LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)
Data Management

Host Galaxy Association for `DIAObjects`

M. L. Graham et al.

DMTN-TBD

Latest Revision: 2020-05-10

DRAFT

Abstract

This document argues that, in order to better enable extragalactic transient science with brokers, two new `DIAObject` catalog elements should be computed and included in the alert packets: (1) the `objectId` for the three `Object` catalog galaxies with the lowest separation distance (based on the galaxy's 2D luminosity profile) from the `DIAObject`, and (2) the separation distances for those three `Objects`.

Change Record

Version	Date	Description	Owner name
0	2020-02-26	Inception.	Melissa Graham

Document source location: `https://github.com/lsst-dm/dmtn-xxx`

Draft

Contents

1 Introduction **1**

2 Current DM Plans for Host Galaxy Association **1**

3 Options to Improve Host Galaxy Association **2**

3.1 Effective Radius 2

3.2 Second Moments 3

3.3 2D Algorithms 3

3.4 Hostless Transients 4

3.5 Galaxy/Transient Types 4

4 Recommendations **5**

4.1 Draft RFC 5

5 Appendix A: Separation Distance from Second Moments **7**

6 Appendix B: Overcoming Background Galaxy Contamination **9**

Draft

Host Galaxy Association for DIAObjects

1 Introduction

LSST will issue alert packets within 60 seconds for all sources detected during difference image analysis (DIA; DIASources), which are associated by sky coordinate into objects (DIAObject). Individuals and brokers processing alerts will use the information in these packets to rapidly evaluate and prioritize DIAObjects for follow-up with limited resources. Thus the contents of the alert packet have been designed to contain a sufficient amount of LSST data about each DIAObject to enable immediate analysis.

One important piece of information is the association of each DIAObject with a static-sky Object from the Data Release catalogs. Brokers will use the Object association to obtain data about the static-sky object from the DR catalogs, such as whether it might be galactic or extra-galactic, at high- or low-redshift, nuclear or offset from a host, etc. All of this information can help an alert stream user identify and prioritize their targets of interest, and delivering alerts *with* the static-sky association already completed avoids the situation of multiple users cross-matching in real time. The Object association for DIAObjects will of course also be used by scientists working with the Prompt or Data Release DIAObjects catalogs on longer timescales, not just alert consumers. However, the main goal of this document is to assess the best option – from a scientific perspective – for Object association that can be completed during the 60 second Alert Production timescale.

Below, Section 2 describes the current plan for associating DIAObjects with the DR Objects catalog; Section 3 presents and discusses options that have been used successfully by other surveys; and Section 4 makes recommendations for additions to the DIAObject catalog to improve associations for extragalactic transients and their host galaxies.

2 Current DM Plans for Host Galaxy Association

With respect to associations between Prompt DIAObjects and Data Release Objects, the contents of the alert packet as defined in the Data Products Definitions Document (DPDD; LSE-163) includes the following:

- `nearbyObj` (`unit64[6]`), the *"closest* `Objects` *(3 stars and 3 galaxies) in Data Release database"*
- `nearbyObjDist` (`float[6]`), the *"distances to* `nearbyObj`*"* in arcseconds
- `nearbyObjLnP` (`float[6]`), the *"natural log of the probability that the observed* `DIAObject` *is the same as the nearby* `Object`*"*

For the latter, there is a footnote that says *"This quantity will be computed by marginalizing over the product of position and proper motion error ellipses of the* `Object` *and* `DIAObject`*, assuming an appropriate prior"*.

The current definitions of `nearbyObj`, `nearbyObjDist`, and `nearbyObjLnP` are not as useful as they could be for transients in host galaxies. For extragalactic transients, the three nearest galaxies are not always the three most likely host galaxies, and the radial distance in arcseconds matters less than a *separation distance* which accounts for the galaxies' spatial luminosity profiles. Furthermore, the definition of `nearbyObjLnP` is only appropriate for static variable point sources (stars): for transients in host galaxies, the observed `DIAObject` will never be "the same as the nearby `Object`".

3 Options to Improve Host Galaxy Association

Statistically, the most likely host for a given transient is the galaxy which contributes the most optical flux at the transient's location. This is usually estimated by first calculating a *separation distance* from the nearby galaxies to the transient which is expressed in terms of the galaxy's spatial luminosity profile, and then assuming the galaxy with the lowest separation distance is the host. The following are several options for estimating which nearby galaxy is the most likely host of an extragalactic transient.

3.1 Effective Radius

For the separation distance, use the radial distance from the core of the galaxy to the location of the transient, divided by the effective radius of the galaxy. The DR `Objects` table is already planned to contain suitable effective radii such as the parameter `kronRad90` [LSE-163]. This option for the separation distance would not require any additional processing aside from dividing the radial distance from transient to `Object` by the effective radius of the `Object`.

Although this kind of separation distance would account for the relative sizes of the potential host galaxies, it does not account for their position angles, and so would not be as accurate for assessing potential high-inclination host galaxies.

3.2 Second Moments

Calculate a separation distance based on the two-dimensional luminosity profile of the nearby galaxies. For example, Sullivan et al. (2006) describe the method applied to the Supernova Legacy Survey (SNLS), using a separation distance of $R^2 = C_{xx}x_r^2 + C_{yy}y_r^2 + C_{xy}x_r y_r$, where C_{xx}, C_{yy}, and C_{xy} are ellipse parameters derived from the second moments of the galaxy luminosity profile and x_r,y_r are the on-sky distances between the centroids of the transient and the galaxy. The DR `Objects` table is already planned to contain the second moments of the galaxy luminosity profiles (`Ixx`, `Iyy`, and `Ixy`; LSE-163). A step-by-step description of how this separation distance can be calculated from planned `DIAObject` and `Object` table elements is provided in Section 5.

Multiple recent surveys have used this method (or similar) to associate transients with their host galaxies, such as Sako et al. (2018) for SDSS supernovae, and Gupta et al. (2016) who use real and simulated data to evaluate the optimal method for host association. This option for the separation distance requires slightly more computational steps, but would account for both the relative sizes and position angles of the potential host galaxies.

3.3 2D Algorithms

Aside from adopting a separation distance, there are more complicated methods for identifying the most likely host for a given transient. For example, the nearby galaxy with the smallest fraction of light interior to an isophot through the transient's location, where the isophot shape is given more degrees of freedom and not constrained to concentric ellipticals as in the second moment method above. Another example is to use an algorithm that provides deblended footprints for nearby extended objects, and that can estimate the fraction of light in given pixel that should be attributed to each (e.g., as the SCARLET deblender can do, Melchior et al. (2018)). The most likely host galaxy would be the one which contributes the most flux at the pixel location of a transient. While all DR `Objects` will be associated with a footprint (a region of connected pixels), the footprint information will not be stored in the `Object` table. The use of footprints in identifying potential host galaxies would require more

computational resources during Prompt processing, but would probably return a more accurate host association for only a very small fraction of DIAObjects.

3.4 Hostless Transients

For some scientific analyses, transients which are >3-5 effective radii away from the nearest galaxy, or for which $> 99\%$ of the potential host's luminosity is within the radial distance between transient and host center, are considered "hostless" (e.g., Sand et al. 2011). Such a cutoff has been appropriate for past samples of ~hundreds of transients, but will not be appropriate for the LSST sample size. Furthermore, the decision of whether and how to consider a transient "hostless" is best left as a scientific decision for the end-user. Thus, no such cut should be applied during the association of DIAObjects and Objects, and the most likely hosts should still be reported, even if the probability is low.

3.5 Galaxy/Transient Types

The association of transients with their host galaxy can be more accurate if their properties are also considered. For example, the potential host galaxy's redshift can be used to calculate separation distances in physical units, or to estimate the absolute brightness of the transient and consider whether it is physically plausible. Priors based on the established correlations between transient types and host galaxy morphology or color can also be used to refine a probabilistic host association, such as how core collapse supernovae are almost always associated with star formation (except for a few notable cases, e.g., Graham et al. 2012; Irani et al. 2019). A demonstration that these correlations between host and transient types are so robust that the host type can be used to provide a statistical classification of the transient type was presented by Foley & Mandel (2013).

However, making science-informed associations between transients and galaxies based on any kind of derived properties is beyond the scope of Prompt processing, and is best left to the users on a case-by-case basis. Thus, properties of the transients and/or the nearby galaxies (beyond their coordinates and luminosity profile) should not be used during the association of DIAObjects and Objects. For a very thorough assessment of an optimized, science-driven system for associating supernovae and their host galaxies – including the role and performance of machine learning methods – we direct the reader to Gupta et al. (2016).

Draft

4 Recommendations

1: Two new `DIAObject` catalog elements should be added: `potentialHost`, containing the `objectId` for the three galaxies with the lowest separation distances, and `potentialHostSeparation`, containing the separation distances for those three galaxies.

2: These two new `DIAObject` catalog elements should use a separation distance calculated with respect to the transient location using the second moments of each galaxy's luminosity profile (as described in Section 5).

These new catalog elements add unit64[3] and float[3] to the `DIAObject` catalog and to each alert, which is a small and worthwhile addition. Including the three galaxies (instead of the two or one) with the lowest separation distances is motivated by the high-reward science case of correctly associating transients in the outskirts of large nearby host galaxies, as described in Appendix 6. This science case also requires that separation distances be initially calculated for all galaxies within ~200 arcsec (i.e., $4R_e$ of a large nearby galaxy); how to identify this sub-set of galaxies in the vicinity (e.g.,. radial on-sky distance, HEALpix) is left as a technical implementation detail. Large bright nearby galaxies which are not in the `Object` catalogs should have their coordinates and second moments imported so that they are included in this potential host association process.

The existing `DIAObject` catalog elements `nearbyObj` and `nearbyObjDist` can remain unchanged and an analog for the existing `DIAObject` catalog element `nearbyObjLnP`, which represents the probability of association for static but variable point sources (stars), is not necessary for potential host galaxies.

4.1 Draft RFC

The following text should be posted as a Request For Comments (RFC) in Jira, and at the same time this DMTN should be made official and available. The RFC would spawn an LSST Change Request on the Data Products Definitions Document.

In order to better enable extragalactic transient science with brokers, it is proposed that two new `DIAObject` table elements be computed during Alert Production: (1) `potentialHost`, containing the `objectId` for the three `Object` catalog extended sources with the lowest separation

distances, and (2) potentialHostSeparation, the separation distances for those three Objects. The separation distance should be calculated with respect to the transient location using the second moments of each Object's luminosity profile, as described in DMTN-XXX. All parameters required for the calculation of the separation distances are already planned to be in the DIAObject and Object tables, and this change would only add unit64[3] and float[3] per DIAObject catalog entry, and per alert.

The change request on the DPDD would be to add two rows to DIAObject Table (Table 3), and a footnote.

potentialHost	unit64[3]	Three extended Objects with lowest separations in Data Release database[FN].
potentialHostSeparation	float[3]	Separations of potentialHost.

[FN] Separations should be calculated with respect to the transient location using the second moments of each Object's luminosity profile, as described in DMTN-XXX.

Draft

5 Appendix A: Separation Distance from Second Moments

From the LSST catalogs the following `Object` table elements are used to define the parameters needed to calculate the separation distance [LSE-163]:

Parameter	Unit	Table Element	Description
$x_{\rm trans}$, $y_{\rm trans}$	degrees	`DIAObject` `radec`	transient centroid
$x_{\rm gal}$, $y_{\rm gal}$	degrees	`Object` `radec`	galaxy centroid
$\overline{x^2}$,$\overline{y^2}$,$\overline{xy}$	arcsec^2	`Object` `Ixx`, `Iyy`, `Ixy`	galaxy second moments

There might be an issue with using the `Object` catalog second moments: the `Ixx`, `Iyy`, and `Ixy` are defined with with respect to the local tract/patch and not sky coordinates (Jira DM-19519). Although the local tangent projection will probably work fine for this application of the second moments, this should be verified at the time of implementation.

As described in Section 10 of E. Bertin's Source Extractor manual[1] (and presumably many other places), the unitless ellipse parameters C_{xx}, C_{yy}, C_{xy} can be calculated from the second moments via:

$$C_{xx} = \frac{\overline{y^2}}{\sqrt{\left(\frac{\overline{x^2}-\overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \tag{1}$$

$$C_{yy} = \frac{\overline{x^2}}{\sqrt{\left(\frac{\overline{x^2}-\overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \tag{2}$$

$$C_{xy} = -2\frac{\overline{xy}}{\sqrt{\left(\frac{\overline{x^2}-\overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \tag{3}$$

The sky distances between the transient and galaxy centroids are calculated as follows, and include the cos-dec factor and a conversion from units of degrees to arcseconds:

[1] Version 2.3: `https://www.astromatic.net/pubsvn/software/sextractor/trunk/doc/sextractor.pdf`

$$x_r = 3600(x_{\rm SN} - x_{\rm gal}) \tag{4}$$

$$y_r = 3600(y_{\rm SN} - y_{\rm gal}) \cos y_{\rm gal} \tag{5}$$

Finally, the separation distance R in arcseconds is calculated as:

$$R^2 = C_{xx} x_r^2 + C_{yy} y_r^2 + C_{xy} x_r y_r. \tag{6}$$

Draft

6 Appendix B: Overcoming Background Galaxy Contamination

In this appendix, the probability of failed host galaxy associations for nearby transients with large host offsets due to interloping background galaxies is quantified, and used to evaluate the minimum number of potential hosts that should be stored in a `DIAObect` record. For simplicity, this appendix uses the effective radius as the separation distance (Section 3.1).

The 10-year `Object` catalogs will include ~4 billion galaxies with $i < 25$ mag across the ~20000 deg^2 main survey area, known as the “gold" sample, especially in the context of weak lensing studies [LPM-17]. However, the 10-year coadded depths will detect galaxies down to 5σ limiting magnitudes of 26.1, 27.4, 27.5, 26.8, 26.1, and 24.9 mag in filters *ugrizy*; this is ~3 times as many as in the “gold" sample, or ~10 billion galaxies. This high density of background galaxies complicates the process for associating large nearby host galaxies with their transients, especially the rare transients in their outskirts.

Consider a transient at $3R_e$ from the center of a nearby galaxy with $z = 0.01$ and an effective radius of $R_e = 10$ kpc (~50 arcsec). In order for this transient to be associated with it's true host, the separation distance for all `Objects` within an area of at least $A_{3R_e} = \pi(3R_e)^2 = 0.0052$ deg^2 would need to be considered. Based on the final 10-year number of detected galaxies (10 billion) and the total survey area (20000 deg^2), that is $\sim (10^10/20000) \times 0.0052 \approx 2600$ galaxies. Furthermore, the true host galaxy must have a lower separation distance than the N nearest background galaxies, where N is the number of potential hosts that will be listed in the `DIAObject` parameters `potentialHost` and `potentialHostSeparation`.

To investigate the probability of host-association failure for nearby transients, we simulate a mock catalog of randomly distributed random background interloper galaxies. It is based on the same LSST-like mock galaxy catalog used by Graham et al. (2018) for studies of photometric redshifts. Here, the catalog is limited to galaxies with at least a 5σ detection in the *griz* bands at the projected 10-year depths listed above. The catalog contains redshifts and apparent *ugrizy* magnitudes, which are used to approximate intrinsic absolute r-band magnitudes. The absolute magnitudes are used to synthesize approximate galaxy radii based on the relationship between absolute magnitude and radius for late-type galaxies defined in Figure 3 of Shen et al. (2003). These synthesized radii are over-estimates because late-types are generally larger than early-types, and because this magnitude-radius relation was defined for SDSS galaxies at lower redshifts than the LSST high-z galaxies it is being applied to. This is deliberate, because it will result in upper limits on the rate of interlopers. The characteristics

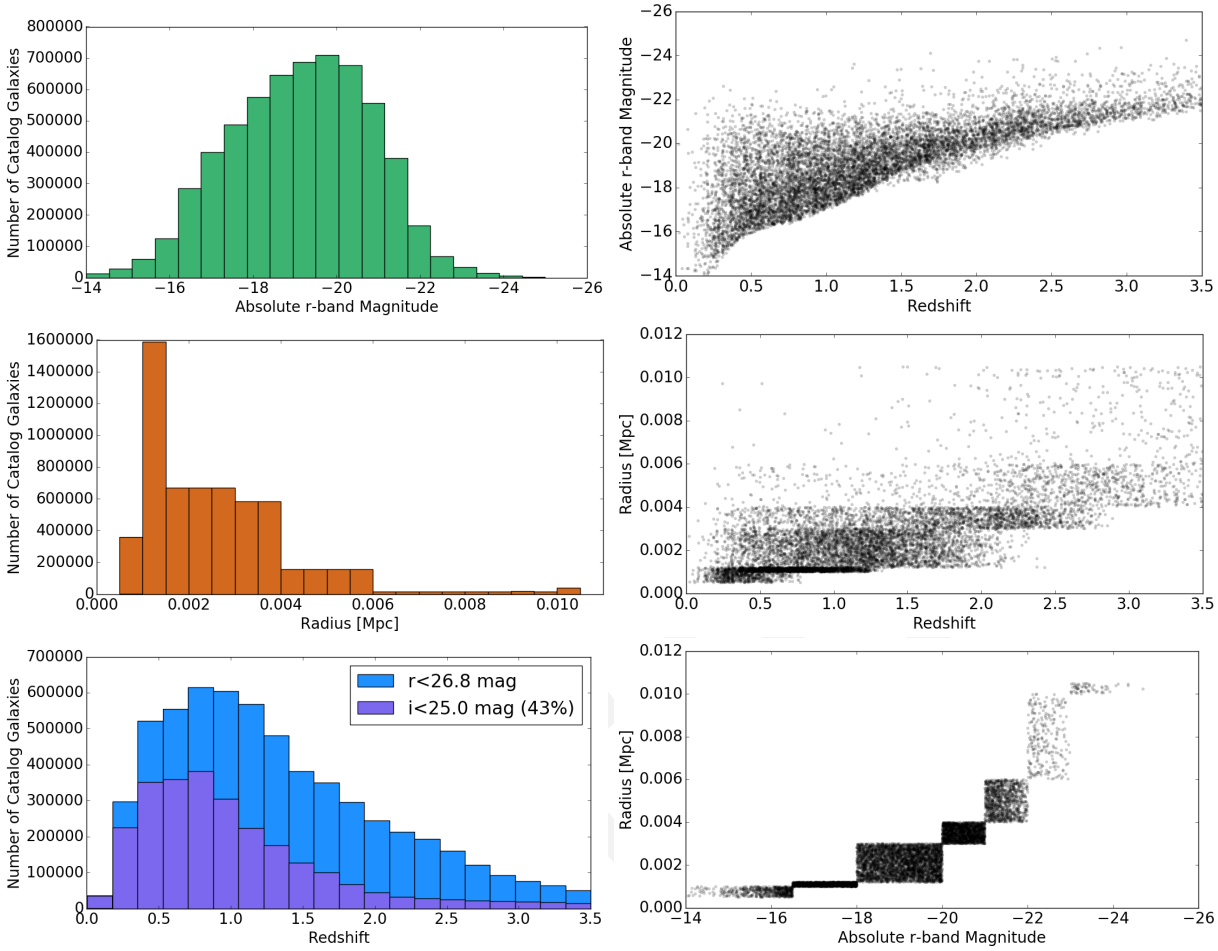

FIGURE 1: *Left:* Histograms of the synthesized absolute r-band intrinsic magnitude (top) and radius (middle) for catalog galaxies, and the simulated galaxy redshifts (bottom). *Right:* Correlation with redshift of the synthesized intrinsic magnitude (top) and radius (middle), and the approximate relation between radius and intrinsic magnitude (bottom).

of this crude galaxy catalog are illustrated in Figure 1.

Consider again the large nearby galaxy with $R_e = 10$ kpc at a redshift of $z = 0.01$, for which the sky area within $3R_e$ is $A(3R_e) = 0.0052$ deg^2 and contains ~2600 background galaxies. From the simulated catalog described above, 100 sets of background galaxies are randomly selected. For each set, the fraction of the nearby galaxy's area which is covered by the area of interloping background galaxies is calculated: $f_A = \sum A(3R_{e,\rm bkg})/0.0052$. This fraction is equivalent to the probability that a transient at $3R_e$ from this large nearby host will be within $3R_{e,\rm bkg}$ (i.e., closer to) a background interloper. The probability that this transient is closer to N interlopers than to its true host is $P_{\rm fail} = (f_A)^N$. This is the probability of a failed host

association, where “failed" means that the DIAObject record of the N galaxies with the lowest separation distances does not include the true host. For this large nearby galaxy, average values of f_A and $P_{\rm fail}$ from the 100 simulated background sets are $f_A = 0.181 \pm 0.008$ and $P_{\rm fail}(N = 3) = 0.006 \pm 0.0008$

Since this probability of failure is based on the sky density of background galaxies, it is independent of the radius and redshift of the nearby galaxy. However, it does depend on the factor applied to effective radius (i.e., the offset of transient considered) and N, the number of potential hosts recorded in the DIAObject catalog. Figure 2 shows the probability of failure as a function of R_e and N. These are **upper limits** on the probability of failure, as they are derived from upper-estimates of galaxy radii and the sky density of the final 10-year LSST galaxy catalog.

To assist in the interpretation of Figure 2, the following describes several conclusions drawn from points in this plot:

- Transients with low host offsets, $1R_e$, are closer to (within $1R_{e,\rm bkg}$ of) one background interloper $\sim 2\%$ of the time (purple point at $1R_e$). (However, given the high surface brightness of nearby galaxies, such a background galaxy may well be undetected.)
- Transients with high host offsets, $5R_e$, are closer to (within $5R_{e,\rm bkg}$ of) one background interloper $\sim 50\%$ of the time (purple point at $5R_e$), and closer to six background interlopers $> 1\%$ of the time (magenta point at $5R_e$).
- In order to achieve a 1% probability of failure for transients offset by $3R_e$ from large nearby galaxies, the DIAObject catalog record should include the $N = 3$ galaxies with the lowest separation distances (green point at $3R_e$).

Based on Figure 2, in order to reduce the probability of failure in associating nearby, large-offset transients with their true hosts:

1. The separation distances for all galaxies within at least $4R_e \approx 200$ arcsec should be calculated and considered.
2. The 3 galaxies with the lowest separation distances should be included in the DIAObject catalog record.

Adopting these recommendations would cause up to 1% of the transients at $3R_e$ from large nearby galaxies to experience a failed host association, where the true host is not listed in the `DIAObject` record. Since $3R_e$ encompasses $\sim 99\%$ of a galaxy's light, and most transient types are distributed proportional to the light, **the upper limit** on the host association failure rate for nearby transients should be $\sim 0.01\%$.

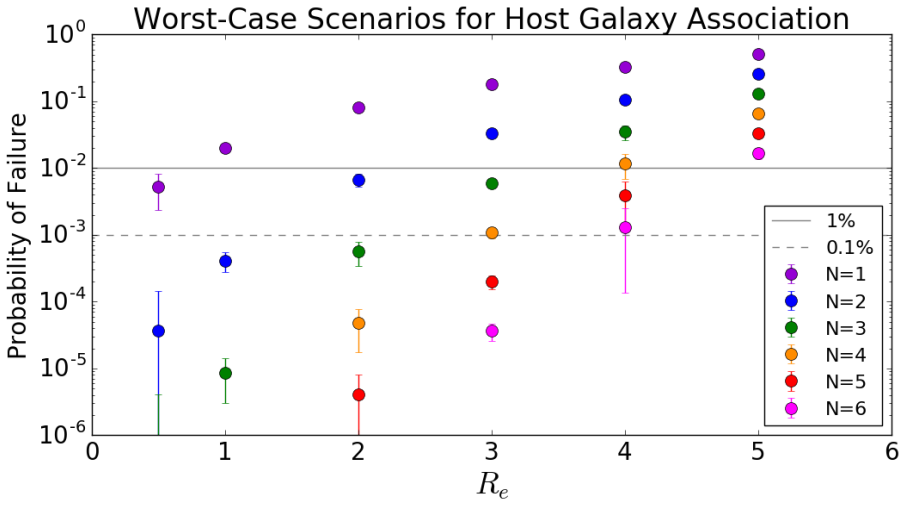

FIGURE 2: The probability that a transient will fail to be associated with a large ($R_e = 10$ kpc) nearby host galaxy due to background interlopers, as a function of the transient's offset in effective radii from galaxies in the vicinity (including the true host), where "failure" means the true host's separation distance is not in the top N nearest galaxies. **This is a "worst case scenario" because it applies to an exceptionally large nearby galaxy, and all background galaxy radii estimates are upper limits** (as described in the text). Error bars show the standard deviation from the 100 randomly-generated sets of background galaxies.

References

Foley, R.J., Mandel, K., 2013, ApJ, 778, 167 (arXiv:1309.2630), doi:10.1088/0004-637X/778/2/167, ADS Link

Graham, M.L., Sand, D.J., Bildfell, C.J., et al., 2012, ApJ, 753, 68 (arXiv:1205.0015), doi:10.1088/0004-637X/753/1/68, ADS Link

Graham, M.L., Connolly, A.J., Ivezić, Ž., et al., 2018, AJ, 155, 1 (arXiv:1706.09507), doi:10.3847/1538-3881/aa99d4, ADS Link

Gupta, R.R., Kuhlmann, S., Kovacs, E., et al., 2016, AJ, 152, 154 (arXiv:1604.06138), doi:10.3847/0004-6256/152/6/154, ADS Link

Irani, I., Schulze, S., Gal-Yam, A., et al., 2019, ApJ, 887, 127 (arXiv:1904.01425), doi:10.3847/1538-4357/ab505d, ADS Link

[LPM-17], Ivezić, Ž., The LSST Science Collaboration, 2018, *LSST Science Requirements Document*, LPM-17, URL https://ls.st/LPM-17

[LSE-163], Jurić, M., et al., 2017, *LSST Data Products Definition Document*, LSE-163, URL https://ls.st/LSE-163

Melchior, P., Moolekamp, F., Jerdee, M., et al., 2018, Astronomy and Computing, 24, 129 (arXiv:1802.10157), doi:10.1016/j.ascom.2018.07.001, ADS Link

Sako, M., Bassett, B., Becker, A.C., et al., 2018, PASP, 130, 064002 (arXiv:1401.3317), doi:10.1088/1538-3873/aab4e0, ADS Link

Sand, D.J., Graham, M.L., Bildfell, C., et al., 2011, ApJ, 729, 142 (arXiv:1011.1310), doi:10.1088/0004-637X/729/2/142, ADS Link

Shen, S., Mo, H.J., White, S.D.M., et al., 2003, MNRAS, 343, 978 (arXiv:astro-ph/0301527), doi:10.1046/j.1365-8711.2003.06740.x, ADS Link

Sullivan, M., Le Borgne, D., Pritchet, C.J., et al., 2006, ApJ, 648, 868 (arXiv:astro-ph/0605455), doi:10.1086/506137, ADS Link

Draft