# Tensor-Based Analysis of High-Order Chromatin Interactions in Pore-C Data

By Will Lounsbery-Scaife

Columbia University, Department of Biomedical Informatics

New York Genome Center

Gursoy Lab

## I. Abstract

The three-dimensional structure of the genome within the nucleus plays an important role in regulating gene expression. Traditional assays for studying 3D chromatin structure, such as Hi-C, have improved our understanding of chromatin features like topologically associated domains (TADs), CTCF loops, and A/B compartments. However, these assays predominantly capture pairwise interactions between genomic loci, limiting their usefulness in studying high-order chromatin interactions. Pore-C enables the profiling of proximal high-order interactions by combining chromatin conformation capture with nanopore sequencing, but the matrix-based methods used on Hi-C data do not extend to Pore-C data. In this study, we demonstrate a new method to efficiently analyze Pore-C data by representing reads in multidimensional tensors and applying Canonical Polyadic Decomposition (CPD) to extract biologically meaningful insights from these tensors. Our work here serves as an initial step in demonstrating how CPD can be applied to Pore-C tensors, and we hope that future research can build upon our methodology in order to develop a comprehensive approach to analyzing Pore-C data.


## II. Introduction

The spatial configuration of the genome within the cell nucleus is a critical factor in gene regulation, dictating how genetic instructions are accessed and executed (Dekker et al., 2012; Misteli et al., 2007). The arrangement of the 3D genome represents a dynamic interplay between DNA, RNA, and various proteins, resulting in a complex network of chromatin interactions. Techniques such as Chromatin Conformation Capture (3C) and Hi-C (Dekker et al., 2002; Lieberman et al., 2009) have shed light on features such as topologically associated domains, or TADS (Dixon et al. 2012), chromatin loops mediated by CCCTC-binding factor (CTCF) (Rao et al., 2014), and the partitioning of the genome into active (A) and inactive (B) compartments (Lieberman et al., 2009, Fortin et al., 2015). However, such features represent only a fraction of the chromatin interaction landscape.

Previous methods for studying chromatin conformation have primarily captured pairwise interactions. In a typical 3C experiment, after DNA is cross-linked, cut, and ligated, the ligation products represent interactions between two DNA fragments. Thus, most chromatin conformation data is limited in its ability to reveal complex, multi-locus interactions within the

3D genome. The advent of Pore-C, an approach that combines chromatin conformation capture techniques with the high-throughput capabilities of nanopore sequencing, improves our ability to study higher-order interactions (Deshpande et al., 2022). Pore-C is able to profile proximal chromatin contacts (contacts within a 200-nm contact radius) between multiple genomic loci at the genome scale, thus enabling the sequencing of reads that contain multiple DNA fragments, and providing a more holistic view of 3D chromatin structure. However, to take advantage of this new sequencing modality, new analytical methods capable of capturing the high-order nature of Pore-C reads must be developed.

To address this need, we propose a tensor-based method for efficiently analyzing the Pore-C data. Inspired by matrix-based methods for analyzing Hi-C data, our approach is capable of handling multi-way chromatin interactions in addition to simple pairwise interactions. Tensors, or multidimensional matrices, are adept at representing Pore-C data, preserving the intricate network of chromatin interactions in a structured format. By representing Pore-C data in higher-order tensors and applying tensor decomposition methods, we can transform these complex data sets into biologically interpretable formats.

In this study, we employ Canonical Polyadic Decomposition (CPD) (Rabanser et al., 2017) to decompose, or factorize, tensors constructed from Pore-C reads. CPD, a technique grounded in multilinear algebra, factorizes a tensor into a series of rank-one tensors. Analyzing these lower-rank factor tensors has the potential to reveal new insights on how chromatin conformation informs gene regulation and function. Similar to how dimensionality reduction techniques such as PCA, SVD, and NMF, when applied to Hi-C matrices, can reveal structural features such as TADs and A/B compartments (Prive et al., 2018; Wall et al., 2003), we propose that tensor decomposition is a viable approach for extracting novel biological insights from Pore-C data. By incorporating high dimensional data that captures multi-way chromatin interactions, our tensor-based approach should be able to expand our knowledge of the 3D structure of the genome beyond what is possible with traditional matrix decomposition methods. To ensure the validity of our approach, we initially applied Eigendecomposition using CoolTools (Abdennur et al., 2022) to two-dimensional Pore-C matrices containing only pairwise interactions, thereby creating a baseline by mirroring existing matrix decomposition approaches for analyzing traditional Hi-C interaction matrices. We further annotated these 2D Pore-C interaction matrices with CTCF enrichment and histone modification peaks. Subsequently, we applied CPD to higher-dimensional tensors containing multi-way interactions.

Our work here serves as an initial step in demonstrating how CPD can be applied to Pore-C tensors, and we hope that future research can build upon our methodology in order to develop a comprehensive approach to analyzing Pore-C data.

## II. Background

Before proceeding with a discussion of our tensor methodology, let us first define several important terms in the context of Pore-C experiments.

(1) **Read:** In genomic sequencing, a read refers to the sequence of nucleotides obtained from a single sequencing event. In the case of Pore-C, a read encompasses the sequence data derived from concatenated DNA fragments reflecting chromatin interactions.

(2) **Cardinality**: Refers to the number of distinct genomic loci represented within a single read. It indicates the complexity of the interaction captured, with higher cardinalities pointing to more intricate chromatin interactions.

(3) **Concatemer:** A concatemer in Pore-C is a long DNA molecule formed by the joining of multiple chromatin fragments. These concatemers are the result of proximity ligation events during the chromatin conformation capture process and are essential for revealing high-order chromatin interactions.

(4) **Fragment:** A fragment is a section of DNA that has been isolated and identified during sequencing. In Pore-C, fragments within a concatemer represent individual chromatin segments that have been brought into proximity and ligated together.

Chromunity, introduced in the Pore-C methodology (Deshpande et al., 2022), identifies sets of genomic loci, or bin-sets, that appear together within high-order contacts at frequencies significantly higher than background bin-sets. A group of genomic loci belonging to the same “Chromunity” implies that the chromatin regions to which these loci belong are participating in cooperative interactions across the genome. Such cooperativity could play a significant role in gene regulation and the overall functioning of the cell. However, Chromunity has certain limitations in its application and scope. Firstly, the method's computational complexity becomes a significant challenge when dealing with very large datasets. The process of detecting

concatemer communities in Chromunity involves constructing a graph with quadratic complexity ($O(|\tilde{D}|^2)$), making it computationally demanding, especially for datasets of considerable size (e.g., $|\tilde{D}| > 10^7$). This complexity limits the portion of the dataset that can be effectively used for community detection and may subsequently restrict the power for discovering new synergies. Secondly, while Chromunity excels in identifying synergistic interactions within specific genomic regions, it may not capture the full spectrum of chromatin's 3D structure, especially when it comes to non-synergistic interactions that are also biologically significant.

The tensor-based approach introduced here seeks to improve our ability to effectively analyze Pore-C data by addressing some of the limitations of Chromunity. By representing chromatin interactions as multidimensional tensors and applying Canonical Polyadic Decomposition (CPD), this research demonstrates an efficient and scalable alternative for Pore-C analysis compared to the original Chromunity method developed by Deshpande et al (2022).

**Tensors**

A Pore-C tensor is a multi-dimensional array that represents chromatin interactions across various genomic loci. The equations and definitions described in this section are paraphrased from Rabanser et al. (2017), as well as from W. H. Greub's textbook *Linear Algebra* (2012). Refer to these sources for a more thorough explanation of the concepts discussed below.

A tensor product space $V \otimes W$ is a mathematical construct formed by combining two vector spaces $V$ and $W$, resulting in a new vector space where each element represents all possible combinations of pairs of vectors from $V$ and $W$. The dimensionality of this new space is the product of the dimensions of $V$ and $W$, allowing for the exploration of complex, multidimensional relationships.

Mathematically, a tensor $T$ in $N$-dimensions can be seen as a generalization of matrices to higher dimensions:

$$T = \sum_{i_1,i_2,\ldots,i_k=1}^{n_1,n_2,\ldots,n_k} t_{i_1 i_2 \ldots i_k}(e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}) \quad (1)$$

In the context of our Pore-C tensors, each element $t_{i1}, t_{i2}, \ldots, t_{iN}$ quantifies the interaction strength between genomic loci and $e_{ij}$ are basis vectors. For two-dimensional chromatin interactions, such as those observed in traditional Hi-C assays, we can simplify the tensor model to matrices, where a matrix $M$ represents pairwise interactions. In this case, $M$ can be represented as:

$$M = \sum_{i,j=1}^{n,m} m_{ij}(e_i \otimes e_j) \tag{2}$$

This matrix representation is useful for identifying pairwise genomic interactions but lacks the depth provided by higher-dimensional tensor analysis. To further understand the underlying genomic structure, we employ matrix diagonalization and eigen-decomposition techniques. Matrix diagonalization facilitates the simplification of matrices into their constituent parts, revealing underlying patterns in genomic interactions.

$$M = QDQ^T \tag{3}$$

Similarly, eigen-decomposition, allows us to explore these interactions in terms of eigenvalues and eigenvectors, offering insights into the dominant interaction patterns.

$$M = \sum_{i=1}^{r} \lambda_i v_i v_i^T \tag{4}$$

Applying Canonical Polyadic Decomposition (CPD) to Pore-C tensors presents a promising avenue for efficiently analyzing higher-order reads captured by this new data modality. By decomposing a tensor $T$ as:

$$T = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r \tag{5}$$

We aim to unravel high-order chromatin interactions. This decomposition simplifies the data analysis as well as potentially uncovers multifaceted interaction patterns, contributing to a more comprehensive understanding of genomic regulation.

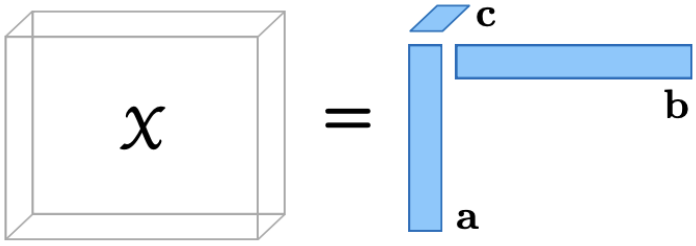


**Figure 1.** A rank-1 mode-3 tensor, where each element is the product of corresponding scalar components from vectors across three dimensions.

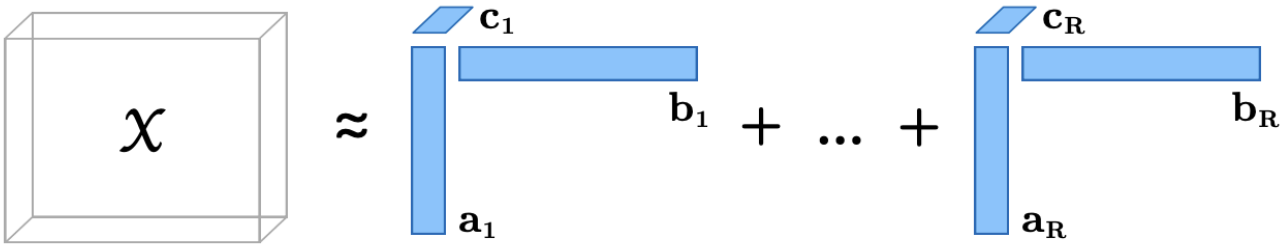


**Figure 2.** Illustration of Canonical Polyadic Decomposition (CPD) for a third-order tensor from Rabanser et al. (2017). The tensor $\mathcal{X}$ is approximated as a sum of rank-one tensors, represented here by the outer products of vectors $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$. Mathematically, the CPD is expressed as $\mathcal{X} \approx \sum r = 1^R \lambda_r \cdot (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$, where *R* is the rank of the decomposition and $\lambda_r$ are the scaling factors for each rank-one tensor.

## III. Materials and Methods

### Creating tensors

The Pore-C data used in our tensors was derived from cell line GM12878 with restriction enzyme NlaIII. Each tensor represents all reads from a single chromosome; no inter-chromosomal fragments were included in any tensor. Chromosomes were segmented into 100kb (kilobase) genomic bins, and the fragments in each read were mapped by their midpoints to the corresponding genomic bin. Reads containing fragments from multiple chromosomes were broken up into single-chromosome subreads. For example, if a read of cardinality 4 contained two fragments from chromosome one and two fragments from chromosome two, this read would be treated as two cardinality-2 reads. If two fragments within a read mapped to the same bin, only one of these fragments was kept. Only fragments that passed the filtering algorithm developed by Deshpande et al. (2022) were included in each tensor. Another inclusion criterion was that a read must have a cardinality of at least two for the tensor's corresponding chromosome; if a read contains only one fragment that maps to a given chromosome, the read is not included in the corresponding tensor.

| | Read ID | Chrom | Pass filter | Frag ID | Frag Start | Frag End | Bin |
|---|---|---|---|---|---|---|---|
| ✓ | 74 | chr1 | True | 470239 | 210,002,811 | 210,004,521 | 5251 |
| ⊘ | 74 | **chr5** | True | 474320 | 88,310,645 | 88,310,947 | - |
| ⊘ | 74 | chr1 | True | 470266 | 210,005,240 | 210,006,101 | **5251** |
| ⊘ | 74 | chr1 | **False** | 471851 | 112,397,968 | 112,399,849 | - |
| ✓ | 74 | chr1 | True | 471835 | 192,541,351 | 192,541,964 | 4814 |
| ✓ | 74 | chr1 | True | 474333 | 60,202,479 | 60,204,325 | 1506 |

**Table 1.** An example of a 6-cardinality read with fragments mapping to chromosomes one and five. Read 74 is stored as a 3-cardinality read in the chromosome one tensor, at coordinates (1506, 4814, 5251). The fragment that fails to pass the filtering algorithm, the duplicate fragment mapping to bin 5251, and the fragment mapping to chromosome five are discarded. For Read 74 to be included in the chromosome five tensor, at least one additional fragment would need to map to chromosome 5.

The number of dimensions in a tensor determines the maximum cardinality that the tensor is capable of storing; a 4D tensor can store reads with cardinalities of 2, 3, and 4, while a 3D tensor can store reads with cardinalities of 2 and 3. The dimensions in a tensor are symmetric, with the size of each dimension depending on the number of bins that fit in that tensor's chromosome. For example, with 100kb bins, a 3D tensor for chromosome 21 would have the shape $468 \times 468 \times 468$. Ideally, we would want to create tensors that can store cardinalities beyond just three or four; Pore-C data contains some high-order reads with cardinalities exceeding 20. However, with each additional dimension in the tensor, the memory required to store the tensor increases exponentially. Higher-dimensional tensors are also *much* more sparse than their lower-dimensional counterparts.

To understand how each read is stored in a tensor, consider a scenario involving a read with a cardinality of 2, with fragments mapping to bins 5 and 8 on chromosome 21. When this read is processed by our Python script, the tensor coordinates (5, 8, 0, 0) and all permutations thereof would be incremented by one. That is, the values stored at (5, 8, 0, 0), (5, 0, 8, 0), (8, 5, 0, 0), and so on, would all be incremented by one. Similarly, a read with a cardinality of 3, mapping to bins 3, 4, and 8, would have its corresponding coordinates (3, 4, 8, 0) incremented, along with all permutations. This symmetrical tensor construction is analogous to the symmetry that can be seen in Hi-C matrices.

**Data**

To determine which reads within Pore-C data are especially informative or interesting, we investigated the presence of "subread support" for various reads across our Pore-C tensors. This metric evaluates the quality of a high-cardinality read based on the presence of its lower cardinality subreads within the data. For example, consider a cardinality-4 read (Read-A) with fragments mapping to bins 13, 14, 15, and 25. Any read with a lower cardinality that contains a subset of these fragments would be considered a subread of Read-A. For example, a cardinality-3 read with fragments mapping to bins 13, 15, and 25 would be considered a subread of Read-A. A lack of subread support for any given high-order read suggests that the read may be considered low quality. Such reads may contain more noise than signal; their presence may be attributable to artifacts in the Pore-C data collection process, rather than actual trends in chromatin structure. Conversely, if a higher-order read has a high degree of subread support, this implies that the read is of higher quality and that it is more likely to contain biologically interesting information. The concept of subread support could be useful for future filtering strategies aiming to enhance data quality by ensuring that only biologically significant reads are stored in the tensors.

We found that high-order reads with higher frequencies tended to have greater subread support than high-order reads with low frequencies (Figure 3). This suggests that high-order reads with low frequency (particularly those that occur only once) may be present due to chance rather than due to any interesting biological phenomena. Thus, when creating our tensors, we decided to only include reads that exceeded some minimum frequency threshold. Performing this frequency filter leads to higher quality, albeit more sparse, tensors.

To determine which reads are more informative or interesting, subread support is not the only metric we should consider. In general, we expect regions of chromatin that are close together in one-dimensional space to appear together more frequently than distant regions. If regions that are distant in one-dimensional space repeatedly appear within the same reads, this could indicate that these distant regions are interacting in a biologically meaningful way. In particular, we are interested in genomic regions that, while distant from each other in 1D space, appear in reads at a disproportionately higher rate than what we would expect given their distance from each other.

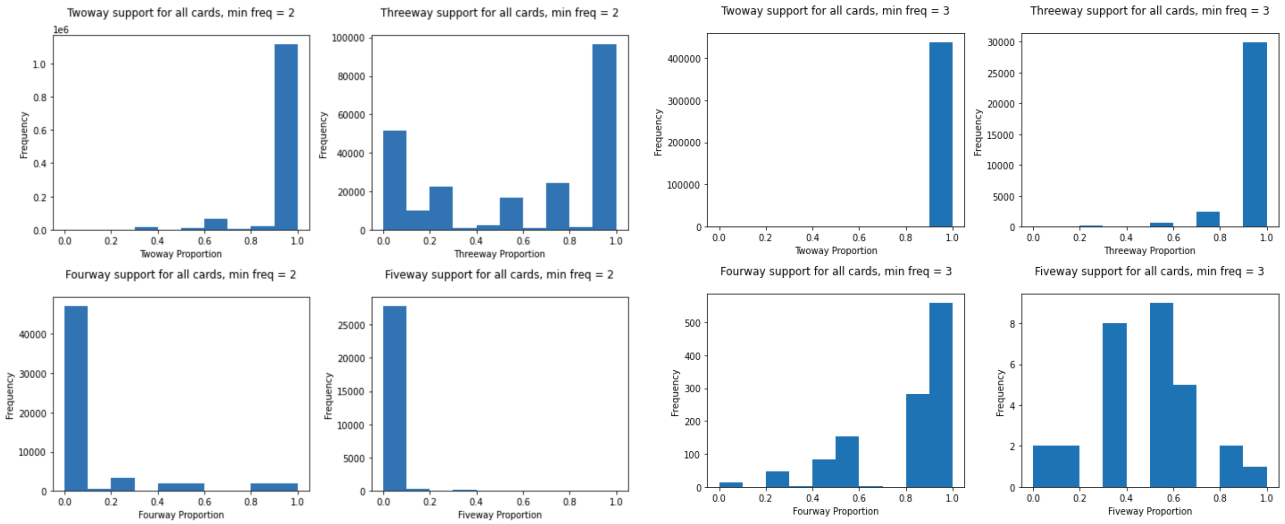


**Figure 3:** Subread support frequency for Pore-C interactions. Histograms show subread support across two-way to five-way interactions at minimum frequencies of 2 and 3. Subread support reflects validation of higher-cardinality interactions by lower-cardinality ones. Higher minimum frequency thresholds demonstrate increased support.

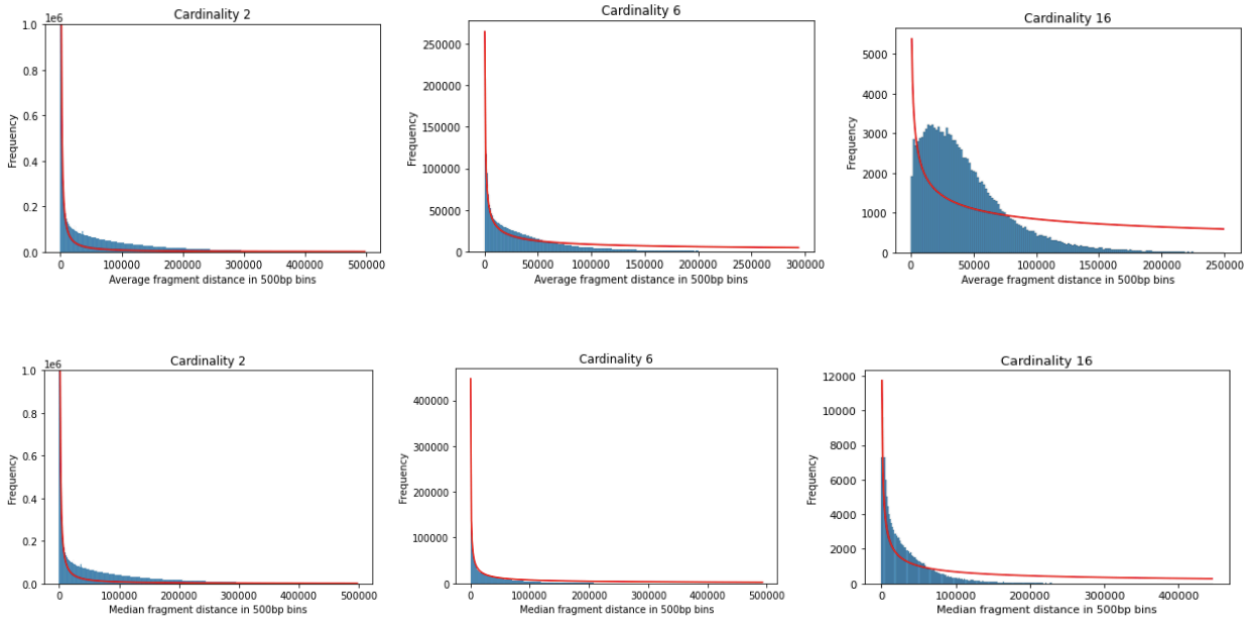


**Figure 4:** Frequency distributions of inter-fragment distances within Pore-C reads for cardinalities 2, 6, and 16. Top and bottom rows represent average and median distances, respectively, within 500kb bins. The rapid decline in frequency suggests a dominance of short-range over long-range interactions.

**Validation with Low-Cardinality Data: Pore-C Matrices**

To validate our tensor methodology for Pore-C data, we initially concentrated on cardinality-2 data to generate matrices akin to Hi-C. This step allowed to create interaction matrices comparable to matrices to established Hi-C matrices. First, we created balanced '.cool' files from our Pore-C tensors with Cooler (Abdennur et al., 2020), using default normalization settings for the matrix balancing. Then, using CoolTools, we applied eigendecomposition to our Pore-C matrices, isolating primary interaction patterns and structural domains indicative of chromatin's three-dimensional organization. This process dimensionality reduction technique highlights the main axes of variation and compartmentalization within the chromosomal landscape. The first eigenvector was used to annotate the interaction matrix with the dark horizontal and vertical lines that appear on the Pore-C interaction matrix. These lines, which indicate where the first eigenvector's sign changes, create compartments that delineate the boundaries between active ("A compartments") and inactive ("B compartments") chromatin regions. Boundaries between these compartments mark transitions between areas of varying interaction intensities, reflecting the segregation of chromosomal regions into compartments that preferentially associate with like compartments—active with active, and inactive with inactive.

We further annotated our interaction matrices with structural and regulatory elements by including tracks showing CTCF binding peaks and H3K27ac histone modification peaks using CoolBox (Xu et al., 2021). This annotation provided a functional layer to the interaction patterns, linking chromatin proximity with regulatory activity and suggesting mechanisms of chromatin folding's effect on gene regulation.

**Tensor Decomposition**

Pore-C tensors tend to be both large in size while containing many zeros. Thus, sparse tensors are crucial for representing Pore-C data in tensors. Briefly, a dense tensor is a tensor where most of the elements are non-zero. In practical terms, this means that in a dense tensor, one must store every single element, including the zeros. The key point is that the memory storage for a dense tensor is proportional to the product of its dimensions, regardless of the actual values. A sparse tensor, on the other hand, is characterized by most of its elements being zero. This is a common scenario in many real-world datasets where you have a large number of possible elements, but only a few of them have meaningful, non-zero values. In sparse tensors, storing every single element (including a vast majority of zeros) is highly inefficient. Therefore,

sparse tensors are stored using special formats that only record the non-zero elements and their positions. This approach significantly reduces the memory required to store the tensor.

Unfortunately, performing sparse CPD is significantly more challenging than performing dense CPD. Most widely-used CPD implementations, such as those in the Tensorly and PyTorch, perform CPD on dense tensors. While these libraries do technically have methods for performing sparse CPD, these methods tend to rely on temporarily converting sparse tensors into a dense representation performing CPD. For Pore-C tensors, this is not feasible due, as converting to dense representation will cause memory overflow (even when we allocate the maximum memory allowance on our cluster of 1000 GB).

To construct our Pore-C tensors, we used the TensorFox library in Python (Bottega, 2023). TensorFox is an efficient package specializing in multilinear algebra and tensor routines, with an emphasis on CPD. Tensorfox's ability to perform operations on sparse tensors makes it well-suited for decomposing Pore-C tensors. TensorFox provides a wide range of parameters for fine-tuning the CPD process. In our research, we only begin to explore different parameter configurations, and we find that different configurations can lead to different results on the same tensor (Figure 7)**.** Thus, future work should include optimizing the parameters for CPD in order to determine which settings allow us to recover the most biologically relevant information.

**Hyperparameters and Rank Selection**

In our CPD process for 4D tensor decomposition, the tensor rank (R) was varied (3 to 7) to evaluate its impact on decomposition detail and computational complexity. In addition to Tensorfox's default CPD hyperparameters, we experimented with a variety of options to try and find an optimal configuration. When experimenting with different hyperparameters, we took advantage of parallel processing, using four CPU cores to reduce runtime. Significant challenges arose during CPD, especially when decomposing tensors with more than four dimensions, as well as tensors for chromosomes larger than chromosome 20. Most of these challenges came in the form of memory overflow and index overflow errors. With the index errors in particular, the solution is not immediately obvious. It may be necessary to modify TensorFox's source code in order to work with data structures as large as Pore-C tensors. Future efforts should focus on developing methods to handle large tensors more effectively. This will include optimizing sparse tensor representations and exploring efficient unfolding methods to manage larger chromosomes without encountering index errors.

## IV. Results and Discussion

Subread support analysis indicated a strong prevalence of high-order interactions with full subread support (proportion 1.0), affirming the robustness and potential biological relevance of these interactions (Figure 3). These findings show that high-cardinality reads tend to have lower levels of subread support. This is not surprising because higher-order reads have more possible sub-reads; there would need to be more sub-reads present to support a 6-cardinality read than a 4-cardinality read. We also found that increasing the minimum frequency threshold significantly improves subread support across all cardinalities. Since subread support acts as a proxy for interaction validity, this suggests that increasing the minimum frequency threshold for including a set of coordinates in the tensor will increase the quality of reads included in the tensor (albeit increasing the sparsity of the tensor).

Distance measurements between fragments within reads yielded a skewed distribution, with short-range interactions being more frequent than long-range ones (Figure 4). This distribution aligns with the current understanding of chromosomal folding, where loci in closer linear proximity tend to engage in more frequent interactions. The median and mean distance plots underscore the presence of spatially proximal chromatin clusters, which could be indicative of functional genomic domains such as TADs.

### 2D Matrix Analysis

The two-dimensional interaction matrix (Figure 5) derived from Pore-C data from chromosome 20 for cell line GM12878 demonstrates that Pore-C data captures interesting patterns reflective of chromatin organization. The central diagonal's prominence confirms the prevalence of short-range interactions, supporting conventional models of chromosomal folding. Notably, the presence of darker spots beyond the diagonal implies the presence of long-range interactions, which may correspond to regulatory elements, including enhancer-promoter interactions and boundaries of TADs. Eigendecomposition analysis has revealed distinct patterns within the Pore-C matrices, suggesting chromosomal A/B compartmentalization into active and inactive regions. This is illustrated by the gray squares of varying sizes along the diagonal.

Future work should involve comparing the compartments from these Pore-C matrices to those for GM12878 Hi-C matrices.

Though difficult to tell conclusively, the annotations for CTCF enrichment and H3K27ac binding appear to correlate with the compartments obtained from the first eigenvector. Further analysis across multiple chromosomes should be conducted to determine whether these features show an association with regions of increased interaction in Pore-C data, and whether these correlations are supported by Hi-C analysis as well. Similar to Hi-C matrices, the Pore-C matrices only capture two-way interactions, so A/B compartments and their correlations with features such as CTCF enrichment and histone binding should be the same across Pore-C and Hi-C matrices.

The analysis of 2D Pore-C matrices, therefore, provides a complementary approach to Hi-C in delineating chromatin interactions. While the results do validate our overall goal of using tensors to represent high-dimensional Pore-C data, they also highlight the complexity of chromatin organization and the influence of chromosomal features, such as centromeres, on interaction patterns. Determining how to normalize Hi-C data is a complex problem (Bottega, 2023). Thus, given the increased dimensionality and complexity of Pore-C data, we expect that developing methods for normalizing/regularizing Pore-C data will be an important area of research in order to maximize the effectiveness of any computational Pore-C analysis.

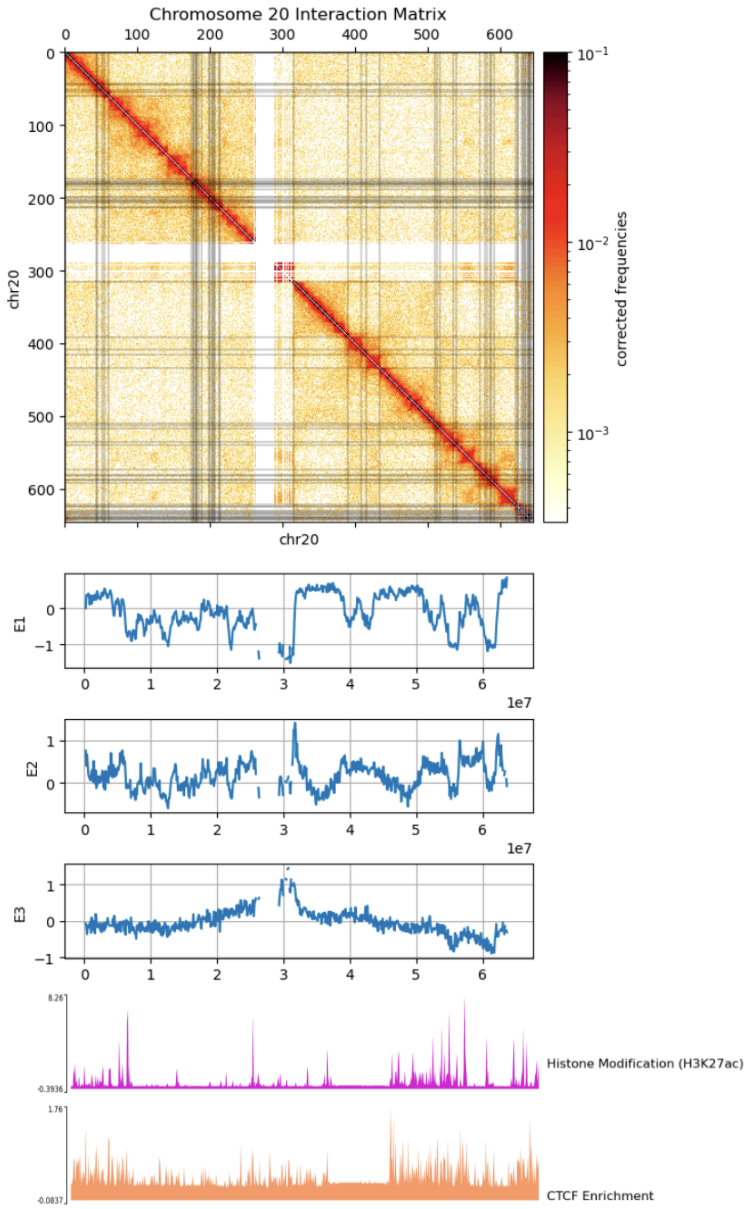


**Figure 5**: *Top*: Pore-C interaction matrix for human chr20 with 100kb bin size, displaying the log-normalized frequency of interactions between bins across the chromosome. Darker colors indicate higher interaction frequencies on a logarithmic scale. Gray lines indicate locations along chr20 where the first Eigenvector (E1) changed signs. Constructed using CoolTools. *Middle Panels*: Plots of the first three eigenvectors obtained from eigendecomposition of the chr20 interaction matrix. The x-axes are labeled with genomic bin indices, and the y-axes represent the normalized eigenvector values, indicating the relative positioning of chromatin within nuclear space. . *Lower Panels:* Pink: H3K27ac histone modification peaks along chr20. The vertical axis represents the level of enrichment based on the negative log of the p-value. Orange: CTCF enrichment peaks aligned along chr20. The vertical axis shows the negative log of the p-value for CTCF binding signal, identifying regions of significant CTCF interaction. Constructed using CoolBox.

**CPD Decomposition**

The application of Canonical Polyadic Decomposition (CPD) on the Pore-C data tensors resulted in decomposed factor matrices that varied with the selection of CPD parameters. Variations in the factor matrices across different parameter sets underscore the sensitivity of CPD to its hyperparameters. These discrepancies highlight the complexity of accurately capturing the multi-faceted nature of chromatin interactions as well as the need for careful parameter optimization when performing sparse CPD in TensorFox. The most obvious way to optimize TensorFox's CPD hyperparameters is to identify settings that yield factor matrices that correlate with the eigenvectors produced by eigendecomposition of 2D Pore-C matrices. Ideally, we would want the peaks of the factor values like those in Figure 6 to align with the peaks of the eigenvectors like those in Figure 5. However, such a comparison is not yet possible. This is because the eigenvectors were calculated from a normalized, or balanced, Pore-C matrix (using the default matrix balancing settings from the Cooler package), while the CPD factor matrix values were calculated from an unbalanced Pore-C tensor. Currently, no established methods exist for normalizing or balancing a Pore-C tensor.

CPD experiments exhibited an expected decrease in performance when using higher-dimensional tensors, as these tensors were more sparse than their 2D counterparts. CPD operations on 3-order Pore-C tensors had both lower accuracy and higher relative error, respectively, than CPD operations on 2D tensors. Similarly, CPD operations on 4-order tensors had worse performance metrics than operations on 3-order tensors.

In constructing Pore-C tensors, an inclusive strategy incorporating all subreads yields more dense tensors and a more comprehensive dataset, but at the cost of increased memory requirements, CPD runtime, and tensor sparsity. The balance between data inclusivity and computational feasibility/accuracy is therefore an important consideration. Our subread support data provides a measure of data quality for Pore-C reads, with more supported reads being higher quality, and less likely to be the result of noise, than less supported reads. Reads with lower subread support may still be biologically meaningful, but additional methods should be developed to ensure that such reads are valid.

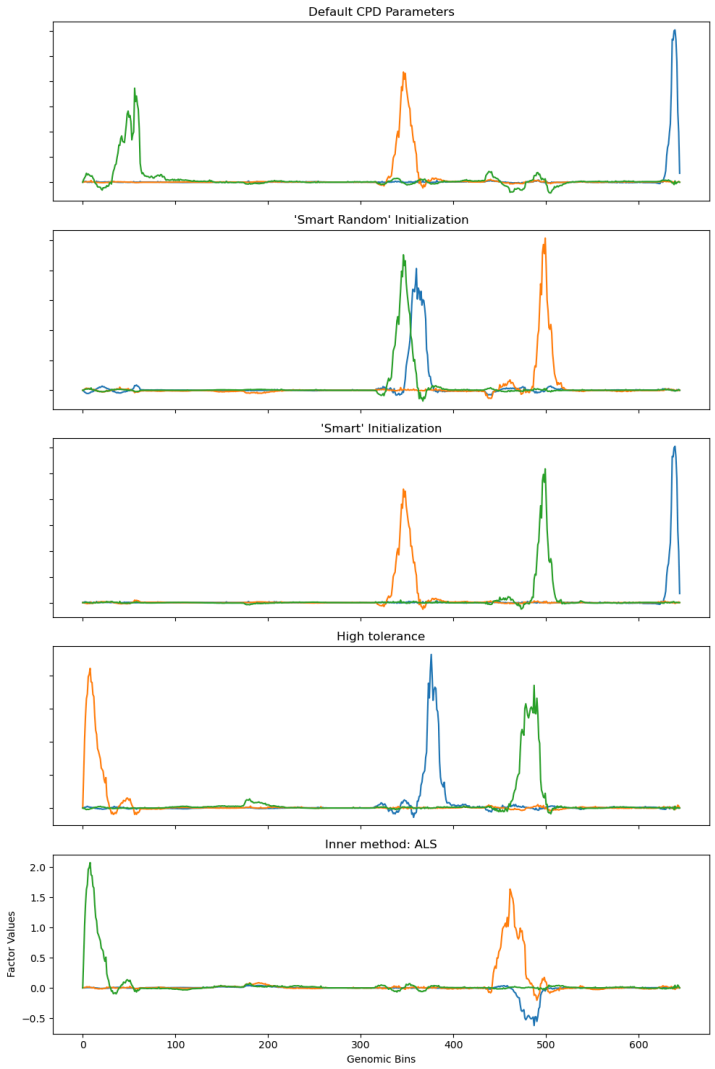


**Figure 6**: Plots of Factor matrices from CPD of chr20 under varied TensorFox hyperparameters. Y-axis shows factor values; x-axis indicates genomic bins. Ranks 1 to 3 are represented in blue, orange, and green, respectively. Top to bottom: default parameters, 'Smart Random' and 'Smart' initializations, high tolerance settings, and ALS method application. Each sub-plot reflects the impact of these hyperparameters on the tensor decomposition output.

The approach of applying CPD to Pore-C tensors presents a promising avenue for analyzing multiway chromatin interactions, and potentially revealing complex patterns in chromatin structure that are not apparent from traditional Hi-C analysis. However, there are still numerous challenges associated with our tensor construction and decomposition methods. Future research should focus on addressing issues such as data quality, data sparsity, memory constraints, and parameter selection. Tensor construction should be refined through more robust means of determining which reads should be included in each tensor, as well as potentially

giving differential weighting or bias to reads that are especially interesting or informative. Reads with high frequency, high subread support, and high inter-fragment distance (ideally measured by geometric mean) could be weighted more heavily in the tensor through a robust scoring system. Normalization of tensor frequencies will also be crucial to correct for biases and improve the comparability of interaction strengths. Additionally, excluding centromeric regions from future analysis may prevent skewing in principal component analysis and CPD, allowing these methods to pick up on more subtle structural patterns rather than focusing only on centromere location. Finally, alternative tensor decomposition methods, such as non-negative decomposition and tucker decomposition, could be explored in order to further improve our ability to analyze high-order reads in Pore-C data.

## V. Conclusion

Our study introduces a novel approach to analyzing Pore-C data by employing Canonical Polyadic Decomposition (CPD) on multidimensional tensors, capturing a more complex array of chromatin interactions than traditional Hi-C methods. We demonstrated that applying CPD to Pore-C tensors can potentially provide a richer, more intricate depiction of chromatin organization, though this is contingent upon rigorous optimization of CPD parameters and computational strategies. A key finding is the necessity of balancing data inclusivity with computational demands, emphasizing the importance of a strategic approach to tensor construction and the interpretation of high-cardinality reads.

Should the CPD method prove reliable upon further validation, it would represent a significant advance in the field of epigenetics and the study of the 3D genome. The ability to discern high-order interactions with greater specificity could shed light on the complex mechanisms of gene regulation and chromatin dynamics, paving the way for a deeper understanding of genomic architecture. The current stage of our research serves as a foundational step, indicating a promising direction rather than conclusive outcomes. The broader implications of this method, if fully realized, could contribute to the identification of novel chromatin configurations and their roles in gene expression and cellular function, subject to the constraints and considerations highlighted in our analysis.

## VIII. References


1. Dekker J. (2008). Gene regulation in the third dimension. *Science* (New York, N.Y.), 319(5871), 1793–1794. https://doi.org/10.1126/science.1152850
2. Misteli T. (2007). Beyond the sequence: cellular organization of genome function. *Cell*, 128(4), 787–800. https://doi.org/10.1016/j.cell.2007.01.028
3. Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science* (New York, N.Y.), 295(5558), 1306–1311. https://doi.org/10.1126/science.1067799
4. Lieberman-Aiden, Erez et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* (New York, N.Y.) vol. 326,5950 (2009): 289-93. https://doi.org/10.1016/j.cell.2007.01.028
5. Fortin, JP., Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 16, 180 (2015). https://doi.org/10.1186/s13059-015-0741-y

6. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376–380. https://doi.org/10.1038/nature11082
7. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021
8. Deshpande, A. S., Ulahannan, N., Pendleton, M., Dai, X., Ly, L., Behr, J. M., Schwenk, S., Liao, W., Augello, M. A., Tyer, C., Rughani, P., Kudman, S., Tian, H., Otis, H. G., Adney, E., Wilkes, D., Mosquera, J. M., Barbieri, C. E., Melnick, A., Stoddart, D., … Imieliński, M. (2022). Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature biotechnology*, 40(10), 1488–1499. https://doi.org/10.1038/s41587-022-01289-z
9. Zhen, C., Wang, Y., Geng, J., Han, L., Li, J., Peng, J., Wang, T., Hao, J., Shang, X., Wei, Z., Zhu, P., & Peng, J. (2022). A review and performance evaluation of clustering frameworks for single-cell Hi-C data. *Briefings in bioinformatics*, 23(6), bbac385. https://doi.org/10.1093/bib/bbac385
10. Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* (Oxford, England), 34(16), 2781–2787. https://doi.org/10.1093/bioinformatics/bty185
11. Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis* (pp. 91-109). Boston, MA: Springer US.
12. Abdennur, N., & Mirny, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics (Oxford, England), 36(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540
13. Open2C, Abdennur, N., Abraham, S., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., ... & Venev, S. V. (2022). Cooltools: enabling high-resolution Hi-C analysis in Python. BioRxiv, 2022-10. https://doi.org/10.1101/2022.10.31.514564
14. Xu, W., Zhong, Q., Lin, D., Zuo, Y., Dai, J., Li, G., & Cao, G. (2021). CoolBox: a flexible toolkit for visual analysis of genomics data. BMC bioinformatics, 22(1), 489. https://doi.org/10.1186/s12859-021-04408-w
15. Rabanser, S., Shchur, O., & Günnemann, S. (2017). Introduction to tensor decompositions and their applications in machine learning. arXiv preprint arXiv:1711.10781.
16. Greub, Werner H. Linear algebra. Vol. 23. Springer Science & Business Media, 2012.

17. Bottega, F. (2023). Tensor-Fox [Software]. GitHub. https://github.com/felipebottega/Tensor-Fox
18. Lyu, H., Liu, E., & Wu, Z. (2020). Comparison of normalization methods for Hi-C data. BioTechniques, 68(2), 56–64. https://doi.org/10.2144/btn-2019-0105