Ocular Drift Transforms Retinal Image Statistics and Spatiotemporal Receptive Fields

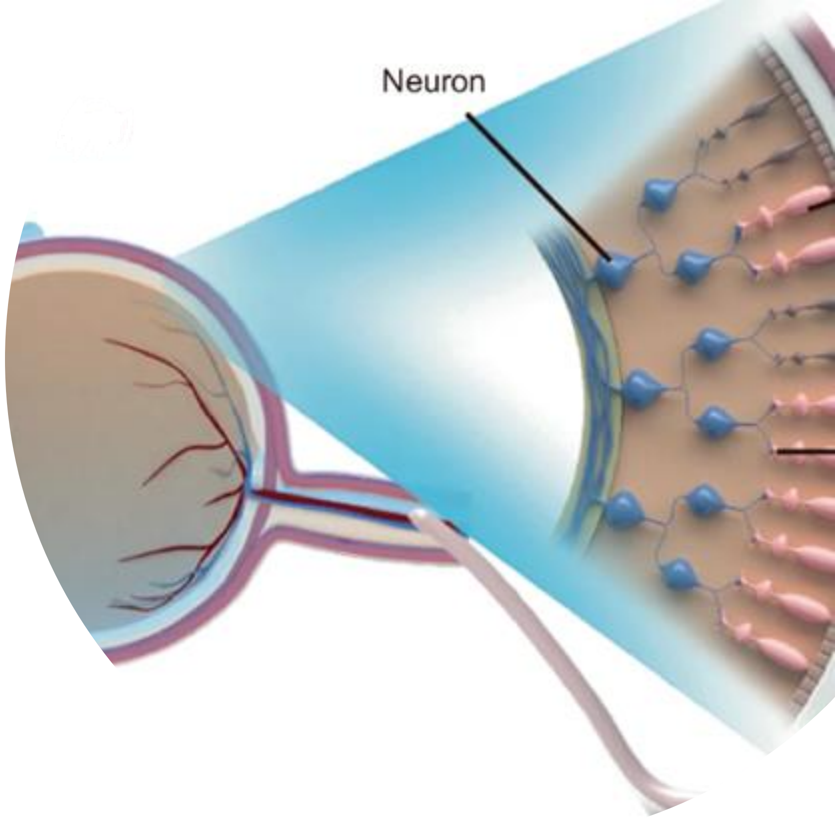
Dennis Perez, Alexander Belsten, and Jacob Yates

Introduction

Constrained visual system & seemingly complex world

Environmental statistics

Efficient Coding



Vision begins at the retina

Operates under constraints

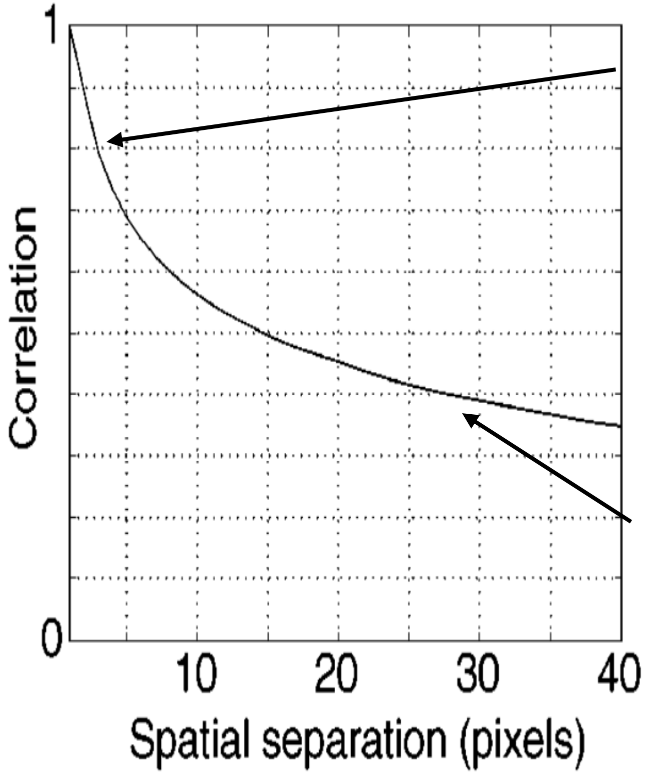
Understanding the statistical structure of the natural world may help explain the design of neural coding strategies in the retina

Evolved in the natural world

Neurons are tuned to the **statistical regularities** of natural scenes

What can we learn from natural scene statistics?

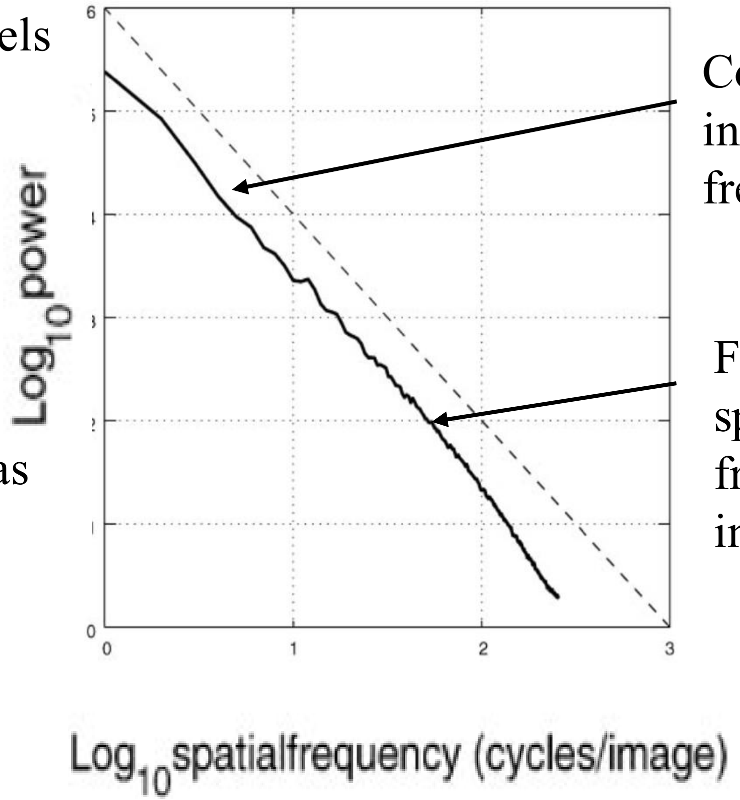
Autocorrelation



Nearby pixels are highly correlated

Decreases as distance increases

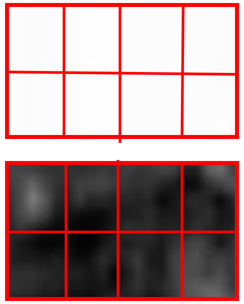
Power Spectrum



Concentrated in lower frequencies

Falls off as spatial frequency increases

$1/f^2$



Predictable!

Not so much

Lower spatial frequencies dominate the signal

Natural scenes are not random - they have a lot of predictable structure

Simoncelli & Olshausen, 2001

Autocorrelation

Neighboring regions of space are highly correlated

Power Spectrum

Large, smooth regions with slow changes in intensity are more common than sharp jumps

The prominent statistical property of retinal input is predictable structure – or **redundancy**.

How is redundancy relevant to the design of our visual system?

Patterns can be leveraged to form an efficient representation

Simoncelli & Olshausen, 2001

Efficient coding hypothesis

- Retinal neurons adapted to the statistics of their input

- Maximize the information between their inputs and outputs, with respect to their constraints

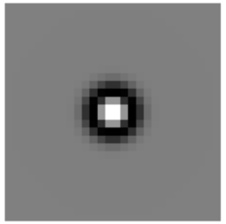
A theoretical framework for retinal design

- This process can be formalized mathematically, and it results in clear predictions
  - Sensory systems will try to reduce redundancy inherent in the signal
  - Transform visual input into a statistically independent basis
  - Reduces signals that transmit same info at the cost of more energy
  - Spreads power such that unlikely, meaningful features, are emphasized

Predicts center-surround receptive fields as a strategy to decorrelate spatial input

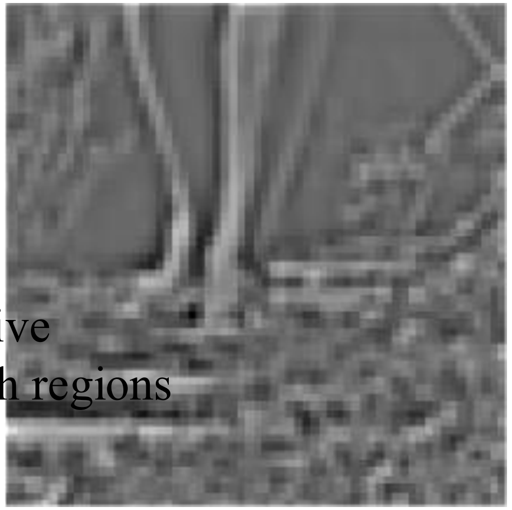
Original Image

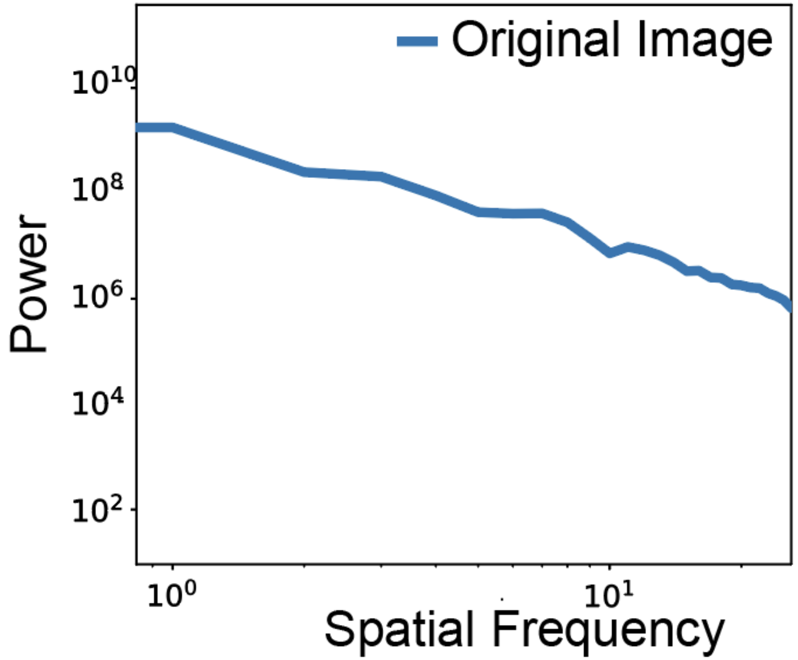
Filter



Acts as a spatial derivative
Attenuating slow/smooth regions
(lower freqs)

Filtered Image





power is dominated by low frequencies

Edges remain
& no freq is over-represented

Signal’s power is flattened

- Redundancy reduction:
  - Remove predictable parts of the signal
  - Remaining signal is more statistically independent across channels
  - Each channel now carries unique information
  - Shows how neurons leverage patterns to form an efficient representation

- Models based on efficient coding predict RFs that match biological recordings
- Highlights convergence zone of theoretical & experimental neuroscience
- Literature showing the connection between properties of natural stimuli and neural processing is extensive
- Reinforces our belief in the connection between the properties of the world and retinal design

Karamanlis, 2022

Puzzle

- Redundancy reduction as a primary encoding strategy suggests retinal input statistics match the environment

- Retinal input is not static & has altered statistics

- Prime example: eye movements
  - Alter retinal image statistics
  - Alternative explanation for the whitening process

Fixational Eye Movements: Ocular Drift

- Your eyes do not sit still!

- Slow, irregular, constant movement

- Occurs between microsaccades

- Displacement < 10th of a degree with speed of ~50 arcmins/s (varies)

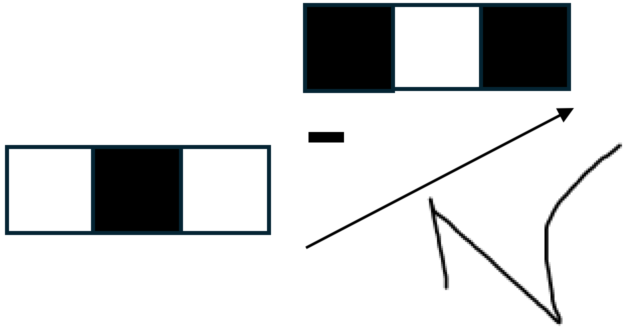
Barlow, 1952; Martinez-Conde et al., 2004; Rolfs, 2009; Fang et al., 2018; Alexander & Martinez-Conde, 2019

Ocular Drift shifts retinal image statistics

- Eyes are always moving

- Signal on the retina is constantly being shifted

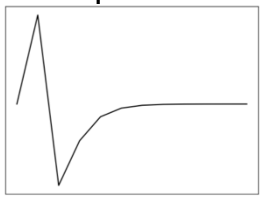
- Retinal input statistics are constantly changing



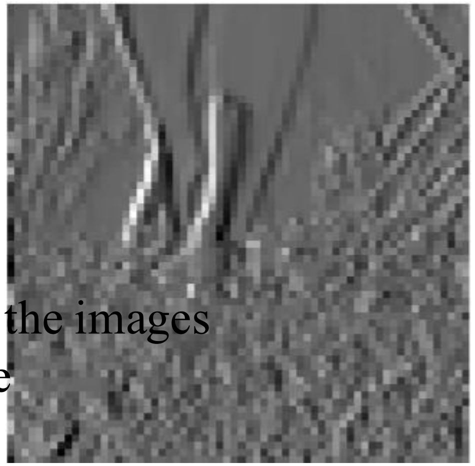
Eye movements, when combined with temporal filtering, also produce a whitened power spectrum

Original Images

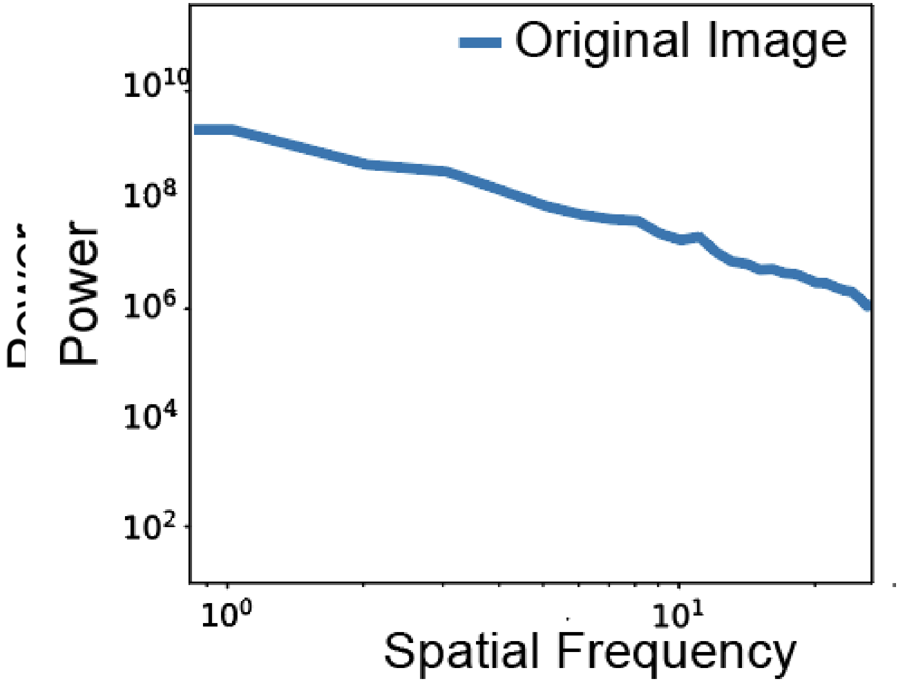
Filte
r



Filtered Images



Take a difference between the images
using a temporal derivative

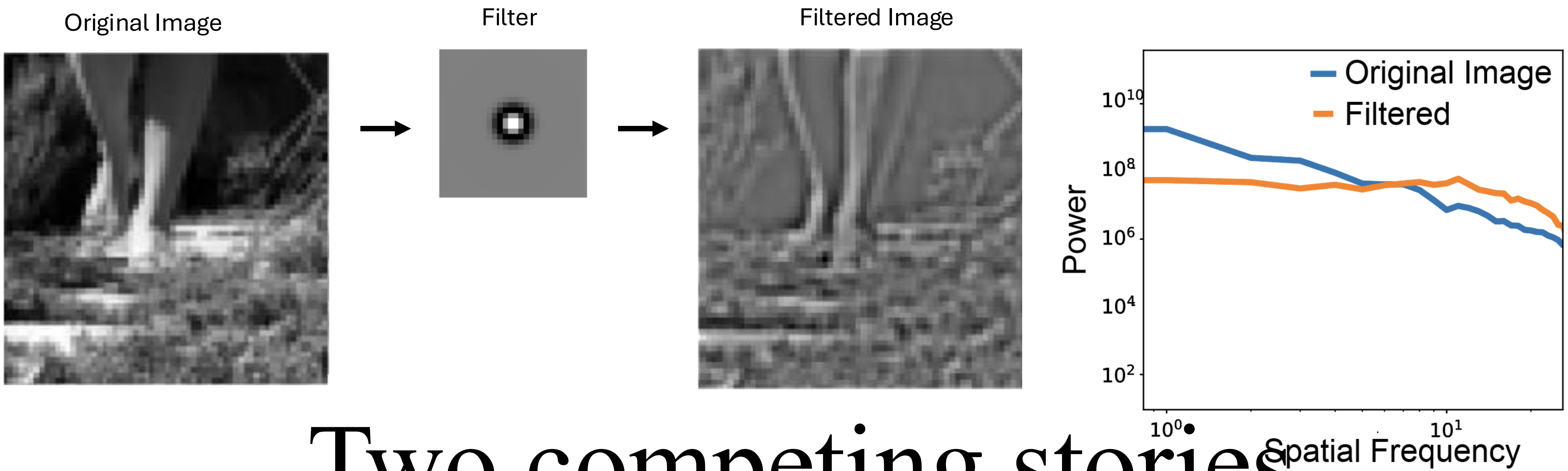


Signal is dominated by lower freqs

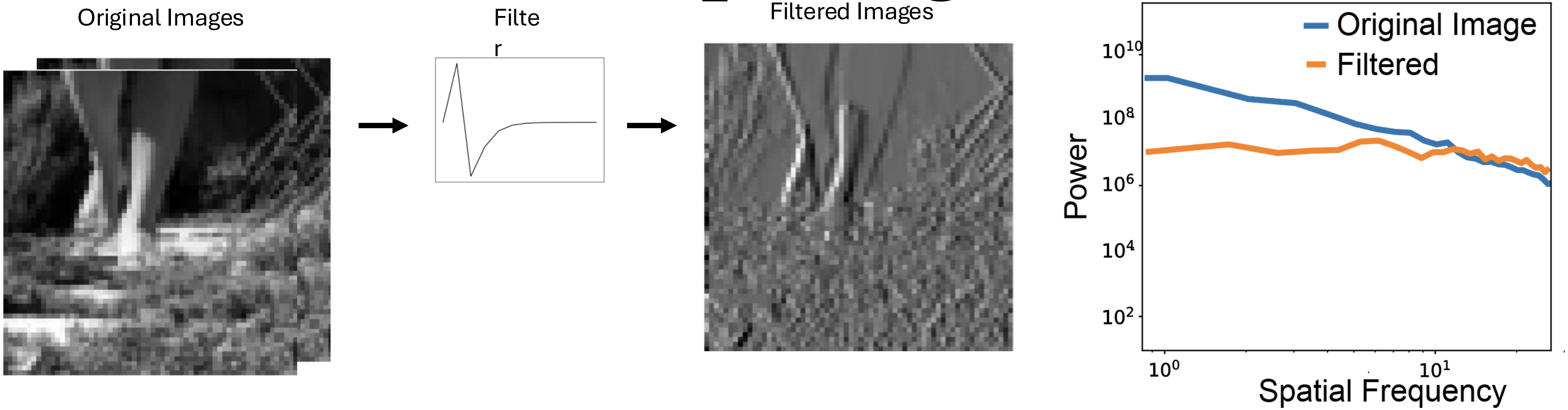
Edges remain

Power is flattened
& pairwise correlations removed

Dong & Atick, 1995; Dan et al. 1996; Pitkow & Meister 2012; Kuang et al, 2012; Rucci & Victor, 2015; Anderson et al, 2020; Karamanlis, 2022



Two competing stories



Present Work

- Models based on efficient coding predict center-surround receptive fields when trained on natural images

- If we train the same model on eye movements, do the learned representations change?

- To test this, we will train an efficient coding model on movies with and without fixational eye movements & compare the learned representations

Methods

Data

Ocular drift simulation

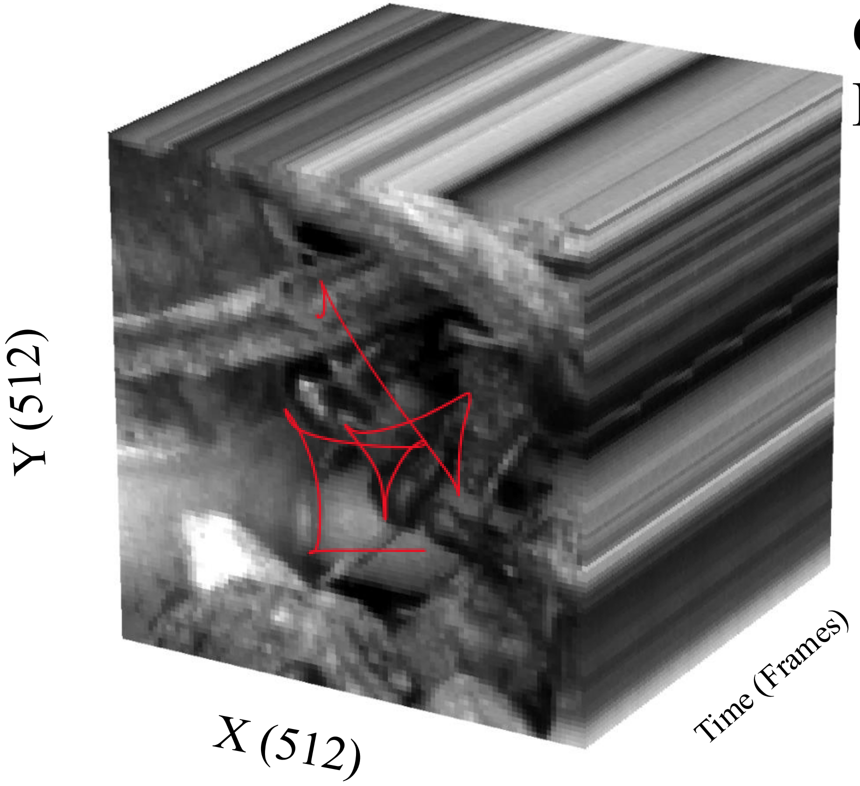
Model

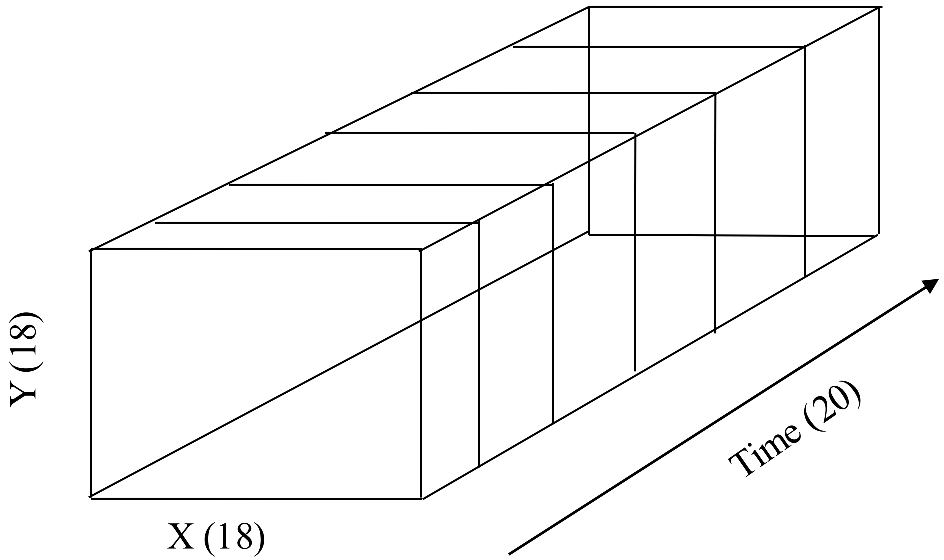
Chicago Motion Database

- Palmer Group at the University of Chicago in Chicago, IL
- Fixed-camera recordings of moving objects filmed primarily outdoors in the Chicagoland area, or through a dissecting microscope
- Clips contain constant and in many cases full-frame motion of animals, plants, and water

https://cmd.rcc.uchicago.edu

Ocular Drift Simulation via Brownian Motion





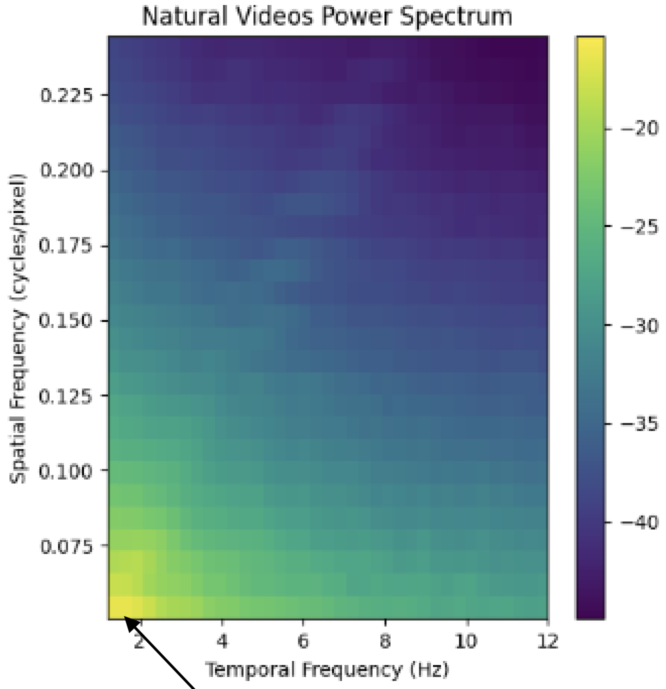
- "Drift trajectories can be statistically modelled as a self-avoiding random walk"
- Recent study found Brownian motion provided a best fit for actual drift recordings
- Simulation of ocular drift by sampling natural movies along Brownian trajectories parameterized by recorded drift statistics

- Random walk over pixel space
- At each time step t, sampling position coordinates $\boldsymbol{x(t)}$ and $\boldsymbol{y(t)}$ are displaced by a Gaussian perturbation
  - $\boldsymbol{x(t)=x(t-1)+\sqrt{2D\Delta t}\cdot\Delta X}$
  - $\boldsymbol{y(t)=y(t-1)+\sqrt{2D\Delta t}\cdot\Delta Y}$
    - $\boldsymbol{D}$ is a diffusion constant (40 $arcmin^2/s$)
    - Δt is the frame rate of 30 fps (1/30)
    - ΔX and ΔY are draws from a random normal distribution ~ $\mathcal{N}(0,1)$

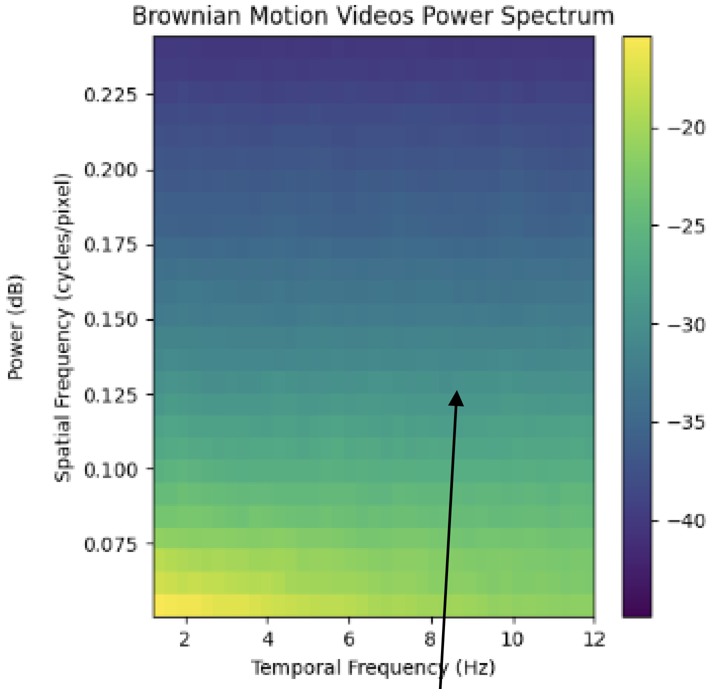
Engbert, et al., 2011; Kuang et al., 2012; Alexander & Martinez-Conde, 2019

Movies with and without eye movements demonstrate differences in power

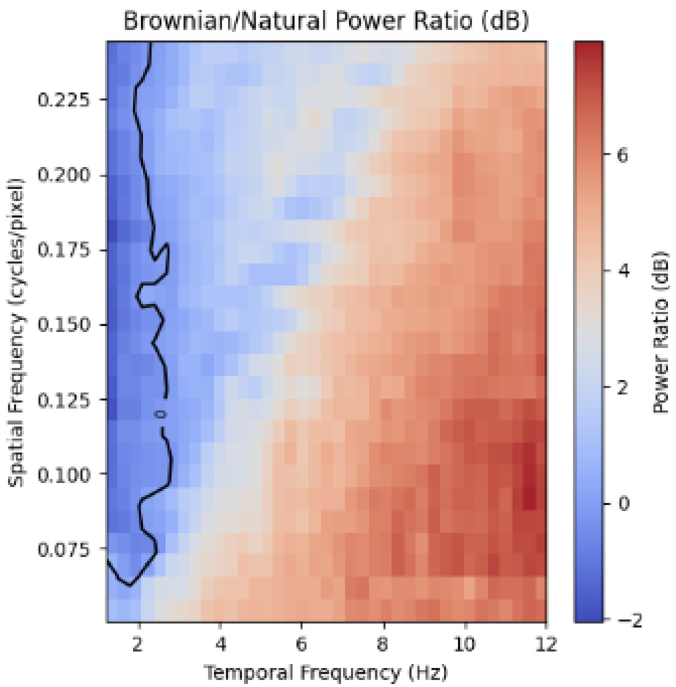
Movies **without** eye movements



Movies **with** eye movements



Difference in power



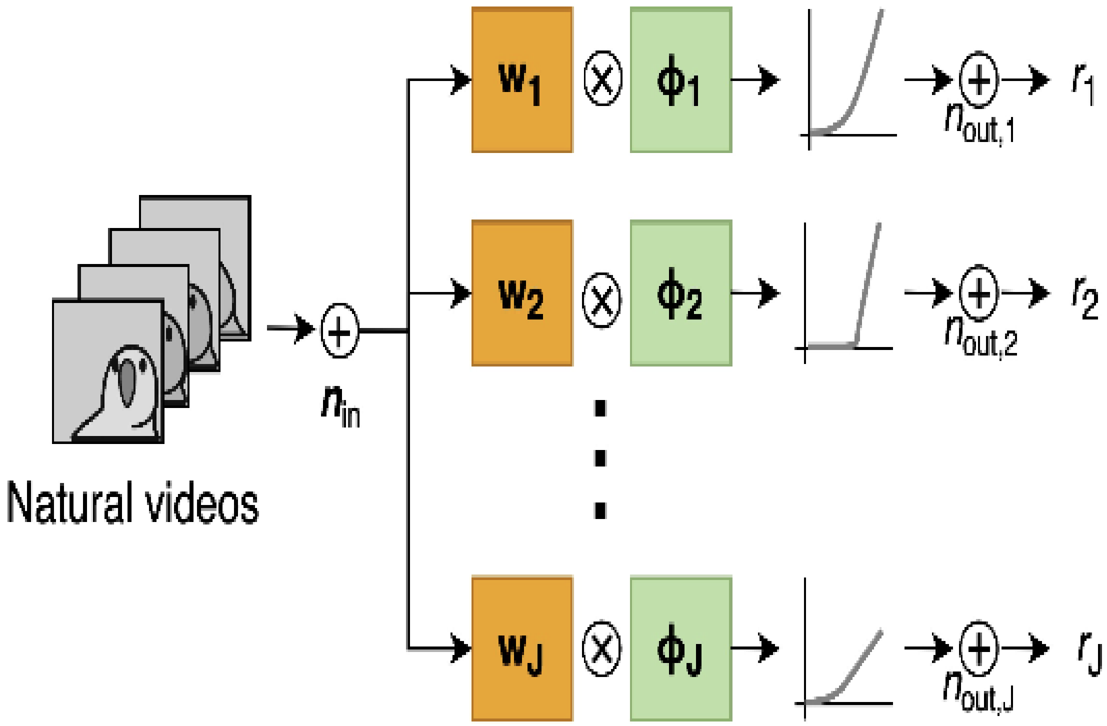
Power is strongest in lower spatial frequencies at zero-temporal frequency

Power is redistributed into time & across spatial frequencies

Anything to the right of the black line means the eye movement videos have more power

Goal: maximize information transfer between input x and the neural response r, subject to metabolic cost of firing spikes

$$I(X;R) - \sum_j \lambda_j \langle r_j \rangle$$



$$\begin{aligned} \text{maximize} \quad & \log \frac{\det\left(\mathbf{G}\mathbf{W}^\top\left(\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{n_{\text{in}}}\right)\mathbf{W}\mathbf{G} + \mathbf{C}_{n_{\text{out}}}\right)}{\det\left(\mathbf{G}\mathbf{W}^\top\mathbf{C}_{n_{\text{in}}}\mathbf{W}\mathbf{G} + \mathbf{C}_{n_{\text{out}}}\right)} \\ \text{subject to} \quad & \mathbb{E}[r_j] = 1. \end{aligned}$$

Generator Signal

$$y_j = \mathbf{w}_j^T\left(\mathbf{x} + \mathbf{n}_x\right)$$

Unit norm weight vector of neuron j

Input signal

Gaussian noise (input/sensory noise)

Scalar output

Neural Response

$$r_j = f_j\left(y_j\right) + n_r$$

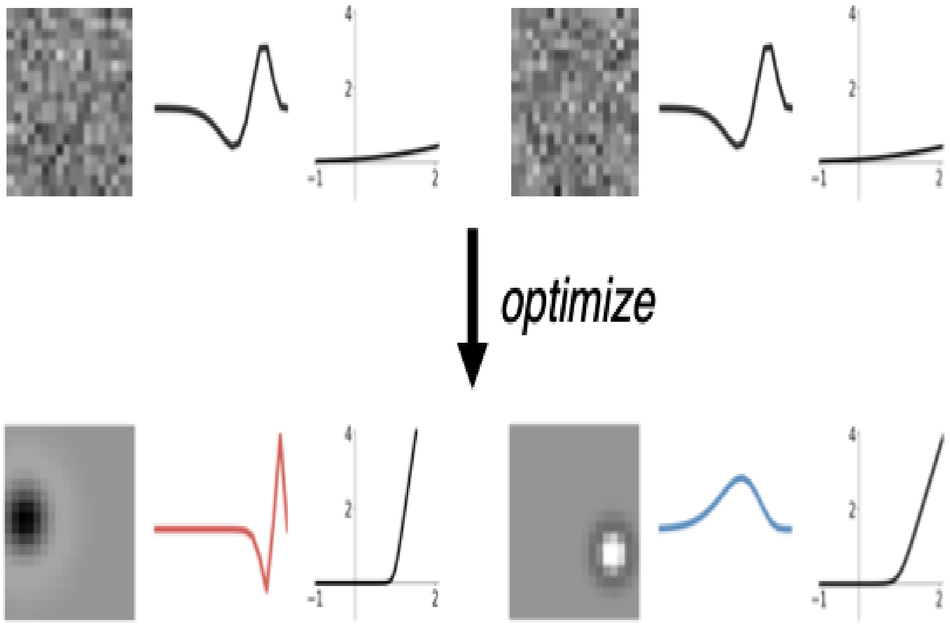
point-wise NL

Generator Signal

Gaussian noise (output/neural noise)

The Model

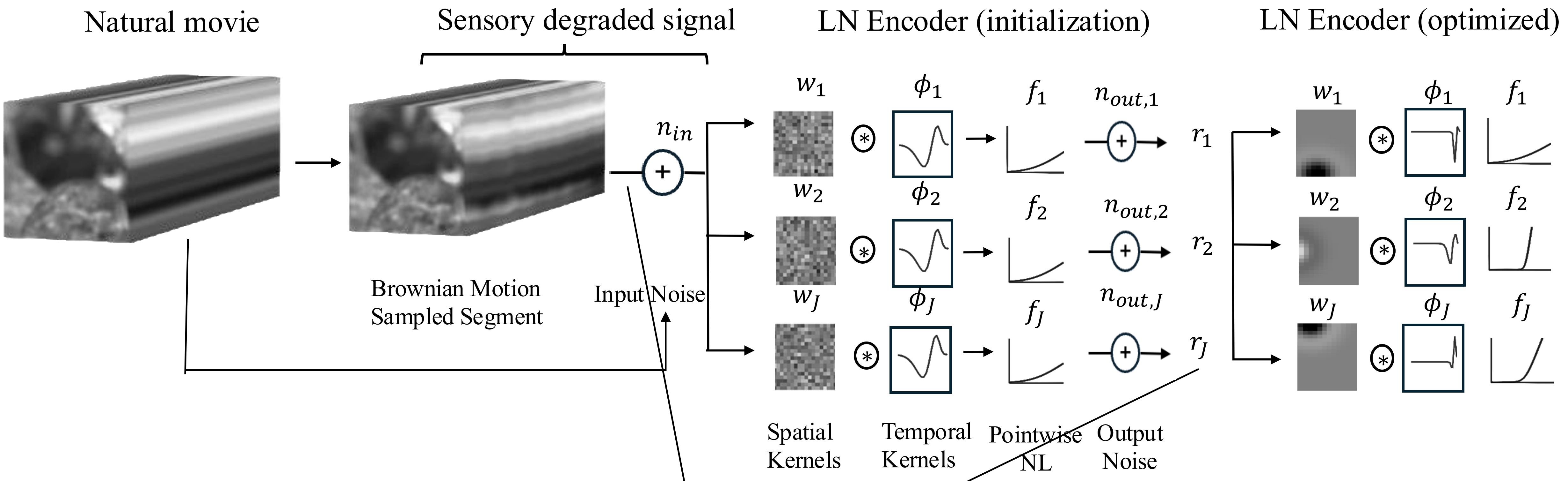
Jun et al, 2022



- When trained on natural movies
- Model predicts conventional receptive fields
- Predictions are consistent with previous literature and observed neural data

Karklin & Simoncelli, 2011; Doi & Lewicki, 2014; Ocko et al, 2019; Jun et al, 2022

Add Eye Movements to Training



input x

neural response r

Maximize information transfer

$$I(X;R) - \sum_j \lambda_j \langle r_j \rangle$$

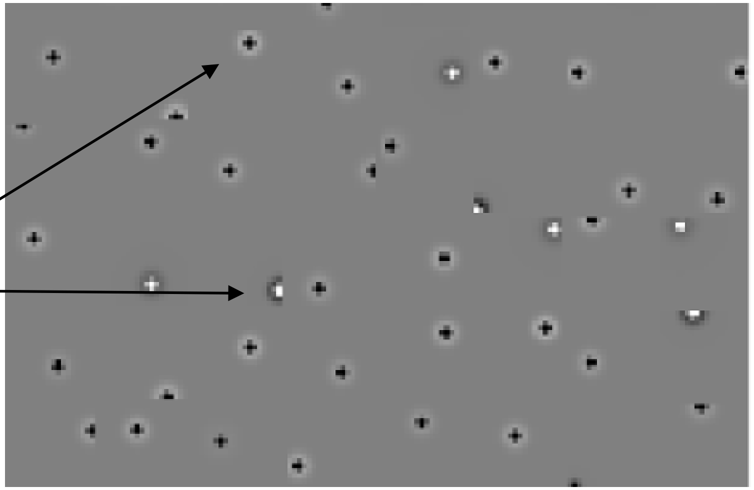
subject to metabolic cost of firing spikes

Results

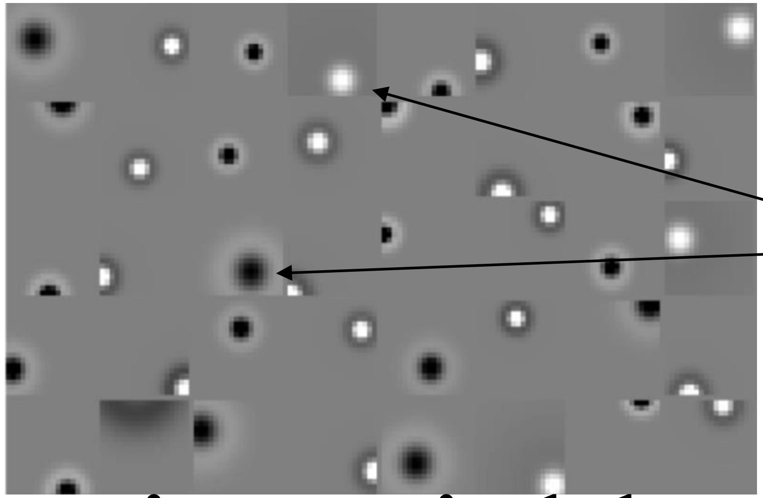
Training on movies with eye movements leads to representations that emphasize temporal changes

Learned Receptive Fields

Natural Condition

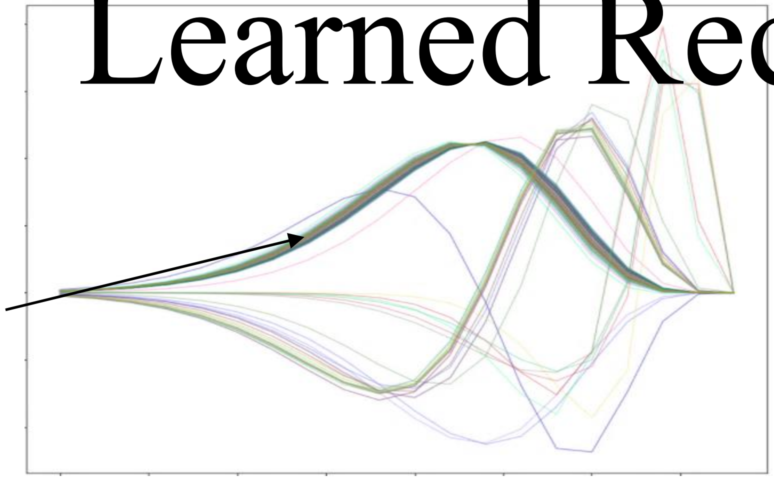


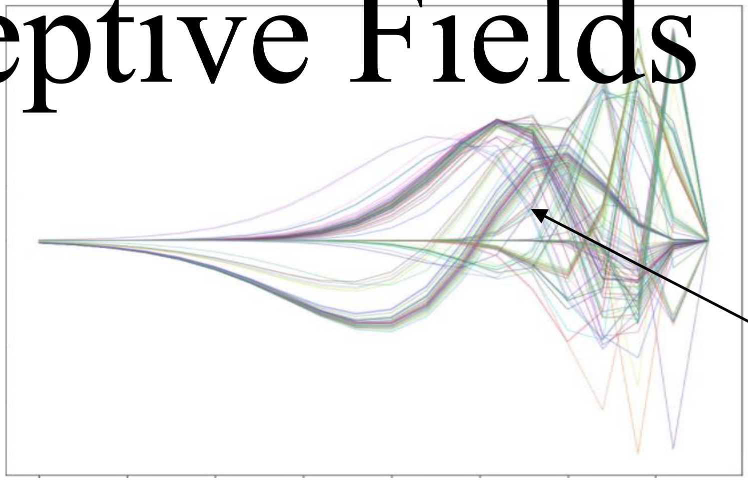
Brownian Condition



Compact spatial filters

Diffuse spatial filters



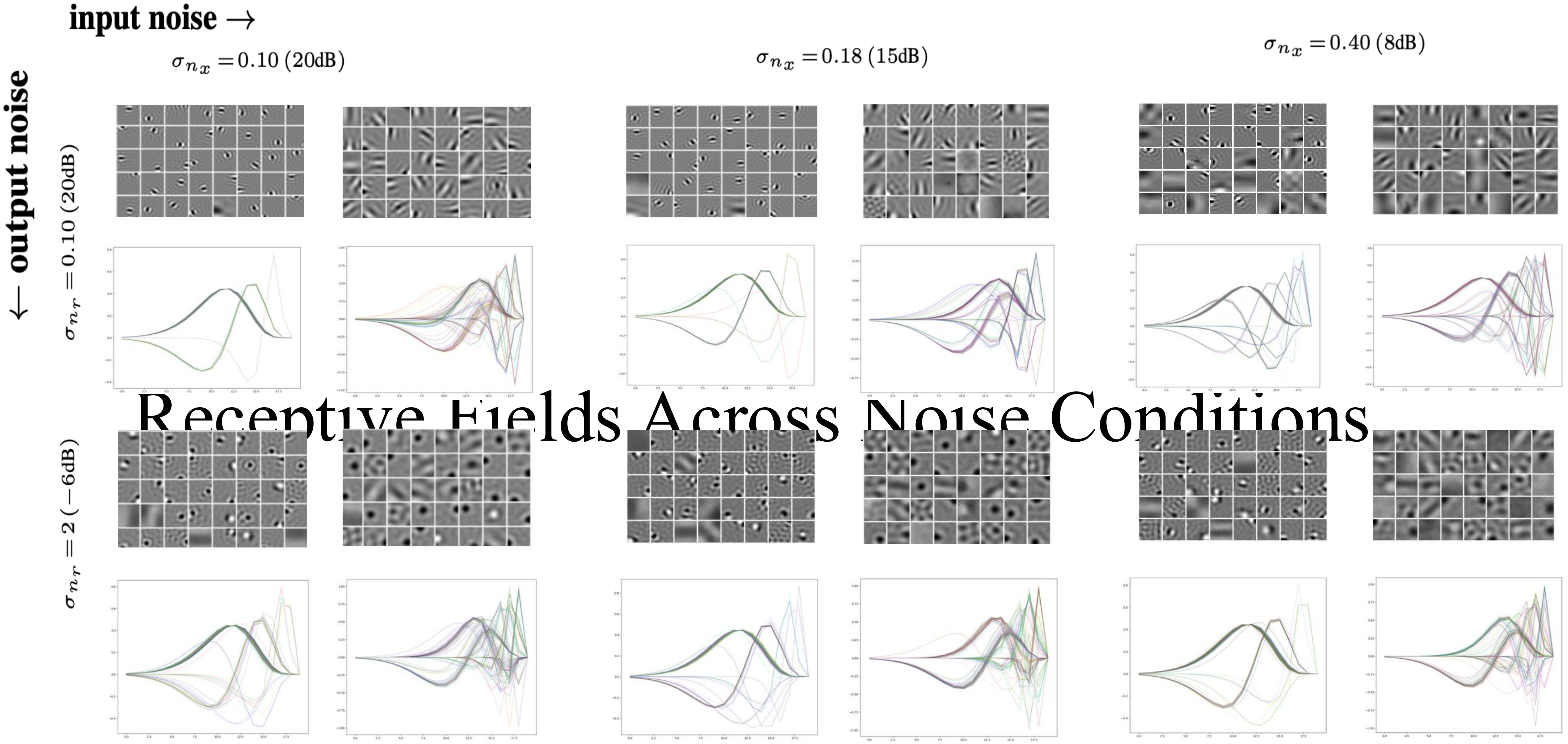


Higer convergence toward monophasic filters

Higer convergence toward biphasic filters

Take away: Both predict Center-surround like RGCs; Brownian filters show less spatial selectivity & increased temporal selectivity

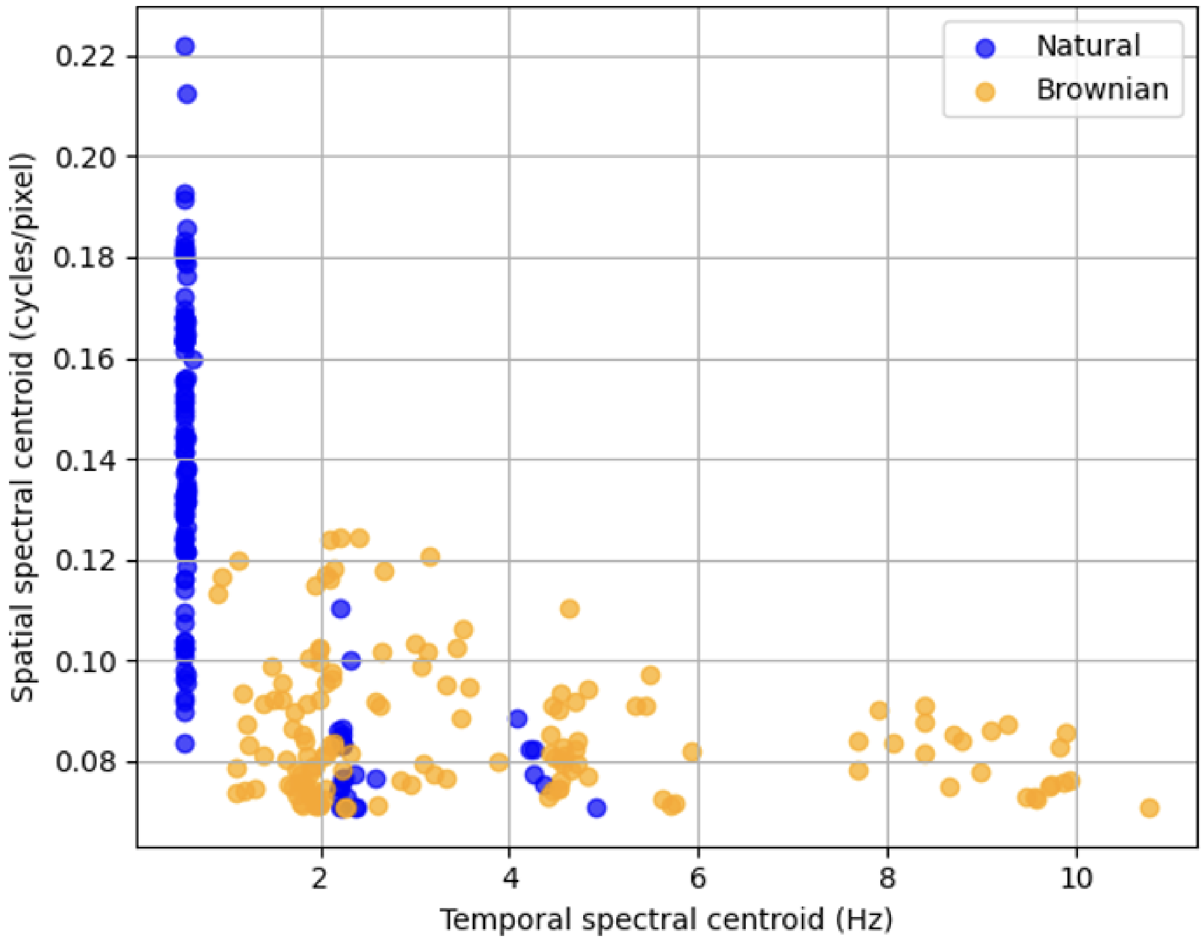
Receptive Fields Across Noise Conditions



Take away: spatial RFs become more diffuse, & temporal RFs become faster, more biphasic, and have more cell type diversity

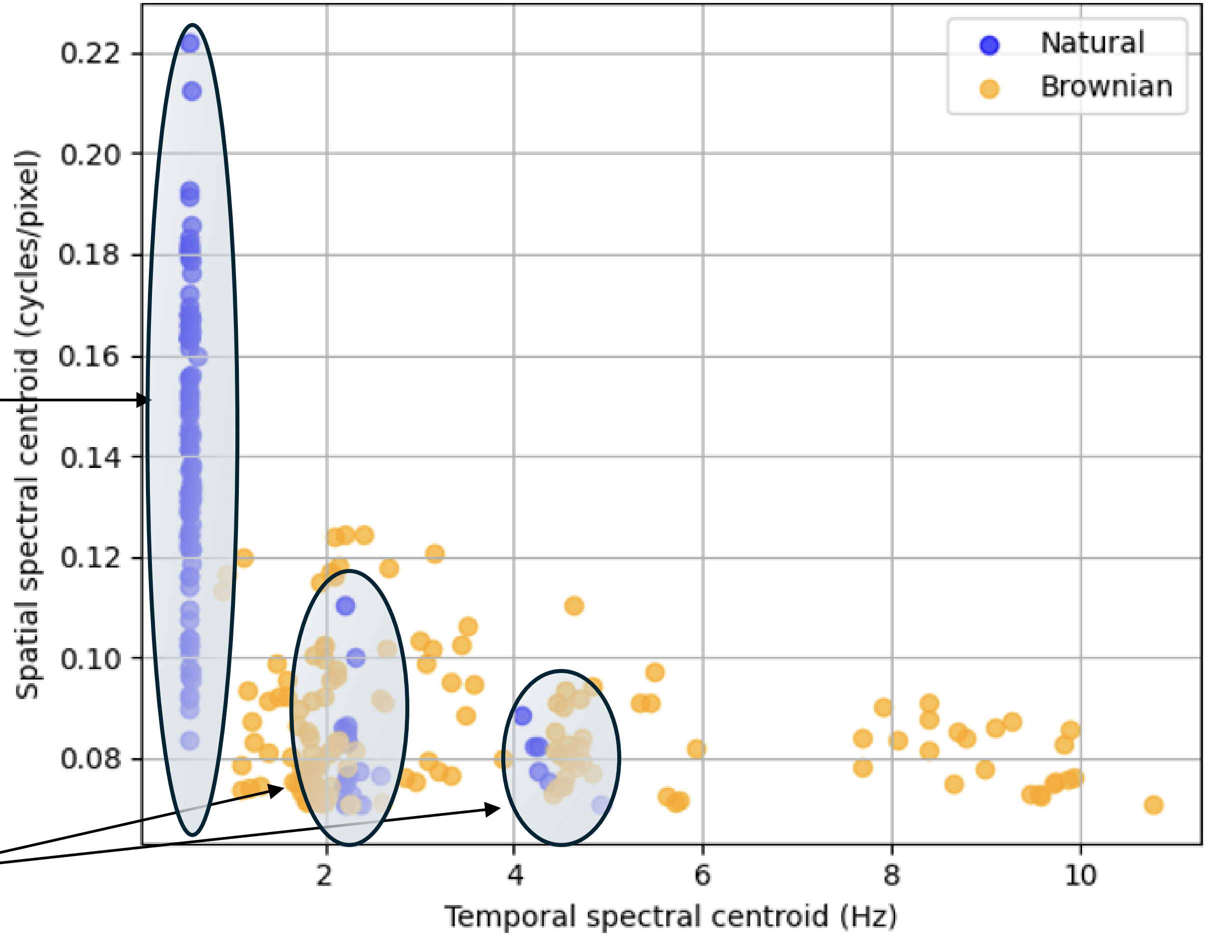
Spectral Profiles

This plot shows where a given cell's spectral sensitivity is most concentrated



Natural Movie Training

Filters form strong clusters for **spatial frequency selectivity**

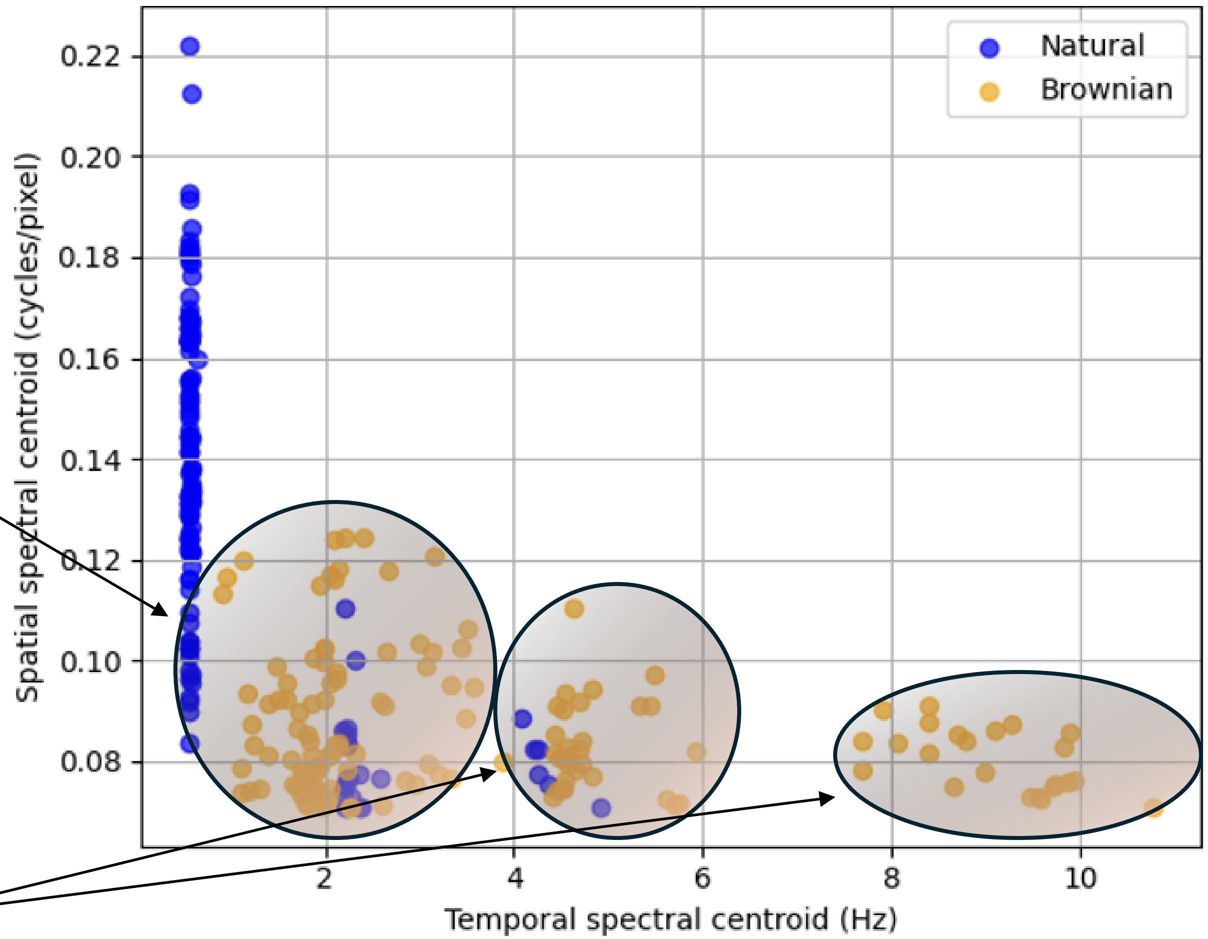


And less for **temporal frequency selectivity**

You may notice the clusters form in the regions where power is most concentrated (**space**)

Brownian Motion Training

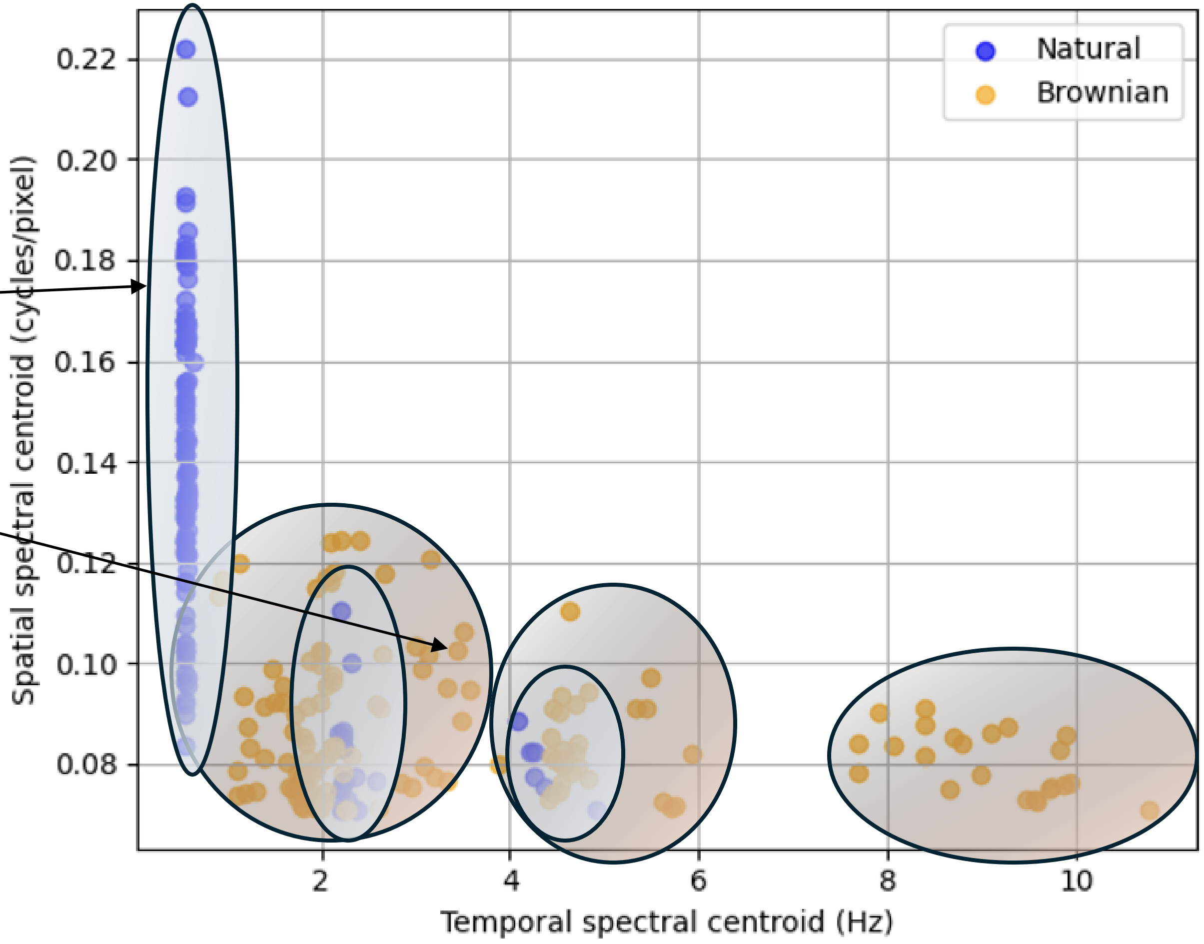
Less clustering for **spatial frequency selectivity**



And more for **temporal frequency selectivity**

Cells cluster in the regions where power is most concentrated (**time**)

Eye movements shift learned receptive fields from spatial to temporal selectivity



Spectral Profiles

$$\bar{f}_{\text{spatial}} = \frac{\sum_i r_i \cdot \text{PSD}(r_i)}{\sum_i \text{PSD}(r_i)}$$

r_i is the radial frequency bin center

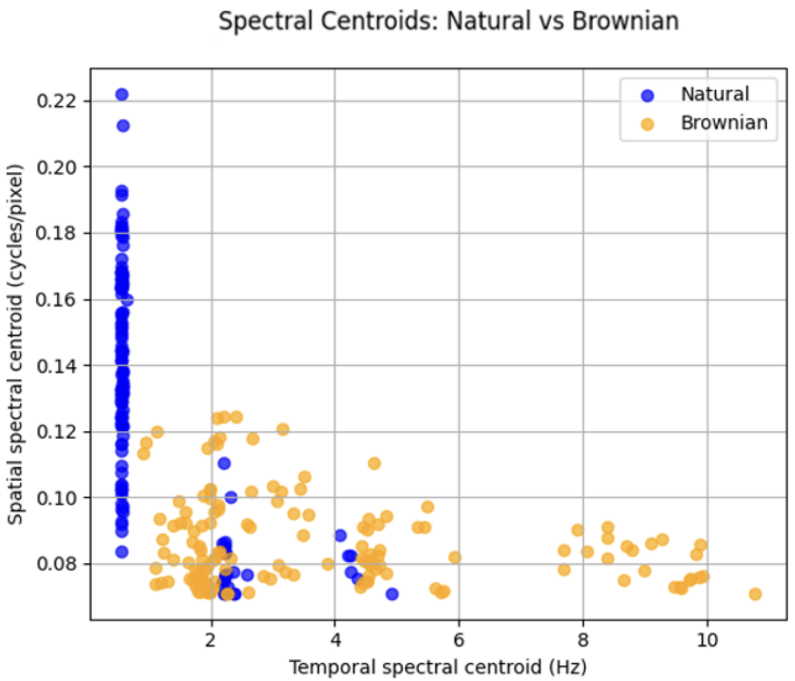
$\text{PSD}(r_i)$ is the mean power at that bin

$\bar{f}_{spatial}$ is the spatial spectral centroid, which tells you where in the radial frequency spectrum the receptive field concentrates most of its energy.

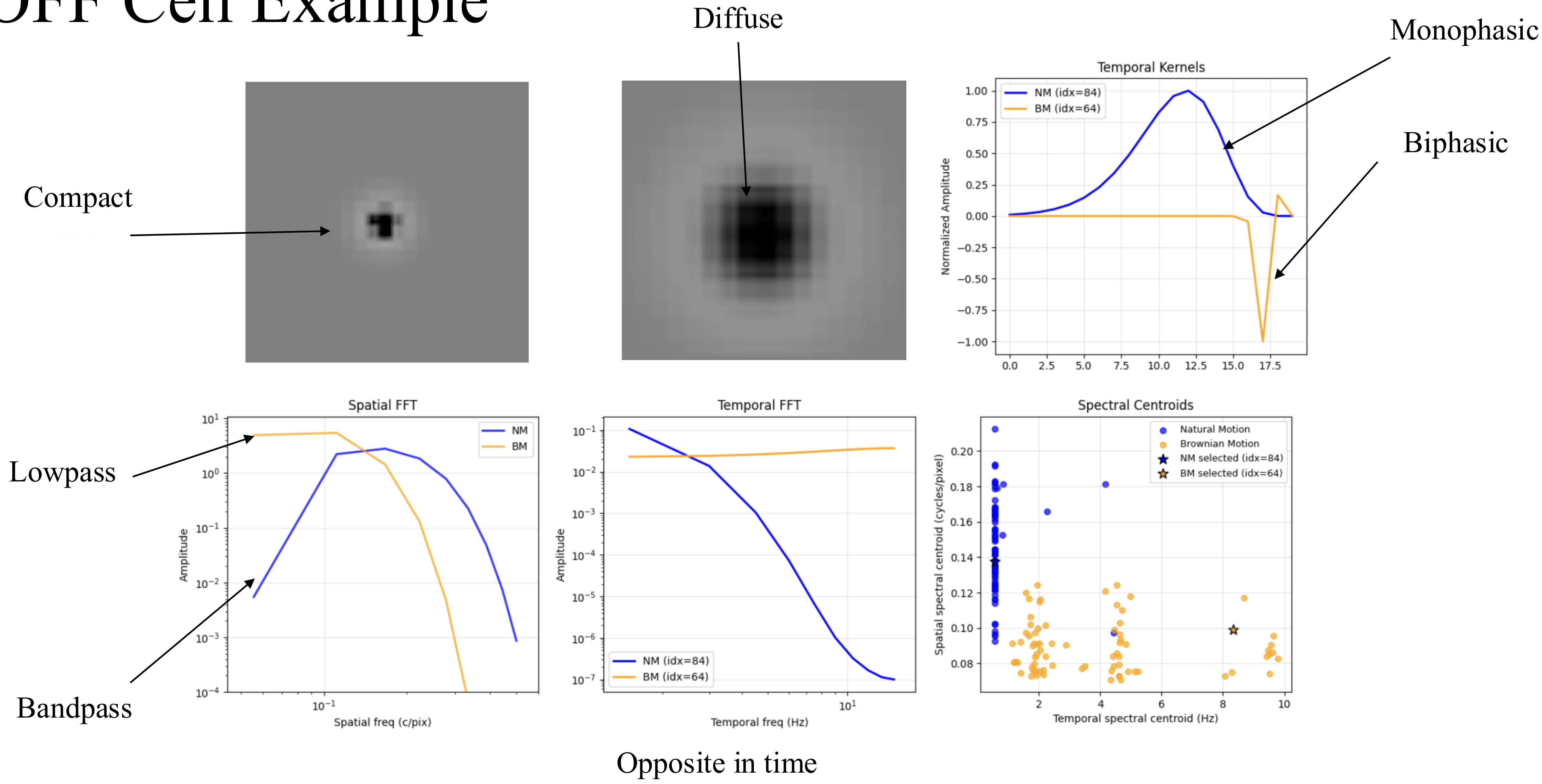
$$\bar{f}_{\text{temporal}} = \frac{\sum f_i \cdot P(f_i)}{\sum P(f_i)}$$

f_i = frequency in Hz
$P(f_i)$ = power at that frequency
$\bar{f}_{temporal}$ gives you a single number (in Hz) that tells you where in the frequency spectrum the temporal kernel concentrates its energy.

$$\frac{\text{cycles}}{\text{degree}} \div \frac{120\text{ px}}{\text{degree}} = \frac{\text{cycles}}{\text{pixel}}$$

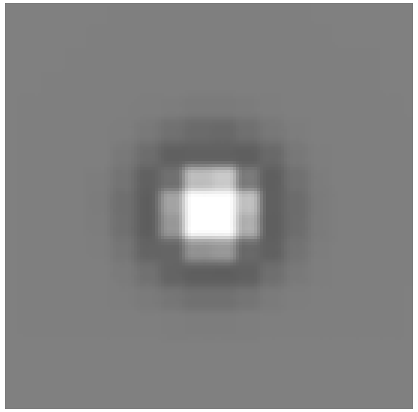


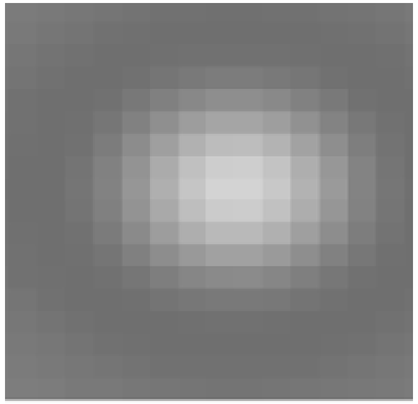
OFF Cell Example

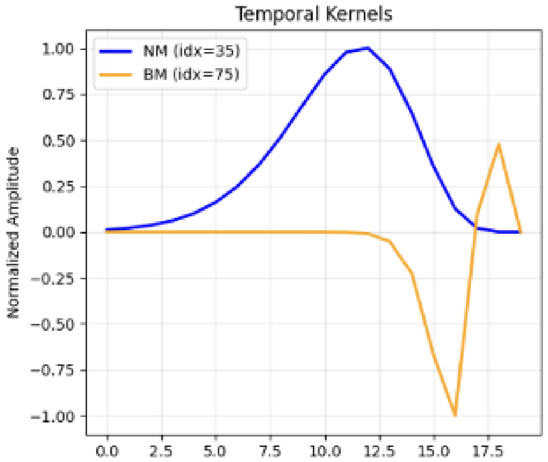


ON Cell Example

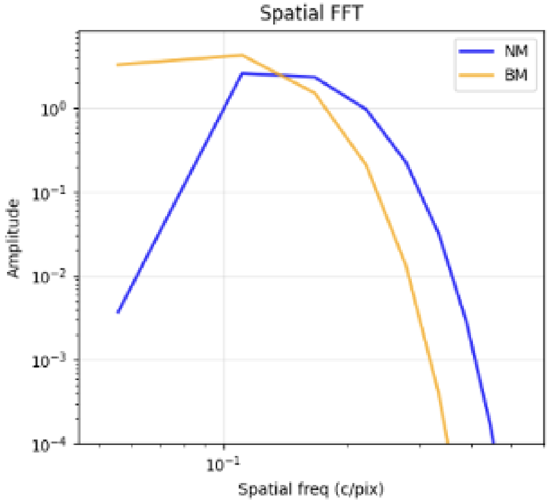
Space:
Compact to diffuse

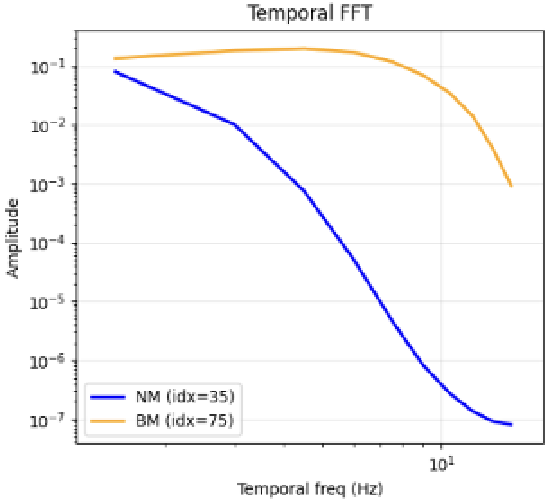


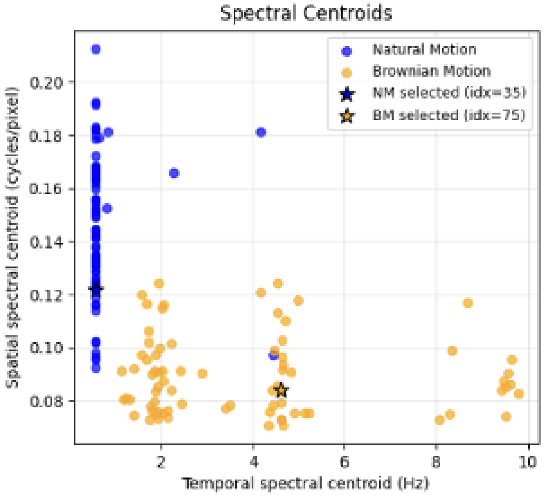




Time:
Mono to biphasic

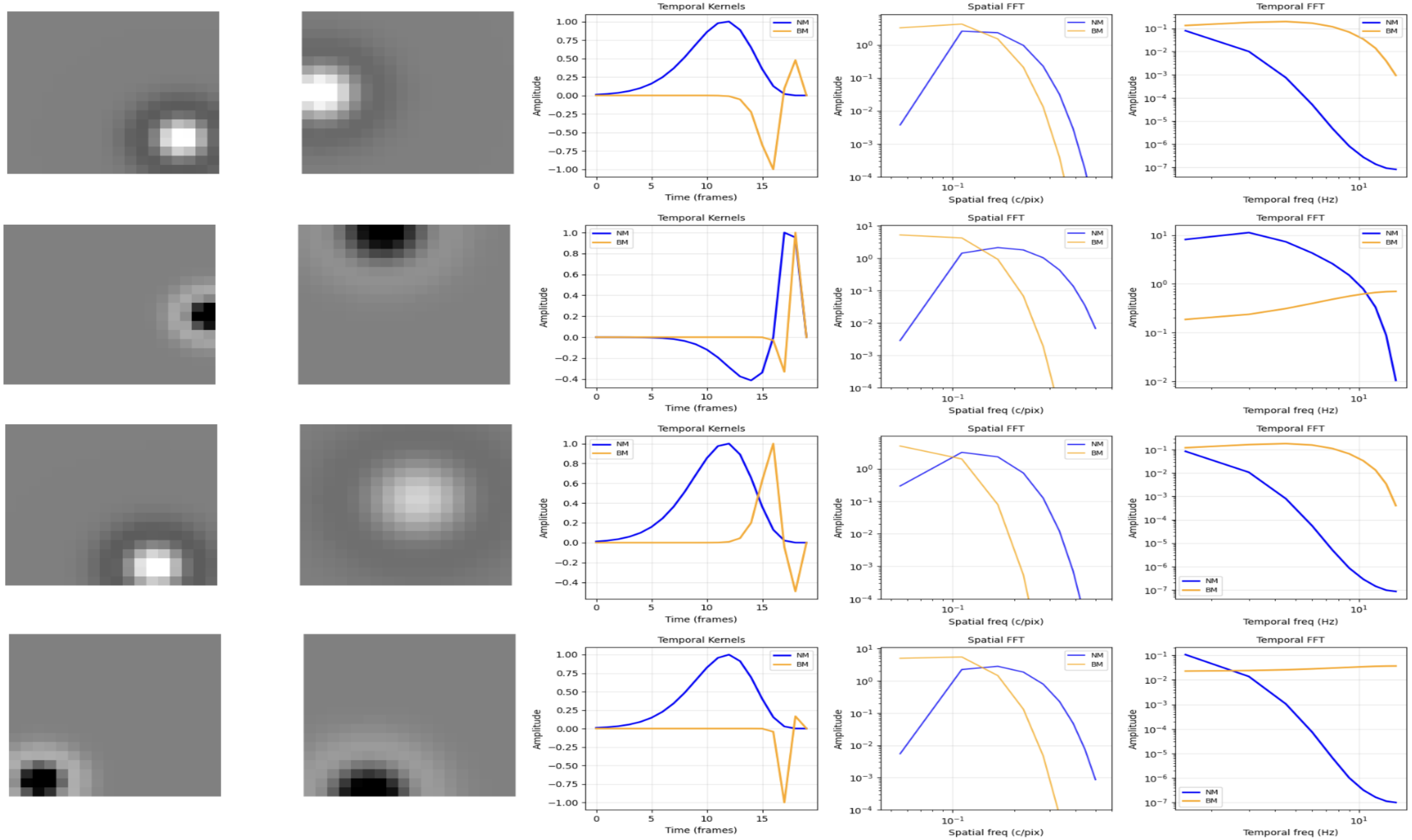






SF:
Band to lowpass

TF:
Lowpass to flat


Temporal Kernels
Amplitude
Time (frames)
NM
BM
Spatial FFT
Spatial freq (c/pix)
Temporal FFT
Temporal freq (Hz)

- If we train the same model on eye movements, do the learned representations change?

- To test this, we trained an efficient coding model on movies with and without fixational eye movements

- Training with eye movements predicted receptive fields
  - Lower spatial selectivity
  - Higher temporal selectivity

- Suggests eye movements introduce another layer to the encoding scheme of the retina, one we hope to discover through further exploration

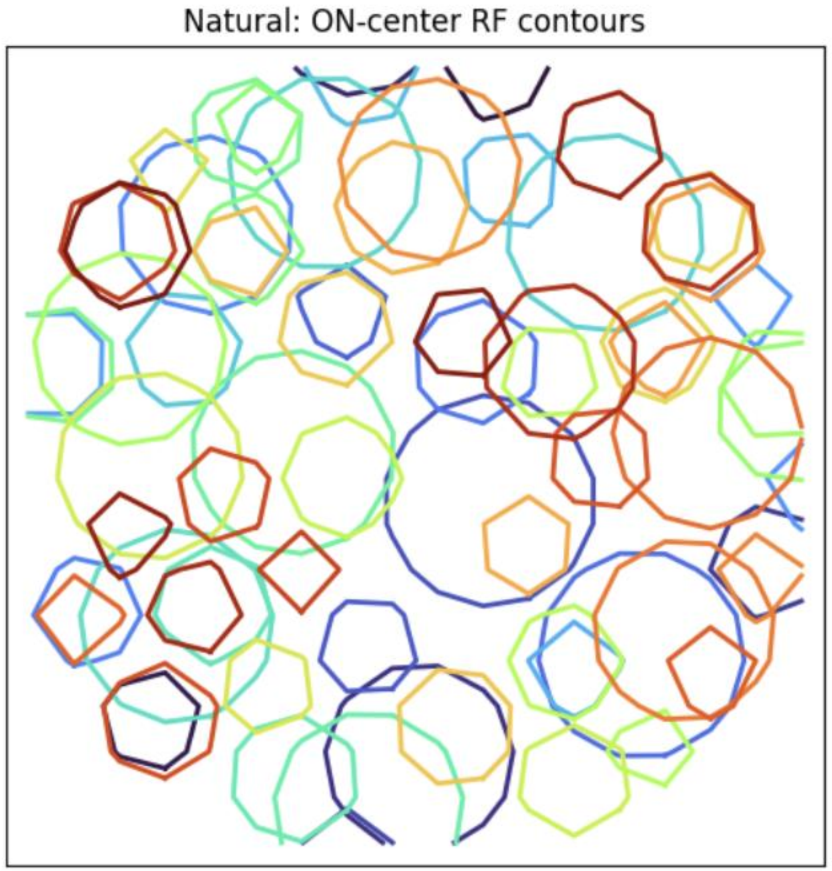
What’s next?

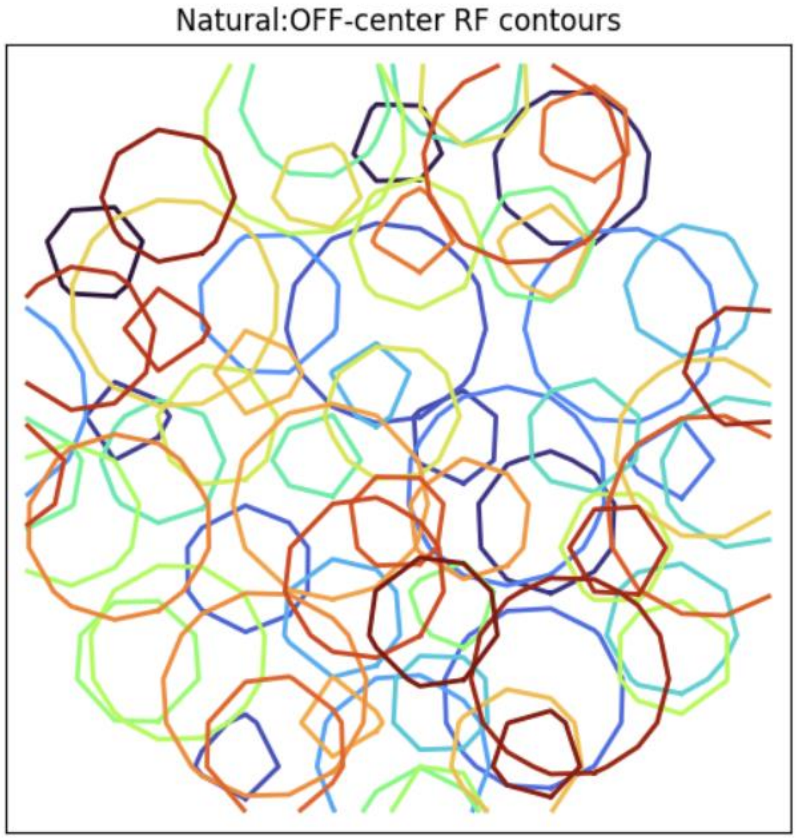
Do BM predicted filters match actual RGCs?

Do cells whiten; if so, how?

Do cells form mosaics?

Natural Movies Condition





Brownian Movies Condition