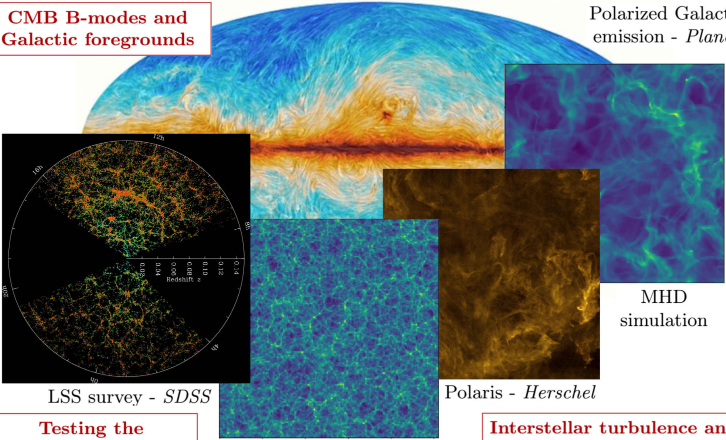
Generative models and component separations for physical fields with Scattering Transforms

Erwan Allys - ENS, Paris
(Physics laboratory and Center for data science)

Journées de l'Action Spécifique Numérique
December 15$^{\text{th}}$ 2025

CMB B-modes and Galactic foregrounds

Polarized Galactic emission - *Planck*



MHD simulation

LSS survey - *SDSS*

Polaris - *Herschel*

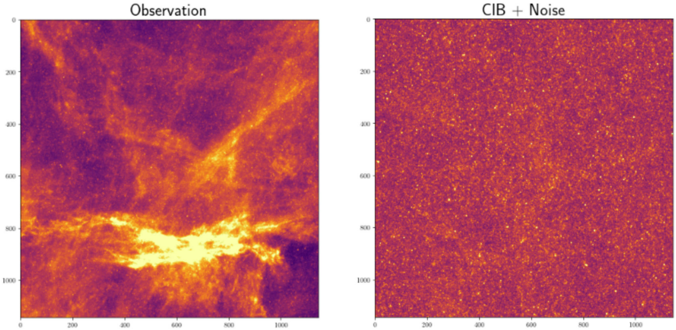
Testing the cosmological model

LSS simulation - *Quijote*

Interstellar turbulence and galactic cycle of matter

→ Common difficulty: non-linearity ⇒ non-Gaussian structures
→ Sometimes no model and limited data regime...
→ How to still leverage these structures for scientific objectives?

Example: CIB/Galactic dust emission



- **Galactic dust emission and Cosmic Infrared Background (CIB)**
  - Thermal dust emission in the interstellar medium
  - Same emission from Milky Way and other galaxies
  - Cosmic background dominates a smaller scales

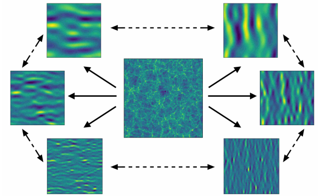
→ Characterization of Galactic dust on those scales?
→ Challenge of low-data regime + lack of prior model

Outline

1. **Scattering Transforms and generative models**
2. **Component separation and modeling from the data**

Scattering transform (ST) statistics

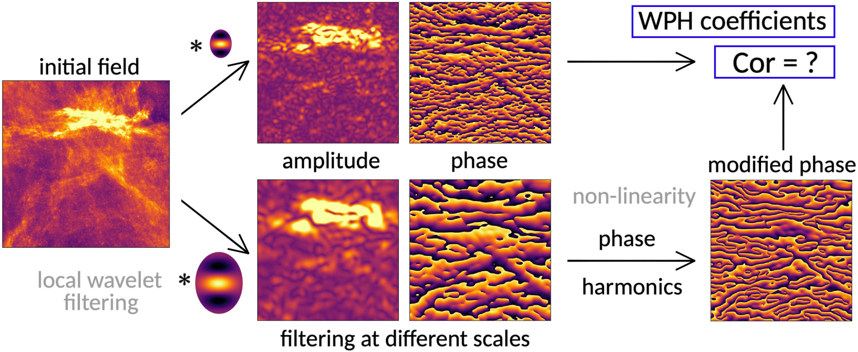
- **Scattering transform statistics** **(Mallat+, 2010+)**
  - Initially developed in data science
    $\rightarrow$ suited for non-linear/non-Gaussian processes
  - Inspired from neural networks
    $\rightarrow$ efficient characterization and reduced variance
  - Do not need any training stage
    $\rightarrow$ explicit mathematical form and interpretability



$\rightarrow$ **Wavelet filters separating the different scales**
$\rightarrow$ **Coupling between scales with non-linearities**

Scattering Transform (ST) statistics

- **Wavelet Phase Harmonics and phase alignment** *(EA+20)*



→ 1 coeff / pair of scales / type of interaction
→ Can be extended to cross-statistics between maps

Scattering Transform (ST) statistics

- **A family of statistics**
  - Different generations of statistics
    → Wavelet Scattering Transforms (WST) *(EA+19)*
    → Wavelet Phase Harmonics (WPH) *(EA+20)*
    → Scattering covariances/spectra *(Cheng+23)*
  - All share the same framework

Scattering Transform (ST) statistics

- **A family of statistics**
  - Different generations of statistics
    → Wavelet Scattering Transforms (WST) ***(EA+19)***
    → Wavelet Phase Harmonics (WPH) ***(EA+20)***
    → Scattering covariances/spectra ***(Cheng+23)***
  - All share the same framework

- **Characterization and parameter inference**
  - Interstellar medium ***(EA+19, Regaldo+20, Saydjari+20, Lei+22)***
  - Weak lensing ***(Cheng+20, 21, 24)***
  - Large scale structures ***(EA+20, Eickenberg+22, Valogiannis+22a, 22b)***
  - 21cm epoch of reionization ***(Greig+22, Hothi+23)***
  - ...

→ Very informative (sometimes on par with CNN!)
→ Wide range of applicability (generic, training-less)

Generative models from Scattering transforms

- **Generative model from ST statistics** *(Bruna, Mallat, 19)*
  - From the ST statistics $\Phi(s)$ of data s
  - Maximum entropy model under ST constraints
  - Quantitative non-Gaussian modeling of *physical processes*

$$p(s) \longrightarrow s_0 \longrightarrow \phi(s_0) \longrightarrow p^{\mathrm{m.e.}}_{\phi(s_0)}(\tilde{s}) \longrightarrow \tilde{s}$$

Generative models from Scattering transforms

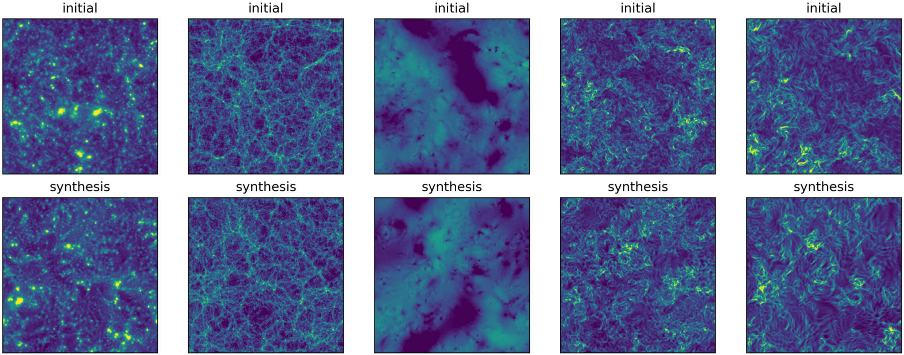
- **Generative model from ST statistics** *(Bruna, Mallat, 19)*
  - From the ST statistics $\Phi(s)$ of data s
  - Maximum entropy model under ST constraints
  - Quantitative non-Gaussian modeling of *physical processes*

$$p(s) \longrightarrow s_0 \longrightarrow \phi(s_0) \longrightarrow p^{\mathrm{m.e.}}_{\phi(s_0)}(\tilde{s}) \longrightarrow \tilde{s}$$

- **Practical implementation (microcanonical)**
  - Constraints $\Phi(s)$ from a (set of) data s
  - Sampled with a gradient-descent algorithm
    $\rightarrow$ from a white noise realization
    $\rightarrow$ Pixel-space optim. of $\tilde{s}$ such that $\Phi(\tilde{s}) \simeq \Phi(s)$

Generative models from Scattering transforms

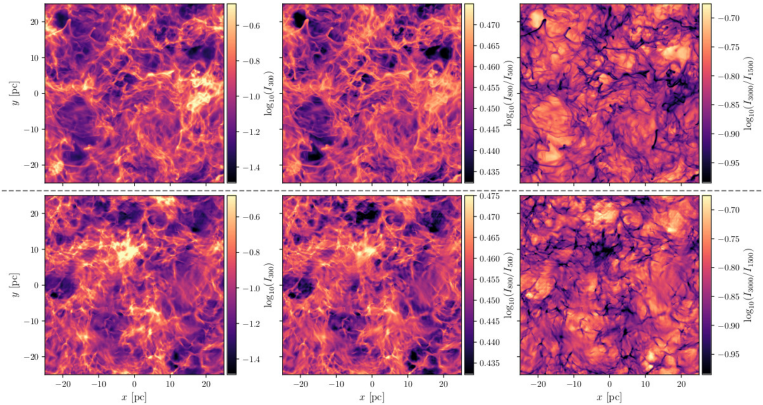
- **Generative model from a single image** *(Cheng+24)*
  - Scattering spectra + physical dimensionality reduction



→ Realistic NG models from a few hundreds coefficients!
→ Usual (NG) statistics very well reproduced (up to 1-10 %)

Generative models from Scattering transforms

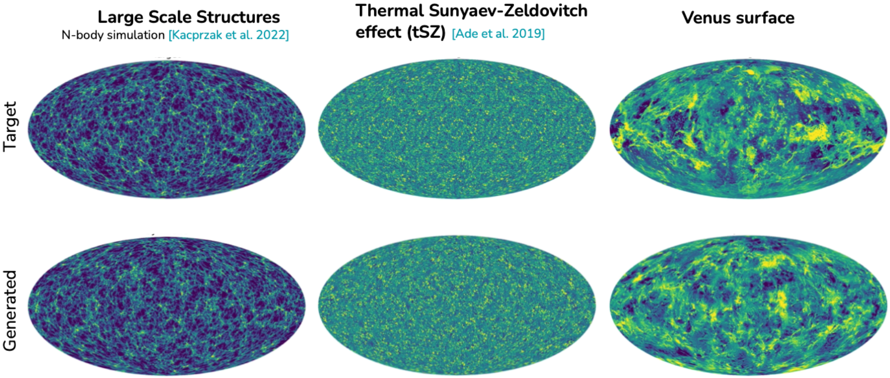
- **Multi-frequency dust emission models** *(Regaldo+22)*
  - Cross-WPH, simulated dust intensity, 300/500/800/1500/3000GHz



→ Extension to cross-statistics beyond linear correlation

Generative models from Scattering transforms

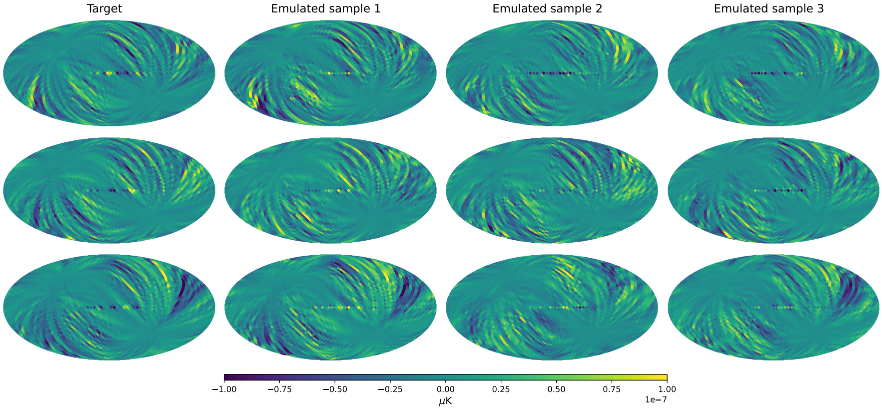
- Scattering transform on the sphere *(Mousset+24)*



→ Extension possible to different types of data
→ 1D, 3D, 2+1D data... *(Morel+25, Hothi+25)*

Generative models from Scattering transforms

- **Inhomogeneous processes on the sphere** *(Campeti+25)*
  - Full-sky CMB instrumental systematics



→ Extension (partially) possible to non-homogeneous data

Application of ST generative models

- **Modeling a known physical process**
  - At least one sample of the process of interest
  - Interest in having many realizations:
    $\rightarrow$ data augmentation e.g. for machine learning *(Jeffrey+, 22)*
    $\rightarrow$ bootstrapping for various statistical estimations
  - ST models are not perfect 1-to-1 replacement

Application of ST generative models

- **Modeling a known physical process**
  - At least one sample of the process of interest
  - Interest in having many realizations:
    $\rightarrow$ data augmentation e.g. for machine learning *(Jeffrey+, 22)*
    $\rightarrow$ bootstrapping for various statistical estimations
  - ST models are not perfect 1-to-1 replacement

- **Low dim. parametric models of physical fields**
  - Parameters are the ST coefficients
  - As parameter space for inverse problems
    $\rightarrow$ Allow to work without physically-driven models *$\rightarrow$ see S. Pierre's talk*
  - As latent space for machine-learning tasks
    $\rightarrow$ Modeling of component from unlabeled mixture

Application of ST generative models

- **Modeling a known physical process**
  - At least one sample of the process of interest
  - Interest in having many realizations:
    $\rightarrow$ data augmentation e.g. for machine learning *(Jeffrey+, 22)*
    $\rightarrow$ bootstrapping for various statistical estimations
  - ST models are not perfect 1-to-1 replacement

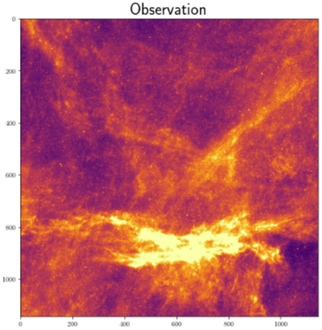
- **Low dim. parametric models of physical fields**
  - Parameters are the ST coefficients
  - As parameter space for inverse problems
    $\rightarrow$ Allow to work without physically-driven models *$\rightarrow$ see S. Pierre's talk*
  - As latent space for machine-learning tasks
    $\rightarrow$ Modeling of component from unlabeled mixture

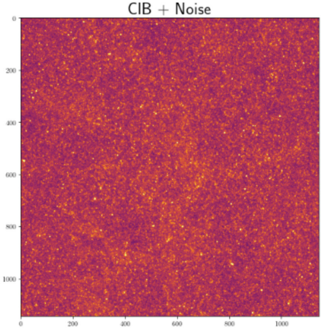
$\rightarrow$ Lot of new possibilities in low-data regimes or without prior models

Outline

1. Scattering Transforms and generative models
2. **Component separation and modeling from the data**

Separating CIB and Galactic dust emission

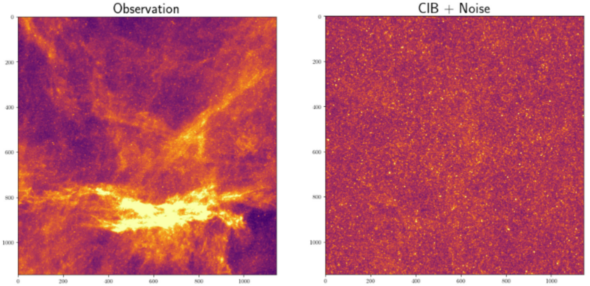




- **Galactic dust emission and Cosmic Infrared Background (CIB)**
  - Thermal dust emission in the interstellar medium
  - Same emission from Milky Way and other galaxies
  - Cosmic background dominates a smaller scales

→ Characterization of Galactic dust on those scales?
→ Challenge of low-data regime + lack of prior model

Separating CIB and Galactic dust emission

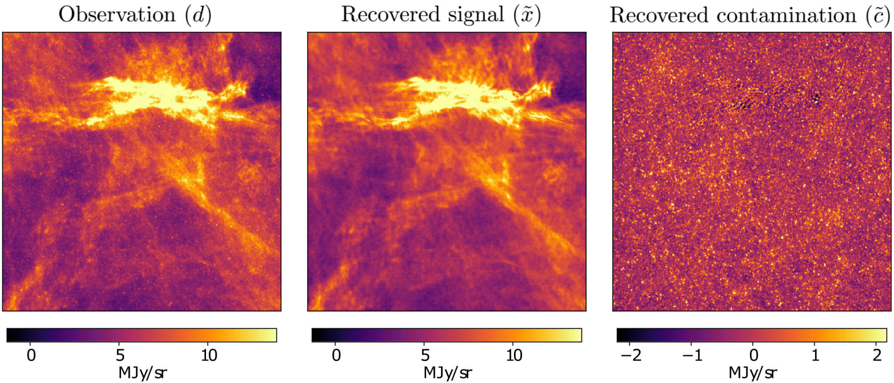


- **Dust emission/Cosmic Infrared Background (Auclair+, 24)**
  - $d = s + c$, s thermal dust emission, c CIB
  - CIB model from separate observation (cosmological ⇒ homogeneous)
  - Gradient descent under two constraints, using a ST-model for $\{c_i\}_i$

$$\left\langle \Phi(\tilde{s} + c_i) \right\rangle_i \simeq \Phi(d), \qquad \Phi(\tilde{c}) = \Phi(c)$$

Separating CIB and Galactic dust emission

- **Recovered components (Auclair+24)**



→ **Statistical component separation solely from obs. data**
→ **Thermal dust is statistically recovered up to the beam**

A general framework for component separation

- **Constraints from ST statistics**
  - Constraints written directly from available data
  - Can rely on ancillary data with cross-statistics
  - Complete knowledge of all components not needed

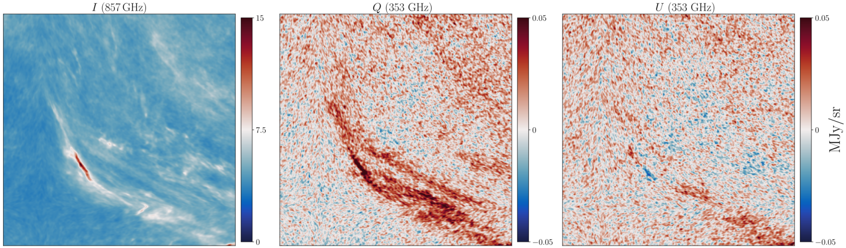
A general framework for component separation

- **Constraints from ST statistics**
  - Constraints written directly from available data
  - Can rely on ancillary data with cross-statistics
  - Complete knowledge of all components not needed

- **Optimization scheme**
  - Sampling e.g. with gradient descent in pixel space
  - Degrees of freedom can be adapted
  - Complementary weighting schemes for the losses

→ Versatile framework for component separation
→ Can include various statistical constraints

Application to polarized Galactic dust foregrounds

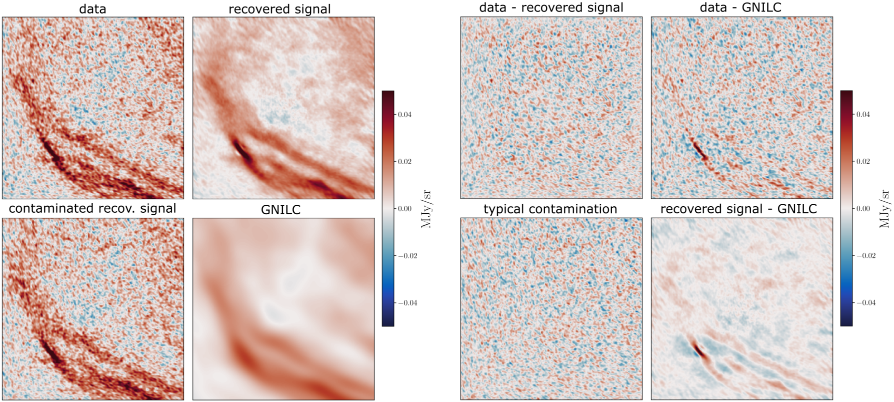


- **Application on polarized Planck data (Tsouros+, in prep.)**
  - Improvement from **Régaldo+21** and **Delouis+22**
  - Dust Q/U emission at 353 GHz + noise + CMB
  - 7 constraints including correlation with total intensity (857 GHz)

→ Generate new Q/U dust map directly from the data?
→ No prior model for the polarized dust emission

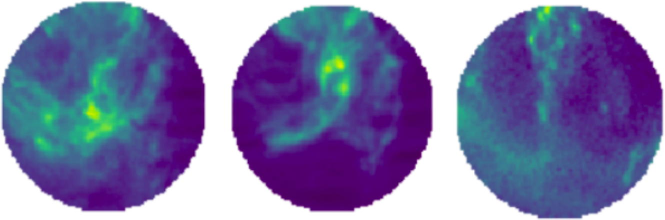
Application to polarized Galactic dust foregrounds

- Application on polarized Planck data (Tsouros+, in prep.)



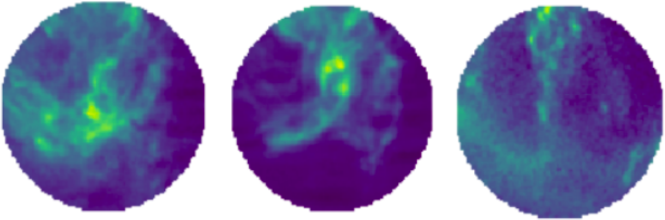
→ Maps that pass all compatibility tests with the data
→ To be extended full-sky and multi-frequency
→ see S. Pierre's talk for a full Bayesian framework!

Unsupervised separation of HI data



- **HI observations of the interstellar medium**
  - Warm (WNM) and Cold (CNM) Neutral Media
  - Two phases with different spectral/spatial properties

Unsupervised separation of HI data



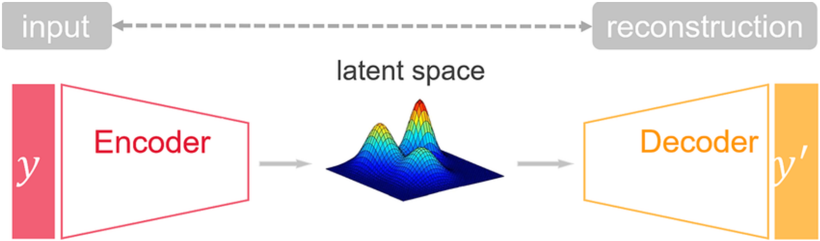
- **HI observations of the interstellar medium**
  - Warm (WNM) and Cold (CNM) Neutral Media
  - Two phases with different spectral/spatial properties

- **Modeling Galactic WNM/CNM directly from the data?**
  - GALFA-HI in 3 km/s bins (treated as 2D maps)
  - High-latitude + $|v| < 40$ km/s, 4' angular resolution
  - $\sim$ 36k 256^2 patches with CNM, WNM, noise

→ First step with only spatial morphology
→ Lack of prior model + interface with Machine Learning

Unsupervised separation of HI data

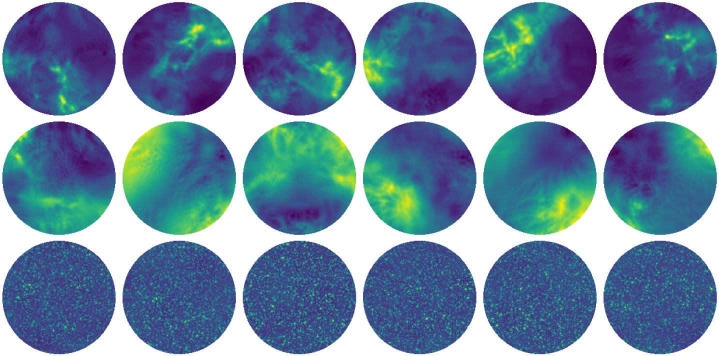
- **Variational Auto-Encoder (VAE) in ST space (Siahkoohi+, 23)**
  - Learn the identity in ST space over the dataset
  - Gaussian mixture model in latent space
    $\rightarrow$ one Gaussian per component (hyperparameter)



$\rightarrow$ Unsupervised learning of components in ST space
$\rightarrow$ ST model for each component after training

Unsupervised separation of HI data

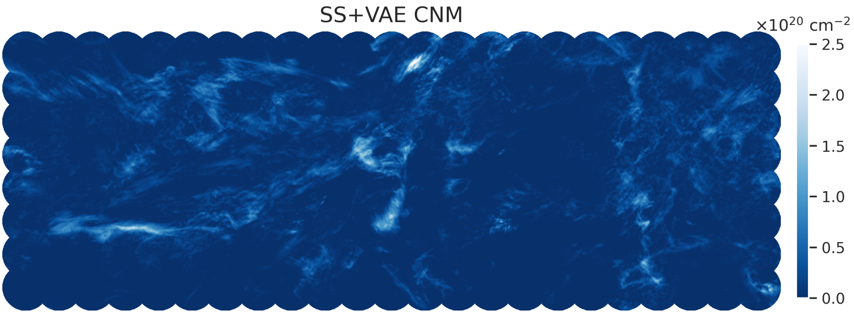
- **Application to GALFA-HI data** **(Lei, Clark+, 2025)**
  - Unsupervised identification of 3 components



→ **WNM/CNM/noise well modeled!**
→ **Interfacing ST models with other ML algorithms**

Application to CNM mapping

- **Component separation from learned models (Lei, Clark+, 2025)**
  - ST-based component separation (other could be used)
  - 19°x51° footprint, LOS-integrated CNM column density



→ Phase separation directly from the data (preliminary!)
→ From spatial structure only, spectral information next

Conclusion

- **Scattering Transforms**
  → Efficient non-Gaussian statistics inspired from neural network
  → Characterize interaction between scales in non-linear processes
- **New tools for (astro-)physics**
  → Modeling and component separations
  → Ability to work with a very limited amount of data
  → Ability to work without prior data model
- **Applications to come are very exciting!**
  → Versatile and powerful tools: happy to discuss :-)
  → We are developing a versatile library: stay tuned!

Thanks for your attention!