Integrative Analysis of Epigenetic, Transcriptomic, and Metabolomic Responses to Arsenic Exposure Using Coupled Matrix Factorization

Sujit Silas Armstrong Suthahar

University of California, Los Angeles

1 Introduction and Motivation

Arsenic (As), a naturally occurring element, poses significant toxicity risks in its inorganic form (iAs). It is a pervasive environmental toxin associated with severe health implications, including cancer, cardiovascular disease, and endocrine disruption [1]. Its global distribution and bioaccumulation make it a critical public health concern. While extensive research has been conducted to study arsenic-induced toxicity, the mechanisms by which it influences molecular pathways remain poorly understood. This complexity arises from arsenic's ability to induce widespread epigenetic reprogramming, alter gene expression, and disrupt metabolic homeostasis.

Despite advances in high-throughput techniques, existing studies often analyze epigenomic, transcriptomic, and metabolomic data independently. This approach overlooks the intricate interplay among these layers of regulation. Integrating such multi-omics data provides an opportunity to uncover novel regulatory mechanisms underlying arsenic toxicity. Previous works have explored matrix factorization for dimensionality reduction and integration, but few have applied coupled matrix factorization (CMF) to simultaneously model diverse omics datasets [2].

Previous studies have explored similar computational approaches for integrative data analysis but have not used the same methods. Bagherian et al. developed coupled matrix–matrix and coupled tensor–matrix completion methods to predict drug–target interactions [3]. Similarly, Erbe et al. applied the CoGAPS matrix factorization algorithm and transfer learning with projectR to integrate single-cell ATAC-seq datasets to uncover common regulatory patterns across datasets [4]. Duren et al. leverages coupled non-negative matrix factorization for the joint analysis of scRNA-seq and scATAC-seq data, revealing the underlying cell types in heterogeneous populations [5]. Collectively, these papers address the potential of matrix factorization and coupled methods for integrating single-cell with other modalities but not RNA-seq, reduced representation bisulfite sequencing (RRBS), and metabolics. In our study, we apply CMF using the parafac2_aoadmm model to integrate and analyze RRBS, RNA-seq, and metabolomics data from mouse ESCs and EpiLCs treated with arsenic [6].

2 Problem Definition

This study aims to investigate how arsenic exposure influences molecular pathways across the epigenome, transcriptome, and metabolome. Specifically, we seek to uncover general trends in global methylation dysregulation that drive transcriptional and metabolic changes in ESCs and EpiLCs. Additionally, we aim to identify distinct regulatory networks and mechanisms underlying arsenic toxicity.

To address this as a data analytics problem, we employed CMF, which enables the joint factorization of datasets sharing common features (columns) but differing in dimensions (row counts). This approach can model arsenic's effects as a coupled decomposition problem, where each dataset can provide unique yet complementary insights. The broader objective is to link these molecular alterations to the pathophysiological outcomes associated with arsenic-related diseases, including cancer and developmental disorders.

3 Methods

3.1 Algorithm Description

The analysis was conducted using the `parafac2_aoadmm` model, a specialized implementation of Coupled Matrix Factorization (CMF) [2]. CMF is designed to jointly factorize multiple matrices that share column features but differ in row dimensions. The model follows the formulation:

$$X^{(i)} \approx B^{(i)} D^{(i)} C^T$$

where $X^{(i)}$ represents the i-th input matrix (e.g., RRBS, RNA-seq, or metabolomics), $B^{(i)}$ is a collection of factor matrices for each dataset capturing sample-specific patterns, $D^{(i)}$ consists of diagonal matrices indicating the signal strength of each dataset, and C is a shared factor matrix representing common features across the datasets.

To achieve this decomposition, PARAFAC2 relies on alternating optimization combined with the alternating direction method of multipliers (AO-ADMM) [6]. To fit a coupled matrix factorization for integrating **RRBS**, **RNA-seq**, and **metabolomics** data, we solve the following optimization problem:

$$\min_{\mathbf{A}, \{\mathbf{B}^{(i)}\}_{i=1}^{3}, \mathbf{C}} \frac{1}{2} \sum_{i=1}^{3} \frac{\|\mathbf{B}^{(i)} \mathbf{D}^{(i)} \mathbf{C}^{\top} - \mathbf{X}^{(i)}\|^2}{\|\mathbf{X}^{(i)}\|^2},$$

where $\mathbf{A}$ is the matrix constructed by stacking the diagonal entries of all $\mathbf{D}^{(i)}$-matrices ($i = 1, 2, 3$ corresponding to **RRBS**, **RNA-seq**, and **metabolomics** data, respectively). However, this optimization problem does not yield a unique solution, as fitting coupled matrix factorization can produce different factor matrices. This makes it challenging to interpret the factor matrices directly.

To address this, we introduce regularization terms, leading to the following revised optimization problem:

$$\min_{\mathbf{A}, \{\mathbf{B}^{(i)}\}_{i=1}^{3}, \mathbf{C}} \frac{1}{2} \sum_{i=1}^{3} \frac{\|\mathbf{B}^{(i)} \mathbf{D}^{(i)} \mathbf{C}^{\top} - \mathbf{X}^{(i)}\|^2}{\|\mathbf{X}^{(i)}\|^2} + \sum_{n=1}^{N_A} g_n^{(A)}(\mathbf{A}) + \sum_{n=1}^{N_B} g_n^{(B)}(\{\mathbf{B}^{(i)}\}_{i=1}^{3}) + \sum_{n=1}^{N_C} g_n^{(C)}(\mathbf{C}),$$

where the g-functions represent regularization penalties, and N_A, N_B, and N_C denote the number of regularization terms applied to $\mathbf{A}$, $\{\mathbf{B}^{(i)}\}_{i=1}^{3}$, and $\mathbf{C}$, respectively.

3.2 Software Implementation

We implemented the parafac2_aoadmm model using the MatCouply library in Python, which supports flexible CMF implementations [2]. The existing implementation provided by MatCouply was customized to handle the decomposition of RRBS, RNA-seq, and metabolomics data. Additional visualization modules were developed to interpret the factor matrices ($B^{(i)}, D^{(i)}, C$), enabling insights into the underlying patterns of the data. Additionally, the L1 norm was enforced to introduce a sparsity constraint to help with interpretability by reducing the complexity of the model. Non-negativity constraints were explicitly disabled to explore a broader solution space and the nature of the data. The maximum number of iterations was set to 100, ensuring adequate computational time for the algorithm to converge to a stable solution. The complete code implementation has been uploaded to GitHub for reproducibility purposes.

3.3 Data and Preprocessing

Each dataset was normalized according to its specific requirements. For the RRBS data, beta values (methylation values normalized for coverage) were obtained. RNA-seq data was processed to produce transcripts per million (TPM) normalized counts, and the metabolomics data was normalized using total ion count. Rows in each dataset represent molecular entities (e.g., CpG sites, genes, or metabolites), while columns correspond to experimental conditions or replicates. Additionally, genes with low expression levels and genomic regions with low-coverage methylation were filtered out to reduce analytical noise.

1. **RRBS (414,234 × 12)**: Profiles DNA methylation patterns in ESCs and EpiLCs under arsenic treatment at a resolution of CpG sites.

2. **RNA-seq (26,210 × 12)**: Quantifies transcriptomic changes in response to arsenic exposure.

3. **Metabolomics (137 × 12)**: Measures metabolic fluxes in treated and control samples.

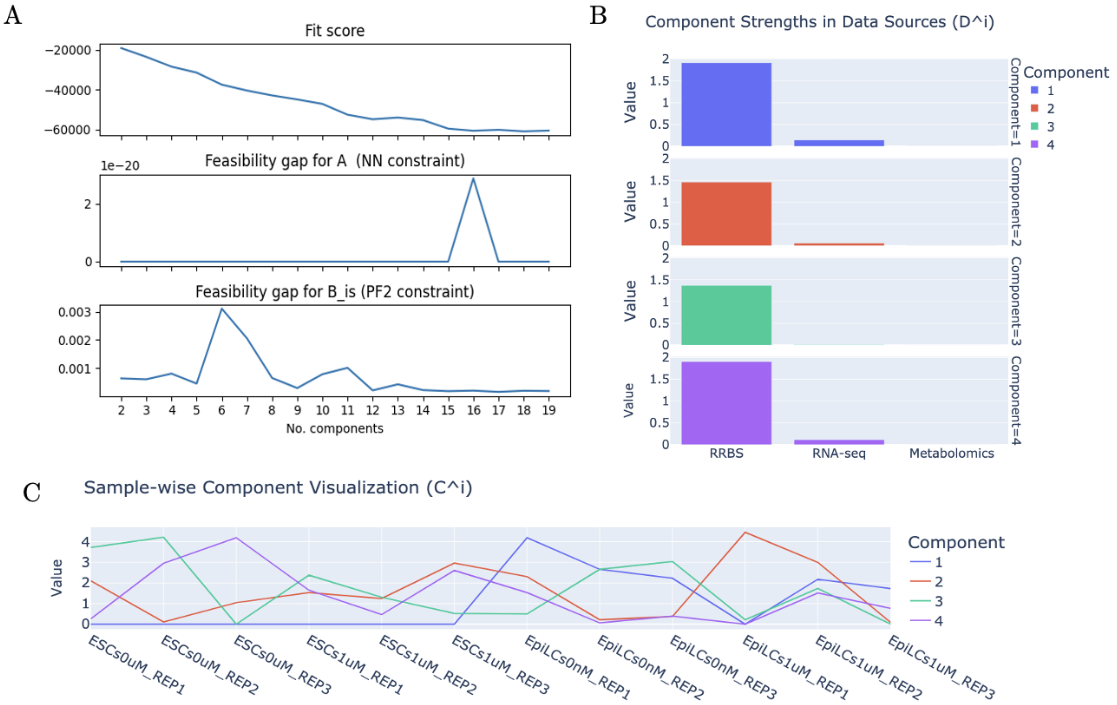

Figure 1: **A.** Model fitting performance for `parafac2_aoadmm` model across varying component numbers, including the fit score and feasibility gaps for constraints. **B.** Component strengths derived from RRBS, RNA-seq, and metabolomics datasets, highlighting the contribution of each data source to distinct components. **C.** Sample-wise visualization of component values across replicates and experimental conditions, illustrating trends in the underlying global structure of the data.

4 Results and Discussion

Using Matcouply's `parafac2_aoadmm` model, we determined that a model with three components provided the best fit. This decision was based on evaluating fit scores (analogous to the sum of squared errors) and feasibility gaps, which assess whether the model successfully enforces constraints (Figure 1A). As the number of components increased, the model's ability to reconstruct the original matrix diminished, as indicated by decreasing fit scores. Additionally, fluctuating feasibility gap scores and instability were observed, particularly when fitting 5–9 and 15–17 components. Based on these observations, we decided to proceed with a three-component model. The decomposed matrices were further analyzed by plotting them individually. The $\mathbf{D}^{(i)}$ matrix, representing the weights, was initially visualized to examine the weight assigned to each modality. The bar plot revealed that the model consistently prioritized RRBS over the other modalities, with RNA-seq data receiving relatively lower weightage for some components (Figure 1B). Previous studies have highlighted the significance of epigenetic changes in understanding the regulatory landscape [11]. Consistent with these findings, our model effectively captured such patterns.

Next, we investigated the shared factor matrix $\mathbf{C}^{T}$ by plotting its values across all samples (ESCs and EpiLCs) for both treated and untreated conditions to observe general trends in the first four components (Figure 1C). Initial observations revealed discernible trends captured by the model. For instance, global hypomethylation was observed in ESCs compared to EpiLCs, irrespective of arsenic treatment, suggesting a mechanism to preserve their epigenomic landscape and maintain pluripotency under arsenic exposure [7]. Independent analyses of the RRBS data and line plots of the sharred factor matrix confirmed that this trend was captured by the first component. Similarly, components 2 and 3 revealed additional methylation trends of interest.

We examined the unique factor matrix $\mathbf{B}^{(i)}$ for each datatype. Given the size of each matrix, we narrowed our focus to the top 25 CpG sites, genes, or metabolites by identifying the most variable features across the four components. Several interesting relationships were uncovered, aligning with existing literature. For instance, low levels of 5-oxoproline is known to be an indicator of arsenic exposure. This appears to be a highly variable element in the metabolite factor matrix [8]. *Gapdh*, a housekeeping gene

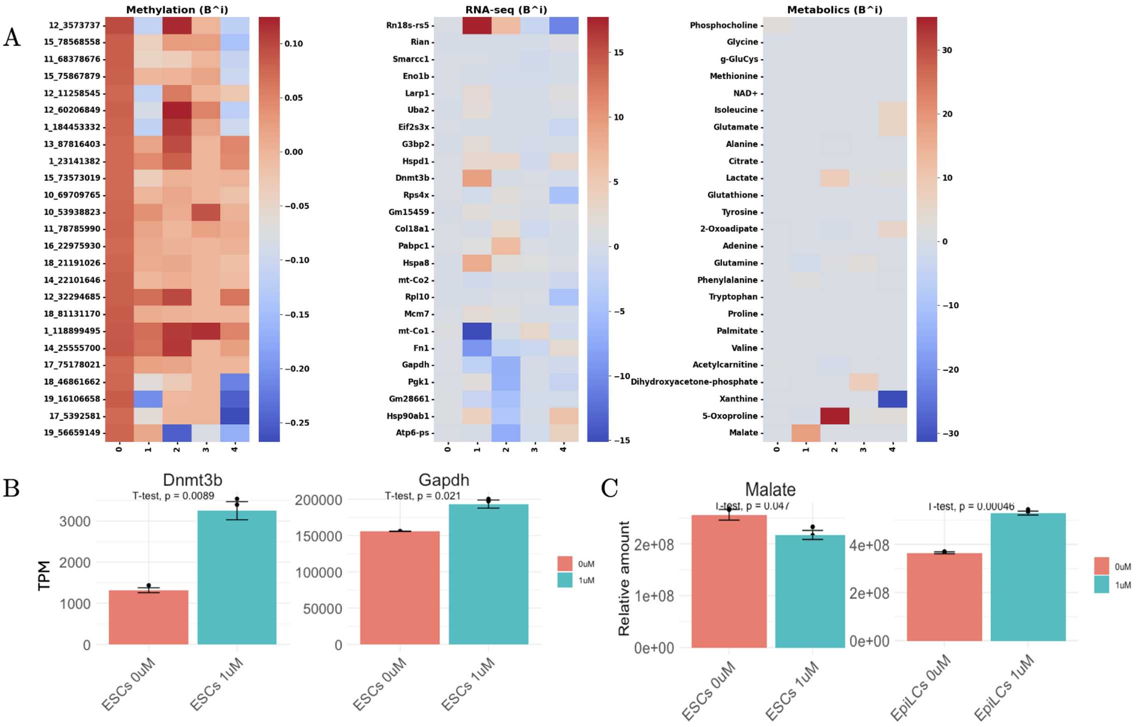

Figure 2: **A.** Heatmaps showcasing the top 25 variable elements from the $\mathbf{B}^i$ matrix for the methylation data (left), RNA-seq data (middle), and metabolomics data (right). **B.** Transcript quantification for *Dnmt3b* and *Gapdh* across conditions in transcripts per million (TPM). A simple two sample t-test indicates statistical significance. **C.** Relative abundance of malate across conditions. Again, a two samples t-test shows statistical significance.

associated with arsenic resistance, was identified to be highly variable in the factor matrix, and its TPM levels were evaluated [9]. Additionally, *Dnmt3b* , a gene crucial for establishing embryonic methylation patterns and implicated in tumor occurrence, was highlighted by the analysis and was further evaluated by looking at the TPM counts [10]. These findings demonstrate the model's ability to capture biologically relevant patterns and provide insights into the underlying regulatory mechanisms.

Overall, this study demonstrates the utility of CMF in integrating multi-omics data to elucidate the effects of arsenic exposure. By capturing shared and unique features across RRBS, RNA-seq, and metabolomics datasets, CMF provides a comprehensive view of arsenic-induced toxicity. Due to limited time and computational resources, we did not quantitatively evaluate our model's performance using cross-validation. Moving forward, we aim to address this limitation by leveraging UCLA's high-performance computing resources for a more robust evaluation of the model, as the random initialization of matrix factorization techniques can lead to varied results. Future work will also focus on annotating CpG sites and exploring their functional implications through experimental validation. This research underscores the potential of CMF for integrative omics analyses, paving the way for novel discoveries in toxicology and systems biology.

4.1 References

1. Mohammed Abdul, K. S., Jayasinghe, S. S., Chandana, E. P. S., Jayasumana, C., De Silva, P. M. C. S. (2015). Arsenic and human health effects: A review. *Environmental Toxicology and Pharmacology*, *40*(3), 828–846. https://doi.org/10.1016/j.etap.2015.09.016

2. Roald, M. (2023). MatCoupLy: Learning coupled matrix factorizations with Python. *SoftwareX*, *21*, 101292. https://doi.org/10.1016/j.softx.2022.101292

3. Bagherian, M., Kim, R. B., Jiang, C., Sartor, M. A., Derksen, H., Najarian, K. (2021). Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions. *Briefings in Bioinformatics*, *22*(2), 2161–2171. https://doi.org/10.1093/bib/bbaa025

4. Erbe, R., Kessler, M. D., Favorov, A. V., Easwaran, H., Gaykalova, D. A., Fertig, E. J. (2020). Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell ATAC-seq data sets. *Nucleic Acids Research*, *48*(12), e68. https://doi.org/10.1093/nar/gkaa349

5. Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., Wang, Y., Wong, W. H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, *115*(30), 7723–7728. https://doi.org/10.1073/pnas.1805681115

6. PARAFAC2-based Coupled Matrix and Tensor Factorizations with Constraints. (n.d.). Retrieved November 13, 2024, from https://arxiv.org/html/2406.12338v1

7. Pecori, F., Yokota, I., Hanamatsu, H., et al. (2021). A defined glycosylation regulatory network modulates total glycome dynamics during pluripotency state transition. *Scientific Reports*, *11*, 1276. https://doi.org/10.1038/s41598-020-79666-4

8. Gasser, M., Lenglet, S., Bararpour, N., Sajic, T., Vaucher, J., Wiskott, K., Augsburger, M., Fracasso, T., Gilardi, F., Thomas, A. (2023). Arsenic induces metabolome remodeling in mature human adipocytes. *Toxicology*, *500*, 153672. https://doi.org/10.1016/j.tox.2023.153672

9. Némethi, B., Csanaky, I., Gregus, Z. (2006). Effect of an inactivator of glyceraldehyde-3-phosphate dehydrogenase, a fortuitous arsenate reductase, on disposition of arsenate in rats. *Toxicological Sciences*, *90*(1), 49–60. https://doi.org/10.1093/toxsci/kfj058

10. Okano, M., Bell, D. W., Haber, D. A., Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257. https://doi.org/10.1016/s0092-8674(00)81656-6

11. Xu, W., Xu, M., Wang, L. et al. Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. Sig Transduct Target Ther 4, 55 (2019). https://doi.org/10.1038/s41392-019-0081-6