A Non-Negative Tensor Factorization Approach to Deconvolute Microenvironment in Breast Cancer

Min Shi[1,2], Liubou Klindziuk[2] and Shamim Mollah[1,2]*

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO, 63110, USA and
[2]Institute for Informatics, Washington University School of Medicine, St. Louis, MO, 63110, USA

*To whom correspondence should be addressed.

Abstract

Motivation: An in-depth understanding of epithelial breast cell responses to the growth-promoting ligands is required to elucidate how the microenvironment (ME) signals affect cell-intrinsic regulatory networks and the cellular phenotypes they control, such as cell growth, progression, and differentiation. This is particularly important in understanding the mechanisms of breast cancer initiation and progression. However, the current mechanisms by which the ME signals influence these cellular phenotypes are not well established.

Results: We developed a high-order correlation method using proteomics data to reveal the regulatory dynamics among proteins, histones, and six growth-promoting ligands in the MCF10 cell line. In the proposed method, the protein-ligand and histone-ligand correlations at multiple time points are first encoded in two three-way tensors. Then, a non-negative tensor factorization model is used to capture and quantify the protein-ligand and histone-ligand correlations spanning all time points, followed by a partial least squares regression process to model the correlations between histones and proteins. Our method revealed the onset of specific protein-histone signatures in response to growth ligands contributing to distinct cellular phenotypes that are indicative of breast cancer initiation and progression.

Contact: smollah@wustl.edu

1 Introduction

There is increasing evidence that alterations in the cellular microenvironment affect normal tissue morphogenesis and promote breast cancer formation and progression (Spill *et al.*, 2016). Extracellular matrix (ECM), a major component of the cellular microenvironment, interacts with growth factors to influence cell behaviors contributing to cell growth, progression, and differentiation (Walker *et al.*, 2018). Details about these interactions within the complex and continuously changing breast microenvironment are not well understood and are important during tumor formation. Therefore, an in-depth understanding of these interactions within the tumor microenvironment is critical for early detection and designing effective anti-cancer therapies. These complex interactions are controlled by inter- and intra-pathways higher-order cell signaling involving a cascade of biochemical reactions and physical interactions among proteins, RNAs, DNA, and other biomolecules. These interactions are typically represented as a network encoding pairwise correlations between nodes optionally with some mathematical formulation of their dynamics. Usual mathematical formulations of the dynamics typically assume that each node v in the network has an associated "state" variable (representing, for example, expression of the corresponding signaling protein) as a function of the "time" variable t, describing the value of this variable at a node ("state" of the node) depends on the state of the nodes interacting with it. A major drawback of this type of network is that it does not consider the higher-order correlations of the interactions, which may lead to an inaccurate or incomplete view of the global interconnectivity of the signaling proteins and their functional contributions to cellular phenotype. For example, a network only encodes pairwise correlations of node state variables and thus cannot represent a joint k-way correlation among k state variables for any $k > 2$. In a recent study, Mollah et al., developed a 3-way correlation network model, iPhDNet, to represent causal connectivities among phosphoproteins, histones, and drugs using a hybrid machine learning approach (Mollah and Subramaniam, 2020). In this study, we developed a higher-order correlation model using tensor decomposition from measured proteomic signals. Our method constructs a signaling proteins-ligands-time 3-way tensor from signaling data and a histones-ligands-time 3-way tensor from histones data to understand the influences of various growth ligands on cell signaling pathways.

A large-scale initiative, Library of Integrated Network-Based Cellular Signatures (LINCS) (http://www. lincsproject.org), has carried out the multi-omics characterization of breast epithelial cells to six growth-promoting ligands, through measurement of reverse phase protein array (RPPA) signaling proteins (Creighton and Huang, 2015), and global chromatin profiles (GCP) (Litichevskiy *et al.*, 2018). These measurements were carried out at multiple time points post-treatment of cells. The RPPA and GCP are antibody base and Mass Spectrometry (MS)-based targeted proteomics assays that include a representative set of signaling proteins and different combinations of histone modifications treated by multiple growth ligands, respectively. This higher-order tensor model study offers the scope for obtaining key signatures of growth ligands response based on detailed mechanisms reconstructed from the measurements. Specifically, we present a novel approach to elucidate how microenvironment signals affect cell-intrinsic intracellular protein signals and histone networks to generate experimentally observable cellular phenotypes. Our methods identify histone modification fingerprints, which are endpoints of complex

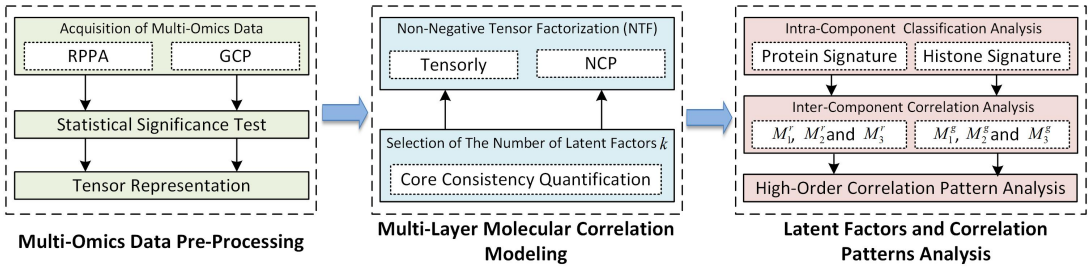


Fig. 1: Workflow of the proposed approach, which involves three major components including multi-omics data preparation, high-order data correlation modeling and results interpretation.

signaling events following growth ligand treatment. Chromatin topology changes, caused by the histone signatures, reflect the altered cellular state. For characterization of the breast microenvironment responses, we used the breast epithelial MCF10A cell line from the LINCS study that profiled 295 signaling proteins and 58 histone marks profiled 4, 8, 24, and 48 hours after treatment with six growth-promoting ligands, namely, epidermal growth factor (EGF), hepatocyte growth factor (HGF), Oncostatin M (OSM), bone morphogenetic protein2 (BMP2), transforming growth factor beta (TGFB), Interferon Gamma-1b (IFNG). In the following sections, we describe the method, results and discussion, and future work.

2 Materials and Methods

This section proposes a non-negative tensor factorization based approach to model the high-order correlations among proteins, ligands and histones based on the collected multi-omics data containing two time-serial proteomic datasets.

2.1 Workflow

Fig.1 shows the overall workflow of of our method which consists of three subsequent components as follows:

1. **Multi-Omics Data Pre-Processing.** The time-serial multi-omics datasets are firstly collected, followed by a statistical significance test to retain all significant samples with p values less than a threshold. Then, we calculate the fold change abundance for all data samples w.r.t their control measurements. where the resulting fold changes are finally organized in two three-way tensors with each representing a different time-serial proteomic dataset.
2. **Multi-layer Molecular Correlation Modeling.** In this part, a computational algorithm is adopted to explore the collected multi-omics data. Specifically, a Non-negative Tensor Factorization (NTF) based method (Kolda and Bader, 2009) is used to model the high-order molecular correlations encoded in the three-way tensors, where the number (k) of latent factors for NTF is determined based on the widely-used core consistency metric (Bro and Kiers, 2003).
3. **Latent Factors and Correlation Patterns Analysis.** This part focuses on the interpretation of latent factors derived from the multi-omics data through NTF. First, to examine the similarities among proteins or among histones from their respective tensor representations, we perform the clustering w.r.t the protein-ligand component and histone-ligand component (e.g., intra-component), respectively. Then, we perform an inter-component correlation analysis, *i.e.*, three correlation pattern matrices $\mathbf{M}_1^r$, $\mathbf{M}_2^r$ and $\mathbf{M}_3^r$ ($\mathbf{M}_1^g$, $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$ for GCP) can be computed for RPPA, to gain insights into the interactions among proteins (or histones) and ligands. Finally, we combine the correlation patterns from RPPA and GCP to uncover the high-order dynamic interactions among proteins, ligands and histones over time.

The description of above workflow is organized as follows. The first and second components regarding multi-omics data pre-processing and multi-layer molecule correlation modeling are described in Section 2.2 and Section 2.3, respectively. The third component regarding latent factors and correlation pattern analysis is presented in Section 3. It is necessary to note that we perform the pre-processing and NTF on RPPA and GCP datasets respectively, and then combine their resulting latent factors of NTF to collectively understand the extracellular matrix of breast tissue.

2.2 Multi-Omics Data Pre-Processing

Breast cancers are incredibly complicated at the molecular level, which necessitates multi-omics data at different levels to gain a complete insight into the interaction mechanism among molecules such as proteins, histones and ligands. To understand the dynamic regulations of growth promoting factors in the breast cancer microenvironment, we specifically collected two datasets named RPPA and GCP respectively.

2.2.1 Data Acquisition

The level3 (log 2 normalized) Reverse Phase Protein Array (RPPA) data were downloaded from the Synapse platform (https://www.synapse.org/#!Synapse:syn12555331) in June 2019. The dataset consists of highly sensitive and selective 295 antibody-based protein and phosphoprotein measurements of six growth ligands, namely, EGF, HGF, OSM, BMP2, TGFB, IFNG. Each data sample contains three replicates for each of these ligands and a control (PBS) measured at five time points (1, 4, 8, 24, and 48 hours).

The level 3 (log2 normalized) Global Chromatin Profiling assay (GCP) against 65 probes that monitor combinations of post-translational modification on histones data using in MCF10A healthy breast epithelial cell line were downloaded from (https://www.synapse.org/#!Synapse:syn18491838) in August 2020. These assays were treated with six growth ligands, namely, EGF, HGF, OSM, BMP2, TGFB, IFNG at various concentrations. The data contains three replicates for each of these ligands and a control (PBS). These samples were measured at four time points (4, 8, 24, and 48 hours).

To perform consistent analysis on above two datasets, we removed all samples in the 1st hour from the RPPA data. As a result, both RPPA and GCP data involve 4 distinct time points of 4, 8, 24 and 48 hours.

2.2.2 Statistical Significance Test

For both RPPA and GCP datasets, each data sample (e.g., protein or histone) has three replicates representing three independent experiments at each time point. To validate the quality of these sample values, *i.e.*, noisy samples may exist due to human errors or other reasons, we applied

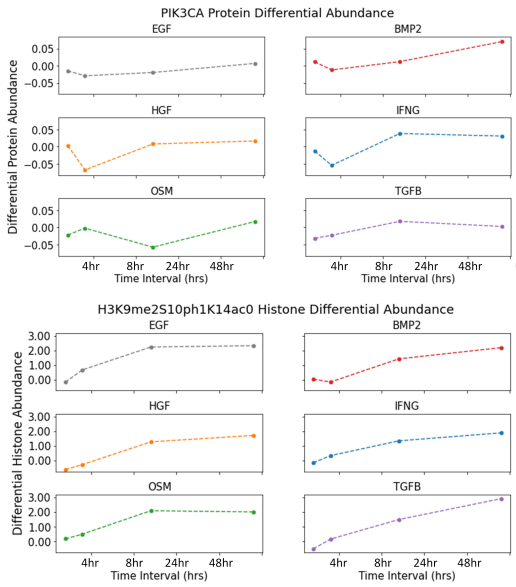


Fig. 2: The differential abundances of protein PIK3CA and histone H3K9me2S10ph1K14ac0 after treated with 6 different ligands, respectively.

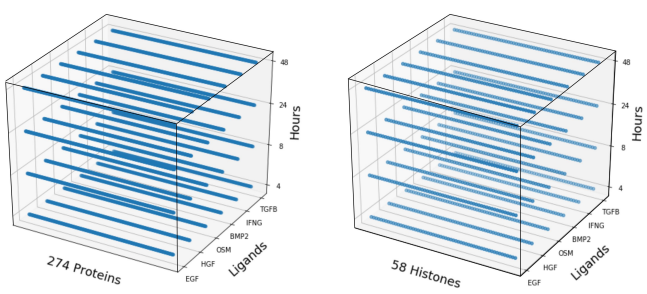


Fig. 3: The 3-way tensor representations of RPPA (left) and GCP (right) datasets, where each tensor has three dimensions or modes, e.g., protein, ligand and hour for RPPA, and histone, ligand and hour for GCP.

an unpaired Student's t-test (Ruxton, 2006) between each sample and the respective controls measurements. We kept all significant samples with $p < 0.05$, where 274 proteins in RPPA and 58 histones in GCP were finally retained in the experiment.

2.2.3 Tensor Representation of Data

After the t-test and sample screening processes, the three replicates for each sample were averaged. Then, the differential abundance (a.k.a fold change) of proteins and histones w.r.t their respective controls were calculated, *i.e.*, the quantity change between each sample measurement and its control measurement. Fig.2 shows the differential abundance of an example protein from RPPA data and an example histone from the GCP data, *i.e.*, after treated with ligand IFNG, histone H3K9me2S10ph1K14ac0 was first down-regulated through 4th to 8th hours and then up-regulated till to the 48th hour, while protein PIK3CA showed consistent up-regulation through 4th to 48th hours.

The RPPA fold-change data at each time point i can be represented in a matrix $\mathbf{L}_i^r \in \mathbb{R}^{274\times 1}$. By stacking all matrices observed at four subsequent time points (e.g., 4, 8, 24 and 48 hours), we are able to represent the RPPA data with a 3D tensor (a.k.a 3-way array) denoted by $\mathbf{T}^r = \{\mathbf{L}_4^r, \mathbf{L}_8^r, \mathbf{L}_{24}^r, \mathbf{L}_{48}^r\}$. Similarly, The GCP fold-change data can be represented as a 3D tensor $\mathbf{T}^g = \{\mathbf{L}_4^g, \mathbf{L}_8^g, \mathbf{L}_{24}^g, \mathbf{L}_{48}^g\}$, where $\mathbf{L}_j^g \in \mathbb{R}^{58\times 1}$ denotes the fold-change matrix at time point j.

We integrated $\mathbf{T}^r \in \mathbb{R}^{274\times 4}$ and $\mathbf{T}^g \in \mathbb{R}^{58\times 4}$ to form the multi-omics data $D = \{\mathbf{T}^r, \mathbf{T}^g\}$ studied in this paper, which are visualized in two 3-way tensors shown in Fig.3. In the following, we proposed a NTF-based method to respectively decompose the two 3-way tensors and integrated their resulting latent factors to collectively reveal the regulations of growth promoting factors in the breast cancer microenvironment.

To maintain the significance of both negative and positive fold-changes (e.g., both can indicate that proteins or histones have been impacted by ligands) in the non-negative factorization, we first applied an absolute transformation over all the fold-change values. For example, we constructed two binary indicator matrices $I^r = \{-1, 1\}^{274\times 4}$ and $I^g = \{-1, 1\}^{58\times 4}$ to respectively indicate the directions (e.g., either positive or negative) of protein and histone fold-changes. Then, the absolute transformations for RPPA and GCP data were performed through $\mathbf{T}^r = \mathbf{T}^r * I^r$ and $\mathbf{T}^g = \mathbf{T}^g * I^g$, where $*$ represents the multiplication operation.

2.3 Multi-Layer Molecular Correlation Modeling

From the tensor representations shown in Fig.3, we can observe multi-layer correlation/interaction relationships between proteins and ligands (or histones and ligands) along the time axis, where the variation of molecule correlations at early times (e.g., 4 hour) could incur an influence to molecule correlations at later times (e.g., 24 and 48 hours) (Li and Ngom, 2010). In the following, we present a method to model the multi-layer molecular correlations using the non-negative tensor factorization technique.

2.3.1 Non-Negative Tensor Factorization (NTF)

Tensor factorization can be viewed as the high-order version of matrix factorization (Lee and Seung, 2001) by extending the two-dimensional matrices to tensors. A tensor is a multi-way array or multi-dimensional matrix (Cichocki *et al.*, 2009) formally defined as:

Definition 2.1 (Tensor) *Let $I_1, I_2, \cdots, I_N$ denote index upper bound. A tensor $\mathbf{T} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ of order N is a N-way array where elements $\mathbf{T}_{i_1,i_2,\cdots,i_n}$ are indexed by $i_n \in \{1, 2, \cdots, I_n\}$ for $1 \le n \le N$.*

Tensors are generalizations of vectors and matrices, where the rank of a tensor is the minimum number of rank-one tensors defined as:

Definition 2.2 (Rank-One Tensor) *A tensor $\mathbf{T} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ of order N has rank-one if it can be written as the outer product (represented by notation $\circ$) of N vectors, i.e., $\mathbf{T} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}$ or $\mathbf{T}_{i_1,i_2,\cdots,i_N} = \mathbf{a}_{i_1}^{(1)}\mathbf{a}_{i_2}^{(2)}\cdots\mathbf{a}_{i_N}^{(N)}$ for all $1 \le i_n \le I_n$.*

Tensor representations are powerful to encode the complex relations across different modes such as histone, ligand and time factors for RPPA data, while tensor factorization provides a way to decompose a given tensor into k components or latent factors in different domains with common latent factors reflecting the correlations across these different domains. A model which imposes non-negativity on all resulting components is called NTF. The two widely used tensor factorization techniques are PARAFAC (Carroll and Chang, 1970) and Tucker (Tucker, 1966), which have shown their promises in many real-world applications such as cancer survival prediction (Fang, 2019) and image inpainting (Zhou *et al.*, 2017).

2.3.1.1 NTF Techniques

The non-negative PARAFAC was proposed by Carroll *et al.* (Carroll *et al.*, 1989) and then developed by Bro (Bro *et al.*, 1997) with a more efficient method to calculate the non-negative components. Fig.4 (a) shows the graphical representation of PARAFAC, which aims to decompose a given

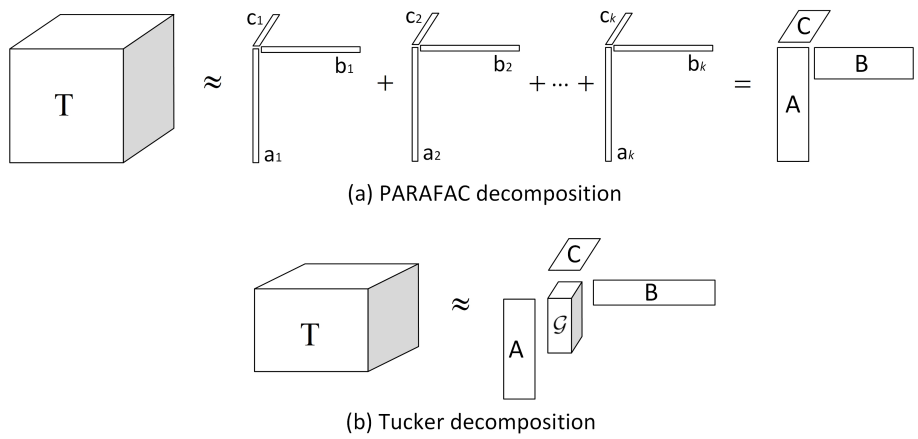


Fig. 4: The graphical representations of PARAFAC and Tucker for a 3-way tensor decomposition, where PARAFAC is a special case of Tucker with super-diagonal core tensor $\mathcal{G}$.

3-way tensor $\mathbf{T} \in \mathbb{R}^{I_1,I_2,I_3}$ to a sum of component rank-one tensors as:

$$\mathbf{T} \approx \sum_{r=1}^{k} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{1}$$

where $\mathbf{a}_r \in \mathbb{R}^{I_1}$, $\mathbf{b}_r \in \mathbb{R}^{I_2}$ and $\mathbf{c} \in \mathbb{R}^{I_3}$. k represents the number of latent components or factors, which is often assumed to be around the actual rank of tensor $\mathbf{T}$ thereby data relationships in the original tensor could be fully preserved by the resulting components. Alternatively, the PARAFAC decomposition can be represented as:

$$\mathbf{T} = [[\mathbf{A}, \mathbf{B}, \mathbf{C}]], \tag{2}$$

where the loading matrices **A**,**B**,**C** can be understood as the linear mapping from the respective modes to the latent spaces defined as:

$$\begin{aligned} \mathbf{A} &= [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k] \in \mathbb{R}^{I_1 \times k} \\ \mathbf{B} &= [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_k] \in \mathbb{R}^{I_2 \times k} \\ \mathbf{C} &= [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k] \in \mathbb{R}^{I_3 \times k}. \end{aligned} \tag{3}$$

Tucker decomposition is another popular NTF technique (Tucker, 1966). As shown in Fig.4(b), compared with PARAFAC, in Tucker an additional core tensor $\mathcal{G} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ is calculated together with the three loading matrices **A**, **B** and **C** represented as:

$$\begin{aligned} \mathbf{T} &\approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} \sum_{r_3=1}^{k_3} g_{r_1 r_2 r_3} \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \end{aligned} \tag{4}$$

where $k_1 \times k_2 \times k_3$ rank-one tensors are produced compared with PARAFAC which only has k rank-one tensors. k_1, k_2 and k_3 are numbers of components for each mode of the 3-way tensor.

It is obvious to conclude that PARAFAC is a special case of Tucker with super-diagonal core tensor $\mathcal{G}$. Studies (Kim and Choi, 2007; Karami *et al.*, 2012) show that PARAFAC is suitable for problems where the latent factor interpretation is essential, whereas Tucker is often used for data compression purpose since the core tensor can be viewed as a compression of the original tensor. Another difference is the rank-one decomposition tensors of PARAFAC are often unique, whereas Tucker decomposition is not unique (Kolda and Bader, 2009). For these reasons, we choose to use non-negative PARAFAC to analyse and understand the cancer microenvironment in this work. In the following, two different implementations of 3-way PARAFAC based on Tensorly (Kossaifi *et al.*, 2019) and NCP (Kim *et al.*, 2014) are respectively introduced and used for RPPA and GCP tensor factorization in this paper.

2.3.1.2 Non-Negative PARAFAC Based on Tensorly

Tensorly is a python library to support tensor operations such as tensor decomposition, tensor learning and tensor algebra. Tensorly adopts the widely used Alternating Least Square (ALS) (Harshman *et al.*, 1970; Carroll and Chang, 1970) method to solve the non-negative PARAFAC decomposition, where the objective is to learn k components that best approximate the input 3-way tensor $\mathbf{T}$:

$$\begin{aligned} &\min_{\hat{\mathbf{T}}} \parallel \mathbf{T} - \hat{\mathbf{T}} \parallel_F^2, \\ &\text{with } \hat{\mathbf{T}} = \sum_{r=1}^{k} \lambda_r \circ \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [[\boldsymbol{\lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]], \\ &\text{subject to: } \mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r \geq 0, \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_k]$ are the scaling factors which are set to force a normalization (e.g., $l-2$ norm) on the resulting latent factors. For the ALS method, the loading matrices **A**, **B** and **C** are alternatively optimized with every time optimizing one matrix while holding the other two matrices. For example, if we optimize **A** with fixed **B** and **C**, the objective in Eq.5 can be represented as:

$$\min_{\hat{\mathbf{A}}} \parallel \mathbf{T}_{(1)} - \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} \parallel_F^2, \tag{6}$$

where $\mathbf{T}_{(1)}$ represents the mode-1 of $\mathbf{T}$ and notation $\odot$ denotes the Khatri-Rao product. $\hat{\mathbf{A}}$ is the temporal version of **A** without normalization. Then, the optimal solution of $\hat{\mathbf{A}}$ can be calculated by (Kolda and Bader, 2009):

$$\hat{\mathbf{A}} = \mathbf{T}_{(1)}[(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}}]^{\dagger}, \tag{7}$$

where notation † represents the Moore–Penrose pseudoinverse operation (Barata and Hussein, 2012) which typically involves a high computation complexity. To reduce the complexity, Eq.7 can be simplified as (Van Loan, 2000):

$$\hat{\mathbf{A}} = \mathbf{T}_{(1)}(\mathbf{C} \odot \mathbf{B})(\mathbf{C}^{\mathrm{T}}\mathbf{C} * \mathbf{B}^{\mathrm{T}}\mathbf{B})^{\dagger}. \tag{8}$$

Finally, a normalization process is applied on each column $\hat{\mathbf{a}}_r$ in $\hat{\mathbf{A}}$ to obtain the respective column $\mathbf{a}_r$ in the loading matrix **A**, *i.e.*, $\mathbf{a}_r = \hat{\mathbf{a}}_r / \lambda_r$,

where $\lambda_r = \| \hat{\mathbf{a}}_r \|$. The above procedure also applies to optimize the loading matrices **B** and **C**. The iterations for the alternating optimization of **A**, **B** and **C** repeat until the stopping criterion, *i.e.*, the maximum number of iterations and minor change between two iterations, are satisfied.

2.3.1.3 Non-Negative PARAFAC Based on NCP

NCP is another widely used tool to solve the non-negative PARAFAC decomposition, which is implemented based on the Block Principal Pivoting (BPP) method (Kim and Park, 2012). For a given 3-way tensor $\mathbf{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, NCP aims to minimize a similar decomposition objective as Tensorly by:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \| \mathbf{T} - [[\mathbf{A}, \mathbf{B}, \mathbf{C}]] \|_F^2 \quad \text{subject to: } \mathbf{A}, \mathbf{B}, \mathbf{C} \geq 0 \tag{9}$$

To avoid confusion, we use $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ and $\mathbf{A}^{(3)}$ to respectively denote **A**, **B** and **C** in the paragraph. For the GCP method, the tensor approximation $\mathbf{T} \approx [[\mathbf{A}, \mathbf{B}, \mathbf{C}]]$ can be reformulated as:

$$\begin{aligned} &\mathbf{T}_{(n)} \approx \mathbf{A}^{(n)} \times \left(\mathbf{D}^{(n)}\right)^{\mathrm{T}}, \text{for } \forall n \in \{1, \cdots, N\} \\ &\text{with } \mathbf{D}^{(n)} = \mathbf{A}^{(N)} \circ \cdots \circ \mathbf{A}^{(n+1)} \circ \mathbf{A}^{(n-1)} \circ \cdots \circ \mathbf{A}^{(1)} \end{aligned} \tag{10}$$

where $N = 3$ in this paper and $\mathbf{T}_{(n)}$ represents the mode-n matricization of tensor $\mathbf{T}$. Then, the optimization in Eq.9 can be viewed as solving the following subproblems:

$$\min_{\mathbf{A}^{(n)}} \| \mathbf{D}^{(n)} \times \left(\mathbf{A}^{(n)}\right)^{\mathrm{T}} - \left(\mathbf{T}_{(n)}\right)^{\mathrm{T}} \|_F^2, \text{for } \forall n \in \{1, \cdots, N\} \quad \text{subject to: } \mathbf{A}^{(n)} \geq 0 \tag{11}$$

It is easy to observe from Eq.11 that NCP actually reduces the NTF to N independent non-negative matrix factorization problems, where each problem can be simplified as:

$$\min_{\mathbf{X} \geq 0} \| \mathbf{V}\mathbf{X} - \mathbf{W} \|_F^2, \tag{12}$$

where $\mathbf{V} \in \mathbb{R}^{P \times Q}$, $\mathbf{X} \in \mathbb{R}^{Q \times L}$ and $\mathbf{W} \in \mathbb{R}^{P \times L}$. Then, the BBP algorithm with multiple right-hand sides is utilized to solve these subproblems. The BBP algorithm (Kim and Park, 2012) to solve each subproblem is summarized in Algorithm 1, where $\mathbf{x}_{\mathcal{F}_l}$ and $\mathbf{y}_{\mathcal{G}_l}$ represents the subsets of $l-$th column of **X** and **Y** indexed by $\mathcal{F}_l$ and $\mathcal{G}_l$, respectively.

2.3.2 Selection of the Number of Latent Factors *k*

Before applying the non-negative PARAFAC on the multi-omics data, the number of latent components or factors k needs to be determined, which should be around the actual rank of the target tensor. Determining the number of latent factors is critical. If k is smaller than the actual rank, the resulting latent factors cannot fully preserve the data relationships in the original tensor (a.k.a the under-fitting problem). On the contrary, if k is larger than the actual rank of the target tensor, the latent factors are able to span the full information revealed in the tensor, but it would meanwhile learn redundant information thereby causing the over-fitting problem.

It is non-trivial to determine the best number of latent factors k, which normally necessitates multiple trails and fine-tuning back and forth (Gauvin *et al.*, 2014). In the past, the *core consistency* metric has been widely used to evaluate the fitness of a candidate k (Bro and Kiers, 2003), where the main idea is that the components of an appropriate PARAFAC model are desired to reflect trilinear variation in the data. As demonstrated in section 2.3.1.1, the PARAFAC decomposition is actually a special case of the Tucker decomposition (in Eq.4) since Eq.1 can be rewritten as:

$$\mathbf{T} \approx \sum_{r_1=1}^{k} \sum_{r_2=1}^{k} \sum_{r_3=1}^{k} g^d_{r_1,r_2,r_3} \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \tag{13}$$

Algorithm 1 The block principle pivoting (BPP) algorithm

Input: $\mathbf{V} \in \mathbb{R}^{P \times Q}$, $\mathbf{W} \in \mathbb{R}^{P \times L}$
Output: $\mathbf{X}(\in \mathbb{R}^{Q \times L}) = \arg\min_{\mathbf{X} \geq 0} \| \mathbf{V}\mathbf{X} - \mathbf{W} \|_F^2$

Compute $\mathbf{V}^{\mathrm{T}}\mathbf{V}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{W}$;
Initilize $\mathcal{F}_l = \emptyset$, $\mathcal{G}_l = \{1, \cdots, q\}$ for all $l \in \{1, \cdots, L\}$, $\mathbf{X}$, $\mathbf{Y} = -\mathbf{V}^{\mathrm{T}}\mathbf{W}$, $\alpha = 3$ and $\beta = q + 1$;
Compute all $x_{\mathcal{F}_l}$ from $\mathbf{V}^{\mathrm{T}}_{\mathcal{F}_l}\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l}$ using column grouping;
Compute all $\mathbf{y}_{\mathcal{G}_l} = \mathbf{V}^{\mathrm{T}}_{\mathcal{G}_l}(\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l} - \mathbf{w}_l)$
while any $(\mathbf{x}_{\mathcal{F}_l}, \mathbf{y}_{\mathcal{G}_l})$ is infeasible **do** /** $\mathbf{x}_{\mathcal{F}_l} < 0$ or $\mathbf{y}_{\mathcal{G}_l} < 0$ **/
  Find the indices of columns in which the solution is infeasible;
  $I = \{j : (\mathbf{x}_{\mathcal{F}_l}, \mathbf{y}_{\mathcal{G}_l})$ is infeasible$\}$;
  for $l \in I$ **do**
    $\mathcal{H}_l = \{q \in \mathcal{F}_l : \mathbf{x}_q < 0\} \cup \{q \in \mathcal{G}_l : \mathbf{y}_q < 0\}$
  end for
  for $l \in I$ with $|\mathcal{H}_l| < \beta_l$ and $\alpha_l \geq 1$ **do**
    $\beta_l = |\mathcal{H}_l|, \alpha_l = 3, \hat{\mathcal{H}}_l = \mathcal{H}_l$;
  end for
  for $l \in I$ with $|\mathcal{H}_l| < \beta_l$ and $\alpha_l = 0$ **do**
    $\hat{\mathcal{H}}_l = \{q : q = \max\{q \in \mathcal{H}_l\}\}$;
  end for
  for $l \in I$ **do**
    $\mathcal{F}_l = (\mathcal{F}_l - \hat{\mathcal{H}}_l) \cup (\hat{\mathcal{H}}_l \cap \mathcal{G}_l), \mathcal{G}_l = (\mathcal{G}_l - \hat{\mathcal{H}}_l) \cup (\hat{\mathcal{H}}_l \cap \mathcal{F}_l)$;
    Compute $\mathbf{x}_l$ from $\mathbf{V}^{\mathrm{T}}_{\mathcal{F}_l}\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l}$ using column grouping;
  end for
end while

Algorithm 2 RPPA tensor decomposition based on Tensorly

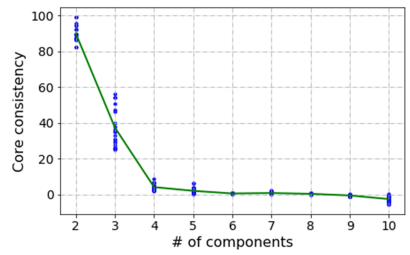
Input: RPPA tensor $\mathbf{T}^r$, the number of latent factors k.
Output: The loading matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$.

Initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ with SVD;
while (conditions not satisfied) **do**
  $\hat{\mathbf{A}} = \mathbf{T}^r_{(1)}(\mathbf{C}^r \odot \mathbf{B}^r)(\mathbf{C}^{r\mathrm{T}}\mathbf{C}^r * \mathbf{B}^{r\mathrm{T}}\mathbf{B}^r)^{\dagger}$;
  $\hat{\mathbf{B}} = \mathbf{T}^r_{(1)}(\mathbf{C}^r \odot \mathbf{A}^r)(\mathbf{C}^{r\mathrm{T}}\mathbf{C}^r * \mathbf{A}^{r\mathrm{T}}\mathbf{A}^r)^{\dagger}$;
  $\hat{\mathbf{C}} = \mathbf{T}^r_{(1)}(\mathbf{B} \odot \mathbf{A}^r)(\mathbf{B}^{r\mathrm{T}}\mathbf{B}^r * \mathbf{A}^{r\mathrm{T}}\mathbf{A}^r)^{\dagger}$;
  if (normalization) **then**
    Normalize columns of $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ to obtain $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$.
  end if
end while
Return $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$.

Algorithm 3 RPPA tensor decomposition based on NCP

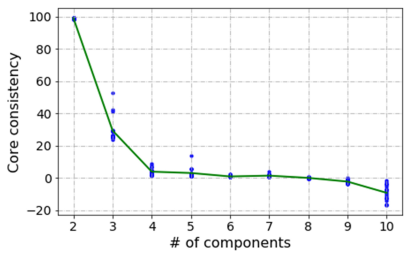
Input: RPPA tensor $\mathbf{T}^r$, the number of latent factors k.
Output: The loading matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$.

Initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ with NNDSVD;
$\mathbf{A}^r$ = *BPP*($\mathbf{W} = \mathbf{B}^r \odot \mathbf{C}^r, \mathbf{X} = \mathbf{A}^{r\mathrm{T}}, \mathbf{W} = (\mathbf{T}^r_{(1)})^{\mathrm{T}}$);
$\mathbf{B}^r$ = *BPP*($\mathbf{W} = \mathbf{A}^r \odot \mathbf{C}^r, \mathbf{X} = \mathbf{B}^{r\mathrm{T}}, \mathbf{W} = (\mathbf{T}^r_{(2)})^{\mathrm{T}}$);
$\mathbf{C}^r$ = *BPP*($\mathbf{W} = \mathbf{A}^r \odot \mathbf{B}^r, \mathbf{X} = \mathbf{C}^{r\mathrm{T}}, \mathbf{W} = (\mathbf{T}^r_{(3)})^{\mathrm{T}}$);
Return $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$.



k	Avg. CC
2	89.9
3	37.3
4	4.13
5	2.05
6	0.58
7	0.82
8	0.38
9	-0.53
10	-2.55

(a) Core consistency (CC) *vs.* k on RPPA data



k	Avg. CC
2	99.0
3	29.4
4	3.91
5	3.09
6	0.95
7	1.45
8	0.05
9	-2.22
10	-9.18

(b) Core consistency (CC) *vs.* k on GCP data

Fig. 5: The core-consistency *vs.* the number of components (k) tested on the RPPA tensor (top (a)) and GCP tensor (bottom (b)), respectively. The core consistency for each k was computed 100 times independently. The top 20% core consistencies (five blue dots) together with their averages (solid line) are shown in the left panel. The actual values for average core consistencies (CC) are shown in the right panel

where $\mathcal{G}^d = \{g^d_{r_1,r_2,r_3}\}_{r_1,r_2,r_3=1,\cdots,k} \in \mathbb{R}^{k\times k\times k}$ is a super-diagonal tensor with all diagonal elements (e.g., $r_1 = r_2 = r_3$) equal to 1 and other off-diagonal elements equal to 0. The Tucker core tensor $\mathcal{G}$ can be considered as the optimal representation of the target tensor $\mathbf{T}$ *w.r.t.* the subspace defined by the latent factors (Bro and Kiers, 2003). In this regard, if a specific k is optimal for PARAFAC to fully span trilinear variations in the data, the Tucker core tensor $\mathcal{G}$, computed with fixed loading matrices output from PARAFAC upon k, should be close to the super-diagonal. Then, the *core consistency* (CC) is defined as the difference between $\mathcal{G}$ and $\mathcal{G}_d$ to quantify the goodness of a particular k:

$$\text{CC} = 100\left(1 - \frac{\sum_{r_1=1}^{k}\sum_{r_2=1}^{k}\sum_{r_3=1}^{k}(g_{r_1,r_2,r_3} - g^d_{r_1,r_2,r_3})^2}{k}\right). \quad (14)$$

The CC reflects how well a particular Tucker core fits the assumption of a good PARAFAC model upon k. A CC above 50 can be interpreted as an acceptable model (Bro and Kiers, 2003; Gauvin *et al.*, 2014).

We leverage core consistency as the metric to determine the suitable number of latent factors k for RPPA tensor and GCP tensor, respectively. Fig.5 shows that the core consistencies change with k ranging from 2 to 10. We can observe for both RPPA and GCP tensors, the average core consistencies are around 90 at $k = 2$. Therefore, in the experiment, we consider $k = 2$ as the suitable number of latent components for PARAFAC decomposition on both RPPA and GCP data.

2.3.3 NTF on Multi-Omics Data

After determining the suitable number of components by core consistency, we applied the non-negative PARAFAC model implemented in Tensorly and NCP to decompose the RPPA tensor $\mathbf{T}^r$ and GCP tensor $\mathbf{T}^g$, respectively.

2.3.3.1 RPPA Tensor Decomposition

The non-negative PARAFAC decomposition based on the Tensory implementation is summarized in Algorithm 2, which takes the 3-way RPPA tensor $\mathbf{T}^r$ as input and then produces three loading matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$. To achieve robust results, we adopted a Singular Value Decomposition (SVD) algorithm embedded in Tensorly to initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$. The stopping conditions used in the experiment are the maximum number of iterations and the difference between two subsequent iterations, where the training will stop when either of these two criterion are satisfied.

The non-negative factorization based on NCP implementation is summarized in Algorithm 3. We adopted the Nonnegative Double Singular Value Decomposition (NNDSVD) (Boutsidis and Gallopoulos, 2008) to initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$, which is based on two SVD processes, with one approximating the data matrix and the other approximating positive sections of the resulting partial SVD factors utilizing an algebraic property of unit rank matrices. *BPP* represents the block principle pivoting presented in Algorithm 1.

2.3.3.2 GCP Tensor Decomposition

Similarly, we are able to decompose the GCP tensor $\mathbf{T}_g$ and derive the loading matrices $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ based on PARAFAC and NCP respectively in the same ways described in section 2.3.3.1.

3 Latent Factors and Correlation Patterns Analysis

This section focuses on the analysis of the latent factors ($\mathbf{A}^r$, $\mathbf{B}^r$, $\mathbf{C}^r$, $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ shown in Fig.6). To explore the interactions among proteins, ligands and histones in the RPPA and GCP data, we proposed to construct correlation pattern matrices $\mathbf{M}^r_1$, $\mathbf{M}^r_2$, $\mathbf{M}^r_3$, $\mathbf{M}^g_1$, $\mathbf{M}^g_2$ and $\mathbf{M}^g_3$, respectively. We also incorporated these correlation pattern matrices to collectively understand the correlation patterns over time in the breast cancer microenvironment based on a partial least squares regression model.

3.1 Latent Factors Analysis

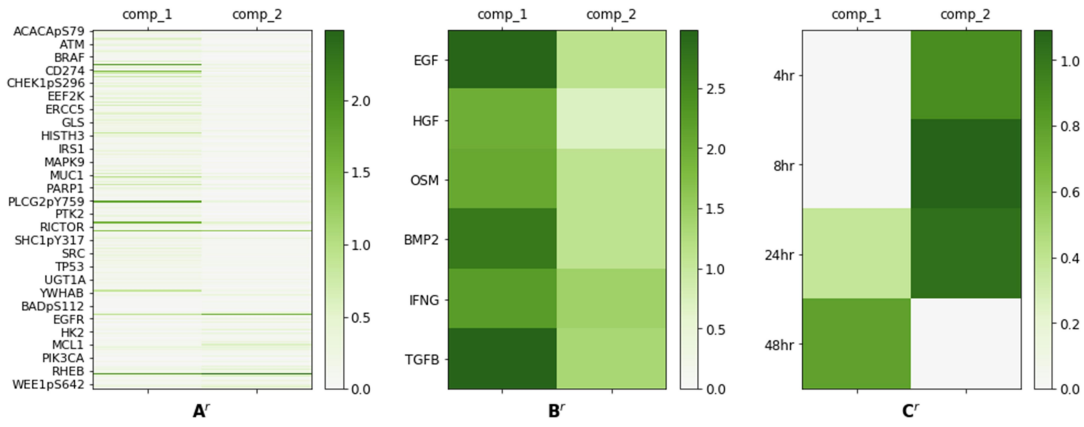
Fig.6 (a) demonstrates the heatmaps of $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ matrices generated from the RPPA tensor factorization. $\mathbf{A}^r$ and $\mathbf{B}^r$ represent that proteins and ligands are mapped in a space with two latent factors or components, whereas $\mathbf{C}^r$ interprets the temporal significance of proteins and ligands in the two respective components.

Fig.6 (b) demonstrates the the heatmaps of $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ matrices from GCP tensor decomposition based on NCP. Analogically, $\mathbf{A}^g$ and $\mathbf{B}^g$ denote that the interactions of histones and ligands are mapped in a two-dimensional space, where the temporal interaction information is reflected by $\mathbf{C}^g$.

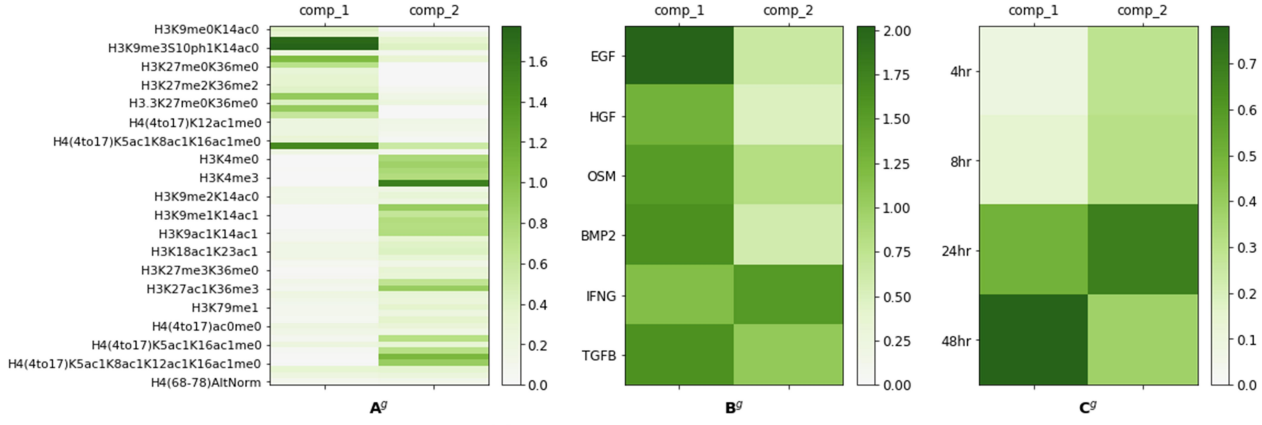
3.2 Generation of Protein and Histone Signatures

We further performed the classification or clustering analysis over all proteins and histones based on the latent factor matrices $\mathbf{A}^r$ and $\mathbf{A}^g$, respectively.

Each protein is mapped into a space with two latent factors (e.g., comp_1 and comp_2 shown in $\mathbf{A}^r$) after NTF, which enables us to classify all proteins in two clusters, where each protein is assigned to a respective cluster with the largest membership value across the two latent factors. The left panel of Fig.7 (a) shows the two protein clusters (also called protein signatures in this paper) identified, where each cluster corresponds to a latent factor in $\mathbf{A}^r$. In the tumor microenvironment, the protein signatures can be interpreted as the categorization of proteins demonstrating similar regulation behaviors after treated with growth ligands. In addition to above hard clustering, we also applied a Softmax transformation over $\mathbf{A}^r$, which reveals the probability distributions of different proteins belonging to the two clusters shown in the right panel of Fig.7 (a).

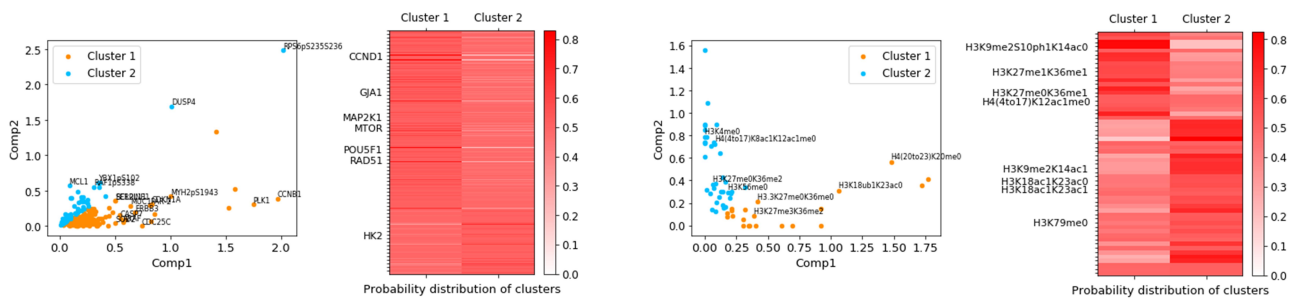


(a) Latent factor matrices from RPPA tensor based on NCP



(b) Latent factor matrices from GCP tensor based on NCP

Fig. 6: Latent factors generated from RPPA and GCP tensor decomposition.



(a) Protein signatures (b) Histone signatures

Fig. 7: The identified protein and histone signatures.

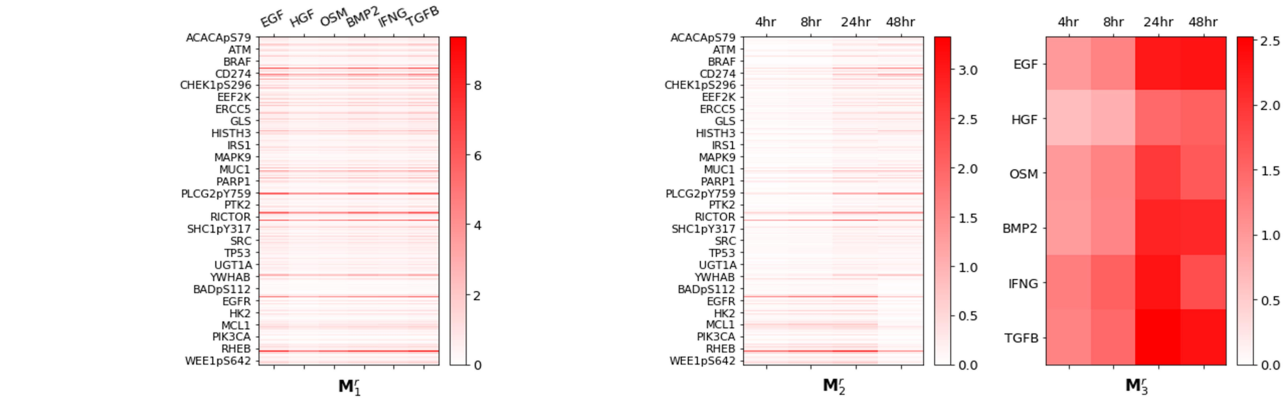
Similarly, the left panel of Fig.7 (b) shows the two identified histone signatures generated from loading matrix $\mathbf{A}^g$. The probability distributions of histones in the two signatures are shown in the right panel of Fig.7 (b). Compared with protein signatures, we can observe that histone signatures are more discriminative (e.g., the two histone signatures have a clearer boundary) after treated with the same growth ligands.

3.3 Correlation Patterns Between Proteins and Ligands

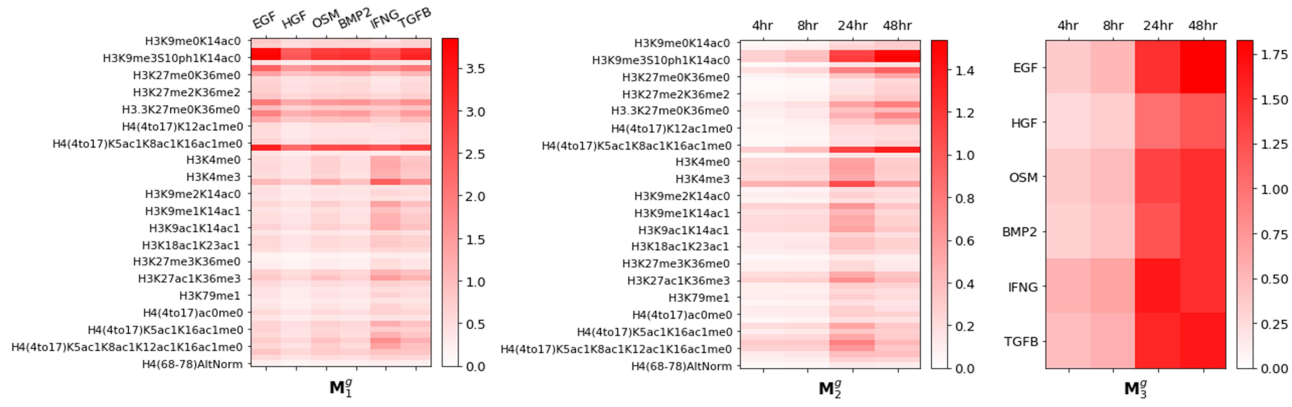
To gain insights into the interactions between proteins and ligands over time, we proposed to construct three auxiliary matrices (shown in Fig.8 (a)) through the combinations of $\mathbf{A}_1^r$, $\mathbf{B}_2^r$ and $\mathbf{C}_3^r$ as:

$$\begin{aligned} \mathbf{M}_1^r &= \mathbf{A}^r \otimes (\mathbf{B}^r)^{\mathrm{T}}, \\ \mathbf{M}_2^r &= \mathbf{A}^r \otimes (\mathbf{C}^r)^{\mathrm{T}}, \\ \mathbf{M}_3^r &= \mathbf{B}^r \otimes (\mathbf{C}^r)^{\mathrm{T}}, \end{aligned} \tag{15}$$

where $\otimes$ denotes the outer product. $\mathbf{M}_1^r$ reveals the correlation intensity between proteins and ligands, *i.e.*, if a high intensity value (e.g., the value is larger than a selected threshold) is observed between a protein and a ligand in the matrix, we can understand that the protein has been significantly



(a) Correlation pattern matrices from RPPA tensor based on NCP



(b) Correlation pattern matrices from GCP tensor based on NCP

Fig. 8: (a) reveals the correlation patterns between proteins and ligands. (b) reveals the correlation patterns between histones and ligands.

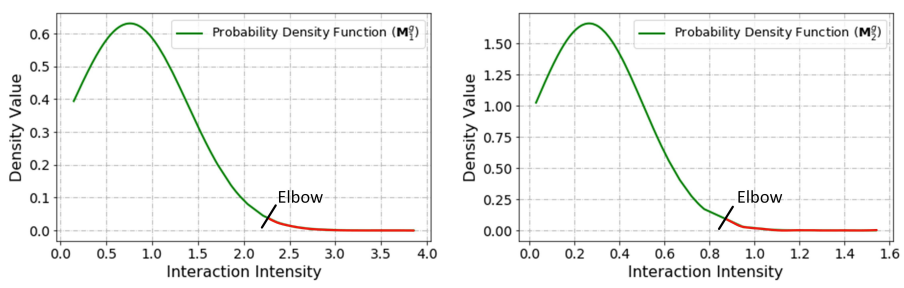


Fig. 9: The distributions of intensity values for correlation pattern matrices $\mathbf{M}_1^g$ and $\mathbf{M}_2^g$.

regulated by the corresponding ligand. $\mathbf{M}_2^r$ and $\mathbf{M}_3^r$ reveal the intensities of dynamic regulations through 4th hour to 48th hour, which actually explain the onsets of proteins being regulated by ligands in $\mathbf{M}_1^r$ as well as the trend of these regulations changes over time.

3.4 Correlation Patterns between Histones and Ligands

Similarly, we created three matrices (shown in Fig.8(b)) as follows to reveal the correlations between histones and ligands over time:

$$\begin{aligned} \mathbf{M}_1^g &= \mathbf{A}^g \otimes (\mathbf{B}^g)^{\mathrm{T}}, \\ \mathbf{M}_2^g &= \mathbf{A}^g \otimes (\mathbf{C}^g)^{\mathrm{T}}, \\ \mathbf{M}_3^g &= \mathbf{B}^g \otimes (\mathbf{C}^g)^{\mathrm{T}}, \end{aligned} \tag{16}$$

where $\mathbf{M}_1^g$ uncovers the intensities of correlation between histones and ligands, with their temporal trends reflected in $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$. It is interesting to observe from $\mathbf{M}_3^g$ that all 6 ligands are active in the 24th hour and maintains in the 48th hour, meaning that ligands have shown impacts on histones through 24th to 48th hours.

3.5 Correlation Patterns between Proteins, Ligands and Histones

To gain insights into the mechanism of how ligands regulate proteins and further modify histones, we first obtained all the significant histones based on their interactions with ligands shown in $\mathbf{M}_1^g$. We performed two distribution learning (e.g., fitting two Gaussian distributions) over all intensity values in $\mathbf{M}_1^g$ and $\mathbf{M}_2^r$ to determine the significant histones

Histone\Protein	ANXA7	BID	COL6A1	EGFRpY1173	EIF4EBP1	ESR1pS118	HIF1A	MCL1	PRKCDpS664
H3K9me2S10ph1K14ac0	-0.56	-0.58	-0.77	-0.81	-0.69	-0.69	-0.63	-0.95	-0.74
H3K9me3S10ph1K14ac0	-0.52	-0.54	-0.71	-0.75	-0.64	-0.64	-0.59	-0.89	-0.69
H3K18ub1K23ac0	-0.21	-0.22	-0.30	-0.32	-0.27	-0.27	-0.24	-0.38	-0.29
H4(20to23)K20me0	-0.09	-0.09	-0.15	-0.16	-0.13	-0.13	-0.11	-0.21	-0.14
H3K4ac1	2.44	2.49	3.11	3.25	2.86	2.86	2.66	3.71	3.04

(a) PLSR coefficients of proteins w.r.t each histone (NCP-based NTF)

Histone\Protein	COL6A1	EGFRpY1173	EIF4EBP1	ESR1pS118	MCL1	PRKCDpS664
H3K9me2S10ph1K14ac0	-0.39	-0.40	-0.36	-0.35	-0.45	-0.38
H3K9me3S10ph1K14ac0	-0.37	-0.39	-0.35	-0.34	-0.44	-0.37
H3K18ub1K23ac0	-0.12	-0.12	-0.11	-0.10	-0.14	-0.11
H4(20to23)K20me0	-0.04	-0.04	-0.03	-0.03	-0.06	-0.04
H3K4ac1	1.67	1.72	1.60	1.57	1.87	1.66

(b) PLSR coefficients of proteins w.r.t each histone (Tensorly-based NTF)

Fig. 10: The coefficient matrix between the selected histones and proteins.

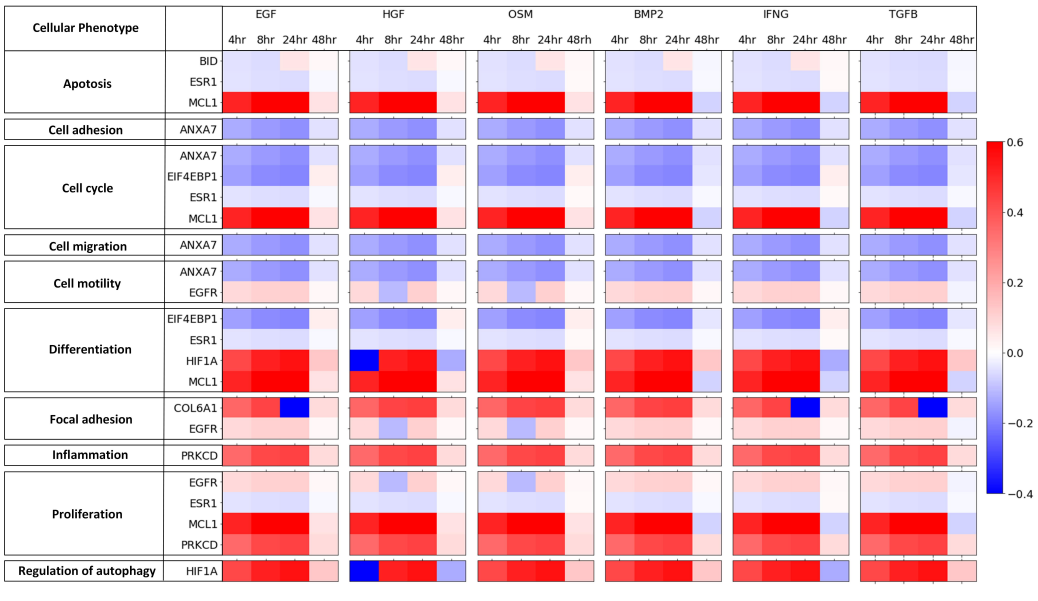


Fig. 11: The Canonical signaling pathways.

and their onsets of regulation, where Fig.9 shows their probability density function curves. We chose a cut-off value at the elbow as the threshold to determine the significance of histones. If the interaction intensity between a histone and a corresponding ligand in $\mathbf{M}_1^g$ is larger than the threshold, then we consider that the histone has been regulated by the respective ligand. Similar rule applies to $\mathbf{M}_2^g$ that determines the times when histones have been regulated. By considering above two factors, we finally filtered out the following 5 significant histones: 1) H3K9me2S10ph1K14ac0, 2) H3K9me3S10ph1K14ac0, 3) H3K18ub1K23ac0, 4) H4(20to23)K20me0, and 5) H3K4ac1, where 1), 2) and 4) were impacted by all 6 ligands, 3) was impacted by EGF and TGFB, and 5) was impacted by the only ligand IFNG. All histones are active in 24th and 48th hours observed except for H3K4ac1 which is active only in the 24th hour. It is worth to note that all 6 ligands are active in 24th and 48th hours based on the intensity distribution in $\mathbf{M}_3^g$ (see supplementary materials). Among the selected histones, H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0 and H3K18ub1K23ac0 are repressive histone marks, while H4(20to23)K20me0 is an active histone mark.

Then, we adopted a Partial Least Squares Regression (PLSR) model (Geladi and Kowalski, 1986) to learn the correlations between histones and proteins based on the two matrices $\mathbf{M}_1^g$ and $\mathbf{M}_1^r$, *i.e.*, to identify which proteins have contributed to the modifications of above 5 significant histones. PLSR is a predictive model that aims to predict a set of dependent variables (a.k.a responses) from a set of independent variables (a.k.a factors). To train a PLSR model, we considered the protein-ligand interactions as factors and the histone-ligand interactions as responses, and aimed to learn the contributions of different proteins to the significance of histone regulations by:

$$\mathbf{M}_1^g = \mathbf{M}_1^r \mathbf{W}_{coef} + \mathbf{E} \tag{17}$$

where $\mathbf{W}_{coef}$ is the coefficient matrix revealing the contributions of proteins to the levels of histone regulations after treated with ligands, and $\mathbf{E}$ is the residual or error matrix for the PLSR model. We performed the Gaussian distribution fitting over the average PLSR coefficients of proteins w.r.t all the histones (for both NCP and Tensorly based results), where all proteins with coefficients over the cut-off value at the elbow on the curve were considered as significant. As a result, we obtained 9 and 6 most relevant proteins for the 5 selected histones and Fig.10 shows the pair-wise PLSR coefficients between the 5 significant histones and the corresponding proteins with NCP (left) and Tensorly (right) based NTF, respectively.

We can observe from Fig.10 that NCP and Tensorly-based PLSR modelings produce very similar results, *i.e.*, 6 significant proteins are shared for both methods, which empirically verified the roles of identified proteins in triggering the overall histone modifications when treated with ligands. We selected the 6 shared proteins in Fig.10 (a) and (b) and demonstrated their regulation mechanisms in two groups shown in Fig.12. On the top, all 6 proteins were regulated after treating with 6 ligands, which then led to the modifications of the three repressive histone marks and further inhibited the chromatin methylation (e.g., closed status of chromatin). On the other side, the two ligands EGF and TGFB showed regulation on all 6 proteins. Then, these proteins modified the only active histone mark H4(20to23)K20me0 which led to the open status of chromatin.

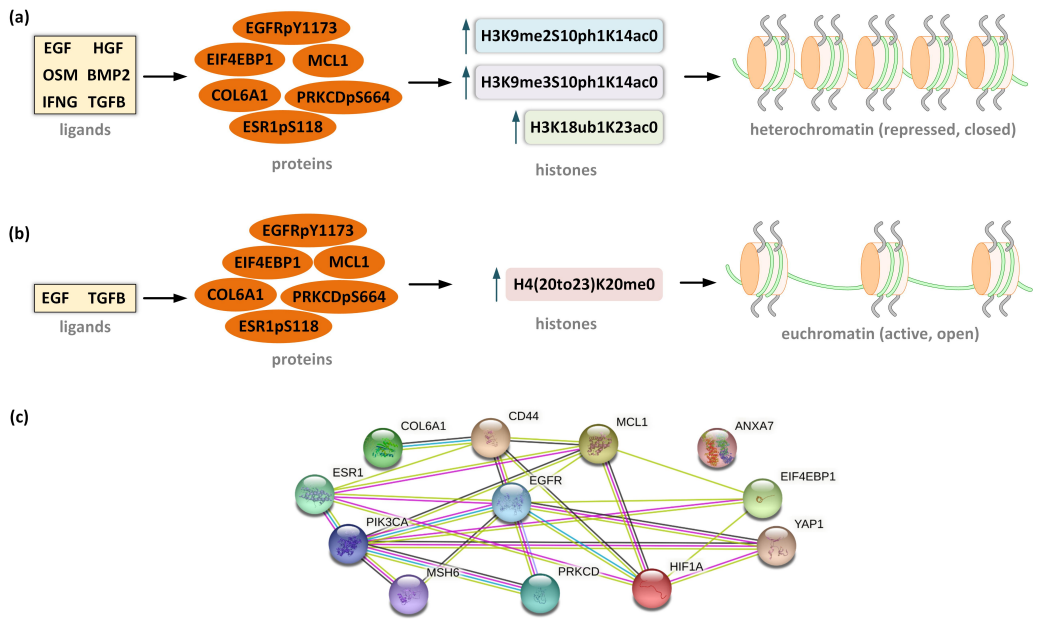


Fig. 12: (a) and (b) show the ligand mediated regulation and (c) shows the string interaction network among enriched proteins.

3.6 Comparison with Existing Findings

Many of above identified proteins have been proven to be significant and paly a role in the breast tumour microenvironment. The expression level of ANXA7 has been reported to be correlated with the occurrence, progression and survival of BC (Sønder *et al.*, 2019). Some results suggest that Anxa7 influences the expression of hormone receptors, which is associated with the poor prognosis of BC development (Guo *et al.*, 2013). EGFR expression is associated with aggressive phenotypes in preclinical breast cancer models (Rimawi *et al.*, 2006). The overexpression of EIF4EBP1 is associated with reduced relapse free patient survival across all breast tumor subtypes (Cha *et al.*, 2015). ESR1 amplification could be a common mechanism in proliferative breast disease and a very early genetic alteration in a large subset of breast cancers (Holst *et al.*, 2007). HIF1A plays a role in escaping tamoxifen's restraining effects on breast cancer (Jögi *et al.*, 2019). MCL-1 is a prognostic indicator and drug target in breast cancer (Campbell *et al.*, 2018). Overexpression of PRKCD is related to the low patient survival in breast cancer (Belle *et al.*, 2015).

In addition, when we selected top 50 most significant proteins for histones based on PLSR, we have observed more important proteins related to the breast cancer for both Tensorly and NCP-based NTFs, including the two chromatin remodellers MSH6 and PIK3CA. MSH6 is a mismatch repair gene which is associated with a statistically significant risk for breast cancer (Roberts *et al.*, 2018). Clinical evidence shows that patients with PIK3CA-mutated triple-negative breast cancer present a better overall survival (Mosele *et al.*, 2020).

3.7 Mapping of Enriched Proteins to Canonical Signaling Pathway

We further mapped the enriched proteins (identified from the NCP-based NTF) and phosphoproteins to their corresponding canonical signaling pathways using KFGG and Reactome databases (Croft *et al.*, 2010) shown in Fig.11, where we observed the effects of all six growth ligands on these proteins and phosphoproteins regulating specific cellular phenotype (behaviors). To generate the regulation profile of these proteins for each of the 6 growth ligands, we first obtained the correlation pattern matrix $\mathbf{M}_1^r$ between proteins and ligands, which can be viewed as the absolute regulation values. Then, we determined the directions of the regulations for different proteins by each of the 6 ligands based on directions of the respective differential abundances of proteins., *i.e.*, if the fold-change of a protein by a particular ligand is negative, then its corresponding value in Fig.11 means down-regulated (negative).

4 Discussion

An NTF-based model is a new way to represent higher-order correlations among various biomolecular signals in the breast tumor microenvironment. Further, elucidating the regulatory principle governing the higher-order signaling analysis among phosphoproteins and histones may provide insight into the structural basis of ECM and chromatin remodeling, uncovering a paradigm in cancer cell signaling. Higher-order assemblies are an essential aspect of many biological processes because they enable the formation of precisely organized molecular machines from constituents present in inactive states at low concentrations, such as protein abundances, to promote biochemical reactions in cells. We have introduced an efficient new tensor decomposition, NTF, based on proteome scale signals of, e.g., RPPA, MS expressions, and histones' phosphoprotein-binding relations among proteins, histones to separate proteome scale nondirectional tensors into mathematically defined decorrelated components. We have demonstrated the utility of NTF on a three-dimensional higher correlation data to uncover higher-order protein-ligand and histone-ligand sub-tensors and their tripling relations among them with biological and statistical significance. This approach focuses mainly on identifying higher-order signaling in intra-pathways in the cellular system and higher-order signaling among these transduction signaling pathways (inter-pathway signaling) and multiple histone signatures. The proposed tensor decomposition was a simple approach to extract the orthogonal decomposition in analyses of higher-order proteins and histones correlations in proteome-scale breast cell line data. These components and tripling among them include reconstruction of higher-order signaling of inter and intra-pathways from nondirectional tensor of higher correlations. The result shows the possible onset of

breast cancer initiation at 24 hour of treatments by all the growth ligands (Fig.8). The result also uncovers new higher-order coordinated differential relations among the effects of specific growth ligands on proteins contributing to specific cellular phenotypes, including cell cycle, cell adhesion, cell migration, cell motility, apoptosis, differentiation, focal adhesion, inflammation, proliferation, and regulation of autophagy (Fig.11).

Our analysis also reveals possible effects of the growth ligands on phosphoprotein mediated histone regulations that may alter chromatin structure. Specifically, we observe significant effects of all ligands on enriched phosphoproteins, namely EGFR, EIF4EBP1, COL6A1, PRKCD, and ESR1 regulating H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0, and H3K18ub1K23ac0 marks to closed conformation (heterochromatin status) starting at 24 hour (Fig.12(a)). We also observe specific effects of EGF and TGFB on these same phosphoproteins regulating H4(20to23)K20me0 mark to an open conformation (Fig.12(b)). From protein-protein interaction (PPI) analysis we further infer that YAP1 protein, a Hippo signaling coactivator and a major regulator of cancer stemness, binds directly with PIK3CA, a chromatin remodeler, and HIF1A, suggesting its possible role in chromatin remodeling affecting the regulation of autophagy, and differentiation when treated with all ligands (Fig.12(c) and Fig.11). We also observe CD44, a cell-surface glycoprotein that plays a role in the degradation of extracellular matrix interacts with EGFR suggesting its role in proliferation, focal adhesion, and cellular motility when treated with EGF, BMP2, IFNG, and TGFB (Fig.12(c), Fig.11). In this study, we demonstrated that a tensor-based higher-order correlation model could be used to capture three-dimensional inter- and intra- signaling pathway interactions of the tumor microenvironment in breast tissue. In the future, we plan to extend this model to capture four or higher-dimensional data, using other omics data sets such as gene expression data (RNASeq) and DNA- protein binding data (ATACSeq).

Contributions

Original concept: S.M.; experimental design: S.M.; data analysis: M.S., K.L., S.M.; implementation: M.S., K.L., S.M.; data interpretation: M.S., S.M.; manuscript writing: M.S., K.L., S.M.

Acknowledgements

We want to thank LINCS consortium to make these data available.

References

Barata, J. C. A. and Hussein, M. S. (2012). The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, **42**(1-2), 146–165.

Belle, L. *et al.* (2015). The tyrosine phosphatase ptpn14 (pez) inhibits metastasis by altering protein trafficking. *Science signaling*, **8**(364), ra18–ra18.

Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, **41**(4), 1350–1362.

Bro, R. *et al.* (1997). Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, **38**(2), 149–172.

Bro, R. and Kiers, H. A. (2003). A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **17**(5), 274–286.

Campbell, K. J. *et al.* (2018). Mcl-1 is a prognostic indicator and drug target in breast cancer. *Cell Death & Disease*, **9**(2), 1–14.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, **35**(3), 283–319.

Carroll, J. D. *et al.* (1989). Fitting of the latent class model via iteratively reweighted least squares candecomp with nonnegativity constraints. In *Multiway data analysis*, pages 463–472.

Cha, Y.-L. *et al.* (2015). Eif4ebp1 overexpression is associated with poor survival and disease progression in patients with hepatocellular carcinoma. *PloS one*, **10**(2), e0117493.

Cichocki, A. *et al.* (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.

Creighton, C. J. and Huang, S. (2015). Reverse phase protein arrays in signaling pathways: a data integration perspective. *Drug design, development and therapy*, **9**, 3519.

Croft, D. *et al.* (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, **39**(suppl_1), D691–D697.

Fang, J. (2019). Tightly integrated genomic and epigenomic data mining using tensor decomposition. *Bioinformatics*, **35**(1), 112–118.

Gauvin, L. *et al.* (2014). Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS one*, **9**(1), e86028.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, **185**, 1–17.

Guo, C. *et al.* (2013). Potential role of annexin a7 in cancers. *Clinica chimica acta*, **423**, 83–89.

Harshman, R. A. *et al.* (1970). Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis.

Holst, F. *et al.* (2007). Estrogen receptor alpha (esr1) gene amplification is frequent in breast cancer. *Nature genetics*, **39**(5), 655–660.

Jögi, A. *et al.* (2019). Expression of hif-1α is related to a poor prognosis and tamoxifen resistance in contralateral breast cancer. *Plos one*, **14**(12), e0226150.

Karami, A. *et al.* (2012). Compression of hyperspectral images using discerete wavelet transform and tucker decomposition. *IEEE journal of selected topics in applied earth observations and remote sensing*, **5**(2), 444–450.

Kim, J. and Park, H. (2012). Fast nonnegative tensor factorization with an active-set-like method. In *High-Performance Scientific Computing*, pages 311–326. Springer.

Kim, J. *et al.* (2014). Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, **58**(2), 285–319.

Kim, Y.-D. and Choi, S. (2007). Nonnegative tucker decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51**(3), 455–500.

Kossaifi, J. *et al.* (2019). Tensorly: Tensor learning in python. *The Journal of Machine Learning Research*, **20**(1), 925–930.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Li, Y. and Ngom, A. (2010). Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data. In *2010 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 438–443. IEEE.

Litichevskiy, L. *et al.* (2018). A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. *Cell systems*, **6**(4), 424–443.

Mollah, S. A. and Subramaniam, S. (2020). Histone signatures predict therapeutic efficacy in breast cancer. *IEEE Open Journal of Engineering in Medicine and Biology*, **1**, 74–82.

Mosele, F. *et al.* (2020). Outcome and molecular landscape of patients with pik3ca-mutated metastatic breast cancer. *Annals of Oncology*, **31**(3), 377–386.

Rimawi, M. *et al.* (2006). Egfr expression in breast cancer: association with biologic phenotype, prognosis, and resistance to adjuvant therapy. *Journal of Clinical Oncology*, **24**(18_suppl), 513–513.

Roberts, M. E. *et al.* (2018). Msh6 and pms2 germ-line pathogenic variants implicated in lynch syndrome are associated with breast cancer. *Genetics in Medicine*, **20**(10), 1167–1174.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, **17**(4), 688–690.

Sønder, S. L. *et al.* (2019). Annexin a7 is required for escrt iii-mediated plasma membrane repair. *Scientific reports*, **9**(1), 1–12.

Spill, F. *et al.* (2016). Impact of the physical microenvironment on tumor progression and metastasis. *Current opinion in biotechnology*, **40**, 41–48.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**(3), 279–311.

Van Loan, C. F. (2000). The ubiquitous kronecker product. *Journal of computational and applied mathematics*, **123**(1-2), 85–100.

Walker, C. *et al.* (2018). Role of extracellular matrix in development and cancer progression. *International journal of molecular sciences*, **19**(10), 3028.

Zhou, P. *et al.* (2017). Tensor factorization for low-rank tensor completion. *IEEE Transactions on Image Processing*, **27**(3), 1152–1163.