Structured decomposition improves systems serology prediction and interpretation

This manuscript was automatically generated on July 19, 2021.

## Authors

* **Madeleine Murphy** · Github [murphymadeleine21](https://github.com/murphymadeleine21) Computational and Systems Biology, University of California, Los Angeles
* **Scott D. Taylor** · Github [scottdtaylor95](https://github.com/scottdtaylor95) Department of Bioengineering, University of California, Los Angeles
* **Zhixin Cyrillus Tan** ORCID [0000-0002-5498-5509](https://orcid.org/0000-0002-5498-5509) · Github [cyrillustan](https://github.com/cyrillustan) Bioinformatics Interdepartmental Program, University of California, Los Angeles
* **Aaron S. Meyer** ORCID [0000-0003-4513-1840](https://orcid.org/0000-0003-4513-1840) · Github [aarmey](https://github.com/aarmey) · twitter [aarmey](https://twitter.com/aarmey) Department of Bioengineering, University of California, Los Angeles; Department of Bioinformatics, University of California, Los Angeles; Jonsson Comprehensive Cancer Center, University of California, Los Angeles; Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles

## Results

### Systems serology measurements can be drastically reduced without loss of information

Figure 1: **Systems serology measurements have a consistent multi-dimensional structure.** 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 of the disease-specific antibody fraction that affect immune engagement, such as glycosylation, may be quantified in parallel in an antigen-specific or -generic manner. These measurements have been shown to predict both disease status (see methods) and immune functional properties—ADCD, ADCC, antibody-dependent neutrophil phagocytosis (ADNP), and natural killer cell activation measured by IFNγ, CD107a, and MIP1β expression. 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, antigen-generic 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. While the tensor and matrix differ in their dimensionality, they share a common subject dimension. C) The data is reduced by identifying additively-separable components represented by the outer product of vectors along each dimension. The subjects dimension is shared across both the tensor and matrix reconstruction. D) Venn diagram of the variance explained by each factorization method. Canonical polyadic (CP) decomposition can explain the variation present within either the antigen-specific tensor or glycan matrix on their own (Omberg *et al*, [2007](#ref-oqRObRv8)). CMTF allows one to explain the shared variation between the matrix and tensor (Choi *et al*, [2019](#ref-5XL7PCnt)). In contrast, here we wish to explain the total variation across both the tensor and matrix. This is accomplished with CMTF (see methods).

Figure 1: **Systems serology measurements have a consistent multi-dimensional structure.** 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 of the disease-specific antibody fraction that affect immune engagement, such as glycosylation, may be quantified in parallel in an antigen-specific or -generic manner. These measurements have been shown to predict both disease status (see methods) and immune functional properties—ADCD, ADCC, antibody-dependent neutrophil phagocytosis (ADNP), and natural killer cell activation measured by IFNγ, CD107a, and MIP1β expression. 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, antigen-generic 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. While the tensor and matrix differ in their dimensionality, they share a common subject dimension. C) The data is reduced by identifying additively-separable components represented by the outer product of vectors along each dimension. The subjects dimension is shared across both the tensor and matrix reconstruction. D) Venn diagram of the variance explained by each factorization method. Canonical polyadic (CP) decomposition can explain the variation present within either the antigen-specific tensor or glycan matrix on their own (Omberg *et al*, [2007](#ref-oqRObRv8)). CMTF allows one to explain the shared variation between the matrix and tensor (Choi *et al*, [2019](#ref-5XL7PCnt)). In contrast, here we wish to explain the total variation across both the tensor and matrix. This is accomplished with CMTF (see methods).

Figure 2: **CMTF improves data reduction of systems serology measurements.** A) Percent variance reconstructed (R2X) versus the number of components used in CMTF decomposition. B) CMTF reconstruction error compared to 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.

Figure 2: **CMTF improves data reduction of systems serology measurements.** A) Percent variance reconstructed (R2X) versus the number of components used in CMTF decomposition. B) CMTF reconstruction error compared to 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.

### Factorization accurately imputes missing values

Figure 3: **CMTF accurately imputes missing values.** A) Percent variance predicted (Q2X) versus the number of components used for imputation of 10 randomly held out receptor-antigen pairs. Lines indicate predictions with either antigen (red) or receptor (black) average. B) Percent variance predicted (Q2X) versus the number of components used for 10 randomly held out individual measurements. C) Percent variance predicted (Q2X) with increasing fraction of missing values. Error bars indicate standard deviation with repeated held-out sets.

Figure 3: **CMTF accurately imputes missing values.** A) Percent variance predicted (Q2X) versus the number of components used for imputation of 10 randomly held out receptor-antigen pairs. Lines indicate predictions with either antigen (red) or receptor (black) average. B) Percent variance predicted (Q2X) versus the number of components used for 10 randomly held out individual measurements. C) Percent variance predicted (Q2X) with increasing fraction of missing values. Error bars indicate standard deviation with repeated held-out sets.

### Structured data decomposition accurately predicts functional measurements and subject classes

Figure 4: **Structured data decomposition more accurately predicts functional measurements and subject classes.** (A) Accuracy of prediction (defined as the Pearson correlation coefficient) for different functional response measurements. (B) Prediction accuracy for subject viral and controller status. Model component effects for each function (E) and subject class (F) prediction. Component effects are quantified using the variable weights for a linear model, and the inverse RBF kernel length scale for a Gaussian process model (Paananen *et al*, [2019](#ref-pg5kOUJP)). For the Gaussian process component effects, the component effect is also multiplied by the sign of the corresponding linear model to show whether that variable has an overall positive or negative effect. The component effects are shown scaled to the largest magnitude within each model.

Figure 4: **Structured data decomposition more accurately predicts functional measurements and subject classes.** (A) Accuracy of prediction (defined as the Pearson correlation coefficient) for different functional response measurements. (B) Prediction accuracy for subject viral and controller status. Model component effects for each function (E) and subject class (F) prediction. Component effects are quantified using the variable weights for a linear model, and the inverse RBF kernel length scale for a Gaussian process model (Paananen *et al*, [2019](#ref-pg5kOUJP)). For the Gaussian process component effects, the component effect is also multiplied by the sign of the corresponding linear model to show whether that variable has an overall positive or negative effect. The component effects are shown scaled to the largest magnitude within each model.

### Factor components represent consistent patterns in the HIV humoral immune response

Figure 5: **Factor components represent consistent patterns in the HIV humoral immune response.** Decomposed components along subjects (A), receptors (B), antigens (C), and glycans (D). EC: Elite Controller, TP: Treated Progressor, UP: Untreated Progressor, VC: Viremic Controller (see methods). All plots are shown on a common color scale. Measurements were not normalized, and so magnitudes within a component are meaningful. Antigen names indicate both the protein (e.g., gp120, gp140, gp41, Nef, Gag) and strain (e.g., Mai, BR29).

Figure 5: **Factor components represent consistent patterns in the HIV humoral immune response.** Decomposed components along subjects (A), receptors (B), antigens (C), and glycans (D). EC: Elite Controller, TP: Treated Progressor, UP: Untreated Progressor, VC: Viremic Controller (see methods). All plots are shown on a common color scale. Measurements were not normalized, and so magnitudes within a component are meaningful. Antigen names indicate both the protein (e.g., gp120, gp140, gp41, Nef, Gag) and strain (e.g., Mai, BR29).

### Apply tensor factorization to human coronavirus (hCoV) systems serology

Figure 6: **Apply tensor factorization to human coronavirus (hCoV) systems serology.** (A) Schematics, (B) Variance explained. Factor components: (C) subjects, (D) antigens, (E) receptors, (F) weeks

Figure 6: **Apply tensor factorization to human coronavirus (hCoV) systems serology.** (A) Schematics, (B) Variance explained. Factor components: (C) subjects, (D) antigens, (E) receptors, (F) weeks

## References

Choi D, Jang J-G & Kang U (2019) S3CMTF: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization. *PLoS One* 14: e0217316

Omberg L, Golub GH & Alter O (2007) A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc Natl Acad Sci U S A* 104: 18371–6

Paananen T, Piironen J, Andersen MR & Vehtari A (2019) Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution arXiv

## Expanded View Figures

Figure EV1: **Decision boundaries of subject classification.** (A) Viremic and non-viremic decision is mostly dependent on components 2 and 4. (B) Controller and progressor on components 3 and 5.

Figure EV1: **Decision boundaries of subject classification.** (A) Viremic and non-viremic decision is mostly dependent on components 2 and 4. (B) Controller and progressor on components 3 and 5.

Figure EV2: **Demonstrating the instability of Alter et al’s method of elastic net prediction** Plots comparing the generated models of prediction from 3 identical trainings using elastic net, focusing on ADCD prediction (a-c). While one would expect similar models as they are all trained on the same dataset, the models vary significantly with respect to the number of receptors used, and their assigned values.

Figure EV2: **Demonstrating the instability of Alter et al’s method of elastic net prediction** Plots comparing the generated models of prediction from 3 identical trainings using elastic net, focusing on ADCD prediction (a-c). While one would expect similar models as they are all trained on the same dataset, the models vary significantly with respect to the number of receptors used, and their assigned values.

Figure EV3: **Simple methods for investigating gp120/p24 antigen ratio progression predictions yield different results based on IgG** A) Raw gp120/p24 measurement ratios against IgG, separated by subject class. Lines in boxes indicate median. The boxes show the quartiles of the dataset, while error bars indicate the rest of the distribution. Points indicate outlier points, which are determined by seaborn as a function of the inter-quartile range. B) Controller vs. Progressor prediction accuracy using gp120 and p24 measurements for each IgG. Predictions were done using logistic regression as described in methods. Accuracy is defined as classification accuracy. The prediction results vary based on which IgG is selected.

Figure EV3: **Simple methods for investigating gp120/p24 antigen ratio progression predictions yield different results based on IgG** A) Raw gp120/p24 measurement ratios against IgG, separated by subject class. Lines in boxes indicate median. The boxes show the quartiles of the dataset, while error bars indicate the rest of the distribution. Points indicate outlier points, which are determined by seaborn as a function of the inter-quartile range. B) Controller vs. Progressor prediction accuracy using gp120 and p24 measurements for each IgG. Predictions were done using logistic regression as described in methods. Accuracy is defined as classification accuracy. The prediction results vary based on which IgG is selected.