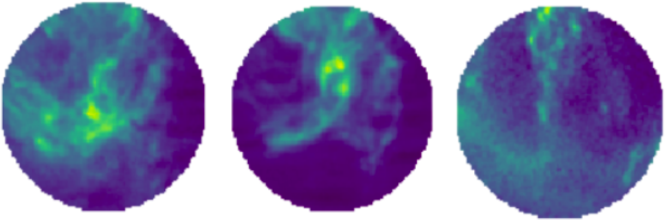
Scattering Transforms for radio data: using spatial morphology for modeling and component separation

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

Journées AtLAST France,
Paris, May 13th 2025

Phase separation in HI data

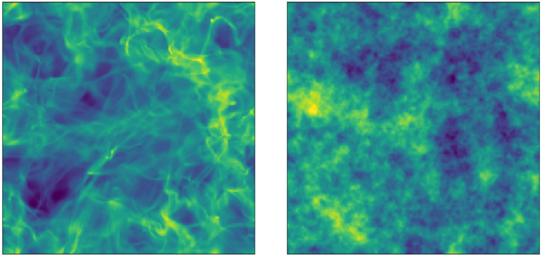


- **HI observations of Galactic WNM/CNM**
  - Warm (WNM) and Cold (CNM) Neutral Media
  - Two phases with different spectral/spatial properties
  - Current separation mostly rely on spectral properties

→ How to model and separate WNM and CNM?
→ By relying on their different spatial morphologies?
→ If possible directly from the data?

Beyond Power Spectrum statistics

- **A generic tool: the Power Spectrum**
  - Square amplitude of Fourier modes
  - Energy/Power in each Fourier mode
  - Most usual statistical tool in astrophysics

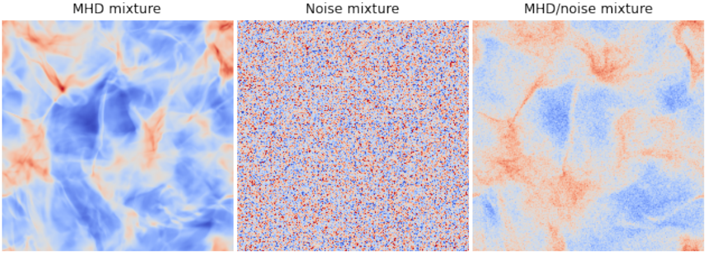


Fields of same power spectrum

→ Does not characterize interaction between scales
→ Need beyond Power Spectrum statistics for NL fields

Non-Gaussian modeling and component separation

- **Non-Gaussianity is not our enemy!**



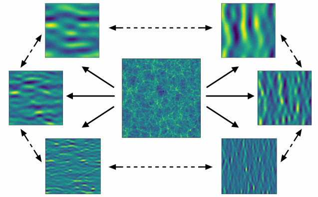
  - Important lever arm for components separation
  - Even from a small amount of data

→ Challenge of using non-Gaussian information
→ Should be possible to work with (very) small dataset

Outline

Scattering transform (ST) statistics

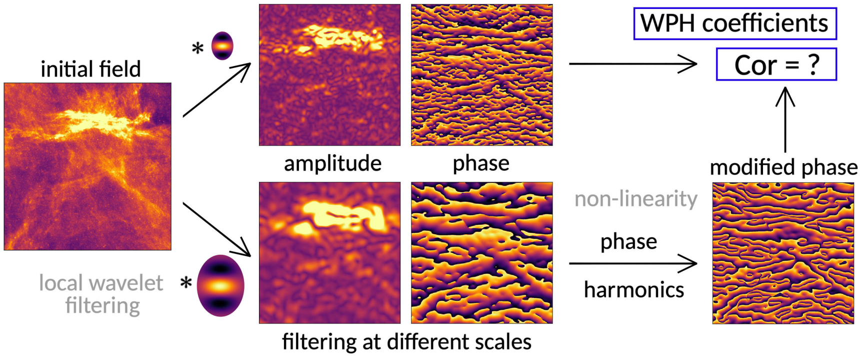
- **Scattering transform statistics** **(Mallat+, 2010+)**
  - Initially developed in data science
  - Inspired from neural networks
    → efficient characterization and reduced variance
  - Do not need any training stage
    → explicit mathematical form and interpretability



→ Wavelet filters separating the different scales
→ 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)*
  - 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

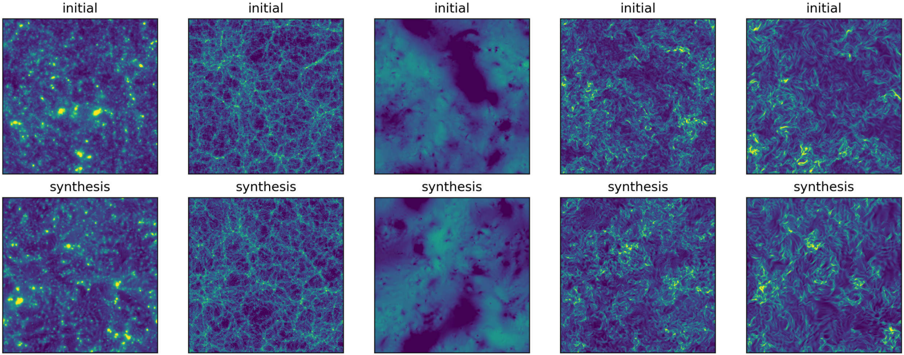
- **Generative model from ST statistics** *(Bruna, Mallat, 19)*
  - From the ST statistics $\Phi(s)$ of data s
  - Maximum entropy model under ST constraints
  - Quantitative non-Gaussian modeling of *physical processes*

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

- **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

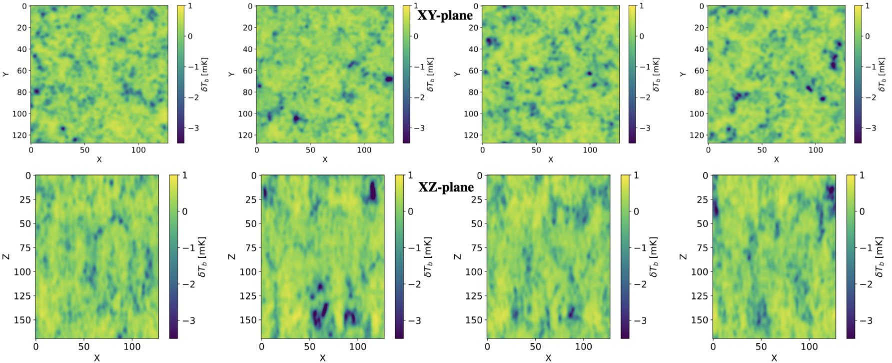
- **Generative model from a single image** *(Cheng+24)*
  - Scattering spectra + physical dimensionality reduction



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

Generative models from Scattering transforms

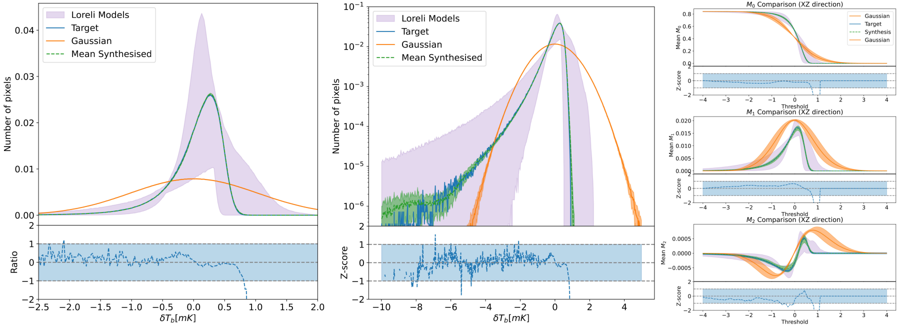
- **ST generative models for radio data** *(Hothi+, in prep)*
  - Epoch of Reionization spectroscopic data-cube



(a) Target Field (b) Synthesised Field 1 (c) Synthesised Field 2 (d) Synthesised Field 3

Generative models from Scattering transforms

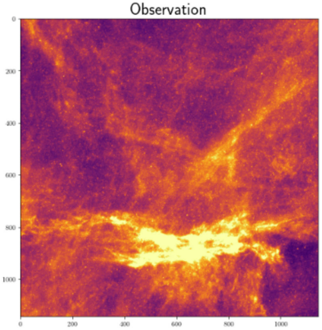
- **ST generative models for radio data** *(Hothi+, in prep)*
  - Statistical validation for pdf (lin/log) and xz-Minkowski functional

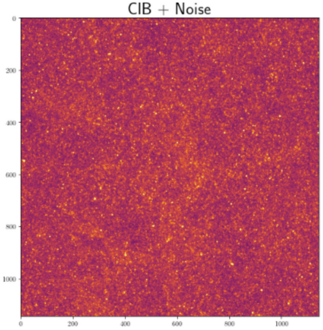


→ Very good modeling from one single data-cube
→ Extend ST applications to spectroscopic radio data

Outline

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

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

From generative model to comp. sep.

- **Generative model from available sample**
  - Estimate $\phi(s)$ from sample s
  - Generate maps $\tilde{s}$ such that

$$\Phi(\tilde{s}) \simeq \Phi(s)$$

  - Sampled with gradient descent from white noise

From generative model to comp. sep.

- **Generative model from available sample**
  - Estimate $\phi(s)$ from sample s
  - Generate maps $\tilde{s}$ such that

$$\Phi(\tilde{s}) \simeq \Phi(s)$$

  - Sampled with gradient descent from white noise

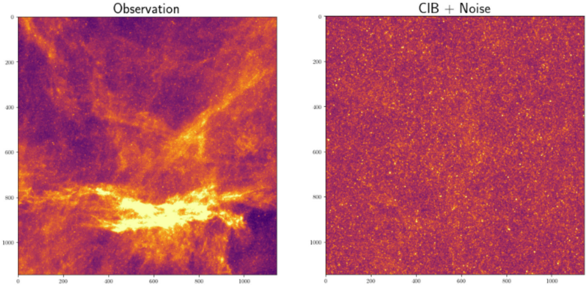
- **Indirect observation with know contamination**
  - $d = s + c$, assume we have $\{c_i\}_i$
  - Generate a map $\tilde{s}$ such that

$$\langle \Phi(\tilde{s} + c_i) \rangle_i \simeq \Phi(d)$$

  - Gradient descent from d (for instance)

$\rightarrow$ Framework for component separation
$\rightarrow$ Can include various other statistical constraints

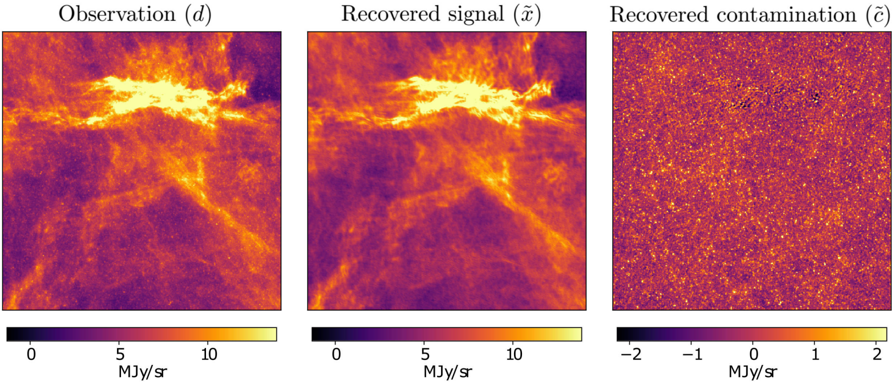
Separation solely from observational data



- **Dust emission/Cosmic Infrared Background (Auclair+, 24)**
  - $d = s + c$, s thermal dust emission, c CIB
  - CIB model from separate observation (cosmological $\Rightarrow$ homogeneous)
  - Two constraints, with $\{c_i\}_i$ from ST model

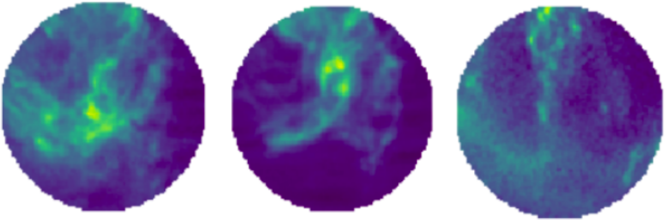
$$\left\langle \Phi(\tilde{s} + c_i) \right\rangle_i \simeq \Phi(d), \qquad \Phi(\tilde{c}) = \Phi(c)$$

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



→ Statistical component separation solely from obs. data
→ Thermal dust is recovered at an unprecedented resolution
→ Soon in a Bayesian framework! (Pierre+, in prep.)

Unsupervised separation of HI data

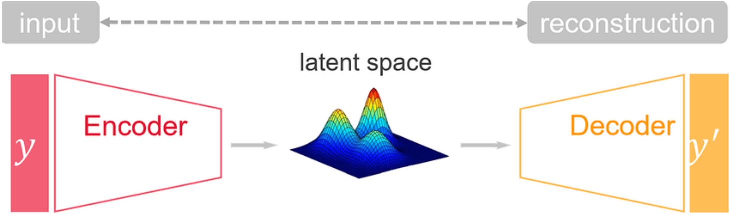


- **Modeling Galactic WNM/CNM from HI data**
  - GALFA-HI in 3 km/s bins (treated as 2D maps)
  - High-latitude + $|v| < 40$ km/s, 4' angular resolution
  - $\sim$ 36k 256^2 patches with CNM, WNM, noise

→ Learn WNM and CNM ST models from these patches?
→ First step with only spatial morphology

Unsupervised separation of HI data

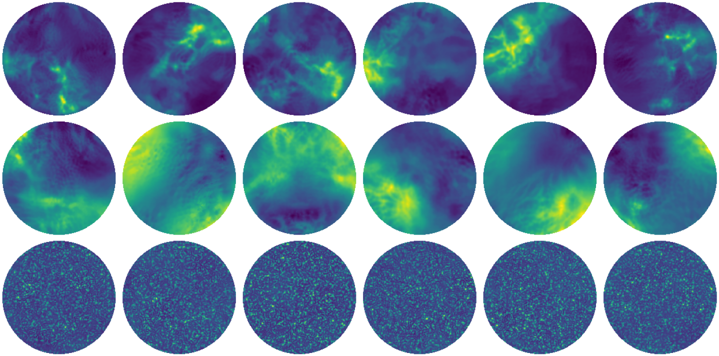
- **Variational Auto-Encoder (VAE) in ST space** **(Siahkoohi+, 23a,b)**
  - Learn the identity in ST space over the dataset
  - Gaussian mixture model in latent space
    $\rightarrow$ one Gaussian per component (hyperparameter)



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

Unsupervised separation of HI data

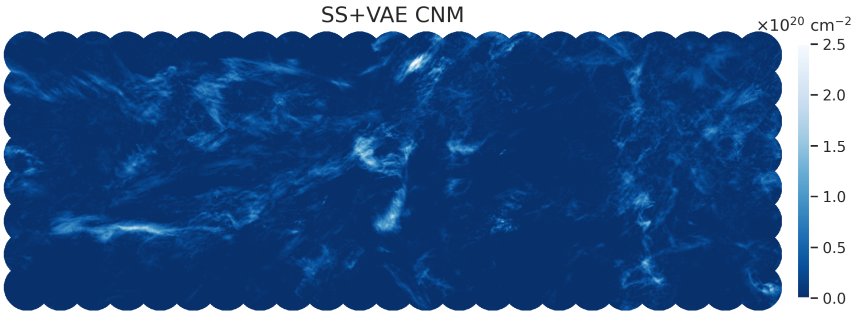
- **Application to GALFA-HI data (Lei, Clark+, in prep.)**
  - Unsupervised identification of 3 components



→ WNM/CNM/noise seem well modeled! (in progress)
→ Interfacing ST models with other ML algorithms

Application to CNM mapping

- **Component separation from learned models** **(Lei, Clark+, in prep.)**
  - 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!**
  → Complete ST modeling of spectroscopic datacubes
  → Multiple applications to radio data
  → Versatile and powerful tools: happy to discuss :-)

Thanks for your attention!