HOCMO: A Tensor-based Higher-Order Correlation Model to Deconvolute Epigenetic Microenvironment in Breast Cancer

Charles Lu[1,2+], Rintsen Sherpa[2,3+], Liubou Klindziuk[2,4], Stefanie Kriel[1,2], Shamim Mollah[1,2*]

1 Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, MO, 63110, USA
2 Institute for Informatics, and Data Science, Washington University in St. Louis School of Medicine, St. Louis, MO, 63110, USA
3 Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, 48109, USA
4 Broad Institute, Cambridge, MA, 02142, USA

+These authors contributed equally to this work
* Corresponding author

* 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 these cellular processes. However, the mechanisms by which epigenetic microenvironment signals influence these cellular phenotypes are complex and currently not well established. To deconvolute the complexity of the epigenetic microenvironment signals in breast cancer, we developed a novel tensor-based correlation method: HOCMO (Higher-Order Correlation Model), applying to proteomics time series data to reveal the four-way regulatory dynamics among signaling proteins, histones, and growth-promoting ligands across multiple time points in the breast epithelial cells. HOCMO reveals two functional modules and the onset of specific protein-histone signatures in response to growth ligands contributing to distinct cellular phenotypes indicative of breast cancer initiation and progression. We evaluate robustness of our tensor model against baseline method TensorLy and achieved slight improvement in terms of reconstruction error and execution time. HOCMO is a data independent self-supervised learning method with superior interpretability that can capture the strength of complex interactions such as inter- and intra-pathway cellular signaling networks in any diseases or biological systems.

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 which encodes pairwise correlations between nodes with some mathematical formulation of their dynamics. Usual mathematical formulations of the dynamics typically assume that each node in the network has an associated "state" variable (representing, for example, expression of the corresponding signaling protein) as a function of the "time" variable. 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 connectivity among phosphoproteins, histones, and drugs using a hybrid machine learning approach[3]. In this study, we develop a higher-order (k >= 3-way) correlation model using tensor decomposition from measured proteomic signals. In this study we develop a ligands-signaling proteins-histones-time 4-way correlation network by constructing a signaling proteins-ligands-time 3-way tensor from protein expression data and a histones-ligands-time 3-way tensor from histone modification data to understand the influences of various growth ligands on cell signaling pathways. We developed a Non-negative Tensor Factorization (NTF)[4] based approach to capture the higher-order correlations among proteins, ligands and histones across multiple time points from the collected multi-omics data containing two time-series proteomic datasets. Unlike deep learning models, NTF learns the underlying complex representation of data in an unsupervised manner, offers better interpretability, scalability and data-efficiency when dataset is limited. Unlike deep learning models, NTF does not encounter the issue of being a black box model due to more transparent architecture. The overall workflow of our method consists of three components (Figure 1). First, we collect data from reverse phase protein array (RPPA) and global chromatin profiling (GCP) time series data and represent them as two 3D tensors. Then, we adopt an NTF based method to elicit the higher-order dynamic 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 to collectively reveal the significant interactions among proteins, ligands, and histones at the respective time states. We introduce a novel metric called Higher-Order

Correlation score (HOC-score) to quantify the correlation among ligands, phosphoproteins, and histones causing the chromatin structure changes.

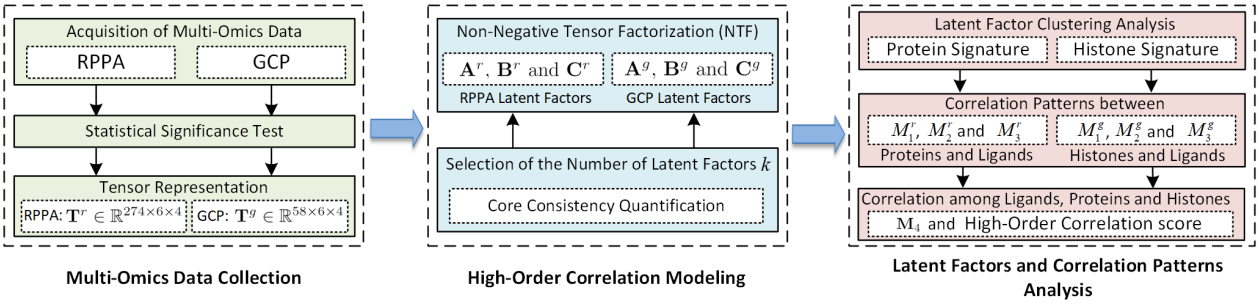


Figure 1. Workflow of the proposed approach. It involves three major components: i) multi-omics data preprocessing, ii) higher-order data correlation modeling using non-negative tensor factorization, and iii) interpretation of tensor factorization results.

A large-scale NIH 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 RPPA signaling proteins[5], and GCP[6]. 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 offers a method for obtaining key signatures of cellular response to growth ligands based on detailed mechanisms reconstructed from these datasets. 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 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. To evaluate our results from proteomics analyses, we further performed differential expression analysis and gene set enrichment analysis to capture gene activities at the transcriptomic level using RNAseq data.

STAR Methods

Multi-Omics Datasets

We utilized the RPPA and GCP datasets from LINCS to capture the dynamic regulations of growth ligands influencing phosphoproteins and histone modifications in the breast tumor microenvironment. We then used the RNAseq dataset from LINCS to further verify our phosphoprotein-histone signatures results with the respective transcriptional profiles.

RPPA and GCP Data Acquisition

We downloaded the level3 (log 2 normalized) RPPA data from the Synapse platform (https://www.synapse.org/#!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 growth promoting 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:syn18804314) 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 harmonize data and provide analyses that are consistent across the above two datasets, we removed all samples in the first hour from the RPPA data. This resulted in RPPA and GCP data consisting uniform four distinct time points of 4, 8, 24 and 48 hours. In both RPPA and GCP datasets, each data sample (e.g., protein or histone mark) has three replicates representing three independent experiments at each time point. We applied an unpaired Student's t-test to retain statistically significant (p-value<0.05) 274 proteins in RPPA and 58 histones in GCP 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[8] version 2.7.3 and output as BAM files. A count matrix of all samples and their genes

was generated using the featureCounts program[9] from the R package RSubread[10] version 2.6.1 and the ENSEMBL GRCh38 gene annotations.

Tensor Representation of RPPA and GCP Data

After performing the significant test and sample screening processes, the differential abundance (fold change, *i.e.*, the ratio of each sample measurement to its control measurement) of signaling proteins and histones relative to their respective controls (PBS) are calculated. Figure 2A illustrates the differential abundance of an example protein PIK3CA from RPPA data and an example histone mark H3K9me2S10ph1K14ac0 from the GCP data. For example, after treatment with 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.

The RPPA fold-change data at each time point i with 274 proteins and 6 ligands can be represented in a matrix $\mathbf{L}_i^r \in \mathbb{R}^{274\times 6}$, where the superscript r denotes RPPA data. By stacking all matrices observed at four sequential 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 $\mathbf{T}^r = \{\mathbf{L}_4^r, \mathbf{L}_8^r, \mathbf{L}_{24}^r, \mathbf{L}_{48}^r\} \in \mathbb{R}^{274\times 6\times 4}$. Similarly, The GCP fold-change data with 58 histones and 6 ligands can be represented as a 3D tensor $\mathbf{T}^g = \{\mathbf{L}_4^g, \mathbf{L}_8^g, \mathbf{L}_{24}^g, \mathbf{L}_{48}^g\} \in \mathbb{R}^{58\times 6\times 4}$ (with superscript g denoting GCP data), where $\mathbf{L}_j^g \in \mathbb{R}^{58\times 6}$ denotes the fold-change matrix at time point j. $\mathbf{T}^r$ and $\mathbf{T}^g$ are visualized in two 3-way tensors shown in Figure 2B (left) & (right) respectively.

The differential abundance analysis on RPPA and GCP data sets captures fold-change values reflecting negative and positive regulations of phosphoproteins and histone abundances incurred by the treatment of the six growth ligands. To satisfy the non-negativity constraint of the NTF method, we first transformed these fold-change values to their absolute values. Using these absolute fold-change values we then construct two tensors $\mathbf{T}^r = \mathbf{T}^{274\times 6\times 4}$ for RPPA representing fold-changes of 274 phosphoproteins and $\mathbf{T}^g = \mathbf{T}^{58\times 6\times 4}$ for GCP representing fold-changes of 58 histone modifications after the treatment of six growth ligands across four (4-, 8-, 24-, and 48-hour) time points (Figure 2B). We used these tensors as inputs for the subsequent non-negative tensor factorization steps.

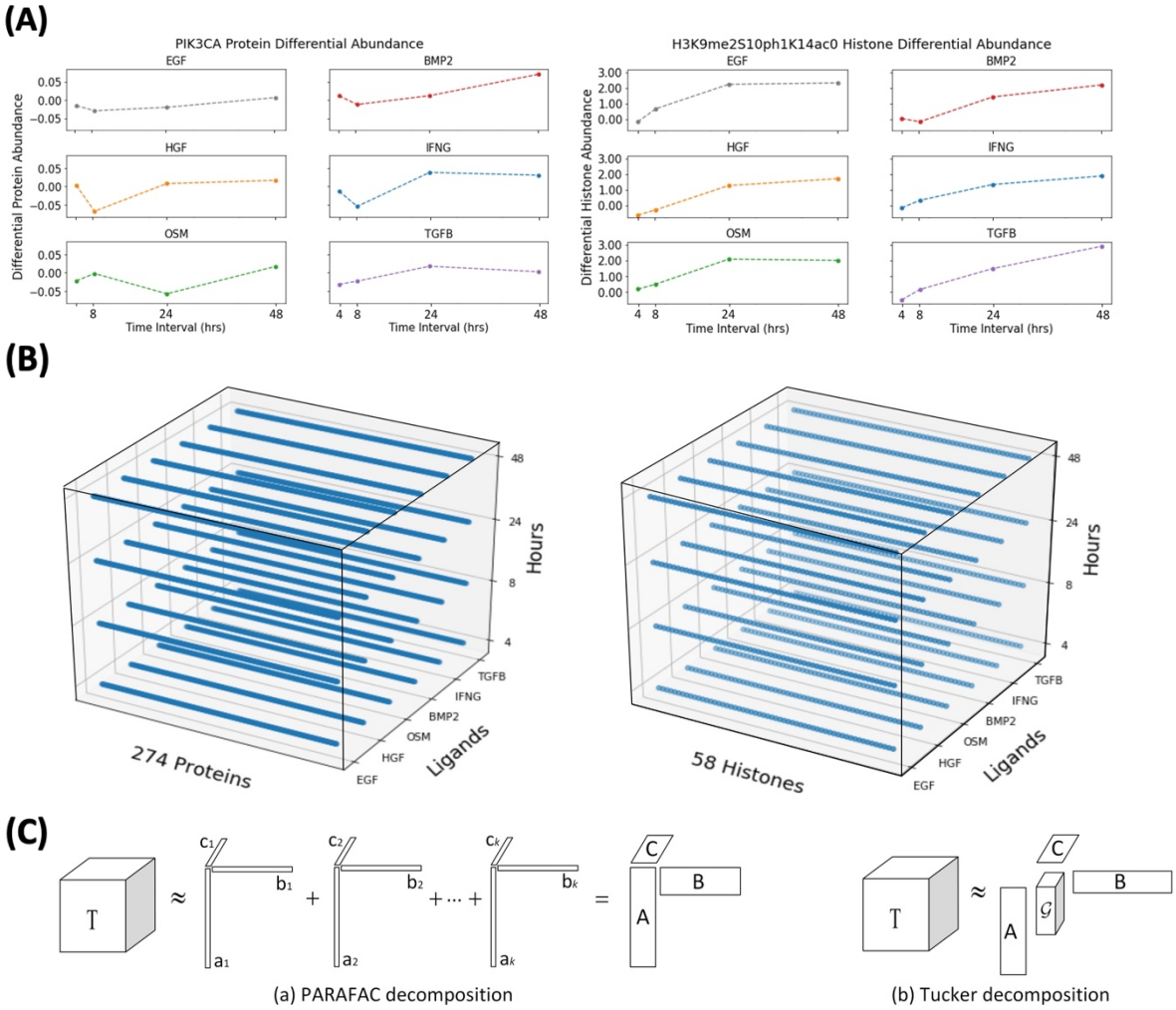


Figure 2. (A) Differential abundance changing over time. The differential abundances of protein PIK3CA and histone H3K9me2S10ph1K14ac0 after treatment with six growth ligands at various time points. **(B) Visualizations of the 3-way RPPA tensor and GCP tensor.** Each tensor has three dimensions or modes, protein, ligand and time (hour) for RPPA (left), and histone, ligand and time (hour) for GCP (right). **(C) 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.

Higher-Order Correlation Modeling

The tensors encode for the correlations or interactions between proteins and ligands (or between histones and ligands) changing over time (Figure 2B). In this section, we present a framework using NTF to model multi-layer correlations (4-way correlations), such as those among ligands, proteins, histones, and time shown above.

Non-Negative Tensor Factorization (NTF)

Tensor factorization can be viewed as the higher-order version of matrix factorization[4] by extending the two-dimensional matrices to tensors. A tensor is a multi-way array or multi-dimensional matrix[11] 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 $1 \leq i_n \leq I_N$.*

Tensors can capture the complex relations between a larger number of factors, such as the interplay between protein, ligand and time factors in RPPA data, with higher fidelity. In turn, 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)[12] and Tucker[13], which have shown promises in many real-world applications such as cancer survival prediction[14] and image processing[15].

PARAFAC and Tucker

The non-negative PARAFAC was proposed by Carroll *et al.*[16] and then developed by Bro[17] with a more efficient method to calculate the non-negative components. Figure 2C (a) 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:

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

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 approximating the actual rank of tensor $\mathbf{T}$. Data relationships in the original tensor ($\mathbf{T}$) are thereby fully preserved by the resulting components $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$. Alternatively, the PARAFAC decomposition can be represented as:

$$\mathbf{T} = \big[[\mathrm{A,B,C}]\big], \qquad (2)$$

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

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

Tucker decomposition is another popular NTF technique[13]. As shown in Figure 2C (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:

$$\begin{aligned} \mathbf{T} &\approx \mathcal{G} \underset{1}{\times} \mathbf{A} \underset{2}{\times} \mathbf{B} \underset{3}{\times} \mathbf{C} \\ &= \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} \sum_{r_3=1}^{k_3} g_{r_1 r_2 r_3}\, \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \end{aligned} \quad (4)$$

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.

Therefore, PARAFAC is a special case of Tucker with the super-diagonal core tensor $\mathcal{G}$. Studies[18,19] 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[20]. For these reasons, we chose to use non-negative PARAFAC to analyze and study the tumor microenvironment. There are many methods to implement the 3-way PARAFAC including Nonnegative CANDECOMP/PARAFAC (NCP)[21] and TensorLy[22]. In this study we chose to implement NTF using NCP for our RPPA and GCP tensors given its lower reconstruction error as shown in Figure 3A, and the importance of interpretability in a biological context. The NTF package as part of TensorLy was used to validate our 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[23]. 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:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \| \mathbf{T} - [[\mathbf{A}, \mathbf{B}, \mathbf{C}]] \|_F^2 \quad \text{subject to: } \mathbf{A}, \mathbf{B}, \mathbf{C} \geq 0, \quad (5)$$

In other words, NCP aims to solve latent factor matrices $\mathbf{A} \in \mathbb{R}^{I_1 \times k}$, $\mathbf{B} \in \mathbb{R}^{I_2 \times k}$ and $\mathbf{C} \in \mathbb{R}^{I_3 \times k}$ (k is the number of latent factors) to approximate the input tensor $\mathbf{T}$. In the following, we briefly describe the principle of solving $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ in NCP. For mathematical simplicity, 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:

$$\begin{aligned} &\mathbf{T}_{(n)} \approx \mathbf{A}^{(n)} \times \left(\mathbf{D}^{(n)}\right)^{\mathsf{T}}, \text{for } \forall n \in \{1, \cdots, N\} \\ &\text{with } \mathbf{D}^{(n)} = \mathbf{A}^{(N)} \circ \cdots \circ \mathbf{A}^{(n+1)} \circ \mathbf{A}^{(n-1)} \circ \cdots \circ \mathbf{A}^{(1)}, \end{aligned} \quad (6)$$

where $N = 3$ representing the 3-way tensor factorization 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:

$$\min_{\mathbf{A}^{(n)}} \| \mathbf{D}^{(n)} \times \left(\mathbf{A}^{(n)}\right)^{\mathsf{T}} - \left(\mathbf{T}_{(n)}\right)^{\mathsf{T}} \|_F^2 \text{,for } \forall n \in \{1, \cdots, N\} \quad \text{subject to: } \mathbf{A}^{(n)} \geq 0, \tag{7}$$

where we need to solve $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ and $\mathbf{A}^{(3)}$ in this paper. 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:

$$\min_{\mathbf{X} \geq 0} \| \mathbf{V}\mathbf{X} - \mathbf{W} \|_F^2, \tag{8}$$

where $\mathbf{V} \in \mathbb{R}^{P \times Q}$ corresponds to $\mathbf{D}^{(n)}$, $\mathbf{X} \in \mathbb{R}^{Q \times L}$ corresponds to $\left(\mathbf{A}^{(n)}\right)^{\mathsf{T}}$, and $\mathbf{W} \in \mathbb{R}^{P \times L}$ corresponds to $\left(\mathbf{T}^{(n)}\right)^{\mathsf{T}}$ where P, Q, L are integers. Then, the BPP algorithm with multiple right-hand sides is utilized to solve these subproblems.

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 to approximate the actual rank of the target tensor.

It is non-trivial to determine the best number of latent factors k, which typically necessitates multiple trials and fine-tuning[24]. The *core consistency* (CC) metric has been widely used to evaluate the fitness of a candidate k,[25] where the resulting latent components of an appropriate PARAFAC model are reflective of the trilinear variation of the data. Since the Tucker decomposition can represent the trilinear variation in data by the shared core tensor and various loading matrices, the definition of CC is naturally achieved by examining the difference between PARAFAC and Tucker based tensor decomposition. As demonstrated in the previous section, the PARAFAC decomposition is a special case of the Tucker decomposition (Eq.4) since Eq.1 can be rewritten as:

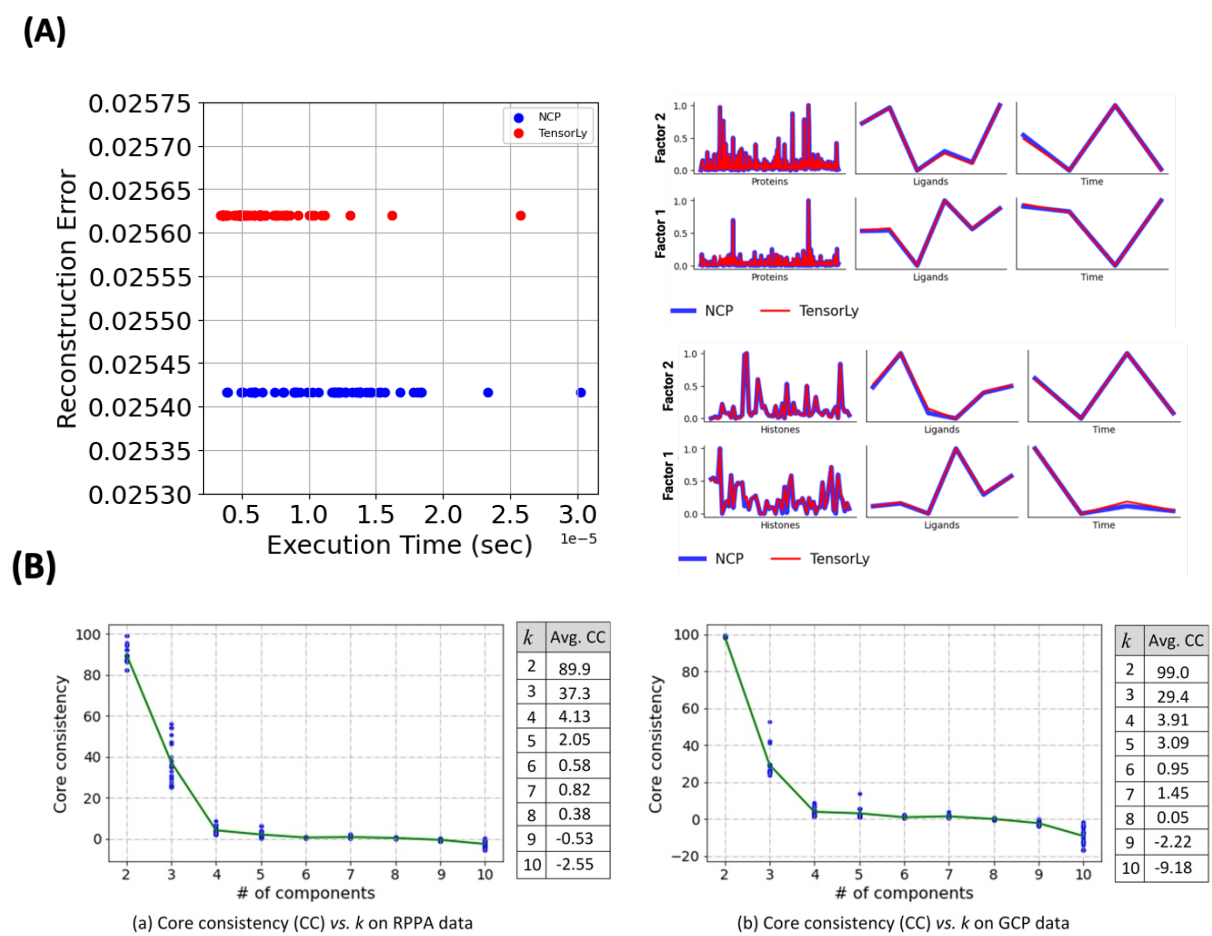
$$\mathbf{T} \approx \sum_{r_1=1}^{k} \sum_{r_2=1}^{k} \sum_{r_3=1}^{k} g^d_{r_1,r_2,r_3} \, \mathbf{a}_{r_1} \circ \mathbf{b}_{r_2} \circ \mathbf{c}_{r_3}, \tag{9}$$

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[25]. 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:

$$\mathrm{CC} = 100\left(1 - \frac{\sum_{r_1=1}^{k}\sum_{r_2=1}^{k}\sum_{r_3=1}^{k}(g_{r_1,r_2,r_3} - g^{d}_{r_1,r_2,r_3})^2}{k}\right). \quad (10)$$

The CC reflects how well a particular Tucker core fits the assumption of a good PARAFAC model for a given k. In order to accurately represent the data, typically a larger value of k is chosen such that a high CC value is yielded (max value = 100). A CC above 50 is interpreted as an acceptable model[24,25].

Instead of choosing the number of latent factors randomly, we leverage CC as the metric to determine the optimal number of latent factors k for RPPA and GCP tensors. Using Eq.10, we calculated CC values for k ranging from 2 to 10 (Figure 3B). We observed the best average core consistencies for RPPA and GCP tensors at 89.9 and 99.0 respectively when $k = 2$. Although a larger k would be desirable, the CC values for all $k > 3$ for both RPPA and GCP data are less than 50 (a value that is considered as a not acceptable model). 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

Figure 3. (A) *Evaluation* of *the* model. Factors derived from the tensor decomposition processes (TensorLy and NCP) were used to reconstruct a tensor. The mean squared error was measured between

this reconstructed tensor and our original tensor and plotted against the execution time of the tensor decomposition process (left). Comparison of the performance of NCP and TensorLy implementations using two factors to estimate protein, histone, ligands and time in RPPA and GCP data(right). **(B) The core-consistency vs. the number of components (k)**. The core consistency for each k tested on the RPPA tensor (left) and GCP tensor (right) was computed 100 times independently. The top 20% core consistencies (blue dots) with their averages (solid line) are shown in the left panel (graph). The actual values for average core consistencies (CC) are shown in the right panel (table).

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 tensor $\mathbf{T}^r \in \mathbb{R}^{274\times6\times4}$ and GCP tensor $\mathbf{T}^g \in \mathbb{R}^{58\times6\times4}$, respectively. The non-negative factorization on RPPA data based on NCP is summarized in Algorithm 1, which takes the 3-way RPPA tensor $\mathbf{T}^r$ as an input and then decomposes it to three latent factor matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ (r represents RPPA tensor). Nonnegative Double Singular Value Decomposition (NNDSVD)[26] is applied to initialize $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$, 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. Then the block principle pivoting (BPP) is applied to generate the final loading (latent factor) matrices $\mathbf{A}^r$, $\mathbf{B}^r$ and $\mathbf{C}^r$ (Algorithm 1).

Algorithm 1 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 $\mathbf{C}^r$.

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

The same algorithm is applied to decompose the GCP tensor $\mathbf{T}_g$ and derive the latent factor matrices $\mathbf{A}^g$, $\mathbf{B}^g$ and $\mathbf{C}^g$ (g represents GCP tensor).

Higher-Order Correlations Representation

To gain insights into the interactions between proteins and ligands in the RPPA data over time, we further constructed three auxiliary matrices through the combinations between $\mathbf{A}^r_1$, $\mathbf{B}^r_2$ and $\mathbf{C}^r_3$ as:

$$\mathbf{M}^r_1 = \mathbf{A}^r \cdot (\mathbf{B}^r)^{\mathsf{T}},\ \mathbf{M}^r_2 = \mathbf{A}^r \cdot (\mathbf{C}^r)^{\mathsf{T}},\ \mathbf{M}^r_3 = \mathbf{B}^r \cdot (\mathbf{C}^r)^{\mathsf{T}}, \quad (11)$$

where $\cdot$ denotes the matrix multiplication. $\mathbf{M}^r_1$ (Supplemental Table 1) stores the pairwise interaction intensities between 274 proteins and 6 ligands, while $\mathbf{M}^r_2$ (Supplemental Table 2) and $\mathbf{M}^r_3$ (Supplemental Table 3) store the pairwise interaction intensities of 4 time states with

the 274 proteins and 6 ligands respectively. To reveal the dynamic correlations between histones and ligands in the GCP data, we created three matrices as follows:

$$\mathbf{M}_1^g = \mathbf{A}^g \cdot (\mathbf{B}^g)^{\mathsf{T}},\ \mathbf{M}_2^g = \mathbf{A}^g \cdot (\mathbf{C}^g)^{\mathsf{T}},\ \mathbf{M}_3^g = \mathbf{B}^g \cdot (\mathbf{C}^g)^{\mathsf{T}}, \quad (12)$$

$\mathbf{M}_1^g$ (Supplemental Table 4) captures the pairwise interaction intensities between the 58 histones and 6 ligands, while $\mathbf{M}_2^g$ (Supplemental Table 5) and $\mathbf{M}_3^g$ (Supplemental Table 6) store the pairwise interaction intensities of 6 time states with 58 proteins and 6 ligands respectively. Further, based on the two matrices $\mathbf{M}_1^g$ and $\mathbf{M}_1^r$, *i.e.*, we constructed $\mathbf{M}_4$ (Supplemental Table 7) that correlates histones and proteins from the RPPA and GCP data to represent their pairwise correlations by:

$$\mathbf{M}_4 = \mathbf{M}_1^g \cdot (\mathbf{M}_1^r)^{\mathsf{T}} \quad (13)$$

In this study, we use $\mathbf{M}_4$ (e.g., positive values) to quantify the relative interaction or correlation intensity between proteins and histones after the treatment of six growth ligands. The directions (e.g., either positive or negative) of protein and histone regulations incurred by the growth ligands at respective time points can be obtained from the indicator tensors $I^r = \{-1,1\}^{274\times6\times4}$ for RPPA and $I^g = \{-1,1\}^{58\times6\times4}$ for GCP. To better understand the epigenetic tumor environment of breast tissue, we aim to elucidate the regulatory mechanisms governing histone modifications mediated by ligand-induced phosphoproteins and their consequential impact on changes in chromatin structure. To quantify the correlation amongst ligands, phosphoproteins, and modifying histones, we define a Higher-Order Correlation (HOC) score for any triplet relation <i, j, k> representing the ith ligand, j^{th} phosphoprotein and kth histone as:

$$HOC - score(i,j,k) = p_{ij} \cdot q_{jk} \quad (14)$$

where p_{ij} is the interaction intensity between the ith ligand and jth protein stored in $\mathbf{M}_1^r$, and q_{jk} is the interaction intensity between the jth protein and kth histone stored in $\mathbf{M}_4$. As the value of HOC approaches 1, the strength of correlation among triplets at a specific time point increases.

Gene Set Enrichment Analysis

For validation, we performed a gene set enrichment analysis to find out which canonical pathways associated with the significant phosphoproteins identified from NTF 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 for each individual ligand at a specific time point compared to PBS at the same time point. GSEA[27] 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.

Results

HOCMO captures histone signatures of tumor microenvironment (TME) structures

Latent Factor Analysis: In order to find patterns within the protein and histone modification datasets, we decomposed the RPPA and GCP tensors into their latent factors. These latent factors were mapped onto a shared two-dimensional space and represented by matrices for each variable (Figure 4A). We observed variable affinities towards the two latent components within each of the matrices. Distinct groups can be easily seen. For both RPPA (Figure 4A a $\boldsymbol{C}^r$) and GCP (Figure 4A b $\boldsymbol{C}^g$) datasets, the 4, 8, and 24 hour time states appear to map closely to the second latent factor, while the 48hr time state has a greater affinity for the first latent factor. In the histone latent matrix (Figure 4A b $\boldsymbol{A}^g$), two groups of histones can be seen separated by the latent factors. However, not all affinities are immediately clear from the latent matrices, and in order to properly define groups of related proteins and histones, we must further perform clustering analysis.

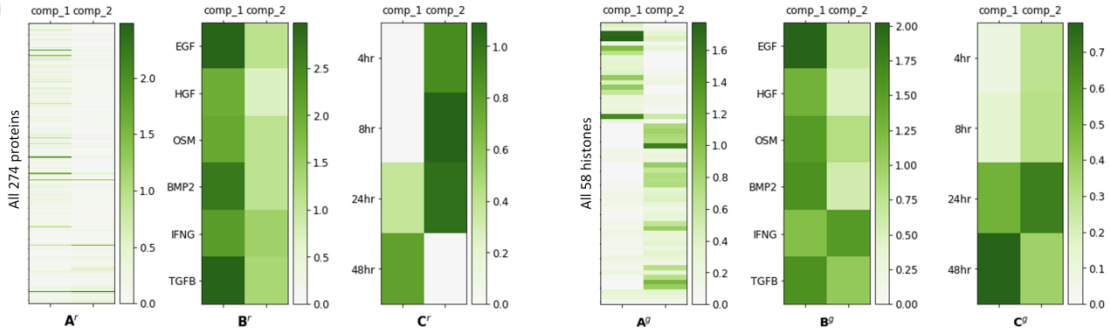
Distinct clusters for phosphoproteins and histone modification signatures through latent structure: To visualize related groups of proteins and histones based on their latent factors, we clustered the protein latent matrix $\mathbf{A}^r$ (Figure 4A a) and the histone latent matrix $\mathbf{A}^g$ (Figure 4A b). Each protein/histone was assigned to a cluster based on which of the two latent factors it had a greater affinity towards, resulting in two clusters within proteins (Figure 4B a) and histones (Figure 4B b). In the tumor microenvironment, the protein signatures can be interpreted as the categorization of proteins demonstrating similar responses to the growth ligands. Similarly, histones in the same signature may correspond to co-occurring histone modifications in response to the six growth ligands. In addition to the above hard clustering, we applied a Softmax transformation, visualized as a heatmap, which reveals the probability distributions of different proteins/histones belonging to each cluster.

Correlation patterns between proteins and ligands and histones and ligands reveal potential onset at 24th hour

To further quantify the interactions between proteins and ligands, as well as the variations of their interactions over time, we constructed pairwise interaction matrices for each of the three pairs between proteins, ligands, and time states (Figure 4C top panel, Supplemental Table 8) based on their latent factor matrices (Figure 4A a). We can see that proteins are variably regulated by the six growth ligands (Figure 4C top panel $\mathbf{M}_1^r$), *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. Additionally, we see that proteins are regulated by ligands at different time states, where most proteins and ligands are significantly active at the 24-hour time point (Figure 4C top panel $\mathbf{M}_2^r$ and $\mathbf{M}_3^r$). Similarly, we constructed pairwise interaction matrices for each of the three pairs between histones, ligands, and time states (Figure 4C bottom panel, Supplemental Table 9) based on their latent factor matrices (Figure 4A a). We can see that some histones are significantly regulated by the respective ligands (Figure 4C bottom panel $\mathbf{M}_1^g$). We also see parallels to correlations found in the protein data, as histones and ligands

appear to be significantly active after the 24-hour time point (Figure 4C bottom panel $\mathbf{M}_2^g$ and $\mathbf{M}_3^g$).

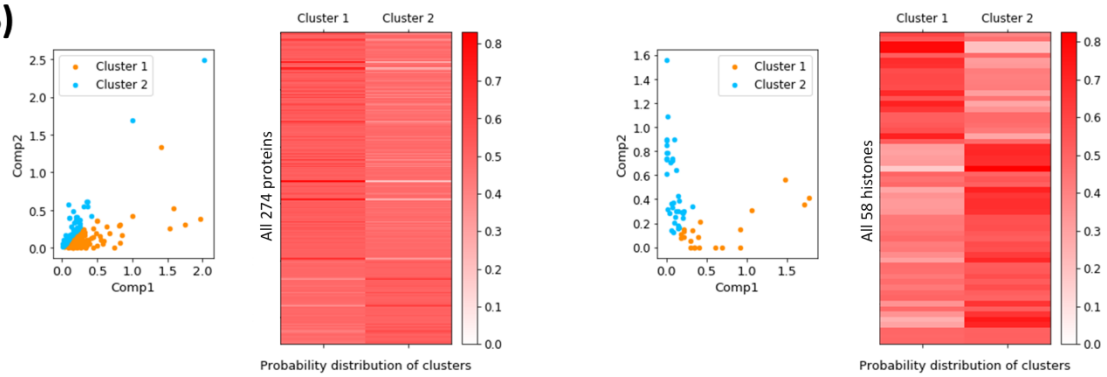
(A)



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

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

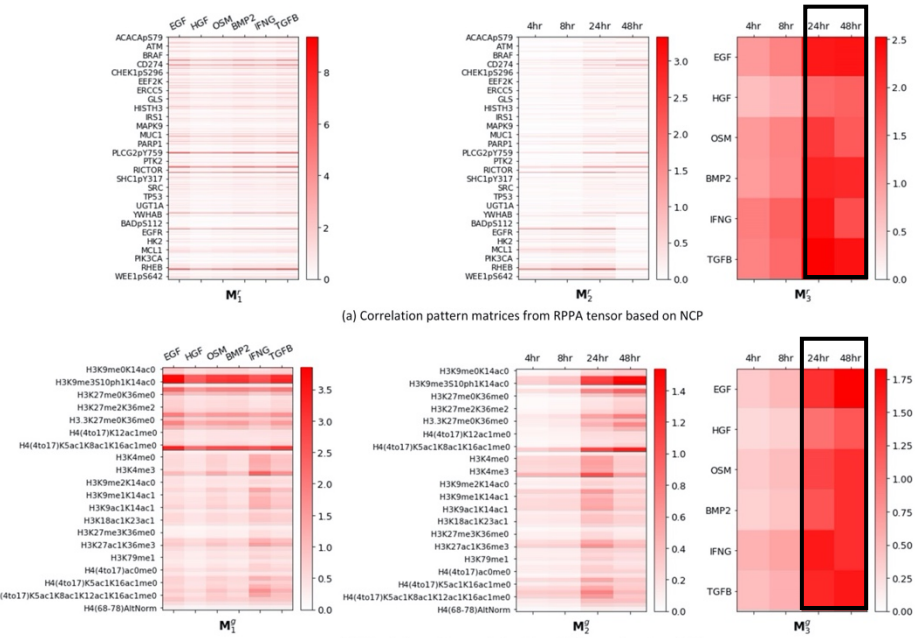
(B)



(a) Protein signatures

(b) Histone signatures

(C)



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

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

(D)

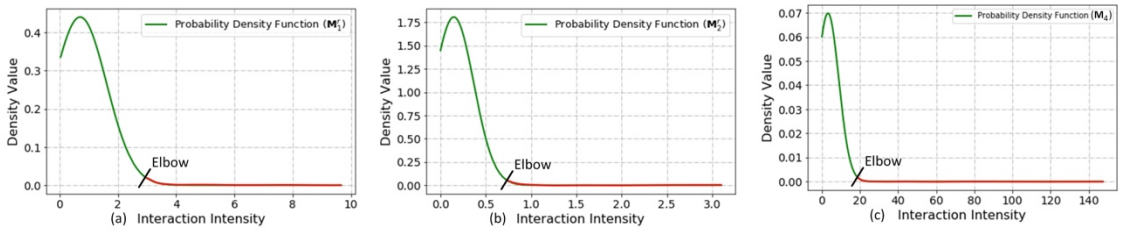


Figure 4. **(A) Latent factors generated from NTF.** (left) A^r, B^r and C^r represent proteins, ligands and hours from RPPA that are mapped to a two-dimensional latent space. Rows in A^r, B^r and C^r represent all 274 proteins, 5 ligands and 4 time states from the RPPA data respectively, and columns represent the two shared latent factors obtained by NTF. (right) A^g, B^g and C^g represent histones, ligands and hours from GCP that are mapped to a two-dimensional latent space. Rows in A^g, B^g and C^g represent all 58 histones, 5 ligands and 4 time states from the GCP data respectively, and columns represent the two shared latent factors obtained by NTF. **(B) The identified (left) protein signatures and (right) histone signatures**. We performed clustering analysis based on the factor matrices $\boldsymbol{A}^r$ and $\boldsymbol{A}^g$ respectively. 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 heatmaps in (left) and (right) show the actual probability distributions of each protein and histone to the respective clusters, where rows contain 274 proteins (only 28 protein labels shown) and 58 histones (only 20 histone labels shown) respectively. The protein signatures can be interpreted as the categorization of proteins demonstrating similar responses after treated with six growth ligands. Similarly, histones in the same cluster demonstrate similar responses after treated with these same growth ligands. **(C) The generated pattern matrices from the latent factor matrices.** Rows in M_1^r and M_2^r represent all 274 proteins, while their columns represent 6 ligands and 4 time states respectively (top panel). M_1^r reveals the correlation patterns between 274 proteins and six ligands, where the onsets are revealed in M_2^r and M_3^r over the 4 time points in the RPPA data. (bottom) Rows in M_1^g reveals the correlation patterns between 58 histones and 6 ligands, where the onsets are revealed in M_2^g and M_3^g over the 4 time points in the GCP data. **(D) The distribution of intensity values for correlation pattern matrices $\boldsymbol{M_1^r}$ and $\boldsymbol{M_2^r}$.** A density distribution learning is performed on (a) M_1^r between signaling proteins and growth ligands, (b) M_2^r between signaling proteins and time points, and (c) M_4 between signaling proteins and histones. Correlations with interaction intensities above the elbow are considered to be significant.

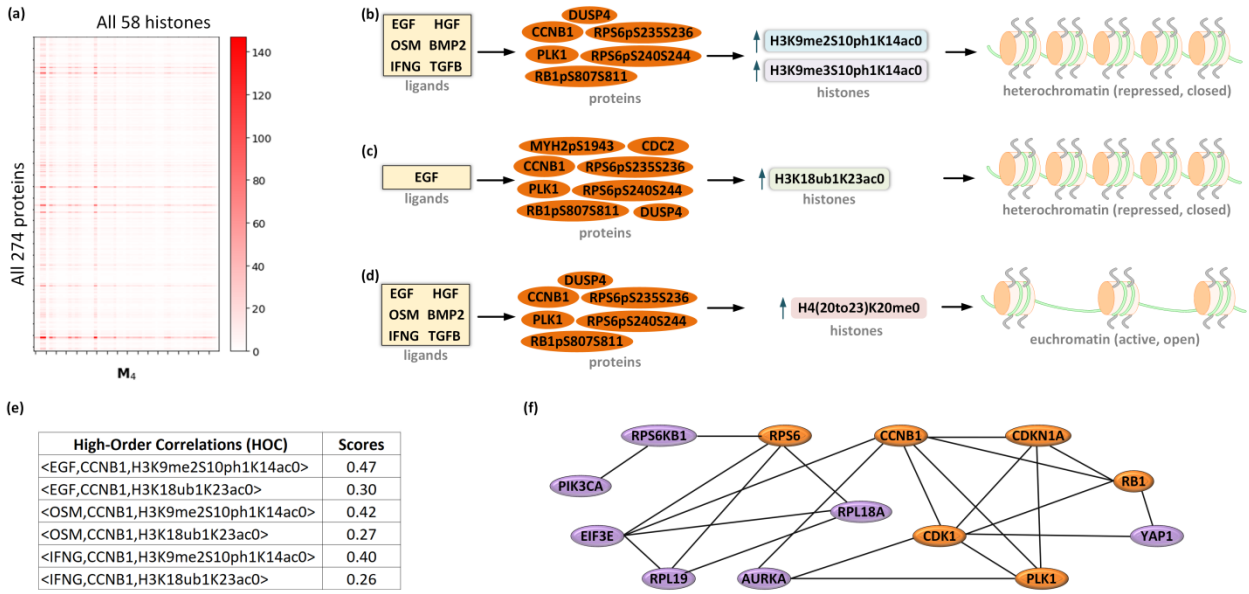
Identify significant interactome of ligands, signaling proteins, and histones triplets

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 (Figure 4C top panel $\mathbf{M}_1^r$ and Figure 4C bottom panel $\mathbf{M}_1^g$). We are interested in investigating the most significant interactions and the respective driver proteins, histones, and ligands. First, to identify significantly interacting proteins and histones, we performed a density distribution learning (e.g., fitting a Gaussian distribution, Figure 4D (a)) over the intensity values in the interaction matrix between proteins and histones (Figure 4C top panel $\mathbf{M}_1^r$). Note that since all the original values in the correlation matrices are non-negative, the fitted curve presents as a truncated Gaussian distribution displaying only the non-negative portion. We chose a cut-off value at the elbow as the threshold to determine significance, which has been a widely used method to decide the threshold in existing works[28,29]. If the interaction intensity between a protein and a corresponding ligand is larger than the threshold, we then consider that the protein has been regulated by the respective ligand. By considering the above two factors, we finally filtered out eight significantly regulated 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. We then perform a similar density distribution learning (Figure 4D (b)) on the interaction matrix between

proteins and time states (Figure 4C top panel $\mathbf{M}_2^r$) and found that the above eight phosphoproteins are all significant after the 24-hour time point.

Similarly, to determine significantly active histone marks in response to the growth ligands, we once again performed density distribution learning, this time over all intensity values in the interaction matrix between histones and ligands (Figure 4C bottom panel $\mathbf{M}_1^g$), as well as the interaction matrix between histones and time states (Figure 4C bottom panel $\mathbf{M}_2^g$). 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 significantly 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 significantly active at the 24th and 48th hour time points based on the intensity distributions on the respective interaction matrices (Figure 4C top panel $\mathbf{M}_3^r$ and Figure 4C bottom panel $\mathbf{M}_3^g$). 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 interactions between proteins and ligands (Figure 4C top panel $\mathbf{M}_1^g$) together with the interactions between histones and ligands (Figure 4C bottom panel $\mathbf{M}_1^r$), we constructed an additional matrix ($\mathbf{M}_4$: Figure 5a) to represent the pairwise interactions between all 274 proteins and 58 histones. We then performed a Gaussian distribution fitting the density distribution of the interaction intensities, where all interactions with coefficients over the cut-off value at the elbow on the curve were considered as significant (Figure 4D (c)). This resulted in eight significant proteins that have contributed to the modification of the five identified significant histones. In order to verify the above correlations, we also performed NCP and TensorLy based tensor factorization, which have produced the same results.



High-Order Correlations (HOC)	Scores
<EGF,CCNB1,H3K9me2S10ph1K14ac0>	0.47
<EGF,CCNB1,H3K18ub1K23ac0>	0.30
<OSM,CCNB1,H3K9me2S10ph1K14ac0>	0.42
<OSM,CCNB1,H3K18ub1K23ac0>	0.27
<IFNG,CCNB1,H3K9me2S10ph1K14ac0>	0.40
<IFNG,CCNB1,H3K18ub1K23ac0>	0.26

Figure 5. Significant signaling proteins and histones interactions. (a) Showing the HOC-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 examples of the higher-order correlation (HOC) scores for six triplet relations among ligands, proteins and histones. (f) Showing the protein-protein interaction network among six enriched proteins (orange, obtained from HOCMO) and seven inferred proteins (purple) obtained from STRINGdb.

Finally, we summarized the identified correlations between proteins and histones after the treatment of growth ligands in three groups as hypothetical pathways: 1) Six proteins (CCNB1, PLK1, RB1pS807S811, RPS6pS240S244, DUSP4 and RPS6pS235S236) were significantly regulated after treatment with the six ligands, which then led to the modifications of the two repressive histone marks H3K9me2S10ph1K14ac0 and H3K9me3S10ph1K14ac0 resulting a repressed chromatin state (e.g., closed status of chromatin) (Figure 5b); 2) The growth ligand EGF regulated all eight proteins, which then mediated the active histone mark H3K18ub1K23ac0 which led to a closed chromatin conformation (Figure 5c); 3) Six growth ligands regulated six proteins (CCNB1, PLK1, RB1pS807S811, RPS6pS240S244, DUSP4 and RPS6pS235S236) which mediated histone mark H4(20to23)K20me0 resulting in an open chromatin conformation (Figure 5d). In order to further quantify these chain-based regulation, we developed a metric called Higher-Order Correlation score (HOC-score) which considered the correlation among ligands, proteins, and histones at a given time point. HOC-score reveals how much the respective interaction among ligands, proteins and histones contributes to the chromatin structure changes (Figure 5e). We further inferred interactive proteins of the identified seven enriched proteins using STRINGdb[30] (Figure 5f).

Identify signaling networks that controls cellular phenotypes: mapping to canonical signaling pathway

In order to contextualize these findings, we further mapped the eight enriched proteins and phosphoproteins to their corresponding canonical breast cancer signaling pathways using the

KEGG and Reactome databases[31] (Figure 6). We observed strong upregulation of all these eight proteins and phosphoproteins across all ligands after the 24-hour mark. Differential expression within these canonical signaling pathways infers disruption of healthy cellular function.

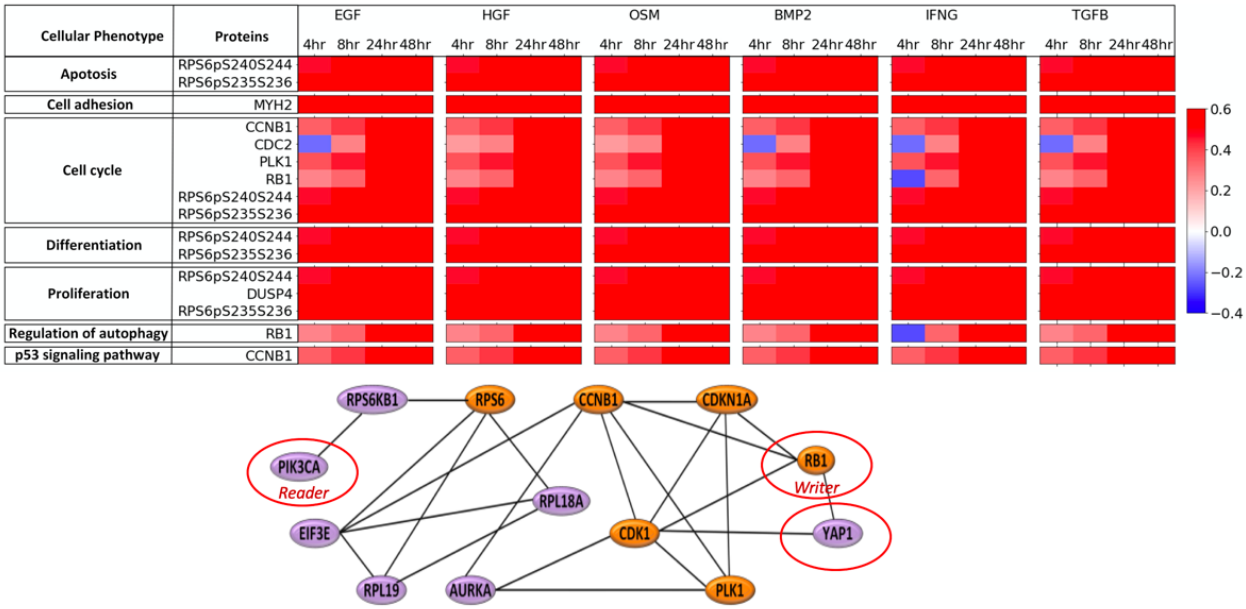


Figure 6. Canonical signaling pathways. Ligands specific signaling pathway profile associated with the enriched proteins at various time points, where positive and negative fold changes are indicated by red and blue colors, respectively. Showing the protein-protein interaction network among five enriched proteins (orange, obtained based on NTF) and seven inferred (purple) proteins.

Transcriptomic expression analyses show enrichment of cell cycle pathway and corroborate with proteomics results

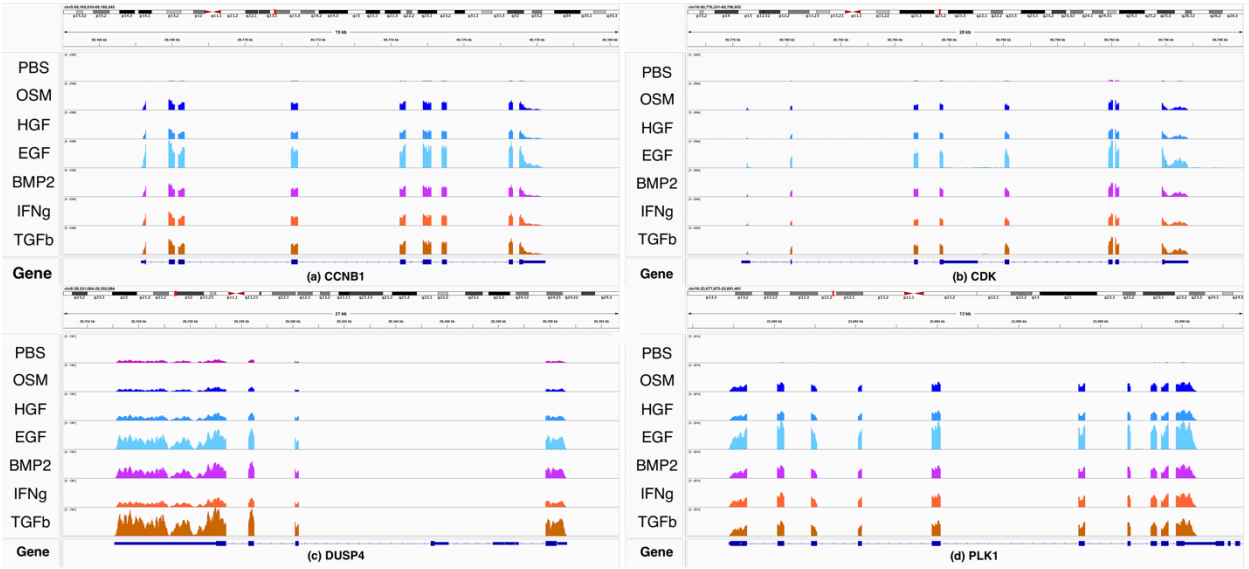
Differential analysis: To provide validation for the ligand-protein correlations produced by the HOCMO, we performed differential expression analysis on RNAseq data to observe the changes in gene expression caused by the growth ligands. We observed CCNB1, CDC2 (also known as CDK1), PLK1, and DUSP4 were significantly ($p < 0.05$) upregulated across all ligands at 24 hours (Figure 7A). We displayed the expression coverage of each ligand at the loci of these genes, where a higher coverage reflects greater gene expression (Figure 7B). We see that coverage is noticeably higher in the ligand-treated samples compared to PBS control. The upregulation of these genes corroborates with the HOCMO results, which found the proteins encoded by these

genes to be correlated with each of the ligands. However, further work must be done in order to completely validate the results of the tensor.

(A)

	EGF		HGF		OSM		BMP2		IFNG		TGFB	
Gene	**Log2 Fold Change**	**p-value**	**Log2 Fold Change**	**p-value**	**Log2 Fold Change**	**p-value**	**Log2 Fold Change**	**p-value**	**Log2 Fold Change**	**p-value**	**Log2 Fold Change**	**p-value**
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

(B)



(C)

(a) **EGF at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_SPLICEOSOME	2.677995	<0.0001
KEGG_CELL_CYCLE	2.611984	<0.0001
KEGG_DNA_REPLICATION	2.456621	<0.0001
KEGG_MISMATCH_REPAIR	2.254064	<0.0001
KEGG_PYRIMIDINE_METABOLISM	2.220961	<0.0001

(b) **HGF at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_SPLICEOSOME	2.216762	<0.0001
KEGG_CELL_CYCLE	2.176947	<0.0001
KEGG_DNA_REPLICATION	2.152190	<0.0001
KEGG_PYRIMIDINE_METABOLISM	1.945159	<0.0001
KEGG_MISMATCH_REPAIR	1.941910	<0.0001

(c) **OSM at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_SPLICEOSOME	2.521423	<0.0001
KEGG_CELL_CYCLE	2.501311	<0.0001
KEGG_DNA_REPLICATION	2.373424	<0.0001
KEGG_HOMOLOGOUS_RECOMBINATION	2.078916	<0.0001
KEGG_PYRIMIDINE_METABOLISM	2.051135	<0.0001

(d) **BMP2 at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_SPLICEOSOME	2.405551	<0.0001
KEGG_DNA_REPLICATION	2.345687	<0.0001
KEGG_CELL_CYCLE	2.312492	<0.0001
KEGG_HOMOLOGOUS_RECOMBINATION	2.006806	<0.0001
KEGG_MISMATCH_REPAIR	1.985067	<0.0001

(e) **IFNG at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	2.393632	<0.0001
KEGG_SPLICEOSOME	2.359967	<0.0001
KEGG_CELL_CYCLE	2.221906	<0.0001
KEGG_DNA_REPLICATION	2.200284	<0.0001
KEGG_PROTEASOME	2.183063	<0.0001

(f) **TGFB at 24 hours vs. PBS at 24 hours**

KEGG Pathway	NES	FDR q-val
KEGG_DNA_REPLICATION	2.195353	<0.0001
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	1.941262	0.004691
KEGG_CELL_CYCLE	1.933115	0.003127
KEGG_GAP_JUNCTION	1.784638	0.022273
KEGG_HOMOLOGOUS_RECOMBINATION	1.758243	0.025389

Figure 7: Transcriptomic analyses. (A) Table showing differential expression analysis results of identified enriched (statistically significant, p-value < 0.05) genes treated with ligands at 24 hours. (B) 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. (C) Top signaling pathways associated with these enriched genes using KEGG pathways.

Gene Set Enrichment Analysis: We further performed gene set enrichment analysis on the transcriptomic (RNAseq) data to find out which pathways were most affected by each ligand treatment. We identified top 5 enriched KEGG pathways for each ligand at 24 hours (Figure 7C). 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 signaling proteins from the HOCMO are key players in mitosis and cell proliferation.

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 method, HOCMO, 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, 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 evaluation on our NCP tensor reconstruction, time, and factors shows slightly better performance than the baseline TensorLy (Figure 3A). The result shows the possible onset of breast cancer initiation at 24 hours of treatments for all growth ligands (Figure 4C). 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 Apoptosis, cell adhesion, cell cycle, differentiation, proliferation, regulation of autophagy and p53 signaling pathway (Figure 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 (Figure 5b). We also observe all ligands on these same phosphoproteins regulating the H4(20to23)K20me0 mark to an open conformation

(Figure 5d). From protein-protein interaction analyses (Figure 5f), 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 (Figure 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[32–36]. 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[37], and cyclin expression levels are even used to subtype breast cancers[38]. 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[39–41]. 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[42] 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.

Code Availability

Code is available on GitHub at https://github.com/smollahlab/HOCMO. This includes the code for the HOCMO package available on PyPi, as well as the analysis of RPPA and GCP LINCS data using HOCMO.

Acknowledgments

This work was supported by Washington University in St. Louis School of Medicine Genetics departmental fund. We want to thank NIH LINCS consortium for making these data available, and Min Shi for helping with figure generation.

Author Contributions

S.M. conceived and designed the project and came up with the conception.
C.L. K.L, and S.M implemented HOCMO. R.S. performed transcriptomic analysis. S.K. performed GSEA analysis. R.S, S.M. performed data interpretation. S.M. wrote the manuscript and supervised the research.

References

1. Jeschke, J., Collignon, E. & Fuks, F. DNA methylome profiling beyond promoters - taking an epigenetic snapshot of the breast tumor microenvironment. *FEBS Journal* **282**, 1801–1814 (2015).
2. Walker, C., Mojares, E. & del Río Hernández, A. Role of Extracellular Matrix in Development and Cancer Progression. *Int J Mol Sci* **19**, 3028 (2018).
3. Mollah, S. A. & Subramaniam, S. Histone Signatures Predict Therapeutic Efficacy in Breast Cancer. *IEEE Open J Eng Med Biol* **1**, 74–82 (2020).
4. Lee, D. H. S. S. Algorithms for non-negative matrix factorization. in *Advances in neural information processing systems* (2000).
5. Huang, S. & Creighton, C. Reverse phase protein arrays in signaling pathways: a data integration perspective. *Drug Des Devel Ther* 3519 (2015) doi:10.2147/DDDT.S38375.
6. Creech, A. L. *et al.* Building the Connectivity Map of epigenetics: Chromatin profiling by quantitative targeted mass spectrometry. *Methods* **72**, 57–64 (2015).
7. Ruxton, G. D. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology* **17**, 688–690 (2006).
8. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
9. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
10. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, e47–e47 (2019).
11. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S. *Nonnegative Matrix and Tensor Factorizations*. (Wiley, 2009). doi:10.1002/9780470747278.
12. Carroll, J. D. & Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283–319 (1970).
13. Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311 (1966).
14. Fang, J. Tightly integrated genomic and epigenomic data mining using tensor decomposition. *Bioinformatics* **35**, 112–118 (2019).
15. Zhou, P., Lu, C., Lin, Z. & Zhang, C. Tensor Factorization for Low-Rank Tensor Completion. *IEEE Transactions on Image Processing* **27**, 1152–1163 (2018).

16. Carroll JD, D. S. G. P. S. Fitting of the latent class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints. *Multiway data analysis* 463–472 (1989).
17. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171 (1997).
18. Kim, Y.-D. & Choi, S. Nonnegative Tucker Decomposition. in *2007 IEEE Conference on Computer Vision and Pattern Recognition* 1–8 (IEEE, 2007). doi:10.1109/CVPR.2007.383405.
19. Karami, A., Yazdi, M. & Mercier, G. Compression of Hyperspectral Images Using Discerete Wavelet Transform and Tucker Decomposition. *IEEE J Sel Top Appl Earth Obs Remote Sens* **5**, 444–450 (2012).
20. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Review* **51**, 455–500 (2009).
21. 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* **58**, 285–319 (2014).
22. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. (2016).
23. Kim, J. & Park, H. Fast Nonnegative Tensor Factorization with an Active-Set-Like Method. in *High-Performance Scientific Computing* 311–326 (Springer London, 2012). doi:10.1007/978-1-4471-2437-5_16.
24. Gauvin, L., Panisson, A. & Cattuto, C. Detecting the Community Structure and Activity Patterns of Temporal Networks: A Non-Negative Tensor Factorization Approach. *PLoS One* **9**, e86028 (2014).
25. Bro, R. & Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *J Chemom* **17**, 274–286 (2003).
26. Boutsidis, C. & Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit* **41**, 1350–1362 (2008).
27. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
28. Shi, C. *et al.* A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J Wirel Commun Netw* **2021**, 31 (2021).
29. Antunes, M., Ribeiro, J., Gomes, D. & Aguiar, R. L. Knee/Elbow Point Estimation through Thresholding. in *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)* 413–419 (IEEE, 2018). doi:10.1109/FiCloud.2018.00066.
30. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res **47**, D607-D613 (2019). doi: 10.1093/nar/gky1131.
31. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691–D697 (2011).
32. 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* **24**, 1809–1819 (2005).

33. 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* **39**, 635–644 (2004).
34. Lv, S. *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. *Int J Biochem Cell Biol* **128**, 105844 (2020).
35. Guo, M. *et al.* miR-656 inhibits glioma tumorigenesis through repression of BMPR1A. *Carcinogenesis* **35**, 1698–1706 (2014).
36. Wang, C. *et al.* Suppression of motor protein KIF3C expression inhibits tumor growth and metastasis in breast cancer by inhibiting TGF-β signaling. *Cancer Lett* **368**, 105–114 (2015).
37. Aaltonen, K. *et al.* High cyclin B1 expression is associated with poor survival in breast cancer. *Br J Cancer* **100**, 1055–1060 (2009).
38. Agarwal, R. *et al.* Integrative Analysis of Cyclin Protein Levels Identifies Cyclin B1 as a Classifier and Predictor of Outcomes in Breast Cancer. *Clinical Cancer Research* **15**, 3654–3662 (2009).
39. Michowski, W. *et al.* Cdk1 Controls Global Epigenetic Landscape in Embryonic Stem Cells. *Mol Cell* **78**, 459-476.e13 (2020).
40. Zeng, X., Chen, S. & Huang, H. Phosphorylation of EZH2 by CDK1 and CDK2. *Cell Cycle* **10**, 579–583 (2011).
41. Wierer, M. *et al.* PLK1 Signaling in Breast Cancer Cells Cooperates with Estrogen Receptor-Dependent Gene Transcription. *Cell Rep* **3**, 2021–2032 (2013).
42. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).