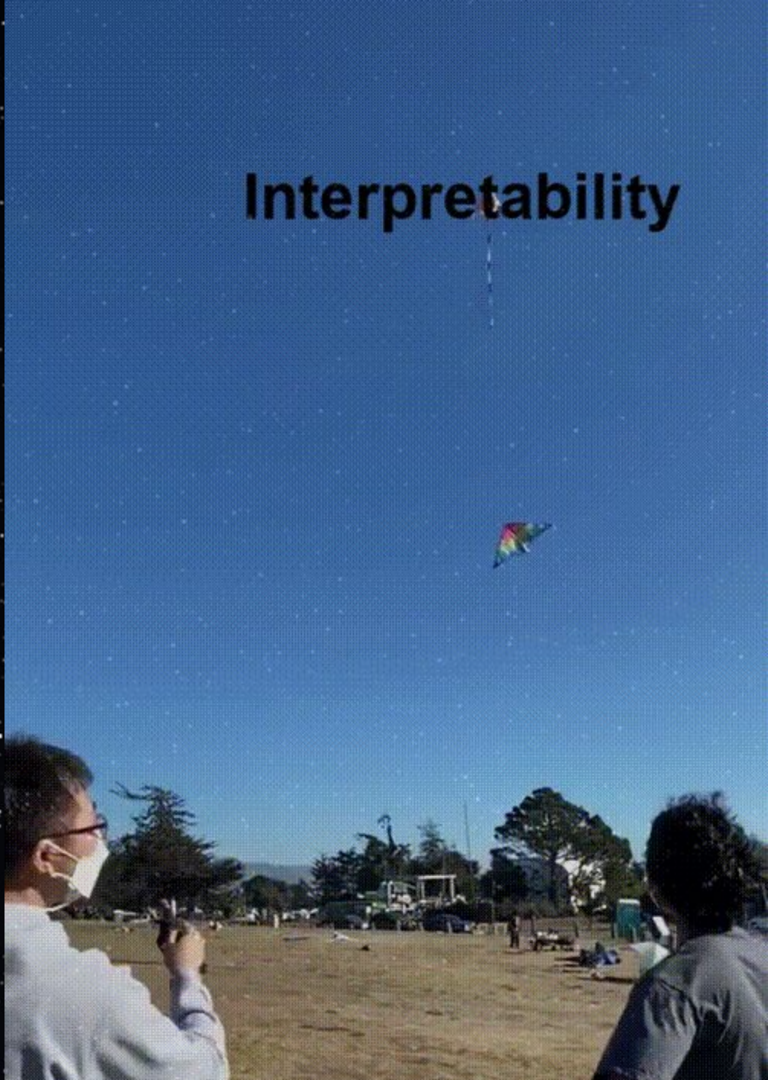
Adaptive wavelet distillation from neural networks through interpretations

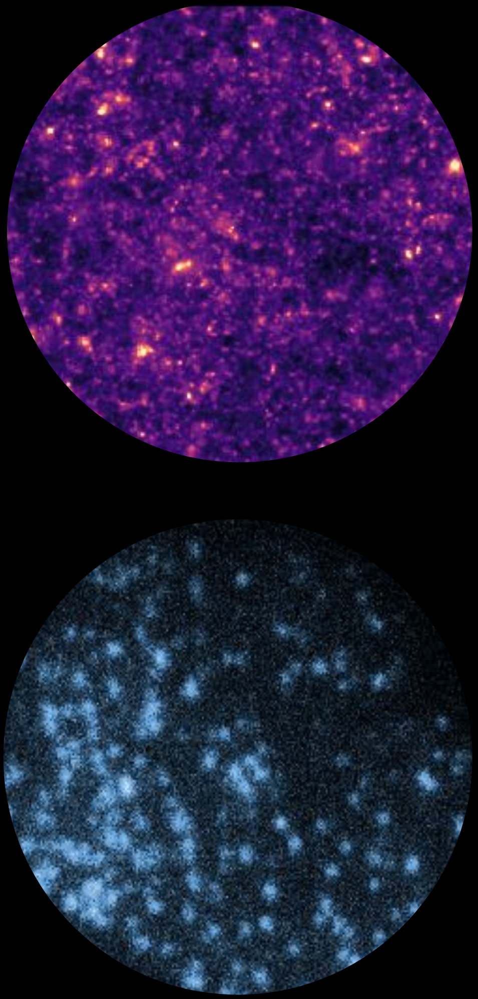


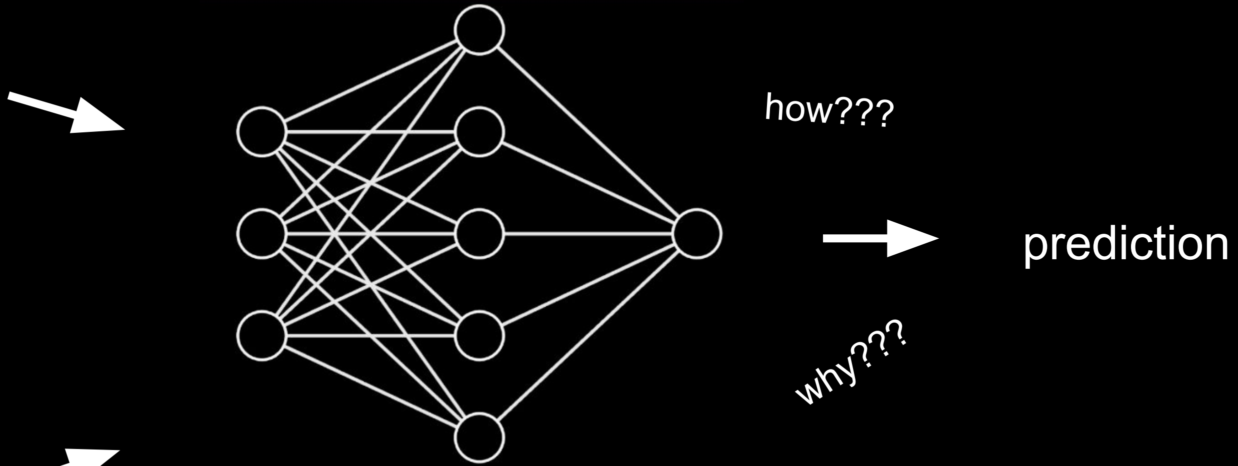


UC Berkeley

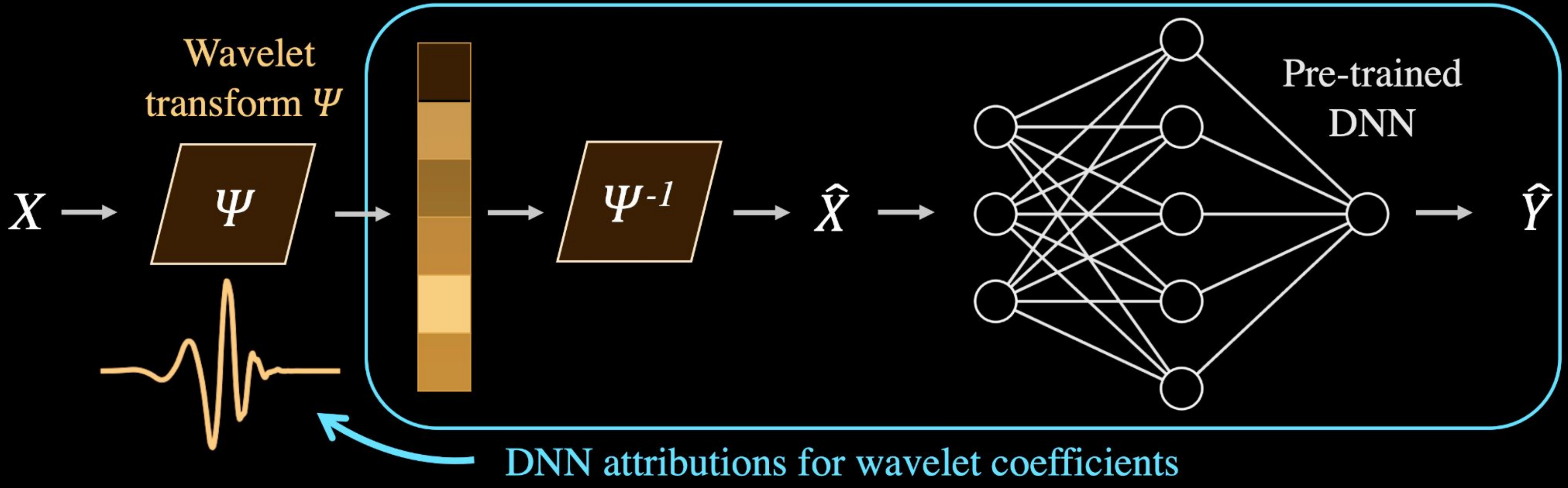


a deep problem: interpretability



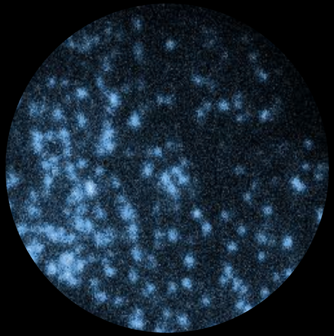


a shallow solution: wavelets



$$\underset{h,g}{\text{minimize}}\ \mathcal{L}(h,g) = \underbrace{\frac{1}{m}\sum_i \|x_i - \widehat{x}_i\|_2^2}_{\text{Reconstruction loss}} + \underbrace{\frac{1}{m}\sum_i W(h,g,x_i;\lambda)}_{\text{Wavelet loss}} + \underbrace{\gamma \sum_i \|\text{TRIM}_{\Psi,f}(\Psi x_i)\|_1}_{\text{Interpretation loss}},$$

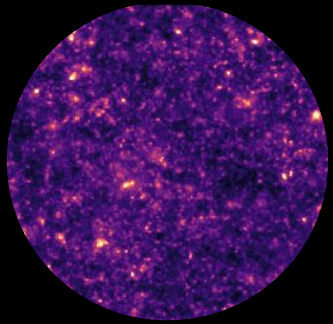
results (spoilers!)



cell-biology: CME event prediction

LSTM	SOA Baseline
0.237	0.197

R^2 (higher is better)



cosmological parameter prediction

Resnet	SOA Baseline
1.156	1.259

RMSE (lower is better)

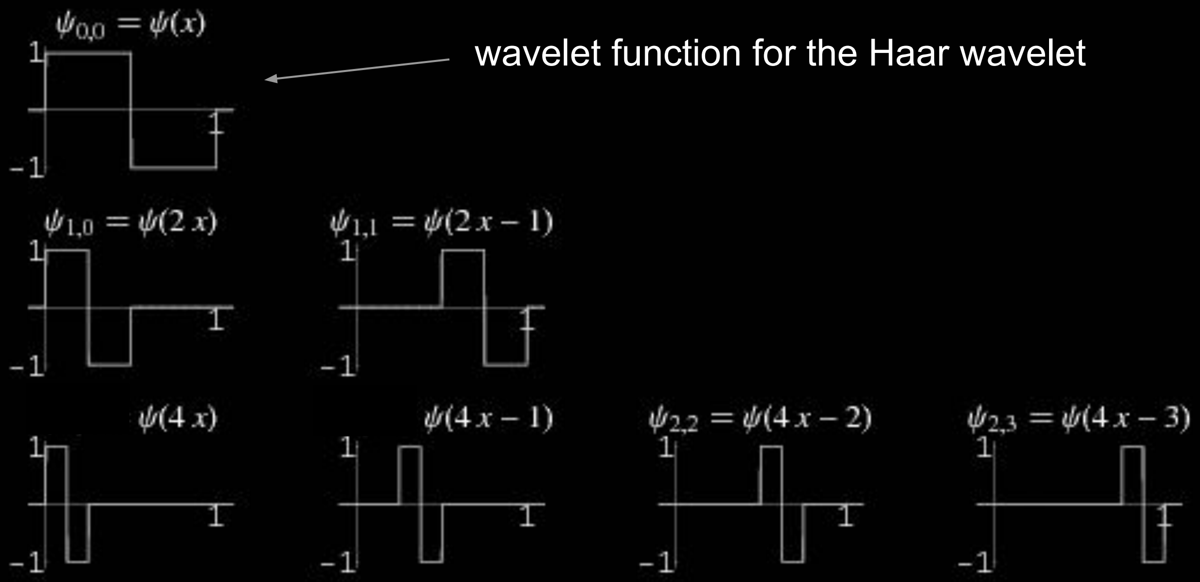
wavelet transform: multi-scale, spatially localized

wavelet function for the Haar wavelet

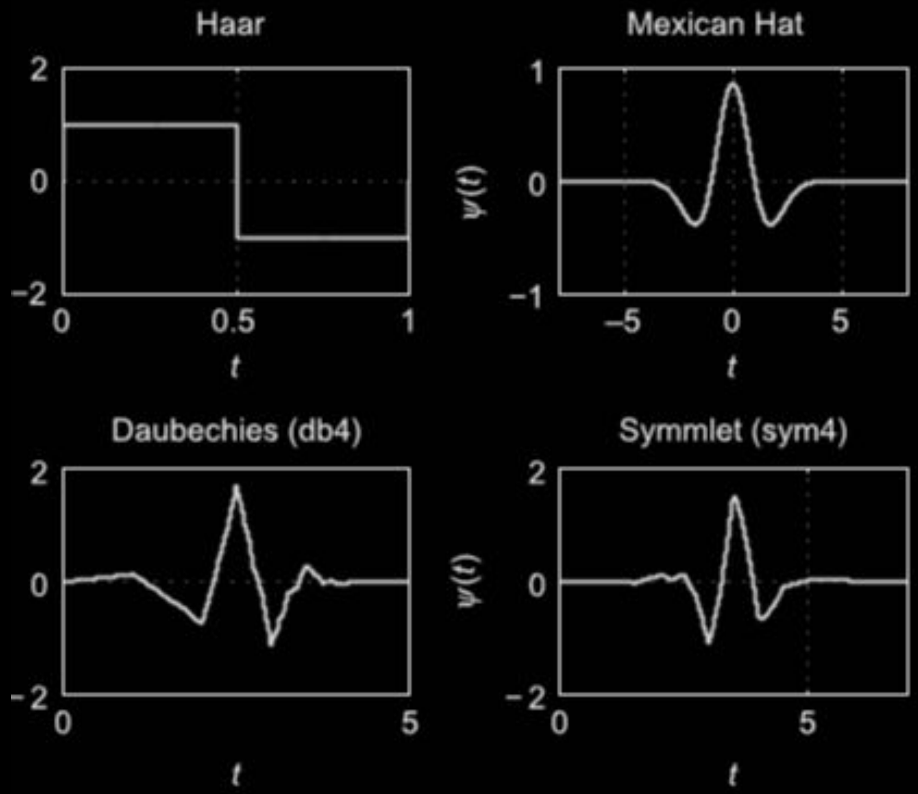
level=0

level=1

level=2

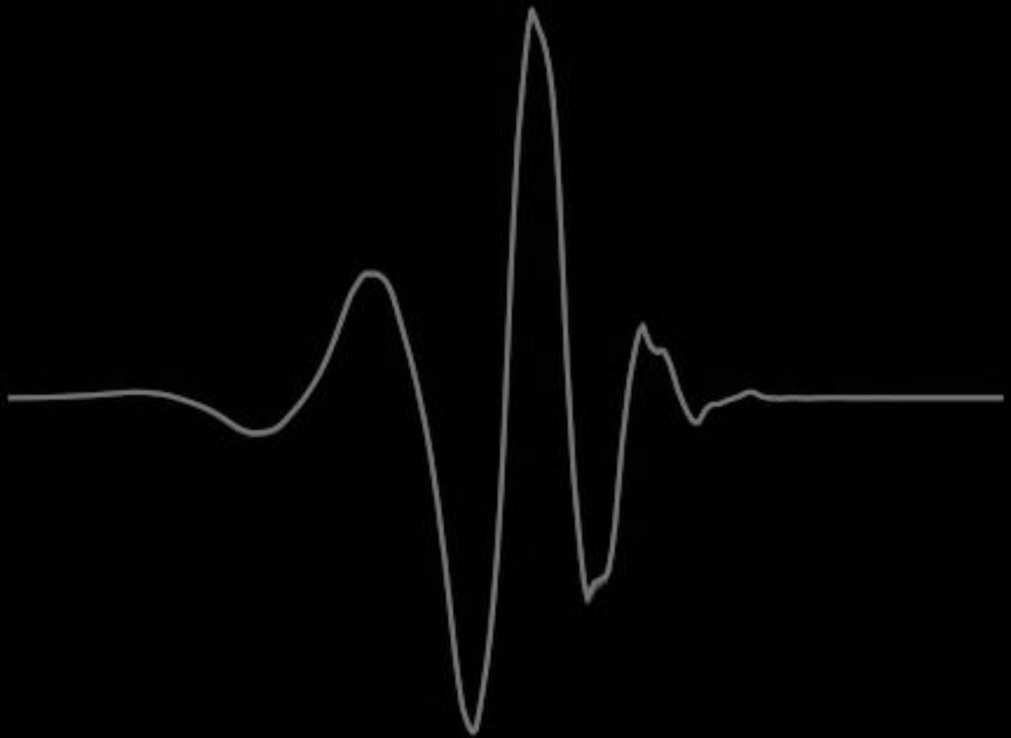


wavelet function can vary



adaptive wavelet

$$\underset{h,g}{\text{minimize}}\ \mathcal{L}(h,g) = \underbrace{\frac{1}{m}\sum_i \|x_i - \widehat{x}_i\|_2^2}_{\text{Reconstruction loss}} + \underbrace{\frac{1}{m}\sum_i W(h,g,x_i;\lambda)}_{\text{Wavelet loss}}$$



orthonormal basis under following conditions (mallat, 1998):

$$|\widehat{h}(w)|^2 + |\widehat{h}(w+\pi)|^2 = 2 \ \ \forall w$$

$$\sum h_i = \sqrt{2}$$

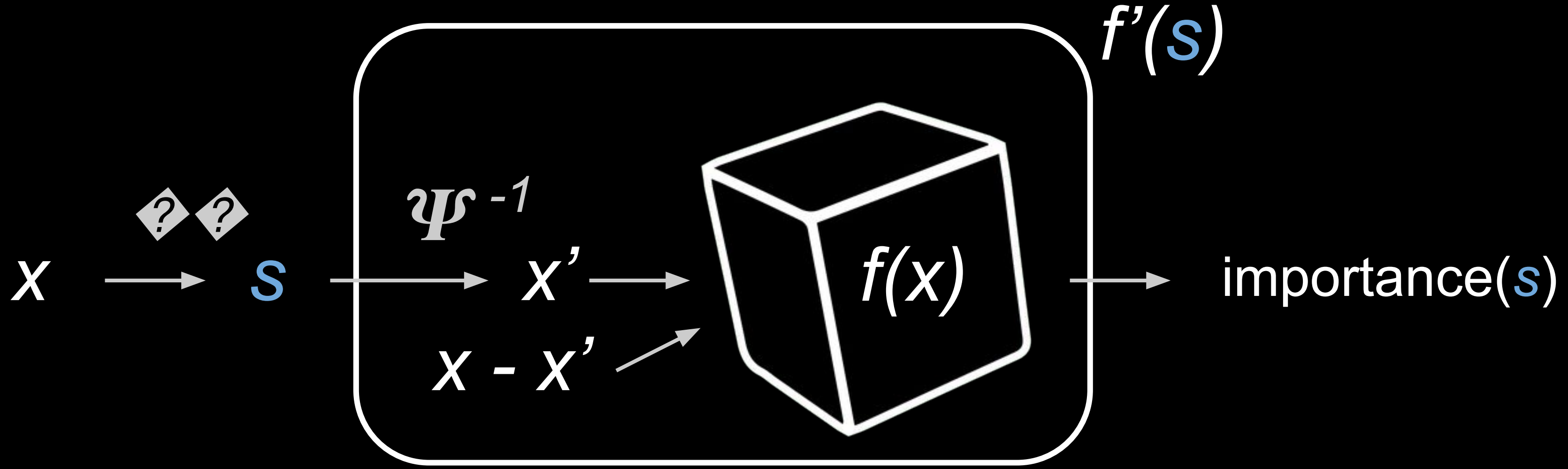
$$g(n) = (-1)^n h(N-1-n)$$

$\|\Phi f\|_1$ small

adaptive wavelet **+ distillation**

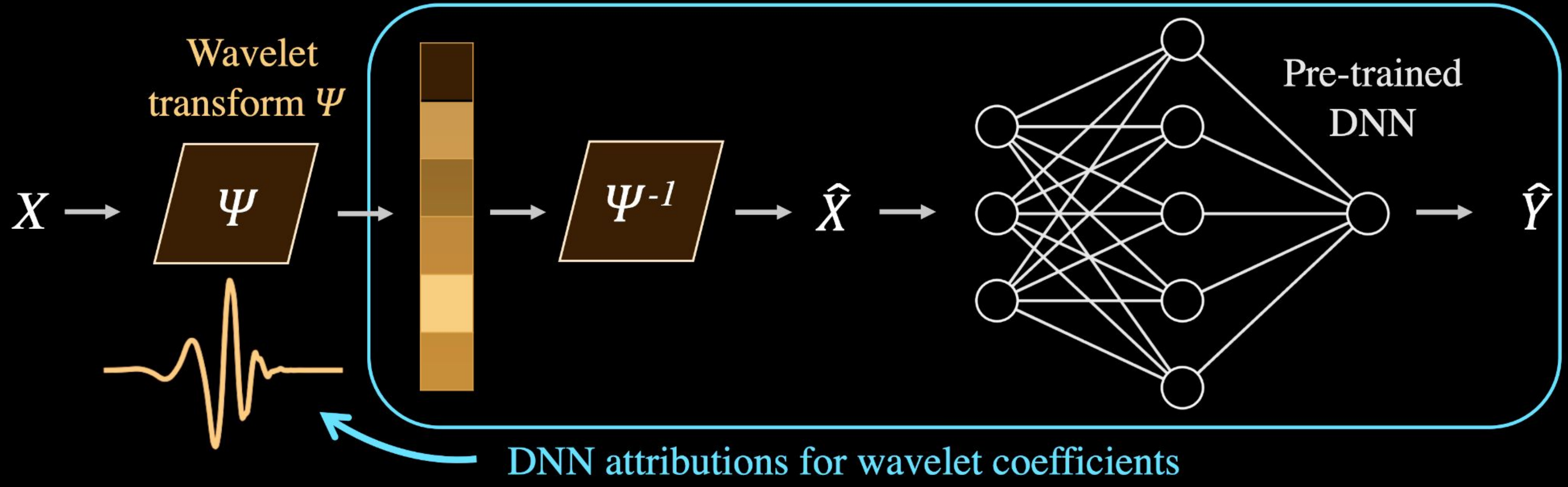
$$\underset{h,g}{\text{minimize}}\ \mathcal{L}(h,g) = \underbrace{\frac{1}{m}\sum_i \|x_i - \widehat{x}_i\|_2^2}_{\text{Reconstruction loss}} + \underbrace{\frac{1}{m}\sum_i W(h,g,x_i;\lambda)}_{\text{Wavelet loss}} + \boxed{\underbrace{\gamma \sum_i \|\text{TRIM}_{\Psi,f}(\Psi x_i)\|_1}_{\text{Interpretation loss}}},$$

transformation importance



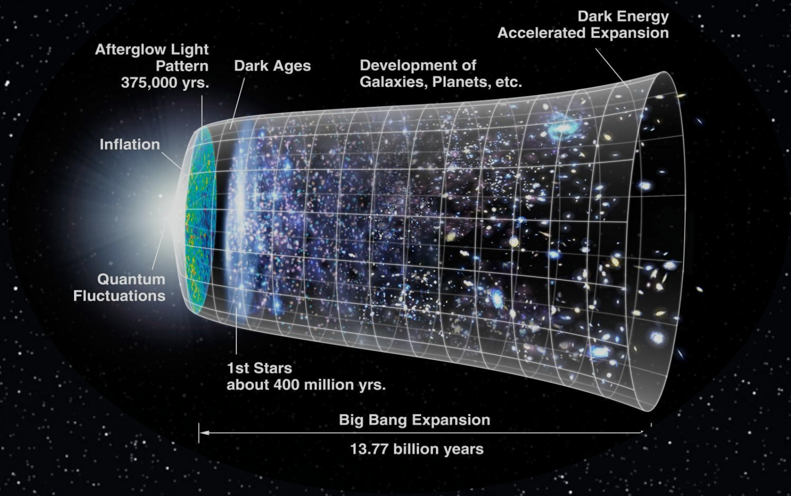
singh, ha, lanusse, boehm, liu, & yu, 2019
“transformation importance with applications to cosmology”

putting it all together

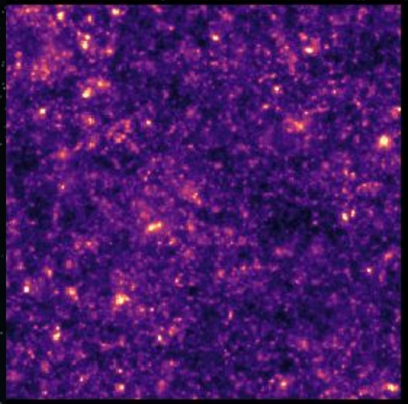


$$\underset{h,g}{\text{minimize}}\ \mathcal{L}(h,g) = \underbrace{\frac{1}{m}\sum_i \|x_i - \widehat{x}_i\|_2^2}_{\text{Reconstruction loss}} + \underbrace{\frac{1}{m}\sum_i W(h,g,x_i;\lambda)}_{\text{Wavelet loss}} + \underbrace{\gamma\sum_i \|\text{TRIM}_{\Psi,f}(\Psi x_i)\|_1}_{\text{Interpretation loss}},$$

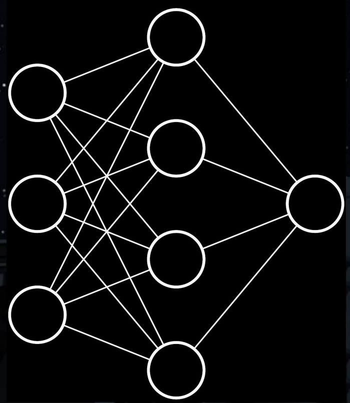
cosmology problem
(more in the paper)



$\Omega_m \rightarrow$

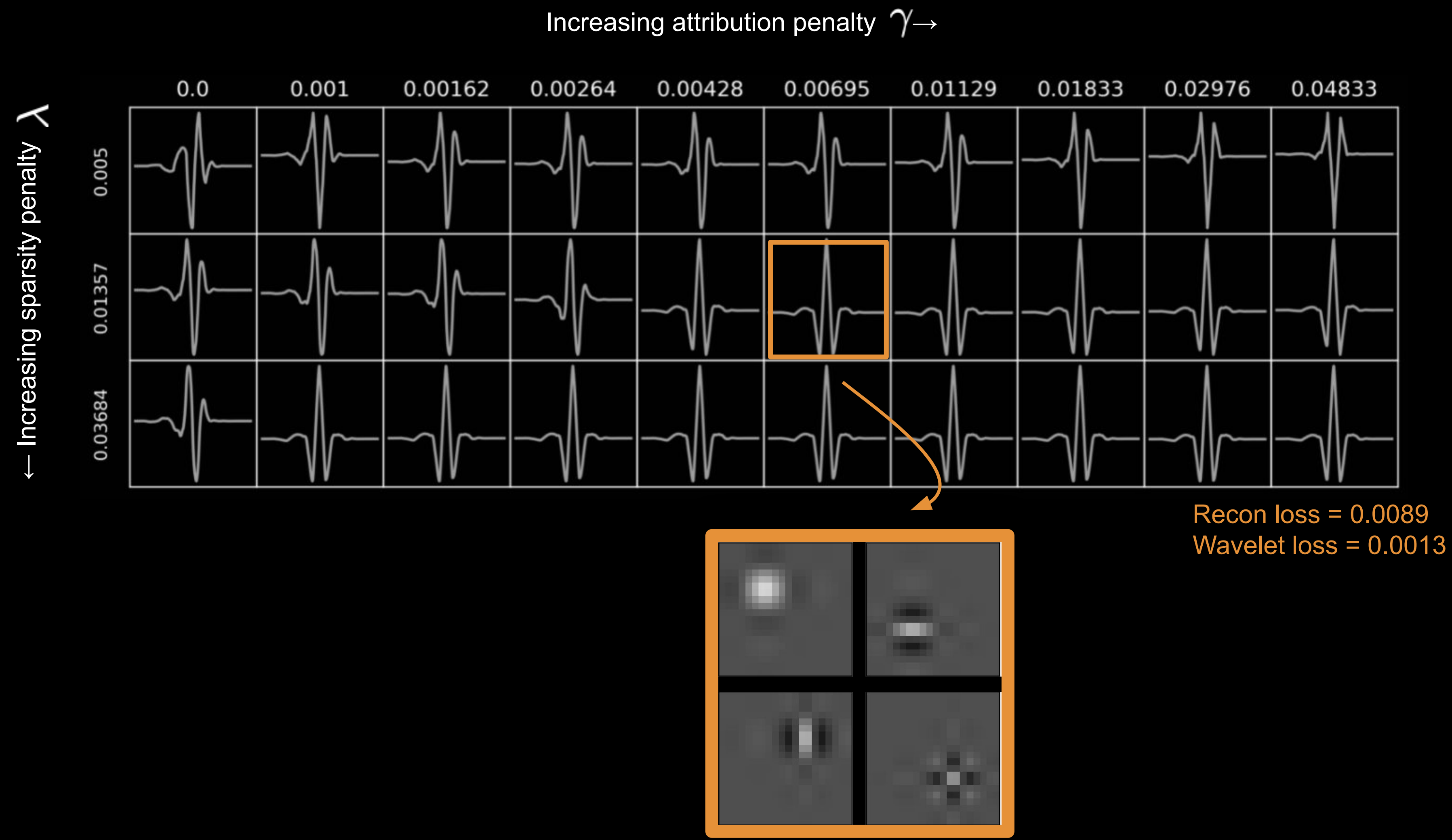


$\rightarrow$



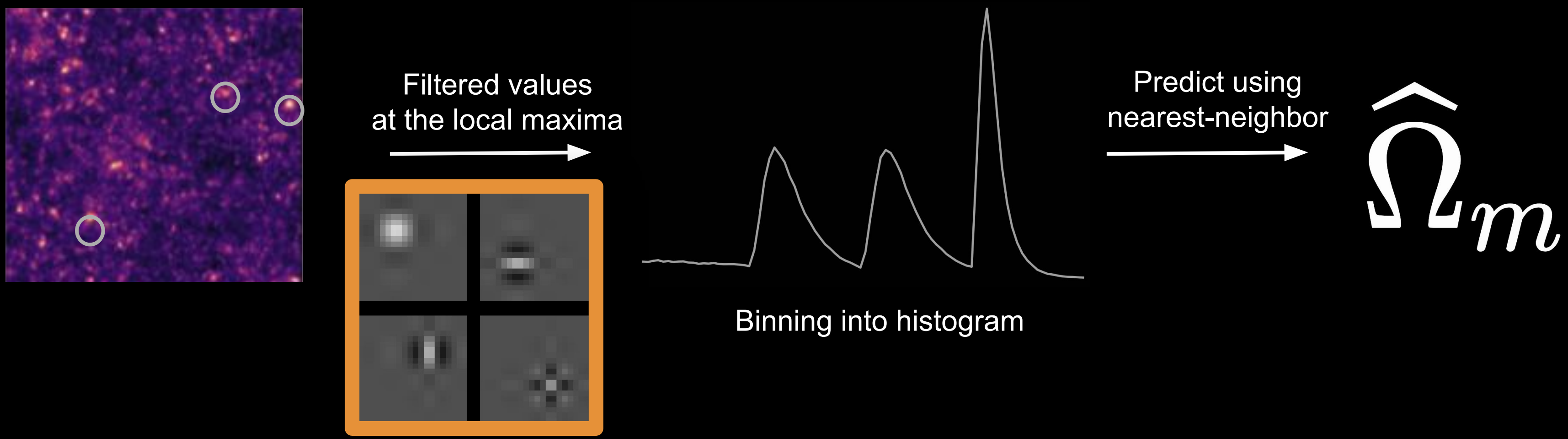
$\rightarrow \hat{\Omega}_m$

Large Synoptic Survey Telescope



Increasing attribution penalty $\gamma \rightarrow$
0.0
0.001
0.00162
0.00264
0.00428
0.00695
0.01129
0.01833
0.02976
0.04833
$\leftarrow$ Increasing sparsity penalty λ
0.005
0.01357
0.03684
Recon loss = 0.0089
Wavelet loss = 0.0013

Predicting via peak counts



Prediction error for Ω_m (RMSE)

AWD	Resnet	AWD no interp. loss	
1.029	1.156	1.354	x 10^{-4}

Peak Height	Laplace*	Roberts-Cross*	DB5	
1.609	1.369	1.259	1.569	x 10^{-4}

*Ribli et al (2019) Nature Astronomy