

Contents

1. Machine Learning for Complex Instrument Design and Optimization	1
1.1 Abstract	1
1.2 Introduction	2
1.3 Machine Learning Challenges and Requirements	4
1.4 Optimizing Instrument Operational Performance	7
1.4.1 Effect of noise transients on gravitational-wave searches	9
1.4.2 Effect of control failures on gravitational-wave detector duty cycle	12
1.4.3 Public Datasets	14
1.5 Optimizing Instrument Design	15
1.5.1 Current simulation & optimization tools	15
1.5.2 Generative models for instrument simulation	16
1.5.3 Public Datasets	18
1.6 Conclusion	18
<i>Bibliography</i>	23

Chapter 1

Machine Learning for Complex Instrument Design and Optimization

Barry C. Barish, Jonathan Richardson, Evangelos E. Papalexakis,
Rutuja Gurav (University of California, Riverside)



Fig. 1.1 Inside the control room of the Laser Interferometer Gravitational Observatory (LIGO) in Livingston, Louisiana. Image credit: Amber Stuver/Wikimedia Commons [URL].

1.1 Abstract

In modern experimental physics, particle accelerators and gravitational-wave observatories enable a wide-range of research at the frontiers of science. These instruments are highly complex consisting of many interacting systems which can face significant operational challenges. Apart from the experiment's main data product, a lot of data about the experimental apparatus and its environment is recorded. Machine learning techniques can analyze this big data at scale and find useful insights into operational faults potentially improving the instrument's performance and achieving

the design goals. Speaking of design, machine learning can also accelerate/augment the expensive physics simulations used during the design phase of such large-scale instruments.

1.2 Introduction

In the era of large-scale scientific experiments, big data management and high performance computing have become indispensable for the fast storage, retrieval, and analysis of the vast amounts of data generated. At the same time, there has been a growing interest in using advanced machine learning (ML) techniques for data analysis to make scientific discoveries. However, the potential of ML to accelerate scientific discovery is not limited to the analysis of an experiment's main data products. Frontier experimental apparatuses like the Large Hadron Collider (LHC), Laser Interferometer Gravitational-Wave Observatory (LIGO) and Electron-Ion Collider (EIC) are highly complex instruments with hundreds of degrees of freedom stabilized by cross-coupled feedback servos and with thousands of auxiliary sensors. The 4 km long detectors of LIGO, for example, consist of six coupled laser cavities formed by dozens of mirrors suspended from quadruple-stage pendula and mounted on active seismic isolation platforms. LIGO's main data product is the *strain*¹ which is used by astrophysicists to search for gravitational-wave signals. In addition to the main strain channel, each LIGO detector records over 10,000 channels monitoring the operation of each subsystem and the seismic, acoustic, and electromagnetic environment (for an overview of LIGO's environmental monitoring system, see [25]). The complexity of large scientific instruments, with their vast quantities of auxiliary data, presents an emerging opportunity to use ML more broadly to learn about the instrument itself.

Leveraging ML tools to diagnose performance limitations, such as poorly understood instrumental noise or control instabilities, and to identify more optimal designs could lead to big scientific returns through improved sensitivity, operational up-time, and data quality. As one example, nonlinear, or nonstationary, noise couplings of mostly unknown origin now limit the Advanced LIGO detectors in several ways. In particular, most of the detector noise in a band of key scientific interest, 20–60 Hz, remains unidentified altogether. The unidentified excess noise is shown in the left panel of Fig-

¹*Strain* is the fractional space change across a 4 km long arm of the interferometer relative to the total length of the arm.

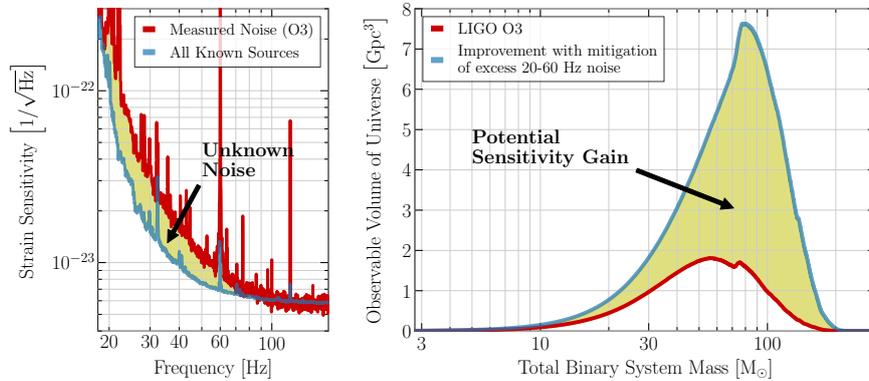


Fig. 1.2 *Left:* Noise floor of the LIGO Hanford gravitational-wave detector during the O3 observing run (red), compared to the budgeted detector noise (blue). As shown, most of the noise in the 20–60 Hz band remains unidentified. *Right:* Projected improvement in the observable volume of Universe for equal-mass binary black hole mergers, as a function of the total mass (in the source frame) of the binary, if this excess noise were identified and fully mitigated.

ure 1.2, indicated by the shaded region. The right panel of Figure 1.2 show the impact of this noise on a key astrophysical metric. As shown, identifying and mitigating this noise would significantly enlarge both the volume of universe and the astrophysical mass range accessible to gravitational-wave science, including enabling observations of mergers of the most massive stellar-origin black holes.

The unidentified excess noise in the LIGO detectors is believed to originate through nonlinear mechanisms because it does not exhibit high coherence with any of the multitude of auxiliary signals. For this reason, it has proven very difficult to pinpoint the origins of nonlinear noise in Advanced LIGO using traditional (linear time-invariant) system identification techniques. However, such challenges are far from being specific to LIGO. In fact, they are common to complex actively controlled instruments, with another example being particle accelerators. In accelerators, ML can be used to optimize the performance of colliding beams, reduce beam losses, and steer the beam.

Developing new tools and approaches for understanding instrumental noise, as well as other operational anomalies, is thus critical to the future of complex instruments such as LIGO, the LHC, and the EIC. New unsupervised ML methods to model and analyze the multitude of auxiliary signals recorded in the detectors, over extended periods of time, offer an emerg-

ing opportunity to generate *actionable insights* that will guide instrument commissioning efforts and future design decisions. Previous applications of ML in complex systems like LIGO [for a through review, see] [Cuoco:2020 or accelerators [7], overwhelmingly employ a supervised learning paradigm, where the desired input-output behavior of a system is known and is being approximated by a ML model (e.g., estimating the masses of the two colliding black holes in an observed gravitational wave). In many of the problems encountered within complex instrument operations, such supervision is scarce or non-existent, calling for the development of novel ML methods to tackle them. Similar challenges also arise in the operation of large complex systems in general [18], so advances in this area will open opportunities for unsupervised and explainable knowledge discovery at-large.

In the remainder of this chapter, we first discuss the key challenges and requirements for developing unsupervised learning models for complex instruments. This section also introduces ML terminology and concepts on which the later sections of this chapter will rely. We then discuss the potential of ML as a tool for diagnosing and optimizing the operational performance of an experimental apparatus, using the LIGO detectors as an illustrative case study. Finally, we discuss the prospect of using ML for instrument design, with applications to optimizing existing accelerators like the LHC as well as the ground-up design of future frontier accelerators such as the EIC².

1.3 Machine Learning Challenges and Requirements

In adapting or developing new ML methods for complex instrument operations management and design, there are a number of fundamental challenges that must be addressed. Below we provide a brief overview of those challenges. Although they are presented in a serial and independent manner, typically a combination of them manifests in any problem we may encounter.

Weak or limited supervision Supervision is a key aspect of machine learning, and the paradigm of *supervised learning* is, most likely, the one to which most readers would have already been exposed: there is an abundance of (input data, desired output) pairs (e.g., an image and its associated label), and the task is to learn a model that maps a representation

²Find more information about the EIC here - <https://www.bnl.gov/eic/>

of the input to the desired label in a manner that *generalizes* to unseen inputs. However, a vast number of problems in the context of instrument diagnostics and design are not afforded the amount of human annotations necessary to train such models. For instance, when predicting operational failures, the annotations that would be most desirable would be ones which would characterize the parts of the instrument which are responsible for the failure. Unfortunately, this presents a *Catch-22*, since those annotations are the knowledge which we seek to extract from our application of machine learning. In such a case, we may resort to *proxy data* which can be obtained cheaply, do not require extensive human annotation, and may help us train models which can shed some light on the ultimate intractable task. We may use the presence or absence of instrument failure as a coarse label that indicates normal or abnormal operation, in hopes that the model that is able to successfully predict such a coarse label is capturing useful information that can enable exploratory analysis to further characterize such a failure. Finally, this challenge is compounded by the fact that, even though the amounts of data generated by an instrument are massive, typically when creating ML-friendly datasets from which a model can learn meaningful and generalizable patterns there is significant cleaning, down-selecting of channels and time periods, and other forms of pre-processing involved, which result in substantially reduced sizes for available curated datasets. This makes learning more challenging.

Explainable models Machine learning models which are tasked with mapping input feature representations (e.g., sensor readings that monitor the state of the instrument for a certain time window) to a desired prediction (e.g., failure at time t_f) can take many different forms, all the way from classification trees and parametric equations to deep neural networks. As we mentioned above, a fundamental requirement for any such model is to generalize well to unseen examples. However, this is only one of the dimensions in which one can examine machine learning models. A dimension that is vital to many problems encountered in the context of this chapter is *explainability*, which measures the degree to which a human can readily understand the reasoning behind the model's decision and attribute such decision to different parts of the input data (in our running example, such attribution could be highlighting different sensors and temporal windows as bearing most of the weight of the prediction of a failure). Simple linear models or tree-based models lend themselves directly to such explanations (a linear model assigns different weights to inputs, and a tree can provide

a set of rules based on the inputs which led to the decision). As machine learning models move towards highly-overparametrized deep neural networks, the ability to generalize successfully increases. However, the ability to provide such explanations and attributions to the input becomes increasingly more challenging. Thus, there is an active research area within machine learning which is concerned with different paradigms and methods of explainability, ranging from readily explainable models to models for which only post-hoc explanations can be provided [13]. In our context, any such form of explanation can be crucially important: coupled with the challenge of weak supervision described above, imbuing explainability to a model tasked with predicting an auxiliary target (e.g., whether a failure occurred) can shine light on parts of the data which can then be further examined systematically by domain experts. In this way, we may gain insight into potential physical mechanisms that can be responsible for the observed outcome (e.g., instrument failure).

Theory/physics-guided models A major promise of modern deep learning models is their purported ability to capture arbitrarily non-linear functions between input and output. However, such a statement, generic as it is, highly depends on a vast number of factors and problem parameters, such as the amount and quality of data available to us, the inherent hardness of the predictive task at hand, and the quality of the supervision, just to name a few. As a result, this promise is not always realized, and in practice successfully training a model that reaches that ideal requires copious amounts of experimentation with different hyperparameters and design choices. Thus, an emerging area in machine learning research is the one of theory-guided, or *physics-guided*, modeling [16]. It aims to leverage well-understood information from a particular scientific domain of interest, incorporating that in the design or the training of the model such that models which are consistent with what theory dictates are preferred over models that are inconsistent (with a varying degree of strictness, depending on how such guidance is incorporated). Doing so may imbue stability and efficiency in how the model is able to perform and generalize, thus making such guided approaches very appealing. In our scenarios, one may envision identifying parts of the instrument whose operation is well-understood and potentially expressed by precise equations, and enforcing consistency of the trained model with respect to those equations. The intended by-product of this action is that the model learned will capture phenomena (both documented and unknown) in the data with higher fidelity, even though full

a-priori modeling of all phenomena may be infeasible or impossible.

Human (expert)-in-the-loop & active learning The typical view of supervision in machine learning considers the process of obtaining human annotations as an offline, slow, and tedious process. However, an emerging trend in machine learning is to blur the boundaries between those two processes by introducing “human-in-the-loop” frameworks, where a model is being continually refined according to human feedback. Key challenges in doing so include the determination of when and how often to solicit feedback, the form of the feedback itself (e.g., it may be easier for a human labeler to identify whether or not two data points are of the same type than to provide a type for a single data point [20]), and in what ways the model should adapt to the new feedback provided by the human. Human-in-the-loop analytics and learning, albeit an active and emerging field, is related closely with ideas previously developed within the field of information retrieval (studying and developing algorithms for efficient search and retrieval of information) and the area of relevance feedback [29]. Furthermore, parts of the desiderata in a human-in-the-loop framework are also the subject of the area of active learning [30], where the objective is to solicit new annotations for a limited number of data points such that the new model’s performance is maximized. Active learning has found applications to science problems, such as anomaly detection in astronomy [15]. In realizing an end-to-end human-in-the-loop framework, however, more advances are necessary in a number of areas including scalability and visualization/human-computer interaction [8, 37, 35].

1.4 Optimizing Instrument Operational Performance

Optimal instrument performance is critical for achieving the scientific goals of a large-scale experiment. However, the complexity of large instruments often makes it difficult to identify the *root cause* of errors encountered during operation. Normal operation typically involves controlling many coupled degrees of freedom simultaneously, with a multitude of potential points of failure and entry for noise. Using instrumental data to identify poor operating conditions and diagnose errors thus requires monitoring many interacting subsystems. Even so, modeling of the instrument, or even individual subsystems, often fails to capture enough realistic detail to reproduce operational errors because (1) the number of possible failure modes is vast and unknown and (2) the state of the instrument and all of

its subcomponents, at any moment, is also not fully known. Thus, the best prospect for uncovering root causes of anomalous operation or failures lies in mining time series from the large number of diagnostic channels³, various subsets of which may be non-trivially correlated, for *interesting* patterns.

Among patterns of interest are the so called anomalies, they are outliers in the data that potentially hint towards processes that might be disrupting nominal operation of the instrument. In [7], the authors present a use-case of modern particle accelerators, a poster child for a large-scale complex instrument and cite that various anomaly detection techniques have been used at the LHC for identifying bad readings from beam position monitors and to assist in automated collimator alignment[34]. A key challenge for human operators is selecting a small subset of relevant channels for investigation from a vast number of channels which record the dynamics of the instrument. In this section, we broadly describe two machine learning pipelines to aid the diagnosis of two types of operational issues - 1. transient noise and 2. control failures. We will first briefly introduce these operational issues and then present detailed real-world examples of these issues encountered in a *state-of-the-art* complex instrument: ground-based gravitational wave detectors.

Transient noise can contaminate the main data product of an experiment and thus lower its quality. For an active instrument like a particle accelerator this means repeating the experimental run which is costly but possible. For a passive observatory, like the ground-based gravitational-wave detectors at LIGO, such transient noise can lead to missing part or whole of a unique astrophysical event that is impossible to observe again. Figure 1.4 shows the now-famous example of a loud noise transient corrupting a portion of the signal from the first ever binary neutron star merger detected by LIGO. Thus, it is critical to understand sources of such noise to potentially eliminate it with upgrades. Using archived data and machine learning methods, we can identify *witnesses* to noise artifacts by looking for correlations between transients in a set of diagnostic channels³. A subset of these diagnostic channels termed as *witness channels* can then be used to categorize, locate and potentially eliminate the *root cause* of a certain noise transients.

³A set of channels used for performing instrument diagnostics, primarily consisting of sensor readout or other quantities related to instrument control.

Control failures render the instrument unable to operate in a nominal state and thus reduces duty-cycle. This operational issue has a relatively larger impact than the **transient noise** issue as no science data can be produced while the instrument is recovering. Diagnosing causes of control failures is crucial to mitigate future adverse events. Such failures are relatively easy to mitigate as they are occurring in real-time but a key task is to *predict* an impending failure. We can address this task with machine learning by using data preceding failure events from a set of diagnostic channels to identify *precursors* to these failure events. In [9], the authors explore the use of a deep neural networks⁴ for predicting operational failures at a particle accelerator facility by modeling precursors to a failure from a set of diagnostic channels³. *Limited Supervision [1.3]*: The machine learning problem is formulated as a binary classification task where data-points corresponding to failures (positive class) and nominal behaviour (negative class) can be automatically labelled without manual human effort. *Explainable Models [1.3]*: Since deep neural networks are often black-box models, the authors employ an explanation technique, Layer-wise Relevance Propagation [2], to highlight a subset of diagnostic channels³ that are relevant to failures.

Before presenting real-world examples of the two aforementioned operational issues in context of our complex instrument of choice, LIGO, let us walk through the generic machine learning pipeline (shown in Figure 1.3 (*top row*)). Raw data from a data archive is processed with a domain expert's guidance to create a *task-specific dataset*. This task-specific dataset is then used to train an appropriate *machine learning model*. The model's performance is tested on a holdout set of data-points that are not part the model's training for *generalization* ability - a key expectation from any machine learning model. Finally, the *validation* of the model output is done using domain expert-defined tests which are often necessary for the downstream application of the model.

1.4.1 *Effect of noise transients on gravitational-wave searches*

For LIGO's online astrophysical search pipelines, one class of nonlinear noise artifact is particularly problematic: Transient noise bursts, of largely unknown origin, known as *noise glitches* [3] [4]. Multiple glitches occur

⁴A class of supervised machine learning models that are particularly good at learning complex, non-linear functions of the input that map it to the prediction target.

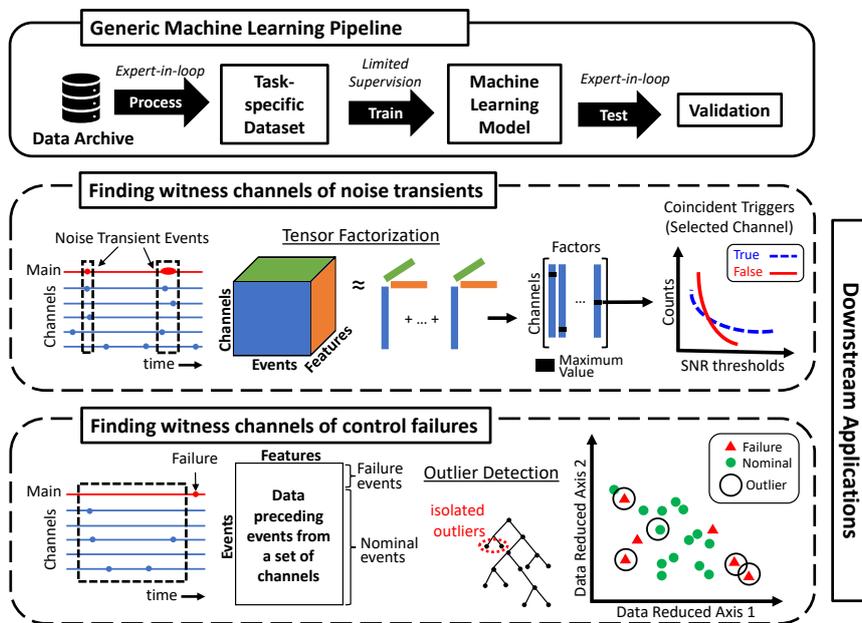


Fig. 1.3 **Top row:** A generic Machine Learning pipeline typically has 4 stages starting at task-specific dataset creation to validation of the modeling output. **Middle row:** End-to-end pipeline for finding witnesses to noise transients that can potentially help diagnose sources of noise relies on selecting transient events of interest from the *main* channel and finding coincidences in a set of diagnostic channels. **Bottom row:** End-to-end pipeline for finding precursors of control failure events relies on isolating data from a set of witness channels preceding such events as outliers.

during a typical hour of observation. Glitches contaminate the astrophysical data streams (see, for example, Fig. 1.4), confusing burst-source gravitational-wave searches and hindering the timely issuance of real-time alerts for electromagnetic follow-up. By introducing a long non-Gaussian tail in the instrumental noise background, glitches also raise the statistical threshold for detecting true astrophysical events. Reducing the frequency of glitches will improve detection rates for all types of events, but most especially for high-mass binary black hole mergers, whose signal-to-noise ratio is the poorest.

In [14], we use matrix and tensor factorization [19] to automatically identify relevant subsets from the set of diagnostic channels that are believed to be potential witnesses to *glitches* present in LIGO data in a 5-day period from the third observing run. Figure 1.3 (*middle row*) shows an end-

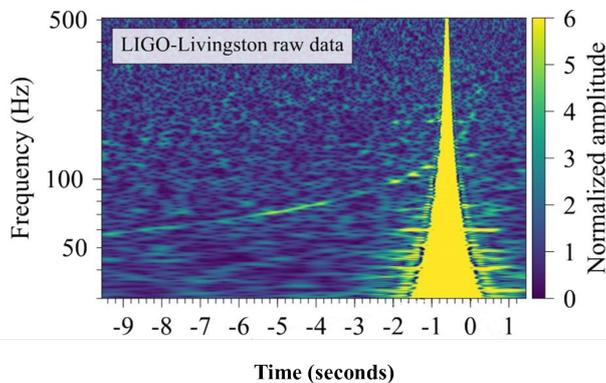


Fig. 1.4 Time-frequency representation of the gravitational-wave event GW170817, as observed by the LIGO Livingston detector. Moments before merger, a noise glitch 1,000 times louder than the gravitational-wave signal occurred. Image reproduced courtesy of LIGO/Caltech/MIT.

to-end pipeline for this task. We select a set of diagnostic channels and a set of *glitches* from LIGO’s main channel where gravitational-waves are observed. We then construct the ***task-specific dataset*** in form of a 3-mode tensor⁵ where mode-1 corresponds to the *glitches* in the main channel, mode-2 corresponds to diagnostic channels selected for this analysis and mode-3 corresponds to features (e.g. duration, peak frequency, bandwidth, signal-to-noise ratio, etc.) of the *glitches* found in the diagnostic channels that are coincident with the main channel *glitches* within a short window around a main channel *glitch*. This tensor essentially encodes the presence or absence of a *glitch* in a diagnostic channel “coincident” with a *glitch* in the main channel. The ***machine learning model*** of choice is tensor factorization. We factorize this tensor into N factors and obtain factor matrices corresponding to the latent space representations of each mode. We use the factor matrix corresponding to the “diagnostic channels” mode to select N channels (one per factor), see Figure 1.5. We ***validate*** the channels thus selected as potential witnesses to the main channel *glitches* in the dataset by examining the true positive rate⁶ and false positive rate⁷ of each selected witness channel for increasing values of signal-to-noise (SNR)

⁵In the context of machine learning, a tensor is a multidimensional array.

⁶Number of times there was a coincident *glitch* in a selected witness channel and the main channel divided by the number of total *glitches* in the witness channel

⁷Number of times there was a *glitch* in a selected witness channel but no coincident *glitch* in the main channel divided by the number of total *glitches* in the witness channel

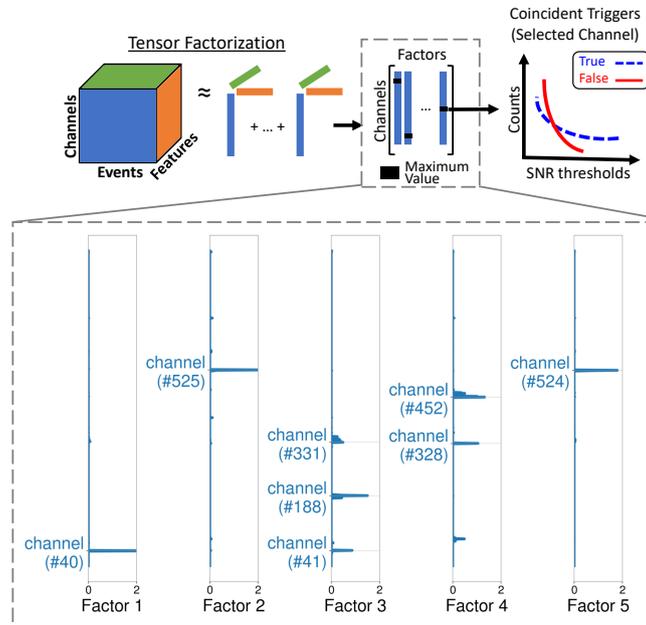


Fig. 1.5 *Finding witness channels to noise transients: Witness Selection* Factors corresponding to diagnostic channels of a $n = 5$ component tensor factorization of $channels \times events \times features$ tensor that encodes presence or absence of coincident noise transients in diagnostic channels. They show high magnitude values in a factor for a few out of the approx. 900 channels used for this analysis. For example, Factor 1 shows a high magnitude value for channel #40 which is then selected as a witness.

thresholds, see Figure 1.6. The channels deemed *good veto candidates* by this validation step can then be considered for downstream applications like, for example, using the channel to veto and remove data segments from the main channel before searching for gravitational-wave signals.

1.4.2 *Effect of control failures on gravitational-wave detector duty cycle*

Achieving resonance in the full LIGO system is a complex process [see, e.g.,] [Staley:2014] typically requiring 30 minutes to complete. Control failure occurs when a disturbance, either internal or external, causes the laser cavities to become “unlocked” from their resonant operating points. This type of control failure is termed as *lock loss*. Due to the frequency of lock losses, all three LIGO-Virgo detectors (LIGO Livingston, LIGO Han-

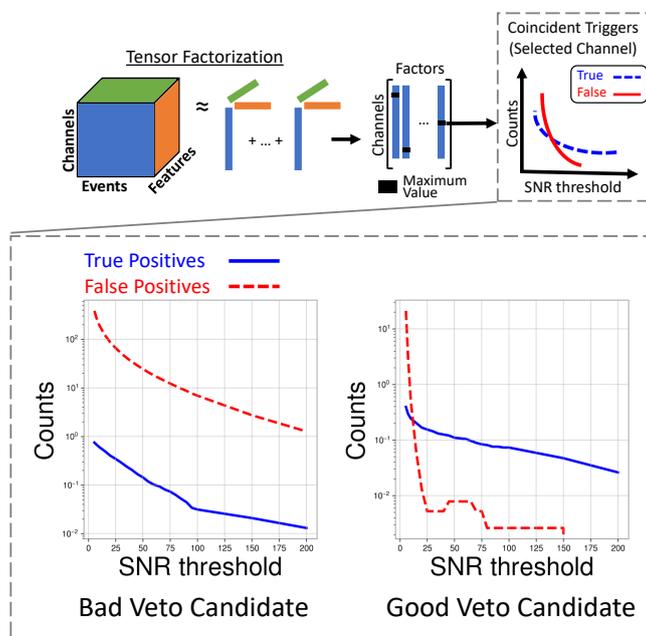


Fig. 1.6 *Finding witness channels of noise transients: Witness Validation* Two different channels selected using factors of the $\text{channels} \times \text{events} \times \text{features}$ tensor show differing true vs false positive characteristics. *Left:* The selected channel has a high false positive rate at increasing SNR thresholds compared to the true positive rate suggesting that it might just be an inherently noisier diagnostic channel and thus cannot be used to veto science data segments for downstream gravitational-wave search pipelines as it will result in large loss of science data. *Right:* There is a sharp drop-off in false positive rate of the selected channel above SNR threshold of 25 while the true positive rate remains relatively stable. We can use this selected channel to veto coincident noise transients in the science data.

ford, and Virgo) were running at the same time only about 50% of the time during the third observing run. The majority of lock losses have no readily identifiable environmental cause Rollins:2017, so may too be triggered by non-Gaussian noise events. Identifying and eliminating triggers of lock loss could potentially make a big improvement in the fraction of calendar time all three detectors are data taking, increasing the number of triple-coincident gravitational-wave detections. Observations by multiple detectors with different antenna patterns are critical for precise sky localization, as needed for targeted electromagnetic follow-up of potential multi-messenger events.

Figure 1.3 (*bottom row*) shows a pipeline that formulates the problem of diagnosing control failures as an anomaly detection task where the ***task-specific dataset*** is constructed by obtaining data preceding control failures from a set of diagnostic channels. We also obtain data during nominal operation periods from these channels. Thus, we do not need any explicit *supervision* as we have automatic *ground-truth (labels)* for the data-points in our dataset as either *failure* or *nominal* events. We hypothesize that data preceding failure events in a subset of channels will have anomalies which might be precursors to the failures. The ***machine learning model*** of choice here is an unsupervised, tree-based Outlier Detection algorithm called Isolation Forest [21] that partitions the dataset using a tree structure and *isolates* anomalies as outliers. We validated the events deemed as anomalies by the algorithm by comparing them against the ground-truth, see Figure 1.7.

There are some key challenges worth highlighting in the machine learning pipelines described in Section 1.3.1 and 1.3.2. Firstly, we assume a preprocessing step of creating the task-specific datasets from the massive amounts of raw, archived time series data. This step has an inevitable need of domain expertise as there can be potentially tens of thousands of diagnostic channels to choose from only a subset of which may be relevant to the task and these channels are often sampled at different rates capturing various phenomenon occurring at different time and frequency scales. Beyond the selection of a set of diagnostic channels for any given analysis, the “*features*” engineered from the raw data-streams can be explicitly hand-crafted to be semantically meaningful to the domain-expert or they can be implicitly learned by the machine learning model. Implicit feature representations of our data, like the ones learned by black-box models like deep neural networks, reduce the burden of hand-crafting good enough features but make the output of our pipelines less interpretable. Moreover, at the end of the pipeline we also seek expert guidance for creating validation tests that examine the utility of the machine learning model’s output for downstream diagnostic applications.

1.4.3 Public Datasets

Most real-world complex instruments have large quantities of raw archival operational data, as described in this section, that is not readily available to be used for training machine learning models. Therefore, machine learning

applications use smaller benchmark datasets curated from the data archive or create synthetic datasets for *proof-of-concept*.

Following are some LIGO-specific public datasets available for machine learning in complex instruments diagnostics.

- (1) **LIGO Gravity Spy:** This dataset, as described in [36], consists of time-frequency spectrogram images of various noise transients in the main data product of LIGO. The machine learning task is to classify these images into a set of classes using state-of-the-art deep learning models like convolutional neural networks. The sample dataset is available here - <https://zenodo.org/record/1476551#.YjoQ-3XMKV5>
- (2) **LIGO diagnostic channels dataset:** This is a multivariate time-series dataset consisting of a 3-hour period centered on a real gravitational-wave event, GW170814, from approx. 1,000 LIGO diagnostic channels. This dataset can be used for identifying noise transients witness channels as described in this section or for subtracting noise from the main data product. The dataset can be found at the Gravitational Wave Open Science Center webpage - <https://www.gw-openscience.org/auxiliary/GW170814/>

1.5 Optimizing Instrument Design

A fundamental requirement in instrument design is the ability to accurately and efficiently simulate different scenarios for a variety of instrument configurations, in pursuit of configurations which optimize a certain target physical quantity. In this section, we first provide a brief overview of the traditional simulation and optimization pipeline, and subsequently we review an emerging machine learning breakthrough which has the potential to revolutionize instrument design.

1.5.1 *Current simulation & optimization tools*

Traditionally, the instrument design pipeline is roughly divided into two parts, simulation, where the behavior of an instrument (or one of its components) is estimated by software, and optimization, where the designer seeks to identify the best combination of parameters that satisfies a set of requirements for the component in question.

Simulation Simulation is the process of estimating the response of an instrument or one of its components through software, thus eliminating the

need to fabricate and physically experiment with that component. Such simulators have been widely used by both accelerators [1] and gravitational wave detectors [10]. Typically, during simulation, the user/designer specifies a thorough description of the component that is being simulated, and provides as input the environmental conditions (e.g., a certain event of interest) for which the component's response is needed. In turn, the simulator computes the response of that component, usually by numerically solving for the equations which govern that component's response. Finally, the last step in the pipeline takes as input the response/state of the component, and outputs a reconstruction of a quantity of interest, which is used as an evaluation metric (e.g., energy).

Optimization During a single run of the simulator, the parameters of the design to be simulated are user-defined. However, manually iterating over different designs in pursuit for one or more designs which satisfy the chosen criteria can be extremely inefficient and incomplete. To that end, in addition to simulators, effective design uses optimizers which are given a user-specified objective to optimize for (e.g., energy) and their goal is to effectively navigate the search space of designs and identify one or more that achieve the optimal objective. Ocelot [32] is such a general-purpose optimization tool which has been very popular in accelerator design. Interestingly, there are connections between optimization tools in instrument design and tools used in modern machine learning to fine-tune the design parameters (also called hyperparameters) of a large model, such as Bayesian Optimization [24, 33].

1.5.2 *Generative models for instrument simulation*

Despite the existence of efficient optimizers, whose goal is to minimize the number of times the simulation must be run, simulation in itself is a very computationally intensive and slow process. Furthermore, in order for a simulation to be accurate, it is typically restricted to very specific components of a larger instrument. As a result, simulating an instrument end-to-end would require the combination of a large number of specialized simulators, which are not necessarily designed to be compatible with each other. Thus, such an endeavor may be technically challenging or even intractable.

On the other hand, machine learning models, and especially deep neural networks, can be used as cheap, generic nonlinear function approximators.

They have been studied for “data analysis” of an instrument’s main data products (e.g., using deep neural networks to quickly estimate astrophysical parameters of a gravitational wave signal in real time, instead of carrying out the exact numerical calculations [11]). However, an emerging trend in instrument design is to replace traditional simulators with deep learning models, specifically leveraging recent advances in generative adversarial networks (GANs) [12]

A GAN is a machine learning model that itself consists of two sub-models, the *generator* and the *discriminator*. Typically those two models are neural network based, so the term “network” is commonly used when referring to them. The generator takes as input a vector, typically drawn at random from a Gaussian distribution, and it is tasked with outputting a data point (e.g., an image or the output of an instrument simulator) that is “realistic.” The discriminator, on the other hand, takes two inputs: a “real” data point (e.g., an actual image in our dataset or an actual output of an instrument simulator) and a generated data point (the output of the generator), and its task is to determine which data point is real and which is fake. During training, those two networks are engaged in a two-player game, where, as the generator gets better at producing realistic outputs, the discriminator improves its ability to detect fake data, and vice-versa. At the end of the training process, the generator will ideally produce data indistinguishable from “real” data, while effectively providing a means of sampling from the distribution of “real” data.

The most well-known use case and success story of GANs is in the generation of images and videos that look extremely realistic [17]. However, the success of GANs extends far beyond this use case, with successful applications in drug discovery [23] and material design [6]. Very recently, there have been successful attempts at using GANs for the simulation of complex scientific instruments, as applied specifically to the simulation of accelerator calorimeters [27, 28, 26]. Figure 1.8 provides an illustrative example. These recent attempts clearly demonstrate the feasibility and potential of harnessing the power of GANs for complex instrument design.

Despite early success on the calorimeter application, broad application of GANs to the problem of instrument simulation and design poses fascinating interdisciplinary research challenges. Those challenges pertain to ensuring that the generated data are of high quality, both as it pertains to how realistic they are as well as to the diversity of generated data points (a known mode of failure for GANs is memorizations of the training set or generation of very similar data points) [31]. Advances in the following

areas, outlined in Sec. 1.3, can pave the way for tackling those challenges:

(a) *Physics-guided models*: In order to ensure that the Generator will obey physical constraints and laws governing the instrument that it models, a physics-guided approach can incorporate this information in the design or training stages of the model. In Paganini et al. [28], the proposed GAN takes such a physics-guided approach, by modifying the loss function that is being optimized during training.

(b) *Limited supervision*: Training data for the GAN come from very expensive simulations. Thus, it may be infeasible to create very large training data sets, a challenge which may amplify issues such as the memorization of the training set by the GAN. Thus, designing solutions that work with limited supervision is imperative.

(c) *Expert-in-the-loop*: A big advantage of applying GANs to instrument simulation is that the definition of “realistic” is objective and can be analytically measured. This is in contrast to the original application of GANs, where judging the realism of a generated face or natural image can be subjective. However, when judging the diversity of generated data, the involvement of human experts can be crucial.

In addition to the immediate efficiency gains of substituting an inefficient simulator with a highly-efficient neural network, a major promise of machine learning driven instrument design is flexibility and modularity: once every component of an instrument can be efficiently simulated by a neural network, this can facilitate the simulation of increasingly larger portions of an instrument, ideally all the way to end-to-end simulation and design.

1.5.3 Public Datasets

Three datasets used in [5] for simulating particle showers using GANs at the ATLAS detector at the LHC are hosted here - <https://opendata.cern.ch/record/15012>

1.6 Conclusion

Instruments at the frontiers of science are becoming increasingly complex and data-intensive. There is a need for intelligent automation to tame this complexity, increase reliability and push the limits of these technological marvels to make new scientific discoveries. This chapter explored the emerging application of machine learning for complex instrument opera-

tions management and design. The current state of the art demonstrates very promising results, with machine learning methods empowering early detection of failures, diagnosing noise sources, and enabling the flexible simulation of different components of an instrument—tasks which may have previously been extremely tedious or downright impossible to do well by hand. There exist a number of fundamental challenges within machine learning research to successfully apply these techniques to solving issues of complex instruments, like the need for explainable and theory-guided modeling. Successful application of these techniques can pave the way for a new generation of complex instruments.

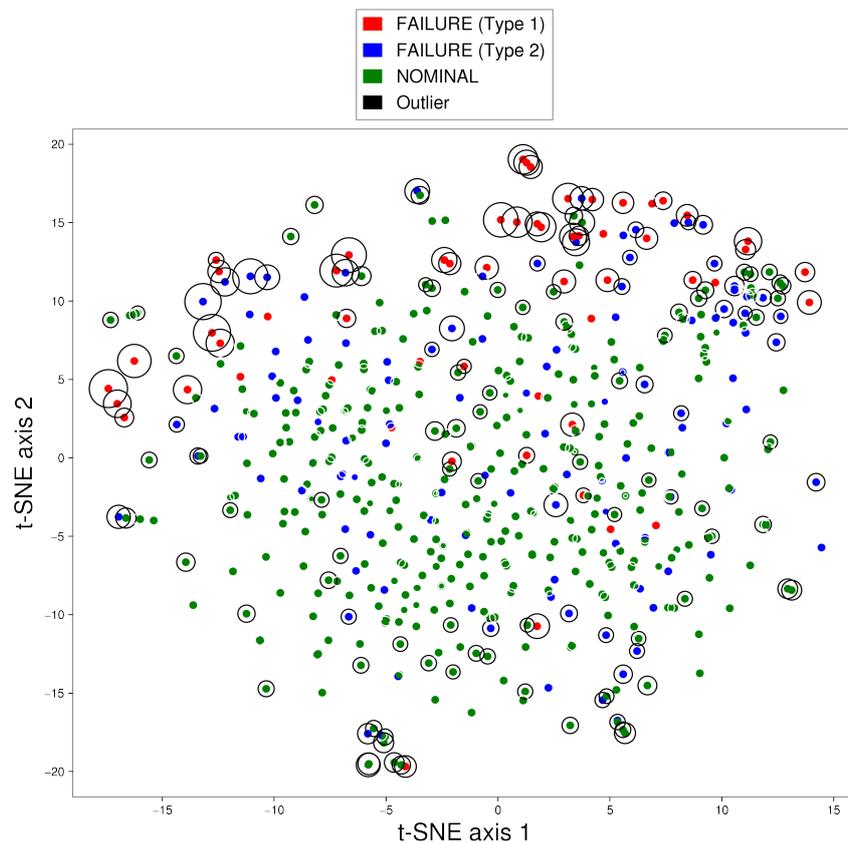


Fig. 1.7 *Finding witness channels of control failures*: (Note: The figure above shows a 2-dimensional projection of the m -dimensional data ($m \gg 2$) using t-Distributed Stochastic Neighbor Embedding (t-SNE) [22] that tries to preserve distances in the higher dimensional space for the purpose of visualization only). We obtain k features from segments of data preceding 2 types of control failures (*red and blue*) from a set of n diagnostic channels ($m = k \times n$). To complete the dataset we also sample data segments during nominal operation (*green*). The outlier detection algorithm is fully unsupervised and does not use the ground-truth labels for the events in the dataset (failure or nominal) to isolate certain events as outliers (*circled in black*). We can see that instances of data preceding control failures (*red, blue*) are identified as outliers. The diameter of the black circle in this plot corresponds to an outlier score (larger score = bigger outlier) and this score can potentially be used as a threshold for reducing false positives (nominal events that are identified as outliers but with a relatively smaller outlier score).

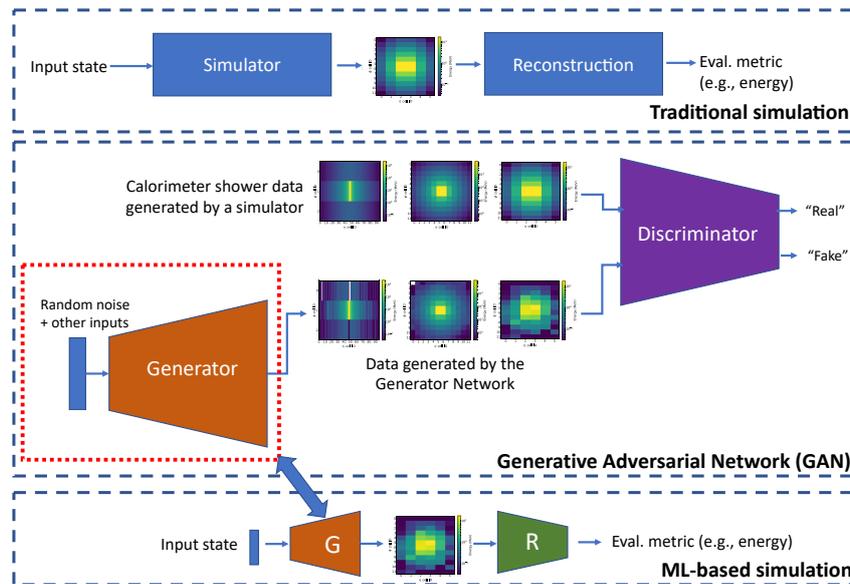


Fig. 1.8 *Top row:* An illustration of the “traditional” simulation pipeline
Middle row: Illustration of a Generative Adversarial Network (GAN): The Generator Network is taking as input random noise and additional information about the state to be simulated and generates “fake” data. The Discriminator Network takes as input “real” data (as simulated by a traditional simulator such as geant4) and “fake” data generated by the Generator, and determines which data point is fake or real.
Bottom row: In the ML-based simulation pipeline, the simulator is replaced by the GAN Generator and a “Reconstruction” neural network is replacing the analytical reconstruction process, which transforms the simulation output to a quantity of interest (e.g., energy). (Data figures show simulated and generated calorimeter showers and are adapted from [28]).

Bibliography

- [1] Sea Agostinelli et al. “GEANT4—a simulation toolkit”. In: *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303.
- [2] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS one* 10.7 (2015), e0130140.
- [3] L Blackburn et al. “The LSC glitch group: monitoring noise transients during the fifth LIGO science run”. In: *Classical and Quantum Gravity* 25.18 (Sept. 2008), p. 184004. DOI: 10.1088/0264-9381/25/18/184004. URL: <https://doi.org/10.1088/0264-9381/25/18/184004>.
- [4] M Cabero et al. “Blip glitches in Advanced LIGO data”. In: *Classical and Quantum Gravity* 36.15 (July 2019), p. 155010. DOI: 10.1088/1361-6382/ab2e14. URL: <https://doi.org/10.1088/1361-6382/ab2e14>.
- [5] ATLAS collaboration et al. “AtlFast3: the next generation of fast simulation in ATLAS”. In: *arXiv preprint arXiv:2109.02551* (2021).
- [6] Yabo Dan et al. “Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials”. In: *npj Computational Materials* 6.1 (2020), pp. 1–7.
- [7] Auralee Edelen et al. “Opportunities in machine learning for particle accelerators”. In: *arXiv preprint arXiv:1811.03172* (2018). URL: <https://arxiv.org/pdf/1811.03172.pdf>.
- [8] Alex Endert et al. “The human is the loop: new directions for visual analytics”. In: *Journal of intelligent information systems* 43.3 (2014), pp. 411–435.
- [9] Lukas Felsberger et al. “Explainable Deep Learning for Fault Prognostics in Complex Systems: A Particle Accelerator Use-Case”. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 139–158.
- [10] Andreas Freise, Daniel Brown, and Charlotte Bond. “Finesse, frequency domain INterferomEter simulation softwarE”. In: *arXiv preprint arXiv:1306.2973* (2013).

- [11] Daniel George and Eliu Antonio Huerta. “Deep Learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data”. In: *Physics Letters B* 778 (2018), pp. 64–70.
- [12] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.
- [13] Riccardo Guidotti et al. “A survey of methods for explaining black box models”. In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.
- [14] Rutuja Gurav et al. “Unsupervised matrix and tensor factorization for LIGO glitch identification using auxiliary channels”. In: *AAAI 2020 Fall Symposium on Physics-Guided AI to Accelerate Scientific Discovery*. 2020.
- [15] Emille EO Ishida et al. “Active Anomaly Detection for time-domain discoveries”. In: *arXiv e-prints* (2019), arXiv-1909.
- [16] Anuj Karpatne et al. “Theory-guided data science: A new paradigm for scientific discovery from data”. In: *IEEE Transactions on knowledge and data engineering* 29.10 (2017), pp. 2318–2331.
- [17] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410.
- [18] Samir Khan and Takehisa Yairi. “A review on the application of deep learning in system health management”. In: *Mechanical Systems and Signal Processing* 107 (2018), pp. 241–265. URL: <https://www.sciencedirect.com/science/article/pii/S0888327017306064>.
- [19] Tamara G Kolda and Brett W Bader. “Tensor decompositions and applications”. In: *SIAM review* 51.3 (2009), pp. 455–500.
- [20] Ramya Korlakai Vinayak and Babak Hassibi. “Crowdsourced clustering: Querying edges vs triangles”. In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 1316–1324.
- [21] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest”. In: *2008 eighth IEEE international conference on data mining*. IEEE. 2008, pp. 413–422.
- [22] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [23] Lukasz Maziarka et al. “Mol-CycleGAN: a generative model for molecular optimization”. In: *Journal of Cheminformatics* 12.1 (2020), pp. 1–18.
- [24] Mitchell McIntire et al. “Bayesian optimization of FEL performance at LCLS”. In: *Proceedings of the 7th International Particle Accelerator Conference*. 2016.
- [25] P Nguyen et al. “Environmental noise in Advanced LIGO detectors”. In: *Classical and Quantum Gravity* 38.14 (June 2021), p. 145001. DOI: 10.1088/1361-6382/ac011a. URL: <https://doi.org/10.1088/1361-6382/ac011a>.
- [26] ATLAS PUB Note. “Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks”. In: (2020).
- [27] Luke de Oliveira, Michela Paganini, and Benjamin Nachman. “Learning particle physics by example: location-aware generative adversarial networks

- for physics synthesis”. In: *Computing and Software for Big Science* 1.1 (2017), pp. 1–24.
- [28] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. “Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters”. In: *Physical review letters* 120.4 (2018), p. 042003.
- [29] Gerard Salton and Chris Buckley. “Improving retrieval performance by relevance feedback”. In: *Journal of the American society for information science* 41.4 (1990), pp. 288–297.
- [30] Burr Settles. “Active learning literature survey”. In: (2009).
- [31] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. “How good is my GAN?” In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 213–229.
- [32] Sergey Tomin et al. “On-line Optimization of European XFEL with OCELOT”. In: *Proc. 16th Int. Conf. on Accelerator and Large Experimental Control Systems (ICALEPCS’17)*. 2017, pp. 1038–1042.
- [33] Ryan Turner et al. “Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020”. In: *arXiv preprint arXiv:2104.10201* (2021).
- [34] Gianluca Valentino et al. “Anomaly detection for beam loss maps in the Large Hadron Collider”. In: *Journal of Physics: Conference Series*. Vol. 874. 1. IOP Publishing. 2017, p. 012002.
- [35] Doris Xin et al. “Accelerating human-in-the-loop machine learning: Challenges and opportunities”. In: *Proceedings of the second workshop on data management for end-to-end machine learning*. 2018, pp. 1–4.
- [36] Michael Zevin et al. “Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science”. In: *Classical and quantum gravity* 34.6 (2017), p. 064003.
- [37] Yutao Zhang et al. “Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.” In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1002–1011.