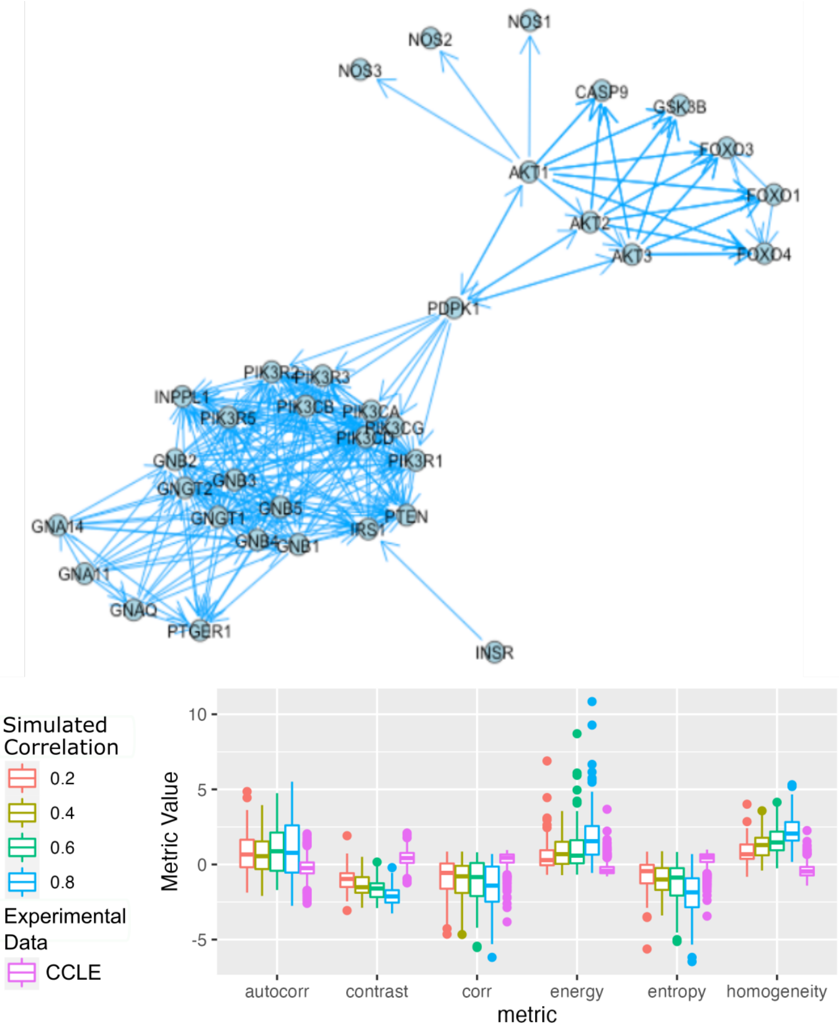


Figure 6. Pi3Kinase gene network. Texture features generated with simulated expression data and experimental CCLE gene expression data (pink). Plots of Haralick features shown for simulated gene expression on the Pi3K gene network with different strengths of co-expression correlation show a trend across node value correlation strength.

metric values between the epithelial (solid - lung, breast, ovary, central nervous system, prostate and skin) and non-epithelial (blood, lymphoctye) cell line samples (**Fig.** 7). EGFR amplification is a particularly common feature of glioblastoma, a large proportion of the CNS tumors and we see this reflected in more extreme metric values for CNS tumors. To assess whether there are differences between the metrics for primary and metastatic samples in tumors with likely EGFR dysregulation, we also analysed the same metrics between primary and metastatic cell-lines with central nervous system origin (**Fig.** S2). We find significant differences in the metric distributions between primary and metastatic cell lines.

Discussion

Network studies in cancer are generating increasing numbers of experimental datasets and provide a rich resource for novel analysis methods. With the generation and availability of many types of cancer-related models and networks, including the generation of fitness landscapes and bulk and single cell gene expression and protein expression datasets, comes a need for

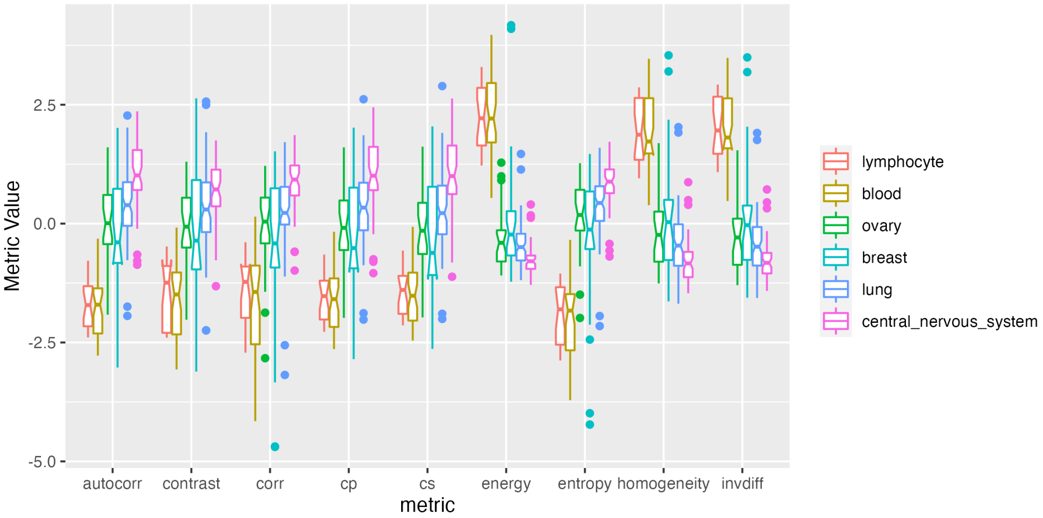


Figure 7. Metrics calculated from gene expression in primary samples of different lineages within the EGFR subnetwork. Gene expression data from the CCLE database was extracted for the genes in the EGFR pathway. Metrics were calculated on this sub-network across 6 of the most represented cancer lineage types in the dataset. Epithelial tumors are separated from non-epithelial tumors in the dataset.

cross-disciplinary analysis methods. Network analysis techniques and summary statistics typically assess edge properties and topology but experiments contain large amounts of additional data about the nodes of a network, for example a gene or protein or cell-line. In order to analyze these node properties in tandem with the network, we must look beyond the most traditional network techniques. We demonstrate, for the first time, the generation of co-occurrence matrices and Haralick texture features as summary features of general networks. Suitable networks for this metric must have ordered node attributes or discrete or continuous node weights.

Our results demonstrate stark differences in texture between network types across social networks, cancer gene expression networks and simulated fitness landscape networks. We also demonstrate differences in texture between cell lines when using experimental data from different cancer types. As this is a new methodology, we decided to present these metrics upon interpretable and well understood network examples, leaving further biological research questions to future work.

Co-occurrence matrices upon networks reflect the relative occurrence of different pairs of node-values that are connected within a network or graph object, examples including gene expression of neighboring genes in a network or neighboring fitness values in a fitness landscapes.

Our method showed that the Haralick features calculated on different landscapes and networks of the same size but with different topologies vary. We demonstrate that these features correspond to properties of node value neighborhoods and graph topological features. The Haralick method can therefore successfully be applied to networks with node attributes and can measure network or fitness landscape topologies. The package provides a framework for the future study of the optimization of parameters such as number of discrete levels chosen to encode node values such as expression or fitness values. Although highly specific methods designed for detecting landscape ruggedness exist, this discretization and co-occurrence matrix method is more generalizable.

Although the GLCM texture features are well characterized in imaging, the true utility of these metrics upon networks has yet to be explored. By utilizing these ideas from image analysis, this method provides a simple analysis and summary technique that is particularly effective for larger network types with node-specific intensities. As the fitness landscape data generated and collected becomes larger, methods such as this that can reduce the dimensionality of complex networks while retaining information about structure may be useful. As such, this package provides an efficient computation of summary statistics for graphs with edges and discretizable node attributes.

We believe that this package can be applied to many network types, not just those represented here and may be able to derive statistics reflective of important network characteristics. This method can be applied, for example, to fitness and growth rate data, gene expression, protein expression, time series data and cross-sectional data. We encourage the use of this package in exploratory network analyses and cancer network analysis and communication of any findings with the authors and the wider community.

References

Barnett, Lionel et al. (1998). “Ruggedness and neutrality: The NKp family of fitness landscapes”. In: *Artificial Life VI: Proceedings of the sixth international conference on Artificial life*, pp. 18–27.

Barretina, Jordi et al. (2012). “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391, pp. 603–607.

Benoit, Kenneth et al. (2018). “quanteda: An R package for the quantitative analysis of textual data”. In: *Journal of Open Source Software* 3.30, p. 774.

Brouillet, S. et al. (Nov. 2015). “MAGELLAN: a tool to explore small fitness landscapes”. In: *bioRxiv*, p. 031583. DOI: 10.1101/031583. eprint: 031583.

Cross, Robert L and Andrew Parker (2004). *The hidden power of social networks: Understanding how work really gets done in organizations*. Harvard Business Press.

Diaz-Uriarte, Ramon (2017). “OncoSimulR: genetic simulation with arbitrary epistasis and mutator genes in asexual populations”. In: *Bioinformatics* 33.12, pp. 1898–1899.

— (2018). “Cancer progression models and fitness landscapes: a many-to-many relationship”. In: *Bioinformatics* 34.5, pp. 836–844.

Franke, Jasper et al. (2011). “Evolutionary accessibility of mutational pathways”. In: *PLoS computational biology* 7.8, e1002134.

Guo, Yipei and Ariel Amir (2021). “Exploring the effect of network topology, mRNA and protein dynamics on gene regulatory network stability”. In: *Nature communications* 12.1, pp. 1–10.

Hajij, Mustafa, Ghada Zamzmi, and Fawwaz Batayneh (2021). “TDA-Net: Fusion of Persistent Homology and Deep Learning Features for COVID-19 Detection From Chest X-Ray Images”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 4115–4119.

Haralick, Robert M (1979). “Statistical and structural approaches to texture”. In: *Proceedings of the IEEE* 67.5, pp. 786–804.

Hosseini, Sayed-Rzgar et al. (2019). “Estimating the predictability of cancer evolution”. In: *Bioinformatics* 35.14, pp. i389–i397.

Hsu, Teng-Kuei et al. (2022). “A general calculus of fitness landscapes finds genes under selection in cancers”. In: *Genome Research*, gr–275811.

Iram, Shamreen et al. (2021). “Controlling the speed and trajectory of evolution with counterdiabatic driving”. In: *Nature Physics* 17.1, pp. 135–142.

Jain, Shweta (2013). “Brain cancer classification using GLCM based feature extraction in artificial neural network”. In: *International Journal of Computer Science & Engineering Technology* 4.7, pp. 966–970.

Kauffman, Stuart A. and Edward D. Weinberger (Nov. 1989). “The NK model of rugged fitness landscapes and its application to maturation of the immune response”. In: *J. Theor. Biol.* 141.2, pp. 211–245. ISSN: 0022-5193. DOI: 10.1016/S0022-5193(89)80019-0.

Kelly, S. Thomas and Michael A. Black (2020). “graphsim: An R package for simulating gene expression data from graph structures of biological pathways”. In: *Journal of Open Source Software* 5.51, p. 2161. DOI: 10.21105/joss.02161. URL: https://doi.org/10.21105/joss.02161.

King, Eshan S et al. (2022). “Fitness seascapes facilitate the prediction of therapy resistance under time-varying selection”. In: *bioRxiv*.

Kumar, Tarun, Leo Blondel, and Cassandra G Extavour (2020). “Topology-driven protein-protein interaction network analysis detects genetic sub-networks regulating reproductive capacity”. In: *Elife* 9.

Lawson, Peter et al. (2019). “Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology”. In: *Scientific reports* 9.1, pp. 1–15.

Levinthal, Daniel A. (July 1997). “Adaptation on Rugged Landscapes”. In: *Manage. Sci.* URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.43.7.934.

Lum, Pek Y et al. (2013). “Extracting insights from the shape of complex data using topology”. In: *Scientific reports* 3.1, pp. 1–8.

Mohanty, Aswini Kumar, Swapnasikta Beberta, and Saroj Kumar Lenka (2011). “Classifying benign and malignant mass using GLCM and GLRLM based texture features from mammogram”. In: *International Journal of Engineering Research and Applications* 1.3, pp. 687–693.

Momtazi, Saeedeh, Sanjeev Khudanpur, and Dietrich Klakow (2010). “A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 325–328.

Muhammad, Ghulam et al. (2017). “Enhanced living by assessing voice pathology using a co-occurrence matrix”. In: *Sensors* 17.2, p. 267.

Nichol, Daniel, Peter Jeavons, et al. (2015). “Steering evolution with sequential therapy to prevent the emergence of bacterial antibiotic resistance”. In: *PLoS computational biology* 11.9, e1004493.

Nichol, Daniel, Joseph Rutter, et al. (2019). “Antibiotic collateral sensitivity is contingent on the repeatability of evolution”. In: *Nature communications* 10.1, pp. 1–10.

Novitasari, Dian Candra Rini et al. (2019). “Application of feature extraction for breast cancer using one order statistic, GLCM, GLRLM, and GLDM”. In: *Advances in Science, Technology and Engineering Systems Journal* 4.4, pp. 115–120.

Ogata, Hiroyuki et al. (1998). “Computation with the KEGG pathway database”. In: *Biosystems* 47.1-2, pp. 119–128.

Opsahl, Tore (2009). “Structure and evolution of weighted networks”. PhD thesis. Queen Mary, University of London.

Saito, Akira et al. (2016). “A novel method for morphological pleomorphism and heterogeneity quantitative measurement: Named cell feature level co-occurrence matrix”. In: *Journal of pathology informatics* 7.1, p. 36.

Sardiu, Mihaela E et al. (2019). “Topological scoring of protein interaction networks”. In: *Nature communications* 10.1, pp. 1–14.

Sayedelahl, Aya et al. (2011). “Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 407–414.

Scott, Jacob and Andriy Marusyk (2017). “Somatic clonal evolution: a selection-centric perspective”. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1867.2, pp. 139–150.

Somasundaram, Eashwar, Adam Litzler, et al. (2021). “Persistent homology of tumor CT scans is associated with survival in lung cancer”. In: *Medical physics* 48.11, pp. 7043–7051.

Somasundaram, Eashwar, Raoul Wadhwa, et al. (2022). “Topology based radiomic feature derived from persistent homology predicts survival in non-small cell lung cancer patients treated with SBRT”. In: *medRxiv*.

Szklarczyk, Damian et al. (2015). “STRING v10: protein–protein interaction networks, integrated over the tree of life”. In: *Nucleic acids research* 43.D1, pp. D447–D452.

Terzopoulos, Demetri (1985). “Co-occurrence analysis of speech waveforms”. In: *IEEE transactions on acoustics, speech, and signal processing* 33.1, pp. 5–30.

Torheim, Turid et al. (2014). “Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines”. In: *IEEE transactions on medical imaging* 33.8, pp. 1648–1656.

Yang, Xiaofeng et al. (Sept. 2012). “Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: An in vivo study of late toxicity”. In: *Med. Phys.* 39.9, pp. 5732–5739. ISSN: 0094-2405. DOI: 10.1118/1.4747526.

Yin, Xin et al. (2021). “Identification of key modules and genes associated with breast cancer prognosis using WGCNA and ceRNA network analysis”. In: *Aging (Albany NY)* 13.2, p. 2519.

Zhang, Jitao David and Stefan Wiemann (2009). “KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor”. In: *Bioinformatics* 25.11, pp. 1470–1471.

Zulpe, Nitish and Vrushsen Pawar (2012). “GLCM textural features for brain tumor classification”. In: *International Journal of Computer Science Issues (IJCSI)* 9.3, p. 354.

Competing interests

The authors declare that they have no competing interests.

Author’s contributions

RB-C conceptualized the project, carried out coding implementation, experiment formulation and analysis. DW assisted with algorithm development and package formulation. JGS assisted with all aspects of conception, writing and development.

Acknowledgements

The authors thank members of Theory Division for their constructive feedback and valuable input in this project, in particular Raoul Wadhwa for assistance with package development in the early stages. J.G.S. was supported by the National Cancer Institute of the National Institutes of Health, under grant R37 CA244613.

Supplementary Information

When we vary the tunable ruggedness of simulated landscapes by varying K in the OncosimulR package, we see changes in the texture metrics (Supplementary Figure S1).

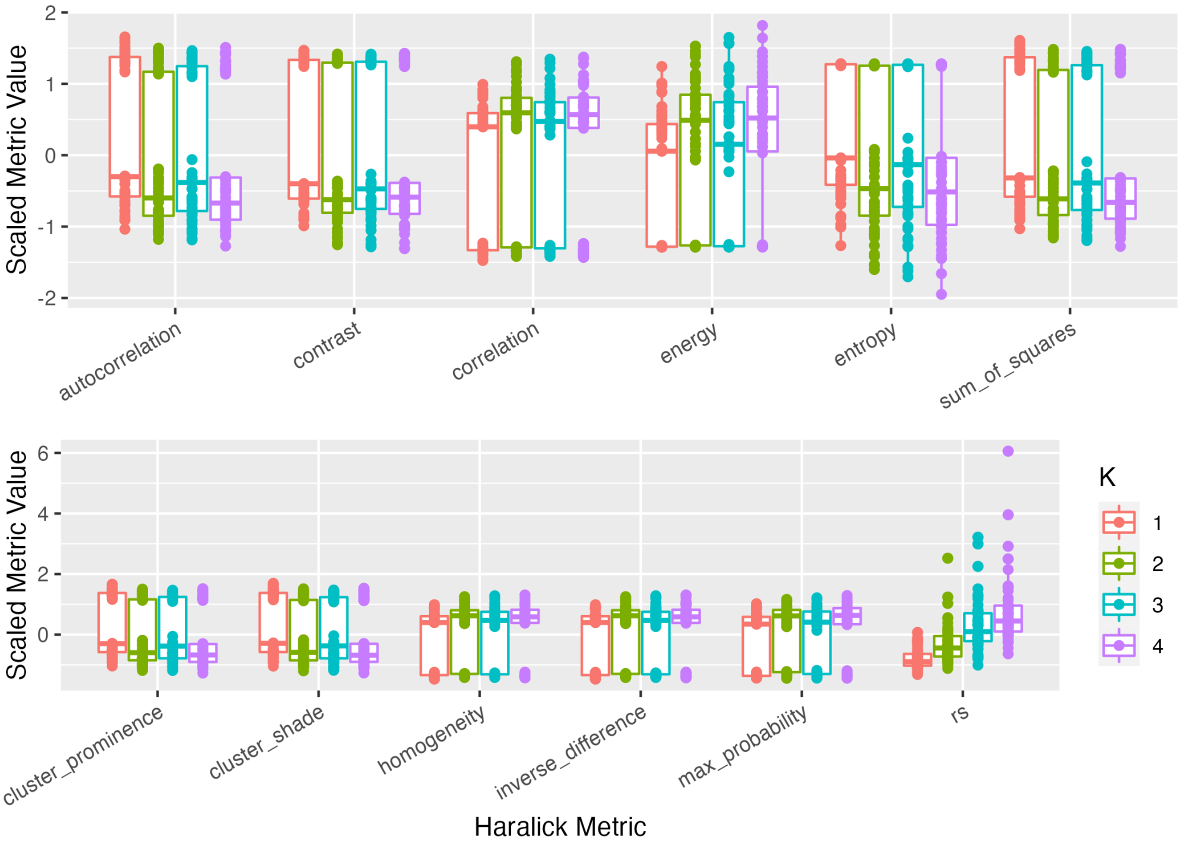


Figure S1. Metrics across NK landscapes. Haralick measures differ across different K for tunably rugged landscapes (5 alleles). Haralick features show distinct bimodal distributions of metrics for tunably rugged landscapes with fitnesses binned into 4 groups. The roughness-slope metric outperforms these in terms of separating landscapes with a single measure, but metrics contain information about landscape structure. Lines connect the same landscapes across different metrics.

The EGFR network and associated expression data shows significantly different texture between primary and metastatic central nervous system tumors (Supplementary Figure S2)

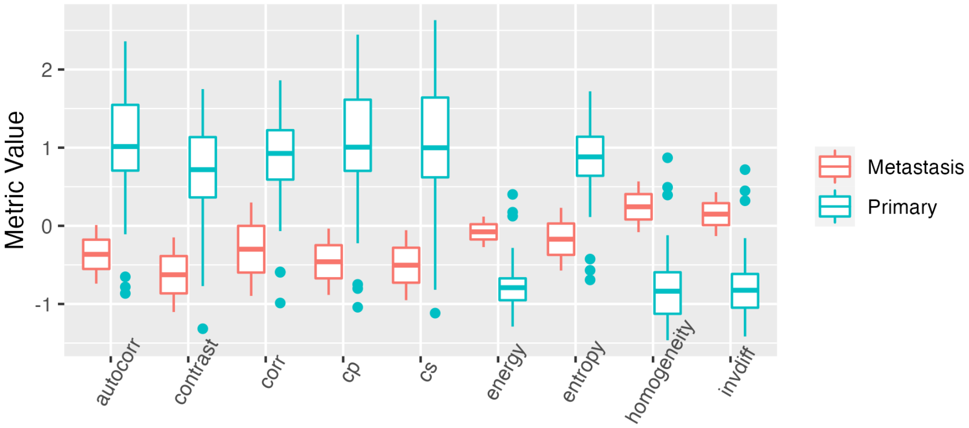


Figure S2. Primary vs Metastatic central nervous system samples for EGFR expression subnetwork using expression values from the CCLE database