



What is a data model?

An anatomy of data analysis in high energy physics

Antonis Antoniou¹ 

Received: 18 August 2020 / Accepted: 29 August 2021 / Published online: 25 October 2021
© The Author(s) 2021

Abstract

Many decades ago Patrick Suppes argued rather convincingly that theoretical hypotheses are not confronted with the direct, raw results of an experiment, rather, they are typically compared with models of data. What exactly is a data model however? And how do the interactions of particles at the subatomic scale give rise to the huge volumes of data that are then moulded into a polished data model? The aim of this paper is to answer these questions by presenting a detailed case study of the construction of data models at the LHCb for testing Lepton Flavour Universality in rare decays of B-mesons. The close examination of the scientific practice at the LHCb leads to the following four main conclusions: (i) raw data in their pure form are practically useless for the comparison of experimental results with theory, and processed data are in some cases epistemically more reliable, (ii) real and simulated data are involved in the co-production of the final data model and cannot be easily distinguished, (iii) theory-ladenness emerges at three different levels depending on the scope and the purpose for which background theory guides the overall experimental process and (iv) the overall process of acquiring and analysing data in high energy physics is too complicated to be fully captured by a generic methodological description of the experimental practice.

Keywords Data · Models · Data models · Lepton Flavour Universality tests · LHC · Philosophy of Science in Practice

1 Introduction

The constantly growing integration of science and technology during the last decades has brought science in the new ‘era of big data’. Modern experimental setups and

✉ Antonis Antoniou
aa17779@bristol.ac.uk

¹ Department of Philosophy, University of Bristol, BS6 6JL, Bristol, UK

other advanced methods of data collection often result in enormous datasets calling for more and more sophisticated methods of data analysis in order to enable the comparison of the experimental results with theoretical hypotheses. The well known Hypothetico-Deductive method whereby theoretical hypotheses are reinforced – or, in Popperian terms, corroborated – in the light of new data nicely captures a large part of the scientific practice, however, at the same time it provides an oversimplified and unrealistic picture in which important details are left aside. How exactly are theoretical hypotheses eventually confronted and tested by experimental results given that the latter are often produced in the form of large datasets and in a language that is not accessible to the theory?

Suppes (1962) answered this question by pointing out that theoretical hypotheses are not directly confronted with the raw unprocessed data from experiments, rather, they are only confronted with *models of data*. What exactly is a model of data however? Suppes explains in an earlier work: ‘The maddeningly diverse and complex experience which constitutes an experiment is not the entity which is directly compared with a model of a theory. Drastic assumptions of all sorts are made in reducing the experimental experience [...] to a *simple entity* ready for comparison with a model of the theory’ (1960, p.297, emphasis added). In Suppes’ mind, a data model is a simple entity that incorporates what is often a very complicated and sophisticated experimental process, into a simple result which is eventually compared with the theoretical predictions of a theory or a model.

While Suppes’ remarks on data models sowed the seeds for further significant work on the philosophy of data, the interest of philosophers of science in data was mainly revived by the seminal works of Bogen and Woodward (1988) and Woodward (1989) and van Fraassen (1980, 1989). In response to van Fraassen’s well known view that the empirical adequacy of a theory is measured by its ability to save the observable phenomena, Bogen and Woodward emphasized the distinction between data and phenomena and argued that theory often saves the non-observable phenomena, rather than the observed data. Within this context, a large part of the discussion that followed on the philosophy of data (e.g. McAllister 1997; Glymour 2000; Harris 2003; and Massimi 2007) has mainly focused on the relationship of data with the physical phenomena that they are often taken to represent.

In the more recent literature on data, this tendency of philosophers to examine the nature of data with respect to the underlying phenomena they represent, has been replaced by a new tendency to closely examine examples of actual scientific practice, in order to explore the methodology of data processing in various scientific fields and the role of data models within them. This approach is most evident in the works of Sabina Leonelli (2015, 2016, 2019) in biology and Alisa Bokulich (2018, 2020) in paleontology which, as one might expect, are highly influenced by works of non-philosophers on data processing, such as Edwards’ book on climate science (2010).

The present article follows this recent tendency and aims in expanding the existing literature on the methodology of data analysis into the field of modern High-Energy Physics (HEP). Given that modern large-scale HEP experiments rely on the production of large volumes of data more than any other scientific field, it is surprising that not much has been said about the methodology of data analysis in this field. As Bokulich notes, although many philosophers have followed Suppes in highlighting

the importance of data models in science, ‘most [of them] have largely black-boxed how data models are produced’ (2020, p.794), and this includes the discussions on the philosophy of HEP.¹ The primary aim of this article is to fill this gap by closely examining the methodology behind the production of data models in HEP in order to facilitate our understanding of the nature and the role of data models in this field.

In particular, the article explores the nature of data models and their place in HEP by providing a detailed case study of experimental tests of Lepton Flavour Universality (LFU) at the LHCb experiment at CERN. The adopted methodology is characterised by a systematic study of real scientific practice, and falls within the emerging framework of Philosophy of Science in Practice which aims in producing ‘productive interactions between philosophical analyses and the study of actual scientific practices’ (Ankeny et al., 2011, p.305). By taking a close look at the work of the experimental physicists of the LHCb collaboration, the idea is to depart from the usual theoretical approach of philosophy of science and re-examine the concept of data models, as well as other related questions, strictly in terms of scientific practice. The ultimate goal is to gain important insights to the question of what a data model is by examining the process by which a data model is constructed in HEP experiments. This thorough examination will nicely demonstrate the way in which theoretical hypotheses are eventually connected with experimental results via data models, and will highlight the importance of considerations regarding the selection criteria, efficiency calculations, data fitting, and uncertainties in the process of constructing a data model that can be compared to a theoretical prediction in HEP. Contrary to the traditional understanding of data models as idealised versions of the raw data perceived by our immediate observations, the proposed understanding of data models does not rely on the problematic distinction between raw and processed data, nor does it involve the process of immediate observation.

The philosophical lessons we can take from the exposition of the particular case study at the LHCb are abundant and are outlined throughout the text. There are however, four main lessons that can be taken to apply in any large scale HEP experiment. The first lesson concerns what Bokulich (*ibid.*) calls *the folk view of data* which, amongst other things, claims that the tampering of data results in their corruption and the decrease of their epistemic reliability. Contra to this seemingly popular conception, it will be shown that not only the epistemic reliability of data often increases via their processing, but also that raw data – i.e. data that comes out of the detector at the early stages of the experiment – are actually useless as they are for the comparison of the experimental results with the theory. One of Bokulich’s central aims is ‘to make plausible the *prima facie counterintuitive claim* that model-filtered data can – in some instances – be more accurate and reliable than so called raw data, and hence beneficially serve the epistemic aims of science’ (*ibid.* p.10, emphasis added). The close examination of our case study in HEP shows that this view – i.e. that model-filtered

¹Massimi (2007) is perhaps one of the few exceptions, since in her attempt to argue for her thesis that data provide evidence for unobservable phenomena in HEP, she follows van Fraassen in providing a logical reconstruction of data models as partially ordered sets. Some further work on the methodology of data processing in HEP comes from Karaca (2018) and will be discussed in the next section.

data are epistemically superior from the raw data obtained from the experiments – not only is not counterintuitive when it comes to the experimental practice of HEP, but rather, it is the *norm* for conducting successful experiments. As will become evident in the following sections, the very nature of experimental HEP makes the interpretation of raw, unprocessed data impossible, and hence, the only way to achieve progress in the field is by collecting and analysing large volumes of processed data.

This also suggests that a clear distinction between raw data and processed data cannot be applied in the context of large-scale HEP experiments. The close examination of the case study on LFU tests illustrates that an understanding of the concept of ‘raw data’ as data perceived directly from our experience is largely irrelevant to the scientific practice in HEP and what is labelled as raw data and processed data is often merely a matter of convention. Instead of placing data into two distinct categories as raw or processed, what best describes the current scientific practice in HEP is the placement of data in a continuous spectrum in which some datasets are more processed than others, without really worrying where to draw the line between raw and processed data.

The second lesson concerns the further distinction between real and simulated data. The careful scrutiny of the experimental practice at the LHCb illustrates that the boundaries between these two types of data are not as sharp as it is often implied in the literature and not particularly important for the completion of successful experiments in HEP. As we shall see, the final datasets that reach the hands of theoretical and experimental physicists for interpretation are essentially consisted of a mixture of real and simulated data that cannot be distinguished, due to the fact that simulated data are often embedded in real measurement outcomes during the various stages of the experiment. The final data model that is eventually compared to the theoretical hypothesis and provides the ‘window’ through which theory makes contact with the real world, is essentially a *co-production* of real and simulated data.

The third lesson concerns the various levels in which theory guides the overall construction of the experiment as well as the extraction, interpretation and the further analysis of the acquired data. In particular, the example of the experimental tests of LFU in rare B-decays nicely illustrates that theory-ladenness emerges at three different levels throughout the experiment depending on the scope and the purpose for which background theoretical assumptions guide the experimental process. The possibility of a vicious circularity due to the theory-ladenness of observation will not be discussed in detail here since this is a well-known problem which has already been thoroughly examined by many authors (Franklin et al., 1989; Brown, 1993; Brewer & Lambert, 2001; Schindler, 2013; Franklin, 2015; Beauchemin, 2017; Ritson & Staley, 2020). The consensus in these discussions is that theory-ladenness is not necessarily vicious and does not lead to a relativist account of contemporary science. In accordance with this view, our case study shows that the various potential threats of circularity are indeed mitigated by the practice of ‘blind analysis’ and the implementation of uncertainties in the final result. The focus here will therefore remain on the different levels and the extend to which various theoretical assumptions affect the physicists’ decisions in triggering data and their overall understanding of the events at the LHC.

Finally, the fourth and most general lesson concerns the overall process in HEP experiments for the construction of a data model to be compared with the theory. In his seminal paper on data models, Suppes defines models of data ‘in terms of possible realizations of the data’ (1962, p.253) in the same way that the models of the theory are possible realizations of the theory in the logician’s sense. This formal characterization of data models by Suppes closely follows his favourite semantic view of theories which sees theoretical models as set-theoretical structures that are deductively derivable from theoretical sentences.² However, as we shall see, the process of building a data model in HEP via the four stages of selection criteria, efficiency calculations, data fitting and uncertainties is way more complicated and less easily formalised than Suppes’ discussion would lead one to believe.

Although the final data model is indeed ‘a simple entity’ as Suppes pointed out, the process of converting the initial data from the detectors into a concise and polished final result in the form of a statistical hypothesis based on the available data is, as we shall see, anything but simple. The complexity of this process mainly stems from the fact that the aforementioned stages do not follow a clear chronological order and cannot always be easily distinguished. Rather, they describe the essential procedures of a long and reiterative process during which data from the experiment are processed and analysed in a number of various ways, including their fusion with data from simulations and the use of highly sophisticated techniques of statistical analysis. During this long process, theory infiltrates the analysis of data at various levels, having clear effects both on the nature of the collected data and their final interpretation. The fact that different theoretical considerations and different techniques of statistical analysis can, in principle, provide slightly different results makes the description of data models as possible realizations of data in a logical sense seem somewhat unsuitable in the context of HEP. Nevertheless, the three types of models in terms of which Suppes described the connection of theory and experiments in his hierarchy of models account – i.e. models of theory, models of experiment and models of data – are useful concepts, and will be used in what follows for sculpting the overall framework of experimental practice in HEP. At the same time, it will be shown that a less stringent version of Suppes’ hierarchy of models account is indeed reflected at the practical level in HEP experiments, despite the criticism that has occasionally received.

The structure of the article is as follows. Section 2 opens the discussion with a defence of Suppes’ hierarchy of models account, which will serve as the basis for the present account on the nature of data models and their relationship with theory. In Section 3 the focus will be shifted to the necessary theoretical framework for understanding the B-anomalies in particle physics and their usage in tests of the theoretical principle of Lepton Flavour Universality. Section 4 will follow with a presentation of the data processing system at the LHCb experiment at CERN, illustrating how theory enters the collection and analysis of data in three different levels. In Section 5 the process of constructing the data model representing the final experimental result will be described in four stages. Finally, Section 6 will follow with a discussion on the distinctions between raw/processed data and real/simulated data. Section 7 concludes

²For a nice review of the syntactic-semantic debate on theories see Lutz (2017) and references therein.

the discussion by drawing together the philosophical insights from the examination of the case study at the LHCb.

2 Stretching the hierarchy of models account

Suppes (1962) begins his analysis of the relationship between theory and experiment by noting that the theoretical principles to be tested do not usually have a direct observable counterpart in the experimental data. Instead, this gap between theoretical predictions and experimental results is filled by a number of different types of models and theoretical principles which Suppes classifies in five levels. At the top level are the models of the theory relevant to the experiment. The main function of these models is to narrow down the typically broad scope of the theory in question into a simple hypothesis H_0 to be tested by the experiment. At the next level one finds the models of experiment. These are models that are ‘closer to the actual situation’ and whose aim is to adjust the theoretical model to the specific features of the particular experimental setup by providing all the necessary details of how the experiment must be designed and how the data can be linked to the hypothesis in question. At the third level, models of data enter. Suppes describes these as the possible realizations of the data that are ‘designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory’ (*ibid.*, p.258). Finally, in the lowest two levels, are the theory of experimental design which deals with various problems of the experiment that are beyond the particular theory being tested, and what Suppes calls the ‘ceteris paribus conditions’ which concern every other ‘intuitive consideration’ of the experimental setup that does not involve formal applications of the theory (e.g. safety rules, control of external disturbances etc.).

Suppes’ account has been further elaborated by Deborah Mayo (1996, Ch.5). Although significantly richer in details, Mayo’s account maintains Suppes’ main idea: theory becomes testable through the models of the theory which provide a distinct primary question or hypothesis to be tested, and experimental results are linked to this hypothesis as models of data. The connection between these two types of models is mediated by the experimental model: ‘If the primary question is to test some hypothesis H , the job of the experimental model is to say, possibly with the aid of other auxiliary hypotheses, what is expected or entailed by the truth of H with respect to the kind of experiment of interest’ (*ibid.*, p.133). For Mayo, the two key functions of the experimental model are (i) to provide an experimental analogue of the primary theoretical model and (ii) to specify the necessary techniques for linking the experimental data to the questions of the experimental model.

What is also common in Suppes’ and Mayo’s approach is their emphasis on the importance of statistical and other formal methods of analysis in the construction of data models, as a necessary tool for the successful transition from the level of the theory to the level of the experiment. Suppes conclusion is that once the experimental results are condensed into a simple data model, ‘every question of systematic evaluation that arises is a formal one’ (Suppes 1962, p.260-1), implying this way that data models are necessarily statistical models, or at least, subject to statistical and mathematical analysis. Drawing on Suppes’ emphasis on statistical methods, Leonelli

(2019, p.22) has recently criticised Suppes' account by identifying three problems. First, Leonelli notes that Suppes' analysis only deals with numerical data, neglecting the fact that there are also cases where data are not quantitative objects and thus are not amenable to statistical analysis. Second, it is hard to see how Suppes' analysis can be applied in situations of exploratory experiments where the research question under investigation is not clearly stated and thus, it cannot be easily compared with the data model. Finally, Suppes' approach presupposes, according to Leonelli, the ability of researchers to identify what constitutes 'raw data' in the experiment, and overlooks the close connection between the activities of data acquisition and data manipulation.

Leonelli's first observation is correct and lends further support to the claim made earlier that the diversity and complexity of data analysis in various scientific fields makes it impossible to come up with a universal philosophical description for the relationship of theories and data in science. The remaining two observations however, are subject to further analysis. Leonelli's point with respect to the application of Suppes' framework on exploratory experiments stems from Suppes' dictum that the theoretical predictions of a theory are typically expressed in the form of an initial hypothesis and are eventually compared with data models. If there is no theoretical hypothesis to be validated via its comparison with a data model, then Suppes' description is inadequate.

This would be true however, only in the unrealistic cases where scientists are blindly looking for new physics in collider experiments from an Archimedean point of view, independently of any sort of background theory. This is hardly the case in large-scale HEP experiments. The description of the methodology of data acquisition and data processing in Sections 4 and 5, clearly shows how the very act of collecting and analysing data in HEP experiments is simply impossible without the presupposition of a clear theoretical hypothesis with respect to which the data models are built. What constitutes an exploratory experiment in HEP is not the fact that the research question is not clearly stated, rather, it is the fact that the question is not part of an already well-formed and established theory to be tested.³ Karaca (2017) also notes that the exploratory nature of HEP experiments concerns the ability of an experiment to achieve a variety of possible outcomes, which as we shall see, can be made possible by the systematic variation of the various experimental parameters.

The example of LFU tests to be used as a case study here, is a clear example of a non-exploratory experiment in which the theoretical prediction of the Standard Model for the R_K ratio is put to the test by building a corresponding data model. However, one might think of a hypothetical situation where either (i) the existing theoretical framework does not provide a precise numerical value of the ratio, or (ii) several competing and not-well established models offer different values of the ratio. In this case, the research hypothesis shifts from 'Is the experimental value of the R_K close enough to the theoretical prediction of the Standard Model?' to the more exploratory question 'What is the value of the R_K ratio?'. Nevertheless, in

³This approach is also compatible with Steinle's account in which exploratory experimentation 'is driven by the elementary desire to obtain empirical regularities' and 'despite its independence from specific theories, the experimental activity may well be highly systematic and driven by typical guidelines' (1997, p.70).

both cases, the final data model is built with respect to a corresponding theoretical question since the ultimate aim is to fit the data into an already existing or a future theoretical framework. If the data is not in a comparable form with a theory, then this task cannot be accomplished.⁴

Leonelli's third objection is very similar to an objection raised by Karaca (2018), regarding the lack of a modelling concept for the data acquisition process in Suppes' account. Using the example of the ATLAS experiment at CERN, Karaca notes that both Suppes' and Mayo's descriptions leave out a significant aspect of the overall process of bringing together theory and experiment in HEP, which is the specification and organization of the necessary experimental procedures in order to select the required data. This is achieved, according to Karaca, via a model of data acquisition whose key function is to specify the operational and technical details during the data acquisition process and to determine the necessary selection criteria for the rejection of non-interesting events in the LHC collision experiments. While Karaca is right to point out that Suppes' description does not explicitly address the process of data acquisition in HEP experiments, the modified version of the hierarchy of models account that I wish to provide here includes these and other related models within the broad concept of experimental models.

Leonelli's additional point to the discussion is that Suppes presupposes a problematic distinction between the 'raw data' that constitute the 'simple datasets' to be processed and the data models that are eventually compared with the theory. Leonelli draws on Harris' (2003) accurate observation that very often the data that are traditionally referred to as 'raw' are in fact data models, and thus, it is not clear how these models can be compared with theoretical hypotheses. However, this confusion comes from a subtle point regarding Suppes' claims. Suppes' definition of data models with respect to their ability to be compared with the theory only applies to the final simple entity which eventually puts the theory to the test. However, Suppes is not saying – or, at least, should not be understood as saying – that *any* data model must necessarily be comparable with a theoretical hypothesis as Leonelli seems to imply. Rather, what Suppes' is saying is that, when it comes to the comparison of theory with experimental results, the entities with which theoretical hypotheses are eventually compared are *necessarily* data models that are subject to statistical analysis. This is a subtle point, but nonetheless it is important for making sense of the fact that very often the various datasets throughout the process of data acquisition and analysis are indeed consisted of – what I shall call – *auxiliary data models*, and whose function is to facilitate the construction of the final data model to be compared with the theory. Moreover, as we shall see in Section 6, the distinction between raw and processed data is indeed not so clear, as Leonelli and Harris have pointed out, however, it is precisely for this reason that it is also not necessary for describing the scientific practice in HEP.

The deeper lesson to be learned here is that the process of theory testing via experiments in HEP is simply too complicated to be fully captured by a sharp tripartite

⁴This view is also nicely supported by Bokulich and Parker (2021) in a recent paper on what they call the 'pragmatic-representational view of data'. By using an example from climate science, Bokulich and Parker highlight the fact that data and data models are representations that should be evaluated in terms of their adequacy for a particular purpose, in which case is the specific research question.

distinction of three types of models. The modified account I wish to present in the following sections is partially a reconciliation of Suppes' approach with Karaca's remarks, which focuses on three different types of models that constitute a research project. There are however three important caveats to keep in mind. First, the concept of experimental models is significantly extended in order to include *every possible* model related activity which facilitates the connection between the theoretical models and the models of data. Second, although it is possible to provide a relatively clear definition (or description) for theoretical and data models, when it comes to the various types of experimental models the boundaries between them and the two aforementioned types cannot be sharply distinguished and whether one wishes to include a specific modelling activity (such as the models for the specification of selection criteria) in one level or another is up to a certain extent a matter of personal choice. Finally, for a given collision experiment in HEP, each one of these three types does not consist of one single entity, rather it should be understood as a cluster of models with similar features serving a common goal.

The three different types of models will thus be understood as follows:

- *Models of theory*: A model of a theory is a mathematical tool whose aim is to narrow down the scope of the background theory by providing an experimentally testable hypothesis H_0 (or a number of hypotheses) concerning a specific type of physical processes or phenomena. The background theory providing the hypothesis need not be a well-established and empirically well-confirmed theory. It can also be an isolated and preliminary theoretical framework based on a small class of observations, which would give rise to a phenomenological model. Depending on the nature of the background theory and the model, the hypothesis may concern the exact numerical value of a theoretical parameter (e.g. the fine structure constant at a given energy level, the electron magnetic moment etc), an estimate of a model parameter in the form of a probability distribution corresponding to a physical quantity (e.g. particle properties such as mass), or a specific relation (in the form of an equation) between two or more physical quantities (e.g. the differential cross section for a given process as a function of the transverse momentum etc).
- *Models of data*: A model of data is the representation of a measurement outcome into a canonical form that allows – directly or indirectly – the comparison of experimental data with the hypothesis under investigation. The construction of a data model involves a variety of data analysis techniques and statistical methods, and as we shall see, it is heavily guided by background theoretical assumptions and other approximations. Depending on the hypothesis in question, a data model can take several forms such as a table, a simple numerical answer with an uncertainty estimate, or as it is most common in HEP, a function represented by a graph. Finally, although the final data model which is eventually compared with theory is typically a 'simple entity' as Suppes pointed out, the construction of this entity often requires a number of auxiliary intermediate data models.
- *Models of experiment*: A model of an experiment is a blanket term referring to every possible modelling activity that facilitates the completion of a measure-

ment process in an experiment, and allows the connection between the final data model and a theoretical hypothesis. In high energy collision experiments this includes the physical models for calculating the interactions of particles with the different parts of the detector, any kind of simulation modelling that provides the basis for necessary calculations (event generators, detector simulations, pseudo-experiments with Monte Carlo simulations etc.), models of data acquisition, and finally, the various statistical models used for the analysis of data.

Measurement is understood in this context as the experimental activity which leads to the quantitative attribution of the value of a targeted physical quantity, typically represented as a theoretical parameter or variable in an idealized experimental model. Eran Tal (2017a, p.240) describes measurement in terms of two levels: the physical interaction between the target of the measurement and the measuring instrument, and the *model of measurement*, which is an abstract and idealized representation of this physical interaction. The attribution of values to various parameters from the sub-detectors of the LHCb to be described in Sections 4 and 5 based on the experimental models of the interactions between the products of the collisions and the detector, nicely illustrates Tal's description of measurement in terms of the concrete physical interactions in the detector and the abstract models of measurement or – as we shall call them here – models of experiment. For the purposes of our discussion, it is also useful to follow Tal (2017b) and distinguish between *instrument indications* and *measurement outcomes*. The former are properties of the final states of measuring instruments after a measurement is completed such as the numerals appearing on the display of a measuring device, and are often understood as providing the raw data of the experiment. The latter are knowledge claims about the value of a physical quantity attributed to a physical process such as the claim that 'the mass of the top quark is $M_{top} = 172.85 \pm 0.714(\text{stat.}) \pm 0.85(\text{syst.}) \text{ GeV}/c^2$ '. As we shall see, it is the measurement outcomes and not the instrument indications that are represented by data models.

This slightly modified version of the hierarchy of models account, even in its crudest form, nicely captures the relationship of theories and experiments in high-energy physics. Background theory, be it the Standard Model, or any other new physics theory or model to be put to test, such as Supersymmetry, provides its predictions via the theoretical models in the form of empirically testable hypotheses. Large scale scattering experiments are then designed and carried out based on these theories, yielding a huge volume of raw data which is eventually turned into a simple data model which is comparable to the theoretical hypothesis. The acquisition of data and the construction of the final data model is unavoidably carried out with the help of various theoretical assumptions and other intermediate experimental models. The purpose of the following sections is to illustrate how this modified account of the hierarchy of models can be applied for the description of experimental tests of the theoretical principle of Lepton Flavour Universality in HEP via the so-called B-anomalies in the rare decays of B-mesons.

3 B-anomalies and Lepton Flavour Universality

The Standard Model (SM) of particle physics is by far the most empirically successful physical theory for the fundamental building blocks of visible matter and the interactions between them. However, despite its tremendous empirical success, the Standard Model is also undermined by a number of experimental results that consistently deviate from its theoretical predictions.⁵ One type of such results, which will be the focus of the present case study, concerns the so-called B-anomalies in the rare decays of B-mesons that are one of the main areas of study in the indirect searches for new physics at the LHCb experiment at CERN.⁶

The term B-anomalies refers to a set of observed experimental results of various observables of B-decays displaying tensions with the SM predictions at the 2-3 sigma level. The overall consistency of these results is interpreted by many physicists as a hint for the presence of new physics in these decays and hence, the accumulation of further data and the precise measurement of these observables via the appropriate data models is of ultimate importance for the development of new physics beyond the Standard Model. A particular observable in these anomalies is the R_K ratio that features in tests of *Lepton Flavour Universality* (LFU) (Bifani et al., 2018; Muller, 2019). LFU is a theoretical principle of the Standard Model which stems from the fact that, apart from their mass differences, the three charged leptons (electrons, muons and taus) are identical copies of each other, and thus the electroweak coupling of the gauge bosons to leptons is independent of the lepton flavour. In practice, this means that according to the SM, electrons couple to photons, Z and W^\pm bosons in the same way the muons and taus do. As Suppes pointed out however, this is a general theoretical principle that does not directly correspond to an experimental observable and thus a theoretical model is needed in order to convert this general theoretical principle into an empirically testable hypothesis.

The most straightforward way to do this is by constructing a theoretical model featuring the ratio of the branching fractions between two different B-decay processes with different flavours of leptons in their final products, such as $B^+ \rightarrow K\mu^+\mu^-$ and $B^+ \rightarrow Ke^+e^-$ (Fig. 1). Since the electroweak couplings of all three charged leptons are the same, the ratio of branching fractions for these two decays is naively expected to be unity, and it can indeed be calculated theoretically with high precision in a given range of the produced dilepton mass squared q^2 . In the low region range for the dilepton mass squared ($1.1 < q^2 < 6.0 \text{ GeV}^2/c^4$), this ratio is predicted by the Standard Model to be unity with $\mathcal{O}(1\%)$ precision (Bordone et al., 2016). This theoretical prediction constitutes the theoretical hypothesis:

$$H_0 : R_{K+[1.1, 6, 0]}^{SM} = 1.00 \pm 0.01_{QED}$$

⁵For more details on the limitations of the Standard Model and the searches for new physics at the LHC see Virdee (2016) and Rappoccio (2019).

⁶As opposed to direct searches, which aim at the observation of hypothetical new particles via their production in scattering experiments, indirect searches for new physics concern the performance of precise measurements of observables in (usually rare) scattering processes, by analysing large volumes of data on observables related to these decays.

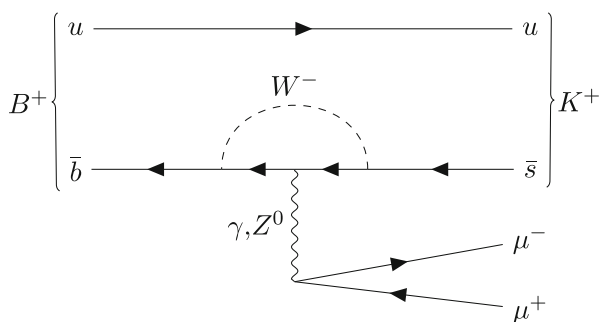


Fig. 1 Feynman diagram of the dominant contribution to the B-meson decay $B^+ \rightarrow K^+ \mu^+ \mu^-$ within the Standard Model. The diagram illustrates the rare process in which a B^+ meson ($u\bar{b}$) decays into a K^+ meson ($u\bar{s}$) and a pair of muons

which is eventually compared to the model of data. The QED subscript indicates the origin of theoretical uncertainties due to QED effects and the numerical interval corresponds to the dilepton mass squared range.

The ultimate aim of ‘the maddeningly diverse and complex experience’ which constitutes the experimental test of LFU is to construct a data model of the R_K ratio: a simple entity in the form of a numerical result, subject to statistical and systematic uncertainties, which is comparable to H_0 in a precise and mathematical manner. The next section provides a brief description of the data processing system of the LHCb in order to facilitate the discussion to follow on the rather complicated process of constructing a data model for the R_K ratio. As we shall see, theory guides the observation and data acquisition process in three different levels: an all-encompassing fundamental level independent of the specifics of the experiment, an intermediate level concerning the physical processes in the detector, and a more restricted level regarding the specifics of the quantities to be measured in a given experiment.⁷

4 Data processing at the LHCb

The LHCb experiment at CERN is currently the largest experiment in physics for the study of rare B-decays. It is specifically designed to profit from the enormous production rate of b quarks in proton-proton collisions at the Large Hadron Collider (LHC) which happen at a rate of around 3×10^{11} per fb^{-1} .⁸ The LHCb detector collects about 25% of the b quarks produced in these collisions, and provides the necessary data for making precise measurements of various observables related to the rare B-decays.

⁷The content of this section was derived from Teubert (2016), Blake et al. (2017), Capdevila et al. (2018), Lionetto (2018), Mauri (2019), Lisovskyi (2019), and Humair (2019).

⁸One inverse femtobarn (fb^{-1}) corresponds to approximately 100 trillion ($\sim 10^{14}$) proton-proton collisions.

The detailed study of these processes requires the determination of the various properties of the final state particles and their kinematics. In order to determine these properties and allow the full reconstruction of an interaction process, a number of different quantities need to be measured including the charge of a particle, its flavour, the momentum vector, and for short lived particles, the production and decay vertex. Since no detector can simultaneously measure these quantities, large detector systems such as the LHCb detector, are typically made of various specialised sub-detectors, each performing a different task. The various sub-detectors of the LHCb detector can be grouped into two complementary sub-systems: the Track Reconstruction system and the Particle Identification system. As the name suggests, the systems involved in track reconstruction aim in reconstructing the trajectories of charged particles in a collision event by combining information from the ‘hits’ recorded in the various sub-detectors. Once the tracks are reconstructed, the Particle Identification (PID) system derives further information from its sub-detectors in order to associate the tracks with a specific particle species. Together with the momentum information provided by the tracking system, the PID also allows the energy of a charged track to be computed using the relativistic energy-momentum relation $E^2 = p^2c^2 + m^2c^4$.⁹

4.1 The three levels of theory-ladenness

Already one may notice here the first and most general level of the theory-ladenness of observation. The overall design and operation of the track reconstruction and particle identification systems at the LHCb (as well as of any other large scale experiment) is based on a number of physical principles that are considered to be fundamental and are expected to hold in any possible new physics theory to be constructed based on these data. These general principles enter the observational process in the form of various implicit and explicit assumptions which lie at the core of almost every experiment in physics and concern the most fundamental facts we know about nature, such as the conservation of energy and Einstein’s mass-energy equivalence principle. This type of theory-ladenness is universal across a particular field of physics and is independent of the aims and quirks of any particular experiment.

The second level of theory-ladenness of observation in collision experiments concerns the physical processes behind the production of ‘hits’ in the detectors and the identification of particles. During a proton-proton collision event, hits are produced in the various trackers by the energy loss of the traversing particles due to their interaction with matter. The two main physical processes that occur in the detectors are inelastic collisions of the products with the atomic electrons and elastic collisions with the nuclei of the atoms of the detectors’ material, leading to the phenomena of ionisation and multiple Coulomb scattering respectively. Theory-ladenness appears at this stage by offering the various physical models for calculating the effects of these physical processes on the detector.

⁹For a detailed description of the LHCb detector see the official publication from the LHCb Collaboration (2008).

As aptly noted by Beauchemin (2017, p.299), quite often there are more than one competing models about the nuclear interactions between charged hadrons and the material of the detector.¹⁰ However, these competing models, although empirically equivalent, affect the simulation of the detector and the selection of data in different ways, giving rise to different results. In other words, the extrapolation and interpretation of data, and consequently the form of the final experimental result as a data model, depends on the choice of the model for the underlying physical processes in the detector. This fact raises the worry of a possible vicious circularity due to the theory ladenness of data selection. If the result depends on the arbitrary choice between several empirically equivalent models, what validates the objectivity of a given result based on a particular model? As will be shown in Section 5.4, the solution to this problem is achieved by separately calculating the effects of each model to the measurement and including them in the systematic uncertainties of the final result.

The third and most specific level of theory ladenness concerns the theoretical principles and assumptions that are specific to the aims of the particular experiment which will be described in the following section. These assumptions basically determine (i) the selection criteria for distinguishing the data from what are considered to be the ‘interesting events’, i.e. events related to the two decays consisting the R_K ratio and (ii) the vast majority of theoretical and mathematical calculations involved in the derivation of the final result. The suggested tripartite distinction of theory-ladenness presented here partially overlaps with Karaca’s (2013) two-fold distinction between the strong and the weak sense of theory-ladenness of experimentation, albeit with an additional intermediate layer. Karaca describes the strong sense of theory ladenness experimentation as the continuous guidance of an experiment by some theoretical account with the aim of ascertaining the conclusions of the same account. This strong sense of theory-ladenness is captured by what I call here the third and most specific level of theory ladenness which essentially determines the collection and further refinement of data at the LHCb trigger system, in order to construct an appropriate data model to be compared with the theoretical hypothesis in question. The weak sense is described by Karaca in a broader context, as the utilization of theoretical considerations that have no guiding power on the progress of the experimental process.

4.2 The LHCb trigger system

Before moving to the analysis of the data modelling process for the R_K ratio it is useful to give a brief description of some technical details regarding the data processing system of the LHCb. The rate of visible collisions at the LHC, i.e. the number of recorded events per second, is currently between 10 and 20 millions (~ 13 MHz

¹⁰The reader may notice here a similarity in the title of the present paper and Beauchemin’s paper due to the presence of the medical terms ‘autopsy’ and ‘anatomy’. Although this is a mere coincidence, the two papers are indeed related with respect to their close examination of the scientific practice at the LHC.

in Run II).¹¹ This number is simply too big to allow every single event to be stored for further analysis and thus, a filtering system is required to select the interesting events by filtering out the events containing various well-studied physical processes that are unrelated to the specific aims of the experiment. For the LHCb experiment, this amounts to the selection of the events that are most likely to contain a B-meson or a D-meson, since, in addition to the study of the rare B-decays, LHCb is also dedicated to the study of D-decays (decays of heavy D-mesons consisting of at least one charm quark/antiquark) and CP violations. The selection of these events is completed in two levels by the LHCb Trigger system, and is based on the information from the various subsystems of the detector.

Practically speaking, the ultimate task of the software algorithms connected to the tracking and the particle identification system of the detector is to attach values to several variables related to the kinematics of the interactions (momentum, energy, mass etc), their topology (scattering angles, flight distance, impact parameter) and the nature of the particles. Two simple examples of such variables are the binary $is_{\mu on}$ variable which depends on the number of hits in the muon stations associated with a track, and the $DLL_x(t)$ variable which corresponds to the likelihood of a track t to belong to a particle species x rather than a pion. These variables can be produced based on information from either a single subdetector or by combining information from several detectors. The job of the trigger system is then to take these variables as inputs and, based on a number of selection criteria that are also known as *cuts*, decide whether a given event is of interest or not.

In order to be able to distinguish the interesting events from the various processes taking place in the LHC, the trigger system of the LHCb is programmed to search for the characteristic signatures of hadrons containing b or c (anti)quarks, which give rise to the heavy flavour decays in which we are interested. The three most significant signatures of these hadrons are (i) their large lifetime, which results in long flight distances compared to the resolution of the detector, (ii) their large mass, which results in high transverse momentum P_T of the product particles,¹² and (iii) the existence of muons in the final state of several key decay modes of these hadrons, such as the $B^+ \rightarrow K^+ \mu^+ \mu^-$ decay in which we are interested in for the measurement of the R_K ratio (Head, 2014). This is where the third – aim specific – level of theory-ladenness becomes apparent: the specification of these signatures for the data selection process is largely driven by various theoretical assumptions for the nature of these decays based on the existing background theoretical knowledge. This fact is also related to what was said earlier in Section 2 about the necessary connection of data acquisition with a clear research question. The fact that the ultimate purpose of this particular experiment is to test LFU via the R_K ratio specifies which events are of interest for this purpose, and eventually determines the choice of the most appropriate selection criteria to distinguish these events.

¹¹In the jargon of particle physics, the recording of an *event* amounts to the recording of all the products from a given collision. Run I and Run II refer to the different periods of operational running for the LHC under different conditions. Run I took place in 2009–2013 and Run II in 2015–2018. Run III is scheduled to take place in the years 2021–2023.

¹²Transverse momentum is the component of momentum transverse (i.e. perpendicular) to the beam line.

The first level of the LHCb trigger system is completed by the Low Level trigger (L0). L0 is a hardware based trigger and its task is to reduce the data output from ~ 13 MHz to approximately 1 MHz at which the LHCb detector can be read out. Contrary to what one might expect, the selection criteria at this first level are not purely theory-laden, rather they are mainly determined by a number of technical limitations. In order to achieve the goal of 1 MHz, the L0 trigger needs to take a decision for every event in a very short amount of time ($4\mu s$), and for this reason, it only receives information from the muon system and the calorimeters, as these are the only sub-detectors able to provide information in such a short amount of time. Once this information is received, the trigger algorithm discards all events with too many hits in the SPD detector since such high occupancy events would require an excessive fraction of the available processing time at the next level of the data process. After these criteria are applied, L0 proceeds to a coarse-graining of the interesting events by selecting muons with a high transverse momentum p_T and other events with high energy deposits in the calorimeter. The thresholds for these cuts are not fixed, rather they are constantly changing according to the data-taking conditions of the experiment, even during the same year or Run.

The second level of the trigger system is completed in two stages by the High Level triggers (HLT1 and HLT2) of the detector. These are software based algorithms and their task is to further reduce the amount of data in order to be stored onto servers at the CERN Data Centre and distributed to physicists for analysis. In Run II, HLT is programmed to reduce the data rate from the 1MHz output of the L0 trigger to 12,5 kHz which is low enough to be permanently stored on disks. During the first stage of the High Level triggering, HLT1 receives information from the tracking system and proceeds to a partial event reconstruction by applying various selection criteria based on the impact parameter of the events, the quality of the tracks and the transverse momentum.¹³ This process reduces the data output to ~ 70 kHz and passes the selected tracks to HLT2. The HLT2 algorithm then performs a full reconstruction of all the selected events that satisfy $P_T > 0.3$ GeV independently of their impact parameter or matching hits in the muon chambers. The overall process of reducing the amount of data from the LHC collisions to a manageable dataset to be distributed widely is illustrated schematically in Fig. 2.

This pragmatic dimension of the data acquisition process is nicely captured by Bokulich and Parker in their discussion of the *problem space* in data modelling, in which the goal is to achieve a particular purpose of interest guiding the construction of the model (2021, p.12). As Bokulich and Parker note, the final properties of the data model are jointly determined by the different dimensions of the problem space, namely the representational relationship between the data and the target of the experiment, the data users, the adopted methodology and the background circumstances of the experiment. Given that the purpose of our case study is to test LFU via the R_K ratio, the final form of the data model is indeed jointly determined by a number of

¹³The impact parameter of a particle, typically denoted by χ^2 , is related to the angle of scattering, i.e. the angle at which a particle is deflected by a another particle after collision, and is used in particle identification to tag flavours to the particles.

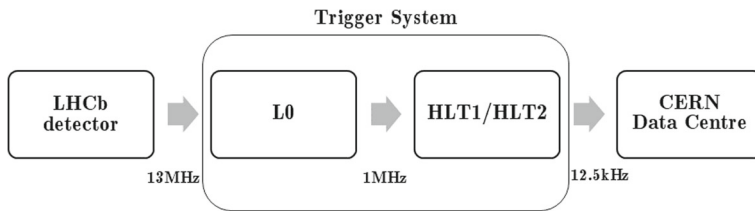


Fig. 2 Schematic representation of the various stages of the data acquisition process at the LHCb. The numbers below the arrows indicate the size of data before and after each level

theoretical and pragmatic factors including the choice of models for the interactions of product particles with the detector, the available computational time and power, the storage capacity at the CERN Data Centre, the reconstruction of events so that they are amenable to statistical analysis and so on.

The completion of the High Level triggering process marks a significant milestone where the vast majority of the available data from the proton-proton collisions at the LHC is discarded *irretrievably*, mainly due to the technical limitations of the data processing system both in terms of the data-processing time and the store capacity of CERN's Data Centre. The reduction of data from 13 MHz to the final 12,5 kHz that eventually becomes available to the users means that about 99,9% of the available data from visible collisions never reaches the physicists' desks for further analysis. Add this to the fact that only about 1% of the actual collisions in the LHC provide products that end up in the detectors, and it is not hard to see that the otherwise huge amount of data that eventually gets stored and analysed by physicists is only a minute fraction of the potentially available information provided by the proton-proton collisions at the LHC. Even though extreme care is taken to make sure that the data collected correspond to the events containing new physics, it is widely acknowledged by the physics community that a large amount of information containing hints to new physics is permanently lost during this process.

This brings us to the final aim of the overall experimental process which is the acquisition and organisation of data for the construction of the data model representing the results of the experiment. Part of this process concerns the determination of the selection criteria for the collection of data, whereas another part takes place only once a sufficient amount of data for the study of a particular phenomenon becomes available at the CERN Data Centre. The datasets are then widely distributed to the scientific community on an international scale by the Worldwide LHC Computing Grid (WLCG) for further statistical and mathematical analysis. Given the huge amount of data required to produce reliable results, it should be stressed that the overall process of deriving an experimental result from the available data is typically a *non-linear* and laborious activity of constant refinement and revision, which usually takes years of collaborative work to complete. Nevertheless, it can be characterised by four main stages which will be the focus of the next section. Although in practice these stages do not follow a clear chronological order and are not always clearly distinguishable, they nicely capture the most essential procedures for the construction of a data model in HEP.

5 Constructing data models for the R_K ratio

This section provides a description of the four main stages for the construction of the data model of the R_K ratio as it was recently presented by the LHCb collaboration (2019; 2021): (i) data selection (ii) efficiency calculations (iii) data fitting and (iv) uncertainty calculations. The analysis of these four stages illustrates the importance of considerations for the construction of a data model with respect to the data acquisition criteria, the complicated calculations of the performance and efficiency of the detector with the help of simulation, the fitting of finite data to continuous functions via statistical analysis, and the evaluation of possible errors during the measurement process. As we shall see, each stage of this procedure is, in its own way, replete with various underlying theoretical assumptions, giving further credence to the idea that observation in HEP is highly theory-laden.

5.1 Selection criteria

The first stage in the construction of the R_K data model concerns the determination of the selection criteria to be applied to the trigger system in order to distinguish the signal – i.e. the events of interest that contribute to the ratio – from the background – i.e. the unrelated events in the collision with similar signatures. The involved strategies during this stage are determined at each step according to the source and the specific characteristics of each type of background based on the existing theoretical knowledge. For instance, a particularly invasive form of background comes from the mis-identification of pions as leptons in the $B^+ \rightarrow K^+\pi^+\pi^-$ decays, which are 30 times more frequent than the B-decays constituting the ratio. The suppression of this background is achieved by applying a combination of cuts for the $isMuon$, the DLL_μ , and the DLL_e variables in the particle identification algorithms. In general, the choice of these cuts is based on a combination of both theoretical and pragmatic criteria regarding the expected behaviour of the detector with respect to each type of background. The underlying assumption is that there is sufficient knowledge of the nature of different types of the background processes which produce signals that could potentially be mis-identified by the detector as coming from the rare B-decays.

The main challenge at this stage, is what Franklin (1998, 2015) calls ‘the problem of cuts’, which stems from the possibility that the experimental result simply reflects the choices of the particular cuts on the triggering system. In other words, the worry is that certain combinations of cuts will give rise to different sets of results and there is simply no way of knowing which of these combinations provides a genuine unbiased result. The situation becomes worse in cases where the effects of the cuts to the result are known to the experimenter in advance, and hence, the idea of producing a desired outcome may distort the objectivity of the experimental results. As Franklin notes, the experimenter’s bias is mitigated by applying the practice of ‘blind analysis’, in which the experimenters analysing the data do not know the result until the analysis method is finalized, following an extended peer review within the collaboration. Beauchemin (2017) adds that the solution to these problems also comes from the implementation of systematic uncertainties in the result, which will be further discussed in Section 5.4.

Once the first stage of calculating and applying the selection criteria for distinguishing the relevant decays for the R_K ratio is completed, the measurement of the ratio requires the calculation of two types of quantities: the efficiencies, ε , for selecting each one of the decays and the yield, N , of each decay mode, which is the number of recorded events contributing to the ratio. The calculation of these two types of quantities constitutes the second and third stage respectively for the construction of the R_K data model.

5.2 Efficiency calculations

The second stage of the data modelling process concerns the calculation of the detector's efficiency during the triggering, reconstruction, and identification processes. These efficiencies are usually integrated in the total efficiency of the detector, ε_{tot} , which can be defined as the fraction of the events registered and correctly identified at the detector, with respect to the actual number of events produced by the proton-proton collision in the LHC. The knowledge of these efficiencies is essential, since in order to know the true value of the ratio between the two yields, it is clear that we must first be in a position to know how many of the rare B-decays that actually occur are eventually recorded by the detector and become available for analysis. The calculation of the 'true number' of rare B-decays is a crucial yet challenging aspect at this stage. Given that the only way to detect and count these decays is via the – imperfect – detectors, how is it possible to know how many of these decays are eventually recorded? The answer is via simulation.

The overall process of calculating the efficiency of the detector by simulation can be described in three steps. The first step is to provide a complete list of all the particles that come out of a certain physical process, including the ones that are stable enough to interact with the detector. This is made possible by various software algorithms that are known as *event generators*. When combined, event generators provide a complete description of all the particles that come out from a collision between protons in which a B-meson is produced, providing this way the necessary knowledge for the expected yield of rare B-decays.

Once the events are generated, the next step in this stage is to simulate the path of the produced particles in the various parts of the detector, in order to model the detector's response. This process requires the construction of a detailed digital map of the LHCb detector in a language that is readable to the software. Ideally, this map would include every single wire and pipe of the detector ensuring that the simulation provides accurate results, however, this would require an unrealistic amount of processing time, and thus various approximations are used. This part of the simulation also involves the implementation of various physics models in the software, describing the different physical processes that are expected to take place in the detector (bremsstrahlung, ionization, multiple scattering etc.) according to the background theory. Once again, it should be noted that it is not possible to include in the simulation every single physical process that is expected to take place (this would require the simulation to run for a tremendous amount of time) and thus, the physical models are chosen on a pragmatic basis, taking into account limitations on time and computational power. The final output of this second step in the simulation is a large

database with information about energy deposits in the detector including their times and locations.

The third and final step in simulation is the *digitization* of data. This is the process whereby the available data from simulation is converted in the same format as the data provided by the experiment electronics and the detector's data acquisition system. The idea is to convert the information about the energy deposits from the simulation into whatever it is that the detector actually reads – i.e. voltages, currents and times. Moreover, this is the stage where various other interesting detector effects are also taken into consideration with the help of various models, such as the difference in light collected from a scintillator tile in the calorimeter depending on whether the energy is deposited in the middle or in the edge of the tile. The final result is a simulated dataset that has the exact same format as the data coming out of the detector's data acquisition system, and for which, as opposed to the real data, there is precise knowledge of the physical processes that generate them. This allows the calculation of the efficiencies of the tracking and particle identification systems of the detector. After digitization, the simulated data follow the exact same path through the trigger system just as the real data, allowing this way the calculation of the efficiency of the trigger system as well.

This procedure is not immune to problems either however. Even though the simulation is considered to provide a good estimate of the detector's efficiency in real data acquisition, it is still possible to have discrepancies between the simulation-calculated efficiency and the true efficiency of the detector. This may happen for instance due to technical problems during data acquisition from real collisions that are not taken into account in the simulation, or poor modelling of certain aspects of the detector in the simulation software (for instance, it is known that the performance of the RICH detectors and calorimeters is not accurately simulated by the LHCb software). These discrepancies are often corrected by a data-driven method called 'tag & probe' whereby the simulation efficiency is revised based on data calibration samples from other well-studied decays.¹⁴

The various models described during this stage – i.e. detector layout models, models of the physical processes and other effects in the detector, models of data flow in the detector etc. – are all part of the class of experimental models described in Section 2, whose task is to facilitate the connection between the theoretical and the final data model of the experiment.

5.3 Data fits

As already mentioned, in addition to the detector efficiencies, the measurement of the R_K ratio requires the calculation of the yields N of the decays of the ratio. This is achieved in the third stage of data modelling via the process of data fitting. In general, data fitting is the mathematical process of finding a function that *best fits* a number of data points (i.e. the process of 'fitting the curve'), with the aim of determining or estimating the values of various unknown parameters affecting the collection of data.

¹⁴For a detailed description of the 'tag & probe' method see Archilli et al. (2013).

As noted by Suppes (1962, p.253), one of the most profound complications in the reconciliation of data and theoretical predictions is that the former are of a discrete and finite nature, whereas the latter are typically continuous functions or infinite sequences. Data fitting is the mathematical tool for solving this tension by finding the most appropriate (continuous) function that best describes the finite sequence of data collected in the experiment.

In the context of the measurement of the R_K ratio, the fits are performed to the data for the combined mass $m(K^+\ell^+\ell^-)$ in each decay, providing this way a probability distribution for the mass of the B-meson. This distribution is considered to be the best description of the set of observations x_i , given that these observations are also affected by the presence of residual background (i.e. background that evades the data selection process). Once the fit is performed, the probability density function is re-parametrised so that it is a function of the relevant yield N and R_K , and maximum likelihood estimations are then performed to find the values of the yields for the signal N_s and background N_b for which it is most likely to observe the given masses $m(K^+\ell^+\ell^-)$ in each decay process. For instance, Fig. 3 illustrates the fit performed to the $m(K^+\mu^+\mu^-)$ data in order to extract the yield of the nonresonant decay $B^+ \rightarrow K^+\mu^+\mu^-$. This is an example of an auxiliary data model needed for the construction of the final data model representing the R_K ratio. The extraction of the yield from these fits involves the use of specialized software algorithms both for the determination of the shape of the curve and the maximum likelihood estimation of the parameters, taking into account all possible contaminations to the fit from background contributions.

Maximum likelihood estimation (MLE) is one of the most popular statistical methods for calculating unknown parameters such as the yields of decay processes in high energy physics experiments. Roughly speaking, given a probability density function $f(x_i; \theta_i)$ describing a set of observations x_i , that are characterised by a set of parameters θ_i , MLE is a method of finding the values of θ_i that make the data most

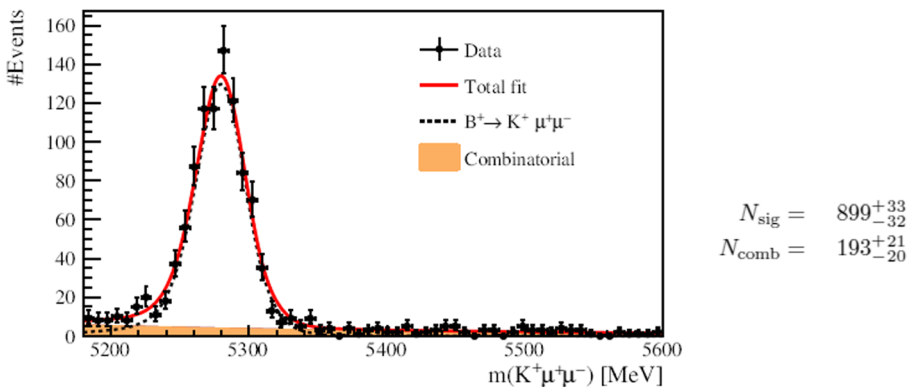


Fig. 3 Fit to $m(K^+\ell^+\ell^-)$ for $B^+ \rightarrow K^+\ell^+\ell^-$ events in the Run II data, along with the contribution from combinatorial background. The extracted values for the signal yield N_{sig} and the background N_{comb} are displayed at the right of the figure (Humair, 2019)

likely. What this means in practice, is that the final value of R_K which is eventually compared to the theoretical hypothesis H_0 , is *itself* a hypothesis as well, which nonetheless, is derived from the available experimental data on the basis of various mathematical criteria. What makes MLE a popular method in HEP is the fact that compared to other estimation methods, it is characterized by a number of ‘good’ statistical properties such as consistency, small bias and robustness.¹⁵

For completeness, let us note here that the most likely value of the R_K ratio given the available data for the most recent measurement (LHCb Collaboration, 2021) was found to be

$$R_K = 0.846^{+0.042+0.039}_{-0.013-0.012} \quad (1)$$

where the first uncertainty is statistical and the second systematic. The fourth and last part in the construction of data models concerns the determination of these uncertainties, which, as we shall see, are a very important and indispensable part of a data model.

5.4 Uncertainty calculations

The attribution of statistical and systematic uncertainties in a HEP experimental result can be understood as a way of quantifying possible errors in the data taking process. This understanding reflects the seemingly more popular ‘error approach’ in HEP, whose objective is to determine an estimate of a quantity which is as close as possible to the unique true value of the quantity. This is opposed to the ‘uncertainty approach’ whose objective is to determine an interval of values which can be equally assigned to a quantity with relatively high confidence, and can be understood as a way of quantifying doubt during a measurement process.¹⁶ In the case of the R_K ratio, the preference to the error approach is reflected by the expression of the result as a single numerical value – which is supposedly as close as possible to the real value of the ratio – associated with statistical and systematic uncertainties.

Generally speaking, in a HEP experiment, there are six main sources of uncertainty: (i) the intrinsic probabilistic nature of the underlying quantum field theory, (ii) the theoretical uncertainties involved in the calculation of various quantities due to highly complicated (usually QCD related) theoretical calculations (iii) the various measurement errors that are present in the data taking process even without taking into account any quantum effects, (iv) the variability in the selection of different models and different measurement methods in the experiment (v) the experimenter’s insufficient knowledge about various aspects of the experiment due to limitations of

¹⁵In short, an estimator of a parameter is said to be *consistent* if it converges, in probability, to the true value of the unknown parameter as the number of measurements tends to infinity. The *bias* of an estimator is the average deviation of the estimate from the true value over an infinitely large number of repeated experiments. *Robustness* is the property of an estimator to have limited sensitivity to the presence of outliers in the data. The full mathematical definitions of these properties can be found in Lista (2016, Ch.5).

¹⁶See Mari and Giordani (2014) for an illuminating discussion on the error approach and the uncertainty approach in science.

cost, computational time, computational power and so on, and (vi) the simple fact that a repeated measurement may yield different results for the same quantity.¹⁷

These and other possible sources of error give rise to two different types of uncertainty that are typically accompanying a HEP result in the form of the data model: statistical uncertainties and systematic uncertainties. A possible way to distinguish between these two types of uncertainty on the semantic level, is to understand the former as expressing the possible fluctuations in a measurement result even when all input quantities and other factors affecting the measurement are perfectly known and stable. This means that the presence of statistical uncertainty can be attributed to the probabilistic nature of quantum field theory and other purely statistical factors, and thus its minimisation is quite often merely a matter of collecting additional data in future runs. Systematic uncertainties on the other hand, can be seen as resulting from our imperfect knowledge on various aspects of the experiment, the mis-modeling of detectors in the simulations, and the possible defects and biases of measuring instruments during the data taking procedures. A large part of the data analysis process therefore concerns the precise calculation and mitigation of systematic uncertainties to the extent allowed by the available funds and the available time, which in turn will provide extra security and robustness to the final result.

In a recent study on uncertainties in HEP, Staley (2020) aptly notes that the distinction between statistical and systematic uncertainties in HEP is rather opaque.¹⁸ The main reason for this ambiguity stems from the fact that the sources of systematic uncertainty in a measurement are often unknown and difficult to distinguish from statistical uncertainties. Moreover, they often require a different method of evaluation, which in turn makes the combination of systematic and statistical uncertainties in the final result problematic. In order to resolve this lack of consensus, particle physicists have developed an extensive literature on the treatment of systematic uncertainties providing possible definitions and practical guidance on methods of statistical evaluation.¹⁹ Barlow (2002) for instance, provides two conflicting definitions of systematic error by ‘widely read and accepted authors’ and shows how different measurements in HEP reflect these two definitions. He then concludes his paper with a set of practical advice for practitioners.

Given this ambiguity, in practice, the lack of consensus on the distinction between statistical and systematic uncertainties is usually resolved by simply stating the sources of statistical and systematic uncertainties in a published result.²⁰ In the case

¹⁷I am grateful to an anonymous referee for pointing out that this list – or any list – is, of course, non exhaustive.

¹⁸The following remarks from experimental physicist Pekka Sinervo confirm this: ‘the definition of these two sources of uncertainty in a measurement is in practice not clearly defined, which leads to confusion and in some cases incorrect inferences. [...] The definition of such uncertainties is often ad hoc in a given measurement, and there are few broadly-accepted techniques to incorporate them into the process of statistical inference’ (2003, p.122).

¹⁹See for example Barlow (2002), Sinervo (2003), Lyons (2006), Wanke (2016), and Bailey (2017) and references therein. Staley (2020) offers a very illustrative philosophical analysis of the various aspects of this debate.

²⁰This, of course does not solve the problem of how one should evaluate and combine these two types of uncertainty as noted by Staley (2020).

of the R_K ratio, ‘by convention, the uncertainty on R_K arising from the statistical fluctuations affecting [the ratio of the yields] $\frac{N_{K\mu\mu}}{N_{Kee}}$ is referred to as statistical uncertainty’ (Humair 2019, p.133, emphasis added). All other sources of uncertainty are integrated as systematic uncertainties and are listed below (*ibid.*, p.138):

1. Calibration samples size
2. Kinematic reweighting
3. PID calibration
4. Trigger calibration
5. Occupancy proxy
6. Tracking efficiency
7. q^2 and mass resolution
8. Decay model
9. Fit shape

Some of these uncertainties are related to the finite nature of the data samples while others come from various limitations in the detector, and the presence of physical effects like bremsstrahlung which significantly complicates the identification of electron tracks. The calculation of each type of uncertainty follows a different methodology according to the nature of the source, but the main idea remains the same. As we have seen, the overall data taking process for the extraction of the R_K ratio involves the utilisation of various auxiliary models and other assumptions that are necessary for carrying out the calculations leading to the final result. However, given that there is often no theoretical or empirical justification for (i) the use of one experimental model over another, or (ii) the assignment of a particular value in a parameter of a model or an assumption (e.g. a specific threshold value in the triggering system) the result is extracted several times either by varying the auxiliary experimental model or the value of a parameter within the selected model (Staley 2020, p.102). The variance in the result due to the use of different models and different parameters is then recorded as a systematic uncertainty.

Uncertainties are a crucial and indispensable part for the reliability of an experimental result but a further discussion on their nature and exact role requires a much deeper analysis which is beyond the scope of this paper. As a closing remark, let us simply note that in addition to being a quantifiable measure of comparison between different results from different experiments, uncertainties are also a solid way of determining the accuracy and precision of a specific result. This point has been nicely illustrated by Beauchemin (2017) who emphasizes the critical role of uncertainties in determining the robustness and the validity of measurements. A measurement is robust insofar as the systematic uncertainties on the final results are ‘sufficiently small’ regardless of the source of these uncertainties. Sufficiently small is to be understood here as being significantly smaller than the order of magnitude of the physical effect to be measured. How much smaller is significantly smaller is not written in stone, however, the main idea is that the smaller the uncertainty, the more robust the result will be.

Beauchemin also notes, rather interestingly, that in cases of small uncertainties, the allegedly vicious circularity of theory-ladenness in observation is not problematic, precisely because ‘its impact on the physics conclusions will be small and fully accounted for’ (*ibid.*, p.303). Beauchemin’s remarks have been further elaborated by Ritson and Staley (2020) who nicely illustrate how the identification of the assumptions on which a result depends and the further quantification of the dependence of this result on the various assumptions in terms of uncertainty calculations, jointly control the possibility of a vicious circularity at the practical level. The determination of the dependence of the result on the various theoretical assumptions in terms of uncertainty serves in discriminating amongst those model assumptions that have the highest impact on the uncertainty of the result and those whose variation introduces negligible changes. The clear separation between the statistical and systematic uncertainties, and the identification of the different sources of uncertainties in the published result as presented in the above list, nicely demonstrates how Ritson and Staley’s observations can be applied at the example of the R_K ratio.

6 Two dubious distinctions

Now that we have seen how the available data are treated in different ways during the various stages of the construction of a data model, we are in a position to make some remarks about the two distinctions between (i) raw and processed data and (ii) simulated and real (or signal) data. Although in both cases, the two extremes in these distinctions can be clearly defined, the transition between the two types of data in each case is, as we have seen, quite blurry. Regarding the first distinction, raw data are often defined as objects that are directly perceived by our experience without any mediating processing or influence by theory (*cf.* Harris, 2003, p.1511). If this definition is taken seriously, then it is not clear at all what should be counted as raw data in a large-scale HEP experiment. In practice, physicists tend to refer to the electronic signals produced by the physical processes in the various parts of the detector as the ‘raw data’ given to us by the proton collisions, whereas the output of the triggering system that eventually gets stored in the data centre and reaches the hands of researchers is referred to as ‘reconstructed events’.

However, none of the signals produced in the detectors is actually directly perceived by the researchers at CERN. Before reaching the hands of physicists, the data from the electric signals produced at the early stages of the experiment at the Track Reconstruction and Part Identification systems described in Section 4, undergo a long process of refinement and reconstruction by the computer algorithms of the LHCb detectors and triggering systems. Hence, the ‘first points of contact’ – i.e. the reconstructed events – are long lists of numerical data about energy deposits on detectors, momenta etc, but as we have seen, these data are far from being unmanipulated and clear from any theoretical influence. The very nature of particle physics therefore makes it impossible to talk about raw data in this field in a strong sense.

This point also illustrates that the basic definition of data models as ‘a corrected, rectified, regimented, and in many instances idealized version of the data we gain from immediate observation’ given by Frigg and Hartmann (2016) does not really

apply in the case of HEP. Nevertheless, the data in the reconstructed events are, in a sense, also ‘raw’, since they still need to undergo a long process of further analysis by scientists in order to reach their final form as a data model which is comparable to the theory. A more appropriate way to describe this situation is thus to say that data follow a long ‘ripening’ journey which starts from their birth as electric signals in the heart of detectors, and goes all the way up to the final polished form of a data model, without really worrying at which stage the data should be considered to be raw.²¹ It is precisely for this reason that the novel definition of data models in Section 2, does not depend on a clear-cut distinction between raw and processed data and thus avoids the relevant objection discussed by Leonelli.

Regarding the folk view that sees the tampering of data as an act of decreasing their epistemic reliability, it should be obvious from our discussion so far that this does not apply to HEP experiments. Generally speaking, a dataset is epistemically reliable if the information it provides for the physical phenomenon it represents is correct. In the context of LFU tests, to say that the processing of data decreases their epistemic reliability is therefore to say that the processed datasets provide less accurate information about the possible violation of LFU in B-decays compared to their less processed counterparts. This is not true however. The successful completion of a large-scale experiment in HEP and the extraction of meaningful and reliable conclusions about the empirical adequacy of various theoretical claims, *necessarily* requires the processing of data by statistical methods and computer simulations. For instance, as we have seen, the calculation of detector efficiencies (Section 5.2) involves the introduction of simulated data in the datasets which can be seen as a form of tampering the initial data. This step however, is taken to ensure that the calculated number of yields in the data fitting stage reflects the actual number of B-decays occurring in the collider and not the number of yields detected by the LHCb. Hence, the processing of data in some cases increases the reliability of the datasets in that it mitigates the impact of possible errors in the less processed datasets due to poor detector performance, computational limitations and so on.

Moreover, it is safe to say that the so-called raw data from these experiments, are not just epistemically less reliable than the processed data in some cases, but when it comes to their comparison with theoretical predictions, they are also practically useless in their pure form. The successful comparison of a theoretical hypothesis with data necessarily requires that the raw data extracted from the detectors are moulded into an appropriate form that makes them comparable to theoretical predictions in order to serve the purpose for which they are extracted. However, the raw data extracted from the first level of the triggering system are far from fulfilling this requirement. Hence, the seemingly counterintuitive claim that processed data are epistemically more reliable and more useful than the raw data obtained by experiments is actually a platitude when it comes to HEP.

²¹Bokulich’s comment on the blurriness of this distinction is characteristic: ‘I will not engage the difficult question here of where exactly to draw the line between (raw) data and a data model. It may very well be that the distinction is one of degree with vague boundaries, rather than a difference of kind; [...] and where the line is drawn may further be context dependent’ (2018, fn.25).

As for the distinction between simulated and real data in HEP, this has already been discussed in detail by Margaret Morrison (2015, Ch.8). Morrison uses the example of the Higgs discovery to emphasize the absolute necessity of simulation, not only in calculating the efficiency of a detector, but also in almost every other aspect of the LHC experiments. Her main conclusion is that given that simulation and signal data are essentially combined during the data analysis process, the sharp distinction between simulation and experiment is practically meaningless, and that ‘simulation is as much part of the experiment as the signal data’ (*ibid.*, p.289). Parker (2017) reaches a similar conclusion in her attempt to show that the results of computer simulations that are often embedded in measurement practices can be understood as measurement outcomes of equal epistemic importance to the outcomes of real measurements.

The calculation of the efficiencies via simulation provided in Section 5.2 is a clear example of such cases, where simulation results are actually embedded in real measurement outcomes in a way that makes it practically impossible to distinguish between the two. This example however, illustrates only one out of the many applications of simulation in a large-scale HEP experiment such as the LHCb. In addition to the calculation of the detector efficiency, simulation is also involved in the very early stages of the experiment to design and optimize the detectors for best physics performance, as well as in the calculation of the performance of the detector which is crucial for the extraction and interpretation of the available data (indeed, the numbers provided at the beginning of Section 4 regarding the performance of the LHCb detector can only be estimated by simulation). It is also heavily used for the estimation of background signal in the extracted data and the evaluation of the possible physical processes in the various parts of the sub-detector in order to assess their impact on the final data model via the calculation of uncertainties.²²

For the purposes of our discussion, it therefore suffices to say that although what counts as simulation data and what counts as signal data in the experiment is quite straightforward, the data that reaches the experimenters as reconstructed events for further analysis, is in effect an indistinguishable amalgamation of these two types. Along with a number of additional factors, simulation data therefore have a clear influence on the final properties of the data model either directly via their presence in the processed datasets that reach the scientists’ desks, or indirectly via their effects on the various aforementioned stages and procedures of the experiment. Although the discovery of a new particle or the presence of new physics in a physical process cannot – of course – be claimed based solely on simulation data, the final data model that is eventually compared to theory to make such claims is in effect a *co-production* of real and simulated data. The extent to which each type of data contributes to the final results depends on the specific details of the experiment. This further suggests that the question whether real data are more reliable than simulated data does not really apply in the case of HEP, since in practice, there is rarely a case in which a dataset is exclusively constructed from real data.

²²For a detailed review of the impact of simulation to collider experiments in general, including the discussion of cases where the use of simulation samples made a difference in the precision of the physics results, see Elvira (2017). For a description of the LHCb simulation system see Clemencic et al. (2011).

7 Concluding remarks

The description of the four stages in HEP data modelling and the following remarks on the two distinctions between raw/processed data and real/simulated data bring us to the end of our discussion. As we have seen, the construction of a data model in HEP typically proceeds via a four stage process in which (i) the selection criteria for reducing the available data are defined and applied at the trigger systems, (ii) the efficiency of the detector in recording the relevant events is calculated, (iii) the yields of the decays are determined by data fitting, and (iv) the uncertainties accompanying the final result are determined and calculated. It is important to note once again, that in practice, these four stages are not clearly separated during the data analysis, nor do they follow a linear path in which one stage follows after the completion of another. Rather, the activity of constructing a data model is a long and iterative process of trial and error, in which several attempted algorithms for extracting the result go back and forth a peer review process until they reach the necessary standards for publication. The breakdown of these procedures in four different stages only aims in giving an overview of the main tasks that need to be accomplished in order to compress the huge amount of information hidden in the available data into a simple data model to be compared with the theory.

Figure 4 perfectly captures Suppes' dictum that theoretical hypotheses are eventually compared with 'a simple entity' – i.e. the data model – which incorporates all the relevant information extracted from the many and various procedures that constitute the LHCb experiment. The graph is taken from the LHCb's most recent announcement of the result, and shows the comparison between the theoretical prediction of the Standard Model (vertical dashed line) and various experimental results for the R_K ratio (horizontal lines). Compared to the previous result at the LHCb Collaboration (2019) and the results from the Belle and BaBar experiments, the 2021 LHCb result by far has the smallest associated uncertainty which makes it the most precise and robust measurement of the R_K ratio to date.

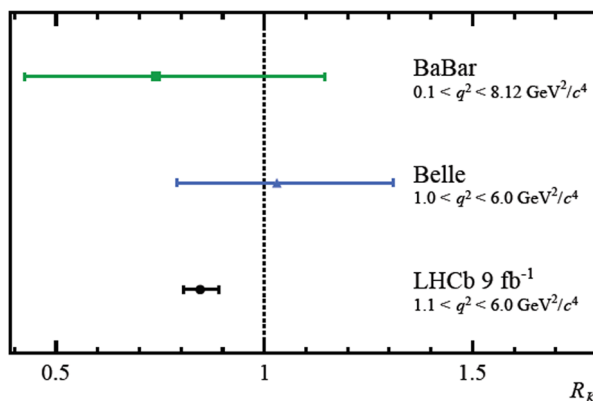


Fig. 4 Comparison of the Standard Model theoretical prediction and various R_K ratio results from different experiments (LHCb Collaboration, 2021)

This result is consistent with the Standard Model prediction at the level of 3.1 standard deviations, which corresponds to a p-value of $\sim 0.1\%$. In practice, this means that if the ‘null hypothesis’ is correct – i.e. if there is no violation of Lepton Flavour Universality – then the probability of obtaining any data yielding a discrepancy from the Standard Model prediction that is at least as great as that obtained with these data is about 1 in 100000. The 3.1 sigma level is still far away from the golden 5 sigma level for claiming a new discovery in particle physics, corresponding to the much lower p-value of approximately 1 in 3.5 million. This, however, is a significant improvement to the 2.5 standard deviation of the 2019 measurement with a p-value of 1 in 166, in that it comes with even smaller uncertainties and makes the possibility of discovering new physics in rare B-decays more credible. Future measurements of the ratio based on larger data samples are expected to both reduce the total uncertainty and increase the sigma level in order to reach a more definite conclusion for the possible violation of LFU in B-decays.

It is also worth noting that the plethora of experimental results showing potential anomalies at the 2-3 sigma level has already led to the development of various phenomenological models containing new physics in the form of additional interactions that allow the violation of LFU. The most promising types of such models involve the existence of additional particles such as the so-called ‘leptoquarks’ (Bečirević et al., 2016) or a new heavy neutral Z' boson (Celis et al., 2015). The precise way in which indirect searches in HEP, such as the measurement of the R_K ratio, give rise to new models extending the Standard Model of particle physics, and the impact of these models on future research in HEP is an interesting topic that deserves to be explored further in future work.

To summarise, the main objective of this article was to explore the connection of theory with experimental results via the concept of data models, by studying in detail an example of experimental practice in HEP. Our discussion began with a brief presentation of Suppes’ hierarchy of models account and his distinction between models of theory, models of experiment and models of data. The following section focused in providing the theoretical framework of the rare decays of B-mesons at the LHC in order to understand the experimental process of LFU tests at the LHCb for which the data model of the R_K ratio is constructed. The discussion continued with a presentation of the LHCb trigger system, followed by the presentation of the four main stages for the construction of the data model of the R_K ratio and some remarks on the two distinctions between raw/processed data and real/simulated data in support of the four main conclusions of this paper.

The first conclusion is that the first data collected at the early stages of the experiment, which can be characterised as the raw data of the experiment, are useless as they are for the comparison between theory and experimental results, since they necessarily need to undergo a process of refinement in order to be transformed into a language that is comparable to theory. This also indicates that raw data in HEP cannot be understood in the traditional sense as data directly perceived from human experience and that, contra to popular perception, the process of refining the data sometimes makes the processed datasets epistemically more reliable than non-processed data.

The second conclusion concerns the fact that the final datasets that reach the hands of physicists for analysis consist of a mixture of simulated and real data that cannot

be distinguished. The use of simulation and its data are essentially involved directly or indirectly in almost every step of the data acquisition and data analysis process and hence, one can safely say that the final data model of the R_K ratio that is eventually compared to the theoretical prediction of the Standard Model is a co-production of data coming from the physical interactions of particles in the detector and computer simulations.

The third conclusion is that theory guides the observation and the derivation of results in three different levels: a fundamental level which is universal across all experiments in HEP, an intermediate level regarding the various processes throughout the experiment which are not directly involved with the physical phenomenon under investigation, and a third and most specific level which explicitly guides the overall experimental procedure based on the specific research question of the experiment. In the core of these three levels lies the fundamental assumption that new physics will resemble known physics. This means that the anticipated models and theories that go beyond the Standard Model are expected to respect all the fundamental laws of current physics, and new physics will only appear in extremely short distance/high energy scales and in rare processes such as the decays of B-mesons which have not yet been studied in detail.

Finally, the detailed study of the LFU tests at the LHCb showed that Suppes' categorization is not as rigid as one might first think, in that the three types of models cannot always be easily distinguished. Nonetheless, this categorization remains a useful conceptual tool for describing the otherwise extremely complicated structure of large scale experiments in HEP. In this context, a data model can be understood as the representation of an experimental result in the form of a graph, table or numerical answer that allows the comparison of experiment with theory. While this straightforward answer to the question of what a data model is does not differ from what Suppes and others have said, what is of special philosophical interest is the complicated and extremely laborious process of constructing a data model in HEP, which has largely been overlooked by philosophers of science. The detailed analysis of the necessary considerations regarding the determination of cuts, the calculation of efficiencies and uncertainties and the fitting of data with sophisticated algorithms shows that the process of constructing a data model in HEP involves much more than the mere collection and organization of raw data, and cannot be easily formalized as Suppes implied.

In addition to these main conclusions, the detailed description of the idiosyncrasies of the LHCb experiment for the test of LFU and the various challenges faced by physicists in their attempt to derive the experimental results also reveals a number of further issues worthy of philosophical attention. The pragmatic dimension of the experimental process regarding the determination of selection criteria based on time limitations, computational power and store capacity, and the fact that the LHCb detector is able to collect only 25% of the b-quarks that are produced in the proton collisions of the LHC means that the otherwise huge amount of data that eventually gets stored for further analysis is only a tiny fraction of the potentially available data from the proton-proton collisions in the LHC. Although special attention is given to collect the most relevant data with respect to a research question, it is a widely

accepted fact that the data that are irretrievably thrown away at the LHC contain evidence for new physics and hence, the final data model of the R_K ratio, as well as most of the results in HEP, is not a solid and flawless representation of reality in the microscopic scale as one might think. Rather, it is itself a hypothesis based on our best estimation given the small fragment of data we are able to collect from particle collisions.

Acknowledgements I am greatly appreciative to Jonas Rademacker and Konstantinos Petrides from the Bristol Flavour Physics Group for their valuable insights on various aspects of the LHCb experiment at CERN and for their guidance through the vast literature on Lepton Flavour Universality tests. Many thanks also to Karim Thebault and Adam Toon for their constructive feedback on earlier versions of the paper, to the audience of the Boston graduate conference on the philosophy of data, and to two anonymous referees for their significant contributions to the improvement and finalisation of this work.

Funding This research was supported by an Arts and Humanities Research Council Ph.D. Studentship, awarded through the South, West and Wales Doctoral Training Partnership.

Declarations

Ethics approval This article does not contain any studies with human or animal subjects.

Informed consent This article does not contain any studies with human subjects.

Conflict of Interests The author declares that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ankeny, R., Chang, H., Boumans, M., & Boon, M. (2011). Introduction: philosophy of science in practice. *European Journal for Philosophy of Science*, 1(3), 303.
- Archilli, F., Baldini, W., Bencivenni, G., Bondar, N., Bonivento, W., Cadeddu, S., Campana, P., Cardini, A., Ciambone, P., & Vidal, X.C. (2013). Performance of the muon identification at LHCb. *Journal of Instrumentation*, 8(10), P10020.
- Bailey, D. C. (2017). Not Normal: the uncertainties of scientific measurements. *Royal Society Open Science*, 4(1), 160600.
- Barlow, R. (2002). Systematic errors: facts and fictions. arXiv:[hep-ex/0207026](https://arxiv.org/abs/hep-ex/0207026).
- Beauchemin, P.-H. (2017). Autopsy of measurements with the ATLAS detector at the LHC. *Synthese*, 194(2), 275–312.
- Bečrević, D., Fajfer, S., Košnik, N., & Sumensari, O. (2016). Leptoquark model to explain the B-physics anomalies, R_K and R_D . *Physical Review D*, 94(11), 115021.
- Bifani, S., Descotes-Genon, S., Vidal, A. R., & Schune, M.H. (2018). Review of Lepton Universality tests in B decays. *Journal of Physics G: Nuclear and Particle Physics*, 46(2), 23001.

- Blake, T., Lanfranchi, G., & Straub, D.M. (2017). Rare B decays as tests of the Standard Model. *Progress in Particle and Nuclear Physics*, 92, 50–91.
- Bogen, J., & Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, 97(3), 303–352.
- Bokulich, A. (2018). Using models to correct data: paleodiversity and the fossil record. *Synthese*, 1–22.
- Bokulich, A. (2020). Towards a taxonomy of the model-ladenness of data. *Philosophy of Science*, 87(5), 793–806.
- Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science*, 11(1), 1–26.
- Bordone, M., Isidori, G., & Pattori, A. (2016). On the standard model predictions for R_K and R_K^* . *The European Physical Journal C*, 76(8), 440.
- Brewer, W. F., & Lambert, B. L. (2001). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philosophy of Science*, 68(S3), S176–S186.
- Brown, H. I. (1993). A theory-laden observation can test the theory. *The British Journal for the Philosophy of Science*, 44(3), 555–559.
- Capdevila, B., Crivellin, A., Descotes-Genon, S., Matias, J., & Virto, J. (2018). Patterns of New Physics in $b \rightarrow s\ell^+\ell^-$ transitions in the light of recent data. *Journal of High Energy Physics*, 2018(1), 93.
- Celis, A., Fuentes-Martin, J., Jung, M., & Serodio, H. (2015). Family nonuniversal Z' models with protected flavor-changing interactions. *Physical Review D*, 92(1), 15007.
- Clemencic, M., Corti, G., Easo, C., Jones, R., Milgioranzi, C. S., Pappagallo, M., & Robbe, P. (2011). The LHCb simulation application, Gauss: Design, evolution and experience. *Journal of Physics: Conference Series*, 331, 32023.
- Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge: MIT Press.
- Elvira, V. D. (2017). Impact of detector simulation in particle physics collider experiments. *Physics Reports*, 695, 1–54.
- Franklin, A. (2015). The theory-ladenness of experiment. *Journal for General Philosophy of Science*, 46(1), 155–166.
- Franklin, A. (1998). Selectivity and the production of experimental results. *Archive for history of exact sciences*, 53(5), 399–485.
- Franklin, A., Anderson, B., Coleman, D., Gruvander, L., & Neal, P. (1989). Price Can a Theory-Laden Observation Test the Theory? *British Journal for the Philosophy of Science*, 40(2), 229–231.
- Frigg, R., & Hartmann, S. (2016). Models in science. In E. Zalta (Ed.) *Stanford encyclopedia of philosophy*. Spring 202 ed.
- Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Erkenntnis*, 52(1), 29–37.
- Head, T. (2014). The LHCb trigger system. *Journal of Instrumentation*, 9(09), C09015.
- Harris, T. (2003). Data models and the acquisition and manipulation of data. *Philosophy of Science*, 70(5), 1508–1517.
- Humair, T. (2019). Testing lepton universality in penguin decays of beauty mesons using the LHCb detector. Phd thesis, Imperial College London.
- Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context*, 26(1), 93–136.
- Karaca, K. (2017). A case study in experimental exploration: exploratory data selection at the Large Hadron Collider. *Synthese*, 194(2), 333–354.
- Karaca, K. (2018). Lessons from the Large Hadron Collider for model-based experimentation: the concept of a model of data acquisition and the scope of the hierarchy of models. *Synthese*, 195(12), 5431–5452.
- Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, 82(5), 810–821.
- Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago: University of Chicago Press.
- Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science*, 9(2), 22.
- LHCb Collaboration (2008). The LHCb detector at the LHC. *Journal of Instrumentation*, 3(8), S08005.
- LHCb Collaboration (2019). Search for Lepton-Universality Violation in $B^+ \rightarrow K^+ \ell^+ \ell^-$ Decays. *Physical Review Letters*, 122.
- LHCb Collaboration (2021). Test of lepton universality in beauty-quark decays. arXiv:2103.11769.
- Lionetto, F. (2018). Measurements of angular observables of $B^0 \rightarrow K^{*0} \mu^+ \mu^-$ and $B^0 \rightarrow K^{*0} e^+ e^-$ decays and the upgrade of LHCb. Phd thesis, University of Zurich.

- Lisovskyi, V. (2019). Study of rare b-baryon decays and test of Lepton Universality at LHCb. Phd thesis, Université Paris-Saclay.
- Lista, L. (2016). *Statistical methods for data analysis in particle physics*. New York: Springer.
- Lutz, S. (2017). What was the syntax-semantics debate in the philosophy of science about? *Philosophy and Phenomenological Research*, 95(2), 319–352.
- Lyons, L. (2006). *Statistical problems in particle physics astrophysics and cosmology*. Oxford: Imperial College Press.
- Mari, L., & Giordani, A. (2014). Modelling measurement: Error and uncertainty. In M. Boumans, G. Hon, & A. Petersen (Eds.) *Error and uncertainty in scientific practice* (pp. 79–96). London: Pickering & Gatto.
- Massimi, M. (2007). Saving unobservable phenomena. *British Journal for the Philosophy of Science*, 58(2), 235–262.
- Mauri, A. (2019). Direct and indirect searches for new physics in $b \rightarrow s\ell\ell$ Decays. Phd thesis, University of Zurich.
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
- McAllister, J. W. (1997). Phenomena and patterns in data sets. *Erkenntnis*, 47(2), 217–228.
- Morrison, M. (2015). *Reconstructing reality : models, mathematics, and simulations Oxford studies in philosophy of science*. New York: Oxford University Press.
- Muller, K. (2019). Tests of Lepton Flavour Universality at LHCb. *Journal of Physics Conference Series*, 1271.
- Parker, W. S. (2017). Computer simulation, measurement, and data assimilation. *The British Journal for the Philosophy of Science*, 68(1), 273–304.
- Rappoccio, S. (2019). The experimental status of direct searches for exotic physics beyond the standard model at the Large Hadron Collider. *Reviews in Physics* 4.
- Ritson, S., & Staley, K. (2020). How uncertainty can save measurement from circularity and holism. *Studies in History and Philosophy of Science Part A*.
- Schindler, S. (2013). Theory-laden experimentation. *Studies in History and Philosophy of Science Part A*, 44(1), 89–101.
- Sinervo, P. K. (2003). Definition and treatment of systematic uncertainties in high energy physics and astrophysics. *Statistical Problems in Particle Physics Astrophysics, and Cosmology*, 122–129.
- Staley, K. W. (2020). Securing the empirical value of measurement results. *The British Journal for the Philosophy of Science*, 71(1), 87–113.
- Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, 64, S65–S74.
- Suppes, P. (1960). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese*, 12(2-3), 287–301.
- Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.) *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261). Stanford University Press: Stanford.
- Tal, E. (2017a). A model-based epistemology of measurement. In N. Mößner, A. Nordman, & E. Tal (Eds.) *Reasoning in measurement* (pp. 233–253). Taylor & Francis.
- Tal, E. (2017b). Measurement in science. In E. N. Zalta (Ed.) *Stanford encyclopedia of philosophy*. (Fall 2017 ed).
- Teubert, F. (2016). *Rare decays probing physics beyond the standard theory*, (pp. 357–370). Singapore: World Scientific.
- van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon Press.
- van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.
- Virdee, T. S. (2016). Beyond the standard model of particle physics. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences* 374(2075).
- Wanke, R. (2016). How to deal with systematic uncertainties, *Data Analysis in High Energy Physics*, 263–296.
- Woodward, J. (1989). Data and phenomena. *Synthese*, 79(3), 393–472.