# Accelerated spike-triggered non-negative matrix factorization reveals coordinated ganglion cell subunit mosaics in the primate retina

Sören J. Zapp[1,2], Mohammad H. Khani[1,2,3,§], Helene M. Schreyer[1,2,§], Shashwat Sridhar[1,2], Varsha Ramakrishna[1,2,3], Steffen Krüppel[1,2,4], Matthias Mietsch[5,6], Dario A. Protti[7], Dimokratis Karamanlis[1,2,3,%], Tim Gollisch[1,2,4,*]

1. University Medical Center Göttingen, Department of Ophthalmology, Göttingen, Germany
2. Bernstein Center for Computational Neuroscience, Göttingen, Germany
3. International Max Planck Research School for Neurosciences, Göttingen, Germany
4. Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen, Germany
5. German Primate Center, Laboratory Animal Science Unit, Göttingen, Germany
6. German Center for Cardiovascular Research, Partner Site Göttingen, Göttingen, Germany
7. The University of Sydney, School of Medical Sciences (Neuroscience), Sydney, NSW, Australia

* tim.gollisch@med.uni-goettingen.de

[§] current address: Institute of Molecular and Clinical Ophthalmology Basel, Basel, Switzerland

[%] current address: University of Geneva, Department of Basic Neurosciences, Geneva, Switzerland

## ABSTRACT

A standard circuit motif in sensory systems is the pooling of sensory information from an upstream neuronal layer. A downstream neuron thereby collects signals across different locations in stimulus space, which together compose the neuron's receptive field. In addition, nonlinear transformations in the signal transfer between the layers give rise to functional subunits inside the receptive field. For ganglion cells in the vertebrate retina, for example, receptive field subunits are thought to correspond to presynaptic bipolar cells. Identifying the number and locations of subunits from the stimulus–response relationship of a recorded ganglion cell has been an ongoing challenge in order to characterize the retina's functional circuitry and to build computational models that capture nonlinear signal pooling. Here we present a novel version of spike-triggered non-negative matrix factorization (STNMF), which can extract localized subunits in ganglion-cell receptive fields from recorded spiking responses under spatiotemporal white-noise stimulation. The method provides a more than 100-fold speed increase compared to a previous implementation, which can be harnessed for systematic screening of hyperparameters, such as sparsity regularization. We demonstrate the power and flexibility of this approach by analyzing populations of ganglion cells from salamander and primate retina. We find that subunits of midget as well as parasol ganglion cells in the marmoset retina form separate mosaics that tile visual space. Moreover, subunit mosaics show alignment with each other for ON and OFF midget as well as for ON and OFF parasol cells, indicating a spatial coordination of ON and OFF signals at the bipolar-cell level. Thus, STNMF can reveal organizational principles of signal transmission between successive neural layers, which are not easily accessible by other means.

## INTRODUCTION

In sensory pathways, neuronal signals typically carry spatiotemporal information about the outside world. The information is transmitted by an array of neurons, each encoding a constrained region of the stimulus space, defining its receptive field. A postsynaptic neuron can pool signals across multiple neurons in this array, constituting a receptive field that is larger than in the presynaptic layer. Owing to synaptic rectification or other signal transformations in the signal transmission, the integration of presynaptic signals is often nonlinear, which gives rise to substructure within the receptive field, for example, in the form of functional subunits that together compose the receptive field (Zapp et al. 2022). Nonlinear integration of subunit signals then sets the stage for complex computations. For example, in the retina, the neural network where vertebrate visual processing starts, subunits convey sensitivity of ganglion cells to spatial contrast or to different types of motion stimuli (Demb and Singer 2015; Gollisch and Meister 2010; Kerschensteiner 2022).

For retinal ganglion cells, receptive field subunits are thought to correspond to presynaptic bipolar cells (Borghuis et al. 2013; Demb et al. 2001; Liu et al. 2017). Their excitatory signals may be integrated nonlinearly by the ganglion cells (Borghuis et al. 2013; Demb et al. 2001), which plays a substantial role in sensitivity to high spatial frequency (Enroth-Cugell and Robson 1966; Schwartz et al. 2012; Victor and Shapley 1979) as well as in local dynamics like contrast adaptation (Jarsky et al. 2011; Manookin and Demb 2006; Rieke 2001). Understanding how subunits compose the receptive field and how their signals are integrated can thus, on the one hand, provide essential information about the functional connectivity and the propagation of information in sensory networks, like how ganglion cells pool signals over an array of bipolar cells. These insights may then aid in explaining how ganglion cell receptive fields implement specific neural computations. On the other hand, identifying the layouts of subunits for different neuronal cell types can complement morphological findings by identifying differences in connectivity patterns and shared connections. In particular, the subunits of ganglion cells of a single type may be expected to form a mosaic-like arrangement, as is commonly observed for receptive fields of retinal cells of a given type. This has been reported across species for photoreceptors (Marc and Sperling 1977), bipolar cells (Cohen and Sterling 1990a, 1990b; Sterling et al. 1988), and ganglion cells (Anishchenko et al. 2010; DeVries and Baylor 1997; Gauthier et al. 2009).

The retina subjected to light stimulation has proved to be a suitable model for investigating the functional connectivity in sensory systems. Nevertheless, direct experimental investigations of the neuronal connections that constitute the subunits, for example by simultaneously stimulating and recording pre- and postsynaptic neuronal layers, remain challenging. In light of the experimental limitations, several methods have been proposed to infer subunits computationally from recorded ganglion cell responses to visual stimuli. The methods include data-driven fitting of cascade networks that model subunits as linear filters with a nonlinear transfer function converging onto a ganglion cell (Freeman et al. 2015; Maheswaranathan et al. 2018; McFarland et al. 2013). Similarly, convolutional neural networks, which offer a comparable model of signal convergence with potentially deeper layering have been used to capture how local processing inside the receptive field of a ganglion cell shapes the cell's visual responses (Maheswaranathan et al. 2023; McIntosh et al. 2016; Tanaka et al. 2019). On the other side of the spectrum, there are statistical analyses of a cell's stimulus-response relationship, such as spike-triggered clustering (Shah et al. 2020) and spike-triggered non-negative matrix factorization (STNMF; Liu et al. 2017).

For example, STNMF relies on visual stimulation with spatial or spatiotemporal white-noise patterns and on analyzing correlations in the set of those stimulus patterns that elicited spikes (spike-triggered stimulus ensemble), while placing a non-negativity and a sparsity constraint on the extracted structures. Subunits identified with STNMF from salamander ganglion cells have been shown to match receptive fields of simultaneously recorded bipolar cells (Liu et al. 2017). However, the proposed implementation is computationally expensive, requiring on the order of several hours of analysis time per ganglion cell on a standard desktop computer. This has limited the control over the effects of hyperparameters, such

as the strength of sparsity regularization, which needs to be adjusted to the right range for obtaining spatially localized subunits.

Here, we develop a new version of STNMF that addresses the shortcomings of previous implementations. We introduce a novel combination of state-of-the-art algorithms of non-negative matrix factorization (NMF) and complement it with guided initialization procedures to improve speed and reliability. The resulting method allows subunit recovery for a cell in a matter of seconds on a standard desktop computer. We demonstrate how clustering consensus methods can be applied for hyperparameter selection and offer tools that reliably provide suitable values for sparsity regularization. This makes STNMF more versatile across cells of different functional classes and across datasets of different species. In this work, we firstly demonstrate the superior speed of the new STNMF algorithm and its flexibility in analyses of ganglion cell populations in both salamander and primate retina. Secondly, we show that the versatile implementation of STNMF with hyperparameter tuning allows matching subunit properties to anatomical characteristics of parasol and midget ganglion cells in the primate retina. Thirdly, we use STNMF to investigate subunit mosaic arrangements of ON- and OFF-type pathways, demonstrating that these are spatially aligned between ON- and OFF-pathways, in contrast to the anti-alignment of ganglion cell receptive field mosaics (Roy et al. 2021).

## RESULTS

STNMF is a type of spike-triggered analysis that aims at extracting spatial subunits from the structure of spike-eliciting stimulus segments under white-noise stimulation. Spike-triggered analyses have long been used for assessing receptive fields via computation of the spike-triggered average (STA; Bryant and Segundo 1976; Chichilnisky 2001; De Boer and Kuyper 1968) as well as for obtaining multiple (typically temporal) stimulus filters via spike-triggered covariance (STC) analysis (Cantrell et al. 2010; Fairhall et al. 2006; Gollisch and Meister 2008; Samengo and Gollisch 2013; Schwartz et al. 2006). As an extension of these approaches, STNMF can identify localized subunits, based on the statistical structure (e.g. correlations) of spike-triggered stimuli, leading to a parts-based (Lee and Seung 1999) decomposition of the receptive field. For detecting the relevant structure evoked by the subunits, STNMF requires a spatiotemporally uncorrelated stimulus (white noise) with a spatial resolution finer than the expected subunit size.

The analysis starts with recording spikes of a neuron under white-noise stimulation, for example, a checkerboard layout with light intensities of the checkerboard squares flickering randomly and independently in the case of visual stimulation. From the recorded data, one then extracts the spike-triggered stimuli, that is, the stimulus sequences preceding a spike over a time window given by the temporal sensitivity of the neuron. Since subunits of a given cell typically share the same temporal dynamics (Maheswaranathan et al. 2018), focusing on the spatial profile by integrating out time has been a viable option to reduce input dimensions (Liu et al. 2017; Shah et al. 2020). To do so, the temporal component of the spike-triggered stimuli, obtained from the STA, is used to compress each spike-triggered stimulus into a single effective spike-triggered spatial pattern (Kaardal et al. 2013; Liu et al. 2017). Concretely, for each spike-triggered stimulus sequence, we take the average of the stimulus frames weighted with the temporal component. Furthermore, to focus the analysis on the relevant region in space, the time-averaged spike-triggered stimuli are cropped to roughly the region of a cell's receptive field, here done by selecting the smallest rectangle containing the three-standard-deviation ellipse of a Gaussian function fitted to the spatial component of the STA.

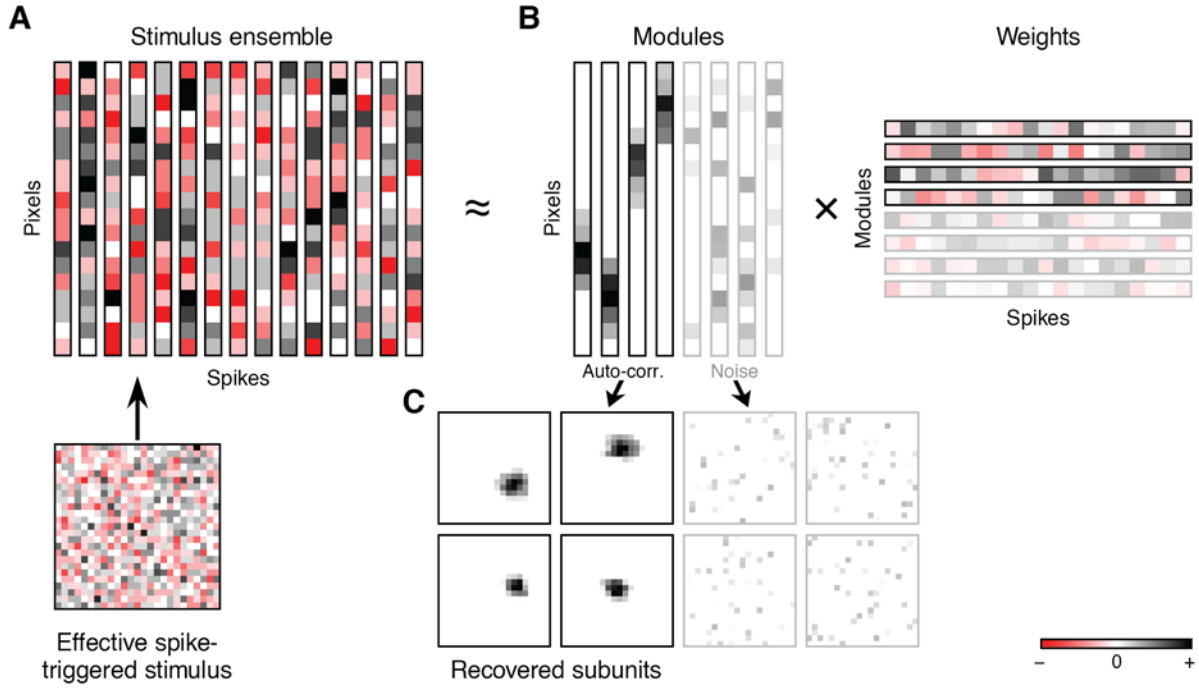


**Figure 1. Conceptual sketch of spike-triggered non-negative matrix factorization.** (**A**) The stimulus ensemble (top) is a matrix of pixels by spikes. Each column is the effective spike-triggered stimulus of a given spike (bottom). The stimulus ensemble serves as the input to the matrix factorization (dimensions not to scale). (**B**) Using semi-NMF, the stimulus ensemble is decomposed into two smaller matrices, the non-negative spatial modules and their corresponding weights. Together, these represent a lower-dimensional approximation of the stimulus ensemble. (**C**) The models can be reshaped into two-dimensional spatial layouts. Some modules exhibit localized structure and are identified via their spatial autocorrelation as subunits, whereas others (here displayed with reduced saturation) capture noise.

The collection of these time-averaged, cropped spike-triggered stimulus patterns gives us the effective spike-triggered stimulus ensemble (STE) for a given cell. We arrange the STE into a pixel-by-spikes matrix (**Figure 1A**), where each column collects all the pixel values of the effective spike-triggered stimulus for a given spike. We then perform semi-non-negative matrix factorization (semi-NMF; Ding et al. 2010) on this STE matrix. The use of semi-NMF, rather than a full NMF with non-negativity constraints on all matrices, enforces non-negativity only on one of the two matrices in the factorization. This accommodates the negative contrast values contained in our STE matrix.

The semi-NMF decomposes the STE matrix into two smaller matrices, which contain the spatial modules and their corresponding weights for each spike (**Figure 1B**). Any stimulus that led to a spike is approximated by a linear combination of the modules, with weights specific to that spike. The non-negativity constraint is here applied to the module matrix. Thus, there are no negative pixel values in the modules, which could cancel out with positive values in the linear combination. As a result, the modules emerge as purely additive building blocks. By limiting the number of modules, these building blocks must be shared between the different spike-triggered stimuli and therefore reflect the prevailing structure (the pixel correlations) in the STE. Along with a sparsity constraint on the module matrix, STNMF extracts spatially localized modules that capture the spatial patterns behind the spiking response. These are taken as the estimated subunits within the receptive field of the ganglion cell (**Figure 1C**).

The optimal number of modules is not known a priori, and its selection poses a common hurdle in the field of NMF and of subunit recovery. Much like for clustering algorithms, the obtained modules might merge separate structures or oversplit if their number is not chosen well. For identifying the localized subunits, however, STNMF is relatively robust to changing the number of modules as long as it is somewhat larger than the number of (expected) subunits (Liu et al. 2017). The reason for this is that the excess modules beyond the ones that capture the actual subunits turn into non-localized, noise-like modules. Allowing for more modules then increases the number of noise-like modules, whereas the

number of localized subunit-like modules remains largely unaffected. The noise-like modules are typically dominated by pixels outside the receptive field, which do not contribute to activating the cell and are thus not particularly correlated with pixels that compose the localized subunits. The noise-like subunits therefore capture spurious correlations within the STE, essentially reflecting noise in the STE from finite sampling. For the detection of subunits, these noise-like modules can then be ignored, making the identification of localized subunits robust to varying the number of modules. As previously suggested (Liu et al. 2017), we find 20 modules to be a good number for our application on retinal ganglion cells in the datasets analyzed here. The recovered modules of interest are visibly distinguishable from the noisy excess modules by their spatially localized structure. We automate this differentiation by defining those modules as recovered subunits that have a sufficiently high spatial autocorrelation (**Figure 1C**; see Methods).

*Accelerated fast hierarchical alternating least squares speeds up subunit recovery*

We investigated different NMF algorithms to provide the most suitable implementation in context of subunit recovery. Due to its non-negativity constraint, NMF is NP-hard (Vavasis 2010). To obtain a good decomposition, NMF is generally performed by iterative improvements. Various algorithms facilitate these iterations under the non-negativity constraint. Many of them are accurate but computationally costly algorithms (Kim and Park 2011), as they often seek exact solutions at each iteration when updating either the modules or the weights and because they consider all data at once. As an alternative, hierarchical alternating least squares (HALS; Cichocki et al. 2007; Ho 2008; Li and Zhang 2009) has gained popularity in recent years. HALS replaces accurate with approximated solutions at each iteration of NMF, but exploits the thereby gained iteration speed. Over more but faster iterations, it outperforms previous methods in many applications of NMF (Cichocki et al. 2007; Cichocki and Phan 2009; Gillis and Glineur 2012) while ensuring convergence (Gillis and Glineur 2008). Instead of inferring all modules from their weights at once, HALS updates the modules sequentially. This breaks down large matrix multiplications into smaller, more efficient vector operations. The speed of this approach compensates for this approximation because more iterations are possible in shorter time. Various modifications of the HALS algorithm have been proposed recently to further increase its speed or its accuracy.

Here, we combine two HALS algorithms into what we call Accelerated Fast HALS (AF-HALS). Fast HALS (Cichocki and Phan 2009) achieves a speed-up through computational simplification by employing vector normalization to cancel out terms in the update equation. Accelerated HALS (Gillis and Glineur 2012) improves factorization accuracy per iteration by performing multiple consecutive HALS updates as long as the number of floating point operations stays below that of the heavy matrix operations preceding the next iteration. With Fast HALS, each iteration in NMF becomes less complex, requiring less computation time. With Accelerated HALS, each iteration becomes more accurate, requiring fewer iterations.

To demonstrate a direct comparison in speed and accuracy in the context of STNMF, we selected the popular active-set method (Lawson and Hanson 1974) as an alternative to AF-HALS. It is representative for current state-of-the-art NMF algorithms (Kim and Park 2011; Zhang et al. 2014) as it prevails for its high accuracy at each iteration. A previous implementation of STNMF was based on the active-set method (Liu et al. 2017). We demonstrate the differences exemplarily on a retinal ganglion cell from the salamander retina, which had been recorded under white-noise stimulation as part of a previous study of subunit recovery with STNMF (Gollisch and Liu 2018; Liu et al. 2017). We find that, using the same data, number of modules, and initial conditions, AF-HALS converges on localized modules both faster and within fewer iterations (**Figure 2**). Not only are the individual NMF iterations considerably faster with AF-HALS, reaching 10 iterations within a fraction of a second rather than the several seconds required by the active-set implementation, but the achieved error and the quality of the identified subunit structure is comparable or even slightly better for AF-HALS, despite the approximation at each iteration.

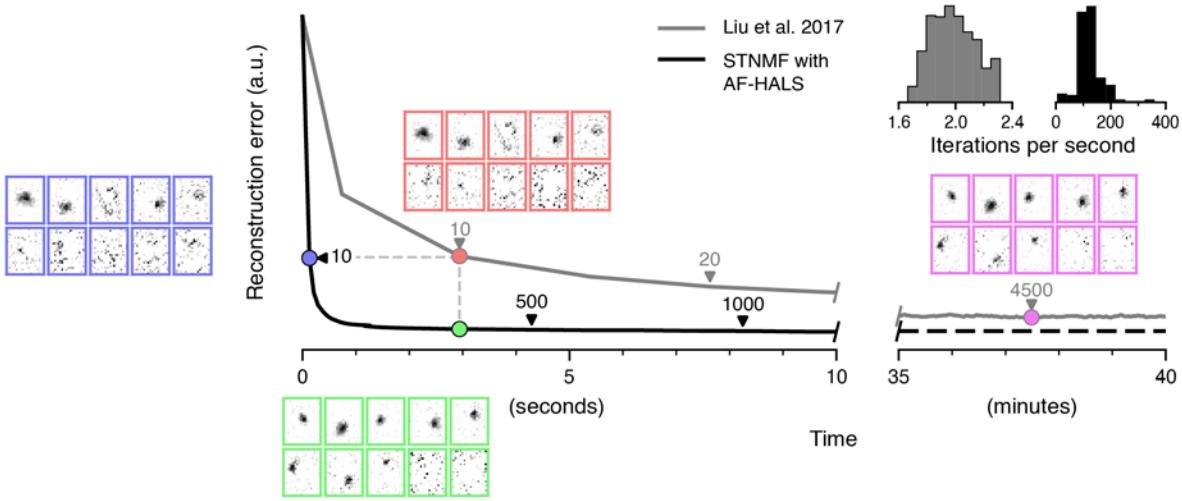


**Figure 2. STNMF with Accelerated Fast HALS (AF-HALS) converges faster than the previous implementation of STNMF.** Reconstruction error of STNMF implemented with AF-HALS (black) and STNMF based on the active-set method used in Liu et al. (2017; gray) for a sample salamander ganglion cell. Arrows denote the number of iterations. Insets allow visual inspection of the recovered modules at different states during the iterations. Although the active-set method calculates more accurately at each step, the reconstruction process is similar after ten iterations in both methods (blue and red insets). The active-set method does not reach the low error that AF-HALS accomplishes within a few seconds. The visualized subunits of the active-set method after 45 minutes (purple inset) arrive at the state that AF-HALS reached within five seconds (green inset). Top right: Distributions of iteration speeds (measured as inverse of each iteration duration) of the active-set method (gray) and of AF-HALS (black) for the sample cell.

Thus, a clear subunit decomposition already emerges within the first few hundred iterations, taking only a few seconds, with a remaining error that is smaller than what is reached by the active-set implementation even after several tens of minutes of runtime (**Figure 2**). In the end, both methods arrive at similar subunit estimates, but convergence for AF-HALS is reached much sooner by orders of magnitude in absolute computation time.

*Guided initialization can replace repeated runs*

As NMF is a non-convex optimization problem (Vavasis 2010), there are no guarantees for any given method to arrive at the globally optimal solution. In typical iterative implementations, solutions correspond to local optima and therefore depend on the starting point of the iterative search (Lee and Seung 1999), that is, the initial values of the modules. A popular way to compensate for the dependence on the starting point is to perform multiple repetitions of NMF starting from different random initializations of the modules and select the best of the obtained decompositions, the one with minimal residual reconstruction error (Berry et al. 2007; Cichocki et al. 2009).

However, random initialization generally does not offer a strong starting point for efficient convergence of the iterative procedures (Wild et al. 2004), and relying on many repetitions with different initializations poses a high demand on computational time. Furthermore, the selection of the best solution may depend on the particular measure of comparison. To circumvent these obstacles, we substitute the time-consuming repeated runs by implementing a guided initialization, based on singular value decomposition (SVD; Boutsidis and Gallopoulos 2008). Applying SVD to the matrix of the STE already yields structured components, akin to modules after several iterations of randomly initialized NMF, albeit with both positive and negative entries. Furthermore, SVD-based initialization has been shown to yield earlier convergence and smaller residual errors of NMF (Qiao 2015). To accommodate for the non-negativity constraint of NMF, methods of SVD-based initialization typically set negative values in the SVD components to zero before supplying these as initial conditions to the NMF algorithm.

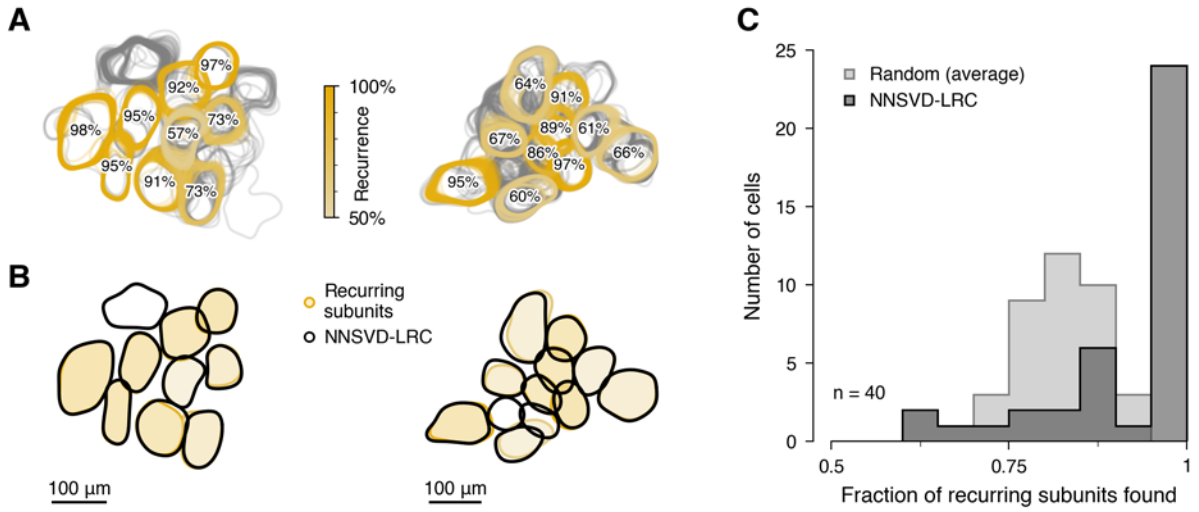


**Figure 3. A single run of NNSVD-LRC-initialized STNMF yields subunits consistent with recurring subunits from 100 different, randomly initialized runs**. (**A**) Contour outlines of subunits of 100 randomly initialized STNMF runs for two sample cells (fast-OFF type) from salamander retina. Subunits of different runs are regarded as identical if they exhibit substantial spatial overlap (Jaccard index > 0.5; see Methods). Subunits that emerged in more than half of the runs are colored, with saturation corresponding to the level of recurrence (see colorbar) and marked with a percentage of recurrence. (**B**) Comparison of subunit layouts recovered with NNSVD-LRC (gray) and with randomly initialized runs and a 50%-recurrence-criterion (yellow). Each yellow outline is the contour of the mean of the recurring subunits in (**A**). For the two sample cells, a single NNSVD-LRC-based STNMF run found all recurring subunits of the randomly initialized runs. (**C**) Distributions over 40 analyzed fast-OFF cells of the fraction of recurring subunits recovered by a single run with NNSVD-LRC (dark gray) or by 100 runs with random (light gray) initialization. Recovered subunits are determined via the Jaccard index (see above and Methods). NNSVD-LRC-based STNMF reliably finds more of the recurring subunits than the individual randomly initialized runs (averaged over 100 runs for each cell).

As a further improvement, one can take advantages of the sign ambiguity of SVD and duplicate selected components with opposite sign before setting the negative values in all components to zero, as is done in non-negative singular value decomposition with low-rank correction (NNSVD-LRC; Atif et al. 2019). This offers potentially twice the amount of information from the selected SVD components. Furthermore, the procedure produces initial modules with around half of their values at zero, offering around 50% sparsity before even starting NMF. NNSVD-LRC subsequently uses the low-rank approximation of the input matrix constructed from the selected components to perform a few NMF iterations. The low-rank factorization reduces the computational complexity of the matrix multiplications involved, providing a computationally cheap head start for the full NMF. We modified NNSVD-LRC to be suitable for semi-NMF and improved its numerical stability (see Methods). This includes keeping the previously skipped duplicate of the component with the highest singular value and replacing the elements of columns whose entries are all zeros with a small non-zero constant to avoid convergence issues in HALS-based NMF.

We find that a single run of STNMF initialized with NNSVD-LRC can typically recover all recurring subunits that had been identified via 100 repetitions of randomly initialized STNMF. To show this, we compared STNMF results for an available dataset of salamander retinal ganglion cells (Gollisch and Liu 2018). For this comparison, we aimed at determining those spatially localized subunits that were robustly identified across multiple randomly initialized STNMF runs. We defined them as subunits that could be repeatedly identified across a majority of runs by their high spatial overlap (see Methods). We found that those robustly identified subunits matched the subunits recovered with NNSVD-LRC-based STNMF quite closely (**Figure 3**). This demonstrates the viability of the NNSVD-LRC initialization and supports the possibility of an additional 100-fold speed increase by omitting the repeated runs with random initialization, in addition to the already manifold speed increase obtained from applying AF-HALS.

*Effects of sparsity regularization on subunit layout*

In contrast to regular NMF, where the non-negativity constraint is placed on all matrices, semi-NMF only forces one matrix to be non-negative, in our case the modules, not the weights. Due to the loosened constraint, semi-NMF typically fails to retain additive, parts-based solutions without additional constraints (Ding et al. 2010). This crucial aspect of NMF can be recovered in semi-NMF through sparsity regularization on one of the factor matrices (Hoyer 2002, 2004; Kim and Park 2007, 2008). Unlike other methods of subunit recovery or machine learning in general, where regularization serves as an optional step to counteract overfitting or to alleviate the lack of sufficient data, sparsity regularization in STNMF based on semi-NMF is therefore an essential part of its functionality when parts-based decompositions are desired. To obtain localized subunits, we enforce the modules to be both non-negative and sparse. We achieve this by $\ell 1$-norm-based regularization, including a term in the objective function that penalizes large numbers of non-zero-pixel values in the modules.

As is customary, the strength of regularization is controlled via a regularization parameter. Increasing the value of the regularization parameter drives more pixel values in the modules to zero. Once regularization becomes too strong, subunits are forced to shrink in size or will be split into multiple smaller subunits until they eventually vanish. The optimal amount of sparsity depends on the dimensionality of the data. Accessible properties, like the number of pixels and number of spikes, as well as unknown characteristics prior to subunit recovery, like the true number of subunits and their sizes, play a role. Consequently, the regularization is difficult to control and may have to be individually adjusted for cells that differ in their properties. On the other hand, for cells that share characteristics, similar strengths of regularization should be suitable.

We examined the impact of sparsity regularization on the subunit decomposition for data recorded from ganglion cells of a salamander retina (Gollisch and Liu 2018). We ran STNMF initialized with NNSVD-LRC on the data of selected cells at varying sparsity regularization. Without sparsity regularization, we obtained modules that span the entire receptive field with strong overlap (example cell in **Figure 4C insets**). As regularization strength increases, these modules first break up into large subunits that subsequently split into multiple smaller subunits. For a certain range, the subunit layout becomes more stable, with variations in smaller subunits further away from the receptive field center. Although prominent subunits are fairly robust to changes in sparsity, good control to provide the right amount of sparsity regularization is beneficial to extract a subunit layout that is most likely to represent the underlying structure of the receptive field.

*Consensus analyses aid in selection of regularization parameter*

The regularization requires the choice of a hyperparameter value to control the amount of sparsity. With our tractable implementation of STNMF, we here provide a method to do so in reasonable computation time. A common approach to finding the most suitable weight for regularization is through cross-validation. However, techniques that rely on held-out data are problematic in applications of unsupervised learning like NMF (Bro et al. 2008; Owen and Perry 2009). The weight matrix in STNMF is highly dependent on the spikes from the specific input and cannot be used to compare training and test data sets, but would have to be recomputed for the test data. Relatedly, the reconstruction error of the factorization does not pose a suitable metric in the context of sparsity regularization as it increases monotonically with regularization (see Methods). Adjusted cross-validation-like procedures for matrix factorization have been proposed (Fu and Perry 2017; Owen and Perry 2009) based on selectively leaving out data entries across both rows and columns of the input matrix (pixels and spikes) and keeping these randomly scattered matrix elements as test data. However, using missing or masked entries in the input matrix does not generalize well to accelerated NMF algorithms like Accelerated HALS. This is due to optimizations regarding vectorized calculations and matrix multiplications that are factored out of the column-wise updates.

Instead of cross-validation, we therefore turn to the solution of a similar parameter-selection problem. A common hurdle for clustering techniques is the choice of the optimal number of clusters. This problem also translates to NMF and may be addressed by using so-called consensus methods (Monti et al. 2003). To explore this approach to find a suitable regularization parameter for sparsity, we temporarily neglect the aforementioned SVD-based initialization and revert to repeated runs of NMF with different random initializations. The variability in the solutions across runs is then exploited to find the most consistent decomposition among the sets of solutions. The more appropriate the choice of the sparsity regularization parameter, the more the subunit structure becomes similar across solutions, that is, the more the different solutions consent regarding which spikes relate to the same subunits (Brunet et al. 2004).

In clustering, the consensus is typically analyzed via the consensus matrix, which contains for each pair of samples the fraction of times that the two samples were clustered together. In STNMF, the spikes take the role of the samples, and the weight factor matrix denotes their affiliation with the different modules, analogous to different clusters. This, however, is not a hard cluster assignment, but a soft weighting of how much each module contributes to reconstructing a given spike-triggered stimulus. Thus, to build the consensus across repeated runs, we determine the module most strongly affiliated with each spike as the module with the maximum weight for the spike. For our purposes, we further slightly modify the calculation of the consensus matrix by only allowing localized, subunit-like modules to contribute positively to the consensus. That is, if a pair of spikes is affiliated with the same non-localized module for a given STNMF run, this is treated like a case where affiliations do not match. We found that this adjustment reduced noise in the consensus analysis that likely comes from spurious coincidences of spikes with noise-like modules.

The consensus matrix then contains, for all pairs of spikes, the fraction of randomly initialized STNMF runs for which the pair is affiliated with the same localized subunit (**Figure 4B**). As proposed previously (Brunet et al. 2004), we quantify the dispersion of values in the consensus matrix by its cophenetic correlation coefficient (CPCC). The cophenetic correlation measures the degree of clustering in the consensus matrix in a hierarchical clustering distance manner. This provides a scalar metric between zero and unity, with higher values indicating increased stability of the decomposition.

For a given sparsity parameter, we run STNMF repeatedly with different random initializations. The consensus analysis reflects how robust the parameter is to the initializations and describes the stability of the subunit decomposition, which can then be compared via the CPCC across different sparsity parameter values (**Figure 4C**). We find that stability first increases with stronger sparsity regularization, as it increasingly constrains the solution and decreases variability among the decompositions, and then plateaus or declines again when regularization becomes too strong. Beyond this, a high coefficient can eventually also arise at the extreme end of strong regularization when very few or even only a single localized subunit emerges, causing near perfect consensus. In particular, this occurs as our computation of the consensus matrix only considers localized subunits. In practice, however, this extreme region can easily be avoided by observing the corresponding subunit structures.

The most important feature of the stability curve is that it exhibits a distinctive bend after the initial sharp increase before it plateaus or decreases. The location of the bend indicates a suited strength of sparsity regularization, providing a solution that is as stable as possible while affecting its structure as little as possible. Different techniques can be used to extract a specific regularization parameter value from the location of the bend, such as via a coefficient obtained from a fitted saturation curve. Here, we simply selected a parameter value by visual inspection. After determining a suitable regularization parameter, we run STNMF in that configuration initialized with NNSVD-LRC to obtain the final set of localized subunits.

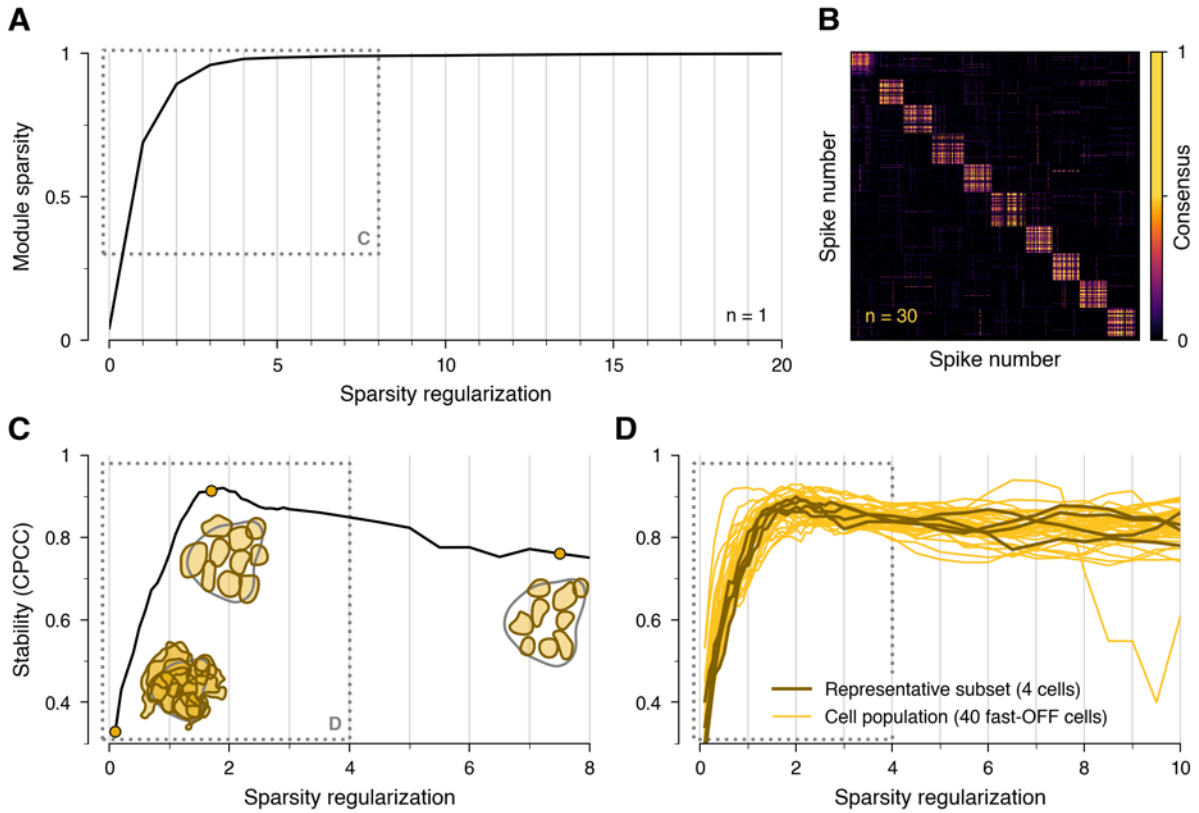


**Figure 4. Consensus analysis for determining the suitable weight of sparsity regularization.** (**A**) Module sparsity (fraction of zeros in the modules) of the decomposition with increasing sparsity regularization for a sample salamander cell. The first step in the analysis is finding the window (dotted frame) in which regularization starts to dominate the decomposition, that is, module sparsity approaches unity. Within that range, the consensus of decompositions is probed. (**B**) Reordered consensus matrix showing the pairwise agreement between spike pairs across 30 decompositions. High consensus is indicated by the ten clusters on the diagonal corresponding to ten reliably recovered subunits as visualized in the second subunit layout inset of (**C**). (**C**) Stability curve for the same cell as obtained from consensus matrices at different regularization strength. The diagonal clustering structure in (**B**) is measured with a scalar value, the cophenetic correlation coefficient (CPCC). The correlation increases with increasing sparsity regularization, supported by refined subunit outlines (first and second inset). After a peak in stability, subunits become too sparse (third inset) until they eventually vanish when regularization becomes too strong. (**D**) The stability curves for a wide range of regularization parameters of a subset of cells (dark) compared to the curves of all 40 fast-OFF cells (yellow) from a salamander retina. Dotted frame indicates the range of interest as specified in (**C**).

*Suitable regularization parameters can be inferred from a subset of cells*

We applied the consensus analysis to the salamander data of retinal ganglion cell recordings. To obtain a suitable sparsity regularization parameter for subsequently recovering the subunits for all cells, we proceeded as follows. Among the cells belonging to the same functional type, we selected a representative subset that resides near the population average for properties like receptive field diameter, number of spikes, and amount of noise in the STA by visual inspection. Here, with our computationally fast implementation of STNMF at hand, we were able to probe the effect of sparsity regularization quickly as a first step to narrow down the range of interest. We ran STNMF with NNSVD-LRC with increasing sparsity regularization to find an upper bound after which all values in the modules approached zero (**Figure 4A**). As our AF-HALS implementation of STNMF imposes sparsity at multiple sub-steps per full iteration, we found 200 iterations to give a sufficient estimate of the sparsity, measured as the fraction of zero values in the modules. Within the determined range, we investigated the solution stability for a coarse selection of regularization parameter values using the consensus across repeated STNMF runs of random initialization (**Figure 4C**). As this merely serves to further narrow down suitable bounds of regularization, we only performed five repetitions at each sparsity value to save time. For cells with nonlinear receptive fields that do in fact exhibit subunits, the stability curve rises

sharply and shows one initial peak before plateauing or decreasing (**Figure 4C**), as expected. The area around the peak is the range of interest as it captures both ends of the impact of sparsity regularization. Within this window, we then performed the full consensus analysis at high resolution of regularization coefficients on the subset of cells (**Figure 4D** gray frame). We propose to perform between 20 to 30 repetitions for each regularization parameter value with 1000 STNMF iterations or until convergence to obtain a reasonable estimate of stability. Finally, we extract the suitable sparsity regularization parameter from the bend in the average of the stability curves by visual inspection. Beyond the subset of cells used for tuning the regularization, we then applied that regularization parameter to NNSVD-LRC-based STNMF to the entire population of cells of that type to identify the final subunit layouts for all of them. The process was repeated for each cell type.

By applying consensus analysis to the full population over a wide range of sparsity regularization, we verified that parameters inferred from a subset of cells can be extrapolated to the entire population effectively. For the sake of thoroughness, we analyzed a range of 44 sparsity parameter values between zero and ten, using a comparatively large number of repetitions (50) for each of them (**Figure 4D** bright curves). Note, that this extensive analysis takes substantially more time than the procedure described in the paragraph above and was performed here for comparison purposes only. Occasionally, the stability curve can abruptly decrease and approach low values, as seen in **Figure 4D** for one example. This occurs due to excessive regularization strengths beyond the appropriate level, breaking subunits arbitrarily into smaller structures and thereby preventing consensus across repeated decompositions. Otherwise, the stability curves are generally well in agreement with each other so that an arbitrary subset of cells is representative of the population here (**Figure 4D**).

*Subunits match previous analyses of bipolar cell receptive fields*

We recovered subunits of the fast-OFF ganglion cells in the salamander data given the procedure described above. To estimate the subunits of all cells, STNMF took a total of five minutes on a conventional office computer averaging at around 9 seconds per ganglion cell. This is well beyond a 100-fold reduction in computation time compared to the earlier implementation based on the active-set method. Yet, the obtained subunits match the results from that previous analysis (Liu et al. 2017). We observe a close resemblance in the set of subunits of individual cells (**Figure 5A**) and find the subunit outlines to align for the cell population (**Figure 5B**). Furthermore, we were able to reproduce previously reported overlaps between subunits across cell pairs (**Figure 5C**). For some of the receptive fields, we recovered additional subunits to the otherwise matching subunit layout that had not been detected before. The differences we observe typically concern subunits with relatively small average weights in the STNMF decomposition. One possible explanation for these unmatched subunits may be the differing procedures to determine the final subunit layout. In the present work, the subunits correspond to the localized modules from STNMF without further curation. In the previous analysis (Liu et al. 2017), the final subunits were obtained from the average of reoccurring subunits over multiple repetitions. Lower-weight subunits that occurred less frequently across the repetitions might have been more likely to be excluded by the previous method.

Experiments combining multi-electrode recordings of ganglion cells with single-cell recordings of bipolar cells had been performed previously in salamander retinas to investigate the relationship between computationally inferred subunits and measured bipolar cell receptive fields (Liu et al. 2017). Here, we confirm the results using the data of the respective ganglion cells. We find similar overlaps between subunits and bipolar cell receptive fields (**Figure 5D**). These findings suggest that our implementation can reveal bipolar cell receptive fields as subunits.

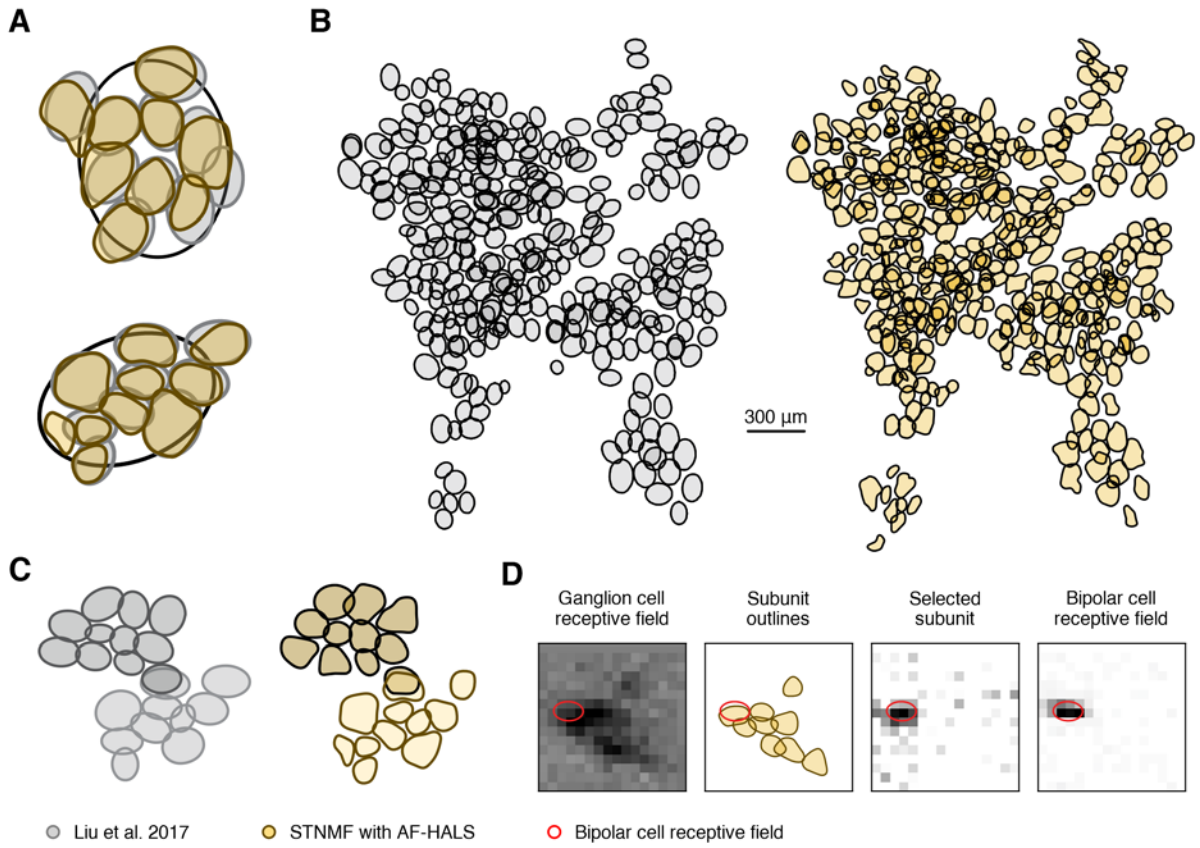


**Figure 5. Subunits of AF-HALS-based STNMF match previous findings and bipolar cell receptive fields.** Subunit layouts of salamander cells (yellow) are compared to Liu et al. (2017; gray). Subunits are represented as 1.5-sigma ellipses of Gaussian fits in gray, and as yellow contour lines (see Methods). (**A**) Decomposition for two example cell receptive fields. Receptive field outlines are 1.5-sigma ellipses of Gaussian fits. The subunit layouts resemble the previously estimated subunits. (**B**) Comparison of subunit mosaics of a population of fast-OFF cells, colors like in (**A**). Most of the previously identified subunits as well as some additional ones are recovered. (**C**) Examples of subunit layouts for adjacent fast-OFF ganglion cells with overlapping (shared) subunit, as obtained with both methods. (**D**) Comparison of recovered subunits with receptive field of a simultaneously recorded bipolar cell. One of the subunits (yellow outlines) matches the bipolar cell receptive field (red outline). Panels (**A** gray), (**B** left), (**C** left), and (**D** rightmost) are adapted from Liu et al. (2017) licensed under CC BY 4.0.

*STNMF recovers subunit mosaics for all of the four major cell types in the primate retina*

With the new, fast implementation of STNMF in place, we next aimed at performing subunit analyses of ganglion cell populations in the primate retina. To this end, we recorded spiking activity from ganglion cells in isolated marmoset retina under white-noise stimulation. We identified the four major cell types of the primate retina, ON and OFF parasol as well as midget cells, according to their receptive field sizes and temporal filters (see Methods). STNMF was able to identify subunits from the spiking data of individual cells for each of the four cell types (**Figure 6A**). Furthermore, the separation of the cells into four distinct types with at least partial tiling of receptive fields (**Figure 6B**) allowed us to identify the layout of subunits on a cell-type-specific population level (**Figure 6C**). By doing so, we found that subunits associated with cells of the same functional type also tile the retina in a mosaic-like fashion. Subunit layouts from different ganglion cells of the same type intertwine nearly seamlessly so that it is not easily visible where the boundaries between ganglion cell receptive fields are.

The availability of four distinct cell-type populations also allowed us to investigate the generality of the consensus analysis and to probe whether optimal sparsity regularization is cell-type-specific. As cells of the same functional type share properties like receptive field size and firing rate more so than cells of different types, the required sparsity may differ across cell types.

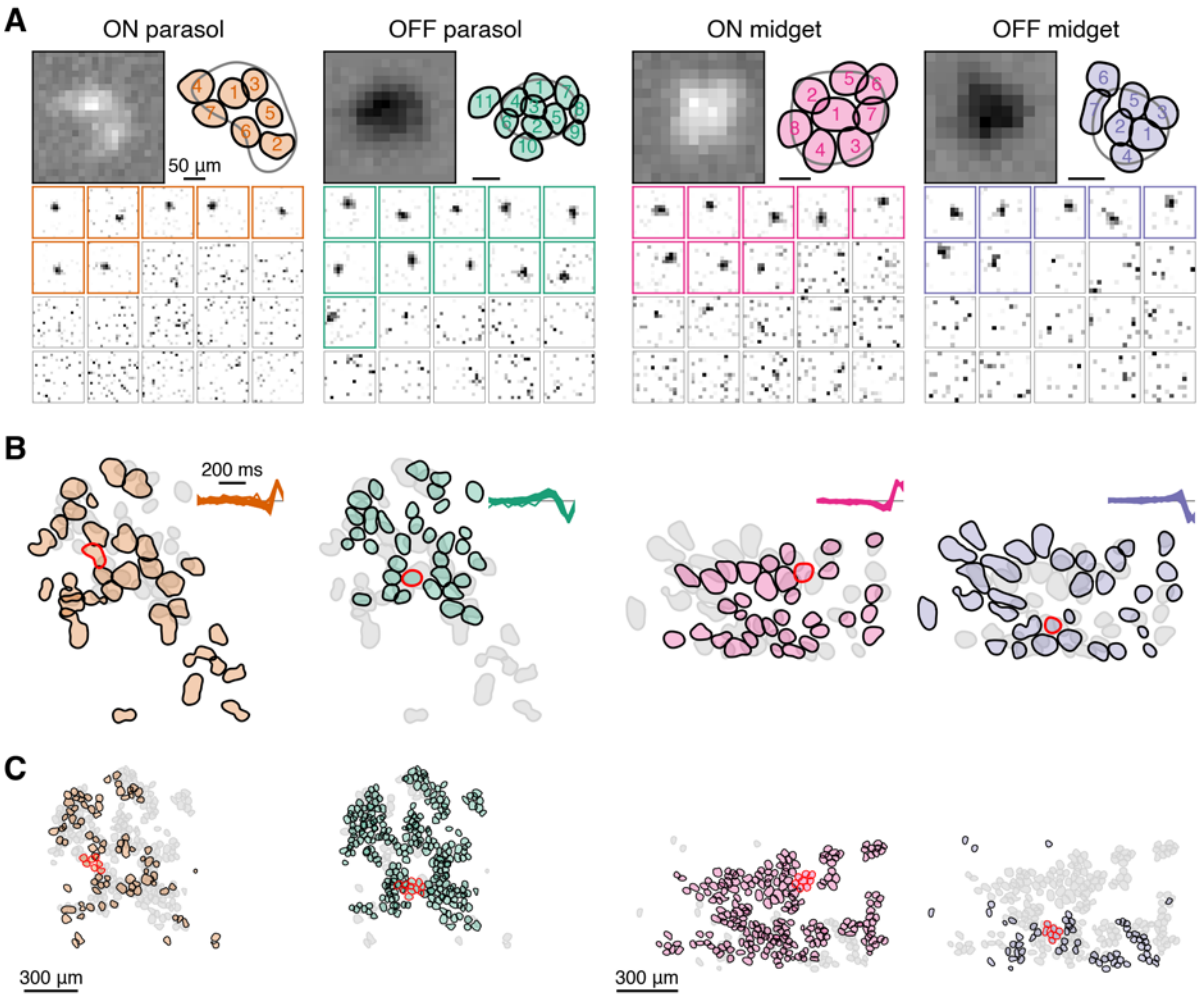


**Figure 6. STNMF recovers subunits in all four major cell types of the marmoset retina.** Data is shown from ON and OFF parasol cells (one retina; mid-periphery) and ON and OFF midget cells (one retina; periphery). Note that the scale bars differ between the four sample cells. (**A**) Receptive fields, STNMF modules, and subunit layouts of sample ON and OFF parasol cells and ON and OFF midget cells. Modules identified as localized subunits by their autocorrelation (see Methods) are marked by colored frames and depicted together in a colored subunit layout of contours superimposed on the receptive field contour (gray; top right). The outlines are numbered in descending order according to the mean STNMF weight. (**B**) Receptive field mosaic of the four main cell populations (colored). Gray outlines correspond to the receptive fields of the corresponding type with reversed polarity of preferred contrast. The temporal dynamics of the receptive fields are depicted by the temporal filters of the spike-triggered average (insets; scale bar 200 ms). The sample cells of (**A**) are marked in red. (**C**) Subunit mosaics corresponding to (**B**). Sample cells marked in red. Not all OFF midget cells have nonlinear subunits, as reported previously (Karamanlis et al. 2023). Scale bars (300 µm) correspond to both (**B**) and (**C**).

To analyze whether this is reflected in the optimal sparsity regularization parameter, we examined the regularization-dependent consensus and compared the resulting stability curves for a wide range of regularization parameter values. For each of the four types, we found stability curves as expected, with a bend that indicates the optimal regularization strength (**Figure 7A**). Within each type, the shapes of the stability curves exhibit similarities, but this does not hold across types.

Comparing, for example, the stability curves for ON versus OFF midget cells reveals slight, but systematic differences (**Figure 7B**). The slope of the stability curves for OFF midget cells tends to be somewhat shallower around the bend than for ON midget cells. The difference can be illustrated by analyzing the family of curves with principal component analysis and comparing the shapes of the curves according to the projections on the first two principal components (**Figure 7C**). This shows that curves of cells for the same type tend to group together but diverge across types, despite some overlap of the two populations. It is likely that differences in the levels of noise and in the subunit nonlinearities contribute to these cell-type differences of regularization effects.

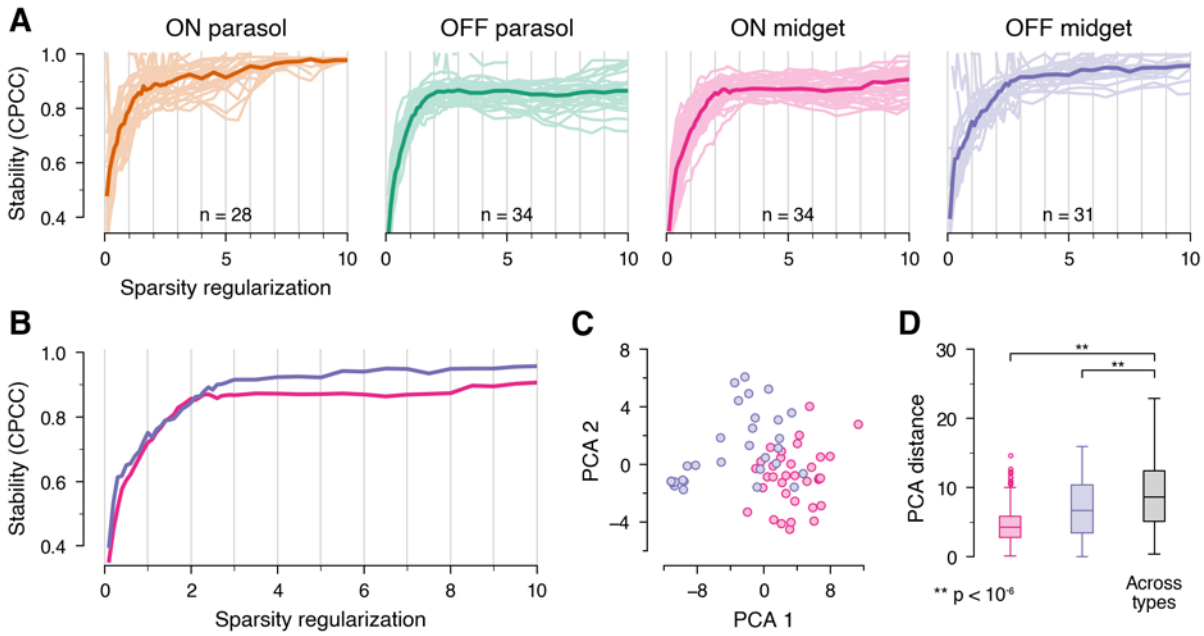


**Figure 7. Stability curves align well among cells of same functional type.** (**A**) Stability defined by the cophenetic correlation coefficient (CPCC) for different sparsity regularization strength for ON and OFF parasol cells (one retina) and ON and OFF midget cells (one retina). Consensus for cells without nonlinear integration in their receptive fields may be undefined as indicated by incomplete stability curves. The dark curves represent the medians (excluding not defined values) of the corresponding population. Stability curves show the expected increase and plateau. (**B**-**D**) Consensus comparison of ON and OFF midget cells. (**B**) Superimposed stability curves of the different cell types, showing slight but systematic differences. (**C**) Stability curves projected onto the first two principal components of all curves from the two populations. (**D**) Inter- and across-type Euclidean distances in the two-dimensional PCA space, showing significant differences of distances across versus within cell types. Central line and box represent the median and the interquartile range (IQR; 1st to 3rd quartile), respectively. Whiskers extend to most extreme values within 1.5 IQR, and dots indicate outliers.

More quantitatively, we find that the distances between pairs of curves are significantly larger across cell types than for curves of the same type (**Figure 7D;** two-tailed Wilcoxon rank-sum test, $p < 10^{-6}$ for both ON and OFF cells versus across types). This indicates that the stability analysis is similar for cells within a type, but can differ across types, suggesting, along with the tight alignment of the stability curves observed in the salamander dataset (**Figure 4D**), that a suited sparsity regularization parameter can be inferred from a representative subset of cells for each functional cell type, but that different types may require different levels of regularization. The subunit populations in **Figure 6C** are based on cell-type specific regularization parameters.

*Subunits reflect anatomical differences between parasol and midget cell circuits*

The sizes of subunits are fairly homogeneous within functional types, but differ between parasol and midget cells (**Figure 8A;** subunit diameters mean ± standard deviation: 60.5±16.7 µm for ON parasol cells, 56.3±10.9 µm for OFF parasol cells, 44.6±12.2 µm for ON midget cells, 38.1±12.9 µm for OFF midget cells, $p < 10^{-6}$ for both ON parasol versus ON midget and OFF parasol versus OFF midget, two-tailed Wilcoxon rank-sum test, data from three marmoset retinas), suggesting distinct presynaptic inputs. Parasol cells in primates receive excitatory inputs from diffuse bipolar cells (Calkins and Sterling 2007; Jacoby et al. 2000; Wässle 1999), whereas midget cells are primarily connected to midget bipolar cells (Jusuf et al. 2006a, 2006b; Kolb and Dekorver 1991). We found the average midget subunit diameter of around 42 µm to match the receptive field size of midget bipolar cells (31–51 µm) as measured in the peripheral macaque retina (Dacey et al. 2000). The average parasol subunit diameter of around 58 µm lies between reported anatomical dendritic tree sizes (30–50 µm) (Boycott and Wässle 1991) and measured receptive field sizes (74–114 µm) (Dacey et al. 2000) of diffuse bipolar cells in the peripheral macaque retina.

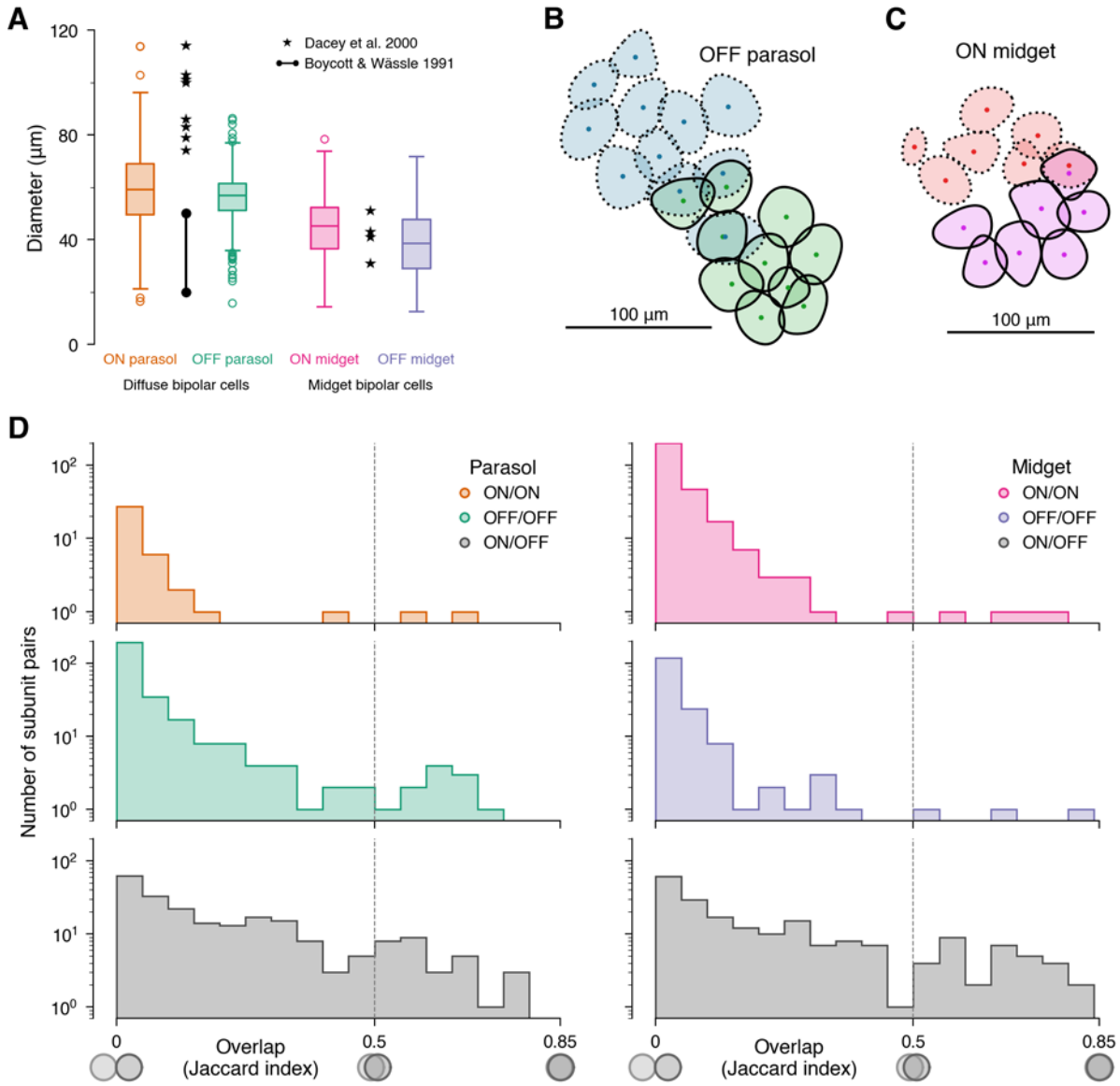


**Figure 8. Subunits match anatomical properties of bipolar cells.** (**A**) Subunit diameters of the four major cell types collected from three marmoset retinas. (Box plots like in **Figure 7D**.) Parasol and midget subunits are comparable in size to diffuse bipolar and midget bipolar cells, respectively. Bipolar cell sizes were previously reported in the macaque retina according to physiological measurements (Dacey et al. 2000) and to calculations based on cone inputs (Boycott and Wässle 1991). (**B-C**) Examples of the subunits of adjacent parasol (**B**) and midget (**C**) cells. Markers represent centers of mass. While the parasol subunits can display substantial overlap from adjacent cells, the midget cell subunit layouts are typically distinct with no or few overlapping subunits. (**D**) Histograms of subunit pair overlap. Overlap between subunit pairs within ON (top), OFF (middle), and across ON and OFF (bottom) parasol and midget cells is measured with the Jaccard index. Only overlap pairs with an index greater than zero are visualized. The dashed line indicates the lower bound of what we consider here as substantial overlap, suggesting shared bipolar cell inputs. Consistent with the examples in (**C**), we observed few overlapping pairs for midget cells, but a considerable amount (**B**) for OFF parasol cells.

Even though these values and our subunit analyses come from different primate species, the matching sizes, in particular the relative differences between parasol and midget inputs, suggest a correspondence between subunits and bipolar cell inputs.

Midget cells in our analysis yielded on average 5.7±2.3 subunits and parasol cells had 6.3±2.7 subunits (mean ± standard deviation). Furthermore, to analyze how the subunits of different cells relate to each other, we identified all pairs of subunits from different individual ganglion cells that displayed an overlap and quantified the relative overlap by the Jaccard index. The maximum value of this index is unity, which would indicate identical, fully overlapping subunits, whereas values near zero correspond to minimal overlap. Particularly high relative overlap of pairs of subunits occurred mostly for OFF

parasol cells, as illustrated in an example of subunits from two cells in **Figure 8B**. Midget ganglion cells, on the other hand, showed far fewer near identical subunits obtained from two neighboring ganglion cells of the same type. Rather, subunit layouts of neighboring cells complemented each other, jointly tiling space with little overlap (see example in **Figure 8C**). Consistent with anatomical evidence, we find a higher degree of subunit sharing among parasol cells compared to midget cells. Dendritic trees of midget ganglion cells do not overlap independent of eccentricity, while parasol cells do show overlap in the periphery primate retina (Dacey 1993; Lee et al. 2010), suggesting the possibility of systematically shared input.

*Different ganglion cell types display distinct subunit mosaics*

Our results indicate that subunits of ganglion cells belonging to the same functional type arrange in a tiling pattern across the retinal space (**Figure 6C**). Within functional types, we thus expect pairwise overlap of subunits to typically remain small. To test this, we computed the relative overlap for each pair of subunits for parasol cells (**Figure 8D** left) and for midget cells (**Figure 8D** right). Within subunit pairs of ganglion cells from the same type, we observe only few cases of relative overlap beyond 50% of overlap, indicating that violations of subunit tiling are rare. However, some subunit pairs display particularly strong overlap, especially for pairs obtained from OFF parasol cells. These may indicate shared subunits, as would be expected from joint input to the two cells from the same bipolar cells. Unlike within types, subunits do not arrange regularly across ganglion cell types. For parasol versus midget cells, this is already indicated by differences in subunit size (**Figure 8A**). For ON versus OFF parasol cells as well as for ON versus OFF midget cells, we observe a continuum of overlap values with many violations of tiling (overlap larger than about 50%) in subunit pairs of one ON- and one OFF-type ganglion cell (**Figure 8D** bottom). As the violations occur substantially more often across than within ON and OFF types, they suggest the existence of distinct ON and OFF subunit mosaics.

*ON- and OFF-type subunit mosaics are aligned*

Having characterized large-scale subunit mosaics of ON- as well as OFF-type ganglion cells of corresponding types, we next asked whether these mosaic pairs display any type of spatial coordination with each other. For example, the layouts of subunits from ON and OFF parasol cells might be aligned with each other, with approximately matching subunit locations between the two mosaics. Alternatively, the mosaics might be anti-aligned with subunit locations of one population lying preferentially in between several subunits of the other population. Or the two mosaics might be independent of each other, displaying no specific coordination. On the level of receptive fields of ganglion cells, a recent study had demonstrated that the mosaics of ON and OFF cells of corresponding types (parasol ganglion cells in macaque) tend to be anti-aligned, following a prediction from efficient coding theory (Roy et al. 2021). We thus asked whether a corresponding coordination can be observed at the preceding processing layer, as the subunits reflect the organization of the inputs coming to the ganglion cells, which stem from their presynaptic bipolar cells.

For both parasol and midget cells, we selected a recording with particularly dense coverage of receptive fields in order to investigate alignment between ON- and OFF-type mosaics within regions with good coverage of identified subunits (**Figure 9A**). To quantify the level of alignment, we followed the procedure proposed by Roy et al. (2021). We reduced the contour of each receptive field to a single coordinate point, defined by its center-of-mass. We then computed the inter-mosaic coordination energy (IMCE), which describes the pairwise spacing between centroids across mosaics of different cell types (Roy et al. 2021). In this analysis, each heterotypic pair of centroids contributes energy to the IMCE depending on the distance between the two centroids, with smaller distances corresponding to higher energies. Thus, higher values of the IMCE indicate stronger alignment between the mosaics.

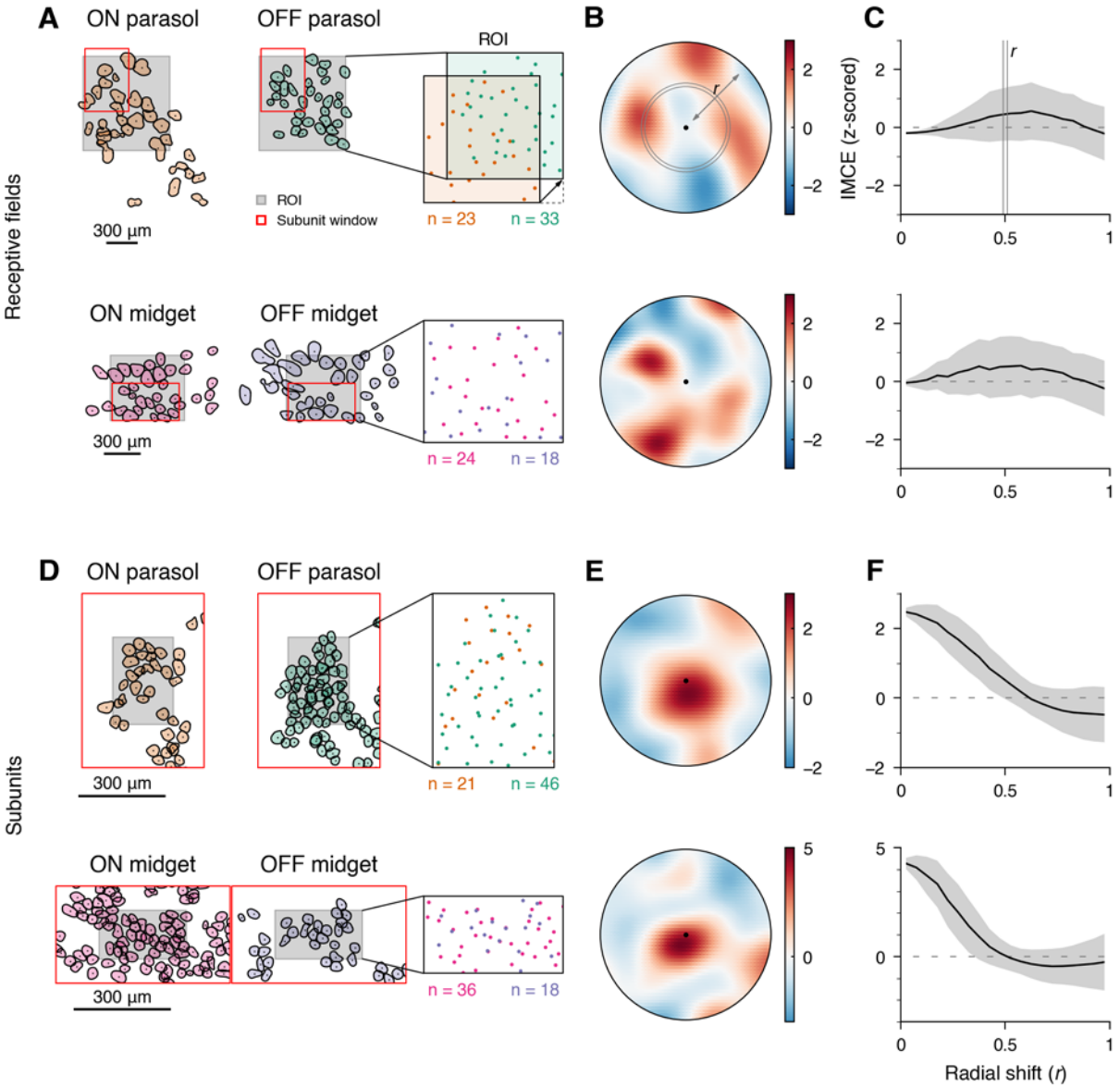


**Figure 9. ON- and OFF-type subunit mosaics are aligned, while receptive field mosaics show anti-alignment.** (**A**) Contour mosaics of ON- and OFF-type parasol (top) and ON- and OFF-type midget (bottom) ganglion cell receptive fields, each from one retina, mid-periphery and periphery, respectively. Centers of mass (black markers) inside the region of interest (ROI, shaded area) are considered for analysis (colored markers inside black frame, right). One center-of-mass mosaic is shifted relative to the other, as schematically indicated (right); n = numbers of enclosed subunits. (**B**) Topographical map visualizing the inter-mosaic coordination energy (IMCE) at different mosaic shifts around the original position in the center. (**C**) Dependence of the IMCE on radial distance by averaging across angles in (**B**), displaying mean (black) and standard deviation (shaded gray). IMCE topographical map and radial average curves are z-scored, and the shift distance *r* is normalized to the median homotypic nearest-neighbor distance of the ON-type mosaic. (**D-F**) Same as (**A-C**), but for the subunit mosaics of the cell populations analyzed in (**A-C**). The red rectangles in (**D**) correspond to the red rectangles of the corresponding cell populations in (**A**). The decreasing IMCE curves in (**F**) indicate that corresponding ON and OFF subunit mosaics tend to be aligned.

The analysis consists of observing how the IMCE changes as one mosaic is shifted vertically and horizontally relative to the other (**Figure 9A** right). For aligned mosaics, paired centroids are close to one another and the energy is initially high and decreases with shifting. For anti-aligned mosaics, the energy rises, as centroids shift closer to each other. Finally, for independent mosaics, there is no distinct increase or decrease. We visualize the IMCE for different mosaic offsets as a topographic map (**Figure 9B**). To start, we consider the alignment of ganglion cell receptive fields themselves. For both parasol and midget receptive field mosaics, the energy increases with the relative shift between the ON and OFF populations, as evident in the radial average of the topographic map (**Figure 9C**). This indicates

that the ON- and OFF-type receptive field mosaics are anti-aligned, as previously reported for parasol cells in the macaque retina (Roy et al. 2021).

We then applied the procedure also to the corresponding subunit mosaics (**Figure 9D**) of the same parasol and midget cell populations. Unlike for the receptive fields, the IMCE of the subunit mosaics was found to be largest for zero shift, and a decrease with mosaic shift is apparent for both parasol and midget subunit mosaics (**Figure 9E-F**). This indicates that, in contrast to receptive field mosaics, subunit mosaics of corresponding ON and OFF cells are aligned. Note that mechanistically, such a switch from aligned subunits to anti-aligned receptive fields at the next layer of neuronal processing is not contradictory; ON and OFF ganglion cells may simply pool their signals from (aligned) subunits over distinct, anti-aligned spatial regions.

Note further that the mosaics are partially depleted due to missing receptive fields of non-recorded cells. Although the procedure is robust to subsampling to a certain extent (Roy et al. 2021), we had manually chosen regions of interest for both parasol and midget populations, in which we find few gaps in the tiling for both ON- and OFF-type mosaics (**Figure 9D** right). The tested portion of the mosaic pairs include around 20–50 subunits each. Also with larger regions of interest defined by the convex hull of centroid pairs excluding outliers, as proposed (Roy et al. 2021), the IMCE remained sharply decreasing with relative shifts of the subunit mosaics, corroborating the alignment of the ON- and OFF-type subunits.

## DISCUSSION

We propose a method for subunit identification based on spike-triggered non-negative matrix factorization (STNMF), which finds subunits within a receptive field from recorded spike-train data in a matter of seconds of computation time. The approach enables, for example, large-scale analyses of the primate retina, recovering the nonlinear subunits of the four major ganglion cell types: ON and OFF parasol, and ON and OFF midget cells. This allows in-depth cell-type-specific analyses of subunits on a cell population level, comparison to type-specific anatomical properties, and new insights about the spatial coordination of ON- and OFF-pathway signal transmission.

### *Recovered subunits resemble bipolar cell receptive fields in the primate retina*

We applied the method to recorded spike trains from marmoset retinas and analyzed their subunit properties. The disparities we reported on the subunits of parasol versus midget ganglion cells align with existing anatomical data. Parasol cells exhibited larger subunits and a higher degree of overlap among neighboring cells compared to midget cells, reflecting their overlapping dendritic trees and shared presynaptic inputs (Dacey 1993; Lee et al. 2010).

There are suggestions that functional subunits of primate parasol cell receptive fields correspond to diffuse bipolar cell inputs (Crook et al. 2008, 2014). Since our subunits from marmoset do not exceed receptive field and dendritic tree size estimates of macaque diffuse bipolar cells (Boycott and Wässle 1991; Dacey et al. 2000), it is unlikely that the subunits are composed of combinations of multiple bipolar cell inputs. To the extent of our comparison of midget and parasol cells, the retina of marmoset and macaque are reported to be functionally equivalent (Lee et al. 2010), though absolute comparisons of subunit and dendritic tree sizes should be taken with caution. Yet, the matching relative sizes of larger subunits and larger presynaptic bipolar cells for parasol compared to midget cells resonate with the idea of subunits reflecting bipolar cell input. Together, our findings concerning subunit density, uniform tiling, and potential sharing of subunits between pairs of parasol cells support the notion that the parasol subunits may correspond to individual bipolar cell inputs. Note, though, that these bipolar cell inputs do not occur in isolation, but are likely shaped and modulated by interactions with amacrine cells (Jacoby

et al. 1996; Kolb and Dekorver 1991) and gap junctions between individual bipolar cells (Appleby and Manookin 2020; Jacoby et al. 2000; Manookin et al. 2018).

For midget cell subunits, the correspondence to bipolar cell input may seem more difficult. On average, midget ganglion cells appear to receive input from around eight (mid-periphery) up to 13 midget bipolar cells (periphery) in the marmoset retina (Jusuf et al. 2006b). We typically found fewer subunits per midget ganglion cell. Yet, these subunits of midget cells form a mosaic at the population level and subunit sizes seem consistent with what one might expect for receptive fields of individual bipolar cells. It is possible that the discrepancy in the numbers of subunits and connected bipolar cells reflects that some subunits do not correspond to individual bipolar cells or that some subunits remain undetected. Alternatively, some anatomically connected bipolar cells might not have a strong functional connection to the postsynaptic midget cell under the applied spatiotemporal white-noise stimulus.

*Subunits arrange in a uniform mosaic by cell type*

Our subunit analysis with STNMF revealed distinct mosaics of subunits for ON and OFF parasol, and ON and OFF midget cells in the marmoset retina. The distinct subunit mosaics are in line with ganglion cell types stratifying in different levels of the inner plexiform layer (Watanabe and Rodieck 1989) and receiving input from distinct bipolar cell types (Masland 2001; Wässle 2004). The observed uniformity of the subunit mosaics matches reports of bipolar cell input density to be independent of the size of the dendritic tree of the ganglion cell in the marmoset retina (Eriköz et al. 2008).

There is recent evidence that the mosaics of ON and OFF parasol ganglion cell receptive fields in the macaque retina are spatially anti-aligned (Roy et al. 2021). Models of efficient coding indicate that anti-alignment is the optimal arrangement between populations of ON and OFF cells in a high-noise/high-threshold coding regime, in which ganglion cells likely operate, as it allows nearby ON and OFF cells to encode non-redundant information (Jun et al. 2021). Increased resolution is an additional benefit of anti-alignment, and it may have advantages for down-stream convergence (Roy et al. 2021). Our present analysis confirmed the anti-alignment of ganglion cell receptive fields. By contrast, the same analysis applied to the identified subunit mosaics indicated alignment rather than anti-alignment between ON- and OFF-type subunit mosaics for midget as well as for parasol cells. Thus, there seems to be a switch from aligned to anti-aligned spatial coordination when going from the bipolar cell layer to the ganglion cell layer. A reason for this switch could be that the conditions that make anti-alignment efficient for ganglion cell receptive fields do not hold at the level of bipolar cells. Due to the finer mosaics of bipolar cells, anti-alignment may offer little additional benefit in spatial resolution. Furthermore, the accumulated noise at the bipolar cell level, where signals are carried by graded potentials rather than stochastic spikes, is likely lower than after spike generation in ganglion cells (Berry et al. 1997), and – in contrast to the high-noise scenario – alignment is thought to be optimal for efficient coding in the low-noise regime (Jun et al. 2021).

*Accelerated Fast HALS as a fast and versatile method for STNMF*

The large-scale analyses were made possible by the improved methodology of different aspects in the STNMF approach. Through purpose-built initialization, the method is steered towards a suitable solution, and by means of a tailored NMF algorithm, it is scalable to large-scale population recordings. Furthermore, by adjusting consensus methods to our needs, we demonstrated a viable method for hyperparameter selection. Our findings suggest that the selection of the sparsity regularization parameter based on a subset of neuronal cells is transferable to cells of the same functional type. Subsequent analyses for cells of the same type reduce to a parameter-free procedure that can be applied in a plug-and-play fashion.

Solving a non-negative least squares problem (NNLS) sits at the core of NMF. To solve it accurately, typically, costly algorithms are employed iteratively (Kim and Park 2011). These include multiplicative

updates (Lee & Seung, 1999), the projected quasi-Newton method (D. Kim et al., 2007), the projected gradient method (Lin, 2007), non-negative quadratic programming (Zdunek & Cichocki, 2008), and the active-set method (Lawson & Hanson, 1974). The algorithms strive to provide exact solutions at each iteration by taking into account all data at once. Here, we avoid the high computational demand with hierarchical alternating least squares (HALS) by solving smaller, non-restricted least squares problems followed by setting all negative values to zero (Cichocki et al. 2007; Ho 2008; Li and Zhang 2009). While Fast HALS (Cichocki and Phan 2009) simplifies computational operations for speed, A-HALS (Gillis and Glineur 2012) refines the coarse approximations of HALS by performing multiple cycles of updates within one NMF iteration. The intended purpose of A-HALS is to accelerate the convergence by reducing the number of necessary iterations, which lends it its name. However, by incorporating sparsity regularization, we make use of another advantage of this procedure. Additional constraints like sparsity regularization are typically integrated into the equations of the iterative update (Kim and Park 2007; Liu et al. 2017). As we instead perform it multiple times, that is, at each A-HALS cycle within one NMF iteration, the continuously applied sparsity constraint refines the subsequent repetitions by giving more weight to prominent structures and progressively suppressing the effect of uncorrelated noise. Consequently, the result after one full iteration offers relatively high accuracy and sparsity. Our AF-HALS method with sparsity regularization may prove to be a valuable tool not only for the estimation of subunits, but also for other applications of NMF.

*Sparsity regularization as a crucial ingredient of STNMF*

In the field of machine learning, sparsity regularization has become an important tool to combat overfitting and has earned the role of an enhancement strategy to allow generalization when data is limited. In many applications of signal processing and learning representations, however, sparsity has proved to be an even more essential and critical player (Candès et al. 2006; d'Aspremont et al. 2007; Donoho 2006; Hoyer 2004). Sparsity in NMF offers interpretable solutions (d'Aspremont et al. 2007) and even inherently reduces the impact of noise by decreasing the effects of uncorrelated patterns in inputs (Donoho 1995; Elad 2006; Hyvärinen 1999). Aside from these benefits, sparsity is an integral part of STNMF. It replaces the dropped non-negativity constraint of one factor matrix in semi-NMF (Hoyer 2002, 2004; Kim and Park 2007, 2008), as otherwise the parts-based reconstruction can collapse (Ding et al. 2010). In the context of STNMF, sparsity is thus a vital component for producing localized subunits.

Beyond simple $\ell 1$-norm-based sparsity, there have been other efforts towards regularization techniques that favor localized solutions in the fields of NMF and clustering such as decomposing widefield calcium imaging signals into localized functional brain regions (Saxena et al. 2020). The prior definition of the localized regions based on a brain atlas in this previous study, however, does not easily translate to our problem of identifying subunits of unknown number, locations, and shapes. A different approach to enforce localized solutions is to apply a locally normalized $\ell 1$-regularization, which was introduced with spike-triggered clustering (Shah et al. 2020) to instill spatial information into the $\ell 1$-regularization in subunit estimation. The penalty of sparsity regularization at a given location of a subunit was adjusted based on the spatially neighboring values. For our application, the forced locality is not directly applicable, as it makes a module either contain a localized pattern or collapse to no non-zero values. Noise in the input cannot easily escape to excess modules, and the number of included modules would have to be precisely determined instead of using an upper bound on the number of expected subunits as is done here. An alternative approach to regularization with localized structures comes from using basis functions to parameterize stimulus filters. For example, using spline-based parametrizations can help estimate receptive fields and has also been combined with subunit-based model fits (Huang et al. 2021).

*Parameter selection by consensus analysis*

To find a suitable amount of sparsity regularization, we investigated how robust the decomposition is with respect to varying initialization. To measure how stable the decomposition is, we here adjust the consensus procedure proposed by Brunet et al. (2004) to be applied to the problem of identifying regularization parameters. An alternative to using consensus analysis is L-curve fitting. The L-curve technique (Hansen 1999a, 1999b) aids in determining a suitable parameter for a regularized least squares problem. The method inspects the relation between two norms, the fitted data and the regularization term. With too much regularization, the data will not be properly accounted for, and with too little regularization, the fit will be good but dominated by undesired effects. Both terms are visualized against each other for a range of regularization values, exhibiting a curve shaped like the letter "L". Determining a corner point can then be used to obtain a suitable trade-off between the two terms for the least squares optimization and the regularization penalty. An advantage over our proposed consensus analysis is the conceptual simplicity and that it does not require multiple repetitions of NMF. It has been applied to regularized NMF with HALS previously, although for a different type of regularization (Cichocki and Phan 2009). Nevertheless, when applied to STNMF, we did not find a sharp corner with the data available, and also the dedicated method of L-corner fitting (Hansen 1994) did not succeed. We therefore turned towards the more involved consensus analysis, which we found to provide a good alternative.

*NNSVD-LRC offers reliable initialization*

We demonstrate that a single run of STNMF with SVD-based initialization can replace the selection of the most suitable solution among 100 randomly initialized runs. Specifically, NNSVD-LRC is advantageous, as it converges faster to a stationary point due to low error and increased sparsity in the initialization (Atif et al. 2019). Notably, on the one hand, techniques based on SVD do not guarantee NMF to find the global optimum (Atif et al. 2019). On the other hand, however, seeking the global optimum in network structures bears the risk of overfitting (Choromanska et al. 2015). Furthermore, the alternative of selecting a most suitable among many randomly initialized solutions raises challenges because a measure of suitability is not easily defined and a manual selection of a best solution or a particular criterion is subject to human bias. To alleviate these problems, an adapted consensus clustering procedure (Zhou et al. 2020) poses a viable alternative for initialization. Instead of choosing the best decompositions among 100 randomly initialized NMF runs, the resulting modules are concatenated to serve as the input to another NMF run that yields a final set of modules. We found that decompositions initialized with NNSVD-LRC are comparable with solutions of that approach while requiring only one run of NMF.

*Limitations*

We identify three primary limitations of the proposed method. Firstly, the method poses strict constraints on the stimulus with which the spiking data is obtained. To infer subunits from pixel correlations in the spiking response, a white-noise stimulus with uncorrelated pixel contrast across space is necessary. This demands to allocate time to a specific stimulus during the experimental recording. Here, we were able to recover subunits successfully from less than one hour of recording time. Another caveat is the pixel size of the stimulus. Although STNMF makes few assumptions on the data, the stimulus design must provide stimulus pixels of suitable size to fit bipolar cell receptive fields, whose spatial scales can vary with species and with location on the retina. To aid stimulus design, estimates of dendritic-tree size or measured receptive fields in the literature can be consulted, or relevant spatial scales of nonlinear integration can be estimated from responses to standard reversing-grating stimuli at varying spatial frequency. The third drawback is the inability to produce subunits with both positive and negative components. Although the unconstrained sign of the STNMF weights allows the recovery of subunits that have either purely positive or purely negative entries, mixed structures, such as subunits with an antagonistic surround, are not consistent with the non-negativity constraint. One might speculate that

antagonistic structures could be recovered as separate, additional subunits if their effects are strong enough, but we have not observed such cases in our datasets.

*Comparison to other methods of subunit inference*

Popular computational methods of subunit inference in recent years have been linear-nonlinear cascade models (Maheswaranathan et al. 2018; Real et al. 2017), approaches of convolutional neural network models (Maheswaranathan et al. 2023; McIntosh et al. 2016; Tanaka et al. 2019), and methods of statistical inference (Liu et al. 2017; Shah et al. 2020). Linear-nonlinear-linear-nonlinear (LNLN) models consist of a layer of linear filters of the spatial or spatiotemporal subunits with nonlinear transfer functions that additively converge into another nonlinear transfer function to produce a firing rate or spiking probability. Recent work has demonstrated how the filters and nonlinearities can be trained in a supervised manner on the stimulus input and spiking response (Maheswaranathan et al. 2018). This approach directly yields a generative model for predicting responses to novel stimuli. It was successfully applied to data from retinal ganglion cells of salamanders. To combat complexity, sparsity regularization was applied as well as restricting the stimulus to one spatial dimension. Training may struggle here when going to two spatial dimensions, as the dimensionality of the optimization task will be strongly increased.

Convolutional neural networks are similarly trained end-to-end. The models are often optimized on data from a population of neurons and provide convolutional filters of subunit types rather than individual subunits for each neuron. Training can be done with natural images instead of artificial stimuli like white noise or flickering bars. For models trained with ganglion cell data from the salamander retina, the obtained convolutional filters shared properties with actual bipolar and amacrine cells (Maheswaranathan et al. 2023; Tanaka et al. 2019), although it remains unclear to what extent the model architecture resembles the actual neural circuitry. Spike-triggered clustering (Shah et al. 2020), on the other hand, recovers subunits on stimulus correlations much like STNMF. Subunit number and regularization are here determined through cross-validation, and the obtained models were demonstrated to be capable of predicting spiking responses to natural scenes for primate retinal ganglion cells.

A common trait of the state-of-the-art subunit inference techniques is their high demand on computational power and time. The training of parameters of a convolutional neural network requires powerful GPUs and hours of training. Extensive cross-validation to find the most suitable number of subunits is a reoccurring obstacle as it is time-consuming and often has to be performed for every neuron as the number of subunits may vary. In comparison, our implementation of STNMF offers an analysis of only a few seconds on a conventional office CPU and provides the number of subunits automatically. Hyperparameter search is reduced to a subset of neurons for each cell type making a subunit population analysis considerably faster than previously available.

*Beyond retinal subunits*

Nonlinear signal integration is a ubiquitous feature of neural systems and of sensory signal processing in particular. Analyses based on receptive field subunits are thus applicable to many other sensory areas and have been applied in visual cortex (Almasi et al. 2020; Bartsch et al. 2022; Lehky et al. 1992; Vintch et al. 2015), in medial superior temporal visual cortex (Beyeler et al. 2016; Mineault et al. 2012), and in the auditory system (Ahrens et al. 2008; Harper et al. 2016; Keshishian et al. 2020; McFarland et al. 2013). The methodology presented in the present work is directly applicable to these systems, given sufficiently long recordings with white-noise-like stimuli. In the case of the auditory system, for example, subunits might then correspond to filters in spectro-temporal space, whereas subunits in higher, motion-sensitive visual cortical areas may be partly defined by their direction and speed tuning. Applying STNMF in these systems should aid in developing nonlinear stimulus–response models that link the functional circuitry to the performed neural computations.

In addition to sensory systems, the implementation of NMF developed here together with the presented techniques for controlling hyperparameters and initialization should be of interest for analyzing neuronal population activity. Uncovering hidden structure and dynamics in the joint activity of large ensembles of neurons is seen as an important part of understanding neural computations and information flow in the brain, and a common approach is to identify latent variables of the population activity via dimensional-reduction techniques (Cunningham and Yu 2014; Gallego et al. 2017; Gao and Ganguli 2015; Koh et al. 2022). NMF provides an appealing approach to this challenge, as one may hope that the non-negativity constraint helps infer latent variables—the equivalent of the subunits of the ganglion cell receptive fields—with interpretable structure (Carbonero et al. 2023; Devarajan 2008; Nagayama et al. 2022; Onken et al. 2016). The computational speed achieved by the combination of fast and accelerated HALS as demonstrated in this work should be advantageous for applications of NMF to neural population data, in particular in order to control regularization and other hyperparameters.

## ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 724822) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project numbers 432680300 (SFB 1456, project B05) and 515774656.

## AUTHOR CONTRIBUTIONS

SJZ developed and implemented the AF-HALS-based STNMF method, collected and curated the marmoset retina data, analyzed the data, and prepared figures and the initial draft of the manuscript. DK, MHK, HMS, SS, VR, SK, MM, and DAP all contributed to experiments and data collection. SJZ, DK, MHK, HMS, and TG supplied methodology. TG supervised the project and contributed to writing the manuscript. All authors helped revise the manuscript.

## DISCLOSURE STATEMENT

The authors declare no competing interests.

## DATA AND SOFTWARE AVAILABILITY

The salamander spike trains analyzed in this work come from a publicly accessible repository (https://gin.g-node.org/gollischlab/Liu_etal_2017_RGC_spiketrains_for_STNMF; DOI: 10.12751/g-node.62b65b). We have made the marmoset spike trains from the present study publicly available in this repository: https://gin.g-node.org/gollischlab/Zapp_Gollisch_2023_Marmoset_RGC_spiketrains_under_spatiotemporal_white_noise (DOI: 10.12751/g-node.zpj6rc). The implemented software for STNMF analysis is provided as a Python package. It includes the NMF algorithm and initialization procedures optimized with NumPy for multi-threaded CPU computation. Tools for visualization are also available as well as routines for the consensus analysis and custom extensions via callback functions. The software, along with an extensive documentation and example code for running the analysis, is available in a GitHub repository: https://github.com/gollischlab/STNMF_with_AFHALS.

## METHODS

### *Ethics statement*

All experimental procedures were performed in strict accordance with national and institutional guidelines. Recordings from marmoset retinas were performed on tissue obtained from animals used by other researchers, as approved by the institutional animal care committee of the German Primate Center and by the responsible regional government office (Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit, permit number 33.19-42502-04-17/2496).

### *Electrophysiology*

The salamander data stem from a publicly available repository (Gollisch and Liu 2018) and had been described and analyzed in a previous publication (Liu et al. 2017). Here, we analyzed spike trains of the first retina recording in the repository ("cell_data_01_NC.mat"). The spike trains had been recorded from ganglion cells in isolated axolotl salamander (*Ambystoma mexicanum*) retinas, mounted on 252-channel planar multi-electrode arrays (MEAs) and provided with oxygenated Ringer's solution at a constant temperature of around 22°C.

Data from common marmoset (*Callithrix jacchus*) were recorded from retinas of three animals of different sex and age (female 2.3 years, male 18.1 years, and male 13.2 years). The eyes were removed between three and twelve minutes after time of death. Left and right eye were labeled and dark-adapted for at least one hour. The retina was isolated and cut into pieces, classified at time of dissection by their quadrant (nasal, superior, and dorsal) and eccentricity (central, mid-periphery, and periphery). The temporal quadrant was not considered due to the optic disc and the high density of cell bodies and axons at the fovea. Dissection was performed under infrared illumination on a stereo-microscope equipped with night-vision goggles. Individual retina pieces were removed carefully from the sclera and choroid and placed on MEAs (MultiChannel Systems, Reutlingen, Germany). Pieces not recorded immediately were stored in darkness in oxygenated (95% O2 and 5% CO2) Ames' solution, supplemented with 6 mM D-glucose, and buffered with 22 mM NaHCO3 to maintain a pH of 7.4. The data analyzed in this work were recorded from three pieces (one mid-periphery, two periphery), using either a conventional 60-channel MEA (10 µm electrode diameter, 100 µm electrode distance; one recording) or a 60-channel perforated MEA (30 µm electrode diameter, 100 µm electrode distance; two recordings), containing holes between the electrodes in order to keep the retina piece in place by slight suction from below. On the conventional MEA, the piece was held in place by coating the MEA with poly-D-lysine (Millipore, 1 mg/mL), applied to the array for one to two hours and then rinsed off prior to mounting.

During recording, the retina pieces were perfused with oxygenated Ames' solution and kept at a constant temperature of around 32–34°C, using an in-line heater and, in the case of the conventional MEA, additionally a heating plate below the array. Bandpass-filtered (300 Hz to 5 kHz) voltage signals from each electrode were sampled at 25 kHz. Spikes were extracted using Kilosort (Pachitariu et al. 2016) adjusted to accommodate for MEA data (https://github.com/dimokaramanlis/KiloSortMEA). The automated extraction was followed by visual inspection and curation using Phy2 (https://github.com/cortex-lab/phy). Clusters that displayed good separation from other clusters as well as a uniform spike waveform and a clear refractory period were considered as representing individual ganglion cells.

### *Visual stimulation*

Spiking activity was recorded in response to visual stimulation. Stimuli were projected onto the photoreceptor side of the retina using a gamma-corrected white-light OLED projector (eMagin, USA) at 800 × 600 square pixels with a refresh rate of 60 Hz (salamander and two marmoset experiments) or 85 Hz (one marmoset experiment). The pixel size (width and height) translated to 2.5 µm or 7.5 µm on

the retina, depending on the configuration. The projector was controlled by a custom-built software based on C++ and OpenGL to generate the stimulus frames. The spatiotemporal white-noise stimulus was presented with a pixel size of 15 µm on the retina (30 µm for salamander) and temporally updated at 30 Hz (salamander), 20 Hz, or 21.25 Hz (marmoset). The light intensity of each stimulus pixel was drawn independently from a binary distribution with 100% contrast and a mean light level of about 2.5 mW/m$^2$ (salamander), 3.33 mW/m$^2$, or 4.89 mW/m$^2$ (marmoset). For analysis, the stimulus was described by its contrast values of +1 (bright) or –1 (dark) for each stimulus pixel. The white-noise stimulus was presented in (non-repeating) sections of 50–180 seconds (depending on the experiment), which were periodically interleaved with a presentation of the same, fixed white-noise sequence (10–31 seconds), intended as held-out data for model validation. The data from the fixed white-noise sequence were not considered in this work. The recording durations of the marmoset experiments ranged between 43 minutes and three and a half hours of non-repeated stimuli (3:28, 0:43, and 1:42 hours).

*Receptive field analysis*

The receptive field of a given ganglion cell was estimated using reverse correlation to obtain the spike-triggered average (STA; Chichilnisky 2001). The stimulus frames within the 700 ms time window preceding a spike were collected and averaged across all spikes. The STA summarizes the spatiotemporal dynamics of the receptive field. To describe the temporal filter of the STA, we selected the pixel that exhibits the maximum absolute value across all STA frames. We defined the temporal filter as the value of that pixel across time, normalized to unit Euclidean norm. The spatial profile was defined as the stimulus frame of the STA that contained the highest absolute pixel value.

The receptive field diameter was estimated by fitting a two-dimensional Gaussian function to the spatial profile of the STA. The effective receptive field diameter $d = \sqrt{ab}$ was obtained from the major axis $a$ and minor axis $b$ of the ellipse at two standard deviations.

For inspection of subunit shape and tiling, contour outlines of the receptive field and subunits were produced. First, the image of each spatial filter was up-scaled by a factor of eight with nearest neighbor sampling and then smoothed with a circular Gaussian filter with a standard deviation appropriate to the spatial size (1.2 original stimulus pixels for receptive fields and 0.5 stimulus pixels for subunits). The contour was determined with the function “find_contours” of the Python module “skimage.measure” at 47.5% of the maximum value of the smoothed filter, equivalent to the 1.22-sigma contour of a circular Gaussian fit. To counter possible expansion effects by the Gaussian smoothing, two-dimensional Gaussians were fit to the smooth and non-smooth filter image. The ratio in their effective diameters was then used to correct and rescale the contour by scaling the contour-level sigma. Contour islands and holes were excluded by selecting the positive closed contour of largest area. Contours were used for visualization of the subunit layouts and mosaics, as well as for determining the relative overlap of two subunits.

*Functional cell-type classification*

Ganglion cells from marmoset recordings were classified based on their temporal filters and effective receptive field diameters. For each cell, the filter and diameter values were concatenated into a vector and then standardized to zero mean and unit variance across all cells. The vectors’ first few principal components that explained 0.9 of the variance were clustered using K-means clustering. The cells grouped by the cluster labels were checked by visual inspection, and occasional non-matching cells that violated receptive field tiling and temporal filter shape were removed from the cell cluster. The clusters were manually labeled as ON and OFF parasol and ON and OFF midget cells based on the sign and shape of the temporal filter and the diameter of the receptive field. Cells in an additional fifth cluster were considered unclassified and excluded from the analysis (excluded cells for each of the three recordings: 12 out of 109 cells total, 32 out of 127, and 39 out of 103).

*Spike-triggered stimulus ensemble*

The spatial region of the stimulus relevant to a given ganglion cell's response was defined as the smallest rectangular window, $x \times y$, that encloses the three-sigma outline of the Gaussian fit of the receptive field (see above). The set of stimulus sequences that lie within this spatial window and within the 700 ms time window preceding a spike compose the spike-triggered stimulus ensemble. Frames in each spike-triggered stimulus were weighted by the temporal filter of the STA and then summed to one effective stimulus frame per spike. For multiple spikes occurring in one time bin, the identical effective stimulus was repeated accordingly. The effective stimuli of all spikes compose the effective spike-triggered stimulus ensemble (STE). The stimulus frames of the STE with $n = x \times y$ pixels were flattened into column vectors to form the STE matrix with dimensions $n \times m$, corresponding to $n$ pixels by $m$ spikes.

*Sparse semi-non-negative matrix factorization*

Semi-non-negative matrix factorization (semi-NMF) decomposes a matrix $\mathbf{V} \in \mathbb{R}^{n \times m}$ into two factor matrices $\mathbf{W} \in \mathbb{R}^{n \times r}$ and $\mathbf{H} \in \mathbb{R}^{r \times m}$ with $r < m, n$, such that

$$\mathbf{V} \approx \mathbf{WH}.$$

These variable names are common to NMF terminology and are related to STNMF in the following way: $\mathbf{V}$ is the STE, where each column $v_i \in \mathbb{R}^n$ is the effective stimulus frame of spike $i \in 1, \dots, m$. The columns of $\mathbf{W}$, $w_k \in \mathbb{R}^n$, are the $k \in 1, \dots, r$ spatial modules, and each column of $\mathbf{H}$, $h_i \in \mathbb{R}^r$, contains the corresponding $r$ weights for each spike $i \in 1, \dots, m$. The stimulus $v_i$ that corresponds to spike $i$ is approximated by the linear combination of the modules $\mathbf{W}$, as basis vectors, and the corresponding weights, $h_i$, as coefficients or encodings. The product of $\mathbf{W}$ and $\mathbf{H}$ forms a lower-dimensional representation of $\mathbf{V}$ based on recurring structures within $\mathbf{V}$. This gives rise to the basis spatial modules $\mathbf{W}$, among which we find the resulting subunits.

Semi-NMF is an NMF variant that relaxes the non-negativity constraint on $\mathbf{V}$ and $\mathbf{H}$, while $\mathbf{W}$ remains non-negative. To retain parts-based factorizations in semi-NMF, the loosened non-negativity constraint has to be replaced (Ding et al. 2010). Typically this is achieved by sparsity regularization on $\mathbf{W}$ (Hoyer 2002, 2004; Kim and Park 2007, 2008). Here, this resulted in the use of sparse semi-NMF.

NMF implementations generally improve the approximation $\mathbf{WH}$ iteratively by minimizing an objective function. As is common for NMF, we chose the Euclidean distance between $\mathbf{V}$ and the reconstruction $\mathbf{WH}$, because it is simple to differentiate (Lee and Seung 1999). Sparsity was implemented with an $\ell 1$-norm penalty on the modules as an additional term in the objective function, weighted by a regularization parameter λ

$$\frac{1}{2}\|\mathbf{V} - \mathbf{WH}\|_F^2 + \lambda \sum_{j=1}^{n} \sum_{k=1}^{r} \left\|w_{jk}\right\|_1, \text{ subject to } \mathbf{W} \geq 0,$$

where $w_{jk}$ for $j = 1, \dots, n$ and $k = 1, \dots, r$ are the elements of $\mathbf{W}$, $\|\cdot\|_F^2$ is the squared Frobenius norm and $\|\cdot\|_1$ is the $\ell 1$-norm.

To distinguish the localized subunits from noisy modules in $\mathbf{W}$, the spatial autocorrelation was calculated using Moran's I (Li et al. 2007; Moran 1950). Given module $w_k$ for $k = 1, \dots, r$ as the vector $s \in \mathbb{R}^n$, the autocorrelation was computed with

$$I = \frac{n}{\sum_a \sum_b l_{ab}} \frac{\sum_a \sum_b l_{ab}(s_a - \bar{s})(s_b - \bar{s})}{\sum_a (s_b - \bar{s})^2},$$

where $a, b = 1, \dots, n$ are pairs of pixel indices, $\bar{s}$ is the mean of vector $s$, and $\mathbf{L} \in \mathbb{R}^{n \times n}$ is the neighbor matrix with $l_{ab}$ equal to unity if $a$ and $b$ correspond to spatially adjacent pixels in the layout $n = x \times y$ and zero otherwise. The scalar $I$ ranges from negative to positive unity with higher positive values denoting denser localization in space. We chose an autocorrelation threshold of $I = 0.25$ to distinguish

localized subunits from noise modules. Additionally, one can consult the mean spike weight of a given module $k$ by averaging the $k$th row of $\mathbf{H}$ to exclude candidate subunits from among the modules when their contribution to the spike generation is small, e.g., smaller than that of non-localized modules. We encountered such low-weight localized modules only rarely and excluded them from the final set of displayed subunits in a few cases based on this criterion.

*Accelerated fast hierarchical alternating least squares*

The objective function was minimized by alternating optimization of the two factor matrices. Given a prior initialization of $\mathbf{W}$, $\mathbf{H}$ was updated while $\mathbf{W}$ was held fixed, and $\mathbf{W}$ was subsequently updated while the newly computed $\mathbf{H}$ was held fixed. This process is generally repeated until a termination criterion is reached. We here chose a fixed number of 1000 iterations (5000 for selected checks) as most cells showed convergence much earlier. The iterations concluded with a final update of $\mathbf{H}$ to provide the appropriate weights for the final set of modules. As the update of $\mathbf{H}$ is unconstrained, unlike the update of $\mathbf{W}$, this also helps in obtaining an accurate assessment of the achieved reconstructions error.

The alternating updates of $\mathbf{W}$ and $\mathbf{H}$ split the optimization into two sub-problems. In the context of semi-NMF, $\mathbf{H}$ is not constrained to be non-negative. Given the Euclidean distance objective function, the update of $\mathbf{H}$ reduces to a least squares problem (Ding et al. 2010) and can be solved accurately using the Moore-Penrose inverse or pseudoinverse $(\cdot)^{\dagger}$

$$\mathbf{H} \leftarrow \mathbf{W}^{\dagger}\mathbf{V}.$$

Subsequently, we applied an $\ell 2$-normalization of the rows of $\mathbf{H}$. This normalization step is generally performed in Euclidean-distance-based NMF to counteract non-uniqueness of solutions (since a scaling of a row in $\mathbf{H}$ can be compensated by an opposite scaling of the corresponding column in $\mathbf{W}$ without affecting the reconstruction) and to prevent exploding coefficients during the alternating updates of $\mathbf{W}$ and $\mathbf{H}$ (Ho 2008). The update of the non-negative matrix $\mathbf{W}$, on the other hand, poses a non-negative least squares (NNLS) problem, in fact a regularized NNLS in the case of sparse semi-NMF. HALS algorithms (Cichocki et al. 2007) approximate it by updating the modules sequentially. We incorporated our sparsity constraint into Fast HALS (Cichocki and Phan 2009) with the local update for $w_k$ for $k \in 1, \dots, r$

$$w_k \leftarrow [w_k + (\mathbf{V}\mathbf{H}^{\top})_k - \mathbf{W}(\mathbf{H}\mathbf{H}^{\top})_k - \lambda \mathbf{1}_n]_+,$$

where $w_k$ and $(\cdot)_k$ are the $k$th column of $\mathbf{W}$ and of the result of the matrix multiplication in parentheses, respectively, $\mathbf{1}_n \in \mathbb{R}^n$ is a column vector of length $n$ containing all ones, and $[\cdot]_+ = \max\{\cdot, 0\}$ is the elementwise maximum with zero of a vector. The terms $\mathbf{V}\mathbf{H}^{\top}$ and $\mathbf{H}\mathbf{H}^{\top}$ remain fixed, whereas $\mathbf{W}$ is updated in-between the column updates. This update rule is based on Fast HALS with additional sparsity regularization. Fast HALS simplifies the update equation of HALS and results in a significant speed up (Cichocki and Phan 2009). It allows terms to be dropped from the HALS equation facilitated by the rows of $\mathbf{H}$ being $\ell 2$-normalized prior to the update. Fast HALS takes advantage of this normalization step.

While Fast HALS improves the local update rule on the scale of one module, A-HALS (Gillis and Glineur 2012) optimizes the computational cost on the scale of all modules in $\mathbf{W}$. While $\mathbf{W}$ is updated continuously, A-HALS exploits the fact that the terms $\mathbf{V}\mathbf{H}^{\top}$ and $\mathbf{H}\mathbf{H}^{\top}$ in the local update remain unchanged over the course of iterating over the modules. These matrix multiplications outweigh the other terms of the local update rule in the number of floating-point operations, making the update of $\mathbf{W}$ costly compared to the outcome. A-HALS wraps an outer loop around the local updates of each column of $\mathbf{W}$, effectively updating $\mathbf{W}$ multiple times per one NMF iteration, in order to tradeoff the floating-point operations in the pre-computed matrix multiplications and in the local updates. This increases the approximation at comparably little additional cost and improves the accuracy of a single update of $\mathbf{W}$.

To determine the number of outer loop cycles, we found the previously suggested parameters to be suitable (Gillis and Glineur 2012). The maximum number of outer loop cycles is determined by $\alpha \times \rho_{\boldsymbol{W}}$, where $\rho_{\boldsymbol{W}} = 1 + \frac{mn+nr}{mr+m}$ is the number of floating-point operations in the update and $\alpha = 0.5$ is a scaling factor. An additional dynamic stopping criterion checks whether the last improvement of $\mathbf{W}$ is smaller than 10% of the first improvement, that is, in the first cycle of the outer loop. The improvement of $\mathbf{W}$ is measured as the Frobenius norm of the difference to the previous cycle of the outer loop, $\left\|\mathbf{W} - \mathbf{W}^{(previous)}\right\|_F$. Aside from the acceleration by performing multiple update cycles, A-HALS improves numerical stability by ensuring no zero-columns in $\mathbf{W}$ to avoid convergence issues of HALS (Gillis and Glineur 2008; Ho 2008). Columns of all zeros are replaced by columns whose entries all take a small positive constant (here $10^{-16}$). In the context of semi-NMF, we observed this to benefit the convergence of an SVD-based pseudoinverse $\mathbf{W}^{\dagger}$ for the subsequent update of $\mathbf{H}$. We combined Fast HALS and A-HALS by replacing the local update in A-HALS with the equations from Fast HALS and complementing this with sparsity regularization. All analyses were carried out on computers with an Intel Xeon CPU E3-1270 v5 @ 3.60 GHz.

*Non-negative singular value decomposition with low-rank correction*

The aforementioned alternating optimization starts with an update of the weights $\mathbf{H}$ inferred from the STE $\mathbf{V}$ and an initialization of $\mathbf{W}$. We initialized the modules $\mathbf{W}$ using a slightly modified non-negative singular value decomposition with low-rank correction (NNSVD-LRC), which was introduced for NMF (Atif et al. 2019). Here, we describe a modified version for semi-NMF with slight optimizations. Given $r$ modules, we extracted half that number of components from a truncated SVD of $\mathbf{V}$. Specifically, we used $p = \left\lceil \frac{r}{2} \right\rceil$ components from the decomposition $\mathbf{V} \approx \mathbf{U\Sigma A}$ with the matrices $\mathbf{U} \in \mathbb{R}^{n \times p}$ and $\mathbf{A} \in \mathbb{R}^{p \times m}$ and the diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$. To obtain a two-factor form as in NMF, the singular values were integrated into the left and right singular vectors with their element-wise square root to provide $\mathbf{U} \leftarrow \mathbf{U}\sqrt{\mathbf{\Sigma}}$ and $\mathbf{A} \leftarrow \sqrt{\mathbf{\Sigma}}\mathbf{A}$, respectively, so that $\mathbf{V} \approx \mathbf{UA}$ with the modified $\mathbf{U}$ and $\mathbf{A}$. Unlike the original NNSVD-LRC, we only initialized $\mathbf{W}$ (via $\mathbf{U}$) and thus discarded $\mathbf{A}$. The components of $\mathbf{U}$ sorted by decreasing singular values served as the columns of $\mathbf{W}$. In doing so, each component was used twice, with and without a sign-flip, ordered so that the one with the largest positive element came before its sign-inverted counterpart to ensure reproducible component ordering. (In case of an odd $r$, the component with the lowest singular value was only added once.) In the end, all negative entries were set to zero to make the initialization of $\mathbf{W}$ non-negative. This resulted in the initialization $\mathbf{W} = \left[u_1, -u_1, u_2, -u_2, \dots, u_p, -u_p\right]_+$. Inspired by A-HALS, we implemented an additional step to avoid numerical instability (Gillis and Glineur 2012). Any columns of $\mathbf{W}$ that contained only zeros were set to a small positive value ($10^{-16}$). Any zeros residing in nonzero columns remained to accelerate matrix multiplications and to allow sparsity to emerge.

The solution can be improved further at low computational cost by a few NMF iterations on the low-rank approximation $\mathbf{V}^{(p)} \approx \mathbf{UA}$. As $\mathbf{V}^{(p)}$ has a factorization with low-rank matrices, the matrix multiplications for the NMF iterations become less complex, and NMF iterations reduce from a complexity of $\mathcal{O}(nmr)$ to $\mathcal{O}((n+m)r^2)$ (Atif et al. 2019; Gillis and Glineur 2012). In the original NNSVD-LRC, the low-rank NMF was performed using A-HALS until the difference in relative reconstruction error $\left\|\mathbf{V}^{(p)} - \mathbf{WH}\right\|_F / \left\|\mathbf{V}^{(p)}\right\|_F$ to the previous iteration fell below 5% of the initial error, which typically occurred within ten iterations (Atif et al. 2019). Here, since we used sparse semi-NMF instead, we observed that the proposed termination criterion was always reached within one AF-HALS update in the case of STNMF. We thus fixed the number of iterations with $\mathbf{V}^{(p)}$ at one and omitted the termination criterion. Without the need for computing reconstruction errors, it suffices to initialize $\mathbf{W}$ from the SVD components, leaving $\mathbf{H}$ to be subsequently computed from the initialization of $\mathbf{W}$. This further reduces computational cost and provides a fast initialization procedure.

For comparison and for consensus methods, we applied random initialization of **W** by sampling its values independently from a uniform distribution between zero and unity. We used the Mersenne Twister (MT19937) pseudorandom number generator (Matsumoto and Nishimura 1998) for potential compatibility across Python, R, and MATLAB. For reproducibility, we applied a specified set of seeds, here ranging from 0 to 49 for the 50 repetitions of the consensus analysis.

*Consensus analysis*

We used consensus analyses (Brunet et al. 2004; Monti et al. 2003) to determine the best tradeoff between solution stability and sparsity. Consensus methods allow judging the robustness of the parameter when there is a lack of measurable error. A typical proxy for goodness of decomposition in the field of NMF is the reconstruction error, or residual, as it is non-increasing over the NMF iterations (Cichocki et al. 2007). However, with additional constraints like sparsity regularization, the residual does not correspond to the used objective function. Furthermore, in the context of STNMF, the noisy excess modules may influence the residual, but neglecting them instead, favors solutions that exhibit more localized modules over those with fewer. Consequently, decompositions based on different parameters cannot be compared easily by their residual. This is one of the reasons, why cross-validation is not suitable for STNMF.

For applying consensus methods, we performed 30 randomly initialized repetitions of sparse semi-NMF (50 repetitions for the detailed comparative analysis) for a range of parameters. The parameters correspond to different sparsity regularization strengths. Semi-NMF was run for 1000 iterations, which we found to typically be enough to reach convergence. Similarity between solutions was assessed by treating the encodings in **H** as cluster labels and comparing them across solutions. We define cluster labels as the index of the module with the highest absolute weight for a given spike. Although a spike may be elicited by a combination of subunit activations, considering the maximum associated module suffices for sake of comparison. Because the ordering of the emerging modules is arbitrary, the cluster labels cannot be compared across solutions directly. To remove this ambiguity, pairwise comparisons of cluster labels among the spikes within each solution are computed in a connectivity matrix **G**. Each entry $g_{ij}$ in the square matrix is unity, if spikes *i* and *j* share the cluster label and zero otherwise. As the number of unique pairs grows according to $m(m-1)/2$ for increasing number of spikes *m*, we limited ourselves to 25000 spikes selected at random with a persistent selection across the analysis. That amounts to more than one gigabyte of memory at 32-bit floating-point precision for one decomposition. The number may be increased with more memory available or with memory-efficient implementations. Nevertheless, we found that this subsample provides a sufficient representation of the resulting cluster memberships. The elementwise average of the connectivity matrices of all 30 (or 50) solutions expresses the consensus across solutions. This consensus matrix contains values between zero and unity, describing the fraction of agreement between the encodings across the 30 (or 50) runs. We calculated the cophenetic correlation using the function “linkage” to perform hierarchical clustering and “cophenet” to calculate the cophenetic distances within the clustering from the Python module “scipy.cluster.hierarchy”. The coefficient is a real value between zero and unity, with unity denoting perfect consensus and identical solutions (Brunet et al. 2004).

STNMF reserves excess modules to capture noise in the data. These modules are not considered localized subunits and are not representative of the solution. As uncorrelated noise is distributed arbitrarily across the excess modules, they are likely to vary across solutions. Thereby they skew the described metric. Consequently, we only considered modules containing localized subunits as valid cluster labels. Connectivity pairs involving non-localized modules were set to zero in the connectivity matrices. These entries do not reach consensus across solutions. This causes a lower dispersion in the consensus matrix, quenching the cophenetic correlation coefficients. The metric becomes more sensitive to differences in detected subunits. In the case of no localized subunits in the decomposition, we did not

compute the cophenetic correlation. This results in gaps in the stability curve (see **Figure 7A** for an example).

The cophenetic correlation coefficient corresponds to the similarity of the recovered subunits across the repeated decompositions. It represents the stability of the solutions, and we used it to compare the different sparsity regularization parameters. The optimal trade-off between high stability and little regularization was found at the bend of the curve of regularization strength versus stability (**Figure 4C**). Here, we determined the inflection point by visual inspection and defined it as the most suitable sparsity parameter. To identify it systematically instead, we propose that one could describe the curve as a piecewise linear function with two or three components and select the intersection of the functions.

*Subunit comparison*

To compare recovered subunits on a geometrical basis, we computed their relative spatial overlap. Given the contour outlines (described above) of two subunits *i* and *j*, we calculated their relative overlap $o_{ij}$ with the Jaccard index (Jaccard 1912), which is defined as the ratio of the intersection by the union of their areas $a_i$ and $a_j$:

$$o_{ij} = \frac{a_i \cap a_j}{a_i \cup a_j}$$

We considered subunit pairs with $o_{ij} > 0.5$, that is, more than 50% area overlap, to be strongly overlapping and thus candidates of identical subunits.

For the purpose of determining prevalent subunits across multiple randomly initialized runs of STNMF (**Figure 3**), the relative overlap was compared across decompositions. A subunit that showed overlap with $o_{ij} > 0.5$ across more than 50% of the runs (colored outlines in **Figure 3**) was considered a robustly recovered subunit.

*Inter-mosaic coordination*

The spatial coordination of ON- and OFF-type subunit mosaics was determined using the inter-mosaic coordination energy (IMCE) as proposed previously (Roy et al. 2021). In brief, contour mosaics were reduced to subunit centroids by their center of mass. Within a region of interest, one of the mosaics (center of mass points) was shifted in horizontal and vertical position relative to the other. The squared inverse distance between each heterotypic pair of centroids $r_{ij}$ was measured with

$$e_{ij} = \frac{1}{\max\{r_{ij}, r_{min}\}^2}$$

where $r_{min}$ is a minimum distance set to 0.2 times the median heterotypic nearest-neighbor distance. The values of $e_{ij}$ were averaged across all pairs to obtain the IMCE. The IMCE was computed for different offsets of the OFF-type mosaic relative to the ON-type mosaic on a 50-by-50 Cartesian grid before being z-scored. The shift distances were normalized to the median homotypic nearest neighbor distance of the ON-type (frozen) mosaic. The generated topographic map of IMCE values was radially averaged to examine whether the IMCE increased or decreased with radial shift distance. We selected the analyzed region of interest in two ways. First, we manually chose a region where the density of subunits was approximately homogeneous on both mosaics to minimize effects of gaps. We then confirmed the determined IMCE by following the proposed implementation of finding a suitable convex hull that maximized the number of enclosed heterotypic pairs while excluding outliers (Roy et al. 2021).

## REFERENCES


Ahrens, M. B., Linden, J. F., and Sahani, M. (2008). Nonlinearities and Contextual Influences in Auditory Cortical Responses Modeled with Multilinear Spectrotemporal Methods. *Journal of Neuroscience*, *28*(8), 1929–1942.

Almasi, A., Meffin, H., Cloherty, S. L., Wong, Y., Yunzab, M., and Ibbotson, M. R. (2020). Mechanisms of Feature Selectivity and Invariance in Primary Visual Cortex. *Cerebral Cortex*, *30*(9), 5067–5087.

Anishchenko, A., Greschner, M., Elstrott, J., Sher, A., Litke, A. M., Feller, M. B., and Chichilnisky, E. J. (2010). Receptive field mosaics of retinal ganglion cells are established without visual experience. *Journal of Neurophysiology*, *103*(4), 1856–1864.

Appleby, T. R., and Manookin, M. B. (2020). Selectivity to approaching motion in retinal inputs to the dorsal visual pathway. *ELife*, *9*, e51144.

Atif, S. M., Qazi, S., and Gillis, N. (2019). Improved SVD-based initialization for nonnegative matrix factorization using low-rank correction. *Pattern Recognition Letters*, *122*, 53–59.

Bartsch, F., Cumming, B. G., and Butts, D. A. (2022). Model-based characterization of the selectivity of neurons in primary visual cortex. *Journal of Neurophysiology*, *128*(2), 350–363.

Berry, M. J., Warland, D. K., and Meister, M. (1997). The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(10), 5411–5416.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, *52*(1), 155–173.

Beyeler, M., Dutt, N., and Krichmar, J. L. (2016). 3D Visual Response Properties of MSTd Emerge from an Efficient, Sparse Population Code. *Journal of Neuroscience*, *36*(32), 8399–8415.

Borghuis, B. G., Marvin, J. S., Looger, L. L., and Demb, J. B. (2013). Two-Photon Imaging of Nonlinear Glutamate Release Dynamics at Bipolar Cell Synapses in the Mouse Retina. *Journal of Neuroscience*, *33*(27), 10972–10985.

Boutsidis, C., and Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, *41*(4), 1350–1362.

Boycott, B. B., and Wässle, H. (1991). Morphological Classification of Bipolar Cells of the Primate Retina. *European Journal of Neuroscience*, *3*(11), 1069–1088.

Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. L. (2008). Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, *390*(5), 1241–1251.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(12), 4164–4169.

Bryant, H. L., and Segundo, J. P. (1976). Spike initiation by transmembrane current: a white-noise analysis. *The Journal of Physiology*, *260*(2), 279–314.

Calkins, D. J., and Sterling, P. (2007). Microcircuitry for Two Types of Achromatic Ganglion Cell in Primate Fovea. *Journal of Neuroscience*, *27*(10), 2646–2653.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*(2), 489–509.

Cantrell, D. R., Cang, J., Troy, J. B., and Liu, X. (2010). Non-Centered Spike-Triggered Covariance Analysis Reveals Neurotrophin-3 as a Developmental Regulator of Receptive Field Properties of ON-OFF Retinal Ganglion Cells. *PLoS Computational Biology*, *6*(10), e1000967.

Carbonero, D., Noueihed, J., Kramer, M. A., and White, J. A. (2023). Non-Negative Matrix Factorization for Analyzing State Dependent Neuronal Network Dynamics in Calcium Recordings. *BioRxiv*, 561797v3.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*(2), 199–213.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. Ben, and LeCun, Y. (2015). The Loss Surfaces of Multilayer Networks. In G. Lebanon, & S. V. N. Vishwanathan (Eds.), *Proceedings of Machine Learning Research* (Vol. 38, pp. 192–204). PMLR.

Cichocki, A., and Phan, A. H. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E92-A*(3), 708–721.

Cichocki, A., Zdunek, R., and Amari, S. (2007). Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization. In *Independent Component Analysis and Signal Separation* (pp. 169–176). Springer Berlin Heidelberg.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations* (1st ed.). Wiley.

Cohen, E., and Sterling, P. (1990a). Convergence and divergence of cones onto bipolar cells in the central area of cat retina. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *330*(1258), 323–328.

Cohen, E., and Sterling, P. (1990b). Demonstration of cell types among cone bipolar neurons of cat retina. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *330*(1258), 305–321.

Crook, J. D., Packer, O. S., and Dacey, D. M. (2014). A synaptic signature for ON- and OFF-center parasol ganglion cells of the primate retina. *Visual Neuroscience*, *31*(1), 57–84.

Crook, J. D., Peterson, B. B., Packer, O. S., Robinson, F. R., Troy, J. B., and Dacey, D. M. (2008). Y-Cell Receptive Field and Collicular Projection of Parasol Ganglion Cells in Macaque Monkey Retina. *Journal of Neuroscience*, *28*(44), 11277–11291.

Cunningham, J. P., and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, *17*(11), 1500–1509.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. G. (2007). A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, *49*(3), 434–448.

Dacey, D. M. (1993). The mosaic of midget ganglion cells in the human retina. *Journal of Neuroscience*, *13*(12), 5334–5355.

Dacey, D. M., Packer, O. S., Diller, L., Brainard, D., Peterson, B., and Lee, B. (2000). Center surround receptive field structure of cone bipolar cells in primate retina. *Vision Research*, *40*(14), 1801–1811.

De Boer, E., and Kuyper, P. (1968). Triggered Correlation. *IEEE Transactions on Biomedical Engineering*, *15*(3), 169–179.

Demb, J. B., and Singer, J. H. (2015). Functional Circuitry of the Retina. *Annual Review of Vision Science*, *1*(1), 263–289.

Demb, J. B., Zaghloul, K. A., Haarsma, L., and Sterling, P. (2001). Bipolar Cells Contribute to Nonlinear Spatial Summation in the Brisk-Transient (Y) Ganglion Cell in Mammalian Retina. *Journal of Neuroscience*, *21*(19), 7447–7454.

Devarajan, K. (2008). Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. *PLoS Computational Biology*, *4*(7), e1000029.

DeVries, S. H., and Baylor, D. A. (1997). Mosaic arrangement of ganglion cell receptive fields in rabbit retina. *Journal of Neurophysiology*, *78*(4), 2048–2060.

Ding, C., Tao Li, and Jordan, M. I. (2010). Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, *41*(3), 613–627.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306.

Elad, M. (2006). Why Simple Shrinkage Is Still Relevant for Redundant Representations? *IEEE Transactions on Information Theory*, *52*(12), 5559–5569.

Enroth-Cugell, C., and Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of Physiology*, *187*(3), 517–552.

Eriköz, B., Jusuf, P. R., Percival, K. A., and Grünert, U. (2008). Distribution of bipolar input to midget and parasol ganglion cells in marmoset retina. *Visual Neuroscience*, *25*(1), 67–76.

Fairhall, A. L., Burlingame, C. A., Narasimhan, R., Harris, R. A., Puchalla, J. L., and Berry, M. J. (2006). Selectivity for Multiple Stimulus Features in Retinal Ganglion Cells. *Journal of Neurophysiology*, *96*(5), 2724–2738.

Freeman, J., Field, G. D., Li, P. H., Greschner, M., Gunning, D. E., Mathieson, K., Sher, A., Litke, A. M., Paninski, L., Simoncelli, E. P., and Chichilnisky, E. J. (2015). Mapping nonlinear receptive field structure in primate retina at single cone resolution. *ELife*, *4*, e05241.

Fu, W., and Perry, P. O. (2017). Estimating the number of clusters using cross-validation. *Journal of Computational and Graphical Statistics*, *29*(1), 162–173.

Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978–984.

Gao, P., and Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, *32*, 148–155.

Gauthier, J. L., Field, G. D., Sher, A., Greschner, M., Shlens, J., Litke, A. M., and Chichilnisky, E. J. (2009). Receptive Fields in Primate Retina Are Coordinated to Sample Visual Space More Uniformly. *PLoS Biology*, *7*(4), e1000063.

Gillis, N., and Glineur, F. (2008). Nonnegative Factorization and The Maximum Edge Biclique Problem. *ArXiv*, 0810.4225v1.

Gillis, N., and Glineur, F. (2012). Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization. *Neural Computation*, *24*(4), 1085–1105.

Gollisch, T., and Liu, J. K. (2018). *Data: Salamander retinal ganglion cells under finely structured spatio-temporal white noise* [Dataset]. G-Node. https://doi.org/10.12751/g-node.62b65b

Gollisch, T., and Meister, M. (2008). Modeling convergent ON and OFF pathways in the early visual system. *Biological Cybernetics*, *99*(4–5), 263–278.

Gollisch, T., and Meister, M. (2010). Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron*, *65*(2), 150–164.

Hansen, P. C. (1994). Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, *6*, 1–35.

Hansen, P. C. (1999a). Regularization Tools Version 3.0 for Matlab 5.2. *Numerical Algorithms*, *20*, 195–196.

Hansen, P. C. (1999b). *The L-curve and its use in the numerical treatment of inverse problems: Vol. 1999:15*. IMM, Department of Mathematical Modelling, Technical University of Denmark.

Harper, N. S., Schoppe, O., Willmore, B. D. B., Cui, Z., Schnupp, J. W. H., and King, A. J. (2016). Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. *PLOS Computational Biology*, *12*(11), e1005113.

Ho, N.-D. (2008). *Nonnegative matrix factorization algorithms and applications* [Diss. PhD thesis]. Université catholique de Louvain.

Hoyer, P. O. (2002). Non-negative sparse coding. *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 557–565.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, *5*, 1457–1469.

Huang, Z., Ran, Y., Oesterle, J., Euler, T., and Berens, P. (2021). Estimating smooth and sparse neural receptive fields with a flexible spline basis. *Neurons, Behavior, Data Analysis, and Theory*, *5*(3), 1–30.

Hyvärinen, A. (1999). Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation. *Neural Computation*, *11*(7), 1739–1768.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytologist*, *11*(2), 37–50.

Jacoby, R. A., Wiechmann, A. F., Amara, S. G., Leighton, B. H., and Marshak, D. W. (2000). Diffuse bipolar cells provide input to OFF parasol ganglion cells in the macaque retina. *The Journal of Comparative Neurology*, *416*(1), 6–18.

Jacoby, R., Stafford, D., Kouyama, N., and Marshak, D. (1996). Synaptic inputs to ON parasol ganglion cells in the primate retina. *Journal of Neuroscience*, *16*(24), 8041–8056.

Jarsky, T., Cembrowski, M., Logan, S. M., Kath, W. L., Riecke, H., Demb, J. B., and Singer, J. H. (2011). A synaptic mechanism for retinal adaptation to luminance and contrast. *Journal of Neuroscience*, *31*(30), 11003–11015.

Jun, N. Y., Field, G. D., and Pearson, J. (2021). Scene statistics and noise determine the relative arrangement of receptive field mosaics. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(39), e2105115118.

Jusuf, P. R., Martin, P. R., and Grünert, U. (2006a). Synaptic connectivity in the midget-parvocellular pathway of primate central retina. *The Journal of Comparative Neurology*, *494*(2), 260–274.

Jusuf, P. R., Martin, P. R., and Grünert, U. (2006b). Random wiring in the midget pathway of primate retina. *Journal of Neuroscience*, *26*(15), 3908–3917.

Kaardal, J., Fitzgerald, J. D., Berry, M. J., and Sharpee, T. O. (2013). Identifying Functional Bases for Multidimensional Neural Computations. *Neural Computation*, *25*(7), 1870–1890.

Karamanlis, D., Khani, M. H., Schreyer, H. M., Zapp, S. J., Mietsch, M., and Gollisch, T. (2023). Natural stimuli drive concerted nonlinear responses in populations of retinal ganglion cells. *BioRxiv*, 523412v1.

Kerschensteiner, D. (2022). Feature Detection by Retinal Ganglion Cells. *Annual Review of Vision Science*, *8*(1), 135–169.

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2020). Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *ELife*, *9*, e53445.

Kim, H., and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, *23*(12), 1495–1502.

Kim, J., and Park, H. (2008). Sparse Nonnegative Matrix Factorization for Clustering. *Georgia Institute of Technology, Tech. Rep. GT-CSE-08-01*.

Kim, J., and Park, H. (2011). Fast Nonnegative Matrix Factorization: An Active-Set-Like Method and Comparisons. *SIAM Journal on Scientific Computing*, *33*(6), 3261–3281.

Koh, T. H., Bishop, W. E., Kawashima, T., Jeon, B. B., Srinivasan, R., Mu, Y., Wei, Z., Kuhlman, S. J., Ahrens, M. B., Chase, S. M., and Yu, B. M. (2022). Dimensionality reduction of calcium-imaged neuronal population activity. *Nature Computational Science*, *3*(1), 71–85.

Kolb, H., and Dekorver, L. (1991). Midget ganglion cells of the parafovea of the human retina: A study by electron microscopy and serial section reconstructions. *The Journal of Comparative Neurology*, *303*(4), 617–636.

Lawson, C. L., and Hanson, R. J. (1974). *Solving Least Squares Problems*. Prentice-Hall.

Lee, B. B., Martin, P. R., and Grünert, U. (2010). Retinal connectivity and primate vision. *Progress in Retinal and Eye Research*, *29*(6), 622–639.

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lehky, S., Sejnowski, T., and Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *The Journal of Neuroscience*, *12*(9), 3568–3581.

Li, H., Calder, C. A., and Cressie, N. (2007). Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. *Geographical Analysis*, *39*(4), 357–375.

Li, L., and Zhang, Y.-J. (2009). FastNMF: highly efficient monotonic fixed-point nonnegative matrix factorization algorithm with good applicability. *Journal of Electronic Imaging*, *18*(3), 033004.

Liu, J. K., Schreyer, H. M., Onken, A., Rozenblit, F., Khani, M. H., Krishnamoorthy, V., Panzeri, S., and Gollisch, T. (2017). Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature Communications*, *8*(1), 149.

Maheswaranathan, N., Kastner, D. B., Baccus, S. A., and Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. *PLOS Computational Biology*, *14*(8), e1006291.

Maheswaranathan, N., McIntosh, L. T., Tanaka, H., Grant, S., Kastner, D. B., Melander, J. B., Nayebi, A., Brezovec, L. E., Wang, J. H., Ganguli, S., and Baccus, S. A. (2023). Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, *111*(17), 2742–2755.

Manookin, M. B., and Demb, J. B. (2006). Presynaptic Mechanism for Slow Contrast Adaptation in Mammalian Retinal Ganglion Cells. *Neuron*, *50*(3), 453–464.

Manookin, M. B., Patterson, S. S., and Linehan, C. M. (2018). Neural Mechanisms Mediating Motion Sensitivity in Parasol Ganglion Cells of the Primate Retina. *Neuron*, *97*(6), 1327–1340.

Marc, R. E., and Sperling, H. G. (1977). Chromatic organization of primate cones. *Science*, *4288*(196), 454–456.

Masland, R. H. (2001). The fundamental plan of the retina. *Nature Neuroscience*, *4*(9), 877–886.

Matsumoto, M., and Nishimura, T. (1998). Mersenne twister. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30.

McFarland, J. M., Cui, Y., and Butts, D. A. (2013). Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. *PLoS Computational Biology*, *9*(7), e1003143.

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. A. (2016). Deep Learning Models of the Retinal Response to Natural Scenes. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc.

Mineault, P. J., Khawaja, F. A., Butts, D. A., and Pack, C. C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), E972–E980.

Monti, S., Tamayo, P., Mesirov, J. P., and Golub, T. R. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1), 91–118.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, *37*(1–2), 17–23.

Nagayama, M., Aritake, T., Hino, H., Kanda, T., Miyazaki, T., Yanagisawa, M., Akaho, S., and Murata, N. (2022). Detecting cell assemblies by NMF-based clustering from calcium imaging data. *Neural Networks*, *149*, 29–39.

Onken, A., Liu, J. K., Karunasekara, P. P. C. R., Delis, I., Gollisch, T., and Panzeri, S. (2016). Using Matrix and Tensor Factorizations for the Single-Trial Analysis of Population Spike Trains. *PLOS Computational Biology*, *12*(11), e1005189.

Owen, A. B., and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Annals of Applied Statistics*, *3*(2), 564–594.

Pachitariu, M., Steinmetz, N. A., Kadir, S., Carandini, M., and Harris, K. D. (2016). Fast and accurate spike sorting of high-channel count probes with KiloSort. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc.

Qiao, H. (2015). New SVD based initialization strategy for non-negative matrix factorization. *Pattern Recognition Letters*, *63*, 71–77.

Real, E., Asari, H., Gollisch, T., and Meister, M. (2017). Neural Circuit Inference from Function to Structure. *Current Biology*, *27*(2), 189–198.

Rieke, F. (2001). Temporal Contrast Adaptation in Salamander Bipolar Cells. *Journal of Neuroscience*, *21*(23), 9445–9454.

Roy, S., Jun, N. Y., Davis, E. L., Pearson, J., and Field, G. D. (2021). Inter-mosaic coordination of retinal receptive fields. *Nature*, *592*(7854), 409–413.

Samengo, I., and Gollisch, T. (2013). Spike-triggered covariance: geometric proof, symmetry properties, and extension beyond Gaussian stimuli. *Journal of Computational Neuroscience*, *34*(1), 137–161.

Saxena, S., Kinsella, I., Musall, S., Kim, S. H., Meszaros, J., Thibodeaux, D. N., Kim, C., Cunningham, J., Hillman, E. M. C., Churchland, A., and Paninski, L. (2020). Localized semi-nonnegative matrix factorization (LocaNMF) of widefield calcium imaging data. *PLOS Computational Biology*, *16*(4), e1007791.

Schwartz, G. W., Okawa, H., Dunn, F. A., Morgan, J. L., Kerschensteiner, D., Wong, R. O., and Rieke, F. (2012). The spatial structure of a nonlinear receptive field. *Nature Neuroscience*, *15*(11), 1572–1580.

Schwartz, O., Pillow, J. W., Rust, N. C., and Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*(4), 484–507.

Shah, N. P., Brackbill, N., Rhoades, C. E., Kling, A., Goetz, G., Litke, A. M., Sher, A., Simoncelli, E. P., and Chichilnisky, E. J. (2020). Inference of nonlinear receptive field subunits with spike-triggered clustering. *ELife*, *9*, e45743.

Sterling, P., Freed, M. A., and Smith, R. G. (1988). Architecture of rod and cone circuits to the on-beta ganglion cell. *Journal of Neuroscience*, *8*(2), 623–642.

Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L. T., Baccus, S. A., and Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Vavasis, S. A. (2010). On the Complexity of Nonnegative Matrix Factorization. *SIAM Journal on Optimization*, *20*(3), 1364–1377.

Victor, J. D., and Shapley, R. M. (1979). The nonlinear pathway of Y ganglion cells in the cat retina. *Journal of General Physiology*, *74*(6), 671–689.

Vintch, B., Movshon, J. A., and Simoncelli, E. P. (2015). A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *Journal of Neuroscience*, *35*(44), 14829–14841.

Wässle, H. (1999). Parallel pathways from the outer to the inner retina in primates. In K. R. Gegenfurtner, & L. T. Sharpe (Eds.), *Color Vision: From Genes to Perception* (pp. 145–162). Cambridge University Press.

Wässle, H. (2004). Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, *5*(10), 747–757.

Watanabe, M., and Rodieck, R. W. (1989). Parasol and midget ganglion cells of the primate retina. *The Journal of Comparative Neurology*, *289*(3), 434–454.

Wild, S., Curry, J., and Dougherty, A. (2004). Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, *37*(11), 2217–2232.

Zapp, S. J., Nitsche, S., and Gollisch, T. (2022). Retinal receptive-field substructure: scaffolding for coding and computation. *Trends in Neurosciences*, *45*(6), 430–445.

Zhang, C., Jing, L., and Xiu, N. (2014). A New Active Set Method For Nonnegative Matrix Factorization. *SIAM Journal on Scientific Computing*, *36*(6), A2633–A2653.

Zhou, T., Kang, J., Cong, F., and Li, D. X. (2020). Early childhood developmental functional connectivity of autistic brains with non-negative matrix factorization. *NeuroImage: Clinical*, *26*, 102251.