4-6-2023

The CAMELS Project: Public Data Release

Francisco Villaescusa-Navarro
Simons Foundation

Shy Genel
Simons Foundation

Daniel Anglés-Alcázar
Simons Foundation

Lucia A. Perez
School of Earth and Space Exploration

Pablo Villanueva-Domingo
Universitat de València

See next page for additional authors

Recommended Citation

Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., Perez, L., Villanueva-Domingo, P., Wadekar, D., Shao, H., Mohammad, F., Hassan, S., Moser, E., Lau, E., Machado Poletti Valle, L., Nicola, A., Thiele, L., Jo, Y., Philcox, O., Oppenheimer, B., Tillman, M., Hahn, C., Kaushal, N., & et al. (2023). The CAMELS Project: Public Data Release. *Astrophysical Journal, Supplement Series, 265*(2). http://doi.org/10.3847/1538-4365/acbf47
Retrieved from: https://digitalcommons.mtu.edu/michigantech-p/17074

Follow this and additional works at: https://digitalcommons.mtu.edu/michigantech-p
Part of the Physics Commons

Authors

Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, Lucia A. Perez, Pablo Villanueva-Domingo, Digvijay Wadekar, Helen Shao, Faizan G. Mohammad, Sultan Hassan, Emily Moser, Erwin T. Lau, Luis Fernando Machado Poletti Valle, Andrina Nicola, Leander Thiele, Yongseok Jo, Oliver H.E. Philcox, Benjamin D. Oppenheimer, Megan Tillman, Chang Hoon Hahn, Neerav Kaushal, and et al.

The Astrophysical Journal Supplement Series, 265:54 (14pp), 2023 April https://doi.org/10.3847/1538-4365/acbf47

The CAMELS Project: Public Data Release

Francisco Villaescusa-Navarro[1,2], Shy Genel[1,3], Daniel Anglés-Alcázar[1,4], 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], Megan Tillman[18], ChangHoon Hahn[2], Neerav Kaushal[19], Alice Pisani[1,2,20], 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[11,30,31], Nicholas Battaglia[12], Daisuke Nagai[32], David N. Spergel[1,2], Lars Hernquist[13], Blakesley Burkhart[1,18], Desika Narayanan[33,34], Benjamin Wandelt[1,35], Rachel S. Somerville[1], Greg L. Bryan[1,36], Matteo Viel[25,26,27,37], Yin Li[1,22], Vid Irsic[38,39], Katarina Kraljic[40], Federico Marinacci[41], and Mark Vogelsberger[42]

[1] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA; camel.simulations@gmail.com
[2] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA
[3] Columbia Astrophysics Laboratory, Columbia University, New York, NY 10027, USA
[4] Department of Physics, University of Connecticut, 196 Auditorium Road, Storrs, CT 06269, USA
[5] Arizona State University, School of Earth and Space Exploration, 781 Terrace Mall, Tempe, AZ 85287, USA
[6] Instituto de Física Corpuscular (IFIC), CSIC-Universitat de València, E-46980, Paterna, Spain
[7] Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA
[8] School of Natural Sciences, Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540, USA
[9] Waterloo Center for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[10] Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[11] Department of Physics & Astronomy, University of the Western Cape, Cape Town 7535, South Africa
[12] Department of Astronomy, Cornell University, Ithaca, NY 14853, USA
[13] Center for Astrophysics–Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[14] Institute for Particle Physics and Astrophysics, ETH Zürich, Wolfgang-Pauli-Strasse 27, CH-8093 Zürich, Switzerland
[15] Department of Physics, Princeton University, Princeton, NJ 08544, USA
[16] Center for Theoretical Physics, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Republic of Korea
[17] CASA, Department of Astrophysical and Planetary Sciences, University of Colorado, 389 UCB, Boulder, CO 80309, USA
[18] Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854, USA
[19] Department of Physics, Michigan Technological University, Houghton, MI 49931, USA
[20] The Cooper Union for the Advancement of Science and Art, 41 Cooper Square, New York, NY 10003, USA
[21] Department of Astronomy, University of Massachussets Amherst, Amherst, MA 01003, USA
[22] Center for Computational Mathematics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[23] Dipartimento di Matematica e Fisica. Universitá Roma Tre, via della Vasca Navale 84, I-00146, Roma, Italy
[24] INFN—National Institute for Nuclear Physics, via della Vasca Navale 84, I-00146 Roma, Italy
[25] Scuola Internazionale Superiore di Studi Avanzati. via Bonomea, 265, I-34136, Trieste, Italy
[26] INAF-OATs, Osservatorio Astronomico di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy
[27] IFPU—Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34151 Trieste, Italy
[28] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[29] Department of Physics and Astronomy, Tufts University, Medford, MA 02155, USA
[30] Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK
[31] South African Astronomical Observatories, Observatory, Cape Town 7925, South Africa
[32] Department of Physics, Yale University, New Haven, CT 06520, USA
[33] Department of Astronomy, University of Florida, Gainesville, FL, USA
[34] University of Florida Informatics Institute, 432 Newell Drive, CISE Bldg E251, Gainesville, FL, USA
[35] Sorbonne Universite, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis boulevard Arago, F-75014 Paris, France
[36] Department of Astronomy, Columbia University, 550 W 120th Street, New York, NY 10027, USA
[37] INFN—National Institute for Nuclear Physics, Via Valerio 2, I-34127 Trieste, Italy
[38] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[39] Cavendish Laboratory, University of Cambridge, 19 J.J. Thomson Avenue, Cambridge CB3 0HE, UK
[40] Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France
[41] Dipartimento di Fisica e Astronomia "Augusto Righi," Università di Bologna, via Gobetti 93/2, I-40129, Bologna, Italy
[42] Kavli Institute for Astrophysics and Space Research, Department of Physics, MIT, Cambridge, MA 02139, USA

Received 2022 January 10; revised 2023 January 30; accepted 2023 February 22; published 2023 April 6

Abstract

The Cosmology and Astrophysics with Machine Learning Simulations (CAMELS) project was developed to combine cosmology with astrophysics through thousands of cosmological hydrodynamic simulations and machine learning. CAMELS contains 4233 cosmological simulations, 2049 *N*-body simulations, and 2184 state-of-the-art hydrodynamic simulations that sample a vast volume in parameter space. In this paper, we present the CAMELS public data release, describing the characteristics of the CAMELS simulations and a variety of data products generated from them, including halo, subhalo, galaxy, and void catalogs, power spectra, bispectra, Lyα spectra, probability distribution functions, halo radial profiles, and X-rays photon lists. We also release over 1000 catalogs that contain billions of galaxies from CAMELS-SAM: a large collection of N-body simulations that have been combined with the Santa Cruz semianalytic model. We release all the data, comprising more than 350 terabytes and containing 143,922 snapshots, millions of halos, galaxies, and summary statistics. We provide further technical details on how to access, download, read, and process the data at https://camels.readthedocs.io.

Key words: Cosmology – Hydrodynamical simulations – Astrostatistics – Galaxy formation

1. Introduction

Recent advances in deep learning are triggering a revolution across fields, and cosmology and astrophysics are not left behind. Applications include parameter inference (Ravanbakhsh et al. 2017; Peek & Burkhart 2019; Hassan et al. 2020; Mangena et al. 2020; Ntampaka et al. 2020; Cole et al. 2022; Villaescusa-Navarro et al. 2021a, 2021c), superresolution (Kodi Ramanah et al. 2020; Li et al. 2021; Ni et al. 2021), generation of mock data (Zamudio-Fernandez et al. 2019; Han et al. 2021; Hassan et al. 2022), painting hydrodynamic properties on N-body simulations (Jo & Kim 2019; Yip et al. 2019; Zhang et al. 2019; Alves de Oliveira et al. 2020; Kasmanoff et al. 2020; Thiele et al. 2020; Moews et al. 2021; Wadekar et al. 2021; Bernardini et al. 2022; Harrington et al. 2022; Horowitz et al. 2022), improving the halo–galaxy connection (Wadekar et al. 2020; Lovell et al. 2021; Moster et al. 2021; Xu et al. 2021; Delgado et al. 2022), removing and/or cleaning astrophysical effects (Gagnon-Hartman et al. 2021; Makinen et al. 2021; Villanueva-Domingo & Villaescusa-Navarro 2021), emulating nonlinear evolution and speeding up numerical simulations (He et al. 2019; Kodi Ramanah et al. 2019; Dong et al. 2021; Kaushal et al. 2022), learning functions to interpolate among simulation properties (Giusarma et al. 2019; Chen et al. 2020), estimating masses of dark matter halos (Calderon & Berlind 2019; Man et al. 2019; Lucie-Smith et al. 2020; Villanueva-Domingo et al. 2021, 2022) and galaxy clusters (Ho et al. 2019; Ntampaka et al. 2019; Gupta & Reichardt 2020; Kodi Ramanah et al. 2020, 2021; Yan et al. 2020; de Andres et al. 2022), finding universal relations in subhalo properties (Shao et al. 2022), generating realistic galaxy images (Fussell & Moews 2019), model selection and classification (e.g., Hassan et al. 2019), and improving spectral energy distribution fitting techniques (Lovell et al. 2019; Gilda et al. 2021), among many others (see Stein 2020[43] for a comprehensive compilation). At its core, many of these results are based on using neural networks to approximate complex functions that may live in a high-dimensional space. These techniques have the potential to revolutionize the way we do cosmology and astrophysics.

From the cosmological side, we have now a well-established and accepted model: the ΛCDM model. This model not only describes the laws and constituents of our universe but also is capable of explaining a large variety of cosmological observables, from the temperature anisotropies of the cosmic microwave background to the spatial distribution of galaxies at low redshift. The model has free parameters characterizing fundamental properties of the universe such as its geometry, composition, the properties of dark energy, the sum of the neutrino masses, etc. One of the most important tasks in cosmology is to constrain the values of the these parameters with the highest degree of accuracy. In that way, we may be able to provide answers to fundamental questions such as "What is the nature of dark energy?" and "What are the masses of the neutrinos?"

Many studies have shown that there is a wealth of cosmological information located on mildly to highly nonlinear scales that need summary statistics other than the power spectrum to be retrieved (Allys et al. 2020; Dai et al. 2020; de la Bella et al. 2021; Friedrich et al. 2020; Hahn et al. 2020; Uhlemann et al. 2020; Villaescusa-Navarro et al. 2020a; Banerjee & Abel 2021a, 2021b; Hahn & Villaescusa-Navarro 2021; Hortua 2021; Massara et al. 2021; Naidoo et al. 2022; Porth et al. 2023). Extracting the maximum amount of information from these scales presents two main challenges. First, the optimal summary statistics that fully characterizes non-Gaussian density fields is currently unknown. Second, these scales are expected to be affected by astrophysical effects, such as feedback from supernovae (SNe) and active galactic nuclei (AGNs), in a poorly understood way (e.g., Somerville & Davé 2015; Naab & Ostriker 2017). Due to this uncertainty, cosmological analyses are typically carried out avoiding scales that are affected by astrophysical processes.

On the other hand, the cosmological dependence on astrophysical processes such as the formation and evolution of galaxies is typically neglected. Thus, while intrinsically linked, cosmology and galaxy formation tend to progress in parallel with limited interactions. Building bridges between cosmology and galaxy formation will thus benefit the development of both branches and contribute to an unified understanding.

Unfortunately, the interplay of cosmology and astrophysics takes places on many different scales, including nonlinear ones. This implies that cosmological hydrodynamic simulations are among the best tools to model and study the interactions between cosmology and astrophysics. However, given the uncertainties in both cosmology and galaxy formation models, it would be desirable to run simulations for different values of the cosmological parameters and also for different astrophysical models. Finally, if the number of simulations is large enough, one can make use of machine-learning techniques to extract the maximum amount of information from the simulations while at the same time being able to develop high-dimensional interpolators to explore the parameter space without having to run additional simulations.

The Cosmology and Astrophysics with Machine Learning Simulations (CAMELS) project (Villaescusa-Navarro et al. 2021b) was conceived to combine cosmology and astrophysics through numerical simulations and machine learning. At its core, CAMELS consists of a set of 4233 cosmological simulations that have different values of the cosmological parameters and different astrophysical models. All these virtual universes can be used as a large data set to train machine-learning algorithms.

The CAMELS project was first introduced and described in detail in Villaescusa-Navarro et al. (2021b). The theoretical

[43] https://github.com/georgestein/ml-in-cosmology

justification behind some of its main features (e.g., the use of a latin-hypercube covering a big volume in parameter space) was presented in Villaescusa-Navarro et al. (2020b). Since then, a number of different works have made use of the CAMELS simulations to carry out a large and diverse variety of tasks:

1. In Shao et al. (2022), CAMELS was used to identify a universal relation between subhalo properties using neural networks and symbolic regression.
2. In Mohammad et al. (2022), CAMELS was used to train convolutional neural networks to inpaint masked regions of highly nonlinear 2D maps from different physical fields.
3. In Villaescusa-Navarro et al. (2021a), CAMELS was used to show that neural networks can extract cosmological information and marginalize over baryonic effects at the field level using multiple fields simultaneously.
4. In Villaescusa-Navarro et al. (2021c), CAMELS was used to show that neural networks can place robust, percent level, constraints on $\Omega_{\rm m}$ and σ_8 from 2D maps containing the total matter mass of hydrodynamic simulations.
5. In Villaescusa-Navarro et al. (2022b), the CAMELS Multifield Data Set, a collection of hundreds of thousands of 2D maps and 3D grids for 13 different fields was presented and publicly released.
6. In Hassan et al. (2022), CAMELS was used to train a generative model that can produce diverse neural hydrogen maps by end of reionization ($z\sim 6$) as a function of cosmology.
7. In Villanueva-Domingo et al. (2022), a model based on graph neural networks (GNNs) was trained on the data from the CAMELS simulations to predict the total mass of a dark matter halo given its galactic properties while accounting for astrophysical uncertainties.
8. In Villanueva-Domingo et al. (2021), the GNN models proposed in Villanueva-Domingo et al. (2022) and trained on CAMELS data were used to obtain the first constrain on the mass of the Milky Way and Andromeda using artificial intelligence.
9. In Nicola et al. (2022), CAMELS was used to investigate the potential of auto- and cross-power spectra of the baryon distribution to robustly constrain cosmology and baryonic feedback.
10. In Wadekar et al. (2022), CAMELS was used to reduce the scatter in the Sunyaev–Zeldovich (SZ) flux–mass relation, Y–M, to provide more accurate estimates of cluster masses.
11. In D. Wadekar et al. (2023, in preparation), CAMELS was used to study deviations from self-similarity in the Y–M relation due to baryonic feedback processes, and to find an alternative relation that is more robust.
12. In Thiele et al. (2022), CAMELS was used to demonstrate the strong constraints that next-generation measurements of the y-distortions could provide on feedback models.
13. In Moser et al. (2022), CAMELS was used to compute thermal and kinetic SZ profiles. A Fisher analysis was performed to forecast the constraining power of observed SZ profiles on the astrophysical models varied in the simulations.
14. In Villaescusa-Navarro et al. (2022a), CAMELS was used to investigate whether the value of the cosmological parameters can be constrained using properties of a single galaxy.
15. In Y. Jo et al. et al. (2023, in preparation), CAMELS has been exploited to infer the full posterior on the combinations of cosmological and astrophysical parameters that reproduce observations such as cosmic star formation history and stellar mass functions using simulation-based inference.
16. L. Perez et al. (2023, in preparation) created CAMELS-SAM, a third larger "hump" of CAMELS by combining N-body simulations with the Santa Cruz (SC) semianalytic model (SAM) of galaxy formation. CAMELS-SAM contains billions of galaxies and represents a perfect tool to investigate and quantify the amount of cosmological information that can be extracted with galaxy redshift surveys.

In this paper, we describe the characteristics of the CAMELS simulations together with a variety of data products obtained from them, and we publicly release all available data. This paper is accompanied by the online documentation hosted at https://camels.readthedocs.io, containing further technical details on how to access, read, and manipulate CAMELS data. We believe that the CAMELS data will trigger new developments and findings in the fields of cosmology and galaxy formation.

This paper is organized as follows. In Section 2, we briefly describe the simulations of the CAMELS project and their scientific goals. The specifications of the data release are outlined in detail in Section 3. In Section 4, we describe how to access and download the data together with the overall data organization. We conclude in Section 5.

2. Simulations

2.1. Overview

CAMELS consists of a set of 4233 cosmological simulations: 2049 N-body and 2184 hydrodynamic. All simulations follow the evolution of 256^3 dark matter particles and 256^3 fluid elements (only the hydrodynamic simulations) from $z=127$, down to $z=0$ in a periodic box of $(25\,h^{-1}\,{\rm Mpc})^3$ volume. The initial conditions were generated at $z=127$ using second-order perturbation theory (2LPT).[44] The linear power spectra were computed using CAMB (Lewis et al. 2000). The mass resolution is approximately $1.27\times 10^7\,h^{-1}\,M_\odot$ per baryonic resolution element, and the gravitational softening length is approximately 2 kpc comoving. For each simulation, we saved 34 snapshots from $z=6$, to $z=0$, which are spaced piecewise equally in redshift-space at $z>3$ and equally in cosmological scale factor log space at $z<3$, and have intervals that overall range between ≈150 and 650 Myr. All simulations share the value of these cosmological parameters: $\Omega_{\rm b}=0.049$, $h=0.6711$, $n_s=0.9624$, $\sum m_\nu=0.0\,{\rm eV}$, $w=-1$. However, the value of $\Omega_{\rm m}$ and σ_8 varies from simulation to simulation.

The state-of-the-art hydrodynamic simulations have been run using two different codes, AREPO (Springel 2010; Weinberger et al. 2019) and GIZMO (Hopkins 2015), and they made use of the IllustrisTNG (Weinberger et al. 2017; Pillepich et al. 2018b) and SIMBA (Davé et al. 2019) galaxy formation models, respectively. However, the values of four astrophysical parameters vary from simulation to simulation. Two

[44] https://cosmo.nyu.edu/roman/2LPT/

parameters, $A_{\rm SN1}$ and $A_{\rm SN2}$, control the efficiency of SN feedback, while the other two parameters, $A_{\rm AGN1}$ and $A_{\rm AGN2}$, parameterize the efficiency of feedback from supermassive black holes, as described in more detail below. In Figure 1, we illustrate visually the effect of changing one of the astrophysical parameters in one simulation. As can be seen, while the large-scale structure remains unchanged, changing the efficiency of SN feedback has a large effect on both small and large galaxies.

For each hydrodynamic simulation, CAMELS includes its *N*-body counterpart. The *N*-body simulations have been run with GADGET-III (Springel 2005). With the simulation snapshots and initial conditions, we also release the GADGET parameter files, CAMB parameters files, and linear power spectra used to run the simulations.

2.2. Organization

The CAMELS simulations are divided into three different *suites*:

1. *IllustrisTNG*. All simulations run with the AREPO code and employing the IllustrisTNG model belong to this suite. There are 1092 IllustrisTNG simulations in CAMELS.
2. *SIMBA*. All simulations run with the GIZMO code and employing the SIMBA subgrid model belong to this suite. There are 1092 SIMBA simulations in CAMELS.
3. N*-body*. All *N*-body simulations belong to this suite. There are 2049 *N*-body simulations in CAMELS.

We provide further details on each suite below. Each simulation suite contains four different *sets*, depending on the way the values of the cosmological parameters, astrophysical parameters, and initial conditions random phases are organized:

1. *LH* stands for *latin-hypercube*. This set contains 1000 simulations, each with different values of $\Omega_{\rm m}$, σ_8, $A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, $A_{\rm AGN2}$, and the initial conditions random phases, with no repetitions of any of the above across the hydrodynamical suites. In the case of the *N*-body suite, this set contains 2000 simulations varying $\Omega_{\rm m}$, σ_8, and the initial conditions random phases, such that they match those from the IllustrisTNG and SIMBA LH sets.
2. *1P* stands for *one parameter at a time*. This set contains 61 simulations with the same values of the initial conditions random seed. The simulations only differ in the value of a single cosmological or astrophysical parameter at a time, with 11 variations for each, including the set of fiducial values. In the case of the *N*-body suite, this set contains 21 simulations varying $\Omega_{\rm m}$ and σ_8.
3. *CV* stands for *cosmic variance*. This set contains 27 simulations that share the values of the cosmological and astrophysical parameters. The simulations only differ in the value of the initial conditions random seed, which are however matched across the suites. There are 27 *N*-body counterpart simulations for this set.
4. *EX* stands for *extreme*. This set contains four simulations that have the same value of the initial conditions random seed (matched across the suites) and the same value of the cosmological parameters. One of them represents a model with no feedback, while the other two have either extremely large SN or AGN feedback. The *N*-body suite only contains one simulation.

For further details on the CAMELS simulations, we refer the reader to Villaescusa-Navarro et al. (2021b) and references therein.

2.3. Parameters

Both the IllustrisTNG and SIMBA simulation suites model galaxy formation by following Newtonian gravity in an expanding background, hydrodynamics, radiative cooling, star formation, stellar evolution and feedback, SMBH growth, and AGN feedback. IllustrisTNG also follows magnetic fields in the MHD limit, and SIMBA follows dust grains. The implementations of gravity and hydrodynamics solvers differ between the codes, as well as the parameterizations of radiative cooling, star formation, and stellar evolution. However, the most consequential differences between the suites are in the implementations of feedback in the form of galactic winds and from AGN, because the physics of these processes are the least theoretically understood as well as least observationally constrained. Therefore, these are also the parts of the physical modeling that we have chosen to apply variations to, through the parameters mentioned above, $A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, $A_{\rm AGN2}$, as described next.

CAMELS was designed to sample a large volume in parameter space. Thus, the value of both the cosmological and astrophysical parameters is varied within a very broad range:

$$\Omega_{\rm m} \in [0.1, 0.5], \tag{1}$$

$$\sigma_8 \in [0.6, 1.0], \tag{2}$$

$$A_{\rm SN1} \in [0.25, 4.0], \tag{3}$$

$$A_{\rm SN2} \in [0.5, 2.0], \tag{4}$$

$$A_{\rm AGN1} \in [0.25, 4.0], \tag{5}$$

$$A_{\rm AGN2} \in [0.5, 2.0]. \tag{6}$$

In both the LH and 1P sets, the value of $\Omega_{\rm m}$ and σ_8 is sampled linearly, while the value of the astrophysical parameters is varied in logarithmic scale.

In both models, $A_{\rm SN1}$ represents a normalization factor for flux of the galactic wind feedback. In IllustrisTNG it is implemented as a prefactor for the overall energy output per unit star formation (Pillepich et al. 2018b), while in SIMBA it is implemented as a prefactor for the mass-loading factor (galactic wind mass outflow rate per unit star formation rate, hereafter SFR) relative to that predicted by higher-resolution simulations (Anglés-Alcázar et al. 2017b). In both models, $A_{\rm SN2}$ represents a normalization factor for the speed of the galactic winds. This implies that, for a fixed $A_{\rm SN1}$, changes in $A_{\rm SN2}$ in IllustrisTNG affect the galactic wind speed in concert with the mass-loading factor (to keep a fixed energy output), while in SIMBA changes in $A_{\rm SN2}$ affect the galactic wind speed in concert with the galactic wind energy flux (with a fixed mass-loading factor).

In both models, $A_{\rm AGN1}$ represents a normalization factor for the energy output of AGN feedback while $A_{\rm AGN2}$ affects the specific energy of AGN feedback. However, the implementations of AGN feedback are quite significantly different between the suites and so is the effect of those parameters. In IllustrisTNG, $A_{\rm AGN1}$ is implemented as a prefactor for the overall power injected in the “kinetic” feedback mode (Weinberger et al. 2017), while in SIMBA it is implemented as a prefactor for the momentum flux of mechanical outflows

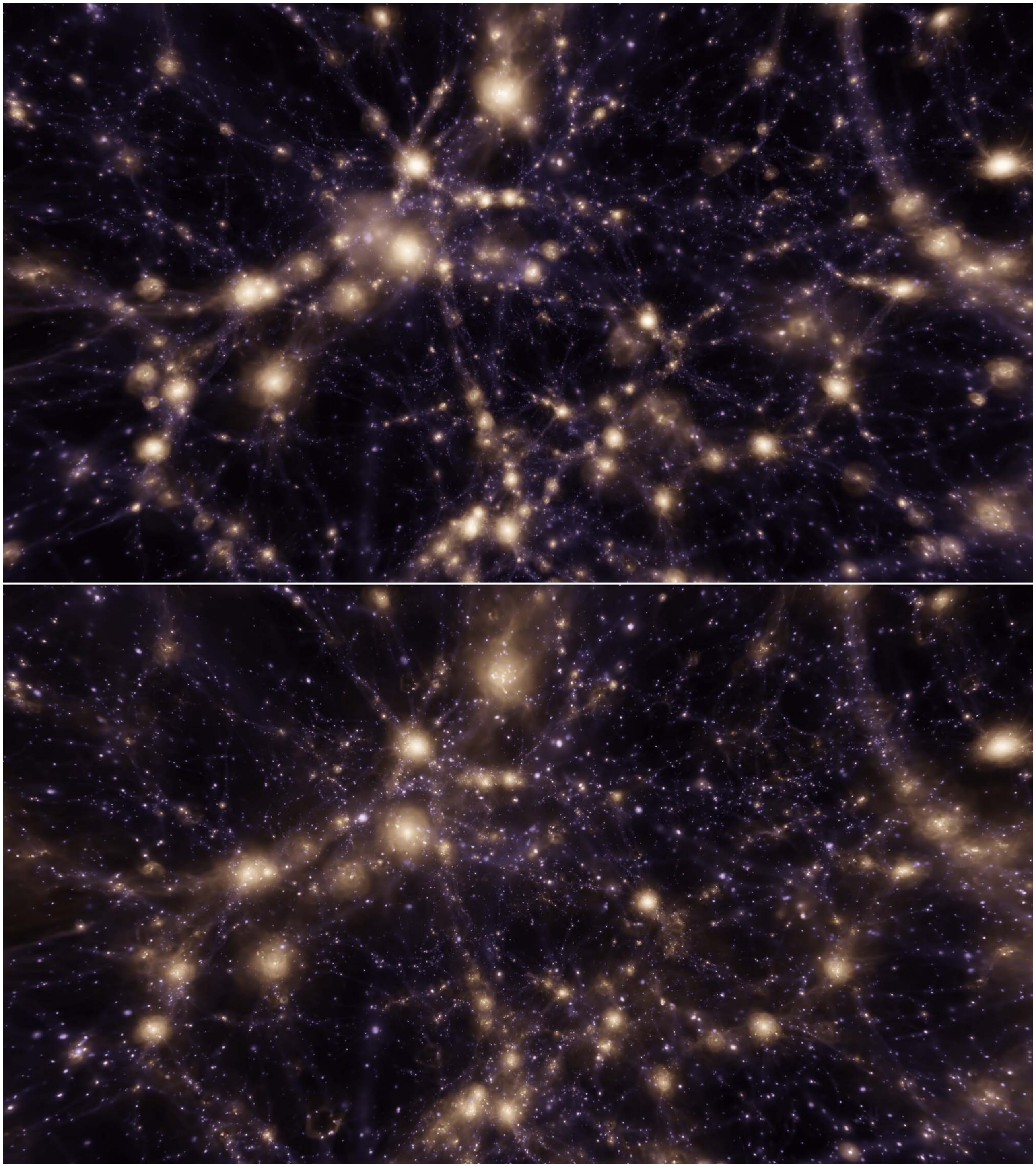

Figure 1. We show two images of the gas distribution of two distinct IllustrisTNG simulations. The one on the top displays the results for a simulation with high supernova feedback strength, while the one on the bottom is from a simulation with low supernova feedback. The color represents gas temperature, while its brightness corresponds to the gas density. Finally, we apply an extinction based on gas metallicity. As can be seen, the effect of feedback is very pronounced: it not only affects the gas abundance and temperature on the smallest galaxies but also changes the gas distribution in the most massive galaxies.

(Anglés-Alcázar et al. 2017a) in the "quasar" and "jet" feedback modes. In IllustrisTNG, $A_{\rm AGN2}$ directly parameterizes the burstiness and the temperature of the heated gas during AGN feedback "bursts," while in SIMBA it controls the speed of continuously driven AGN jets. We refer the reader to Villaescusa-Navarro et al. (2021b) for a detailed description of the feedback parameter variations in CAMELS.

It is very important to remark that, in light of the discussion above, while the cosmological parameters in the *N*-body, IllustrisTNG, and SIMBA suites represent the very same physical effect, the astrophysical parameters in the SIMBA and IllustrisTNG suites do not. The reason is that these parameters characterize similar physical processes but in different subgrid models. Thus, one should not attempt to match these parameters across suites. In other words, when doing, e.g., parameter inference from some observable to the value of the cosmological and astrophysical parameters, and the model is trained on IllustrisTNG simulations, one can attempt to test the model to see if it is able to recover the correct cosmology from SIMBA simulations. On the other hand, one should not try to infer the value of the astrophysical parameters of IllustrisTNG simulations from a model trained on SIMBA simulations.

We also emphasize that IllustrisTNG and SIMBA differ substantially on a variety of subgrid physics ingredients in addition to the feedback parameter variations considered in CAMELS, which should be taken into account when interpreting similarities or differences between the two simulation suites. For example, IllustrisTNG implements radiative cooling following Katz et al. (1996), Wiersma et al. (2009) assuming the spatially uniform ionizing background of Faucher-Giguère et al. (2009) while SIMBA radiative cooling and photoionization heating are modeled using the GRACKLE-3.1 library (Smith et al. 2017) assuming the spatially uniform ionizing background of Haardt & Madau (2012). In both cases, hydrogen self-shielding is modeled following Rahmati et al. (2013). IllustrisTNG employs the pressurized subgrid interstellar medium model of Springel & Hernquist (2003) with stars forming from gas above a hydrogen number density threshold of $n_{\rm H} \sim 0.13\,{\rm cm}^{-3}$, while star formation in SIMBA follows the molecular gas-based prescription of Krumholz & Gnedin (2011), and a minimum level of pressurization is included to resolve the Jeans mass in star-forming gas (Davé et al. 2016). Both IllustrisTNG and SIMBA assume a Chabrier (2003) stellar initial mass function and track chemical enrichment from Type II SNe, Type Ia SNe, and asymptotic giant branch stars following nine individual elements (H, He, C, N, O, Ne, Mg, Si, and Fe) but assume different yields (for details, see Davé et al. 2016; Pillepich et al. 2018a). Massive black holes accrete gas following the spherical Bondi (1952) parameterization in IllustrisTNG while SIMBA implements a two-phase prescription where cold gas accretion follows the gravitational torque accretion model (Hopkins & Quataert 2011; Anglés-Alcázar et al. 2013, 2017a), and hot gas accretion follows the Bondi (1952) parameterization.

To illustrate the differences between the IllustrisTNG and SIMBA simulations, we have taken all galaxies of all simulations belonging to the LH sets of both suites. For each galaxy, we consider eight different properties at $z = 0$, and Figure 2 shows their corresponding 1D and 2D distributions. As can be seen, while the distributions roughly overlap in all cases, there are some noticeable differences. SIMBA appears to be more efficient in forming galaxies at the low-mass end, but the overall SFR, gas metallicity, and stellar metallicity distributions are similar to IllustrisTNG. On the other hand, at fixed stellar mass, SIMBA tends to produce galaxies with higher maximum circular velocity compared to IllustrisTNG and smaller stellar effective radii at intermediate to low masses. Other interesting differences are seen in the black hole mass distributions, which show evidence for the higher seed mass used in IllustrisTNG along with lower galaxy mass threshold for seeding compared to SIMBA.

3. Data Description

In this section, we describe the different data products we release. We note that for each of the below data products we provide a code and/or instructions to read them. The reader can find all details on the associated website: https://camels.readthedocs.io.

3.1. Snapshots

We release the full snapshots generated by the GADGET-III, AREPO, and GIZMO codes. For each simulation, we have 34 snapshots from $z = 6$, down to $z = 0$ (we provide further details in the online documentation about the redshifts of the snapshots). We also release the initial conditions of each simulation.

All initial condition files and the snapshots of all simulations contain the positions, velocities, and IDs of the particles. The snapshots of the hydrodynamic simulations contain additional fields that store properties of the gas, stars, and black hole particles. Examples are the masses of the particles, the electron fraction from gas, or the age of the star particles. We note that the simulations from the IllustrisTNG and SIMBA suites are not identical in terms of the fields they store. The differences reflect the different subgrid models employed in these two simulations. The structure and contents of the IllustrisTNG snapshots are the same as in the publicly released IllustrisTNG simulation data (Nelson et al. 2019). Likewise, the SIMBA snapshots are identical in format to that available on the publicly released SIMBA database.[45]

The snapshots are stored as `hdf5` files, and we provide details in the online documentation on how to read and manipulate the data from them.

3.2. Halo and Subhalo Catalogs

We release the halo and subhalo catalogs generated from the CAMELS simulations. The halos and subhalos have been identified using SUBFIND (Springel et al. 2001; Dolag et al. 2009), ROCKSTAR (Behroozi et al. 2013a), and the Amiga halo finder (AHF; Knollmann & Knebe 2009). The codes have been run on top of all snapshots of all simulations. In total, we release 506,022 catalogs that contain millions of halos, subhalos, and galaxies. We now describe the catalogs in more detail.

3.2.1. Subfind

SUBFIND (Springel et al. 2001; Dolag et al. 2009) was run on-the-fly for the IllustrisTNG simulations while for the SIMBA and *N*-body simulations it was run in postprocessing. SUBFIND identifies both halos and subhalos, and computes several physical quantities for them in both the *N*-body and hydrodynamic simulations. We release all SUBFIND catalogs (one per simulation and redshift) for all simulations and all redshifts. The data is stored as `hdf5` files, and we provide details on how to read the data and the information stored in them in the online documentation.

3.2.2. Amiga Halo Finder

AHF (Knollmann & Knebe 2009) was run in postprocessing for the IllustrisTNG and SIMBA simulations. AHF utilizes isodensity contours to locate halo centers. Halo virial radii are defined to represent spherical overdensity regions with 200 times the critical density. We release the AHF catalogs for all simulations and redshift snapshots, including (1) global halo

[45] http://simba.roe.ac.uk

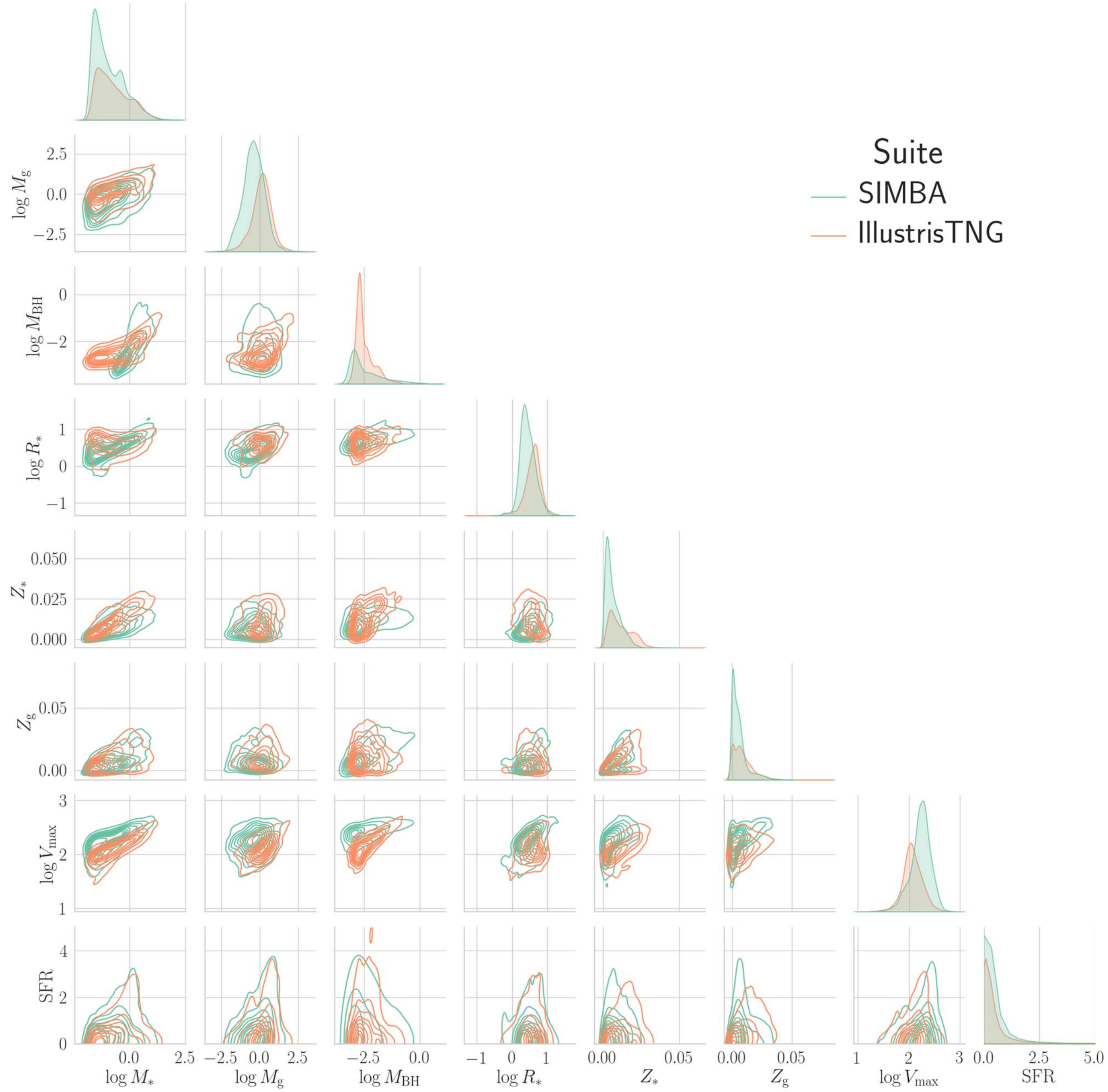

Figure 2. In this plot, we illustrate the similarities and differences between the IllustrisTNG and SIMBA suites considering eight different properties of the subhalos: (1) stellar mass, M_*, (2) gas mass, M_g, (3) black hole mass, M_{BH}, (4) stellar half-mass–radius, R_*, (5) stellar metallicity, Z_*, (6) gas metallicity, Z_g, (7) maximum circular velocity, V_{max}, and (8) star formation rate, SFR. We show the 1D and 2D distribution of these properties for all galaxies in the LH sets of the IllustrisTNG (orange) and SIMBA (green) suites. Masses are in units of $10^{10}/(M_\odot\, h^{-1})$, R_* in kiloparsecs per hour, V_{max} in kilometers per second, and SFR in $M_\odot$ yr^{-1}.

properties, (2) radial profiles, and (3) particle ID lists to identify the host halo of each particle. We provide further details on the format and how to read these catalogs in the online documentation.

3.2.3. Rockstar

In addition to the SUBFIND and AHF halo catalogs, we also release halo catalogs constructed using the ROCKSTAR halo finder (Behroozi et al. 2013a). ROCKSTAR identifies dark matter halos based on an adaptive hierarchical refinement of friends-of-friends (FoF) in 6D phase space plus time. Substructures are identified using a successively smaller linking length, and particles are assigned to the innermost substructures, which are defined as halo seeds, based on their phase-space proximity. We release ROCKSTAR halo catalogs of all 34 snapshots from $z = 6$ to 0 for all simulations.

Furthermore, we use CONSISTENT-TREES (Behroozi et al. 2013b) to generate merger trees from the ROCKSTAR halo catalogs. We note that CONSISTENT-TREES ensures consistency of halo mass, position, and velocity across time steps. Since all CAMELS simulations have only 34 snapshots, we perform the following exercise to quantify its validity. We have compared the ROCKSTAR + CONSISTENT-TREES outputs at $z = 0$ from two CAMELS simulations that have the same initial conditions but different time resolution (34 versus 200 snapshots). We find good agreement between the outputs for certain proxies of merger history such as peak mass and half-mass assembly time. However, we caution readers when using more detailed properties of the halo merger histories, such as accretion history, which are affected by the lower time sampling. All ROCKSTAR catalogs and CONSISTENT-TREES merger trees occupy 1.2 terabytes of data.

3.2.4. CAESAR

We release full crossmatched galaxy–halo catalogs for each snapshot generated using the YT extension package CAESAR.[46] CAESAR identifies halos based on a 3D FoF algorithm using a linking length of 0.2 times the mean interparticle spacing, and within each halo it identifies galaxies based on a 6D FoF with a linking length of 0.0056 times the mean interparticle spacing applied only to dense gas (hydrogen number density $n_{\rm H} > 0.13$ cm^{-3}) and star particles.

CAESAR then calculates a huge number of properties for each halo and galaxy, including physical properties such as masses and sizes for each component, dynamical properties such as velocity dispersions, and photometric properties using the Flexible Stellar Population Synthesis (Conroy et al. 2009; Conroy & Gunn 2010) package.[47] There are over 130 bands precomputed, both apparent and absolute magnitudes, both with and without dust extinction. Extinction is calculated via the line-of-sight dust content to each star along a chosen viewing axis (for IllustrisTNG, a Milky Way–like dust-to-metal ratio is assumed), providing pseudoradiative transfer that generally agrees with full radiative transfer calculations within $\lesssim$0.1 mag. An extinction law is assumed that is a composite of Milky Way for low galaxy specific star formation rates (sSFR $< 10^{-10}$ yr^{-1}), Calzetti for high (sSFR $> 10^{-9}$ yr^{-1}), and interpolated in between, with a further interpolation in galaxy stellar metallicity to incorporate a Small Magellanic Cloud (SMC) law such that at above solar metallicity no SMC law is folded in, while for metallicities below one-tenth solar it is fully SMC (regardless of sSFR).

All this information is stored in a single `hdf5` file for each snapshot, called a CAESAR catalog. Quantities from the catalog can be loaded into CAESAR using simple Python list comprehension, and it is straightforward to access halo information for a given galaxy and vice versa. CAESAR also provides particle membership lists for each galaxy/halo, so that one can compute any user-desired quantity by loading the particles from the original snapshot.[48] CAESAR also provides the functionality to compute progenitors and descendants of galaxies and/or halos across different snapshots; though this information has not been precomputed.

CAESAR catalogs typically are roughly 1% of the size of the corresponding snapshots, so they provide a compact and manageable way to access galaxy and halo data quickly and conveniently. CAESAR also interfaces seamlessly with `yt` for further analysis and visualization. See the online documentation and `caesar.readthedocs.io` for more details.

3.3. Void Catalogs

We release void catalogs built from the CAMELS simulations with the Void Identification and Examination toolkit (`VIDE`) from Sutter et al. (2015). `VIDE`, based on `ZOBOV` (Neyrinck 2008), has been widely used to find voids both in data—e.g., voids from the SDSS BOSS (Hamaus et al. 2016, 2020) and eBOSS (Aubert et al. 2022) data sets, or data from the Dark Energy Survey (Pollina et al. 2019)—and simulations (e.g., Kreisch et al. 2019, 2022; Verza et al. 2019; Contarini et al. 2021). Furthermore, `VIDE` has also been applied to hydrodynamic simulations (Habouzit et al. 2020; Panchal et al. 2020), showing its suitability for the CAMELS data set.

`VIDE` was run on top of CAMELS galaxies that were defined as subhalos containing more than 20 star particles. Given the size of the CAMELS simulations, and the extended size of cosmic voids (that usually span sizes from 5 to 100 h^{-1} Mpc), the number of voids for each CAMELS simulation is relatively small. The `VIDE` catalogs store information about the positions, sizes, ellipticities, and member galaxies of each void. In the online documentation, we provide further details on how to read and manipulate the `VIDE` catalogs.

3.4. Lyα Spectra

We release mock Lyα spectra generated using a public, well-tested code exhibited in Bird et al. (2015) and used previously for studies of the Lyα Forest in Gurvich et al. (2017). The spectra is generated for 5000 sight lines randomly placed through the simulation box. This spectral data was generated for the IllustrisTNG and SIMBA suites for all simulation sets at all redshifts. The locations of the random sight lines vary across snapshots.

The total absorption along a sight line is the sum of the absorption from all the nearby gas cells. The simulated spectra has a spectral resolution of 1 km s^{-1}, and any lines with an optical depth of $\tau < 10^{-5}$ are neglected. For further details on the artificial spectra calculation, we direct the reader to Bird et al. (2015). In the online documentation, we provide details on how to read and manipulate the Lyα spectra.

3.5. Summary Statistics

We release a large set of summary statistics, containing power spectra, bispectra, and probability distribution functions (PDFs). This data can be used for a large variety of tasks such as carrying out parameter inference and building emulators.

3.5.1. Power Spectrum

The power spectrum is the most prominent summary statistic of cosmology. The procedure used to carry out this task is the following. First, the positions and masses of the considered particles are read from the snapshots. Next, the masses of the particles are deposited into a regular grid with 512^3 voxels using the cloud-in-cell mass-assignment scheme (MAS). We then Fourier transform that field and correct modes amplitudes to account for the MAS. Finally, the power spectrum is computed by averaging the square of the modes amplitudes

$$P(k_i) = \frac{1}{N_i}\sum_{k\in k_{\rm bin}} |\delta(\boldsymbol{k})|^2, \qquad (7)$$

[46] https://caesar.readthedocs.io

[47] See https://caesar.readthedocs.io/en/latest/catalog.html for the full list of quantities.

[48] CAESAR works most straightforwardly with the PYGADGETREADER package for this; see the CAESAR documentation for examples.

where the k-bins have a width equal to the fundamental frequency, $k_F = 2\pi/L$ (L is the box size), and N_i is the number of modes in the k-bin. The wavenumber associated with each bin is

$$k_i = \frac{1}{N_i}\sum_{k\in k_{\rm bin}} k\ . \tag{8}$$

We have computed the power spectra of the total matter for both the N-body and the hydrodynamic simulations. Besides, for the hydrodynamic simulations, we have also computed the power spectra of the gas, dark matter, stars, and black hole components. We have done this for all snapshots of each simulation. We have made use of PYLIANS[49] to carry out the calculation. In total, we release 440,946 power spectra. All power spectra occupy $\simeq$10 gigabytes of data.

The above methods are inefficient if we wish to compute the power spectrum at large k, because they require a unwieldy fast Fourier transform (FFT) grid. In this regime, alternative methods such as configuration-space power spectrum estimators (Philcox & Eisenstein 2020) can be of use, because their computational cost decreases as the minimum scale increases. We provide power spectrum multipoles computed up to $k = 1000\,h\,{\rm Mpc}^{-1}$, and $\ell_{\rm max} = 4$, using a combination of the above PYLIANS code and the HIPSTER pair-counting approach package (Philcox 2021), switching between the two at $k = 25\,h\,{\rm Mpc}^{-1}$, and convolving the small-scale spectra with a window of size $R_0 = 1\,h^{-1}\,{\rm Mpc}$ for efficiency. Spectra are computed at $z = 0$ for all matter species listed above, and we include results from each simulation of the LH set from the IllustrisTNG, SIMBA, and N-body suites in both real- and redshift-space, with the latter using three choices of redshift-space axis. In total we compute 44,000 power spectra up to $k = 1000\,h\,{\rm Mpc}^{-1}$, requiring $\simeq$14,000 CPU-hours and occupying $\simeq$0.6 gigabytes of storage.

For all spectra, we store the value of k in each k-bin, the value of $P(k_i)$, and (for the large-scale spectra) the number of modes in each bin. We provide further details on how to read and manipulate these files in the online documentation.

3.5.2. Bispectrum

On large scales, the first non-Gaussian statistic of interest is the bispectrum, encoding the three-point average of the density field. In this release, we provide bispectrum measurements from gas, dark matter, and total matter for the 1000 simulations of the LH set of the IllustrisTNG and SIMBA suites, as well as 1000 N-body simulations. These are performed at redshift zero, both in real-space and redshift-space (for three choices of line of sight). Additional data can be computed upon request.

On large scales, bispectra are computed analogously to Section 3.5.1, first gridding the data with 128^3 voxels using a triangular-shaped-cloud MAS scheme. We then use the PYLIANS estimator (Villaescusa-Navarro 2018), implementing the approach of Watkinson et al. (2017), which practically computes the following sum via a series of FFTs:

$$B(k_1,\, k_2,\, \mu) = \frac{\sum_{\boldsymbol{k}_1}\sum_{\boldsymbol{k}_2}\delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2)\delta(-\boldsymbol{k}_1 - \boldsymbol{k}_2)}{N_T(k_1,\, k_2,\, \mu)}. \tag{9}$$

The bispectrum is parameterized by two lengths, k_1 and k_2, and an internal angle $\mu \equiv \hat{\boldsymbol{k}}_1 \cdot \hat{\boldsymbol{k}}_2$, with $N_T(k_1,\, k_2,\, \mu)$ giving the number of triangles per bin. We use 20 k-bins with $\Delta k = 0.25\,h\,{\rm Mpc}^{-1} \approx k_F$, and ten linearly spaced μ bins.

The above method becomes prohibitively expensive as $k_{\rm max}$ (and thus the FFT grid) increases. To ameliorate this, we compute the bispectra at large k using the HIPSTER code, as for the small-scale power spectra, here convolving the spectra with a smooth window of scale $R_0 = 2\,h^{-1}\,{\rm Mpc}$. This computes the Legendre multipoles of the bispectrum, related to Equation (9) by

$$B(k_1,\, k_2,\, \mu) = \sum_{\ell=0}^{\infty} B_\ell(k_1,\, k_2) L_\ell(\mu), \tag{10}$$

for Legendre polynomial $L_\ell(\mu)$, and uses 25 linearly spaced k-bins in the range $[0,\, 50]h\,{\rm Mpc}^{-1}$ for $\ell \leqslant 5$, subsampling to 10^5 particles for efficiency. These bispectra are computed for the same simulations as before, and will allow information to be extracted from very small scales. In total, 28,000 bispectra are estimated using each method, requiring $\simeq$70,000 CPU-hours and $\simeq$2.1 gigabytes of storage.

3.5.3. Probability Distribution Function

We estimate PDFs for 13 different physical fields using the 3D grids of the CAMELS Multifield Data set (CMD; see Section 3.8). The PDFs are calculated for all the fields: (1) gas temperature, (2) gas pressure, (3) neutral hydrogen density, (4) electron number density, (5) gas metallicity, (6) gas density, (7) dark matter density, (8) total mass density, (9) stellar mass density, (10) magnetic fields, (11) ratio between magnesium over iron, (12) gas velocity, and (13) dark matter velocity, for all the grid sizes, i.e., 128, 256, and 512 at redshifts 0.0, 0.5, 1.0, 1.5, and 2.0. The PDFs are calculated as follows. First, the 1000 3D grids from all simulations in the LH set are read into memory. We then calculate the minimum value across grids, and if it equals 0, a small offset is added to all voxels of all grids. The offset, ε, is given by

$$\varepsilon = \frac{\min_{\text{non-zero}}}{10^{20}}, \tag{11}$$

where $\min_{\text{non-zero}}$ denotes the nonzero minimum of all the 1000 grids. Then we log-transform the entire field (to the base 10) and construct a histogram of 500 bins between the minimum and maximum values of the entire field. Finally, we save to disk the number of counts in each bin for each grid in the considered field.

3.6. Profiles

We provide 3D spherically averaged profiles of gas density, thermal pressure, gas mass-weighted temperature, and gas mass-weighted metallicity for the 1P, LH, and CV sets of both the IllustrisTNG and SIMBA suites. We follow Moser et al.

[49] https://pylians3.readthedocs.io

(2021) in extracting halo information and construction of the profiles. Specifically, we use `illstack_CAMELS`[50] (a CAMELS-specific version of the original, more general code `illstack` used in Moser et al. 2021), to generate the 3D profiles, extending radially from 0.01 to 10 Mpc in 25 $\log_{10}$ bins. The profiles are stored in `hdf5` format, which can be read with the Python script provided in the `illstack_CAMELS` repository.

3.7. X-Rays

We provide mock X-ray photon lists in the form of SIMPUT fits files for all halos above $10^{12}\,M_\odot$ across all hydrodynamic CAMELS runs at redshift $z = 0.05$ obtained from the snapshot 032. The SIMPUT files are generated using the pyXSIM package[51] and contain positional coordinates in R.A. and decl. coordinates and energy in units of keV. These files serve as inputs into other software packages, including SOXS[52] and SIXTE (Dauser et al. 2019) that generate mock observations for specific telescopes using custom instrument profiles. These SIMPUT files can also represent idealized observations by an X-ray telescope, and we also provide a single collated file with projected radial surface brightness profiles for all halos for the soft X-ray band (0.5–2.0 keV) in units of ergs per second per squared kiloparsec. This file holds 1.6×10^5 radial profiles across the 2190 1P, CV, LH, and EX simulations.

3.8. CAMELS Multifield Data Set

The CMD is a collection of hundreds of thousands of 2D maps and 3D grids generated from CAMELS data. CMD contains 15,000 2D maps for 13 different fields at $z = 0$, and 15,000 3D grids, at three different spatial resolutions and at five different redshifts. The data was generated by assigning particles positions and properties (e.g., mass and temperature for the temperature field) to either 2D maps or 3D grids. There are many possible machine-learning applications of this data set, e.g., (1) parameter inference (Villaescusa-Navarro et al. 2021a, 2021c), (2) summary or field level emulation, (3) mapping N-body to hydrodynamic simulations, (4) super-resolution, and (5) time evolution. In total, CDM represents over 70 terabytes of data. We refer the reader to Villaescusa-Navarro et al. (2022b) and the CMD online documentation[53] - for further details on this data set.

3.9. CAMELS-SAM

CAMELS-SAM represents a newer third "hump" of CAMELS, mimicking its construction and purpose but using larger N-body volumes that are populated with galaxies using the SAM (Somerville et al. 2008, 2015) of galaxy formation. The N-body simulations are run with AREPO (Weinberger et al. 2019), and follow the evolution of 640^3 dark matter particles over a periodic box of $(100\ \mathrm{h}^{-1}\ \mathrm{cMpc})^3$ volume from $z = 127$, to $z = 0$. For each simulation, we save 100 snapshots, corresponding to those in the public IllustrisTNG data release,[54] which are spaced approximately equally in log space of the scale factor and correspond to ~100–200 Myr intervals. The initial conditions were otherwise generated as described in Section 2, with the same underlying cosmology, and a newly generated latin-hypercube varying Ω_m, σ_8, and three SAM parameters. Those parameters were chosen as the ones closest to the astrophysical parameters varied in CAMELS. Two parameters control the amplitude and rate of mass outflow from massive stars out of a galaxy, and the third parameter broadly controlling the strength of the radio jet mode of AGN.

Like CAMELS, CAMELS-SAM has an LH set containing 1000 simulations. The values of the cosmological and astrophysical parameters in the set are organized in a latin-hypercube. We additionally have five simulations in the CV set where the value of the initial random seed varies, and the five parameters are held fixed to their fiducial values. Finally, a 1P set with 12 simulations exists, where the SC-SAM was run at the smallest and largest value of each SAM parameter for two of the CV simulations.

It is important to emphasize the differences between the original CAMELS and the CAMELS-SAM simulations. First, CAMELS-SAM consists of N-body simulations with a volume 64× larger than that of the former, while CAMELS contains both N-body and hydrodynamic simulations. Second, CAMELS-SAM stored 100 snapshots while CAMELS only kept 34. Third, galaxies are modeled in very different ways: in CAMELS they arise from the hydrodynamic simulations while in CAMELS-SAM they are modeled through the SAM.

For all CAMELS-SAM simulations, we release the following:

1. the halo and subhalo catalogs from both SUBFIND and ROCKSTAR;
2. the merger trees generated from CONSISTENT -TREES;
3. the galaxy catalogs from the SAM.

The galaxy catalogs are stored as *.dat* text files with comma-separated values. These files contain information about the halo and galaxies from all snapshots of a given simulation. The exact available properties, their organization and units, and the example code to open these files can be found on the CAMELS-SAM online documentation.[55] The total size of these data products is around 35 terabytes.

The raw data (compressing full N-body snapshots across redshifts) has been stored on tape, and its content can be retrieved upon request. We refer the reader to L. Perez et al. (2023, in preparation) for further details on CAMELS-SAM, as well as a proof-of-concept of its power using clustering summary statistics to constrain cosmology and astrophysics with neural networks.

4. Data Access and Structure

In this section, we describe the different methods to access the data and its structure.

4.1. Data Access

We provide access to CAMELS data through four different platforms:

1. *Binder*. Binder is a system that allows users to read and manipulate data that is hosted at the Flatiron Institute through either a Jupyter notebook or a unix shell. The system provides access to the entire CAMELS data and

[50] https://github.com/emilymmoser/illstack_CAMELS
[51] http://hea-www.cfa.harvard.edu/~jzuhone/pyxsim/
[52] http://hea-www.cfa.harvard.edu/~jzuhone/soxs/
[53] https://camels-multifield-dataset.readthedocs.io
[54] https://www.tng-project.org
[55] https://camels-sam.readthedocs.io

allows users to perform calculations that do not require large amounts of CPU power. We note that heavy calculations are not supported by this system, and we recommend the user to download the data locally and work with it accordingly. We provide the link to the Binder environment in the online documentation. All CAMELS data can be accessed, read, and manipulated through Binder. We provide further technical details on Binder usage in the online documentation.

2. *Globus*. Globus[56] is a system designed to transfer large amounts of data in an efficient way. All CAMELS data can be transferred through Globus. We provide the Globus link in the online documentation.[57] Users can transfer the data to either another cluster or directly to their personal computer.
3. *URL*. We also provide a uniform resource locator (url) to access the data through a browser. We do not recommend transferring large quantities of data using this procedure, as both the speed and its reliability is much worse than those of Globus. On the other hand, to download small amounts of data, such as a particular power spectrum or a halo catalog, it may be useful. All CAMELS data can be accessed and downloaded through the url. We provide the url link in the online documentation where it will be always updated.
4. *FlatHUB*. FlatHUB is a platform that allows users to explore and compare data from different simulations by browsing and filtering the data, making simple preview plots, and downloading subsamples of the data. We provide access to the SUBFIND halo and subhalo catalogs of the IllustrisTNG and SIMBA suites through this platform. We provide a link to FlatHUB in the online documentation.

4.2. Data Structure

The data is organized in different folders that contain similar type of data:

1. *Sims*. This folder contains the raw data from the simulations, such as initial conditions, snapshots, and parameter files. This folder contains 205 terabytes of data.
2. *FOF_Subfind*. This folder contains the SUBFIND halo and subhalo catalogs described in Section 3.2.1. This folder contains 4 terabytes of data.
3. *AHF*. This folder contains the AHF halo catalogs described in Section 3.2.2. This folder contains 6 terabytes of data.
4. *Rockstar*. This folder contains the ROCKSTAR halo and subhalo catalogs together with the merger trees from CONSISTENT-TREES as described in Section 3.2.3. This folder contains 1 terabyte of data.
5. *Caesar*. This folder contains the CAESAR halo and galaxy catalogs described in Section 3.2.4. This folder contains around 1 terabyte of data.
6. *Pk*. This folder contains the power spectra described in Section 3.5.1. This folder contains approximately 10 gigabytes of data.
7. *Bk*. This folder contains the bispectra measurements described in Section 3.5.2. This folder contains approximately 2.6 gigabytes of data.
8. *CMD*. This folder contains the CMD. This folder contains 76 terabytes of data.
9. *VIDE_Voids*. This folder contains the void catalogs described in Section 3.3. This folder contains 200 megabytes of data.
10. *Lya*. This folder contains the Lyα spectra described in Section 3.4. This folder contains 14 terabytes of data.
11. *PDFs*. This folder contains the PDF measurements described in Section 3.5.3. This folder contains more than 1 gigabyte of data.
12. *Profiles*. This folder contains the spherically averaged 3D profiles described in Section 3.6. This folder contains 48 gigabytes of data.
13. *X-rays*. This folder contains the X-rays photon lists described in Section 3.7. This folder contains over 100 gigabytes of data.
14. *SCSAM*. This folder contains all CAMELS-SAM data products described in Section 3.9. This folder contains more than 50 terabytes.
15. *Utils*. This folder contains additional files that can be useful to the user, including a file with the value of the scale factors corresponding to simulation snapshots and files indicating the values of the cosmological and astrophysical parameters of each simulation.

When possible, we have organized the data in the different folders in a self-similar way. We show the generic data structure scheme in Figure 3. The data is first organized into folders that contain the following: (1) the IllustrisTNG hydrodynamic simulations, (2) the SIMBA hydrodynamic simulations, (3) the *N*-body counterparts of (1), and (4) the *N*-body counterparts of (2). Inside each of these folders, the user can find many different subfolders whose names refer to the specific simulation set and realization: e.g., the first simulation of the LH set is denoted as LH_0. Finally, inside each of those folders, the user can find the data with the particular characteristics of each data product. We note that these folders may contain data products for a particular CAMELS simulation at all redshifts.

For some data products, e.g., CMD and CAMELS-SAM, the data organization is slightly different than the one outlined above. In those cases, we provide further details in the online documentation.

5. Summary

The goal of the CAMELS project is to connect cosmology with astrophysics via thousands of state-of-the-art cosmological hydrodynamic simulations and extract the maximum amount of information from them via machine learning. CAMELS contains 4233 cosmological simulations, 2049 *N*-body simulations and 2184 state-of-the-art hydrodynamic simulations sampling a vast volume in parameter space using two independent codes that solve hydrodynamic equations and implement subgrid physics in very distinct ways. CAMELS data have already been used for a large variety of tasks, from providing the first constraints on the mass of the Milky Way and Andromeda galaxies using artificial intelligence to show that neural networks can extract information from vastly different physical fields while marginalizing over astrophysical effects at the field level.

[56] https://www.globus.org

[57] Since this link may change with time, we make it available in the online documentation, where it can be updated if needed.

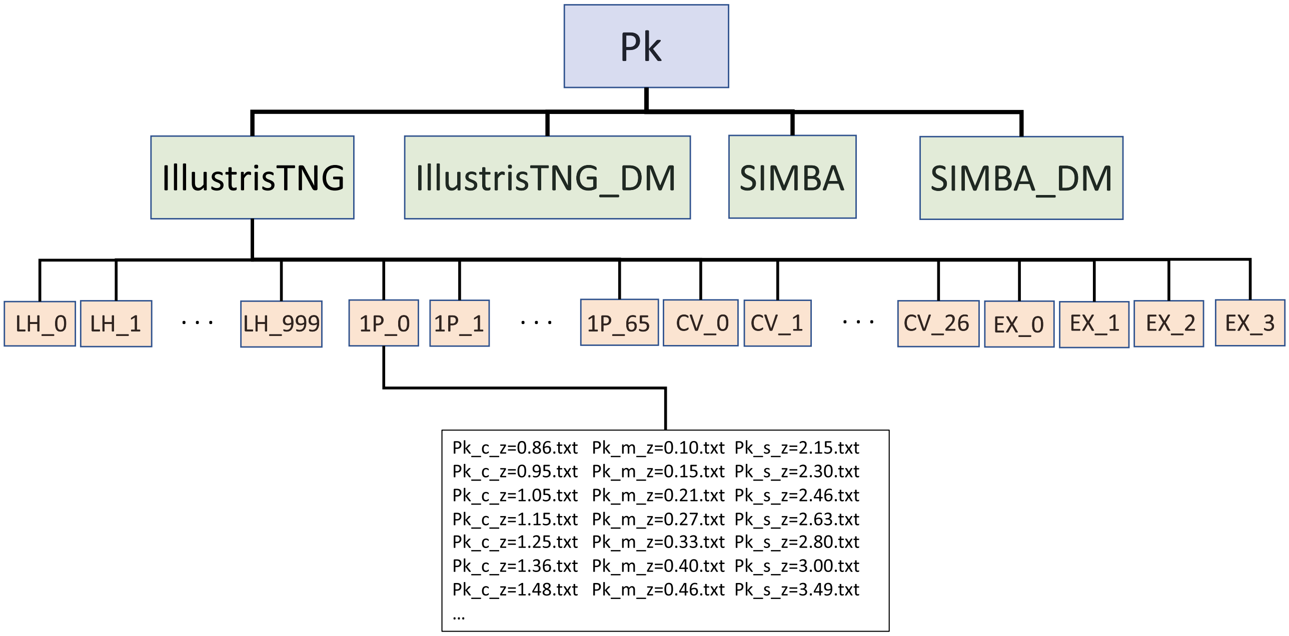
Pk
IllustrisTNG
IllustrisTNG_DM
SIMBA
SIMBA_DM
LH_0
LH_1
· · ·
LH_999
1P_0
1P_1
· · ·
1P_65
CV_0
CV_1
· · ·
CV_26
EX_0
EX_1
EX_2
EX_3
Pk_c_z=0.86.txt Pk_m_z=0.10.txt Pk_s_z=2.15.txt
Pk_c_z=0.95.txt Pk_m_z=0.15.txt Pk_s_z=2.30.txt
Pk_c_z=1.05.txt Pk_m_z=0.21.txt Pk_s_z=2.46.txt
Pk_c_z=1.15.txt Pk_m_z=0.27.txt Pk_s_z=2.63.txt
Pk_c_z=1.25.txt Pk_m_z=0.33.txt Pk_s_z=2.80.txt
Pk_c_z=1.36.txt Pk_m_z=0.40.txt Pk_s_z=3.00.txt
Pk_c_z=1.48.txt Pk_m_z=0.46.txt Pk_s_z=3.49.txt
...

Figure 3. This scheme shows the generic structure of CAMELS data. The top level represents the type of data it contains (power spectra in this case). Inside that folder, there are typically four folders containing the data for the three different simulation suites: IllustrisTNG, SIMBA, and their *N*-body counterparts (`IllustrisTNG_DM` and `SIMBA_DM`). Within each of those folders, there are numerous folders, containing the data from the different simulations belonging to each suite; i.e., the simulations from the four sets: LH, 1P, CV, and EX. Finally, inside each of those folders, the user can find the data products themselves. In this particular case, the power spectra for the different component.

In this paper, we have described the characteristics of the CAMELS simulations and a variety of additional data generated from them, including halo, subhalo, galaxy, and void catalogs, power spectra, bispectra, Lyα spectra, PDFs, radial profiles, and X-rays photon lists. We have also described CAMELS-SAM, a collection of more than 1000 galaxy catalogs created by applying the SAM to a set of hundreds of *N*-body simulations. We have made all this data publicly available, comprising hundreds of terabytes. We provide access to the data through different platforms, including a Binder environment for interactive data manipulation with Jupyter notebooks, a Globus link for efficient transfer of large amounts of data, and the FlatHUB platform for quick exploration of SUBFIND (sub)halo catalogs. We emphasize that the information outlined in this paper may become outdated as additional data products become available over time. However, the online documentation located at https://camels.readthedocs.io will always be updated accordingly.

It is also important to be aware of the limitations associated to the CAMELS simulations. First, the volume sampled by each individual simulation is relatively small, $(25\ h^{-1}\ \mathrm{Mpc})^3$, inhibiting the formation of the most extreme objects in the universe such as galaxy clusters and large voids. Second, while CAMELS covers a large volume in parameter space, it would be desirable to make it even larger by including other cosmological and astrophysical parameters. Third, CAMELS only contains two distinct suites of hydrodynamic simulations: IllustrisTNG and SIMBA. Ideally, we would like to expand CAMELS to simulations performed with additional codes employing different subgrid models. Fourth, the resolution of CAMELS may not be high enough for some astrophysical problems. Future versions of CAMELS will be designed to tackle these limitations.

We believe that CAMELS data will become a powerful tool for the community.

We are indebted to the high-performance computing system administrators at the Flatiron Institute and Princeton University for their invaluable help and work accommodating the CAMELS storage needs. The authors are pleased to acknowledge that the work reported in this paper was partially performed using the research computing resources at Princeton University, which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's Research Computing Division. Further technical details on the CAMELS simulations and instructions to download the data can be found at https://camels.readthedocs.io and https://www.camel-simulations.org. The work of F.V.N., S.G., D.A. A., S.H., O.P., A.P., K.W., W.C., M.E., U.S., D.S., B.B., B.W., R.S., and G.B. was supported by the Simons Foundation. D.A. A. acknowledges support by NSF grants AST-2009687 and AST-2108944, Chandra X-Ray Observatory grant TM2-23006X, and Simons Foundation award CCA-1018464. P.V. D. acknowledges support from the Generalitat Valenciana through the GenT program (CIDEGENT/2018/019, CPI-21-108). B.B. is grateful for generous support by the David and Lucile Packard Foundation and Alfred P. Sloan Foundation. The Flatiron Institute is supported by the Simons Foundation. The CAMELS project is supported by the Simons Foundation and NSF grant AST 2108078.

ORCID iDs

Francisco Villaescusa-Navarro https://orcid.org/0000-0002-4816-0455
Shy Genel https://orcid.org/0000-0002-3185-1540
Daniel Anglés-Alcázar https://orcid.org/0000-0001-5769-4945
Lucia A. Perez https://orcid.org/0000-0002-8449-1956
Pablo Villanueva-Domingo https://orcid.org/0000-0002-0936-4279

Digvijay Wadekar https://orcid.org/0000-0002-2544-7533
Helen Shao https://orcid.org/0000-0002-0152-6747
Faizan G. Mohammad https://orcid.org/0000-0001-9243-7434
Sultan Hassan https://orcid.org/0000-0002-1050-7572
Emily Moser https://orcid.org/0000-0003-1593-1505
Erwin T. Lau https://orcid.org/0000-0001-8914-8885
Luis Fernando Machado Poletti Valle https://orcid.org/0000-0002-1948-3562
Leander Thiele https://orcid.org/0000-0003-2911-9163
Yongseok Jo https://orcid.org/0000-0003-3977-1761
Oliver H. E. Philcox https://orcid.org/0000-0002-3033-9932
Benjamin D. Oppenheimer https://orcid.org/0000-0002-3391-2116
Megan Tillman https://orcid.org/0000-0002-1185-4111
ChangHoon Hahn https://orcid.org/0000-0003-1197-0902
Neerav Kaushal https://orcid.org/0000-0003-4786-2348
Alice Pisani https://orcid.org/0000-0002-6146-4437
Joyce Caliendo https://orcid.org/0000-0001-9719-7177
Christina Kreisch https://orcid.org/0000-0002-5061-7805
Kaze W. K. Wong https://orcid.org/0000-0001-8432-7788
William R. Coulton https://orcid.org/0000-0002-1297-3673
Gabriele Parimbelli https://orcid.org/0000-0002-2539-2472
Ulrich P. Steinwandel https://orcid.org/0000-0001-8867-5026
Romeel Dave https://orcid.org/0000-0003-2842-9434
Daisuke Nagai https://orcid.org/0000-0002-6766-5942
David N. Spergel https://orcid.org/0000-0002-5151-0006
Lars Hernquist https://orcid.org/0000-0001-6950-1629
Blakesley Burkhart https://orcid.org/0000-0001-5817-5944
Desika Narayanan https://orcid.org/0000-0002-7064-4309
Benjamin Wandelt https://orcid.org/0000-0002-5854-8269
Rachel S. Somerville https://orcid.org/0000-0002-6748-6821
Greg L. Bryan https://orcid.org/0000-0003-2630-9228
Matteo Viel https://orcid.org/0000-0002-2642-5707
Yin Li https://orcid.org/0000-0002-0701-1410
Vid Irsic https://orcid.org/0000-0002-5445-461X
Katarina Kraljic https://orcid.org/0000-0001-6180-0245
Federico Marinacci https://orcid.org/0000-0003-3816-7028
Mark Vogelsberger https://orcid.org/0000-0001-8593-7692

References

Allys, E., Marchand, T., Cardoso, J. F., et al. 2020, PhRvD, 102, 103506
Alves de Oliveira, R., Li, Y., Villaescusa-Navarro, F., Ho, S., & Spergel, D. N. 2020, arXiv:2012.00240
Anglés-Alcázar, D., Davé, R., Faucher-Giguère, C.-A., Özel, F., & Hopkins, P. F. 2017a, MNRAS, 464, 2840
Anglés-Alcázar, D., Faucher-Giguère, C.-A., Kereš, D., et al. 2017b, MNRAS, 470, 4698
Anglés-Alcázar, D., Özel, F., & Davé, R. 2013, ApJ, 770, 5
Aubert, M., Cousinou, M.-C., Escoffier, S., et al. 2022, MNRAS, 513, 186
Banerjee, A., & Abel, T. 2021a, MNRAS, 500, 5479
Banerjee, A., & Abel, T. 2021b, MNRAS, 504, 2911
Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013a, ApJ, 762, 109
Behroozi, P. S., Wechsler, R. H., Wu, H.-Y., et al. 2013b, ApJ, 763, 18
Bernardini, M., Feldmann, R., Anglés-Alcázar, D., et al. 2022, MNRAS, 509, 1323
Bird, S., Haehnelt, M., Neeleman, M., et al. 2015, MNRAS, 447, 1834
Bondi, H. 1952, MNRAS, 112, 195
Calderon, V. F., & Berlind, A. A. 2019, MNRAS, 490, 2367
Chabrier, G. 2003, PASP, 115, 763
Chen, C., Li, Y., Villaescusa-Navarro, F., Ho, S., & Pullen, A. 2020, arXiv:2012.05472
Cole, A., Miller, B. K., Witte, S. J., et al. 2022, JCAP, 2022, 004
Conroy, C., & Gunn, J. E. 2010, ApJ, 712, 833
Conroy, C., Gunn, J. E., & White, M. 2009, ApJ, 699, 486
Contarini, S., Marulli, F., Moscardini, L., et al. 2021, MNRAS, 504, 5021
Dai, J.-P., Verde, L., & Xia, J.-Q. 2020, JCAP, 2020, 007
Dauser, T., Falkner, S., Lorenz, M., et al. 2019, A&A, 630, A66
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827
Davé, R., Thompson, R., & Hopkins, P. F. 2016, MNRAS, 462, 3265
de Andres, D., Cui, W., Ruppin, F., et al. 2022, EPJWC, 257, 00013
de la Bella, L. F., Tessore, N., & Bridle, S. 2021, JCAP, 2021, 001
Delgado, A. M., Wadekar, D., Hadzhiyska, B., et al. 2022, MNRAS, 515, 2733
Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, MNRAS, 399, 497
Dong, X., Ramachandra, N., Habib, S., et al. 2021, arXiv:2112.05681
Faucher-Giguère, C., Lidz, A., Zaldarriaga, M., & Hernquist, L. 2009, ApJ, 703, 1416
Friedrich, O., Uhlemann, C., Villaescusa-Navarro, F., et al. 2020, MNRAS, 498, 464
Fussell, L., & Moews, B. 2019, MNRAS, 485, 3203
Gagnon-Hartman, S., Cui, Y., Liu, A., & Ravanbakhsh, S. 2021, MNRAS, 504, 4716
Gilda, S., Lower, S., & Narayanan, D. 2021, ApJ, 916, 43
Giusarma, E., Reyes Hurtado, M., Villaescusa-Navarro, F., et al. 2019, arXiv:1910.04255
Gupta, N., & Reichardt, C. L. 2020, ApJ, 900, 110
Gurvich, A., Burkhart, B., & Bird, S. 2017, ApJ, 835, 175
Haardt, F., & Madau, P. 2012, ApJ, 746, 125
Habouzit, M., Pisani, A., Goulding, A., et al. 2020, MNRAS, 493, 899
Hahn, C., & Villaescusa-Navarro, F. 2021, JCAP, 2021, 029
Hahn, C., Villaescusa-Navarro, F., Castorina, E., & Scoccimarro, R. 2020, JCAP, 2020, 040
Hamaus, N., Pisani, A., Choi, J.-A., et al. 2020, JCAP, 2020, 023
Hamaus, N., Pisani, A., Sutter, P. M., et al. 2016, PhRvL, 117, 091302
Han, D., Sehgal, N., & Villaescusa-Navarro, F. 2021, PhRvD, 104, 123521
Harrington, P., Mustafa, M., Dornfest, M., Horowitz, B., & Lukić, Z. 2022, ApJ, 929, 160
Hassan, S., Andrianomena, S., & Doughty, C. 2020, MNRAS, 494, 5761
Hassan, S., Liu, A., Kohn, S., & La Plante, P. 2019, MNRAS, 483, 2524
Hassan, S., Villaescusa-Navarro, F., Wandelt, B., et al. 2022, ApJ, 937, 83
He, S., Li, Y., Feng, Y., et al. 2019, PNAS, 116, 13825
Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, ApJ, 887, 25
Hopkins, P. F. 2015, MNRAS, 450, 53
Hopkins, P. F., & Quataert, E. 2011, MNRAS, 415, 1027
Horowitz, B., Dornfest, M., Lukić, Z., & Harrington, P. 2022, ApJ, 941, 42
Hortua, H. J. 2021, arXiv:2112.11865
Jo, Y., & Kim, J.-h. 2019, MNRAS, 489, 3565
Kasmanoff, N., Villaescusa-Navarro, F., Tinker, J., & Ho, S. 2020, arXiv:2012.00186
Katz, N., Weinberg, D. H., & Hernquist, L. 1996, ApJS, 105, 19
Kaushal, N., Villaescusa-Navarro, F., Giusarma, E., et al. 2022, ApJ, 930, 115
Knollmann, S. R., & Knebe, A. 2009, ApJS, 182, 608
Kodi Ramanah, D., Charnock, T., & Lavaux, G. 2019, PhRvD, 100, 043515
Kodi Ramanah, D., Charnock, T., Villaescusa-Navarro, F., & Wandelt, B. D. 2020, MNRAS, 495, 4227
Kodi Ramanah, D., Wojtak, R., Ansari, Z., Gall, C., & Hjorth, J. 2020, MNRAS, 499, 1985
Kodi Ramanah, D., Wojtak, R., & Arendse, N. 2021, MNRAS, 501, 4080
Kreisch, C. D., Pisani, A., Carbone, C., et al. 2019, MNRAS, 488, 4413
Kreisch, C. D., Pisani, A., Villaescusa-Navarro, F., et al. 2022, ApJ, 935, 100
Krumholz, M. R., & Gnedin, N. Y. 2011, ApJ, 729, 36
Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473
Li, Y., Ni, Y., Croft, R. A. C., et al. 2021, PNAS, 118, 2022038118
Lovell, C. C., Acquaviva, V., Thomas, P. A., et al. 2019, MNRAS, 490, 5503
Lovell, C. C., Wilkins, S. M., Thomas, P. A., et al. 2021, MNRAS , 509, 5046
Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., & Thiyagalingam, J. 2020, arXiv:2011.10577
Makinen, T. L., Lancaster, L., Villaescusa-Navarro, F., et al. 2021, JCAP, 2021, 081
Man, Z.-Y., Peng, Y.-J., Shi, J.-J., et al. 2019, ApJ, 881, 74
Mangena, T., Hassan, S., & Santos, M. G. 2020, MNRAS, 494, 600
Massara, E., Villaescusa-Navarro, F., Ho, S., Dalal, N., & Spergel, D. N. 2021, PhRvL, 126, 011301
Moews, B., Davé, R., Mitra, S., Hassan, S., & Cui, W. 2021, MNRAS, 504, 4024
Mohammad, F. G., Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., & Vogelsberger, M. 2022, ApJ, 941, 132
Moser, E., Amodeo, S., Battaglia, N., et al. 2021, ApJ, 919, 2
Moser, E., Battaglia, N., Nagai, D., et al. 2022, ApJ, 933, 133

Moster, B. P., Naab, T., Lindström, M., & O'Leary, J. A. 2021, MNRAS, 507, 2115
Naab, T., & Ostriker, J. P. 2017, ARA&A, 55, 59
Naidoo, K., Massara, E., & Lahav, O. 2022, MNRAS, 513, 3596
Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 2
Neyrinck, M. C. 2008, MNRAS, 386, 2101
Ni, Y., Li, Y., Lachance, P., et al. 2021, MNRAS, 507, 1021
Nicola, A., Villaescusa-Navarro, F., Spergel, D. N., et al. 2022, JCAP, 2022, 046
Ntampaka, M., Eisenstein, D. J., Yuan, S., & Garrison, L. H. 2020, ApJ, 889, 151
Ntampaka, M., ZuHone, J., Eisenstein, D., et al. 2019, ApJ, 876, 82
Panchal, R. R., Pisani, A., & Spergel, D. N. 2020, ApJ, 901, 87
Peek, J. E. G., & Burkhart, B. 2019, ApJL, 882, L12
Philcox, O. H. E. 2021, MNRAS, 501, 4004
Philcox, O. H. E., & Eisenstein, D. J. 2020, MNRAS, 492, 1214
Pillepich, A., Nelson, D., Hernquist, L., et al. 2018a, MNRAS, 475, 648
Pillepich, A., Springel, V., Nelson, D., et al. 2018b, MNRAS, 473, 4077
Pollina, G., Hamaus, N., Paech, K., et al. 2019, MNRAS, 487, 2836
Porth, L., Bernstein, G. M., Smith, R. E., & Lee, A. J. 2023, MNRAS, 518, 3344
Rahmati, A., Pawlik, A. H., Raicevic, M., & Schaye, J. 2013, MNRAS, 430, 2427
Ravanbakhsh, S., Oliva, J., Fromenteau, S., et al. 2017, arXiv:1711.02033
Shao, H., Villaescusa-Navarro, F., Genel, S., et al. 2022, ApJ, 927, 85
Smith, B. D., Bryan, G. L., Glover, S. C. O., et al. 2017, MNRAS, 466, 2217
Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51
Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E., & Hernquist, L. 2008, MNRAS, 391, 481
Somerville, R. S., Popping, G., & Trager, S. C. 2015, MNRAS, 453, 4337
Springel, V. 2005, MNRAS, 364, 1105
Springel, V. 2010, MNRAS, 401, 791
Springel, V., & Hernquist, L. 2003, MNRAS, 339, 289
Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726
Stein, G. 2020, georgestein/ml-in-cosmology: Machine Learning in Cosmology, v1.0, Zenodo, doi:10.5281/zenodo.4024768
Sutter, P. M., Lavaux, G., Hamaus, N., et al. 2015, A&C, 9, 1
Thiele, L., Villaescusa-Navarro, F., Spergel, D. N., Nelson, D., & Pillepich, A. 2020, ApJ, 902, 129
Thiele, L., Wadekar, D., Hill, J. C., et al. 2022, PhRvD, 105, 083505
Uhlemann, C., Friedrich, O., Villaescusa-Navarro, F., Banerjee, A., & Codis, S. 2020, MNRAS, 495, 4006
Verza, G., Pisani, A., Carbone, C., Hamaus, N., & Guzzo, L. 2019, JCAP, 2019, 040
Villaescusa-Navarro, F. 2018, Pylians: Python Libraries for the Analysis of Numerical Simulations, Astrophysics Source Code Library, ascl:1811.008
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021a, arXiv:2109.09747
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021b, ApJ, 915, 71
Villaescusa-Navarro, F., Ding, J., Genel, S., et al. 2022a, ApJ, 929, 132
Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2021c, arXiv:2109.10360
Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2022b, ApJS, 259, 61
Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. 2020a, ApJS, 250, 2
Villaescusa-Navarro, F., Wandelt, B. D., Anglés-Alcázar, D., et al. 2020b, arXiv:2011.05992
Villanueva-Domingo, P., & Villaescusa-Navarro, F. 2021, ApJ, 907, 44
Villanueva-Domingo, P., Villaescusa-Navarro, F., Anglés-Alcázar, D., et al. 2022, ApJ, 935, 30
Villanueva-Domingo, P., Villaescusa-Navarro, F., Genel, S., et al. 2021, arXiv:2111.14874
Wadekar, D., Thiele, L., Villaescusa-Navarro, F., et al. 2022, arXiv:2201.01305
Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, arXiv:2012.00111
Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2021, ApJ, 916, 42
Watkinson, C. A., Majumdar, S., Pritchard, J. R., & Mondal, R. 2017, MNRAS, 472, 2436
Weinberger, R., Springel, V., Hernquist, L., et al. 2017, MNRAS, 465, 3291
Weinberger, R., Springel, V., & Pakmor, R. 2019, arXiv:1909.04667
Wiersma, R. P. C., Schaye, J., Theuns, T., Dalla Vecchia, C., & Tornatore, L. 2009, MNRAS, 399, 574
Xu, X., Kumar, S., Zehavi, I., & Contreras, S. 2021, MNRAS, 507, 4879
Yan, Z., Mead, A. J., Van Waerbeke, L., Hinshaw, G., & McCarthy, I. G. 2020, MNRAS, 499, 3445
Yip, J. H. T., Zhang, X., Wang, Y., et al. 2019, arXiv:1910.07813
Zamudio-Fernandez, J., Okan, A., Villaescusa-Navarro, F., et al. 2019, arXiv:1904.12846
Zhang, X., Wang, Y., Zhang, W., et al. 2019, arXiv:1902.05965