Snowmass2021 - Letter of Interest

[Dark Energy Discovery with Multi-Survey Cross-Correlations]

Thematic Areas: (check all that apply □/■)

□ (CF1) Dark Matter: Particle Like
□ (CF2) Dark Matter: Wavelike
□ (CF3) Dark Matter: Cosmic Probes
■ (CF4) Dark Energy and Cosmic Acceleration: The Modern Universe
□ (CF5) Dark Energy and Cosmic Acceleration: Cosmic Dawn and Before
■ (CF6) Dark Energy and Cosmic Acceleration: Complementarity of Probes and New Facilities
□ (CF7) Cosmic Probes of Fundamental Physics
□ (Other) *[Please specify frontier/topical group]*

Contact Information: (authors listed after the text)
Submitter Name/Institution: Andrew Hearin (ANL), Alexie Leauthaud (UCSC), Daisuke Nagai (Yale)
Contact Email: ahearin@anl.gov

Abstract: Cosmological surveys of galaxies and clusters in the 2020s will measure large-scale structure growth with the statistical precision needed to address some of the most fundamental questions in physics. To meet the modeling challenges presented by such high-precision measurements, cosmological cross-correlations have emerged as a hallmark of modern galaxy surveys, in large part due to the stringent control these measurements offer on systematics. Numerous groups have demonstrated the potential for dramatic gains in cosmological constraining power that can be reaped from such measurements, particularly in joint analyses of galaxies and clusters that include information from the nonlinear regime. However, present-day modeling techniques are ill-suited to achieve these gains. While extensions of perturbation theory offer robust predictions with relatively few assumptions, analyses of spatial scales smaller than the quasi-linear regime will remain a formidable challenge for such methods for the foreseeable future. Alternative approaches based on the connection between galaxies and dark matter halos can in principle unlock cosmological information from far smaller scales, but *all* such efforts are based on highly simplistic models of the galaxy-halo connection *of a single tracer population*. The deficiencies of contemporary theoretical prediction tools are brought into sharp focus by the goal to truly leverage the full power of survey data in the 2020s with Cross-Survey Cosmological Cross-Correlations (CSC3). In the CSC3 program, multi-wavelength measurements of large-scale structure are utilized in a *joint* analysis of data from multiple surveys at once. The linchpin to this effort is a simulation-based forward modeling methodology that transforms high-resolution N-body simulations into synthetic skies with all three components of the density field: dark matter, baryonic gas, and stellar mass. Modeling techniques befitting the quality and richness of data in the 2020s will additionally harness the differentiability and computational performance of deep learning algorithms, replacing existing approaches based on classical machine learning that do not scale to the problem sizes of CSC3. The signature deliverables of CSC3 include both robust cosmological posteriors as well as tight constraints on the physics of galaxy formation.

Scientific Context: Astronomical surveys spanning a wide range of wavelengths are poised to tackle a number of fundamental questions in cosmology and astrophysics. What is the root physical origin of the dark energy responsible for cosmic acceleration at late times? What is the sum of the neutrino masses? Is there evidence to rule out single-field inflation? What is the particle nature of dark matter? What drives the observed bimodality in galaxy color and star formation rate? Where are the so-called "missing baryons"?

Through measurements made with a panoply of instruments (e.g., Roman, Euclid, Rubin, Simons, CMB-S4, eROSITA, DESI, SPHEREx), upcoming surveys will probe all three components of the universe: dark matter, gas, and galaxies. Furthermore, for the first time, the spatial extent of these datasets is such that *simultaneous information on all three components will be available over large swaths of the sky*. This new era of large scale, multi-wavelength, and overlapping surveys will open up the exciting prospect of *using all data sets simultaneously to constrain all three components*. We will refer to this ambitious program as *Cross-Survey Cross-Correlation Cosmology (CSC3)*.

The power of CSC3 will come from the fact that it maximally utilizes information from all surveys, from all components of the universe, as well as information from both linear and non-linear scales. Salcedo et al. [13] show that percent-level constraints on cosmological parameters can be obtained with measurements of the clustering and lensing of galaxies and clusters, as well as their cross-spectra, even when marginalizing over uncertainty in galaxy formation physics [7;15]. Similar points were also emphasized by two Astro2020 Science White Papers [2;3]. Furthermore, cosmological constraints are improved by a factor of 3 to 4 when small scales are included [12;16] – roughly the same scientific gain that would be accomplished by covering 15 times more sky! *CSC3 is a holy grail in cosmology for the upcoming decade, and most, if not all, of the fundamental questions listed above, will only be fully addressed when the CSC3 vision is realized.*

Although considerable progress towards the goals of CSC3 have been made on the observational side [1;2;6;8;9;11;14], joint survey analyses are still quite limited in terms of the utilized range of spatial scales, and also the simplistic assumptions about the connection between tracer gas and galaxies and the underlying cosmic density field. There is a long-standing effort to address some of these shortcomings using machine learning to accelerate simulation-based cosmological predictions [5], which has now become a standard technique adopted by many groups building fast "emulators" [4;16]. Such efforts typically rely upon classical machine learning techniques such as Gaussian Process that scale poorly with the problem sizes and dimensions of joint, multi-survey analysis, and are commonly built upon simplistic empirical models such as the Halo Occupation Distribution (HOD). Due to the very formulation of HOD-type models, incorporating new constraints from additional tracer galaxy populations requires a significant expansion of the parameter space, and/or reliance upon plausibly-violated assumptions about the galaxy-halo connection. Thus conventional halo occupation models actually *penalize* attempts to incorporate new constraining data, and in this sense these models bear the mark of the era of single-survey analysis in which they were developed.

Historically, generating physically realistic multi-wavelength predictions has required modeling approaches such as *hydrodynamical simulations* or *Semi-Analytic Models* (SAMs). Although such models are irreplaceable in the effort to understand a detailed picture of the physical processes operating within galaxies and clusters, neither approach would be feasible for inference based on direct simulation-based predictions of multi-survey cross-correlations. The scientific payload of CSC3 can only be delivered with simulations of sufficient size to contain large statistical samples of clusters (requiring simulated volumes of $\mathcal{O}(1)\mathrm{Gpc}^3$), and of sufficient resolution to accurately track the internal structure and evolutionary history of Milky Way mass halos ($M_{\rm halo} \gtrsim 10^{12} M_\odot$). Critically, our primary science target is *cosmological inference*, in which we will derive confidence intervals on model parameters, and any attempt at rigorous Bayesian inference requires evaluating the likelihood of *millions* of proposed universes, using $\sim$ 100 survey-scale simulations run over a range of cosmological parameters.

Hydrodynamical simulations meeting our specifications are so computationally expensive that any *in-*

dividual simulation demands $\sim 10-100$M cpu-hours. While state-of-the-art SAMs optimized to run on contemporary supercomputers have recently achieved capability to derive confidence intervals on galaxy formation parameters using a single simulation of volume $\sim 0.1\mathrm{Gpc}^3$, the ability to run converged Markov Chain Monte Carlos (MCMC) for SAMs implemented in high-resolution, Gpc-sized simulated volumes of varying cosmology is likely out of reach for precision-cosmology in the 2020s. Considerable recent progress has been made by a new generation of empirical models that bridge the gap between the level of complexity achieved by SAMs and the computational efficiency of empirical models, e.g., UniverseMachine[3] and EMERGE[10]. The ability of these models to make CPU-efficient multi-wavelength predictions is quite promising, but significant further advances are needed on both the modeling and computation side for this new approach to conduct multi-wavelength inference with survey-scale simulations. We conclude this section by noting that no individual simulation, regardless of its quality, can achieve the aims of the proposed CSC3 program. Thus, *theoretical techniques with practical capability to conduct multi-survey cosmological inference currently do not exist, and so the field of theoretical cosmology is ill-equipped for the quality, richness, and volume of cosmological data that will arrive in the 2020s.*

Key Requirements for this program: A full CSC3 program will leverage lensing measurements (Rubin, Roman), IR/optical measurements (SPHEREx, Rubin, Roman), microwave observations (SO, S4), X-ray data (eROSITA), and will dramatically improve on the constraining power of any individual survey. The largest constraining power will come from cross-correlations between dark matter, gas, and massive galaxies at $z < 1$ where the effects of late time dark energy are the most important. For example, DESI will measure redshifts for 15 million massive galaxies at $z < 1$ and SPHEREx will provide redshifts for 50 million galaxies with $M_\star > 10^{11} M_\odot$ and with a redshift precision of $\delta z < 0.03$. Such galaxies live in dark matter halos with masses $M_h > 10^{12}$, where both the impact of AGN and stellar feedback must be modeled.

In order to harness the power of CSC3, the desired model should be realistic and flexible enough to predict the spatial distributions of the three components without biasing cosmological constraints. A CSC3 program must also model all radial scales and account for "baryonic effects" (modifications to gas and dark matter distributions due to feedback from AGN and stellar winds). To achieve the cosmological goals of CSC3, *it is not critical for the fine-grained physics of galaxy- and cluster-formation be precisely known and modeled*. Instead, the models need only be accurate enough to capture the cosmological-dependence of structure formation. Computationally, CSC3 models must be able to generate predictions for observable data vectors fast enough to be able to run MCMC chains to derive likelihoods.

In order to meet the predictive needs associated with the incoming flood of multi-wavelength astronomical data, and to maximize the scientific returns of the upcoming surveys, we consider it critical and urgent for the cosmology community to invest in the development of a new generation of modeling approaches that address the limitations of contemporary techniques. Such development will provide the community with the capability to carry out numerous studies:

- *Theoretical CSC3 modeling of new generation multi-wavelength data (DESI, SPHEREx, Roman, Rubin, Euclid, SO/S4, eROSITA) for improved dark energy constraints.*
- *Accurate mock generation with self-consistent treatment of gas, galaxies, groups and clusters.*
- *Computation of data covariance matrices from accurate multi-wavelength mocks.*
- *Capability to study survey design and optimal combinations of multi-wavelength statistics.*
- *Capability to forward model systematic errors that arise from model limitations.*

A new modeling capability outlined here will be a critical step forward towards achieving the science returns of CSC3 in the 2020s.

References

[1] Abbott, T. M. C. & DES Collaboration, 2019: Cosmological Constraints from Multiple Probes in the Dark Energy Survey. , **122(17)**, 171301.

[2] Battaglia, N., J. C. Hill, S. Amodeo, J. G. Bartlett, K. Basu, J. Erler, S. Ferraro, L. Hernquist, M. Madhavacheril, & M. McQuinn, 2019: Probing Feedback in Galaxy Formation with Millimeter-wave Observations. In , vol. 51, p. 297.

[3] Behroozi, P., M. Becker, F. C. v. d. Bosch, C. Conroy, M. Dickinson, C. M. Hirata, A. Hearin, A. Leauthaud, C. Ly, & Y.-Y. Mao, 2019: Empirically Constraining Galaxy Evolution. In , vol. 51, p. 125.

[4] Euclid Collaboration, M. Knabenhans, J. Stadel, S. Marelli, D. Potter, R. Teyssier, L. Legrand , A. Schneider, B. Sudret, L. Blot, S. Awan, C. Burigana, C. S. Carvalho, H. Kurki-Suonio, & G. Sirri, 2019: Euclid preparation: II. The EUCLIDEMULATOR - a tool to compute the cosmology dependence of the nonlinear matter power spectrum. , **484(4)**, 5509–5529.

[5] Heitmann, K., D. Higdon, C. Nakhleh, & S. Habib, 2006: Cosmic Calibration. *ApJL*, **646**, L1–L4.

[6] Hildebrandt, H., F. Köhlinger, J. L. van den Busch, B. Joachimi, C. Heymans, A. Kannawadi, A. H. Wright, M. Asgari, C. Blake, H. Hoekstra, S. Joudaki, K. Kuijken, L. Miller, C. B. Morrison, T. Tröster, A. Amon, M. Archidiacono, S. Brieden, A. Choi, J. T. A. de Jong, T. Erben, B. Giblin, A. Mead, J. A. Peacock, M. Radovich, P. Schneider, C. Sifón, & M. Tewes, 2020: KiDS+VIKING-450: Cosmic shear tomography with optical and infrared data. , **633**, A69.

[7] Krause, E. & T. Eifler, 2017: cosmolike - cosmological likelihood analyses for photometric galaxy surveys. , **470(2)**, 2100–2112.

[8] Krolewski, A., S. Ferraro, E. F. Schlafly, & M. White, 2020: unWISE tomography of Planck CMB lensing. , **2020(5)**, 047.

[9] Leauthaud, A., S. Saito, S. Hilbert, A. Barreira, S. More, M. White, S. Alam, P. Behroozi, K. Bundy, J. Coupon, T. Erben, C. Heymans, H. Hildebrandt, R. Mandelbaum, L. Miller, B. Moraes, M. E. S. Pereira, S. A. Rodríguez-Torres, F. Schmidt, H.-Y. Shan, M. Viel, & F. Villaescusa-Navarro, 2017: Lensing is low: cosmology, galaxy formation or new physics? , **467(3)**, 3024–3047.

[10] Moster, B. P., T. Naab, & S. D. M. White, 2017: EMERGE - An empirical model for the formation of galaxies since $z \sim 10$. *ArXiv:1705.05373*.

[11] Münchmeyer, M., M. S. Madhavacheril, S. Ferraro, M. C. Johnson, & K. M. Smith, 2019: Constraining local non-Gaussianities with kinetic Sunyaev-Zel'dovich tomography. , **100(8)**, 083508.

[12] Reid, B. A., H.-J. Seo, A. Leauthaud, J. L. Tinker, & M. White, 2014: A 2.5 per cent measurement of the growth rate from small-scale redshift space clustering of SDSS-III CMASS galaxies. , **444(1)**, 476–502.

[13] Salcedo, A. N., B. D. Wibking, D. H. Weinberg, H.-Y. Wu, D. Ferrer, D. Eisenstein, & P. Pinto, 2020: Cosmology with stacked cluster weak lensing and cluster-galaxy cross-correlations. , **491(3)**, 3061–3081.

[14] Shirasaki, M., E. T. Lau, & D. Nagai, 2020: Probing cosmology and cluster astrophysics with multi-wavelength surveys - I. Correlation statistics. , **491(1)**, 235–253.

[15] Zentner, A. R., E. Semboloni, S. Dodelson, T. Eifler, E. Krause, & A. P. Hearin, 2013: Accounting for baryons in cosmological constraints from cosmic shear. , **87(4)**, 043509.

[16] Zhai, Z., J. L. Tinker, M. R. Becker, J. DeRose, Y.-Y. Mao, T. McClintock, S. McLaughlin, E. Rozo, & R. H. Wechsler, 2019: The Aemulus Project. III. Emulation of the Galaxy Correlation Function. , **874(1)**, 95.