Current progress and challenges from the CAMELS project

(Cosmology and Astrophysics with MachinE Learning Simulations)

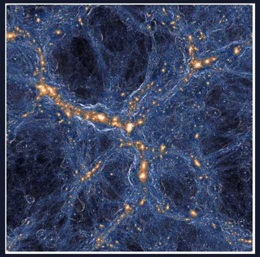
Core team:
Francisco Villaescusa-Navarro
Daniel Anglés-Alcázar
Shy Genel

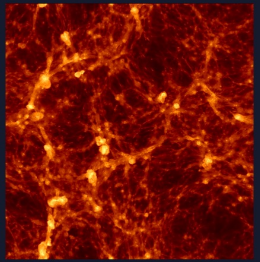
Daniel Anglés-Alcázar
Department of Physics, University of Connecticut

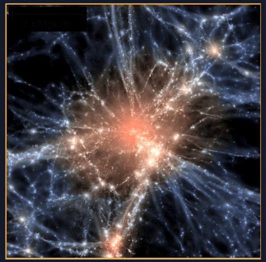
Debating the potential of machine learning in astronomical surveys #2 - ML-IAP/CCA-2023
Flatiron institute / IAP, November 27th - December 1st, 2023

Cosmology and Astrophysics with MachinE Learning Simulations

Core team:
Francisco Villaescusa-Navarro
Daniel Anglés-Alcázar
Shy Genel

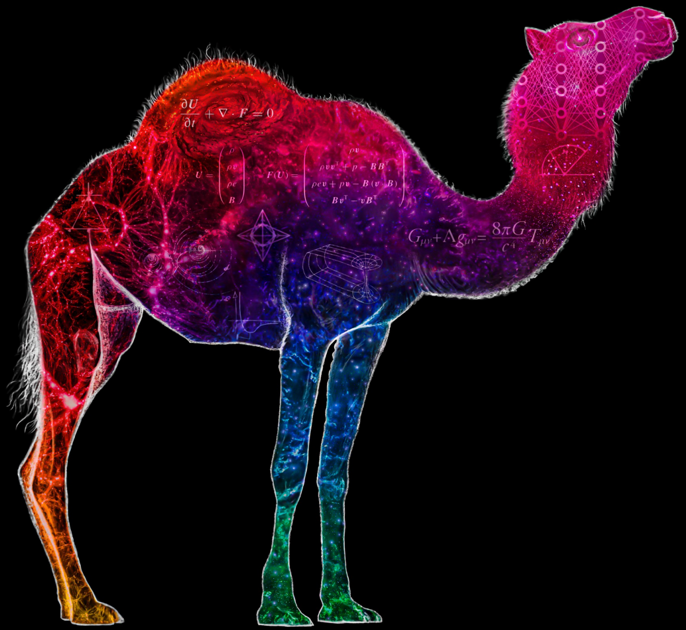

IllustrisTNG team


SIMBA team


Astrid team

SMAUG collaboration

+ SWIFT-Eagle
+ Magneticum
+ Ramses
+ Enzo
+ ...



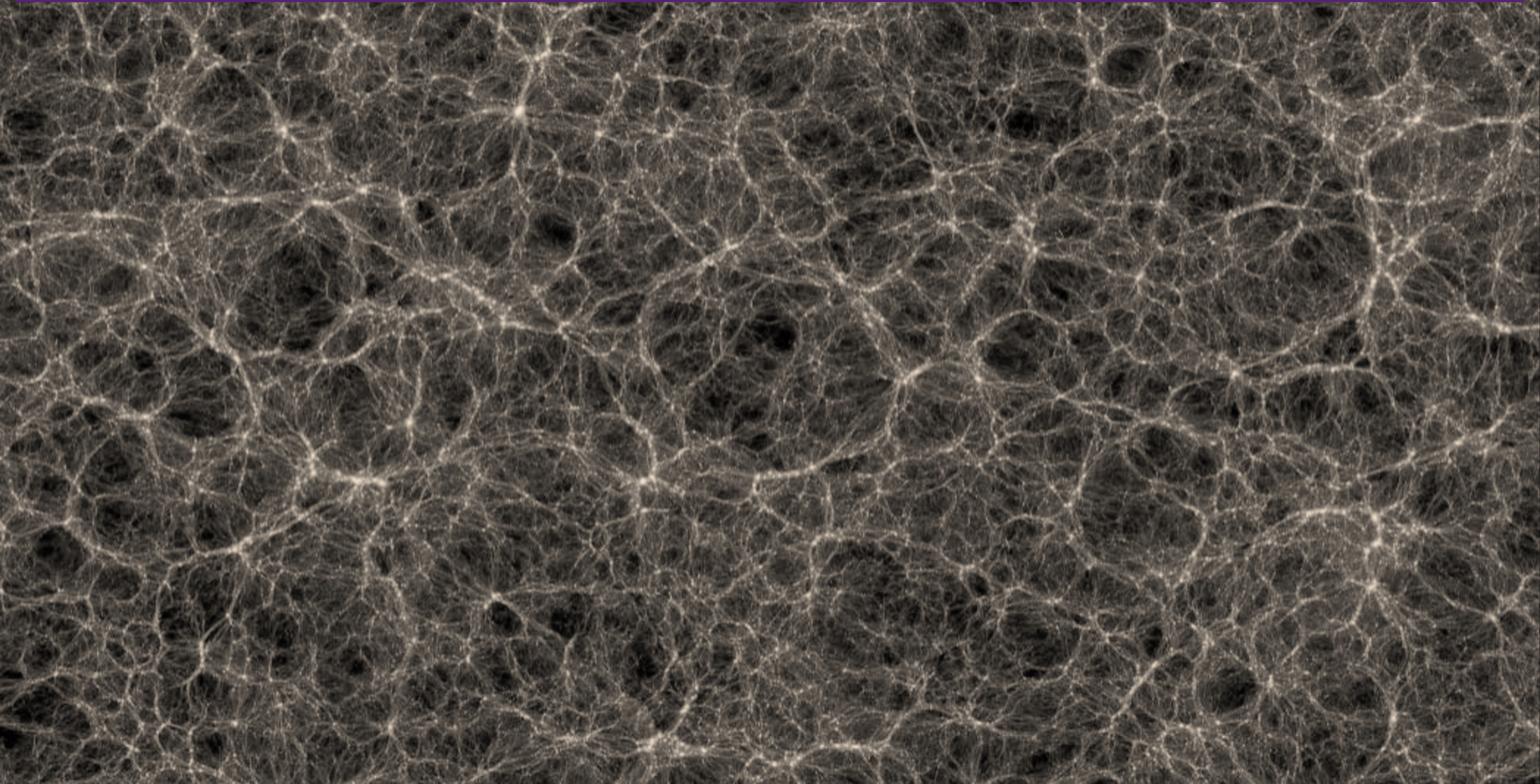
CAMELS

Adrian Bayer
Alex Barreira
Ana Maria Delgado
Andrina Nicola
Alice Pisani
Benjamin Oppenheimer
Benjamin Wandelt
Blakesley Burkhart
ChangHoon Hahn
Colin Hill
Core Francisco Park
Daisuke Nagai
Desika Narayanan
David Spergel
Emily Moser
Erwin T. Lau

Faizan Mohammad
Gabrielle Parimbelli
Greg Bryan
Gabriella Contardo
Helen Shao
Jay Wadekar
Jingjing Shi
Joyce Caliendo
Lucia Perez
Lars Hernquist
Leander Thiele
Luis F. Machado Poletti
Matteo Viel
Matthew Gebhardt
Megan Tillman
Michael Eickenberg

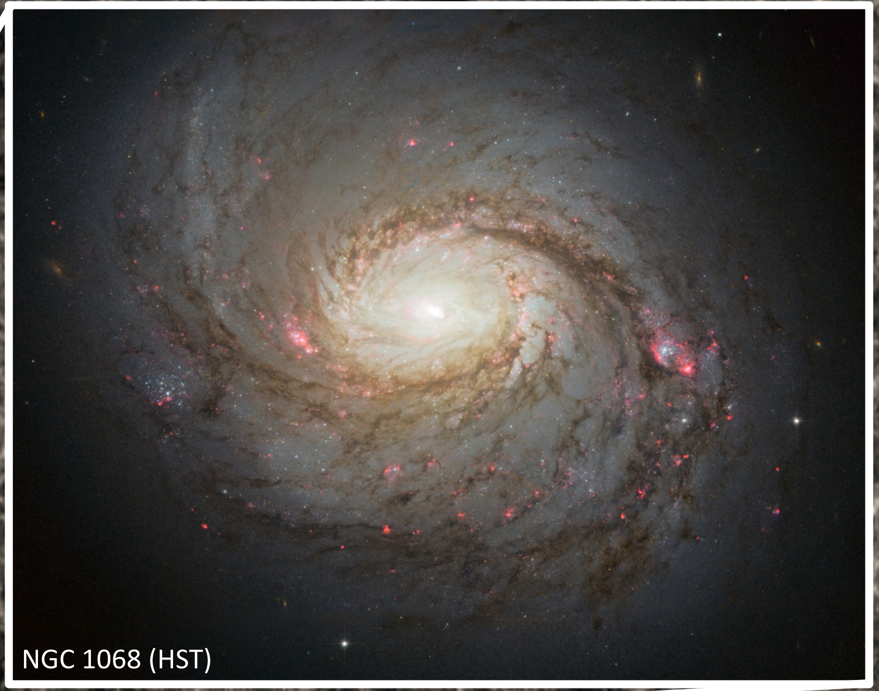
Neerav Kaushal
Nicholas Battaglia
Oliver Philcox
Pablo Villanueva-Domingo
Rachel Somerville
Romeel Dave
Stephanie Tonnensen
Sultan Hassan
Romain Teyssier
Ulrich Steinwandel
Valentina La Torre
Vid Irsic
William Coulton
Yin Li
Yongseok Jo
Yueying Ni

Large scale structure → …galaxy formation physics… → cosmology?



Large scale structure → …galaxy formation physics… → cosmology?

We only see the tip of the iceberg!

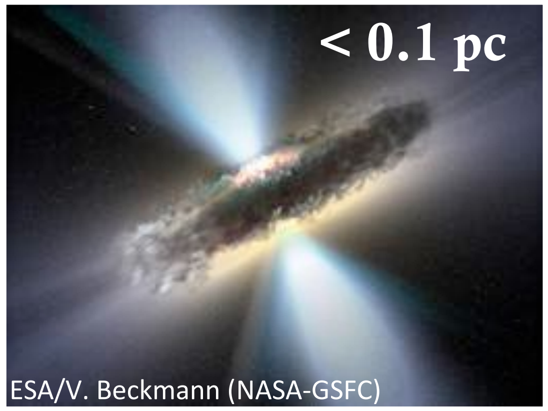


NGC 1068 (HST)

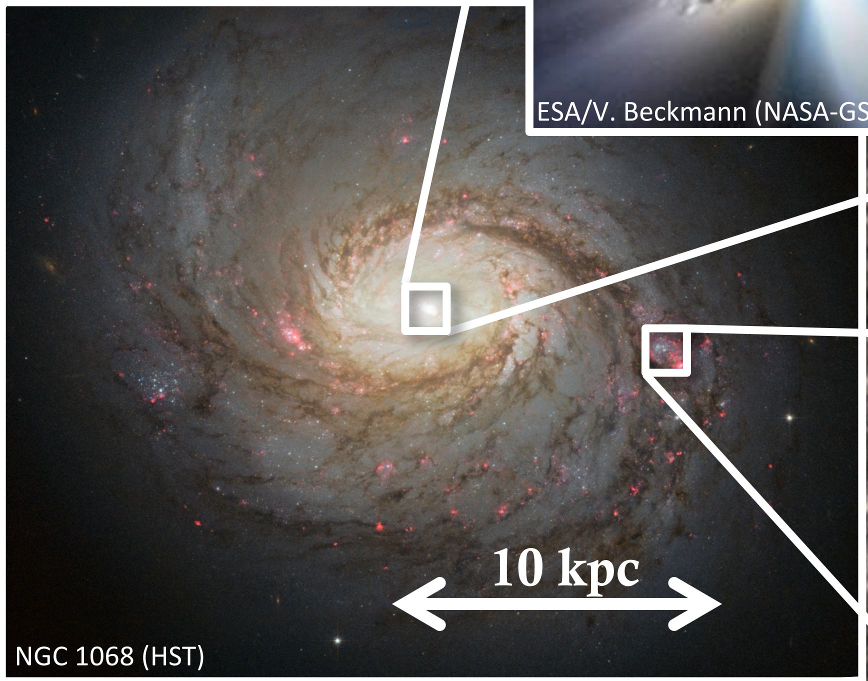
Galaxies form at the centers of dark matter halos

Large scale structure → …galaxy formation physics… → cosmology?

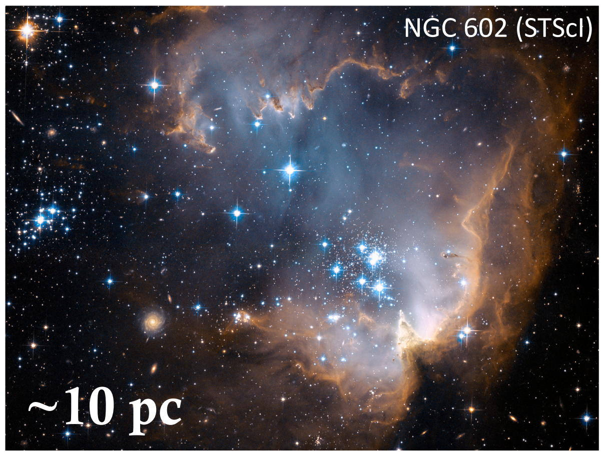
We only see the tip of the iceberg!



Supermassive black holes (SMBH) grow at the centers of galaxies and likely affect their evolution via radiation, winds, jets…



Massive stars affect their surrounding interstellar medium through supernovae, radiation, winds…



Galaxies form at the centers of dark matter halos

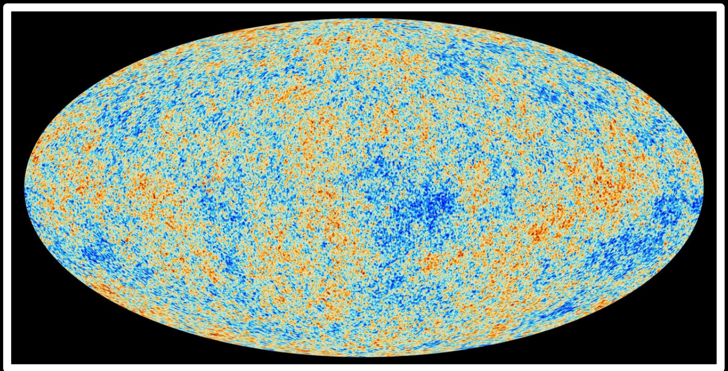
25 Mpc

Dream goal in galaxy formation simulations:

predict detailed properties of millions of galaxies starting from cosmological initial conditions using ‘ab-initio’ physics

Galaxy surveys

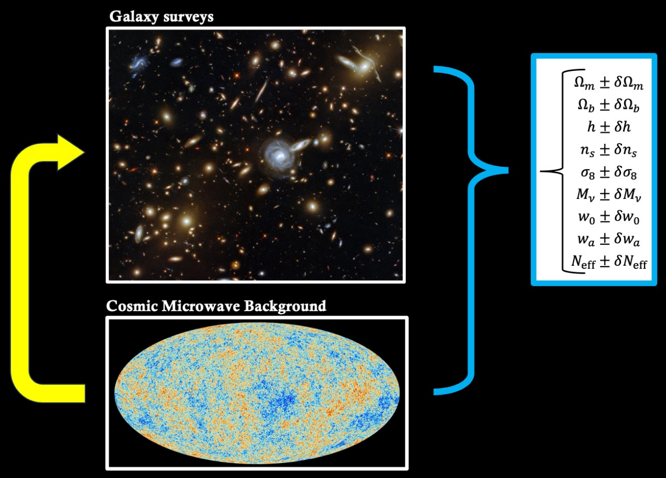
Cosmic Microwave Background



$$
\left\{
\begin{array}{c}
\Omega_m \pm \delta\Omega_m \\
\Omega_b \pm \delta\Omega_b \\
h \pm \delta h \\
n_s \pm \delta n_s \\
\sigma_8 \pm \delta\sigma_8 \\
M_\nu \pm \delta M_\nu \\
w_0 \pm \delta w_0 \\
w_a \pm \delta w_a \\
N_{\rm eff} \pm \delta N_{\rm eff}
\end{array}
\right.
$$

Dream goal in Cosmology:

Infer accurately and without bias the cosmological parameters using the full amount of information in the observable Universe

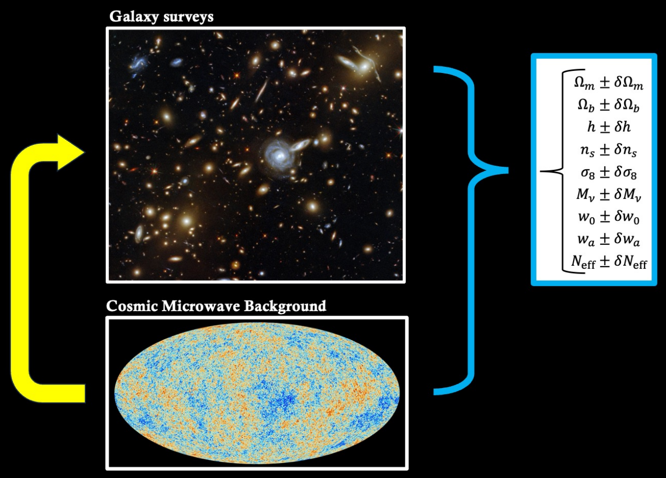


PROBLEM 1: Galaxy formation simulations rely on "sub-grid" models for unresolved processes that are still poorly understood

PROBLEM 2: Lots of cosmological information on small scales inaccessible due to impact of uncertain astrophysical processes

PROBLEM 3: The optimal summary statistic to extract cosmological information is unknown

PROBLEM 4: Need to speed up simulations to predict cosmological observables for large cosmological volumes



PROBLEM 1: Galaxy formation simulations rely on "sub-grid" models for unresolved processes that are still poorly understood

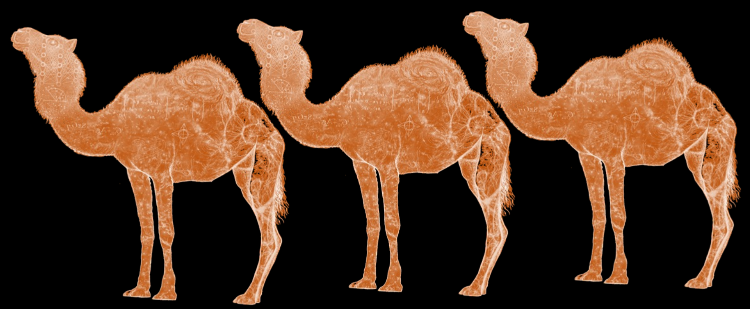
PROBLEM 2: Lots of cosmological information on small scales inaccessible due to impact of uncertain astrophysical processes

PROBLEM 3: The optimal summary statistic to extract cosmological information is unknown

PROBLEM 4: Need to speed up simulations to predict cosmological observables for large cosmological volumes

The CAMELS approach

Run thousands of simulations spanning the full range of uncertainty in galaxy formation physics and train machine learning algorithms to extract the maximum amount of cosmological information at the field level while marginalizing over uncertainties in baryonic effects



The CAMELS suites

>10,000 cosmological boxes of $(25\mathrm{Mpc}/h)^3$

>5,000 variations of TNG, SIMBA, and ASTRID

- cosmological params (Ω_m , σ_8, …)
- astrophysical params (feedback)

>5,000 corresponding DM-only simulations

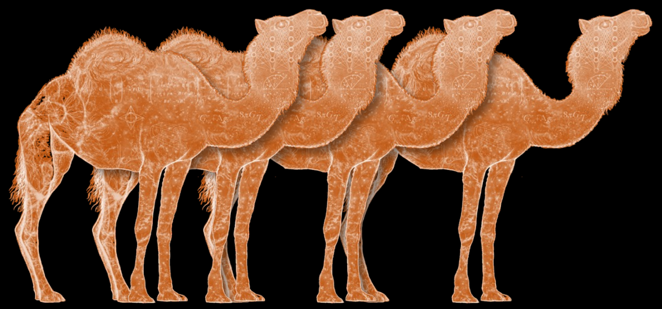
Additional simulation sets:

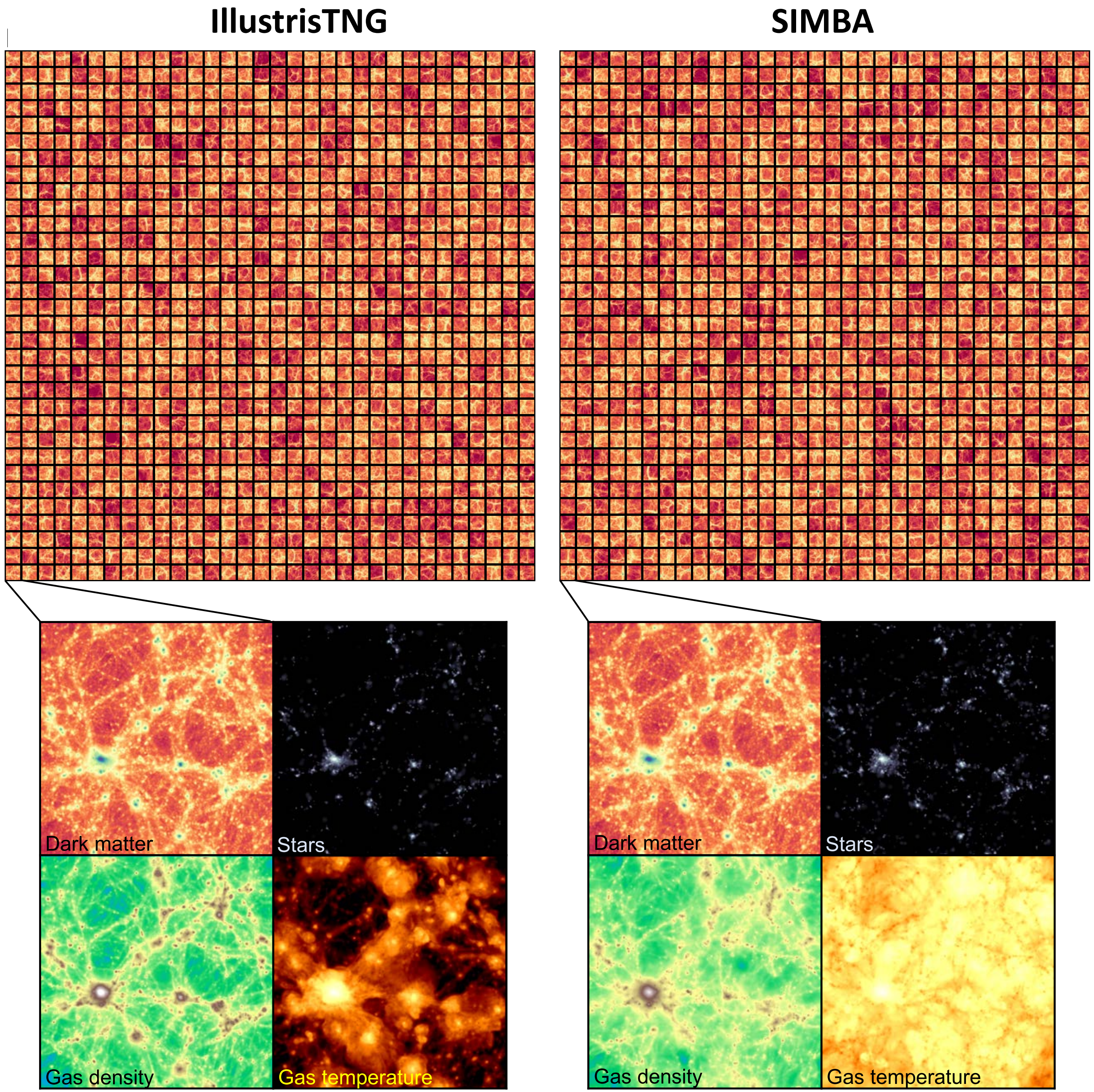
- Same ICs, varying one parameter
- Fiducial model, varying the ICs

CAMELS-SAM (Lucía Perez)

1,000 DM-only simulations of $(100\mathrm{Mpc}/h)^3$

- 2 cosmological params (Ω_m , σ_8)
- Santa-Cruz SAM parameter variations





CAMELS SAM

Lucia A. Perez

Princeton Future Faculty in the Physical Sciences Fellow & CCA Flatiron Research Fellow

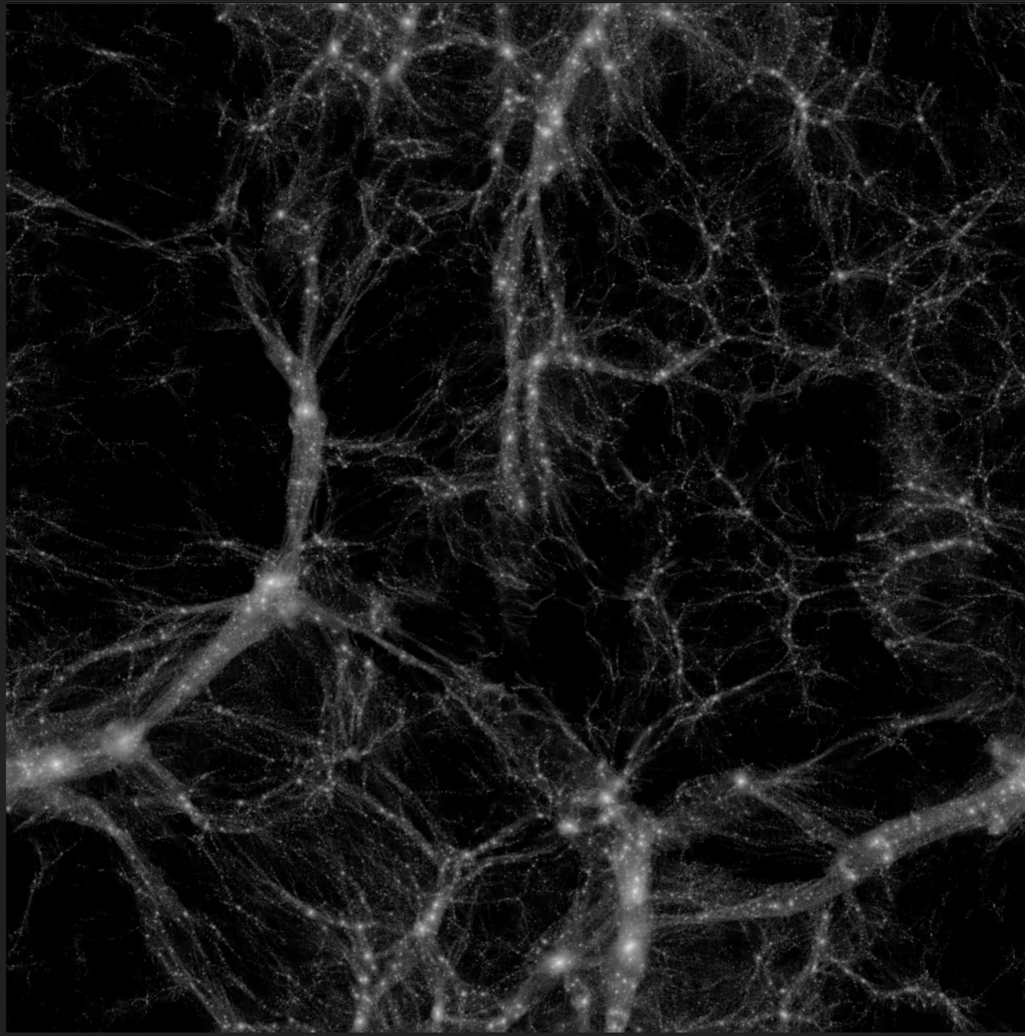
New large-volume simulation 'hump' of CAMELS project

- CAMELS (Cosmology & Astrophysics with MachinE Learning Simulations): machine learning data sets to create predictions for observations, marginalize over astrophysics to learn cosmology, and identify useful summary statistics and analyses
- 1000+ N-body simulations: $(100\ h^{-1}\ \mathrm{Mpc})^3$ large ; $N=640^3$ particles of ~1-6 x $10^8\ h^{-1}\ M_{sol}$; 100 snapshots between $0<z<27$
- Cosmological parameter space: Ω_m (fraction of energy density in DM+baryons) & σ_8 (~amplitude of density fluctuations)
- Run through the Santa Cruz Semi-Analytic Model:

  "A_{SN}": mass outflow + reheating rates of cold gas due to SNe + stars
  "A_{AGN}": AGN feedback, how much mass ejected in radio jets?

Data is public! camels-sam.readthedocs.io | arxiv.org/abs/2204.02408

Proof-of-concept in Perez+2022: constraining power of galaxy clustering statistics
(3D two-point correlation function, count-in-cells, Void Probability Function)



LH_643: $\Omega_m = 0.131$; $\sigma_8 = 0.986$

CAMELS public data repository

https://camels.readthedocs.io

The CAMELS project: public data release

Francisco Villaescusa-Navarro,[1,2] Shy Genel,[1,3] Daniel Anglés-Alcázar,[4,1] Lucia A. Perez,[5] Pablo Villanueva-Domingo,[6] Digvijay Wadekar,[7,8] Helen Shao,[2] Faizan G. Mohammad,[9,10] Sultan Hassan,[1,11] Emily Moser,[12] Erwin T. Lau,[13] Luis Fernando Machado Poletti Valle,[14] Andrina Nicola,[2] Leander Thiele,[15] Yongseok Jo,[16] Oliver H. E. Philcox,[2,8] Benjamin D. Oppenheimer,[17,13] Megan Tillman,[18] ChangHoon Hahn,[2] Neerav Kaushal,[19] Alice Pisani,[1,20,2] Matthew Gebhardt,[4] Ana Maria Delgado,[13] Joyce Caliendo,[4,21] Christina Kreisch,[2] Kaze W.K. Wong,[1] William R. Coulton,[1] Michael Eickenberg,[22] Gabriele Parimbelli,[23,24,25,26,27] Yueying Ni,[28] Ulrich P. Steinwandel,[1] Valentina La Torre,[29] Romeel Dave,[30,11,31] Nicholas Battaglia,[12] Daisuke Nagai,[32] David N. Spergel,[1,2] Lars Hernquist,[13] Blakesley Burkhart,[18,1] Desika Narayanan,[33,34] Benjamin Wandelt,[35,1] Rachel S. Somerville,[1] Greg L. Bryan,[36,1] Matteo Viel,[25,27,26,37] Yin Li,[1,22] Vid Irsic,[38,39] Katarina Kraljic,[40] and Mark Vogelsberger[41]

- Large number of labeled data products in the form of 1D, 2D, and 3D arrays
- Full documentation and metadata available
- Designed to enable a broad range of creative AI applications
- Public access to full data, (limited) local computing, and tutorials

arXiv:2201.01300

CAMELS

Search docs

CAMELS

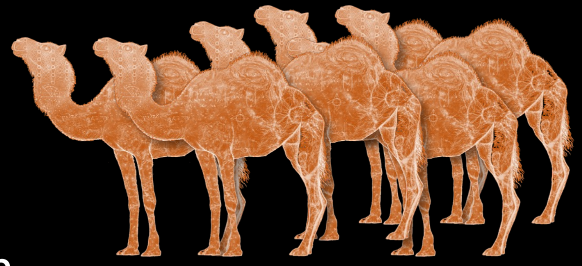
CAMELS stands for **C**osmology and **A**strophysics with **M**achin**E** **L**earning **S**imulations, and it is a project that aims at building bridges between cosmology and astrophysics through numerical simulations and machine learning. CAMELS contains 10,680 cosmological simulations –5,164 N-body and 5,516 state-of-the-art (magneto-)hydrodynamic– and more than 700 Terabytes of data. CAMELS is the largest set of cosmological hydrodynamic simulations ever run.

Type	Code	Subgrid model	Simulations
Hydrodynamic	Arepo	IllustrisTNG	2,143
	Gizmo	SIMBA	1,092
	MP-Gadget	Astrid	2,116
	OpenGadget	Magneticum	77
	SWIFT	EAGLE	77
	Ramses		5
	Enzo		6
N-body	Gadget-III	–	5,164

Introductory video to the CAMELS project:

The video below shows an example of a CAMELS hydrodynamic simulation run with the Ramses code. Gas density and gas temperature are shown in blue and red, respectively as a function of time. CAMELS contains thousands of simulations like this one.

First CAMELS results very encouraging!



Cosmology inference… from 2D maps: (Paco, Yueying Ni, Jonah Rose)

- 2D projected maps of 27 fields (dark matter, gas, stars) with 100 kpc/pixel resolution
- It works! 2-3% error in Ω_{m} and σ_8 with all fields combined (3-4% with HI only)
- Extracting information down to the smallest scales (1 pixel), marginalizing over baryonic effects
- But… only the total mass field is robust to differences in galaxy formation model (TNG vs SIMBA)

… from summary statistics (Andrina Nicola, Lucía Perez, Ana María Delgado)

… from galaxy positions/velocities with GNN (Natalí de Santi, Helen Shao)

… and from a single galaxy! (Paco, Nicolas Echeverri-Rojas, Chaitanya Chawak)

Constraining feedback…

… with SZ (Emily Moser, Pandey, Shivam), **spectral distortions** (Leander Thiele), **Lyα** (Megan Tillman, Blakesley Burkhart)

Predicting galaxy/halo properties:

- Finding universal Relations in (sub)halo properties (Helen Shao)
- Inferring halo masses from galaxy properties with GNN (Pablo Villanueva-Domingo)
- Halos mass and CGM properties from X-ray and HI maps (Naomi Gluck)
- Reducing the scatter in the SZ flux-mass relation (Digvijay Wadekar)

Not an exhaustive list!

Emulation (Sultan Hassan, Chris Lovell, Yongseok Jo, Max Lee, Matt Gebhardt), **Inpainting** (Faizan Mohammad)

Cosmological inference at the field level

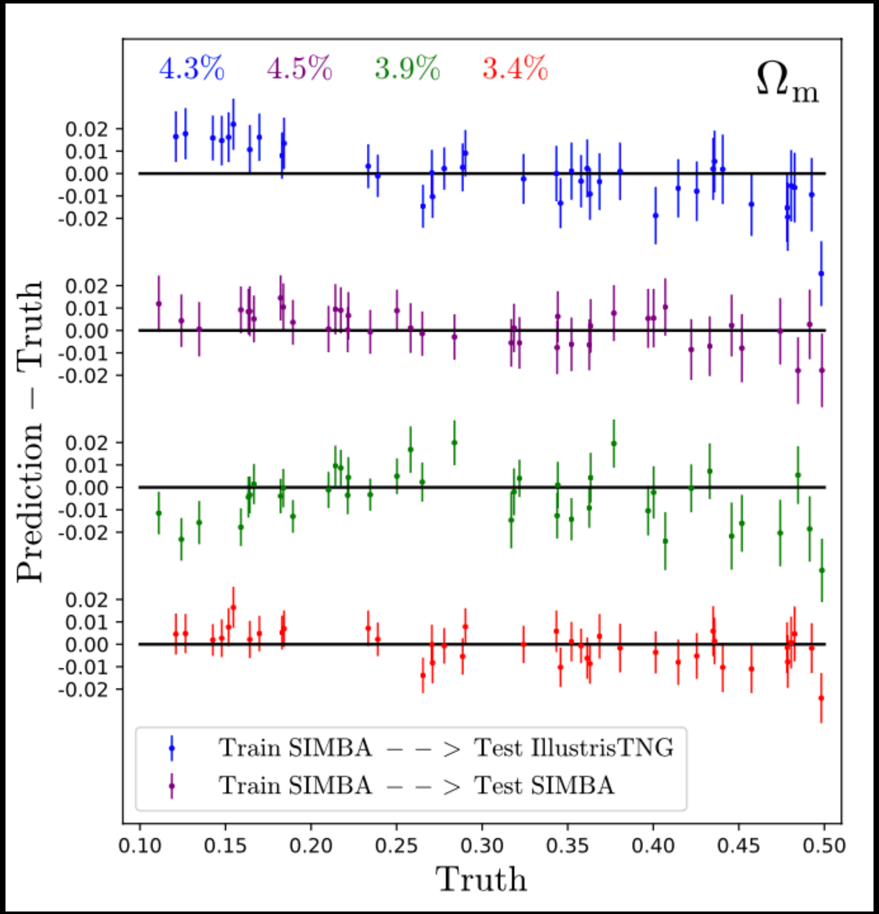
Villaescusa-Navarro, Anglés-Alcázar, Genel, et al. (2021a,b,c)

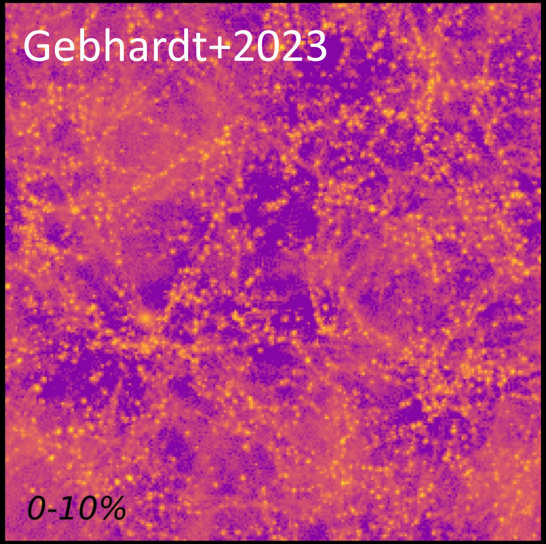
The 2D total mass field is a robust predictor

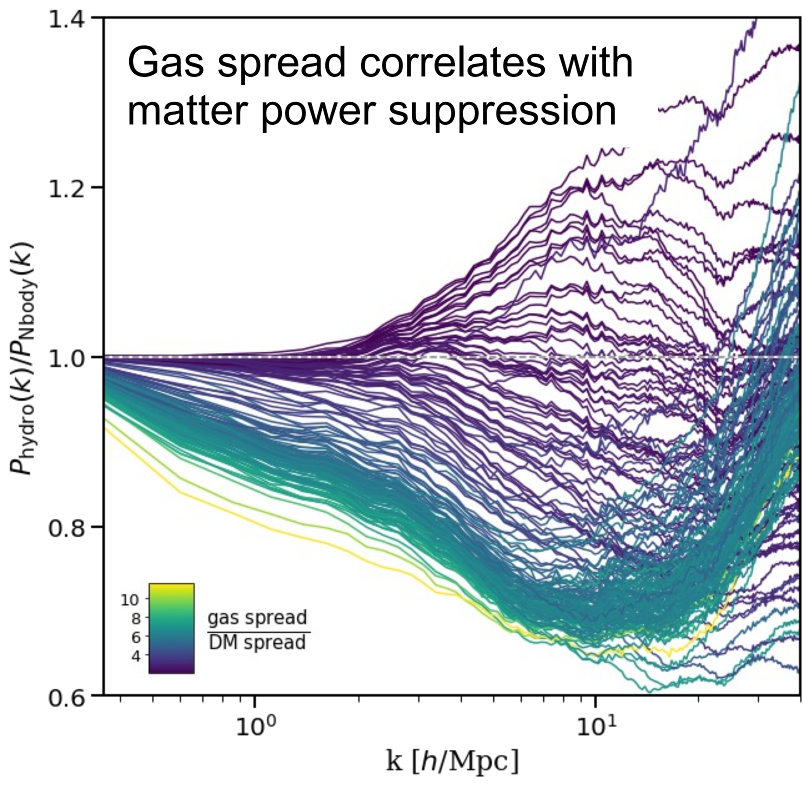
- Extracting more information than power spectra
- Down to smallest scale (100 kpc/pixel)
- Marginalizing over baryonic effects

Despite the large impact of baryons on the matter power spectrum

Delagado+2023, Gebhardt+2023, Pandey+2023



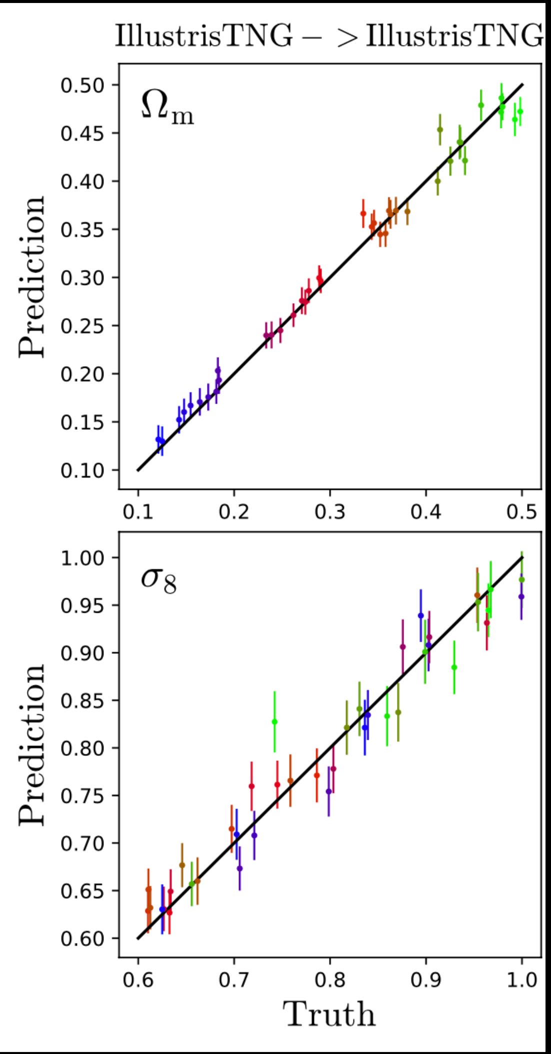


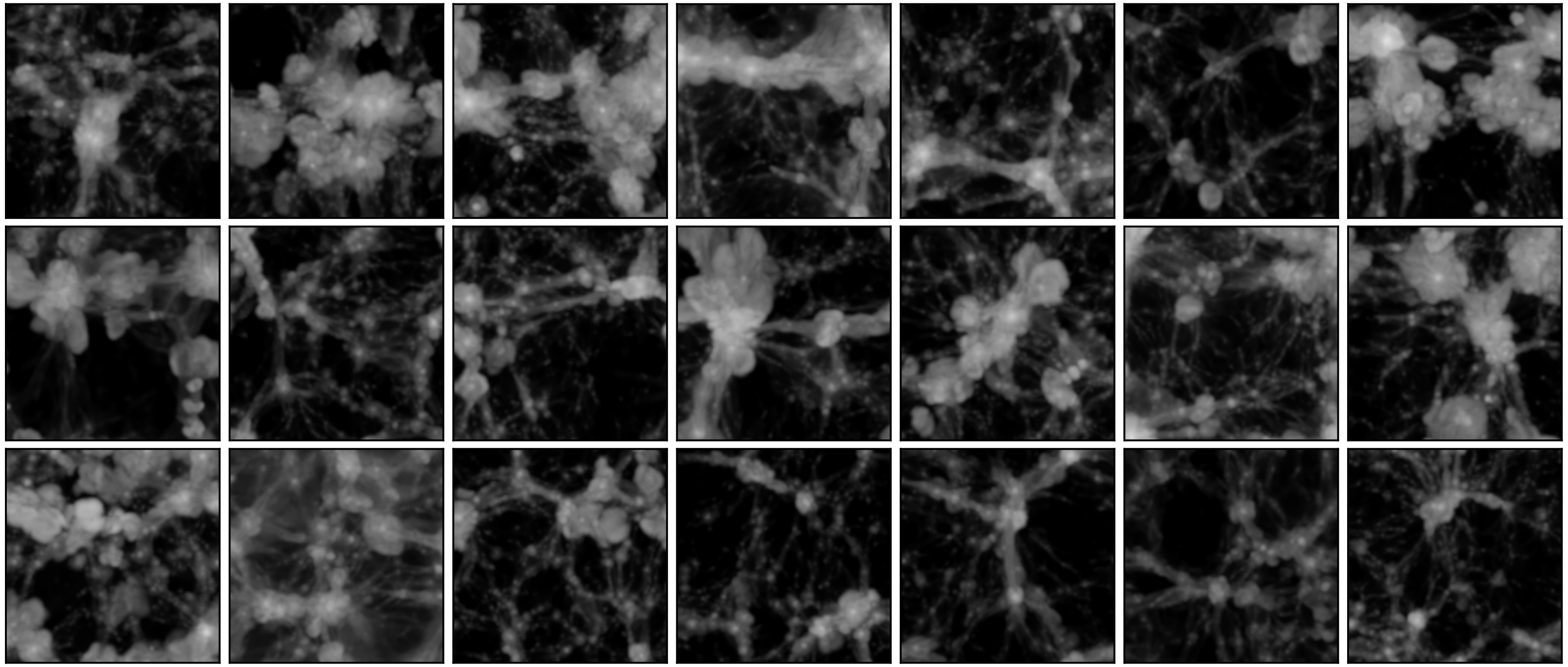


Cosmological inference at the field level

Villaescusa-Navarro, Anglés-Alcázar, Genel, et al. (2021a,b,c)

→Train neural network on temperature maps to predict input cosmological parameters while marginalizing over sub-grid physics



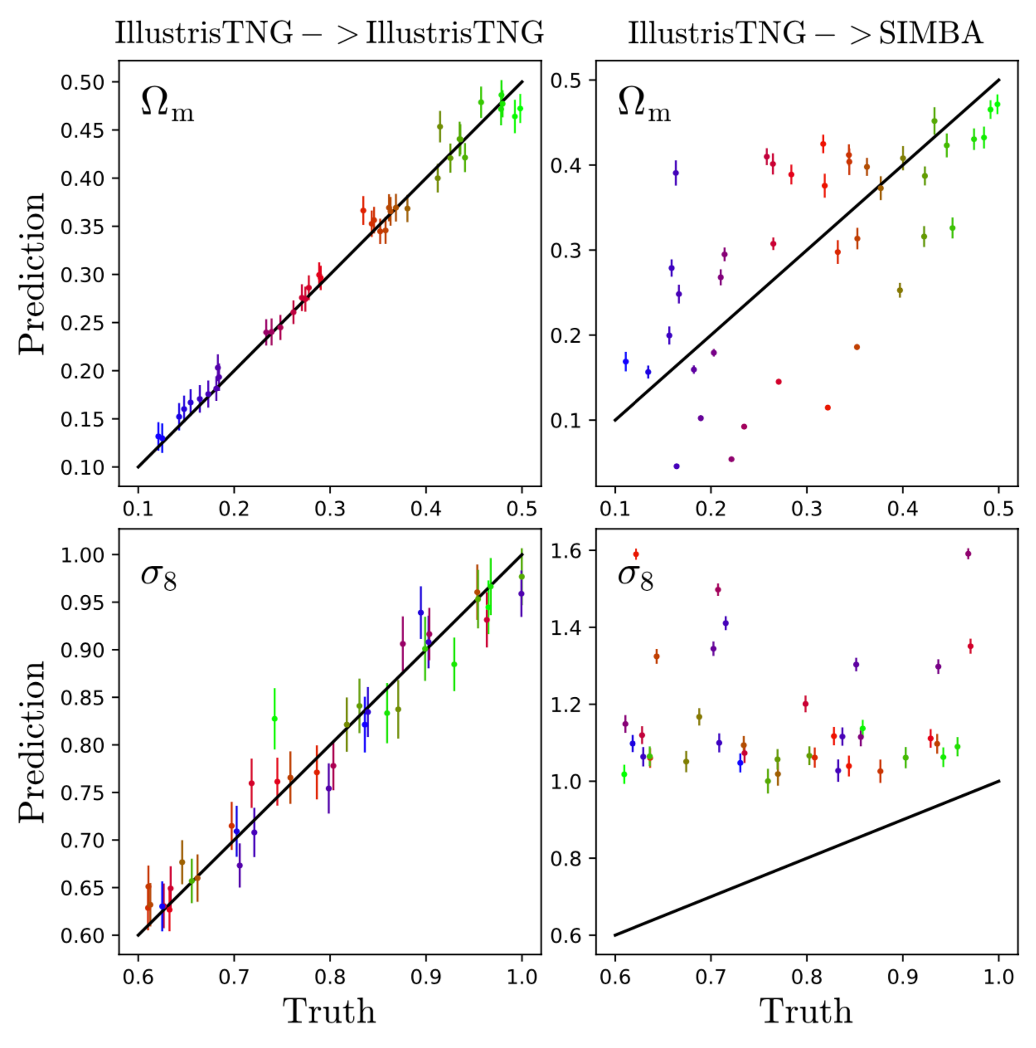


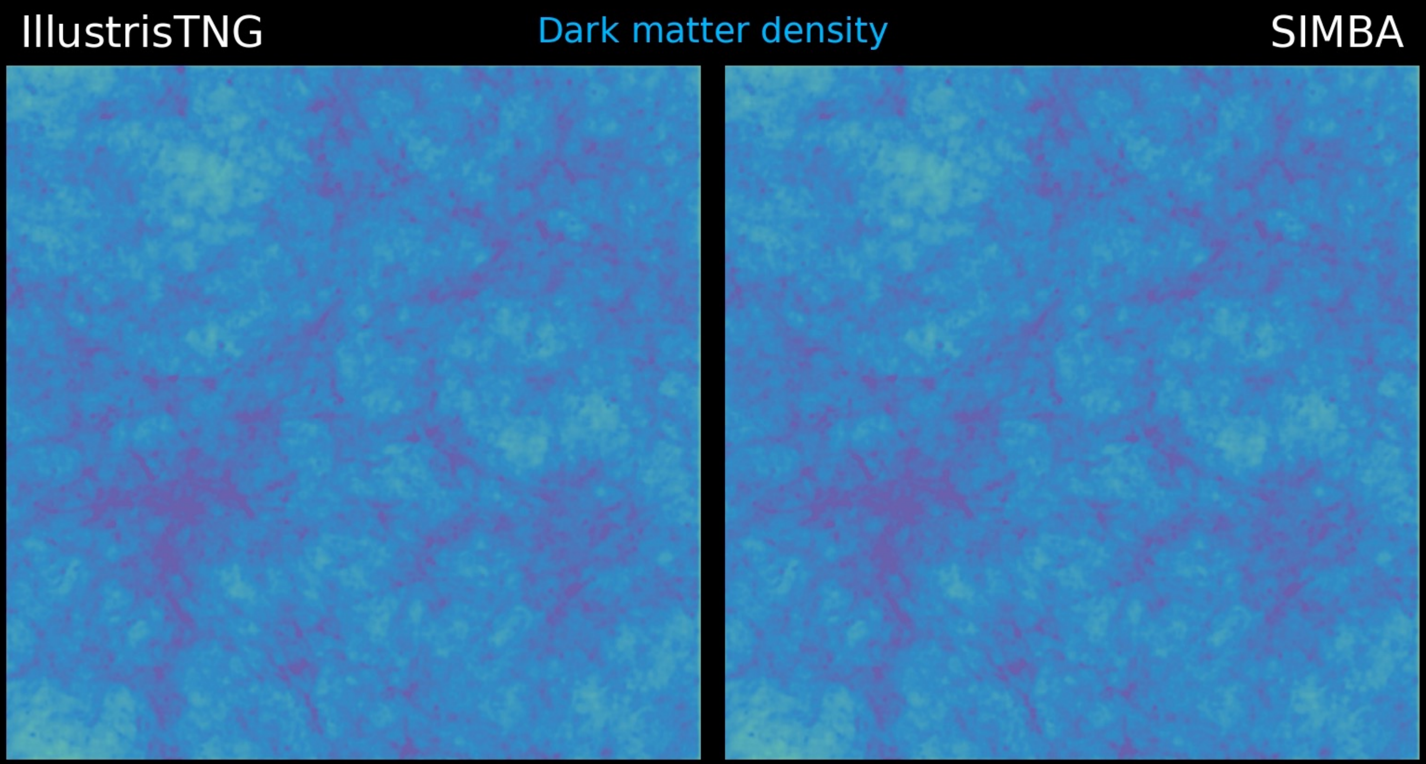
Every map has 256×256 pixels, covers an area of 25×25 $(h^{-1}\mathrm{Mpc})^2$, and has a different cosmology & astrophysics. 15,000 images in total.

Cosmological inference at the field level

Villaescusa-Navarro, Anglés-Alcázar, Genel, et al. (2021a,b,c)

→ Inference from 2D Temperature maps is not robust to galaxy formation physics implementation





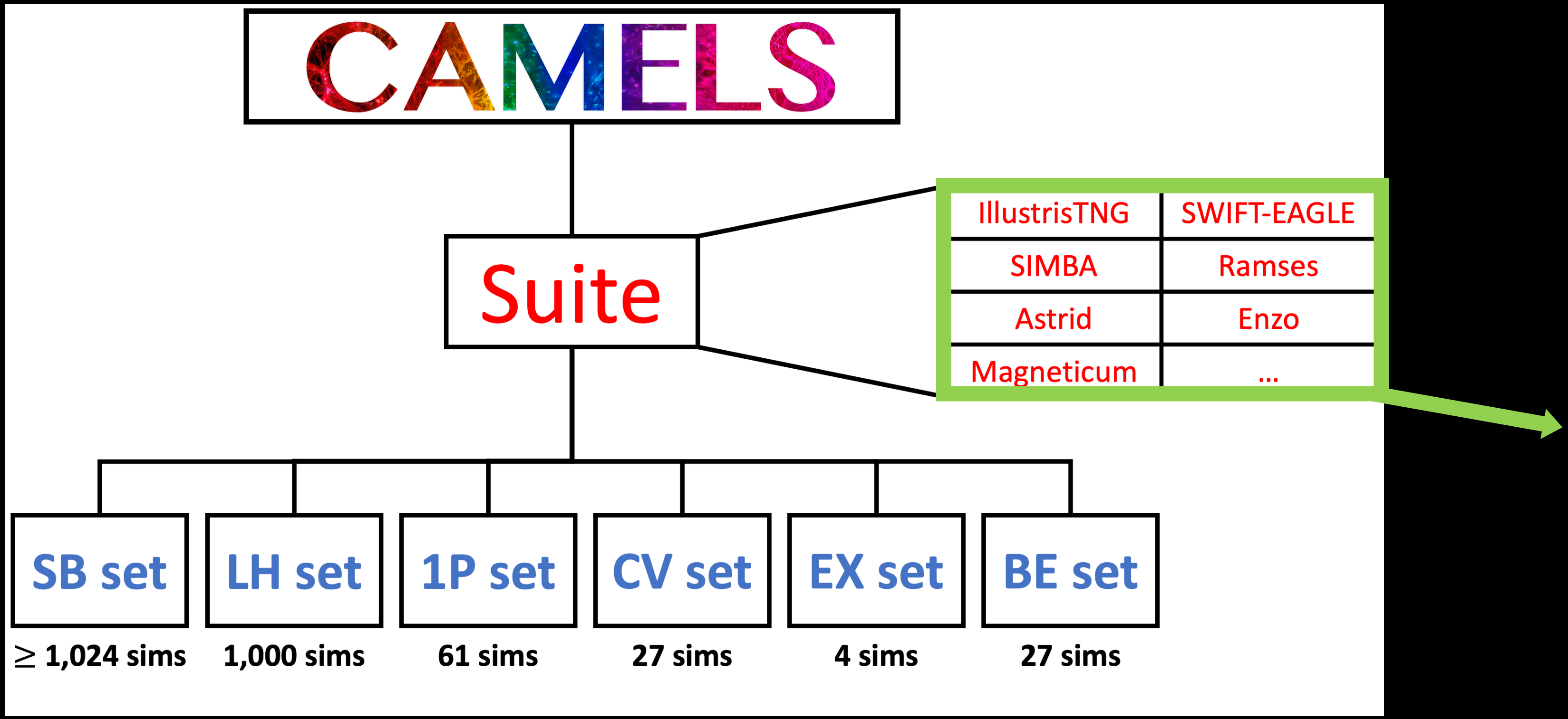
Differences between sub-grid models limit learning across them

Crucial to expand the range of models in the training set

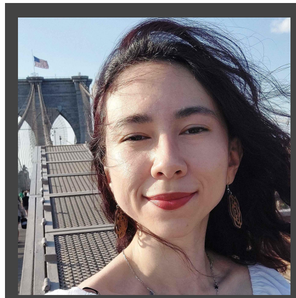
The CAMELS project: Expanding the galaxy formation model space with new ASTRID and 28-parameter TNG and SIMBA suites

Yueying Ni,[1,2] Shy Genel,[3,4] Daniel Anglés-Alcázar,[5,3] Francisco Villaescusa-Navarro,[3,6] Yongseok Jo,[3] Simeon Bird,[7] Tiziana Di Matteo,[2,8] Rupert Croft,[2,8] Nianyi Chen,[2] Natalí S. M. de Santi,[3,9] Matthew Gebhardt,[5] Helen Shao,[6] Shivam Pandey,[10,11] Lars Hernquist,[1] and Romeel Dave[12]

“Building robust machine-learning models favors training and testing on the largest possible diversity of galaxy formation models”



Reach out if you would like to bring your model into the public CAMELS dataset!



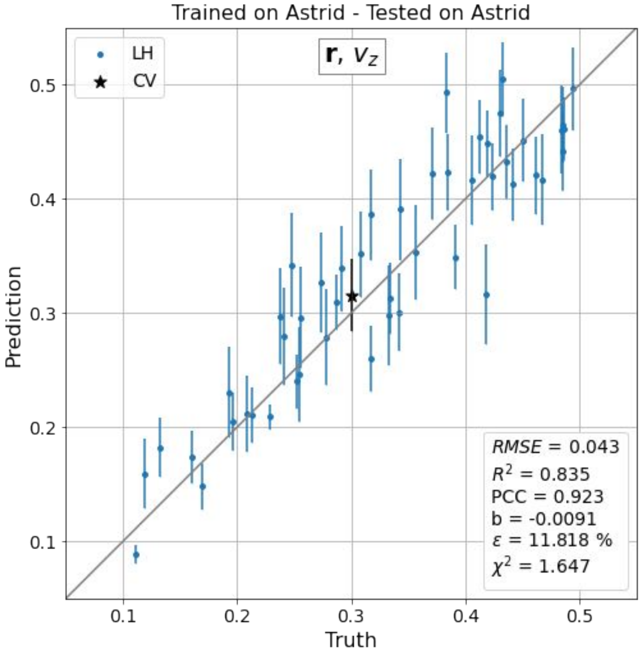
Natalí de Santi

Flatiron Institute
University of São Paulo

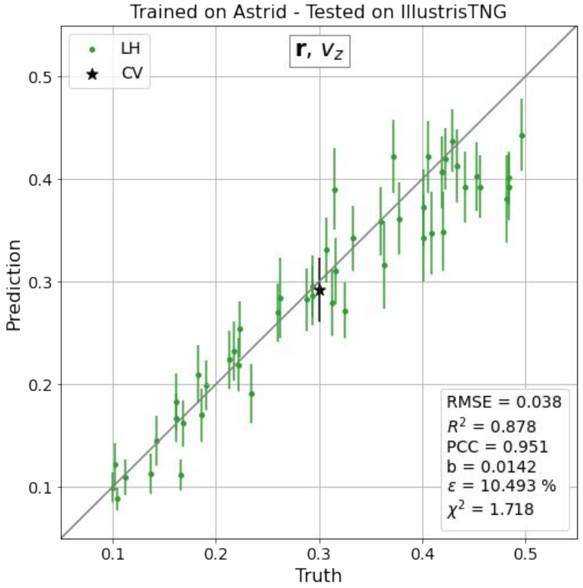
natalidesanti@gmail.com

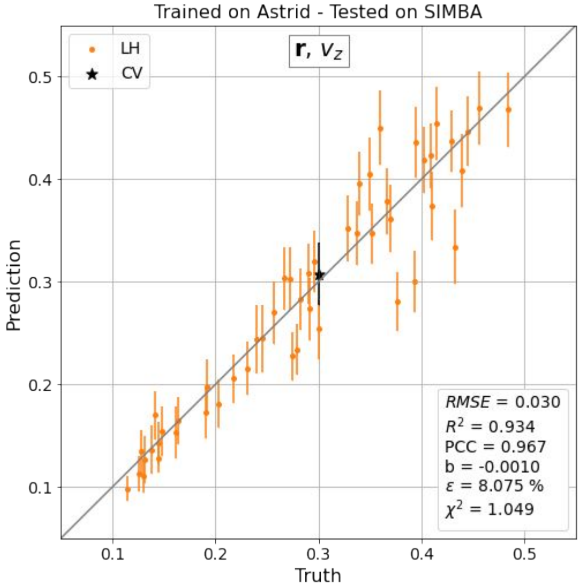
Robust field-level likelihood-free inference with galaxies

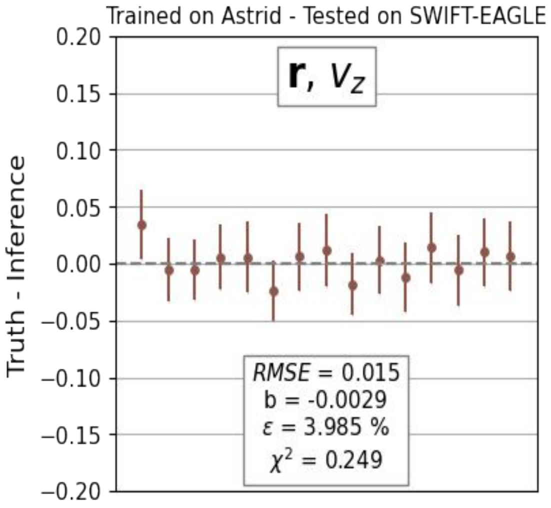
Dataset: Galaxies from Astrid
Machine Learning Method: Graph Neural Networks
Objective: Ω_m inference

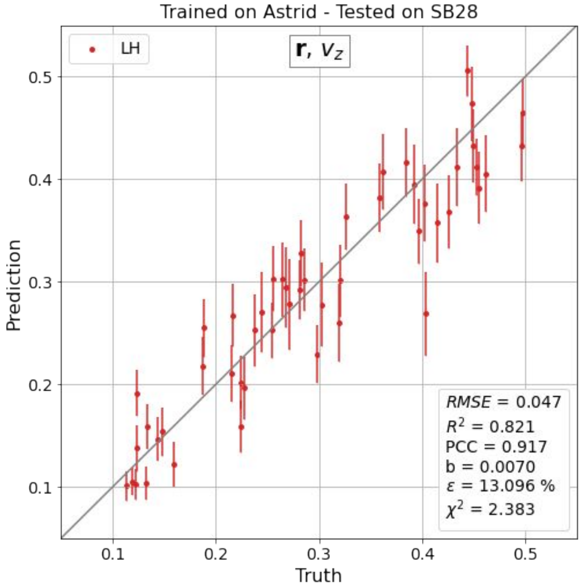


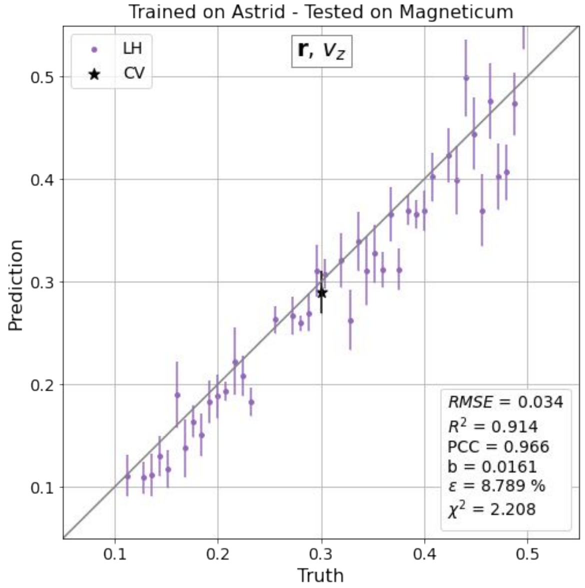
arXiv: 2302.14101







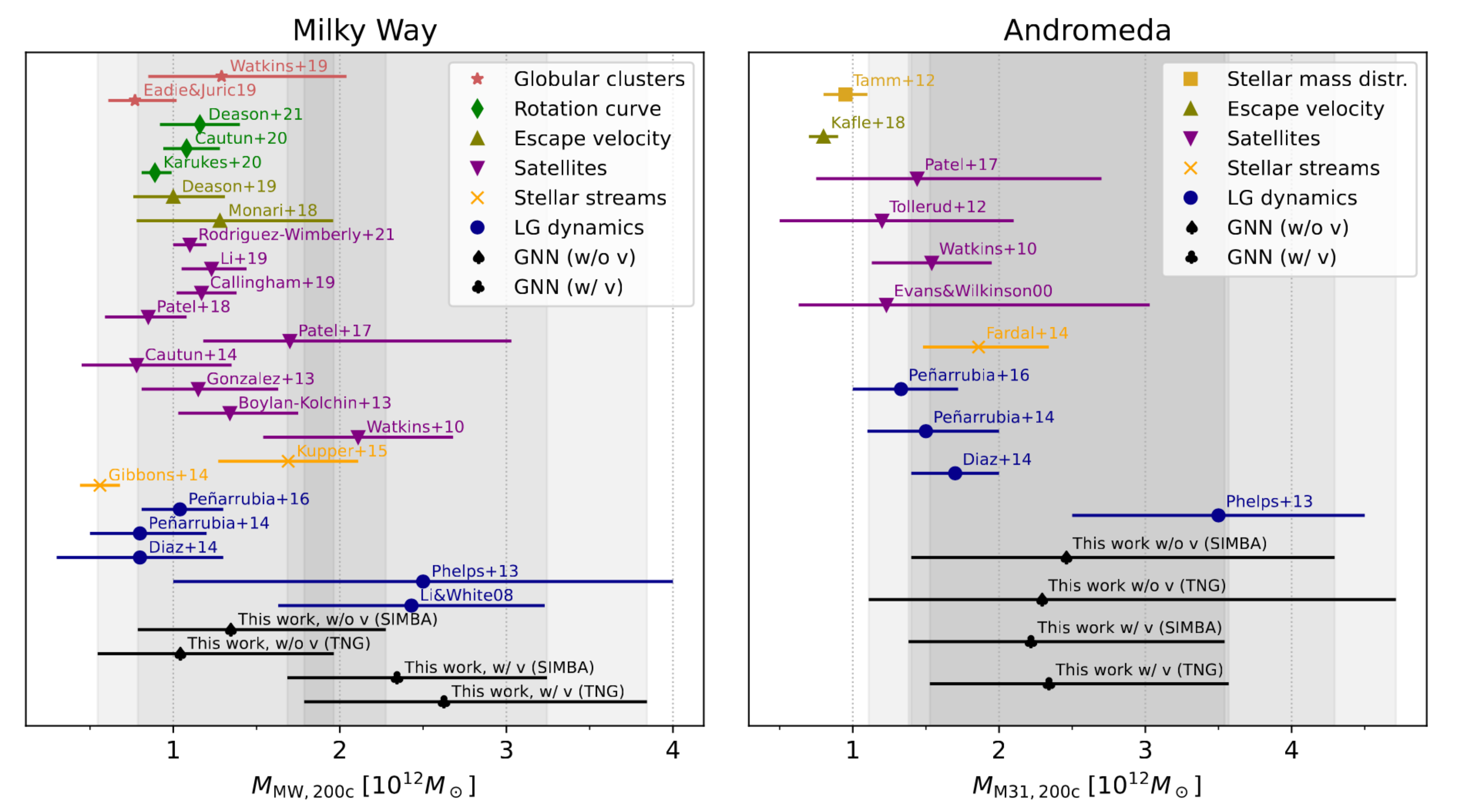




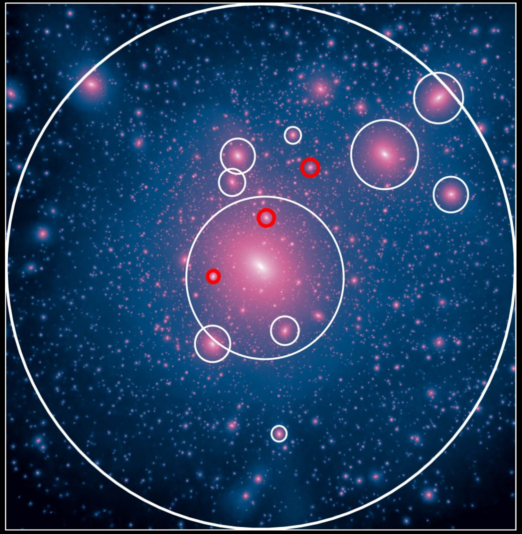
- Information came from galaxy positions and velocities;
- The broader variation in Astrid allowed a robust model across 5 different sub-grid physics sets;
- First steps to apply this machinery on real data.

Weighing the Milky Way and Andromeda with Artificial Intelligence

Pablo Villanueva-Domingo,[1,*] Francisco Villaescusa-Navarro,[2,3,†] Shy Genel,[2,4] Daniel Anglés-Alcázar,[5,2] Lars Hernquist,[6] Federico Marinacci,[7] David N. Spergel,[2,3] Mark Vogelsberger,[8] and Desika Narayanan[9,10]



Graph Neural Networks trained on positions, velocities, and stellar masses of galaxies to predict halo mass



THE NEW YORKER

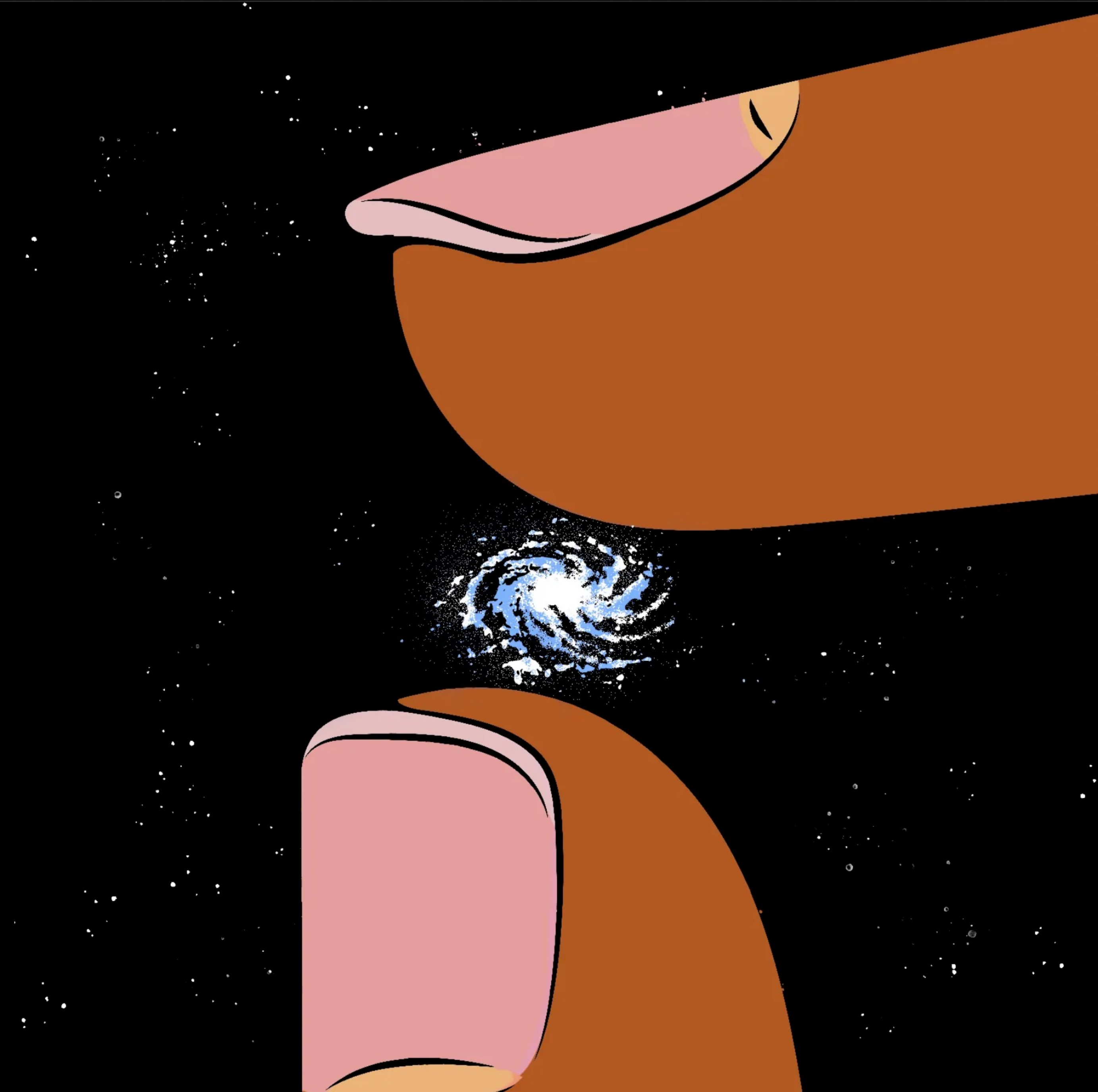
ELEMENTS

WHAT CAN WE LEARN ABOUT THE UNIVERSE FROM JUST ONE GALAXY?

In new research, begun by an undergraduate, William Blake's phrase "to see a world in a grain of sand" is suddenly relevant to astrophysics.

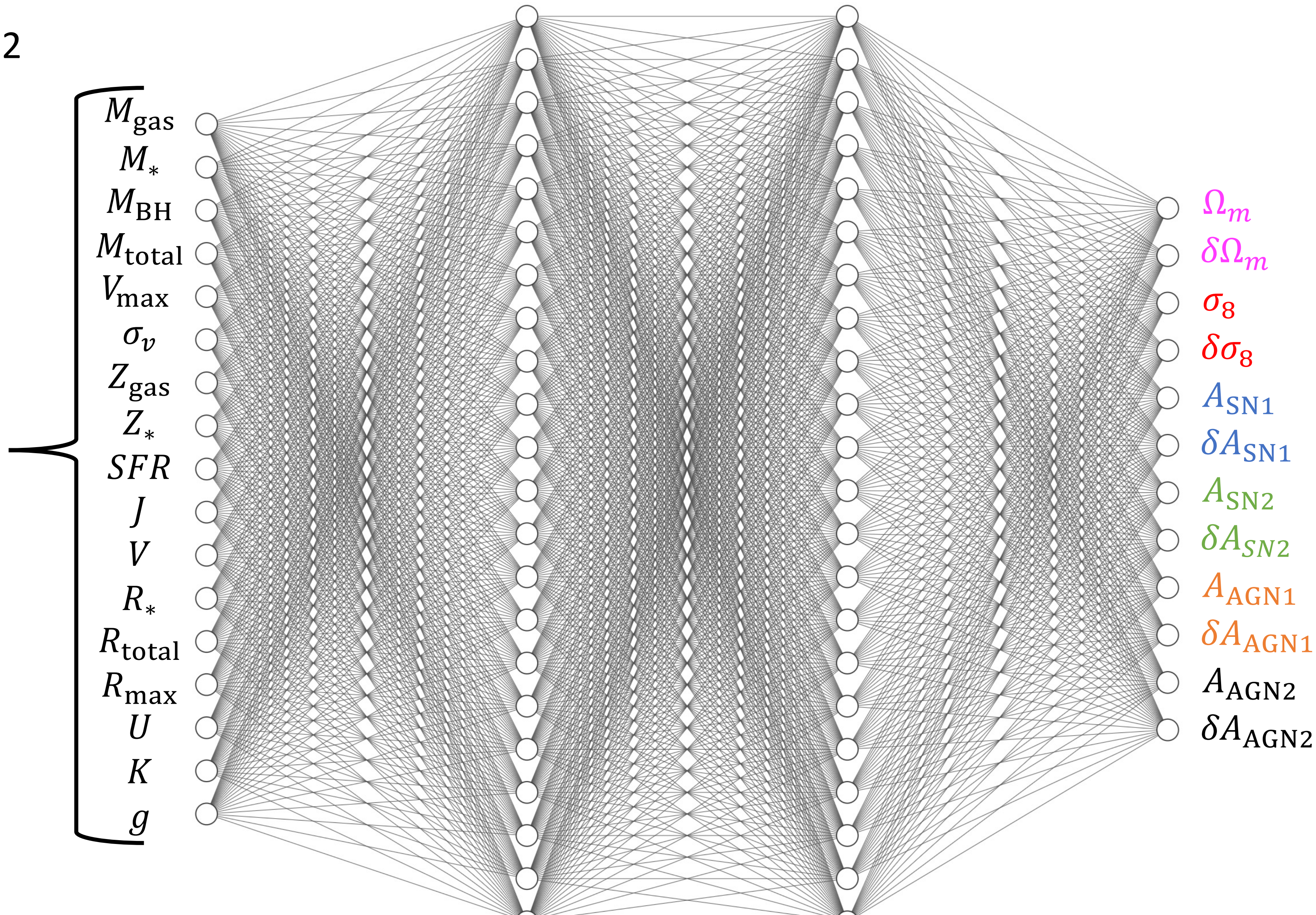
By Rivka Galchen

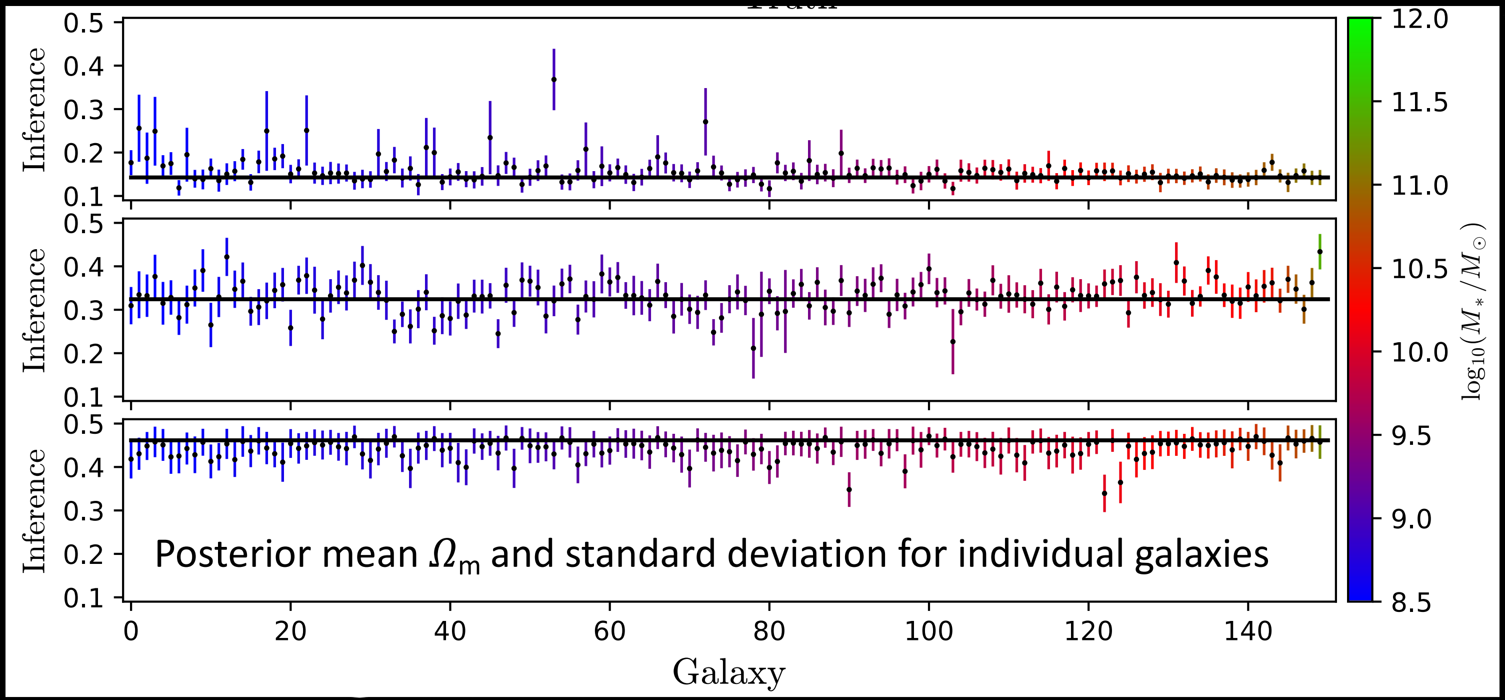
March 23, 2022

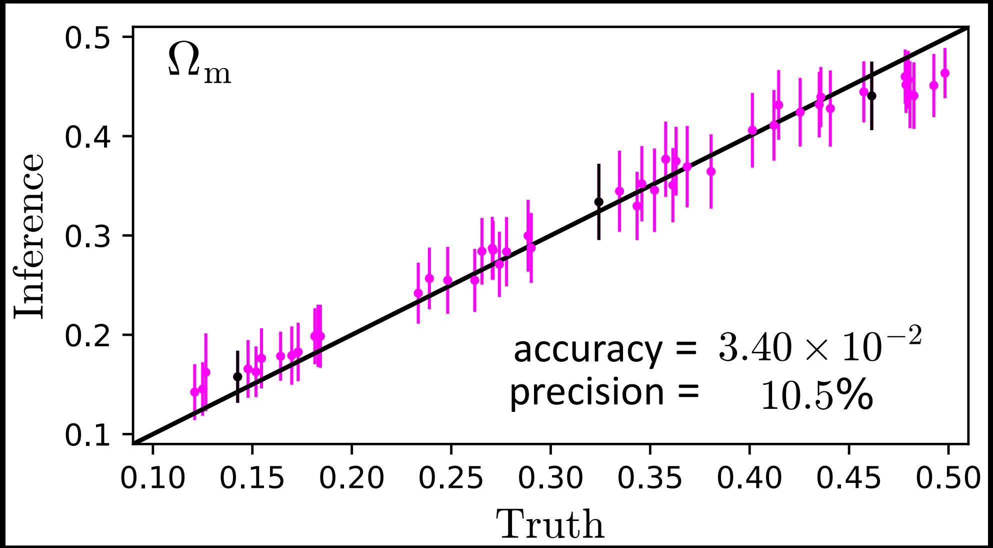


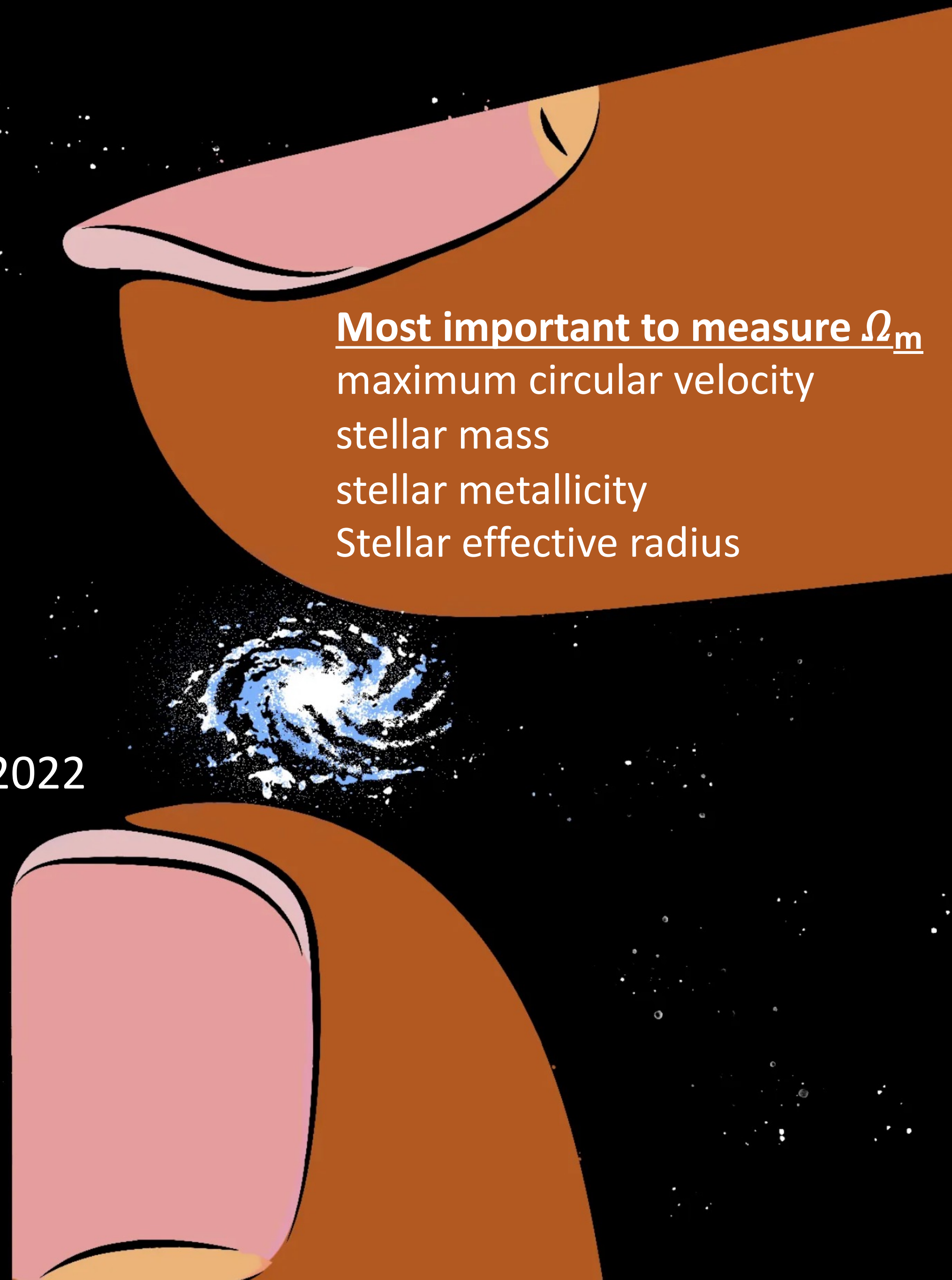
CAMELS enables testing new ideas: **Cosmology with a single galaxy?**

Villaescusa-Navarro+2022
Echeverri-Rojas+2023
Chawak+2023









Most important to measure Ω_m

maximum circular velocity

stellar mass

stellar metallicity

Stellar effective radius

Villaescusa-Navarro+2022

Echeverri-Rojas+2023

Chawak+2023

The CAMELS project:

Cosmology and Astrophysics with MachinE Learning Simulations

Villaescusa-Navarro, Anglés-Alcázar, Genel, et al. (2021,2022,2023)
Perez+2023 (arXiv:2204.02408), Ni+2023 (arXiv:2304.02096)

- Largest suite of cosmological hydrodynamic simulations with thousands of model variations designed for machine learning applications

- Encouraging results extracting cosmological information at the field level down to small scales even where astrophysical effects are significant

- Many possible applications in galaxy formation and cosmology (inference, emulating/accelerating simulations, learning physics with AI,...)

- Full dataset publicly available: https://camels.readthedocs.io

- Challenges: larger-volume simulations, extending parameter space, interpolation between models, robustness, synthetic observations…

CAMELS