Specific Aims — Computational & Predictive Modeling Core (CPM)

Methicillin-resistant *Staphylococcus aureus* (MRSA) and *Candida albicans* (CA) are common commensal organisms and the most common life-threatening bloodstream infections, arising from both community-acquired and healthcare-associated settings. Despite the existence of effective therapies, both infections are associated with poor outcomes[1,2]. Treatment outcomes are poorly predicted by either *in vitro* drug effectiveness or the pathogen's genetic features, indicating that outcomes are primarily driven either by features of the host response or pathogen-host interactions.

Recent progress has been made in identifying the determinants of outcome in both infections. Several demographic and clinical factors are weakly associated with outcome but are insufficiently predictive in individual patients[1,2]. Based upon the hypothesis that outcome is determined by the confluence of host, pathogen, and antibiotic features, we have previously used broad molecular profiling to identify the molecular variation in infection response. This systems-level analysis previously identified genetic[3,4], methylation[5], transcriptional, and cytokine[4,6,7] predictors of MRSA persistence. However, the extent to which these signatures operate as distinct molecular mechanisms of immune response, or reflect shared underlying molecular mechanisms, is yet unclear. We expect that pan-signature patterns might also improve predictions of outcome if they help refine our view of shared molecular mechanisms. 20% of all CA bloodstream infections are polymicrobial with *Staphylococcus aureus* as a common co-isolated organism[8]. Thus, one might expect there are either host immunologic deficiencies or commonalities in immune evasion mechanisms that create a shared permissive niche.

Dimensionality reduction is an essential tool in identifying patterns across datasets. These techniques, such as principal component analysis, visualize variation, reduce noise, impute missing values, and reduce dimensionality. When integrating data across additional modes (e.g., subjects, genes, and time in a longitudinal RNAseq study) higher-order tensor-based dimensionality reduction methods exist with the same benefits[9–12]. These methods are naturally suited to combining different types of measurements, as each dataset mode (e.g., each subject) is effectively isolated[10,12]. Tensor-based dimensionality reduction supplies enhanced benefits when data can be organized in tensor form by reducing the data to a greater extent, improving factor interpretation by isolating effects along each mode, and effectively handling confounding factors like batch effects[10,12,13]. However, tensor-based dimensionality reduction methods currently have significant limitations in the scope of their use. Most critically, it must be possible to perfectly align measurements along each dataset mode which is not feasible in many situations such as longitudinal studies[9].

A central focus of the computational core will be cross-project data integration. We will start by focusing on developing new methodologies for longitudinal data dimensionality reduction and data integration using pre-existing profiling in CA infection. Next, we will identify common patterns across MRSA and CA infection subjects. Finally, we will coordinate closely with each project to identify how mechanisms identified in model systems are reflected in the clinical cohorts. These data integration efforts will allow us to test a series of hypotheses about the nature of CA progression, similarities in the determinants of MRSA and CA infection and outcome, and how signatures of immunologic and pathogen response from the mechanistic studies are reflected in the human infection samples.

Aim 1: Evaluate the integrated longitudinal determinants of CA outcome

Hypothesis: CA infection outcome is associated with distinct immunological trajectories.

a) Integrate proteomics and gene expression through a continuous mode coupled tensor decomposition
b) Express longitudinal dynamics as a dynamical system to infer regulatory component interactions
c) Test for association between CA immunological trajectories and mortality

Aim 2: Identify shared and distinct molecular and cellular patterns across MRSA and CA

Hypothesis: Shared molecular patterns exist across molecular measurement and infection types.

a) Improve factorization of clinical cytokine panels
b) Integrate MRSA and CA transcriptomic, methylation, and cytokine measurements
c) Identify pattern associations with age, sex, infection outcome, and isolate properties

Aim 3: Integrate outcomes from experimental models and human infection

Hypothesis: Deep murine model & isolate response profiling will identify mechanisms found in human infection.

a) Identify interactions between MRSA epigenetic plasticity & murine/human immune responses
b) Characterize interactions between CA phenotypic and immunologic response features
c) Infer the effects of PAMPs and corresponding trained innate immunity

In executing these activities, the CPM core will contribute to the impact of each project. It will also lead key methodologic advancements in the systems analysis of infection and immunologic response.

Significance

Both methicillin-resistant *Staphylococcus aureus* (MRSA) and *Candida albicans* (CA) are commonly found commensal organisms. While these organisms coexist with their human hosts in most circumstances, in some individuals, most often those immunocompromised in some form, these organisms cause deadly, invasive infections[2]. Effective therapeutic agents exist against both organisms. However, even when isolates from a subject are effective *in vitro*, therapy can be ineffective in resolving infection *in vivo*. Thus, there is a critical need to understand the immunological determinants of infection risk, and how interactions between the organisms and host determine effective clearance, persistence, or mortality.

Recent progress has been made in identifying the clinical and molecular determinants of susceptibility and outcome to both infections[3–7,14]. While demographic and clinical factors are associated with infection and outcome, they are insufficiently predictive in individuals and do not reveal a mechanistic picture[1,4,7,8]. Molecular signatures of infection susceptibility, resolution, and mortality have helped to reveal variation in the immunologic response of subjects. However, it is yet unclear to what extent these signatures are distinct or reflect a common immunologic niche.

We hypothesize that identifying commonalities in molecular signatures between molecular modalities in the same subjects will help to both provide a more mechanistic view of infection and improve predictions of infection outcome. Pan-signature patterns will help to improve our understanding of infection by providing a more complete view of the immunologic differences (Fig. 1). They can correspondingly help to improve outcome predictions in two ways: First, if patterns turn out to be reflections of the same underlying immunologic features, the fusion of these patterns should provide a more accurate estimate of its presence within individual subjects. Second, if two predictive patterns reflect independent variation, recognizing this will help to stratify subjects into separate groups, with potentially differing responses to therapeutic interventions.

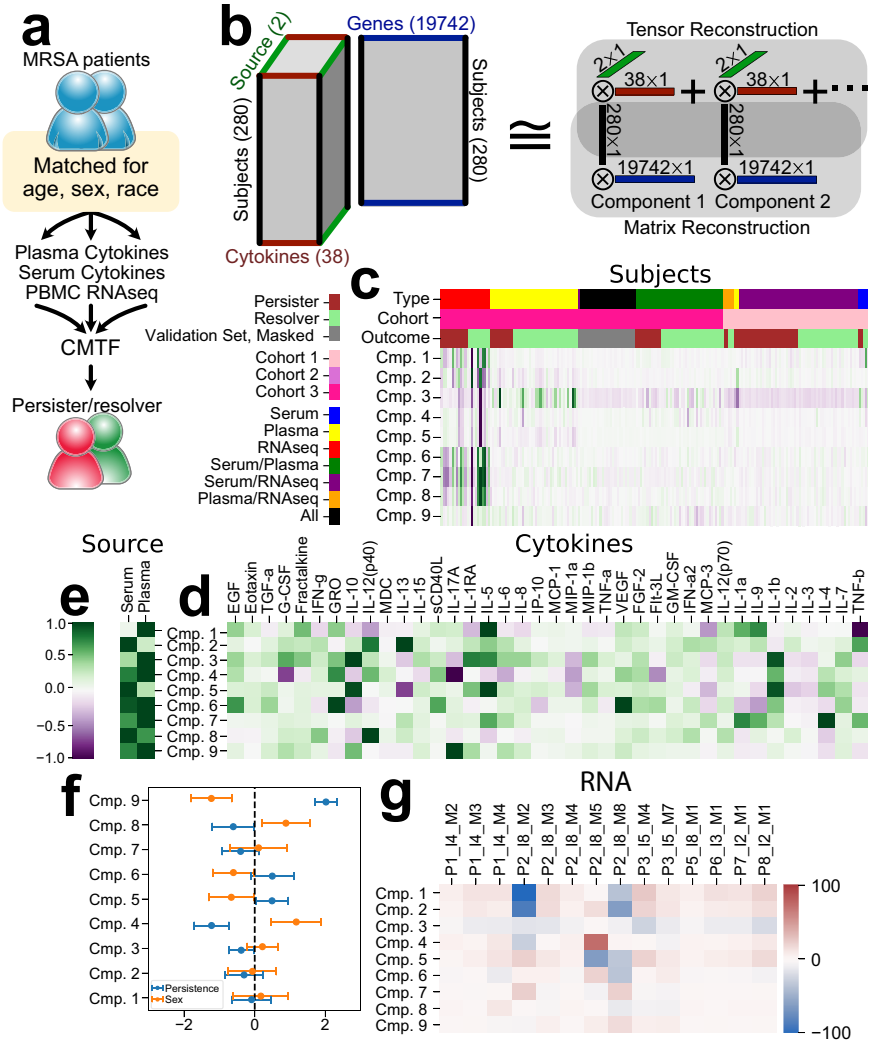

Figure 1. MRSA preliminary data integration through coupled matrix-tensor factorization. a) The UCLA immunogenomics center profiled samples from MRSA bacteremia subjects. b) Cytokines from two sources can be arranged in tensor form, while the RNAseq gene modules are a matrix that shares a subjects mode. Decomposition takes the form of vector outer products, with shared subject factors. c–g) Subject (c), cytokine (d), cytokine source (e), and expression (g, subset shown) factors describe the results. As the data is estimated by the outer product of the factors, the resulting factorization is interpretable. For example, component 9 explains an increase in IL-17A (d) that is preferentially present in the plasma (e). Through the logistic regression weights, we can see that component 9 is associated with persistence and female sex. Gene module P2_I8_M8, weighted on component 9 (g), includes strong enrichment of Th17 signatures. More importantly, the combined factors are more predictive of outcome (ROC 0.85 vs. 0.7/0.75 for each data type separately).

Identifying shared patterns across the response to both MRSA and CA infections will similarly improve our understanding of both diseases. 20% of all CA bloodstream infections are polymicrobial with *Staphylococcus aureus* as a common co-isolated organism[8]. Further, IL-10 production has been identified has an indicator in both infections of a non-productive immune response and poor pathogen eradication[3,5,6]. Thus, one might expect there are either host immunologic deficiencies or commonalities in immune evasion mechanisms that create a shared permissive niche. Beyond a shared niche, integrating the molecular profiling from both diseases can help to characterize the molecular variation associated with shared clinical covariates, such as immunosuppressive drugs, age, and sex. Most accurately defining these molecular patterns of variation can in

turn help delineate whether a predictive molecular marker might be directly implicated in the response to infection, or indirectly implicated via these common characteristics.

Finally, data integration will aid the integration of the basic/mechanistic and clinical studies. Basic studies are essential to mechanistically untangle the regulation behind the immunologic responses to MRSA and CA infection. By contrast, the clinical samples are necessary to ensure what is found in basic studies occurs in patients. Therefore, our tight integration will ensure the relevance and impact of our core's findings.

Innovation

Extend coupled tensor decompositions to continuous representations

Data tensors are organized multi-modal arrays. Their one-mode and two-mode variants are vectors and matrices (Fig. 2). Tensor decomposition is a suite of dimensionality reduction methods that decompose tensors into factors[9]. For instance, CP (canonical polyadic or PARAFAC) decomposition (CPD) approximates a tensor as the sum of several components, each represented as the cross product of mode-specific vectors. CPD can be thought of as a generalization of matrix factorization methods like principal component analysis (PCA).

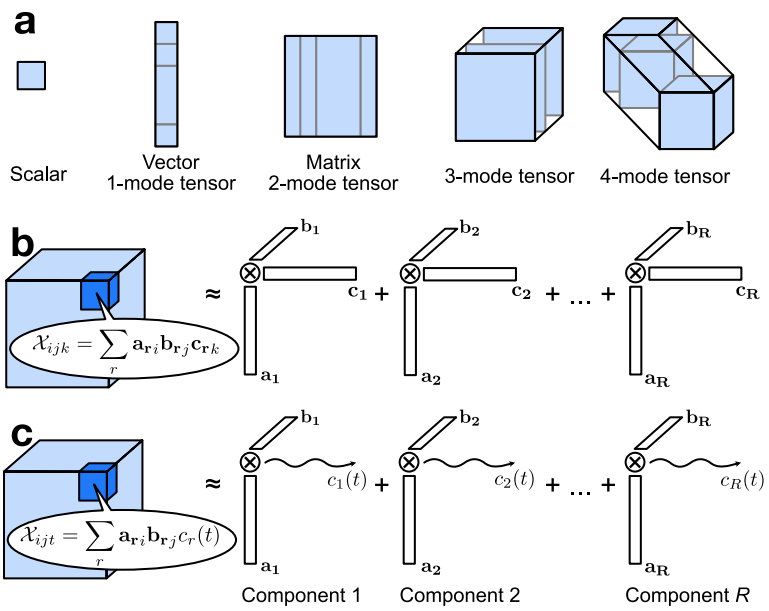

Figure 2. Description of tensor decomposition. (a) Vectors and matrices are 1- and 2-mode tensors, respectively. (b) In CP-structured decompositions, a tensor is approximated by the outer product of vectors representing the variation along each mode. (c) In Aim 1, we will modify this process to allow a functional representation of certain modes.

However, tensor structured data requires each entry on every mode to align perfectly, which limits its use in longitudinal and other studies. For example, longitudinal sample collections usually happen sporadically over time (e.g., days 1, 3, 7, 13 for one patient, but days 1, 5, 8, 11, 17 for another). This prevents one from organizing collection time as a separate mode in a tensor. Other continuous quantities such as the concentration of drug applied may not be matched across studies or may have an effect we wish to constrain to a specific functional form (e.g., Hill curve)[15–17]. Therefore, we propose to extend the definition of CP to work with unaligned modes by expressing them as continuous functions. For example, entry $\mathcal{X}_{ijk}$ will be represented as the sum of $a_i b_j c(k)$ in each component, where $c(k)$ is a continuous function, instead of $a_i b_j c_k$. The continuous function can take many forms, can help to constrain the solution, and allows samples along the mode to take unique positions. Relaxing this limitation will open application of tensor-based methods to a wide array of profiling studies and improve our ability to extract and interpret patterns from the data. Here, we will apply this advancement to study the longitudinal dynamics of CA infection (**Aims 1 & 2**). We will also then integrate this dimensionality reduction with dynamical systems analysis to infer interactions between component patterns that give rise to the observed dynamic responses (**Aim 1B**). Through these applications, this advancement will help us to derive more detailed insights into the mechanisms of CA infection.

Improved methods for factor analysis of low abundance cytokines

Matrix and tensor factorizations are extremely powerful tools for reducing the dimensionality of complex data. In fact, matrix and tensor completion methods can be some of the most accurate missing value imputation approaches, enabling use of other modeling techniques that require complete and uncensored datasets[10,18]. However, these methods are most effective when they account for confounding factors in the data generation process, such as censorship or missing values.

The limited sensitivity of multiplexed cytokine measurements is one of the most common challenges in their modeling[19]. For instance, in our earlier MRSA profiling work (Fig. 1), individual cytokines were below the limit of quantitation 0–100% of the time, and subjects on average had 20% of their cytokines below the limit of quantitation, in line with others' observations with this assay[19]. Useful information is still contained within these measurements, both through the abundance of the measured cytokines and in the pattern of which cytokines are below the limit of detection. We will develop a tailored imputation scheme to handle these censored values. Working with the **Immunomics core**, we will experimentally validate the imputation accuracy of our method using Simoa assays with fg/mL sensitivity[20].

Tensor analysis as a universal approach to data integration/fusion

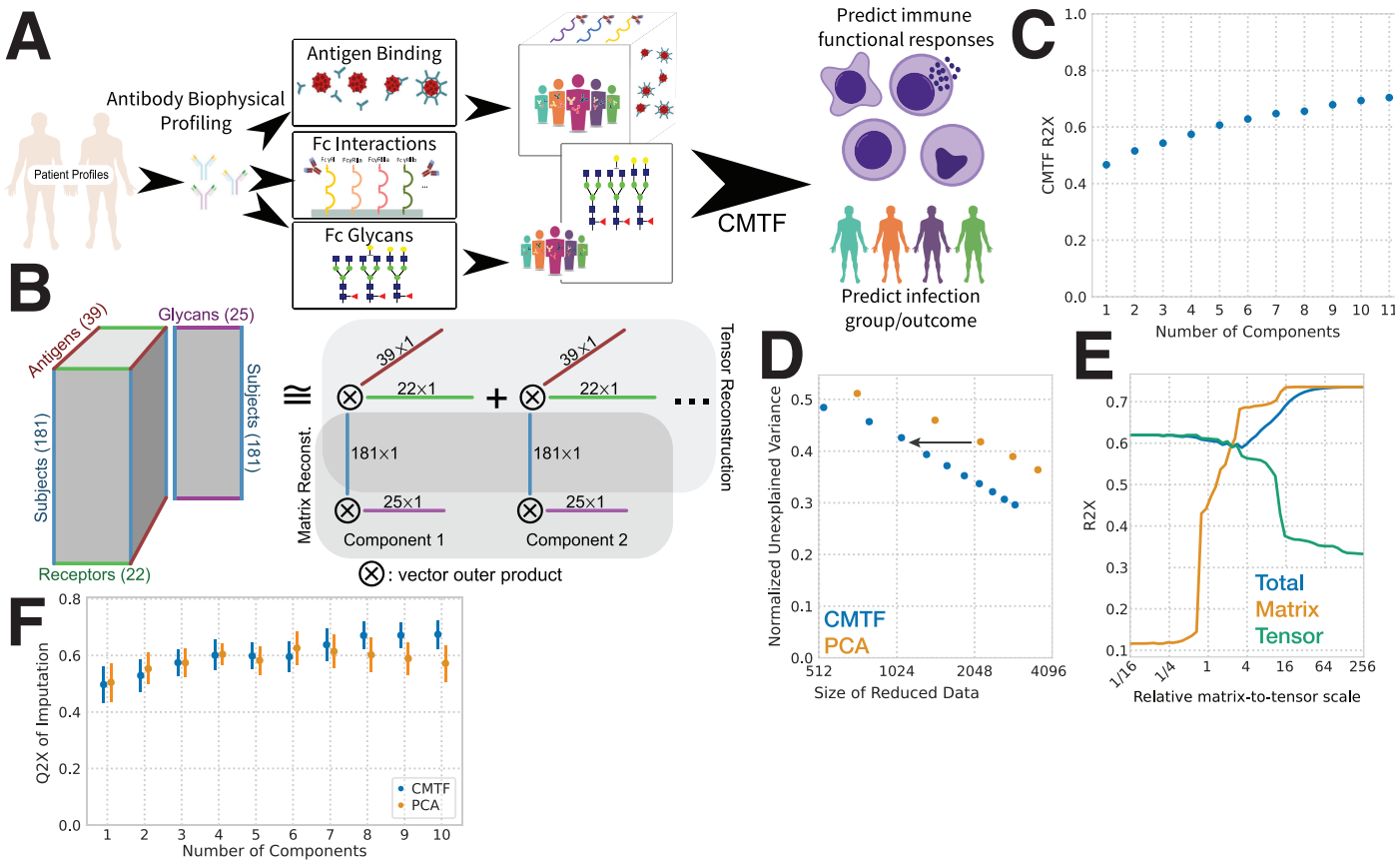

Figure 3. Tensor-based dimensionality reduction improves modeling of systems serology measurements and integration of biophysical and glycan profiling. A) General description of the data. Antibodies are first separated based on their binding to a panel of disease-relevant antigens. Next, the binding of those immobilized antibodies to a panel of immune receptors is quantified. Other molecular properties, such as glycosylation, may be quantified in parallel in an antigen-specific or -generic manner. These measurements predict both disease status and immune functional properties. B) Overall structure of the data. Antigen-specific measurements can be arranged in a three-dimensional tensor wherein one dimension each indicates subject, antigen, and receptor. In parallel, non-antigen-resolved measurements such as quantification of glycan composition can be arranged in a matrix with each subject along one dimension, and each glycan feature along the other. Although the tensor and matrix differ in their dimensionality, they share a common subject dimension. The data are reduced by identifying additively separable components represented by the outer product of vectors along each dimension. The subject dimension is shared across both the tensor and matrix reconstruction. C) Percent variance reconstructed (R2X) versus the number of components used in CMTF decomposition. D) CMTF reconstruction error compared with PCA over varying sizes of the resulting factorization. The unexplained variance is normalized to the starting variance. Note the log scale on the x-axis. CMTF consistently led to a similar variance explained with half the resulting factorization size compared with PCA. For example, as indicated by the arrow, to obtain a normalized unexplained variance of 0.45, PCA required ~2,048 values, and CMTF needed only ~1,024 values. E) The overall and matrix- or tensor-specific R2X with varied relative scaling. F) Percent variance predicted (Q2X) on imputation. From Tan *et al*, *Mol Sys Biol*, 2021.

Tensor-based dimensionality reduction is especially useful for multi-modal data by virtue of (1) more effectively reducing the data and (2) separating variation according to the mode over which it occurs[9]. The latter property enables tensor-based analysis as a natural solution to integration or “fusion” of datasets with shared modes[21,22]. Integrative analysis of datasets is a common challenge, and a variety of solutions have been proposed[21–26]. Tensor-based dimensionality reduction has several unique benefits over alternative approaches. First, as a linear method it is eminently interpretable, and can be readily interpreted by non-practitioners with effective visualization tools. Second, as a purely data-driven approach, its analysis results are pure reflections of the collected data and consequently not susceptible to problems in biased characterization of biological pathways and other prior knowledge. Finally, the algorithmic simplicity of the methods mean that they are readily extensible and can be integrated with other modeling modalities (e.g., **Aim 1B**). As we explore in recent work, while variable selection methods like LASSO or elastic net can be used to build predictive models, the variables selected by these models lose significance when variables are highly correlated, as is often the case with biological measurements[10]. Consequently, dimensionality reduction is essential to try and make mechanistic inferences about large-scale data. In this proposal, we will extensively employ these benefits to integrate data across several different modes. We expect that demonstrating these strengths will considerably broaden the application of these methods for data integration efforts.

Data Sharing Plan: All analysis by the Meyer lab is performed with code and data openly available before publication on Github. Data and code repositories will be designed following FIAR guiding principles, such that both are available and readily reusable for other efforts. For instance, data will be linked to deposition identifiers and metadata available in documented, searchable form. Deposition of code and data will be performed upon publication working with the **BDM Core**. The Meyer lab is a core developer of TensorLy, an open-source Python package for tensor learning, and will make methods advancements available there[27].

Approach

The core will operate as a hub of data integration and exploration across the cores/projects and has three Aims with this in mind. Beginning with a subset of preexisting data[28], the core will lead development of new methods for data integration and deriving multi-modal patterns across longitudinal studies. These will be applied as the CA cytokine profiling is completed to identify common patterns within and between CA and MRSA infections and their outcomes. Finally, the core will work closely with each project in select multivariate and predictive modeling tasks, integrating the basic studies and clinical measurements.

Aim 1. Evaluate the integrated longitudinal determinants of CA outcome

Motivation: A complete picture of CA infection requires integrating the available data, which includes measurements of both cytokine and gene expression response. Longitudinal samples additionally provide unique information about the dynamics of response. A tailored approach toward dimensionality reduction of these data will help to reveal those patterns that are predictive of infection outcome.

Aim 1A: Integrate proteomics and gene expression through a continuous mode coupled tensor decomposition

To start, we will work with the pre-existing transcriptional data of the CA infection cohort[28]. Briefly, 48 subjects with CA infection had blood drawn at admission and longitudinally thereafter. PAX genes were derived from those samples and submitted for RNAseq. 319 samples were successfully analyzed, or roughly 7 per subject (GSE176262). Gene expression will first be reduced into co-expressed modules by WGCNA[29]. Serum samples are available for these samples, and cytokines will be profiled (**Project 3**). We will start with analysis of the expression data and then integrate the cytokines as that data becomes available (Fig. 4).

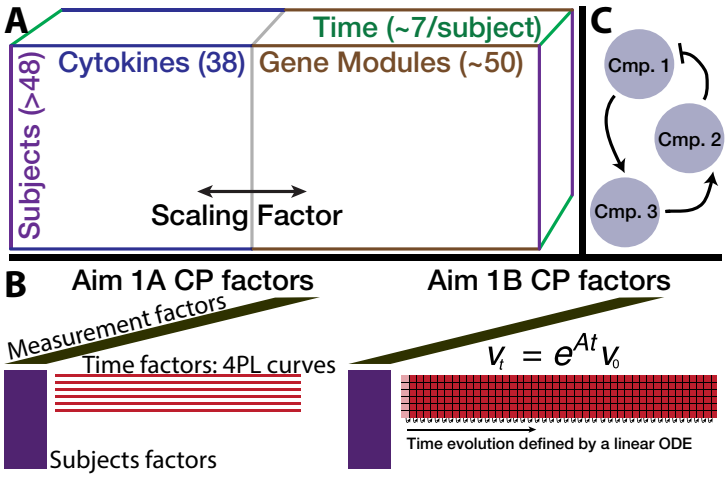

Figure 4. Layout of the CA cohort. A) At least 48 subjects have previously been profiled for their PAX gene expression. The same samples, along with a wider panel, will be profiled for their cytokine abundances. Subjects have longitudinal sampling, with an average of 6.6 samples per subject, and so time will be included as an additional mode. B–C) We will compare constructing the effect of time using a curve for each component, versus a linear dynamic system. The latter will allow us to infer dynamic influences between components (C).

Representing some tensor modes as continuous functions instead of vectors allows for a more flexible alignment of samples along that mode. While standard tensor-structured data mandates that samples be measured at matching positions, we will generalize our approach to allow samples to be unaligned. To start, we will still store a dataset as a multi-dimensional array, with one dimension of that array representing the continuous mode (for simplicity's sake we will assume this is time). With traditional tensor structure, there will be many positions with missing values due to samples being measured at unique times (Figs. 4 & 5). We will store the time value for each sample as a separate vector. The factorization process will overcome missing values through key adjustments.

To allow a continuous representation of time, we must first extend current decomposition methods. Tensor decompositions can be calculated using a variety of schemes. They most generally fall into alternating least squares (ALS) methods, wherein the best fit along each mode is calculated iteratively, or direct fitting (DF) methods wherein the overall quality of the fit is optimized using traditional optimization. Each approach is better suited to certain situations. For example, ALS can be extremely efficient, accurate, and avoid local minima[9]. However, ALS can become trapped in fitting "swamps" wherein all the factor modes are highly interdependent, partly alleviated by line search schemes[30]. DF is simple to implement and can benefit through use of standard optimization methods[31,32]. However, it can be slower to obtain accurate fitting results and more easily becomes trapped in local minima. Because of these tradeoffs, it is helpful to investigate both approaches.

We will implement both the ALS and DF methods for continuous representations of a decomposition mode. For DF, this simply requires changing the optimization variables of the continuous mode to be the variables representing the mode's curve rather than the individual mode positions themselves. To provide the same implementation for ALS, we will replace the factor solving step along the continuous mode. Briefly, during ALS the data tensor is unfolded along the given mode, and the decomposed representation of the other modes is calculated using the Khatri-Rao product. The new factors for the mode of interest are then solved using linear least squares. Two steps must be replaced to accomplish our goal: First, during the Khatri-Rao product we must build a standard matrix representation of our continuous dimension by calculating the value of the curve at each point. Second, during ALS solving for the continuous mode we will use standard non-linear minimization rather than linear least squares solving. Solving the other modes in ALS will be largely unchanged. The factors representing the continuous mode can be discretized by solving for the factors value at each position. This allows one to derive the Khatri-Rao product and perform least-squares solving along those modes. During the tensor unfolding, missing values due to the time mode will constitute an entire column of the matrix; those columns can therefore be removed from the least squares solving in ALS without affecting the result.

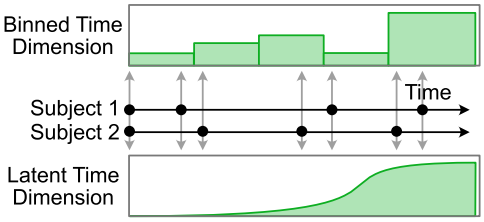

Figure 5. Illustration of the problem. Traditional tensor decompositions require that data be arranged into a single multidimensional array. However, longitudinal studies typically involve samples measured at unaligned timepoints. Binning these samples (top) leads to inconsistent coverage and loss of time resolution. A continuous latent dimension instead takes advantage of the non-alignment for better resolution and allows one to force the factors to represent consistent trends.

We will verify our method is correct in a couple ways: (1) We will check that during each iteration of fitting the quality of the fit strictly increases. (2) We will generate synthetic data, with varied amounts of error introduced, and verify we can recover a correct representation. Next, we will compare each method's performance. Both approaches will be tested with synthetic data of varied size, mode number, rank number, degrees of cross-mode correlation, and missing values fractions. These benchmarking results will help others apply our extension themselves and ensure we fully understand the situations in which either ALS or DF is better given their tradeoffs, alongside any complications introduced by the continuous mode.

As an application of our method, we will factor the CA data to discover consistent patterns across both the cytokines and gene expression modules. For the continuous representation of time, we will use a four-parameter logistic curve. This provides a flexible representation of trends over time, while enforcing a unidirectional effect over time for each component. We have used it to successfully represent trends in serology responses over time[10,33]. Longitudinal trends that are not unidirectional can still be represented through the effect of multiple components. As the tensor decomposition process simply maximizes the variance explained, the scaling between the gene module and cytokine data influences which variance is explained first (Figs. 3E & 4). Therefore, in all subsequent analysis we will vary this scaling to maximize a modeling property, such as predicting clinical outcomes. Finally, to evaluate the benefit of our decomposition approach, we will evaluate: (1) The data reduction possible with our approach, compared to the data in "flattened" matrix form (Fig. 3D). (2) The influence of scaling on explaining each of the data types (Fig. 3E). (3) The ability of our approach to impute missing values (Fig. 3F). Each of these will help to confirm that our dimensionality reduction successfully captures meaningful patterns within the data.

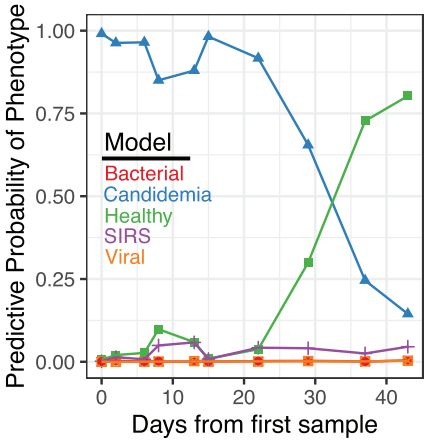

Figure 6. Consistent dynamic trends exist in the immunologic response to CA infection. Samples are derived from a single subject with CA infection that eventually resolves. Each color indicates the prediction output of a multi-class logistic regression model predicting the type of infection sample. As CA resolves, the subject's samples transition to classified as healthy. From Steinbrink *et al*, *Genome Med*, 2021.

Aim 1B: Express longitudinal dynamics as a dynamical system to infer regulatory component interactions

Dynamical systems are uniquely helpful for identifying regulatory interactions from dynamic measurements[17,34,35]. However, their use can be challenging with omic-scale studies because there is insufficient power to determine how individual species influence one another. Dimensionality reduction provides a solution to this by working in terms of correlated patterns rather than individual molecules. Using the tools developed above, we will link these two approaches.

To integrate dynamical systems analysis with tensor decomposition of the molecular factors, we will instead express the time dimension as a linear dynamic system, with a starting factors vector and matrix that

defines the time evolution of the system (**Fig. 4B/C**). As above, solving for the other modes of the tensor decomposition can proceed as usual, since the time factors can be arranged into a standard matrix. The time factor can then be solved using standard fitting approaches for dynamical models[17,35]. Fitting will be performed for the time dimension by optimizing the overall sum of squares error for the tensor reconstruction. Solving the linear dynamical system will be performed using the matrix exponential[36]. The gradient of the time factor with respect to the unknown initial vector and system matrix will be derived using Jax[37]. This quantity will then be transformed into an expression for the overall fit with respect to the unknown parameters using an analytical expression for the gradient of the sum of squared error, based on the n-mode unfolding times the Khatri-Rao product[22,32].

We will compare this formulation of the continuous mode to the previous one in its (1) extent of data reduction (Fig. 3D), (2) imputation performance for held-out data (Fig. 3F), and (3) the stability of the factors upon bootstrapping the subjects[10]. We expect an improvement in all these quantities due to the improved time factor expressiveness. Factor interactions revealed by this analysis will be validated in collaboration with **Project 2**. For instance, a component A may represent a high abundance of IL-17A early in infection and component B a high abundance of neutrophils. This analysis would provide the inference that component A directly contributes to an increase in B. We could then test this using an antibody to deplete IL-17A and then immunophenotyping.

Aim 1C: Test for association between CA immunological trajectories and mortality

We hypothesize that, while single snapshots of each subject are likely predictive, the dynamics of CA infection contain unique information that improves predictions of whether one mounts a successful immune response (Fig. 6). To test this, we will build a model using the subject factors derived in **Aim 1A & B** that explain both the gene expression and cytokine data. Whether a subject ultimately resolves or succumbs to their infection will be predicted using a logistic regression classifier with the subject factors as input and outcome (mortality versus resolution) as the predictor. The prediction accuracy of the model will be assessed by a receiver operator characteristic (ROC) curve using 10-fold cross-validation. As mentioned above, in the coupled factor analysis, the relative scaling (and thus priority in explaining the data) between the gene expression and cytokine data is a parameter of the model, and so this will be adjusted to maximize prediction accuracy. With a predictive model, we will inspect the coefficients of the logistic regression model to determine which factors are associated with outcome. Variation in the coefficients will be quantified by bootstrapping. Associations with isolates' phenotypic features from **Project 2** will be evaluated in a similar manner (see **Aim 3**).

The entirety of the data is almost certain to predict outcome effectively, because many samples were collected long after admission, where subjects were already resolving or progressing in their infection. Therefore, several additional model inquiries are necessary to confirm the specific value of the longitudinal information. First, we will repeat the steps above using factors derived from a progressively earlier subset of the data. For example, rather than factor the data with all samples, we will only take samples at least 2 days prior to mortality or resolution, 5 days prior to mortality or resolution, etc. We expect that removing later samples will reduce but not eliminate the model's ability to predict outcome. A reasonable intermediate cutoff will be determined where most of the model's predictive ability is retained but predictions are made for outcomes observed at least a week into the future. This will reenforce that the model is able to distinguish differences in immunological response that determine outcome.

Finally, we will test that longitudinal dynamics play a significant role in predicting infection outcome. To do so, we will compare outcome predictions with our dynamics-aware model to those derived from two different baselines: First, we will build a corresponding principal component analysis (PCA) and logistic regression model predicting outcome using either the first sample on admission, or alternatively the last sample used in the full model above. By representing each subject with one sample time point, we remove dynamics as a feature the model can use to predict outcome. Second, we will shuffle the timepoints in the full model, so that the overall approach is the same, but the signal of sample ordering is no longer available. We expect that both these models lacking dynamics information will be less predictive of outcome, indicating that dynamics is important to predicting infection outcome.

Expected results: We expect to find that the integrated molecular signatures can predict both the persistence and mortality of CA infection. Furthermore, the dynamics of response to CA infection and integrating both data types will help to predict these clinical parameters more accurately. Finally, we expect that arranging the samples in tensor form will (1) reduce the measurements to a greater extent, (2) effectively separate the contribution of subject-to-subject differences, dynamics, and each measurement, and (3) more accurately represent the data (quantified through imputation accuracy).

Within the factors, we expect to find several known patterns of CA response. First, the response of peripheral blood cells to CA is defined by TLR and interferon (IFN) responses[38]. While all the samples in this Aim will have disseminated infections, we expect they will vary in this response signature due to variation in the severity of infection. Resolving samples should also smoothly transition to a healthy state (Fig. 4)[28].

We expect that there will be clear correspondence between the transcriptional and cytokine signatures for those components that primarily explain cytokine variation. To evaluate this, we will test for correlations between the transcriptional factors and CytoSig, a database of known transcriptional cytokine signatures[39]. We will quantify which components have a significant correlation with each signature (Spearman correlation with permutation test) and, in the presence of such an association, whether there is the expected weighting of that cytokine in the cytokine factors. Agreement between both data sources will be compared to the amount of agreement with the factors shuffled.

Synergy with other Aims and Projects: Aim 1 lays the groundwork for an integrated representation of CA infection. This representation is used throughout **Projects 2 & 3** as a basis of comparison for their findings. This data integration is built upon in **Aim 2**, with the integration of MRSA cohort samples. In **Aim 3** findings from *in vitro* and mouse studies are compared to these patterns. Working with **Project 2**, we will have the opportunity to validate the mechanism of key interactions we identify in **Aim 1C**.

Alternative approaches: We do not anticipate challenges in implementing this Aim. The approach taken here is very similar to that applied in our preliminary analysis integrating gene expression and cytokine measurements to predict MRSA persistence (Fig. 1). There, we have been able to observe that our approach improves outcome prediction and effectively identifies patterns across both datasets. We also have extensive experience with this and alternative data-driven approaches[10,17,40]. If, for some unanticipated reason, the continuous mode approach is not possible, we can always use timepoint binning (Fig. 5).

Like PCA and other matrix decomposition methods, tensor decomposition can also be used to impute missing values[31]. Therefore, we will keep track of missing values within the study and adjust our dimensionality reduction approach based on the extent and pattern of missingness.

We will similarly keep track of the batch properties of experiments and monitor the extent to which our reduced data reflects study design patterns. Like PCA, tensor-based dimensionality reduction can help to separate batch effects from other variation in the data[12]. If batch effects appear within many components of our decomposition, we can directly correct for batch effects using standard, well-developed methods[41].

Dimensionality reduction can become challenging to calculate at extreme scales. While we do not expect challenges at the scales proposed in this study, a variety of solutions exist to improve algorithmic scalability if this were to become an issue. These include line search routines to improve convergence[30]; streaming and batching methods to work with data that is larger than can be loaded into memory[42,43]; problem conditioning like orthogonalization[44]; and alternative algorithms to trade-off accuracy, scalability, and convergence speed[45].

Aim 2. Identify shared and distinct molecular and cellular patterns across MRSA and CA

Motivation: The commonalities between MRSA and CA infection and their occasional co-isolation suggest a shared immunologic niche. We will examine the common molecular patterns across both cohorts, and test for their association with outcome and other clinical parameters.

Aim 2A: Improve factorization of clinical cytokine panels

A common challenge with proteomic cytokine analysis is that many samples can be below or above the limit of quantitation (LoQ). This means that, while we know the cytokine amount in that sample is below or above a known value, we do not otherwise know its quantity. Common approaches, such as filling in the LoQ boundary value, introduce bias to the data. While this may not be an issue with a small number of such values, it is common to have cytokines with half of their values outside the LoQ in multiplexed cytokine samples. Sample and resource limitations preclude experimental solutions like more sensitive technologies[20].

To investigate solutions to this problem, we will use a series of cytokine panel measurements made from MRSA infection subjects. This dataset is an excellent testbed for several reasons: (1) The LoQ is a known confounder in the data. The cytokine measurements were made across two independent cohorts, where the LoQ was different for each. As a result, the cohorts subtly separate along certain components during dimensionality reduction (Fig. 1, e.g., component 1). We can therefore directly examine whether this confounding effect is resolved. (2) Many of these samples are still available within the **Immunogenomics Core**, and so we will use the single-molecule quantitation technology Simoa as a more sensitive assay to experimentally validate that we are correctly imputing even the censored cytokine amounts[20]. (3) Many subjects had both their plasma and serum analyzed in parallel, sometimes with differences in the LoQ or the abundance of the cytokine. As a result, the dataset provides a testbed in which the inference process is optionally made simpler by having information about each individual subject (captured within the tensor arrangement of the data; Fig. 1B).

To correct for the LoQ effect, we will modify our cytokine factorization process. Like in previous work, we will factor these data as a 3-mode tensor, with modes representing cytokine, subject, and whether samples were derived from serum or plasma (Fig. 1B). 9 components can explain roughly 80% of the data variance using standard CPD (Fig. 1). We will modify the ALS fitting process by updating the values in the cytokine tensor after each iteration of ALS. Two separate tensors, L and H, will have NaN values anywhere a cytokine falls within the range of quantitation or is missing. If a cytokine is below the range, the L tensor will store the

LoQ and H will have a NaN value. Cytokines above the LoQ will have the limit stored in H and L will have a NaN value. Using these, the cytokines falling outside the LoQ will be updated by filling in their values with those predicted by reconstructing the data tensor. Iterative filling is a common technique for missing and censored value handling. After filling in the value, these values will additionally be adjusted based on the information we do know about their quantity. If the filled value is above the LoQ for a cytokine that was below the LoQ, for example, we will update the value to be the LoQ. In this way we will fill in the values with the best inference given what we know about the factors and how these values fell outside the LoQ. This expectation-maximization process will proceed until convergence.

We will test this new approach to cytokine factor analysis in several ways. First, this updated process will free the latent factors from having to explain variance due to cytokines below the LoQ being held exactly at the LoQ (Fig. 7A). Therefore, we expect that fewer components will be needed to explain the same amount of variation in the dataset. Second, we expect this improvement will enhance our ability to predict unseen cytokine abundances. To test this, we will remove random cytokine-subject measurements and test the ability of the factorization process to impute these values (Fig 3F)[10]. We expect this imputation error will be reduced when properly accounting for the LoQ. Finally, we expect that the cohort-to-cohort difference seen in some factors (e.g., cohorts 1 versus 2 in components 1 and 2 in Fig. 1C) will be removed by no longer holding many values at the LoQ. This is because, as illustrated in Fig. 7A, the LoQ differs between cohorts, and therefore produces a series of values that differ in a cohort-specific manner. We will test for cohort-to-cohort differences using a Kruskal–Wallis test with each component and expect any components that are significantly difference between cohorts will no longer have such a difference with this enhancement.

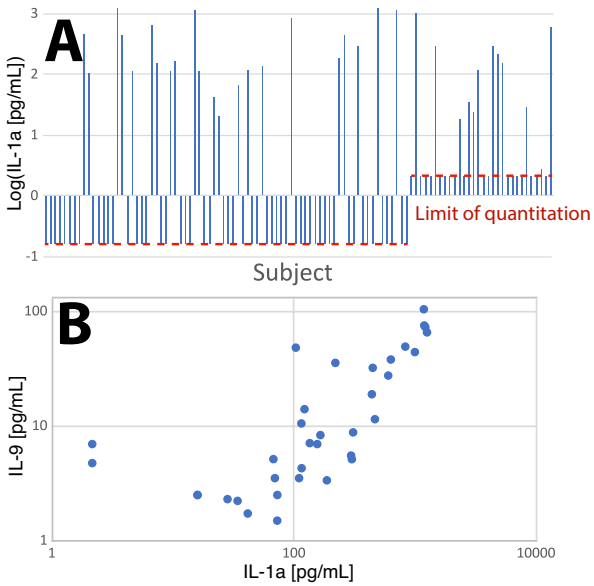

Figure 7. Illustration of limit of quantitation effect within the MRSA dataset. A) Plot of IL-1a abundance for a subset of the MRSA subjects, one of the cytokines with the most values below LoQ. Note the log axis. The lower limit of quantitation is different between two cohorts because the samples were run at separate times. Because many subjects are set to the LoQ, the difference between cohorts can drive cohort-batch patterns. B) IL-9 and IL-1a abundance are highly correlated, suggesting that imputation based on dimensionality reduction should be highly effective as the pattern of one can be used to infer the other.

With a cytokine factorization method fully developed, we will apply this approach throughout the study, including with the CA cohort and methodological improvements in Aims 1A/B. As experimental validation of our approach, we will also identify cytokines with consistent inferred variation in abundance below the LoQ (uncertainty quantified by bootstrapping subjects). Working with the **Immunogenomics Core** we will experimentally validate the variation in three cytokines by measuring their abundance in 20 subjects' samples by Simoa[20]. We expect to find quantitative agreement between the inferred and measured cytokine abundances as quantified by the Pearson correlation between the inferred and measured abundance, only for those samples below the LoQ for the multiplexed assay. These computational and experimental validation tests will demonstrate that this is a generally valuable approach for multiplexed cytokine analysis.

Aim 2B: Integrate MRSA and CA transcriptomic, methylation, and cytokine measurements

With an improved approach for cytokine factor analysis, we will then apply decomposition of the transcriptomic and cytokine data across both MRSA and CA. Decomposition will be applied using the same ALS approach as we previously used for more typical tensor-matrix factorization[10,22], but with coupling between five sets of axes (Fig. 8). First, the subject patterns will be constrained to be shared across cytokine and transcriptomic datasets within each cohort. Next, the cytokine patterns of each decomposition component will be shared across the MRSA and CA studies. Finally, the gene module and methylation patterns will be shared across the MRSA and CA studies. Coupling is accomplished by solving the least squares step along that mode using a concatenation of the two coupled datasets[10,22,31].

Dataset coupling introduces a necessary new free parameter in how the resulting factors are derived (Fig. 3E & 8). Because the factorization process minimizes the sum of squared error, the relative scaling between datasets changes the amount of variance explained by the factorization process. For example, one could make the scaling of the cytokine measurements much larger than the gene expression data. In this case, the derived factors would be almost entirely dependent on the cytokine measurements alone, and the variance explained in the gene expression would only be whatever happens to correlate with the cytokine data. In practice,

however, we generally want to derive factors that are a function of both datasets. To explore the effect of scaling, we will vary the study-to-study scaling and the relative cytokine and gene expression scaling (Fig. 3E & 8). Scaling values will be picked such that the resulting factors are responsive to both datasets. We can also elect to identify a scaling factor that maximizes resulting predictions of infection outcome.

We expect that coupled dimensionality reduction will show that there are shared component patterns across the datasets. To verify this, we will compare the size of the resulting reduced data when factored in coupled form versus when each dataset is reduced on its own. That is, we will break each coupling so that we have six individual tensors, and then approximate them using CPD individually. By calculating the percent variance explained (R2X) and factorization size for each dataset, we can produce a plot indicating the effectiveness of the dimensionality reduction (Fig. 3D). This will then be compared to the R2X and size of the reduction achieved in coupled form. We expect that greater reduction can be achieved in coupled form because the coupled factors matrices are reused and thus not duplicated across datasets. A second way in which we will verify that there are shared coupled patterns is by permuting the positions of the coupled mode in one of the datasets. For instance, we will permute the positions of the subjects only within the gene modules tensors. This preserves the data, as the ordering of the subjects is arbitrary, but breaks the subject associations between the cytokine and gene modules measurements. We expect that this will lead to a striking reduction in the effectiveness of the dimensionality reduction. The same process can be performed with the ordering of the cytokines or gene modules across studies.

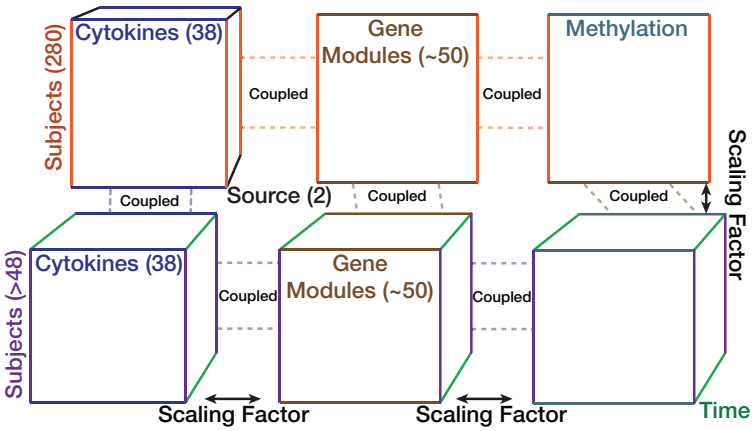

Figure 8. Diagram of the factorization layout and coupling. Both the MRSA and CA data are structured as described above. The MRSA cytokine tensor includes slices from serum and plasma samples. All three CA data types include a third dimension representing time. The datasets are coupled such that the cytokine, gene expression, methylation, MRSA subject, and CA subject factors are shared. Data coupling introduces three scaling factors, which influence the scaling between data types and between the cohorts.

Aim 2C: Identify pattern associations with age, sex, infection outcome, and isolate properties
With a reduced representation of all the molecular data, we will then test for associations of the subject factors with subject age, sex, and infection outcome. Sex and infection outcome will be predicted from the molecular data factors using regularized logistic regression. Age will be predicted using regularized least squares regression. Infection outcome will be a binary outcome, indicating infection persistence (see **Project 1**) in MRSA, and mortality (see **Projects 2 & 3**) in CA. The models will be separately constructed for MRSA and CA in case there are distinct factor-to-output relationships. However, we will still be able to identify common regulatory events through the model weights.

The prediction models will be evaluated in several ways. First, the predictive accuracy of the models will be evaluated using 10-fold cross-validation. Briefly, 10% of the subject factors will be left out of the data, and the model will be built to identify factor-output associations. The model predictions will then be derived for the left-out data using the molecular factors. This process will be repeated until each subset of data has been predicted. Prediction accuracy will be quantified using the area under the receiver operator curve (AUC). As described above, in addition to the regularization strength of the prediction models, we will adjust the two scaling factors within the coupled decomposition, and the number of factored components, to maximize the prediction AUC. Separately, the variance in the prediction-to-factor component associations will be quantified by bootstrapping the subjects. In total, these activities will help to identify robust associations between the factor components and outcome/clinical properties. As the factors are shared, the factorization process naturally allows us to identify shared associations across molecular measurements and infection types.

Expected results: We expect to find that the experimental and computational validation will show that our factorization-imputation scheme in **Aim 2A** will more accurately summarize the patterns found within proteomic cytokine profiling and can impute the abundance of cytokines below the LoQ. In **Aim 2B**, we expect to find consistent, multi-omic and pan-cohort patterns of immunologic response across CA and MRSA infection, as well as across measurement modalities. Evidence of this will include the extent of dimensionality reduction, our comparison to an independent dataset of cytokine transcriptional response, and permutation experiments. We expect that a subset of these patterns will be associated with age, sex, infection outcome, and isolate properties (Fig. 1).

Synergy with other Aims and Projects: Aim 2 produces a fully integrated view of CA and MRSA infections, identifying their shared and unique patterns. As outlined more within **Aim 3**, *every **Project** will use this as a*

basis of comparison for the molecular patterns found within human infection. Aim 2.1 additionally leverages the unique capabilities of the **Immunogenomics Core** to experimentally validate the methodologic advancements.

Alternative approaches: The data integration here differs from our preliminary and previously published analysis only in its scale (Fig. 1)[10], and so we do not anticipate challenges in its execution. However, if, for example, it became difficult to factor all the data in a combined form, we could factor each tensor individually in uncoupled form (Fig. 8). Patterns between each dataset could then be linked by looking for correlations between the factor matrices. Another possibility is to apply a tensor-based partial least squares regression technique[46,47]. Rather than maximizing the variance explained within the molecular data, this would maximize covariance with between the data and prediction of interest. Doing so would reduce the number of components needed to make each prediction, although different factors would exist for each prediction.

Aim 3: Integrate outcomes from experimental models and human infection

Motivation: As the last part of its data integration role, the CPM core will help with the multivariate analysis of data throughout the projects, along with their integration with clinical and immunologic profiling.

Methodology: Details of several methods used by the CPM core throughout the projects are outlined here.

For several applications we will employ **principal component (PC) analysis (PCA)**. This is a dimensionality reduction technique that finds the principal directions of variation within a dataset. An associated benefit of this reduction is that the PC patterns are orthogonal, often aiding downstream analysis. PCA can be extremely effective at dimensionality reduction, particularly for biological data that tends to have highly correlated measurements. To examine the optimal number of PCs for our analysis, we will examine the variance reconstructed (R2X) by the method with differing numbers of components. PCA is most effective when a small number of PCs explain a large fraction of the variance.

A second method we will apply is **partial least squares regression (PLSR)**. This is a supervised method, meaning it can be used to make predictions of a certain quantity given training inputs. In contrast to PCA, which maximizes the variance explained in an input dataset, PLSR identifies components that maximally explain the covariance between the input and predictor datasets. By performing dimensionality reduction during the process of model construction, PLSR can provide accurate predictions in situations where there are a greater number of input variables than samples. Like PCA, it is highly effective for biological data wherein many input variables tend to be highly correlated. We will plot the loadings of PLSR models to identify the relationships between input variables and output predictions of interest. The scores of the PLSR model can also be used to identify how samples are related to these input variables.

We will use **cross-validation** to estimate the predictive ability of our modeling efforts. Briefly, a subset of the data will be held out during the model building process, and the model will be fit to the remaining data. The model will then be used to predict the outcome of interest using the held-out samples' input values. The model's predictions will then be compared to the actual, held-out answers. This process will be repeated with subsets of held-out data until all samples have been held out once. Cross-validation is critical to prevent model overfitting, wherein a model is tailored to the exact training dataset, rather than identifying generalizable trends that can be used to predict new samples. Therefore, model parameters, such as the number of components used within PLSR, will be determined by minimizing the cross-validation-estimated prediction error. However, this model adjustment approach can lead to model bias because the model parameters are adjusted in a dataset-specific way[48]. To ensure this is not a problem, we will employ a nested cross-validation strategy, where model parameters are set within a second cross-validation process, where needed.

To examine the variation in outputs for any of the modeling techniques, we will use **bootstrapping**. Briefly, this is a method that effectively simulates how a model would change if an entirely new replicate of the current dataset were produced. This is accomplished by resampling the independent replicates of the dataset, with replacement, to generate a new dataset with some entries duplicated and others removed. This technique can be applied to any model output, including graphical plots, and enables hypothesis testing using model outputs.

Where a multi-modal structure exists in our data (Fig. 2), we will apply **canonical polyadic (CP) decomposition (CPD)**, also known as Parallel Factors Analysis (PARAFAC). Decomposition will be performed using TensorLy[27], using ALS and line search. Factors will be normalized after factorization such that each component has variance of 1, and sign indeterminacy will be corrected by ensuring all but one mode is positive on average[9]. Finally, we will variance-order the components so that there is a consistent ordering of the factors. Where applied, our tensor approach will be justified by examining the imputation error and extent of dimensionality reduction compared to PCA (Fig. 3D/F)[10].

In both **Projects 1 and 2**, we will need to identify a subset of isolates for deeper investigation. To do so, we will take a **variation-maximizing subsampling** approach. The initial profiling of all isolates will be used to visualize the panel of isolates in PCA space. A subset of isolates will be chosen so that they are spread out in PCA space for each clinical outcome. In this way, we will ensure that the subset of isolates chosen retain the phenotypic diversity of the original cohort.

Aim 3A: Identify interactions between MRSA epigenetic plasticity & murine/human immune responses

We will work closely with **Project 1** to use multivariate analysis to address each of their hypotheses. To start, one of the initial goals of **Project 1** is to profile phenotypic, genotypic and epigenotypic characteristics of persistent and resolving MRSA isolates in a panel of environments, including the cellular, serum, and plasma fractions of human blood. Briefly, previous work from our group found that host molecular measurements could be predictive of bacterial persistence, but that standard lab cultures did not differentiate the outcomes of infection[3,5,7]. With the hypothesis that alternative environments can improve the ability of diagnostic cultures to predict persistence, we will examine the colony formation of 75 persistent and 75 resolving strains across the culture environments.

We expect a high degree of strain-to-strain and condition-to-condition correlations. For instance, certain strains will be susceptible to vancomycin (VAN) across all environments, and certain conditions might hinder or promote growth across all strains. Therefore, to explore these data in a multivariate way, we will examine the colony responses using PCA. Colony counts will be log-transformed and centered before dimensionality reduction. We expect that a small number of PCs will be able to capture >90% of the dataset variation and help to visualize these multivariate patterns. PCA will also help us identify a subset of strains to select for further analysis that have representative responses among the overall pattern, by helping to visualize the overall response patterns across each culture environment.

Next, we will use discriminant PLSR (dPLSR) to determine whether the colony response behaviors can predict resolving versus persistence status. (dPLSR is simply PLSR using a binary predictor variable). We expect to find that the confluence of colony responses will be able to predict resolver status more accurately than responses in the standard assay conditions alone. In parallel, we will also evaluate the ability of each culture condition to predict resolver status on its own using logistic regression. We expect that certain culture conditions will also be able to predict resolver status better than standard lab diagnostic conditions. Comparing dPLSR and logistic regression will allow us to determine to what extent predictions can be made more accurately by combinations of response patterns across conditions.

Also in **Project 1**, we will help with the multi-modal analysis of the *in vivo* treatment interventions. Briefly, four proposed interventions for overcoming VAN persistence will be applied, across a panel of strains, with or without VAN treatment, with mouse replicates in each group. At the end of each experiment mice will be sacrificed and the colony forming units (CFU) of several tissues will be quantified. To help visualize the variation among these data, we will arrange it into a four-mode tensor (VAN treatment, strain, intervention, tissue site). The data will be averaged across replicates, centered and variance scaled across all other modes within each tissue site. We will use CPD to reduce this data into component patterns and visualize its variation. As described above, we will compare the extent of dimensionality reduction as validation of our tensor approach compared to PCA and adjust our approach if needed. Uncertainty in the resulting factors will be quantified and visualized by bootstrapping.

In **Project 1, Aim 3**, we will help explore the broad profiling collected through multi-modal tensor analysis. Briefly, the host response to infection will be collected at early and late timepoints, across a series of isolates, in six tissues, for both sexes, with and without VAN treatment, with animal replicates, using a series of molecular profiling technologies. We will arrange this data into a multi-modal (time, isolate, tissue, sex, VAN, molecular measurement) tensor. Measurements will be averaged across animal replicates, and then centered and variance scaled across conditions for each measurement. We will then reduce this data into component patterns using CPD, comparing the extent of dimensionality reduction to that achieved with PCA. Uncertainty in the resulting factors will be quantified and visualized by bootstrapping. Given that the data reduction is unaware of the resolving status of the isolate strain, and we expect that the data will be explained by a relatively small number of component patterns, we will be able to look for patterns that show a difference in abundance between persister and resolver status, even among a small number of isolates. These molecular signatures can then be compared to those in **Aim 2** where we identify human subjects patterns.

Aim 3B: Characterize interactions between CA phenotypic and immunologic response features

Working with **Project 2**, we will help to explore the variation among CA isolates, and how this variation relates to their genomic features and the immune response to infection. Like **Project 1** does with MRSA, an initial focus of **Project 2** will be the phenotypic characterization of CA isolates. Each isolate will be profiled in a variety of ways within **Project 2, Aim 1**, including: growth rate, anti-fungal susceptibility, filamentation, biofilm formation, pathogen-associated molecular patterns (PAMPs) expression (through staining), invasion and damage of the endothelium, and genomic features as quantified by the BDM core. We expect that many of these behaviors will be correlated or anti-correlated and will therefore explore the phenotypic variation by PCA. Associations with mortality and infection timepoint (to look at adaptation) will be explored by PLSR. We will also identify correlations between the properties here and the isolate factors in **Aim 2B**. (Per-isolate factor associations can be derived by taking the Khatri-Rao product of the subjects and time factors in **Aim 2B**[9].)

Projects 2 and 3 will coordinate to measure several whole blood responses CA. Briefly, whole blood from 2 male and 2 female donors will be used. The CA isolates will be introduced, and cytokine/chemokine

responses measured. Intracellular cytokine staining will be performed on the exposed cells. CA killing, phagocytosis, and survival will be measured. These data will be arranged into a 3-mode tensor, with modes representing CA isolate, measurement, and blood donor. We expect that this arrangement will be extremely helpful for identifying overall patterns because there will be similar patterns across donors, as well as donor-specific effects across isolates. The patterns in these data will then be identified and visualized using CPD. The in-filling technique from **Aim 2A** will be applied to accurately model the cytokine measurements. Associations with mortality will be explored by logistic regression of the isolates factor matrix. We will also identify correlations between the isolate factors derived here and those from **Aim 2B**.

Aim 3C: Infer the effects of PAMPs and corresponding trained innate immunity

Finally, working with **Project 3** we will help to model the mechanistic determinants of trained immunity. Working with **Project 3**, we will build predictive models of the presence and function of PAMP/PRR signaling within systemic CA patients. To start, we will integrate the measurements in **Project 2, Aim 1** that characterize the abundance of PAMPs in each isolate. These will be compared to pattern recognition receptor (PRR) responses, and the cytokine and gene expression measurements collected in the same subjects (**Project 3**). PLSR will be used to link PAMP abundance to PRR response. PAMP abundance and PRR response will then be used to predict cytokine, gene expression, and methylation responses by PLSR. Gene expression will first be reduced into co-expressed modules by WGCNA[29].

Separately, we will help to explore the induction of specific enhancer elements by PAMPs, and their pharmacologic manipulation by a panel of compounds. Briefly, **Project 3** will profile the macrophage (Mf) states induced by 10 distinct PAMPs. 5 of these will be selected for more in-depth analysis, where the effect of 10 compounds is profiled. Mf state will be profiled through open chromatin, methylation, and gene expression profiling. The sequencing results will be processed into region or gene summaries by the **BDM core**. We will arrange these data into tensor form for exploration, where the data is represented in 3-mode form (PAMP state, drug intervention, measurement). After factorization, each molecular pattern will be related to functional responses by regression. We will also test for correlations between each of the gene expression and methylation patterns derived here, and those identified within the human subjects in **Aim 2B**. We will test the robustness of derived signatures through bootstrapping across replicates (**Project 3**) and subjects (**Aim 2B**).

Expected results: We expect to find that the more extensive profiling of isolates and their interactions with the immune system performed here will reveal properties of the organisms that correlate with persistence and (in the case of CA) mortality. We expect that these organismal properties will have correlates with cytokine, gene expression, and methylation signatures from **Aim 2B** reflecting the immune response to these properties. Through the analysis in concert with **Project 1, Aim 3**; **Project 2, Aim 2**; and **Project 3**, we will identify the mechanistic underpinnings of these molecular patterns, again mapped back to the unified view of human infection in **Aim 2B**. Using predictive models of sex, alongside sex as an experimental parameter, will identify signatures that are associated with sex.

Synergy with other Aims and Projects: ***Aim 3*** *connects to every part of each* ***Project*** *in the center.* Through these activities, the center shares a fully unified approach toward the multivariate and multi-modal analysis of molecular signatures. Findings using *in vitro* and murine model systems will be tested for their representation within the human subject cohorts in **Aim 2**. This unified analysis similarly serves as a bridge and eventually a hypothesis-generation engine for common mechanisms of persistence in both CA and MRSA.

Alternative approaches: As described above in the description of our tensor approach, dimensionality reduction can proceed with PCA after tensor-structured data is flattened into matrix form. In each application of tensor analysis, we will explicitly justify its structure into tensor form by comparing to PCA. Another possibility is to apply a tensor-based partial least squares regression technique[46,47]. Rather than maximizing the variance explained within the molecular data, this would maximize covariance with between the data and prediction of interest. Doing so would reduce the number of components needed to make each prediction, although different factors would exist for each prediction.

Timeline & Publication Plan: The CPM core will begin in Year 1 with the methodologic developments in **Aim 1A & 2A**, as these can initially rely on pre-existing data from the MRSA and CA cohorts. We plan for each of these to be described in methods-focused publications to help promote their wide-spread use. The other parts of this proposal will then proceed as the data becomes available. For instance, we expect that the CA cytokine measurements would become available in Year 2, to finish the work in **Aim 1**. Eventually, we anticipate that the CPM core activities will be reflected in all the center publications (**Aim 3**), and that a publication describing the fully unified view of CA and MRSA (**Aim 2**) will be produced in Year 5.

Power analysis: Power analysis of all studies has been included within the **BDM core** to allow the CPM core to focus on the development and application of innovative systems-level analyses. As described above, we use non-parametric methods throughout, including bootstrapping and permutation, to rigorously assess model uncertainty. Model prediction will be benchmarked through cross-validation, with nesting where necessary.

References

1. Blot SI, Vandewoude KH, Hoste EA, Colardyn FA. Outcome and Attributable Mortality in Critically Ill Patients With Bacteremia Involving Methicillin-Susceptible and Methicillin-Resistant Staphylococcus aureus. Arch Intern Med. 2002 Oct 28;162(19):2229. PMID: 12390067

2. Pappas PG, Lionakis MS, Arendrup MC, Ostrosky-Zeichner L, Kullberg BJ. Invasive candidiasis. Nat Rev Dis Primer. 2018 Jun 7;4(1):18026. PMID: 29749387

3. Mba Medie F, Sharma-Kuinkel BK, Ruffin F, Chan LC, Rossetti M, Chang Y-L, Park LP, Bayer AS, Filler SG, Ahn R, Reed EF, Gjertson D, Yeaman MR, Fowler VG, the MRSA Systems Immunobiology Group. Genetic variation of DNA methyltransferase-3A contributes to protection against persistent MRSA bacteremia in patients. Proc Natl Acad Sci. 2019 Oct 1;116(40):20087–20096. PMCID: PMC6778225

4. Matzaraki V, Le KTT, Jaeger M, Aguirre-Gamboa R, Johnson MD, Sanna S, Rosati D, Franke L, Zhernakova A, Fu J, Withoff S, Jonkers I, Li Y, Joosten LAB, Netea MG, Wijmenga C, Kumar V. Inflammatory Protein Profiles in Plasma of Candidaemia Patients and the Contribution of Host Genetics to Their Variability. Front Immunol. 2021 Aug 26;12:662171. PMCID: PMC8428519

5. Chang Y-L, Rossetti M, Gjertson DW, Rubbi L, Thompson M, Montoya DJ, Morselli M, Ruffin F, Hoffmann A, Pellegrini M, Fowler VG, Yeaman MR, Reed EF, with the MRSA Systems Immunology Group. Human DNA methylation signatures differentiate persistent from resolving MRSA bacteremia. Proc Natl Acad Sci. 2021 Mar 9;118(10):e2000663118. PMCID: PMC7958259

6. Johnson MD, Plantinga TS, van de Vosse E, Velez Edwards DR, Smith PB, Alexander BD, Yang JC, Kremer D, Laird GM, Oosting M, Joosten LAB, van der Meer JWM, van Dissel JT, Walsh TJ, Perfect JR, Kullberg B-J, Scott WK, Netea MG. Cytokine Gene Polymorphisms and the Outcome of Invasive Candidiasis: A Prospective Cohort Study. Clin Infect Dis. 2012 Feb 15;54(4):502–510. PMCID: PMC3269308

7. Chan LC, Rossetti M, Miller LS, Filler SG, Johnson CW, Lee HK, Wang H, Gjertson D, Fowler VG, Reed EF, Yeaman MR, the MRSA Systems Immunobiology Group. Protective immunity in recurrent *Staphylococcus aureus* infection reflects localized immune signatures and macrophage-conferred memory. Proc Natl Acad Sci. 2018 Nov 20;115(47):E11111–E11119. PMCID: PMC6255181

8. Carolus H, Van Dyck K, Van Dijck P. Candida albicans and Staphylococcus Species: A Threatening Twosome. Front Microbiol. 2019 Sep 18;10:2162. PMCID: PMC6759544

9. Kolda TG, Bader BW. Tensor Decompositions and Applications. SIAM Rev. 2009;51(3):455–500.

10. Tan ZC, Murphy MC, Alpay HS, Taylor SD, Meyer AS. Tensor-structured decomposition improves systems serology analysis. Mol Syst Biol. 2021 Sep;17(9). PMCID: PMC8420856

11. Omberg L, Golub GH, Alter O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. Proc Natl Acad Sci. 2007 Nov 20;104(47):18371–18376. PMCID: PMC2147680

12. Martino C, Shenhav L, Marotz CA, Armstrong G, McDonald D, Vázquez-Baeza Y, Morton JT, Jiang L, Dominguez-Bello MG, Swafford AD, Halperin E, Knight R. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. Nat Biotechnol. 2021 Feb;39(2):165–168. PMCID: PMC7878194

13. Chitforoushzadeh Z, Ye Z, Sheng Z, LaRue S, Fry RC, Lauffenburger DA, Janes KA. TNF-insulin crosstalk at the transcription factor GATA6 is revealed by a model that links signaling and transcriptomic data tensors. Sci Signal. 2016 Jun 7;9(431):ra59–ra59. PMCID: PMC4914393

14. Jaeger M, Matzaraki V, Aguirre-Gamboa R, Gresnigt MS, Chu X, Johnson MD, Oosting M, Smeekens SP, Withoff S, Jonkers I, Perfect JR, van de Veerdonk FL, Kullberg B-J, Joosten LAB, Li Y, Wijmenga C, Netea MG, Kumar V. A Genome-Wide Functional Genomics Approach Identifies Susceptibility Pathways to Fungal Bloodstream Infection in Humans. J Infect Dis. 2019 Jul 31;220(5):862–872. PMCID: PMC6667794

15. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013 Jan;41(Database issue):D955-961. PMCID: PMC3531057

16. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 29;483(7391):603–607. PMCID: PMC3320027

17. Farhat AM, Weiner AC, Posner C, Kim ZS, Orcutt-Jahns B, Carlson SM, Meyer AS. Modeling cell-specific dynamics and regulation of the common gamma chain cytokines. Cell Rep. 2021 Apr;35(4):109044. PMCID: PMC8179794

18. Liu J, Musialski P, Wonka P, Ye J. Tensor Completion for Estimating Missing Values in Visual Data. IEEE Trans Pattern Anal Mach Intell. 2013 Jan;35(1):208–20. PMID: 22271823

19. Chaturvedi AK, Kemp TJ, Pfeiffer RM, Biancotto A, Williams M, Munuo S, Purdue MP, Hsing AW, Pinto L, McCoy JP, Hildesheim A. Evaluation of Multiplexed Cytokine and Inflammation Marker Measurements: a Methodologic Study. Cancer Epidemiol Biomarkers Prev. 2011 Sep;20(9):1902–1911. PMCID: PMC3400264

20. Rissin DM, Kan CW, Campbell TG, Howes SC, Fournier DR, Song L, Piech T, Patel PP, Chang L, Rivnak AJ, Ferrell EP, Randall JD, Provuncher GK, Walt DR, Duffy DC. Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. Nat Biotechnol. 2010 Jun;28(6):595–599. PMCID: PMC2919230

21. Acar E, Papalexakis EE, Gürdeniz G, Rasmussen MA, Lawaetz AJ, Nilsson M, Bro R. Structure-revealing data fusion. BMC Bioinformatics. 2014 Dec;15(1):239. PMCID: PMC4117975

22. Acar E, Kolda TG, Dunlavy DM. All-at-once Optimization for Coupled Matrix and Tensor Factorizations. ArXiv11053422 Phys Stat [Internet]. 2011 May 17 [cited 2021 May 13]; Available from: http://arxiv.org/abs/1105.3422

23. Erbe R, Kessler MD, Favorov AV, Easwaran H, Gaykalova DA, Fertig EJ. Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell ATAC-seq data sets. Nucleic Acids Res. 2020 Jul 9;48(12):e68–e68. PMCID: PMC7337516

24. Tuncbag N, Braunstein A, Pagnani A, Huang S-SC, Chayes J, Borgs C, Zecchina R, Fraenkel E. Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem. J Comput Biol. 2013 Feb;20(2):124–136. PMCID: PMC3576906

25. Kelley DZ, Flam EL, Izumchenko E, Danilova LV, Wulf HA, Guo T, Singman DA, Afsari B, Skaist AM, Considine M, Welch JA, Stavrovskaya E, Bishop JA, Westra WH, Khan Z, Koch WM, Sidransky D, Wheelan SJ, Califano JA, Favorov AV, Fertig EJ, Gaykalova DA. Integrated Analysis of Whole-Genome

ChIP-Seq and RNA-Seq Data of Primary Head and Neck Tumor Samples Associates HPV Integration Sites with Open Chromatin Marks. Cancer Res. 2017 Dec 1;77(23):6538–6550. PMCID: PMC6029614

26. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018 Oct;34(10):790–805. PMCID: PMC6309559

27. Kossaifi J, Panagakis Y, Anandkumar A, Pantic M. TensorLy: Tensor Learning in Python. J Mach Learn Res. 2019;20(26):1–6.

28. Steinbrink JM, Myers RA, Hua K, Johnson MD, Seidelman JL, Tsalik EL, Henao R, Ginsburg GS, Woods CW, Alexander BD, McClain MT. The host transcriptional response to Candidemia is dominated by neutrophil activation and heme biosynthesis and supports novel diagnostic approaches. Genome Med. 2021 Dec;13(1):108. PMCID: PMC8259367

29. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec;9(1):559. PMCID: PMC2631488

30. Rajih M, Comon P, Harshman RA. Enhanced Line Search: A Novel Method to Accelerate PARAFAC. SIAM J Matrix Anal Appl. 2008 Jan;30(3):1128–1147.

31. Acar E, Dunlavy DM, Kolda TG, Mørup M. Scalable tensor factorizations for incomplete data. Chemom Intell Lab Syst. 2011 Mar;106(1):41–56.

32. Acar E, Dunlavy DM, Kolda TG. A scalable optimization approach for fitting canonical tensor decompositions. J Chemom. 2011 Feb;25(2):67–86.

33. Kaplonek P, Wang C, Bartsch Y, Fischinger S, Gorman MJ, Bowman K, Kang J, Dayal D, Martin P, Nowak RP, Villani A-C, Hsieh C-L, Charland NC, Gonye ALK, Gushterova I, Khanna HK, LaSalle TJ, Lavin-Parsons KM, Lilley BM, Lodenstein CL, Manakongtreecheep K, Margolin JD, McKaig BN, Rojas-Lopez M, Russo BC, Sharma N, Tantivit J, Thomas MF, Sade-Feldman M, Feldman J, Julg B, Nilles EJ, Musk ER, Menon AS, Fischer ES, McLellan JS, Schmidt A, Goldberg MB, Filbin MR, Hacohen N, Lauffenburger DA, Alter G. Early cross-coronavirus reactive signatures of humoral immunity against COVID-19. Sci Immunol [Internet]. 2021 Sep 9; Available from: https://www.science.org/doi/10.1126/sciimmunol.abj2901 PMCID: PMC8132219

34. Yuan B, Shen C, Luna A, Korkut A, Marks DS, Ingraham J, Sander C. CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy. Cell Syst. 2020 Dec;S2405471220304646. PMID: 33373583

35. Meyer AS, Zweemer AJM, Lauffenburger DA. The AXL Receptor Is a Sensor of Ligand Spatial Heterogeneity. Cell Syst. 2015 Jul 29;1(1):25–36. PMCID: PMC4520549

36. Sidje RB. Expokit: a software package for computing matrix exponentials. ACM Trans Math Softw. 1998 Mar;24(1):130–156.

37. Frostig R, Johnson MJ, Leary C. Compiling machine learning programs via high-level tracing. Stanford, CA; 2018. p. 3.

38. Salgado RC, Fonseca DLM, Marques AHC, da Silva Napoleao SM, França TT, Akashi KT, de Souza Prado CA, Baiocchi GC, Plaça DR, Jansen-Marques G, Filgueiras IS, De Vito R, Freire PP, de Miranda GC, Camara NOS, Calich VLG, Ochs HD, Schimke LF, Jurisica I, Condino-Neto A, Cabral-Marques O. The network interplay of interferon and Toll-like receptor signaling pathways in the anti-Candida immune response. Sci Rep. 2021 Dec;11(1):20281. PMCID: PMC8514550

39. Jiang P, Zhang Y, Ru B, Yang Y, Vu T, Paul R, Mirza A, Altan-Bonnet G, Liu L, Ruppin E, Wakefield L, Wucherpfennig KW. Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. Nat Methods. 2021 Oct;18(10):1181–1191. PMCID: PMC8493809

40. Meyer AS, Miller MA, Gertler FB, Lauffenburger DA. The Receptor AXL Diversifies EGFR Signaling and Limits the Response to EGFR-Targeted Inhibitors in Triple-Negative Breast Cancer Cells. Sci Signal. 2013;6(287):ra66–ra66. PMCID: PMC3947921

41. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2020 Sep 1;2(3):lqaa078. PMCID: PMC7518324

42. Gujral E, Pasricha R, Papalexakis EE. SamBaTen: Sampling-based Batch Incremental Tensor Decomposition. ArXiv170900668 Cs Stat [Internet]. 2017 Sep 18 [cited 2021 May 15]; Available from: http://arxiv.org/abs/1709.00668

43. Choi D, Jang J-G, Kang U. S3CMTF: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization. Wang J, editor. PLOS ONE. 2019 Jun 28;14(6):e0217316.

44. Sharan V, Valiant G. Orthogonalized ALS: A Theoretically Principled Tensor Decomposition Algorithm for Practical Use. Proc 34th Int Conf Mach Learn. 2017;70:3095–3104.

45. Cheng D, Peng R, Liu Y, Perros I. SPALS: Fast Alternating Least Squares via Implicit Leverage Scores Sampling. Adv Neural Inf Process Syst. Curran Associates, Inc.; 2016. p. 9.

46. Zhang X, Li L. Tensor Envelope Partial Least-Squares Regression. Technometrics. 2017 Oct 2;59(4):426–436.

47. Zhao Q, Caiafa CF, Mandic DP, Chao ZC, Nagasaka Y, Fujii N, Zhang L, Cichocki A. Higher-Order Partial Least Squares (HOPLS): A Generalized Multi-Linear Regression Method. IEEE Trans Pattern Anal Mach Intell. 2013 Jul;35(7):1660–1673.

48. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res. 2010 Mar 1;11:29.