Strong Lens Challenge Data Release

The Strong Lensing Data Challenge and `slsim` team:

August 25, 2025

1 Introduction

The Strong Lens Data Challenge is a data challenge aimed at providing both real and simulated data to foster the development of Machine Learning algorithms for the classification of Strong Gravitational Lenses as observed by the LSST. Our main objective is to provide training and test datasets generated in accordance to the expected Vera C. Rubin Observatory capabilities [1] and rank classification algorithms applied to such data. As an extra objective, we also provide lens parameters for testing regression metrics on lens modeling. We use this technical release to detail the challenge's datasets, structure and scientific background.

2 Motivation and background

With the advent of now current generation telescopes and surveys, the computational power and cost needed to process the amount of data generated has grown exponentially, with the LSST alone generating 20terabytes of data per night alone, and the data management facility requiring 150 teraflops, the same as the world's most powerful supercomputer around 20 years ago [2]. Such amount of data and processing power has demanded new management, processing and modeling methods from scientists and engineers working in Astrophysics.

The revolution in Machine Learning methods brought by Deep Learning, mainly for the tasks of classification, regression and representation of large datasets, has allowed the processing and modeling of large amounts of data without the need for direct human intervention and validation, in the case of unsupervised learning. In the realm of physical science, the impact has already been felt and the methods have become standard in the literature [3, 4],particularly in the case of galaxy surveys [5].

Data Challenges aimed at preparing the scientific community for the release of future and current survey data and fostering the development of algorithms and machine learning models have been effective in stimulating collaboration and innovation and introducing new methodologies in data classification and regression, proving to be one of the best ways to engage the community with a common objective.

The LSST Strong Lensing Data Challenge is one in a series of Strong Lensing Data challenges [6] aimed at developing and classification algorithms in the task of identifying strong gravitational lenses with a focus on data generated to be as accurate as possible to the upcoming data of the LSST Survey. In this data challenge, we also offer the possibility of testing regression algorithms, providing a manifold of parameters for lens modeling and inference.

For this data challenge, we have simulated gravitational lenses using data from LSST DP0 and HSC PDR3, while using the `slsim` code to inject lenses into the DP0 data, and the `simct` code to inject into the HSC PDR3 data, processes which are detailed in this note.

We hope that this Challenge provides a solid basis for the further development of Machine Learning techniques in Strong Lenses classification and modeling in the LSST era.

3 `SLSim` and DP0 injection

We simulate a strong lens population, consisting of galaxy-galaxy lenses, using the `SLSim` pipeline, a pipeline capable of modeling realistic gravitational lens populations including galaxy-galaxy lenses, lensed quasars, lensed supernovae, and cluster-scale lenses. The lens galaxies are drawn from a population of massive elliptical galaxies, and source galaxies are drawn from a population of blue galaxies, consistent with observational data and theoretical models. Each lens system includes parameters such as the deflector mass profile, the source light profile, and the lensing geometry. The simulation produces image cutouts with realistic lensing features such as arcs and multiple images.

To enable lens discovery in Rubin data-like environments, we inject these simulated lens systems into the DP0 1 year coadd images. The injection process overlays the simulated lens image cutouts onto selected regions in the DP0 sky images, taking care of PSF convolution and match noise levels. For each injection, a location is chosen that avoids bright sources and masked areas. The injected images are saved as new FITS files and visualized to ensure a realistic appearance and data compatibility. This framework supports large-scale injection of simulated lenses into LSST-like data and provides a testing and training sets for developing and validating lens-finding algorithms in realistic survey conditions. For more detailed information, please visit `SLSim` github repo: https://github.com/LSST-strong-lensing/slsim

4 Degraded real HSC data

HSC images of lenses and galaxies are degraded to match the characteristics of LSST 1-year coadd data using our custom pipeline. This degradation includes adjustments to the zero-point magnitude, PSF full width at half maximum (FWHM), pixel scale, Poisson noise, and background noise. The PSF is modified by convolving it with a Gaussian kernel such that the average FWHM along the

x and y axes matches the target LSST FWHM. The pixel scale of the HSC image is resampled to $0.2''$ using the `reproject` Python package, which performs flux-preserving transformations of astronomical images. Finally, we add Gaussian background noise and source Poisson noise to match the expected RMS and exposure time of the LSST DP0 1-year coadds.

Using this pipeline, we degrade both simulated HST lenses generated with the `simct` code and real galaxy and lens images. The real lenses are drawn from categories A and B of the SuGOHI sample [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], including cluster lenses with images within $4''$ from the core. For non-lenses, the sample includes: (i) 33% spiral galaxies from [20] in the HSC Wide field, selected to have a Kron radius smaller than $2''$ to match the typical Einstein radius of galaxy-scale lenses; (ii) 30% luminous red galaxies (LRGs) from the HSC COSMOS Deep field, selected using color cuts similar to those in [21] but extended by 0.5 mag towards the blue to include galaxies with blue features or companions; (iii) 5% galaxies from the *GalaxyZoo: Hubble* COSMOS catalog [22], exhibiting at least one of the following morphological features: spiral arms, rings, clumpiness, or mergers; and (iv) 32% randomly selected galaxies from the HSC COSMOS Deep field with $r < 24$. For all galaxies in the HSC COSMOS Deep field, we apply an additional redshift filter, selecting only those with COSMOS photometric redshifts $z > 0.2$.

4.1 SIMCT lenses

We generate simulated galaxy-galaxy lens images around potential lensing galaxies from the Hyper Suprime Cam (HSC) Public data release-3. We have used the SIMCT [1] tool, which can generate in situ realistic simulated lensed images for all types of static lens systems from real or simulated galaxy (or groups/clusters) catalogs [23].

We start with a population of massive elliptical galaxies taken from HSC PDR3 to identify which galaxies can act as potential lenses based on the background source luminosity function, and their angular lens cross section, and the lens mass model is defined by taking the light properties of the galaxy (position, ellipticity, magnitudes and redshift). We assume mass follows light and use the relation between luminosity and velocity dispersion to obtain the latter for a given lens galaxy.

The background source is defined with a parametric model with parameters drawn from the redshift and luminosity function distributions from [24]. To produce realistic looking lens systems, only blue galaxies (g - $i < 1.5$) have been chosen and the colors of the background galaxies are drawn from the HSC PDR3 galaxies, matched in the *magnitude - redshift* plane. The source positions with respect to the lens are drawn randomly from an area inside the caustic, and the source size is estimated from the luminosity–size relation in [25]. This allows us to define all of the parameters of the Sersic model for the lensed galaxy.

We use GRAVLENS [26] to simulate lensed images using the lens and source

[1] `https://github.com/anumore/simct/tree/revised/code/gal_lens`

models as discussed above. We apply a magnitude cut for the second brightest lensed image to be brighter than 25 to ensure that at least two of the lensed images are detectable. We add shot noise to the images and convolve them with a median point spread function obtained for the respective HSC fields for each band. These images are then superposed on the respective real galaxy images from HSC in each band. To avoid using an already known lens within the HSC survey, we remove all Grade A and B lenses from the SuGOHI sample [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. The superposed images are saved to multi-extension FITS files. Using this approach, we generate $> 10^5$ galaxy-galaxy lenses from the HSC deep-ultradeep and Wide fields.

5 Dataset Description

The Challenge will consist of two datasets: the validation dataset, used for training and validation, which we'll release at the start of the dataset and contains labeled data for the different objects; and the blind test dataset, which will be released at a later date and will contain unlabeled data for the final test of the classification and regression models provided by the teams.

5.1 Validation Dataset

The validation dataset will consist of $200k$ objects, equally divided in 4 different classes:

- Simulated `slsim` lenses injected in DP0 data,
- Simulated `slsim` non-lenses injected in DP0 data,
- Degraded simulated HSC lenses,
- Degraded HSC non-lenses,

generated as described in this document. For each class of objects, one shall find a `.fits` file containing the parameters for all of the objects in the given class, and a directory containing, for each object, 5 `.fits` files corresponding to each band. Each .fits file contains a primary empty `hdu` and a secondary `hdu` containing two columns, one for the Object's ID, `Lens ID`, or `Object ID` in the case of non-lenses, and one column for the image of the object at the corresponding band `band_`. Each column contains a single row, describing the Object's ID and containing the 41×41 image of the object in the given band.

The naming convention for the files is of the form `DX_I000YYYYY_b.fits`, where `X=1` for `slsim` data and `X=2` in the case of HSC data. `I=L` denotes lenses and I=N non lenses, `YYYYY` denotes the label of the indexing of the object, and `b=g,r,i,z,y` denotes the corresponding band .

The parameter file, `parameters.fits` contains a primary empty `hdu` and a secondary `hdu` containing the parameters for each object (lenses and non-lenses). In the case of lenses, the columns are

- **Lens ID**: Object ID.
- **RA**: Right Ascension of the center of the object cutout in degrees.
- **Dec**: Declination of the center of the object cutout in degrees.
- **zlens**: Redshift of deflector.
- **vel_disp**: Velocity dispersion of deflector object given in km/s.
- **ell_m**: Ellipticity of deflector's mass.
- **ell_m_PA**: Position angle of the deflector's mass ellipticity.
- **sh**: Shear of the deflector.
- **sh_PA**: Position angle of the deflector's shear.
- **Rein**: Einstein radius of the deflector object, given in arcsec.
- **ell_l**: Ellipticity of the deflector's light.
- **ell_l_PA**: Position angle of the deflector's light.
- **Reff_l**: Effective radius of deflector, given in arcsec.
- **n_l_sers**: Sérsic index of the deflector's Sérsic profile.
- **mag_lens{}**: Magnitude of the deflector for a given band (g, r, i, z, y).
- **zsrc**: Redshift of source.
- **srcx, srcy**: Extended source object position in the cutout.
- **mag_src{}**: Unlensed magnitude of the source for a given band (g, r, i, z, y).
- **ell_s**: Ellipticity of source.
- **ell_s_PA**: Position angle of the source's ellipticity.
- **Reff_s**: Effective radius of source, given in arcsec.
- **n_s_sers**: Sérsic index of the source's Sérsic profile.

And in the case of non-lenses, the columns are

- **Object ID**: The object's ID.
- **RA**: Right Ascension of the center of the object cutout in degrees.
- **Dec**: Declination of the center of the object cutout in degrees.
- **z_central**: Redshift of the object.

- **mag_lens{}**: Magnitude of the object for a given band (g, r, i, z, y).

Some of the columns contain nonphysical values such as -999, which means that in the process of simulating the lens, the value could not be obtained by the lens modeling or the physical parameter was not properly measured. The resulting submitted .csv file with the predictions must be filled with -999 for such parameters or parameters which are not found for a certain type of object.

Some objects may have their id duplicated. Such objects won't be found in the parameters files, and won't be included in the predictions for the validation datasets.

5.2 Test Dataset

The final blind test dataset will contain $100k$ unlabeled objects, generated in accordance to the codes described above. The naming of the objects will follow a uniform convention and follow a similar format as the validation dataset, now only containing the images themselves and the ID column in order to identify the object.

References

[1] Anowar J. Shajib†, Graham P. Smith, Simon Birrer, Aprajita Verma, Nikki Arendse, Thomas E. Collett, Tansu Daylan, and Stephen Serjeant. Strong gravitational lenses from the Vera C. Rubin Observatory. *Phil. Trans. Roy. Soc. Lond. A*, 383(2295):20240117, 2025.

[2] Lsst data management. `https://www.lsst.org/about/dm`, 2025. Accesed 2025-07-9.

[3] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91(4):045002, 2019.

[4] Cora Dvorkin et al. Machine Learning and Cosmology. In *Snowmass 2021*, 3 2022.

[5] Marc Huertas-Company and François Lanusse. The DAWES review 10: The impact of deep learning for the analysis of galaxy surveys. *Publ. Astron. Soc. Austral.*, 40:e001, 2023.

[6] Lsst strong lensing science collaboration:science and activities. `https://sites.google.com/view/lsst-stronglensing/science`, 2025. Accesed 2025-07-9.

[7] Alessandro Sonnenfeld, James H. H. Chan, Yiping Shu, Anupreeta More, Masamune Oguri, Sherry H. Suyu, Kenneth C. Wong, Chien-Hsiu Lee, Jean Coupon, Atsunori Yonehara, Adam S. Bolton, Anton T. Jaelani,

Masayuki Tanaka, Satoshi Miyazaki, and Yutaka Komiyama. Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). I. Automatic search for galaxy-scale strong lenses. *PASJ*, 70:S29, January 2018.

[8] Kenneth C. Wong, Alessandro Sonnenfeld, James H. H. Chan, Cristian E. Rusu, Masayuki Tanaka, Anton T. Jaelani, Chien-Hsiu Lee, Anupreeta More, Masamune Oguri, Sherry H. Suyu, and Yutaka Komiyama. Survey of Gravitationally Lensed Objects in HSC Imaging (SuGOHI). II. Environments and Line-of-Sight Structure of Strong Gravitational Lens Galaxies to z $\sim$ 0.8. *ApJ*, 867(2):107, November 2018.

[9] Alessandro Sonnenfeld, Anton T. Jaelani, James Chan, Anupreeta More, Sherry H. Suyu, Kenneth C. Wong, Masamune Oguri, and Chien-Hsiu Lee. Survey of gravitationally-lensed objects in HSC imaging (SuGOHI). III. Statistical strong lensing constraints on the stellar IMF of CMASS galaxies. *A&A*, 630:A71, October 2019.

[10] James H. H. Chan, Sherry H. Suyu, Alessandro Sonnenfeld, Anton T. Jaelani, Anupreeta More, Atsunori Yonehara, Yuriko Kubota, Jean Coupon, Chien-Hsiu Lee, Masamune Oguri, Cristian E. Rusu, and Kenneth C. Wong. Survey of Gravitationally lensed Objects in HSC Imaging (SuGOHI). IV. Lensed quasar search in the HSC survey. *A&A*, 636:A87, April 2020.

[11] Anton T. Jaelani, Anupreeta More, Masamune Oguri, Alessandro Sonnenfeld, Sherry H. Suyu, Cristian E. Rusu, Kenneth C. Wong, James H. H. Chan, Issha Kayo, Chien-Hsiu Lee, Dani C. Y. Chao, Jean Coupon, Kaiki T. Inoue, and Toshifumi Futamase. Survey of Gravitationally lensed Objects in HSC Imaging (SuGOHI) - V. Group-to-cluster scale lens search from the HSC-SSP Survey. *MNRAS*, 495(1):1291–1310, June 2020.

[12] Alessandro Sonnenfeld, Aprajita Verma, Anupreeta More, Elisabeth Baeten, Christine Macmillan, Kenneth C. Wong, James H. H. Chan, Anton T. Jaelani, Chien-Hsiu Lee, Masamune Oguri, Cristian E. Rusu, Marten Veldthuis, Laura Trouille, Philip J. Marshall, Roger Hutchings, Campbell Allen, James O'Donnell, Claude Cornen, Christopher P. Davis, Adam McMaster, Chris Lintott, and Grant Miller. Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). VI. Crowdsourced lens finding with Space Warps. *A&A*, 642:A148, October 2020.

[13] Anton T. Jaelani, Cristian E. Rusu, Issha Kayo, Anupreeta More, Alessandro Sonnenfeld, John D. Silverman, Malte Schramm, Timo Anguita, Naohisa Inada, Daichi Kondo, Paul L. Schechter, Khee-Gan Lee, Masamune Oguri, James H. H. Chan, Kenneth C. Wong, and Kaiki T. Inoue. Survey of Gravitationally Lensed Objects in HSC Imaging (SuGOHI) - VII. Discovery and confirmation of three strongly lensed quasars. *MNRAS*, 502(1):1487–1493, March 2021.

[14] Kenneth C. Wong, James H. H. Chan, Dani C. Y. Chao, Anton T. Jaelani, Issha Kayo, Chien-Hsiu Lee, Anupreeta More, and Masamune Oguri. Survey of Gravitationally lensed objects in HSC Imaging (SuGOHI). VIII. New galaxy-scale lenses from the HSC SSP. *PASJ*, 74(5):1209–1219, October 2022.

[15] James H. H. Chan, Kenneth C. Wong, Xuheng Ding, Dani Chao, I. Non Chiu, Anton T. Jaelani, Issha Kayo, Anupreeta More, Masamune Oguri, and Sherry H. Suyu. Survey of gravitationally lensed objects in HSC imaging (SuGOHI) - IX. Discovery of strongly lensed quasar candidates. *MNRAS*, 527(3):6253–6275, January 2024.

[16] Anton T. Jaelani, Anupreeta More, Kenneth C. Wong, Kaiki T. Inoue, Dani C. Y. Chao, Premana W. Premadi, and Raoul Cañameras. Survey of gravitationally lensed objects in HSC imaging (SuGOHI) - X. Strong lens finding in the HSC-SSP using convolutional neural networks. *MNRAS*, 535(2):1625–1639, December 2024.

[17] R. Cañameras, S. Schuldt, Y. Shu, S. H. Suyu, S. Taubenberger, T. Meinhardt, L. Leal-Taixé, D. C. Y. Chao, K. T. Inoue, A. T. Jaelani, and A. More. HOLISMOKES. VI. New galaxy-scale strong lens candidates from the HSC-SSP imaging survey. *A&A*, 653:L6, September 2021.

[18] Yiping Shu, Raoul Cañameras, Stefan Schuldt, Sherry H. Suyu, Stefan Taubenberger, Kaiki Taro Inoue, and Anton T. Jaelani. HOLISMOKES. VIII. High-redshift, strong-lens search in the Hyper Suprime-Cam Subaru Strategic Program. *A&A*, 662:A4, June 2022.

[19] S. Schuldt, R. Cañameras, I. T. Andika, S. Bag, A. Melo, Y. Shu, S. H. Suyu, S. Taubenberger, and C. Grillo. HOLISMOKES: XIII. Strong-lens candidates at all mass scales and their environments from the Hyper-Suprime Cam and deep learning. *A&A*, 693:A291, January 2025.

[20] Ken-ichi Tadaki, Masanori Iye, Hideya Fukumoto, Masao Hayashi, Cristian E. Rusu, Rhythm Shimakawa, and Tomoka Tosaki. Spin parity of spiral galaxies II: a catalogue of 80 k spiral galaxies using big data from the Subaru Hyper Suprime-Cam survey and deep learning. *MNRAS*, 496(4):4276–4286, August 2020.

[21] K. Rojas, E. Savary, B. Clément, M. Maus, F. Courbin, C. Lemon, J. H. H. Chan, G. Vernardos, R. Joseph, R. Cañameras, and A. Galan. Search of strong lens systems in the Dark Energy Survey using convolutional neural networks. *A&A*, 668:A73, December 2022.

[22] Kyle W. Willett, Melanie A. Galloway, Steven P. Bamford, Chris J. Lintott, Karen L. Masters, Claudia Scarlata, B. D. Simmons, Melanie Beck, Carolin N. Cardamone, Edmond Cheung, Edward M. Edmondson, Lucy F. Fortson, Roger L. Griffith, Boris Häußler, Anna Han, Ross Hart, Thomas Melvin, Michael Parrish, Kevin Schawinski, R. J. Smethurst, and Arfon M.

Smith. Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. *MNRAS*, 464(4):4176–4203, February 2017.

[23] Anupreeta More, Aprajita Verma, Philip J. Marshall, Surhud More, Elisabeth Baeten, Julianne Wilcox, Christine Macmillan, Claude Cornen, Amit Kapadia, Michael Parrish, Chris Snyder, Christopher P. Davis, Raphael Gavazzi, Chris J. Lintott, Robert Simpson, David Miller, Arfon M. Smith, Edward Paget, Prasenjit Saha, Rafael Küng, and Thomas E. Collett. SPACE WARPS- II. New gravitational lens candidates from the CFHTLS discovered through citizen science. *MNRAS*, 455(2):1191–1210, January 2016.

[24] C. Faure, J. P. Kneib, S. Hilbert, R. Massey, G. Covone, A. Finoguenov, A. Leauthaud, J. E. Taylor, S. Pires, N. Scoville, and Anton M. Koekemoer. On the Contribution of Large-Scale Structure to Strong Gravitational Lensing. *ApJ*, 695(2):1233–1243, April 2009.

[25] Mariangela Bernardi, Ravi K. Sheth, James Annis, Scott Burles, Daniel J. Eisenstein, Douglas P. Finkbeiner, David W. Hogg, Robert H. Lupton, David J. Schlegel, Mark SubbaRao, Neta A. Bahcall, John P. Blakeslee, J. Brinkmann, Francisco J. Castander, Andrew J. Connolly, István Csabai, Mamoru Doi, Masataka Fukugita, Joshua Frieman, Timothy Heckman, Gregory S. Hennessy, Željko Ivezić, G. R. Knapp, Don Q. Lamb, Timothy McKay, Jeffrey A. Munn, Robert Nichol, Sadanori Okamura, Donald P. Schneider, Aniruddha R. Thakar, and Donald G. York. Early-Type Galaxies in the Sloan Digital Sky Survey. III. The Fundamental Plane. *ApJ*, 125(4):1866–1881, April 2003.

[26] Charles R. Keeton, Daniel Christlein, and Ann I. Zabludoff. What Fraction of Gravitational Lens Galaxies Lie in Groups? *ApJ*, 545(1):129–140, December 2000.