Probabilistic Mapping of Dark Matter with Neural Score Matching

astro-ph.CO arXiv:2011.08271

Benjamin Remy

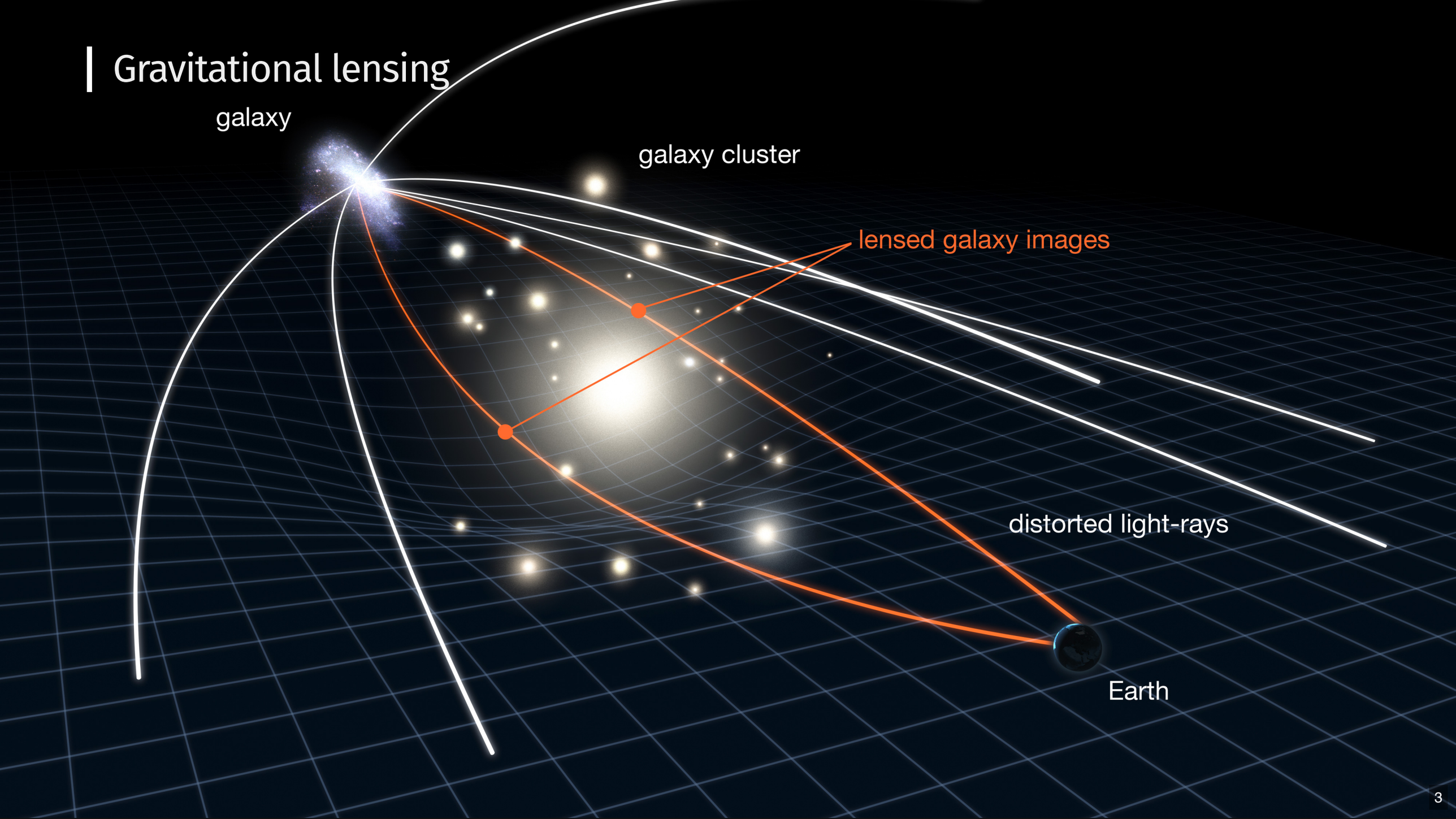
With : Francois Lanusse, Niall Jeffrey, Jia Liu, J.-L. Starck, Ken Osato

CosmoStat

cnes

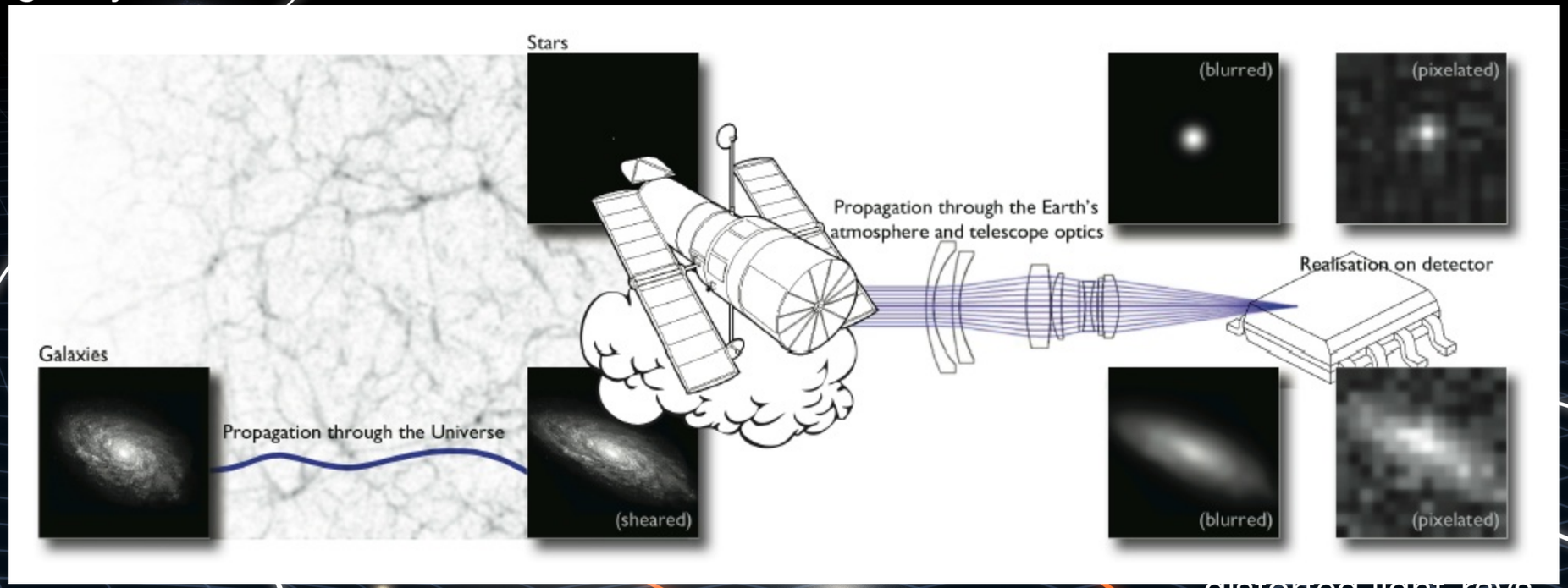
université PARIS-SACLAY

Gravitational lensing



Gravitational lensing

galaxy



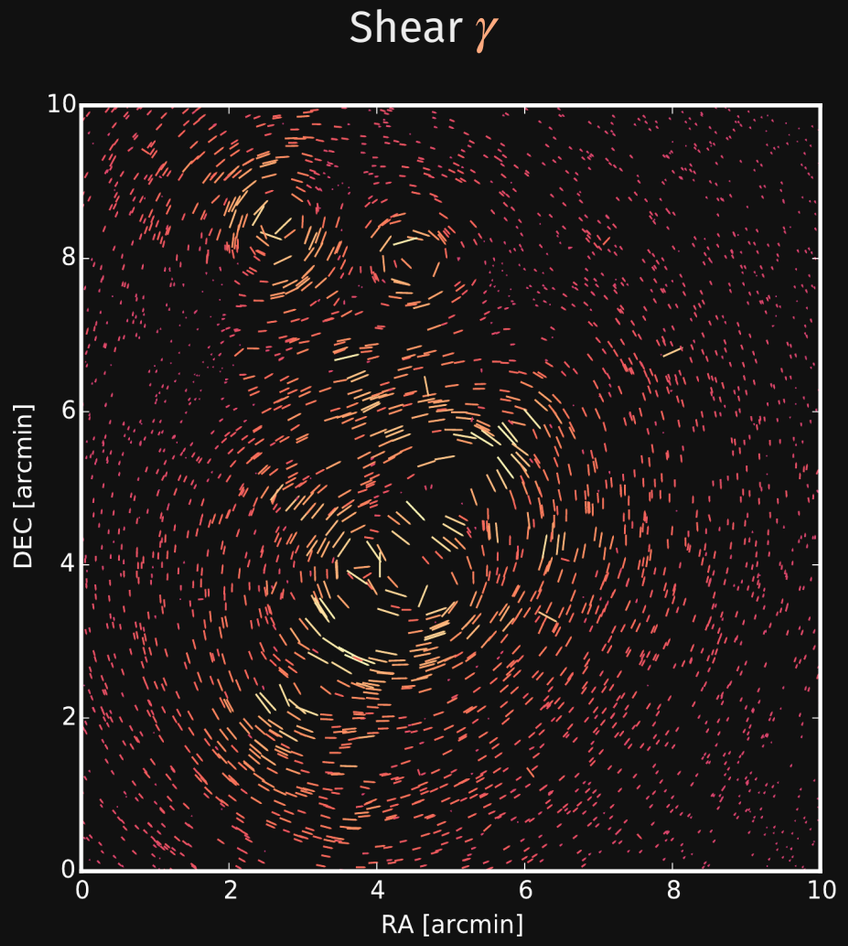
distorted light-rays

Galaxy shapes as estimators for gravitational shear

$$e = \gamma + e_i \qquad \text{with} \qquad e_i \sim \mathcal{N}(0, I)$$

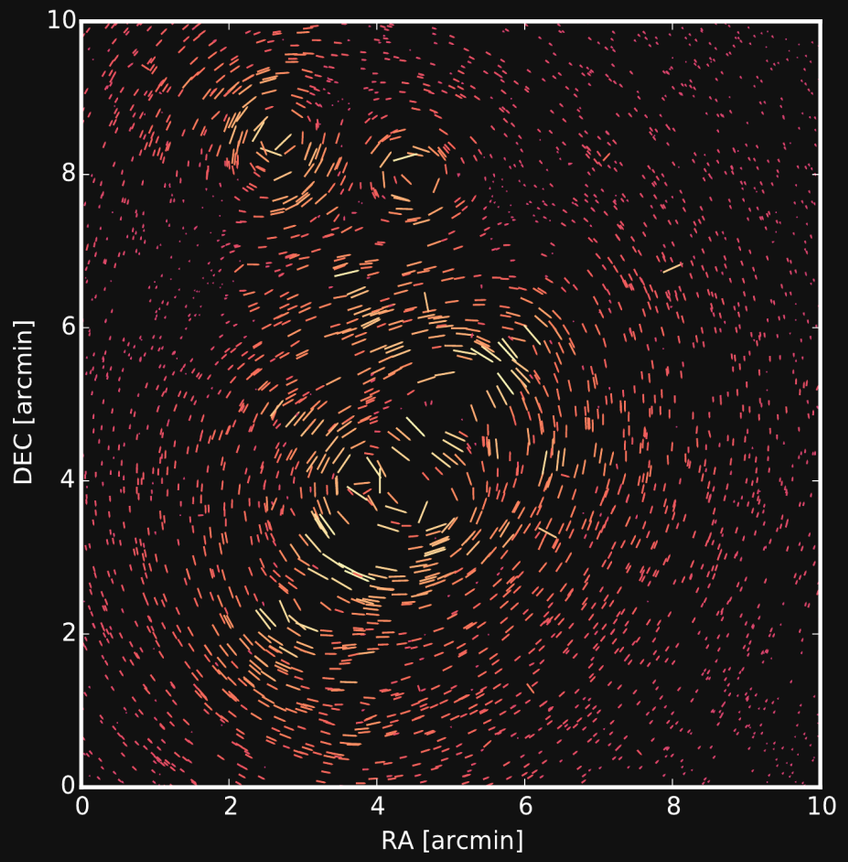
- We are trying the measure the **ellipticity** e of galaxies as an estimator for the **gravitational shear** γ

The Weak Lensing Mass-Mapping as an Inverse Problem

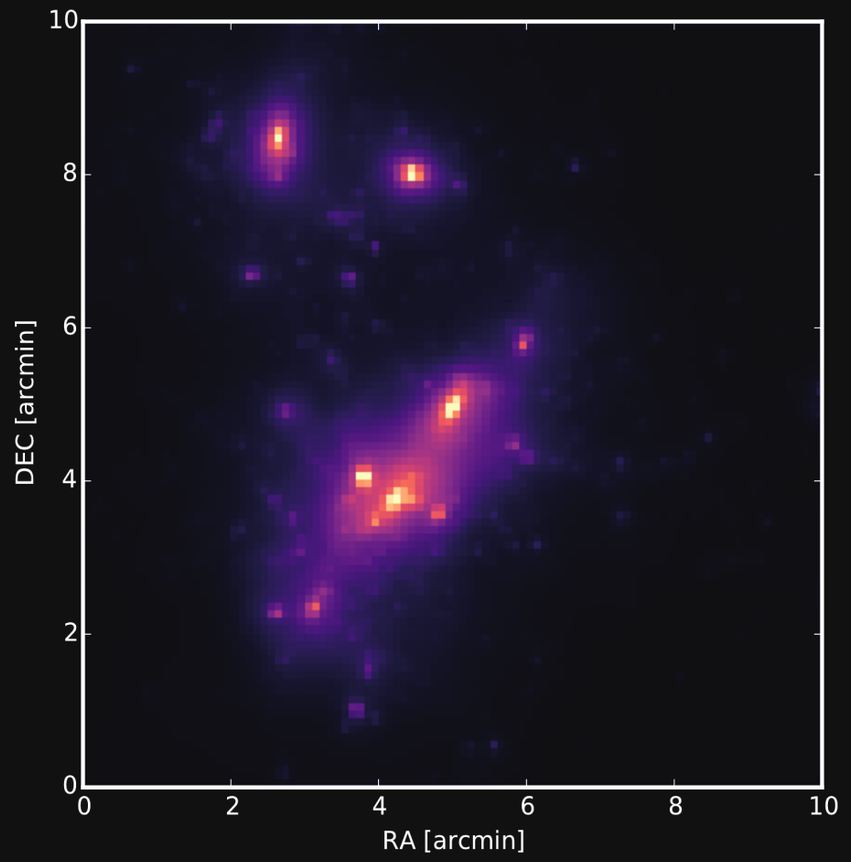


The Weak Lensing Mass-Mapping as an Inverse Problem

Shear γ

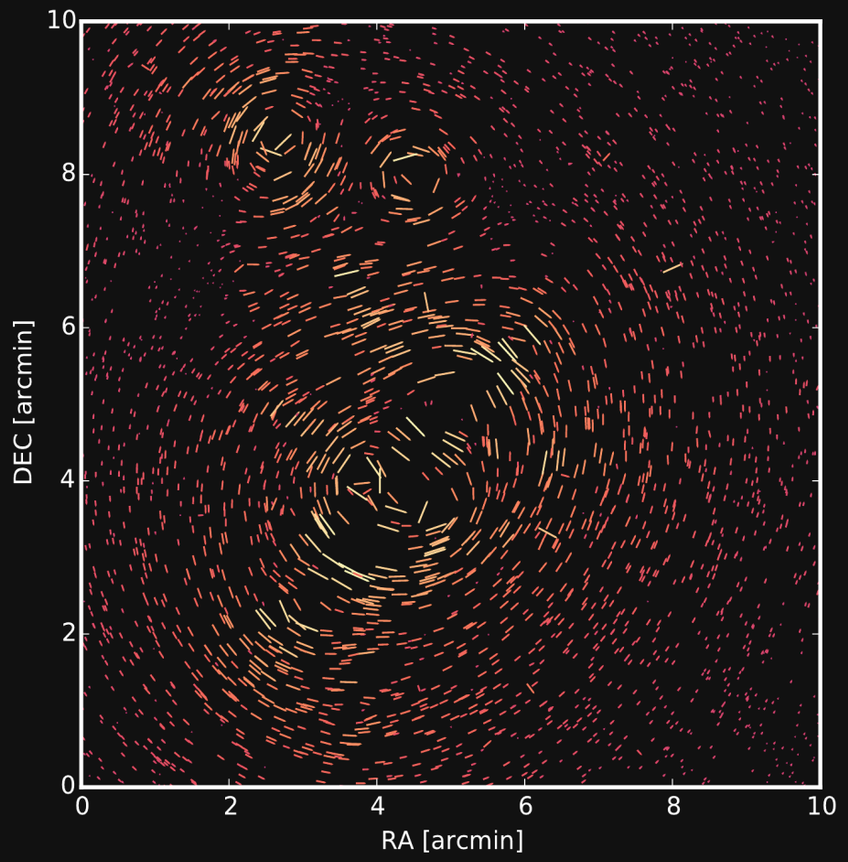


Convergence κ

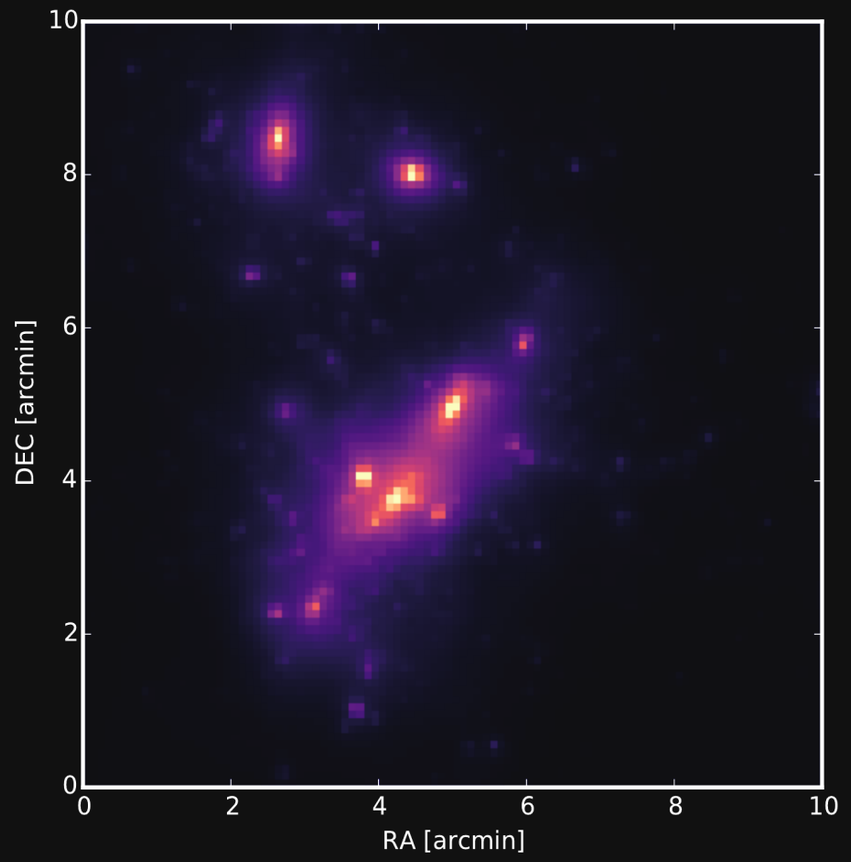


The Weak Lensing Mass-Mapping as an Inverse Problem

Shear γ



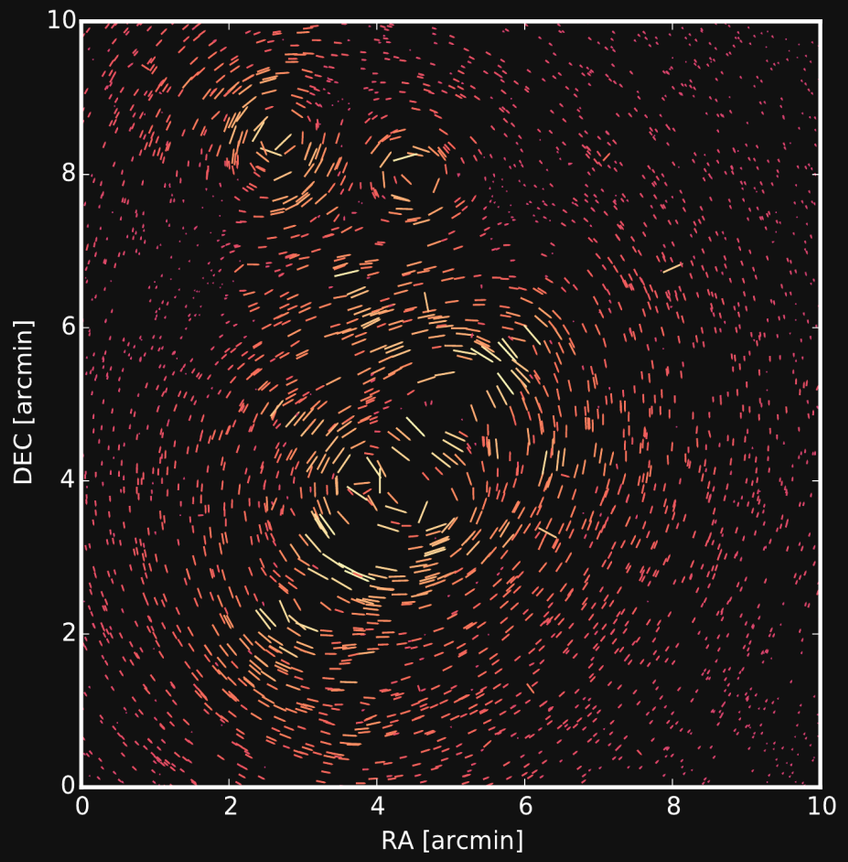
Convergence κ



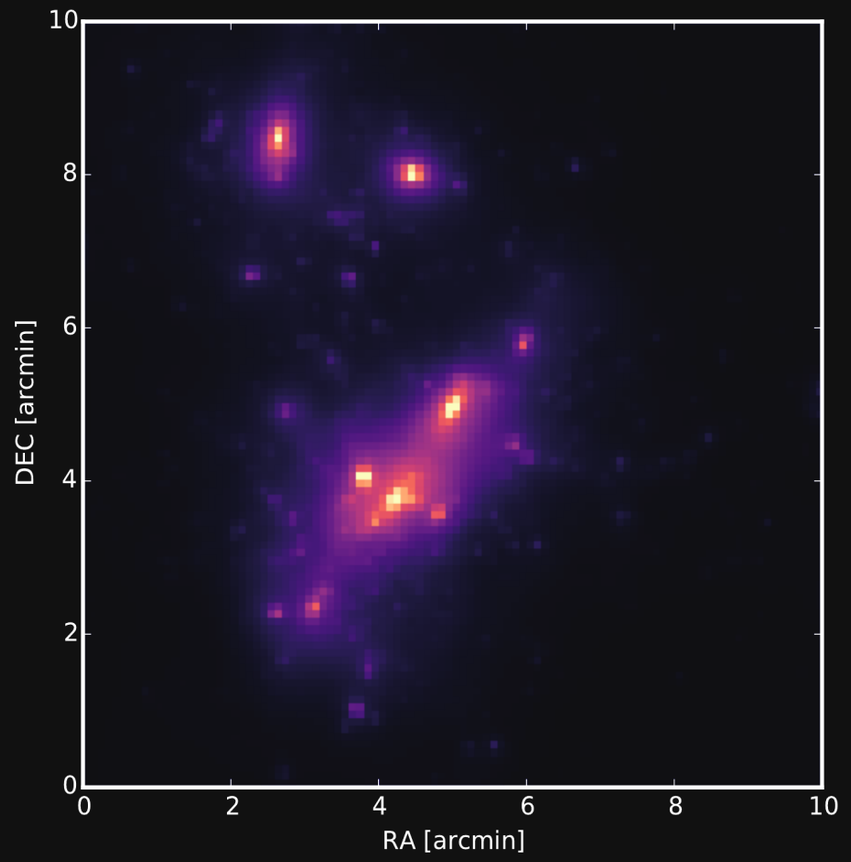
$$\gamma_1 = \frac{1}{2}(\partial_1^2 - \partial_2^2)\,\Psi \quad ; \quad \gamma_2 = \partial_1\partial_2\,\Psi \quad ; \quad \kappa = \frac{1}{2}(\partial_1^2 + \partial_2^2)\,\Psi$$

The Weak Lensing Mass-Mapping as an Inverse Problem

Shear γ



Convergence κ



$$\gamma = \mathbf{P}\kappa$$

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

- $p(\gamma|\kappa)$ is the data likelihood, which **contains the physics**

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

- $p(\gamma|\kappa)$ is the data likelihood, which **contains the physics**
- $p(\kappa)$ is the prior knowledge on the solution.

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.

$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

- $p(\gamma|\kappa)$ is the data likelihood, which **contains the physics**
- $p(\kappa)$ is the prior knowledge on the solution.

We can estimate for instance the Maximum A Posteriori solution:

$$\hat{\kappa} = \arg\max_{\kappa} \; \log p(\gamma \mid \kappa) + \log p(\kappa)$$

$$\hat{\kappa} = \arg\max_{\kappa} \; -\tfrac{1}{2} \parallel \gamma - \mathbf{A}x \parallel^2_{\mathbf{\Sigma}} + \log p(\kappa)$$

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.
$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

- $p(\gamma|\kappa)$ is the data likelihood, which **contains the physics**
- $p(\kappa)$ is the prior knowledge on the solution.

We can estimate for instance the Maximum A Posteriori solution:

$$\hat{\kappa} = \arg\max_{\kappa} \; \log p(\gamma \mid \kappa) + \log p(\kappa)$$

$$\hat{\kappa} = \arg\max_{\kappa} \; -\tfrac{1}{2} \parallel \gamma - \mathbf{A}x \parallel^2_{\mathbf{\Sigma}} + \log p(\kappa)$$

Or **estimate from the full posterior** $p(\kappa|\gamma)$ with MCMC or Variational Inference methods.

Linear inverse problems

$$\boxed{\gamma = \mathbf{A}\kappa + n}$$

$\mathbf{A}$ is known and encodes our physical understanding of the problem.
$\Longrightarrow$ When non-invertible or ill-conditioned, the inverse problem is ill-posed **with no unique solution** x

The Bayesian view of the problem:

$$p(\kappa|\gamma) \propto p(\gamma|\kappa)\, p(\kappa)$$

- $p(\gamma|\kappa)$ is the data likelihood, which **contains the physics**
- $p(\kappa)$ is the prior knowledge on the solution.

We can estimate for instance the Maximum A Posteriori solution:

$$\hat{\kappa} = \arg\max_{\kappa} \; \log p(\gamma \mid \kappa) + \log p(\kappa)$$

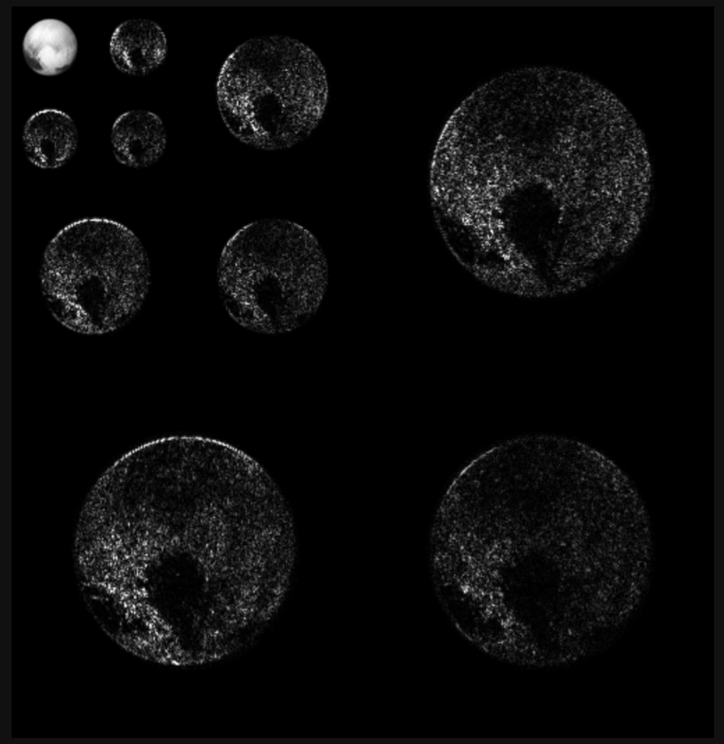
$$\hat{\kappa} = \arg\max_{\kappa} \; -\tfrac{1}{2} \parallel \gamma - \mathbf{A}x \parallel^2_{\mathbf{\Sigma}} + \log p(\kappa)$$

Or **estimate from the full posterior** $p(\kappa|\gamma)$ with MCMC or Variational Inference methods.

$\boxed{\log p(\kappa)}$??

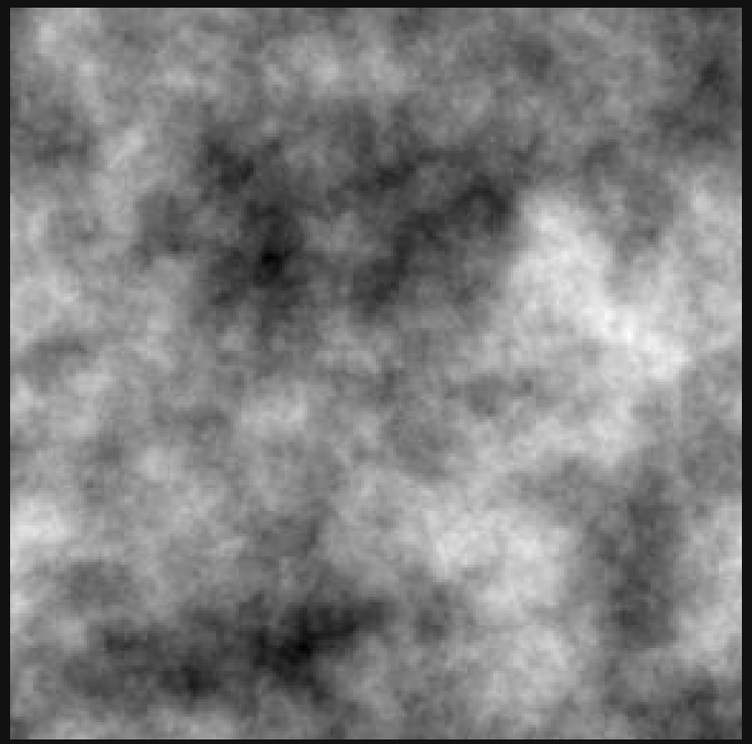
Classical examples of signal priors

Sparse



$$\log p(x) = \| \mathbf{W}x \|_1$$

Gaussian



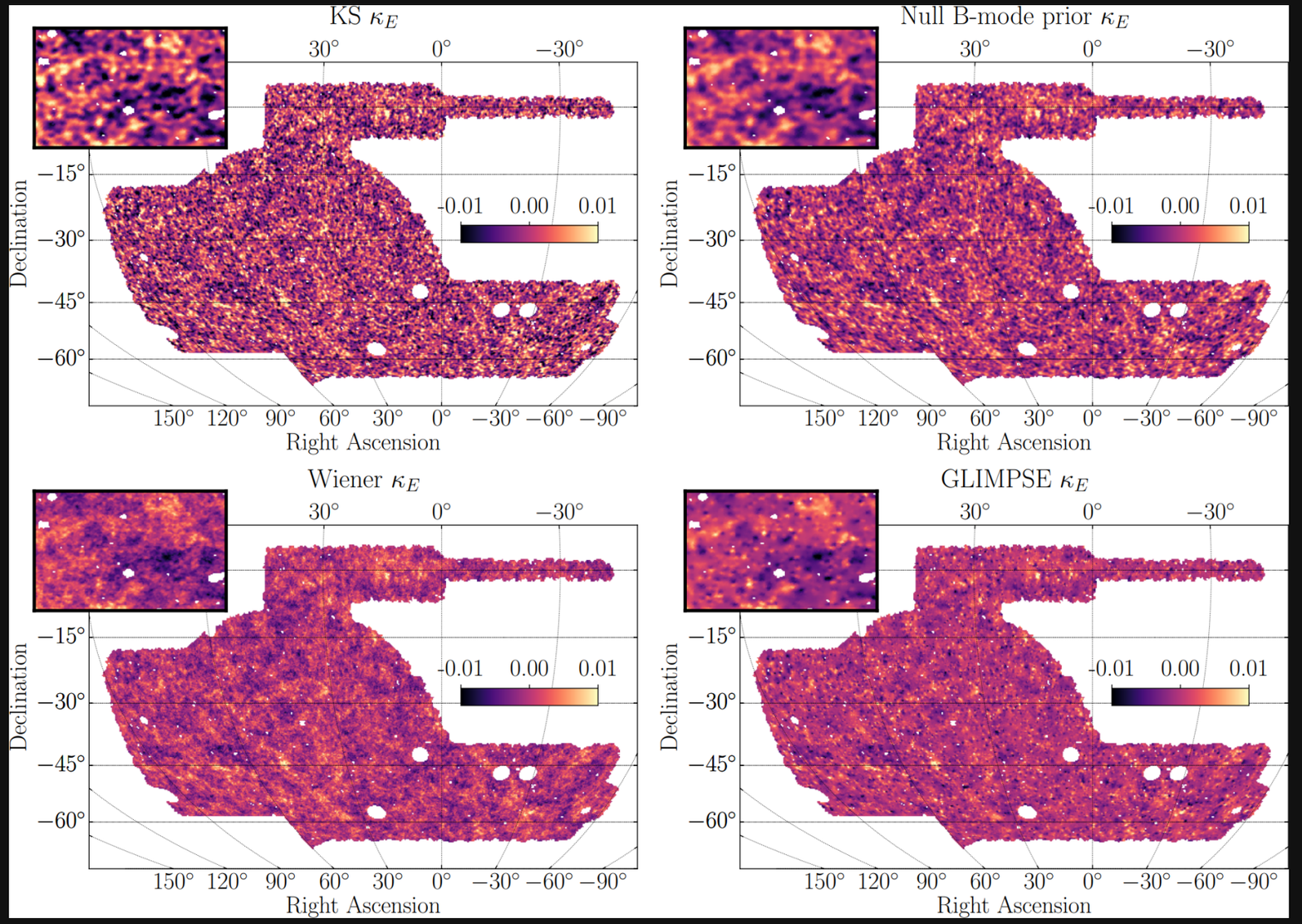
$$\log p(x) = x^t \mathbf{\Sigma^{-1}} x$$

Total Variation

$$\log p(x) = \| \nabla x \|_1$$

Illustration on the Dark Energy Survey (DES) Y3

Jeffrey, et al. (2021)



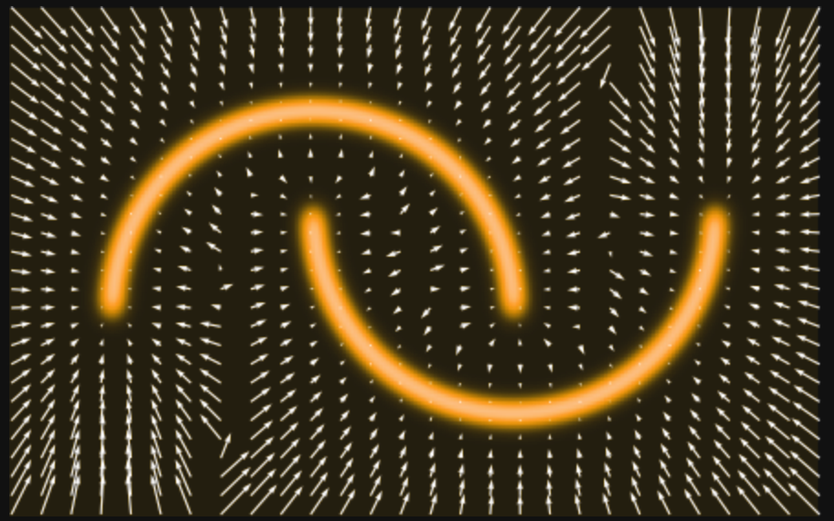
But what about learning the prior
with **deep generative models**?

The score is all you need!

- Whether you are looking for the MAP or sampling with HMC or MALA, you **only need access to the score** of the posterior:

$$\frac{d \log p(x|y)}{dx}$$

  - Gradient descent: $x_{t+1} = x_t + \tau \nabla_x \log p(x_t|y)$
  - Langevin algorithm:
    $x_{t+1} = x_t + \tau \nabla_x \log p(x_t|y) + \sqrt{2\tau} n_t$

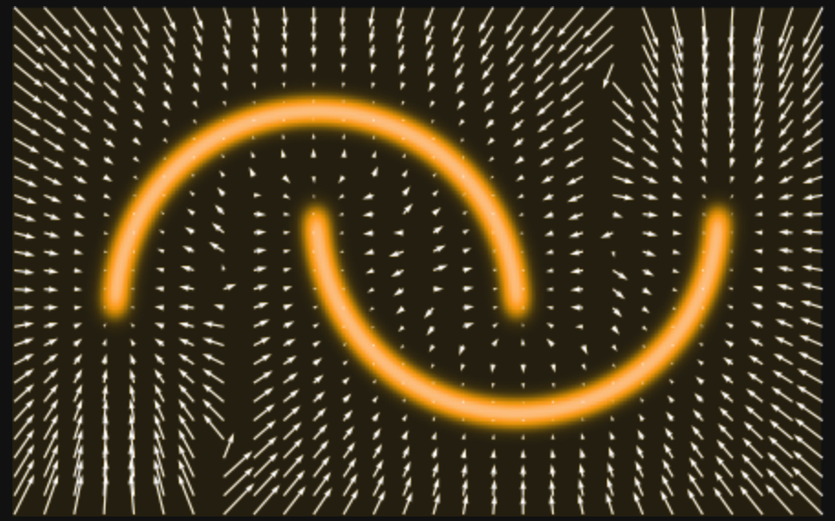


The score is all you need!

- Whether you are looking for the MAP or sampling with HMC or MALA, you **only need access to the score** of the posterior:

$$\frac{d \log p(x|y)}{dx}$$

  - Gradient descent: $x_{t+1} = x_t + \tau \nabla_x \log p(x_t|y)$
  - Langevin algorithm: $x_{t+1} = x_t + \tau \nabla_x \log p(x_t|y) + \sqrt{2\tau} n_t$



- The score of the full posterior is simply:

$$\nabla_\kappa \log p(\kappa|\gamma) = \underbrace{\nabla_\kappa \log p(\gamma|\kappa)}_{\text{known}} + \underbrace{\nabla_\kappa \log p(\kappa)}_{\text{can be learned}}$$

$\Longrightarrow$ all we have to do is **model/learn the score of the prior**.

Neural Score Estimation by Denoising Score Matching

- **Denoising Score Matching**: An optimal Gaussian denoiser learns the score of a given distribution.

Neural Score Estimation by Denoising Score Matching

- **Denoising Score Matching**: An optimal **Gaussian denoiser learns the score** of a given distribution.
  - If $x \sim \mathbb{P}$ is corrupted by additional Gaussian noise $u \in \mathcal{N}(0, \sigma^2)$ to yield

$$x' = x + u$$

Neural Score Estimation by Denoising Score Matching

- **Denoising Score Matching**: An optimal **Gaussian denoiser learns the score** of a given distribution.
  - If $x \sim \mathbb{P}$ is corrupted by additional Gaussian noise $u \in \mathcal{N}(0, \sigma^2)$ to yield

$$x' = x + u$$

  - Let's consider a denoiser r_θ trained under an ℓ_2 loss:

$$\mathcal{L} = \| x - r_\theta(x', \sigma) \|_2^2$$

Neural Score Estimation by Denoising Score Matching

- **Denoising Score Matching**: An optimal **Gaussian denoiser learns the score** of a given distribution.
  - If $x \sim \mathbb{P}$ is corrupted by additional Gaussian noise $u \in \mathcal{N}(0, \sigma^2)$ to yield

$$x' = x + u$$

  - Let's consider a denoiser r_θ trained under an ℓ_2 loss:

$$\mathcal{L} = \| x - r_\theta(x', \sigma) \|_2^2$$

  - The optimal denoiser $r_{\theta^\star}$ verifies:

$$\boxed{r_{\theta^\star}(\boldsymbol{x}', \sigma) = \boldsymbol{x}' + \sigma^2 \nabla_{\boldsymbol{x}} \log p_{\sigma^2}(\boldsymbol{x}')}$$

Neural Score Estimation by Denoising Score Matching

- **Denoising Score Matching**: An optimal **Gaussian denoiser learns the score** of a given distribution.
  - If $x \sim \mathbb{P}$ is corrupted by additional Gaussian noise $u \in \mathcal{N}(0, \sigma^2)$ to yield

$$x' = x + u$$

  - Let's consider a denoiser r_θ trained under an ℓ_2 loss:

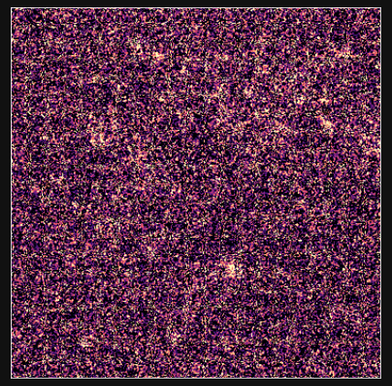
$$\mathcal{L} = \| x - r_\theta(x', \sigma) \|_2^2$$

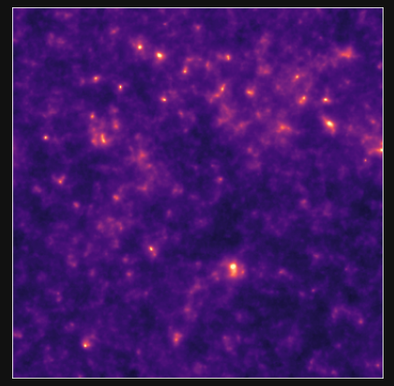
  - The optimal denoiser $r_{\theta^\star}$ verifies:

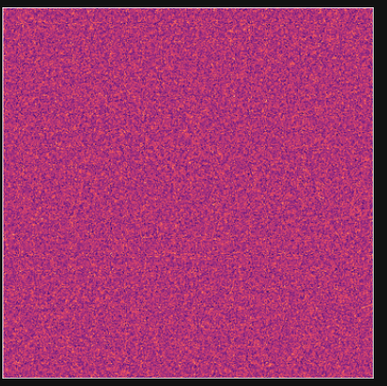
$$\boxed{r_{\theta^\star}(\boldsymbol{x}', \sigma) = \boldsymbol{x}' + \sigma^2 \nabla_{\boldsymbol{x}} \log p_{\sigma^2}(\boldsymbol{x}')}$$

$\boldsymbol{x}'$ $\quad$ $\boldsymbol{r}_{\theta^\star}(\boldsymbol{x}', \sigma)$ $\quad$ $\nabla_{\boldsymbol{x}} \log p_{\sigma^2}(\boldsymbol{x}')$

$$\Big(\quad - \quad \Big) / \sigma^2 =$$







Efficient sampling by Annealed HMC

- Even with gradients, **sampling in high number of dimensions is difficult!** Because of:
  - Curse of dimensionality
  - Highly correlated chains

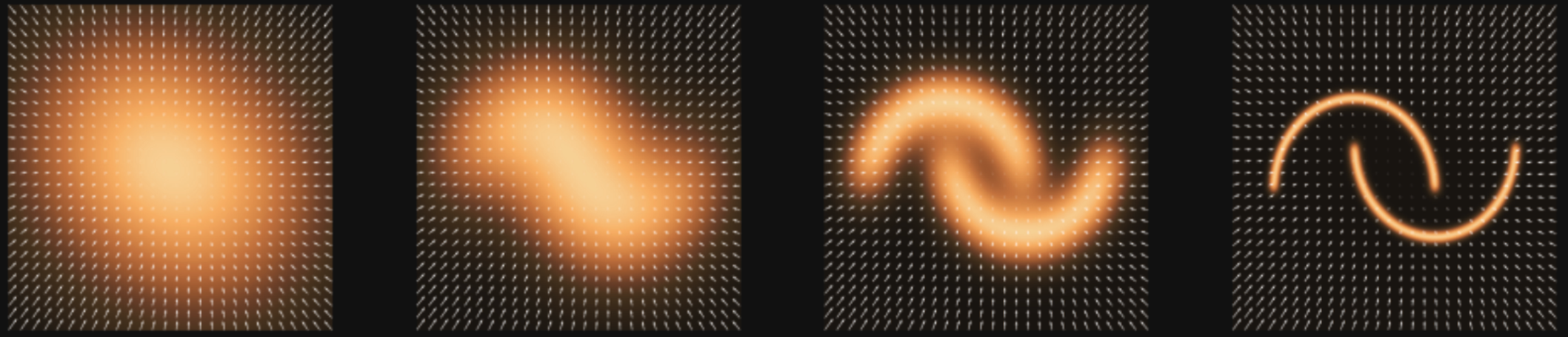
Efficient sampling by Annealed HMC

- Even with gradients, **sampling in high number of dimensions is difficult!** Because of:
  - Curse of dimensionality
  - Highly correlated chains
- $\Longrightarrow$ Use a **parallel annealing strategy** to effectively sample from full distribution.

Efficient sampling by Annealed HMC

- Even with gradients, **sampling in high number of dimensions is difficult!** Because of:
  - Curse of dimensionality
  - Highly correlated chains
- $\Longrightarrow$ Use a **parallel annealing strategy** to effectively sample from full distribution.
- We use the fact that our score network $\mathbf{r}_\theta(x, \sigma)$ is learning a noise-convolved distribution $\nabla \log p_\sigma$, where

$$p_\sigma(x) = \int p_{\text{data}}(x')\mathcal{N}(x|x', \sigma^2)dx', \qquad \sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$$

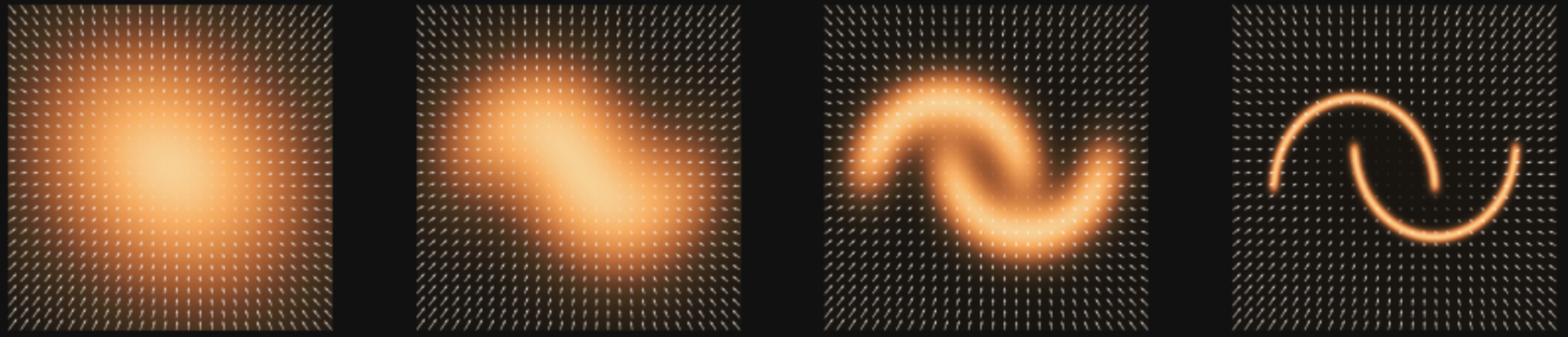


σ_1 σ_2 σ_3 σ_4

Efficient sampling by Annealed HMC

- Even with gradients, **sampling in high number of dimensions is difficult!** Because of:
  - Curse of dimensionality
  - Highly correlated chains
- $\Longrightarrow$ Use a **parallel annealing strategy** to effectively sample from full distribution.
- We use the fact that our score network $\mathbf{r}_\theta(x, \sigma)$ is learning a noise-convolved distribution $\nabla \log p_\sigma$, where

$$p_\sigma(x) = \int p_{\text{data}}(x')\mathcal{N}(x|x', \sigma^2)dx', \qquad \sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$$

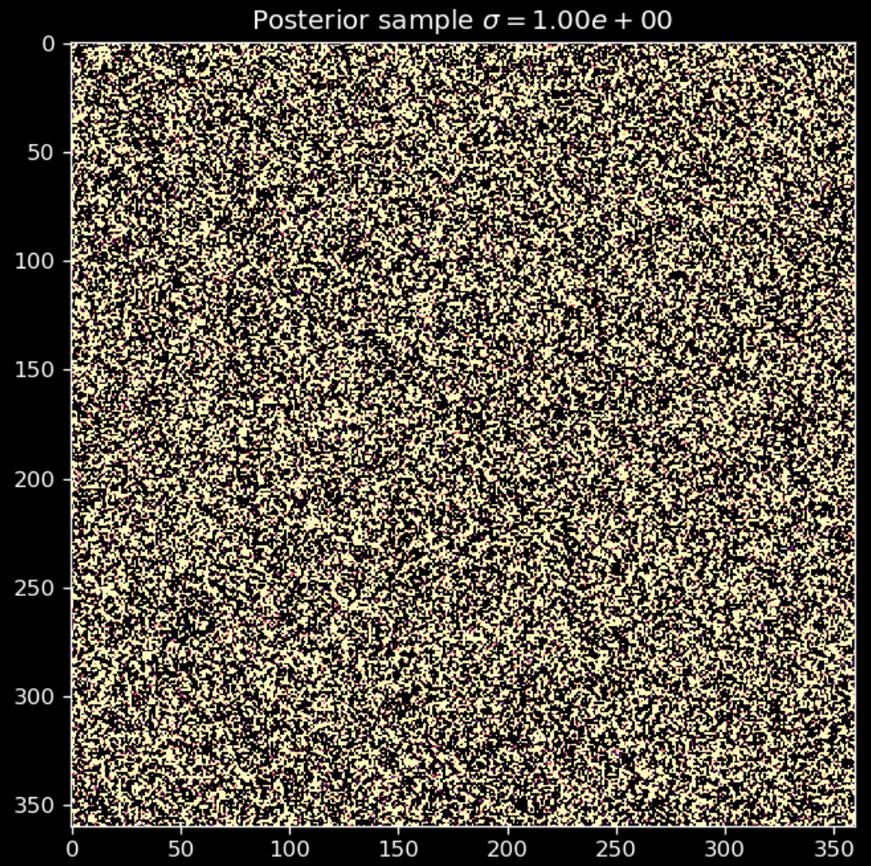


σ_1 σ_2 σ_3 σ_4

- Run many HMC chains in parallel, progressively annealing the σ to 0, **keep last point in the chain as independent sample**.

Illustration on κ-TNG simulations

$$\nabla_\kappa \log p_\sigma(\kappa|\gamma) = \nabla_\kappa \log p_\sigma(\gamma|\kappa) \quad + \quad \nabla_\kappa \log p_\sigma(\kappa)$$


Posterior sample $\sigma = 1.00e+00$
0
50
100
150
200
250
300
350

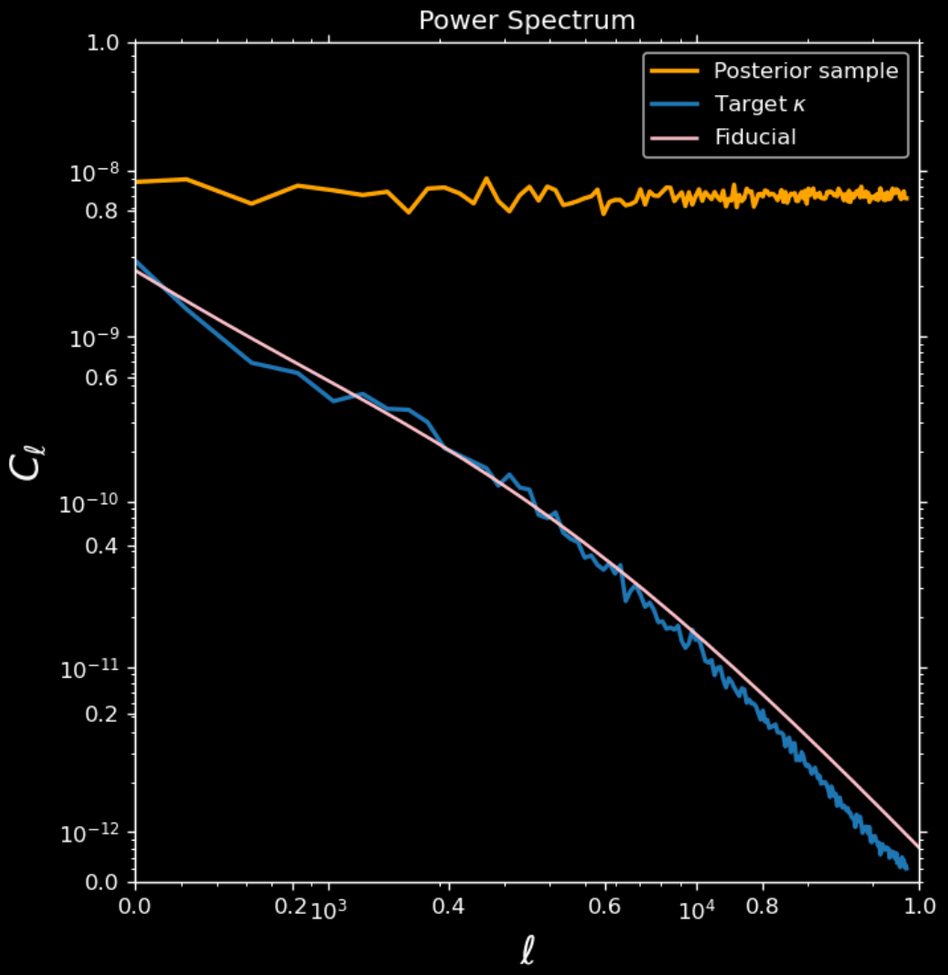
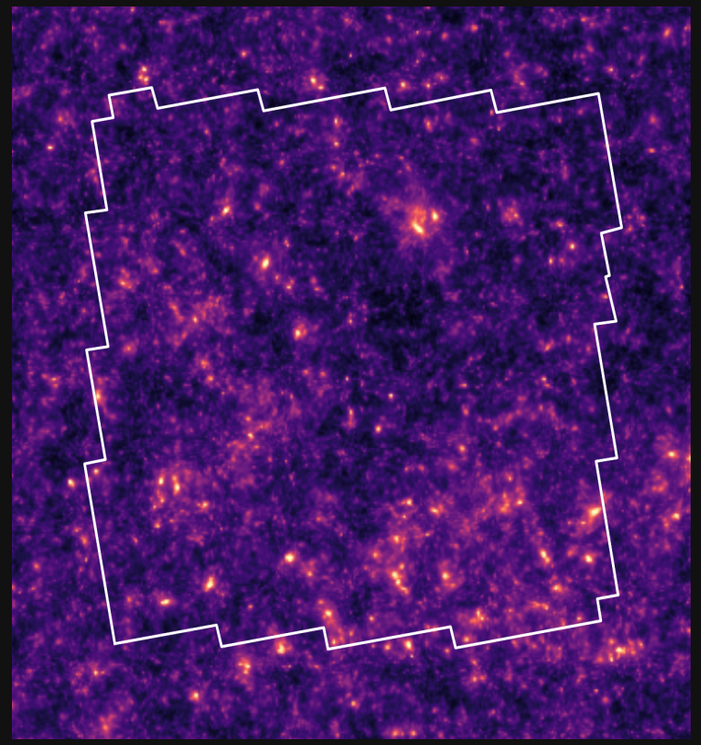
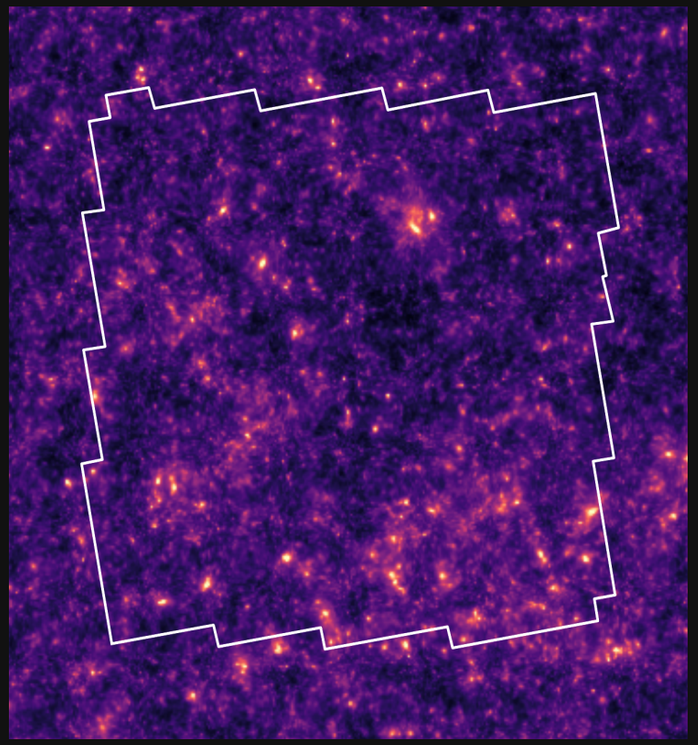

Power Spectrum
Posterior sample
Target κ
Fiducial
C_ℓ
ℓ
1.0
10^{-8}
0.8
10^{-9}
0.6
10^{-10}
0.4
10^{-11}
0.2
10^{-12}
0.0
0.0
0.2
10^3
0.4
0.6
10^4
0.8
1.0

Illustration on κ-TNG simulations

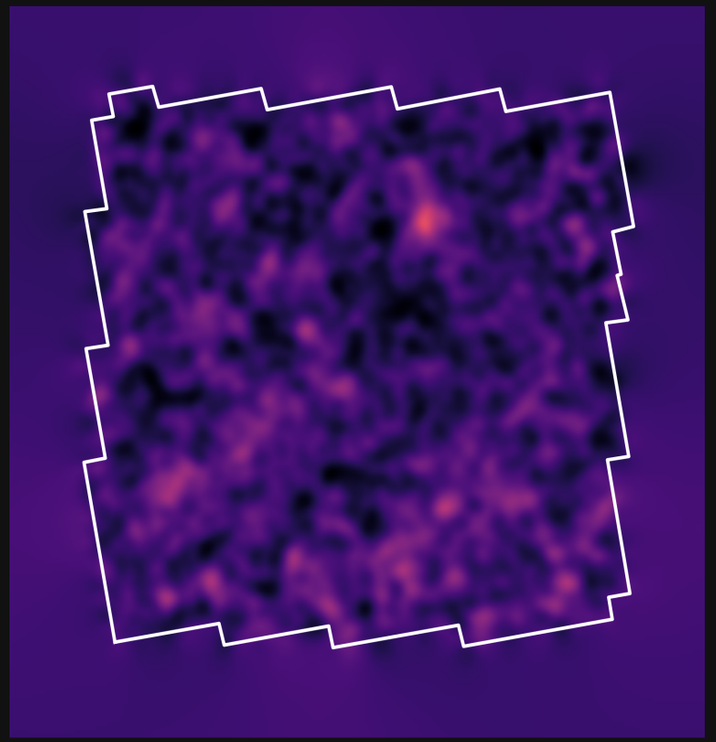


True convergence map

Illustration on κ-TNG simulations

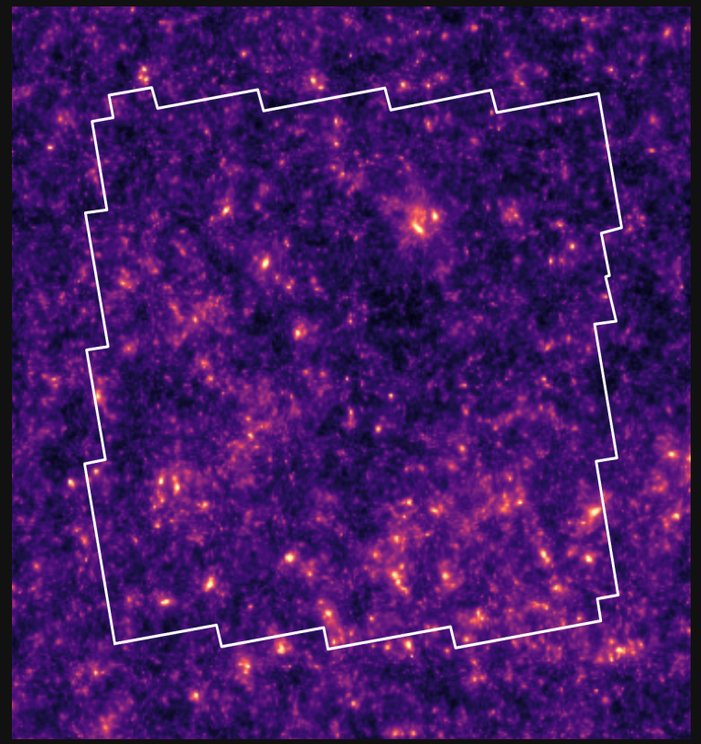


True convergence map

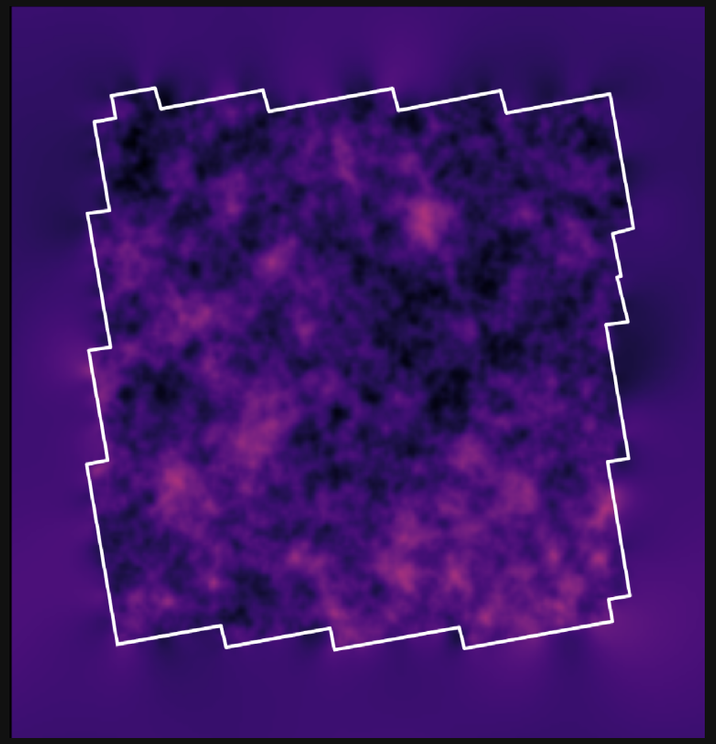


Traditional Kaiser-Squires

Illustration on κ-TNG simulations

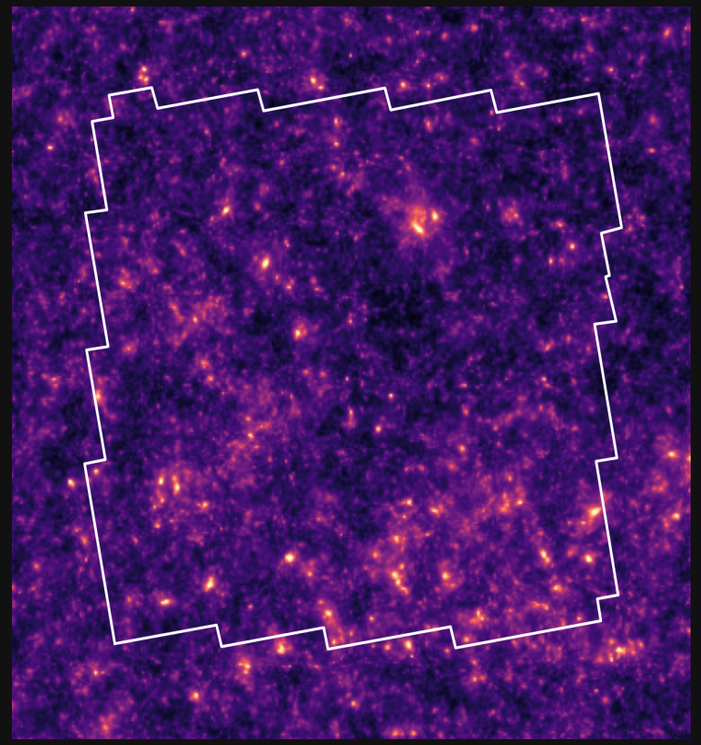


True convergence map

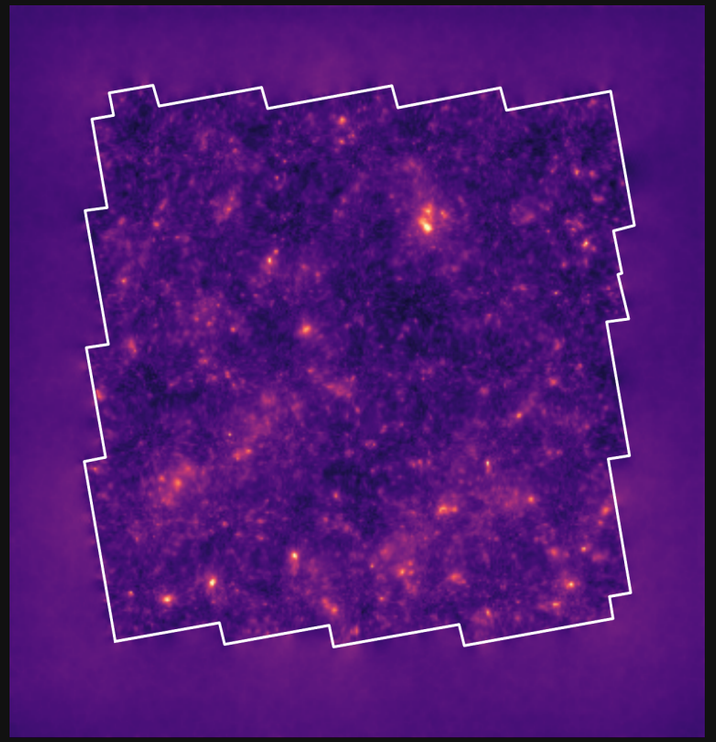


Wiener Filter

Illustration on κ-TNG simulations

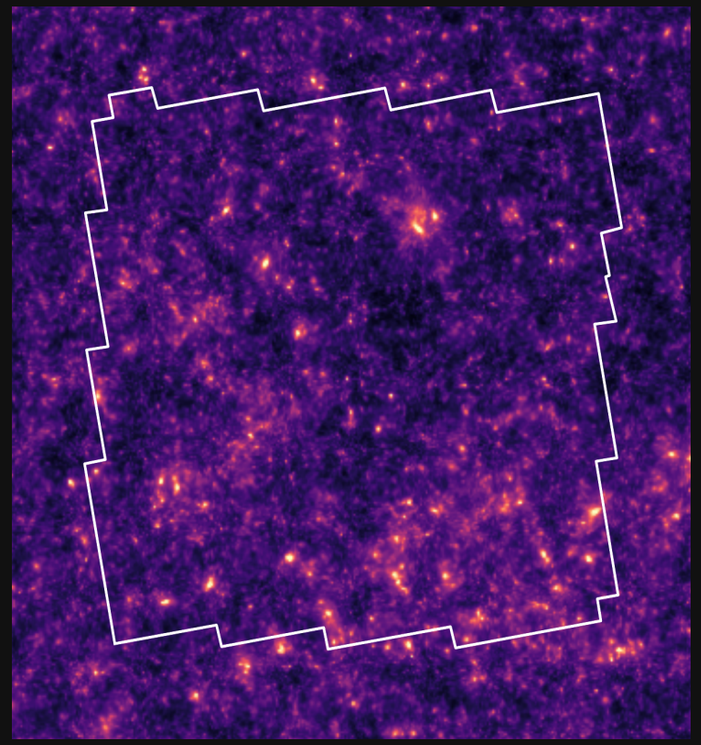


True convergence map

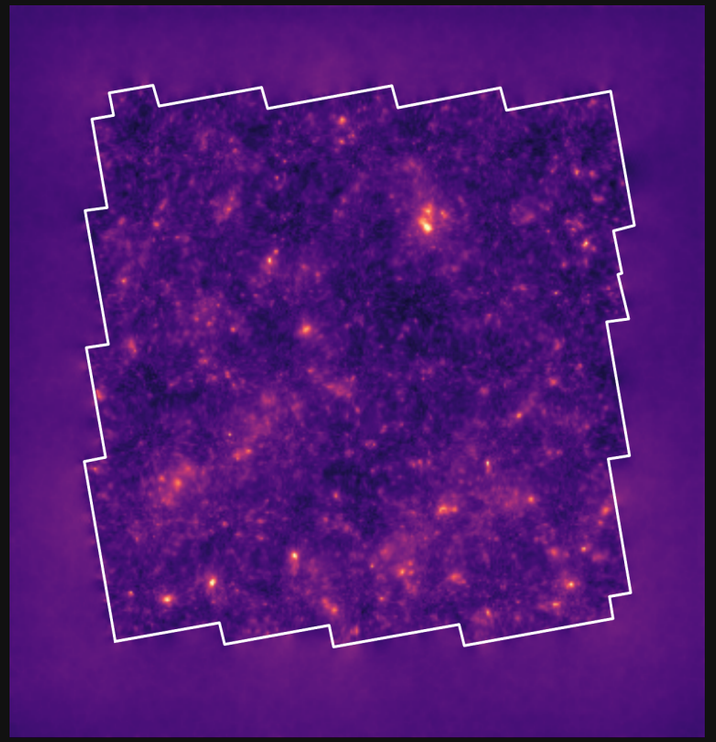


Posterior Mean (ours)

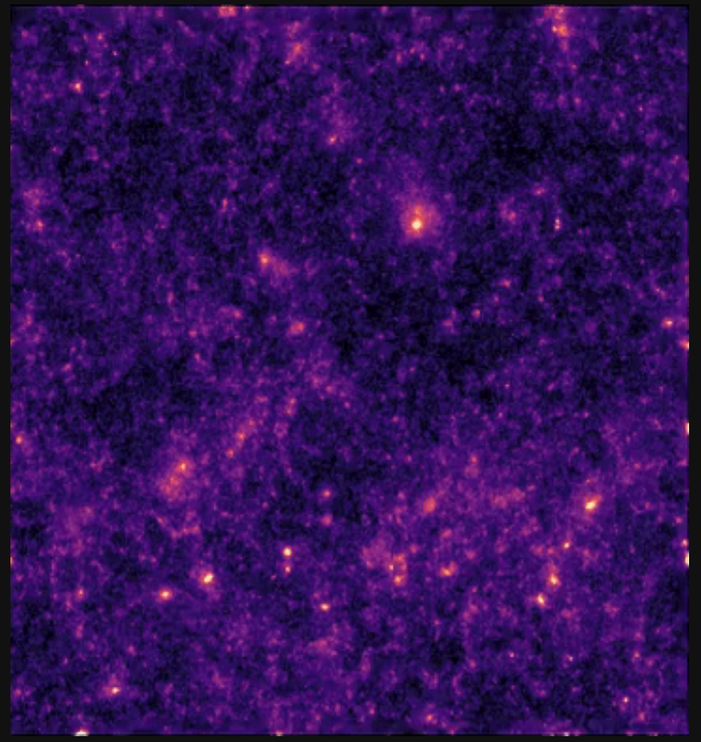
Illustration on κ-TNG simulations



True convergence map

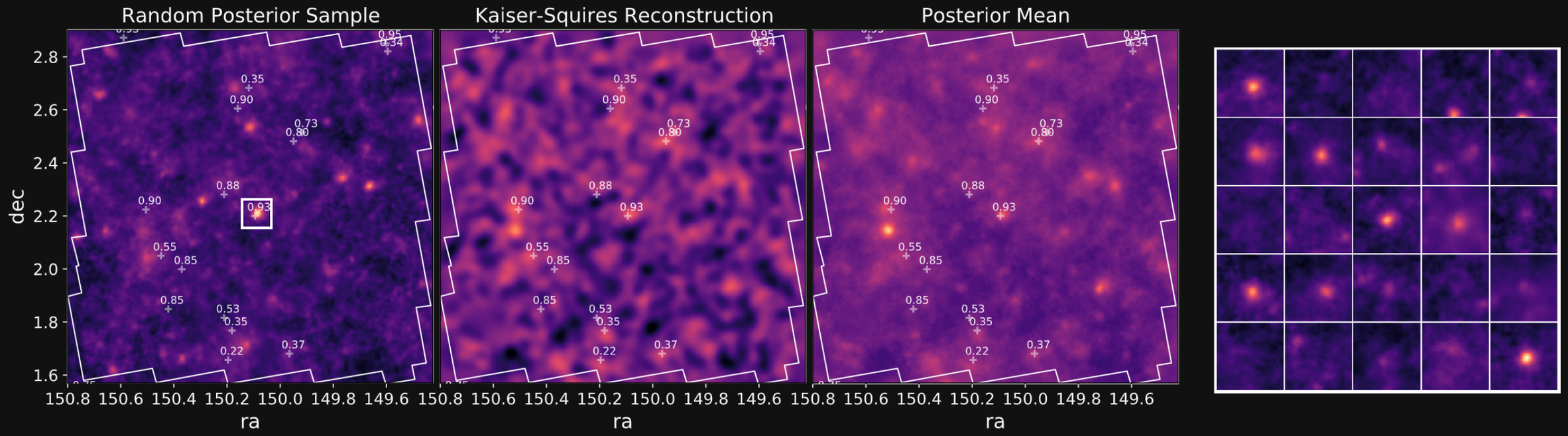


Posterior Mean (ours)



Posterior samples

Probabilistic Mass-Mapping of the HST COSMOS field



- COSMOS shear data from Schrabback et al. 2010
- Prior learned from MassiveNuS at fiducial cosmology (320x320 maps at 0.4 arcsec resolution).
- Known massive X-ray clusters indicated with crosses, along with their redshifts, right pannel shows cutouts of central cluster from multiple posterior samples.

Takeaways

- Hybrid physical/deep learning modeling:
  - Deep generative models can be used to provide **data driven priors**.
  - **Explicit likelihood**, uses of all of our physical knowledge.
    $\Longrightarrow$ The method can be applied for varying PSF, noise, or even different instruments!

Takeaways

- Hybrid physical/deep learning modeling:
  - Deep generative models can be used to provide **data driven priors**.
  - **Explicit likelihood**, uses of all of our physical knowledge.
    $\Longrightarrow$ The method can be applied for varying PSF, noise, or even different instruments!
- Neural Score Estimation is a **scalable approach** to learn a prior score.

Takeaways

- Hybrid physical/deep learning modeling:
  - Deep generative models can be used to provide **data driven priors**.
  - **Explicit likelihood**, uses of all of our physical knowledge.
    $\Longrightarrow$ The method can be applied for varying PSF, noise, or even different instruments!
- Neural Score Estimation is a **scalable approach** to learn a prior score.
- Knowledge of the posterior score is all we need for Bayesian inference aka uncertain quantification.

Takeaways

- Hybrid physical/deep learning modeling:
  - Deep generative models can be used to provide **data driven priors**.
  - **Explicit likelihood**, uses of all of our physical knowledge.
    $\Longrightarrow$ The method can be applied for varying PSF, noise, or even different instruments!
- Neural Score Estimation is a **scalable approach** to learn a prior score.
- Knowledge of the posterior score is all we need for Bayesian inference aka uncertain quantification.
- We implemented a new class of mass mapping method, providing the full posterior
  $\Longrightarrow$ recovered a very high quality convergence map of the COSMOS field.

Takeaways

- Hybrid physical/deep learning modeling:
  - Deep generative models can be used to provide **data driven priors**.
  - **Explicit likelihood**, uses of all of our physical knowledge.
    $\Longrightarrow$ The method can be applied for varying PSF, noise, or even different instruments!
- Neural Score Estimation is a **scalable approach** to learn a prior score.
- Knowledge of the posterior score is all we need for Bayesian inference aka uncertain quantification.
- We implemented a new class of mass mapping method, providing the full posterior
  $\Longrightarrow$ recovered a very high quality convergence map of the COSMOS field.

Thank you!