DATA ANALYSIS CHALLENGES IN 21CM LIM: A HERA PERSPECTIVE

Steven Murray

LIM25, Annecy, June 2025

Funded by the European Union

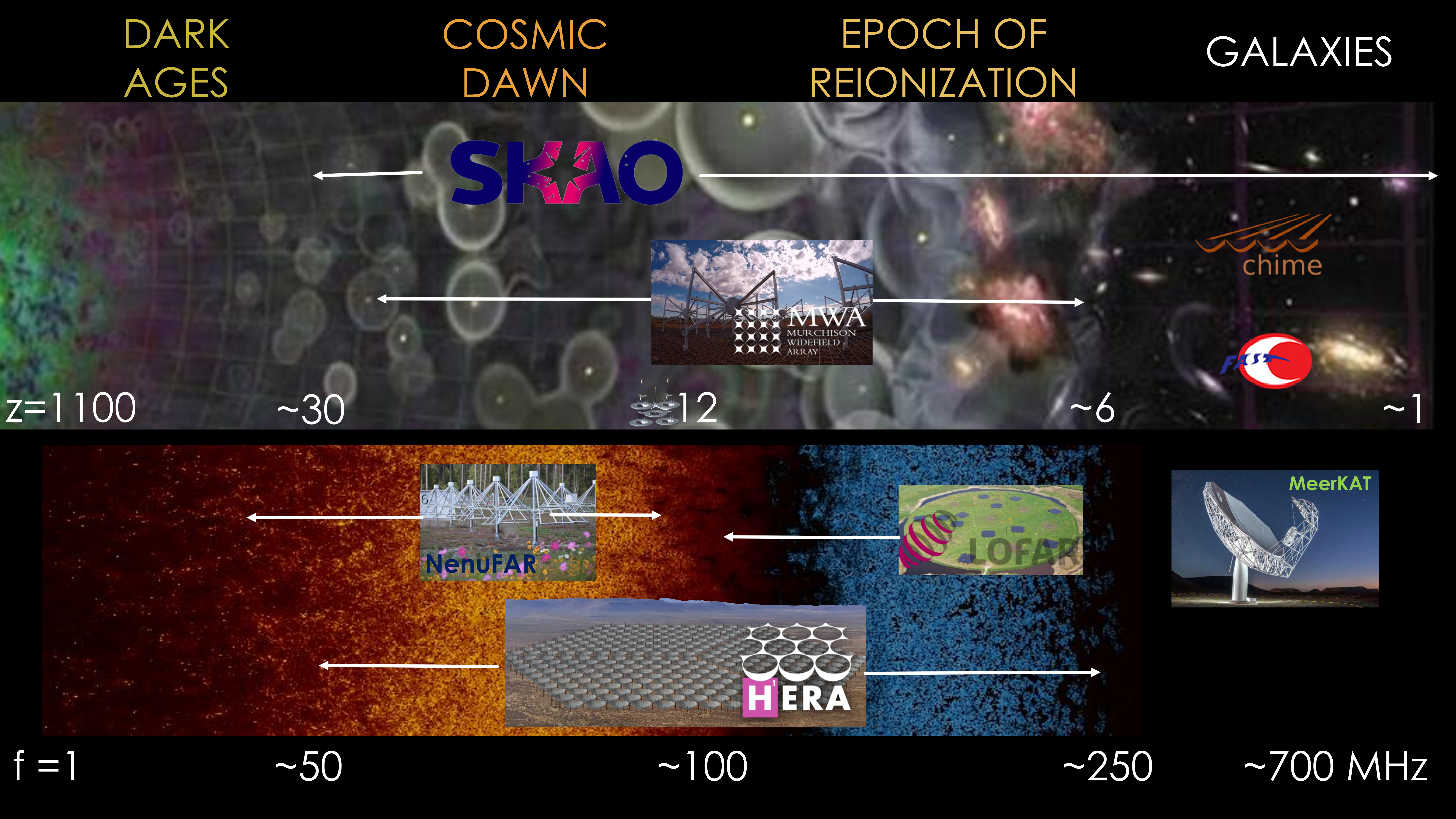
European Research Council
Established by the European Commission

Scuola Normale Superiore

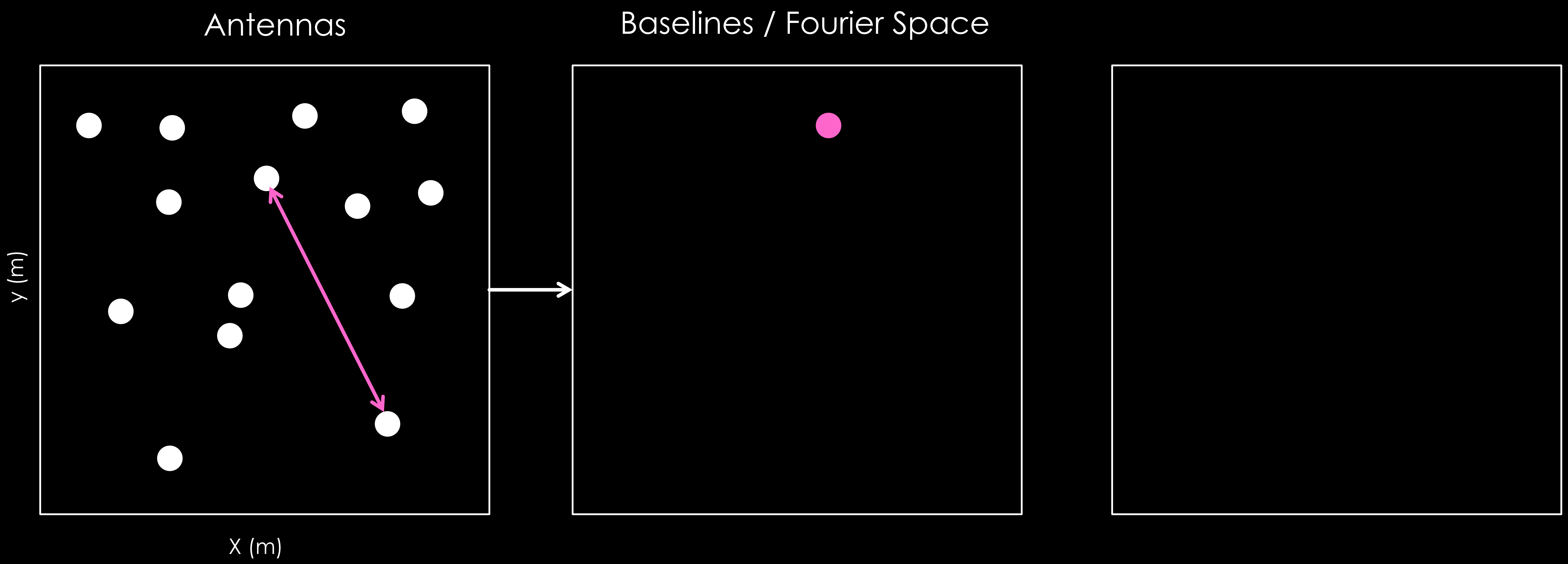
DALL-E3 *"HERA gazing at the first stars, with a rubber duck"*

OUTLINE

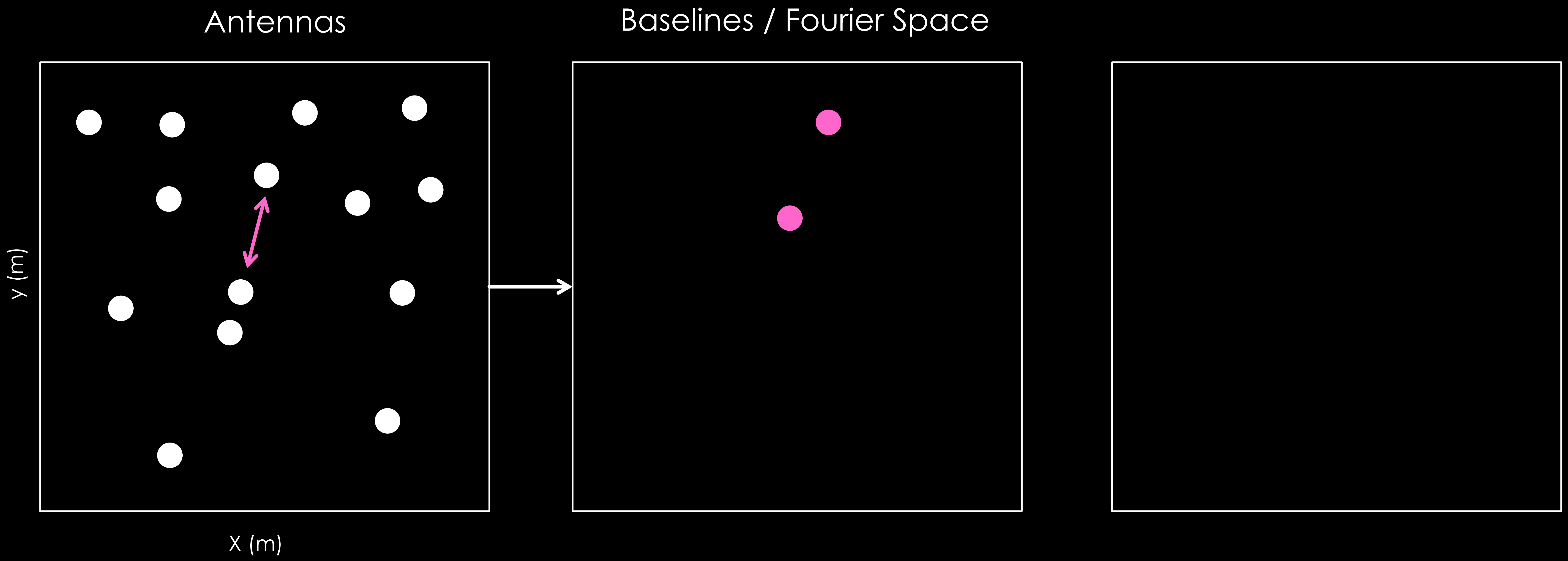
1. Why the 21cm line?
2. Big-picture challenges + Approaches
3. Specific Challenges:
    1. RFI and Missing data
    2. Mutual Coupling
    3. Propagation of Systematics
    4. Simulations and Validation


DARK AGES
COSMIC DAWN
EPOCH OF REIONIZATION
GALAXIES
SKAO
MWA
MURCHISON WIDEFIELD ARRAY
chime
FAST
z=1100
~30
~12
~6
~1
NenuFAR
LOFAR
MeerKAT
HERA
f =1
~50
~100
~250
~700 MHz

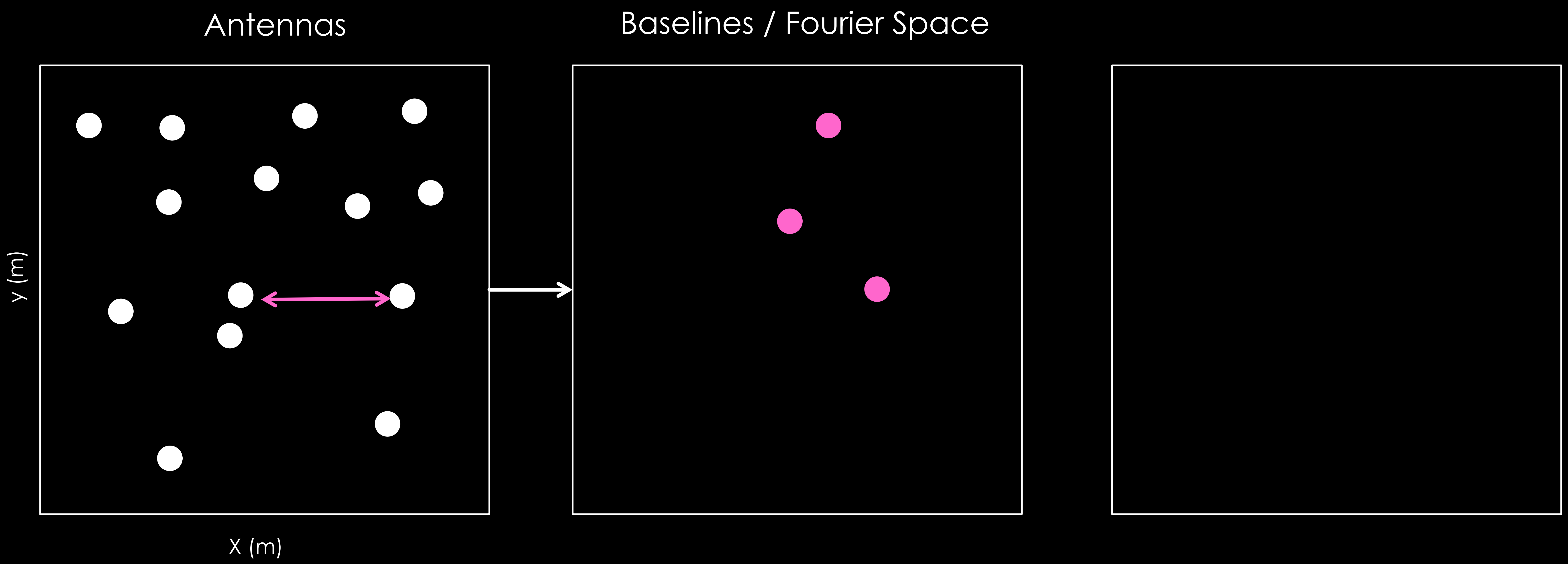
INTERFEROMETERS



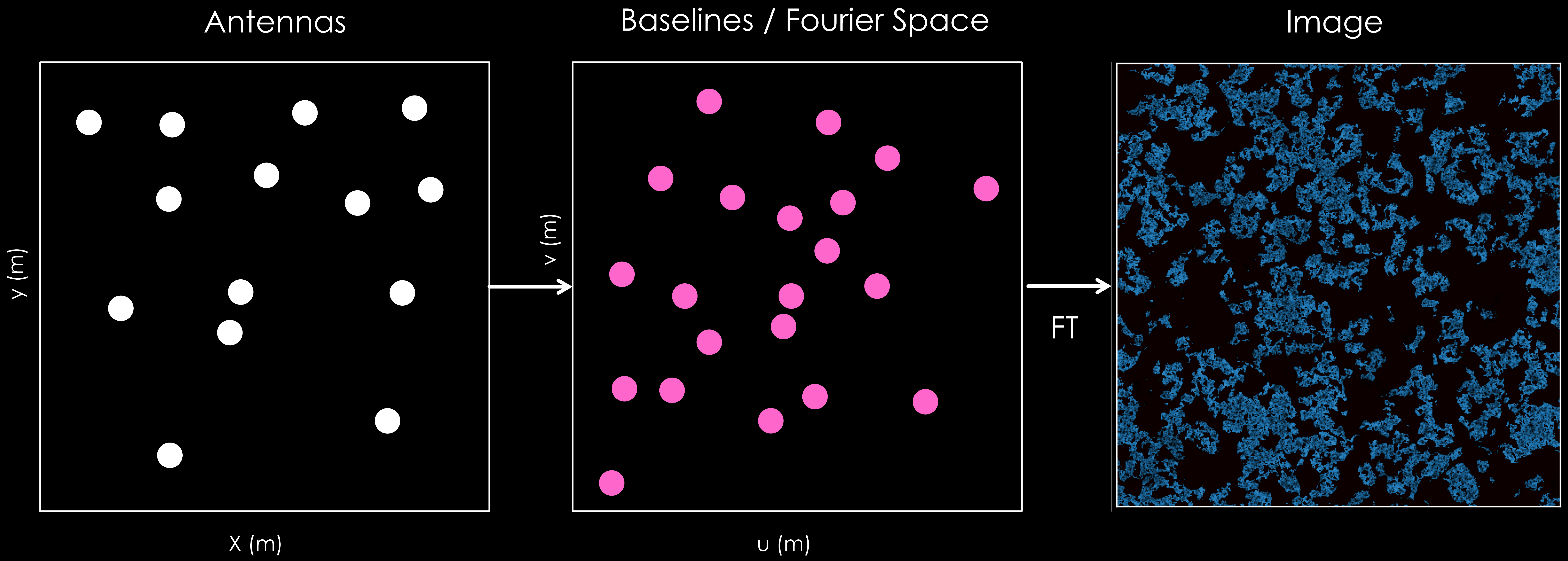
INTERFEROMETERS



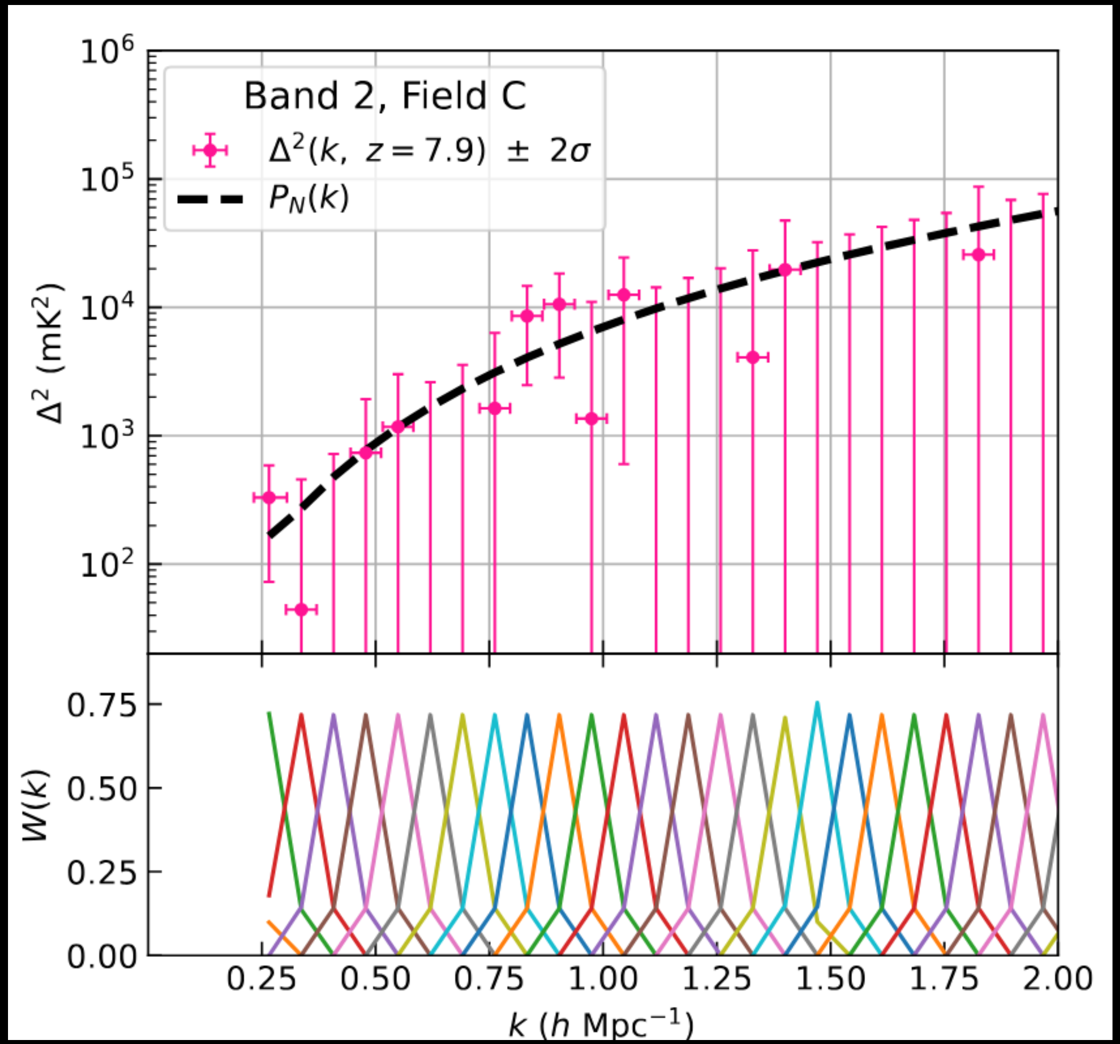
INTERFEROMETERS



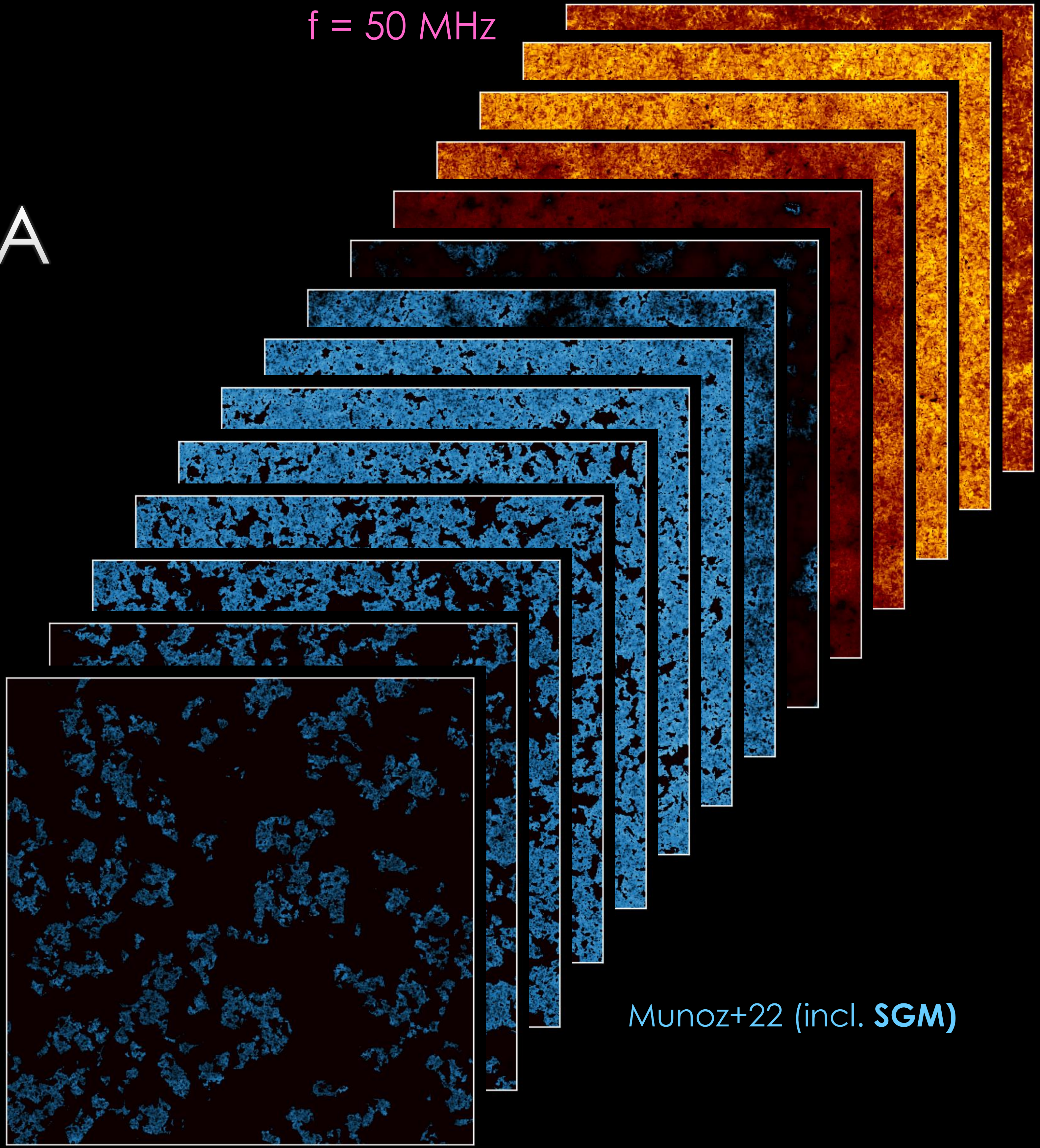
INTERFEROMETERS



3D MAPS & POWER SPECTRA

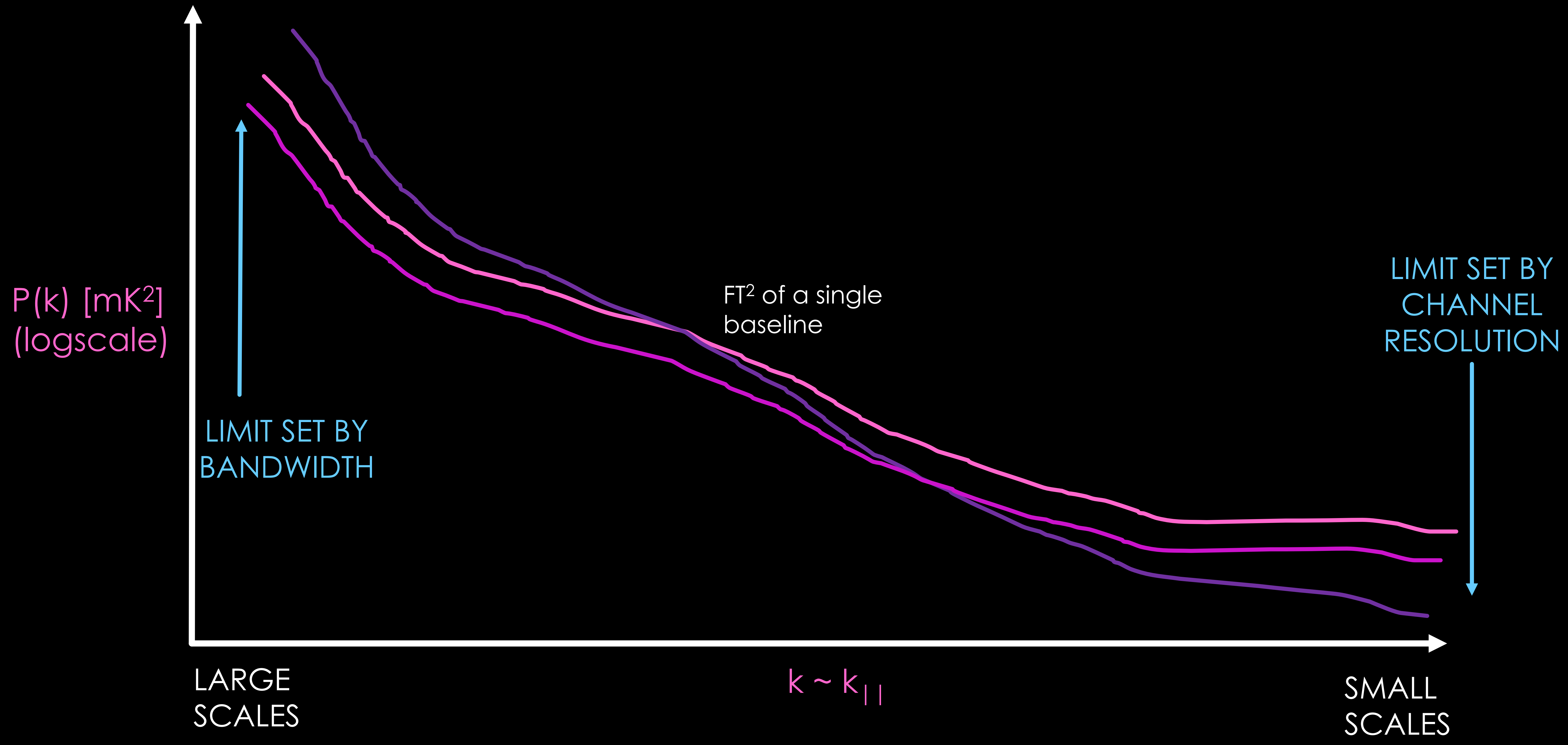


HERA Collaboration 2022
(incl. **SGM)**

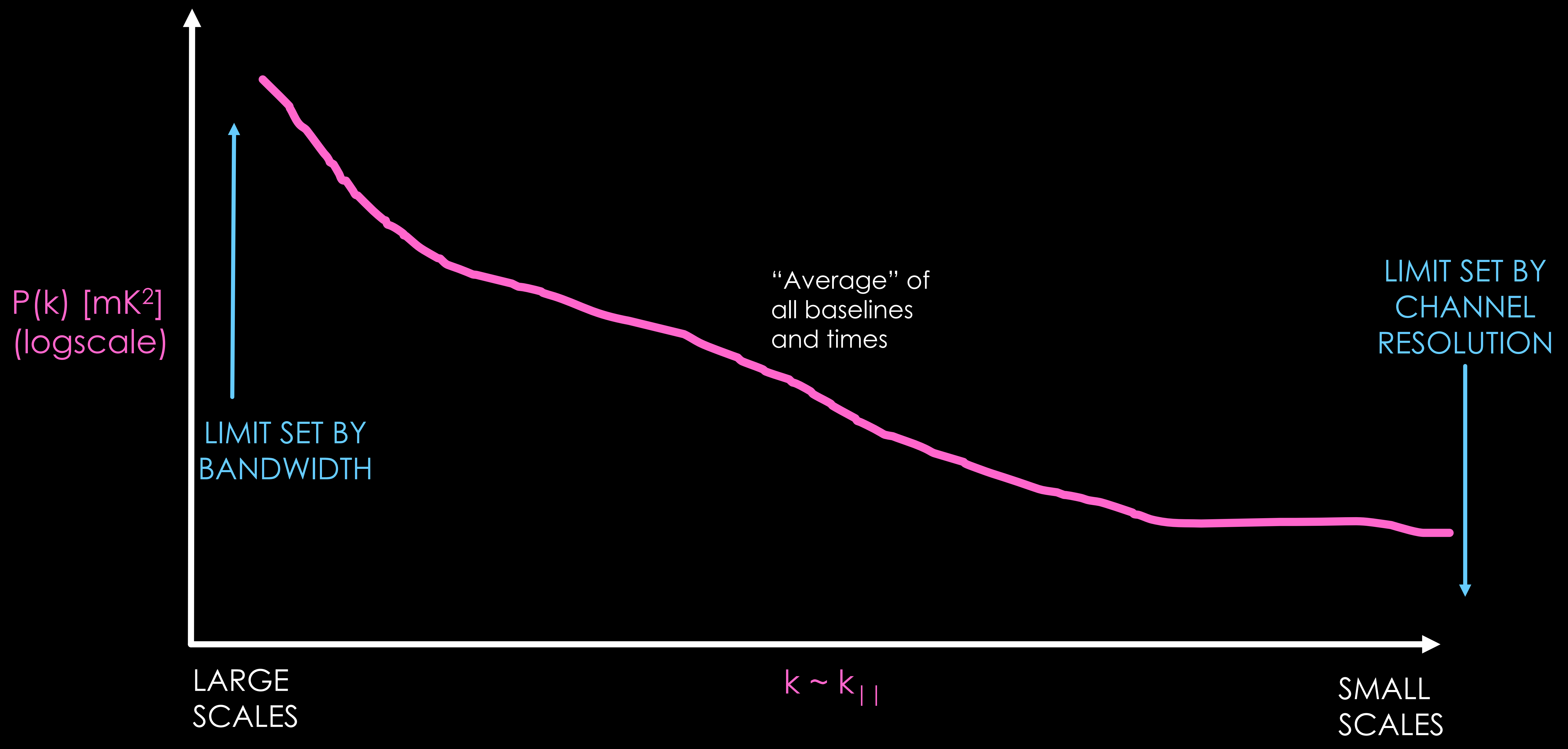


Munoz+22 (incl. **SGM)**

THINKING IN POWER SPECTRUM SPACE

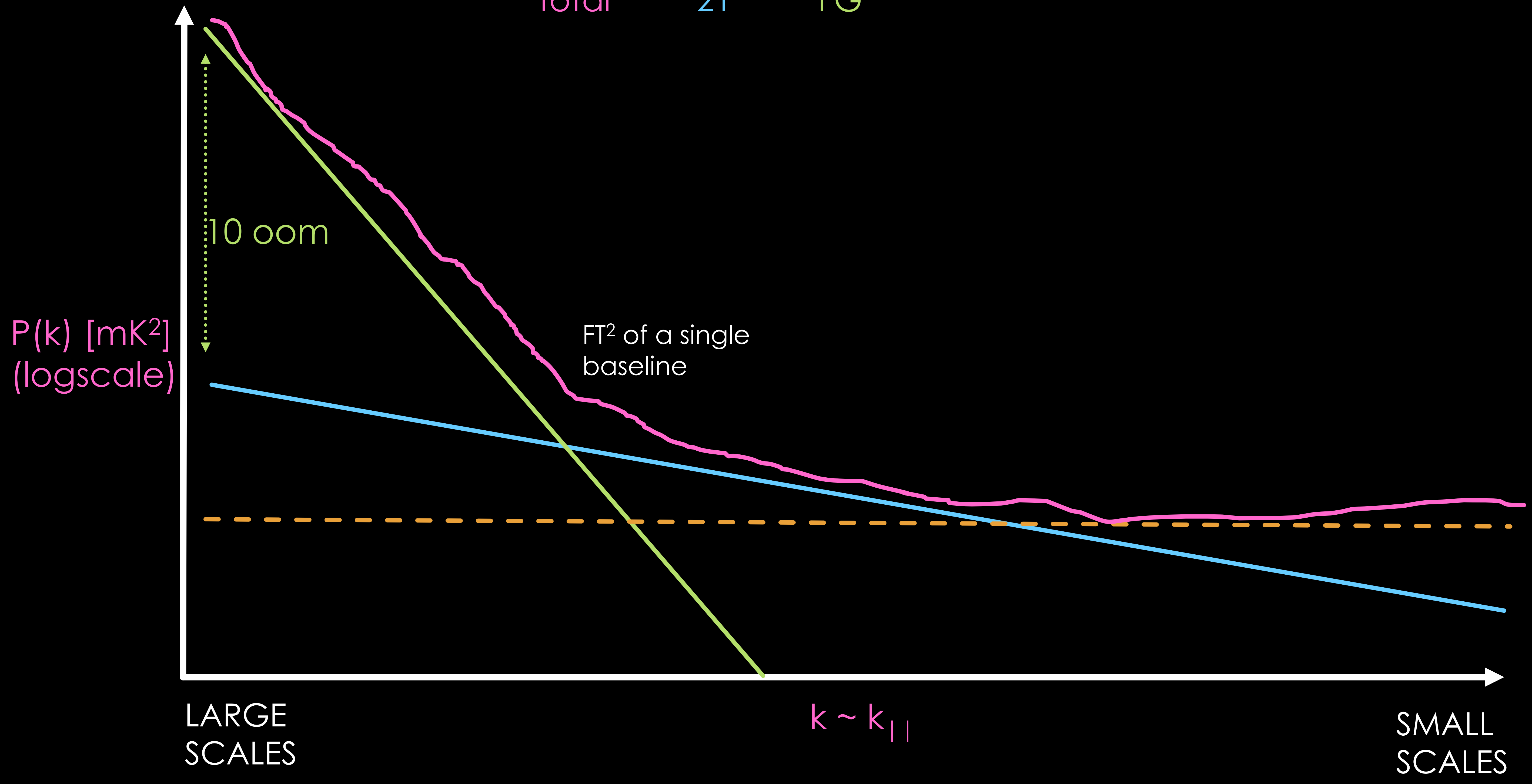


THINKING IN POWER SPECTRUM SPACE



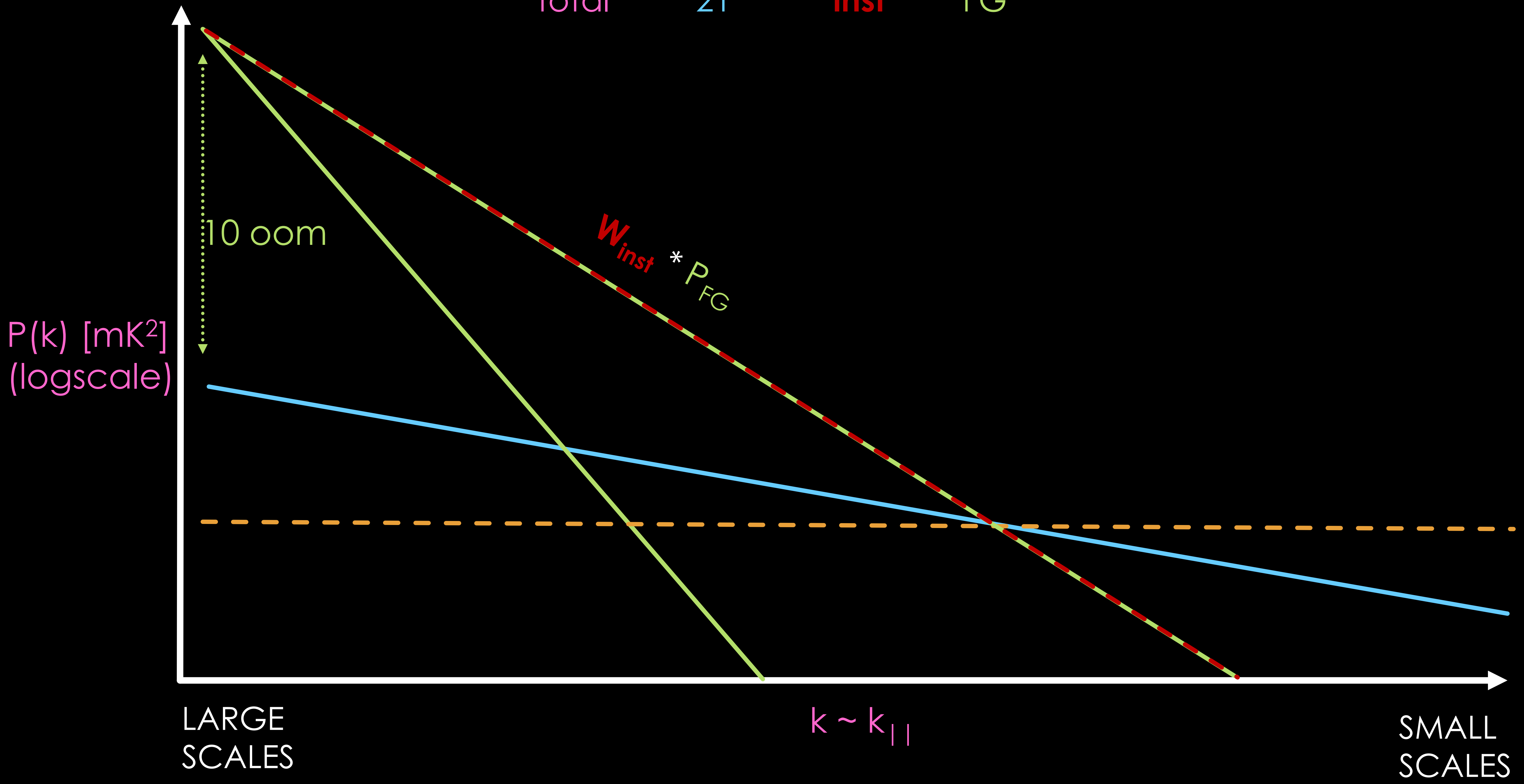
THINKING IN POWER SPECTRUM SPACE

$P_{total} = P_{21} + P_{FG} + \text{Noise}$



THINKING IN POWER SPECTRUM SPACE

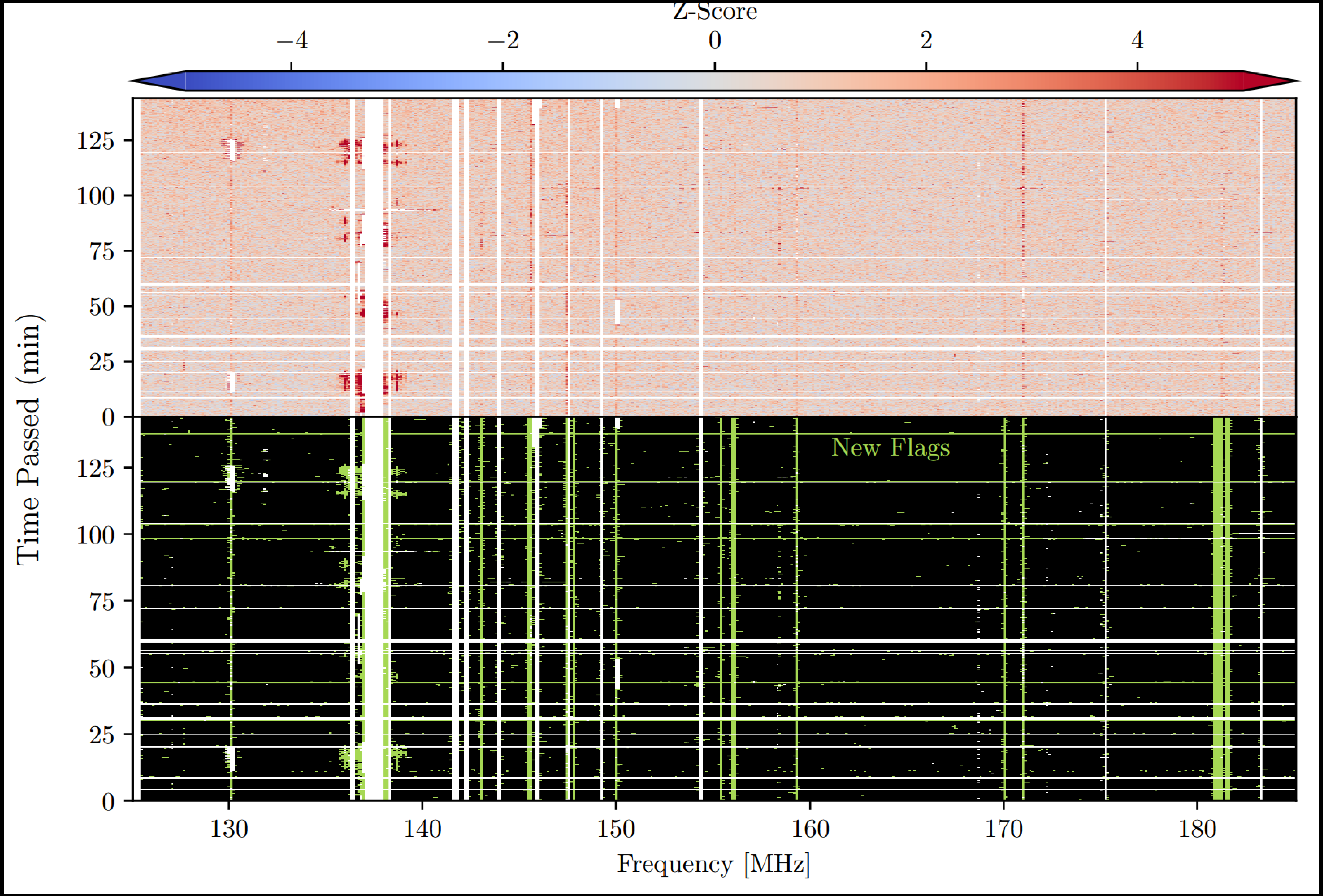
$$P_{total} = P_{21} + \mathbf{W}_{\mathbf{inst}} * P_{FG} + \text{Noise}$$



SPECIFIC CHALLENGES

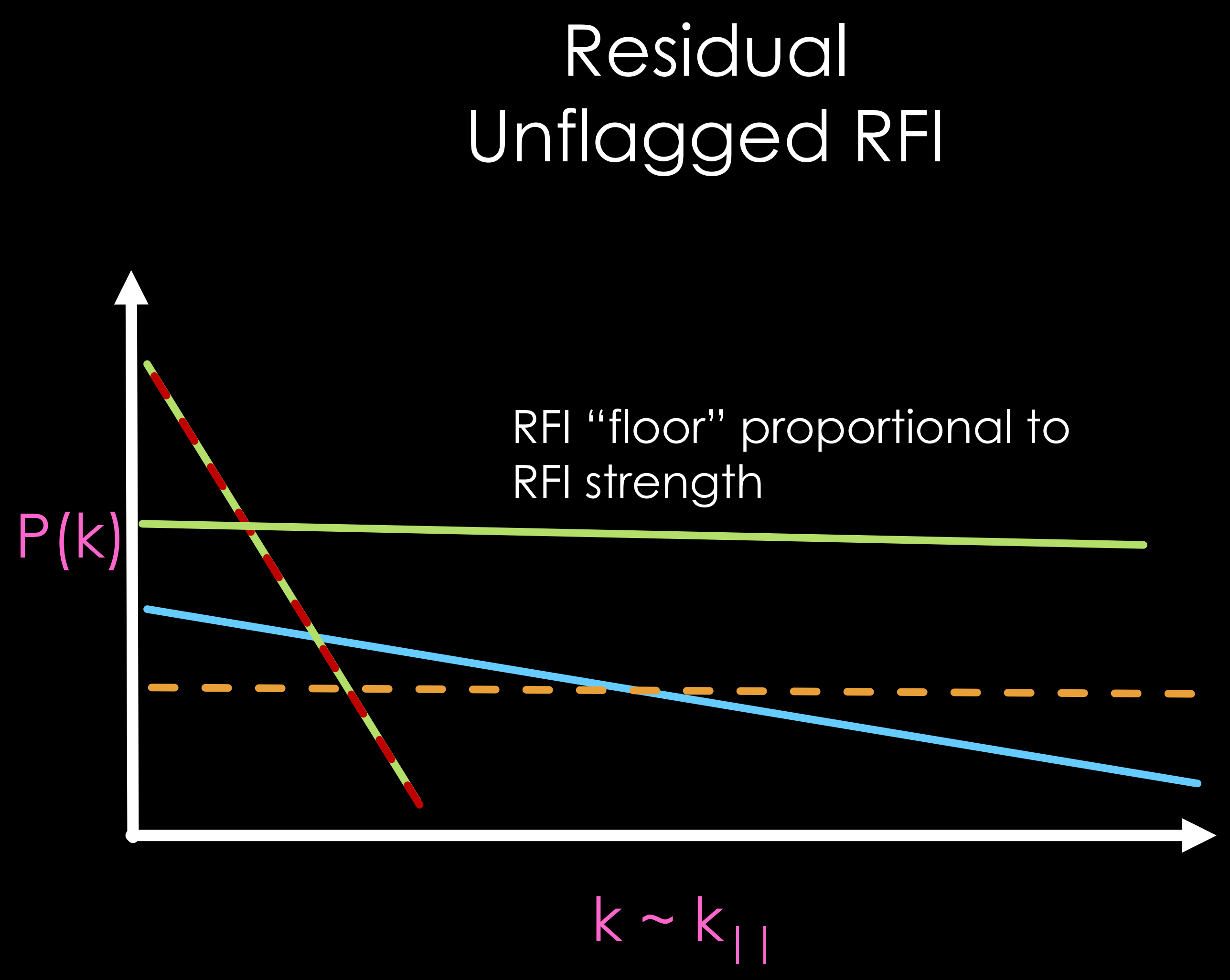
#1 – RFI & MISSING DATA

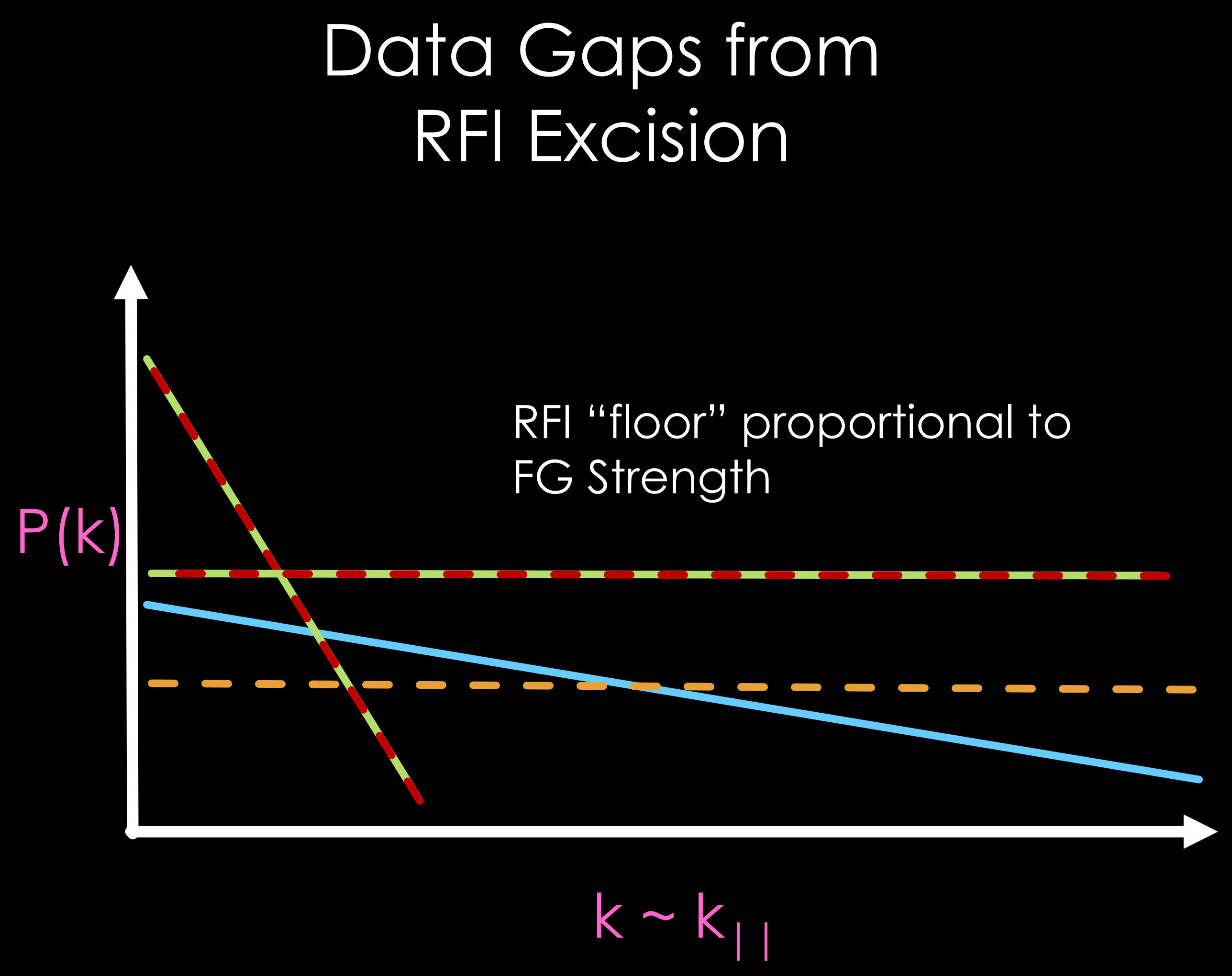
RFI IN HERA DATA



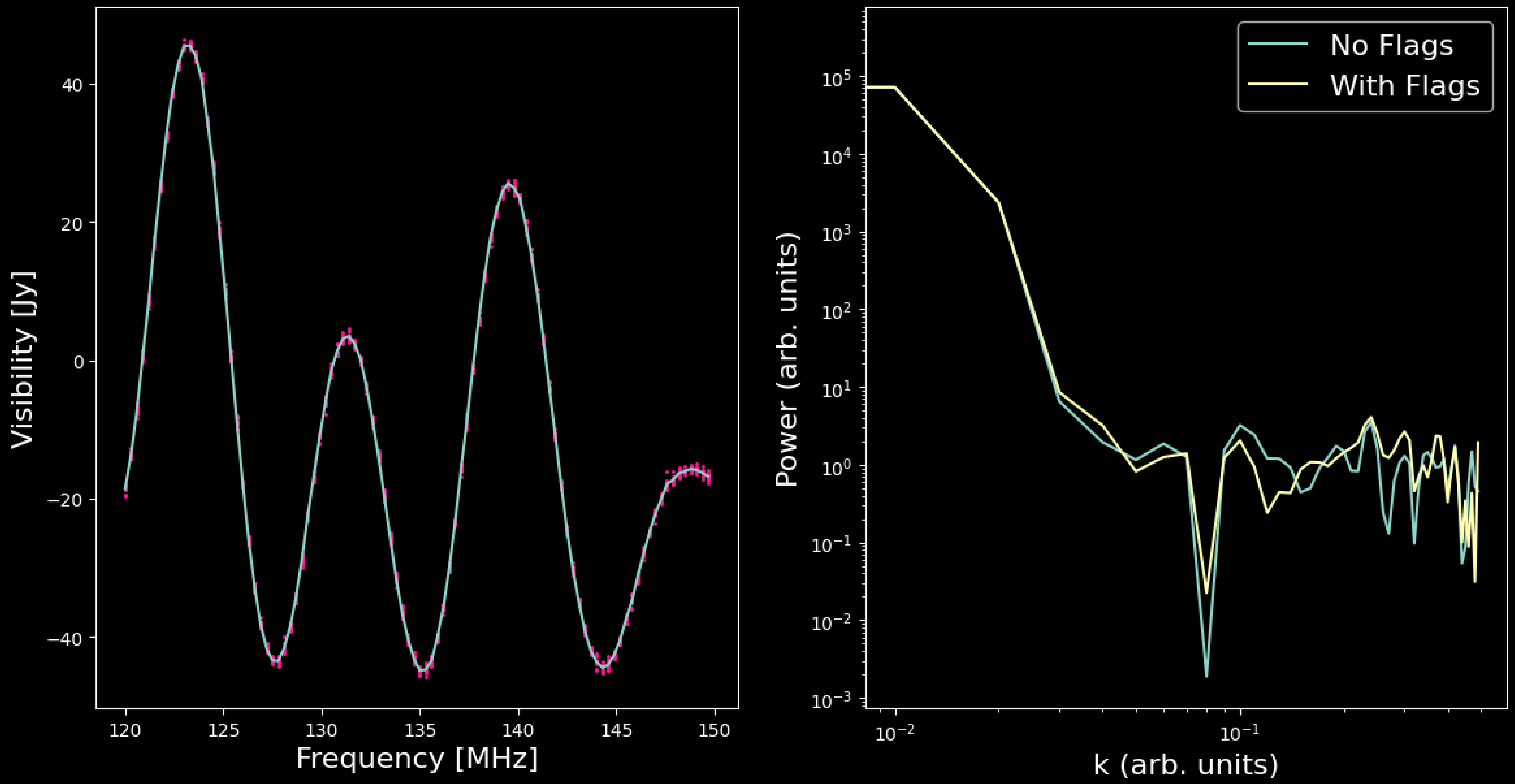
HERA Collaboration, 2025 (in prep)

EFFECT OF RFI

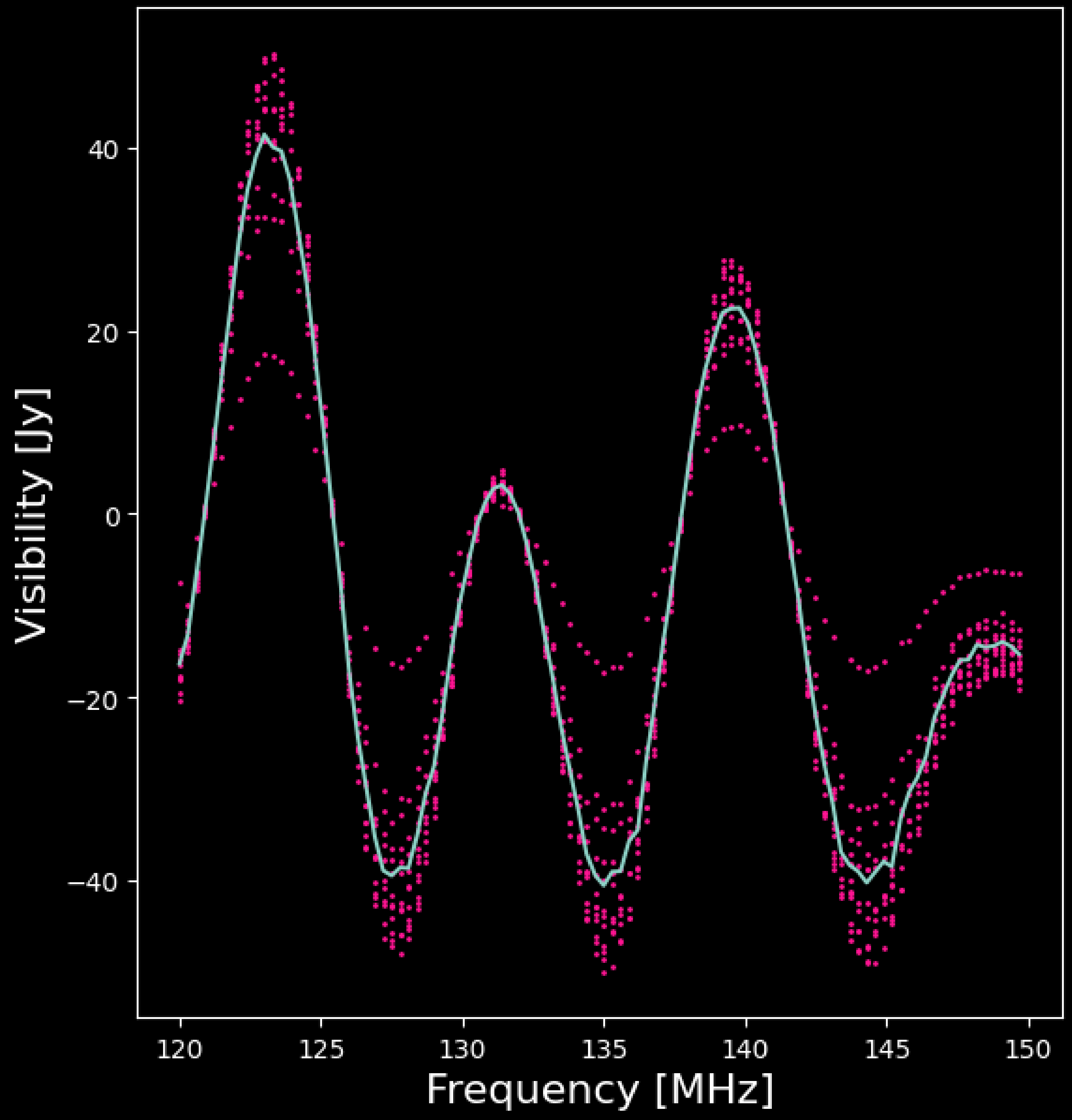


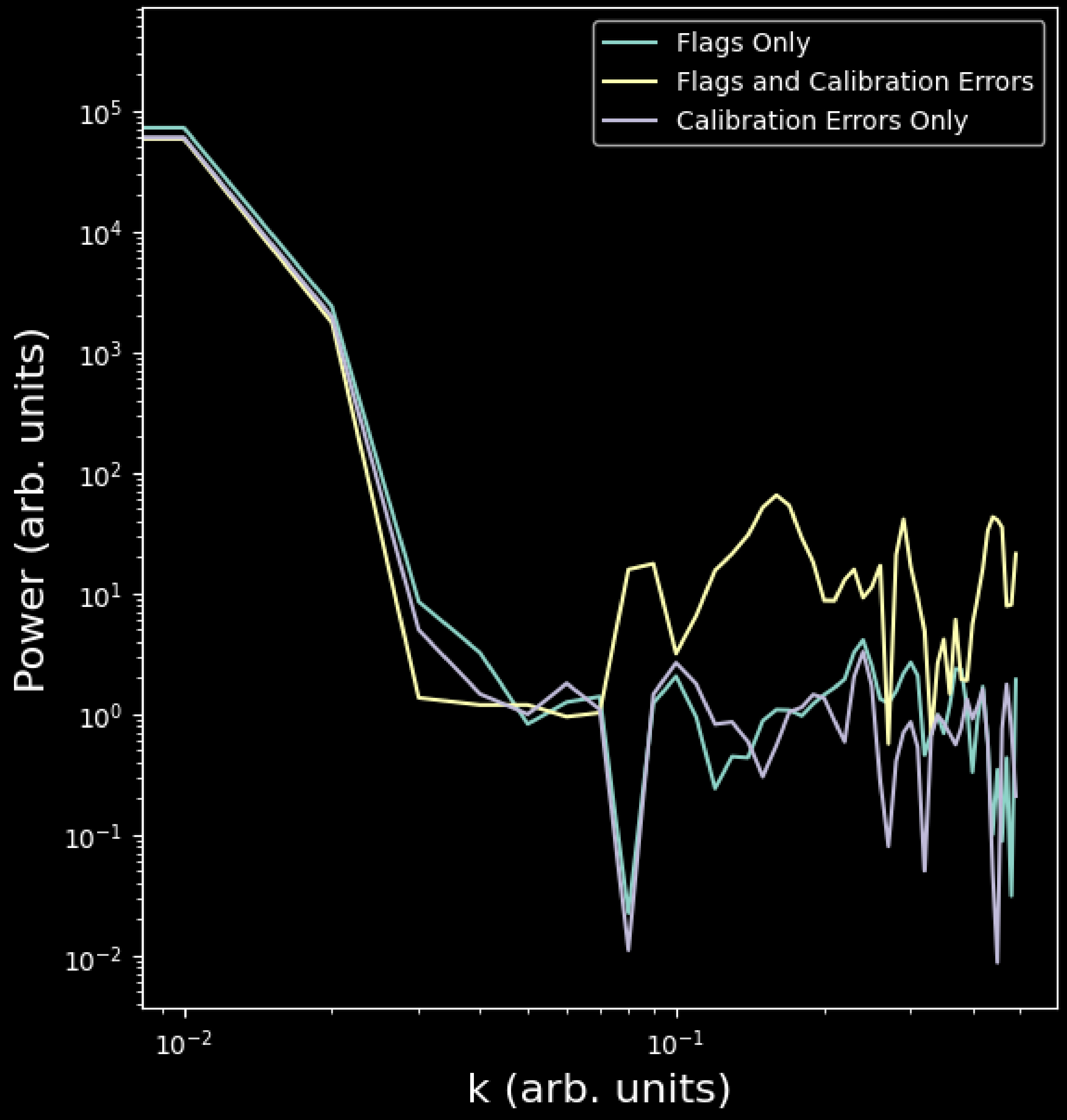


AVERAGING THEN IN-PAINTING

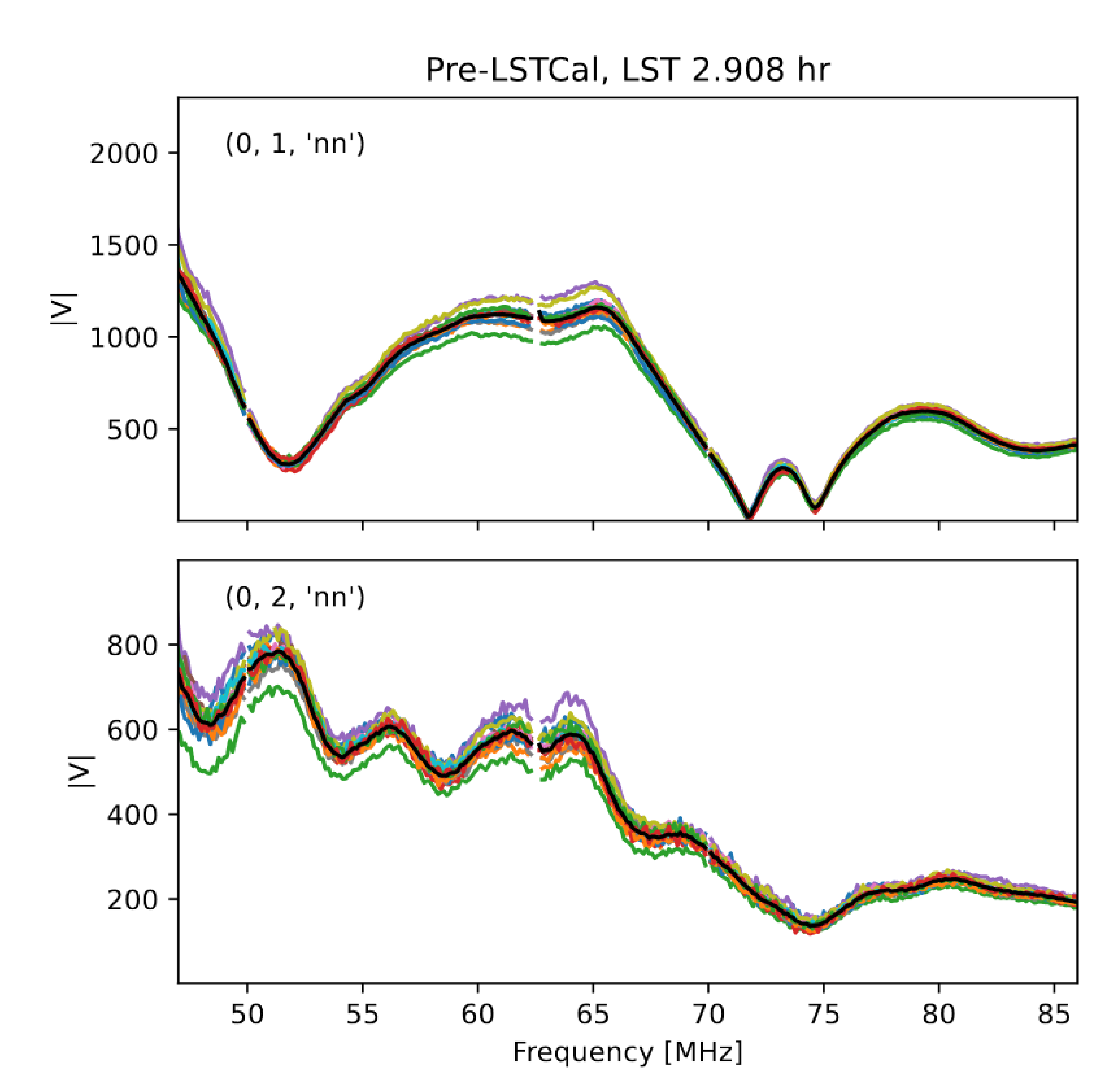


AVERAGING THEN IN-PAINTING



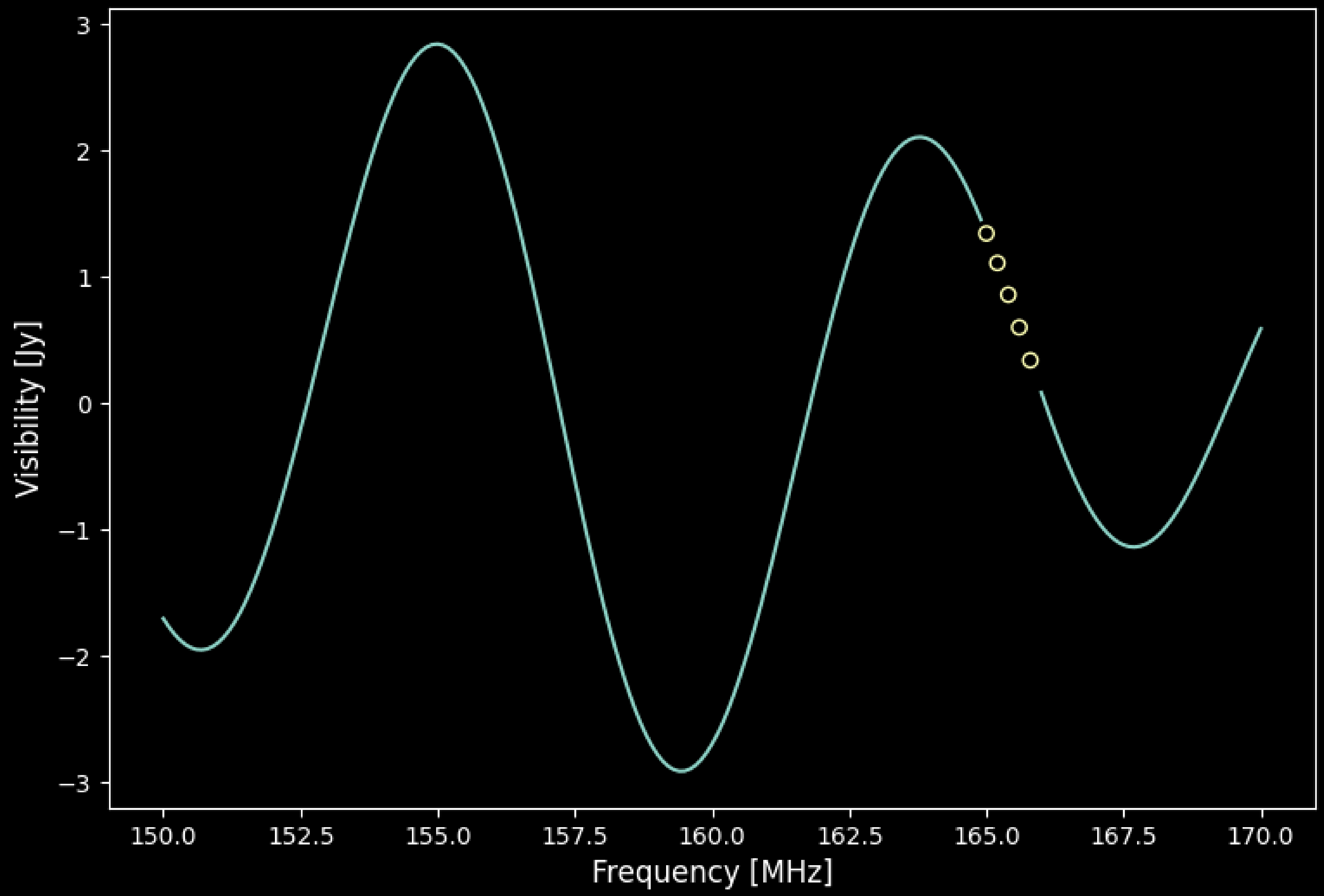


AVERAGING THEN IN-PAINTING



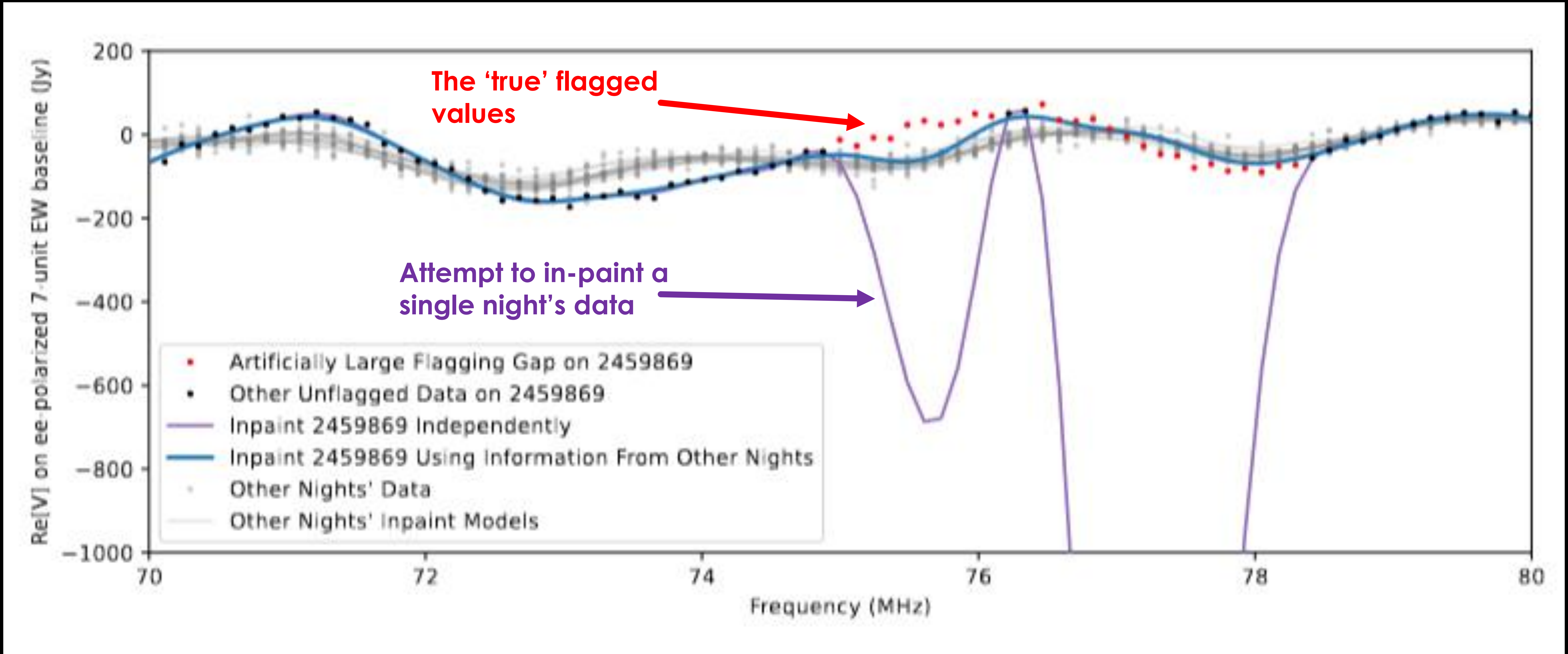
PER-NIGHT IN-PAINTING

"IN-PAINTING" INSERTS A VALUE INTO THE GAP **ONLY** TO MAINTAIN SMOOTHNESS



PER-NIGHT IN-PAINTING

… But It's Easy to get it very wrong…

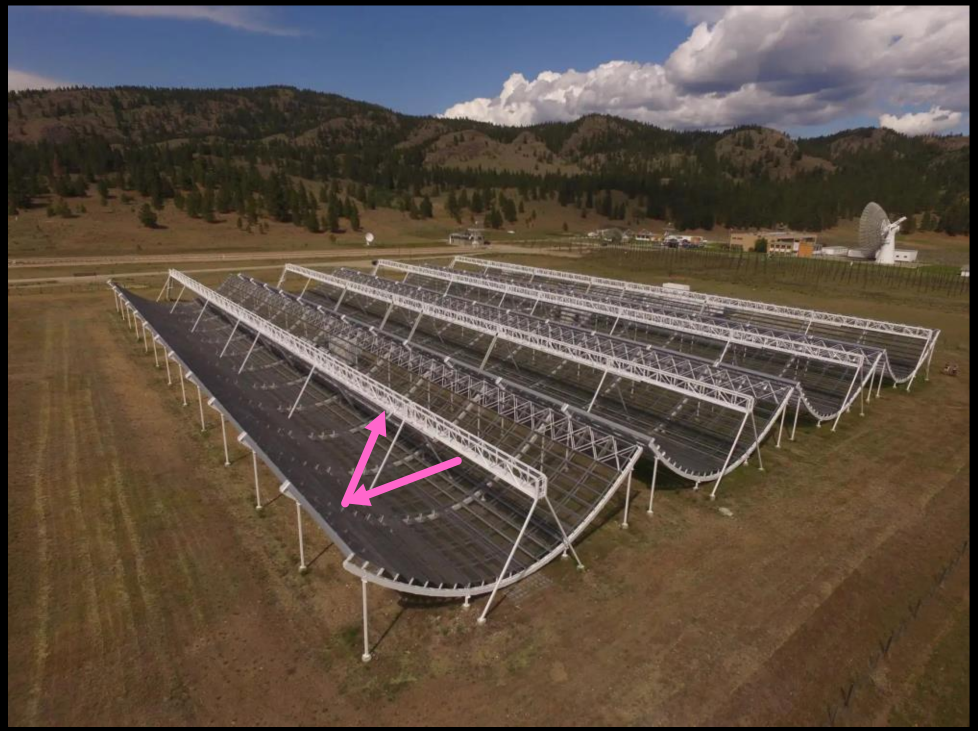


See Khandakar's talk later today!

#2: MUTUAL COUPLING

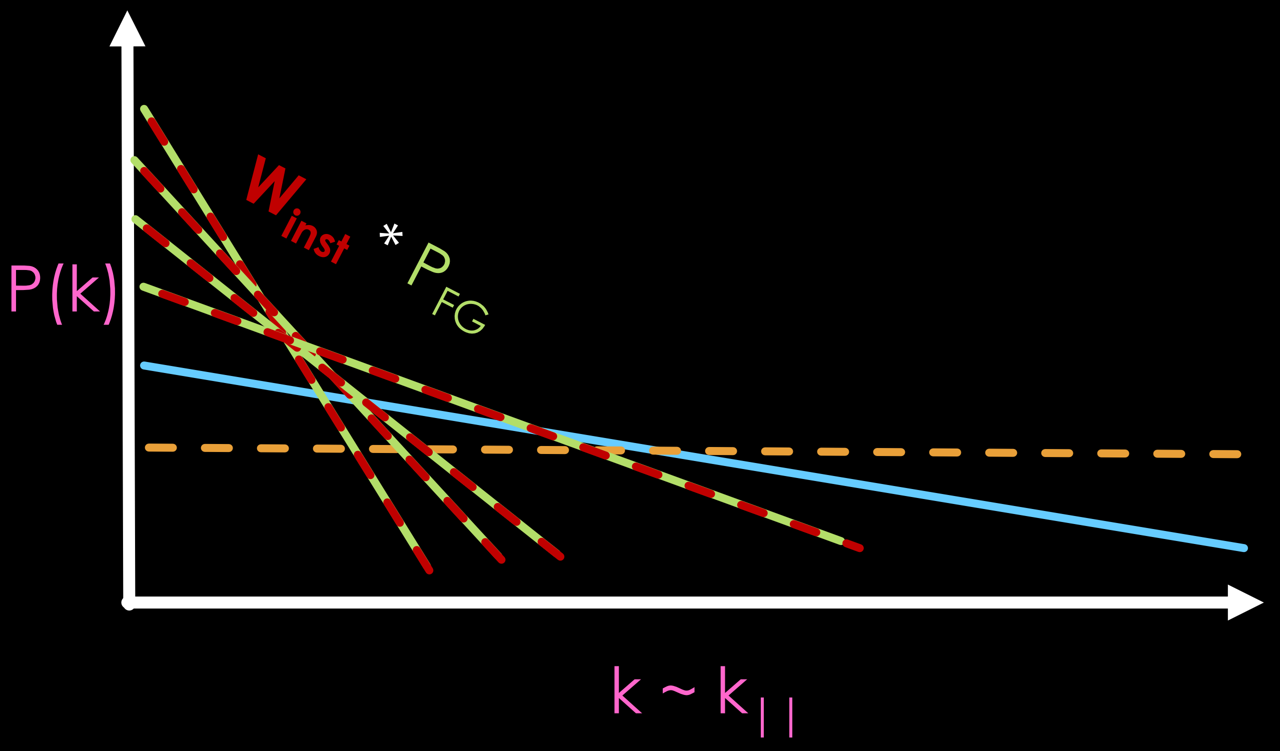
WHAT IS MUTUAL COUPLING?

WHEN SKY SIGNAL ENTERS ONE ANTENNA/FEED & **RE-RADIATES** INTO OTHER ANTENNAS.

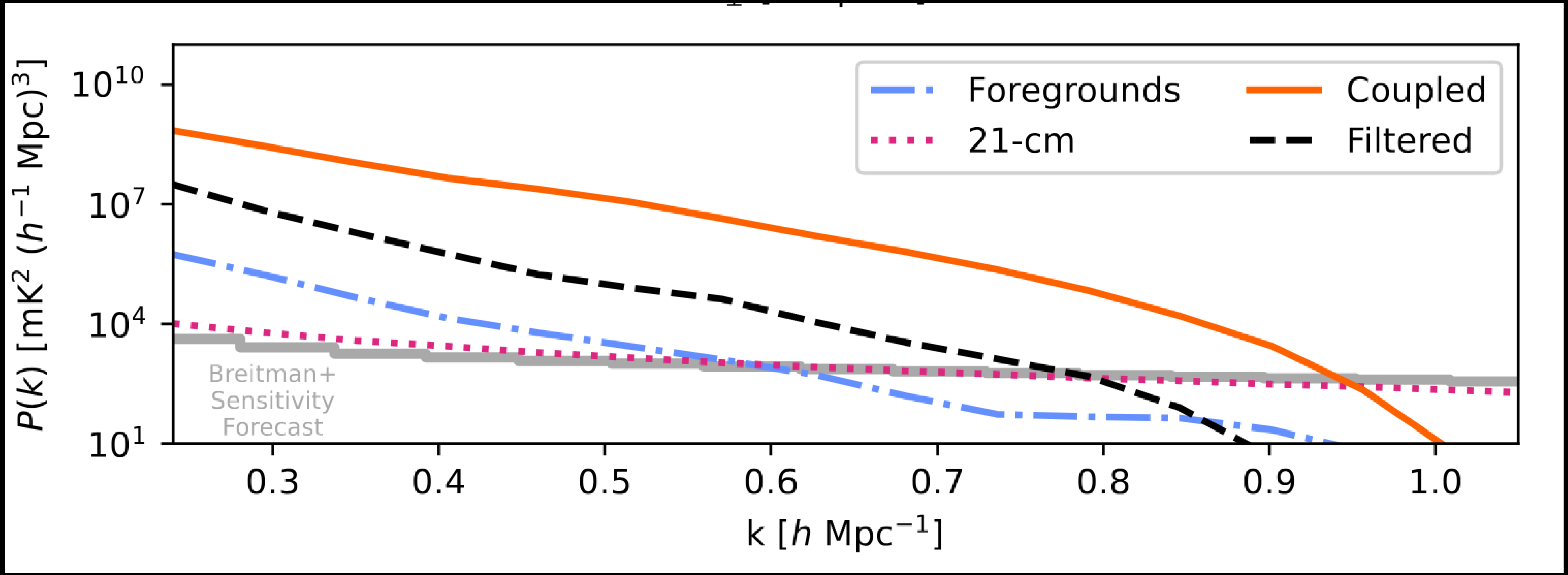


WHAT IS MUTUAL COUPLING?

WHEN SKY SIGNAL ENTERS ONE ANTENNA/FEED & **RE-RADIATES** INTO OTHER ANTENNAS.



WHAT IS MUTUAL COUPLING?



Rath, Pascua+2025 (incl. **SGM**)

MODELING MUTUAL COUPLING

FIRST-ORDER SEMI-ANALYTIC MODEL DEVELOPED BY:

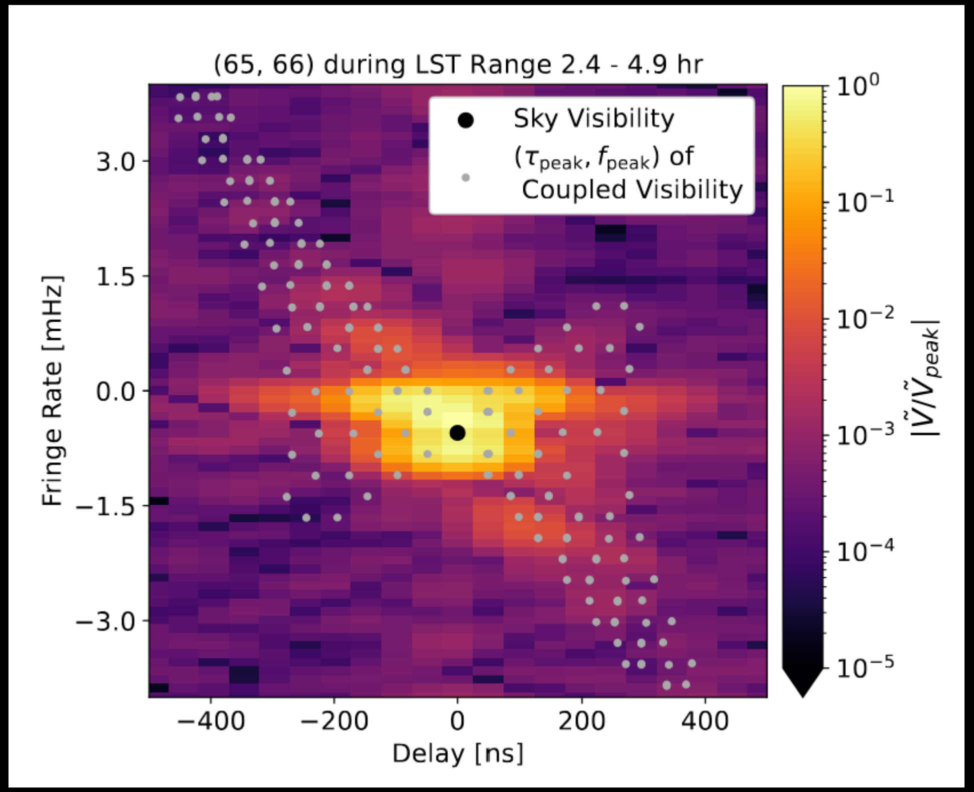
KERN+2019
JOSAITIS+2021
RATH+2024

"Virtual" baselines formed with each antenna

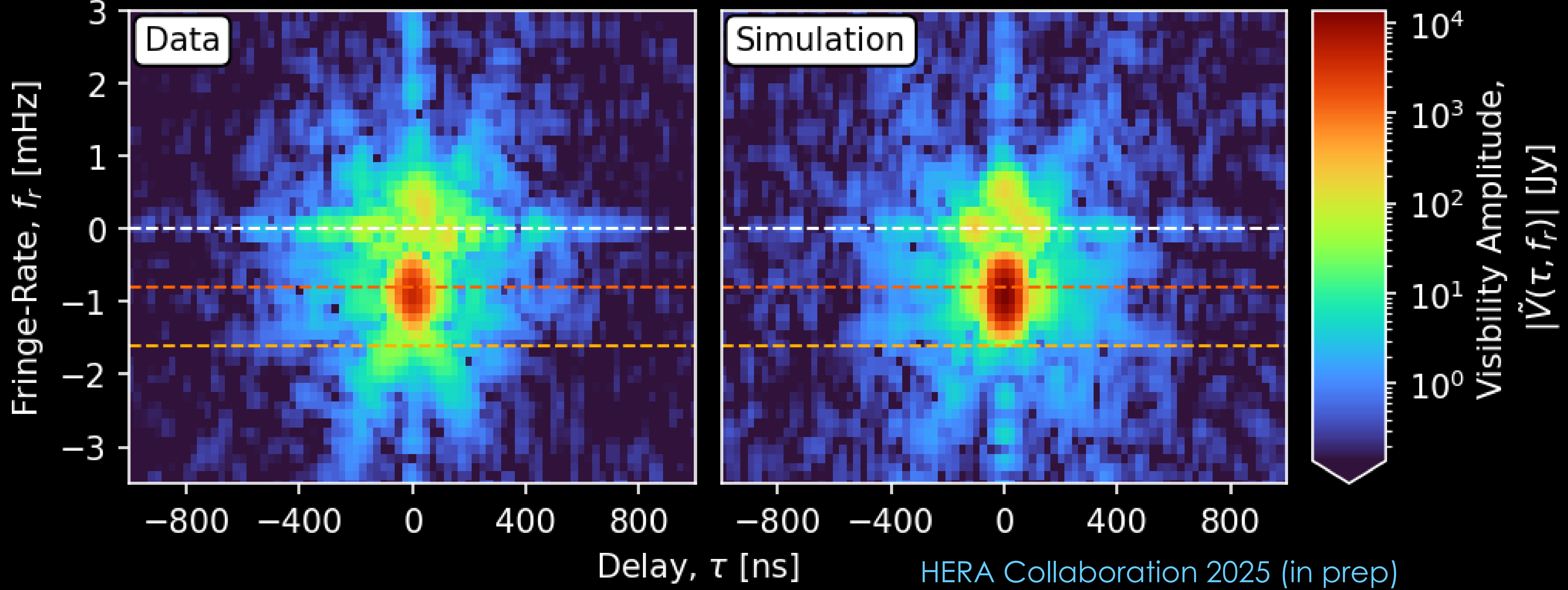
$$\mathbf{V}_{ij}^{(1)} = \mathbf{V}_{ij}^{(0)} + \sum_k \left(\mathbf{V}_{ik}^{(0)}\mathbf{X}_{jk}^{\dagger} + \mathbf{X}_{ik}\mathbf{V}_{kj}^{(0)}\right).$$

Un-coupled

Coupling Coefficients based on geometry and antenna properties



Rath+2024



HERA Collaboration 2025 (in prep)

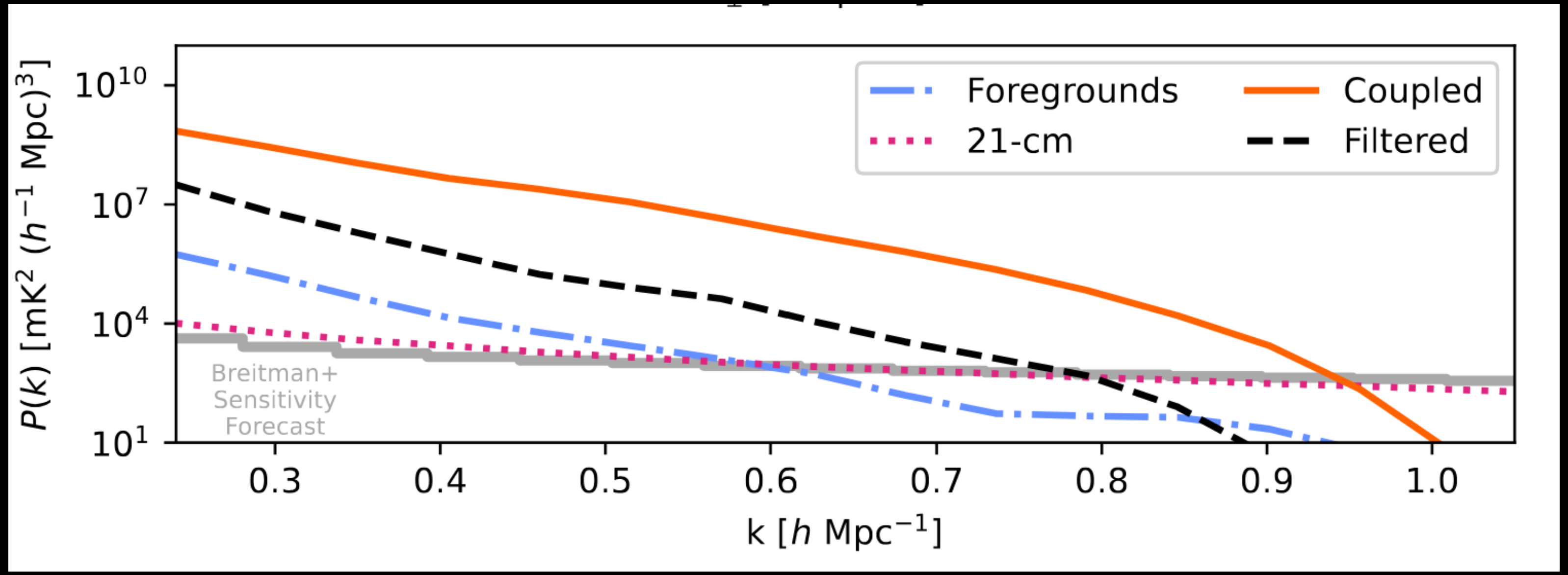
MITIGATING MUTUAL COUPLING

"Virtual" baselines formed with each antenna

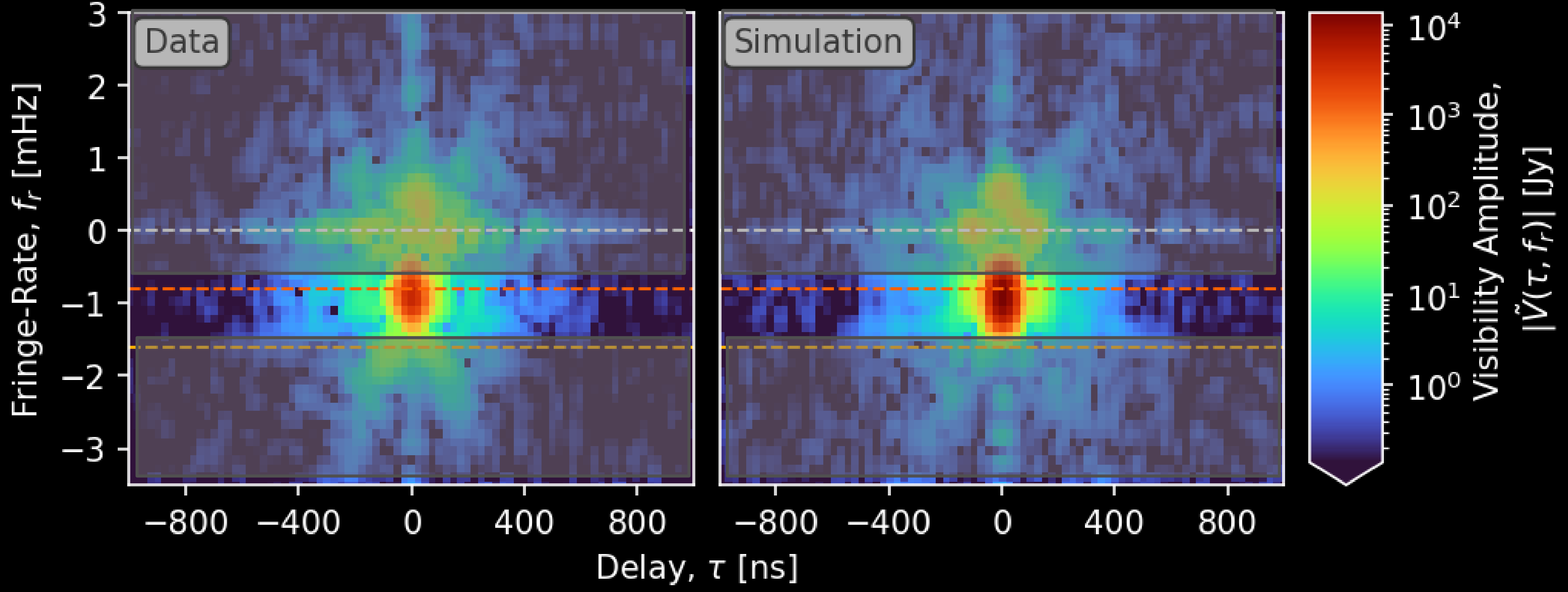
$$\mathbf{V}_{ij}^{(1)} = \mathbf{V}_{ij}^{(0)} + \sum_k \left(\mathbf{V}_{ik}^{(0)} \mathbf{X}_{jk}^{\dagger} + \mathbf{X}_{ik} \mathbf{V}_{kj}^{(0)} \right).$$

Un-coupled

Coupling Coefficients based on geometry and antenna properties



Rath, Pascua+2025 (incl. **SGM**)



HERA Collaboration 2025 (in prep)

FRINGE-RATE FILTERING DEVELOPED BY:
PARSONS+2016
PASCUA+2025 (SIGNAL LOSS)

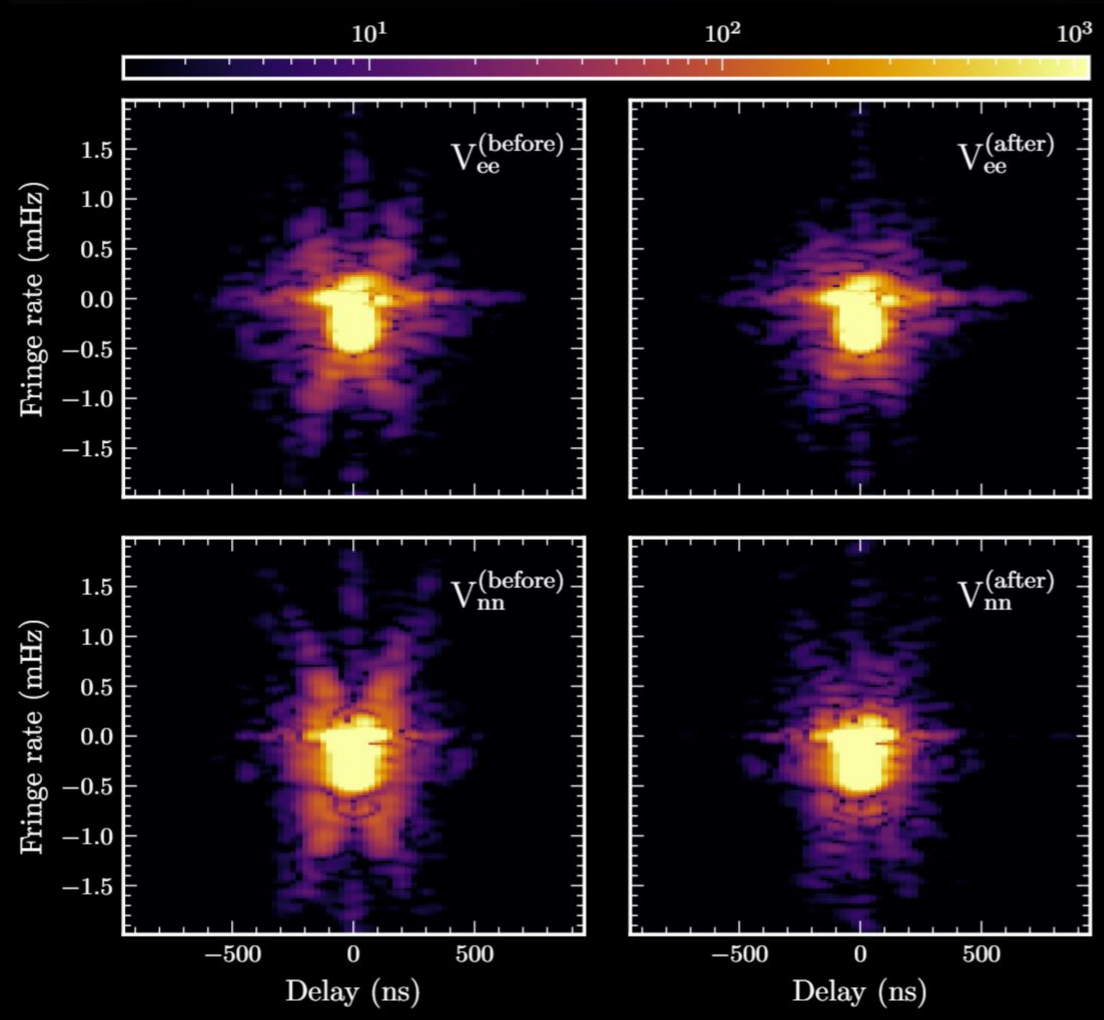
MITIGATING MUTUAL COUPLING

"Virtual" baselines formed with each antenna

$$\mathbf{V}_{ij}^{(1)} = \mathbf{V}_{ij}^{(0)} + \sum_k \left(\mathbf{V}_{ik}^{(0)} \mathbf{X}_{jk}^{\dagger} + \mathbf{X}_{ik} \mathbf{V}_{kj}^{(0)} \right).$$

Un-coupled

Coupling Coefficients based on geometry and antenna properties



(Cox+ in prep)

More direct inversion requires determining **X** from *data*.

Some different approaches:

- Model **X** based on geometry + smooth parametric model of horizon-sensitivity. Fit to data under assumptions about uncoupled power at high fringe rate and/or delay. (Rath+ in prep, Hewitt+ in prep)
- Estimate every element of **X** directly (high-dimensional inference) assuming constancy over time. (Cox+ in prep)

#3: PROPAGATING SYSTEMATICS TO PARAMETER INFERENCE

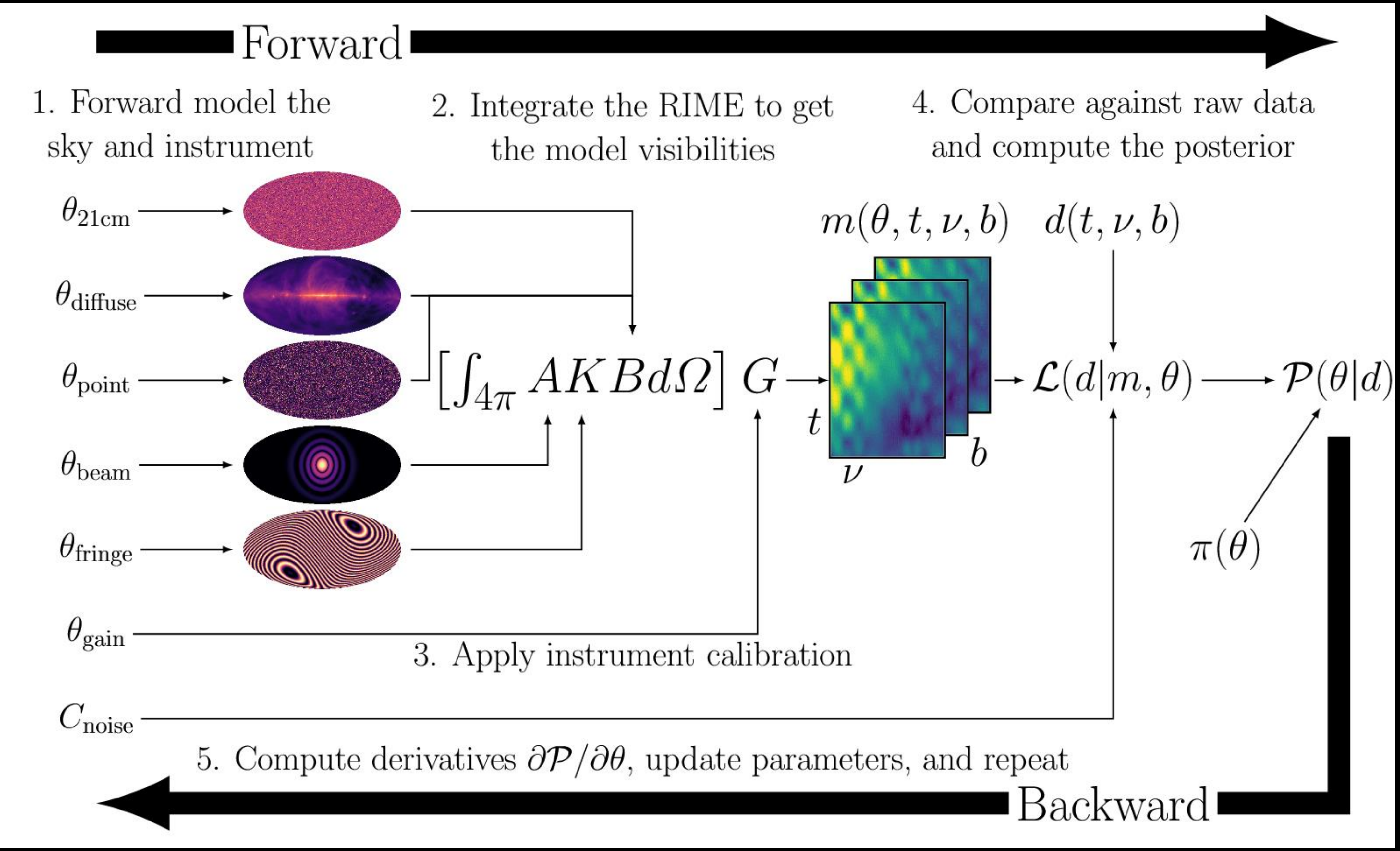
NO ESCAPE FROM UPPER LIMITS

NO DETECTION CAN BE MADE UNLESS WE CAN AGREE THAT RESIDUAL SYSTEMATICS ARE SUB-DOMINANT.

THIS REQUIRES A CONSISTENT *MODEL* OF RESIDUAL SYSTEMATICS PROPAGATED THROUGH TO THE POWER SPECTRUM.

PROPAGATING SYSTEMATICS: BAYESIAN FORWARD MODELS

- FULL FORWARD MODELS ARE *HARD*.
- SOME INITIAL EFFORTS THAT CONDITION ON CERTAIN COMPONENTS:
  - **BAYESEOR:** (SIMS+2019, BURBA+2023)
  - **GIBBS SAMPLING:** (KENNEDY+2023, GLASSCOCK+2024, WILENSKY+2024)
- NEW 'FULL' END-TO-END HIGH-DIMENSIONAL FRAMEWORK FROM KERN 2019.
  - AUTO-DIFFERENTIABLE
  - NO SUPPORT FOR MUTUAL COUPLING (YET!)

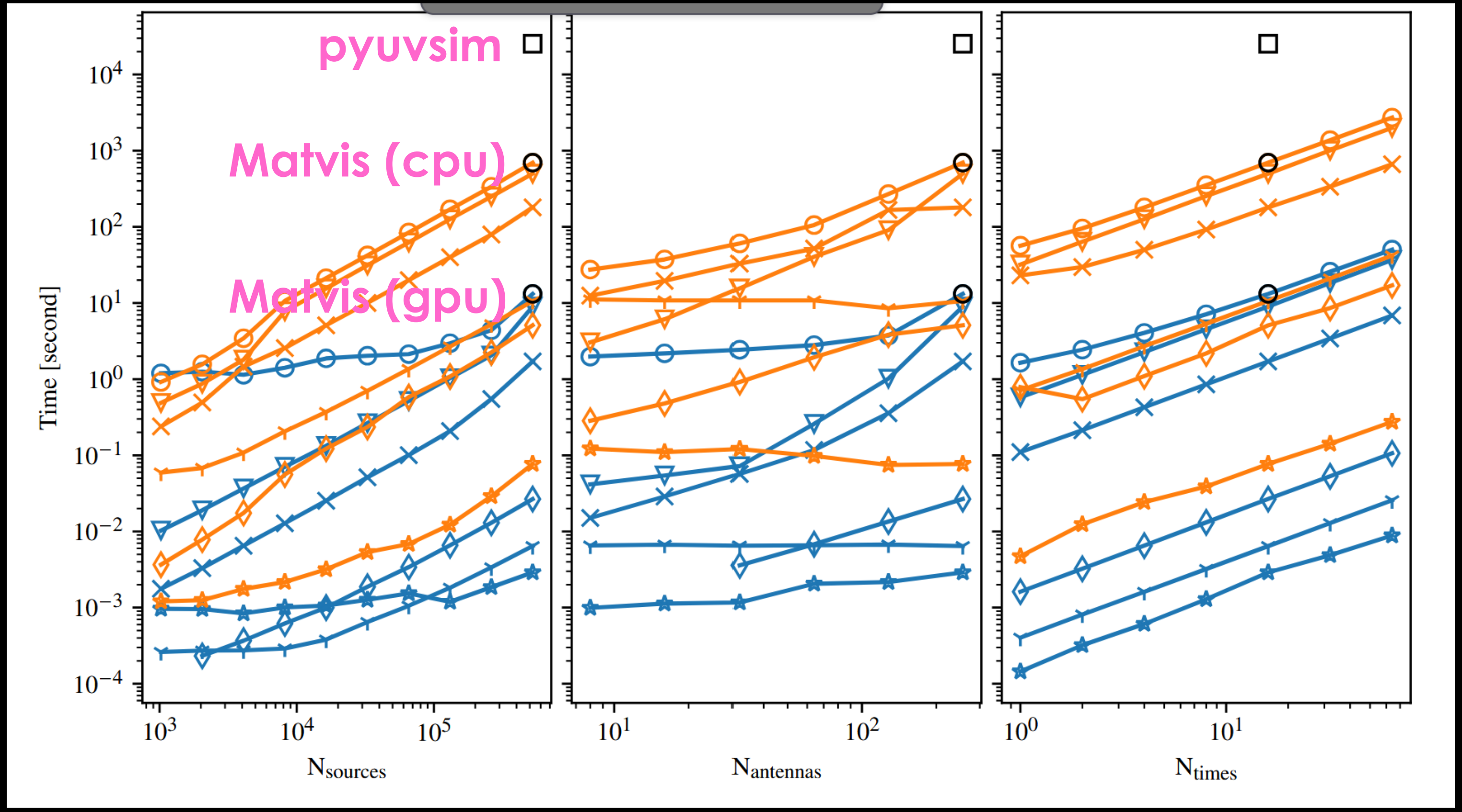


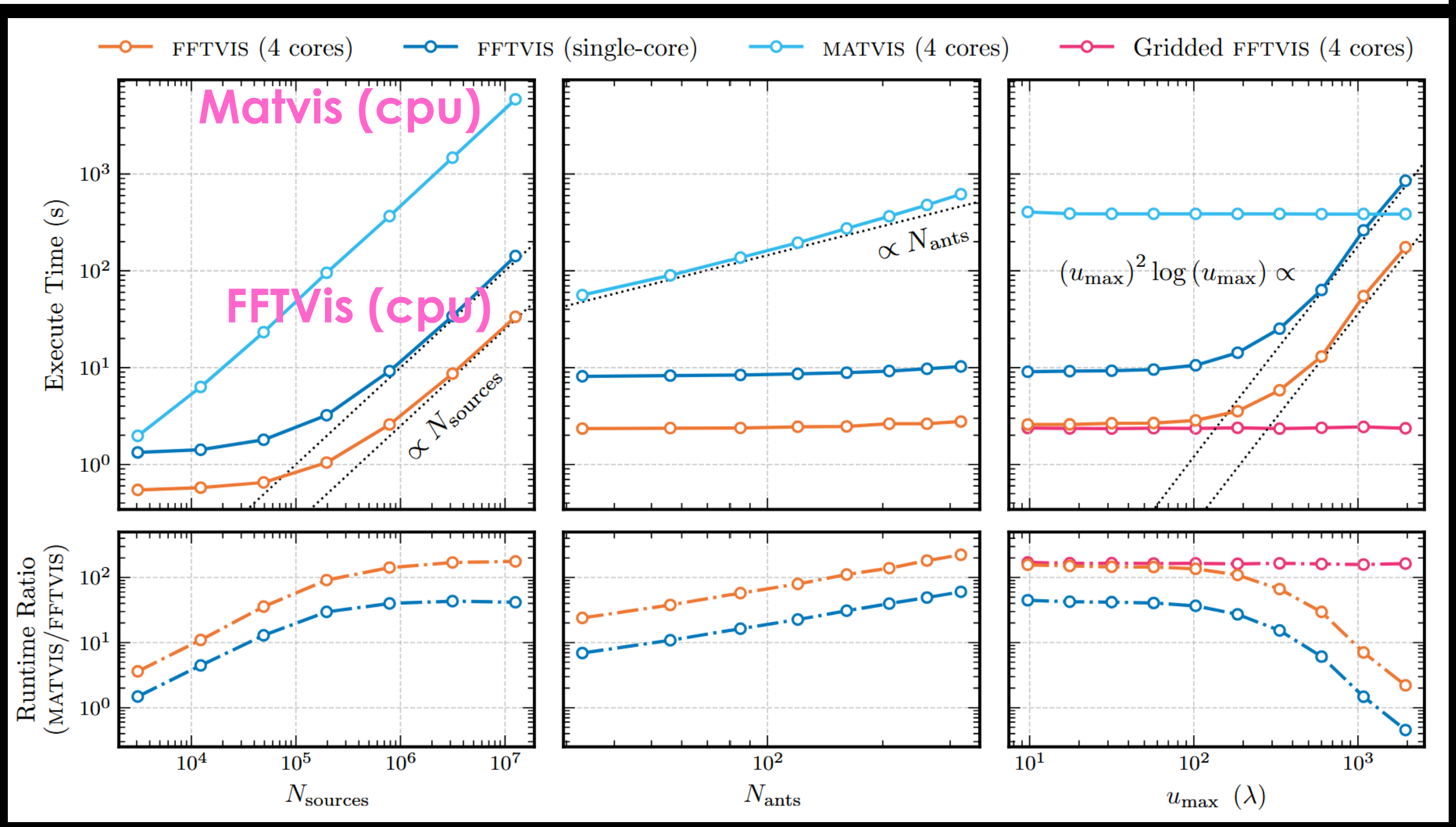
Kern+2025

#4: DETAILED SIMULATIONS FOR VALIDATION

NEW INSTRUMENT SIMULATORS

(Kittiwisit, murray+2025)



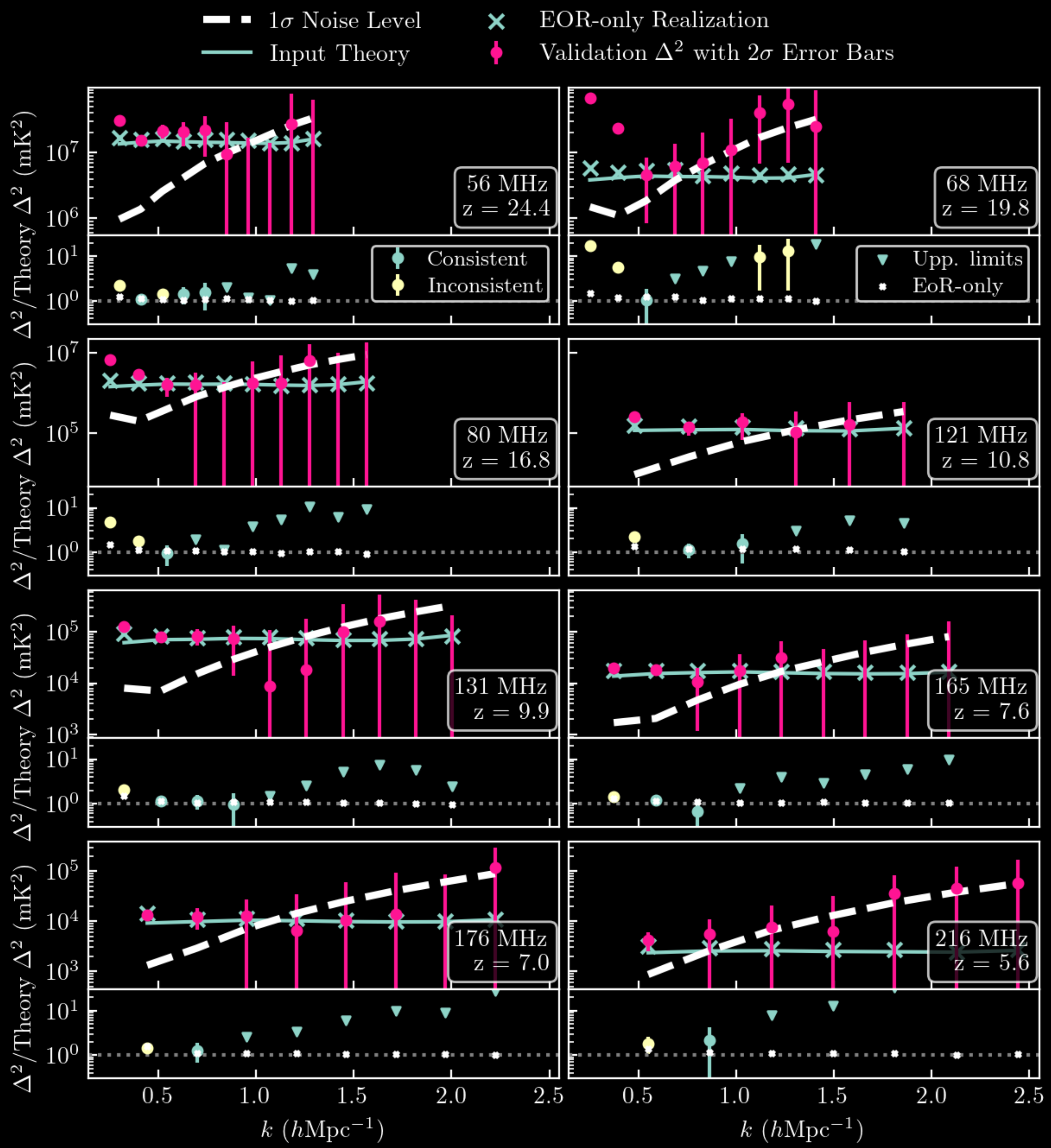


(Cox, Murray+2025)

- New instrument simulators are abundant, from slow-and-accurate, to fast-and-bespoke:
  - Pyuvsim (lanman+2019)
  - Woden (Line+2022)
  - Oskar
  - Matvis (Kittiwisit, murray+2025)
  - FFTVis (Cox, Murray+2025)

END-TO-END VALIDATION SIMULAITONS

- ENABLE DETAILED VALIDATION TESTS OF ANALYSIS PIPELINES
- BECOMING MORE ACHIEVABLE WITH FASTER SIMULATORS
- LATEST HERA ANALYSIS SUPPORTED BY HUGE MULTI-TB MOCK DATA SIM.



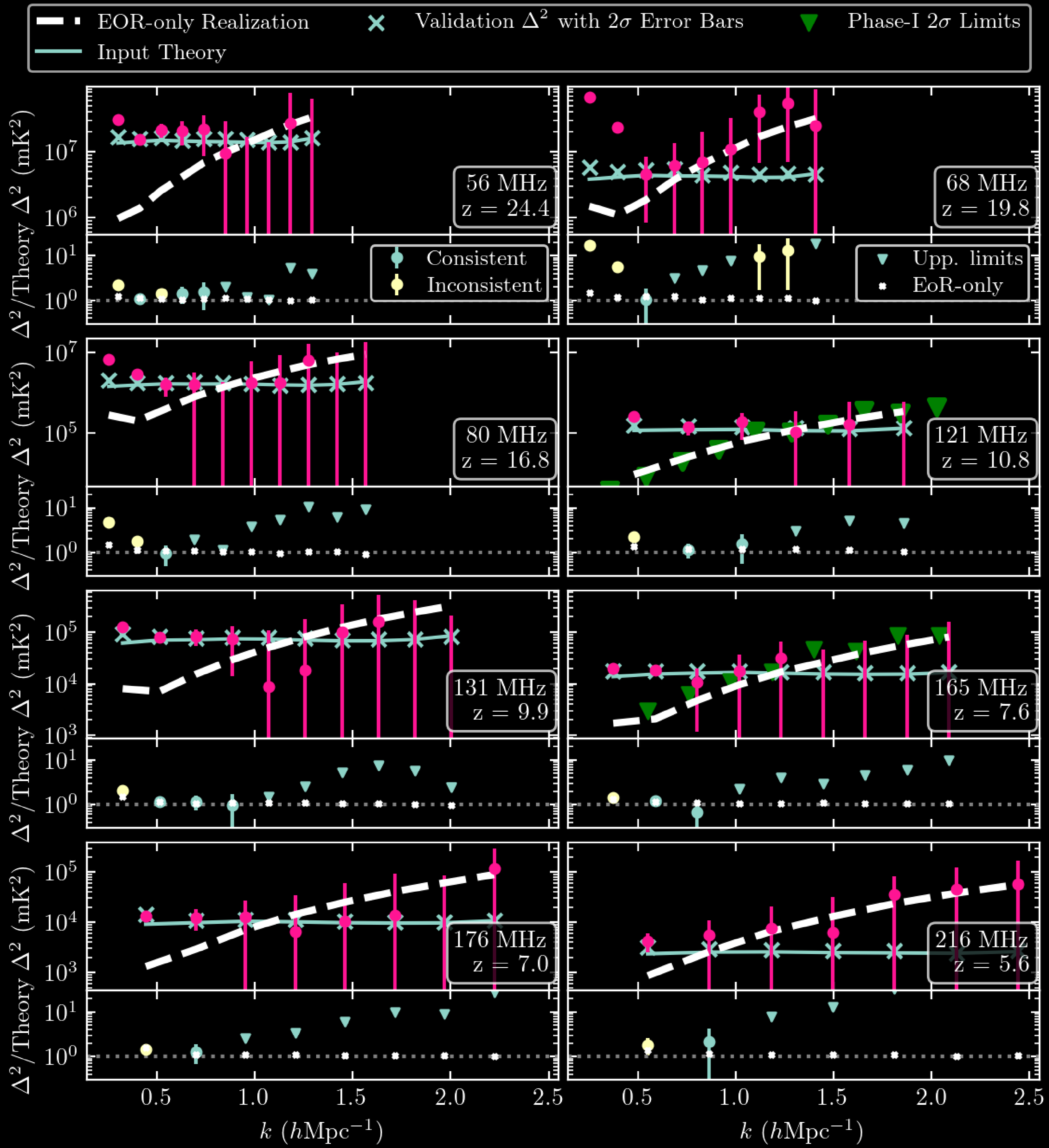
See Shabbir's talk later today!

SUMMARY

- 21cm Data analysis challenges focused on **preserving spectral smoothness at all costs.**
- A la carte menu of specific challenges:
  - RFI and Missing Data
  - Mutual Coupling
  - Propagation of systematics
  - Instrument simulation

BONUS: HERA SNEAK-PEEK

- **Phase II** instrument covers broader frequency range (redshifts ~5.6 – 25)
- **>300 nights** of data taken. **14 nights** in upcoming limits.
- ~140 antennas (compared to ~30 in Phase I)
- **Redundant Averaging** (enhanced SNR)
- Similar sensitivity to previous release with 14 nights instead of 100.
- **Noise-limited** at k>~0.6 h/Mpc.
- **Mutual coupling-dominated** from k=0.3-0.6 h/Mpc



HERA Collaboration 2025 (in prep) **PRELIMINARY**

BONUS SLIDES

NEXT-GEN VALIDATION SIMULATIONS: **MATVIS**

MATVIS USES A 'TRICK' TO TURN THE VISIBILITY INTEGRAL INTO A SINGLE MATRIX-MATRIX PRODUCT.

THIS IS FAST ON BOTH CPU AND GPU.

$$V_{ij}^{pq}(\nu_a,t) = \sum_{nkk'} \mathbf{A}_i^{pk}(\nu_a,\theta^{(n)}_{\rm hrz}(t)) \cdot \mathbf{A}_j^{qk'*}(\nu_a,\theta^{(n)}_{\rm hrz}(t)) \times \mathbf{C}^{(n)}_{kk'}(\nu_a,\theta^{(n)}_{\rm eq}) \overbrace{\exp\left(-2\pi i \nu_a \tau^{(n)}_{ij}\right)}^{\text{Per-baseline 'fringe' term}}.$$

Per-antenna 'fringe' terms

$$\begin{aligned} V_{ij}^{pq}(\nu_a,t) &= \sum_{k,k',n} \mathbf{A}_{ipkn}\mathbf{F}_{in}\mathbf{C}_{kk'n}\mathbf{F}^*_{jn}\mathbf{A}^*_{jqk'n}, \\ &= \sum_{k,k',k''n} \underbrace{\mathbf{A}_{ipkn}\mathbf{F}_{in}\mathbf{M}_{kk''n}}_{\mathbf{Z}}\underbrace{\mathbf{M}^*_{k'k''n}\mathbf{F}^*_{jn}\mathbf{A}^*_{jqk'n}}_{\mathbf{Z}^*}, \end{aligned}$$

$$= \underbrace{\mathbf{Z}\mathbf{Z}^\dagger}_{\text{FAST MATRIX MULTIPLY}},$$

NEXT-GEN VALIDATION SIMULATIONS: **MATVIS**

MATVIS USES A 'TRICK' TO TURN THE VISIBILITY INTEGRAL INTO A SINGLE MATRIX-MATRIX PRODUCT.

THIS IS FAST ON BOTH CPU AND GPU.

Per-baseline 'fringe' term

$$V_{ij}^{pq}(\nu_a,t) = \sum_{nkk'} \mathbf{A}_i^{pk}(\nu_a,\theta^{(n)}_{\rm hrz}(t)) \cdot \mathbf{A}_j^{qk'*}(\nu_a,\theta^{(n)}_{\rm hrz}(t)) \times \mathbf{C}^{(n)}_{kk'}(\nu_a,\theta^{(n)}_{\rm eq})\exp\left(-2\pi i \nu_a \tau^{(n)}_{ij}\right).$$

Per-antenna 'fringe' terms

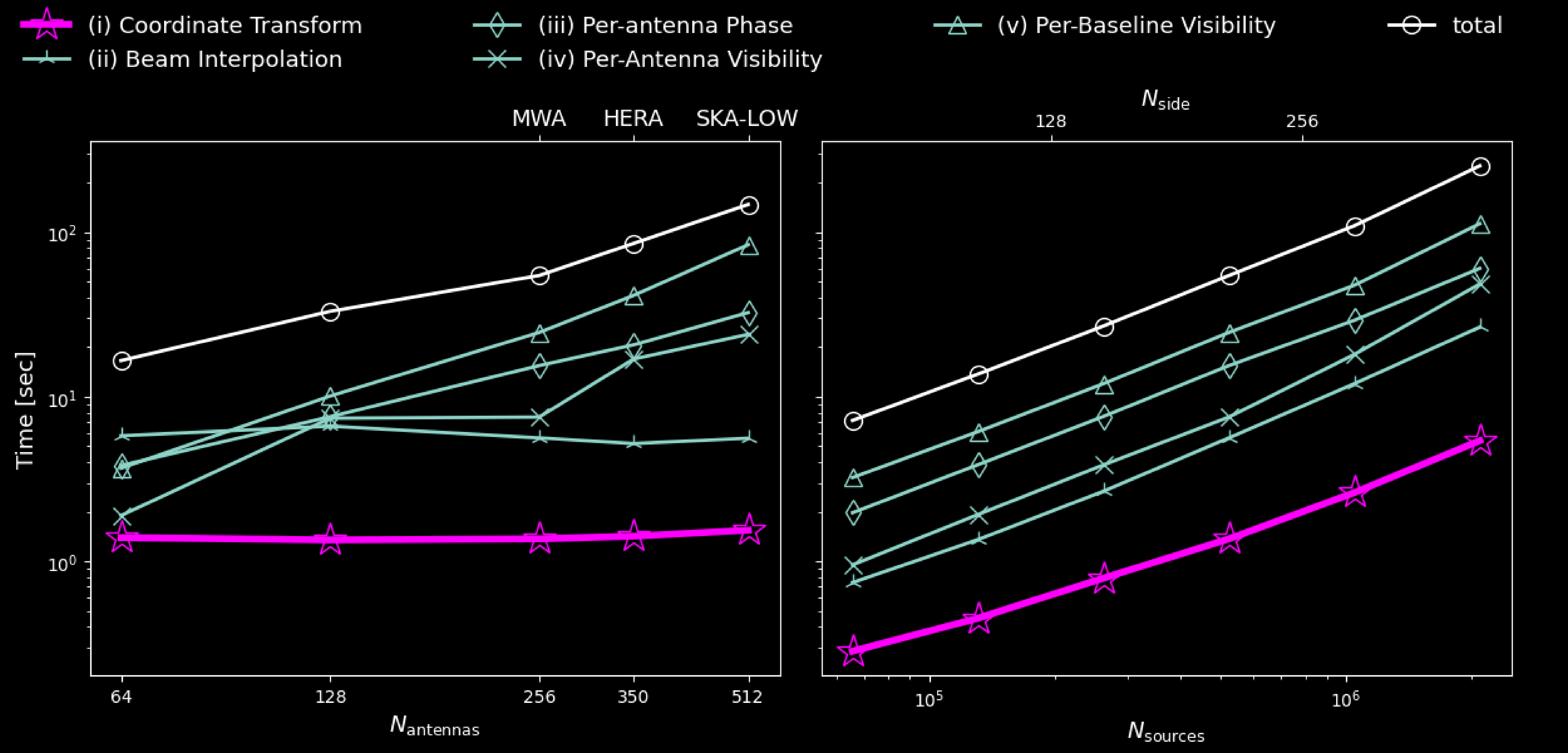
$$\begin{aligned} V_{ij}^{pq}(\nu_a,t) &= \sum_{k,k',n} \mathbf{A}_{ipkn}\mathbf{F}_{in}\mathbf{C}_{kk'n}\mathbf{F}^*_{jn}\mathbf{A}^*_{jqk'n}, \\ &= \sum_{k,k',k''n} \underbrace{\mathbf{A}_{ipkn}\mathbf{F}_{in}\mathbf{M}_{kk''n}}_{\mathbf{Z}}\underbrace{\mathbf{M}^*_{k'k''n}\mathbf{F}^*_{jn}\mathbf{A}^*_{jqk'n}}_{\mathbf{Z}^*}, \end{aligned}$$

$$= \mathbf{Z}\mathbf{Z}^\dagger,$$

FAST MATRIX MULTIPLY

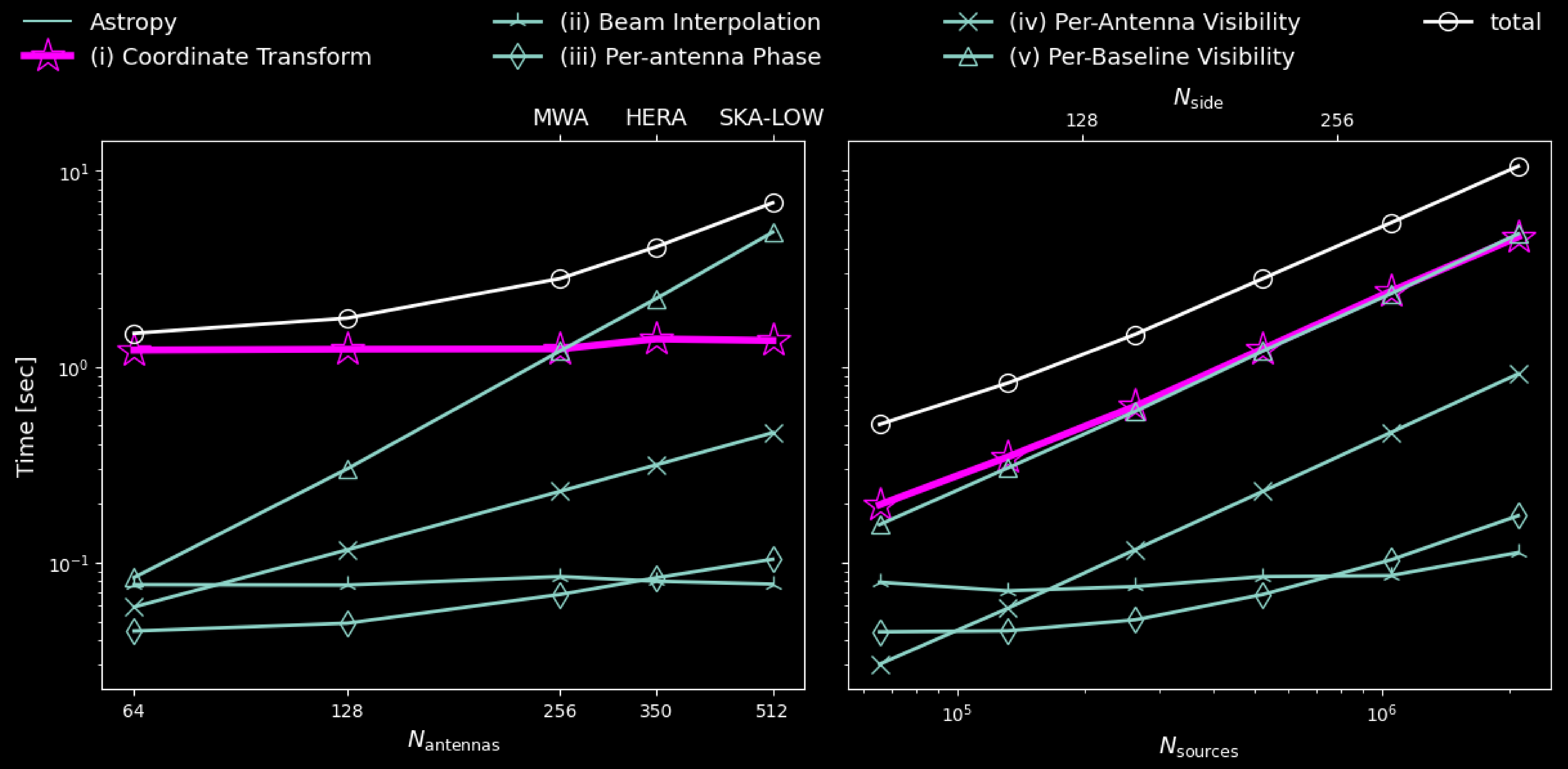
NEXT-GEN VALIDATION SIMULATIONS: **MATVIS**

Coordinate transformations between Ra/Dec and Alt/Az are not a bottleneck on the CPU



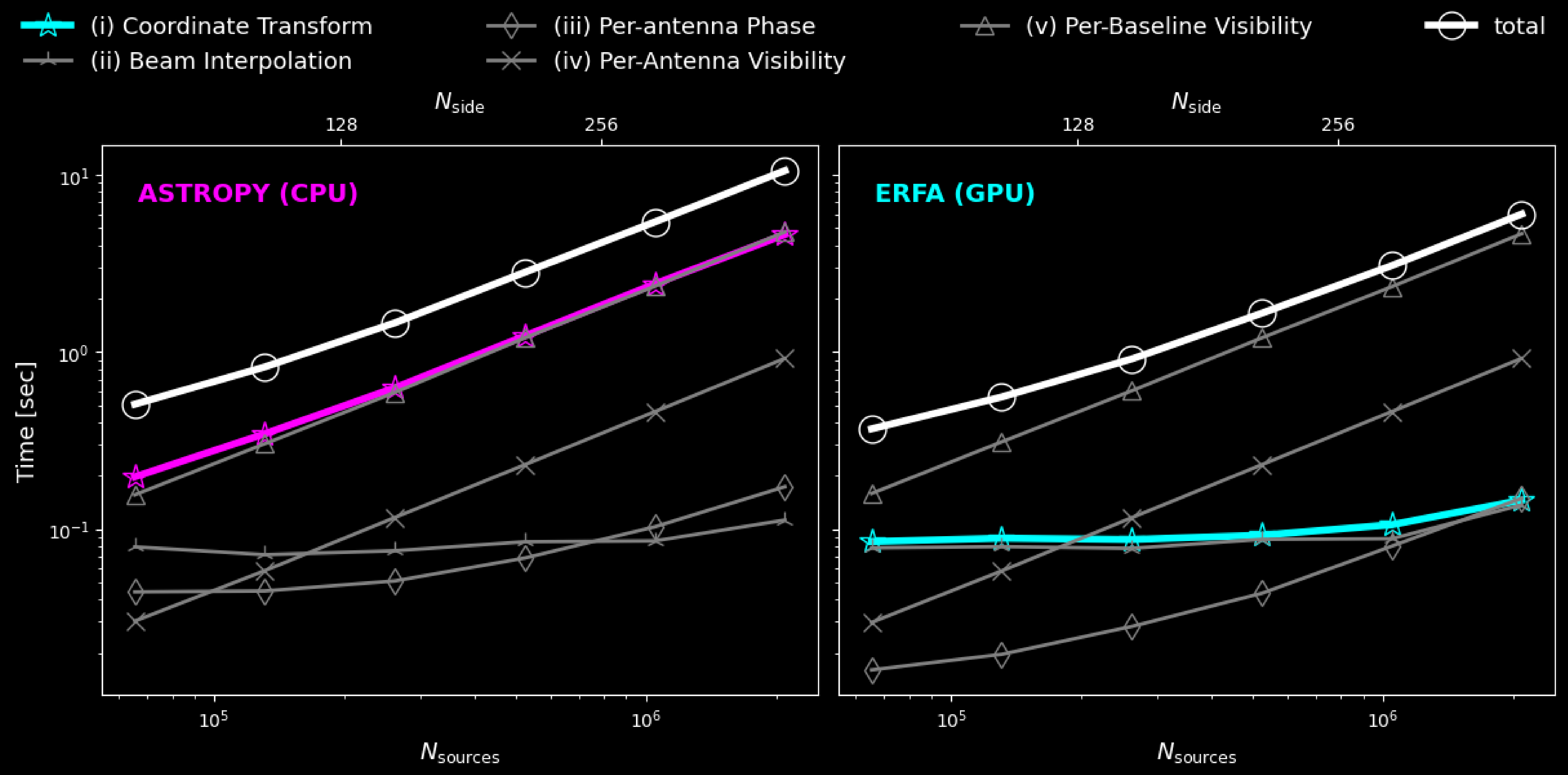
NEXT-GEN VALIDATION SIMULATIONS: **MATVIS**

BUT WHEN THE MATRIX-MULTIPLY IS DONE ON GPU, THEY BECOME NON-NEGLIGIBLE



NEXT-GEN VALIDATION SIMULATIONS: **MATVIS**

So I re-wrote the guts of the astropy coordinate transform on the GPU:



BUT EVEN THIS IS NOT FAST ENOUGH…

NEXT-GEN VALIDATION SIMULATIONS: **FFTVIS**

We've developed a new simulator that is 3x as fast on the **CPU** as matvis on the **GPU**.

It utilizes non-uniform FFT's using **FINUFFT** (**F**lat-**I**ron **N**on-**U**niform **FFT**).

It has the limitation that all beams must be identical.

We need 43k CPU-Hours per simulation!

GPU VERSION IS COMING…

Cox, **SGM+**2024 (in prep)