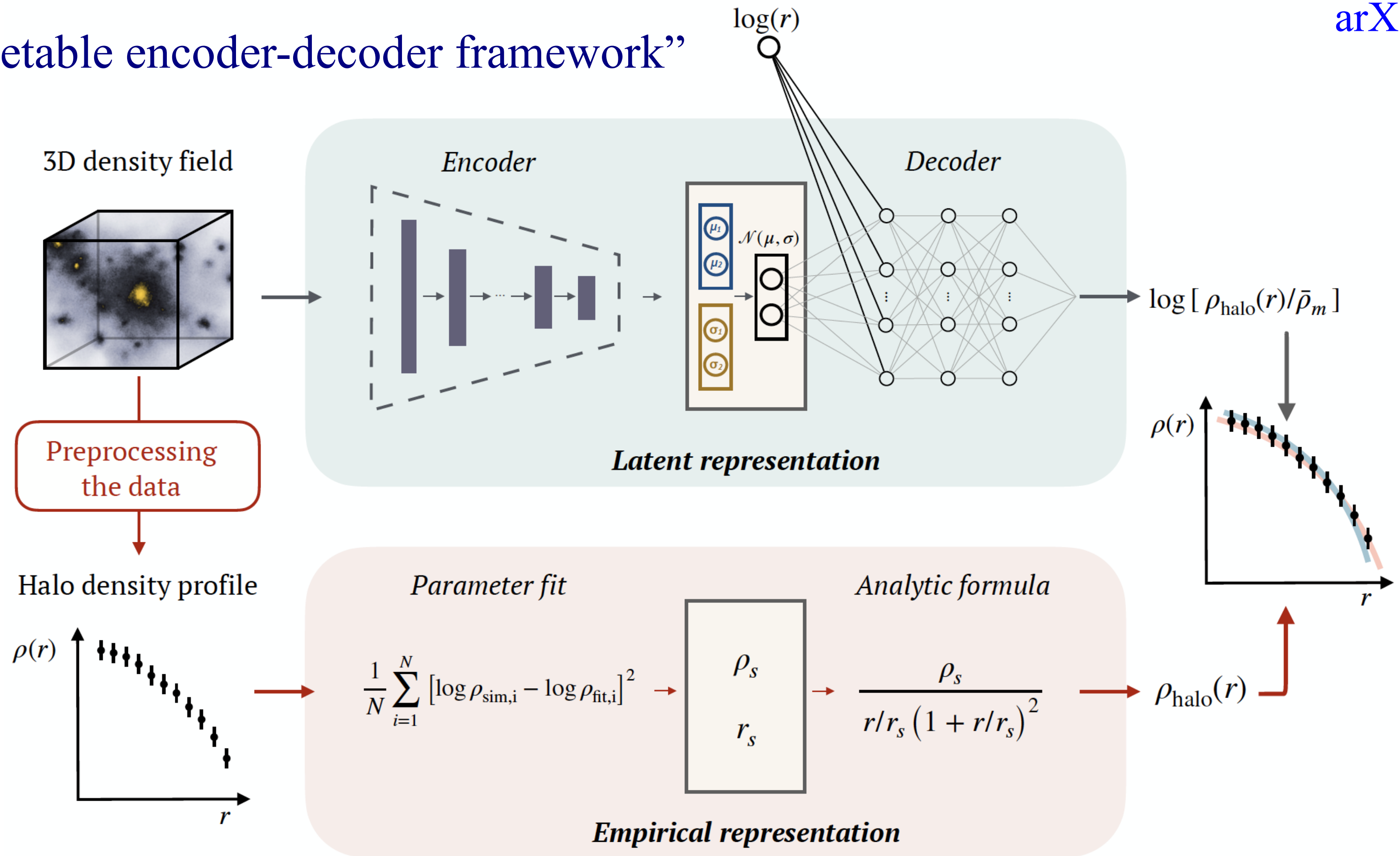
Interpretable machine learning?

Discovering the building blocks of dark matter halo density profiles with neural networks

Luisa Lucie-Smith,[1,*] Hiranya V. Peiris,[2,3] Andrew Pontzen,[2] Brian Nord,[4,5,6] Jeyan Thiyagalingam,[7] and Davide Piras[2]

arXiv:2203.0082

"An interpretable encoder-decoder framework"



Goodness of fit

$$\mathcal{L}=\mathcal{L}_{\rm pred}(\boldsymbol{\rho}_{\rm true},\boldsymbol{\rho}_{\rm pred})+\beta\,\mathcal{D}_{\rm KL}[p_\phi(\boldsymbol{z}|\boldsymbol{x});q(\boldsymbol{z})],$$

Disentanglement of latents

$$\mathcal{L}_{\rm pred}=\frac{1}{N}\sum_{i=1}^{N}\left(\log_{10}\rho_{i,\rm true}-\log_{10}\rho_{i,\rm pred}\right)^2,\qquad \mathcal{D}_{\rm KL}(\mathcal{N}(\mu_{\boldsymbol{z}},\sigma_{\boldsymbol{z}});\mathcal{N}(0,1))=-\frac{1}{2}\sum_{i=1}^{L}\left[1+2\log\sigma_i-\mu_i^2-\sigma_i^2\right].$$

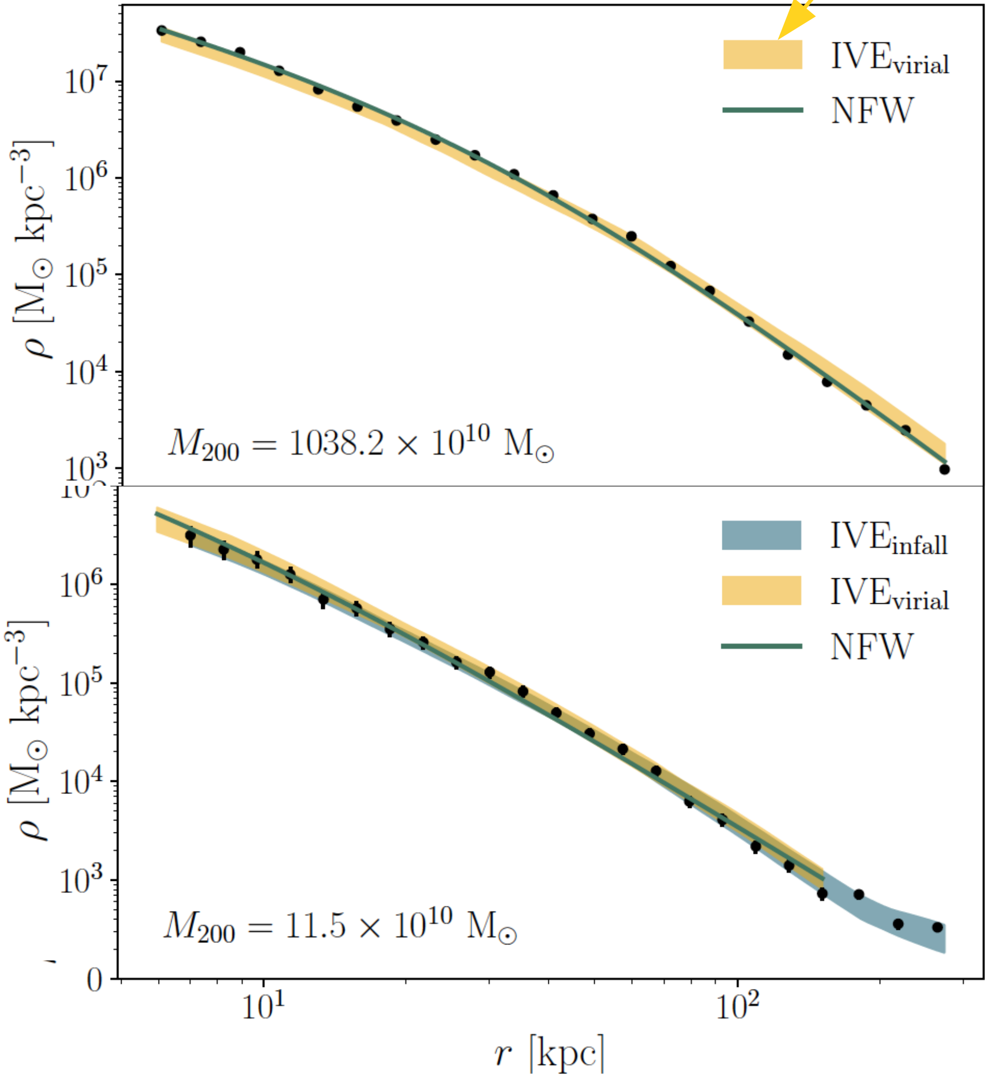
NFW96 insights into z=0 halo density profiles

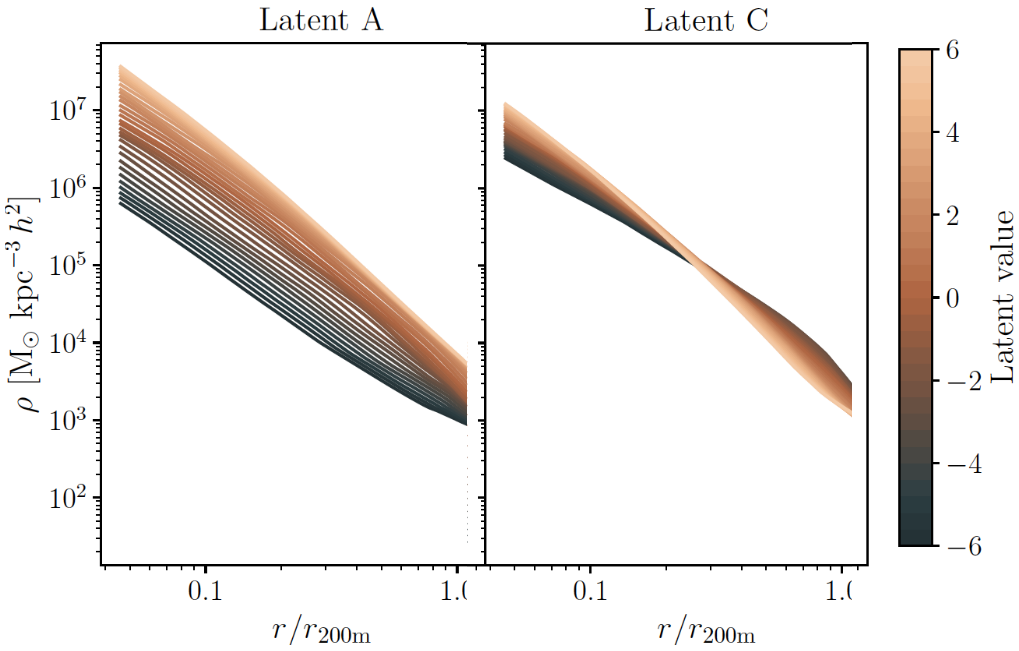
I. Spherically averaged halo density profiles within R200 can be fit over the resolved radial range to within the noise due to substructure and counting statistics by a smooth function of just two variables.

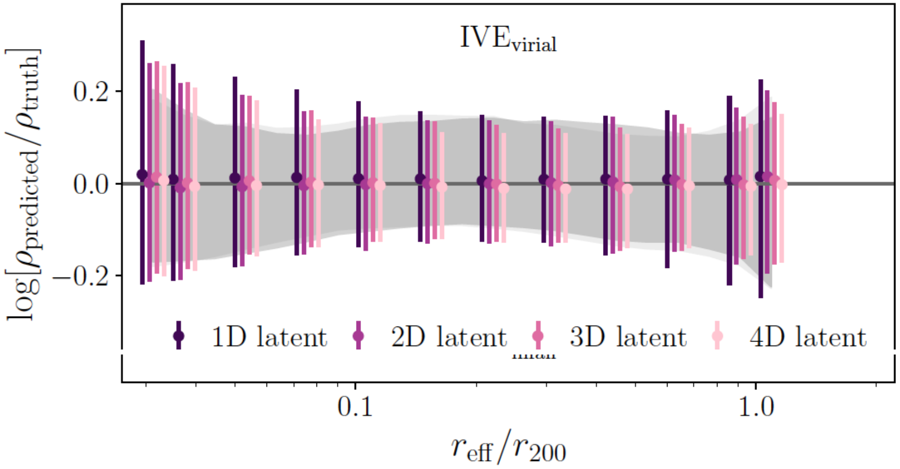
II. Profiles are homologous: they can be fit by a "universal" curve, with the two parameters corresponding to a characteristic radius and a characteristic density, hence to offsets of the universal curve parallel to the *x*- and *y*-axes in a log-log plot.

III. The characteristic densities and radii are correlated: bigger halos are less dense.

DL, r < R200


IVE$_{\rm virial}$
NFW
$M_{200} = 1038.2 \times 10^{10}$ M$_\odot$
IVE$_{\rm infall}$
IVE$_{\rm virial}$
NFW
$M_{200} = 11.5 \times 10^{10}$ M$_\odot$
ρ [M$_\odot$ kpc^{-3}]
10^7
10^6
10^5
10^4
10^3
0
10^1
10^2
r [kpc]


Latent A
Latent C
ρ [M$_\odot$ kpc^{-3} h^2]
10^7
10^6
10^5
10^4
10^3
10^2
0.1
1.0
$r/r_{\rm 200m}$
Latent value
6
4
2
0
−2
−4
−6


IVE$_{\rm virial}$
$\log[\rho_{\rm predicted}/\rho_{\rm truth}]$
0.2
0.0
−0.2
1D latent
2D latent
3D latent
4D latent
0.1
1.0
$r_{\rm eff}/r_{200}$

NFW96 insights into z=0 halo density profiles

I. Spherically averaged halo density profiles within R200 can be fit over the resolved radial range to within the noise due to substructure and counting statistics by a smooth function of just two variables. ✓

II. Profiles are homologous; they can be fit by a "universal" curve, with the two parameters corresponding to a characteristic radius and a characteristic density, hence to offsets of the universal curve parallel to the *x*- and *y*-axes in a log-log plot.

III. The characteristic densities and radii are correlated: bigger halos are less dense.

NFW96 insights into z=0 halo density profiles

I. Spherically averaged halo density profiles within R200 can be fit over the resolved radial range to within the noise due to substructure and counting statistics by a smooth function of just two variables. ✓

II. Profiles are homologous; they can be fit by a "universal" curve, with the two parameters corresponding to a characteristic radius and a characteristic density, hence to offsets of the universal curve parallel to the *x*- and *y*-axes in a log-log plot. ✗

III. The characteristic densities and radii are correlated: bigger halos are less dense.

NFW96 insights into z=0 halo density profiles

I. Spherically averaged halo density profiles within R200 can be fit over the resolved radial range to within the noise due to substructure and counting statistics by a smooth function of just two variables. ✓

II. Profiles are homologous; they can be fit by a "universal" curve, with the two parameters corresponding to a characteristic radius and a characteristic density, hence to offsets of the universal curve parallel to the *x*- and *y*-axes in a log-log plot. ✗

III. The characteristic densities and radii are correlated: bigger halos are less dense. ✗

“Failure” is, in part, a consequence of the disentanglement requirement