The Astrophysical Journal Supplement Series, 258:2 (29pp), 2022 January https://doi.org/10.3847/1538-4365/ac3bca

Preparing to Discover the Unknown with Rubin LSST: Time Domain

Xiaolong Li[1], Fabio Ragosta[2], William I. Clarkson[3], and Federica B. Bianco[1,4,5,6]
[1] Department of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA; lixl@udel.edu
[2] INAF and University of Naples "Federico II," via Cinthia 9, I-80126 Napoli, Italy
[3] Department of Natural Sciences, University of Michigan—Dearborn, 4901 Evergreen Road, Dearborn, MI 48128, USA
[4] Joseph R. Biden, Jr., School of Public Policy and Administration, University of Delaware, Newark, DE 19717, USA
[5] Data Science Institute, University of Delaware, Newark, DE 19717, USA
[6] CUSP: Center for Urban Science and Progress, New York University, Brooklyn, NY 11201, USA
Received 2021 June 9; revised 2021 November 19; accepted 2021 November 22; published 2021 December 22

Abstract

Perhaps the most exciting promise of the Rubin Observatory Legacy Survey of Space and Time (LSST) is its capability to discover phenomena never before seen or predicted: true astrophysical novelties; but the ability of LSST to make these discoveries will depend on the survey strategy. Evaluating candidate strategies for true novelties is a challenge both practically and conceptually. Unlike traditional astrophysical tracers like supernovae or exoplanets, for anomalous objects, the template signal is by definition unknown. We approach this problem by assessing survey completeness in a phase space defined by object color and flux (and their evolution), and considering the volume explored by integrating metrics within this space with the observation depth, survey footprint, and stellar density. With these metrics, we explore recent simulations of the Rubin LSST observing strategy across the entire observed spatial footprint and in specific Local Volume regions: the Galactic Plane and Magellanic Clouds. Under our metrics, observing strategies with greater diversity of exposures and time gaps tend to be more sensitive to genuinely new transients, particularly over time-gap ranges left relatively unexplored by previous surveys. To assist the community, we have made all of the tools developed publicly available. While here we focus on transients, an extension of the scheme to include proper motions and the detection of associations or populations of interest will be communicated in Paper II of this series. This paper was written with the support of the Vera C. Rubin LSST Transients and Variable Stars and Stars, Milky Way, Local Volume Science Collaborations.

Unified Astronomy Thesaurus concepts: Transient detection (1957); Sky surveys (1464); Surveys (1671); Peculiar variable stars (1202)

1. Introduction

The Rubin Observatory Legacy Survey of Space and Time (hereafter LSST) is an ambitious project that promises to monitor the entire Southern Hemisphere sky over a continuous 10 yr interval starting in 2024. It will deliver high sensitivity, high (seeing-limited) spatial resolution, and high temporal cadence ($\geqslant$1 image per night, ~few days repeat on each field). While other surveys have stretched into one or two directions in this feature space,[7] delivering observations at high cadence over small fields of view (e.g., SNLS; Guy et al. 2010) or monitoring large fields of view but at low spatial resolution (e.g., ASAS-SN; Kochanek et al. 2017), the combination of high spatial resolution, high cadence, and high sensitivity places the Rubin LSST in a unique position to contribute to nearly all fields of astronomy with an unprecedentedly rich data set.

Perhaps the most exciting promise of Rubin LSST is thus its potential to discover as-yet unknown phenomena. This work focuses on assessing the potential of LSST to discover "true novelties": phenomena that have neither been observed, nor predicted, under different choices of observing strategy.

The Rubin LSST observing strategy is designed to accomplish several science goals within four science themes: (1) probing dark energy and dark matter; (2) taking an inventory of the solar system; (3) exploring the transient optical sky; and (4) mapping the Milky Way. These diverse goals lead to strict interlocking constraints including requirements on image quality, depth—single-visit depth and number of visits per field—filters system, and total sky coverage. The science drivers and technical requirements are described in detail in Ivezić et al. (2019, hereafter I19)[8] and summarized in Bianco et al. (2022).

While the survey strategy (and indeed facility design) is thus mostly specified by the main science goals, these constraints still allow for a significant flexibility in the details of the survey strategy. For example: while the reference design (e.g., Claver et al. 2014; Kahn et al. 2010, I19) leads to a revisit time of 3 days on average for 18,000 deg^2 of sky, with two visits per night, this still allows for a large distribution and even a significant range of median values for the internight time gaps, as seen in Figure 1 (see also Figure 2 in Bianco et al. 2022, the opening paper in this focus issue).

LSST will include several "surveys," each helping to address the four key science pillars as well as other science goals in different ways. The majority of the 10 yr will be spent on a survey designed explicitly to meet the requirements specified in Ivezić et al. (2019): the "Wide Fast Deep" survey (hereafter WFD). It is expected that this will take between 75% and 85%

[7] Table 1 in Graham et al. (2019) presents a comparison of the characteristics of LSST and several precursor synoptic surveys.

[8] For the Science Requirements Document (SRD) itself, see Ivezić & the LSST Science Collaboration (2013).

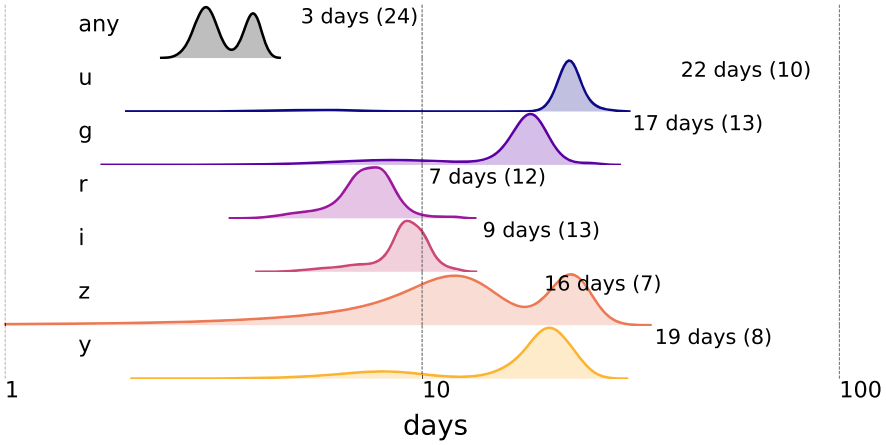


Figure 1. Distribution of median time gaps between observations in different nights for 86 simulations of the Rubin LSST Wide Fast Deep Survey `OpSim v1.5` (see Section 2.4). Here, observations are spatially grouped by `healpixel` using resolution parameter `NSIDE` = 64, with pixel area 0.84 square degrees (roughly corresponding to 55 arcminute resolution). The top row shows the distribution of gaps between observations of the same field in *any* filter, and the rows beneath show the time gaps between observations in the same filter, as labeled. The distributions are normalized by peak height, and the modal value over the 10 yr survey simulation across all `OpSims` is indicated to the right of each distribution, along with the number (in parentheses) of observations in a 1.5 day-wide bin around the peak value. The distributions are smoothed via kernel density estimation (using the `Python Seaborn` package with default settings). The breadth of the distribution (and at times its complex shape with multiple peaks) demonstrates how simulated Rubin LSST cadences, all developed in compliance with the SRD requirements that make specifications for each filter, can still be very different even in core features such as the median time gap, with ranges that can span tens of days (e.g., *z* band). For reference, the median time gap for the baseline survey `baseline_v1.5` is 22, 18, 7, 9, 11, 19 for *ugrizy*, respectively, and 3 days for images in any filter. This figure is discussed in Section 1. Also see Bianco et al. (2022).

of the time on-sky. The remaining time will be spent on "special programs." These include "mini-surveys" and "micro-surveys": programs that cover specific, extended areas of sky with a cadence designed to pursue a specific science goal (such as the detection of main-belt comets in the North Eclicptic Spur [9]); and "Deep Drilling Fields" (DDFs): single pointings that will be visited on multiple nights during the survey at an enhanced cadence designed to both reach a higher cumulative depth in the stacked images and a more dense time sampling, and of which four have been selected.[10] Additionally there is the potential to spend some of the time not spent in WFD observation on "Targets of Opportunity," for example, to follow-up of multimessenger triggers (Margutti et al. 2018).

This loose division of LSST into the different flavors of (sub)surveys implies different levels of flexibility for the observing strategies for different regions of the sky. For example, while the expected range of per-visit exposure times in the WFD region is tightly constrained (∼30 s) to achieve the goals of the four LSST science pillars and given observing efficiency constraints (I19), some mini-surveys may be better served by (or even require) different exposure times, extending to potentially much shorter and/or longer exposures than the WFD program exposure time.

Perhaps uniquely among modern surveys, the Rubin Observatory has embedded community involvement in the design of the survey (Bianco et al. 2022). To that end, it has shared its extensive simulations framework with the scientific community at a very high level of detail, including (but not limited to) detailed hardware specifications, facility operations models (including detailed observatory and instrument overheads), atmospheric transmission, and also models for astrophysical populations and interstellar dust, which together allow simulated recovery of tracer populations (Connolly et al. 2014). For most users in the scientific community, it is the metadata of the predicted observing strategies (e.g., observing time, expected seeing, instantaneous depth to 5σ photometric precision) that is most relevant to the evaluation of survey strategies: Rubin has made a large number (many hundreds, to date; see Bianco et al. 2022) of simulated LSST surveys available to the community. The Operations Simulator (Delgado & Reuter 2016) generates the metadata for a full 10 yr period of operation under specified desiderata for the run characteristics.

Led by Lynne Jones and Peter Yoachim at the University of Washington, the project has also developed a dedicated Metrics Analysis Framework (`MAF`; Jones et al. 2014),[11] and continues to work with the community in the development of the tools to extract the scientific utility of the `OpSims` for various scientific cases. Standard metrics run on all `OpSims` by the project fall under the main `sims_maf` package,[12] while community-contributed metrics are curated at the `maf-contrib` project.[13] We discuss the Operations Simulator and `MAF` in a little more detail from the point of view of the true novelties we seek, in Section 2.1. See also Bianco et al. (2022) for more details.

Recent community input on the LSST survey strategy is roughly divided into three phases. The first phase concluded with the development of the "Community Observing Strategy Evaluation Paper" (COSEP; LSST Science Collaboration et al. 2017), which attempted to distill the requirements of a wide range of science cases into specifications (and in some cases evaluations) of simple quantitative measures of scientific effectiveness that could be compared between science cases. In the second phase, the community was asked by the project to prepare cadence whitepapers to suggest alternatives to the baseline cadence; 46 whitepapers were ultimately submitted.[14]

[9] https://docushare.lsstcorp.org/docushare/dsweb/Get/Document-30596/schwamb_sso_nes.pdf
[10] https://www.lsst.org/scientists/survey-design/ddf
[11] Also available at https://www.lsst.org/content/lsst-metrics-analysis-framework-maf.
[12] https://github.com/lsst/sims_maf
[13] https://github.com/LSST-nonproject/sims_maf_contrib
[14] https://www.lsst.org/submitted-whitepaper-2018

In the current phase, the community and Rubin are working to implement the quantitative scoring for the scientific yield of the LSST survey, to allow its effectiveness to be evaluated on a timescale commensurate with the ultimate decisions by the project on the survey strategy to adopt. This paper forms part of this third phase of community input.

A word on notation is in order. We refer to a simulated 10 yr survey as an OpSim. Following the naming convention of the COSEP, we use the acronym "MAF" (or sometimes "metric") to refer to a piece of code that measures properties of an OpSim on a per-field basis. The overall evaluation of a strategy requires assessing its power over a large number of scientific goals. Therefore, in order to be useful for comparison, the MAFs must themselves be summarized into Figures of Merit (FoMs): single numbers that convey the power of a survey (as simulated) to achieve a specific science goal. These FoMs empower the designated Rubin Committee, the Survey Cadence Optimization Committee[15] (SCOC), to make recommendations about which strategy to embrace, as described in Bianco et al. (2022). An example of a MAF(*metric*) might be a characterization of the time gaps between repeat observations of each field in a particular filter pair of interest, while the associated FoM would collapse the distribution into a single number that captures the sensitivity of the strategy to the detection of transients in some range of parameter space. For more on the operational definitions of metrics and FoMs, see the COSEP.

As discussed in the introduction to the COSEP, ideally the FoMs would be measured in bits of information that the survey would contribute in excess of the previously available information on a phenomenon. While clear in principle, this information-theory inspired definition of an FoM is challenging to achieve in practice. Not all science cases easily translate into a measurement on a quantity. For cosmology, for example, one could conceivably quantify the scientific power of a survey by the decrease in the uncertainty on the scientific parameter of interest, for example H_0. However, the survey power even for identifying particular tracers becomes more ambiguous. The power of a survey in identifying progenitors of supernovae, for example, is less easily quantifiable: as additional qualifiers are placed on the phenomena to be measured (for example: sensitivity to different types of progenitors), the translation of survey sensitivity into bits of additional information becomes increasingly difficult. Following this logic, measuring the power of a survey to discover truly novel phenomena would be impossible. The assessment of the ability of a survey realization (OpSim run) to discover true novelties requires a model-free approach (otherwise we would by default limit ourselves to unobserved, but predicted, phenomena; Chandola et al. 2009).

We have set out to define metrics and FoMs that will allow for comparison of simulated LSST strategies based on their potential to discover *true novelties*, in terms of discovery parameter space that is well covered (or not!) by the simulated surveys. We focus here on stochastic transients that appear once (during the duration of the main survey) in association with a source. We expect repeated transients, including periodic and quasiperiodic signals, would benefit from the strategies that benefit the discovery of transients as well, with increase potential for discovery associated with the repetition of the signal, and refer the users to the several papers discussing the potential for LSST surveys to characterize signal periodicity (Johnson et al. 2019; Lund et al. 2018; LSST Science Collaboration et al. 2017, Chapter 5). We make these MAFs and FoMs publicly available to aid Rubin survey strategy decisions.

We remain agnostic on the accuracy with which the OpSims actually implement the desired strategies, but focus instead on the output: how well the resulting OpSims support the detection of true anomalies as quantified in our metrics and figures of merit. We point out that we are not in this paper attempting to weight science cases against each other, which is a task that takes place at a higher level in the observing strategy determination (Bianco et al. 2022). Nevertheless, the detection of anomalies (whether predicted or otherwise) does underlie a number of science cases (such as the detection of supernovae characterized by a particular light-curve template), and thus improved sensitivity to true anomalies will also improve sensitivity to a variety of science cases. We therefore expect that OpSims that are advantageous for the detection of true anomalies will also be advantageous for many other science cases, but that comparison is beyond the scope of the present work.

The OpSims are continually under development based on input from the project and the scientific community, and thus the suite of available simulations is continuously evolving. Improvements made between releases include general strategy updates (such as changes to the recommended exposure time per visit), improvements in the implementation of engineering constraints (such as the time required for filter changes), and improvements in the implementation of the observing strategies themselves (such as the way in which special cadences are implemented). Bianco et al. (2022) provides more detail.[16]

We selected the OpSim v1.5 family of simulations (a major release from 2020 May with 86 simulated strategies)[17] to develop and demonstrate the metrics and figures of merit, as it contains sufficient variety among the simulations to elucidate the various requirements for detecting true anomalies.

As the simulations evolve, application of the figures of merit to more recent releases is then straightforward. As an example, we present the evaluation of our figures of merit to the OpSim v1.7 (2021 January) and OpSim v1.7.1 (2021 April) releases, which implement an updated exposure time per visit (2×15 s instead of the 1×30 s used in OpSim v1.5).

This publication is one of a pair: in this communication (Paper I), we focus on detecting individual objects of interest in a multidimensional feature space that includes time coverage, filter coverage, star density, and total footprint on the sky. Inclusion of constraints from *proper motion*, which is rather more involved and also lends itself naturally to detection of previously unknown *populations* and structures, is deferred to F. Ragosta et al., in preparation (Paper II). The present paper therefore does NOT directly address proper-motion anomalies: the reader is referred to Paper II for those issues.

This paper is organized as follows: Section 2 summarizes the simulations and methods, and describes the feature space we use. Sections 3 through 5 then communicate the metrics and figures of merit we have developed, and present the evaluations

[15] https://www.lsst.org/content/charge-survey-cadence-optimization-committee-scoc

[16] The release details are also announced on the LSST Community web forum, e.g., https://community.lsst.org/t/survey-simulations-v1-7-1-release-april-2021/4865.

[17] https://community.lsst.org/t/fbs-1-5-release-may-update-bonus-fbs-1-5-release/4139.

of the figures of merit over the WFD main survey area, on a wide range of simulated observing strategies. Here we consider the following metrics: color and time evolution (Section 3), integrated depth (Section 4), and spatial footprint (Section 5). Section 6 then applies the set of the figures of merit to the OpSims chosen, first to the WFD region (Section 6.1), then to the mini-surveys (Section 6.2). The application of the figures of merit to the more recent OpSim v1.7 set of simulations is presented in Section 6.3. In Section 7 we conclude with some recommendations on the usage and interpretation of the metrics and figures of merit we have developed. In Appendices A and B and , we present a discussion of the impact of choices we made throughout the construction of our figure of merit and a description of the interactive tools we have developed to facilitate exploration of the multidimensional feature space.

2. Methodology

Here we summarize the simulations and methods used. In Section 2.1 we summarize the Operations Simulator and Metric Analysis Framework, both provided by Rubin Observatory, in the context of our work. In Section 2.2 we briefly discuss the tools by which we accomplish spatial selection, to isolate regions such as the Galactic Plane and Magellanic Clouds. The usage of *feature space* to identify discovery space for true novelties is introduced in Section 2.3, and the output metadata produced by OpSim in this context is summarized in Section 2.4.

2.1. MAF and OpSim

It is beyond the scope of this paper to describe in detail the software that Rubin Observatory has made available to the community to contribute to the survey design. The reader is referred to the opening paper of this Focus Issue for more details including a historical account of the software genesis (Bianco et al. 2022), and to the references therein for a detailed discussion of the workings of the software, in particular Naghib et al. (2019), Yoachim et al. (2016), Delgado & Reuter (2016), and Delgado et al. (2014) for the Operation-Simulator and Jones et al. (2014) for the Metric Analysis Framework. Here, we briefly summarize the core functionality.

The Operations Simulator software[18] OpSim allows the generation of a simulated strategy based on a series of strategy requirements: for example, total number of images per field per filter, including simulated weather, telescope downtimes, etc. The input of an OpSim run is the survey requirements (survey strategy) and the output is a database of observations with associated characteristics (e.g., image 5σ depth), which specify a sequence of simulated observations for the 10 yr survey. The Rubin OpSim went through several versions since its initial creation (Delgado et al. 2014) that primarily differ in the algorithms with which sequences of pointings and filters are optimized to achieve the prescribed survey characteristics (Bianco et al. 2022).

The Metric Analysis Framework (MAF[19]) application programming interface (API) is a software package created by Rubin Observatory (Jones et al. 2014) to facilitate the evaluation of simulated LSSTs to achieve specific science goals as measured by the strategy's ability to obtain observations with specified characteristics. The MAF interacts with databases. The MAF has been made public upon its creation to facilitate community input in the strategy design. The MAF enables selections of observations within an OpSim primarily by an SQL constraint, which allows the user to select, for example, filters or time ranges (e.g., the first year of the survey). Further, the choice of slicers allows the user to group observations. For example, one may "slice" the survey by equal-area spatial regions, using the HEALPIX scheme of Górski et al. (2005). Throughout, we choose a HealpixelSlicer with resolution parameter NSIDE = 16, corresponding to a pixel area of 13.4 square degrees (and thus the choice that most closely matches the size of the Rubin LSST field of view).

2.2. Spatial Selection

In practice, OpSim generates the synthetic observations using "proposals" for various assumed programs, including WFD, DDFs, and "special" programs such as the Galactic midplane, Magellanic Clouds, and the North Ecliptic Spur, with the "proposalID" parameters preserved in the OpSim output. When evaluating our figures of merit for the WFD region, we use this "proposalID" to select the relevant observations. We start by evaluating all of our metrics on the WFD by selecting proposalID=1.

Later on, when evaluating our FoMs on specific regions (Section 6.2), we select the simulated observations spatially (considering all observations taken within a specific region, regardless of their proposalID provenance), as the science that can be extracted from observations of a particular spatial region depends only on what was observed, and not on the proposalID with which each observation was originally identified. For example, some regions of moderate stellar density may be covered *both* by observations associated with the WFD survey and with the Galactic Plane mini-survey. Selecting by one or another proposalID would miss one or another of the set of observations of these regions. This is particularly relevant when considering simulated strategies that extend the WFD region to encompass regions that would be classified as "mini-surveys" in the other simulated strategies. Olsen et al. (2018a), for example, discussed some possible strategies that would do this. This leads us to define the Magellanic Cloud "mini-survey" region as the spatial area occupied by the Magellanic Clouds, and to run our metric on all images in that spatial region within an OpSim. This is somewhat different from usual practice in other Rubin survey-strategy-related work that focuses on evaluating the efficiency of a specific proposed mini-survey (i.e., a specific proposalID). The spatial selector we have developed is quite flexible—regions can be specified programmatically or by hand—and we have made it publicly available.[20]

2.3. Feature Space

Anomaly detection is an important field of research with deep methodological ramifications (Chandola et al. 2009; Martínez-Galarza et al. 2020). Notable advances in the field have been achieved in recent years across disciplines: from threat detection in defense and security (e.g., Sultani et al. 2018), to astrophysics (Baron & Poznanski 2017; Pruzhinskaya

[18] https://www.lsst.org/scientists/simulations/opsim
[19] https://www.lsst.org/scientists/simulations/maf
[20] https://github.com/xiaolng/healpixSelector

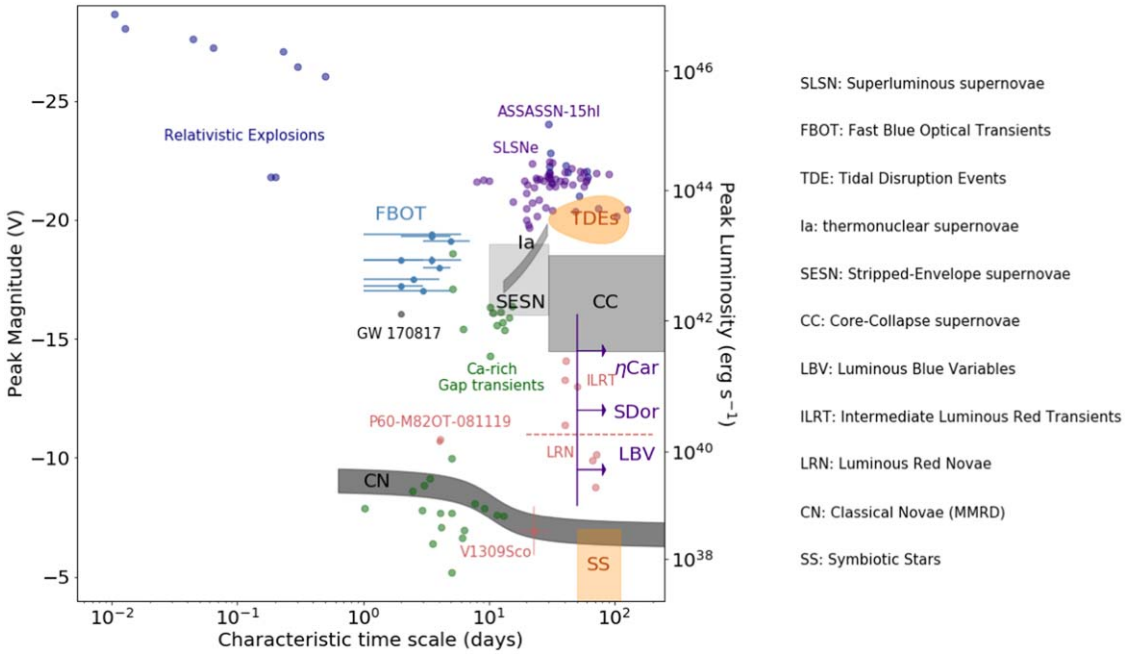


Figure 2. The phase space of transients, reproduced with permission from Ivezić et al. (2019) with minor modifications: the intrinsic brightness is plotted against the characteristic timescale of evolution. Shaded areas indicate the region of this phase space occupied by various classes of objects, with individual objects indicated for some of the less-populated classes. The notable gap at all intrinsic magnitudes fainter than −20 is likely due, at least in part, to an observational bias, as surveys are typically not able to probe large volumes of the universe at short timescales down to faint apparent magnitudes. Data for superluminous supernovae (SLSNe) and fast blue optical transients (FBOTs) that were not included in the original plot are collected from Inserra (2019) and Drout et al. (2014), respectively. See Section 2.3.

et al. 2019; Martínez-Galarza et al. 2020; Aleo et al. 2020; Soraisam et al. 2020; Doorenbos et al. 2021; Ishida et al. 2021; Lochner & Bassett 2021; Storey-Fisher et al. 2021; Vafaei Sadr et al. 2021) with the discovery of rare and possibly unique astrophysical phenomena (Lintott et al. 2009; Micheli et al. 2018; Boyajian et al. 2018; we note that two of these "true novelties" were detected through crowd-sourced data analysis).

The literature has not yet converged on a strict definition of the meaning of "anomalies" or "anomaly detection"—which tends to vary depending on the application, as indicated by the works cited above—so we adopt the following loose definition for the present paper: anomaly detection is the detection of an event (e.g., rapid brightness variation; variation in color) and/or astrophysical characteristic (e.g., extreme temperature; unusual motion through the sky) that is in some quantifiable way "unusual" as compared to the typical object or predicted population. We strive to remain as agnostic as possible throughout this work about the nature of these anomalies in order to avoid as much as possible a bias against any regions of the feature space. We therefore do not place limits on exactly how an "unusual" object would be quantified, as different users of the LSST data set will have different scientific goals in mind. Furthermore, we focus on the discovery of phenomena that are not only unusual, but not predicted from theory before their detection, and we refer to these anomalies as to "true novelties."

Anomaly detection is generally approached either through unsupervised or supervised learning techniques (e.g., Bishop 2006; Yang et al. 2006; Hastie et al. 2009). In unsupervised learning, or *clustering*, a similarity metric is defined in the available feature space enabling the grouping of similar objects together, as well as the identification of objects that do not belong to any existing group (the anomalous objects). The supervised approach identifies groups in a latent lower-dimensional space based on experts' classifications in the original feature space, which, in the case of supervised learning applied to anomaly detection, is typically a binary classification of "normal" versus anomalous.

We expect the end-user of LSST data will employ some combination of supervised and unsupervised anomaly detection, utilizing some specific training set if they have a particular phenomenon in mind, or more generally, some combination of feature space. But, by definition, we do not know what templates or feature space combinations will be used to discover true novelties. We therefore attack the problem more generally, by considering the volume of feature space covered by the candidate strategies. Gaps in the observing strategy (e.g., gaps in spatial coverage; gaps in timescales to which the survey is sensitive; gaps in color coverage) impact the discovery of true novelties by both increasing the risk that an anomaly would go undetected, if it falls in a gap, and making its anomalous nature harder to assess. In this series of papers, we focus on survey design to maximize the throughput of algorithms for anomaly detection, regardless of the nature of the algorithmic approach.

As measured by imaging surveys, astronomical objects are characterized by brightness, brightness ratio in different portions of the energy spectrum (color), position, shape, and the rate and direction of change in any of those features. The collection of properties defines a multidimensional phase space, with different categories of phenomena lying in different regions of this phase space (see Figure 2). Accordingly, we identified the following features that can be measured in the Rubin Observatory data:

1. Color
2. Time evolution
3. Motion
4. Morphology

5. Association.

We set morphology aside, as largely, the power of the survey to measure morphological anomalies does not depend on the survey strategy, but rather on the image system design (e.g., resolution and depth). We assume that the discovery of anomalous associations of objects depends on our accuracy in measuring the properties of each object; one easy example might be the discovery of faint, spatially extended stellar populations that have heretofore gone undetected. To measure dynamical anomalies in a completely model-independent way proves to be more involved, because it requires comparison of measured proper motions to those of established Galactic dynamical parameters. *Motion* is thus deferred to Paper II, where we also develop an FoM for the detection of previously unknown *stellar populations.*

Having identified features that can be extracted from the Rubin Observatory LSST data such as color information or light-curve evolution, we measure the completeness of the survey in a hypercube in the feature space as a model-independent measure of the power to detect *novel transients* or novel modes of variability, defining transients as objects whose observational *and physical* properties are changed by some event, usually as the result of some kind of eruption, explosion, or collision, whereas variables are objects whose nature is not altered significantly by the event (e.g., flaring stars). Furthermore, some objects vary not because they are intrinsically variable, but because some aspect of their viewing geometry causes them to vary (e.g., eclipsing binaries).

One further parameter that influences our ability to detect anomalies is the sky footprint. Trivially, a larger sky footprint will lead to a higher event rate for anomalies. If one wants to maximize the chance of detecting anomalies whose sources are approximately isotropically distributed on the sky (i.e., anomalies associated with extragalactic populations and the Galactic halo), then a larger footprint would be favorable. For Galactic anomalies, for which the source distribution is presumably highly non-isotropic, the probability of detection will still scale weakly with the total area covered, but the scaling will be dominated by the density of objects in the sky. And both will scale with the depth over which the footprint is observed.

Ultimately, we define a set of metrics that can simply be summed to generate an FoM for *true novelties*:

$$\mathrm{FoM} = \sum_{i=c,s,d,A_{\mathrm{sky}},D_{\mathrm{Star}}} w_i \, \mathrm{FoM}_i \qquad (1)$$

where c, s, d, A_{sky}, and D_{Star}, represent the color, light-curve shape, magnitude depth, footprint, and star density, respectively, and w are weights that can be assigned to favor the discovery of, for example, transients over nonevolving objects, or Galactic over extragalactic transients.

The weights w_i allow the investigator to imprint their own judgment on the relative scientific importance of the different metrics. Because we wish to remain as phenomenon-agnostic as possible, we refrain from assigning weights. Instead, we normalize each `MAF` to the best of our ability in a 0–1 range where 1 is optimal, so as to provide a "neutral" comparison of the existing LSST simulations.

2.4. `OpSim` Data

We base our results primarily on `OpSim v1.5`, a recent `OpSim` run that contains 86 databases in 20 families as listed in Table 1.

A more detailed description and discussion of the simulations can be found in Bianco et al. (2022) and on the Community LSST discussion forum.[21] Here we simply note that `baseline` refers to the straightforward implementation of the requirements in Ivezić & the LSST Science Collaboration (2013); the acronyms that were mentioned in this work to refer to different surveys within LSST, such as WFD and DDFs, that are mirrored in the names of the families of `OpSim`s. We note that `rolling` refers to "rolling cadence," a WFD strategy implementation where fields are not observed homogeneously in time over the survey lifetime, but rather some fields are observed more frequently early in the survey and, to different degrees, abandoned later on, to focus on other fields. These strategies provide a denser cadence on each field for some fraction of the survey time, while preserving the overall cumulative depth requirements, and are generally beneficial to the study of rapid-timescale transients (including supernovae). The `footprint` family of observations modifies the survey footprint according to different recommendations.[22] `Pair strategy` is a family of `OpSim`s that explores different approaches to pairing in time filters and observations. The `filterdist` varies the filter distribution across WFD, and `third` adds a third observation at the end of the night. The `goodseeing` family explores different requirements on weather to enable observations. The remaining `OpSim` families explore exposure time (e.g., `short` or `var_expt`), specific observing times (exclusively or in combination with the regular surveys) such as `twilight`, explicit observing phenomena that can be enhanced by cadence choices such as `dcr`, differential chromatic diffraction, or synergy with other surveys, such as Euclid. Lastly, `u60` has longer exposures in the u band (60 s compared to the standard 30 s) and some `OpSim`s explore modifications of the single exposure time by implementing a single 30 s observation instead of 2 x 15 "snaps" (that get combined into a single image to produce the standard Rubin data products; see also Section 6.3 for an assessment of the impact of this choice across `OpSim`s).

3. Color and Time Evolution

Astrophysical transients and variable phenomena have captured humanity's curiosity through the history of science. Modern astrophysics and particularly the use of digital equipment in the last half century enabled extremely fast-paced advances in this field. Figure 2, reproduced and updated from Figure 27 in Ivezić et al. (2019), shows the phase space of known astrophysical transients: transients and variable phenomena occupy different regions of this phase space of intrinsic brightness versus characteristic timescales. At the beginning of the 20th century, essentially only supernovae were known to exist, and the phase space has populated rapidly with many different classes of transients since. It is worth noting the gap for timescales shorter than ~ 1 day: while it is possible that this region is scarcely populated *intrinsically*, it is also true that an observational bias impairs discovery in this region. To be

[21] https://community.lsst.org/t/fbs-1-5-release-may-update-bonus-fbs-1-5-release/4139, released in 2020 May.
[22] https://www.lsst.org/call-whitepaper-2018

Table 1
OpSim v1.5 Databases

Family	Name
agn	agnddf
alt	alt_dust
	alt_roll_mod2_dust_sdf_0.20
baseline	baseline_2snaps
	baseline_samefilt
	baseline
bulges	bulges_bs
	bulges_bulge_wfd
	bulges_cadence_bs
	bulges_cadence_bulge_wfd
	bulges_cadence_i_heavy
	bulges_i_heavy
daily	daily_ddf
dcr	dcr_nham1_ug
	dcr_nham1_ugr
	dcr_nham1_ugri
	dcr_nham2_ug
	dcr_nham2_ugr
	dcr_nham2_ugri
descddf	descddf
filterdist	filterdist_indx1
	filterdist_indx2
	filterdist_indx3
	filterdist_indx4
	filterdist_indx5
	filterdist_indx6
	filterdist_indx7
	filterdist_indx8
footprint	footprint_add_mag_clouds
	footprint_big_sky_dust
	footprint_big_sky_nouiy
	footprint_big_sky
	footprint_big_wfd
	footprint_bluer_footprint
	footprint_gp_smooth
	footprint_newA
	footprint_newB
	footprint_no_gp_north
	footprint_standard_goals
	footprint_stuck_rolling
goodseeing	goodseeing_gi
	goodseeing_gri
	goodseeing_griz
	goodseeing_gz
	goodseeing_i
greedy	greedy_footprint
roll	roll_mod2_dust_sdf_0.20
rolling	rolling_mod2_sdf_0.10
	rolling_mod2_sdf_0.20
	rolling_mod3_sdf_0.10
	rolling_mod3_sdf_0.20
	rolling_mod6_sdf_0.10
	rolling_mod6_sdf_0.20
short	short_exp_2ns_1expt
	short_exp_2ns_5expt
	short_exp_5ns_1expt
	short_exp_5ns_5expt
spider	spiders
third	third_obs_pt120
	third_obs_pt15
	third_obs_pt30
	third_obs_pt45
	third_obs_pt60
	third_obs_pt90
twilight neo	twilight_neo_mod1
	twilight_neo_mod2
	twilight_neo_mod3
	twilight_neo_mod4
u60	u60
var	var_expt
wfd	wfd_depth_scale0.65_noddf
	wfd_depth_scale0.65
	wfd_depth_scale0.70_noddf
	wfd_depth_scale0.70
	wfd_depth_scale0.75_noddf
	wfd_depth_scale0.75
	wfd_depth_scale0.80_noddf
	wfd_depth_scale0.80
	wfd_depth_scale0.85_noddf
	wfd_depth_scale0.85
	wfd_depth_scale0.90_noddf
	wfd_depth_scale0.90
	wfd_depth_scale0.95_noddf
	wfd_depth_scale0.95
	wfd_depth_scale0.99_noddf
	wfd_depth_scale0.99

Note. The description of these OpSims can be found in the release notes of OpSim v1.5.

effective in discovery and characterization at these timescales, surveys need to reach high depth and high cadence simultaneously, while also surveying a large volume if phenomena in this region of the phase space are truly rare.

Due to their diversity in timescales, color, and evolution, the study of transients, and particularly studies that aspire to discover new transient phenomena, requires dense space *and* time coverage. The LSST has both high photometric sensitivity and a large footprint, enabling the surveying of a large volume of universe. This offers tremendous opportunities to study the variable sky. LSST's capability to discover novel transients then largely depends on its observational cadence.

Different transients will benefit from different observational strategies because of the different phenomenological expression of their intrinsic physics. To make sure the observational strategies under design maximize our chances to discover *any* novel transient, we created the *filterTGapsMetric*. This MAF evaluates the ability of LSST's observational strategies to capture information about color and its time evolution at multiple timescales. We know some timescales remain unexplored in the present collection of LSST simulations, as

discussed, for example, in the work of Bellm (2021) and Bianco et al. (2019).

3.1. The filterTGapsMetric

Rubin LSST will image the sky in six filter bands: *u, g, r, i, z, and y*. The filterTGapsMetric measures all time gaps between two filters in an OpSim, i.e., *ug, gr, ri*, and so on. The filterTGapsMetric FoM evaluates the coverage of time gaps for each filter pair.

On a field-by-field basis, for each filter pair, the metric and FoM are evaluated as follows:

1. Select the survey (e.g., WFD in this paper) and the observation time range using SQL constraint and slice the sky with HealpixelSlicer (see Section 2.1);
2. Fetch observation times for each healpixel for all visits in either of the two filters;
3. Compute all possible time gaps that can be constructed from pairs of visits.

Figure 3 shows the distribution of time gaps for all filters pairs for the baseline v1.5.

Armed with field-by-field time-gap distributions, the FoM for the entire candidate survey strategy is then computed by measuring how well the distribution of time gaps matches an ideal distribution. We use the Kullback–Leibler (KL) divergence (or relative entropy; Kullback & Leibler 1951) to measure the discrepancy between the ideal and observed distribution. The KL divergence provides an information-criteria based measure of the difference between two distributions: the KL divergence from Q to P is defined as $D_{\mathrm{KL}}(P||Q) = \sum P \log\left(\frac{P}{Q}\right)$. The KL divergence is not a distance (in the sense that it does not satisfy the triangle inequality), it is in general not symmetric (under exchange of Q and P), and it is not normalized. To derive a normalized quantity from D_{KL}, we use $e^{-D_{\mathrm{KL}}}$, where two identical distributions, with $D_{\mathrm{KL}} = 0$ would contribute 1 to the sum, while $D_{\mathrm{KL}} > 0$ would contribute < 1. Thus a larger FoM would indicate a lower discrepancy from the "ideal" distribution and thus a scientifically preferable simulation. This FoM is naturally normalized between 0 and 1 for each field.

All that remains is to choose the "ideal" distribution of time gaps against which candidate strategies will be compared. We choose different "ideal" distributions depending on whether color evolution or the light-curve shape is being probed (Figure 4 shows an example). Bianco et al. (2019) have shown that color can be measured reliably even for rapid explosive transients, for time gaps as long as 1.5 hr. Of course this is not necessarily true for novel phenomena, but we will use this as a fiducial time interval. Yet, as Figure 2 highlights, few known phenomena are known on sub-hour timescales (examples include stellar flare, blazars, and few others) and we believe that these timescales may harbor unknown phenomena just because they are less explored than hour-, day-, and month-long timescales. Furthermore, the shorter the time gap between observations, the more reliable the color measurement is. Thus, to favor observations at short timescales, we choose for the ideal distribution a uniform distribution in $\log_{10}(\Delta t)$ that extends down to the minimum possible repeat time of a few seconds set by the shutter and readout electronics, and on the other end extends to the nominal 1.5 hr required to measure the color of known rapidly evolving transients. In Appendix A we investigate our level of sensitivity to the particular choice of distribution for the *filterTGapsMetric*. To probe light-curve shapes via pairs of observations in the same filter, we want to measure evolution at all timescales. For observation pairs in the same filter, then, we adopt a uniform distribution in $\log_{10}(\Delta t)$ for the entire 10 yr survey.

The steps of the calculation of the FoM for time gaps, are thus:

1. Compute the discrepancy measure $e^{-D_{\mathrm{KL}}}$ between the distribution of time gaps and an "ideal" distribution, for each filter pair. An example comparison of the "ideal" and observed distribution is shown in Figure 3.
2. Sum the discrepancy measures over the filter pairs, weighted by the number of visit pairs over the whole sky in each filter pair N_k, and optionally by a "scientific" weight-factor w that allows certain filter pairs and/or spatial fields to be (de-)emphasized.

In practice, we adopt $w = 1$ throughout to avoid pre-judging the outcome, but investigators with particular preferences for subclasses of anomaly could express this preference via the weights w. In general, some filters and filter combinations may well be more useful than others to discover anomalies. Trivially, the value of w_k could be set by the limiting magnitude for the shallowest filter in a filter pair.

This weighted sum, over the filter pairs and over the positions in the sky, is the FoM for the OpSim of interest. The process is summarized in the relation:

$$\mathrm{FoM}_{\mathrm{tGaps}} = \sum_{i}^{\mathrm{fields}} \sum_{k=ug,gr,\ldots}^{\mathrm{pairs}} w_{k,i} N_{k,i} e^{-D_{\mathrm{KL},\,k,\,i}}, \quad (2)$$

where $0 \leqslant w_{k,i} \leqslant 1.0$, N_k stands for the number of visits in the OpSim for each of the filter pairs, and the index i runs through the healpixels (Section 2.1).

In practice, we use a simplified version of the above relationship where the metrics are summed over the sky for each filter pair before computing the KL divergence, since we will embed preferences in the pointing with subsequent components of the FoM (see Section 5):

$$\mathrm{FoM}_{\mathrm{tGaps}} = \sum_{k=ug,gr,\ldots}^{\mathrm{pairs}} w_k N_k e^{-D_{\mathrm{KL},k}}. \quad (3)$$

3.2. Results

Figure 5 shows the $\mathrm{FoM}_{\mathrm{tGaps}}$ calculated in Equation (3) for all OpSim runs in OpSim v1.5. Because the "ideal" comparison distribution is different for the color (different-filter) and light-curve shape (same-filter) pairs, Equation (3) is evaluated twice for each OpSim: once over the 15 different-filter pairs (for color) and once over the six same-filter pairs, with the results presented separately.

The light-curve shape $\mathrm{FoM}_{\mathrm{tGaps}}$ (Figure 5(a)) shows the short and twilight families of OpSims rising among the top performers. This can be explained by the fact that these OpSims contain short exposures that fill in the distributions at short timescale. After a significant performance gap, we then see filterdist, rolling, and dcr examples as the next best options. While seeing, for example, rolling strategies near the top is not surprising, as they naturally provide a log-like coverage that supports the discovery and study of

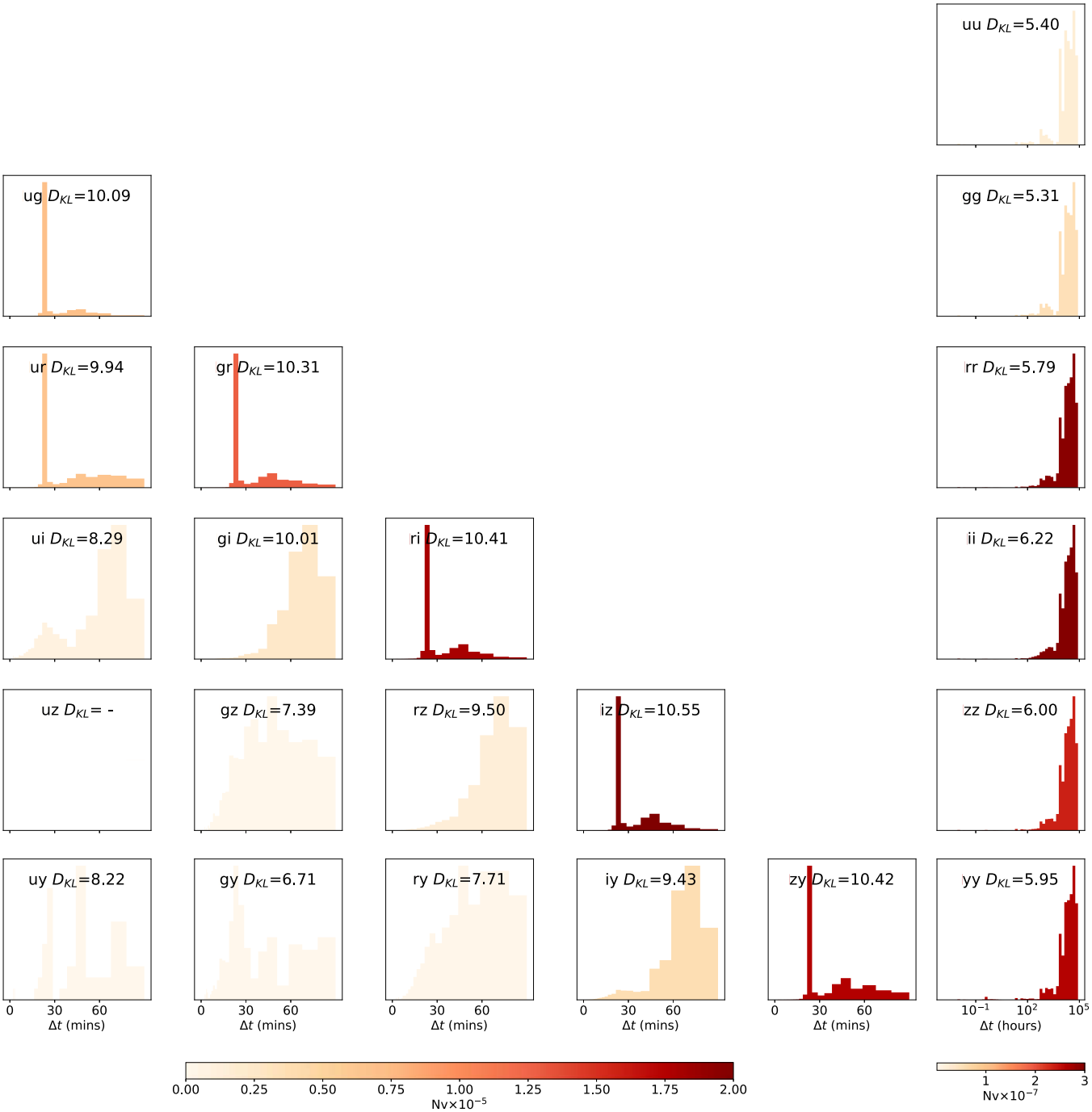


Figure 3. The distribution of all time gaps for the `baseline_v1.5 OpSim`. The triangle of plots on the left shows all time gaps between different filters (which enable the measurement of color) within 1.5 hr. The column of plots on the right shows the distribution of time gaps in the same filter for the 10 yr survey, which enables the measurement of brightness changes. The filters are indicated in each quadrant: from *u* to *y* moving from top to bottom and left to right. All histograms are normalized, but the intensity of the color is proportional to the total number of observations in that filter pair, as indicated by the color bar. In each quadrant, the value of D_{KL} is reported (see Section 3.1). We note that the majority of observations are taken with adjacent filters, which gives a narrow leverage on the spectral energy distribution, and less power to measure color. Color is in fact better measured with filters that are more separated in wavelength, for example *g*-*i* or *r*-*z*, as described in Bianco et al. (2019).

transients at multiple scales, the same filter $\mathrm{FoM}_{\mathrm{tGaps}}$ score is similar ($\mathrm{FoM}_{\mathrm{tGaps}} < 0.2$) for all `OpSims` after the `short` and `twilight` families.

Two families of `OpSims` rise to the top of the list when ranked by $\mathrm{FoM}_{\mathrm{tGaps}}$ for the color diagnostics: `short` and `rolling` (Figure 5(b)). Somewhat surprisingly, `twilight` is in the top quartile, but not near the top. This can be explained if the short exposures are primarily taken with a single filter at twilight. Unsurprisingly, `baseline_samefilt`, designed for minimal filter changes, ranks last.

4. Depth Metrics

Because our time-gap metrics are essentially based on the number of images that meet some criteria in an `OpSim`, it is important to ensure that the images that are counted are also meeting some quality standards. In particular, we need to include

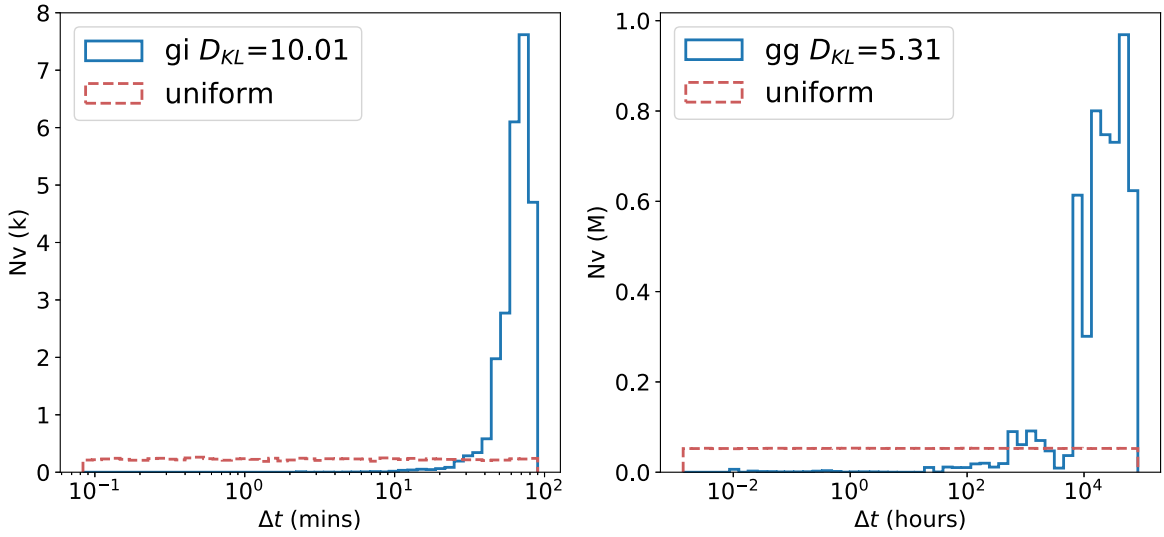


Figure 4. Time gaps in $g-i$ (left) and $g-g$ (right) from `baseline_v1.5` compared to the "ideal" distribution, plotted in red. Shown is a uniform distribution in log space up to 1.5 hr for the colors (left, with the y-axis in units of 1000 observations) and up to 10 yr for the light-curve shape (right, with the y-axis in units of 1 million observations). See Section 3.1.

information about the image depth (i.e., limiting magnitude), so that we compare the discovery potential within the same volume of the universe. Some `OpSim`s augment the WFD survey with short exposures (see Section 2.4). In fact, we noted in the $\mathrm{FoM}_{\mathrm{tGaps}}$ analysis (Section 3) that `OpSim`s that include short exposures rise to the top of the ranked list of `OpSim`s: while these `OpSim`s meet the nominal criteria and provide valuable image pairs at short time gaps, if shallow, they may fail to extend the survey *volume* to unexplored regions, which is the most important contribution LSST will make in the anomaly discovery space. To account for this, we add a metric component that measures the depth of the images collected by an `OpSim`.

We can start by comparing the depth distribution of the `OpSim`s for each filter with the apparent magnitude limits specified in the Science Requirements Document (Table 6 in Ivezić & the LSST Science Collaboration 2013). In practice, the main contributor to the difference in the distribution of depths between the `OpSim`s seems to be the time allocated to short exposures. Short exposures are typically designed for specific purposes, such as the detection of near Earth objects (e.g., the `twilight_neo` family) or decreasing the saturation limit so as to enable calibrations with shallower surveys (Gizis 2019). We want to reward surveys that include short exposures, as they add timescales, and this is what the $\mathrm{FoM}_{\mathrm{tGaps}}$ does, but we want to penalize surveys where these short exposures come at a cost of deeper images.

Figure 6 compares the per-image 5σ magnitude limit distribution for two `OpSim`s with short-exposures (`twilight_neo_mod1` and `short_exp_2ns_1expt`) with the baseline survey (blue filled histogram). The `OpSim`s including short exposures show a bimodal distribution of limiting magnitudes. The short exposures contribute to a cluster that peaks at magnitude brighter than 21 in any band ($u=20.45$, $g=20.95$, $r=20.95$, $i=20.95$, $z=20.75$, and $y=19.95$ for `short_2ns` and $r=20.95$, $i=20.85$, $z=20.25$, and $y=20.95$ for `twilight_neo_mod1`). However, while for `short_exp_2ns_1expt` the distribution of faint (fainter than magnitude $\sim$21.5) images is not substantially different from that of the baseline survey, `twilight_neo_mod1` has fewer faint images in the r, i, and z bands, and more in the y band.

We can calculate the per-image relative depth of an `OpSim`. If we calculate it as the difference between the median of the distribution and the survey specification in the Science Requirements Document (Ivezić & the LSST Science Collaboration 2013), we have a single easily interpretable number that tells us if an `OpSim` *over*-performs or *under*-performs the survey requirements (and therefore the baseline `OpSim`):

$$\mathtt{OpSim}_{\mathrm{depth}} = \sum_{k=u,g,r,i,z,y}^{\mathrm{filters}} (m_{\mathrm{median},\,k} - m_{\mathrm{goal},\,k}), \qquad (4)$$

where the sum extends to the six filters. To generate the $\mathrm{FoM}_{\mathrm{depth}}$ from this starting point, the range of the FoM for each filter is shifted and scaled to $[0, \frac{1}{6}]$ such that for each filter k, the shallowest `OpSim` has $\mathrm{FoM}_{\mathrm{depth},k}=0$ and the deepest $\mathrm{FoM}_{\mathrm{depth},k}=\frac{1}{6}$, and $\mathrm{FoM}_{\mathrm{depth}} \leqslant 1$. This has the effect of treating the contributions from each filter equally.

This leads to the ranking of the `OpSim`s shown in Figure 7. It is important to note that, aside from the `u60`, `short`, and `footprint` families, all other `OpSim`s have a similar score, between $\sim$0.8 and $\sim$0.9, so essentially there is little difference between these `OpSim`s. Nonetheless, the `short_exp` family of images ranks low, balancing for the high rank conferred in the earlier FoM by the higher number of images within short time gaps. The `u60`, which produces 60 s u-band exposures instead of the standard 30 s, ranks near the top. We also note that several `OpSim`s and families of `OpSim`s outperform the `baseline`, including the `filterdist` family and the `wfd` family. However, specifically in the case of the `wfd`, whether the `OpSim` is designed to include DDFs or not is what determines the performance decrease, even if the DDFs themselves are not included in our metric calculation (for a given scale parameter, `wfd` simulations are lower by as much as 20 ranks from their `_noddf` counterpart). We explain this as follows: the inclusion of DDFs puts additional constraints on the survey that compete with, for example, airmass and weather constraints leading to a slight decrease in image quality. We also note that the `rolling` and

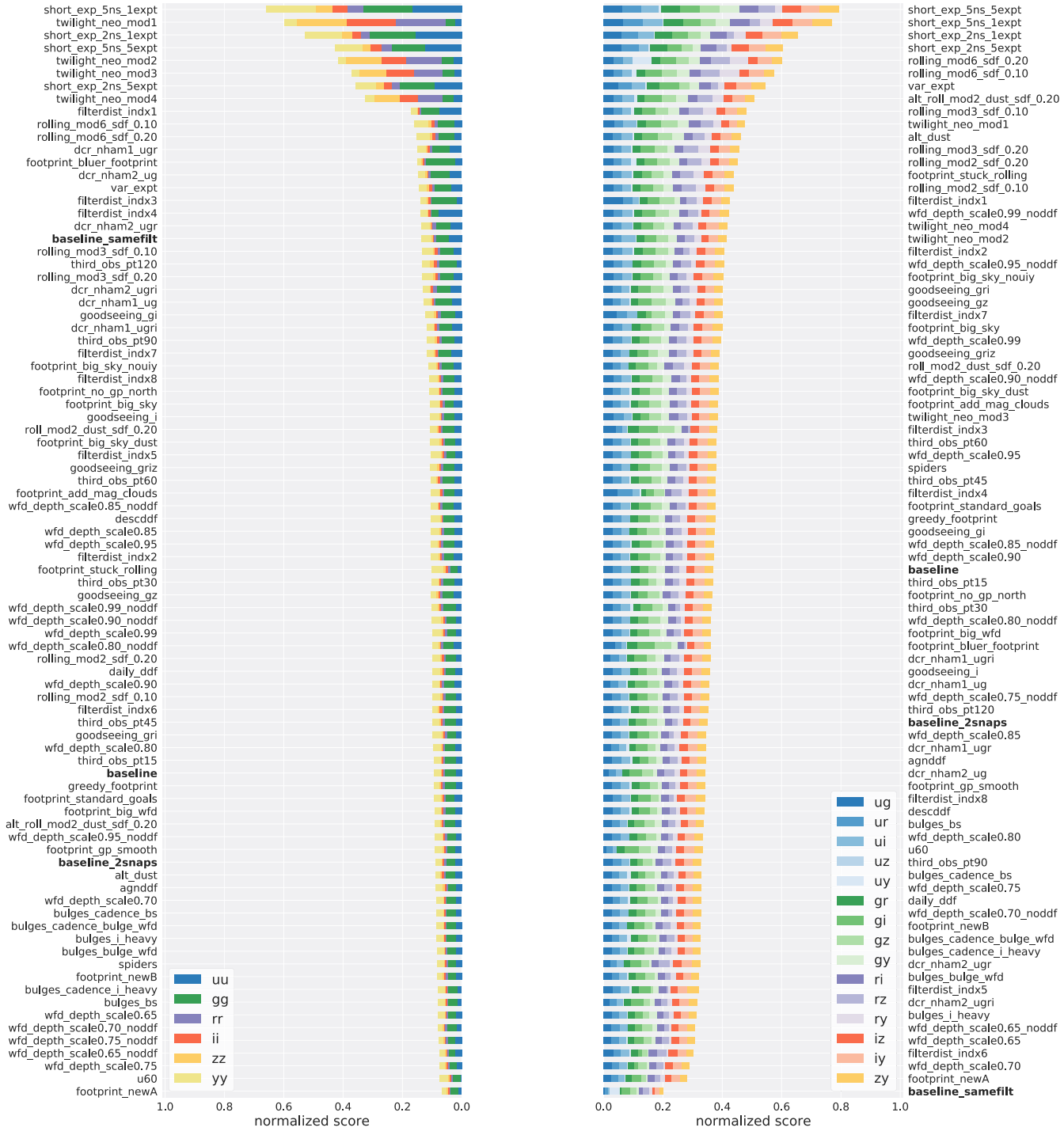


Figure 5. Figures of merit $\mathrm{FoM}_{\mathrm{tGaps}}$ for the `OpSim v1.5` runs based on the distribution of time gaps. The $\mathrm{FoM}_{\mathrm{tGaps}}$ is calculated as described in Equation (3). The plot on the left (a) shows the FoM for repeat visits in the same filter. The plot on the right (b) shows the value for observations in pairs of different filters. Each `OpSim` is presented as a bar whose length corresponds to the value of the FoM: the FoMs for different filters are concatenated horizontally. For example, on the left, the different color bars represent the time gap FoM for different filters from *u* to *y*. The `OpSims` are sorted by the total FoM. In panel (a) the bars grow toward the left, in panel (b) the grow toward the right, so that asymmetries in the plot can give intuition on the overall distribution of the two different metrics across the set of the `OpSims`. See Section 3.1.

`footprint_big_sky` families are penalized in this metric. This may be a consequence of the added constraints on pointing competing with the constraints on image quality (which relate to weather, airmass, etc.).

5. Footprint

Footprint coverage is another important factor that plays a crucial role in determining LSST's ability to discover anomalous and unusual phenomena.

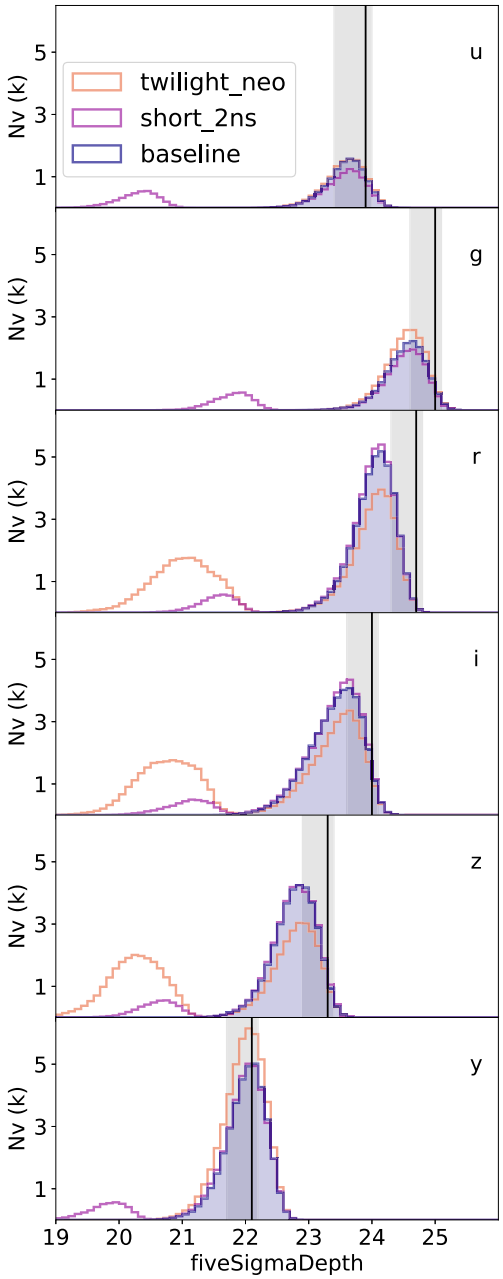


Figure 6. Distribution of depth for images in three different `OpSims`. The survey specifications are indicated by the a gray band (minimum requirement to stretch goal) and the vertical line (design specification) as per Ivezić & the LSST Science Collaboration (2013), their Table 6, for each filter (as indicated in the top left of each panel). Some `OpSims` are designed to include in the WFD short exposures and may perform well in metrics based on number of exposures taken. But shallower images generate lower signal-to-noise ratio measurements and only allow for the exploration of a smaller volume of the universe. This would have a negative impact on the discovery of anomalies if it came at the cost of long exposures. We show the distribution of 5σ depth for images in three `OpSims`: the baseline (blue filled histogram), `twilight_neo_mod1`, and `short_exp_2ns_1expt`. `twilight_neo_mod1` and `short_exp_2ns_1expt` have additional short-exposures leading to a bimodal distribution. However, while for `short_exp_2ns_1expt` `OpSim` the distribution of faint (fainter than ∼21.5 in each band) images is similar to the that of the baseline, `twilight_neo_mod1` has fewer faint images in the *r*, *i*, and *z* bands, and more in the *y* band. See Section 4.

For the purpose of our analysis, we define "footprint" as the extent of the sky (number of `healpixels` in the sky) that is "well observed" for each filter or filter pair of interest. This approach is agnostic about the location of the fields in the sky, as we do not know where true novelties may be. To decide if a field is "well observed," we compare the number of relevant observations to the median number obtained in a chosen baseline LSST implementation (here, `baseline_1.5`, all visits excluding DDFs), under the motivation that the strategy ultimately adopted by the project should outperform this baseline (see Tables 2–5).

In this context, "relevant observations" are defined slightly differently depending on whether one is measuring brightness evolution or color. For single filters that measure brightness evolution, all observations in that filter are relevant as they measure different timescales of evolution (so the comparison count is just the number of observations in the 10 yr surveys). For filter pairs that measure the color, observations in a pair are only relevant if they occur within a window of time small enough to measure color in spite of temporal flux evolution, so the comparison count is the number of observation pairs constructed from images in different filters and collected within two days of each other. This is a softer constraint that the 1.5 hr maximum gap to measure color adopted in Section 3 (set by the fastest observed transient evolution timescales; e.g., Bianco et al. 2019; Bellm 2021), but we choose it in order to separate the footprint and timescale considerations to some extent, while still requiring that two observations making up a color measurement are separated by a sufficiently narrow gap to follow reasonably rapid color evolution. Our sensitivity to this choice is discussed further in Appendix A.

We acknowledge that this choice of threshold is somewhat arbitrary and that this will influence the result of this component of our FoM. We will return to the choice of threshold, and its impact on the science figures of merit, when we extend our analysis to other versions of the `OpSim` strategies in Section 6.3. For the present, we emphasize that this thresholding is entirely relative to the baseline simulation: we are *not* imposing a requirement that the threshold must guarantee a significant probability of detection. Consider for example the $u-y$ filter pair: since there are not $u-y$ observations in the `baseline_1.5` survey,[23] a field with nonzero $u-y$ pairs would be considered "well observed" by us for that filter combination. By choosing a threshold relative to a fiducial implementation of the LSST survey, we seek to identify survey strategies that expand the potential of LSST.

With these considerations in mind, the footprint figures of merit are calculated following the steps below. For each filter pair *k*:

1. count the number of visits per `healpixel`; for same-filter pairs, consider all possible time gaps; for different-filter pairs, consider time gaps within two days;
2. count the number of `healpixels` with more visits than the median number of visits in `baseline_1.5`.

The top panel of Figure 8 shows sky maps with healpixels (`NSIDE` = 16) colored by the result of this metric for four `OpSims` of interest.

Depending on whether a scientist's focus is on extragalactic or galactic anomalies, the preferred footprint would be different. For extragalactic anomalies, one would simply want

[23] The *u* and *y* filters benefit from very different sky conditions, and since the Rubin filter wheel can house only five out of the six filters at once, these two filters are likely to never be available in the same night.

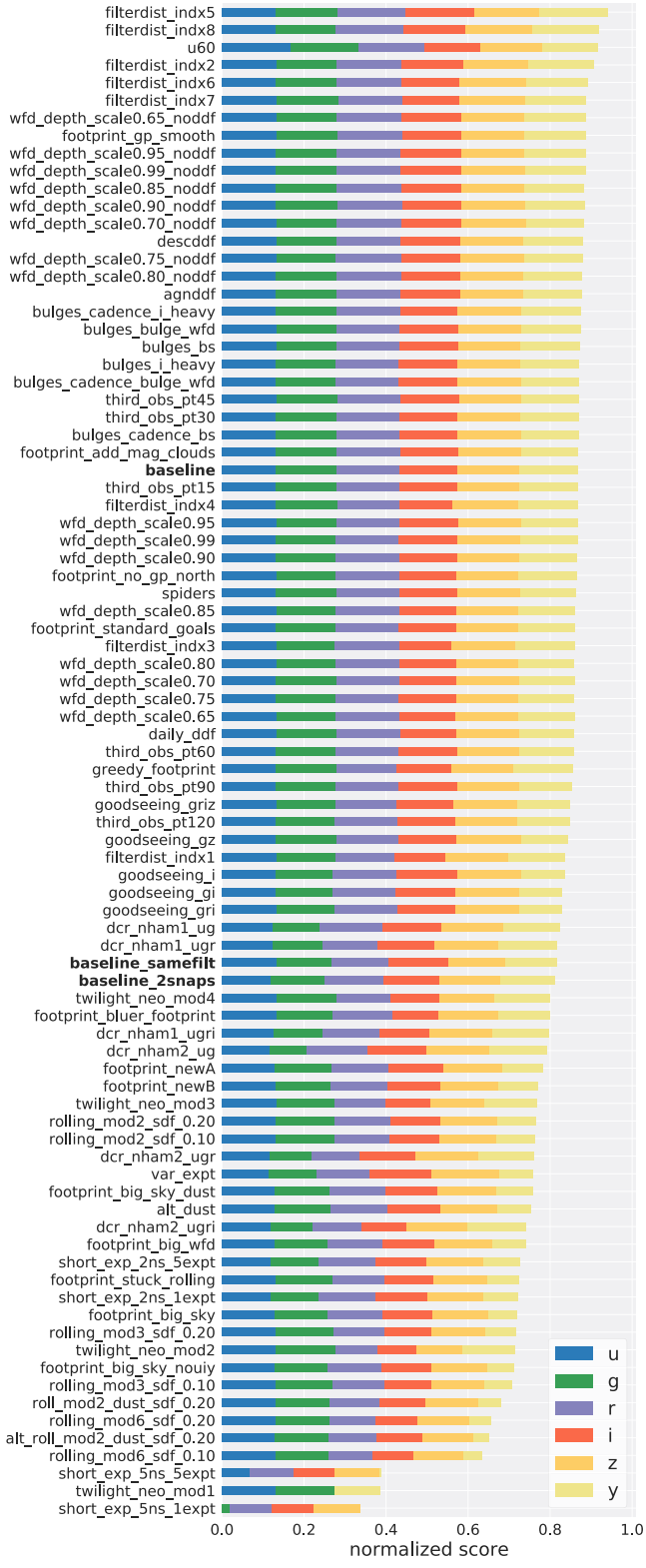


Figure 7. Ranking of OpSims based on the depth of the exposure as discussed in Section 4. u60, which produces a 60 s *u*-band exposure instead of the standard 30 s, extends the observed volume slightly in the *u* band and ranks highly in this metric; however, it was performing poorly in both FoM_{tGaps} (Figures 5(a) and (b)). Some (but not all) short_exp and twilight OpSims perform poorly, indicating the short exposures come at the cost of high-quality, deep, exposures at least in some filters. Otherwise, for the most part, the families of OpSims are clustered together in this diagram, all with a similar FoM_{depth} score: 90% of the OpSims generate values within 10% of each other in this metric.

Table 2
Thresholds for the Footprint Figures of Merit Based on the Count of Visits in Pairs in baseline_v1.5

	u	*g*	*r*	*i*	*z*	*y*
u	1711					
g	67	3570				
r	76	130	20301			
i	24	45	185	20582		
z	2	13	37	200	16470	
y	0	5	18	92	220	18431

Note. If the number of observations is larger than the number reported in this table, a healpix field is considered "well observed" as it exceeds the expectation that is set by the baseline strategy. See Section 5.

Table 3
Thresholds for the Footprint Figures of Merit Based on baseline_v1.5 Visit Count for GP Fields; See Section 6.2

	u	*g*	*r*	*i*	*z*	*y*
u	780					
g	58	1081				
r	41	61	1275			
i	11	19	46	1176		
z	2	8	12	67	1126	
y	0	4	8	31	61	1830

Table 4
Thresholds for the Footprint Figures of Merit Based on baseline_v1.5 Visit Count for the LMC; See Section 6.2

	u	*g*	*r*	*i*	*z*	*y*
u	685					
g	46	862				
r	37	47	882			
i	7	11	27	842		
z	1	2	6	67	925	
y	0	2	2	15	38	1458

to maximize the sky coverage under the assumption of isotropy, whereas for Galactic science, the probability of discovering an anomalous object or phenomenon would scale with the number of objects in the Galaxy in that observing field (Street et al. 2021).[24] Therefore, in addition to the FoM just described, which focuses on extragalactic science and which we call FoM_{EG}, hereafter, we include one further footprint figure of merit, FoM_{Gal}, which scales with the field's star density. FoM_{Gal} is the sum of the product of the binary metric described above and the number of stars in that field (itself obtained from a realization of the TRILEGAL models of Girardi et al. 2005 accessed via MAF), i.e., the sum of number of stars extended only to "well observed" healpixels. Maps resulting from this metric are shown in the bottom panel of Figure 8. The threshold $N_{\mathrm{median},k}$ can be read off of the appropriate table (Table 2 through Table 5) with *k*, a two-filter index (e.g., $k = gg$ or k=*ui*, which for the WFD would lead to

[24] While the richest Galactic regions (including the inner Plane and the Magellanic Clouds) will be subject to some degree of spatial confusion at LSST apparent magnitudes, this confusion will be similar, to first order, between all observing strategies.

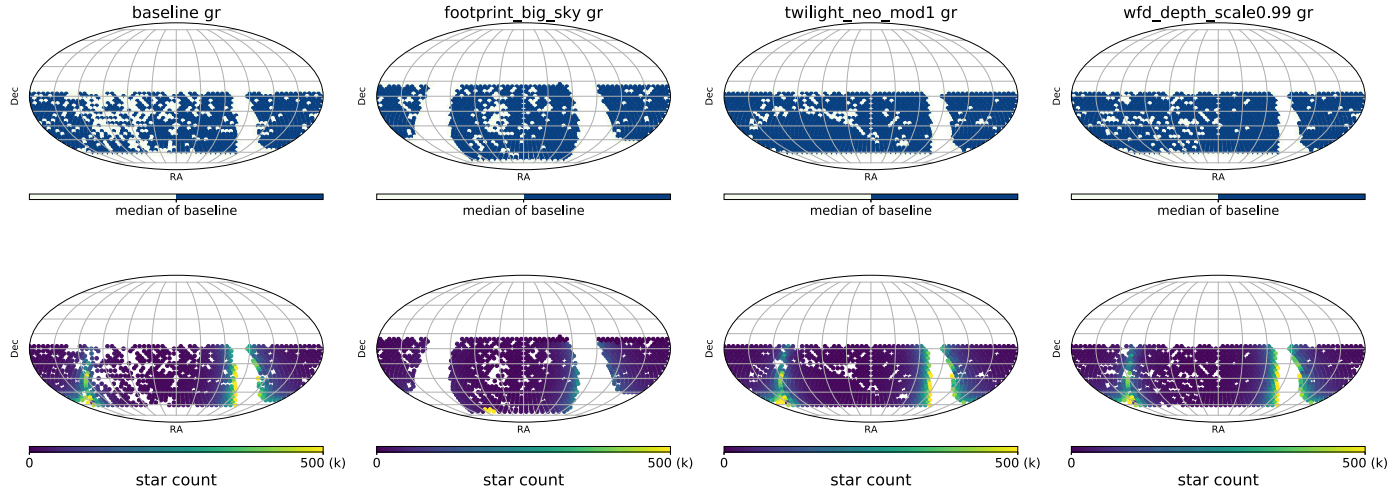


Figure 8. Maps of the footprint FoMs for four OpSims: from left to right, baseline_v1.5, footprint_big_sky, twilight_new_mod1, and wfd_depth_scale0.99. In the top panel, the colored heapixels ($N = 16$) are the locations where the number of observations in g and r within a 2 day interval is greater than the median count for such observations in baseline_1.5: the metric for extragalactic footprint ($\mathrm{FoM_{EG}}$). The bottom panel shows the results for $\mathrm{FoM_{Gal}}$ where each heapixels is weighted by the corresponding star count. The footprint_big_sky extends the footprint to higher latitude than baseline_1.5. The twilight allocates additional visits near twilight. See Section 5.

Table 5
Thresholds for the Footprint Figures of Merit Based on baseline_v1.5 Visit Count for the SMC; See Section 6.2

	u	g	r	i	z	y
u	561					
g	32	741				
r	26	39	780			
i	5	9	18	861		
z	3	4	4	66	903	
y	2	4	5	19	40	1225

$N_{\mathrm{median},gg} = 3570$ and $N_{\mathrm{median},ui} = 24$). For an OpSim, these FoMs are therefore defined as:

$$
\begin{aligned}
p_{i,k} &= 1 \text{ if } N_{i,k} > N_{\mathrm{median},k} \text{ else } 0 \\
\mathrm{FoM_{EG}} &= \textstyle\sum_k^{\mathrm{filters}} P_k \sum_i^{\mathrm{fields}} p_{i,k}, \\
\mathrm{FoM_{Gal}} &= \textstyle\sum_k^{\mathrm{filters}} S_k \sum_i^{\mathrm{fields}} s_i p_{i,k}, \qquad (5)
\end{aligned}
$$

where i is an index that ranges over all observed fields, s_i is the number of stars for the ith field, and $p_{i,k}$ is set to 1 or 0 as indicated. Similarly to the depth figure of merit (Section 4), the normalization factors P_k and S_k are the reciprocal of the maximum value across all of the OpSims of the sum over fields (the inner sums in Equation (5)) in the kth filter pair divided by the number of filter pairs. This normalization serves to treat all of the filter pairs on an equal footing: an OpSim must be simultaneously top-ranked in all filter pairs under consideration to achieve an FoM value of 1.0. The two footprint FoMs are thus each evaluated as a sum over all filters and filter pairs (by Equation (5)), and are presented in Figures 9 and 10 for all 86 simulations in OpSim v1.5.

While some OpSims were designed to cover a large footprint (such as footprint_bigsky), other OpSims perform better under the footprint figure of merit we develop here, which includes visit count thresholding in addition to simply evaluating the area covered. So we see again the short and rolling cadences rising to the top.

6. Discussion

We have created a series of MAFs and FoMs to assess the ability of Rubin Observatory LSST to discover completely novel astrophysical objects and phenomena.

In Figure 2 we showed that some regions of the transients' phase space are underpopulated and argued this is likely due to observational biases. Nonetheless, the huge survey grasp of LSST should lead to the discovery of true anomalies even in regions of this phase space that are not observationally underpopulated. As traced by photometry, LSST will expand the observed *volume* of space, such that rare events will be detected at a higher rate than before, which motivates our approach in assessing the coverage of the feature space relevant to the discovery of transients.

Thus, in an attempt to remain agnostic to what specific characteristic may render an object or phenomenon anomalous and therefore which kind of anomalies we could discover, we choose to assess the completeness of coverage achieved in a phase space quantified by figures of merit exploring the following observables:

1. Flux change, parameterized as $\mathrm{FoM_{tGaps}}$-magnitude;
2. Color, parameterized as $\mathrm{FoM_{tGaps}}$-color;
3. Depth, parameterized as $\mathrm{FoM_{depth}}$;
4. Sky footprint, parameterized as $\mathrm{FoM_{EG}}$;
5. Star counts, parameterized as $\mathrm{FoM_{Gal}}$.

The five elements enumerated above are added straightforwardly to one another (Equation (1)), although the final FoM could be fine-tuned to some phenomenological expectations (for example, to the discovery of *Galactic*, as opposed to *extragalactic* transients) by choosing the weights in the sum over the FoM components.

We note that the weights are thus formally somewhat arbitrary but scientifically rather crucial to the balance of scientific considerations imprinted on the sum figure of merit by the investigator. Remaining "agnostic," we opt to strive for

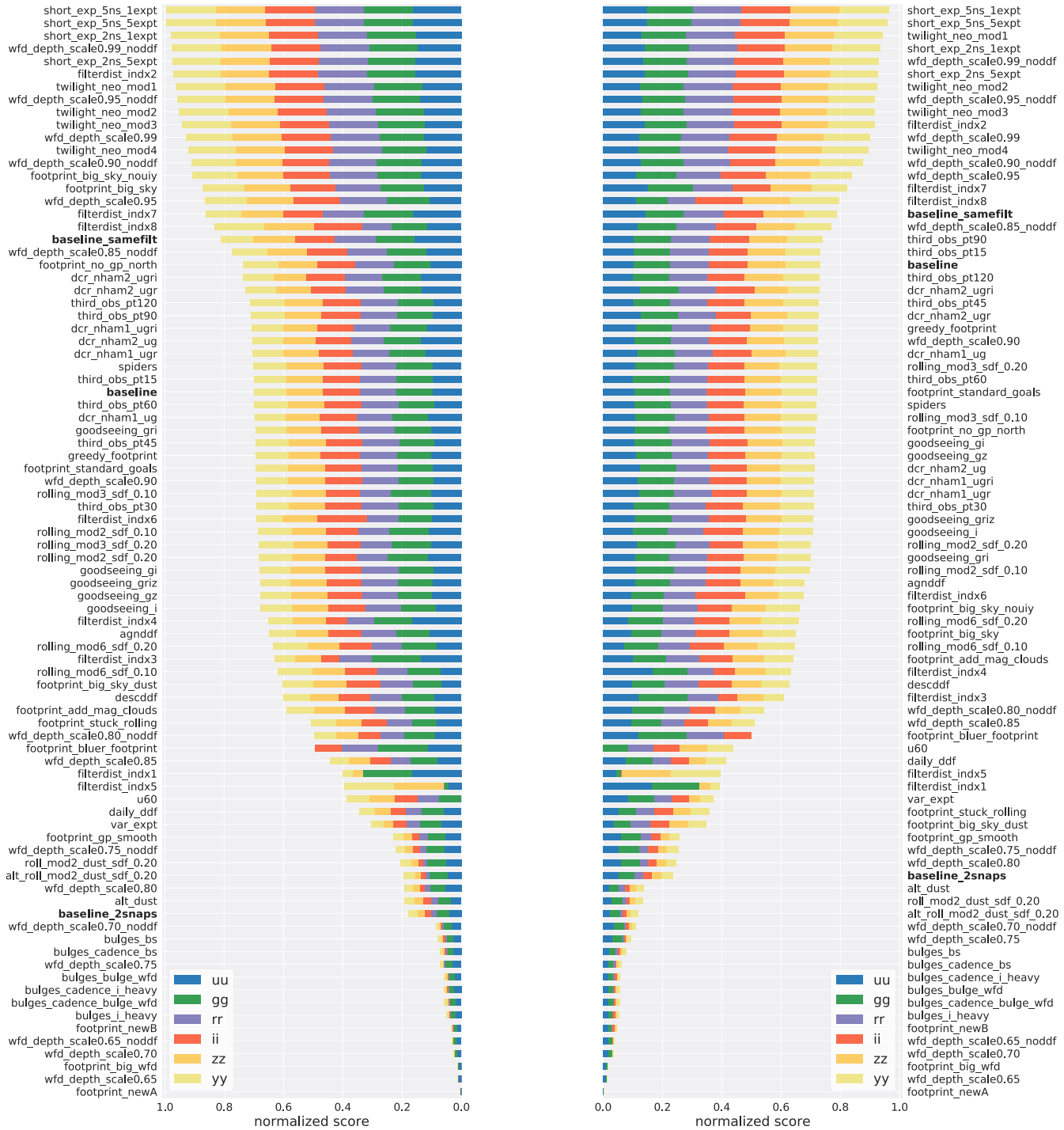
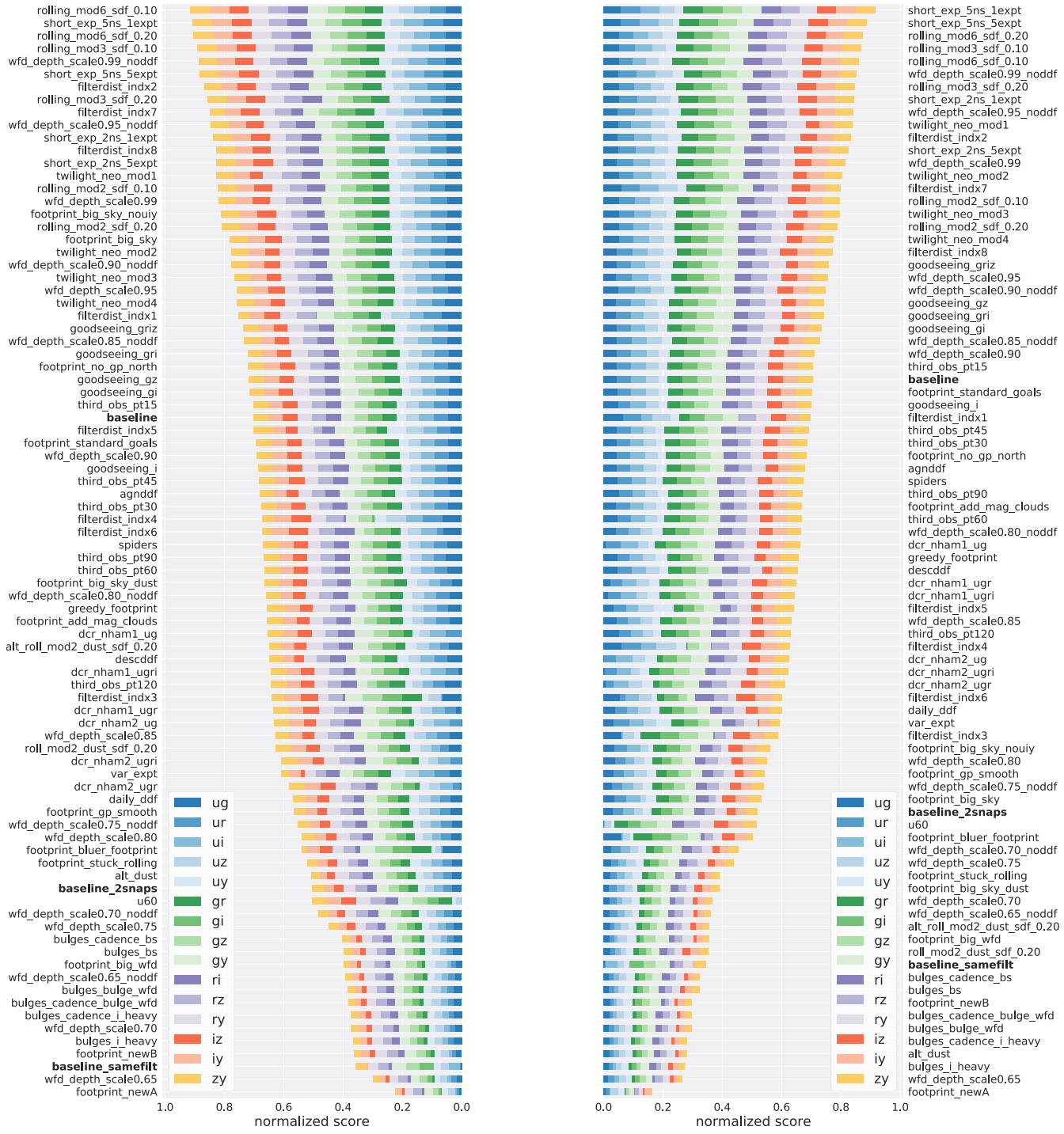


Figure 9. The figure of merit FoM_{EG} (a) and FoM_{Gal} (b) for all `OpSim` runs (for the WFD survey, selected as `proposalId=1` in the `SQL` query; see Section 2) based on footprint coverage and star count with image pairs in the same filter (measuring light-curve shape) as described in Section 5 (Equation (5)). Colors and symbols denote filter combinations using the same conventions as in Figure 5. The two FoMs go hand in hand, with small differences in the ranking.

balance in the normalization and relative weighting of each element of the FoM. The individual figures of merit are each normalized so that they essentially rank all of the `OpSim`s on a 0.0–1.0 scale for that particular dimension in feature space, where an `OpSim` must be top-ranked simultaneously in each filter (or filter pair) to achieve a maximum 1.0 score (Sections 3–5). We then choose weighting factors (w_i in Equation (1)) to weight each of the five FoMs equally.

6.1. Main Survey

First, we want to summarize some considerations arising from our analysis of the performance of different LSST



(a) (b)

Figure 10. The same as Figure 9 but for image pairs in different filters (measuring color) as described in Section 5 (Equation (5)). Colors and symbols denote filter combinations using the same conventions as in Figure 5.

simulations from `OpSim v1.5`. These considerations, however, should be read in light of the discussion of the different `OpSim` versions in Bianco et al. (2022) and rely on the reader having familiarity with the `OpSim v1.5` set, as described there in and in more detail on the Rubin Community forum.[25] We will also extend this discussion to other versions of `OpSims` briefly in Section 6.3.

The bar charts included in this work (Figures 5, 7, 9, and 10) provide an intuitive way to understand how sensitive an FoM is to observing cadence choices. We note that:

1. The $\mathrm{FoM}_{\mathrm{tGaps}}$ for flux evolution (Figure 5(a)) is very sensitive to `OpSim` details: `OpSims` that included short exposures are critically improved as they provide visibility

[25] https://community.lsst.org/t/fbs-1-5-release-may-update-bonus-fbs-1-5-release/4139

into timescales that are otherwise not accessible to the survey. Going back to the phase space of transients presented in Figure 2 and the discussion of existing observational biases, rapid evolutionary timescales are quite likely to host unobserved, unexpected phenomena: *true novelties*. Our metric reflects this expectation.

2. This effect is mitigated by the depth metric that down-weights `OpSims` where the short exposures come at the cost of overall survey depth. Otherwise, this metric does not differ much across most `OpSims` as the median observation depth is well defined by I19.
3. The galactic and extragalactic footprint metrics as defined by us are somewhat less sensitive to observing choices, as indicated by the more gentle slope of the silhouette of the bar chart in Figures 9 and 10. However, in Figure 9 three regimes are visible: `OpSims` that include short exposure (`twilight`, `short_exp`, and some `wfd_depth` implementations with a large fraction of observations included in the WFD survey, i.e., a large "scale" parameter) raise to the top. A number of specific implementations from nearly all families, however, sink to the bottom and perform very poorly (some footprint implementation and `wfd_depth` surveys with small scale parameter). The ranking of the `OpSims` is similar for Figures 9 and 10.

Figure 11 shows the performance for the combined FoM as described above, organized by `OpSim` family. Observations associated with the WFD proposal are shown in Figure 11(a), and the results including the mini-surveys are shown in panel (b). The mini-surveys themselves are discussed in more detail in Section 6.2. This visualization provides a synoptic look at our FoM. Individual components of the FoM can still be identified by the color of the bar element. Furthermore, this visualization allows us to identify the performance range for a family of `OpSims`, providing a more intuitive way to assess the reason why `OpSims` may rank differently, but also a way to assess how the detail of an implementation can affect results. For example, the `short` and `twilight` families are among the top performers, with little sensitivity to the details of the implementation. Conversely, the `wfd` and `footprint` families (the former ranking third overall, the latter in the middle, ranking seventh) provide a range of results, from excellent to poor, depending on the implementation details. For both `wfd` and `footprint`, the variation in performance within the family is dominated by differences in the footprint and star-count FoMs (the purple and orange portions of the bars). For the `wfd`, the result of both footprint FoMs scales with the "scale" parameter, the number of visits allocated to the WFD survey (but see also Section 6.2). It should be noted that these are core families of simulations, with a range of implementation details that can be tweaked, so it is not surprising that they result in a range of measured performance. See Section 2.4, Table 1, and Bianco et al. (2022) for more details.

In Section 6.2 we will discuss Figure 11(b) and address the question of what the mini-surveys add to the science performed in the WFD regions, by considering together all of the exposures not identified with a DDF.

Applying our metrics to `OpSim v1.5`, we note that:

1. The $\mathrm{FoM}_{\mathrm{tGaps}}$-magnitude-evolution component (see also Section 3) is pushing entire families of `OpSims` to the top, namely those that include short observations and thus expand the LSST feature space to short timescales.
2. Within an `OpSim` family, the most significant contribution in determining the ranking of `OpSims` is the $\mathrm{FoM}_{\mathrm{EG}}$ and $\mathrm{FoM}_{\mathrm{Gal}}$ that are, however, strongly correlated (see also Section 5).
3. Overall, the top-performing `OpSims` in each family are all within a score of $\sim$0.3 of each other, demonstrating that all `OpSim` families have the potential of being implemented in a way that is favorable to the discovery of true novelties, with the exception of specialized surveys such as `bulge` and `alt_dust`. These families that typically allocate visits to focus areas of the sky are penalized in the footprint portion of our FoM. We refrain from discussing the `rolling` family of `OpSims` until Section 6.3.

Figure 12 presents radar plots for selected `OpSims`, to help assess the balance between the figures of merit when designing the final strategy. The best-performing `OpSims` for each of the top four families are plotted (panel (a) for the WFD), along with `short_exp_2ns_1exp` (which stands out as the most well-balanced `OpSim` for our set of metrics when applied to WFD).

With this visualization, we can see the substantial impact of the flux-change component of the metric, which measures completeness in pairs of observations in the same filter, on the overall result, and how this metric is, however, compensated by the depth $\mathrm{FoM}_{\mathrm{depth}}$.

We provide an interactive widget that allows the reader to explore the radar plot for our set and other sets of metrics in Appendix B.

6.2. Mini-surveys

In addition to the primary WFD survey, LSST has the capability of conducting mini-surveys including but not limited to the Galactic Plane, Magellanic Clouds, and DDFs. These mini-surveys enhance science cases that yield greater science return with greater density of targets, including (but not limited to) the detection of stellar-mass black holes, dwarf novae and Type Ia supernova progenitors, and gravitational microlensing at various timescales. Because these mini-survey regions tend to cover regions of high density of stellar sources, they are therefore more likely to discover phenomena never observed before.

To assess the coverage achieved in areas of interest to the mini-surveys, we select observations by spatial footprint (rather than by `proposalID`), as discussed in Section 2.

Figure 13 shows the adopted mini-survey regions: Galactic Plane (GP), Large Magellanic Cloud (LMC), and Small Magellanic Cloud (SMC). The adopted GP footprint is a cosine function of Galactic longitude, with amplitude $|b| = 10°$ and first zero at $|l| = \pm 85°$, plus a strip at constant thickness $b \leqslant 2.°5$ to accommodate the thin disk at all longitudes.[26] For the Magellanic Clouds, we select all `healpix` fields (resolution parameter `NSIDE`=16) within 3.5 degrees (the side of the LSST field of view) of any of the 12 fields covering the

[26] This GP footprint is similar to the "zone of avoidance" from high-density GP regions defined in method `_plot_mwZone()` in the `MAF` module `spatialPlotters.py`, except that, there, the first zero occurs at $l = \pm 80°$.

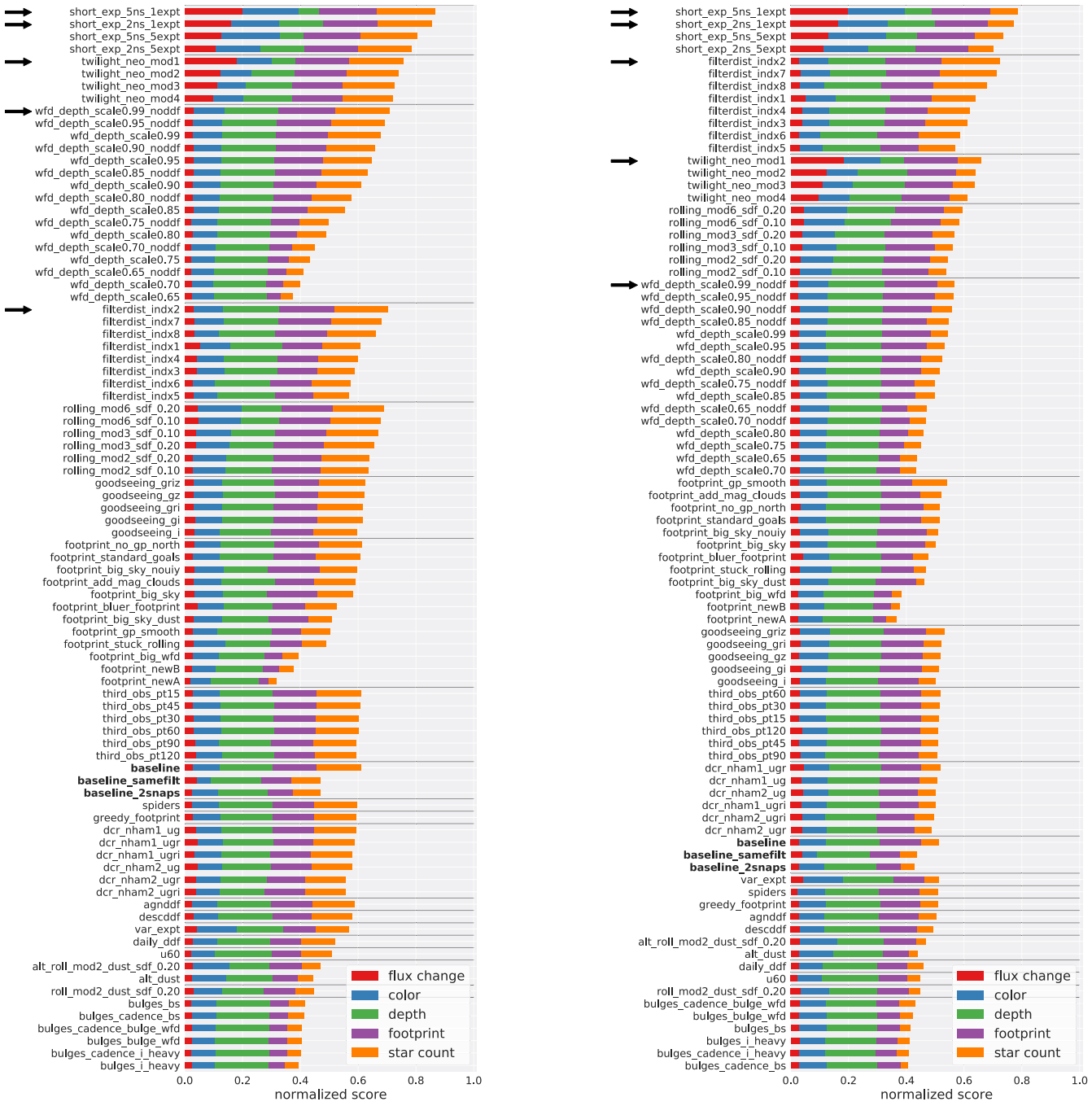


Figure 11. Bar plot showing the performance for our final five-fold FoM ranked by family's top-performing `OpSim`: (a) for WFD observations selected by setting `proposalId=1`, and (b) all observations not identified with DDF. Arrows point to the `OpSims` that are also shown in the radar plots in Figure 12. This plot is discussed in Section 6.

cloud main bodies proposed in the Olsen et al. (2018) cadence white paper.

We also provide code for the user to choose a specific region of the sky of their interest (see Appendix B) either by setting a formula from coordinate parameters, or by interactively selecting pixels.

The individual figures of merit in these regions are normalized following similar schemes as for the main survey, but with thresholds or maximum values evaluated over the spatial regions of interest. We normalize the FoM of the time-gap metric by its maximum value across the `OpSims`. For the footprint, as with Section 5, we choose the median number of visits from `baseline v1.5` within the defined footprint, normalized by the total selected number of fields within (254 for GP, 12 for the LMC, and five for the SMC), as a threshold to decide whether to classify a field as "well observed" (see Tables 3–5 for the comparison $N_{\mathrm{median},k}$ counts for the mini-survey spatial regions).

Figures 14 and 15 present the evaluations of the figures of merit on the mini-survey regions. Figure 14 shows the figure of

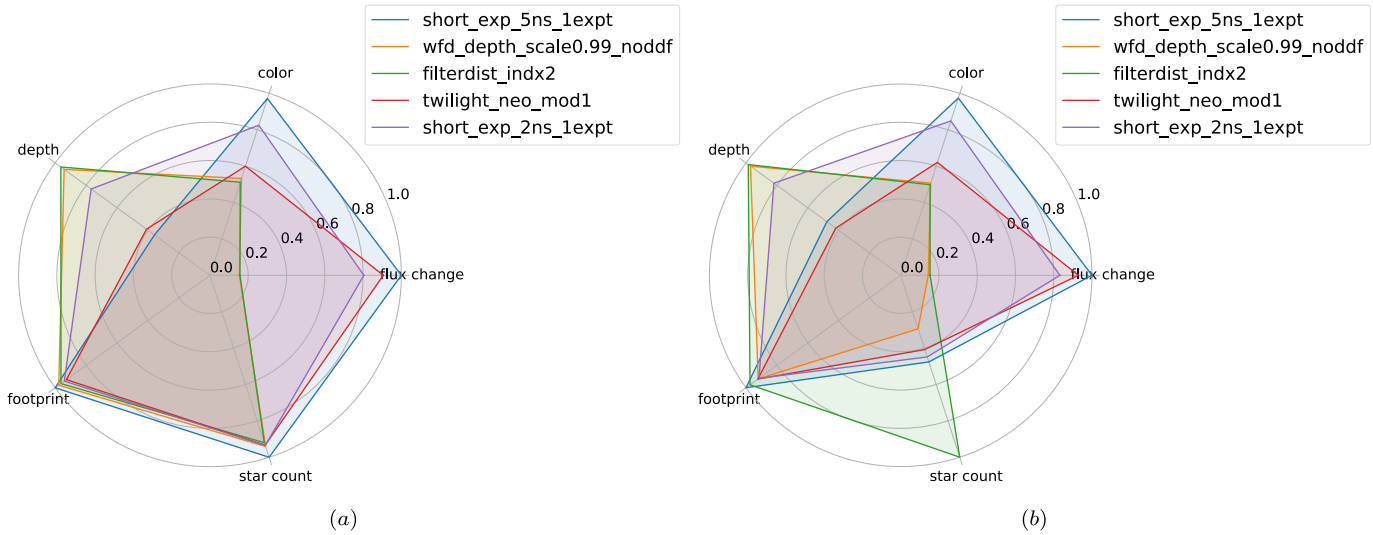


Figure 12. Radar plots showing the highest-performing `OpSim` in each of the top four `OpSim` families, plus `short_exp_2ns_1exp` (in purple: this is the most well-balanced of the `OpSim`s for WFD), for (a) the WFD survey (`proposalId=1`), and (b) for all regions excluding the DDFs. An interactive version of this plot is available at https://xiaolng.github.io/widgets/radar.html (see Appendix B). These present the metrics as the vertices of a polygon with the metric value mapped to the distance from the center of the polygon. With multiple `OpSim`s plotted in the same radar plot, we can compare the tensions between FoM components, while the total area inside of the polygon is a measure of the overall quality of the `OpSim`. For our FoM, which is the simple sum of five components, this visualization is well suited to provide a synoptic view. This plot is discussed in Section 6.1.

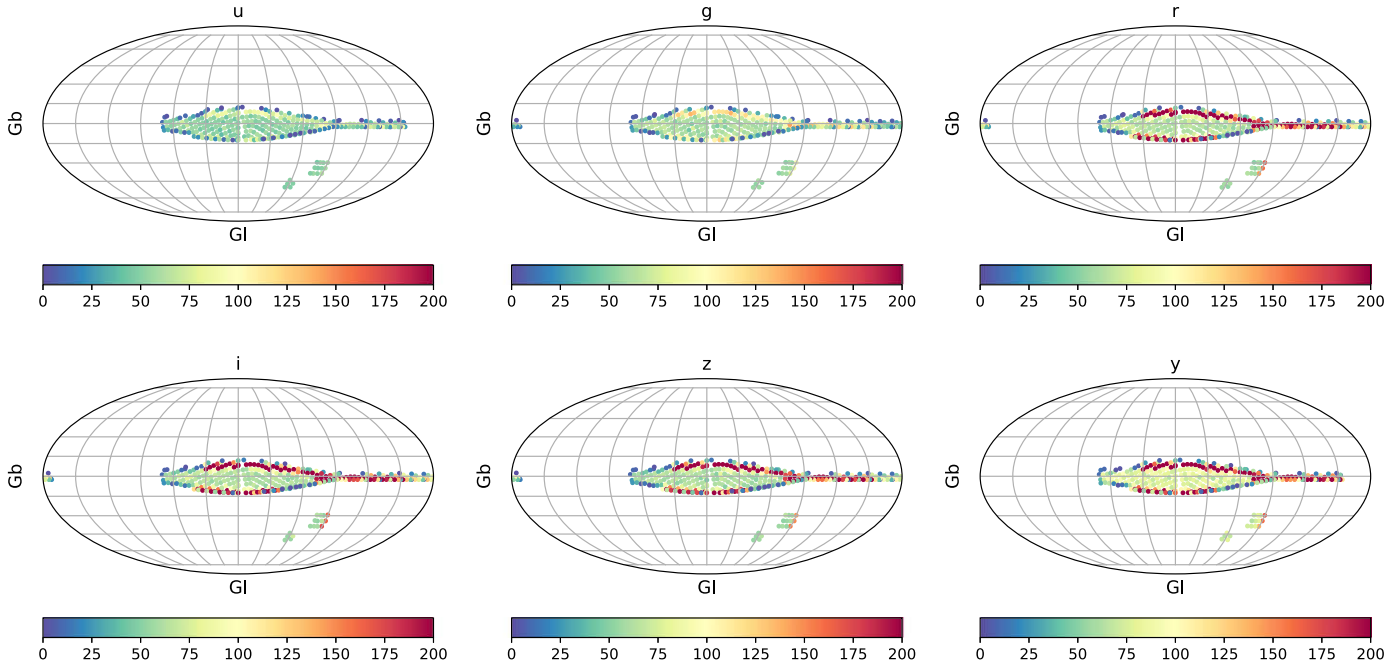


Figure 13. The footprint of the GP and Magellanic Clouds. The definition of these spatial regions and the selection of the corresponding footprint are described in Section 6.2. We select 254 fields for the GP, 12 fields for the LMC, and five fields for the SMC. The color shows the number of visits in `baseline_v1.5` in six bands in each field.

merit evaluation for the three spatial regions for fields observed as part of the WFD coverage (i.e., observations with `proposalID=1`). This demonstrates quite dramatically that the Magellanic Clouds are not allocated WFD-like coverage in most of the strategies considered. Figure 15 widens the evaluation to include all exposures except those associated with DDFs.

Strong variation is apparent between the families of `OpSim`s, as expected for families that experiment with the areas of coverage on-sky. The `footprint` family shows strong variation within the same family, depending on which region is favored: `footprint_gp_smooth` performs the best for the GP, but is in the bottom quartile of all of the `OpSim`s for the Magellanic Clouds. Conversely, `footprint_add_mag_clouds` is ranked high for both Magellanic Clouds, and the GP regions.

The `alt_` implementations perform quite badly for the GP regions, but allocate favorable observations to the Magellanic Clouds. Curiously, the `bulges` family of `OpSim`s are among the *worst*-performing families for the GP regions, though they are in the top three for the Magellanic Clouds. The baseline

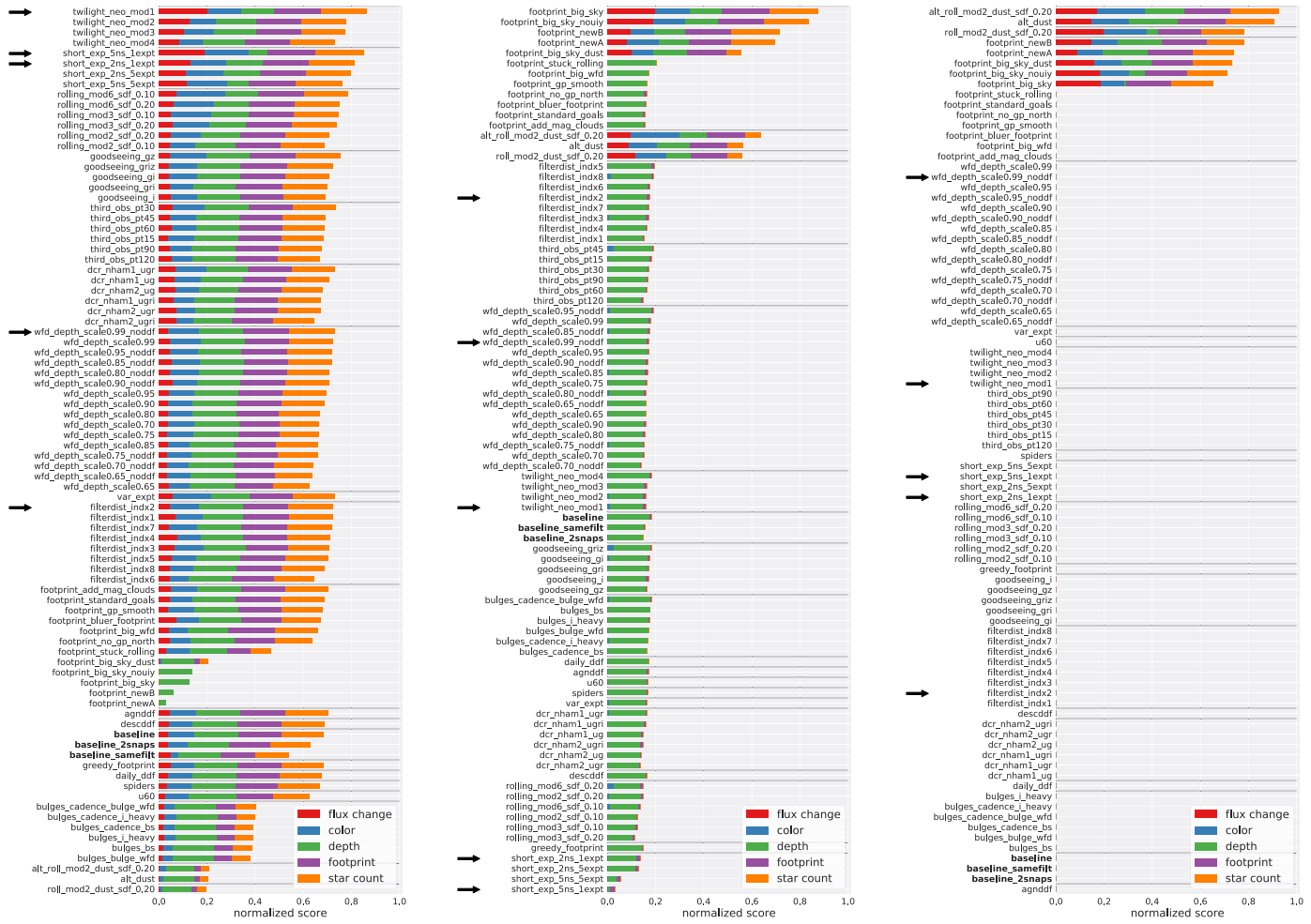


Figure 14. Bar plot, as Figure 11, showing OpSims ranked by family, but this time for three selected spatial regions: (a) the GP, (b) the LMC, and (c) the SMC. Only visits allocated to the WFD (labeled as proposalId=1) are counted. Arrows point to the OpSims that are also shown in the radar plots in Figure 12. The twilight and short families of OpSims perform best on the GP, as they did over the entire WFD footprint, while wfd_depth, formerly ranked third, is now ranked the seventh. But in reality, the top-performing OpSims in most families all perform similarly. The main differences are generally driven by $\mathrm{FoM}_{\mathrm{tGaps}}$ in the same filter. Only eight OpSims cover the SMC, and only five cover both the LMC and SMC with WFD-identified observations. See Section 6.2.

strategies appear near the middle of the distribution for the mini-survey regions.

Since the regions are to some extent competing with each other in terms of allocation, Figures 14 and 15 may be best interpreted in terms of which OpSims to avoid due to their being problematic for particular regions of scientific importance. From that perspective, OpSims alt_dust and footprint_new are unlikely to satisfy those interested in the GP, while the filterdist family serves the Magellanic Clouds particularly poorly, at least for the purpose of anomaly detection. We remind the reader that an observational strategy that performs well for one science case does not necessarily perform well for others, and vice versa (Bianco et al. 2022).

By comparing Figures 11(a) and (b), we can address the question of what the mini-surveys add to the science performed in the WFD regions by comparing the results of our FoM for the WFD-exposures only (Figure 11(a)) and for all of the exposures in an OpSim, excluding only those identified with a DDF (Figure 11(b)). Most of the FoMs remain relatively unchanged by the inclusion of the mini-survey exposures. The exception is the "star density" FoM: $\mathrm{FoM}_{\mathrm{Gal}}$ returns systematically *lower* values for most OpSims when the mini-surveys are included. However, the lower metric value is an artifact of normalization and thresholding. Adding the mini-survey observations can add star-dense fields that lead to a significant performance improvement in $\mathrm{FoM}_{\mathrm{Gal}}$. Since the best OpSim in this metric returns $1.0/N_{\mathrm{filter\_combinations}}$ by construction for each filter combination (i.e., all OpSims are normalized by the best performer for each filter combination), this performance gap causes the metric for the other OpSims to drop. Aside from standout OpSims, the ordering of the families changed somewhat, though not radically: the short_exp family remains the top performer, for example, and the baseline strategies remain in the bottom quartile when OpSims are ranked by family.

6.3. Comparison with v1.7

Our work is based on OpSim v1.5, the version of OpSim simulations released in 2020 May. However, since then, more simulations have been released. We briefly inspected the performance of OpSim v1.7 (74 simulations at the time of writing) and OpSim v1.7.1 (10 simulations), the most recent simulations at the time of writing.

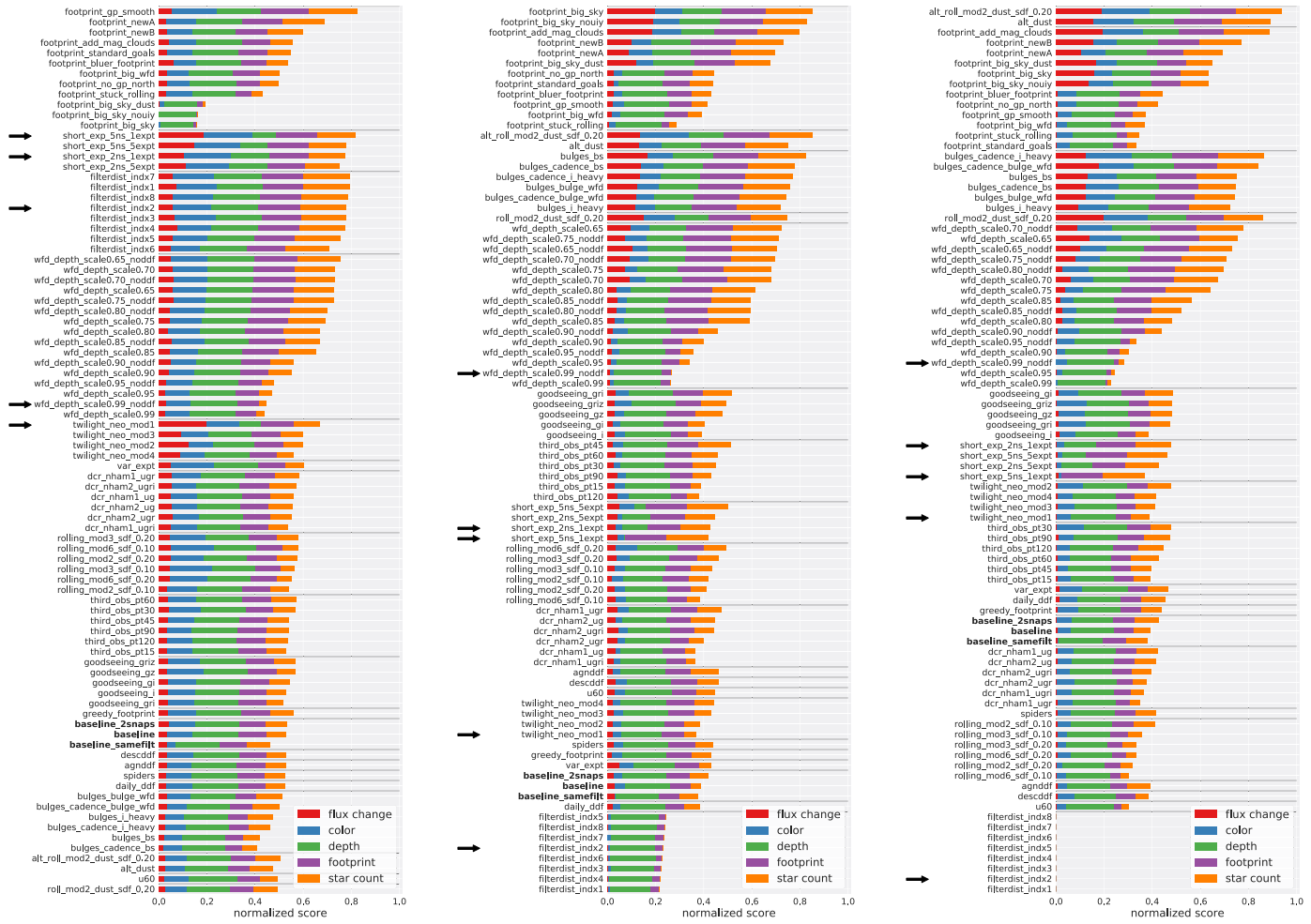


Figure 15. The same as Figure 14 but with all visits excluding DDFs being counted for the mini-survey spatial regions: (a) GP, (b) LMC, and (c) SMC. Arrows point to the OpSims that are also shown in the radar plots in Figure 12. While the footprint family has the best-performing OpSim overall for the GP (footprint_gp_smooth), it also has the worst-performing OpSims, showing the largest dynamical range due to the combined effects of both the footprint and time-gap metrics. Meanwhile, all filterdist OpSims perform well. However, the LMC and SMC are now covered with most OpSims, with the exception of the filtdist family. See Section 6.2.

It is important to note some key differences between OpSim v1.5, OpSim v1.7, and OpSim v1.7.1 (however, a thorough description of these simulations is outside the scope of this paper, and the reader is reminded that details are available on the Rubin Community web forum.[27])

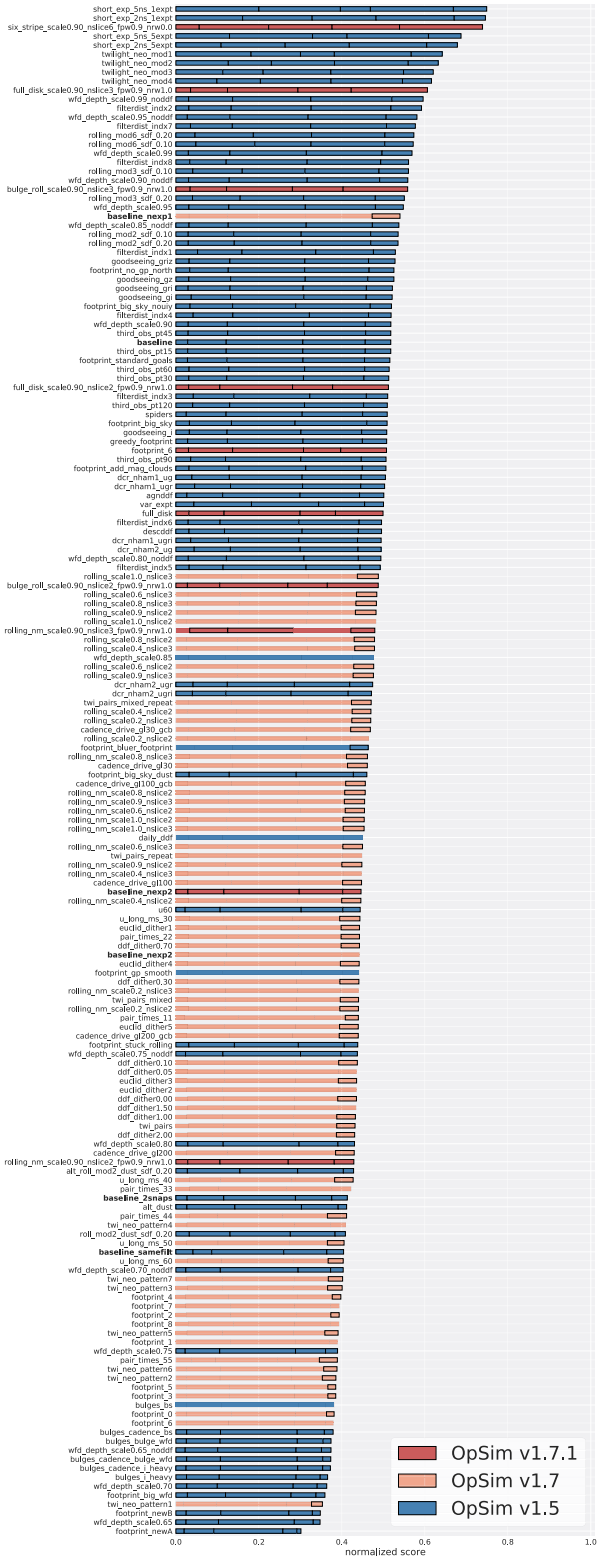
OpSim v1.5 uses 1×30 s exposures for almost all simulations, while OpSim v1.7 and OpSim v1.7.1 use 2×15 s exposures per visit. It is estimated that this would lead to a loss of efficiency of $\sim 9\%$.[28] It is also expected that the rolling family of OpSims would display significant changes compared to OpSim v1.5, due to improvements in the way rolling cadences are implemented to more closely match their specifications. Versions 1.7 and later of the rolling OpSims are considered a more reliable implementation of rolling cadence than v1.5 (Lynne Jones, private communication).

Figures 16 and 17 show our FoM for all OpSims with the three OpSim versions side by side, color-coded by OpSim. Figure 16 shows the results for observations identified with the WFD survey (proposalId=1), while Figure 17 shows the results for all observations except those identified with DDFs (and thus addressing the impact of the inclusion of the mini-surveys in the overall science figures of merit).

When run on our final FoM, OpSim v1.5 leads in general to larger FoM values (and thus suggests greater scientific yield). In Figure 16(a), we can observe how almost all OpSim v1.5s (blue) outperform OpSim v1.7s (orange), while OpSim v1.7.1 simulations populate all regions of the chart, with six_stripe_scale0.90_nslice6_fpw0.9_nmw0.0 outperforming all others. This is a rolling cadence, with six decl. stripes as the rolling scheme. This OpSim performs well on all components of our metrics except the piece that measures flux change ($\text{FoM}_{\text{tGaps}}$ in the same filter) where this OpSim is outperformed, as discussed in Sections 3 and 6, by OpSims that include short exposures. However, the performance on measuring color (i.e., the

[27] OpSim v1.5 https://community.lsst.org/t/fbs-1-5-release-may-update-bonus-fbs-1-5-release/4139, OpSim v1.7 https://community.lsst.org/t/survey-simulations-v1-7-release-january-2021/4660, OpSim v1.7.1 https://community.lsst.org/t/survey-simulations-v1-7-1-release-april-2021/4865.

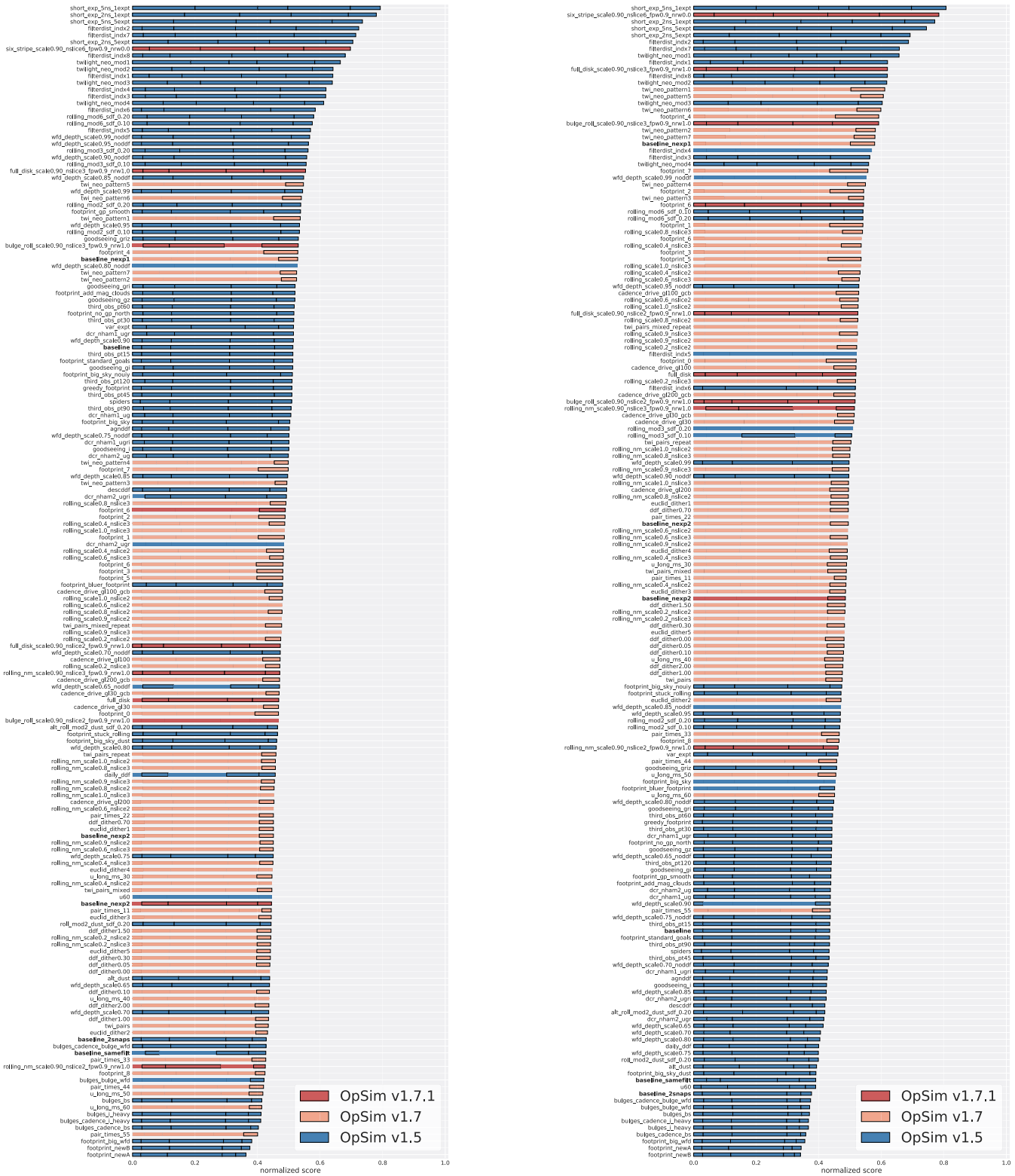
[28] See for example https://community.lsst.org/t/october-2019-update-fbs-1-3-runs/3885.



(a)

(b)

Figure 16. Bar plot showing the ranking of `OpSim`s based on our five-fold FoM for WFD visits (selected as `proposalId=1`). All simulations from `OpSim v1.5`, `OpSim v1.7`, and `OpSim v1.7.1` are included. Panel (a) shows the result of our FoM while (b) shows the result after scaling down the number of visits in `OpSim v1.5` by 9% to isolate the impact of small differences in survey efficiency associated with the single-visit collection strategy (1×30 s vs. 2×15 s). The contribution of each component of our FoM is shown in the same order as in Figures 11, 14, and 15: *flux change, color, depth, footprint*, and *star count* from left to right. This plot is discussed in Section 6.3.



(a) (b)

Figure 17. The same as Figure 16 but for all regions of the sky that do not correspond to DDFs. This plot is discussed in Section 6.3.

number of observations within 1.5 hr in different filters) and the footprint components of the FoM are sufficient to compensate for this and place the `OpSim` at the top.

We suspect that much of the difference between `OpSim` versions may be due to the sensitivity of the figures of merit to the total number of observations collected, through the $\sim 9\%$ reduction from versions `1.5` to `1.7` in the number of observations per field, as noted above (particularly considering that our footprint FoMs are based on a threshold). To test this hypothesis, we scale down by 9% the number of visits in the

calculations of the FoM_{tGaps} and footprint FoM elements (FoM_{EG} and FoM_{Gal}) applied to OpSim v1.5. The results of this exercise are shown in Figures 16(b) and 17(b).

This in fact does mitigate the almost binary split in performance between the two OpSim versions seen in panel (a), although the twilight and short simulations from version OpSim v1.5 continue to be at the top. Correcting for the 9% depth effect, the OpSim v1.7.1 release also improves relative to OpSim v1.5 (as expected), and now all nine of the 10 simulations in OpSim v1.7.1 are in the top 50%. Indeed, once we control for the overall number of observations, the OpSim v1.7 and OpSim v1.7.1 evaluations tend to populate the upper half of the distribution. However, most of the very highest-performing OpSims by our FoMs still belong to OpSim v1.5. We note that seven of the top eight-performing OpSims in OpSim v1.5 include exploration of short exposures (Figure 16(b)).

Inclusion of the mini-surveys seems to mitigate slightly the preference for 1×30 s exposures, with a handful of OpSims from OpSim v1.7 now appearing in the top quartile (Figure 17(a)). As with the WFD-only observations, the overall number of exposures seems to explain most of the discrepancy between OpSim v1.5 and the newer releases (Figure 17(b)).

The rolling_ family, which, as we mentioned, is most realistically implemented in OpSim v1.7 and OpSim v1.7.1, suffers a slight performance decrease in the later versions compared to OpSim v1.5.

While the performance across OpSim v1.7.1 simulations is diverse, six_stripe_scale0.90_nslice6_fpw0.9-_nmw0.0 stands out as a high-throughput observing scheme for our science.

7. Conclusion

Rubin LSST is designed to transform entire fields of astronomy by collecting an unprecedentedly large and rich photometric data set. Yet one of the most promising aspects of LSST is its potential to discover completely novel phenomena, never before observed or predicted from theory. We created a five-fold FoM that relies on a set of MAFs that assesses the ability of Rubin Observatory LSST to discover novel astrophysical objects, but instead of selecting known anomalies (e.g., Boyajian et al. 2018) or theoretically predicted unusual phenomena to benchmark our results, as more commonly done in the field (Soraisam et al. 2020; Pruzhinskaya et al. 2019; Aleo et al. 2020; Martínez-Galarza et al. 2020; Lochner & Bassett 2021; Doorenbos et al. 2021; Ishida et al. 2021; Vafaei Sadr et al. 2021), we attempted to remain true to the premise that a *true novelty* is something that fundamentally cannot be predicted. This exercise is conceptually difficult, as by definition we do not know what we are looking for. We can however rely on the completeness of the feature space derived from the survey's data: if all measurable features are exhaustively sampled, anomalies can be detected.

We thus created a series of MAFs and FoMs that measure the completeness in the space of observables derived from LSST data. Completeness to color and magnitude (and their evolution) was probed by measuring the number of observations and time gaps between observations in pairs of different filters and in the same filter, respectively (Section 3). We scaled a survey quality by the survey's sky coverage, choosing to benchmark this component of the metric to a fiducial implementation of LSST, baseline_v1.5, and by the number of objects observed, scaling the footprint itself by the number of stars in each field (Section 5). These metrics were then summed into a single FoM. Finally, since the FoM so far assembled largely relies on the number of observations, an FoM element was needed that considers the *quality* of the observations. For this, we added an FoM_{depth} to measure LSST images' magnitude depth, penalizing, for example, OpSims that include short-exposure observations if these take time from high-quality, deep observations (Section 4). Proper-motion considerations are reserved for Paper II.

While the main purpose of this paper is to conceptualize a nonparametric way to explore a survey's potential for anomaly detection, these considerations will ultimately need to be applied to current and future Rubin LSST candidate strategies. To illustrate how this can be done, we performed the comparisons for recent suites of simulations.

We identified some high-performing families within OpSim v1.5 and justified their high rank as measured by our FoM (Sections 6.1 and 6.2). Generally, families of OpSim that maximize the diversity of the observations (in terms of time gaps, footprint, and exposure time) seem to be preferred, but there is considerable variation within each family.

To first order, as expected, the mini-surveys seem to be led by footprint considerations. Since fundamentally the allocation of observations to mini-survey regions is a zero-sum game, we point out here that there are high-performing OpSims for the mini-surveys that do not dramatically impact the science obtained in the main survey—so allocating a modest number of exposures to the mini-surveys does not seriously impact the scientific goals of the main survey.

We briefly inspected the most recent (at the time of writing) versions of the OpSim, OpSim v1.7 and OpSim v1.7.1, and found that their performance is impacted, in general, by collecting exposures in two snapshots (2×15 s versus 1×30 s). The loss in survey efficiency when collecting exposures in two snaps is understood and expected; however, the two snaps may be necessary due to image-quality-related considerations, limiting saturation, etc.[29] While we see the effects of increased survey efficiency in our metrics when a 1×30 s strategy is used (OpSim v1.5), it should be emphasized that none of our metrics include considerations on the impact of this choice on image quality or on the capability to open up *intravisit* timescales by treating the two exposures in a visit separately. Combining the impact of Rubin's LSST data volume with visibility into short timescales is potentially transformational for rare phenomena (e.g., relativistic explosions; see Figure 2). We note, however, that any analysis based on the individual snaps that make the 30 s exposure would require custom pipelines.

Even correcting for this loss in efficiency, some families of OpSim v1.5 simulations are the best performers for our science case: namely those that provide visibility into additional timescales by adding short exposures to the observing plan, but planning them when long exposures are unfeasible, so that they do not come at the cost of an overall loss of survey depth (e.g., twilight). We point out that any extension of the feature space is advantageous to the discovery of true novelties, and thus we are not bound to the minimum allocation of short exposures required for other goals (such as

[29] See the following discussion in the Rubin Community public forum: https://community.lsst.org/t/scientific-impact-of-moving-from-2-snaps-to-a-single-exposure/3266.

cross-calibration of LSST to external catalogs with brighter saturation limits; e.g., Gizis 2019).

A comprehensive discussion of the detailed reasons why a specific OpSim achieves a certain performance is beyond the scope of this paper. We encourage instead the use of our metrics to evaluate existing and new OpSims to implement an LSST survey that maximizes the throughput of Rubin Observatory in its four science pillars with particular care given to the discovery novelties, which has the potential to advance or transform all of these fields.

The code on which this analysis is based is available in its entirety in a dedicated GitHub repository.[30]

This work was supported by the Preparing for Astrophysics with LSST Program, funded by the Heising Simons Foundation through grant 2021-2975, and administered by Las Cumbres Observatory.

This paper was created in the nursery of the Vera C. Rubin Legacy Survey of Space Time Science Collaborations[31] and particularly of the Transient and Variable Star Science Collaboration[32] (TVS SC) and Stars, Milky Way, and Local Volume Science Collaboration[33] (SMWLV SC). The authors acknowledge the support of the Vera C. Rubin Legacy Survey of Space and Time TVS SC and SMWLV SC that provided opportunities for collaboration and exchange of ideas and knowledge.

The authors are thankful for the support provided by the Vera C. Rubin Observatory MAF team in the creation and implementation of MAFs.

The authors acknowledge the support of the LSST Corporations that enabled the organization of many workshops and hackathons throughout the cadence optimization process through private fundraising.

The authors thank Dr. Edward Ajhar, who emphasized the importance of an evaluation of the effectiveness of the Rubin survey strategy in the discovery of unknown phenomena at the 2019 LSST (Rubin) Project Community Workshop.

We used the following software packages:

Software: We made use of python including the following libraries: numpy (Harris et al. 2020), maplotlib (Hunter 2007), scikit-learn (Pedregosa et al. 2011), pandas (McKinney 2010), seaborn (Waskom 2021); the glasbey package to generate maximally separable colors. (Glasbey et al. 2007); rsmf (right-size my figures);[34] d3.js[35] to create spatial selection tools and interactive radar/parallel plots.

Appendix A
Sensitivity to Specific Choices in the Definitions of Metrics and FoMs

In Section 3 we measured the properties of the distribution of observations within an OpSim. We settled on a log-uniform distribution between seconds and 1.5 hr in time to be the designated ideal distribution of time gaps to give more weight to short time gaps that can measure color even if the energy of a transient changes rapidly. However, we recognize that while this choice supports our reasoning for the properties of observations that enable color and flux-change measurements, it is not necessarily "ideal," nor it is the only distribution that we can choose that would reward the desired observational properties we have identified. We are, however, relatively insensitive to the choice of this distribution: changing from log-uniform, to linear, to log-normal, to just counting the number of visits within 1.5 hr, most OpSims do not change rank significantly (most ranks change by fewer than 10 places). The exceptions are some OpSims in the short and twilight families that generally benefit from a distribution that emphasizes short time gaps such as the log-uniform. In addition, the var family shows a significant sensitivity to this choice, dropping over 30 ranks when its effectiveness is measured by the number of visits. This single-OpSim family experiments with variable exposure time for every visit in the range 20–100 s to maintain the single-image depth constant. However, the Rubin LSST image processing pipeline is not set up to process images collected with this scheme. Therefore, the OpSim is provided as an interesting exploratory exercise with no expectation that such a strategy could eventually be adopted.[36] See Figure A1 for more details.

In the $\mathrm{FoM}_{\mathrm{tGaps}}$, we had set a tight 1.5 hr constraint on the time between observations in different filters to provide an unambiguous measure of color (based on the evolution timescales of known fast transients as assessed in Bianco et al. 2019). In Section 5, however, we relaxed this requirement to a two-day window for $\mathrm{FoM}_{\mathrm{EG}}$ and $\mathrm{FoM}_{\mathrm{Gal}}$. While ideal to measure color, a 1.5 hr limit competes with significant observational and technical constraints that limit the filters to be used based on lunar illumination and mechanical system constraints on filter changes (Bianco et al. 2022). Yet, observations with filters broadly spaced in wavelength are still valuable to constrain the physical properties (e.g., temperature) of a transient and, if time evolution can be inferred in combination with measurements in multiple filters, color can be constrained. When evaluating the footprint of an OpSim, we do not want to be overly prescriptive (or we would risk having too few observations to effectively measure differences between OpSims' footprints). Our results are not, however, insensitive to this choice: while we report the final footprint metric value setting a maximum time gap of two days, to measure our sensitivity to the specific time gaps between observations counted to build the present FoM, we compare the results of the footprint-color component of $\mathrm{FoM}_{\mathrm{EG}}$ with results that would have been obtained with a 1.5 hr time gap. We find that: (1) the overall value of the metric is significantly smaller for all filter pairs for 1.5 hr than two days; this is explained by the fact that within the 1.5 hr range some filter combinations are going to be unavailable, but this fact does not affect the metric ranking; and, (2) most of the metrics are ranked similarly with the following exceptions (six OpSims change rank quartile):

1. alt_dust and alt_roll_mod2_dust_sdf_0.20 rise in rank from the bottom quartile to the top and second quartiles, respectively;

[30] https://github.com/fedhere/LSSTunknowns
[31] https://www.lsstcorporation.org/science-collaborations
[32] https://lsst-tvssc.github.io/
[33] https://milkyway.science.lsst.org/
[34] https://github.com/johannesjmeyer/rsmf
[35] https://d3js.org
[36] See https://community.lsst.org/t/fbs-1-5-release-may-update-bonus-fbs-1-5-release/4139.

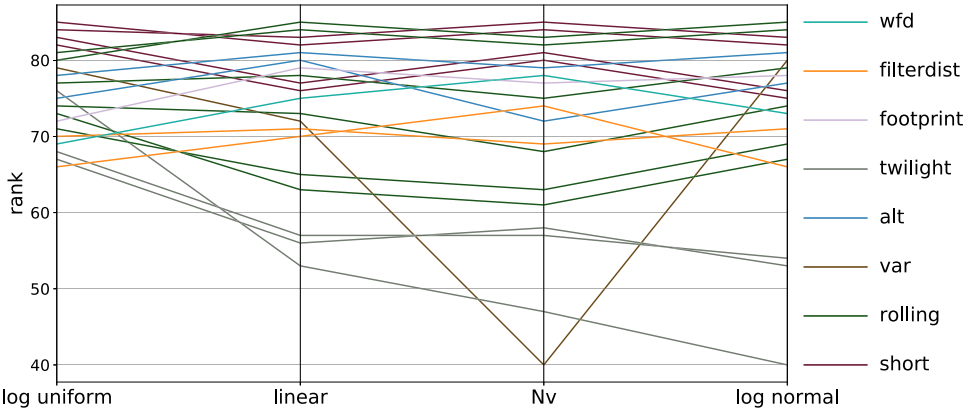


Figure A1. Parallel coordinate plots showing the change in rank for the top 15 OpSims when using a different "ideal" distribution for time gaps between observations in different filters (leading to color measurements). The OpSims are color-coded by family, as indicated in the legend. This figure is described in Appendix A.

2. filterdist_indx1 and filterdist_indx4 drop from the second and first to the third quartile, respectively. Other OpSims within the filtdist family also drop rank, but less significantly.
3. twilight_neo_mod1 and twilight_neo_mod2 drop from the second and third to the fourth and last quartiles, respectively; twilight_neo_mod3 also changes rank significantly but within the third quartile.

Altogether, 24 OpSims (35%) change rank by more than 10 spots. Figure A2 shows the rank changes for all OpSims in OpSim v1.5, with the OpSims that change rank quartile highlighted in red.

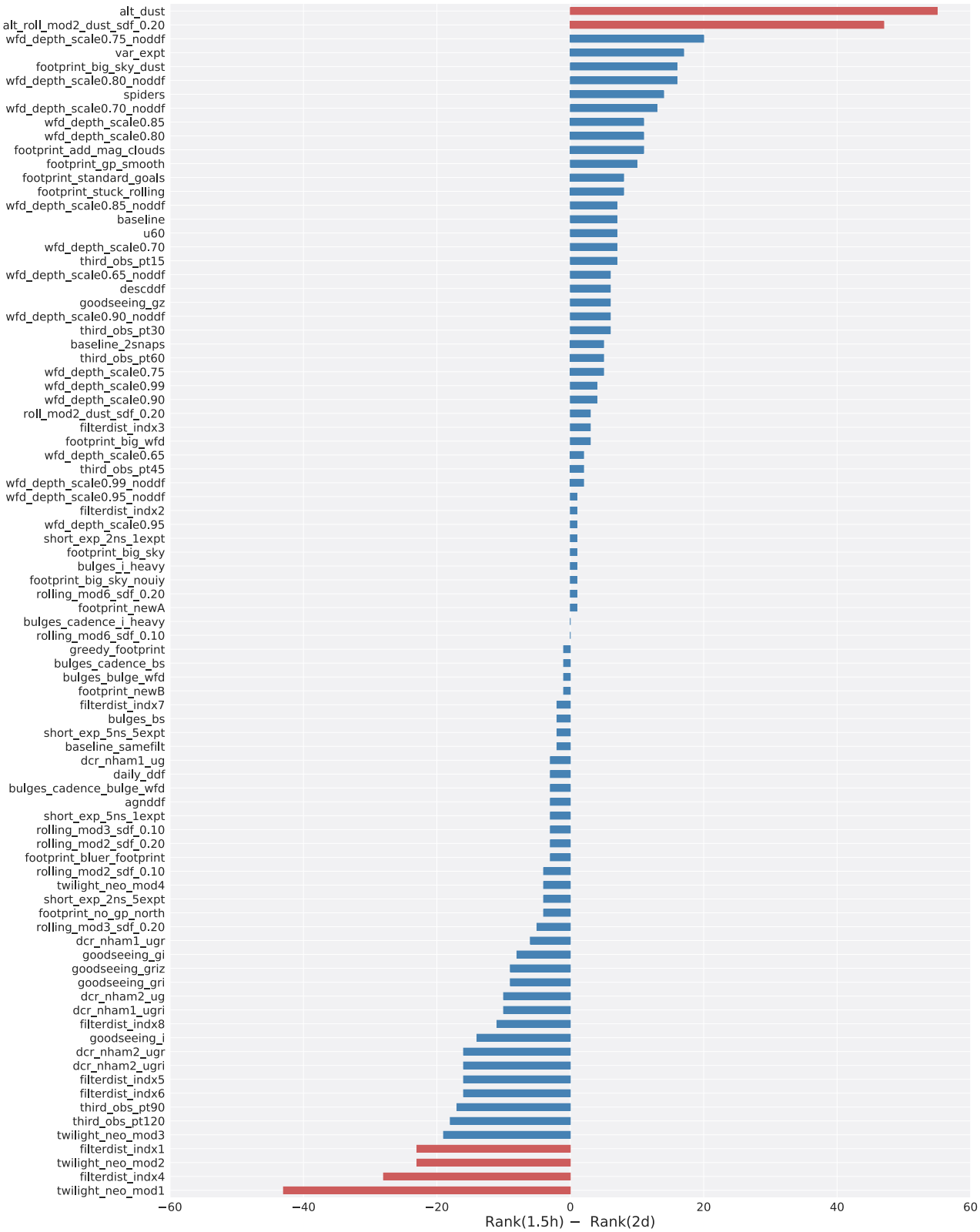


Figure A2. Rank changes after modifying the maximum time gap from two days to 1.5 hr in the $\mathrm{FoM}_{\mathrm{EG}}$ (see Section 5). `OpSims` that changed rank quartile are marked in red. This figure is described in Appendix A.

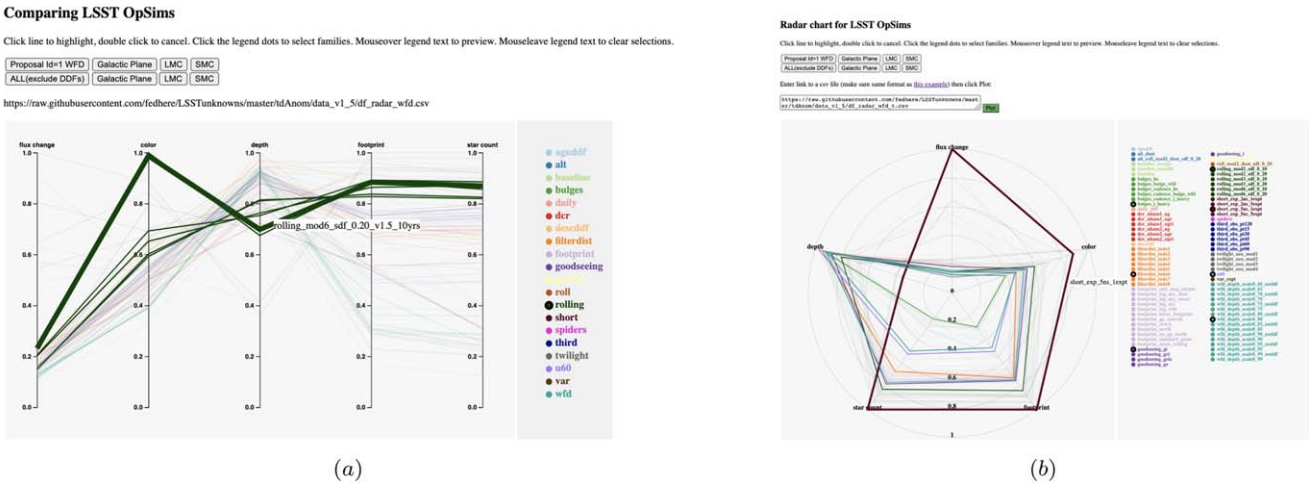
Appendix B
Interactive Tools

We provide three interactive tools that support the analysis performed in this work.

We make javascript-D3 (Bostock et al. 2011) interactive versions of two synoptic visualizations of the results of our FoM available: a parallel coordinate plot and a radar plot.

The parallel coordinate plot[37] (Figure B1(a)) allows the user to follow the performance of an `OpSim` across the components of the FoM. Toggling between families of `OpSim`s to highlight particular `OpSim`s of interest, while keeping all other `OpSim`s in the background, the user can easily identify "standout" `OpSim`s by FoM element. By selecting the "cumulative" option, the viewer can follow the evolution of an `OpSim` across components of the FoM while retaining information about the overall performance.

The radar plot[38] (Figure B1(b), also discussed in Section 6) is a synoptic visualization that maps multiple elements of an FoM to a polygon, with the distance from the center of the polygon representing the result of the FoM element. It allows the user to visualize tension between components of the FoM as well as the overall quality of an `OpSim`, which maps roughly to the area of the polygon (the reader is advised, however, that the area of the polygon slice will vary if the order of the metrics is modified. Thus, the area of the slice is an intuitive, but not a quantitatively accurate proxy for metric



(a) (b)

Figure B1. Snapshots of the interactive versions of two kinds of summary plots: a parallel coordinate plot (a) and a radar plot (b). These interactive visualizations are made available to the reader to explore our metrics or load their own. See Appendix B.

[37] https://xiaolng.github.io/widgets/parallel.html

[38] https://xiaolng.github.io/widgets/radar.html

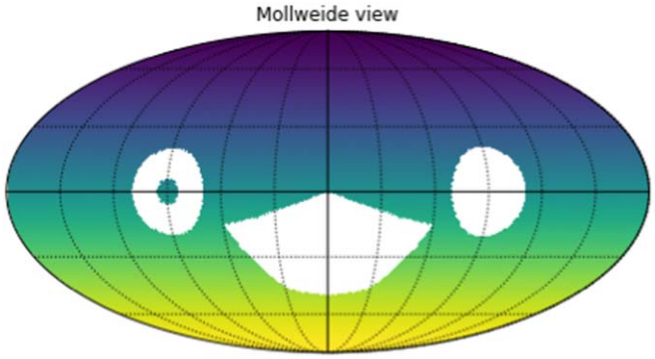


Figure B2. Healpixel-based selection of sky areas performed with our widget: convex, concave, and hollow regions can be selected on multiple spatial projections (Mollweide shown). See Appendix B.

quality). It is however hard to include many `OpSims` in the same radar plot without compromising readability. This widget allows the reader to toggle between `OpSims`, which are color-coded by family.

In both widgets, it is also possible to select the survey or sky area that the user wants to inspect (e.g., WFD, LMC, etc.; see Section 2.3).

While both widgets come pre-loaded with the metrics developed in this paper, the user can easily visualize their own metrics by uploading a comma-separated-value format file with the result of the `MAF`s containing the following columns: `db` (the database name), m_1 (numerical value for the first element of the metric), m_2 (numerical value for the second element of the metric), ... , m_n (numerical value for the last element of the metric).

We offer a `python`-based widget to select regions of sky (see Figure B2 for an example) based on a specific pixelization (e.g., healpix), which was used to select the GP, LMC, and SMC regions in Section 6.2. This tool is available in a dedicated `GitHub` repository[39] as a `jupyter` notebook and interactive webtool (Li 2021).

ORCID iDs

Xiaolong Li https://orcid.org/0000-0002-0514-5650
Fabio Ragosta https://orcid.org/0000-0003-2132-3610
William I. Clarkson https://orcid.org/0000-0002-2577-8885
Federica B. Bianco https://orcid.org/0000-0003-1953-8727

References

Aleo, P. D., Ishida, E. E. O., Kornilov, M., et al. 2020, RNAAS, 4, 112
Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530
Bellm, E., Burke, C. J., Coughlin, M. W., et al. 2021, arXiv:2110.02314
Bianco, F., Ivezić, Ž., Jones, L., et al. 2022, ApJS, 258, 1
Bianco, F. B., Drout, M. R., Graham, M. L., et al. 2019, PASP, 131, 068002
Bishop, C. M. 2006, Pattern Recognition and Machine Learning Information Science and Statistics (Berlin: Springer)
Bostock, M., Ogievetsky, V., & Heer, J. 2011, IEEE Trans. Vis. Comput. Graph., 17, 2301
Boyajian, T. S., Alonso, R., Ammerman, A., et al. 2018, ApJL, 853, L8
Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Comput. Surv., 41, 1
Claver, C. F., Selvy, B. M., Angeli, G., et al. 2014, Proc. SPIE, 9150, 91500M
Connolly, A. J., Angeli, G. Z., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9150, 915014
Delgado, F., & Reuter, M. A. 2016, Proc. SPIE, 9910, 991013
Delgado, F., Saha, A., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9150, 915015
Doorenbos, L., Cavuoti, S., Brescia, M., D'Isanto, A., & Longo, G. 2021, in Intelligent Astrophysics, ed. I. Zelinka et al. (Cham: Springer)
Drout, M. R., Chornock, R., Soderberg, A. M., et al. 2014, ApJ, 794, 23
Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E., & da Costa, L. 2005, A&A, 436, 895
Gizis, J. 2019, Calibrating Milky Way Maps: An LSST Bright(ish) Star Survey, https://docushare.lsstcorp.org/docushare/dsweb/Get/Document-30579/gizis_brightstar_minisurvey
Glasbey, C., van der Heijden, G., Toh, V. F., & Gray, A. 2007, Color Res. Appl., 32, 304
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Graham, M. J., Kulkarni, S., Bellm, E. C., et al. 2019, PASP, 131, 078001
Guy, J., Sullivan, M., Conley, A., et al. 2010, A&A, 523, A7
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Berlin: Springer)
Hunter, J. D. 2007, CSE, 9, 90
Inserra, C. 2019, NatAs, 3, 697
Ishida, E. E. O., Kornilov, M. V., Malanchev, K. L., et al. 2021, A&A, 650, A195
Ivezić, Ž. & the LSST Science Collaboration 2013, LSST Science Requirements Document, http://ls.st/LPM-17
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Johnson, M. A. C., Gandhi, P., Chapman, A. P., et al. 2019 MNRAS, 484, 19
Jones, R. L., Yoachim, P., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9149, 91490B
Kahn, S. M., Kurita, N., Gilmore, K., et al. 2010, Proc. SPIE, 7735 77350J
Kochanek, C. S., Shappee, B. J., Stanek, K. Z., et al. 2017, PASP, 129, 104502
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79
Li, X. 2021, xiaolng/healpixSelector Zenodo, doi:10.5281/zenodo.4914714
Lintott, C. J., Schawinski, K., Keel, W., et al. 2009, MNRAS, 399, 129
Lochner, M., & Bassett, B. A. 2021, A&C, 36, 100481
LSST Science Collaboration, Marshall, P., Anguita, T., et al. 2017, arXiv:1708.04058
Lund, M. B., Pepper, J. A., Shporer, A., & Stassun, K. G. 2018, arXiv:1809.10900
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. vanderWalt & J. Millman, 56
Margutti, R., Cowperthwaite, P., Doctor, Z., et al. 2018, arXiv:1812.04051
Martínez-Galarza, J. R., Bianco, F., Crake, D., et al. 2020, arXiv:2009.06760
Micheli, M., Farnocchia, D., Meech, K. J., et al. 2018, Natur, 559, 223
Naghib, E., Yoachim, P., Vanderbei, R. J., Connolly, A. J., & Jones, R. L. 2019, AJ, 157, 151
Olsen, K., Di Criscienzo, M., Jones, R. L., et al. 2018a, arXiv:1812.02204
Olsen, K., Szkody, P., Cioni, M.-R., et al. 2018b, Mapping the Periphery and Variability of the Magellanic Clouds, https://docushare.lsstcorp.org/docushare/dsweb/Get/Document-30645/olsen_mc_mini.pdf
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pruzhinskaya, M. V., Malanchev, K. L., Kornilov, M. V., et al. 2019, MNRAS, 489, 3591
Soraisam, M. D., Saha, A., Matheson, T., et al. 2020, ApJ, 892, 112
Storey-Fisher, K., Huertas-Company, M., Ramachandra, N., et al. 2021, MNRAS, 508, 2946
Street, R. A., Bachelet, E., Tsapras, Y., et al. 2021, LSST Survey Footprint in the Galactic Plane and Magellanic Clouds, https://docushare.lsst.org/docushare/dsweb/Get/Document-37639/Galactic_Plane_Footprint.pdf
Sultani, W., Chen, C., & Shah, M. 2018, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (Piscataway, NJ: IEEE), 6479
Vafaei Sadr, A., Bassett, B. A., & Kunz, M. 2021, Neural Comput. Appl., doi:10.1007/s00521-021-05839-5
Waskom, M. L. 2021, JOSS, 6, 3021
Yang, H., Xie, F., & Lu, Y. 2006, in Fuzzy Systems and Knowledge Discovery, ed. L. Wang et al. (Berlin: Springer)
Yoachim, P., Coughlin, M., Angeli, G. Z., et al. 2016, Proc. SPIE, 9910, 99101A

[39] https://github.com/xiaolng/healpixSelector