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

Min Shi[1‡], Rintsen Sherpa[2‡], Liubou Klindziuk[2], Stefanie Kriel[2], Shamim Mollah[1,2*]

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

‡These authors contributed equally to this work.
* smollah@wustl.edu

Abstract

An in-depth understanding of epithelial breast cell response to growth-promoting ligands is required to elucidate how signals from the microenvironment affect cell-intrinsic regulatory networks and their resultant cellular phenotypes, such as cell growth, progression, and differentiation. Understanding the cellular response to these signals is particularly important in understanding the mechanisms of breast cancer initiation and progression. There is increasing evidence that aberrant epigenetic marks are present in cells of the breast tumor microenvironment and are known to affect primary cellular processes like proliferation, differentiation, and apoptosis. However, the current mechanisms by which epigenetic microenvironment signals influence these cellular phenotypes are not well established. We developed a high-order correlation method using proteomics data to reveal the regulatory dynamics among signaling proteins, histones, and growth-promoting ligands in the breast epithelial cells. 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 (NTF) model is used to capture and quantify the protein-ligand and histone-ligand correlations spanning all time points, followed by a correlation analysis to reveal the pairwise 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. Furthermore, the differential and enrichment analyses of transcriptomic data corroborated with the correlations found between ligands and signaling proteins by the NTF-based model.

Author summary

Breast cancer is the most common malignancy, which has become the dominant cause of cancer-relevant death in women worldwide. Breast cancer comprises not only tumor cells but also epigenetic alterations in the surrounding microenvironment (ME). Increasing evidence shows that aberrant epigenetic marks are present in cells of the ME, contributing to cancer initiation, growth, progression, invasion, and metastasis. As a result, epigenetic ME responses are key indicators of the therapeutic efficacy of various clinical treatments. Due to the importance of epigenetic ME responses in various cancer cellular activities, recent years have seen an increasing interest in developing therapies targeting specific epigenetic biomarkers in the tumor microenvironment (TME).

However, communications among different biomolecules such as growth-promoting ligands, signaling proteins, and epigenetic marks in TME are complex. How their interactions affect various cellular behaviors causing cancer initiation, growth and progression are not well understood. In this study, we focus on the interaction mechanisms among ligands, proteins, and histone modifications, aiming to identify likely biomarkers and the onset of tumorigenesis through computational modeling. The findings from our study will help identify potential biomarkers for developing epigenetic-directed breast cancer therapies in the future.

Introduction

There is increasing evidence that epigenetic alterations in the cellular microenvironment affect normal tissue morphogenesis and promote breast cancer formation and progression [1]. The extracellular matrix (ECM), a major component of the cellular microenvironment, interacts with growth factors to influence cell behaviors contributing to growth, progression, and differentiation [2]. Details about these interactions within the complex and continuously changing breast microenvironment are important during tumor formation but are not well understood. Therefore, an in-depth understanding of these interactions within the tumor microenvironment is critical for the early detection and design of effective anti-cancer therapies. These complex interactions are controlled by higher-order inter- and intra-pathway cell signaling networks involving a cascade of biochemical reactions and physical interactions among proteins, genes, 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 state variable at a node ("state" of the node) depends on the state of other 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 thus of 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 [3]. In this study, we develop 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.

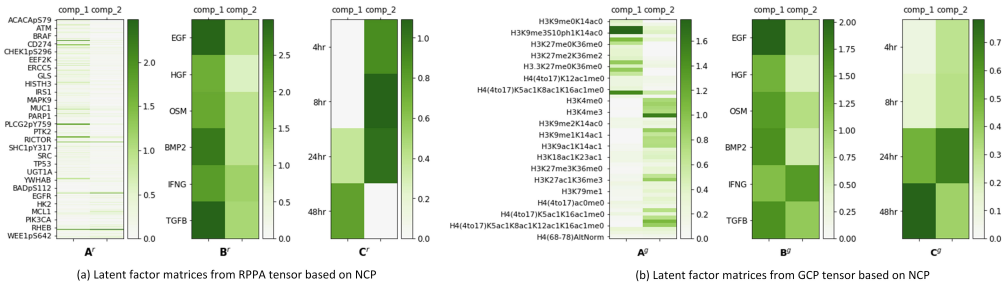
A large-scale initiative, the library of integrated network-based cellular signatures (LINCS) (http://www.lincsproject.org), has carried out the multi-omics characterization of breast epithelial cells to six growth-promoting ligands, through measurement of reverse phase protein array (RPPA) signaling proteins [4], and global chromatin profiles (GCP) [5]. These measurements were carried out at multiple time points post treatment of cells. The RPPA and GCP are antibody and mass spectrometry (MS)-based targeted proteomics assays, respectively, where multiple growth ligands are used to treat a representative set of signaling proteins (RPPA) and different combinations of histone modifications (GCP). This higher-order tensor model study offers the scope for obtaining key signatures of cellular response to growth ligands based on detailed mechanisms reconstructed from LINCS measurements. Specifically, we present an

effective approach to elucidate how microenvironment signals affect cell-intrinsic intracellular protein signals and histone networks, leading to experimentally observable cellular phenotypes. Our methods identify histone modification fingerprints, which are endpoints of the complex signaling events following growth ligand treatment. Chromatin topology changes, caused by histone signature modifications, reflect the altered cellular state.

For characterization of the breast microenvironment responses, we used the human breast epithelial cell line (MCF10A) from the LINCS study which profiled 295 signaling proteins and 65 histone marks at 4, 8, 24, and 48 hours after treatment with six growth-promoting ligands: epidermal growth factor (EGF), hepatocyte growth factor (HGF), oncostatin M (OSM), bone morphogenetic protein 2 (BMP2), transforming growth factor beta (TGFB), and interferon gamma-1b (IFNG). These ligands target diverse receptor types and activate various canonical downstream pathways and were chosen because they elicited multiple phenotype changes affecting proliferation, migration, and morphology. In addition, we performed differential expression and gene set enrichment analysis on RNAseq data from the LINCS study. Observations from changes in gene expression caused by the growth ligands validated the ligand-protein correlations generated by our high-order tensor decomposition method.

Results

This section focuses on the analysis of the factor matrices $\mathbf{A}^r$, $\mathbf{B}^r$, $\mathbf{C}^r$, $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ generated from NTF. Fig.1 shows the factor matrices based on NCP. To explore the interactions among proteins, ligands and histones, we constructed correlation pattern matrices $\mathbf{M}_1^r$, $\mathbf{M}_2^r$, $\mathbf{M}_3^r$, $\mathbf{M}_1^g$, $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$, respectively. We also incorporated these correlation pattern matrices to collectively understand the correlation patterns over time in the breast cancer microenvironment.



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

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

Fig 1. Latent factors generated from NTF. (a) $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ represent proteins, ligands and hours from RPPA that are mapped to a two-dimensional latent space. (b) $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ represent histones, ligands and hours from GCP that are mapped to a two-dimensional latent space.

Latent Factors Analysis

Fig.1a demonstrates the heatmaps of $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ 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 dynamics of the two latent factors change over the 4-, 8-, 24- and 48-hour time points. Fig.1b demonstrates that the heatmaps of $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ from GCP tensor decomposition. Analogically, $\mathbf{A}^g$ and $\mathbf{B}^g$ denote that histones and ligands are mapped in a two-dimensional space, where the dynamic information of the two factors is reflected in $\mathbf{C}^g$.

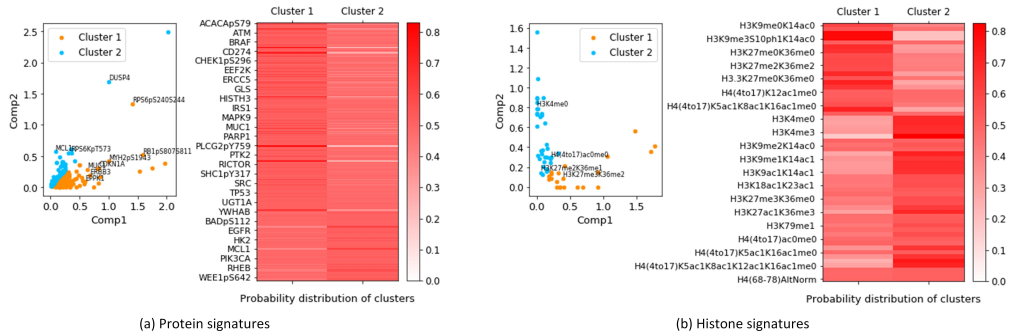


Fig 2. The identified protein and histone signatures. We performed clustering analysis based on latent factor matrices (a) $\mathbf{A}^r$ and (b) $\mathbf{A}^g$. In the scatter plots, each cluster corresponds to a latent component and each protein or histone belongs to the cluster with largest membership value. The matrices show the actual probability distributions of each protein and histone to the respective clusters.

Generation of Protein and Histone Signatures

We further performed the clustering analysis over all proteins and histones based on the latent factor matrices $\mathbf{A}^r$ and $\mathbf{A}^g$. 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.2a shows the two identified protein clusters (also called protein signatures), 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 responses to the 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.2a.

Correlation Patterns Between Proteins and Ligands

To gain insights into the interactions between proteins and ligands over time, we constructed three auxiliary matrices (Fig.3a) through the combinations between $\mathbf{A}_1^r$, $\mathbf{B}_2^r$ and $\mathbf{C}_3^r$ as:

$$\mathbf{M}_1^r = \mathbf{A}^r \otimes (\mathbf{B}^r)^{\mathrm{T}},\ \mathbf{M}_2^r = \mathbf{A}^r \otimes (\mathbf{C}^r)^{\mathrm{T}},\ \mathbf{M}_3^r = \mathbf{B}^r \otimes (\mathbf{C}^r)^{\mathrm{T}}, \tag{1}$$

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

Correlation Patterns between Histones and Ligands

Similarly, we created three matrices (Fig.3b) as follows to reveal the dynamic correlations between histones and ligands:

$$\mathbf{M}_1^g = \mathbf{A}^g \otimes (\mathbf{B}^g)^{\mathrm{T}},\ \mathbf{M}_2^g = \mathbf{A}^g \otimes (\mathbf{C}^g)^{\mathrm{T}},\ \mathbf{M}_3^g = \mathbf{B}^g \otimes (\mathbf{C}^g)^{\mathrm{T}}, \tag{2}$$

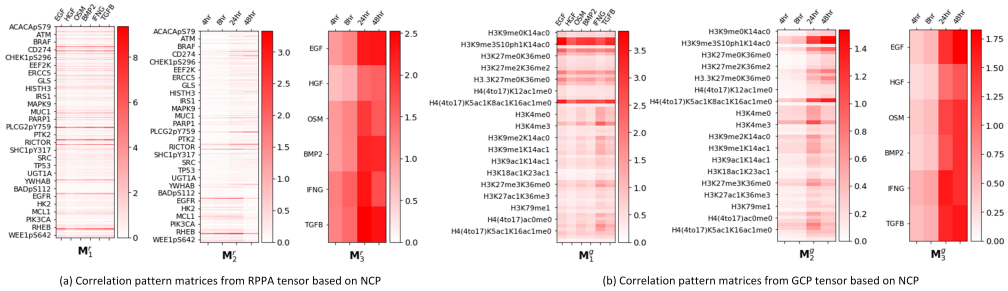


Fig 3. The correlation pattern matrices. (a) $\mathbf{M}_1^r$ reveals the correlation patterns between proteins and ligands, where their onsets are revealed in $\mathbf{M}_2^r$ and $\mathbf{M}_3^r$. (b) $\mathbf{M}_1^g$ reveals the correlation patterns between histones and ligands, where their onsets are revealed in $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$.

where $\mathbf{M}_1^g$ uncovers the intensities of correlation between histones and ligands, with their onsets reflected in $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$.

Correlation Patterns between Proteins, Ligands and Histones

To gain insights into the mechanism of how ligands regulate proteins and further mediate histones, we obtained all the significant proteins and histones based on their interactions with ligands. We performed two density distribution learning (e.g., fitting two Gaussian distributions) over all intensity values in $\mathbf{M}_1^r$ and $\mathbf{M}_2^r$ to determine the significant proteins and their onsets of regulation, where Fig.4a and Fig.4b show their probability density function curves. We chose a cut-off value at the elbow as the threshold to determine the significance of proteins. If the interaction intensity between a protein and a corresponding ligand in $\mathbf{M}_1^r$ is larger than the threshold, we then consider that the protein has been regulated by the respective ligand. Similar rule applies to $\mathbf{M}_2^r$ which determines the times when proteins have been regulated by ligands. By considering above two factors, we finally filtered out eight significant proteins and phosphoproteins: 1) CCNB1, CDC2, PLK1, RB1pS807S811, RPS6pS240S244, DUSP4 and RPS6pS235S236 were regulated by all six ligands, and 2) MYH2 was regulated by EGF, BMP2 and TGFB.

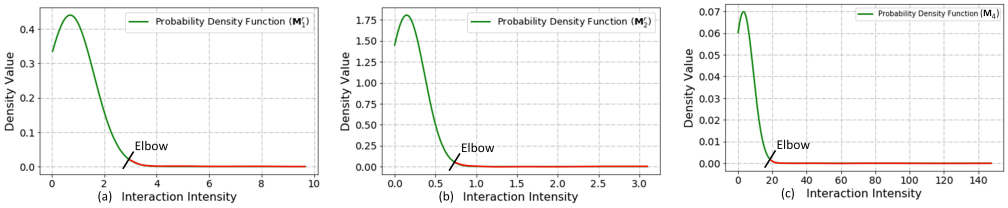


Fig 4. The distribution of intensity values for correlation pattern matrices. We performed a density distribution learning on (a) $\mathbf{M}_1^r$ between signaling proteins and growth ligands, (b) $\mathbf{M}_2^r$ between signaling proteins and time points, and (c) $\mathbf{M}_4$ between signaling proteins and histones. Correlations with interaction intensities above the elbow are considered to be significant.

Similarly, to decide significant histones, we performed two distribution learning (e.g., fitting two Gaussian distributions) over all intensity values in $\mathbf{M}_1^g$ and $\mathbf{M}_2^g$ to determine the significant histones and their onsets of regulation. We finally filtered out the following five significant histone marks: H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0, H3K18ub1K23ac0, H4(20to23)K20me0 and H3K4ac1.

Histones H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0 and H4(20to23)K20me0 were regulated by all six growth ligands, while histone H3K18ub1K23ac0 was regulated by EGF and TGFB, and histone H3K4ac1 was specifically regulated by IFNG. All histones were active in the 24th and 48th hours observed except for H3K4ac1 which was active only in the 24th hour. Among the selected histones, H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0 and H3K18ub1K23ac0 are repressive histone marks, while H4(20to23)K20me0 is an active histone mark.

It is worth noting that all six ligands were active in the 24th and 48th hours based on the intensity distribution in $\mathbf{M}_3^r$ and $\mathbf{M}_3^g$. Then, we further explored the correlations between histones and proteins to identify which of the eight identified proteins contributed to the modifications of the five identified histones listed above. Based on the two matrices $\mathbf{M}_1^g$ and $\mathbf{M}_1^r$, *i.e.*, we constructed $\mathbf{M}_4$ to represent the interaction intensities between histones and proteins by:

$$\mathbf{M}_4 = \mathbf{M}_1^g \otimes (\mathbf{M}_1^r)^{\mathrm{T}} \tag{3}$$

We then performed a Gaussian distribution fitting the density distribution of the interaction intensities in $\mathbf{M}_4$ (Fig.5a), where all interactions with coefficients over the cut-off value at the elbow on the curve were considered as significant (see Fig.4c). The elbow method has been widely used to decide the threshold in existing works [6, 7]. This resulted in eight significant proteins that have contributed to the modification of the five identified significant histones. It is worth noting that both NCP and Tensorly based tensor factorization have produced the same results and verified above correlations between proteins and histones.

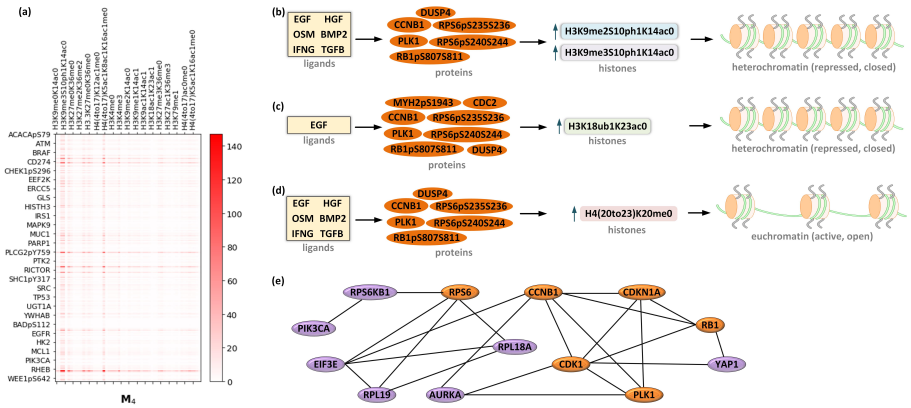


Fig 5. Significant signaling proteins and histones interactions. (a) Showing the NTF-based interactions between signaling proteins and histones. (b), (c) and (d) Showing specific ligands mediated protein and histone regulations, where four significant histones and eight proteins are examined. (e) Showing the protein-protein interaction network among five enriched proteins (orange, obtained based on NTF) and seven inferred (purple) proteins.

Further, we summarized the identified correlations between proteins and histones after treatment with growth ligands in three groups (Fig.5b, 5c and 5d): 1) Six proteins (CCNB1, PLK1, RB1pS807S811, RPS6pS240S244, DUSP4 and RPS6pS235S236) were regulated after treating with six ligands, which then led to the modifications of the two repressive histone marks H3K9me2S10ph1K14ac0 and H3K9me3S10ph1K14ac0 causing the repressed chromatin (e.g., closed status of chromatin); 2) The growth ligand EGF regulated all eight proteins, which then mediated the only active histone mark H3K18ub1K23ac0 which led to the closed status of chromatin; 3) Six growth ligands

regulated six proteins (CCNB1, PLK1, RB1pS807S811, RPS6pS240S244, DUSP4 and RPS6pS235S236) which mediated histone mark H4(20to23)K20me0 resulting in the open status of chromatin.

Mapping of Enriched Proteins to Canonical Signaling Pathway

We further mapped the eight enriched proteins and phosphoproteins to their corresponding canonical signaling pathways using the KEGG and Reactome databases [8] shown in Fig.6, where we observed the effects of all six growth ligands on these proteins and phosphoproteins in regulating specific cellular phenotypes (behaviors). To generate the regulation profile of these proteins for each of the six growth ligands, we first obtained the correlation pattern matrix $\mathbf{M}_1^r$ between proteins and ligands, which can be viewed as the relative regulation values. Then, we determined the directions of the regulations for various proteins and phosphoproteins by their fold-changes; upregulated if positive fold-change, downregulated otherwise (Fig.3).

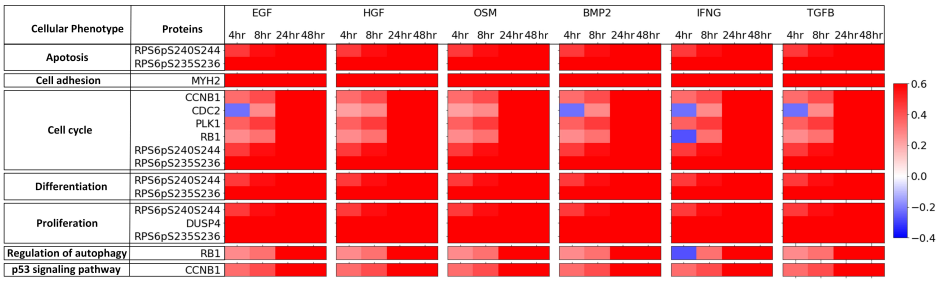


Fig 6. Canonical signaling pathways. Ligands specific signaling pathway profile associated with the enriched proteins at various time points, where positive and negative fold changes are indicated by red and blue colors, respectively.

Differential Expression Analysis

To provide some validation for the ligand-protein correlations produced by the NTF-based model, we performed differential expression analysis on RNAseq data to observe the changes in gene expression caused by the growth ligands. Differential expression analysis was performed using DESeq2 [9] version 1.32.0 in R to compare expression levels in each ligand-treated sample at each time point to the control at the respective time point (e.g. EGF at 24 hours versus PBS at 24 hours).

CCNB1, CDC2 (also known as CDK1), PLK1, and DUSP4 were significantly (FDR < 0.05) upregulated across all ligands at 24 hours (Fig.7). Fig.8 displays the coverage of each ligand at the loci of these genes, where a higher coverage reflects greater gene expression. We see that coverage is noticeably higher in the ligand-treated samples compared to PBS. The upregulation of these genes is in agreement with the NTF-based model, which found the proteins encoded by these genes to be correlated with each of the ligands.

	EGF		HGF		OSM		BMP2		IFNG		TGFB	
Gene	**Log2 Fold Change**	**FDR**	**Log2 Fold Change**	**FDR**	**Log2 Fold Change**	**FDR**	**Log2 Fold Change**	**FDR**	**Log2 Fold Change**	**FDR**	**Log2 Fold Change**	**FDR**
CCNB1	3.51	<0.001	2.90	<0.001	3.32	<0.001	2.97	<0.001	3.43	<0.001	2.41	<0.001
CDK1	3.45	<0.001	2.91	<0.001	3.28	<0.001	3.04	<0.001	3.41	<0.001	2.42	<0.001
PLK1	3.74	<0.001	3.04	<0.001	3.38	<0.001	3.24	<0.001	3.53	<0.001	2.64	<0.001
DUSP4	3.33	<0.001	2.16	<0.001	2.81	<0.001	3.37	<0.001	3.01	<0.001	4.53	<0.001

Fig 7. Differential expression analysis results for each ligand at 24 hours.

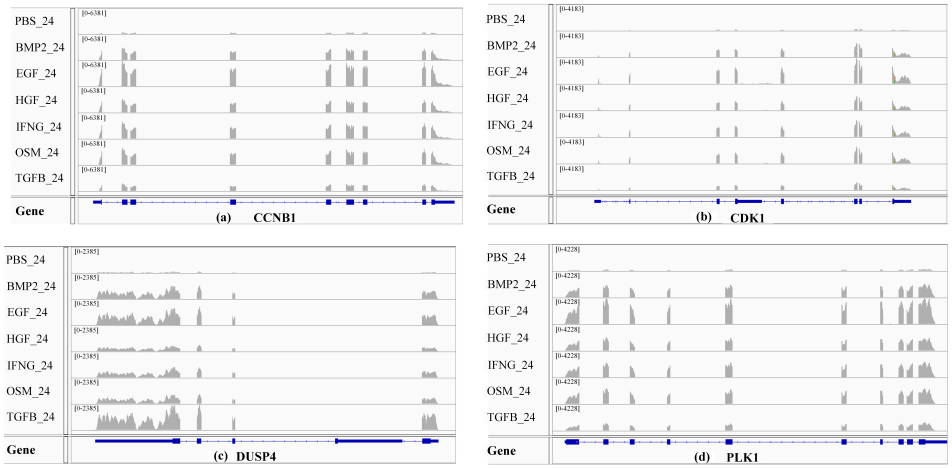


Fig 8. Differential expression analysis. Coverage histograms for each ligand and control at 24 hours showing loci of (a) CCNB1, (b) CDK1, (c) DUSP4, and (d) PLK1. Control is shown on the top track, with the ligands below. Peak height corresponds with expression levels.

Gene Set Enrichment Analysis

We further performed gene set enrichment analysis on the RNAseq data to find out which pathways were most affected by each ligand treatment. Using the RNAseq raw count matrix data, an expression dataset file and a phenotype label file were created as specified in the GSEA User Guide [10] for each individual ligand at a specific time point compared to PBS at the same time point. GSEA [11] version 4.1.0 was run for each comparison using 'c2.cp.kegg.v7.2.symbols.gmt' as the gene set database, 'Human_ENSEMBL_Gene_ID_MSigDB.v7.4' as the chip platform, and 'phenotype' as the permutation type, set at 1000 permutations.

Fig.9 shows the top 5 enriched KEGG pathways for each ligand at 24 hours. Many of the same pathways are highly enriched across all ligands, suggesting that the ligands affect the phenotype of breast cells through similar mechanisms. Cell cycle in particular stands out because the proteins from the NTF-based model are key players in mitosis and cell proliferation [12, 13].

Discussion

A 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 method, NTF, based on proteome scale signals of MS expressions, and on histones' phosphoprotein-binding relations among proteins, thus using histones to separate proteome scale nondirectional tensors into mathematically defined decorrelated components. We have demonstrated the utility of NTF on three-dimensional higher correlation data to uncover higher-order protein-ligand and histone-ligand sub-tensors,

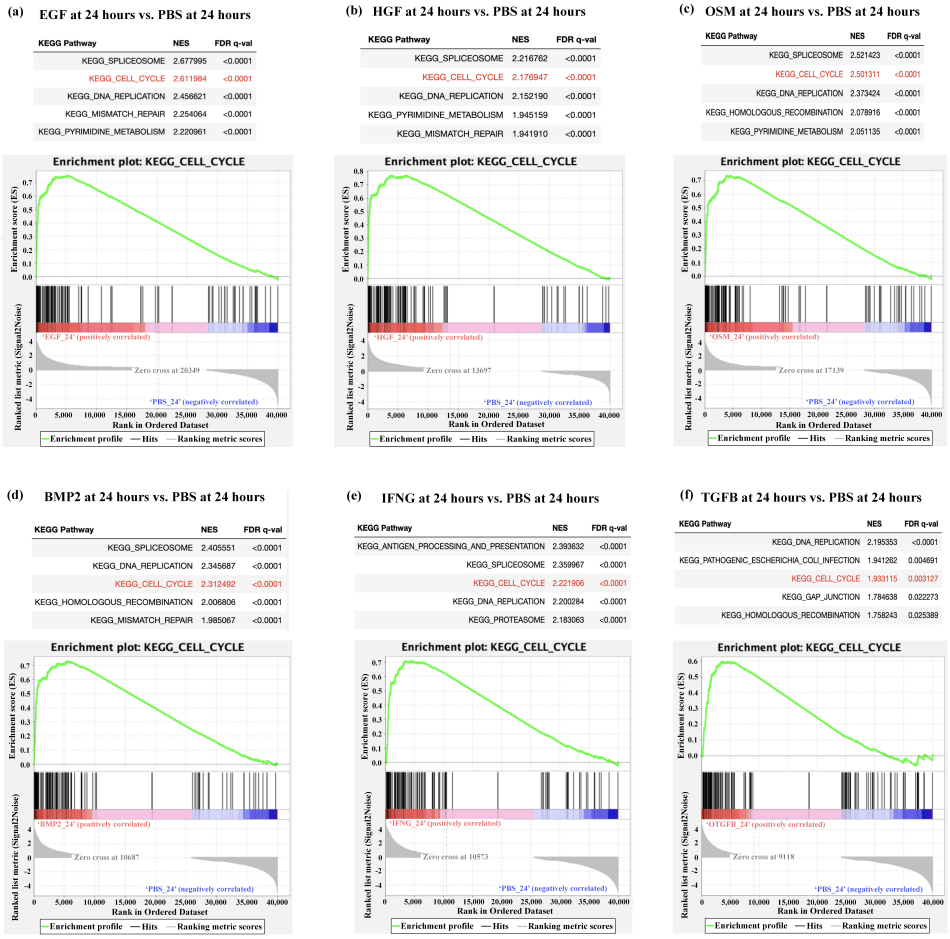


Fig 9. Gene set enrichment analysis. Top 5 enriched KEGG pathways for (a) EGF, (b) HGF, (c) OSM, (d) BMP2, (e) IFNG, and (f) TGFB at 24 hours sorted by normalized enrichment score (NES) and the enrichment plots for the KEGG cell cycle pathway. Enrichment plots show a walk down the list of ranked genes, which is based on the differential expression with respect to the two phenotypes. The black bars in the mid-section represent the location of the cell cycle genes in the ranked gene list. The bottom plot displays the ranked list metric.

and the tripling relations among them with biological and statistical significance. This approach focuses mainly on identifying higher-order signaling within individual pathways in the cellular system (intra-pathway signaling) and higher-order signaling between these transduction signaling pathways (inter-pathway signaling), as well as on identifying multiple histone signatures. The proposed tensor decomposition is 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 within and between pathways from nondirectional tensors of higher correlations. The result shows the possible onset of breast cancer initiation at 24 hours of treatments for all growth ligands (Fig.3). 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 Apotosis, cell adhesion, cell cycle, differentiation, proliferation, regulation of autophagy and p53 signaling pathway (Fig.6).

Our analysis also reveals possible effects of the growth ligands on phosphoprotein mediated histone regulations that may alter chromatin structure. For instance, we observe significant effects of all ligands on enriched phosphoproteins, namely CCNB1, PLK1, RB1, DUSP4 and RPS6 regulating H3K9me2S10ph1K14ac0 and H3K9me3S10ph1K14ac0 marks to closed conformation (heterochromatin status) starting at 24 hour (Fig.5b). We also observe all ligands on these same phosphoproteins regulating the H4(20to23)K20me0 mark to an open conformation (Fig.5d). From protein-protein interaction analyses (Fig.5e), we further infer that the YAP1 protein, a Hippo signaling coactivator and a major regulator of cancer stemness binds directly with RB1 (regulator involves in heterochromatin formation), indirectly with PIK3CA (chromatin remodeler), by directly interacting with RPS6 (ribosomal protein), suggesting its possible role in chromatin remodeling affecting the cell proliferation, regulation of autophagy, and differentiation in breast cancer when treated with all the ligands (Fig.5b, c, d).

Through the differential gene expression analysis, we further corroborated with the correlations found between ligands and proteins by the NTF-based model. Namely, genes CCNB1, CDK1, PLK1, and DUSP4 were significantly upregulated across all 6 ligands at 24 hours, and the model predicted that all 6 ligands have a significant effect on the proteins encoded by these genes. Similar gene expression profiles have also been reported in other studies, providing further support for the reliability of the model [14–18]. However, for robust support for the model to be achieved we must analyze additional datasets from the LINCS study and conduct validation experiments. Dysregulation of cell cycle genes is a hallmark of tumor cells across various cancer types including breast cancer, and cyclins and cyclin-dependent kinases in particular are important markers. High CCNB1 expression is associated with poor survival in breast cancer [19], and cyclin expression levels are even used to subtype breast cancers [20]. While much is known about their expression profiles and how they interact with other proteins in breast tumor cells, their involvement in histone modification and downstream epigenetic effects on the tumor microenvironment are relatively unknown. The protein-histone correlations produced by the NTF-based model suggest that these genes have some sort of interaction with the histone modifications H3K9me2S10ph1K14ac0, H3K9me3S10ph1K14ac0, H4(20to23)K20me0, H3K18ub1K23ac0, and H3K4ac1. Based on current knowledge, there are a few possible mechanisms by which cell cycle genes may be involved in histone modification. CDK1 and PLK1 have been found to affect expression of histone methyltransferases via phosphorylation [21–23]. The presence or absence of methyl groups on histones can alter the conformation of chromatin, causing downstream regulatory effects on many genes and drastically altering the tumor microenvironment. The genes may be indirectly modifying histones, like in the above-mentioned studies, or it may be direct. In order to elucidate the details of protein-histone correlations, we plan to use the ATACseq data from the LINCS study to understand the chromatin accessibility landscape of the cells, and perform gene knockout experiments with CUT&RUN [24] to find the direct effects of these genes on the histones from the model. Ultimately, we will be able to create a network of these genes, their transcription factors, and the histones.

Materials and Methods

This section proposes a Non-negative Tensor Factorization (NTF) [25] based approach to model the high-order correlations among proteins, ligands and histones based on collected multi-omics data containing two time-serial proteomic datasets. Fig.10 shows the overall workflow of our method which consists of three subsequent components. First, we collect the time-serial multi-omics datasets which are represented as two 3D

tensors (RPPA and GCP). Then, we adopt a NTF based method to model the high-order correlations encoded in the RPPA and GCP tensors. Finally, we focus on the interpretation of latent factors derived from the RPPA and GCP tensor factorization.

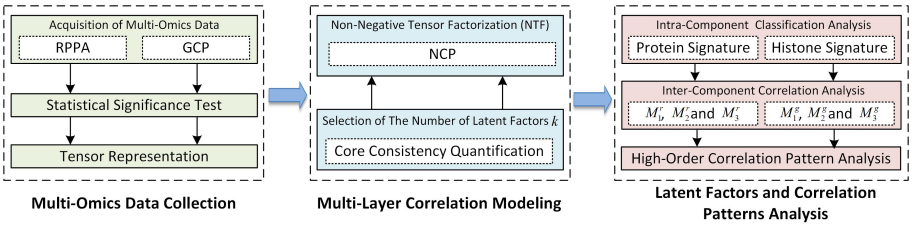


Fig 10. Workflow of the proposed approach. The proposed method involves three major components including multi-omics data collection, high-order data correlation modeling using non-negative tensor factorization, and interpretation of tensor factorization results.

Multi-Omics Datasets

Breast cancers are incredibly complicated at the molecular level, which necessitates multi-omics data at different levels to gain a complete insight into the interactive mechanisms among biomolecules such as proteins, histones and ligands. To understand the dynamic regulations of growth ligands in the breast cancer microenvironment, we specifically collected two datasets named RPPA and GCP respectively. In addition, we also collected a RNAseq dataset to validate the results produced by our method at the gene expression level.

RPPA and GCP Data Acquisition

We downloaded the level3 (log 2 normalized) RPPA data from the Synapse platform (https://www.synapse.or g/#!Synapse:syn12555331) in June 2020. The dataset consists of highly sensitive and selective antibody-based measurements of 295 proteins and phosphoproteins after individual treatments with six different growth ligands, namely, EGF, HGF, OSM, BMP2, TGFB, and IFNG. Each data sample contains three replicates for each of these ligands and a control treated with phosphate-buffered saline (PBS) measured at five time points (1, 4, 8, 24, and 48 hours).

We downloaded the level 3 (log2 normalized) GCP data from the Synapse platform (https://www.synapse.org/#!Synapse:syn 18491838) in August 2020. This dataset consists of 65 probes monitoring combinations of post-translational modification on histones in MCF10A healthy breast epithelial cell line. Cells in these assays were treated with EGF, HGF, OSM, BMP2, TGFB, or IFNG at various concentrations. The data contains three replicates for each of these ligands and a control treated with PBS. These samples were measured at four time points (4, 8, 24, and 48 hours).

To perform consistent analysis on the above two datasets, we removed all samples in the first hour from the RPPA data. As a result, both RPPA and GCP data involve four distinct time points of 4, 8, 24 and 48 hours. For both the 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, we applied an unpaired Student's t-test [26] between each sample and its respective control measurements. We kept all significant samples with p-value<0.05, and 274 proteins in RPPA and 58 histones in GCP were retained in the experiment.

RNAseq Data Acquisition

The level 0 (fastq) RNAseq data were downloaded from the Synapse platform (https://www.synapse.org/#!Synapse:syn18518040) in January 2021. 100bp single-end reads were sequenced on an Illumina HiSeq 2500 Sequencer by OHSU MPSSR. The dataset contains the gene expression data of MCF10A breast cells treated with six growth ligands: EGF, HGF, OSM, BMP2, TGFB, and IFNG. Each data sample contains three replicates for each ligand measured at 24 and 48 hours (EGF has four replicates per timepoint). Cells were also treated with a control (PBS) and measurements were taken at 0, 24, and 48 hours. The fastq files were aligned to the ENSEMBL GRCh38 assembly of the human genome using the STAR sequence aligner [27] version 2.7.3 and output as BAM files. A count matrix of all samples and their genes was generated using the featureCounts program [28] from the R package RSubread [29] version 2.6.1 and the ENSEMBL GRCh38 gene annotations.

Tensor Representation of RPPA and GCP Data

After performing the Student's t-test and sample screening processes, the three replicates for each sample were averaged. Then, the differential abundance (fold change, *i.e.*, the quantity change between each sample measurement and its control measurement) of signaling proteins and histones relative to their respective controls were calculated. Fig.11 illustrates the differential abundance of an example protein from RPPA data and an example histone from the GCP data. For example, after treated with ligand IFNG, expression of the protein PIK3CA was initially decreased during the 4th and 8th hours, and was then elevated until the 48th hour. Meanwhile, expression of the histone mark H3K9me2S10ph1K14ac0 showed consistent elevation from the 4th to 48th hour.

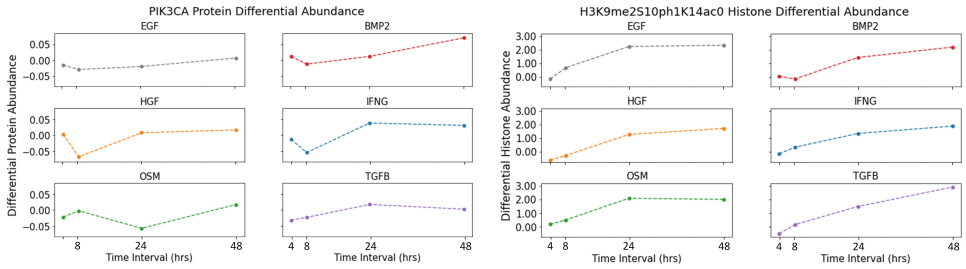


Fig 11. Differential abundance changing over time. The differential abundances of protein PIK3CA and histone H3K9me2S10ph1K14ac0 after treatment with six growth ligands at various time points.

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. $\mathbf{T}^r \in \mathbb{R}^{274\times 4}$ and $\mathbf{T}^g \in \mathbb{R}^{58\times 4}$ are visualized in two 3-way tensors shown in Fig.12.

To capture both significant negative and positive fold-changes in the non-negative factorization, we first applied an absolute value 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 indicate the directions (e.g., either positive or negative) of protein and histone fold-changes, respectively. Then, the absolute value transformations

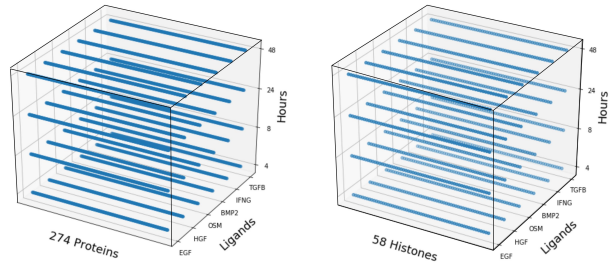


Fig 12. The 3-way tensor representations of RPPA and GCP. Each tensor has three dimensions or modes, e.g., protein, ligand and time (hour) for RPPA (left) , and histone, ligand and time (hour) for GCP (right).

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.

Multi-Layer Correlation Modeling

The tensor representations shown in Fig.12 encode for the correlations or interactions between proteins and ligands (or histones and ligands) changing over time. In this section, we present a method to model above multi-layer (e.g., four time points) correlations using the NTF technique.

Non-Negative Tensor Factorization (NTF)

Tensor factorization can be viewed as the high-order version of matrix factorization [30] by extending the two-dimensional matrices to tensors. A tensor is a multi-way array or multi-dimensional matrix [31] formally defined as:

Definition 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 $1 \leq n \leq N$.*

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 (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 $\mathbf{T}_{i_1, i_2, \cdots, i_N} = \mathbf{a}^{(1)}_{i_1} \mathbf{a}^{(2)}_{i_2} \cdots \mathbf{a}^{(N)}_{i_N}$ for all $1 \leq i_n \leq I_N$.*

Tensor representations are a powerful way to encode the complex relations across different modes such as protein, 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 the Parallel Factor Analysis (PARAFAC) [32] and Tucker [33], which have shown promises in many real-world applications such as cancer survival prediction [34] and image processing [35].

PARAFAC and Tucker

The non-negative PARAFAC was proposed by Carroll *et al.* [36] and then developed by Bro [37] with a more efficient method to calculate the non-negative components. Fig.13a shows the graphical representation of PARAFAC, which aims to decompose a 3-way tensor $\mathbf{T} \in \mathbb{R}^{I_1, I_2, I_3}$ into a sum of component rank-one tensors as:

$$\mathbf{T} \approx \sum_{r=1}^{k} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{4}$$

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 $\mathbf{T}$. Data relationships in the original tensor are thereby fully preserved by the resulting components. Alternatively, the PARAFAC decomposition can be represented as:

$$\mathbf{T} = [[\mathbf{A}, \mathbf{B}, \mathbf{C}]], \tag{5}$$

where the loading matrices $\mathbf{A}$,$\mathbf{B}$,$\mathbf{C}$ can be understood as the linear mapping from the respective modes to the latent spaces defined as:

$$\begin{aligned} \mathbf{A} &= [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k] \in \mathbb{R}^{I_1 \times k} \\ \mathbf{B} &= [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_k] \in \mathbb{R}^{I_2 \times k} \\ \mathbf{C} &= [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k] \in \mathbb{R}^{I_3 \times k}. \end{aligned} \tag{6}$$

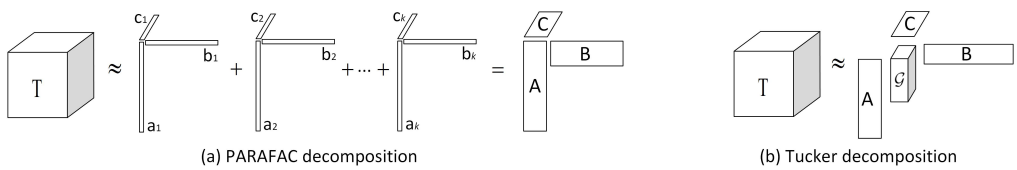


Fig 13. The graphical representations of PARAFAC and Tucker. Tucker contains a core tensor $\mathcal{G}$ and PARAFAC is a special case of Tucker when the core tensor $\mathcal{G}$ is super-diagonal.

Tucker decomposition is another popular NTF technique [33]. As shown in Fig.13b, 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 $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ represented as:

$$\begin{aligned} \mathbf{T} &\approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} \sum_{r_3=1}^{k_3} g_{r_1 r_2 r_3} \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \end{aligned} \tag{7}$$

where $k_1 \times k_2 \times k_3$ rank-one tensors are produced, unlike 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.

From Fig.13 we can see that PARAFAC is a special case of Tucker with the super-diagonal core tensor $\mathcal{G}$. Studies [38, 39] 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 that the rank-one decomposition tensors of PARAFAC are often unique, whereas Tucker decomposition is not [25]. For these reasons, we chose to use non-negative PARAFAC to analyse and study the tumor microenvironment. There are many methods to implement the 3-way PARAFAC including Nonnegative CANDECOMP/PARAFAC (NCP) [40] and Tensorly [41]. In this study, we adopt NCP for RPPA and GCP tensor factorization.We used Tensorly to further validate the results obtained from NCP.

Non-Negative PARAFAC Based on NCP

NCP has been widely used to solve the non-negative PARAFAC decomposition, which is implemented based on the Block Principal Pivoting (BPP) method [42]. For a given 3-way tensor $\mathbf{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, NCP aims to minimize the following objective:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \parallel \mathbf{T} - [[\mathbf{A}, \mathbf{B}, \mathbf{C}]] \parallel_F^2 \quad \text{subject to: } \mathbf{A}, \mathbf{B}, \mathbf{C} \geq 0, \tag{8}$$

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: $\mathbf{X}(\in \mathbb{R}^{Q\times L}) = \arg\min_{\mathbf{X}\geq 0} \parallel \mathbf{VX} - \mathbf{W} \parallel_F^2$

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 $\mathbf{V}_{\mathcal{F}_l}^{\mathrm{T}}\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l}$ using column grouping;
Compute all $\mathbf{y}_{\mathcal{G}_l} = \mathbf{V}_{\mathcal{G}_l}^{\mathrm{T}}(\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l} - \mathbf{w}_l)$
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 $\mathbf{V}_{\mathcal{F}_l}^{\mathrm{T}}\mathbf{V}_{\mathcal{F}_l}\mathbf{x}_{\mathcal{F}_l}$ using column grouping;
    end for
end while

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

$$\begin{aligned} &\mathbf{T}_{(n)} \approx \mathbf{A}^{(n)} \times \left(\mathbf{D}^{(n)}\right)^{\mathrm{T}}, \text{for } \forall n \in \{1, \cdots, N\} \\ &\text{with } \mathbf{D}^{(n)} = \mathbf{A}^{(N)} \circ \cdots \circ \mathbf{A}^{(n+1)} \circ \mathbf{A}^{(n-1)} \circ \cdots \circ \mathbf{A}^{(1)} \end{aligned} \tag{9}$$

where $N = 3$ in this paper and $\mathbf{T}_{(n)}$ represents the mode-n matricization of tensor $\mathbf{T}$. Then, the optimization in Eq.5 can be viewed as solving the following subproblems:

$$\begin{aligned} &\min_{\mathbf{A}^{(n)}} \parallel \mathbf{D}^{(n)} \times \left(\mathbf{A}^{(n)}\right)^{\mathrm{T}} - \left(\mathbf{T}_{(n)}\right)^{\mathrm{T}} \parallel_F^2, \text{for } \forall n \in \{1, \cdots, N\} \\ &\text{subject to: } \mathbf{A}^{(n)} \geq 0 \end{aligned} \tag{10}$$

From Eq.7 we can see that NCP actually reduces the NTF to N independent non-negative matrix factorization problems, where each problem can be simplified as:

$$\min_{\mathbf{X}\geq 0} \parallel \mathbf{VX} - \mathbf{W} \parallel_F^2, \tag{11}$$

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 [42] 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 $\mathbf{X}$ and $\mathbf{Y}$ indexed by $\mathcal{F}_l$ and $\mathcal{G}_l$, respectively.

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, and should be approximately the actual rank of the target tensor.

It is non-trivial to determine the best number of latent factors k, which normally necessitates multiple trials and fine-tuning [43]. In the past, the *core consistency* (CC) metric has been widely used to evaluate the fitness of a candidate k [44], 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 the previous section, the PARAFAC decomposition is actually a special case of the Tucker decomposition (in Eq.4) since Eq.1 can be rewritten as:

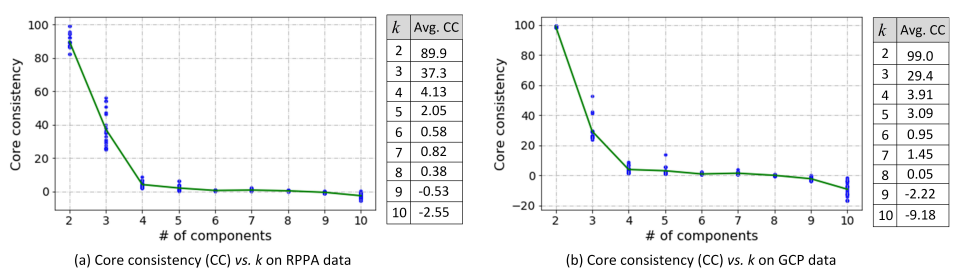
$$\mathbf{T} \approx \sum_{r_1=1}^{k} \sum_{r_2=1}^{k} \sum_{r_3=1}^{k} g^d_{r_1,r_2,r_3} \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \tag{12}$$

where $\mathcal{G}^d = \left\{ g^d_{r_1,r_2,r_3} \right\}_{r_1,r_2,r_3=1,\cdots,k} \in \mathbb{R}^{k\times k\times k}$ 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 [44]. 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 CC is defined as the difference between $\mathcal{G}$ and $\mathcal{G}_d$ to quantify the goodness of a particular k:

$$\mathrm{CC} = 100 \left(1 - \frac{\sum_{r_1=1}^{k} \sum_{r_2=1}^{k} \sum_{r_3=1}^{k} (g_{r_1,r_2,r_3} - g^d_{r_1,r_2,r_3})^2}{k} \right). \tag{13}$$

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 [43, 44].

We leverage CC as the metric to determine the optimal number of latent factors k for RPPA tensor and GCP tensor, respectively. Fig.14 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 this study, we consider $k = 2$ as the optimal number of latent components for PARAFAC decomposition on both RPPA and GCP data.



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 14. The core-consistency *vs.* the number of components (k). The core consistency for each k tested on the RPPA tensor (a) and GCP tensor (b) 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.

NTF on RPPA and GCP Tensors

After determining the optimal number of components by core consistency, we applied the non-negative PARAFAC model implemented in NCP to decompose the RPPA

Algorithm 2 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 $\mathbf{C}^r$.

tensor $\mathbf{T}^r$ and GCP tensor $\mathbf{T}^g$, respectively. The non-negative factorization on RPPA data based on NCP is summarized in Algorithm 2, which takes the 3-way RPPA tensor $\mathbf{T}^r$ as input and then produces three latent factor matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$. We adopted the Nonnegative Double Singular Value Decomposition (NNDSVD) [45] to initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$, which is based on two Singular Value Decomposition (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.

Similar to the RPPA tensor, we were able to decompose the GCP tensor $\mathbf{T}_g$ and derive the latent factor matrices $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ based on NCP.

Acknowledgments

We want to thank LINCS consortium for making these data available. We are grateful to Kevin O'Connor, Olivia Gomez, and Jennie Hazen for the helpful comments.

Author Contributions

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

References

1. Jeschke J, Collignon E, Fuks F. DNA methylome profiling beyond promoters–taking an epigenetic snapshot of the breast tumor microenvironment. The FEBS journal. 2015;282(9):1801–1814.

2. Walker C, Mojares E, del Río Hernández A. Role of extracellular matrix in development and cancer progression. International journal of molecular sciences. 2018;19(10):3028.

3. Mollah SA, Subramaniam S. Histone Signatures Predict Therapeutic Efficacy in Breast Cancer. IEEE Open Journal of Engineering in Medicine and Biology. 2020;1:74–82.

4. Creighton CJ, Huang S. Reverse phase protein arrays in signaling pathways: a data integration perspective. Drug design, development and therapy. 2015;9:3519.

5. Creech AL, Taylor JE, Maier VK, Wu X, Feeney CM, Udeshi ND, et al. Building the Connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry. Methods. 2015;72:57–64.

6. Shi C, Wei B, Wei S, Wang W, Liu H, Liu J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking. 2021;2021(1):1–16.

7. Antunes M, Ribeiro J, Gomes D, Aguiar RL. Knee/elbow point estimation through thresholding. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE; 2018. p. 413–419.

8. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic acids research. 2010;39(suppl_1):D691–D697.

9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):1–21.

10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550.

11. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics. 2003;34(3):267–273.

12. Lindqvist A, Rodríguez-Bravo V, Medema RH. The decision to enter mitosis: feedback and redundancy in the mitotic entry network. Journal of Cell Biology. 2009;185(2):193–202.

13. Owens D, Keyse S. Differential regulation of MAP kinase signalling by dual-specificity protein phosphatases. Oncogene. 2007;26(22):3203–3213.

14. Borlak J, Meier T, Halter R, Spanel R, Spanel-Borowski K. Epidermal growth factor-induced hepatocellular carcinoma: gene expression profiles in precursor lesions, early stage and solitary tumours. Oncogene. 2005;24(11):1809–1819.

15. Nakamura K, Nonaka H, Saito H, Tanaka M, Miyajima A. Hepatocyte proliferation and tissue remodeling is impaired after liver injury in oncostatin M receptor knockout mice. Hepatology. 2004;39(3):635–644.

16. Lv S, Wang X, Bai X, Ning H, Li Y, Wen H, et al. Mesenchymal epithelial transition factor regulates tumor necrosis factor-related apoptotic induction ligand resistance in hepatocellular carcinoma cells through down-regulation of cyclin B1. The International Journal of Biochemistry & Cell Biology. 2020;128:105844.

17. Guo M, Jiang Z, Zhang X, Lu D, Ha AD, Sun J, et al. miR-656 inhibits glioma tumorigenesis through repression of BMPR1A. Carcinogenesis. 2014;35(8):1698–1706.

18. Wang C, Wang C, Wei Z, Li Y, Wang W, Li X, et al. Suppression of motor protein KIF3C expression inhibits tumor growth and metastasis in breast cancer by inhibiting TGF-β signaling. Cancer Letters. 2015;368(1):105–114.

19. Aaltonen K, Amini RM, Heikkilä P, Aittomäki K, Tamminen A, Nevanlinna H, et al. High cyclin B1 expression is associated with poor survival in breast cancer. British journal of cancer. 2009;100(7):1055–1060.

20. Agarwal R, Gonzalez-Angulo AM, Myhre S, Carey M, Lee JS, Overgaard J, et al. Integrative analysis of cyclin protein levels identifies cyclin b1 as a classifier and predictor of outcomes in breast cancer. Clinical cancer research. 2009;15(11):3654–3662.

21. Michowski W, Chick JM, Chu C, Kolodziejczyk A, Wang Y, Suski JM, et al. Cdk1 controls global epigenetic landscape in embryonic stem cells. Molecular cell. 2020;78(3):459–476.

22. Zeng X, Chen S, Huang H. Phosphorylation of EZH2 by CDK1 and CDK2: a possible regulatory mechanism of transmission of the H3K27me3 epigenetic mark through cell divisions. Cell cycle. 2011;10(4):579–583.

23. Wierer M, Verde G, Pisano P, Molina H, Font-Mateu J, Di Croce L, et al. PLK1 signaling in breast cancer cells cooperates with estrogen receptor-dependent gene transcription. Cell reports. 2013;3(6):2021–2032.

24. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife. 2017;6:e21856.

25. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM review. 2009;51(3):455–500.

26. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology. 2006;17(4):688–690.

27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

28. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–930.

29. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic acids research. 2019;47(8):e47–e47.

30. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems; 2001. p. 556–562.

31. Cichocki A, Zdunek R, Phan AH, Amari Si. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons; 2009.

32. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika. 1970;35(3):283–319.

33. Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika. 1966;31(3):279–311.

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

35. Zhou P, Lu C, Lin Z, Zhang C. Tensor factorization for low-rank tensor completion. IEEE Transactions on Image Processing. 2017;27(3):1152–1163.

36. Carroll JD, De Soete G, Pruzansky S. Fitting of the latent class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints. In: Multiway data analysis; 1989. p. 463–472.

37. Bro R, et al. PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems. 1997;38(2):149–172.

38. Kim YD, Choi S. Nonnegative tucker decomposition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2007. p. 1–8.

39. Karami A, Yazdi M, Mercier G. Compression of hyperspectral images using discerete wavelet transform and tucker decomposition. IEEE journal of selected topics in applied earth observations and remote sensing. 2012;5(2):444–450.

40. Kim J, He Y, Park H. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. Journal of Global Optimization. 2014;58(2):285–319.

41. Kossaifi J, Panagakis Y, Anandkumar A, Pantic M. Tensorly: Tensor learning in python. The Journal of Machine Learning Research. 2019;20(1):925–930.

42. Kim J, Park H. Fast nonnegative tensor factorization with an active-set-like method. In: High-Performance Scientific Computing. Springer; 2012. p. 311–326.

43. Gauvin L, Panisson A, Cattuto C. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. PloS one. 2014;9(1):e86028.

44. Bro R, Kiers HA. A new efficient method for determining the number of components in PARAFAC models. Journal of Chemometrics: A Journal of the Chemometrics Society. 2003;17(5):274–286.

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