Opportunities in AI/ML for the Rubin LSST Dark Energy Science Collaboration

LSST DESC Notes

Opportunities in AI/ML for the Rubin LSST Dark Energy Science Collaboration

Version 1.0 – January 2026

The LSST Dark Energy Science Collaboration (DESC), Eric Aubourg,[1] Camille Avestruz,[2,3] Matthew R. Becker,[4] Biswajit Biswas,[4] Rahul Biswas,[5] Boris Bolliet,[6,7] Adam S. Bolton,[8] Clecio R. Bom,[9] Raphaël Bonnet-Guerrini,[10] Alexandre Boucaud,[11] Jean-Eric Campagne,[12] Chihway Chang,[13,14,15] Aleksandra Ćiprijanović,[16,13,15] Johann Cohen-Tanugi,[17] Michael W. Coughlin,[18] John Franklin Crenshaw,[19,20,8] Juan C. Cuevas-Tello,[21] Juan de Vicente,[22] Seth W. Digel,[8,19] Steven Dillmann,[19,23,8] Mariano Javier de León Dominguez Romero,[24] Alex Drlica-Wagner,[16,13,25,15] Sydney Erickson,[20,8] Alexander T. Gagliano,[26,27,28] Christos Georgiou,[29] Aritra Ghosh,[30] Matthew Grayling,[31] Kirill A. Grishin,[11] Alan Heavens,[32] Lindsay R. House,[15,33] Mustapha Ishak,[34] Wassim Kabalan,[11] Arun Kannawadi,[35] François Lanusse,[36] C. Danielle Leonard,[37] Pierre-François Léget,[38] Michelle Lochner,[39] Yao-Yuan Mao,[40] Peter Melchior,[41] Grant Merz,[42] Martin Millon,[43] Anais Möller,[44] Gautham Narayan,[42,15] Yuuki Omori,[13,14,15] Hiranya Peiris,[31] Laurence Perreault-Levasseur,[45,46,47] Andrés A. Plazas Malagón,[19,8,41] Nesar Ramachandra,[4] Benjamin Remy,[13,15] Cécile Roucelle,[11] Jaime Ruiz-Zapatero,[48] Stefan Schuldt,[49,50,51,52] Ignacio Sevilla-Noarbe,[22] Ved G. Shah,[53,54,15] Tjitske Starkenburg,[53,54,15] Stephen Thorp,[31] Laura Toribio San Cipriano,[22] Tilman Tröster,[43] Roberto Trotta,[55,32] Padma Venkatraman,[42] Amanda Wasserman,[42,15] Tim White,[56] Justine Zeghal,[45,47] Tianqing Zhang,[57] *and* Yuanyuan Zhang[58]

[1] *Université Paris Cité, CNRS, CEA, Astroparticule et Cosmologie, F-75013 Paris, France*
[2] *Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[3] *Leinweber Institute of Theoretical Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA*
[5] *Independent*
[6] *Cavendish Astrophysics, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[7] *Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[8] *SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[9] *Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brazil*
[10] *Department of Computer Science, University of Milan, Milan, Italy*
[11] *Université Paris Cité, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France*
[12] *Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France*
[13] *Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[14] *Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[15] *NSF-Simons AI Institute for the Sky (SkAI), 172 E. Chestnut St., Chicago, IL 60611, USA*
[16] *Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA*
[17] *Université Clermont-Auvergne, CNRS, LPCA, 63000 Clermont-Ferrand, France*
[18] *School of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA*
[19] *Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA*
[20] *Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[21] *Engineering Faculty, Universidad Autonoma de San Luis Potosi, Zona Universitaria, San Luis Potosi, 78290, Mexico*
[22] *Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain*
[23] *Stanford Artificial Intelligence Laboratory, Stanford University, Stanford, CA 94305, USA*
[24] *Instituto de Astronomía Teórica y Experimental (IATE - UNC and CONICET CCT Córdoba), Observatorio Astronómico de Córdoba, Universidad Nacional de Córdoba, Laprida 854, X5000BGR, Córdoba, Argentina*
[25] *Kavli Institute of Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[26] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[27] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[28] *Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[29] *Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain*
[30] *Department of Astronomy & DiRAC Institute, University of Washington, Seattle, WA 98195, USA*
[31] *Institute of Astronomy and Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK*
[32] *Imperial Centre for Inference and Cosmology (ICIC), Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*
[33] *Data Science Institute, The University of Chicago, Chicago, IL 60615, USA*

[34] Department of Physics, The University of Texas at Dallas, Richardson, TX 75080, USA
[35] Department of Physics, Duke University, Durham, NC 27708, USA
[36] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, F-91191 Gif-sur-Yvette, France
[37] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom
[38] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[39] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town, 7535, South Africa
[40] Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA
[41] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA
[42] Department of Astronomy, University of Illinois Urbana Champaign, 1002 W. Green St., Urbana, IL, 61801, USA
[43] Institute for Particle Physics and Astrophysics, ETH Zürich, Wolfgang-Pauli-Strasse 27, CH-8093 Zurich, Switzerland
[44] Swinburne University of Technology, Hawthorn, Victoria 3122, Australia
[45] Département de Physique, Université de Montréal, 1375 Avenue Thérèse-Lavoie-Roux, Montréal, QC, H2V 0B3, Canada
[46] Ciela - Montréal Institute for Astrophysical Data Analysis and Machine Learning, Montréal, QC H2V 0B3, Canada
[47] Mila - Quebec Artificial Intelligence Institute, Montréal, QC H2S 3H1, Canada
[48] Advanced Research Computing Centre, University College London, 90 High Holborn, London WC1V 6LJ, UK
[49] Dipartimento di Fisica, Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy
[50] Finnish Centre for Astronomy with ESO (FINCA), University of Turku, FI-20014 Turku, Finland
[51] Department of Physics, P.O. Box 64, University of Helsinki, FI-00014 Helsinki, Finland
[52] INAF - IASF Milano, via A. Corti 12, I-20133 Milano, Italy
[53] Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA
[54] Center for Interdisciplinary Exploration and Research in Astrophysics, Northwestern University, Evanston, IL, USA
[55] Theoretical and Scientific Data Science, International School for Advanced Study, Via Bonomea 265, I-34136 Trieste, Italy
[56] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
[57] Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, Pittsburgh, PA 15260, USA
[58] NSF NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA

The Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) will produce unprecedented volumes of heterogeneous astronomical data—images, catalogs, and alerts—that challenge traditional analysis pipelines. The LSST Dark Energy Science Collaboration (DESC) aims to derive robust constraints on dark energy and dark matter from these data, requiring methods that are statistically powerful, scalable, and operationally reliable. Artificial intelligence and machine learning (AI/ML) are already embedded across DESC science workflows, from photometric redshifts and transient classification to weak lensing inference and cosmological simulations. Yet their utility for precision cosmology hinges on trustworthy uncertainty quantification, robustness to covariate shift and model misspecification, and reproducible integration within scientific pipelines. This white paper surveys the current landscape of AI/ML across DESC's primary cosmological probes and cross-cutting analyses, revealing that the same core methodologies and fundamental challenges recur across disparate science cases. Since progress on these cross-cutting challenges would benefit multiple probes simultaneously, we identify key methodological research priorities, including Bayesian inference at scale, physics-informed methods, validation frameworks, and active learning for discovery. With an eye on emerging techniques, we also explore the potential of the latest foundation model methodologies and LLM-driven agentic AI systems to reshape DESC workflows, provided their deployment is coupled with rigorous evaluation and governance. Finally, we discuss critical software, computing, data infrastructure, and human capital requirements for the successful deployment of these new methodologies, and consider associated risks and opportunities for broader coordination with external actors. Taken together, DESC's combination of community-accessible data, demanding scientific requirements, and mature simulation infrastructure makes the collaboration an excellent testbed for developing and validating robust AI/ML practices for fundamental physics.

Contributors

As a way to provide transparency in a large, multi-author writing effort, this white paper adopts the CRediT[1] taxonomy to report individual contributions. In this white paper, authorship roles are defined as follows:

Conceptualization: Paper- or section-level intellectual framing: defining scope, structure, key messages, and narrative arc.
Project administration: Process coordination: soliciting and organizing inputs, managing timelines and revisions, integrating contributions, and ensuring internal consistency across contributors.
Writing – Original Draft: Substantive preparation of original text, including drafting new material and/or synthesizing multiple contributions into a coherent section or subsection.
Writing – Review & Editing: Substantive review and revision of the manuscript text, including critical feedback, edits for clarity and correctness, and incorporation of comments during iterative drafting.

Where shown, bracketed scope tags (e.g., `[§4.1]`) indicate the section(s) associated with a listed role; `[all]` denotes contributions spanning the full manuscript. Bolded scopes identify primary contributions, and unbolded scopes identify secondary contributions.

Name	Contribution
Eric Aubourg	Writing – Original Draft [§3.9, §5.1]
Camille Avestruz	Writing – Original Draft [**§3.4**, §3.8]
Matthew R. Becker	Conceptualization [all]; Writing – Review & Editing [all]
Biswajit Biswas	Writing – Original Draft [§3.8]
Rahul Biswas	Writing – Original Draft [§3.1]; Writing – Review & Editing [all]
Boris Bolliet	Writing – Original Draft [§5.2]
Adam S. Bolton	Conceptualization [**§6**]; Project administration [**§6**]; Writing – Original Draft [**§6**]; Writing – Review & Editing [**§6**]
Clecio R. Bom	Conceptualization [**§5.2**]; Project administration [**§5.2**]; Writing – Original Draft [**§5.2**]; Writing – Review & Editing [**§5.2**]
Raphaël Bonnet-Guerrini	Writing – Original Draft [§4.1]; Writing – Review & Editing [§4.1]
Alexandre Boucaud	Writing – Original Draft [§3.9, §5.1]
Jean-Eric Campagne	Writing – Original Draft [§3.1, §3.2, §4.1, §4.2, §5.1]
Chihway Chang	Writing – Review & Editing [§3.3]
Aleksandra Ćiprijanović	Writing – Original Draft [§3, §4]; Writing – Review & Editing [§3, §4]
Johann Cohen-Tanugi	Writing – Original Draft [§3.8]; Writing – Review & Editing [all]
Michael W. Coughlin	Writing – Original Draft [§6]
John Franklin Crenshaw	Writing – Original Draft [§3.1]
Juan C. Cuevas-Tello	Writing – Original Draft [§3.2]
Juan de Vicente	Writing – Original Draft [§3.1]

[1] CRediT – Contributor Roles Taxonomy: https://credit.niso.org/

Name	Contribution
Seth W. Digel	Writing – Original Draft [§1]; Writing – Review & Editing [**all**]
Steven Dillmann	Writing – Original Draft [§5.1, §5.2]; Writing – Review & Editing [§5.1, §5.2]
Mariano Javier de León Dominguez Romero	Writing – Review & Editing [all]
Alex Drlica-Wagner	Writing – Original Draft [§3.8]; Writing – Review & Editing [all]
Sydney Erickson	Writing – Original Draft [§3.2]; Writing – Review & Editing [§3.2]
Alexander T. Gagliano	Conceptualization [**all**]; Project administration [**all**]; Writing – Original Draft [**all**]; Writing – Review & Editing [**all**]
Christos Georgiou	Writing – Original Draft [§2, §3.6]
Aritra Ghosh	Writing – Original Draft [§5, §6]
Matthew Grayling	Writing – Original Draft [§3.5]
Kirill A. Grishin	Writing – Original Draft [§3]
Alan Heavens	Writing – Original Draft [§3.3]
Lindsay R. House	Writing – Review & Editing [§1]
Mustapha Ishak	Writing – Original Draft [§3.6]
Wassim Kabalan	Writing – Original Draft [§3.7, §4.1]
Arun Kannawadi	Writing – Original Draft [§3.10]
François Lanusse	Conceptualization [**all**]; Project administration [**all**]; Writing – Original Draft [**all**]; Writing – Review & Editing [**all**]
C. Danielle Leonard	Writing – Original Draft [**§3.6**]
Pierre-François Léget	Writing – Original Draft [§3.5]
Michelle Lochner	Conceptualization [**§5.1**, **§1**]; Project administration [**§5.1**]; Writing – Original Draft [**§1**, **§4.3**, **§5.1**, **§9**]; Writing – Review & Editing [all]
Yao-Yuan Mao	Writing – Original Draft [§8]
Peter Melchior	Conceptualization [**§4**]; Project administration [**§4**]; Writing – Original Draft [**§4**]; Writing – Review & Editing [**§4**]
Grant Merz	Writing – Original Draft [§3.8]
Martin Millon	Writing – Original Draft [§3.2]; Writing – Review & Editing [§3.2]
Anais Möller	Conceptualization [**§4**]; Project administration [**§4**]; Writing – Original Draft [§3, **§4**]; Writing – Review & Editing [**§4**]
Gautham Narayan	Writing – Review & Editing [all]
Yuuki Omori	Writing – Original Draft [§3.3]
Hiranya Peiris	Writing – Original Draft [§3.1, §3.3, §3.7, §4.1, §4.2]; Writing – Review & Editing [all]
Laurence Perreault-Levasseur	Writing – Original Draft [§3.2]; Writing – Review & Editing [§4]
Andrés A. Plazas Malagón	Writing – Original Draft [§3.10]
Nesar Ramachandra	Writing – Original Draft [§3.6]
Benjamin Remy	Writing – Original Draft [§3.3, §4.2]
Cécile Roucelle	Writing – Original Draft [§3.9, §5.1]

Name	Contribution
Jaime Ruiz-Zapatero	Writing – Original Draft [§4.1]
Stefan Schuldt	Writing – Original Draft [**§3.2**]; Writing – Review & Editing [all]
Ignacio Sevilla-Noarbe	Writing – Original Draft [§3, §7, §8]
Ved G. Shah	Writing – Original Draft [§3]; Writing – Review & Editing [all]
Tjitske Starkenburg	Writing – Original Draft [§3, §4]; Writing – Review & Editing [§3, §4]
Stephen Thorp	Writing – Original Draft [§3.1, §3.2, §3.3, §3.7, §4.1, §4.2, §7]; Writing – Review & Editing [all]
Laura Toribio San Cipriano	Writing – Original Draft [§3]
Tilman Tröster	Writing – Original Draft [§3.3, §3.6]
Roberto Trotta	Writing – Original Draft [§3, **§4.1**, §8]; Writing – Review & Editing [all]
Padma Venkatraman	Writing – Original Draft [§3.2]; Writing – Review & Editing [§3.2]
Amanda Wasserman	Writing – Original Draft [§3.5]
Tim White	Writing – Original Draft [§3.9]
Justine Zeghal	Writing – Original Draft [**§3.3**, §4.1]
Tianqing Zhang	Project administration [§3.8, §3.9]; Writing – Original Draft [§3.1]; Writing – Review & Editing [§3.8, §3.9]
Yuanyuan Zhang	Writing – Original Draft [**§3.4**]

Contents

1. Executive Summary

The Legacy Survey of Space and Time Dark Energy Science Collaboration (LSST DESC) is an international collaboration whose mission is to measure the cosmic expansion history and the growth of structure using data from the Vera C. Rubin Observatory, thereby constraining the nature of dark energy and dark matter. Achieving these science goals requires jointly analyzing multiple cosmological probes—weak and strong gravitational lensing, galaxy clusters, Type Ia supernovae, and large-scale structure—each presenting distinct analysis challenges at LSST's unprecedented data volumes. Extracting robust cosmological constraints demands methods that deliver trustworthy uncertainty quantification, remain robust to systematic effects and model misspecification, and scale to the full petabyte-scale survey. These requirements motivate the integration of artificial intelligence (AI) and machine learning (ML) into DESC pipelines. DESC's combination of community-accessible data, mature simulation infrastructure, and rigorous scientific standards makes the collaboration an excellent testbed for developing robust AI/ML practices for fundamental physics.

Recognizing this situation, the DESC formed the *AI for DESC Task Force* with the following charge:

- Catalog the AI/ML needs, use cases, and projects in DESC.
- Identify current gaps in the adoption of AI/ML methodologies by leveraging expert domain knowledge.
- Identify the computational resources, storage, data access, and human research and managerial time needed to take full advantage of AI/ML-related opportunities.
- Identify either qualitatively or quantitatively the projected gains in DESC's science that would result from pursuing AI/ML-related opportunities.

The response to the task force charge is presented in this white paper. It demonstrates the breadth and importance of AI and ML research within DESC, and highlights the challenges and promising pathways for future work.

In this Executive Summary, we synthesize key recommendations and opportunities into a coherent AI/ML strategy for DESC. Three core principles guide this strategy:

- **AI/ML tools should be carefully integrated into DESC pipelines** to facilitate scientific analyses while fulfilling the stringent requirements of precision cosmology and preserving scientific accountability and transparency.
- A **durable AI/ML ecosystem should be built** within DESC and maintained over the survey lifetime, for the collaborative development, validation, and deployment of production-grade AI/ML tools.
- AI/ML must be integrated into DESC in ways that **preserve and support the human-centric nature of research**, improve accessibility, strengthen collaboration quality, and amplify rather than supplant members' contributions.

We have defined a series of recommendations (R) and opportunities (O) in several key areas within DESC in support of these principles. *Recommendations* are actions that the collaboration should undertake to meet

its scientific requirements and ensure robust integration of AI and ML into DESC pipelines. *Opportunities* indicate areas where DESC can extend beyond its requirements and assume a leadership role, influence broader community standards, or explore higher-risk, higher-reward efforts. We summarize these below, along with references throughout the paper where they are discussed.

Advancing Key Methodological Research Directions—Challenges such as uncertainty quantification, robustness to model misspecification, and novelty detection recur across DESC science cases. Progress on these foundational challenges will benefit all probes and merit dedicated effort.

- **R1: Prioritize Fundamental Methodological Research.** Foster collaboration-wide research in several critical areas: quantification of systematic and statistical uncertainties, simulation-based inference robustness, physics-informed modeling (hybrid generative-physical architectures), validation of neural posteriors, and novelty detection. Progress on these fundamental challenges will have an outsized impact across many DESC science cases. (Section 3, Section 4)
- **O1: Methodological Leadership in Trustworthy AI.** The challenges DESC faces (robust inference under misspecification, calibrated uncertainty quantification at scale, physics-informed learning) are frontier problems in machine learning broadly, creating natural opportunities to attract specialist collaborators and position DESC as a leader in trustworthy AI for fundamental science. (Section 4, Section 7)
- **O2: DESC Simulation Assets as Community Benchmarks.** DESC's combination of petabyte-scale community data, stringent scientific requirements, and rich simulation assets—e.g. the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC), Extended LSST Astronomical Time Series Classification Challenge (ELAsTiCC), and Cosmological Data Challenge 2 (CosmoDC2)—makes it an ideal testbed for pioneering robust AI/ML practices. Benchmarks and governance standards developed here can become reference implementations for fundamental physics, and attract colleagues in mathematics and computer science who see DESC's frontier challenges as compelling application areas for new methods. (Section 3, Section 4, Section 7)

Foundation Models—Foundation models, which produce generalizable representations of large-scale, heterogeneous, and multi-modal datasets, are transforming AI capabilities. DESC must develop both the infrastructure to deploy them and the benchmarks to validate them for precision cosmology.

- **R2: Develop Shared Foundation Model Infrastructure.** Build a shared foundation model backbone for DESC, consistent across data modalities and of production-grade quality, and served behind stable APIs. (Section 5, Section 6)
- **R3: Establish DESC-specific Foundation Model Validation Standards.** Create benchmarks that go beyond industry practice: uncertainty calibration, robustness to systematics, sensitivity to training biases, stress tests under distribution shift (temporal, spatial, cross-survey). Develop astronomy-specific interpretability tools to verify the physically meaningful structure preserved within model representations. (Section 5, Section 8)
- **O3: Leadership of Rubin-wide Development of Foundation Models.** DESC could play a central role in coordinating foundation model development across Rubin Science Collaborations and LSST

Interdisciplinary Network for Collaboration and Computing Frameworks (LINCC Frameworks), leveraging distributed computing resources from universities to Department of Energy (DOE) or European High-Performance Computing Joint Undertaking (EuroHPC) facilities. (Section 5, Section 7)

Large Language Models & Agentic AI—Large language models (LLMs) and agentic AI offer avenues to accelerate research and lower the barrier to entry for complex cosmological analyses in DESC. Harnessing this potential responsibly will require thoughtful governance and rigorous validation frameworks.

- **R4: Establish Governance for LLMs and Agentic Systems.** Coordinate DESC-wide activities involving LLMs and agents, establish best practices including evaluation, review, and tiger-team review of pilot studies. Include critical discussions of the technology's limitations and effects on human researchers, with input from experts across domains. Engage with Rubin Data Management to ensure that agentic AI can interface with data products effectively and reliably. (Section 5, Section 8)
- **R5: Build Natural Language Interfaces to DESC Resources.** Develop retrieval-augmented generation (RAG)-based interfaces to DESC documentation, simulations, and data products, lowering onboarding barriers and democratizing access to complex pipelines. (Section 5, Section 6)
- **O4: Pioneering Agentic AI for Scientific Rigor and Reproducibility.** An important application of this work could be "DESC research agents" that automate execution, documentation, and validation of analyses against standardized benchmarks, coupling these systems to clear governance and tiger-team review procedures so that agentic workflows enhance transparency, provenance, and trust in DESC results. (Section 5, Section 8)

Infrastructure & Software—DESC has a mature ecosystem of cosmological analysis pipelines. Building on this foundation, strategic development of AI software stacks, differentiable programming, and computing infrastructure can act as multipliers that benefit all science cases.

- **R6: Establish a Durable AI Software Stack.** Adopt a coherent set of frameworks, tooling, and model export standards. The stack should be portable across DESC computational facilities, sustainable over the 10-year survey, and prioritize open governance to avoid proprietary lock-in. (Section 6)
- **R7: Develop a Differentiable Programming Ecosystem.** Adoption of an interoperable differentiable programming ecosystem (e.g. based on JAX) will act as a multiplier, simultaneously enabling gradient-based sampling, GPU acceleration, hybrid physics-ML models, and end-to-end optimization across DESC pipelines. (Section 3, Section 4, Section 6)
- **R8: Secure Access to Emerging Computing Infrastructure.** Significant new AI-oriented computing is becoming available: DOE infrastructure such as the American Science Cloud (AmSC); the Independent Data Access Center (IDAC) network; and EuroHPC systems such as Leonardo in Italy, Large Unified Modern Infrastructure (LUMI) in Finland, and the Joint Undertaking Pioneer for Innovative and Transformative Exascale Research (JUPITER) exascale system in Germany. DESC should engage early to shape these resources for cosmology and secure allocations for foundation model training at scales infeasible on current systems. (Section 6, Section 7)

Organizational Structure & Governance—The DESC is organized into computing, technical, and analysis working groups (WGs)[2], with analysis working groups primarily aligned with key cosmological probes. Effective AI/ML integration across these groups requires consistent coordination mechanisms and clear standards for development, validation, and deployment.

- **R9: Develop DESC-wide AI/ML Coordination Mechanisms.** Establish structures (e.g., standing working group, cross-WG task forces, regular interchange meetings) to share methodological innovations across probes, tackle common challenges collectively, and minimize duplication. Facilitate rapid dissemination through workshops, tutorials, and methodological discussions. (Section 3)
- **R10: Develop AI/ML Best Practice Guidelines.** Create guidelines to help DESC members develop robust AI/ML analyses, covering topics such as reproducibility, provenance tracking, validation checks, and comprehensive benchmarking—particularly for foundation models and other shared deliverables whose broad applicability demands thorough vetting before widespread adoption. (Section 3, Section 4, Section 5)

Human Capital & Sustainability—The promise of AI/ML for accelerating cosmology with LSST will not be realized without training and support of DESC members. Sustainable adoption of AI/ML also requires attention to the growing computational demands and resulting footprint these methods entail.

- **R11: Focus on AI/ML for Augmenting Rather Than Replacing Understanding.** DESC must strengthen and maintain the technical literacy of the collaboration in AI/ML applications as tools for science rather than supplanting understanding. (Section 8)
- **R12: Track and Optimize Resource Footprint.** DESC should develop tools for monitoring and optimizing computational resource usage of AI/ML models, enabling the collaboration to maximize scientific productivity and make informed decisions about resource allocation and environmental impact. (Section 8)

External Coordination & Partnerships—DESC operates within a rich ecosystem of other Rubin science collaborations, AI institutes, cosmology experiments, and alert brokers that filter streaming Rubin data. Deliberate coordination between these groups will amplify impact and avoid duplicated effort.

- **R13: Coordinate Across Science Collaborations.** Partner with other LSST collaborations and other cosmology experiments. The former include the Informatics and Statistics Science Collaboration (ISSC), Transients and Variable Stars Science Collaboration (TVSSC), Strong Lensing Science Collaboration (SLSC), Active Galactic Nuclei Science Collaboration (AGNSC), and Galaxies Science Collaboration (GSC). The latter include the Dark Energy Spectroscopic Instrument (DESI), the 4-meter Multi-Object Spectroscopic Telescope (4MOST), the European Space Agency (ESA) *Euclid* Mission science teams, and the *Nancy Grace Roman Space Telescope* science collaborations. Areas of coordination should include methodological development, time-series and broker stress-testing, deblending/morphology benchmarks, and sharing tools and best practices. (Section 7)

[2] https://lsstdesc.org/pages/organization.html

- **R14: Engage with AI Institutes and Networks.** National Science Foundation (NSF)–Simons AI Institutes (with explicit LSST/cosmology themes), and European networks such as the European Coalition for AI for Fundamental Physics (EuCAIF)[3] and European Laboratory for Learning and Intelligent Systems (ELLIS)[4], are natural partners. Build systematic engagement through co-funded postdocs, shared workshops, joint proposals, and benchmark datasets. These efforts would connect DESC to the broader AI-for-science ecosystem. (Section 7)
- **R15: Develop the Human-Machine Interface.** Develop close connections between DESC, other LSST science collaborations, in-kind follow-up programs, alert broker teams, LSST data management, and citizen scientists, to facilitate active learning for classification, anomaly detection, and human-in-the-loop interpretability. (Section 4, Section 7)
- **O5: DESC Integration with the Broker Ecosystem.** DESC members are embedded in all seven Rubin Community Broker teams—tight coordination gives direct leverage over SN Ia sample purity, selection effects, and host-galaxy priors, plus an on-ramp from research prototypes to community-facing services. (Section 3, Section 7)

Implementing these recommendations and capitalizing on these opportunities would position DESC to fully exploit LSST's statistical power for cosmology while uncovering unexpected phenomena in the largest optical astronomical dataset ever collected. This will require sustained investment in researchers who bridge domain science and AI/ML methodology. Such investment would benefit not only DESC, but the broader effort to advance AI as a tool for fundamental scientific discovery.

[3] https://eucaif.org/
[4] https://ellis.eu

2. Introduction

The Vera C. Rubin Observatory's LSST will produce unprecedented volumes of heterogeneous data (images, catalogs, alerts) whose full scientific exploitation demands continued methodological innovation. The mission of the DESC is to convert these data into robust constraints on the dark sector by jointly measuring the cosmic expansion history and the growth of structure, thereby shedding much-needed light on dark energy, dark matter, and possible deviations from general relativity. Delivering on these objectives requires methods that are statistically powerful, scalable, and operationally reliable. Recent advances in AI and ML show great promise for critical data analysis roles but still need to meet stringent requirements to be truly useful. artificial intelligence (AI) refers broadly to systems capable of performing tasks that typically require human intelligence, including reasoning, perception, learning, and decision-making. machine learning (ML) is a subfield of AI in which algorithms learn patterns and relationships from data to make predictions or decisions, encompassing both classical methods (e.g., random forests, Gaussian processes) and deep learning (multi-layer neural networks). In the DESC context, ML methods learn mappings between variables (e.g., photometry to redshift, galaxy fields to cosmological parameters) and can be deployed at multiple stages of analysis. Their scientific utility, however, hinges on trustworthy uncertainty quantification and reproducible integration within DESC workflows; without these elements, ML methods cannot meet the stringent requirements of cosmological inference. In parallel, we use AI to denote systems capable of complex cognitive tasks—such as reasoning, knowledge synthesis, and natural language understanding—that can potentially orchestrate tools, generate code, and reshape scientific workflows. As with ML, turning recent AI advances into reliable accelerators of discovery remains an open problem that requires careful evaluation and governance.

The strategic question for DESC is therefore how to develop and integrate AI/ML *the right way*, so that these approaches become dependable components of LSST-era analyses. Intrinsically, this question is relevant to a broad range of scientific endeavors and collaborations. Still, DESC is well positioned to pioneer robust AI/ML practices for fundamental physics as an international collaboration that works with community-accessible data, has a strong open-source culture, and pursues extremely demanding scientific objectives. In this white paper, we set out a strategic framework for how DESC should organize its AI/ML efforts, prioritize methodological investments, and respond effectively to new opportunities arising from rapid AI/ML progress. This paper is structured around four interconnected perspectives on AI/ML within DESC, each building on the previous to articulate a comprehensive strategy:

Section 3: The Current Landscape: ML Across DESC Science—Machine learning is not a future aspiration for DESC: it is already deeply embedded in current science workflows. Here, we survey how ML methodologies intersect with DESC's primary cosmological probes: strong and weak gravitational lensing, galaxy clusters, Type Ia supernova (SN Ia) cosmology, large-scale structure, as well as cross-cutting analysis components including simulations, theory and modeling, deblending, photometric redshift (photo-z) estimation, and shape measurement. This inventory reveals a striking pattern: the same core methodologies (e.g. simulation based inference, differentiable programming, deep learning) appear repeatedly across disparate science cases, while the same fundamental challenges (e.g. uncertainty quantification, robustness to covariate shift and to model misspecification) represent concrete challenges across multiple working groups (see Figure 1).

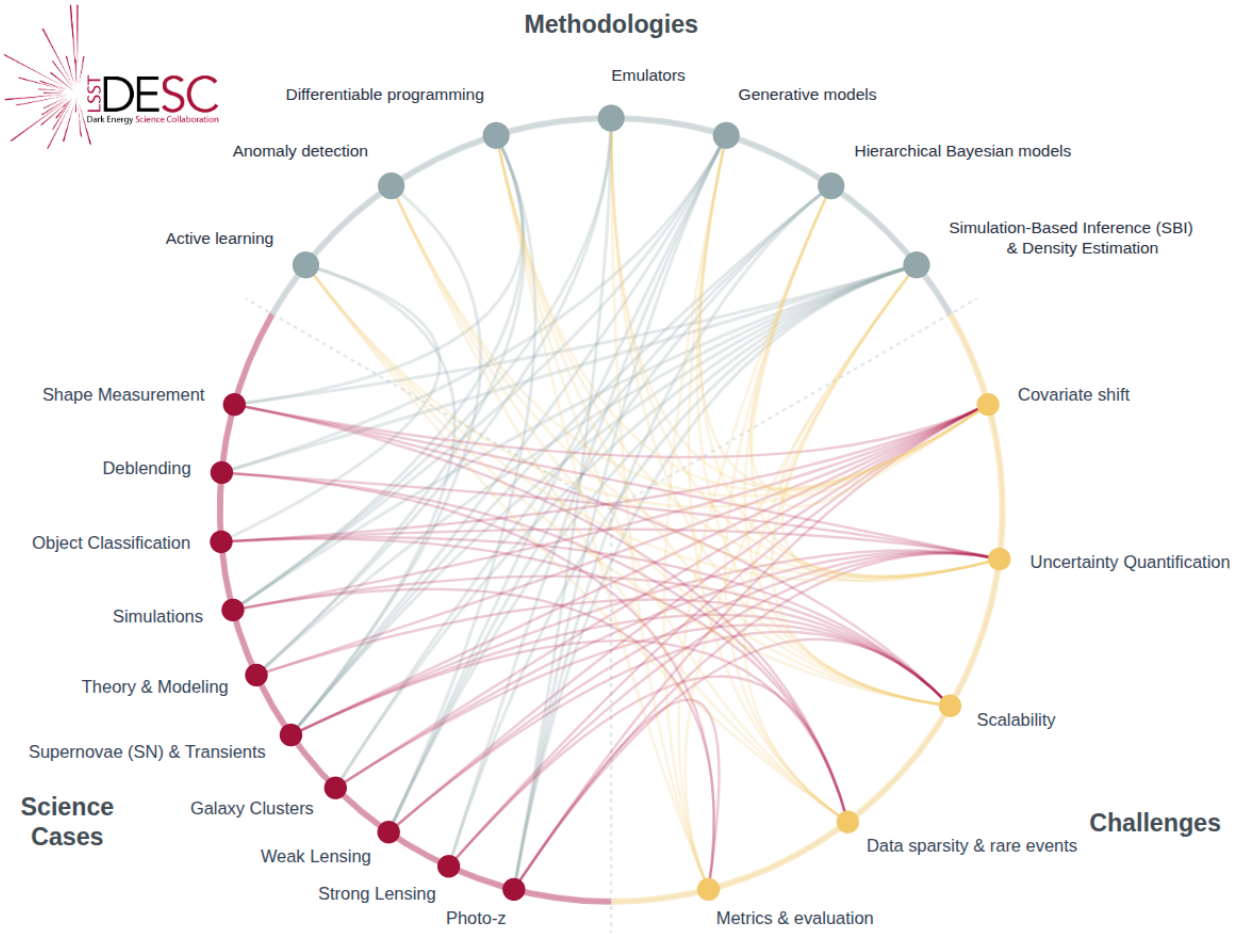

Figure 1. Transversal connections between DESC science applications (left), AI/ML methodologies (top), and shared challenges (right), as surfaced by Section 3. The recurring appearance of the same methods and challenges across disparate science cases motivates collaboration-wide coordination of AI/ML efforts rather than siloed development within individual working groups. An interactive version of this diagram is available at `https://lsstdesc.org/AI_For_DESC/figures/chord-diagram.html`

Section 4: Lifting the Limits of ML: Methodological Research Priorities—Building on the challenges surfaced in Section 3, we identify the key methodological research axes where targeted investment can lift current limitations and enable ML methods to meet the precision and reliability standards demanded by LSST-era cosmology. We organize these priorities around several interconnected themes:

- **Bayesian inference and uncertainty quantification (UQ):** Developing fast and scalable inference techniques that may unlock promising high-dimensional hierarchical models. Improving methods for reliably estimating uncertainty arising from limited training data and model limitations.

- **Simulation-based inference (SBI) and model misspecification:** Advancing neural density estimation (NDE) techniques, optimal summarization methods, and diagnostics for detecting and mitigating

the biases introduced when training simulations imperfectly represent real observations or when training datasets fail to capture the full diversity of LSST data.

- **Physics-informed approaches:** Hybridizing explicit physical models with flexible generative components (flows, diffusion models) and advancing differentiable programming frameworks that embed cosmological theory directly into ML architectures, ensuring that learned components remain interpretable, physically consistent, and robust to extrapolation.
- **Novelty detection and discovery:** Developing representation learning and active human–AI collaboration frameworks capable of identifying rare, previously unmodeled phenomena in LSST's vast data volumes.

These research directions are not purely academic exercises; they directly address the technical barriers that limit the deployment of ML at scale for DESC's most ambitious analyses. We articulate not only what needs to be developed, but why these specific advances matter for cosmological inference and how DESC can contribute to the broader AI research ecosystem by presenting demanding, scientifically motivated benchmarks.

Section 5: Looking Forward: Foundation models and Agentic AI—While Sections 3 and 4 focus on current ML applications and their refinement, Section 5 adopts a forward-looking perspective, examining how two emerging AI paradigms (*data foundation models* and *LLM-based agentic systems*) have the potential to reshape large sections of DESC workflows in ways that go qualitatively beyond incremental improvements to existing methods.

- **Foundation models (FMs)**, trained at scale on heterogeneous data modalities (images, spectra, time series, catalogs), offer the promise of *reusable representations* that can be rapidly fine-tuned or directly deployed across a wide range of downstream tasks (classification, regression, anomaly detection, simulation-based inference) without retraining from scratch for each application. For DESC, this paradigm shift could enable unified, survey-scale feature extractors that serve as common backbones for weak lensing, photometric redshifts, transient classification, and more, dramatically reducing duplication of effort while ensuring cross-probe consistency. However, realizing this vision requires careful attention to uncertainty propagation, robustness to distribution shifts, architectural choices suited to astronomical data, and rigorous, community-governed benchmarking to ensure that foundation models meet DESC's validation standards.
- **LLM-driven agentic systems** are rapidly evolving from research prototypes into tools capable of orchestrating complex scientific workflows: querying databases, generating and executing code, synthesizing literature, and autonomously iterating on analyses. These systems offer tantalizing possibilities for accelerating exploratory research, onboarding new collaboration members, and scaling human oversight across large analysis campaigns. Yet they also introduce new risks: biased recommendations, irreproducible results, and erosion of scientific understanding if deployed without governance. Section 5 outlines both the transformative potential and the implementation requirements (provenance tracking, human-in-the-loop validation, benchmark design, and clear policies on data rights and model transparency) necessary to integrate agentic AI into DESC in ways that guarantee scientific rigor.

The forward-looking stance of Section 5 is deliberate: DESC must not only respond to today's ML methods but actively shape the trajectory of emerging AI technologies by setting clear scientific requirements, contributing demanding use cases to the broader AI research community, and pioneering governance frameworks that other collaborations can learn from.

Section 6, 7, 8: Operationalizing AI/ML: Infrastructure, Coordination, and Risk Management—Even the most sophisticated AI/ML methods will have limited impact if they cannot be reliably deployed, maintained, and integrated into DESC's production pipelines. The final sections of this white paper address the operational foundations required to translate research prototypes into dependable cosmological infrastructure. Section 6 details the *infrastructure requirements* across software, computing, and data:

- **Software:** Establishing a robust, collaboration-endorsed AI software stack (frameworks, experiment tracking, model registries, continuous integration/continuous deployment for models) that ensures reproducibility, portability across DESC computing facilities, and long-term sustainability over the LSST decade. This includes strategies for integrating AI components into DESC analysis pipelines and for managing the rapidly evolving ecosystem of LLMs and agentic frameworks.
- **Computing:** Securing the graphics processing unit (GPU) allocations, distributed training capabilities, and co-located data access necessary for foundation model development, large-scale simulation-based inference campaigns, and real-time alert stream processing. This involves strategic coordination with national labs such as the Argonne Leadership Computing Facility (ALCF) and Oak Ridge Leadership Computing Facility (OLCF), emerging initiatives (AmSC, HPDFs), and international partners (EuroHPC, IDACs).
- **Data:** Ensuring that LSST data products, multi-survey training datasets, and simulation outputs are accessible, well-documented, and equipped with the interfaces—e.g., application programming interfaces (APIs), streaming services, tokenization strategies—required for efficient AI/ML workflows. This includes establishing shared repositories, benchmark datasets, and provenance standards.

Section 7 broadens the scope to examine *opportunities for coordination beyond DESC*: with the other Rubin LSST Science Collaborations, Stage-IV experiments (in particular DESI, 4MOST, Roman, Euclid), AI institutes (e.g., NSF–Simons institutes, CosmicAI), European networks (e.g., EuCAIF, ELLIS), and the Rubin alert broker ecosystem. These partnerships offer opportunities for shared training data, cross-survey foundation models, joint benchmark development, and access to specialized compute resources and expertise. DESC is uniquely positioned to act as both a consumer and a driver of AI methodologies within this broader ecosystem, articulating the demanding requirements of precision cosmology while contributing validated methods and datasets that benefit the wider community.

Finally, Section 8 confronts the *risks and challenges* inherent in DESC's increasing reliance on AI/ML: model miscalibration, opaque failure modes, reproducibility challenges, data governance complexities, and the potential erosion of human scientific understanding. We outline concrete mitigation strategies (validation protocols, redundancy in critical analyses, provenance tracking, training programs, and governance structures) that apply the same rigor to AI components as to any other element of the cosmological inference pipeline.

Section 9: Conclusion—Viewed as a whole, this paper shows that AI/ML is already central to DESC science (Section 3), but unlocking its full potential requires targeted methodological research (Section 4), proactive engagement with emerging technologies (Section 5), and robust operational foundations (Section 6–8). The transversality of methods and challenges across DESC working groups demands deliberate coordination to prevent fragmented effort and ensure that best practices, validated tools, and lessons learned propagate rapidly throughout the collaboration. By articulating this vision (grounded in current capabilities, guided by research priorities, forward-looking in its engagement with foundation models and agentic AI, and operationally realistic about infrastructure and risk) DESC can position itself not only to meet its own science goals but to pioneer robust AI/ML practices for fundamental physics that serve as a model for the broader community.

3. The Current Landscape of ML Across DESC Science

The science goals of DESC place unusually stringent demands on statistical methodology. Extracting percent-level constraints on dark energy, dark matter, and tests of gravity from Rubin LSST data requires not only exquisite control of observational systematics, but also analysis pipelines that can efficiently exploit information distributed across billions of galaxies, multiple probes, and heterogeneous data modalities (images, catalogs, time series, simulations). While AI (in the sense of LLMs and agents) has not yet significantly started to impact DESC, ML is already embedded in many of these workflows and its importance will only grow as analyses become more ambitious and data volumes increase.

In this section, we survey existing intersections of ML methods with DESC science, organized by application area, including photometric redshifts, strong and weak lensing, and galaxy clusters; supernovae and transients; cosmological theory and simulations; deblending; and shape measurement. Each subsection highlights both current capabilities and open challenges, to clarify where targeted methodological investment will yield the most significant scientific returns for DESC.

A note on reading this section: In the subsections that follow, each DESC science application is accompanied by a summary box highlighting the ML methodologies it employs and the challenges it faces. As you read through these applications, pay attention to how often the same methods (SBI, differentiable programming, NDE) and the same challenges (covariate shifts, uncertainty quantification, scalability, data sparsity, metrics) recur across seemingly disparate topics. This reveals a fundamental pattern: a small set of transversal methodologies and challenges cuts across the entirety of DESC's ML applications, as illustrated in Figure 1. This pervasive transversality has direct implications for how DESC should organize its AI/ML efforts, which we synthesize at the end of this section and develop further in Section 4.

3.1. Photometric redshifts

- **Methodology.** Gaussian processes, NDE, SOMs, Transformers, Hierarchical Bayes, Emulators, Neural surrogates, Diffusion models
- **Challenges.** Covariate shifts, UQ, Scalability, Metrics
- **Opportunities.** Multi-survey training, simulation infrastructure, hierarchical inference

The inference of photo-zs represents a foundational challenge for LSST, where the vast majority of tens of billions of detected galaxies will lack spectroscopic redshift measurements due to both observational time constraints and the intrinsic faintness of the sample. Photometric redshifts are derived by establishing empirical or physically motivated mappings between broadband photometry (including colors, magnitudes), morphology, and redshift. This process is fundamentally limited by our incomplete knowledge of galaxy spectral energy distributions (SEDs), the spatial and temporal variations of stellar populations in galaxies, and the nature and distribution of attenuating dust. However, the accuracy and reliability of photo-z estimation is critical across virtually all extragalactic LSST science cases, including weak gravitational lensing, large-scale structure, galaxy cluster cosmology, and supernova surveys. To achieve the DESC goals for constraining the dark energy equation of state, calibration of photo-z estimates must reach the $0.002\times(1+z)$

level for the first year of LSST data (The LSST Dark Energy Science Collaboration et al. 2018). Achieving these benchmarks necessitates not only accurate point predictions but also *well-calibrated uncertainty quantification*, motivating an emphasis on probabilistic methods. In addition, model benchmarking requires a unified framework for per-galaxy photo-z algorithms, ensemble calibration algorithms, mock data generation, and performance evaluation. These tasks are fulfilled by the Redshift Assessment Infrastructure Layers (RAIL; RAIL Team et al. 2025), a photo-z library developed by LINCC Frameworks and DESC.

Supervised Photo-z Estimation—The empirical approach to photo-z inference is to learn a mapping from observed broadband photometry or imaging to redshift, leveraging spectroscopic training samples. From catalog-level photometry, empirical regressors such as random forests in Trees for Photo-z (TPZ; Carrasco Kind & Brunner 2013), gradient boosting machines (`FlexZBoost`; Izbicki & Lee 2017; Dalmasso et al. 2020), and Gaussian processes (GPz; Almosallam et al. 2016) have demonstrated competitive performance by constructing mappings from color-magnitude space to redshift. Neural networks, including both fully connected and specialized architectures, have proven particularly adept at capturing complex relationships in high-dimensional photometric data (Collister & Lahav 2004). On the other hand, nearest-neighbor approaches such as k-nearest neighbors and color-matched nearest-neighbors (kNN and CMNN; Graham et al. 2018) take the training as a reference sample and compute the average redshift of neighbors of the target galaxy. In a similar way, directional neighborhood fitting (DNF; De Vicente et al. 2016) performs a regression on the neighborhood, which allows the construction of a local linear model for each galaxy.
Contemporary approaches have evolved from point estimators to full probabilistic models capable of capturing the full conditional distribution $p(z \mid \text{photometry})$, with NDE techniques (e.g., `PZFlow`; Crenshaw et al. 2024) enabling flexible, well-calibrated redshift probability density functions (PDFs) via maximum likelihood training. Complementing catalog-based methods, image-based inference circumvents the information bottleneck imposed by aperture photometry by operating directly on multi-band pixel data, delegating feature extraction to deep neural networks that leverage morphology and spatial structure inaccessible to catalogs. The Detection, Instance Segmentation, and Classification with Deep Learning (DeepDISC) framework (Merz et al. 2023, 2025) exemplifies this approach, integrating object detection, segmentation, and redshift estimation into a unified pipeline using multiscale vision transformers (MViT; Li et al. 2022) as the backbone feature extractor coupled with mixture density networks for probabilistic PDF estimation. Beyond redshift estimation, tomographic bin assignment for 3×2pt analyses is itself amenable to supervised learning; the DESC Tomography Optimization Challenge (Zuntz et al. 2021) benchmarked multiple algorithms for this task, and subsequent work has shown that neural network classifiers can identify galaxies likely to be correctly binned, improving cosmological constraints (Moskowitz et al. 2023).
Despite their sophistication, *all supervised ML methods remain fundamentally limited by the quality and representativeness of their spectroscopic training samples*: spectroscopic incompleteness, magnitude-limited surveys, and selection biases induce systematic offsets and distortions in the learned photo-z mapping, particularly at faint magnitudes and high redshifts where spectroscopic follow-up is most incomplete (Newman & Gruen 2022). This represents the main challenge for photo-z today and has motivated dedicated calibration strategies. A further problem is the "implicit prior" imposed by each photo-z method (Schmidt et al. 2020). These priors, which have a large impact on photo-z estimates, are opaque and difficult to quantify, making it difficult to compare and combine photo-z posteriors provided by different methods.

Calibration Strategies to Account for Covariate Shifts—Self-organizing maps (SOMs) have emerged as the preeminent unsupervised learning technique for diagnosing and mitigating the biases caused by covari-

ate shifts by performing non-linear dimensionality reduction of photometric feature vectors onto a discrete two-dimensional grid (Masters et al. 2015). SOM-based calibration approaches, such as those deployed in Dark Energy Survey (DES) Year 3 (Myles et al. 2021) and Kilo Degree Survey (KiDS) analyses (Wright et al. 2020a,b, 2025; van den Busch et al. 2022), directly assign photometric galaxies the empirical redshift distribution of spectroscopic galaxies in their SOM cell while down-weighting or even rejecting regions of color space poorly represented in the spectroscopic catalog. More sophisticated SOM-guided data augmentation strategies selectively populate under-represented SOM cells with simulated galaxies from mock catalogs improving ML model performance where spectroscopic coverage is deficient (Moskowitz et al. 2024; Zhang et al. 2025). An alternative approach to SOM for covariate shift mitigation (and without using data augmentation) is represented by stratification by the propensity score (defined as the probability of a covariate vector to be admitted as part of the training set) of both training and target data. Within each propensity score group, supervised photo-z can proceed with any method of choice. This stratified learning (StratLearn) approach is theoretically guaranteed (under some conditions) to cancel covariate shift (Autenrieth et al. 2023). It has demonstrated state-of-the-art performance in the PLAsTiCC (PLAsTiCC team et al. 2018) SN Ia classification challenge, a factor of ~ 2 improvement in photo-z calibration from the cosmic shear KiDS+VIKING-450 dataset (Autenrieth et al. 2024) and a reduced fraction of catastrophic errors and one order of magnitude improvement of the bias for simulated photo-z reconstruction (Moretti et al. 2025).

Hybrid Template-Based Estimators—In contrast to empirical photo-z estimators, a broad class of "template-based" photo-z estimators (e.g., Brammer et al. 2008; Arnouts & Ilbert 2011) attempt to circumvent the problem of covariate shift using physical models of galaxy SEDs. These estimators trade the problem of covariate shift for the problem of model misspecification. Hybrid methods, however, attempt to combine the strengths of empirical and template-based estimators by deriving SED templates in a physics-informed, data-driven manner (Budavári et al. 2000; Csabai et al. 2000). These models have been shown to deliver higher-quality photo-z estimates than traditional template-based estimators while suffering less from covariate shift than pure empirical methods (Crenshaw & Connolly 2020; Li et al. 2025d). They do not perform as well in-distribution as pure empirical methods, however, and still rely on spectroscopic calibration sets. It may be possible to remedy these defects by implementing hybrid, physics-informed models in deep learning frameworks to enable self-supervised learning without reliance on spectroscopic data sets (Boone 2021).

Population-Level Hierarchical Forward Modeling—Traditional photo-z workflows estimate individual galaxy redshifts and aggregate these posteriors to derive population-level quantities such as ensemble redshift distributions $n(z)$ – a computationally expensive bottom-up approach prone to biases when combining noisy individual posteriors (Leistedt et al. 2016; Malz 2021; Malz & Hogg 2022; Alsing et al. 2023). Population-level inference inverts this paradigm by directly targeting the population distribution $P(\boldsymbol{\theta})$ over redshift and physical galaxy parameters (stellar mass, star formation rate, metallicity) as the primary inference objective, leveraging the collective constraining power of the entire photometric dataset while naturally incorporating physical priors on galaxy evolution. These methods rely on forward modeling: generating synthetic photometry from physical parameters via stellar population synthesis (SPS; for reviews, see, e.g., Conroy 2013; Iyer et al. 2026) models and comparing the distribution of model photometry to observed data. Classical SPS calculations—as implemented by, e.g., Flexible Stellar Population Synthesis (FSPS; Conroy et al. 2009, 2010; Conroy & Gunn 2010) and the `prospector` model family (Leja et al. 2017; Johnson et al. 2021; Wang et al. 2023)—are computationally prohibitive for large samples, motivating neural network emula-

tors like `speculator` (Alsing et al. 2020) that achieve $\sim 10^3$–$10^4\times$ speedups with negligible accuracy loss. The `pop-cosmos` framework (Alsing et al. 2024; Thorp et al. 2024a, 2025b; Deger et al. 2025) exemplifies this approach: it defines a probability distribution over a 16-dimensional SPS parameterization using a score-based diffusion model calibrated on ∼420,000 galaxies from COSMOS2020 (Weaver et al. 2022) with 26-band photometry spanning deep ultraviolet (UV) to mid-infrared (IR). This model enables direct estimation of tomographic redshift distributions, and, when used as a data-driven prior in SED fitting, highly accurate individual galaxy redshift inference.

Benchmarking and Evaluation Frameworks—As AI methods become increasingly central to cosmological analyses, it is critical to develop robust frameworks for testing and validation that ensure reproducibility and enable systematic comparison of different approaches. For this purpose, the DESC has developed RAIL, an open-source, Python-based framework to support large-scale photometric-redshift workflows for LSST. RAIL is a library that hosts many per-galaxy algorithms (e.g., `FlexZBoost`, `DNF`, `PZFlow`, `DeepDISC`), ensemble calibration algorithms, (e.g., self-organizing maps). RAIL also provides comprehensive infrastructure that (i) supplies a unified API and modular pipeline stages to train, apply, and compare a broad range of redshift estimators (catalog-based, image-based, probabilistic), (ii) embeds evaluation modules and metrics for both individual-galaxy redshift and ensemble PDFs (RAIL Team et al. 2025), (iii) generates realistic mock data for supervised learning algorithms by applying photometric noise, reference redshift selection, and error to an input catalog, and (iv) enables data challenges to test the robustness of photo-z estimators to a wide array of systematic errors. RAIL's standardized framework facilitates reproducible results and fair benchmarking across different methods, essential for validating AI techniques in preparation for LSST data.

3.2. Strong Lensing

- **Methodology.** Gaussian processes, CNNs, RNNs, Transformers, SBI, Diffusion models, VI
- **Challenges.** Data sparsity, Covariate shifts, UQ
- **Opportunities.** Multi-survey cross-matching (Roman+LSST+Euclid), population-level inference, automated discovery, subhalo constraints from anomalous flux ratios

Strong gravitational lensing is a rare astrophysical phenomenon where the light of a distant object, the source, is deflected by the gravity of an intervening structure, the lens, forming multiple images of the background source. In galaxy-galaxy strong lensing, both the source and the lens are individual galaxies, while on larger scales, the lens could range from a group to an entire galaxy cluster. Despite their rarity (a result of the stringent alignment required between source, lens, and observer), lensed systems are powerful cosmological probes that can constrain dark energy and probe dark matter on sub-galactic scales. Since elliptical galaxies dominate the deflector population, strong lenses also enable studies of their mass profiles, stellar content, and dark matter halos. Furthermore, lensing magnification enables the study of high-redshift sources, offering insights into early galaxy evolution (Schneider et al. 1992).

LSST will be transformative for strong lensing science. Forecasts predict the discovery of $\sim$120,000 systems (Collett 2015), two orders of magnitude more than currently known. This large sample will provide the statistical power required for precision cosmology: time-delay lenses and large samples of static lenses

have been shown to enable competitive dark energy measurements (Shajib et al. 2025). The sample will also include significant numbers of currently rare systems, such as double-source-plane lenses, lensed supernovae, and cluster-scale lenses with multiple background sources. LSST's six-filter imaging (see Section 3.1) will enable photo-z estimation for both lens and source populations, as well as classification of time-delay lenses.

One particularly powerful application is time-delay cosmography, which uses strongly lensed transients to measure the Hubble constant (H_0) (e.g., TDCOSMO Collaboration et al. 2025). This approach yields a geometrical measurement independent of both the local distance ladder (e.g., SH0ES; Riess et al. 2022) and early-universe measurements from the cosmic microwave background (CMB; see, e.g. Planck Collaboration et al. 2020). LSST will provide time-domain coverage for $\sim 100\times$ more systems than current surveys (Wojtak et al. 2019; Goldstein et al. 2019; Arendse et al. 2024; Erickson et al. 2025a; Abe et al. 2025), dramatically increasing the sample of lensed supernovae (SNe) and active galactic nuclei (AGN) available for cosmography. AI/ML models have been proposed for time-delay estimation from light curves, particularly kernel-based methods (e.g. Cuevas-Tello et al. 2006; Cuevas-Tello et al. 2010; AL Otaibi et al. 2016).

Supervised Detection in the Low Data Regime—Given the rarity of strong lensing events, identifying them among billions of cutouts is inherently challenging for visual inspection in wide-field surveys. While early automated detection approaches relied on curvature-based features (Estrada et al. 2007) and arc-characterizing descriptors (Bom et al. 2012, 2017), the advent of convolutional neural networks (CNNs) led to state-of-the-art performance in lens finding (Petrillo et al. 2017; Schaefer et al. 2018; Petrillo et al. 2019; Lanusse et al. 2018). Building on this progress, Metcalf et al. (2019) launched a lens-finding challenge using Bologna Lens Factory simulations based on the Millennium data (Lemson & Virgo Consortium 2006), showing that LSST-like ground-based multi-band images are well suited for this task. A subsequent Euclid-like challenge produced a winning algorithm combining multi-resolution CNNs (Bom et al. 2022), later validated on real data by Melo et al. (2025) using Legacy Survey and *Hubble Space Telescope* (*HST*) images to mimic LSST–*Euclid* synergy. Leveraging multiple networks classifications as ensemble lens classifiers showed improved results over a single network classifier (e.g., Andika et al. 2023; Schuldt et al. 2023b; Gonzalez et al. 2025a), while Holloway et al. (2024) incorporated citizen science annotations to enhance the performance. Because too few real lenses exist for supervised training, realistic mock datasets are essential. Works such as Petrillo et al. (2017) or Schuldt et al. (2021) proposed simulating only the lensing effect on real galaxy images, a practice now standard in the ongoing LSST DESC and SLSC challenge (Bom et al., in prep.). In preparation for LSST, Hyper Suprime-Cam (HSC) data (Aihara et al. 2018; with similar filters and pixel scale) have been used to develop and compare models (see e.g., Shu et al. 2022; Andika et al. 2023; Cañameras et al. 2024; Jaelani et al. 2024; More et al. 2024). While early efforts focused on simple CNN or residual network (ResNet) architectures, more advanced architectures such as vision transformers have also been applied (Gonzalez et al. 2025b). Foundation models such as Zoobot (Walmsley et al. 2023) have also achieved strong results on *Euclid* imaging (Euclid Collaboration: Walmsley et al. 2025; Euclid Collaboration: Lines et al. 2025), and will soon be adopted for LSST (see Sect. 5.1). Finally, ML methods are now expanding beyond galaxy-scale lenses to systems involving entire clusters (Schuldt et al. 2025; Euclid Collaboration: Bazzanini et al. 2025) and galaxy–galaxy lenses within clusters (Angora et al. 2023).

Simulation-Based Inference (SBI)—Beyond lens finding, Hezaveh et al. (2017) pioneered the use machine learning models to predict characteristics of strong lensing systems. Specifically, Hezaveh et al. (2017) showed that simple CNNs can be used to predict parameters of the lens (Einstein radius, complex ellipticity, and the coordinates of the center of the lens) from images from *HST* with a precision comparable to that of traditional methods. Perreault Levasseur et al. (2017) proposed using approximate Bayesian neural networks (BNNs) to obtain calibrated estimations of the marginal posterior of these lens parameters, an approach applied to Atacama Large Millimeter Array (ALMA) observations in Morningstar et al. (2018). Such ML estimates can also be used as starting points for classical model fitting methods (e.g. Euclid Collaboration: Busillo et al. 2025), providing significant speedups. Legin et al. (2021, 2023) compared this approach to neural likeliood estimation (NLE), demonstrating potential for better calibration in 2-stage SBI methods. In Poh et al. (2025), it was demonstrated that BNNs and neural posterior estimation (NPE) can be used to infer parameters that describe the lens system even in ground-based DES-like imagining. In subsequent years, significant progress was made in using HSC images to prepare for LSST (e.g., Pearson et al. 2019; Schuldt et al. 2021; Gentile et al. 2023; Schuldt et al. 2023a,b; Gawade et al. 2025). Following earlier work by Park et al. (2021); Wagner-Carena et al. (2021), Erickson et al. (2025b) applied (sequential) NPE within a hierarchical framework to model strongly lensed quasars, testing on real systems discovered by DES and followed-up with *HST* high-resolution imaging, and Venkatraman et al. (2025) applied hierarchical NPE modeling to simulated LSST i-band coadds. Ongoing DESC work examines how modeling an uncertain sample of static galaxy-galaxy lenses with ML enables new cosmological constraints (Holloway et al. in prep.), leveraging the method demonstrated by Li et al. (2024). And in Jarugula et al. (2024), authors presented a scalable approach for inferring the dark energy equation-of-state parameter from a population of strong gravitational lens images using neural ratio estimation (NRE). Filipp et al. (2025) investigated the robustness of neural ratio and neural posterior estimators to distributional shifts for dark matter substructure inference from strong lensing, finding that these methods can exhibit significant biases when the test data deviates from the training distributions. Initial tests of using domain adaptation (Farahani et al. 2021) for improving robustness of CNNs and neural posterior estimators when predicting characteristics of strong lens systems in the presence of distributional shift were performed in Swierc et al. (2023, 2024) and Agarwal et al. (2024).

High Dimensional Inverse Problem for Lens Modeling—The task of lens modelling, that is, predicting surface brightness of background sources and density maps of foreground lenses is, in its simplest form, a non-linear inverse problem involving a handful of parameters ($\sim 10 - 20$). However, as the quality and resolution of data increases, such parametric description of lensed object becomes too simplistic, and more complex parametrization become necessary to avoid biases. An example of such a parametrization that is particularly well-adapted to ML applications are pixelated images of sources surface brightness and projected densities of lenses. Traditionally, it has been difficult to characterize appropriate priors analytically on such high-dimensional spaces (see, e.g., Suyu et al. 2006; Dye & Warren 2005; Birrer et al. 2015; Vegetti & Koopmans 2009; Nightingale et al. 2018), however, recent advances in high-dimensional inference with deep learning has made progress on this front possible.
An initial attempt at solving the source reconstruction problem was presented in Morningstar et al. (2019), and extended in Adam et al. (2023) to enable joint modeling of generalized pixellated lens densities and sources surface brightness. However, while these models provided high-fidelity maximum a posteriori (MAP) estimates, they lacked the crucial ability to quantify uncertainties. Approaches based on variational inference (VI) have also been proposed in Chianese et al. (2020); Karchev et al. (2022b); Mishra-Sharma & Yang (2022).

Advances of, e.g., Song & Ermon (2019); Ho et al. (2020); Song et al. (2020, 2021); Yang et al. (2023), have shown that generative models used as expressive, data-driven priors are a promising alternative to address this problem. Adam et al. (2022); Karchev et al. (2022a) used score-based models (SBMs) as flexible priors in an explicit inference framework to produce posterior samples of background galaxy sources. In Barco et al. (2025b), this method was extended to allow joint sampling the source and lens parameters for smooth, parametric lenses. Such methods have been shown to alleviate known biases in lens parameters induced by misspecified traditional priors, and methods have been proposed to empirically adapt initially biased SBM priors to correct for, e.g., population-level evolution of galaxy morphologies (Barco et al. 2025a), and to empirically extend misspecified physical models (Payot et al. 2025). More recently, Legin et al. (2025) has shown that SBM priors can be leveraged in a Gibbs sampling scheme to reanalyze *HST* data from lenses observed by Sloan Lens ACS (Advanced Camera for Surveys) Survey (SLACS). Ongoing challenges include increasing the sampling efficiency of these methods to allow modelling a large fraction of the strong lenses expected with LSST.

Leveraging LSST time-series data—LSST will generate an overwhelming number of transient alerts, making the discovery and characterization of strongly lensed short-lived transients (e.g., supernovae) both difficult and time-critical. Kernel-based methods and probabilistic machine learning models such as Gaussian processes will likely play a major role in time-delay inference for lensed quasars (e.g., Cuevas-Tello et al. 2006; Cuevas-Tello et al. 2010; Hojjati & Linder 2014; AL Otaibi et al. 2016; Tak et al. 2017) and supernovae (e.g., Hayes et al. 2024, 2025) discovered by LSST. Temporal deep learning models will also be essential in this area. For instance, Morgan et al. (2022) developed DeepZipper, which integrates long short-term memory (LSTM) networks with CNNs to jointly process temporal and spatial information for identifying strongly lensed supernovae in time-domain surveys. Bag et al. (2024) developed a model using unresolved light-curve data from difference imaging, while Bag et al. (2025) extended this to full multi-band time series with a 2D convolutional LSTM network. Huber & Suyu (2024) instead used an LSTM network to predict time delays between such transients directly from their light curves, whilst Gonçalves et al. (2025) used an ensemble of CNNs to directly estimate H_0 from time series of lensed supernova images and Campeau-Poirier et al. (2023) demonstrated the potential of NREs to infer H_0 from time delays and lens models. Beyond short-lived transients, Jiménez-Vicente & Mediavilla (2025) modeled microlensing in lensed quasar light curves, and Fagin et al. (2024) introduced a latent stochastic differential equation (SDE) framework to jointly model AGN variability and transfer functions, potentially extendable to lensed AGNs for joint inference of time delays and disk parameters.

3.3. Weak Lensing

- **Methodology.** SBI, Neural compression, Differentiable programming, Hierarchical Bayes, Diffusion models
- **Challenges.** Covariate shifts, UQ, Scalability
- **Opportunities.** Multi-resolution joint processing (Roman+LSST), probabilistic deblending, DeepDISC instance segmentation, physics-informed priors for galaxy morphology

As light from background galaxies travels through the Universe, its path is deflected by the gravitational potential of foreground matter, inducing subtle shape distortions of observed galaxies that can be statistically measured. This effect, referred to as weak gravitational lensing, provides a direct probe of the total matter distribution in the Universe, making it a powerful tool for constraining cosmological parameters such as the matter density Ω_m, the amplitude of matter fluctuations σ_8, and the dark energy equation of state w in a cold dark matter with cosmological equation of state (wCDM) model. With its unprecedented depth, image quality, and sky coverage, LSST will provide the most precise mapping of the large-scale structure of the Universe to date. This level of precision has two key implications: (1) It presents a major opportunity to refine cosmological constraints, motivating the development of advanced inference methods that can fully exploit this high-quality data; (2) It demands rigorous control of systematic uncertainties to ensure unbiased cosmological constraints.

Systematics modeling / mitigation—Systematic errors, such as imperfect shear calibration, photo-z uncertainties, spatially varying selection effects, and point-spread function (PSF) residuals, must be well characterized. To date, the precision of current cosmological surveys has permitted the use of state-of-the-art prescriptions that capture the dominant effects of these systematics (e.g., Mandelbaum 2018). However, as forthcoming large-scale structure and weak-lensing data from LSST achieve substantially higher statistical precision, a more accurate and detailed characterization of these systematics will be required, at a level of complexity that renders purely analytical treatments intractable. ML methods offer a complementary pathway by learning complex nonlinear mappings from observational features—e.g., properties of individual galaxies, local image quality metrics, PSF residuals, depth maps, shape measurement parameters—to the resulting systematic bias or residual error (Tewes et al. 2019; Rezaie et al. 2020; Pujol et al. 2020). Neural network or other ML algorithms can be trained on simulated or calibration data for which the true systematic shifts are known, and can then be tuned to predict, flag or correct for the systematic effect when applied to real survey data (Fluri et al. 2022). In doing so, these methods enable rapid and flexible removal of systematic contamination from the cosmological signal, thereby yielding a cleaner, more robustly inferred signal.

Clean catalog construction—Systematics such as Intrinsic Alignment (IA; e.g., Mandelbaum et al. 2006, 2011; Troxel & Ishak 2015; Joachimi et al. 2015) can be mitigated by constructing clean source and/or lens catalogs from galaxy populations where these effects are known to be negligible. Scalable inference of galaxy properties that correlate with active galaxy populations (such as specific star formation rate, sSFR) can enable the construction of IA-mitigated galaxy samples. For instance, machine-learned generative priors can be leveraged (e.g., `pop-cosmos`; Alsing et al. 2024; Thorp et al. 2024a, 2025b; Deger et al. 2025) to estimate per-galaxy sSFR and construct clean catalogs of star-forming galaxies with conservative cuts based on this parameter. This approach, especially when combined with amortized NPE, is scalable to LSST-sized datasets and is expected to outperform color-based selections, which are affected by contamination. Moreover, generative models of the galaxy population can be applied in a weak lensing context to directly infer the redshift distributions of source catalogs subject to tomographic binning and sample selection criteria, provided that the color-redshift relation is realistic and robust. This provides an alternative to approaches such as SOM calibration.

Neural Compression and Simulation-Based Inference—Traditional weak-lensing analyses follow a two-step pipeline: compress high-dimensional shear or convergence fields into summary statistics, then per-

form Bayesian inference on these summaries to obtain posteriors over cosmological parameters. The matter power spectrum and shear–shear correlation functions remain workhorse statistics (KiDS-1000 cosmic shear: Asgari et al. 2021; DES Y3 cosmic shear: Amon et al. 2022; Secco et al. 2022), but in the LSST era significant non-Gaussian information will become available. This has motivated the use of higher-order moments such as the bispectrum and trispectrum (e.g., Gatti et al. 2022), as well as peak counts (e.g., Marques et al. 2024), persistent homology (Prat et al. 2025), and Minkowski functionals (e.g., Kratochvil et al. 2012). While powerful, these handcrafted summaries are not guaranteed to capture all cosmological information. An alternative enabled by ML is to train neural networks that compress the maps directly into low-dimensional summaries. Several strategies have been explored in the weak lensing literature for training such networks (see Lanzieri et al. 2025, for a comparison); among these, information-theoretic criteria (such as those used in Information-Maximizing Neural Networks (IMNNs; Charnock et al. 2018; Makinen et al. 2021) and Variational Mutual Information Maximization (VMIM; Jeffrey et al. 2021)) can yield summaries that closely approximate sufficient statistics, achieving near-optimal compression. Hybrid methods that combine physics-based summaries (e.g., power spectra) with learned summaries can leverage the strengths of both (Makinen et al. 2024). Because the likelihood of these learned summaries is unknown, NDE techniques such as normalizing flows are then used to approximate the posterior within a SBI framework (Alsing et al. 2019). This strategy was first demonstrated on survey data in Jeffrey et al. (2021) and subsequently applied to recent surveys (e.g., Jeffrey et al. 2025; von Wietersheim-Kramsta et al. 2025), with Jeffrey et al. (2025) reporting more than a factor of two improvement in dark energy parameter precision compared to power spectrum inference. However, the practical limit to SBI is not the ability to extract information from the data through learned summaries, but rather the difficulty of producing simulations realistic enough to be compared to observations without incurring biases from model misspecification. In addition, neural summaries are notoriously difficult to interrogate: monitoring them for covariate shifts, unmodeled systematics, and failures in specific regions of the data space is challenging, thereby complicating the construction of robust null tests and diagnostic pipelines. That said, posterior predictive checks against conventional summary statistics (e.g., the power spectrum or higher-order moments) can help detect some forms of model misspecification, though they are not guaranteed to catch all issues.

Hierarchical Bayesian Field-Level Inference—With the advent of GPU-accelerated probabilistic programming, it has become feasible to model the full weak-lensing field in a hierarchical framework that links Gaussian initial conditions of the matter density to observed shear maps through an explicit forward simulation model. Proof-of-concept studies have demonstrated this approach in simplified weak-lensing settings (e.g., Porqueres et al. 2023), showing substantial gains in constraining power relative to power-spectrum analyses. DESC members have contributed key building blocks for such end-to-end pipelines, including differentiable lensing lightcone constructions (Lanzieri et al. 2023) and accurate, differentiable ray-tracing schemes (Zhou et al. 2024b). However, scaling these methods to a full LSST analysis remains extremely challenging: the survey volume and the resolution required for the forward model place stringent demands on memory, compute, and algorithmic efficiency. Ongoing work aims at lifting this bottleneck through distributed simulations across multiple GPUs (Kabalan & Lanusse 2025). In parallel with full forward modeling of the large-scale structure, DESC members have also explored map-based hierarchical inference using lognormal fields (Boruah et al. 2022; Zhou et al. 2024a), which is far less computationally demanding but whose ultimate accuracy is constrained by the limitations of the lognormal approximation. As an alternative approach, DESC members have proposed using diffusion models to learn the forward model of the density field implicitly from simulations and combine this learned prior with an explicit likelihood to constrain observed shear data (Remy et al. 2023), enabling fast reconstruction of high-fidelity mass maps.

3.4. Galaxy Clusters

- **Methodology.** SBI, Object detection, CNNs, Hierarchical Bayes
- **Challenges.** Covariate shifts, UQ, Scalability
- **Opportunities.** Combination of imaging and catalog data, Hierarchical Modeling

Galaxy clusters trace the most massive peaks of the matter density field and form relatively late in cosmic history, making their abundance and internal properties highly sensitive to the growth of structure and to dark energy. Cosmological constraints from clusters have traditionally relied on measurements of the cluster mass function and its redshift evolution, anchored by calibrated relations between mass and observable proxies. ML methods are now entering this pipeline at multiple stages (cluster finding, mass–observable calibration, and population-level inference) offering new ways to combine imaging, catalog, and multi-wavelength data while retaining control over systematics and uncertainties.

Cluster Finding from Images and Catalogs—The first step in cluster cosmology is robust identification of cluster candidates. Non-ML algorithms, such as Red-sequence Matched-filter Probabilistic Percolation (RedMaPPer; Rykoff et al. 2014, 2016), the Wavelet Z-Photometric (WaZP) cluster finder (Aguena et al. 2021a), and the Euclid cluster finders AMICO and PZWAV (Euclid Collaboration: Adam et al. 2019), have been widely used to identify optically selected clusters in galaxy catalogs. In parallel, deep-learning–based cluster finders have emerged in the Sunyaev–Zeldovich (SZ; e.g., Bonjean 2020; Lin et al. 2021; Hurier et al. 2021; Meshcheryakov et al. 2022) and optical domains (Chan & Stott 2019; Grishin et al. 2023, 2025; Tian et al. 2025). A key advantage of these approaches is that they can operate directly on images, rather than on pre-processed catalogs, and thus potentially exploit features (e.g., diffuse emission, subtle color–magnitude structure, environment) that are not captured in standard catalog-level summaries. For example, a You Only Look Once (YOLO; Redmon et al. 2015; Redmon & Farhadi 2016, 2018) architecture trained on images centered on SDSS RedMaPPer clusters was shown to recover not only the training sample but also previously missed systems that were later confirmed in external X-ray catalogs (Grishin et al. 2023). Recent DESC work applied YOLO-CL to DC2 simulations (Grishin et al. 2025), training on both observed SDSS RedMaPPer clusters and simulated massive halos. While the model performs well on images centered on known clusters, blind application to DC2 as a survey shows degraded completeness and purity, highlighting the need for further development and architectural updates, including modern YOLO variants and transformers (Tran et al., in prep.), before deployment on LSST.

Mass-Observable Relations and Weak-Lensing Mass Calibration—Cosmological analyses require accurate and precise relations between cluster mass and observables (richness, SZ signal, X-ray luminosity/temperature, velocity dispersion). Deep neural networks are being explored as flexible mass estimators across multiple wavebands, including X-ray signatures (Ntampaka et al. 2019; Krippendorf et al. 2024; Iqbal et al. 2025), the dynamics of member galaxies (Ho et al. 2019, 2021, 2022; Wang & Thiele 2025), and SZ measurements (de Andres et al. 2022). For LSST, photometric galaxy data contribute primarily through weak-lensing mass estimates that anchor mass–observable relations. The DESC Cluster Lens Mass Modeling tool (CLMM; Aguena et al. 2021b) currently infers weak-lensing masses from radial shear profiles using traditional likelihoods and Markov chain Monte Carlo (MCMC), on the assumption of parametric mass

models such as Navarro–Frenk–White (NFW; Navarro et al. 1997). Ongoing work within DESC explores alternative SBI approaches at this level, which can in principle incorporate more realistic shear profiles, complex noise, and selection effects without requiring an explicit closed-form likelihood.

Simulation-Based Inference for Cluster Cosmology—SBI provides a computationally efficient method for deriving posteriors for cluster- and population-level parameters directly from simulated data vectors. This is particularly attractive for analyses that must jointly model individual clusters and the cluster population via hierarchical frameworks, where traditional likelihood-based methods become increasingly costly and brittle as the parameter space and model complexity grow. Ongoing work within DESC indicates that SBI can recover cluster weak-lensing mass posteriors consistent with those from MCMC, provided that model misspecification is not worse than in the explicit-likelihood case (Gill et al., in prep.). In addition, SBI has been shown to derive relevant constraints directly from cluster counts (Reza et al. 2022, 2024; Zubeldia et al. 2025), and this approach is now being developed on DESC simulations as an alternative, simulation-native pathway to cluster cosmology that can be integrated into the DESC Cluster Cosmology Pipeline. Compared with Stage-III experiments, DESC cluster analyses will require richer astrophysical modeling and the exploration of larger parameter spaces; SBI offers the algorithmic flexibility and, in many regimes, the computational efficiency required to meet these demands.

3.5. Supernova Cosmology and Transients

- **Methodology.** Hierarchical Bayes, RNNs, Transformers, Active learning, Ensembles, SBI, Anomaly detection, VAEs, Gaussian processes
- **Challenges.** Covariate shifts, Data sparsity, Scalability, UQ
- **Opportunities.** PLAsTiCC/ELAsTiCC simulation infrastructure, DESC leadership in alert broker integration (ALERCE, Fink, ANTARES)

LSST is expected to detect ∼10 million transient and variable objects each night, a thousand-fold increase over current surveys. The sheer volume and cadence of detections renders traditional spectroscopic classification infeasible for most events. This presents a major bottleneck to the identification of pure SN Ia samples for cosmological distance measurements, and to constraining the explosion physics of populations of rare and novel phenomena for the first time. To achieve reliable cosmological constraints and meet DESC goals of reducing the systematic uncertainties from light curve modeling below 3% of those obtained from SALT2 (Guy et al. 2007), analysis techniques demand *well-calibrated uncertainty quantification*, *adaptive and scalable performance*, and *robustness to covariate shifts and data corruption*.

Spectrophotometric Modeling—SNe Ia are broadly homogeneous and viable standard candles, but diversity in their spectro-temporal properties and persistent host-dependent effects (Sullivan et al. 2006, 2010; Lampeitl et al. 2010; Kelly et al. 2010; Hayden et al. 2013) still limit standardization precision. Modern ML approaches to standardization now focus on data-driven, differentiable, and multi-modal models rather than hand-engineered linear corrections. Already progress in these directions can be seen in modeling using probabilistic auto-encoders (Stein et al. 2022) and variational autoencoders (VAEs; e.g., `ParSNIP` Boone

2021), that predict time-evolving SEDs from light curves, and use a differentiable forward model to compare in observation space. High quality, well-calibrated data has been instrumental in these endeavors, which can be augmented by the LSST/Rubin samples; however, strategies that account for the shifts induced by calibration errors must still be investigated.

More recent efforts involve ML models as emulators (e.g. Chen et al. 2020; Kerzendorf et al. 2021; Magee et al. 2024) for radiative transfer codes such as TARDIS (Kerzendorf & Sim 2014), that can be used to infer *physical parameters* such as the total luminosity, nickel mass, the composition and the asymmetry of the ejecta given multi-modal inputs such as light curve, time-series spectroscopy, and Rubin and high-resolution space-based imaging (e.g. from the ESA *Euclid* mission and the *Nancy Grace Roman Space Telescope*). These physically motivated emulators are also capable of several tasks that would have previously required individual specialized models. These include predicting the future evolution of SNe of all types, classifying spectra while being agnostic to the imbalance in extant training samples, and can help distinguish SN Ia from impostors that might otherwise contaminate the cosmological sample, as well as actively schedule follow-up spectroscopy.

Photometric Classification—The methodological evolution of photometric classifiers from feature-based approaches (e.g. Narayan et al. 2018) to end-to-end learning has been driven by the challenge of processing irregular, heteroskedastic observations: `SNmachine` (Lochner et al. 2016) leveraged a range of feature sets, from physics-based through to non-parametric approaches, coupled with a variety of traditional ML techniques to achieve high classification accuracy; `SCONE`'s Gaussian process interpolation (Qu et al. 2021) creates regular 2D representations from sparse observations (see also `Avocado`; Boone 2019); while transformer architectures (Pimentel et al. 2023; Allam & McEwen 2024; Cabrera-Vives et al. 2024) leverage self-attention mechanisms to handle missing data naturally. Hybrid, physics-informed approaches have also been explored to extract latent features from light curves using generative modeling, which are then used for classification (Boone 2021). Classification tools have already proven their utility in LSST survey optimization for supernova metrics, with realistic LSST survey cadences (e.g., `SNMachine`, Alves et al. 2022, 2023).

DES pioneered the use of neural networks for photometric classification of SNe Ia for cosmological analysis (Möller & de Boissière 2020; Qu et al. 2021; Möller et al. 2022; Vincenzi et al. 2023; DES Collaboration et al. 2024). Propagating the prediction uncertainties from these models through to cosmological constraints remains an open problem. `SuperNNova` (Möller & de Boissière 2020) addresses the former with a recurrent neural network (RNN) providing calibrated probabilities essential for contamination modeling in dark energy constraints (Vincenzi et al. 2023). In DES, BNNs and ensemble methods were also tested (Möller et al. 2024); for DESC, we can advance fully Bayesian approaches to photometric classification for LSST.

Forward Modeling of the Time-Domain Landscape—Observational modeling led by DESC has catalyzed the development of neural approaches to photometric classification. PLAsTiCC (PLAsTiCC team et al. 2018; Hložek et al. 2023) provided 3.49M test light curves across 18 transient classes using simulations from the Supernova Analysis package (SNANA; Kessler et al. 2009) with cadences and realistic observing conditions from the LSST Operations Simulator (OpSim; Delgado et al. 2014). The challenge established weighted logarithmic loss metrics prioritizing SNe Ia and kilonovae (KNe) for DESC science goals (Malz et al. 2019), with winning solutions employing gradient boosting and ensemble neural networks requiring

engineered features (Boone 2019; Hložek et al. 2023). The dataset remains a foundational benchmark for time-series representation learning in astrophysics years after its development (Fraga et al. 2024; Masson & Bregeon 2024; Cádiz-Leyton et al. 2025b,a; Zivanovic et al. 2025). Building on PLAsTiCC, ELAsTiCC (Narayan & ELAsTiCC Team 2023; Knop & ELAsTiCC Team 2023) stress-tested end-to-end broker infrastructure with ∼50M alerts streamed in real-time to seven community brokers from September 2022–January 2023. DESC remains the only Rubin Science Collaboration to have tested the alert infrastructure from end-to-end on this scale, as it is critical for deploying AI models live, to support online learning and other tasks to optimize the scientific return from Rubin.

Similar to ELAsTiCC and PLAsTiCC, S/N Analysis of Simulated SpectrA for Rubin trAnsientS (SASSAFRAS) is a novel dataset of simulated LSST spectroscopic follow-up with 2.1M spectra across 14 transient types and three telescopes – Gemini, the Southern Astrophysical Research Telescope (SOAR), and 4MOST – using SNANA simulations with realistic noise from each of the three telescopes. These simulations contain an even distribution of spectra per class and have a wide range of redshift distribution between 0.023–1, minimizing the bias inherent from uneven distribution of spectra seen in real data. One group at NSF–Simons AI Institute for the Sky (SkAI) is currently utilizing SASSAFRAS to train a spectroscopic classifier and then transfer learn with real data to create a state of the art classifier. Spectroscopic classifiers trained on SASSAFRAS will be essential for confirming transient labels for active learning algorithms as outlined below.

Online Learning for Spectroscopic Optimization—While archival photometric classification will suffice for the bulk of LSST cosmology analyses, inference over partially obtained data remains crucial for prioritizing spectroscopic targeting before an event has ended. Gated recurrent unit (GRU)-based recurrent and convolutional neural networks have successfully classified partial-phase synthetic light curves (Muthukrishna et al. 2019; Qu et al. 2021; Gagliano et al. 2023; Shah et al. 2025), but performance on observed data remains modest. Transformer-based methods are being increasingly used (Cabrera-Vives et al. 2024), with synthetic pre-training playing a growing role in bridging the simulation gap (Gupta et al. 2025). Within DESC, host-galaxy correlations have been shown to improve early classification (Gagliano et al. 2021); this has driven data-driven modeling of host-galaxy correlations for the ELAsTiCC challenge (Lokken et al. 2023), although spurious host-galaxy associations and the small postage stamps of the field contained within the LSST alert packets may limit utility of real-time inference using these data.

Beyond real-time classification, active learning faces unique astronomical challenges: objects must be selected for spectroscopic follow-up before informative light curve data are obtained, the untargeted population is substantially dimmer than the spectroscopically confirmed sample used in training, and labeling costs vary dramatically with object brightness and sky position. The Recommendation System for Spectroscopic followup (RESSPECT)[5], an initial approach to active learning for transient science, implements uncertainty sampling with random forest classifiers on Bazin (Bazin et al. 2011) parametric features, but requires a minimum of five observations per filter, limiting early-time selection (Kennamer et al. 2020). More recent implementations have refined active learning for early-time SN Ia identification, demonstrating effective follow-up optimization with simulations (Ishida et al. 2019), real-data a posteriori (Leoni et al. 2022) and real-time observational campaigns (Möller et al. 2025), the latter revealing the need for training sets contain-

[5] https://resspect.readthedocs.io/en/latest/

ing events beyond supernovae. `Astronomaly` (Lochner & Bassett 2021) introduces personalized anomaly detection by combining isolation forests with human relevance scoring, addressing the fundamental subjectivity of an anomaly label. The approach has been shown to double the rate of anomaly discovery in radio transients (Andersson et al. 2025). However, active learning remains fundamentally limited by the lack of an informative initial training set, such that early random sampling can produce biased or unrepresentative data that propagates through subsequent iterations of learning. This is a fundamental challenge for novelty detection in LSST data.

Prompt Processing with the Transient Alert Brokers—The seven Rubin Community Brokers implement diverse classification pipelines. Automatic Learning for the Rapid Classification of Events (ALeRCE) employs a CNN for top-level classification from alert postage stamps (Carrasco-Davis et al. 2021), and a hierarchical random forest applied to photometric features for classification along a 15-class taxonomy (Sánchez-Sáez et al. 2021). Alert Management, Photometry, and Evaluation of Light curves (AMPEL) uses a four-tier system with predominantly gradient-boosted random forests (Nordin et al. 2025), and Fink deploys multiple classifiers for early and late-time classification (Fraga et al. 2024; Leoni et al. 2022; Möller & de Boissière 2020; Möller et al. 2025; Möller et al. 2021). Arizona–NOIRLab Temporal Analysis and Response to Events System (ANTARES) employs multi-stage filtering with community-contributed Python classes for tagging sources (Narayan et al. 2018), while Lasair integrates a boosted decision tree classifier from host galaxy properties (Smith et al. 2020) with multi-order coverage (MOC)-based watchmaps for coordination with 4MOST's Time Domain Extragalactic Survey (TiDES), which will be providing 35,000 transient spectra for SN Ia cosmology (Williams et al. 2024). These brokers have demonstrated sub-second latency in processing millions of alerts during the ELAsTiCC campaign, with classification probabilities reported via standardized Avro schemas that enable systematic evaluation of heterogeneous ML architectures. DESC members are involved in all seven Rubin Community Brokers, offering substantial potential for shared software infrastructure for processing the LSST alert stream.

Cosmological Inference using Type Ia Supernovae—Hierarchical Bayesian models have been applied to SN Ia for various goals, including cosmological inference (e.g., March et al. 2011; Shariff et al. 2016; Rubin et al. 2015, 2025; Feeney et al. 2018), constructing empirical SED models (`BayeSN`; Mandel et al. 2009, 2011, 2022; Thorp et al. 2021; Ward et al. 2023; Grayling et al. 2024; Uzsoy et al. 2024), modeling intrinsic colors and dust extinction (Mandel et al. 2017; Thorp & Mandel 2022; Thorp et al. 2024b), handling uncertain photometric classifications (Kunz et al. 2007; Hlozek et al. 2012) and redshifts (Roberts et al. 2017), and modeling the spectrophotometric standards used in photometric calibration (Boyd et al. 2025; Popovic et al. 2025). However, in the LSST era such models will need to incorporate complex effects that cannot easily be treated analytically; e.g., selection effects, photometric classification and photometric redshifts. Work is ongoing to enhance our statistical models using SBI, to leverage the flexibility of neural networks to capture these complex effects (e.g., Boyd et al. 2024; Karchev et al. 2024, 2025). SBI will enable scalable and principled statistical inference of cosmological parameters with LSST.

3.6. Theory and Modeling

· **Methodology.** Emulators, Gaussian processes, Neural surrogates, Differentiable programming, Symbolic regression, SBI
· **Challenges.** Covariate shifts, Scalability
· **Opportunities.** Flexible emulation frameworks, model selection and hypothesis testing, efficient construction of realistic mock datasets, gradient-based sampling.

The role of theory and modeling within DESC is to provide the essential bridge between cosmological parameters and the statistical observables derived from LSST data. Accurate theoretical models are required to translate measured galaxy shapes, positions, and fluxes into constraints on dark energy, dark matter, and gravity. This entails constructing predictive models of large-scale structure, galaxy bias, baryonic physics, and lensing observables that can be robustly compared with data while marginalizing over astrophysical and observational systematics. As the scale and precision of LSST data demand modeling at unprecedented accuracy and speed, ML-based emulators, differentiable theory libraries, and SBI approaches are increasingly central to this effort, enabling fast and robust connections between data and theory.

Fast Surrogates for Cosmological Likelihoods—Emulation and related methods for creating fast-surrogate models using ML and AI will be of crucial importance in accelerating inference pipelines for cosmological analyses in DESC. Emulation is an indispensable tool for integrating aspects of modeling which by nature require simulation too slow to ever consider incorporating directly in sampling (e.g., those requiring N-body or hydrodynamical simulations). At the same time, even for aspects of modeling which are more moderate in evaluation cost (seconds rather than many hours), emulation allows individual likelihood evaluations to be dramatically accelerated. This is one of the key ways we can make computationally feasible the sampling in high-dimensional parameter spaces which will be required for DESC analyses.

Work on these emulation techniques has included directly building emulation tools, particularly outside of wCDM models (Ramachandra et al. 2021), emulating intrinsic alignment correlations (Pandya et al. 2025b), as well as using emulators to efficiently evaluate modeling choices for LSST data (Boruah et al. 2024). The DESC theoretical modeling package `pyCCL` (Chisari et al. 2019) natively supports key matter-power-spectrum emulation tools `baccoemu` (Aricò et al. 2021) and `CosmicEmu` (Lawrence et al. 2017). However, developing frameworks for DESC to analyze models outside wCDM in the nonlinear regime of LSST data remains a challenge that needs to be addressed (Ishak et al. 2019), for which AI can play a major role.

Looking to the future, DESC would benefit from developing mechanisms to enable emulation that has more flexibility with respect to modeling components. Our current methods of building a single emulator from scratch per modeling set-up is high cost (computationally and in terms of person power). This does not scale well for enabling adaptation to new systematics modeling or inference in models outside of wCDM. Considering approaches which use meta-learning (e.g., MacMahon-Gellér et al. 2025) or which are philosophically aligned with foundation models would be of value.

Another area of growth where fast emulators can make major impact is in model selection and hypothesis testing of beyond-wCDM models. Instead of being limited by the cost of theoretical model evaluations, future analyses will be constrained by how efficiently inference pipelines can navigate and compare competing cosmological models. Integrating these emulators within agentic AI systems (Section 5.2)—which can

autonomously refine training data, adapt inference strategies, leverage tools for evidence computation and simulation-based inference, and even propose new model extensions—will further accelerate discovery. For DESC, this synergy will transform the capacity to test gravity, dark energy, and dark-sector interactions, turning high-quality data into a powerful engine for identifying new physics.

Differentiable Programming for Accelerated Sampling—Traditional cosmological inference pipelines often rely on computationally expensive numerical methods such as the Code for Anisotropies in the Microwave Background (CAMB; Lewis et al. 2000) or HaloFit (Smith et al. 2003; Takahashi et al. 2012), limiting the use of gradient-based sampling methods. Differentiable cosmological codes address this by enabling efficient computation of gradients with respect to model parameters, unlocking samplers such as Hamiltonian Monte Carlo (HMC) and the No U-Turn Sampler (NUTS; Hoffman & Gelman 2014).

The `jax-cosmo` library (Campagne et al. 2023) provides a differentiable and hardware-accelerated framework for cosmological computations. With a NumPy-compatible API and close integration with tools such as NumPyro (Phan et al. 2019) and JAXopt (Blondel et al. 2021), `jax-cosmo` offers a practical foundation for building scalable, fully differentiable cosmological models. Piras & Spurio Mancini (2023) combined `jax-cosmo` with neural-network emulators in `CosmoPower-JAX`, enabling high-dimensional Bayesian inference through automatic differentiation and GPU acceleration. The `halox` package (Kéruzoré 2025) uses `jax-cosmo` for cosmological calculations such as power spectra and distance measures in its modeling of dark-matter halo statistics (such as halo mass function and halo bias). Sui et al. (2025) used `jax-cosmo` to test a differentiable Fisher-information approach based on score matching. Recently too, Bartlett & Pandey (2025) have developed a symbolic emulator that leverages genetic programming-based symbolic regression to derive compact, analytic expressions for cosmological observables, including the radial comoving distance, linear growth factor, and nonlinear matter power spectrum.

3.7. Cosmological and Survey Simulations

- **Methodology.** Emulators, Diffusion models, Differentiable programming, SBI, GNNs
- **Challenges.** Covariate shifts, Scalability, Metrics
- **Opportunities.** Joint modeling of galaxies and environments, inference at catalog and field level, modular components for DESC simulators, survey-scale generative models, survey design, systematics mitigation, pipeline stress tests.

Cosmological simulations are a fundamental tool not only for validating analysis pipelines but also, increasingly, for providing "theory" samples for SBI frameworks. Producing mock survey data at the scale and accuracy required for LSST science remains a major challenge, which ML can help address by emulating expensive numerical predictions and by providing data-driven models of otherwise poorly constrained aspects of the galaxy population. Neural emulators have long been used as fast surrogates for non-differentiable components of cosmological forward models, e.g., summary statistics of N-body simulations as in `CosmicEmu` (Heitmann et al. 2016; Moran et al. 2023), `Aemulus` (DeRose et al. 2019), `CosmoPower-JAX` (Piras & Spurio Mancini 2023), or `21cmEMU` (Breitman et al. 2024), and more recently for accelerating SPS calculations via models such as `speculator` and `ProMage` (Alsing et al. 2020; Tortorelli et al. 2025). Extending these

approaches to additional components of the simulation pipeline holds the promise of greatly increasing the dynamical range, realism, and flexibility of LSST mock catalogs at manageable computational cost.

Population-Level Generative Models for Realistic Galaxy Catalogs—Diffusion-based generative models operating in the space of physical galaxy parameters provide a powerful route to building realistic mock catalogs that remain anchored in deep-field observations. The `pop-cosmos` framework (Alsing et al. 2024; Thorp et al. 2025b; Deger et al. 2025) defines a score-based diffusion model over a high-dimensional SPS parameterization—star formation history (SFH), metallicity, dust, nebular emission, etc.—calibrated on $\sim$420,000 galaxies from COSMOS2020 (Weaver et al. 2022) spanning 26 bands from UV to mid-IR. Rather than directly emulating observed fluxes, `pop-cosmos` learns a data-driven prior $p(\boldsymbol{\theta}_{\rm SPS}, z)$ over physical parameters and redshift that reproduces the joint distribution of observed photometry. This model encodes realistic priors on star-formation histories over cosmic time, and learns the evolution of the star-forming sequence. When coupled to survey-specific selection functions and noise models, `pop-cosmos` can therefore generate realistic mock galaxy catalogs that inherit both empirical constraints from deep multi-wavelength data and the flexibility of generative modeling.

Differentiable Empirical Galaxy–Halo Forward Modeling—Complementary to purely catalog-level generative approaches, differentiable galaxy–halo forward models seek to describe galaxy populations as conditional generative processes on top of dark-matter structure. The `Diffsky` framework (OpenUniverse et al. 2025) rebuilds the traditional "halo $\rightarrow$ SFH $\rightarrow$ SED" chain using differentiable, physically interpretable blocks. `Diffmah` (Hearin et al. 2021) provides a JAX-based, few-parameter model of halo mass assembly $M_{\rm halo}(t)$, replacing noisy merger trees with smooth, analytic, differentiable growth histories. On top of this, `Diffstar` (Alarcon et al. 2023) models in situ star-formation histories with a small set of parameters (e.g., star-formation efficiency, gas-consumption timescale, quenching time), while `DiffstarPop` (Alarcon et al. 2025) lifts this to the population level by learning the statistical link between SFH parameters and halo assembly across suites of reference simulations. Finally, Differentiable Stellar Population Synthesis (`DSPS`; Hearin et al. 2023) maps these SFHs and associated metallicity/dust parameters to SEDs and photometry entirely within JAX. Together, `Diffmah` + `Diffstar`/`DiffstarPop` + `DSPS` replace merger trees, non-differentiable semi-analytic recipes, and black-box SPS calls with a modular, probabilistic, fully differentiable stack whose low-dimensional, physically meaningful parameters can be calibrated and explored with gradient-based methods and SBI, while still generating large, realistic synthetic catalogs. AI-based models are in development to generate multiband galaxy images from these synthetic catalogs, conditioned on the parameters of the `Diffsky` suite.

Differentiable Cosmological N-body Solvers—Recent years have seen the emergence of particle–mesh N-body solvers implemented in modern, GPU-accelerated, deep-learning frameworks that support automatic differentiation (e.g., Modi et al. 2021; Li et al. 2022; List et al. 2025; Lanusse et al. 2025). Automatic differentiation enables hierarchical Bayesian inference directly over forward simulations of large-scale structure, opening a path toward full-field inference and near-optimal extraction of cosmological information. The main obstacles to deploying these methods at LSST scale are the computational and engineering demands of simulating survey-sized volumes in a differentiable way. Computing derivatives through the simulation implies non-trivial memory costs that are difficult to satisfy under the constraints of GPU accelerators. Several complementary strategies are being developed to address this challenge, including multi-node domain decomposition for distributed simulations (Kabalan & Lanusse 2025), techniques to reduce the memory

cost of gradient evaluation (Li et al. 2024), and improved time integrators that achieve a given accuracy with fewer time steps (Rampf et al. 2025). Beyond enabling full-field inference, differentiable simulations also enable the combination of physics-based solvers with learned components, yielding hybrid schemes that can improve the speed and accuracy of particle–mesh simulations (Lanzieri et al. 2023; Payot et al. 2023).

Hydrodynamical-Simulation-Based Mappings of Galaxy Properties—A complementary strategy is to treat state-of-the-art hydrodynamical simulations as high-fidelity "teachers" and use ML to distill their complex, small-scale physics into fast, effective models defined directly on dark-matter fields. Rather than specifying parametric galaxy–halo or SPS models, these approaches learn mappings from halo or large-scale-structure descriptors to galaxy properties as realized in the simulations. Recent work on IA illustrates this paradigm. A traditional approach, followed by Van Alfen et al. (2024) is to develop an empirical IA model constrained by hydrodynamical simulations within a flexible halo occupancy distribution (HOD)-like framework. A more ML-oriented approach is to learn an end-to-end emulator of galaxy properties. Jagvaral et al. (2025) introduce a geometric deep-learning approach in which galaxy shapes and orientations from IllustrisTNG (Nelson et al. 2019) are modeled using E(3)-equivariant graph neural networks defined on the cosmic web, capturing the conditional distribution of shapes and orientations given halo mass, environment, and tidal field. This yields a fast, simulation-calibrated surrogate that reproduces intrinsic-alignment statistics at the percent level, enabling embedding of hydro-level realism into DESC mock catalogs without rerunning computationally expensive hydrodynamical simulations.

3.8. Object Classification

- **Methodology.** Ensembles, GNNs, Active learning, Transformers, Self-supervised learning
- **Challenges.** Covariate shifts, Data sparsity, Scalability, UQ
- **Opportunities.** Multi-survey training, Data-driven Priors, Generative Models

The classification of astronomical objects is an area that has seen a surge in ML and AI applications over the past decade. In regards to DESC science, we call out three potentially relevant areas: the detection and removal of bogus/non-astrophysical sources, the classification of galaxy types, and star/galaxy separation. While the first area is critical for minimizing the pollution of source catalogs, it falls mostly under the responsibility of the Rubin Project's data management team and is addressed at the instrument signature removal stage of the Rubin pipeline (Bosch et al. 2018). Galaxy type classification has been a strong application of citizen science and AI/ML applications, with an increasing demand and opportunity expected from the depth that will be achieved by Rubin (e.g., Cao et al. 2024). However, galaxy type classification is not currently identified as a main concern for DESC beyond what is necessary to ensure a clean sample of galaxies for cosmological analyses (e.g., identifying merging galaxies that could confuse some of cosmological measurements).

On the other hand, star/galaxy separation is expected to be crucial to DESC science, as it may influence catalog completeness, weak-lensing shear estimation, galaxy clustering estimates, calibration of photometric redshifts, and object selection for spectroscopy. At the imaging depth of LSST, galaxies vastly outnumber stars, and a nontrivial fraction of those galaxies are compact and blue, making them morphologically and

photometrically similar to point sources, especially under variable seeing and in crowded fields (Fadely et al. 2012).

The nominal approach to star/galaxy separation implemented by the Rubin Science Pipeline uses a cut-based approach based on object "extendedness", which is constructed from a comparison of the PSF- and model-based measurements (see Slater et al. 2020, and references therein). This classifier performs well at bright magnitudes but struggles at the faint end, leading to either very large contamination in the faint star sample (i.e., orders of magnitude more galaxies than stars) or significant (i.e., nearly total) incompleteness for stars. Prior work from DESC members, working on precursor surveys like DES, improved on simple cut-based analyses by using feature-based machine learning—decision trees, random forests, boosted ensembles, and shallow neural networks trained on catalog-level features such as colors, shape moments, and PSF–model magnitude differences (e.g., Soumagnac et al. 2015; Sevilla-Noarbe et al. 2018; Baqui et al. 2021; Bechtol et al. 2025). However, these methods are ultimately limited by deblending errors, incomplete PSF modeling, and the loss of informative spatial structure in catalog summaries.

Parallel investigations have examined models that ingest multi-band image cutouts together with catalog features, allowing networks to learn morphology directly while leveraging color-based information (e.g., proximity to the stellar locus, color-color degeneracies, and uncertainty-aware colors; Kim & Brunner 2017). Efforts increasingly incorporate PSF awareness (e.g. Patel et al. 2025), multi-epoch data (i.e., variability), and deblending context to improve robustness across seeing conditions and sky regions, and explore semi/self-supervised representation learning to exploit LSST's vast unlabeled datasets. Looking ahead, promising methodologies include end-to-end probabilistic deep learning that propagates PSF and noise models into calibrated class probabilities (Burke et al. 2019), which can include prior probabilities based on spatial location and spectrum (López-Sanjuan et al. 2019); vision transformers and graph neural networks (GNNs) that integrate multi-visit information; multi-modal architectures combining images, colors, variability, and proper motion; domain adaptation and label-shift correction to handle spatial and temporal heterogeneity; active learning using sparse spectroscopic labels; and simulation-based training on realistic survey mocks. Emphasis on UQ, continual and federated learning across data releases, and physics-informed constraints should yield classifiers that are both scalable and scientifically reliable at the LSST depth.

3.9. Deblending

- **Methodology.** VAEs, NPE, Instance segmentation, Diffusion models, Normalizing flows, CNNs, Object detection, SOMs
- **Challenges.** Data sparsity, Metrics, UQ
- **Opportunities.** Multi-survey training, Data-driven Priors, Generative Models

Turning pixels into objects is a fundamental problem in astronomical survey pipelines. Object detection and deblending of LSST data is a crucial step in producing catalogs useful for DESC science. Given its unprecedented depth for a ground-based survey, LSST will face new challenges in its detection pipelines compared to previous legacy surveys (Melchior et al. 2021). Blending, or the overlapping of source light profiles, is an imaging systematic that affects all downstream analysis, as it becomes difficult (in reality intractable) to

disentangle photons from a given source in a blend. This problem is exacerbated with increased observing depth, as more light is collected from sources that are overlapping due to line-of-sight projections or physical interactions. Traditional object detection pipelines for wide-field surveys use a maximum likelihood estimator method (Bosch et al. 2018) to identify peaks in intensity corresponding to sources in an image. This method, while statistically justified, is still subject to failure modes, wherein AI can provide alternative and complimentary methods. Similarly, traditional deblending algorithms typically rely on models and assumptions about source light profiles that may not provide sufficient flexibility for the billions of sources LSST will observe (Melchior et al. 2018). DESC has been exploring and developing AI methods to aid in these challenging problems, which are crucial to understand and mitigate.

Catalog-level Blend Identification—While a majority of sources observed by LSST are expected to have some level of blending, a particularly pernicious case is that of unrecognized blends. These are sources in a blended scene that are indeed distinct (determined from high-resolution space-based observations), but are only recognized as a single source by cataloging and deblending pipelines. Unrecognized blends impact measured properties such as galaxy shapes (Dawson et al. 2016), photometric redshifts (Liang et al. 2025), and more. Estimates of the level of unrecognized blends in LSST range from ~ 15–30%, with analysis of early LSST data compared to catalogs from the *HST* Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) yielding an unrecognized blend rate of 18% (Adari & von der Linden 2025). Blends that remain at the catalog level are definitionally unrecognized blends but may still be detectable as outliers via their multi-band photometry or their shapes. Random forests and SOMs along with various anomaly detection algorithms were tested in Liang et al. (2025) who showed that unrecognized blends can be detected at a cost to the sample size. These algorithms were used to assign an unrecognized blend probability, however improvements can be made for specific science cases. For example, using the blend entropy (Ramel et al., in prep) can improve cluster cosmology by removing the most problematic unrecognized blends for cluster analysis. Designing better blending metrics like blend entropy and incorporating them into ML algorithms like GNNs is the main goal of the software package, friendly .

Image-level Deblending Using Deep Learning—Deep learning algorithms designed for object detection and deblending provide an alternative method to traditional pipelines that may help improve catalog completeness and source property measurements. For instance, the Bayesian Light Source Separator (BLISS) framework uses NPE to infer probabilistic catalogs by training a CNN directly on multi-band images (Hansen et al. 2022). BLISS produces well-calibrated posterior approximations for various source properties, and point estimates based on these posterior approximations outperform the standard LSST pipeline in source detection, flux measurement, star/galaxy classification, and galaxy shape estimation (Duan et al. 2025). The method is robust to spatially varying backgrounds and point-spread functions, provided these features are present in the simulated training images (Patel et al. 2025). The DeepDISC instance segmentation (Merz et al. 2023) framework produces object catalogs and segmentation masks from image data, and is being tested with joint *Roman*–Rubin data to incorporate multimodal information for downstream detection and deblending improvement. Both `DebVader` (Arcelin et al. 2021) and Maximum A posteriori with Deep NEural networks for Source Separation (MADNESS; Biswas et al. 2025) use VAEs to handle blending. They use self-supervised training to learn the structure of isolated galaxies. Through additional training of a deblending encoder they learn to isolate a galaxy from a blend. A specialized decoder can directly measure the characteristics of the galaxy (shape, photo-z) without reconstructing explicitly the image of the isolated galaxy. MADNESS adds a normalizing flow to the architecture to improve performance by modelling the

latent-space distribution of galaxies, thereby providing an explicit likelihood for posterior optimization. Ongoing work uses a multimodal VAE to learn both from imaging and spectroscopy, adding more information in the latent space to improve the galaxy characteristics measurement, especially photo-z. The very principles of deblending VAEs alleviate the impact of unrecognized blends, and ongoing work on the use of probabilistic catalogs where the number of detected galaxies is itself non-deterministic will reduce it even more. Even if sources are in such close proximity that LSST imaging will not be able to recognize the overlap and detect the blended group as one source, it is feasible to model the detected sources first, compute the residuals from the fit, and run detection again on the residuals. Because the residuals can have complicated structure, it is beneficial to perform the detection on multi-band residuals, where unrecognized sources appear as colored, localized over- or underdensities. Recognizing them, as well as their likely centers is possible and fairly effective with computer vision architectures like YOLO (Kamath 2020).

Data Driven Priors for Deblending with Explicit Likelihoods—Generative models such as normalizing flows and diffusion models can be trained on unblended galaxies (potentially limited amounts of space-based data) and then serve as data-driven priors for galaxy morphologies (Lanusse et al. 2019). Posterior optimization and sampling becomes possible for inverse problems with explicit likelihood functions (such as inpainting, deconvolution, and deblending). This is particularly effective in low signal-to-noise ratio (SNR) cases where the deblender `scarlet`, (Melchior et al. 2018) which is the default deblending method in the Rubin Science Pipelines, is outperformed by a new, prior-augmented version from `scarlet2` (Sampson et al. 2024). The same approach can also perform transient photometry in the presence of a host galaxy without the need for difference imaging (Ward et al. 2025). Additionally, posterior optimization in latent space by the MADNESS deblender (Biswas et al. 2025), using data-driven priors, also outperformed scarlet.

3.10. Shape Measurement

- **Methodology.** Differentiable programming, Deep networks, SBI
- **Challenges.** Covariate shifts, UQ, Data sparsity, Scalability
- **Opportunities.** Joint optimization of detection, deblending, and shear, hybrid analytic–neural estimators, realistic multi-instrument and multi-epoch scene modeling, active learning, unified shear–photo-z modeling

Weak-lensing shape measurement is one of the most critical and challenging components of the DESC analysis pipeline: small percent-level biases in ensemble shear propagate directly into the cosmological parameters targeted by LSST. Meeting DESC requirements therefore demands methods that are simultaneously accurate enough to control multiplicative and additive shear biases, computationally efficient enough to process billions of galaxies, and amenable to calibration. A unifying theme in recent work is the exploitation of differentiability and GPU acceleration, both in explicit shear calibration schemes and in forward models.

Analytic Calibration and Differentiable Shear Estimators—The Analytic Calibration (AnaCal) framework (Li & Mandelbaum 2023; Li et al. 2025b) demonstrates how differentiability can be used to obtain high-precision

shear responses and noise-bias corrections without relying on large external simulation campaigns. By representing galaxy properties and pixelized images using differentiable basis functions, AnaCal yields analytic shear responses for detection, selection, and shape measurement, achieving LSST-grade accuracy with sub-millisecond inference per galaxy. More broadly, other calibration schemes such as metacalibration (Huff & Mandelbaum 2017; Sheldon & Huff 2017) stand to benefit from differentiable image models and measurement operators: with gradients available throughout the pipeline, shear response and noise-bias corrections can be computed more quickly and robustly.

Deep Learning–Based Shape Estimators—Modern deep-learning architectures provide a natural path to an end-to-end differentiable shear estimator that can simultaneously integrate detection, deblending, shear estimation, and robustness to image defects. Neural networks are inherently GPU-accelerated, highly parallel, and differentiable, making them well suited to high-throughput shear inference at LSST scale. Such an approach was originally demonstrated in (Ribli et al. 2019) and is being explored in a DESC context using the DeepDISC architecture (Merz et al. 2023), which was originally designed as a general purpose architecture for detection and segmentation and which can estimate gravitational shears within a single GPU-resident and differentiable model. Being automatically differentiable, this estimator can be calibrated using the schemes mentioned above.

Hierarchical Forward Modeling with Differentiable Image Simulators—A complementary strategy frames shape measurement as a hierarchical forward-modeling problem, in which cosmological parameters, population-level distributions of galaxy properties, and individual galaxy shapes are inferred jointly from the pixel data (Schneider et al. 2015). In this view, a forward model generates simulated images given a set of hierarchical parameters, and inference proceeds by comparing these simulations to the observed images. Such approaches are made practical thanks to the `JAX-GalSim` effort which re-implements key `GalSim` (Rowe et al. 2015) functionalities in JAX, making this forward model fully differentiable and GPU-accelerated while supporting vectorized batch simulations of thousands of galaxies at once. In ongoing DESC efforts, `JAX-GalSim` is used to implement the hierarchical shear-inference framework of Schneider et al. (2015). Instead of traditional MCMC, gradient-based samplers (NUTS), GPU acceleration, and batching can be used to yield roughly an order-of-magnitude speedup while keeping multiplicative shear biases within LSST requirements. While the aforementioned approach relies on analytic surface brightness profiles to model galaxies, more realism can be achieved through projects like `scarlet2` (Sampson et al. 2024) which extends the modeling to non-parametric morphologies and blended scenes observed with multiple instruments, providing a JAX-based, differentiable scene-modeling framework in which gradients of the likelihood with respect to source parameters and hyperparameters are readily available.

3.11. Synthesis and Recommendations

ML has become a foundational component of the collaboration's scientific infrastructure. It appears at every stage of the analysis pipeline: from pixel-level data processing (deblending, shape measurement) through derived observables (photo-z, cluster masses) to cosmological inference itself (weak lensing, SNe, clusters). Adoption is not driven by bespoke, application-specific methods. Instead, a small set of core methodologies and a consistent set of fundamental challenges recur across the entire application landscape, as

visualized in Figure 1. These cross-cutting patterns have direct implications for how DESC should organize its AI/ML efforts, prioritize methodological investments, and structure collaboration-wide infrastructure.

Among shared approaches, a few key methodologies appear repeatedly: *SBI*, enabling parameter estimation from lensing to supernovae, though constrained by the fidelity of forward-models; *differentiable programming frameworks* such as `jax-cosmo`, `JAX-GalSim`, `scarlet2`, for unlocking gradient-based inference at scale; *NDE and generative models* such as `pop-cosmos`, `DebVader`, and `MADNESS`, which provide flexible probabilistic representations with demonstrated cross-application reusability; *emulators* such as `CosmoPower-JAX` and `speculator`, which trade training cost for orders-of-magnitude speedups; and *active learning*, for maximizing scientific return from limited expert annotations. These techniques recur because the scientific challenges in DESC (extracting maximal information under stringent systematic control, scaling to billions of objects, marginalizing over complex nuisance parameters) demand the same classes of solutions regardless of the specific probe. The fundamental challenges in the use of these techniques are equally shared between domains: *covariate shifts*, spectroscopic selection bias in photo-z, sim-to-real gaps in SNe, model misspecification limiting SBI, and domain adaptation across surveys; *UQ*, obtaining well-calibrated posteriors and propagating uncertainties to cosmological constraints; *scalability*, from billions of galaxies to 10^7 nightly alerts, demanding algorithmic and infrastructure innovations; *data sparsity and rare events*, including limited labeled samples, rare transients, class imbalance, and challenging edge cases like blending; and *metrics and evaluation*, defining task-relevant metrics, validation frameworks, and stress tests aligned with DESC science requirements. Addressing these methodologies and challenges in a general, reusable way (rather than independently within each working group) has multiplicative impact: improving SBI's robustness to covariate shifts benefits not only clusters but also photo-z, lensing, and SNe; developing robust UQ enhances reliability for deblending, shape measurement, and generative simulations simultaneously.

The shared methodological approaches and barriers documented above demand a coordinated response. DESC should *establish collaboration-wide AI/ML coordination mechanisms* (e.g., standing working group, cross-WG task forces, interchange meetings) to ensure methodological innovations are rapidly evaluated for applicability across probes, common challenges are tackled collectively, and duplication is minimized. DESC should also *invest in shared infrastructure and benchmarks* (reusable libraries, e.g., `jax-cosmo` and `JAX-GalSim`, standardized interfaces, e.g., RAIL for photo-z, and validation frameworks that stress-test methods under covariate shift and model misspecification) recognizing that these investments have multiplicative returns. DESC should *develop collaboration-wide best practices and validation standards* for ML methods intended for cosmological inference, establishing requirements for UQ calibration such as coverage tests and probability integral transform (PIT) histograms, distribution-shift diagnostics, stress tests under deliberate misspecification. Finally, mechanisms to *facilitate rapid dissemination of knowledge* (e.g., AI/ML workshops, shared tutorials, method spotlights in collaboration meetings) would accelerate the transfer of innovations across working groups.

4. Methodological Research Priorities to Advance ML for Precision Cosmology

Extracting robust cosmological constraints from LSST requires not only advanced algorithms but also a coherent methodological foundation that bridges simulation, data processing, and inference. Each of these pillars must meet unprecedented demands in scale, accuracy, and interpretability, demands that challenge the limits of both physical modeling and ML. Beyond simply applying existing AI techniques, DESC must develop methods tailored to the structure of astronomical data, the physics of observables, and the statistical rigor required for precision cosmology.

The scientific ambitions of LSST thus motivate AI/ML research in several key areas. First, Bayesian inference and UQ must evolve to handle the high-dimensional, hierarchical models that describe cosmic fields and galaxy populations, while maintaining interpretability and calibration across vast data volumes. Second, SBI and related implicit-likelihood methods must confront the challenge of model misspecification and covariate shifts, ensuring that learned posteriors remain valid when simulations imperfectly represent real observations. Third, physics-informed modeling, through differentiable programming and hybrid generative–physical architectures, offers a path toward interpretable and physically consistent deep learning, capable of representing both known and unknown components of the Universe. Fourth, discovery and anomaly detection are essential to LSST's potential for unexpected science, requiring representation learning and active human-AI collaboration to identify rare and previously unmodeled phenomena.

This section examines these research directions in detail, outlining both recent progress and outstanding challenges. We emphasize not only algorithmic innovation but also the validation, calibration, and interpretability principles required to integrate AI into cosmological analysis pipelines.

The fundamental question is: *What would convince us of a cosmological result obtained with AI?* Answering this question defines the research agenda for AI/ML in DESC and ensures that ML becomes not merely a computational shortcut, but a scientifically trustworthy component of cosmological inference.

4.1. Bayesian Inference and Uncertainty Quantification

UQ represents perhaps the most critical challenge for deploying deep learning in precision cosmology. It is an area where AI in the sciences demands solutions that differ from those in many commercial settings. Robust UQ must distinguish between aleatoric uncertainties (irreducible measurement noise) and epistemic uncertainties (model limitations and incomplete knowledge), as these have fundamentally different implications for cosmological inference and systematic error budgets. ML is poised to be revolutionary for inference, but only if current challenges are satisfactorily addressed.

4.1.1. Explicit Likelihood-Based Bayesian Inference

- **Related Methodologies.** Hierarchical Bayes, VI, Gaussian processes
- **Addresses.** UQ, Scalability

The inferential paradigm in astrophysical and cosmological data analysis has been for the past two decades primarily Bayesian, as this offers conceptual, methodological, and computational benefits (Trotta 2008). The massive increase in data size and complexity afforded by LSST will require a new step forward in inferential methodology, as LSST data will challenge the computational feasibility of current inferential engines. We cover in this section the case of *explicit inference*, where we have the ability to directly evaluate the log-likelihood of our probabilistic models, and potentially its gradients.

Accelerating Posterior Inference—MCMC methods have been the workhorses of likelihood-based Bayesian parameter inference to date. Notable examples are Gibbs sampling (Casella & George 1992) and parallelized versions of Metropolis–Hastings (e.g., the affine-invariant sampler by Goodman & Weare 2010; Foreman-Mackey et al. 2013). For cases where multimodality and/or strong parameter degeneracies are important, nested sampling (Skilling 2006; Ashton et al. 2022) in its many variants (e.g., `MultiNest` Feroz & Hobson 2008; Feroz et al. 2009; `PolyChord` Handley et al. 2015a,b; `dynesty` Speagle 2020; `DNEst4` Brewer & Foreman-Mackey 2018), recently improved with gradients (Lemos et al. 2024), accelerated with neural emulators (Lovick et al. 2025), or normalizing flows (Williams et al. 2021) have been key to ensuring reliable inference.

However, as analyses become increasingly complex, involving large numbers of nuisance parameters and expensive likelihood evaluations, the cost of running cosmological inference with conventional techniques becomes prohibitive.

One avenue to speed up inference is to leverage access to the gradients of the log-posterior. When such gradients are available, a number of inference methods can benefit from them, including the well-established HMC (Neal 2011), as well as more modern generalizations such as NUTS (Hoffman & Gelman 2014). Remarkably, physical ideas continue to lead the development of gradient-based inference techniques. Riemann Hamiltonian Monte Carlo (RHMC Girolami & Calderhead 2011) simulates trajectories in arbitrary geometries by following the geodesics of the likelihood manifold. This makes RHMC schemes extremely robust sampling algorithms for high-dimensional inference in the face of severely non-gaussian posteriors. However, RHMC has not seen a wide spread adoption due to instability of second-order auto-differentiation needed to compute the curvature of the likelihood. In a similar vein, relativity-inspired HMC schemes (Xu & Ge 2024) have recently been proposed that particularly target Minkowski geometries. This effectively introduces a maximum speed for the simulated particle, slowing it down in the areas of most challenging geometry, achieving most of the goals of RMHC without many of its hurdles. The recent Ray-tracing sampler (Behroozi 2025) take a related approach, using light refraction as the guiding analogy to steer samples toward high-likelihood regions while providing a unifying framework in which HMC, LMC, and related methods emerge as special cases. As an alternative to Hamiltonian dynamics, Langevin Monte Carlo (LMC; Rossky et al. 1978) is based on a Langevin diffusion (an SDE whose invariant distribution is the target posterior) and in practice simulates a discretization of this process to generate approximate posterior samples. In this framework, the Metropolis adjustment that ensures the target distribution is sampled can be replaced with a bias requirement on the solution of the SDE, leading to significant speed ups (Grumitt et al. 2022).

A final avenue of improvement is to revise the assumed partition function of the particles simulated by the inference algorithm. Traditional HMC schemes assume that the distribution of particles being simulated follows a canonical partition function. However, more efficient sampling schemes can be constructed by exploring other partition functions. Microcanonical or energy-conserving HMC (MCHMC, Robnik et al. 2023) explores the posterior distribution using a single energy shell. In a way similar to relativistic schemes, this is achieved by modifying the momentum of the particle to slow down at the regions of high-posterior density (Ver Steeg & Galstyan 2021), leading to a far more efficient sampling. Micro-canonical Langevin Monte Carlo (MCLMC; Robnik & Seljak 2024) is a sampling algorithm that combines all the ideas described above to great success. MCLMC already has been deployed to perform inference on physics problems such as lattice field theory simulations (Robnik & Seljak 2024) and even for cosmology where it has been shown to speed up field-level inference by an order of magnitude (Bayer et al. 2023; Simon et al. 2025). This makes MCLMC the cutting edge of gradient-based inference schemes and a promising tool to speed up analyses within DESC.

Alternatively, one can replace a complicated posterior distribution with a more tractable one. VI aims to find an approximation to the posterior distribution p by a "surrogate" parametrized distribution q_ϕ, whose parameters ϕ are trained to minimize the Kullback-Leibler divergence between q and p (see, e.g., Uzsoy et al. 2024; Campagne et al. 2023, using the JAX-powered NumPyro framework, Phan et al. 2019). Here gradients are only needed for q, not for p.

Another area of research focuses on *neural sampling methods*, which leverage in various ways neural networks to accelerate sampling while attempting to preserve asymptotic correctness guarantees. For example, normalizing flows have been used to re-parameterize the sampling space and cure complex geometries (Gabrié et al. 2022). Another recent line of research also leverages ideas from diffusion models and uses a neural score model to accelerate sampling (Havens et al. 2025).

As shown above, such inference strategies depend on differentiable components and will benefit greatly when likelihood codes are rewritten in frameworks that support automatic differentiation. An additional advantage of making probabilistic models compatible with such frameworks is that they usually support GPU acceleration and vectorization, which opens up yet another avenue for acceleration—e.g.,the Numpyro (Phan et al. 2019; Bingham et al. 2019) and BlackJax (Cabezas et al. 2024) libraries written in JAX. Affine-invariant samplers (see Foreman-Mackey et al. 2013) are particularly suited to vectorization on GPU hardware, as has been demonstrated in astronomy contexts (e.g.,Thorp et al. 2024a, 2025b, using the `affine` sampling code).

Bayesian model comparison—Estimation of the Bayesian evidence, the central quantity for model comparison, remains challenging when the models being compared are very high dimensional. Nested sampling has been established as one of the main methods for Bayesian evidence computation, but in its original formulation it suffers from the curse of dimensionality: the efficiency of the constrained sampling step decreases rapidly as the dimensionality of parameter space increases. This has been somewhat mitigated by recent developments such as `PolyChord` (Handley et al. 2015a), which can be used in a few hundreds of dimensions; `dynesty` (Speagle 2020), which uses dynamical allocation of live points (see also Higson et al. 2019); and gradient-guided nested sampling (GGNS; Lemos et al. 2024), which exploits gradients, generative flows and differentiable programming to achieve better efficiency and accuracy in up to ~ 200 dimensions. A suite of other methods for the evaluation of the high-dimensional average of the likelihood

over the prior are also being explored, sometimes combining density estimation with neural techniques (e.g., Heavens et al. 2017; McEwen et al. 2021; Srinivasan et al. 2024). However, they remain confined to moderately low-dimensional parameter spaces, of order a few tens of dimensions.

The frontier represented by evidence estimation in very large dimensional (of order 10^3 or more) parameter spaces from real data remains largely untouched outside of synthetic demonstration examples where the ground truth is known. NRE shows promise in this respect, in that evidence estimation can be obtained from an NRE architecture by adding a suitable inferential head that is trained only on model labels, thus implicitly marginalizing over all parameters in the model. Such an approach naturally also generalizes to performing Bayesian model averaging. An example of this method is Karchev et al. (2023a), where six models for empirical corrections for SN Ia data are compared from Carnegie Supernova Project (CSP) observations (Krisciunas et al. 2017) within a Bayesian hierarchical model setting with $\sim 4,000$ latent variables.

Hierarchical Bayesian Models in Extremely High Dimensions—The manyfold increase in data size requires in many cases a more sophisticated model to capture previously unimportant effects; this in turn increases the dimensionality of the parameter space (especially in hierarchical models, where the latent space dimensionality scales with the number of objects within the model); the likelihood might become intractable, or previously used approximations, such as approximate Gaussianity or linear propagation of errors (Karchev et al. 2023b), neglecting of Eddington bias (Karchev & Trotta 2025), might break down.

Section 3.3 introduced so-called full-field inference for cosmological surveys, in which not only cosmological parameters are inferred but also the initial conditions that seed the evolution of the large-scale structure of the Universe (Porqueres et al. 2023). The fidelity of the forward simulation directly controls the accuracy of posterior constraints on cosmological parameters; consequently, this approach requires exploring extremely high-dimensional parameter spaces (millions to billions of parameters). Sampling such spaces is intractable for traditional MCMC and instead calls for gradient-based methods, as noted above. This, in turn, demands a forward simulation that is both fast and differentiable (see Section 3.7) to make full-field inference at LSST scale attainable.

It is worth noting that such hierarchical full-field inference models are substantially more computationally expensive than alternative SBI methods (see next section 4.1.2), but offer several advantages. First, analyzing statistical errors directly in data space is more interpretable than working with the compressed summary statistics typical of SBI workflows; even with optimal compression, signals can mix and model misspecification becomes difficult to detect. Here, systematic contamination can be treated as additional parameters to be sampled (Porqueres et al. 2019), becoming a machine-aided report of contaminations that have a characteristic pattern on the sky. Second, hierarchical Bayesian inference is designed for extensible, modular models in which new physics can be added—e.g., augmenting the simulation with a baryonification model—whereas SBI would require retraining neural density estimators from scratch. Taken together, these properties make hierarchical Bayesian inference well suited to joint inference of cosmology, systematics, and redshift-distribution uncertainties—capabilities that are considerably more difficult with implicit approaches. Additionally, hierarchical inference provides a digital twin of the Universe, which has multiple scientific applications but also provides a unique way of testing the results by cross-validating with independent data (Stopyra et al. 2024).

4.1.2. Implicit Likelihood Bayesian Posterior Inference

- **Related Methodologies.** SBI, NPE, Normalizing flows
- **Addresses.** UQ, Scalability

The other paradigm is implicit inference, in which we do not assume direct access to the likelihood function, but only have access to samples from the joint distribution $p(x, \theta)$ of data samples x and parameters of interest θ. It should be noted that this situation covers both the case of SBI, and the case where x and θ are available from a training sample of real observations (the canonical example being photo-z estimation from a set of spectroscopic observations).

In particular, SBI is rapidly emerging as a powerful alternative to traditional fitting techniques for Bayesian models. The key idea is to replace an explicit likelihood function by forward simulating (under the model) parameter-data pairs, which are then used to train a neural network to perform inference (e.g., Alsing et al. 2018; Alsing & Wandelt 2018, 2019; Savchenko et al. 2024; Lyu et al. 2025). The advantages are that the (potentially intractable) likelihood can, in principle, incorporate physical effects of arbitrary complexity, which would otherwise be difficult to model (including, e.g., selection effects and complex parameter dependencies). In some variants (Miller et al. 2021) the 1- or 2-dimensional *marginal* distribution for the parameters of interest is targeted directly, thus circumventing the need to evaluate the high-dimensional joint posterior over all parameters in the model; such approaches are naturally suited to Bayesian evidence estimation (Karchev et al. 2023a).

Additionally, inference can be *amortized* within a certain prior range, meaning that once trained the network can deliver almost instantaneous posteriors for a wide range of parameter values, a critical benefit when dealing with billions of galaxies (Hahn & Melchior 2022). This speed-up also permits posterior calibration methods (e.g., guaranteed coverage), which are computationally unfeasible with traditional posterior evaluation methods.

Neural Density Estimation (NDE) methods—The fundamental building blocks of these methods is NDE, where a neural network is used to estimate a distribution, or a ratio of distributions. Various kinds of methods exist: NLE (e.g., Papamakarios et al. 2019; Lueckmann et al. 2019; Alsing et al. 2019), NPE (e.g., Papamakarios & Murray 2016; Lueckmann et al. 2017) and NRE are among the most popular (for an overview see Alsing et al. 2019; Cranmer et al. 2020; Lueckmann et al. 2021). Implementations of NLE and NPE both learn a density based on simulated parameter-data pairs (see, e.g., Alsing et al. 2019), with a variety of different approaches used for learning the multivariate joint or conditional density. Approaches to this include Gaussian mixtures (e.g., Alsing et al. 2018), mixture density networks (e.g., Papamakarios & Murray 2016), and normalizing flows (e.g., Papamakarios et al. 2017, 2019; Alsing et al. 2019; Jeffrey et al. 2021; Hahn & Melchior 2022; for a review see Kobyzev et al. 2021; Papamakarios et al. 2021). More sophisticated density estimators used in generative modeling – such as continuous normalizing flows (Grathwohl et al. 2018; Chen et al. 2018), score-based diffusion models (Song et al. 2020), flow-matching models (Lipman et al. 2022), and transformers (Vaswani et al. 2017) – are also well suited to NLE and NPE tasks (e.g., Diaz Rivero & Dvorkin 2020; Geffner et al. 2023; Wildberger et al. 2023; Gloeckler et al. 2024)

alongside the generative modelling tasks they are commonly used for (e.g., Alsing et al. 2024; Cuesta-Lazaro & Mishra-Sharma 2024; Thorp et al. 2025b).

Optimal Neural Summarization—To ensure the robustness of implicit inference, the process is usually divided into two main steps enabling each neural network to focus on a specific task: (1) compression of high-dimensional data into informative summary statistics, and (2) performing Bayesian inference using neural density estimation methods on this low dimensional but highly informative statistic. To maximize information extraction and improve constraints on cosmological parameters, the community has increasingly adopted neural network–based summarization techniques. While any neural network can be trained on the regression task of inferring parameters given data (e.g., Gupta et al. 2018; Kacprzak & Fluri 2022; Lu et al. 2023), it is unclear how much of the information contained in the data is extracted. In particular, Lanzieri et al. (2025) demonstrate that standard regression loss functions do not guarantee the systematic construction of sufficient statistics. Information-maximizing neural networks (IMNNs; Charnock et al. 2018) directly address this problem by learning summary functions that maximize the Fisher information. They can produce nearly exact posteriors and are thus approximately sufficient statistics of the data. Another approach is to derive a loss function directly from the definition of sufficiency, i.e., by maximizing the mutual information between the summary statistics and parameters of interest (Jeffrey et al. 2021; Chen et al. 2021).

Controlling Epistemic Errors in Inference Results—One fundamental limitation of NDE methods is that their reliance on a neural network to model at some level the likelihood of the data is inherently imperfect. In the asymptotic regime of infinite data and flexible neural network, the approximation to the target posterior will converge, but in practice we are never guaranteed to find ourselves in this regime, and must therefore take into account and mitigate *epistemic errors*. Several strategies have been developed over the years to quantify and mitigate this epistemic uncertainty on inference results. MCMC sampling over network parameters provides gold-standard uncertainty estimates but at usually prohibitive computational cost (e.g. Behroozi 2025). In addition, detecting convergence of the chain remains difficult, usually necessitating drawing more samples than ultimately needed. Because such approach is extremely expensive, other approaches have been developed. *BNNs* approximate the posterior distribution over network parameters through VI, providing principled uncertainty estimates at reduced computational cost. However, the tradeoff between approximation quality and speed remains concerning, especially in the highly multi-modal loss landscapes of deep neural networks. In such settings, scalable variational methods often collapse to a single mode of the posterior rather than exploring the full diversity of solutions (Fort et al. 2019), which limits the quality of their uncertainty estimates. One of the most used VI methods in astronomy is Monte Carlo dropout which utilizes the dropout layer commonly introduced in deep neural networks to prevent correlated activation as one of the computationally cheapest approximations to Bayesian inference; however, its theoretical justification and empirical accuracy are questionable (Le Folgoc et al. 2021). Nonetheless, comparisons with other methods have shown promise for astronomical applications: strong lensing (Perreault Levasseur et al. 2017), supernova time-series classification (Möller & de Boissière 2020, 2022), and star time-series classification (Donoso-Oliva et al. 2023; Cádiz-Leyton et al. 2025b,a). Other VI methods such as Bayes by Backprop and SWAG have been sparsely used for time-series classification and regression with mixed results (Möller & de Boissière 2020; Cranmer et al. 2021). Another strategy is Deep Ensembles in which multiple networks are trained from different random initializations to provide uncertainty estimates via the variance of their predictions (Makinen et al. 2021; Möller et al. 2022, 2024). Unlike variational methods,

ensembles capture uncertainty by effectively sampling from different modes of the loss landscape, resulting in more robust and better-calibrated uncertainty estimates. However, they are more computationally demanding than MC dropout, and do not mitigate errors arising from model misspecification. Comparative studies exploring the trade-offs between these UQ methodologies for achieving superior uncertainty evaluation are a growing focus in the field (Cádiz-Leyton et al. 2025b).

4.1.3. Model Mispecification and Covariate Shifts

- **Related Methodologies.** SBI
- **Addresses.** Covariate shifts

From a technical standpoint, SBI has achieved impressive results: NPE with normalizing flows performs well in low-dimensional regimes (e.g., Srinivasan et al. 2025), NLE scales satisfactorily to higher dimensions, and marginal NRE has shown success across diverse applications (e.g., Alvey et al. 2024; List et al. 2023; Franco-Abellán et al. 2024; Saxena et al. 2024). Yet, these demonstrations rely primarily on simulated data and therefore represent best-case scenarios; direct validations of SBI on real data remain scarce (Karchev et al. 2024; Lüber et al. 2025).

The robustness of SBI depends critically on the realism and completeness of the simulations that underpin it. Simulations must reproduce all relevant aspects of the observations, including astrophysical, instrumental, and observational effects. Any unmodeled process leads to domain shift or model misspecification, which can severely bias inference. This is particularly problematic for NRE, which relies on accurate joint modeling of data and parameters (Filipp et al. 2025), while NLE is comparatively more interpretable since it operates directly in data space. Even when the theoretical model is sound, the observational and noise models must be equally faithful—a condition often unmet given the traditional divide between theoretical and observational cosmology. Bridging this gap is essential for SBI to succeed. Efforts are underway to diagnose and quantify model misspecification through simulation-based calibration and related approaches (Talts et al. 2018; Lemos et al. 2023; Chen et al. 2025; Anau Montel et al. 2025; Kelly et al. 2025). While model misspecification is a key vulnerability of SBI, it is not fundamentally different (if more difficult to diagnose and cure) than the similar risk incurred when using explicit, likelihood-based models: missing components of the model w.r.t. the true data-generating process will lead to potentially severe bias in the resulting inference. SBI is a relatively new technique, and therefore appropriate diagnostic tools are still being developed to ensure its robustness and reliability.

Training Set Representativity—A fundamental challenge across deep learning applications in astronomy is the representativity of training data. Models trained on simulations may fail to generalize to real observations, while those trained on current surveys may struggle with the deeper, higher-resolution LSST data. Techniques such as fake source injection—embedding simulated objects into real images—can mitigate these gaps (Suchyta et al. 2016; Huang et al. 2018), though their success depends on how realistic the injected sources are. The problem is particularly acute for rare phenomena, where training examples are intrinsically limited. To improve generalization, domain adaptation, transfer learning, and hierarchical Bayesian methods are being explored. E.g., Swierc et al. (2024) use domain adaptation to obtain more

robust data summaries that can generalize well between simulated data and mock observations, enabling more accurate neural density estimation. Principled approaches such as stratified learning (discussed in Section 3.1) can mitigate covariate shift with little modification in the learning procedure, but other methods often require substantial experimentation and modifications of training procedures (Krishnaraj et al. 2025). Such corrections must themselves be treated as part of the inference pipeline and undergo rigorous calibration and uncertainty quantification.

Physics Hardening—When available datasets are incomplete or non-representative, physics-informed augmentation can enhance robustness. For example, the DESC ELAsTiCC challenge (Knop & ELAsTiCC Team 2023) injected transients simulated using semi-analytic models (e.g., SNe, KNe) to make classifiers more resilient to underrepresented classes, and Moskowitz et al. (2024) augmented spectroscopically-incomplete training samples with simulated photometry to improve photometric redshift estimation. Latent representations derived from SPS models can also be used to generate synthetic photometry for missing or incomplete observations (e.g., `pop-cosmos` Alsing et al. 2024; Thorp et al. 2025b; Deger et al. 2025), and facilitate comparisons with hydrodynamical simulations without observational systematics. However, these methods inherit the assumptions and uncertainties of the underlying theoretical models—such as uncertain nebular emission strengths in SPS (Byler et al. 2017; Li et al. 2025c; Morisset et al. 2025; Newman et al. 2025)—which can themselves introduce model misspecification (Leistedt et al. 2023; Jespersen et al. 2025). Addressing these limitations requires deeper astrophysical modeling of galaxy formation and evolution, as well as diagnostic tools for identifying misspecification in high-dimensional generative models (e.g., Thorp et al. 2025a). Because physics-informed generative models (e.g., those that capture information within SPS parameterizations) can be used to synthesize observables that the model has not been trained on, such models can be validated not only against unseen data from a test set but also on new types of observations and other surveys (e.g., Alsing et al. 2023, 2024; Thorp et al. 2024a, 2025b; Deger et al. 2025).

4.1.4. Validating Inference Results

- **Addresses.** UQ, Metrics

While the quality of neural posteriors can be tested (Talts et al. 2018; Lemos et al. 2023), and while statistical tests can be performed to determine the probability that the distribution learned by a generative model is identical to that of the training data (Lemos et al. 2024), an open issue is the determination and procurement of a sufficient volume of training data for those tests to be sufficiently sensitive and for the learned distribution to be accurate.

Models trained on the same data but with different algorithms exhibit distinct probability calibration characteristics that must be evaluated and corrected. Similarly, identical algorithms trained on different training sets require independent calibration assessment. Common diagnostics include reliability diagrams used in time-series classification (Möller & de Boissière 2020) or non-conformity scores from conformal inference techniques (Xie et al. 2025), both of which compare predicted probabilities against observed frequencies. Detection of anomalies, i.e., the classification whether a signal is anomalous enough to be reported, is

particularly vulnerable because it probes the tails of a learned distribution. For regression tasks, calibration ensures that predicted uncertainties accurately capture the true error distribution. Poorly calibrated uncertainties can introduce systematic biases in downstream cosmological analyses, leading to incorrect parameter constraints. Recalibration methods for lens modeling are presented by, e.g., Perreault Levasseur et al. (2017), Karchev et al. (2022b), and Gentile et al. (2023). Regarding generative models, Campagne (2025) propose a "two-models" framework to evaluate their statistical consistencies trained on independent subsets of galaxy images. The results emphasize the need for large-enough datasets to enable calibration and validation strategies specific to each generative architecture (e.g., generative adversarial networks, normalizing flows and score-based diffusion), since apparent visual quality and morphological variable distributions alone do not guarantee statistical reliability.

4.2. Physics-Informed Approaches

From a high-level point of view, neural networks are never perfectly trustworthy and are often hardly interpretable. This motivates a general desire to build *physics-informed* models, which can leverage as many explicit physical constraints as possible, thus limiting the potential failure modes of AI components.

4.2.1. Hybridization of Generative Modeling and Physical Models

· **Related Methodologies.** Normalizing flows, Diffusion models, Neural surrogates, Differentiable programming
· **Addresses.** Covariate shifts, Scalability

A promising direction for scientific inference is the hybridization of explicit, physics-based models with generative components that flexibly represent unknown or intractable distributions. In this framework, flow-based or diffusion models serve as probabilistic priors over complex latent variables, such as galaxy morphology or small-scale baryonic processes, while the rest of the model remains physically interpretable and simulation-driven. These generative priors have already proven powerful in astronomical inference, e.g., in the estimation of galaxy properties and photometric redshifts at scale (e.g., `pop-cosmos`; Alsing et al. 2024; Thorp et al. 2024a, 2025b; Deger et al. 2025), and in the generation of high-fidelity, field-level H I maps from dark matter simulations (Mishra et al. 2025). More broadly, generative models naturally support amortized inference frameworks, where neural posterior estimators are trained on samples from the generative process, enabling accurate Bayesian inference without MCMC sampling but at the cost of greater upfront training effort.

Differentiable Programming—To fully integrate such probabilistic components with physical simulations, modern astrophysical codes are increasingly being reimplemented in automatic-differentiation libraries. Differentiable simulators eliminate the approximation errors of emulators and provide exact gradient information for optimization and uncertainty quantification. Examples include differentiable particle-mesh cosmological solvers (Modi et al. 2021; List et al. 2025), theoretical cosmology computations in `jax-cosmo` (Campagne et al. 2023), galaxy–halo connection models in `Diffsky`/`Diffstar` (Alarcon et al. 2023), stellar pop-

ulation synthesis in DSPS (Hearin et al. 2023), halo-model calculations in halox (Kéruzoré 2025), and differentiable image simulations in JAX-GalSim (GalSim Developers 2025).

While the widely used GalSim library (Rowe et al. 2015) produces realistic galaxy images, its non-differentiable design limits efficient gradient-based inference. In contrast, the emerging JAX-GalSim library reimplements core GalSim functionalities in JAX, yielding GPU-accelerated, fully differentiable forward models that enable direct gradient computation for both population-level and individual-level parameters. Similarly, scarlet2 offers a JAX-based, differentiable framework for non-parametric source morphologies and blended scenes observed by multiple instruments. Both libraries support vectorized batch simulations, crucial for large-scale hierarchical inference, and allow gradients of differentiable likelihoods to be computed automatically for maximum-likelihood estimation, variational inference, or gradient-based MCMC.

High-Dimensional Inverse Problems with Explicit Likelihood and Data-Driven Priors—High-dimensional inverse problems are central to cosmology, from galaxy deblending and strong-lensing source reconstruction to recovering the dark matter field from noisy data. In these settings, the forward process, such as instrumental response, noise model, or lensing distortion, is well understood and can be encoded in an explicit likelihood. The underlying components, however, such as galaxy morphologies or the non-Gaussian dark matter structure, lack closed-form descriptions and require expressive statistical models. Generative models, such as diffusion models, can learn realistic priors from high-dimensional observations or simulations. Combining such data-driven priors with explicit likelihoods yields a principled framework: the prior enforces realistic structure, while the likelihood anchors inference to the data, even under low SNR conditions where fully amortized approaches may drift. Recent works have demonstrated this hybrid approach for galaxy source reconstruction, strong lens modeling, and superresolution (Adam et al. 2022; Barco et al. 2025b; Adam et al. 2025) and dark matter field inference (Remy et al. 2023). Moreover, the presence of an explicit likelihood enables learning data-driven priors directly from observations, through iterative refinement using posterior samples (Rozet et al. 2024; Barco et al. 2025a), and can even allow for the correction of model misspecification (Payot et al. 2025). A remaining challenge is efficient posterior sampling, as inference with diffusion priors entails solving ordinary differential equations (ODEs), which is computationally demanding, although it can be performed practically at scale (see, e.g., Thorp et al. 2024a, 2025b).

4.2.2. Imposing Consistency with Physical Equations and Symmetries

- **Related Methodologies.** Physics-informed ML
- **Addresses.** Covariate shifts, Metrics

For trustworthy results, we additionally demand that the AI/ML outputs at various stages of the pipeline satisfy null tests (e.g., B-modes in gravitational lensing or rho-statistics for PSF modeling, Rowe 2010) or obey the laws of physics of the corresponding analysis component rather than merely report final results with high fidelity. The modularity of the pipelines and multi-scale nature of the phenomena asks for validations at every analysis stage. Crucially, our knowledge of physical relations (in the universe, in the atmosphere, in the instrument) permits a form of validation that is typically omitted or impossible in industry applications of AI/ML. This motivates research in areas such as invariant/equivariant representation learning and geometric

learning, with possible interdisciplinary implications beyond the scope of DESC (Ferguson et al. 2025). Furthermore, the use of symmetry-aware equivariant neural networks (ENNs) could help with the extraction of more robustness features from the data and, in combination with domain adaptation, enable easier mitigation of covariate shifts (Pandya et al. 2025a). E.g., with physics-informed neural networks (PINNs) one can, in principle, find solutions for explicitly specified differential equations if their optimization could be made more robust (Rathore et al. 2024).

4.3. Novelty Detection and Discovery

- **Related Methodologies.** Anomaly detection, Self-supervised learning, Active learning
- **Addresses.** Covariate shifts, Data sparsity

A central scientific promise of LSST lies in its potential for unexpected discovery. Many now-fundamental astrophysical phenomena, such as strong lenses, fast radio bursts, and pulsars, as well as singular systems like the Bullet Cluster, were first identified as anomalies. With an anticipated catalog exceeding 20 billion galaxies and roughly 10 million alerts per night, however, detecting novel phenomena in LSST data represents an unprecedented challenge.

Generative models offer a powerful framework for unsupervised discovery by learning the ensemble properties of galaxy images, spectra, photometry, and time-domain behavior, and by enabling the detection of statistically anomalous signals without labeled training data (Liang et al. 2023). In time-series and high-energy astronomy, representation learning has already proven effective for discovery-oriented analyses (Dillmann et al. 2025; Song et al. 2025). However, a common failure mode of generative models, where atypical signals appear highly typical (Nalisnick et al. 2018), will mean that true outliers may not be recognized, a loss if we seek to find them (e.g., rare SN types such as pair instability SNe) and a problem if they contaminate carefully selected samples used in high-precision cosmology (e.g.,unrecognized blends in shape catalogs, Dawson et al. 2016; or catastrophic outliers in photo-z estimates). More specifically, standard unsupervised methods often struggle in the dense and homogeneous latent spaces produced by deep representations (e.g., Baron Perez et al. 2025; Li et al. 2025a). E.g., while Etsebeth et al. (2024) successfully combined Zoobot (a foundation model discussed in Section 5.1) features with the `Astronomaly` framework (Lochner & Bassett 2021) to identify new sources, Walmsley et al. (2022a) found that tailored anomaly-detection techniques were necessary even within the well-studied Galaxy Zoo dataset. This line of work culminated in `Astronomaly Protege` (Lochner & Rudnick 2025), a general-purpose, active anomaly detection system optimized for exploration in deep latent spaces.

The emergence of deep learning and foundation models (Section 5.1) further elevates the importance of active learning, the tight integration of human expertise and machine-driven pattern recognition (Lochner & Rudnick 2025). Self-supervised methods and foundation models promise the ability to generate rich, general-purpose representations, but expert oversight remains essential to interpret their outputs and assess scientific relevance. Automated systems may flag outliers or cluster data effectively, yet human judgment is required to determine which patterns constitute genuine discovery. As AI systems evolve toward agentic operation (Section 5.2), the collaboration between human and machine will become increasingly in-

tertwined. Embedding active-learning capabilities directly within AI infrastructures will therefore be critical to enable rapid, scalable, and participatory scientific discovery—bridging expert analysis and citizen science within the LSST era.

5. Emerging Techniques

The priorities outlined in Section 4 establish the statistical and algorithmic foundations for ML-enabled cosmology, but realizing these priorities at the scale and complexity of LSST demands new approaches. Training distinct models for each science case is neither computationally sustainable nor conducive to the cross-probe consistency required for joint analyses.

This section surveys two classes of emerging AI techniques that may fundamentally change how DESC builds and maintains its analysis infrastructure: data FMs (Section 5.1) and LLM-based agentic systems (Section 5.2). By building general-purpose and reusable models from large unlabeled datasets, data FMs offer the potential for a shared backbone ML infrastructure that can be applied across science cases, greatly reducing time to science for analyses involving ML. LLMs and multi-agent systemss (MASs) target a different bottleneck: the coordination of complex workflows, synthesis of documentation and literature, and accessibility of the DESC software infrastructure to new collaboration members. Together, these techniques provide a framework for scaling the ML efforts in DESC from individual analyses to reproducible, collaboration-wide pipelines.

5.1. Data Foundation Models

FMs, AI systems trained on massive datasets to perform a broad spectrum of tasks (Bommasani et al. 2021), have not only revolutionized AI research but are also rapidly reshaping modern life. Vision FMs are now enabling breakthroughs in robotics (Di Palo & Johns 2024), medical diagnostics (Ma et al. 2024), and remote sensing (Liu et al. 2024). Beyond these applications, their adoption in scientific disciplines such as genetics (Brixi et al. 2025) and heliophysics (Roy et al. 2025) has highlighted their powerful predictive capabilities and their capacity to uncover fundamental relationships in complex data. The burgeoning field of FMs presents a significant opportunity to enable and accelerate cutting-edge astrophysics.

Zoobot (Walmsley et al. 2022b) can be considered the first vision FM in optical astronomy. Having been trained on labels from the Galaxy Zoo citizen science project (Lintott et al. 2008, 2011; Willett et al. 2013), it has demonstrated versatility on a range of downstream tasks. These include broad morphological classification problems critical for studies of galaxy evolution, such as identifying merging galaxies (e.g., Omori et al. 2023; Margalef-Bentabol et al. 2024b; de Graaff et al. 2025), and anomaly detection for finding rare phenomena like strong lenses (Euclid Collaboration: Walmsley et al. 2025; Euclid Collaboration: Lines et al. 2025). Formally released in Walmsley et al. (2023), Zoobot has been adapted to imaging data from multiple surveys, including DESI (Dey et al. 2019), *Euclid* (Euclid Collaboration: Aussel et al. 2024), *HST* (O'Ryan et al. 2023), and *James Webb Space Telescope* (*JWST*) simulations (Margalef-Bentabol et al. 2024a), among others (Holwerda et al. 2024; Heesters et al. 2025; Omori et al. 2025).

FMs have also been designed for astronomical time-series datasets, spurred by the need for automated photometric classification of Galactic and extragalactic transients in the Rubin era. Some examples of these models include Astromer (Donoso-Oliva et al. 2023, 2025), AstroCo (Tan et al. 2025), ATAT (Cabrera-Vives et al. 2024), FALCO (Zuo et al. 2026), and RoMaE (Zivanovic et al. 2025). While the primary requirement for these models is high classification accuracy, they also enable the discovery of new classes of transients

through anomaly detection, and, in some cases, provide lightcurve interpolation for further downstream analysis.

Despite early successes in modality-specific FMs, many astrophysical questions can only be answered by fusing different data types. This is a task for which traditional methods, which rely on reducing data to summary statistics, are quickly being outpaced by powerful multimodal FMs. Though the ideal neural architectures and training objectives for these multi-modal models are areas of active research, models have now been developed for galaxies (Parker et al. 2024), SNe (Zhang et al. 2024), variable and non-variable stars (Leung & Bovy 2024; Rizhko & Bloom 2025), and cosmological simulations (Xia et al. 2025). AION-1 (Parker et al. 2025) represents a significant step toward survey-scale multimodal FMs: trained on over 200 million observations of both stars and galaxies from five major surveys, it integrates images, spectra, and scalar measurements into a billion-parameter model. AION-1 demonstrates strong performance in low-data regimes, enables zero-shot detection of rare objects such as strong gravitational lenses, and produces survey-invariant representations that facilitate knowledge transfer across telescopes.

5.1.1. Foundation Models for DESC Science

Pre-training and reusability—FMs offer a significant advantage over existing techniques by reducing heterogeneous data types into a unified and simplified numerical representation known as a latent space, representation or feature space. Instead of reprocessing a full dataset for each analysis, a single, powerful FM can generate rich data representations once. These representations can then be used directly or rapidly finetuned for numerous specific science cases, resulting in significant savings of computing resources.

This shared representational basis also changes how DESC can approach cosmological inference. Traditional analyses reduce complex image data to limited summary statistics (e.g., moments, colors, or flux ratios), inevitably discarding information. Deep learning enables direct inference from raw observations, but training bespoke networks for each task across the petabyte-scale archive of LSST is infeasible. A general-purpose FM, trained once at scale, can thus act as a reusable "backbone" for all DESC pipelines, propagating consistent representations across tasks.

For time-domain astronomy, the representations produced by FMs are especially powerful. Because models learn features directly from the data, they are not constrained by the a priori assumptions of human-engineered features. This makes them ideal for identifying new or unexpected classes of objects via anomaly and novelty detection. Furthermore, these representations could be highly effective for standard tasks such as early transient classification, which is critical for triggering spectroscopic follow-up.

For simulation-based inference (Section 4.1.2), FM latent spaces can act as highly flexible encoders, providing a more powerful data compression than traditional statistical summaries. In terms of data handling, multimodal models can naturally accommodate missing data when fusing datasets. Moreover, the ability to pre-train on vast unlabeled datasets enables FMs to address the representational bias (dataset shift) often encountered between the training and test sets in supervised learning.

Opportunities offered by multimodality—In astronomy, multi-wavelength models learn a shared latent space from multiple observational wavelengths (e.g., optical and infrared) of the same object. Multimodal

models extend this concept by integrating fundamentally different data types into a shared latent space, often through self-supervised learning techniques. While the specific AI architectures for combining heterogeneous datasets remain in active development, all approaches fundamentally treat different data types as complementary views of the same underlying astrophysical system. The resulting multimodal representations provide a computationally efficient and powerful framework for a broad spectrum of scientific analyses.

In the time domain, joint modeling of photometric light curves and spectra provides a direct path to improving supernova cosmology and our understanding of explosion physics. A multimodal FM can perform cross-modal imputation, predicting the spectral properties of SNe Ia from irregular photometric sequences alone (as has been demonstrated on synthetic data; Shen & Gagliano 2025a,b), thereby recovering physically meaningful features (line velocities or continuum temperatures) that are otherwise inaccessible from broadband imaging data. These inferred spectra can serve as additional standardization parameters for SNe Ia, potentially reducing residuals in the Hubble diagram by incorporating information linked to, e.g., progenitor diversity (Son et al. 2025) or host metallicity (Childress et al. 2013).

Beyond improving standardization, the same cross-modal embeddings enable proactive discovery. By comparing inferred to obtained spectra, outliers can be flagged for long-term monitoring. When combined with host-galaxy properties across the transient samples discovered by Rubin LSST, these models can provide a probabilistic mapping between host galaxy and transient properties, useful for exploring population-level correlations potentially linked to supernova physics and for obtaining sub-populations of highly-standardizable SNe Ia.

Challenges—Despite rapid progress, technical and practical challenges must be addressed before multimodal FMs can be fully integrated into DESC pipelines. Propagating observational uncertainties to all studies conducted downstream of a DESC-wide FM is critical, and the diversity of applications across spatial and temporal scales may not be well matched to a single architecture's inductive biases. Multiple application areas also require capturing the impact of selection effects in extant training samples, which may cause a model to over-weight well-observed populations and under-represent rare but scientifically informative objects. A related challenge is ensuring that learned embeddings disentangle instrumental systematics (e.g., tracking issues, bright sky backgrounds) from true astrophysical signal. Architectures that explicitly factorize instrumental and astrophysical contributions to observed data offer a promising avenue (e.g., Audenaert et al. 2025), and further development of such approaches will be essential for meeting DESC's calibration requirements.

5.1.2. Training Objectives

Training objectives determine whether FMs learn astrophysically meaningful structure or merely reproduce observational correlations. For DESC, these objectives must explicitly promote representations that encode physical invariants (e.g., morphology–redshift relations, color–temperature gradients) while remaining robust to the observational systematics and covariate shifts defined as calibratable in the Science Requirements Document. Self-supervised Learning (SSL) offers the most practical and scalable route toward this goal, as it enables the extraction of representations from vast unlabeled datasets without compromising the requirement for bias quantification and calibration in DESC.

Self-supervised learning—SSL encompasses a range of approaches: reconstructive methods, such as autoencoders (Hinton & Zemel 1993; Hinton & Salakhutdinov 2006), learn to compress data into a low-dimensional representation and then reconstruct the original input; contrastive learning (Chen et al. 2020; Tian et al. 2021) trains models to produce invariant representations for augmented versions of the same data point (e.g., zoomed or rotated); and predictive methods, often implemented via transformer architectures (Vaswani et al. 2017), learn by predicting masked or omitted sections of data based on their surrounding context. The principal advantage of SSL is its scalability, allowing models to be trained on vast datasets without costly manual annotation.

Generative approaches—Beyond these paradigms, generative and diffusion-based models are also being adapted for self-supervised representation learning in the astronomical domain. Although originally designed for data synthesis, diffusion objectives (Yang et al. 2023) can act as powerful denoisers and uncertainty estimators, aligning with the DESC requirement that calibratable systematic errors remain subdominant to statistical uncertainties. How to manage a good generative model in the context of galaxy image synthesis has been investigated for instance using this denoising capability of score-based diffusion models (Campagne 2025). Hybrid diffusion autoencoders (Preechakul et al. 2021) combine reconstructive and generative losses, yielding latent spaces that capture astrophysical variation while marginalizing over observational noise. Methods such as these will be critical for DESC to achieve unbiased shear and photometric-redshift inference with Rubin data.

Unsupervised learning considerations—As highlighted in Section 4.3, the vast LSST dataset will hold immense potential for scientific discoveries. Given this sheer scale, FMs will be critical for creating powerful representations that enable anomaly detection, unsupervised source classification, and similarity searches. However, a significant challenge remains: FMs are typically designed and optimized for supervised tasks, while their use for unsupervised applications is often an afterthought. Recent work by Walmsley et al. (2022a) and Lochner & Rudnick (2025) demonstrates this gap. They found that traditional anomaly detection methods fail when applied to the deep latent features learned by both supervised and self-supervised methods. This indicates a clear need for new research: new unsupervised methods compatible with these features must be developed (such as `Astronomaly Protege`; Lochner & Rudnick 2025) and FMs must be optimized specifically for unsupervised discovery.

5.1.3. Architectural Innovations

Realizing the capabilities of FMs for DESC science requires architectural and training advancements. Astronomical data presents unique barriers to large-scale training, including wide dynamic ranges from faint to bright objects, Poisson noise properties, multi-wavelength observations, and irregular sampling. These characteristics, combined with the science priorities of DESC, necessitate architectures that not only achieve optimal representational power on LSST data but also permit robust uncertainty propagation.

Attention—Efficient attention mechanisms, which allow models to weigh the importance of different parts of the input data, have evolved significantly beyond standard transformers, offering new architectures that can scale to LSST-level data volumes. Methods like Flash Attention (Dao et al. 2022) use techniques such as tiling and kernel fusion to reduce the memory footprint by 10-–100$\times$. This significant reduction for the first

time makes practical application of attention mechanisms at the scale of individual charge-coupled devices (CCDs) or rafts.

Hierarchical Approaches—Astronomical data is inherently hierarchical, with structures ranging from individual galaxies to massive clusters. Sparse attention models like Longformer (Beltagy et al. 2020) and BigBird (Zaheer et al. 2020) are well-suited to this, as their architecture directly mirrors this physical structure. They use local attention (e.g., sliding windows) to model interactions between nearby objects, while global tokens aggregate information about the entire system. These designs are critical for DESC, as they enable the joint modeling of small- and large-scale correlations essential for controlling systematics in shear, clustering, and supernova analyses. This hierarchical approach is well-developed in vision models, with hierarchical vision transformers (e.g., Swin Transformer V2, Liu et al. 2022) being especially promising for DESC imaging. Unlike the token patterns in sparse models, these architectures use multi-resolution attention windows that mirror the multi-scale nature of cosmological information, enabling pixel-level PSF modeling up to large-scale galaxy clustering. This design could enable the unification of weak lensing and large-scale structure analyses using a shared image encoder. Such a model would facilitate end-to-end uncertainty propagation across spatial scales, directly addressing the DESC requirement for cross-probe consistency in systematic control.

Mixture-of-Experts—The principle of a shared, unifying backbone extends to mixture-of-experts (MoE) architectures, which offer natural alignment with DESC objectives. Rather than training independent networks for each object class or redshift regime, sparse MoE layers – such as the Switch Transformer (Fedus et al. 2022) and Mixtral (Jiang et al. 2024) – can dynamically route inputs through specialized sub-networks while preserving a common latent backbone. This paradigm mirrors the DESC software model itself: probe-specific inference modules built atop a common analysis infrastructure. E.g., expert sub-networks could specialize in quiescent versus star-forming galaxies or early- versus late-time transients, while shared latent features will ensure cross-consistency in calibration and selection functions across working groups.

Data Fusion—A final architectural requirement is multimodal data fusion, driven by the operational need to combine LSST data with a continuous stream of ancillary datasets (e.g., *Roman*, DESI, 4MOST/TiDES) as well as managing evolving observing conditions. This challenge can be addressed with early-fusion models, which tokenize heterogeneous data types into a common representation space for joint training (e.g., 4M, Chameleon; Mizrahi et al. 2023; Chameleon Team 2024); AION-1 (Parker et al. 2025) exemplifies this approach in astronomy, using modality-specific tokenizers to convert images, spectra, and scalar measurements into discrete tokens before unified transformer-based masked modeling. Alternatively, late-fusion approaches merge specialized encoders post hoc (Pereira et al. 2023). More generalized architectures, like the Perceiver family (Jaegle et al. 2021a,b), are particularly advantageous. By learning a single, compressed latent array, they are explicitly designed to flexibly accommodate new data modalities. This provides a clear operational path for DESC to update and expand its shared analysis space continuously. For a comprehensive review of fusion strategies and their trade-offs in astronomical applications, see Shao et al. (2026).

Together, these architectural innovations suggest a coherent design philosophy for DESC FMs: hierarchical representations that preserve cosmological structure, modular routing across scientific workflows,

and learned compression layers capable of cross-probe alignment. Each of these desiderata drive toward reproducible, uncertainty-aware analyses within a unified DESC software framework.

5.1.4. Evaluation

In contrast to traditional single-purpose ML models trained end-to-end for a specific science objective, FMs derive their value from serving a broad range of downstream use cases. This generality introduces new evaluation challenges: it is critical to assess performance across the full spectrum of tasks to which the FM will be applied, ensuring that it can be successfully adapted to each. In addition, because FMs are trained on large and difficult-to-characterize datasets, they include implicit biases that must be corrected for downstream scientific applications. It is therefore important to verify the correctness of adaptation and calibration procedures for each task. Establishing a common framework of benchmarks to evaluate this adaptability must be a research priority, particularly given that such validation is not standard practice for many industry-developed FMs.

Benchmarks—The development or deployment of FMs within the DESC ecosystem should be accompanied by a comprehensive suite of benchmarks designed to evaluate predictive performance across a representative range of science tasks. Particular attention should be paid to robustness under survey systematics, sensitivity to biases inherited from the FM's training data, and the ability of adaptation procedures to yield well-calibrated uncertainties for each downstream application.

Interpretability—Mechanistic interpretability will provide a complementary route to validation. Techniques such as attention visualization, activation clustering, or sparse dictionary learning can be adapted to determine whether internal model representations recover known astrophysical relations including the color–magnitude diagram, the fundamental plane, or the Tully–Fisher relation. Developing astronomy-specific interpretability tools would enable DESC to quantify whether FMs encode physically meaningful correlations or merely reproduce empirical correlations.

Distribution shift—Evaluation under distribution shift is also essential for robustness. DESC models must maintain reliability under temporal drift across survey years, spatial variation in observing conditions, and transfer to external datasets such as *Euclid* or *Roman*. Dedicated stress tests should replace the assumption of independent and identically distributed (IID) validation, using importance-weighted calibration errors and worst-group accuracy measures to reveal biases that emerge only under covariate change. These tests will be crucial for ensuring that DESC FMs remain stable as LSST transitions from early to full survey operations.

Long-term impact—Finally, the deployment of large-scale models must consider sustainability and community governance. Training and fine-tuning FMs require substantial computational resources, underscoring the need for shared development, standardized documentation, and reproducible training pipelines. Strategic coordination within DESC and with external collaborations will ensure that these models serve as transparent, responsible, and scientifically verifiable assets for the next generation of cosmological analysis.

5.2. Large Language Models & Agentic AI

Large language models (LLMs) have demonstrated an astonishing capacity for performing cognitive tasks and knowledge work (e.g. synthesizing information, generating text, writing and explaining code) that has triggered an AI revolution over the past few years. When equipped with tools, LLM-powered agentic systems go further: they can develop and execute code, perform exploratory data analysis, or orchestrate multi-step workflows. As of late 2025, complex tasks, especially in software engineering, can increasingly be delegated to such systems and completed with high accuracy. Although integration into scientific workflows remains nascent and faces genuine obstacles, the pace of progress and the potentially transformative impact of the ability to delegate significant fractions of the research workflow to AI suggests DESC should plan for how to effectively integrate these tools and guide the evolution of best practices, in line with the guiding principles in Section 1. The goal of integrating these tools should not be full automation or dehumanization, but empowerment: elevating the level at which researchers can engage within the scientific process, enabling them to tackle more ambitious projects while concentrating on fundamental questions, critical interpretation, and creative reasoning.
Below, we review the existing capabilities and limitations in the adoption of LLMs for science, and examine potential applications for DESC, ranging from mature (e.g. documentation assistance) to aspirational (e.g. AI co-scientists). Overall, this subsection aims to answer the following question: what would successful integration of LLMs and agentic systems within DESC look like?

5.2.1. From LLMs to Agentic AI

LLMs as knowledge tools—At baseline, LLMs are sophisticated systems for natural language processing. They process, synthesize, and generate text, though they remain subject to hallucinations (i.e. false statements expressed with confidence), out-of-date knowledge, and limited problem-solving capabilities. Some of these limitations can be mitigated to engineer reliable production-ready systems using retrieval-augmented generation (RAG), a technique which allows LLMs to ground their responses using context-specific documentation (Lewis et al. 2020; Fan et al. 2024). OpenScholar (Asai et al. 2024), for example, uses RAG over 45 million open-access papers to synthesize citation-backed responses, substantially outperforming general-purpose models while avoiding hallucinated citations. Over the last year, the proficiency of LLMs at problem solving has also significantly improved with "reasoning models" (e.g. OpenAI's o1/o3, DeepSeek-R1 (Guo et al. 2025)) which allocate additional compute at inference time to perform deliberate, multi-step problem-solving (a paradigm sometimes called test-time compute scaling). These models show marked improvements on tasks requiring extended chains of reasoning. On physics tasks specifically, TPBench (Chung et al. 2025) (a benchmark of original theoretical physics problems ranging in difficulty from undergraduate to research-level) finds that newer reasoning models substantially outperform earlier systems, though research-level problems remain largely unsolved. This trajectory suggests that LLM capabilities in scientific reasoning will continue to improve, even as fundamental limitations persist. It is important to note that LLMs alone can inform and assist, but they cannot act: they process information rather than executing tasks.

Agentic AI: from knowledge to action—Agentic systems represent a conceptual leap beyond individual LLMs. Where LLMs generate text, agents take actions: writing and executing code, calling APIs, modifying files, and interacting with other computational infrastructure. Where reasoning models are used, these

agents can also dynamically revise their actions based on feedback and provide interpretable explanations for their decisions. These capabilities make it possible to delegate complex and loosely-defined tasks to agents much as one might delegate them to a colleague. However, not all tasks are equivalent, and a useful mental model is to ask: what could one hand off to a person of varying competency? At the "intern" level, current agentic systems can reliably fix small bugs, run analyses with different parameters, and write simple tests to verify their outputs or those of another code. At the "junior" level, they can implement well-specified features or debug failing pipelines. "Senior"-level capabilities (building entirely new software systems, weighing design trade-offs in solving a specific problem, and validating scientific methodology) remain largely out of reach at present, though the complexity and execution horizon of these systems continues to increase. Tools such as Claude Code[6], Cursor[7], and Devin[8] demonstrate agents completing real software engineering tasks: calling APIs, executing shell commands, maintaining context across dozens of steps, recovering from errors, and iterating toward solutions. On SWE-bench (Jimenez et al. 2023), a benchmark of real GitHub issues from open-source Python repositories, leading agents now resolve over 70% of issues in the human-verified subset; on the harder SWE-bench Pro (Deng et al. 2025), which targets enterprise-level problems, the best systems solve roughly 25%.

Towards Scientific Agentic Systems—Beyond software engineering, agentic systems are increasingly being applied to scientific research. A first generation of agents has focused on literature search, going beyond static RAG systems by using tools and iterative refinement. PaperQA2 (Skarlinski et al. 2024), developed by FutureHouse, exemplifies this approach: rather than simply retrieving from a fixed index, it uses search tools to find relevant papers, traverses citation graphs to discover related work, and iteratively refines its own queries based on retrieved text. This agentic approach matches or exceeds PhD-level researchers on literature retrieval benchmarks. More ambitious systems extend beyond literature search to integrate data analysis and hypothesis generation. Google DeepMind's AI co-scientist (Gottweis et al. 2025), a multi-agent system built on Gemini, uses a "generate, debate, and evolve" approach to produce novel hypotheses; it has identified drug-repurposing candidates for liver fibrosis that were later validated in laboratory experiments. FutureHouse's Robin (Essam Ghareeb et al. 2025), an open-source multi-agent system, autonomously proposed ripasudil as a treatment for dry age-related macular degeneration and validated the hypothesis through wet-lab experiments. Edison Scientific's Kosmos (Mitchener et al. 2025) orchestrates parallel data-analysis and literature-search agents over 12-hour runs, producing detailed scientific reports that independent evaluators found 79% accurate. Sakana AI's AI Scientist v2 (Yamada et al. 2025) uses agentic tree search to autonomously generate hypotheses and run experiments, and demonstrated its ability to run end-to-end machine learning research projects. This list is by no means exhaustive, and these systems remain by and large proofs-of-concept, with little independent data on their utility and reliability for real-world complex scientific research. Nonetheless, these examples illustrate both the interest and rapid advancement in agentic systems for scientific research.

LLM and agentic AI in astronomy—Domain-tuned LLMs such as AstroSage-Llama-3.1-8B (de Haan et al. 2025) demonstrate that compact, astronomy-specific models can match larger general-purpose systems at lower cost. In parallel, LLMs are enabling large-scale knowledge synthesis: Pathfinder (Iyer et al. 2024) applies semantic retrieval and citation-aware summarization across $\sim 350{,}000$ astrophysical papers in the SAO

[6] https://code.claude.com/
[7] https://cursor.com
[8] https://devin.ai/

Astrophysics Data System (ADS), allowing users to move from keyword-centric searches to concept-level exploration. ChatGaia[9] demonstrates the potential for natural-language interfaces to complex astronomical databases: it translates user queries into ADQL for the Gaia Archive, making $\sim$2 billion stellar sources accessible without requiring query-language expertise. Several agentic proofs-of-concept have also emerged. CMBAgent (Laverick et al. 2024; Xu et al. 2025) coordinates RAG with local code execution to run MCMC pipelines for cosmological parameter estimation from Atacama Cosmology Telescope (ACT) data. Mephisto (Sun et al. 2025b) iteratively refines stellar population parameters by orchestrating multi-band galaxy observations with the CIGALE SED-fitting code (Boquien et al. 2019). Denario (Villaescusa-Navarro et al. 2025) extends this paradigm further, spanning a full research lifecycle from idea generation through data selection, modeling, and manuscript drafting. These systems remain demonstrations rather than production tools, with scientific validation of each system's outputs still performed manually.

As a concrete example of the state of the art, and one directly relevant to DESC, the 2025 NeurIPS Weak Lensing Uncertainty Challenge[10] (a competition to infer cosmological parameters from simulated convergence maps while quantifying uncertainty) was won by a team using the CMBAgent system (Bolliet 2025), beating a team of domain experts in weak lensing and ML who placed second without AI assistance. The CMBAgent-assisted team, though not specialists in weak lensing inference, overtook the experts within weeks and held the lead through the end of the competition. This result illustrates how LLM-based agentic systems can accelerate scientific research and lower the barrier to complex analysis techniques.

Fundamental limitations—While early successes have shown the promise of LLMs and agentic systems for science, the current generation of these systems face fundamental constraints that limit their applicability to science. As emphasized by Ilievski et al. (2025), human scientific reasoning relies on abstraction, causal inference, and conceptual transfer, whereas AI systems depend on statistical interpolation. This leads to what has been termed "jagged intelligence": systems that solve Olympiad-level problems while failing at kindergarten logic, with capabilities forming an uneven landscape rather than a coherent skill set. Hallucination (generating confident but incorrect outputs) also remains a fundamental problem for scientific applications of LLMs. Tauman Kalai et al. (2025) provide a theoretical explanation: current training objectives and evaluation benchmarks structurally incentivize guessing over expressing uncertainty. When a model says "I don't know," benchmarks penalize the response in the same way as the model producing an incorrect output, so models learn to guess confidently even when uncertain. Hallucinations, as a result, are a direct consequence of how these systems are trained and evaluated. Hallucinations span a broad taxonomy, from factual errors about well-documented knowledge to fabricated citations to arbitrary confabulations (Ji et al. 2022), and they can be subtle enough to evade casual review. For science, the danger is acute: a hallucinated statistical result or misremembered prior work could propagate through analysis undetected if careful validation frameworks are not imposed on model output. Beyond hallucination, these systems face limitations particularly acute for scientific discovery. They are trained on existing literature and may fail precisely where discovery happens (at the frontier where established statistical patterns break down). They also lack reliable uncertainty quantification: unlike a careful scientist who flags when they have extended beyond their expertise, LLMs produce outputs with uniform confidence regardless of whether they are interpolating within training data or extrapolating beyond it. Finally, complete reproducibility of LLM outputs remains challenging: outputs are stochastic, sensitive to prompt phrasing, and subject to version drift as

[9] https://chatgpt.com/g/g-aYZOjK5zy-chatgaia
[10] https://www.codabench.org/competitions/8934/

underlying models change (challenges that conflict with science's demand for verifiable, repeatable results). For all these reasons, human oversight remains essential for science-critical output.

5.2.2. Potential applications for DESC

The following sections organize potential application areas of LLMs by the outcomes they serve for DESC, rather than by the underlying technology. Each area sits at a different level of technological maturity, and the path toward realizing each application differs accordingly.

Knowledge work and research support—This is the most mature application area, and the one where DESC members can immediately benefit. Success in this area would mean that new members can query DESC documentation in natural language and receive accurate, relevant answers rather than navigating distributed technical documentation and obscure Slack channels. It would mean that literature search moves from keywords to concepts, helping researchers find relevant work faster and discover connections across subfields. Writing assistance for drafts, summaries, and documentation would be readily available, though always with human review. As summarization capabilities improve, collaboration maintenance (meeting summaries, cross-working-group communication) could also become increasingly efficient.
The methodology is relatively mature for these tasks. For literature work, tools like OpenScholar, PaperQA2, and Pathfinder (discussed above) already enable concept-level exploration and synthesis across large scientific corpora. General-purpose LLMs can summarize papers, extract key results, bridge jargon gaps across subfields, and assist with writing tasks. Domain-tuned models like AstroSage (discussed above) have shown that compact, astronomy-specific LLMs can match larger general-purpose systems on astrophysical question-answering at lower cost. RAG techniques allow these systems to ground their responses in specific documentation, reducing hallucinations and enabling citation of sources.
The path forward for DESC involves building a RAG system over collaboration documentation, tutorials, and pipeline code; establishing human review guidelines for any generated content; and starting with low-stakes applications (e.g., onboarding Q&A) before moving to higher-stakes use cases. Looking further ahead, a retrieval-augmented knowledge graph that unifies documentation, code, and literature could provide a continuously evolving record of methodological provenance, making it easier to trace how analysis choices change across the collaboration and how they connect to the broader literature.
The main challenges are hallucination (confident wrong answers erode trust, so citation of sources is essential), the maintenance burden of keeping RAG indices current as documentation evolves, and establishing clear guidelines and training procedures for how to use, review, and acknowledge generated content.

Natural language interaction with data—Natural-language interfaces can dramatically reduce the barrier to accessing complex astronomical databases and archives. Success in this application area would mean that DESC members can query LSST catalogs, simulation products, and image archives dynamically using plain language rather than specialized query languages or custom hard-coded scripts. A researcher could ask "show me all galaxies with strong tidal features in the deep drilling fields" and retrieve relevant candidates without writing SQL or navigating file systems.
At the time of writing, this methodology is now emerging. ChatGaia[11] demonstrates that conversational in-

[11] https://chatgpt.com/g/g-aYZOjK5zy-chatgaia

terfaces to catalog data are feasible: it translates natural-language queries into ADQL for the Gaia Archive, making ~ 2 billion stellar sources accessible without query-language expertise. For imaging data, AION-Search (Koblischke et al. 2024) enables semantic search across 140 million galaxy images by using vision-language models to generate captions and align image embeddings with text queries, allowing researchers to search for morphological features or rare phenomena using free-form descriptions rather than predefined categories.

The path forward for DESC involves building natural-language interfaces to key data products: LSST catalogs, difference imaging outputs, and simulation archives. Such interfaces could be grounded in schema documentation and example queries to reduce the likelihood of hallucinated or malformed outputs. Techniques for constrained generation offer another avenue for guaranteeing that model outputs adhere to a pre-defined schema, and research in this area is ongoing (Mündler et al. 2025). Integration into analysis notebooks would allow seamless transitions between exploratory queries and downstream analysis.

The main challenges to achieving this application area are accuracy (incorrect query translations could return misleading results), coverage (not all queries map cleanly to database operations), and validation (users need techniques for verifying that the system understood their intent). For image search, an additional challenge is that rare phenomena are precisely where training data is sparse, making retrieval of scientifically-interesting objects a non-trivial task.

Software engineering acceleration—This area spans two levels of maturity: interactive coding assistance (mature) and autonomous operation (emerging). Success would mean that DESC developers spend less time on routine implementation and more time on high-level design and scientific validation. Interactive tools would help members implement features, fix bugs, and write tests more efficiently. Autonomous agents would handle low-risk maintenance tasks (updating tutorials when APIs change and package dependencies when softwares do, generating test coverage, fixing CI failures) and open pull requests for human review. Ultimately, the implementation of new analysis pipelines could also be delegated to agents, with researchers specifying the requirements and validation tests in natural language.

The methodology for interactive use is already mature. Tools such as Claude Code, Cursor, and GitHub Copilot are widely deployed and demonstrate real productivity gains on well-specified tasks. Autonomous operation is less mature: agents can be triggered by CI failures to diagnose issues and propose fixes, but reliability remains inconsistent for complex, long-horizon problems due to limiting problem-solving abilities and finite context windows.

The path forward for DESC involves broader adoption of interactive coding assistants for pipeline development, along with collaboration-specific contexts tailored to DESC workflows (e.g., LSST data structures, pipeline conventions, DESC software guidelines). For autonomous agents, the prudent approach is to pilot these systems first on low-risk tasks such as test generation and docstring updates before integrating them into CI workflows.

The main challenges are code correctness (agents can introduce subtle bugs, so review and validation remains essential), security (generated code may have vulnerabilities), and institutional knowledge (members may not understand code they did not write). Observability is also an important point: tracking which agent and version contributed to each change, and ensuring edits stay limited to the relevant code context rather than allowing regular rewrites of entire scripts, will more easily enable debugging and auditing.

Analysis support—This area is less mature and requires additional validation infrastructure to be put in place before deployment should be considered. Success would mean that AI assistants, integrated directly

into analysis notebooks, help researchers move faster through the exploratory phase of analysis: generating visualization code from natural-language requests, suggesting statistical tests, and helping iterate on plots and diagnostics. More ambitiously, such assistants could help construct null tests and systematics checks, lowering the barrier to rigorous validation while freeing the user to explore additional aspects of the data. A junior researcher could more easily implement the suite of tests that an experienced analyst would know to run in a given scientific context, making scientific best practices more accessible and standardizable beyond individual working groups and across the full collaboration. Beyond individual analysis, agents could proactively monitor nightly data streams and flag anomalies, generate human-readable QA summaries, or diagnose pipeline failures. Human-in-the-loop frameworks, where active-learning strategies solicit expert input only for the most ambiguous cases (Settles 2009; Christiano et al. 2017), could further optimize the use of limited expert time for monitoring these systems.

The methodology for LLM-enabled analysis support is nascent. While the building blocks exist (code execution, RAG, multimodal understanding), integrated systems for scientific analysis support are largely experimental at present. Projects like Jupyter AI[12] provide infrastructure for integrating LLMs directly into notebook environments, enabling conversational assistance and code generation within analysis workflows, but robust evaluation frameworks for analysis tasks remain underdeveloped.

The path forward for DESC requires building evaluation benchmarks before deployment: curated test cases where correct answers are known, so that agent outputs can be validated. Piloting on low-stakes analysis tasks (e.g., generating diagnostic plots, summarizing pipeline logs) would build experience before moving to higher-stakes applications. Most importantly, developing clear guidelines for human oversight is critical to ensuring that researchers remain accountable for scientific conclusions: AI can accelerate the work, but the responsibility for validating results and interpreting their meaning must stay with humans.

The main challenges are automation bias (over-trusting fluent outputs that may contain subtle errors), the difficulty of validating agent reasoning on novel analyses, and the risk that diagnostics look valid but miss real issues. For science-critical workflows, the bar for reliability is high, and current systems do not yet meet it. There is also an educational tension: critically evaluating AI-generated analysis requires understanding the underlying methods, yet that understanding has traditionally come from wrestling with implementation details oneself. If AI removes the friction of implementation, how do junior researchers develop the judgment needed to recognize when the AI is wrong? Training the next generation to use these tools effectively, without becoming dependent on them, is an open challenge.

Toward an AI co-scientist—This is the most ambitious and speculative application area. Success would mean an agent capable of PhD-student-level work: running analyses, interpreting results, drafting manuscript sections, and iterating on feedback with sufficient robustness that a PI could trust the output as they would a competent junior colleague. Such an agent might surface connections across DESC working groups that humans miss, generate hypotheses grounded in both literature and data patterns, or systematically explore parameter spaces or analysis choices that would be tedious for humans to cover. The measure of success in this area will be in its ability to amplify existing scientific efforts: researchers would report that these tools help them think more deeply and explore more broadly, not merely execute faster.

The methodology remains largely aspirational. The proofs-of-concept discussed earlier (CMBAgent, Mephisto, Denario, AI co-scientist, Robin) demonstrate that end-to-end workflow orchestration is possible, but none has demonstrated PhD-student-level robustness on real scientific problems. Scientific validation

[12] https://jupyter-ai.readthedocs.io/

of their outputs remains entirely manual. Evaluation frameworks are beginning to emerge: Ye et al. (2025) introduce ReplicationBench, an astrophysics-specific benchmark testing whether agents can faithfully reproduce published analyses; Gravity-Bench (Koblischke et al. 2025) evaluates agents on physics discovery tasks, testing their ability to plan experiments and infer physical laws from simulated gravitational systems; while ScienceBoard (Sun et al. 2025a) and DiscoveryWorld (Jansen et al. 2024) assess compositional scientific reasoning more broadly.

The path forward is necessarily incremental. DESC should build on the higher-maturity applications first (documentation, coding assistance, data access), developing institutional experience and trust before attempting more autonomous scientific work. DESC-specific evaluation benchmarks, perhaps based on reproducing known analyses or detecting injected errors, would help measure readiness. Human oversight must remain central at every stage.

The challenges here are the deepest. Reaching PhD-student-level robustness requires solving the fundamental limitations described earlier: agents must know when they are outside their competence, reason about physics rather than pattern-match on surface features, and produce outputs that are reproducible with full provenance. The bar for scientific discovery is extraordinarily high. A single undetected error in an analysis pipeline could propagate into published results and impede progress instead of enabling it. Finally, proprietary LLMs are trained by companies to be agreeable and affirming; a model valuable for enhancing scientific discovery should be proactive in challenging the user's assumptions, critiquing its own outputs, and pursuing independent (but relevant) lines of inquiry. Whether current architectures can be made reliable enough for this level of autonomy, or whether fundamentally new approaches specific to the physical sciences are needed, remains an open question.

5.2.3. Implementation Considerations

Beyond the application-specific concerns discussed above, three cross-cutting considerations will shape how DESC integrates these tools: evaluation, governance, and infrastructure.

Evaluation and benchmarking—Benchmarking is essential for improving AI systems, and meaningful benchmarks can only be created by domain experts who can accurately evaluate the quality of a provided solution. DESC is well-positioned to contribute here. Benchmarks for DESC applications should measure: (1) factual accuracy on domain literature and documentation; (2) code validity and runtime safety; (3) reproducibility under random seed variation; and (4) robustness to distribution shift and systematic differences (e.g., early vs. late LSST years, cross-survey transfers). Existing efforts like ReplicationBench, Gravity-Bench, and ScienceBoard (discussed above) provide templates for these efforts, but DESC-specific benchmarks (perhaps based on reproducing published analyses, detecting injected systematics, or validating pipeline outputs) would directly measure readiness for collaboration-wide use. Establishing such benchmarks is itself a scientific contribution: it encodes expert knowledge about what problems are worth solving and provides a foundation for systematic improvement of AI.

Governance and reproducibility—Integrating LLMs into scientific workflows raises unresolved questions about reproducibility, provenance, and attribution. Each automated workflow should maintain full provenance metadata: prompts, model versions, retrieved sources, generated code, and execution logs. Observability (logging all agent actions) will enable debugging and audit (Zhou et al. 2025), and increase

credibility in designed agentic systems. Sandboxed execution environments enforce deterministic behavior and ensure that agent-generated code runs in secure, auditable contexts. Reproducibility, however, remains difficult. With the pace of technical improvement rapidly increasing, ensuring that an analysis performed today can be replicated exactly next year is an open problem. Attribution also lacks established norms: how should AI tools be credited in publications? Should they be listed as co-authors or as software tools, and how should their contributions be presented? These questions require collaboration-wide discussion and policy development, but are extremely topical as papers presenting physics research fully conducted by LLMs begin to appear (e.g., Schwartz 2026).

Infrastructure and sustainability—The physical deployment of LLMs within DESC must respect both scale and sustainability. Large models are expensive to run, and the cost-capability trade-off matters for a collaboration operating over LSST's decade-long baseline. Compact open-weight models, fine-tuned on domain-specific data, offer a path to balancing capability with efficiency while avoiding vendor lock-in—AstroSage exemplifies this approach. A dedicated RAG layer connecting DESC documentation, simulation archives, and survey data products could form the backbone of an agentic knowledge infrastructure. But even with efficient models, infrastructure requires ongoing maintenance: RAG indices must be updated as documentation evolves, prompts need revision as models change, and LLM APIs themselves drift over time. Planning for this maintenance burden from the outset is essential.

LLMs and agentic systems will not replace scientific judgment, but they may extend researchers' reach (handling routine tasks, surfacing relevant literature, accelerating code development) so that human effort can concentrate where it matters most. Finding the right interaction patterns between scientists and these systems remains an open challenge. DESC should engage now, building evaluation infrastructure and piloting lower-risk applications, to be ready as capabilities mature.

6. Infrastructure Requirements to Support AI/ML in DESC

Infrastructure is the shared technology needed to realize the methods, models, and scientific opportunities outlined in the preceding sections. It may be deployed in a distributed mode with individuals or small teams using local or institutional resources, or in a coordinated mode within a shared platform or environment. This latter mode is most relevant for the largest-scale AI/ML-enabled model training and inference workflows at the scale of the entire Rubin-LSST data set, including via core infrastructure through DOE-funded high performance computing facilities in the US. The following subsections review the infrastructure elements most relevant to the future of AI/ML in DESC.

6.1. Software

Software is foundational to modern cosmology, especially for DESC, where scientific insight increasingly depends on sophisticated computational workflows. As AI and ML mature into core scientific technologies, software itself becomes a form of research infrastructure. In this context, two tightly connected priorities emerge: first, the development and long-term stewardship of a robust AI software stack that enables model development and experimentation; second, the strategic integration of AI methods into the DESC scientific pipelines, where they will ultimately shape how we reduce data, extract measurements, and perform inference.

6.1.1. The AI Software Stack

DESC has long demonstrated leadership in scientific software development: from collaboration-wide development guidelines[13] and software-oriented publication policies, to a culture of reproducibility and sustained support for collaboration-wide software stacks. As AI becomes a first-class component of scientific analysis, extending that same discipline and strategic thinking to the AI software ecosystem is increasingly important. The goal is to define a modern, durable, and interoperable stack built on best practices for reproducibility and maintainability. This accelerates research while preserving the transparency that has always characterized DESC software.

To meet this goal, we recommend converging on a small set of shared practices and services that make ML development reproducible, portable to the IN2P3 Computing Centre (CC-IN2P3) and National Energy Research Scientific Computing Center (NERSC), and sustainable over the 10-year duration of the LSST survey. Key components include:

- **Frameworks for model development**, likely PyTorch for large models and JAX for differentiable physics.
- **Experiment tracking**, capturing code/data versions, configurations, metrics, and compute environments to ensure reproducibility.

[13] https://lsstdesc.org/assets/pdf/docs/DESC_Coding_Guidelines_latest.pdf

- **Model and artifact registries** to version and archive trained models, datasets, and reports, mirroring survey data-release practices.
- **Standardized export formats** such as Open Neural Network Exchange (ONNX) so models integrate cleanly with Rubin/DESC pipelines and high-performance computing (HPC) environments.
- **Continuous integration (CI) and continuous delivery/deployment (CD) for models** to test and validate training configurations, exported artifacts, and deployment environments.
- **Observability** and drift monitoring so ML components used in production behave predictably and transparently.

These elements are not overhead; they are the operational foundations that allow AI to become reliable scientific machinery rather than episodic experimentation. They also reduce long-term maintenance burden by enforcing shared conventions, minimizing bespoke tooling, and allowing learned models to be audited, reused, and trusted throughout the LSST decade.

In addition to these considerations for ML model development, DESC will increasingly rely on LLMs as flexible interfaces to data, documentation, and tooling. LLMs are unusual compared to conventional models in that they are supplied by a rapidly changing ecosystem of commercial providers, open models, and self-hosted deployments. Versions change frequently, and some use cases may require on-premises or institutionally-hosted models—e.g., at NERSC, CC-IN2P3, the Cambridge Service for Data Driven Discovery (CSD3), or similar facilities—via serving stacks such as vLLM for data-governance or cost reasons. To remain agile in this landscape, DESC should treat LLMs as interchangeable backends behind a stable internal API. Such an abstraction layer enables swapping models without rewriting pipelines and moving workloads between commercial services and collaboration-owned GPU resources as needs evolve. Beyond the LLMs themselves, agentic frameworks (libraries that orchestrate multi-step LLM workflows, tool calls, and human-in-the-loop interactions) are even more volatile, with new options appearing and disappearing on timescales of months. Frameworks such as LangGraph exemplify the current direction, representing agents as graphs of tools, memory, and control logic, and providing execution and tracing engines around them.

Framework sustainability and openness should play an important role in guiding tooling choices. Solutions with strong governance and broad adoption (e.g., PyTorch under the Linux Foundation, ONNX, MLflow or self-hostable experiment-tracking systems) provide long-term stability and avoid dependence on proprietary silos. Finally, coordination with computing partners is essential to ensure portability of the software stack and deployment on more HPC-aligned facilities.

6.1.2. Integration of AI/ML within Analysis Pipelines

As AI components mature, DESC pipelines may evolve from purely sequential “raw-to-reduced” workflows to systems where learned models are first-class, production-grade elements. The emphasis is on embedding AI in ways that enhance measurement fidelity, calibration control, and inference efficiency, while preserving transparency, reproducibility, and smooth integration with existing practices.

Data reduction pipelines—AI/ML is already present in DESC workflows (e.g., photo-z estimation within RAIL). Over time, more components are likely to incorporate learned models at defined stages such as deblending, PSF estimation, artifact rejection, and photometric calibration, tightly coupled to Rubin/DESC data structures and HPC execution. In parallel, DESC has interest in end-to-end approaches that optimize models directly against observational data and simulators, potentially replacing brittle stage boundaries while maintaining provenance through experiment tracking and model registries.

Foundation Models as Services—With the advent of the foundation model paradigm, we envision the possibility of providing "X-as-a-service" (photometry, astrometry, cross-matching, classification, anomaly detection) behind stable APIs, so models can evolve without churn in downstream code.

AI/ML-based cosmological inference—AI-based methods are shifting inference from sampler-centric workflows toward models that learn mappings from data to posteriors or summaries. Foundation-style surrogates can amortize computation and reduce MCMC-heavy workloads, while simulation-based inference trains directly on forward models and field-level methods operate on minimally reduced data. Active-learning loops may trigger simulations on the fly to refine surrogates and concentrate the HPC budget where it most reduces uncertainty. For pipeline use, these elements remain tied to provenance systems and model registries.

Shared infrastructure for emulators and surrogate models—Consistency across emulator efforts can lower reuse costs. A lightweight scaffold could include common interfaces (inputs, units, cosmology/observational context, stochastic controls, uncertainty outputs), a minimal dataset schema for training/evaluation, embedded metadata for provenance and validity domains, and containerized artifacts with dual exports (native framework and ONNX where feasible). A small validation suite (accuracy, coverage, calibration under shift, throughput) running in CI on facility images would help with portability to DESC computing facilities, and hooks for active learning can keep surrogates co-evolving with forward models.

LLMs and agents—Large language models and specialized agents can contribute at multiple levels of DESC work: in notebook environments as integrated assistants that accelerate iteration on analysis code, diagnostics, and documentation/query synthesis; in data discovery and curation by searching heterogeneous archives and DESC repositories, proposing cross-matches, and flagging anomalies; and at facility scale by assembling templated workflows, submitting authenticated jobs, and recording outcomes into experiment tracking and model registries. Adoption pairs capability with governance: standardized interfaces for serving models, authenticated access, audit logs of prompts and actions, and human-in-the-loop checkpoints for any decision with scientific impact.

Likely, investment in general-purpose AI tooling across industry, open-source, and public efforts will continue to exceed resources available for cosmology-specific software. When such tools meet DESC's scientific and operational needs, adopting them can leverage broader community advances while allowing collaboration effort to focus on domain-specific modeling and validation. Finally, AI will also influence our approach to engaging with computing in the future. Computational infrastructure will increasingly be harnessed with the assistance of LLMs and agentic frameworks. Depending on how this transition proceeds, it may enable a larger community of researchers within DESC to engage directly with large-scale scientific computing.

6.2. Computing

6.2.1. Workflows and Scales

Estimating the full scale of resources needed to support the range of DESC AI/ML use cases is beyond the scope of this white paper, and any estimates will evolve with time as new methodologies and science applications are developed. Major resource categories include GPU and CPU time, short- and long-term storage, and network bandwidth both between and within analysis facilities. AI-oriented workloads will range from small-scale research & development (R&D) to training and serving FMs, serving tokens for agents/LLMs, running on-the-fly simulations for active learning loops for SBI, up to potentially running simulations on the fly as part of explicit inference loops. All of these will bring their own requirements for computing cycles, data locality and throughput, interactivity, and orchestration.

At the low end of requirements, resource-augmented instances of the commercial cloud-based Rubin Science Platform (RSP) would provide an accessible route for individuals and small teams to scale up AI/ML workflows that require integration with the Rubin data and software stack. For cost efficiency, these resources would likely need to be managed through either a batch-processing system or through an elastic data-science workflow system. Allocating GPU-based interactive servers (virtual or physical) in the cloud-based RSP context is unlikely to be viable at scale for many users, given the typical idle time in interactive sessions and workflows. Larger AI workflows could also be accommodated through individual or DESC-wide allocations of CPU, GPU, and working storage at NERSC or through proposal-driven allocations under the 10% of computing time reserved for Rubin science users at the Rubin US Data Facility (USDF) at Stanford Linear Accelerator Center (SLAC).

On the high end, significant computing resources may be needed for new simulations to train simulation-based inference approaches to large-scale cosmology analyses. This includes not only the computing needed for simulation but also the storage to save and share large simulation outputs. Another major driver of resource needs would be training of data-oriented FMs at the scale of the full Rubin data set. Again, an accurate estimate of the resources needed for this would require further study of an appropriate reference architecture and training strategy. A rough guide to the scale can be found from the (CPU-based) compute sizing model for Rubin–LSST Data Release Production[14] since it is operating on roughly the same scale of data as a full-scale LSST-based FM would train on. The operations compute model estimates a need for about 50M core-hours in year 1 of the survey, rising linearly year over year (assuming annual data releases) to roughly 10x this amount in year 10. Storage needs are estimated to increase from 30PB in year 1 to over 800PB in year 10 although not all of the associated data products would necessarily be needed for FM training.

In addition to “offline” computing needs, a number of time-critical AI/ML-driven DESC workflows will be driven by the nightly alert stream, including ML-driven algorithms for classification, anomaly detection, and intelligent follow-up observations. Many of these will be implemented within the broker and marshal systems that receive the LSST alert stream; depending on their level of resource-intensiveness, broker/marshal com-

[14] https://dmtn-135.lsst.io

puting capacity may need to be further augmented, and/or integrated with elastic or preemptive compute allocations within research computing facilities or in commercial cloud.

6.2.2. Computing Resource Providers

In the US, DESC computing needs are primarily supported by DOE through its network of computing facilities under the Advanced Scientific Computing Research (ASCR) program within the DOE Office of Science. This includes NERSC as the primary user facility for DESC science analysis, Energy Sciences Network (ESNet) for advanced wide-area networking and data movement, and the ALCF and OLCF for larger-scale computational work. ASCR is also developing the HPDF, led by Jefferson Lab in partnership with Lawrence Berkeley National Laboratory. DOE is also advancing the development of an Integrated Research Infrastructure (IRI) to support flexible, powerful, and accessible implementation of scientific workflows across all these facilities.

Anticipating an increasingly prominent role for AI in the scientific exploitation of data created by Rubin and other DOE-funded facilities across all disciplines, DOE launched the Genesis Mission in November 2025[15], a national effort to accelerate AI-driven scientific discovery across its 17 national laboratories. As infrastructure for this mission, the US Congress has funded the creation of AmSC[16] to develop and deploy the next generation of AI-oriented capabilities on the foundation of the DOE computing platforms noted above. AmSC development is underway now, with funding distributed across an AmSC infrastructure component, a core AI model-development consortium (ModCon), pilot funding for discipline-specific data curation and science benchmarking activities, and seed funding for discipline-specific AI model teams. As a cornerstone of the Genesis Mission, computing and AI model development infrastructure within the AmSC represent a significant opportunity to address ambitious DESC AI/ML goals.

The joint NSF–DOE nature of Rubin Observatory opens the possibility of leveraging significant NSF-supported computing facility resources, especially if pursued in coordination with other Science Collaborations working in areas typically supported by NSF. These resources include the Leadership Class Computing Facility (LCCF) entering production at the Texas Advanced Computing Center. DESC members also participate in both astrophysics-oriented National Artificial Intelligence Research Institutes, funded jointly by NSF and the Simons Foundation (SkAI, led by Northwestern University; and CosmicAI, led by the University of Texas at Austin), and have access to their associated computing resource allocations. Considering interests in Rubin–*Roman*–*Euclid* joint analysis, National Aeronautics and Space Administration (NASA)-funded computing resources may also be a viable option.

More broadly, many DESC members have access to significant campus-level computing at their institutions. Members outside the US have access to their own networks of national resources, including national and regional initiatives prioritizing computing for AI in science. Through the in-kind contribution program that supports LSST data rights for scientists outside of the US and Chile, a network of IDACs is being deployed, with some sites bringing additional CPU and GPU capabilities that DESC would be well positioned to make use of. The UK will host an IDAC, sized to satisfy the resource needs of 20% of the global LSST community during survey operations. The UK IDAC will be connected to UK national research computing facilities, both

[15] https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/
[16] https://science.osti.gov/-/media/grants/pdf/lab-announcements/2025/LAB-25-3555.pdf

traditional (simulation and modeling) supercomputing services and the coming generation of AI-focused Digital Research infrastructure currently being specified and prototyped through the ASCEND process[17].

Hyperscale commercial enterprises may also offer a viable path for certain novel and ambitious Rubin-LSST AI/ML applications, provided that their resources can be engaged through partnership or at significant discount. Potential partners include major cloud providers such as Google, Amazon, and Microsoft; major AI players such as OpenAI, Anthropic, and (again) Google; and GPU manufacturers such as NVIDIA and AMD. Additional effort would be required to develop private-sector partnerships that are mutually beneficial and compatible with the proprietary and non-commercial requirements of the Rubin–LSST data policy[18]. On the positive side, timescales in industry are typically much shorter than in academia: work with an engaged hyperscale commercial partner could potentially deliver large-scale results quickly.

6.3. Data

The primary data products relevant for DESC AI/ML work fall into several categories:

1. LSST data products delivered by Rubin Observatory
2. Derived data products produced through DESC collaboration efforts
3. Data from other major surveys that enhance and expand DESC AI/ML science
4. Simulation data
5. Model weights and biases from AI models trained on the above

Rubin-LSST data products are organized into three categories distinguished by the timescale of their delivery. The most immediate data are the world-public alert packets that will be distributed within minutes of shutter-close, including difference-image detections of transient and variable objects along with associated postage stamps and (for repeat detections) a 1-year time series. “Prompt products” will be released to the LSST data-rights community after 24 hours (for catalogs) and 80 hours (for full focal plane images). On a longer cadence, uniform reprocessing and coaddition across all epochs will deliver annual data releases of catalogs and images. The alert stream will be distributed via the network of LSST Community Brokers, while the prompt products and annual data releases will be accessible via the RSP and also available for bulk download to DESC via the Rubin USDF at SLAC. A workflow based on the Rucio data management software has been implemented to mirror LSST data to NERSC from the USDF, and could be employed for staging data at ALCF and OLCF as well.

Other major data sets of interest for cross-match, co-analysis, and multi-modal FM training in conjunction with Rubin data include: space-based surveys such as *Roman*, *Euclid*, the *Spectro-Photometer for the History of the Universe, Epoch of Reionization, and Ices Explorer* (*SPHEREx*), and the *Wide-field Infrared Survey Explorer* (*WISE*); spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS), DESI, and 4MOST; precursor imaging and time-domain surveys such as DES, the Dark Energy Camera Legacy

[17] A Statement of Community Engagement and Needs for Digital Research Infrastructure; see https://engagementhub.ukri.org/ukri-infrastructure/ascend-process/

[18] http://ls.st/RDO-013

Survey (DECaLS), Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), and Zwicky Transient Facility (ZTF); and CMB data from facilities such as Planck, ACT, the South Pole Telescope (SPT), and Simons Observatory (SO).

Given the multi-petabyte size of the LSST data (and of simulations and other survey data sets on similar scales), both network transfer and disk storage will be limiting factors. DESC would likely benefit from strategic coordination with other LSST science collaborations, LINCC Frameworks (see Section 7), and IDACs regarding which LSST data products are mirrored where, for how long, at what quality of service, in conjunction with which CPU and GPU allocations, and for which analysis purposes.

The current Rubin Data Management system was not primarily designed for large-scale AI/ML work. Hence the data will need to be fitted with additional data interfaces, APIs, and standards that enable efficient use in this new context, such as the following examples:

- Adoption of tools like the Hyrax framework[19] that provides modular components for a full AI/ML workflow tailored to astrophysical data.
- Large-scale cross-match capabilities such as the Hierarchical Adaptive Tiling Scheme (HATS)[20] and Fink Xmatch[21] that are critical to multi-modal dataset construction.
- Performant and scalable services for streaming large batches of image cutouts into AI/ML training and inference workflows.
- Data tokenization and embedding strategies that are well matched to AI/ML model architectures and downstream science tasks.
- Active learning frameworks for alerts and images that maximize the value of limited human expert labeling time with respect to relevant modeling objectives.

In some cases, standards developed by the International Virtual Observatory Alliance[22] may be fit for these purposes in DESC although they are typically conceived around classical astronomy use cases that do not map onto survey-scale AI/ML.

6.4. Benchmarking and Reproducibility

Challenges of reproducibility will only increase as AI/ML-based analyses become more common. Full computational reproducibility requires infrastructure for versioned retention of all input data, any pre-trained models used for inference, all analysis code and frameworks, and all software and environment dependencies. Solutions that allow for some or all of the above elements to be mutable over time fall short of true reproducibility but may be acceptable (or even preferable) to the extent that significant scientific conclusions remain replicable.

Additional challenges posed by the increasing adoption of AI/ML methods include:

[19] https://hyrax.readthedocs.io/en/stable/
[20] https://hats.readthedocs.io/en/latest/
[21] https://fink-portal.org/xmatch
[22] https://www.ivoa.net

- Defining the role and framework of blind analysis.
- Maintaining independence of different experiments in the context of multi-modal FM-based analyses.
- Defining training, validation, and test data sets when we ultimately want to use the full data set for cosmological inference.
- Ensuring that any agentic software used or developed to generate DESC science records the provenance of the operations done within its task in ways that are compatible with scientific standards of reproducibility (explainable AI).

While the above are primarily questions of methodology rather than infrastructure, their solutions will have implications for infrastructure requirements. Some of the required infrastructure elements may include

- Persistent, accessible storage for testbed datasets, SBI training simulations, and pretrained model weights.
- Hosted frameworks for deploying and running against science benchmarks.
- Minified production environments that facilitate use of small-scale development instances to develop methods and benchmarks.
- Standardized APIs and architectures for reproducible large-scale model training and deployment.
- Agentic frameworks for analysis reproduction (e.g., Ye et al. 2025)
- Comprehensive provenance tracking & support for long-term co-archiving of data and analysis.

Our traditional thinking about reproducibility will be further challenged if LLMs and AI agents continue to move us toward a more natural-language approach to dynamically and interactively extracting understanding from data. This trend can bring about a shift away from the traditionally sequential and siloed model of "data, pipeline, results as papers" and toward a more dynamic and interactive world characterized by prompts such as "I want to regenerate the plot in Figure 1 of so-and-so's paper, except I want the magnitude cut at $g = 22$ rather than $g = 23$..." Other disciplines will be experiencing similar transitions and DESC should look for opportunities both to lead by example and to benefit from broader trends and investment.

The vast data scale of LSST has necessitated a fundamental evolution in scientific methodology. This paradigm shift moves from local data processing on individual researchers' computers toward analytical tools designed to operate on shared, high-performance compute platforms. This centralization, combined with the anticipated widespread adoption of general-purpose foundation models, compels the establishment of a common implementation framework to ensure scientific analyses are reproducible, consistent, and interoperable.

While the RSP serves as the primary portal for LSST data access, it was not designed for resource-intensive, large-scale AI applications. To bridge this gap, the Hyrax framework provides a unified platform for exploring the latent space of foundation models. However, their operational deployment remains a significant undertaking that requires dedicated computational resources and a community-governed process for selection and validation.

The rapid pace of innovation in ML means that FMs cannot be considered static, long-term solutions. They must be constructed for a dynamic life cycle that includes systematic processes for review, evaluation, and replacement. This agile strategy is critical for all modalities, especially for time-series models deployed on real-time alert brokers, which must reliably process transient astronomical events to enable rapid discovery.

7. Opportunities for Broader AI/ML Coordination

DESC does not operate in isolation. The broader scientific impact of AI/ML-enabled analysis will depend critically on how well we coordinate with the rest of the Rubin ecosystem, other Stage-IV surveys, AI institutes, and large-scale compute providers. Many of these connections already exist in the form of shared personnel, joint projects, or informal collaborations. Here, we outline a non-exhaustive snapshot of this landscape and highlight opportunities to deepen and systematize these links. This section should be read as a living document: the specific institutes, infrastructures, and programs will evolve over the LSST decade, but the underlying goal will remain positioning DESC as both a demanding scientific user and an influential driver of AI methodologies for fundamental science.

Coordination across the Rubin Community—The Rubin LSST survey provides the most immediate opportunity for DESC coordination. LINCC Frameworks, funded and supported by Schmidt Futures, prioritizes open-source and cross-project infrastructure for faint object detection, time-series data analysis, and photometric redshift estimation. Beyond producing intermediate Rubin data products that will be used in targeted ML pipelines, LINCC will serve a key role in maintaining a software ecosystem which has the potential to substantially accelerate the development of large-scale data foundation models and language models for science. DESC efforts in the initial years of the LSST survey can benefit LINCC Frameworks in optimizing its infrastructure toward these goals: for example, by stress-testing the throughput and accuracy of newly released pipelines and providing benchmark scientific datasets that LINCC can use to develop new detection and analysis algorithms.

In parallel with DESC, other LSST Science Collaborations are advancing AI methods for science, and coordinated work would accelerate progress. The ISSC brings 150+ scientists from both academic and industry positions to shared discussions on large-scale data analysis with LSST. The ISSC could run independent audits of DESC photo-z, shear, and transient pipelines and partner with DESC to explore methodological advancements in AI/ML (Section 4). The TVSSC actively develops time-series analysis tools, and could partner with DESC in stress-testing broker infrastructure and classification benchmarks to evaluate cosmological readiness using photometric LSST samples. The GSC collaboration could co-develop deblending and morphology benchmarks, in order to prevent biases in cosmological inference from mischaracterized shear and clustering constraints. The Stars, Milky Way, and Local Volume Science Collaboration (SMWLVSC) analysis of crowded Milky Way fields will stress test deblending and photometric calibration, while the Solar System Science Collaboration (SSSC) identification of solar-system objects and bogus alerts should inform the use in DESC of image products in large-scale scientific models. Across these collaborations, DESC should also encourage reproducible RSP notebooks, tagged releases, and quarterly cross-collaboration readiness reviews in which independent teams reproduce results and report failure modes, potentially through the Rubin Community Forum[23].

Coordination of Stage IV Experiments—Coordination between DESC (on behalf of the Rubin LSST) and other Stage-IV cosmological experiments will create additional opportunities for mitigating cross-survey systematics and optimizing the cosmological yield of LSST data.

[23] https://community.lsst.org/

DESI Data Release 2 is expected to contain ∼18.7M spectra across 4,000 deg^2 of overlap with the LSST footprint. This spectroscopic sample can be used as a primary calibrator for LSST redshift and lensing systematics. DESC should use DESI's public releases to improve training sets for photometric-redshift models, require uncertainty coverage checks against those sets, and deploy clustering-redshift cross-correlations to validate $n(z)$ across tomographic bins. For weak lensing, DESC could cross-correlate LSST shear measurements with DESI density fields and test magnification and selection effects with controlled changes in DESI target completeness and fiber-assignment weights. The 4MOST TiDES program (Frohmaier et al. 2025) will provide $\sim 30,000$ spectroscopic transients. These data will enable real-time validation of classification algorithms and a sub-2% measurement of the dark energy equation of state. The TiDES survey will also produce >200,000 spectroscopically-confirmed transient host galaxies, providing valuable contextual information for characterizing and marginalizing over environmental differences when curating standardizable SN Ia samples. DESC could leverage these correlations to improve its survey simulations of the extragalactic time-domain sky in successors of the PLAsTiCC and ELAsTiCC challenges. Coordination within the Roman Space Telescope presents possibly the most transformative opportunity for DESC. The OpenUniverse2024 simulations (∼70 deg^2, ∼400 TB publicly available; OpenUniverse et al. 2025) enable immediate testing of how Roman's infrared imaging and superior resolution can be used for full multi-wavelength characterization of static sources in a joint data foundation model. Roman will also reveal blends invisible in Rubin data, allowing for validation of existing deblending/image segmentation methods for Rubin. Characterization with the high-redshift SN Ia population in Roman will also provide a laboratory for exploring any redshift-dependent systematics (e.g., changes in progenitor properties across cosmic time) that should be included in DESC cosmological analysis pipelines. In addition, Schmidt Sciences announced in January 2026 the Eric and Wendy Schmidt Observatory System, a privately funded "system-of-observatories" designed for open-access time-domain and multi-messenger science complementary to LSST. The system comprises four facilities: the Argus Array (Law et al. 2022), a ∼900-telescope optical array delivering ∼8,000 deg^2 instantaneous field of view with cadences down to ∼1 s; the Deep Synoptic Array (DSA; Hallinan et al. 2019), a 1650×6.15 m dish radio interferometer spanning 0.7–2 GHz with real-time imaging; the Large Fiber Array Spectroscopic Telescope (LFAST; Berkson et al. 2024), a scalable fiber-fed array of 0.76 m unit telescopes targeting ELT-class collecting area for photon-starved spectroscopy and rapid follow-up; and the Lazuli Space Observatory (Roy et al. 2026), a 3 m rapid-response optical–NIR facility (400–1700 nm) in lunar-resonant orbit with a wide-field imager and integral-field spectrograph capable of responding to targets of opportunity in <4 hours. With planned operations beginning as early as 2029 and a commitment to open data and shared analysis tools, this privately funded infrastructure could provide valuable cross-wavelength and high-cadence coverage for DESC time-domain and multi-messenger science, particularly for SN Ia cosmology and transient follow-up. As private investment in astronomical infrastructure grows, DESC should monitor these developments for coordination opportunities.

Coordination with AI Institutes—Two NSF-Simons AI institutes have been launched as of September 2024, with funding and explicit scientific themes targeting LSST and cosmology. The SkAI Institute between Northwestern, University of Illinois, and University of Chicago, is developing an FM for transient science that can serve as a precursor to upcoming DESC models. CosmicAI at the University of Texas at Austin is developing LLM-powered AI "copilots" for research. These efforts create opportunities for DESC members to identify DESC pipelines most amenable to automation and provide CosmicAI with datasets for beta testing of their models. Across MIT, Harvard, Northeastern, and Tufts Universities, the NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI) has also explored the use of generative models for field-level inference and multi-modal foundation models for transient science, which DESC can validate with synthetic

Rubin datasets such as CosmoDC2 (Korytov et al. 2019) and PLAsTiCC/ELAsTiCC (PLAsTiCC team et al. 2018; Narayan & ELAsTiCC Team 2023). Further, the focus of DESC on AI integration at multiple stages of data processing will benefit the efforts of these institutes in incorporating realistic atmospheric effects and detector systematics directly into model architectures.

A feasible path toward training generalizable FMs for cosmology is to split the work. Models could be prototyped at individual universities or AI Institutes with support from LINCC Frameworks, and scaled through the pre-training of backbones on national and European supercomputers (pooled across DOE and EuroHPC facilities), since this will likely require hundreds of GPUs and training across multiple days. The models could then be fine-tuned and calibrated near the data on DESC computing facilities such asNERSC, with brokers providing smaller fine-tuned heads as software filters for targeted streaming objectives.

A primary bottleneck to achieving this widespread scientific coordination is the development of robust, well-documented, and well-maintained software infrastructure. LINCC Frameworks, supported by the Schmidt Sciences, provides this support for the Rubin Observatory LSST, but this should be equally supported across all major observatories and international collaborations such as DESC in the coming years to enable the emerging technologies outlined in Section 5.

Coordination with European Networks—EuCAIF coordinates AI infrastructure and research across European institutions, and has produced white papers on infrastructure needs (e.g., Caron et al. 2025), and LLMs/FMs (Barman et al. 2025). DESC members at European institutions can, for instance, join EuCAIF WG4 (machine learning and artificial intelligence infrastructure) to contribute cosmology-specific challenges to EuCAIF's methods repository, or WG1 (foundation models & discovery) to coordinate the development of FMs. These connections may facilitate successful applications for EuroHPC resources where DESC could test whether cutting-edge architectures will scale to LSST volumes. EuroHPC systems (Leonardo with 240 petaflops on NVIDIA A100 GPUs, LUMI with 380 petaflops on AMD MI250x GPUs, or the exascale JUPITER with NVIDIA GH200 superchips) enable training foundation models on billions of galaxy images, computationally infeasible on current NERSC allocations. These systems are already being deployed in astrophysics as a testbed for exascale and GPU-optimized implementations of simulation codes (see e.g. Shukla et al. 2025; Lacopo et al. 2026). EuroHPC access follows a staged pathway from Benchmark (testing code scaling) to Development (algorithm validation) to Extreme Scale (production runs of up to 8M GPU hours). DESC could pursue Benchmark Access to validate early algorithms before committing to larger allocations.

Collaborations with Industry—Tech partnerships can provide expertise, computational resources, and opportunities to stress-test DESC methods at scale. NVIDIA's Academic Grant Program, along with complementary access through Google Cloud and Amazon Web Services, could allow DESC to rapidly prototype architectures and objective functions for foundation models at LSST scale.

Partnerships between DESC and LLM providers (e.g., Anthropic, OpenAI) should also be encouraged. Research credits would allow DESC to simultaneously explore the strengths and failure modes of the current generation of models. This compute could also be used to conduct systematic benchmarking of these models (through, e.g., HuggingFace) on targeted, science-specific use-cases. Any formal arrangement for LLM use across DESC would need to comply with LSST data rights policies (e.g., through private

networking, complete audit trails, and explicit no-train/no-retain clauses). Such an arrangement would yield reproducible evaluation suites that could serve as a case study of language model readiness for science applications.

Rubin Alert Brokers—The seven full-stream Rubin alert brokers are ALeRCE (Förster et al. 2021), AMPEL (Nordin et al. 2019), ANTARES (Matheson et al. 2021), Babamul (Jegou du Laz et al. 2025), Fink (Möller et al. 2021), Lasair (Smith et al. 2019), and Pitt-Google. These systems provide the primary filtering layer between Rubin streams and science-specific transient samples, turning raw alerts into ranked candidates, host associations, and early labels that will drive spectroscopic follow up and downstream analyses. Tight coordination with these brokers will give DESC direct leverage over the quantities that contribute to cosmological systematics: the completeness and purity of SN Ia samples, characterization of selection effects, and calibration of host-galaxy priors. Characterizing the selection functions of alert brokers as part of the analysis pipeline will help align transient discovery with the DESC requirements.

Brokers should maintain rigorous provenance tracking for all derived data features, host-galaxy associations, and classification/anomaly scores so that DESC can understand the selection effects of deployed algorithms. In return, collaboration with DESC can provide the alert brokers with benchmark datasets and the targeted science objectives used to validate their infrastructure and foster additional software development. Algorithms developed in the early years of the Rubin LSST can be ported upstream to broker environments after public release, providing additional metadata (e.g., embeddings from a data foundation model or concise, text-based descriptions of a subset of high-priority alerts) and allowing the broader scientific community and all Science Collaborations to benefit from DESC efforts without violating LSST data rights policies.

8. Risks, Challenges, and Mitigation Strategies for AI/ML in DESC

The increasing reliance on AI/ML within DESC brings not only opportunities but also a set of technical, organizational, and societal risks that must be managed deliberately. The aim is not to discourage the use of AI, but to ensure that methods are deployed in ways that are scientifically robust, sustainable over the LSST decade of operations, compatible with DESC's standards for transparency and reproducibility and with the broader scientific and educational aims of DESC. Below we highlight key challenge areas together with concrete mitigation strategies.

Methodological Robustness and Interpretability—AI/ML models trained directly on data are vulnerable to familiar statistical pitfalls: biased or incomplete training data, overfitting (Huppenkothen et al. 2023; Smith & Geach 2023), and domain mismatch between simulations and real survey data (Ćiprijanović et al. 2023; Pandya et al. 2025a). For methods that sit close to top-level cosmological inferences, there is a justified reluctance to adopt "black-box" results as reference constraints unless they can be thoroughly validated and stress-tested. This is amplified for neural summary statistics and learned emulators (Villaescusa-Navarro et al. 2022; Huppenkothen et al. 2023; Smith & Geach 2023), where it can be difficult to diagnose failure modes or to construct transparent null tests. There is also a subtle "ML-only" risk: some future analyses will be so data- and compute-intensive that no independent non-ML cross-check is feasible, making it even more important that AI-based pipelines be internally well understood and stress-tested.
From a DESC perspective, mitigation rests on *explicit validation and interpretability practices*: (i) defining standardized simulation and challenge suites where AI and traditional methods are compared head-to-head in the regime where both are applicable; (ii) publishing diagnostic plots and ablation studies that isolate which data features drive the constraints; (iii) requiring explicit documentation of model training domains, assumptions, and known failure modes, and discouraging use outside those regimes; (iv) developing approximate surrogate models or interpretable summaries (e.g., response functions, feature attributions) that can be inspected by domain experts; and (v) encouraging redundancy where it matters most—for example, using different architectures, loss functions, or simulation pipelines to cross-check key inferences, even when all are "AI-based". Any AI-based result used as a reference cosmological constraint should be accompanied by a documented validation program and, where possible, benchmarked against simpler baselines. We recognize that imposing these higher standards carries a non-negligible human effort cost, which in turn creates a natural opportunity for agentic AI systems to automate parts of the validation workflow and thereby reduce that burden.

Provenance, Reproducibility, and Integration into Pipelines—As analyses become more complex, certifying full provenance (from raw data through simulations, training runs, and model selection) becomes harder but more important. If DESC cosmology results depend on opaque training pipelines, unversioned models, or undocumented hyperparameters, small bugs or biases can consume a non-negligible fraction of the "tension budget" in precision tests of ΛCDM. While these issues are common to most long-term software infrastructures, the rapid pace of methodological advancement in AI/ML hinders provenance tracking, and the stochastic nature of most neural network training can prevent complete reproducibility. There is also the practical challenge of integrating AI components into mature pipelines that already rely on well-tested codes.
Mitigation here is largely infrastructural and procedural: (i) treat trained models as first-class data products, with versioning, metadata, and model cards describing training data, objectives, and known limitations; (ii)

require that AI components be runnable from containerized environments and integrated into CI pipelines with regression tests; (iii) maintain "shadow" implementations (simpler, slower, or more traditional pipelines) for cross-checks where feasible; and (iv) define clear deprecation and maintenance policies so that AI dependencies do not become unmaintainable over the survey lifetime.

Safe Usage, Data Rights, and External Services—Widespread availability of commercial LLMs and AI services lowers the barrier to experimentation but introduces new questions about data governance and safe usage. Uploading proprietary Rubin/DESC data or unpublished results to third-party services may raise data-rights concerns, and using off-the-shelf models without understanding their limitations can encourage application of techniques outside their domain of validity.
DESC can mitigate these issues by (i) establishing clear guidelines on what kinds of data and metadata may be used with external services, in coordination with Rubin, LSST Discovery Alliance (LSST-DA), and agency policies; (ii) prioritizing self-hosted or collaboration-controlled deployments (e.g., for LLMs and inference services) for sensitive workloads; and (iii) prioritizing services committed to long term availability of their AI models and transparent versioning of these models, for long term reproducibility.

Human Capital, Training, and Sustainability—AI/ML tools (and, increasingly, agentic assistants) can accelerate research by automating repetitive tasks and lowering the barrier to entry for complex workflows. However, there is a real risk that early-career researchers learn to operate pipelines as "black boxes" or solely by prompt engineering, without acquiring a deep understanding of the underlying statistics and physics, and thus weakening long-term scientific literacy and even cognitive abilities (Kosmyna et al. 2025; Trotta 2025). DESC can turn this into an opportunity by (i) framing AI/ML training as an integral part of graduate and postdoctoral education, combining hands-on use of tools with explicit coverage of underlying concepts; (ii) pairing students with mentors who can help them "open the box" at least once (e.g., by reproducing a result from scratch or implementing a simplified version of a method); (iii) encouraging contributions to shared, well-documented libraries rather than bespoke one-off scripts, spreading maintenance across the collaboration and ensuring that successful methods become communal assets; and (iv) explicitly furthering a scientific culture that does not unduly favor efficiency and speed over in-depth engagement with modeling and creative thinking.

Environmental Impact—Large-scale adoption of AI/ML methods comes with significant environmental impact: data centers and computing farms are energy intensive, water hungry, and have generally a negative environmental impact in terms of land usage, noise and e-waste production. The first step toward an environmentally sustainable practice for DESC is to quantify the computing resources used in AI/ML development. This should cover not just the training, validation and deployment costs for methods presented in a research paper (as is common practice), but also track the computational expenditure for hyperparameter search, failed tests, aborted training runs and architectural dead-ends – which often dwarf the compute needed for the ultimate implementation.
To this end, DESC needs to develop easy-to-use, adaptable and constantly reviewed guidelines regarding how to log and report transparently such computational costs, so as to monitor compute usage in DESC over the lifetime of LSST. Taking advantage of more energy-friendly implementations (both software and hardware) should also be encouraged, while training on the matter should be available to all DESC members.
Detailed evaluations of astronomy-specific activities in terms of their carbon footprints are scarce, and usu-

ally limited to research infrastructure (see, e.g., Knödlseder 2025). Stevens et al. (2020) find that even before the rise of AI/ML, supercomputer usage already accounted for the majority of Australian astronomers' carbon footprint. However, compute-related carbon emissions are but one aspect of the multi-faceted environmental impact of AI/ML, which remains understudied. It would be desirable to address this gap by developing in-depth evaluations of the wider environmental cost of LSST-related AI/ML usage, as a way to support sustainability alongside scientific gains (Bashir et al. 2024).

Computing Resources and Infrastructure—Realizing the full potential of AI/ML within DESC will place non-trivial demands on computing resources. Training large foundation models for images, catalogs, or time series, and running large-scale simulation-based inference, require sustained access to GPU clusters at a scale beyond traditional analysis workloads. Similarly, if DESC wishes to host its own LLMs or other generative models for work involving sensitive or proprietary data, these services will need reliable, secure GPU backends and operational support over many years. Without careful planning, AI workloads risk competing destructively with other science uses for scarce accelerators, or fragmenting into ad hoc deployments that are hard to maintain.
Mitigation here is primarily strategic: (i) aligning major AI training campaigns with DESC's existing resource-allocation processes and external partners (e.g., LSST-DA, national and international HPC centers); (ii) prioritizing shared, reusable models and services over one-off experiments; (iii) investing in efficient training and inference schemes (e.g., parameter-efficient finetuning, mixed precision, model distillation); and (iv) treating any self-hosted LLM or foundation-model service as collaboration infrastructure, with clear policies on access, data governance, and long-term support.

Overall, the main risks associated with AI/ML in DESC are not existential but *operational*: biases that are hard to diagnose, results that are difficult to reproduce, methods that are fragile under domain shift, and human capital that is either over- or under-reliant on automation. By treating AI components with the same methodological rigor as any other part of the cosmology pipeline requiring validation, documentation, governance, and training, DESC can reap the benefits of these tools while keeping these risks manageable.

9. Summary and Conclusion

The Vera C. Rubin Observatory LSST will generate heterogeneous data at a scale and complexity that strain traditional analysis pipelines. DESC's mission is to convert these data into robust constraints on the dark sector, which demands methods that are statistically powerful, scalable, and operationally reliable. AI/ML, from NDE for photo-zs to SBI and generative models for field-level cosmology, have *already* demonstrated that they can address key bottlenecks in this program. At the same time, their utility for precision cosmology hinges on trustworthy UQ, explicit treatment of model misspecification and covariate shift, and fully reproducible integration into DESC workflows.

Section 3 and Section 4 demonstrate that DESC is at the forefront of cutting-edge machine learning applications in astronomy. Research into machine learning is now integral to the primary LSST cosmological probes—including strong and weak lensing, supernovae, galaxy clusters, and large-scale structure—as well as cross-cutting topics such as theory, photometric redshifts, simulations, and deblending. Across DESC working groups and the broader cosmology community, several critical themes and methodologies have crystallized:

- **Simulation-Based Inference (SBI):** SBI has emerged as a powerful methodology, enabling analyses of a complexity that typically exceeds the capabilities of traditional forward modeling. This domain offers fertile ground for machine learning research, particularly in the development of emulators to accelerate pipeline components and in extending analyses beyond traditional point statistics. However, SBI remains sensitive to model misspecification, lossy summaries and inaccurate posterior approximations, problems which are particularly challenging to solve in a machine learning paradigm.
- **Bayesian Methodology and Uncertainty Quantification (UQ):** While Bayesian frameworks are ubiquitous in cosmology, machine learning is increasingly being explored to accelerate Bayesian inference on LSST-scale datasets that would otherwise be computationally intractable. Furthermore, the high precision required by cosmology requires accurate uncertainty estimates that go beyond common practice in machine learning. Building on the application of Bayesian neural networks and related methods, DESC is well positioned to drive fundamental developments in this area.
- **Validation and Benchmarking:** For cosmology, rigorous validation is paramount. Algorithms must not only be accurate and unbiased but also capable of correctly propagating uncertainty. Covariate shift, inevitable in many supervised learning contexts, must be mitigated through accurate simulations and techniques for domain adaptation. Benchmarking and validation are particularly important for algorithms used in products intended for broad usage, such as FMs and simulations. The RAIL project (see Section 3.1), developed by DESC specifically to benchmark photometric redshift algorithms, serves as an excellent model for such validation frameworks.
- **Active Learning and Discovery:** Active learning has become an essential part of machine learning and will be crucial in managing LSST data. The RESSPECT project (see Section 3.5), a collaborative initiative developing an active learning pipeline for transient classification, is an example of the comprehensive infrastructure required for effective active learning. Furthermore, human-in-the-loop workflows will be vital for anomaly detection and the identification of rare phenomena within the vast LSST dataset, facilitating novel discoveries that purely automated systems might overlook.

Realizing this potential requires DESC to treat AI/ML as primary components of the measurement pipeline. Section 6–8 of this paper outline the software, computing, and data infrastructure required to support AI/ML at scale. Sustainable integration of emerging tools requires a shared AI software stack, containerized and RSP-compatible workflows, a DESC Data Registry for model and data products, and benchmark suites that tie model performance directly to cosmological and systematic-error budgets (see Section 6). These methods also present opportunities for broader coordination with Rubin operations, community brokers, external AI/ML institutes, and industry, which we outline in Section 7, but present risks ranging from model miscalibration and covariate shift to data rights compliance, environmental cost, and the erosion of human oversight (see Section 7).

On the basis of these insights, we have defined a series of recommendations (R1–R15) and opportunities (O1–O5) in the Executive Summary (Section 1), spanning methodological research, foundation models, LLMs and agentic AI, infrastructure and software, organizational coordination, human capital, and external partnerships. Implementing these recommendations would position DESC to use AI/ML for ambitious and disciplined science. LSST-era cosmology will be limited not by a lack of algorithms, but by our ability to validate, govern, and integrate them. By investing in that infrastructure now, DESC can shape how AI is used for precision cosmology and set a standard for its responsible deployment across the physical sciences.

A. Index of AI/ML Methodologies and Challenges

AI/ML Methods Index

Bayesian hierarchical modeling: Hierarchical probabilistic models that share information across objects or populations. 18, 24, 27, 28, 42

Deep ensembles: Collections of models combined for improved performance and UQ. 28, 35

Gaussian processes: Non-parametric Bayesian regression and emulation for scalar or functional outputs. 18, 21, 28, 32, 42

Neural density estimation (NDE): Neural networks trained to approximate probability densities, often used within SBI. 18

Neural posterior estimation (NPE): Direct neural approximation to the posterior distribution over parameters given simulated data. 36, 45

Simulation-based inference (SBI): Implicit-likelihood Bayesian inference using forward simulations and NDE techniques. 21, 24, 27, 28, 32, 33, 38, 45, 47

Variational inference (VI): Optimization-based approach to approximating posterior distributions. 21, 42

Convolutional neural networks (CNNs): Convolution-based neural networks for image-like data, widely used for classification and regression. 21, 27, 36

Deep neural networks: Generic deep learning architectures not otherwise categorized. 38

Graph neural networks (GNNs): Neural networks designed to operate on graph-structured data, such as cosmic webs or halo catalogs. 33, 35

Recurrent neural networks (RNNs): Sequence models (e.g. LSTMs, GRUs) for time series and light curves. 21, 28

Transformers: Attention-based neural architectures for sequences, time series, or generic sets of tokens (including vision transformers). 18, 21, 28, 35

Diffusion / score-based models: Generative models that iteratively denoise data from a noise process, often used for images and fields. 18, 21, 24, 33, 36, 49

Emulators: Surrogate models (often Gaussian process- or neural network-based) that approximate expensive simulations or likelihoods. 18, 32, 33

Neural surrogates: Neural network models trained to mimic the input–output behavior of complex physical simulations. 18, 32, 49

Normalizing flows: Invertible neural networks used as flexible density models in likelihoods, posteriors, or simulators. 36, 45, 49

Variational autoencoders (VAEs): Latent-variable generative models trained via VI. 28, 36

Differentiable programming: Formulation of simulations and models as differentiable programs amenable to gradient-based inference. 24, 32, 33, 38, 49

Physics-informed ML: Machine learning models that incorporate physical laws or constraints (e.g., PINNs, ENNs). 50

Symbolic regression: Learning analytic mathematical expressions that describe data or physical relationships. 32

Active learning: Strategies that adaptively select the most informative data points for labeling or follow-up. 28, 35, 51

Anomaly detection: Identifying rare or out-of-distribution examples in data, crucial for new physics discovery. 28, 51

Neural compression: Learned low-dimensional representations (summary statistics or latents) of complex data. 24

Self-organizing maps (SOMs): Topology-preserving maps used for exploratory analysis and domain coverage characterization. 18, 36

Self-supervised learning: Learning representations from unlabeled data by solving pretext tasks (e.g., masking, contrastive learning). 35, 51

Instance segmentation: Pixel-level segmentation of individual objects, relevant for deblending crowded scenes. 36

Object detection: Identifying and localizing objects in images (e.g., YOLO). 27, 36

Cross-cutting Challenges Index

Covariate Shifts: Distribution mismatches between training and target data (e.g. spectroscopic selection bias, sim-to-real gaps, model misspecification). Addressed in Section 4.1.3, Section 4.2.1, and Section 4.3. 18, 21, 24, 27, 28, 32, 33, 35, 38, 47, 49, 50, 51

Metrics & Evaluation: Task-relevant metrics, validation frameworks, benchmarking, and stress tests for DESC science. Addressed in Section 4.1.4 and Section 4.2.2. 18, 33, 36, 48, 50

Scalability: Handling LSST-scale data volumes, real-time alert processing, and high-dimensional inference. Addressed in Section 4.1.1, Section 4.1.2, and Section 4.2.1. 18, 24, 27, 28, 32, 33, 35, 38, 42, 45, 49

Data Sparsity & Rare Events: Limited labeled samples, rare transients, class imbalance, and challenging edge cases like blending. Addressed in Section 4.3. 21, 28, 35, 36, 38, 51

Uncertainty Quantification: Obtaining well-calibrated posteriors and propagating uncertainties to cosmological constraints. Addressed in Section 4.1.1, Section 4.1.2, and Section 4.1.4. 18, 21, 24, 27, 28, 35, 36, 38, 42, 45, 48

B. Glossary of Acronyms

Acronyms and Abbreviations

ΛCDM: cold dark matter with cosmological constant. 80

wCDM: cold dark matter with cosmological equation of state. 25, 32

4MOST: 4-meter Multi-Object Spectroscopic Telescope. 11, 16, 30, 57, 72, 77

ACT: Atacama Cosmology Telescope. 61, 73

ADQL: Astronomical Data Query Language. 61, 63

ADS: SAO Astrophysics Data System. 60

AGN: active galactic nuclei. 22

AGNSC: Active Galactic Nuclei Science Collaboration. 11

AI: artificial intelligence. 8, 13, 18, 21, 22, 30, 32, 34, 35, 37, 40, 41, 49, 50, 51, 53, 59, 61, 63, 64, 65, 66, 67, 70, 72, 73, 76, 80, 83

ALCF: Argonne Leadership Computing Facility. 16, 71, 72

ALeRCE: Automatic Learning for the Rapid Classification of Events. 31, 79

ALMA: Atacama Large Millimeter Array. 23

AMPEL: Alert Management, Photometry, and Evaluation of Light curves. 31, 79

AmSC: American Science Cloud. 10, 16, 71

AnaCal: Analytic Calibration. 38

ANTARES: Arizona–NOIRLab Temporal Analysis and Response to Events System. 31, 79

API: application programming interface. 16, 21, 33, 59, 60, 66, 68, 69, 73, 74

ASCEND: A Statement of Community Engagement and Needs for Digital Research Infrastructure. 72

ASCR: Advanced Scientific Computing Research. 71

BLISS: Bayesian Light Source Separator. 37

BNN: Bayesian neural network. 23, 29, 46

CAMB: Code for Anisotropies in the Microwave Background. 33

CANDELS: Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey. 37

CC-IN2P3: IN2P3 Computing Centre. 67

CCD: charge-coupled device. 57

CD: continuous delivery/deployment. 68

CI: continuous integration. 63, 68, 81

CIGALE: Code Investigating GALaxy Evolution. 61

CLMM: Cluster Lens Mass Modeling tool. 27

CMB: cosmic microwave background. 22, 73

CMNN: color-matched nearest-neighbors. 19

CNN: convolutional neural network. 22, 24, 31, 37

CosmoDC2: Cosmological Data Challenge 2. 9, 78

CPU: central processing unit. 70, 71, 73

CSD3: Cambridge Service for Data Driven Discovery. 68

CSP: Carnegie Supernova Project. 44

DECaLS: Dark Energy Camera Legacy Survey. 72

DeepDISC: Detection, Instance Segmentation, and Classification with Deep Learning. 19, 37, 39

DES: Dark Energy Survey. 20, 23, 26, 29, 36, 72

DESC: Dark Energy Science Collaboration. 8, 13, 18, 22, 26, 27, 28, 32, 35, 36, 38, 40, 41, 43, 48, 51, 53, 54, 59, 61, 62, 63, 64, 65, 66, 67, 70, 72, 74, 76, 80, 83

DESI: Dark Energy Spectroscopic Instrument. 11, 16, 53, 57, 72, 77

DNF: directional neighborhood fitting. 19

DOE: Department of Energy. 10, 67, 71, 78

DSPS: Differentiable Stellar Population Synthesis. 34

ELAsTiCC: Extended LSST Astronomical Time Series Classification Challenge. 9, 30, 48, 77

ELLIS: European Laboratory for Learning and Intelligent Systems. 12, 16

ENN: equivariant neural network. 51, 85

ESA: European Space Agency. 11, 29

ESNet: Energy Sciences Network. 71

EuCAIF: European Coalition for AI for Fundamental Physics. 12, 16, 78

EuroHPC: European High-Performance Computing Joint Undertaking. 10, 16, 78

FM: foundation model. 15, 53, 54, 55, 56, 57, 58, 70, 75, 77, 83

FSPS: Flexible Stellar Population Synthesis. 20

GGNS: gradient-guided nested sampling. 43

GNN: graph neural network. 36, 37

GPU: graphics processing unit. 16, 26, 34, 38, 43, 50, 68, 70, 71, 72, 73, 78, 82

GPz: Gaussian Processes for Photo-z. 19

GRU: gated recurrent unit. 30, 85

GSC: Galaxies Science Collaboration. 11, 76

HATS: Hierarchical Adaptive Tiling Scheme. 73

HMC: Hamiltonian Monte Carlo. 33, 42

HOD: halo occupancy distribution. 35

HPC: high-performance computing. 68, 82

HPDF: High Performance Data Facility. 16, 71

HSC: Hyper Suprime-Cam. 22

***HST*:** *Hubble Space Telescope*. 22, 37, 53

IA: Intrinsic Alignment. 25, 35

IAIFI: NSF Institute for Artificial Intelligence and Fundamental Interactions. 77

IDAC: Independent Data Access Center. 10, 16, 71, 73

IID: independent and identically distributed. 58

IMNN: information-maximizing neural network. 46

IN2P3: Institut National de Physique Nucléaire et de Physique des Particules. 67, 87

IR: infrared. 21, 34

IRI: Integrated Research Infrastructure. 71

ISSC: Informatics and Statistics Science Collaboration. 11, 76

JUPITER: Joint Undertaking Pioneer for Innovative and Transformative Exascale Research. 10, 78

***JWST*:** *James Webb Space Telescope*. 53

KiDS: Kilo Degree Survey. 20, 26

KNe: kilonovae. 29, 48

kNN: k-nearest neighbors. 19

LCCF: Leadership Class Computing Facility. 71

LINCC Frameworks: LSST Interdisciplinary Network for Collaboration and Computing Frameworks. 9, 19, 73, 76

LLM: large language model. 10, 18, 53, 59, 60, 61, 62, 64, 65, 66, 68, 69, 70, 74, 77, 78, 81

LMC: Langevin Monte Carlo. 42

LSST: Legacy Survey of Space and Time. 8, 13, 18, 21, 22, 25, 27, 28, 32, 33, 35, 36, 38, 41, 42, 47, 51, 53, 54, 62, 63, 65, 66, 67, 68, 70, 72, 74, 76, 80, 83

LSST-DA: LSST Discovery Alliance. 81

LSTM: long short-term memory. 24, 85

LUMI: Large Unified Modern Infrastructure. 10, 78

MADNESS: Maximum A posteriori with Deep NEural networks for Source Separation. 37

MAP: maximum a posteriori. 23

MAS: multi-agent systems. 53

MCHMC: micro-canonical Hamiltonian Monte Carlo. 43

MCLMC: micro-canonical Langevin Monte Carlo. 43

MCMC: Markov chain Monte Carlo. 27, 39, 42, 46, 49, 61, 69

ML: machine learning. 8, 13, 18, 19, 22, 23, 25, 27, 28, 29, 32, 33, 35, 37, 39, 40, 41, 50, 53, 58, 67, 68, 70, 72, 73, 76, 80, 83

MOC: multi-order coverage. 31

MoE: mixture-of-experts. 57

MViT: multiscale vision transformers. 19

NASA: National Aeronautics and Space Administration. 71

NDE: neural density estimation. 14, 18, 19, 26, 40, 45, 83, 84

NERSC: National Energy Research Scientific Computing Center. 67, 70, 72, 78

NFW: Navarro–Frenk–White. 28

NLE: neural likeliood estimation. 23, 45, 47

NPE: neural posterior estimation. 23, 25, 37, 45, 47

NRE: neural ratio estimation. 23, 44, 45, 47

NSF: National Science Foundation. 12, 71, 77

NUTS: No U-Turn Sampler. 33, 39, 42

ODE: ordinary differential equation. 50

OLCF: Oak Ridge Leadership Computing Facility. 16, 71, 72

ONNX: Open Neural Network Exchange. 68

OpSim: Operations Simulator. 29

Pan-STARRS: Panoramic Survey Telescope and Rapid Response System. 73

PDF: probability density function. 19

photo-z: photometric redshift. 13, 18, 25, 37, 39, 45, 51, 69, 76, 83

PINN: physics-informed neural network. 51, 85

PIT: probability integral transform. 40

PLAsTiCC: Photometric LSST Astronomical Time Series Classification Challenge. 9, 20, 29, 77

PSF: point-spread function. 25, 36, 50, 57, 69

QA: question answering. 64

R&D: research & development. 70

RAG: retrieval-augmented generation. 10, 59, 60, 61, 62, 64, 66

RAIL: Redshift Assessment Infrastructure Layers. 19, 40, 69, 83

RedMaPPer: Red-sequence Matched-filter Probabilistic Percolation. 27

ResNet: residual network. 22

RESSPECT: Recommendation System for Spectroscopic followup. 30, 83

RHMC: Riemann Hamiltonian Monte Carlo. 42

RNN: recurrent neural network. 29

RSP: Rubin Science Platform. 70, 72, 74, 76, 84

SAO: Smithsonian Astrophysical Observatory. 60, 86

SASSAFRAS: S/N Analysis of Simulated SpectrA for Rubin trAnsientS. 30

SBI: simulation-based inference. 14, 18, 23, 26, 28, 31, 32, 33, 40, 41, 44, 45, 47, 70, 83, 84

SBM: score-based model. 24

SDE: stochastic differential equation. 24, 42

SDSS: Sloan Digital Sky Survey. 72

SED: spectral energy distribution. 18, 29, 61

SFH: star formation history. 34

SkAI: NSF–Simons AI Institute for the Sky. 30, 71, 77

SLAC: Stanford Linear Accelerator Center. 70, 72

SLACS: Sloan Lens ACS (Advanced Camera for Surveys) Survey. 24

SLSC: Strong Lensing Science Collaboration. 11, 22

SMWLVSC: Stars, Milky Way, and Local Volume Science Collaboration. 76

SN Ia: Type Ia supernova. 13, 20, 28, 44, 77

SNANA: Supernova Analysis package. 29

SNe: supernovae. 22, 29, 39, 40, 48, 51, 54

SNR: signal-to-noise ratio. 38, 50

SO: Simons Observatory. 73

SOAR: Southern Astrophysical Research Telescope. 30

SOM: self-organizing map. 19, 25, 37

***SPHEREx*:** *Spectro-Photometer for the History of the Universe, Epoch of Reionization, and Ices Explorer*. 72

SPS: stellar population synthesis. 20, 33, 48

SPT: South Pole Telescope. 73

SQL: structured query language. 62

sSFR: specific star formation rate. 25

SSL: self-supervised Learning. 55, 56

SSSC: Solar System Science Collaboration. 76

StratLearn: stratified learning. 20

SZ: Sunyaev–Zeldovich. 27

TiDES: Time Domain Extragalactic Survey. 31, 57, 77

TPZ: Trees for Photo-z. 19

TVSSC: Transients and Variable Stars Science Collaboration. 11, 76

UQ: uncertainty quantification. 14, 36, 40, 41, 47, 83, 84

USDF: US Data Facility. 70, 72

UV: ultraviolet. 21, 34

VAE: variational autoencoder. 28, 37

VI: variational inference. 23, 43, 46, 85

WaZP: Wavelet Z-Photometric. 27

WG: working group. 11, 40

***WISE*:** *Wide-field Infrared Survey Explorer*. 72

YOLO: You Only Look Once. 27, 38, 85

ZTF: Zwicky Transient Facility. 73

C. Acknowledgments

We are thankful to DESC members Prakruth Adari, Keith Bechtol, Simon Birrer, Elisa Chisari, Andy Connolly, Cora Dvorkin, Eric Gawiser, Jimena González, Katrin Heitmann, Xiangchong Li, Simona Mei, Irene Moskowitz, Peter Nugent, Natalia Porqueres, Mara Salvato, Anze Slosar, Crescenzo Tortora, and V. Ashley Villar for their contributions and feedback in the preparation of this manuscript. The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules (IN2P3) in France; the Science & Technology Facilities Council (STFC) in the United Kingdom; and the Department of Energy (DOE), the National Science Foundation (NSF), and the LSST Discovery Alliance (LSST-DA) in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BEIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515. A.T.G. is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). A.M. is supported by the Australian Research Council Discovery Early Career Research Award (DE230100055). M.L. acknowledges support from the South African Radio Astronomy Observatory and the National Research Foundation (NRF) towards this research. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. H.P. and S.T. have been supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programmes (grant agreement no. 101018897 CosmicExplorer). H.P. was additionally supported by the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine. C.D.L. is supported by the Science and Technology Facilities Council (STFC) [grant No. UKRI1172]. The work of Y.Z. is supported by NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the U.S. National Science Foundation. J.d.V., I.S.-N., and L.T.S.C. are partially supported by the Spanish MICINN under grant PID2021-123012 and for the MAD4SPACE-CM TEC-2024/TEC-182 project funded by Comunidad de Madrid. S.S. has received funding from the European Union's Horizon 2022 research and innovation programme under the Marie Skłodowska-Curie grant

agreement No 101105167 — FASTIDIoUS. R.T. acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR "Future Artificial Intelligence Research", as well as Fondazione ICSC, Spoke 3 "Astrophysics and Cosmos Observations", Project ID CN00000013 "Italian Research Center on High-Performance Computing, Big Data and Quantum Computing", and partially supported by INFN INDARK grant. M.I. acknowledges support in part by the U.S. National Science Foundation under grant AST2327245, and also in part by the Department of Energy, office of Science, under Award Number DE-SC0022184. B.R. gratefully acknowledges the support of the NSF-Simons AI-Institute for the Sky (SkAI) via grants NSF AST-2421845 and Simons Foundation MPS-AI-00010513. C.G. is funded by the MICINN project PID2022-141079NB-C32. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. T.Z. is supported by Schmidt Sciences. J.Z. work is partially supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). C.A. acknowledges support from DOE grant DE-SC009193. A.G. is supported by an LSST-DA Catalyst Fellowship, made possible through the support of Grant 62192 from the John Templeton Foundation to LSST-DA. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of LSST-DA or the John Templeton Foundation. T.T. acknowledges funding from the Swiss National Science Foundation under the Ambizione project PZ00P2_193352. M.M. acknowledges support by the SNSF through return CH grant P5R5PT_225598 and Ambizione grant PZ00P2_223738. The work of A.A.P.M. was supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515. M.W.C. acknowledges support from the National Science Foundation with grant numbers PHY-2117997, PHY-2308862 and PHY-2409481. A.Ć. A.D.-W. This work was produced by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. Publisher acknowledges the U.S. Government license to provide public access under the (DOE Public Access Plan). G.M. acknowledges support from Illinois Campus Research Board Award RB25035, NSF grant AST-2308174, and NASA grants 80NSSC24K0219 and 80NSSC25K7739. This work used Delta and DeltaAI at NCSA through allocations PHY240290, PHY250333, PHY250374, PHY250386, PHY250281 and PHY250308 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. M.G. is supported by the European Union's Horizon 2020 research and innovation programme under ERC Grant Agreement No. 101002652. Argonne National Laboratory's work was supported under the US Department of Energy contract DE-AC02-06CH11357. T.S. gratefully acknowledge the support of the NSF-Simons AI-Institute for the Sky (SkAI) via grants NSF AST-2421845 and Simons Foundation MPS-AI-00010513. T.S. was supported by NSF through grant AST-2510183 and by NASA through grants 22-ROMAN22-0055 and 22-ROMAN22-0013. Y.-Y.M. is in part supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, Experimental Research at the Cosmic Frontier program under Award Number DE-SC0009959.

References

Abe, K. T., Oguri, M., Birrer, S., et al. 2025, The Open Journal of Astrophysics, 8, 8, doi: 10.33232/001c.128482

Adam, A., Coogan, A., Malkin, N., et al. 2022, arXiv e-prints, arXiv:2505.19897, doi: 10.48550/arXiv.2211.03812

Adam, A., Perreault-Levasseur, L., Hezaveh, Y., & Welling, M. 2023, ApJ, 951, 6, doi: 10.3847/1538-4357/accf84

Adam, A., Stone, C., Bottrell, C., et al. 2025, AJ, 169, 254, doi: 10.3847/1538-3881/adb039

Adari, P., & von der Linden, A. 2025, SITCOMTN-128: Unrecognized Blends in LSSTComCam Data Preview 1 ECDFS, doi: 10.71929/RUBIN/2570850

Agarwal, S., Ćiprijanović, A., & Nord, B. D. 2024, arXiv e-prints, arXiv:2411.03334, doi: 10.48550/arXiv.2411.03334

Aguena, M., Benoist, C., da Costa, L. N., et al. 2021a, MNRAS, 502, 4435, doi: 10.1093/mnras/stab264

Aguena, M., Avestruz, C., Combet, C., et al. 2021b, MNRAS, 508, 6092, doi: 10.1093/mnras/stab2764

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4, doi: 10.1093/pasj/psx066

AL Otaibi, S., Tiňo, P., Cuevas-Tello, J. C., Mandel, I., & Raychaudhury, S. 2016, MNRAS, 459, 573, doi: 10.1093/mnras/stw510

Alarcon, A., Hearin, A. P., Becker, M. R., et al. 2025, arXiv e-prints, arXiv:2510.27604, doi: 10.48550/arXiv.2510.27604

Alarcon, A., Hearin, A. P., Becker, M. R., & Chaves-Montero, J. 2023, MNRAS, 518, 562, doi: 10.1093/mnras/stac3118

Allam, T., & McEwen, J. D. 2024, RAS Techniques and Instruments, 3, 209, doi: 10.1093/rasti/rzad046

Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016, MNRAS, 462, 726, doi: 10.1093/mnras/stw1618

Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440, doi: 10.1093/mnras/stz1960

Alsing, J., Peiris, H., Mortlock, D., Leja, J., & Leistedt, B. 2023, ApJS, 264, 29, doi: 10.3847/1538-4365/ac9583

Alsing, J., Thorp, S., Deger, S., et al. 2024, ApJS, 274, 12, doi: 10.3847/1538-4365/ad5c69

Alsing, J., & Wandelt, B. 2018, MNRAS, 476, L60, doi: 10.1093/mnrasl/sly029

—. 2019, MNRAS, 488, 5093, doi: 10.1093/mnras/stz1900

Alsing, J., Wandelt, B., & Feeney, S. 2018, MNRAS, 477, 2874, doi: 10.1093/mnras/sty819

Alsing, J., Peiris, H., Leja, J., et al. 2020, ApJS, 249, 5, doi: 10.3847/1538-4365/ab917f

Alves, C. S., Peiris, H. V., Lochner, M., et al. 2022, ApJS, 258, 23, doi: 10.3847/1538-4365/ac3479

—. 2023, ApJS, 265, 43, doi: 10.3847/1538-4365/acbb09

Alvey, J., Bhardwaj, U., Domcke, V., Pieroni, M., & Weniger, C. 2024, PhRvD, 109, 083008, doi: 10.1103/PhysRevD.109.083008

Amon, A., Gruen, D., Troxel, M., et al. 2022, PhRvD, 105, doi: 10.1103/physrevd.105.023514

Anau Montel, N., Alvey, J., & Weniger, C. 2025, PhRvD, 111, 083013, doi: 10.1103/PhysRevD.111.083013

Andersson, A., Lintott, C., Fender, R., et al. 2025, MNRAS, 538, 1397, doi: 10.1093/mnras/staf336

Andika, I. T., Suyu, S. H., Cañameras, R., et al. 2023, A&A, 678, A103, doi: 10.1051/0004-6361/202347332

Angora, G., Rosati, P., Meneghetti, M., et al. 2023, A&A, 676, A40, doi: 10.1051/0004-6361/202346283

Arcelin, B., Doux, C., Aubourg, E., Roucelle, C., & LSST Dark Energy Science Collaboration. 2021, MNRAS, 500, 531, doi: 10.1093/mnras/staa3062

Arendse, N., Dhawan, S., Sagués Carracedo, A., et al. 2024, MNRAS, 531, 3509, doi: 10.1093/mnras/stae1356

Aricò, G., Angulo, R. E., Contreras, S., et al. 2021, MNRAS, 506, 4070, doi: 10.1093/mnras/stab1911

Arnouts, S., & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library, record ascl:1108.009. http://ascl.net/1108.009

Asai, A., He, J., Shao, R., et al. 2024, arXiv e-prints, arXiv:2411.14199, doi: 10.48550/arXiv.2411.14199

Asgari, M., Lin, C.-A., Joachimi, B., et al. 2021, A&A, 645, A104, doi: 10.1051/0004-6361/202039070

Ashton, G., Bernstein, N., Buchner, J., et al. 2022, Nature Reviews Methods Primers, 2, doi: 10.1038/s43586-022-00121-x

Audenaert, J., Muthukrishna, D., Gregory, P. F., Hogg, D. W., & Villar, V. A. 2025, arXiv e-prints, arXiv:2507.05333, doi: 10.48550/arXiv.2507.05333

Autenrieth, M., van Dyk, D. A., Trotta, R., & Stenning, D. C. 2023, Statistical Analysis and Data Mining: The ASA Data Science Journal, 17, doi: 10.1002/sam.11643

Autenrieth, M., Wright, A. H., Trotta, R., et al. 2024, MNRAS, 534, 3808–3831, doi: 10.1093/mnras/stae2243

Bag, S., Canameras, R., Suyu, S. H., et al. 2025, arXiv e-prints, arXiv:2506.22076, doi: 10.48550/arXiv.2506.22076

Bag, S., Huber, S., Suyu, S. H., et al. 2024, A&A, 691, A100, doi: 10.1051/0004-6361/202450485

Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87, doi: 10.1051/0004-6361/202038986

Barco, G. M., Adam, A., Stone, C., Hezaveh, Y., & Perreault-Levasseur, L. 2025a, ApJ, 980, 108, doi: 10.3847/1538-4357/ad9b92

Barco, G. M., Legin, R., Stone, C., Hezaveh, Y., & Perreault-Levasseur, L. 2025b, MLforAstro at ICML Workshop, arXiv:2511.04792. https://arxiv.org/abs/2511.04792

Barman, K. G., Caron, S., Sullivan, E., et al. 2025, European Physical Journal C, 85, 1066, doi: 10.1140/epjc/s10052-025-14707-8

Baron Perez, N., Brüggen, M., Kasieczka, G., & Lucie-Smith, L. 2025, A&A, 699, A302, doi: 10.1051/0004-6361/202554735

Bartlett, D. J., & Pandey, S. 2025, arXiv e-prints, arXiv:2510.18749, doi: 10.48550/arXiv.2510.18749

Bashir, N., Donti, P., Cuff, J., et al. 2024, in An MIT Exploration of Generative AI (MIT), doi: 10.21428/e4baedd9.9070dfe7

Bayer, A. E., Seljak, U., & Modi, C. 2023, arXiv e-prints, arXiv:2307.09504, doi: 10.48550/arXiv.2307.09504

Bazin, G., Ruhlmann-Kleider, V., Palanque-Delabrouille, N., et al. 2011, A&A, 534, A43, doi: 10.1051/0004-6361/201116898

Bechtol, K., Sevilla-Noarbe, I., Drlica-Wagner, A., et al. 2025, arXiv e-prints, arXiv:2501.05739, doi: 10.48550/arXiv.2501.05739

Behroozi, P. 2025, arXiv e-prints, arXiv:2510.25824. https://arxiv.org/abs/2510.25824

Beltagy, I., Peters, M. E., & Cohan, A. 2020, arXiv e-prints, arXiv:2004.05150, doi: 10.48550/arXiv.2004.05150

Berkson, J., Angel, R., Bender, C., et al. 2024, Nanomanufacturing and Metrology, 7, doi: 10.1007/s41871-024-00235-8

Bingham, E., Chen, J. P., Jankowiak, M., et al. 2019, Journal of Machine Learning Research, 20, 1. http://jmlr.org/papers/v20/18-403.html

Birrer, S., Amara, A., & Refregier, A. 2015, ApJ, 813, 102, doi: 10.1088/0004-637X/813/2/102

Biswas, B., Aubourg, E., Boucaud, A., et al. 2025, A&A, 700, A129, doi: 10.1051/0004-6361/202451887

Blondel, M., Berthet, Q., Cuturi, M., et al. 2021, arXiv e-prints, arXiv:2105.15183, doi: 10.48550/arXiv.2105.15183

Bolliet, B. 2025, CMBAGENT: Open-Source Multi-Agent System for Science, latest. https://github.com/CMBAgents/cmbagent

Bom, C. R., Makler, M., & Albuquerque, M. P. 2012, arXiv e-prints, arXiv:1212.1799, doi: 10.48550/arXiv.1212.1799

Bom, C. R., Makler, M., Albuquerque, M. P., & Brandt, C. H. 2017, A&A, 597, A135, doi: 10.1051/0004-6361/201629159

Bom, C. R., Fraga, B. M. O., Dias, L. O., et al. 2022, MNRAS, 515, 5121, doi: 10.1093/mnras/stac2047

Bommasani, R., Hudson, D. A., Adeli, E., et al. 2021, arXiv e-prints, arXiv:2108.07258, doi: 10.48550/arXiv.2108.07258

Bonjean, V. 2020, A&A, 634, A81, doi: 10.1051/0004-6361/201936919

Boone, K. 2019, AJ, 158, 257, doi: 10.3847/1538-3881/ab5182

—. 2021, AJ, 162, 275, doi: 10.3847/1538-3881/ac2a2d

Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103, doi: 10.1051/0004-6361/201834156

Boruah, S. S., Rozo, E., & Fiedorowicz, P. 2022, MNRAS, 516, 4111, doi: 10.1093/mnras/stac2508
Boruah, S. S., Eifler, T., Miranda, V., et al. 2024, arXiv e-prints, arXiv:2403.11797, doi: 10.48550/arXiv.2403.11797
Bosch, J., Armstrong, R., Bickerton, S., et al. 2018, PASJ, 70, S5, doi: 10.1093/pasj/psx080
Boyd, B. M., Grayling, M., Thorp, S., & Mandel, K. S. 2024, arXiv e-prints, arXiv:2407.15923, doi: 10.48550/arXiv.2407.15923
Boyd, B. M., Narayan, G., Mandel, K. S., et al. 2025, MNRAS, 540, 385, doi: 10.1093/mnras/staf629
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503, doi: 10.1086/591786
Breitman, D., Mesinger, A., Murray, S. G., et al. 2024, MNRAS, 527, 9833, doi: 10.1093/mnras/stad3849
Brewer, B. J., & Foreman-Mackey, D. 2018, Journal of Statistical Software, 86, 1, doi: 10.18637/jss.v086.i07
Brixi, G., Durrant, M. G., Ku, J., et al. 2025, bioRxiv, doi: 10.1101/2025.02.18.638918
Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. 2000, AJ, 120, 1588, doi: 10.1086/301514
Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952, doi: 10.1093/mnras/stz2845
Byler, N., Dalcanton, J. J., Conroy, C., & Johnson, B. D. 2017, ApJ, 840, 44, doi: 10.3847/1538-4357/aa6c66
Cañameras, R., Schuldt, S., Shu, Y., et al. 2024, A&A, 692, A72, doi: 10.1051/0004-6361/202347072
Cabezas, A., Corenflos, A., Lao, J., et al. 2024, arXiv e-prints, arXiv:2402.10797, doi: 10.48550/arXiv.2402.10797
Cabrera-Vives, G., Moreno-Cartagena, D., Astorga, N., et al. 2024, A&A, 689, A289, doi: 10.1051/0004-6361/202449475
Cádiz-Leyton, M., Cabrera-Vives, G., Protopapas, P., Moreno-Cartagena, D., & Becker, I. 2025a, arXiv e-prints, arXiv:2507.12611, doi: 10.48550/arXiv.2507.12611
Cádiz-Leyton, M., Cabrera-Vives, G., Protopapas, P., et al. 2025b, A&A, 699, A168, doi: 10.1051/0004-6361/202453388
Campagne, J.-E. 2025, MNRAS, 539, 3445, doi: 10.1093/mnras/staf533
Campagne, J.-E., Lanusse, F., Zuntz, J., et al. 2023, The Open Journal of Astrophysics, 6, 15, doi: 10.21105/astro.2302.05163
Campeau-Poirier, È., Perreault-Levasseur, L., Coogan, A., & Hezaveh, Y. 2023, in Machine Learning for Astrophysics, 6, doi: 10.48550/arXiv.2309.16063
Cao, J., Xu, T., Deng, Y., et al. 2024, A&A, 683, A42, doi: 10.1051/0004-6361/202348544
Caron, S., Ipp, A., Aarts, G., et al. 2025, arXiv e-prints, arXiv:2503.14192, doi: 10.48550/arXiv.2503.14192
Carrasco-Davis, R., Reyes, E., Valenzuela, C., et al. 2021, AJ, 162, 231, doi: 10.3847/1538-3881/ac0ef1
Carrasco Kind, M., & Brunner, R. J. 2013, MNRAS, 432, 1483, doi: 10.1093/mnras/stt574
Casella, G., & George, E. I. 1992, The American Statistician, 46, 167, doi: 10.1080/00031305.1992.10475878
Chameleon Team. 2024, arXiv e-prints, arXiv:2405.09818, doi: 10.48550/arXiv.2405.09818
Chan, M. C., & Stott, J. P. 2019, MNRAS, 490, 5770, doi: 10.1093/mnras/stz2936
Charnock, T., Lavaux, G., & Wandelt, B. D. 2018, PhRvD, 97, 083004, doi: 10.1103/PhysRevD.97.083004
Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. 2018, in Advances in Neural Information Processing Systems 31, ed. S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, 6572–6583, doi: 10.48550/arXiv.1806.07366
Chen, T., Bansal, V., & Scott, J. G. 2025, arXiv e-prints, arXiv:2507.17030, doi: 10.48550/arXiv.2507.17030
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, in Proceedings of Machine Learning Research, Vol. 119, Proceedings of the 37th International Conference on Machine Learning, ed. H. Daumé, III & A. Singh (PMLR), 1597–1607. https://proceedings.mlr.press/v119/chen20j.html
Chen, X., Hu, L., & Wang, L. 2020, ApJS, 250, 12, doi: 10.3847/1538-4365/ab9a3b

Chen, Y., Zhang, D., Gutmann, M. U., Courville, A., & Zhu, Z. 2021, in International Conference on Learning Representations. https://openreview.net/forum?id=SRDuJssQud
Chianese, M., Coogan, A., Hofma, P., Otten, S., & Weniger, C. 2020, MNRAS, 496, 381, doi: 10.1093/mnras/staa1477
Childress, M., Aldering, G., Antilogus, P., et al. 2013, ApJ, 770, 108, doi: 10.1088/0004-637X/770/2/108
Chisari, N. E., Alonso, D., Krause, E., et al. 2019, ApJS, 242, 2, doi: 10.3847/1538-4365/ab1658
Christiano, P. F., Leike, J., Brown, T. B., et al. 2017, in Advances in Neural Information Processing Systems 30, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 4299–4307, doi: 10.48550/arXiv.1706.03741
Chung, D. J. H., Gao, Z., Kvasiuk, Y., et al. 2025, Machine Learning: Science and Technology, 6, 030505, doi: 10.1088/2632-2153/adfcb0
Ćiprijanović, A., Lewis, A., Pedro, K., et al. 2023, Machine Learning: Science and Technology, 4, 025013, doi: 10.1088/2632-2153/acca5f
Collett, T. E. 2015, ApJ, 811, 20, doi: 10.1088/0004-637X/811/1/20
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345, doi: 10.1086/383254
Conroy, C. 2013, ARA&A, 51, 393, doi: 10.1146/annurev-astro-082812-141017
Conroy, C., & Gunn, J. E. 2010, ApJ, 712, 833, doi: 10.1088/0004-637X/712/2/833
Conroy, C., Gunn, J. E., & White, M. 2009, ApJ, 699, 486, doi: 10.1088/0004-637X/699/1/486
Conroy, C., White, M., & Gunn, J. E. 2010, ApJ, 708, 58, doi: 10.1088/0004-637X/708/1/58
Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proceedings of the National Academy of Sciences, 117, 30055, doi: 10.1073/pnas.1912789117
Cranmer, M., Tamayo, D., Rein, H., et al. 2021, Proceedings of the National Academy of Sciences, 118, e2026053118, doi: 10.1073/pnas.2026053118
Crenshaw, J. F., & Connolly, A. J. 2020, AJ, 160, 191, doi: 10.3847/1538-3881/abb0e2
Crenshaw, J. F., Kalmbach, J. B., Gagliano, A., et al. 2024, AJ, 168, 80, doi: 10.3847/1538-3881/ad54bf
Csabai, I., Connolly, A. J., Szalay, A. S., & Budavári, T. 2000, AJ, 119, 69, doi: 10.1086/301159
Cuesta-Lazaro, C., & Mishra-Sharma, S. 2024, PhRvD, 109, 123531, doi: 10.1103/PhysRevD.109.123531
Cuevas-Tello, J. C., Tiňo, P., Raychaudhury, S., Yao, X., & Harva, M. 2010, Pattern Recognition, 43, 1165
Cuevas-Tello, J. C., Tiňo, P., & Raychaudhury, S. 2006, A&A, 454, 695, doi: 10.1051/0004-6361:20054652
Dalmasso, N., Pospisil, T., Lee, A. B., et al. 2020, Astronomy and Computing, 30, 100362, doi: 10.1016/j.ascom.2019.100362
Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. 2022, in Advances in Neural Information Processing Systems 35, ed. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Curran Associates Inc.), 16344–16359, doi: 10.48550/arXiv.2205.14135
Dawson, W. A., Schneider, M. D., Tyson, J. A., & Jee, M. J. 2016, ApJ, 816, 11, doi: 10.3847/0004-637X/816/1/11
de Andres, D., Cui, W., Ruppin, F., et al. 2022, Nature Astronomy, 6, 1325, doi: 10.1038/s41550-022-01784-y
de Graaff, R., Margalef-Bentabol, B., Wang, L., et al. 2025, A&A, 697, A207, doi: 10.1051/0004-6361/202452659
de Haan, T., Ting, Y., Ghosal, T., et al. 2025, Scientific Reports, doi: 10.1038/s41598-025-97131-y
De Vicente, J., Sánchez, E., & Sevilla-Noarbe, I. 2016, MNRAS, 459, 3078, doi: 10.1093/mnras/stw857
Deger, S., Peiris, H. V., Thorp, S., et al. 2025, arXiv e-prints, arXiv:2509.20430, doi: 10.48550/arXiv.2509.20430
Delgado, F., Saha, A., Chandrasekharan, S., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI, ed. G. Z. Angeli & P. Dierickx, 915015, doi: 10.1117/12.2056898
Deng, X., Da, J., Pan, E., et al. 2025, arXiv e-prints, arXiv:2509.16941, doi: 10.48550/arXiv.2509.16941
DeRose, J., Wechsler, R. H., Tinker, J. L., et al. 2019, ApJ, 875, 69, doi: 10.3847/1538-4357/ab1085

DES Collaboration, Abbott, T. M. C., Acevedo, M., et al. 2024, ApJL, 973, L14, doi: 10.3847/2041-8213/ad6f9f
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168, doi: 10.3847/1538-3881/ab089d
Di Palo, N., & Johns, E. 2024, arXiv e-prints, arXiv:2402.13181, doi: 10.48550/arXiv.2402.13181
Diaz Rivero, A., & Dvorkin, C. 2020, PhRvD, 102, 103507, doi: 10.1103/PhysRevD.102.103507
Dillmann, S., Martínez-Galarza, J. R., Soria, R., Stefano, R. D., & Kashyap, V. L. 2025, MNRAS, 537, 931, doi: 10.1093/mnras/stae2808
Donoso-Oliva, C., Becker, I., Protopapas, P., et al. 2025, arXiv e-prints, arXiv:2502.02717, doi: 10.48550/arXiv.2502.02717
—. 2023, A&A, 670, A54, doi: 10.1051/0004-6361/202243928
Duan, Y., Li, X., Avestruz, C., & Regier, J. 2025, arXiv e-prints, arXiv:2510.15315, doi: 10.48550/arXiv.2510.15315
Dye, S., & Warren, S. J. 2005, ApJ, 623, 31, doi: 10.1086/428340
Erickson, S., Millon, M., Venkatraman, P., et al. 2025a, arXiv e-prints, arXiv:2511.13669. https://arxiv.org/abs/2511.13669
Erickson, S., Wagner-Carena, S., Marshall, P., et al. 2025b, AJ, 170, 44, doi: 10.3847/1538-3881/add99f
Essam Ghareeb, A., Chang, B., Mitchener, L., et al. 2025, arXiv e-prints, arXiv:2505.13400, doi: 10.48550/arXiv.2505.13400
Estrada, J., Annis, J., Diehl, H. T., et al. 2007, ApJ, 660, 1176, doi: 10.1086/512599
Etsebeth, V., Lochner, M., Walmsley, M., & Grespan, M. 2024, MNRAS, 529, 732, doi: 10.1093/mnras/stae496
Euclid Collaboration: Adam, R., Vannier, M., Maurogordato, S., et al. 2019, A&A, 627, A23, doi: 10.1051/0004-6361/201935088
Euclid Collaboration: Aussel, B., Kruk, S., Walmsley, M., et al. 2024, A&A, 689, A274, doi: 10.1051/0004-6361/202449609
Euclid Collaboration: Bazzanini, L., Angora, G., Bergamini, P., et al. 2025, arXiv e-prints, arXiv:2511.03064, doi: 10.48550/arXiv.2511.03064
Euclid Collaboration: Busillo, V., Tortora, C., Metcalf, R. B., et al. 2025, arXiv e-prints, arXiv:2503.15329, doi: 10.48550/arXiv.2503.15329
Euclid Collaboration: Lines, N. E. P., Collett, T. E., Walmsley, M., et al. 2025, arXiv e-prints, arXiv:2503.15326, doi: 10.48550/arXiv.2503.15326
Euclid Collaboration: Walmsley, M., Holloway, P., Lines, N. E. P., et al. 2025, arXiv e-prints, arXiv:2503.15324, doi: 10.48550/arXiv.2503.15324
Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15, doi: 10.1088/0004-637X/760/1/15
Fagin, J., Park, J. W., Best, H., et al. 2024, ApJ, 965, 104, doi: 10.3847/1538-4357/ad2988
Fan, W., Ding, Y., Ning, L., et al. 2024, in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, 6491–6501, doi: 10.1145/3637528.3671470
Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. 2021, in Advances in Data Science and Information Engineering, ed. R. Stahlbock, G. M. Weiss, M. Abou-Nasr, C.-Y. Yang, H. R. Arabnia, & L. Deligiannidis (Springer International Publishing), 877–894, doi: 10.1007/978-3-030-71704-9_65
Fedus, W., Zoph, B., & Shazeer, N. 2022, Journal of Machine Learning Research, 23, 1. https://jmlr.org/papers/v23/21-0998.html
Feeney, S. M., Mortlock, D. J., & Dalmasso, N. 2018, MNRAS, 476, 3861, doi: 10.1093/mnras/sty418
Ferguson, A., LaFleur, M., Ruthotto, L., et al. 2025, arXiv e-prints, arXiv:2509.02661, doi: 10.48550/arXiv.2509.02661
Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449, doi: 10.1111/j.1365-2966.2007.12353.x
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601, doi: 10.1111/j.1365-2966.2009.14548.x
Filipp, A., Hezaveh, Y., & Perreault-Levasseur, L. 2025, ApJ, 989, 226, doi: 10.3847/1538-4357/adee20
Fluri, J., Kacprzak, T., Lucchi, A., et al. 2022, PhRvD, 105, 083518, doi: 10.1103/PhysRevD.105.083518
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306, doi: 10.1086/670067

Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al. 2021, AJ, 161, 242, doi: 10.3847/1538-3881/abe9bc

Fort, S., Hu, H., & Lakshminarayanan, B. 2019, arXiv e-prints, arXiv:1912.02757, doi: 10.48550/arXiv.1912.02757

Fraga, B. M. O., Bom, C. R., Santos, A., et al. 2024, A&A, 692, A208, doi: 10.1051/0004-6361/202450370

Franco-Abellán, G., Cañas-Herrera, G., Martinelli, M., et al. 2024, JCAP, 2024, 057, doi: 10.1088/1475-7516/2024/11/057

Frohmaier, C., Vincenzi, M., Sullivan, M., et al. 2025, ApJ, 992, 158, doi: 10.3847/1538-4357/adff4e

Gabrié, M., Rotskoff, G. M., & Vanden-Eijnden, E. 2022, Proceedings of the National Academy of Science, 119, e2109420119, doi: 10.1073/pnas.2109420119

Gagliano, A., Contardo, G., Foreman-Mackey, D., Malz, A. I., & Aleo, P. D. 2023, ApJ, 954, 6, doi: 10.3847/1538-4357/ace326

Gagliano, A., Narayan, G., Engel, A., Carrasco Kind, M., & LSST Dark Energy Science Collaboration. 2021, ApJ, 908, 170, doi: 10.3847/1538-4357/abd02b

GalSim Developers. 2025, JAX-GalSim: A differentiable and GPU-accelerated galaxy image simulation library, https://github.com/GalSim-developers/JAX-GalSim

Gatti, M., Jain, B., Chang, C., et al. 2022, PhRvD, 106, 083509, doi: 10.1103/PhysRevD.106.083509

Gawade, P., More, A., More, S., et al. 2025, MNRAS, 540, 3384, doi: 10.1093/mnras/staf935

Geffner, T., Papamakarios, G., & Mnih, A. 2023, in Proceedings of Machine Learning Research, Vol. 202, Proceedings of the 40th International Conference on Machine Learning, ed. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (PMLR), 11098–11116. https://proceedings.mlr.press/v202/geffner23a.html

Gentile, F., Tortora, C., Covone, G., et al. 2023, MNRAS, 522, 5442, doi: 10.1093/mnras/stad1325

Girolami, M., & Calderhead, B. 2011, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73, 123, doi: https://doi.org/10.1111/j.1467-9868.2010.00765.x

Gloeckler, M., Deistler, M., Weilbach, C. D., Wood, F., & Macke, J. H. 2024, in Proceedings of Machine Learning Research, Vol. 235, Proceedings of the 41st International Conference on Machine Learning, ed. R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (PMLR), 15735–15766. https://proceedings.mlr.press/v235/gloeckler24a.html

Goldstein, D. A., Nugent, P. E., & Goobar, A. 2019, ApJS, 243, 6, doi: 10.3847/1538-4365/ab1fe0

Gonçalves, G., Arendse, N., Kodi Ramanah, D., & Wojtak, R. 2025, arXiv e-prints, arXiv:2504.10553, doi: 10.48550/arXiv.2504.10553

Gonzalez, J., Collett, T., Rojas, K., et al. 2025a, arXiv e-prints, arXiv:2510.23782. https://arxiv.org/abs/2510.23782

Gonzalez, J., Holloway, P., Collett, T., et al. 2025b, arXiv e-prints, arXiv:2501.15679, doi: 10.48550/arXiv.2501.15679

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65, doi: 10.2140/camcos.2010.5.65

Gottweis, J., Weng, W.-H., Daryin, A., et al. 2025, arXiv e-prints, arXiv:2502.18864, doi: 10.48550/arXiv.2502.18864

Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, AJ, 155, 1, doi: 10.3847/1538-3881/aa99d4

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., & Duvenaud, D. 2018, arXiv e-prints, arXiv:1810.01367, doi: 10.48550/arXiv.1810.01367

Grayling, M., Thorp, S., Mandel, K. S., et al. 2024, MNRAS, 531, 953, doi: 10.1093/mnras/stae1202

Grishin, K., Mei, S., & Ilić, S. 2023, A&A, 677, A101, doi: 10.1051/0004-6361/202345976

Grishin, K., Mei, S., Ilic, S., et al. 2025, A&A, 695, A246, doi: 10.1051/0004-6361/202452119

Grumitt, R. D. P., Dai, B., & Seljak, U. 2022, in Advances in Neural Information Processing Systems 35, ed. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Curran Associates, Inc.), 11629–11641, doi: 10.48550/arXiv.2205.14240

Guo, D., Yang, D., Zhang, H., et al. 2025, Nature, 645, 633, doi: 10.1038/s41586-025-09422-z

Gupta, A., Matilla, J. M. Z., Hsu, D., et al. 2018, PhRvD, 97, 103515, doi: 10.1103/PhysRevD.97.103515

Gupta, R., Muthukrishna, D., & Audenaert, J. 2025, arXiv e-prints, arXiv:2510.12958, doi: 10.48550/arXiv.2510.12958

Guy, J., Astier, P., Baumont, S., et al. 2007, A&A, 466, 11, doi: 10.1051/0004-6361:20066930

Hahn, C., & Melchior, P. 2022, ApJ, 938, 11, doi: 10.3847/1538-4357/ac7b84

Hallinan, G., Ravi, V., Weinreb, S., et al. 2019, in Bulletin of the American Astronomical Society, Vol. 51, 255, doi: 10.48550/arXiv.1907.07648

Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015a, MNRAS, 453, 4384, doi: 10.1093/mnras/stv1911

—. 2015b, MNRAS, 450, L61, doi: 10.1093/mnrasl/slv047

Hansen, D. L., Mendoza, I., Liu, R., et al. 2022, in Machine Learning for Astrophysics, 27, doi: 10.48550/arXiv.2207.05642

Havens, A. J., Miller, B. K., Yan, B., et al. 2025, in Proceedings of Machine Learning Research, Vol. 267, Proceedings of the 42nd International Conference on Machine Learning, ed. A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, & J. Zhu (PMLR), 22204–22237. https://proceedings.mlr.press/v267/havens25a.html

Hayden, B. T., Gupta, R. R., Garnavich, P. M., et al. 2013, ApJ, 764, 191, doi: 10.1088/0004-637X/764/2/191

Hayes, E. E., Dhawan, S., Thorp, S., Pierel, J. D. R., & Arendse, N. 2025, arXiv e-prints, arXiv:2509.25350, doi: 10.48550/arXiv.2509.25350

Hayes, E. E., Thorp, S., Mandel, K. S., et al. 2024, MNRAS, 530, 3942, doi: 10.1093/mnras/stae1086

Hearin, A. P., Chaves-Montero, J., Alarcon, A., Becker, M. R., & Benson, A. 2023, MNRAS, 521, 1741, doi: 10.1093/mnras/stad456

Hearin, A. P., Chaves-Montero, J., Becker, M. R., & Alarcon, A. 2021, The Open Journal of Astrophysics, 4, 7, doi: 10.21105/astro.2105.05859

Heavens, A., Fantaye, Y., Mootoovaloo, A., et al. 2017, arXiv e-prints, arXiv:1704.03472, doi: 10.48550/arXiv.1704.03472

Heesters, N., Chemaly, D., Müller, O., et al. 2025, A&A, 699, A232, doi: 10.1051/0004-6361/202554501

Heitmann, K., Bingham, D., Lawrence, E., et al. 2016, ApJ, 820, 108, doi: 10.3847/0004-637X/820/2/108

Hezaveh, Y. D., Perreault Levasseur, L., & Marshall, P. J. 2017, Nature, 548, 555, doi: 10.1038/nature23463

Higson, E., Handley, W., Hobson, M., & Lasenby, A. 2019, Statistics and Computing, 29, 891, doi: 10.1007/s11222-018-9844-0

Hinton, G. E., & Salakhutdinov, R. R. 2006, Science, 313, 504, doi: 10.1126/science.1127647

Hinton, G. E., & Zemel, R. S. 1993, in Advances in Neural Information Processing Systems 6, ed. J. D. Cowan, G. Tesauro, & J. Alspector (Morgan Kaufmann), 3–10

Hložek, R., Malz, A. I., Ponder, K. A., et al. 2023, ApJS, 267, 25, doi: 10.3847/1538-4365/accd6a

Hlozek, R., Kunz, M., Bassett, B., et al. 2012, ApJ, 752, 79, doi: 10.1088/0004-637X/752/2/79

Ho, J., Jain, A., & Abbeel, P. 2020, in Advances in Neural Information Processing Systems, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, Vol. 33 (Curran Associates, Inc.), 6840–6851, doi: 10.48550/arXiv.2006.11239

Ho, M., Farahi, A., Rau, M. M., & Trac, H. 2021, ApJ, 908, 204, doi: 10.3847/1538-4357/abd101

Ho, M., Ntampaka, M., Rau, M. M., et al. 2022, Nature Astronomy, 6, 936, doi: 10.1038/s41550-022-01711-1

Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, ApJ, 887, 25, doi: 10.3847/1538-4357/ab4f82

Hoffman, M. D., & Gelman, A. 2014, Journal of Machine Learning Research, 15, 1593. http://jmlr.org/papers/v15/hoffman14a.html

Hojjati, A., & Linder, E. V. 2014, PhRvD, 90, 123501, doi: 10.1103/PhysRevD.90.123501
Holloway, P., Marshall, P. J., Verma, A., et al. 2024, MNRAS, 530, 1297, doi: 10.1093/mnras/stae875
Holwerda, B. W., Robertson, C., Cook, K., et al. 2024, PASA, 41, e115, doi: 10.1017/pasa.2024.109
Huang, S., Leauthaud, A., Murata, R., et al. 2018, PASJ, 70, S6, doi: 10.1093/pasj/psx126
Huber, S., & Suyu, S. H. 2024, A&A, 692, A132, doi: 10.1051/0004-6361/202449952
Huff, E., & Mandelbaum, R. 2017, arXiv e-prints, arXiv:1702.02600, doi: 10.48550/arXiv.1702.02600
Huppenkothen, D., Ntampaka, M., Ho, M., et al. 2023, arXiv e-prints, arXiv:2310.12528, doi: 10.48550/arXiv.2310.12528
Hurier, G., Aghanim, N., & Douspis, M. 2021, A&A, 653, A106, doi: 10.1051/0004-6361/201730534
Ilievski, F., Hammer, B., van Harmelen, F., et al. 2025, Nature Machine Intelligence, 7, 1378, doi: 10.1038/s42256-025-01109-4
Iqbal, A., Majumdar, S., Rasia, E., et al. 2025, A&A, 704, A334, doi: 10.1051/0004-6361/202555691
Ishak, M., Baker, T., Bull, P., et al. 2019, arXiv e-prints, arXiv:1905.09687, doi: 10.48550/arXiv.1905.09687
Ishida, E. E. O., Beck, R., González-Gaitán, S., et al. 2019, MNRAS, 483, 2, doi: 10.1093/mnras/sty3015
Iyer, K. G., Pacifici, C., Calistro-Rivera, G., & Lovell, C. C. 2026, in Encyclopedia of Astrophysics, ed. I. Mandel, Vol. 4, 236–281, doi: 10.1016/B978-0-443-21439-4.00127-9
Iyer, K. G., Yunus, M., O'Neill, C., et al. 2024, ApJS, 275, 38, doi: 10.3847/1538-4365/ad7c43
Izbicki, R., & Lee, A. B. 2017, arXiv e-prints, arXiv:1704.08095, doi: 10.48550/arXiv.1704.08095
Jaegle, A., Gimeno, F., Brock, A., et al. 2021a, in Proceedings of Machine Learning Research, Vol. 139, Proceedings of the 38th International Conference on Machine Learning, ed. M. Meila & T. Zhang (PMLR), 4651–4664. https://proceedings.mlr.press/v139/jaegle21a.html
Jaegle, A., Borgeaud, S., Alayrac, J.-B., et al. 2021b, arXiv e-prints, arXiv:2107.14795, doi: 10.48550/arXiv.2107.14795
Jaelani, A. T., More, A., Wong, K. C., et al. 2024, MNRAS, 535, 1625, doi: 10.1093/mnras/stae2442
Jagvaral, Y., Lanusse, F., & Mandelbaum, R. 2025, MNRAS, 542, 2560, doi: 10.1093/mnras/staf592
Jansen, P., Côté, M.-A., Khot, T., et al. 2024, in Advances in Neural Information Processing Systems 37, ed. A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang, 10088–10116, doi: 10.52202/079017-0324
Jarugula, S., Nord, B., Gandrakota, A., & Ćiprijanović, A. 2024, arXiv e-prints, arXiv:2407.17292, doi: 10.48550/arXiv.2407.17292
Jeffrey, N., Alsing, J., & Lanusse, F. 2021, MNRAS, 501, 954, doi: 10.1093/mnras/staa3594
Jeffrey, N., Whiteway, L., Gatti, M., et al. 2025, MNRAS, 536, 1303, doi: 10.1093/mnras/stae2629
Jegou du Laz, T., Coughlin, M. W., Bachant, P., et al. 2025, arXiv e-prints, arXiv:2511.00164, doi: 10.48550/arXiv.2511.00164
Jespersen, C. K., Melchior, P., Spergel, D. N., et al. 2025, arXiv e-prints, arXiv:2503.03816, doi: 10.48550/arXiv.2503.03816
Ji, Z., Lee, N., Frieske, R., et al. 2022, arXiv e-prints, arXiv:2202.03629, doi: 10.48550/arXiv.2202.03629
Jiang, A. Q., Sablayrolles, A., Roux, A., et al. 2024, arXiv e-prints, arXiv:2401.04088, doi: 10.48550/arXiv.2401.04088
Jimenez, C. E., Yang, J., Wettig, A., et al. 2023, arXiv e-prints, arXiv:2310.06770, doi: 10.48550/arXiv.2310.06770
Jiménez-Vicente, J., & Mediavilla, E. 2025, MNRAS, 541, 1264, doi: 10.1093/mnras/staf1067
Joachimi, B., Cacciato, M., Kitching, T. D., et al. 2015, SSRv, 193, 1, doi: 10.1007/s11214-015-0177-4
Johnson, B. D., Leja, J., Conroy, C., & Speagle, J. S. 2021, ApJS, 254, 22, doi: 10.3847/1538-4365/abef67
Kabalan, W., & Lanusse, F. 2025, jaxDecomp: JAX Library for 3D Domain Decomposition and Parallel FFTs, v0.2.9. https://github.com/DifferentiableUniverseInitiative/jaxDecomp
Kacprzak, T., & Fluri, J. 2022, PhRvX, 12, 031029, doi: 10.1103/PhysRevX.12.031029
Kamath, S. 2020, Challenges for Dark Energy Science: Color Gradients and Blended Objects, Zenodo, doi: 10.5281/zenodo.3721438

Karchev, K., Anau Montel, N., Coogan, A., & Weniger, C. 2022a, arXiv e-prints, arXiv:2211.04365, doi: 10.48550/arXiv.2211.04365
Karchev, K., Coogan, A., & Weniger, C. 2022b, MNRAS, 512, 661, doi: 10.1093/mnras/stac311
Karchev, K., Grayling, M., Boyd, B. M., et al. 2024, MNRAS, 530, 3881, doi: 10.1093/mnras/stae995
Karchev, K., & Trotta, R. 2025, JCAP, 2025, 031, doi: 10.1088/1475-7516/2025/07/031
Karchev, K., Trotta, R., & Jimenez, R. 2025, arXiv e-prints, arXiv:2508.15899, doi: 10.48550/arXiv.2508.15899
Karchev, K., Trotta, R., & Weniger, C. 2023a, arXiv e-prints, arXiv:2311.15650, doi: 10.48550/arXiv.2311.15650
—. 2023b, MNRAS, 520, 1056, doi: 10.1093/mnras/stac3785
Kelly, P. L., Hicken, M., Burke, D. L., Mandel, K. S., & Kirshner, R. P. 2010, ApJ, 715, 743, doi: 10.1088/0004-637X/715/2/743
Kelly, R. P., Warne, D. J., Frazier, D. T., et al. 2025, arXiv e-prints, arXiv.2503.12315, doi: 10.48550/arXiv.2503.12315
Kennamer, N., Ishida, E. E. O., Gonzalez-Gaitan, S., et al. 2020, arXiv e-prints, arXiv:2010.05941, doi: 10.48550/arXiv.2010.05941
Kéruzoré, F. 2025, arXiv e-prints, arXiv:2509.22478, doi: 10.48550/arXiv.2509.22478
Kerzendorf, W. E., & Sim, S. A. 2014, MNRAS, 440, 387, doi: 10.1093/mnras/stu055
Kerzendorf, W. E., Vogl, C., Buchner, J., et al. 2021, ApJL, 910, L23, doi: 10.3847/2041-8213/abeb1b
Kessler, R., Bernstein, J. P., Cinabro, D., et al. 2009, PASP, 121, 1028, doi: 10.1086/605984
Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463, doi: 10.1093/mnras/stw2672
Knödlseder, J. 2025, arXiv e-prints, arXiv:2507.14510, doi: 10.48550/arXiv.2507.14510
Knop, R., & ELAsTiCC Team. 2023, in American Astronomical Society Meeting Abstracts, Vol. 55, American Astronomical Society Meeting Abstracts, 117.02
Koblischke, N., Jang, H., Menou, K., & Ali-Dib, M. 2025, arXiv e-prints, arXiv:2501.18411, doi: 10.48550/arXiv.2501.18411
Koblischke, N., Parker, L., Lanusse, F., et al. 2024, arXiv e-prints, arXiv:2512.11982, doi: 10.48550/arXiv.2512.11982
Kobyzev, I., Prince, S. J., & Brubaker, M. A. 2021, IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, 3964, doi: 10.1109/TPAMI.2020.2992934
Korytov, D., Hearin, A., Kovacs, E., et al. 2019, ApJS, 245, 26, doi: 10.3847/1538-4365/ab510c
Kosmyna, N., Hauptmann, E., Yuan, Y. T., et al. 2025, arXiv e-prints, arXiv:2506.08872, doi: 10.48550/arXiv.2506.08872
Kratochvil, J. M., Lim, E. A., Wang, S., et al. 2012, PhRvD, 85, 103513, doi: 10.1103/PhysRevD.85.103513
Krippendorf, S., Baron Perez, N., Bulbul, E., et al. 2024, A&A, 682, A132, doi: 10.1051/0004-6361/202346826
Krisciunas, K., Contreras, C., Burns, C. R., et al. 2017, AJ, 154, 211, doi: 10.3847/1538-3881/aa8df0
Krishnaraj, V., Bayer, A. E., Jespersen, C. K., & Melchior, P. 2025, arXiv e-prints, arXiv:2510.19168. https://arxiv.org/abs/2510.19168
Kunz, M., Bassett, B. A., & Hlozek, R. A. 2007, PhRvD, 75, 103508, doi: 10.1103/PhysRevD.75.103508
Lacopo, G., Lepinzan, M. D., Goz, D., et al. 2026, arXiv e-prints, arXiv:2601.01935, doi: 10.48550/arXiv.2601.01935
Lampeitl, H., Smith, M., Nichol, R. C., et al. 2010, ApJ, 722, 566, doi: 10.1088/0004-637X/722/1/566
Lanusse, F., Lanzieri, D., Kabalan, W., Simon-Onfroy, H., & Boucaud, A. 2025, JaxPM: JAX-powered Cosmological Particle-Mesh N-body Solver, v0.1.6. https://github.com/DifferentiableUniverseInitiative/JaxPM
Lanusse, F., Ma, Q., Li, N., et al. 2018, MNRAS, 473, 3895, doi: 10.1093/mnras/stx1665
Lanusse, F., Melchior, P., & Moolekamp, F. 2019, arXiv e-prints, arXiv:1912.03980, doi: 10.48550/arXiv.1912.03980
Lanzieri, D., Lanusse, F., Modi, C., et al. 2023, A&A, 679, A61, doi: 10.1051/0004-6361/202346888
Lanzieri, D., Zeghal, J., Lucas Makinen, T., et al. 2025, A&A, 697, A162, doi: 10.1051/0004-6361/202451535

Laverick, A., Surrao, K., Zubeldia, I., et al. 2024, arXiv e-prints, arXiv:2412.00431, doi: 10.48550/arXiv.2412.00431

Law, N. M., Corbett, H., Galliher, N. W., et al. 2022, PASP, 134, 035003, doi: 10.1088/1538-3873/ac4811

Lawrence, E., Heitmann, K., Kwan, J., et al. 2017, ApJ, 847, 50, doi: 10.3847/1538-4357/aa86a9

Le Folgoc, L., Baltatzis, V., Desai, S., et al. 2021, arXiv e-prints, arXiv:2110.04286, doi: 10.48550/arXiv.2110.04286

Legin, R., Hezaveh, Y., Perreault Levasseur, L., & Wandelt, B. 2021, in Machine Learning and the Physical Sciences Workshop, 95, doi: 10.48550/arXiv.2112.05278

Legin, R., Hezaveh, Y., Perreault-Levasseur, L., & Wandelt, B. 2023, ApJ, 943, 4, doi: 10.3847/1538-4357/aca7c2

Legin, R., Stone, C., Adam, A., et al. 2025, arXiv e-prints, arXiv:2511.19595. https://arxiv.org/abs/2511.19595

Leistedt, B., Alsing, J., Peiris, H., Mortlock, D., & Leja, J. 2023, ApJS, 264, 23, doi: 10.3847/1538-4365/ac9d99

Leistedt, B., Mortlock, D. J., & Peiris, H. V. 2016, MNRAS, 460, 4258, doi: 10.1093/mnras/stw1304

Leja, J., Johnson, B. D., Conroy, C., van Dokkum, P. G., & Byler, N. 2017, ApJ, 837, 170, doi: 10.3847/1538-4357/aa5ffe

Lemos, P., Coogan, A., Hezaveh, Y., & Perreault-Levasseur, L. 2023, in Proceedings of Machine Learning Research, Vol. 202, Proceedings of the 40th International Conference on Machine Learning, ed. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (PMLR), 19256–19273, doi: 10.48550/arXiv.2302.03026

Lemos, P., Malkin, N., Handley, W., et al. 2024, in Proceedings of Machine Learning Research, Vol. 235, Proceedings of the 41st International Conference on Machine Learning, ed. R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (PMLR), 27230–27253, doi: 10.48550/arXiv.2312.03911

Lemos, P., Sharief, S., Malkin, N., et al. 2024, arXiv e-prints, arXiv:2402.04355, doi: 10.48550/arXiv.2402.04355

Lemson, G., & Virgo Consortium. 2006, arXiv e-prints, doi: 10.48550/arXiv.astro-ph/0608019

Leoni, M., Ishida, E. E. O., Peloton, J., & Möller, A. 2022, A&A, 663, A13, doi: 10.1051/0004-6361/202142715

Leung, H. W., & Bovy, J. 2024, MNRAS, 527, 1494, doi: 10.1093/mnras/stad3015

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473, doi: 10.1086/309179

Lewis, P., Perez, E., Piktus, A., et al. 2020, in Advances in Neural Information Processing Systems 33, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Curran Associates, Inc.), 9459–9474, doi: 10.48550/arXiv.2005.11401

Li, T., Collett, T. E., Krawczyk, C. M., & Enzi, W. 2024, MNRAS, 527, 5311, doi: 10.1093/mnras/stad3514

Li, W., Chen, H.-Y., Lin, Q., et al. 2025a, arXiv e-prints, arXiv:2510.06200, doi: 10.48550/arXiv.2510.06200

Li, X., & Mandelbaum, R. 2023, MNRAS, 521, 4904, doi: 10.1093/mnras/stad890

Li, X., Mandelbaum, R., & The LSST Dark Energy Science Collaboration. 2025b, MNRAS, 536, 3663, doi: 10.1093/mnras/stae2764

Li, Y., Leja, J., Johnson, B. D., et al. 2025c, ApJ, 986, 9, doi: 10.3847/1538-4357/adcab4

Li, Y., Modi, C., Jamieson, D., et al. 2024, ApJS, 270, 36, doi: 10.3847/1538-4365/ad0ce7

Li, Y., Wu, C.-Y., Fan, H., et al. 2022, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4794–4804, doi: 10.1109/CVPR52688.2022.00476

Li, Y., Lu, L., Modi, C., et al. 2022, arXiv e-prints, arXiv:2211.09958, doi: 10.48550/arXiv.2211.09958

Li, Y., Fu, L., Chen, Z., et al. 2025d, Research in Astronomy and Astrophysics, 25, 055021, doi: 10.1088/1674-4527/adcc7e

Liang, S., Adari, P., & von der Linden, A. 2025, arXiv e-prints, arXiv:2503.16680, doi: 10.48550/arXiv.2503.16680

Liang, Y., Melchior, P., Lu, S., Goulding, A., & Ward, C. 2023, AJ, 166, 75, doi: 10.3847/1538-3881/ace100

Lin, Z., Huang, N., Avestruz, C., et al. 2021, MNRAS, 507, 4149, doi: 10.1093/mnras/stab2229
Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166, doi: 10.1111/j.1365-2966.2010.17432.x
Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x
Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. 2022, arXiv e-prints, arXiv:2210.02747, doi: 10.48550/arXiv.2210.02747
List, F., Anau Montel, N., & Weniger, C. 2023, arXiv e-prints, arXiv:2310.19910, doi: 10.48550/arXiv.2310.19910
List, F., Hahn, O., Flöss, T., & Winkler, L. 2025, arXiv e-prints, arXiv:2510.05206, doi: 10.48550/arXiv.2510.05206
Liu, F., Chen, D., Guan, Z., et al. 2024, IEEE Transactions on Geoscience and Remote Sensing, 62, 1, doi: 10.1109/TGRS.2024.3390838
Liu, Z., Hu, H., Lin, Y., et al. 2022, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11999–12009, doi: 10.1109/CVPR52688.2022.01170
Lochner, M., & Bassett, B. A. 2021, Astronomy and Computing, 36, 100481, doi: 10.1016/j.ascom.2021.100481
Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31, doi: 10.3847/0067-0049/225/2/31
Lochner, M., & Rudnick, L. 2025, AJ, 169, 121, doi: 10.3847/1538-3881/ada14c
Lokken, M., Gagliano, A., Narayan, G., et al. 2023, MNRAS, 520, 2887, doi: 10.1093/mnras/stad302
López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, A&A, 622, A177, doi: 10.1051/0004-6361/201732480
Lovick, T., Yallup, D., Piras, D., Spurio Mancini, A., & Handley, W. 2025, arXiv e-prints, arXiv:2509.13307, doi: 10.48550/arXiv.2509.13307
Lu, T., Haiman, Z., & Li, X. 2023, MNRAS, 521, 2050, doi: 10.1093/mnras/stad686
Lüber, A., Karchev, K., Fisher, C., et al. 2025, ApJL, 984, L32, doi: 10.3847/2041-8213/adc7aa
Lueckmann, J., Gonçalves, P. J., Bassetto, G., et al. 2017, in Advances in Neural Information Processing Systems 30, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 1289–1299, doi: 10.48550/arXiv.1711.01861
Lueckmann, J.-M., Bassetto, G., Karaletsos, T., & Macke, J. H. 2019, in Proceedings of Machine Learning Research, Vol. 96, Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference, ed. F. Ruiz, C. Zhang, D. Liang, & T. Bui (PMLR), 32–53. https://proceedings.mlr.press/v96/lueckmann19a.html
Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., & Macke, J. 2021, in Proceedings of Machine Learning Research, Vol. 130, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, ed. A. Banerjee & K. Fukumizu (PMLR), 343–351. https://proceedings.mlr.press/v130/lueckmann21a.html
Lyu, H., Alvey, J., Anau Montel, N., Pieroni, M., & Weniger, C. 2025, arXiv e-prints, arXiv:2510.13997, doi: 10.48550/arXiv.2510.13997
Ma, J., He, Y., Li, F., et al. 2024, Nature Communications, 15, 654, doi: 10.1038/s41467-024-44824-z
MacMahon-Gellér, C., Leonard, C. D., Bull, P., & Rau, M. M. 2025, arXiv e-prints, arXiv:2504.00552, doi: 10.48550/arXiv.2504.00552
Magee, M. R., Siebenaler, L., Maguire, K., Ackley, K., & Killestein, T. 2024, MNRAS, 531, 3042, doi: 10.1093/mnras/stae1233
Makinen, T. L., Charnock, T., Alsing, J., & Wandelt, B. D. 2021, JCAP, 2021, 049, doi: 10.1088/1475-7516/2021/11/049
Makinen, T. L., Sui, C., Wandelt, B. D., Porqueres, N., & Heavens, A. 2024, arXiv e-prints, arXiv:2410.07548, doi: 10.48550/arXiv.2410.07548
Malz, A. I. 2021, PhRvD, 103, 083502, doi: 10.1103/PhysRevD.103.083502
Malz, A. I., & Hogg, D. W. 2022, ApJ, 928, 127, doi: 10.3847/1538-4357/ac062f
Malz, A. I., Hložek, R., Allam, Jr., T., et al. 2019, AJ, 158, 171, doi: 10.3847/1538-3881/ab3a2f

Mandel, K. S., Narayan, G., & Kirshner, R. P. 2011, ApJ, 731, 120, doi: 10.1088/0004-637X/731/2/120
Mandel, K. S., Scolnic, D. M., Shariff, H., Foley, R. J., & Kirshner, R. P. 2017, ApJ, 842, 93, doi: 10.3847/1538-4357/aa6038
Mandel, K. S., Thorp, S., Narayan, G., Friedman, A. S., & Avelino, A. 2022, MNRAS, 510, 3939, doi: 10.1093/mnras/stab3496
Mandel, K. S., Wood-Vasey, W. M., Friedman, A. S., & Kirshner, R. P. 2009, ApJ, 704, 629, doi: 10.1088/0004-637X/704/1/629
Mandelbaum, R. 2018, ARA&A, 56, 393, doi: 10.1146/annurev-astro-081817-051928
Mandelbaum, R., Hirata, C. M., Ishak, M., Seljak, U., & Brinkmann, J. 2006, MNRAS, 367, 611, doi: 10.1111/j.1365-2966.2005.09946.x
Mandelbaum, R., Blake, C., Bridle, S., et al. 2011, MNRAS, 410, 844, doi: 10.1111/j.1365-2966.2010.17485.x
March, M. C., Trotta, R., Berkes, P., Starkman, G. D., & Vaudrevange, P. M. 2011, MNRAS, 418, 2308, doi: 10.1111/j.1365-2966.2011.19584.x
Margalef-Bentabol, B., Wang, L., La Marca, A., & Rodriguez-Gomez, V. 2024a, arXiv e-prints, arXiv:2410.01437, doi: 10.48550/arXiv.2410.01437
Margalef-Bentabol, B., Wang, L., La Marca, A., et al. 2024b, A&A, 687, A24, doi: 10.1051/0004-6361/202348239
Marques, G. A., Liu, J., Shirasaki, M., et al. 2024, MNRAS, 528, 4513–4527, doi: 10.1093/mnras/stae098
Masson, M., & Bregeon, J. 2024, arXiv e-prints, arXiv:2412.05061, doi: 10.48550/arXiv.2412.05061
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53, doi: 10.1088/0004-637X/813/1/53
Matheson, T., Stubens, C., Wolf, N., et al. 2021, AJ, 161, 107, doi: 10.3847/1538-3881/abd703
McEwen, J. D., Wallis, C. G. R., Price, M. A., & Spurio Mancini, A. 2021, arXiv e-prints, arXiv:2111.12720, doi: 10.48550/arXiv.2111.12720
Melchior, P., Joseph, R., Sanchez, J., MacCrann, N., & Gruen, D. 2021, Nature Reviews Physics, 3, 712, doi: 10.1038/s42254-021-00353-y
Melchior, P., Moolekamp, F., Jerdee, M., et al. 2018, Astronomy and Computing, 24, 129, doi: 10.1016/j.ascom.2018.07.001
Melo, A., Cañameras, R., Schuldt, S., et al. 2025, A&A, 698, A264, doi: 10.1051/0004-6361/202453195
Merz, G., Liu, Y., Burke, C. J., et al. 2023, MNRAS, 526, 1122, doi: 10.1093/mnras/stad2785
Merz, G., Liu, X., Schmidt, S., et al. 2025, The Open Journal of Astrophysics, 8, 40, doi: 10.33232/001c.136809
Meshcheryakov, A. V., Nemeshaeva, A., Burenin, R. A., Gilfanov, M. R., & Sunyaev, R. A. 2022, Astronomy Letters, 48, 479, doi: 10.1134/S1063773722090055
Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119, doi: 10.1051/0004-6361/201832797
Miller, B. K., Cole, A., Forré, P., Louppe, G., & Weniger, C. 2021, in Advances in Neural Information Processing Systems 34, ed. M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan, 129–143, doi: 10.48550/arXiv.2107.01214
Mishra, S., Trotta, R., & Viel, M. 2025, arXiv e-prints, arXiv:2506.08086, doi: 10.48550/arXiv.2506.08086
Mishra-Sharma, S., & Yang, G. 2022, in Machine Learning for Astrophysics, 34, doi: 10.48550/arXiv.2206.14820
Mitchener, L., Yiu, A., Chang, B., et al. 2025, arXiv e-prints, arXiv:2511.02824, doi: 10.48550/arXiv.2511.02824
Mizrahi, D., Bachmann, R., Kar, O. F., et al. 2023, arXiv e-prints, arXiv:2312.06647, doi: 10.48550/arXiv.2312.06647
Modi, C., Lanusse, F., & Seljak, U. 2021, Astronomy and Computing, 37, 100505, doi: 10.1016/j.ascom.2021.100505
Möller, A., & de Boissière, T. 2020, MNRAS, 491, 4277, doi: 10.1093/mnras/stz3312
Möller, A., & de Boissière, T. 2022, in Machine Learning for Astrophysics, 21, doi: 10.48550/arXiv.2207.04578
Möller, A., Peloton, J., Ishida, E. E. O., et al. 2021, MNRAS, 501, 3272, doi: 10.1093/mnras/staa3602
Möller, A., Smith, M., Sako, M., et al. 2022, MNRAS, 514, 5159, doi: 10.1093/mnras/stac1691

Möller, A., Wiseman, P., Smith, M., et al. 2024, MNRAS, 533, 2073, doi: 10.1093/mnras/stae1953
Möller, A., Ishida, E., Peloton, J., et al. 2025, PASA, 42, e057, doi: 10.1017/pasa.2025.20
Moran, K. R., Heitmann, K., Lawrence, E., et al. 2023, MNRAS, 520, 3443, doi: 10.1093/mnras/stac3452
More, A., Cañameras, R., Jaelani, A. T., et al. 2024, MNRAS, 533, 525, doi: 10.1093/mnras/stae1597
Moretti, C., Autenrieth, M., Serra, R., et al. 2025, The Open Journal of Astrophysics, 8, 50, doi: 10.33232/001c.137525
Morgan, R., Nord, B., Bechtol, K., et al. 2022, ApJ, 927, 109, doi: 10.3847/1538-4357/ac5178
Morisset, C., Charlot, S., Sánchez, S. F., et al. 2025, MNRAS, 538, 1884, doi: 10.1093/mnras/staf143
Morningstar, W. R., Hezaveh, Y. D., Perreault Levasseur, L., et al. 2018, arXiv e-prints, arXiv:1808.00011, doi: 10.48550/arXiv.1808.00011
Morningstar, W. R., Perreault Levasseur, L., Hezaveh, Y. D., et al. 2019, ApJ, 883, 14, doi: 10.3847/1538-4357/ab35d7
Moskowitz, I., Gawiser, E., Bault, A., et al. 2023, ApJ, 950, 49, doi: 10.3847/1538-4357/accc88
Moskowitz, I., Gawiser, E., Crenshaw, J. F., et al. 2024, ApJL, 967, L6, doi: 10.3847/2041-8213/ad4039
Mündler, N., Dekoninck, J., & Vechev, M. 2025, arXiv e-prints, arXiv:2508.10111, doi: 10.48550/arXiv.2508.10111
Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002, doi: 10.1088/1538-3873/ab1609
Myles, J., Alarcon, A., Amon, A., et al. 2021, MNRAS, 505, 4249, doi: 10.1093/mnras/stab1515
Nalisnick, E., Matsukawa, A., Whye Teh, Y., Gorur, D., & Lakshminarayanan, B. 2018, arXiv e-prints, arXiv:1810.09136, doi: 10.48550/arXiv.1810.09136
Narayan, G., & ELAsTiCC Team. 2023, in American Astronomical Society Meeting Abstracts, Vol. 241, American Astronomical Society Meeting Abstracts #241, 117.01
Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, ApJS, 236, 9, doi: 10.3847/1538-4365/aab781
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493, doi: 10.1086/304888
Neal, R. M. 2011, in Handbook of Markov Chain Monte Carlo, 1st edn., ed. S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Chapman and Hall/CRC), 113–162, doi: 10.1201/b10905-6
Nelson, D., Springel, V., Pillepich, A., et al. 2019, Computational Astrophysics and Cosmology, 6, 2, doi: 10.1186/s40668-019-0028-x
Newman, J. A., & Gruen, D. 2022, ARA&A, 60, 363, doi: 10.1146/annurev-astro-032122-014611
Newman, S. L., Lovell, C. C., Maraston, C., et al. 2025, MNRAS, doi: 10.1093/mnras/staf1866
Nightingale, J. W., Dye, S., & Massey, R. J. 2018, MNRAS, 478, 4738, doi: 10.1093/mnras/sty1264
Nordin, J., Brinnel, V., van Santen, J., Reusch, S., & Kowalski, M. 2025, A&A, 698, A13, doi: 10.1051/0004-6361/202452481
Nordin, J., Brinnel, V., van Santen, J., et al. 2019, A&A, 631, A147, doi: 10.1051/0004-6361/201935634
Ntampaka, M., ZuHone, J., Eisenstein, D., et al. 2019, ApJ, 876, 82, doi: 10.3847/1538-4357/ab14eb
Omori, K. C., Bottrell, C., Walmsley, M., et al. 2023, A&A, 679, A142, doi: 10.1051/0004-6361/202346743
Omori, K. C., Bottrell, C., Bellstedt, S., et al. 2025, ApJ, 989, 73, doi: 10.3847/1538-4357/ade989
OpenUniverse, LSST Dark Energy Science Collaboration, Roman HLIS Project Infrastructure, et al. 2025, MNRAS, 544, 3799, doi: 10.1093/mnras/staf1833
O'Ryan, D., Merín, B., Simmons, B. D., et al. 2023, ApJ, 948, 40, doi: 10.3847/1538-4357/acc0ff
Pandya, S., Patel, P., Nord, B. D., Walmsley, M., & Ćiprijanović, A. 2025a, Machine Learning: Science and Technology, 6, 035032, doi: 10.1088/2632-2153/adf701
Pandya, S., Yang, Y., Van Alfen, N., Blazek, J., & Walters, R. 2025b, The Open Journal of Astrophysics, 8, 51749, doi: 10.33232/001c.151749
Papamakarios, G., & Murray, I. 2016, in Advances in Neural Information Processing Systems 29, ed. D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett, 1028–1036, doi: 10.48550/arXiv.1605.06376

Papamakarios, G., Murray, I., & Pavlakou, T. 2017, in Advances in Neural Information Processing Systems 30, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 2338–2347, doi: 10.48550/arXiv.1705.07057

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. 2021, Journal of Machine Learning Research, 22, 1. http://jmlr.org/papers/v22/19-1028.html

Papamakarios, G., Sterratt, D., & Murray, I. 2019, in Proceedings of Machine Learning Research, Vol. 89, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, ed. K. Chaudhuri & M. Sugiyama (PMLR), 837–848. https://proceedings.mlr.press/v89/papamakarios19a.html

Park, J. W., Wagner-Carena, S., Birrer, S., et al. 2021, ApJ, 910, 39, doi: 10.3847/1538-4357/abdfc4

Parker, L., Lanusse, F., Golkar, S., et al. 2024, MNRAS, 531, 4990, doi: 10.1093/mnras/stae1450

Parker, L. H., Lanusse, F., Shen, J., et al. 2025, in The Thirty-ninth Annual Conference on Neural Information Processing Systems. https://openreview.net/forum?id=6gJ2ZykQ5W

Patel, A., Zhang, T., Avestruz, C., Regier, J., & The LSST Dark Energy Science Collaboration. 2025, AJ, 170, 155, doi: 10.3847/1538-3881/adef32

Payot, N., Barco, G. M., Perreault-Levasseur, L., & Hezaveh, Y. 2025, MLforAstro at ICML Workshop. https://ml4astro.github.io/icml2025/assets/camera_ready/35_Blind_Strong_Gravitational_.pdf

Payot, N., Lemos, P., Perreault-Levasseur, L., et al. 2023, arXiv e-prints, arXiv:2311.18017, doi: 10.48550/arXiv.2311.18017

Pearson, J., Li, N., & Dye, S. 2019, MNRAS, 488, 991, doi: 10.1093/mnras/stz1750

Pereira, L. M., Salazar, A., & Vergara, L. 2023, IEEE Access, 11, 84283, doi: 10.1109/ACCESS.2023.3296098

Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJL, 850, L7, doi: 10.3847/2041-8213/aa9704

Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129, doi: 10.1093/mnras/stx2052

—. 2019, MNRAS, 482, 807, doi: 10.1093/mnras/sty2683

Phan, D., Pradhan, N., & Jankowiak, M. 2019, arXiv e-prints, arXiv:1912.11554, doi: 10.48550/arXiv.1912.11554

Pimentel, Ó., Estévez, P. A., & Förster, F. 2023, AJ, 165, 18, doi: 10.3847/1538-3881/ac9ab4

Piras, D., & Spurio Mancini, A. 2023, The Open Journal of Astrophysics, 6, 20, doi: 10.21105/astro.2305.06347

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

PLAsTiCC team, Allam, Jr., T., Bahmanyar, A., et al. 2018, arXiv e-prints, arXiv:1810.00001, doi: 10.48550/arXiv.1810.00001

Poh, J., Samudre, A., Ćiprijanović, A., et al. 2025, JCAP, 2025, 053, doi: 10.1088/1475-7516/2025/05/053

Popovic, B., Kenworthy, W. D., Ginolin, M., et al. 2025, arXiv e-prints, arXiv:2506.05471, doi: 10.48550/arXiv.2506.05471

Porqueres, N., Heavens, A., Mortlock, D., Lavaux, G., & Makinen, T. L. 2023, arXiv e-prints, arXiv:2304.04785, doi: 10.48550/arXiv.2304.04785

Porqueres, N., Kodi Ramanah, D., Jasche, J., & Lavaux, G. 2019, A&A, 624, A115, doi: 10.1051/0004-6361/201834844

Prat, J., Gatti, M., Doux, C., et al. 2025, arXiv e-prints, arXiv:2506.13439, doi: 10.48550/arXiv.2506.13439

Preechakul, K., Chatthee, N., Wizadwongsa, S., & Suwajanakorn, S. 2021, arXiv e-prints, arXiv:2111.15640, doi: 10.48550/arXiv.2111.15640

Pujol, A., Bobin, J., Sureau, F., Guinot, A., & Kilbinger, M. 2020, A&A, 643, A158, doi: 10.1051/0004-6361/202038658

Qu, H., Sako, M., Möller, A., & Doux, C. 2021, AJ, 162, 67, doi: 10.3847/1538-3881/ac0824

RAIL Team, van den Busch, J. L., Charles, E., et al. 2025, arXiv e-prints, arXiv:2505.02928, doi: 10.48550/arXiv.2505.02928

Ramachandra, N., Valogiannis, G., Ishak, M., Heitmann, K., & LSST Dark Energy Science Collaboration. 2021, PhRvD, 103, 123525, doi: 10.1103/PhysRevD.103.123525

Rampf, C., List, F., & Hahn, O. 2025, JCAP, 2025, 020, doi: 10.1088/1475-7516/2025/02/020

Rathore, P., Lei, W., Frangella, Z., Lu, L., & Udell, M. 2024, in Proceedings of Machine Learning Research, Vol. 235, Proceedings of the 41st International Conference on Machine Learning, ed. R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (PMLR), 42159–42191. https://proceedings.mlr.press/v235/rathore24a.html

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2015, arXiv e-prints, arXiv:1506.02640, doi: 10.48550/arXiv.1506.02640

Redmon, J., & Farhadi, A. 2016, arXiv e-prints, arXiv:1612.08242, doi: 10.48550/arXiv.1612.08242

—. 2018, arXiv e-prints, arXiv:1804.02767, doi: 10.48550/arXiv.1804.02767

Remy, B., Lanusse, F., Jeffrey, N., et al. 2023, A&A, 672, A51, doi: 10.1051/0004-6361/202243054

Reza, M., Zhang, Y., Avestruz, C., et al. 2024, arXiv e-prints, arXiv:2409.20507, doi: 10.48550/arXiv.2409.20507

Reza, M., Zhang, Y., Nord, B., et al. 2022, in Machine Learning for Astrophysics, 20, doi: 10.48550/arXiv.2208.00134

Rezaie, M., Seo, H.-J., Ross, A. J., & Bunescu, R. C. 2020, MNRAS, 495, 1613, doi: 10.1093/mnras/staa1231

Ribli, D., Dobos, L., & Csabai, I. 2019, MNRAS, 489, 4847, doi: 10.1093/mnras/stz2374

Riess, A. G., Yuan, W., Macri, L. M., et al. 2022, ApJL, 934, L7, doi: 10.3847/2041-8213/ac5c5b

Rizhko, M., & Bloom, J. S. 2025, AJ, 170, 28, doi: 10.3847/1538-3881/adcbad

Roberts, E., Lochner, M., Fonseca, J., et al. 2017, JCAP, 2017, 036, doi: 10.1088/1475-7516/2017/10/036

Robnik, J., Luca, G. B. D., Silverstein, E., & Seljak, U. 2023, Journal of Machine Learning Research, 24, 1. http://jmlr.org/papers/v24/22-1450.html

Robnik, J., & Seljak, U. 2024, in Proceedings of Machine Learning Research, Vol. 253, Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference, ed. J. Antorán & C. A. Naesseth (PMLR), 111–126, doi: 10.48550/arXiv.2303.18221

Rossky, P. J., Doll, J. D., & Friedman, H. L. 1978, The Journal of Chemical Physics, 69, 4628, doi: 10.1063/1.436415

Rowe, B. 2010, MNRAS, 404, 350, doi: 10.1111/j.1365-2966.2010.16277.x

Rowe, B., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121, doi: https://doi.org/10.1016/j.ascom.2015.02.002

Roy, A., Feldman, S., Klupar, P., et al. 2026, arXiv e-prints, arXiv:2601.02556, doi: 10.48550/arXiv.2601.02556

Roy, S., Schmude, J., Lal, R., et al. 2025, arXiv e-prints, arXiv:2508.14112, doi: 10.48550/arXiv.2508.14112

Rozet, F., Andry, G., Lanusse, F., & Louppe, G. 2024, in Advances in Neural Information Processing Systems 38, ed. A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang, 87647–87682, doi: 10.52202/079017-2783

Rubin, D., Aldering, G., Barbary, K., et al. 2015, ApJ, 813, 137, doi: 10.1088/0004-637X/813/2/137

Rubin, D., Aldering, G., Betoule, M., et al. 2025, ApJ, 986, 231, doi: 10.3847/1538-4357/adc0a5

Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104, doi: 10.1088/0004-637X/785/2/104

Rykoff, E. S., Rozo, E., Hollowood, D., et al. 2016, ApJS, 224, 1, doi: 10.3847/0067-0049/224/1/1

Sampson, M. L., Melchior, P., Ward, C., & Birmingham, S. 2024, Astronomy and Computing, 49, 100875, doi: 10.1016/j.ascom.2024.100875

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141, doi: 10.3847/1538-3881/abd5c1

Savchenko, O., List, F., Abellán, G. F., Anau Montel, N., & Weniger, C. 2024, arXiv e-prints, arXiv:2410.15808, doi: 10.48550/arXiv.2410.15808

Saxena, A., Meerburg, P. D., Weniger, C., Acedo, E. d. L., & Handley, W. 2024, RAS Techniques and Instruments, 3, 724–736, doi: 10.1093/rasti/rzae047

Schaefer, C., Geiger, M., Kuntzer, T., & Kneib, J. P. 2018, A&A, 611, A2, doi: 10.1051/0004-6361/201731201
Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, MNRAS, 499, 1587, doi: 10.1093/mnras/staa2799
Schneider, M. D., Hogg, D. W., Marshall, P. J., et al. 2015, ApJ, 807, 87, doi: 10.1088/0004-637X/807/1/87
Schneider, P., Ehlers, J., & Falco, E. E. 1992, Gravitational Lenses, doi: 10.1007/978-3-662-03758-4
Schuldt, S., Cañameras, R., Shu, Y., et al. 2023a, A&A, 671, A147, doi: 10.1051/0004-6361/202244325
Schuldt, S., Suyu, S. H., Cañameras, R., et al. 2023b, A&A, 673, A33, doi: 10.1051/0004-6361/202244534
Schuldt, S., Suyu, S. H., Meinhardt, T., et al. 2021, A&A, 646, A126, doi: 10.1051/0004-6361/202039574
Schuldt, S., Cañameras, R., Andika, I. T., et al. 2025, A&A, 693, A291, doi: 10.1051/0004-6361/202450927
Schwartz, M. D. 2026, arXiv e-prints, arXiv:2601.02484, doi: 10.48550/arXiv.2601.02484
Secco, L., Samuroff, S., Krause, E., et al. 2022, PhRvD, 105, 023515, doi: 10.1103/physrevd.105.023515
Settles, B. 2009, Active Learning Literature Survey, Tech. Rep. TR1648, University of Wisconsin–Madison. http://digital.library.wisc.edu/1793/60660
Sevilla-Noarbe, I., Hoyle, B., Marchã, M. J., et al. 2018, MNRAS, 481, 5451, doi: 10.1093/mnras/sty2579
Shah, V. G., Gagliano, A., Malanchev, K., et al. 2025, ApJ, 995, 4, doi: 10.3847/1538-4357/ae1130
Shajib, A. J., Smith, G. P., Birrer, S., et al. 2025, Philosophical Transactions of the Royal Society of London Series A, 383, 20240117, doi: 10.1098/rsta.2024.0117
Shao, W., Fan, D., Cui, C., et al. 2026, Information Fusion, 130, 104103, doi: https://doi.org/10.1016/j.inffus.2025.104103
Shariff, H., Jiao, X., Trotta, R., & van Dyk, D. A. 2016, ApJ, 827, 1, doi: 10.3847/0004-637X/827/1/1
Sheldon, E. S., & Huff, E. M. 2017, ApJ, 841, 24, doi: 10.3847/1538-4357/aa704b
Shen, Y., & Gagliano, A. T. 2025a, arXiv e-prints, arXiv:2505.03063, doi: 10.48550/arXiv.2505.03063
—. 2025b, arXiv e-prints, arXiv:2507.16817, doi: 10.48550/arXiv.2507.16817
Shu, Y., Cañameras, R., Schuldt, S., et al. 2022, A&A, 662, A4, doi: 10.1051/0004-6361/202243203
Shukla, N., Romeo, A., Caravita, C., et al. 2025, in Proceedings of the 22nd ACM International Conference on Computing Frontiers: Workshops and Special Sessions, ed. F. Palmubo, A. Tumeo, A. L. Varbanescu, & Y. Simmhan, CF '25 Companion (Association for Computing Machinery), 177–184, doi: 10.1145/3706594.3728892
Simon, H., Lanusse, F., & de Mattia, A. 2025, JCAP, 2025, 039, doi: 10.1088/1475-7516/2025/12/039
Skarlinski, M. D., Cox, S., Laurent, J. M., et al. 2024, arXiv e-prints, arXiv:2409.13740, doi: 10.48550/arXiv.2409.13740
Skilling, J. 2006, Bayesian Analysis, 1, 833. https://doi.org/10.1214/06-BA127
Slater, C. T., Ivezić, Ž., & Lupton, R. H. 2020, AJ, 159, 65, doi: 10.3847/1538-3881/ab6166
Smith, K. W., Williams, R. D., Young, D. R., et al. 2019, Research Notes of the AAS, 3, 26, doi: 10.3847/2515-5172/ab020f
Smith, K. W., Smartt, S. J., Young, D. R., et al. 2020, PASP, 132, 085002, doi: 10.1088/1538-3873/ab936e
Smith, M. J., & Geach, J. E. 2023, Royal Society Open Science, 10, 221454, doi: 10.1098/rsos.221454
Smith, R. E., Peacock, J. A., Jenkins, A., et al. 2003, MNRAS, 341, 1311, doi: 10.1046/j.1365-8711.2003.06503.x
Son, J., Lee, Y.-W., Chung, C., Park, S., & Cho, H. 2025, MNRAS, 544, 975, doi: 10.1093/mnras/staf1685
Song, Y., & Ermon, S. 2019, in Advances in Neural Information Processing Systems 32, ed. H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett, 11895–11907, doi: 10.48550/arXiv.1907.05600

Song, Y., Shen, L., Xing, L., & Ermon, S. 2021, arXiv e-prints, arXiv:2111.08005, doi: 10.48550/arXiv.2111.08005

Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al. 2020, arXiv e-prints, arXiv:2011.13456, doi: 10.48550/arXiv.2011.13456

Song, Y., Villar, V. A., Martínez-Galarza, R., & Dillmann, S. 2025, ApJ, 988, 143, doi: 10.3847/1538-4357/add72e

Soumagnac, M. T., Abdalla, F. B., Lahav, O., et al. 2015, MNRAS, 450, 666, doi: 10.1093/mnras/stu1410

Speagle, J. S. 2020, MNRAS, 493, 3132, doi: 10.1093/mnras/staa278

Srinivasan, R., Barausse, E., Korsakova, N., & Trotta, R. 2025, PhRvD, 112, 103043, doi: 10.1103/shym-w46f

Srinivasan, R., Crisostomi, M., Trotta, R., Barausse, E., & Breschi, M. 2024, PhRvD, 110, doi: 10.1103/physrevd.110.123007

Stein, G., Seljak, U., Böhm, V., et al. 2022, ApJ, 935, 5, doi: 10.3847/1538-4357/ac7c08

Stevens, A. R. H., Bellstedt, S., Elahi, P. J., & Murphy, M. T. 2020, Nature Astronomy, 4, 843, doi: 10.1038/s41550-020-1169-1

Stopyra, S., Peiris, H. V., Pontzen, A., Jasche, J., & Lavaux, G. 2024, MNRAS, 527, 1244, doi: 10.1093/mnras/stad3170

Suchyta, E., Huff, E. M., Aleksić, J., et al. 2016, MNRAS, 457, 786, doi: 10.1093/mnras/stv2953

Sui, C., Pandey, S., & Wandelt, B. D. 2025, arXiv e-prints, arXiv:2507.07833, doi: 10.48550/arXiv.2507.07833

Sullivan, M., Le Borgne, D., Pritchet, C. J., et al. 2006, ApJ, 648, 868, doi: 10.1086/506137

Sullivan, M., Conley, A., Howell, D. A., et al. 2010, MNRAS, 406, 782, doi: 10.1111/j.1365-2966.2010.16731.x

Sun, Q., Liu, Z., Ma, C., et al. 2025a, arXiv e-prints, arXiv:2505.19897, doi: 10.48550/arXiv.2505.19897

Sun, Z., Ting, Y.-S., Liang, Y., et al. 2025b, arXiv e-prints, arXiv:2510.08354, doi: 10.48550/arXiv.2510.08354

Suyu, S. H., Marshall, P. J., Hobson, M. P., & Blandford, R. D. 2006, MNRAS, 371, 983, doi: 10.1111/j.1365-2966.2006.10733.x

Swierc, P., Tamargo-Arizmendi, M., Ćiprijanović, A., & Nord, B. D. 2024, arXiv e-prints, arXiv:2410.16347, doi: 10.48550/arXiv.2410.16347

Swierc, P., Zhao, M., Ćiprijanović, A., & Nord, B. 2023, arXiv e-prints, arXiv:2311.17238, doi: 10.48550/arXiv.2311.17238

Tak, H., Mandel, K., van Dyk, D. A., et al. 2017, Annals of Applied Statistics, 11, 1309, doi: doi:10.1214/17-AOAS1027

Takahashi, R., Sato, M., Nishimichi, T., Taruya, A., & Oguri, M. 2012, ApJ, 761, 152, doi: 10.1088/0004-637X/761/2/152

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. 2018, arXiv e-prints, arXiv:1804.06788, doi: 10.48550/arXiv.1804.06788

Tan, A., Protopapas, P., Cádiz-Leyton, M., et al. 2025, arXiv e-prints, arXiv:2509.24134, doi: 10.48550/arXiv.2509.24134

Tauman Kalai, A., Nachum, O., Vempala, S. S., & Zhang, E. 2025, arXiv e-prints, arXiv:2509.04664, doi: 10.48550/arXiv.2509.04664

TDCOSMO Collaboration, Birrer, S., Buckley-Geer, E. J., et al. 2025, A&A, 704, A63, doi: 10.1051/0004-6361/202555801

Tewes, M., Kuntzer, T., Nakajima, R., et al. 2019, A&A, 621, A36, doi: 10.1051/0004-6361/201833775

The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, arXiv e-prints, arXiv:1809.01669, doi: 10.48550/arXiv.1809.01669

Thorp, S., Alsing, J., Peiris, H. V., et al. 2024a, ApJ, 975, 145, doi: 10.3847/1538-4357/ad7736

Thorp, S., & Mandel, K. S. 2022, MNRAS, 517, 2360, doi: 10.1093/mnras/stac2714

Thorp, S., Mandel, K. S., Jones, D. O., Kirshner, R. P., & Challis, P. M. 2024b, MNRAS, 530, 4016, doi: 10.1093/mnras/stae1111

Thorp, S., Mandel, K. S., Jones, D. O., Ward, S. M., & Narayan, G. 2021, MNRAS, 508, 4310, doi: 10.1093/mnras/stab2849

Thorp, S., Peiris, H. V., Mortlock, D. J., et al. 2025a, ApJS, 276, 5, doi: 10.3847/1538-4365/ad8ebd

Thorp, S., Peiris, H. V., Jagwani, G., et al. 2025b, ApJ, 993, 240, doi: 10.3847/1538-4357/ae0936
Tian, D.-C., Yang, Y., Wen, Z.-L., & Xia, J.-Q. 2025, ApJS, 276, 21, doi: 10.3847/1538-4365/ad8bbd
Tian, Y., Chen, X., & Ganguli, S. 2021, arXiv e-prints, arXiv:2102.06810, doi: 10.48550/arXiv.2102.06810
Tortorelli, L., Fischbacher, S., Robotham, A. S. G., Nussbaumer, C., & Refregier, A. 2025, arXiv e-prints, arXiv:2509.00150, doi: 10.48550/arXiv.2509.00150
Trotta, R. 2008, Contemporary Physics, 49, 71–104, doi: 10.1080/00107510802066753
—. 2025, Nature Astronomy, doi: 10.1038/s41550-025-02738-w
Troxel, M. A., & Ishak, M. 2015, PhR, 558, 1, doi: 10.1016/j.physrep.2014.11.001
Uzsoy, A. S. M., Thorp, S., Grayling, M., & Mandel, K. S. 2024, MNRAS, 535, 2306, doi: 10.1093/mnras/stae2465
Van Alfen, N., Campbell, D., Blazek, J., et al. 2024, The Open Journal of Astrophysics, 7, 45, doi: 10.33232/001c.118783
van den Busch, J. L., Wright, A. H., Hildebrandt, H., et al. 2022, A&A, 664, A170, doi: 10.1051/0004-6361/202142083
Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, arXiv e-prints, arXiv:1706.03762, doi: 10.48550/arXiv.1706.03762
Vegetti, S., & Koopmans, L. V. E. 2009, MNRAS, 392, 945, doi: 10.1111/j.1365-2966.2008.14005.x
Venkatraman, P., Erickson, S., Marshall, P., et al. 2025, arXiv e-prints, arXiv:2510.20778. https://arxiv.org/abs/2510.20778
Ver Steeg, G., & Galstyan, A. 2021, in Advances in Neural Information Processing Systems 34, ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Curran Associates, Inc.), 11012–11025, doi: 10.48550/arXiv.2111.02434
Villaescusa-Navarro, F., Wandelt, B. D., Anglés-Alcázar, D., et al. 2022, ApJ, 928, 44, doi: 10.3847/1538-4357/ac54a5
Villaescusa-Navarro, F., Bolliet, B., Villanueva-Domingo, P., et al. 2025. https://arxiv.org/abs/2510.26887
Vincenzi, M., Sullivan, M., Möller, A., et al. 2023, MNRAS, 518, 1106, doi: 10.1093/mnras/stac1404
von Wietersheim-Kramsta, M., Lin, K., Tessore, N., et al. 2025, A&A, 694, A223, doi: 10.1051/0004-6361/202450487
Wagner-Carena, S., Park, J. W., Birrer, S., et al. 2021, ApJ, 909, 187, doi: 10.3847/1538-4357/abdf59
Walmsley, M., Scaife, A. M. M., Lintott, C., et al. 2022a, MNRAS, 513, 1581, doi: 10.1093/mnras/stac525
Walmsley, M., Lintott, C., Géron, T., et al. 2022b, MNRAS, 509, 3966, doi: 10.1093/mnras/stab2093
Walmsley, M., Géron, T., Kruk, S., et al. 2023, MNRAS, 526, 4768, doi: 10.1093/mnras/stad2919
Wang, B., Leja, J., Bezanson, R., et al. 2023, ApJL, 944, L58, doi: 10.3847/2041-8213/acba99
Wang, B. Y., & Thiele, L. 2025, arXiv e-prints, arXiv:2507.20378, doi: 10.48550/arXiv.2507.20378
Ward, C., Melchior, P., Sampson, M. L., et al. 2025, Astronomy and Computing, 51, 100930, doi: 10.1016/j.ascom.2025.100930
Ward, S. M., Thorp, S., Mandel, K. S., et al. 2023, ApJ, 956, 111, doi: 10.3847/1538-4357/acf7bb
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11, doi: 10.3847/1538-4365/ac3078
Wildberger, J. B., Dax, M., Buchholz, S., et al. 2023, in Machine Learning for Astrophysics, 34, doi: 10.48550/arXiv.2305.17161
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835, doi: 10.1093/mnras/stt1458
Williams, M. J., Veitch, J., & Messenger, C. 2021, PhRvD, 103, 103006, doi: 10.1103/PhysRevD.103.103006
Williams, R. D., Francis, G. P., Lawrence, A., et al. 2024, RAS Techniques and Instruments, 3, 362, doi: 10.1093/rasti/rzae024
Wojtak, R., Hjorth, J., & Gall, C. 2019, MNRAS, 487, 3342, doi: 10.1093/mnras/stz1516
Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020a, A&A, 637, A100, doi: 10.1051/0004-6361/201936782
Wright, A. H., Hildebrandt, H., van den Busch, J. L., et al. 2020b, A&A, 640, L14, doi: 10.1051/0004-6361/202038389
—. 2025, A&A, 703, A144, doi: 10.1051/0004-6361/202554909

Xia, B., Ramachandra, N., Wells, A. I., Habib, S., & Wise, J. 2025, arXiv e-prints, arXiv:2510.07684, doi: 10.48550/arXiv.2510.07684

Xie, T., Zhou, Y., Liang, Z., Favaro, S., & Sesia, M. 2025, arXiv e-prints, arXiv:2510.13037, doi: 10.48550/arXiv.2510.13037

Xu, K., & Ge, H. 2024, in Proceedings of Machine Learning Research, Vol. 235, Proceedings of the 41st International Conference on Machine Learning, ed. R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (PMLR), 54999–55014. https://proceedings.mlr.press/v235/xu24i.html

Xu, L., Sarkar, M., Lonappan, A. I., et al. 2025, arXiv e-prints, arXiv:2507.07257, doi: 10.48550/arXiv.2507.07257

Yamada, Y., Tjarko Lange, R., Lu, C., et al. 2025, arXiv e-prints, arXiv:2504.08066, doi: 10.48550/arXiv.2504.08066

Yang, L., Zhang, Z., Song, Y., et al. 2023, ACM Computing Surveys, 56, 105, doi: 10.1145/3626235

Ye, C., Yuan, S., Cooray, S., et al. 2025, arXiv e-prints, arXiv:2510.24591. https://arxiv.org/abs/2510.24591

Zaheer, M., Guruganesh, G., Dubey, K. A., et al. 2020, in Advances in Neural Information Processing Systems 33, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, 17283–17297, doi: 10.48550/arXiv.2007.14062

Zhang, G., Helfer, T., Gagliano, A. T., Mishra-Sharma, S., & Ashley Villar, V. 2024, Machine Learning: Science and Technology, 5, 045069, doi: 10.1088/2632-2153/ad990d

Zhang, Y.-H., Zuntz, J., Moskowitz, I., et al. 2025, arXiv e-prints, arXiv:2508.20903, doi: 10.48550/arXiv.2508.20903

Zhou, A. J., Li, X., Dodelson, S., & Mandelbaum, R. 2024a, PhRvD, 110, 023539, doi: 10.1103/PhysRevD.110.023539

Zhou, A. J., Li, Y., Dodelson, S., et al. 2024b, JCAP, 2024, 069, doi: 10.1088/1475-7516/2024/10/069

Zhou, L., Ling, H., Fu, C., et al. 2025, arXiv preprint arXiv:2510.09901. https://arxiv.org/abs/2510.09901

Zivanovic, U., Di Gioia, S., Scaffidi, A., et al. 2025, arXiv e-prints, arXiv:2505.20535, doi: 10.48550/arXiv.2505.20535

Zubeldia, Í., Bolliet, B., Challinor, A., & Handley, W. 2025, PhRvD, 112, 083536, doi: 10.1103/drn7-ggqk

Zuntz, J., Lanusse, F., Malz, A. I., et al. 2021, The Open Journal of Astrophysics, 4, 13, doi: 10.21105/astro.2108.13418

Zuo, X., Tao, Y., Huang, Y., et al. 2026, AJ, 171, 10, doi: 10.3847/1538-3881/ae1467