scientific data

OPEN

ARTICLE

A FAIR and AI-ready Higgs boson decay dataset

Yifan Chen[1,2], E. A. Huerta[2,3]✉, Javier Duarte[4], Philip Harris[5], Daniel S. Katz[1], Mark S. Neubauer[1], Daniel Diaz[4], Farouk Mokhtar[4], Raghav Kansal[4,5], Sang Eon Park[6], Volodymyr V. Kindratenko[1], Zhizhen Zhao[1] & Roger Rusack[7]

To enable the reusability of massive scientific datasets by humans and machines, researchers aim to adhere to the principles of findability, accessibility, interoperability, and reusability (FAIR) for data and artificial intelligence (AI) models. This article provides a domain-agnostic, step-by-step assessment guide to evaluate whether or not a given dataset meets these principles. We demonstrate how to use this guide to evaluate the FAIRness of an open simulated dataset produced by the CMS Collaboration at the CERN Large Hadron Collider. This dataset consists of Higgs boson decays and quark and gluon background, and is available through the CERN Open Data Portal. We use additional available tools to assess the FAIRness of this dataset, and incorporate feedback from members of the FAIR community to validate our results. This article is accompanied by a Jupyter notebook to visualize and explore this dataset. This study marks the first in a planned series of articles that will guide scientists in the creation of FAIR AI models and datasets in high energy particle physics.

Introduction

Much of the success of applications of artificial intelligence (AI) to a broad range of scientific problems[1,2] has been due to the availability of well-documented, high-quality datasets[3]; open source, state-of-the-art neural network models[4,5]; highly efficient and parallelizable numerical optimization methods[6]; and the advent of innovative hardware architectures[7].

Across science and engineering disciplines, the rate of adoption of AI and modern computing methods has been varied[2]. Throughout the process of harnessing AI and advanced computing, researchers have realized that the lack of an agreed upon set of best practices to produce, collect, and curate datasets has limited the combination of disparate datasets that with AI may reveal new correlations or patterns[8,9].

From 2014 to 2016, a set of data principles, or best practices, based on findability, accessibility, interoperability, and reusability (FAIR) were defined so that scientific datasets could be readily reused by both humans and machines. The FAIR principles can be applied to address these limitations and increase the potential of AI for discovery in science and engineering. Using high energy physics (HEP) as an example, this article provides a domain-agnostic, step-by-step set of checks to guide in the process of making a dataset FAIR ("FAIRification").

In HEP, there is a long history of the application of machine learning (ML) techniques to find small signals in the presence of large backgrounds. The observation of the Higgs boson at the CERN Large Hadron Collider (LHC) in 2012[10,11] was the result of the extensive use of ML algorithms based on boosted decision trees. Since then, as ML techniques have developed, their use in HEP has become ubiquitous. However, these developments have been largely the result of physicists adopting AI tools developed outside of their field of research.

The authors of this paper are members of the FAIR4HEP collaboration which has representation from the AI community and two of the large LHC collaborations, ATLAS and CMS. We are collaborating to prepare datasets from HEP experiments that meet FAIR data principles[12]. There are several major impediments to this strategy, including, among others, the lack of jargon-free documentation, difficulty of access to, and poor structure of the dataset, and the lack of clear metrics with which to benchmark and compare AI models. A consequence of the FAIR data principles is that they promote the use of open datasets, which in turn supports collaboration

[1]University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA. [2]Argonne National Laboratory, Lemont, Illinois, 60439, USA. [3]University of Chicago, Chicago, Illinois, 60637, USA. [4]University of California San Diego, La Jolla, California, 92093, USA. [5]Halıcıoğlu Data Science Institute, La Jolla, California, 92093, USA. [6]Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA. [7]The University of Minnesota, Minneapolis, Minnesota, 55405, USA. ✉e-mail: elihu@anl.gov

between practitioners of different disciplines (in this case, high-energy physicists and AI researchers) who have overlapping interests in particular datasets. We note that these impediments are common to many disciplines.

The FAIR guiding principles[12], published in 2016, provide guidelines to improve the "FAIRness" of digital assets, such as datasets, code, and research objects. While the principles are valuable for the scientific and engineering fields, they do not include exemplar metrics[13] or define ways to measure how well the FAIR data principles are met for the digital asset.

In HEP, there are currently several efforts available for creating, indexing, and sharing open, public datasets. The CERN Open Data Portal provides access to data and resources from the four major LHC collaborations, ALICE, ATLAS, CMS and LHCb, for both education and research. Previous data releases have already yielded publications using LHC data authored by external researchers unaffiliated with an LHC collaboration[14–19]. However, despite being a repository of LHC open data, it does not allow general members of the HEP research community to upload their own datasets. Zenodo is another platform launched in May 2013, which is part of the OpenAIRE project, in partnership with CERN, that is a catch-all repository for European Commission funded research and is used widely. It allows the community at large to upload data, software, and other artefacts in support of publications, as well as material associated with conferences, projects, or institutions. Citation information is also passed to DataCite and other scholarly aggregators. Zenodo has been used to host several high-profile public HEP datasets including the top quark tagging reference dataset[20], LHC Olympics 2020 Anomaly Detection Challenge dataset[21], hls4ml jet substructure dataset, and the Anomaly Detection Data Challenge 2021 dataset[22]. Other services like Kaggle and Codalab have been used to host HEP challenges like the TrackML accuracy phase[23] and TrackML throughput phase[24]. The Durham High-Energy Physics Database (HEPData)[25] is an open-access repository established for sharing scattering data from experimental particle physics. It mainly comprises the data points from plots and tables related to several thousand publications including those from the LHC.

Despite the widespread availability of public datasets in HEP, these services, and the datasets they host, do not follow FAIR principles. In particular, the interpretation of the FAIR principles in the context of the large datasets available and the specific computing infrastructure needs in HEP is not clear. For instance, the CERN Open Data Portal hosts datasets with sizes approaching 100 TB[26], requiring special versions of software (provided through a virtual machine image) to read and analyze the data. Given these stringent computational, storage, and domain knowledge requirements, the accessibility of these datasets to non-experts and those lacking resources is not completely obvious. To explore how to address these difficulties, we present an analysis of the FAIRness of one of these datasets, the CMS $H(b\bar{b})$ dataset.

This simulated collider dataset contains a selection of proton-proton interactions (events) in which a Higgs boson is produced and decays to two bottom quarks $H(b\bar{b})$ (signal events) as well as background events comprised uniquely of "jets" of particles produced through the strong interaction, referred to as quantum chromodynamics (QCD) multijet events. This dataset was released in the CERN Open Data Portal. By providing the details of the how we evaluate FAIRness of this dataset and the steps taken to meet the FAIR data principles, we can help researchers in other fields create FAIR datasets in a similar manner. To ensure the reliability of our results we have conducted a similar study using the ARDC FAIR self assessment tool. We have found that the steps we have followed and the ARDC assessment tool provide consistent results.

In the following sections we describe how the FAIRness of this dataset is evaluated and present the result in the format of a set of checks. We also describe methods to improve FAIRness, provide a detailed data description, and discuss how we interpret FAIR principles for HEP.

Results

We have assessed the FAIRness of the target dataset, described in the previous section and in more detail below using the related FAIR metrics[13]. The results of this analysis are summarized in Tables 1 and 2. The following subsections summarize the results for each principle, and discuss steps that were, or will be, take to increase the FAIRness of the dataset. We also highlight the difficulties inherent in interpreting and applying these principles to HEP datasets, due to their unique properties of size, complexity, data format, and required domain knowledge.

Findable. The findable principle requires that metadata and data should be easy to find for both humans and machines. For this specific dataset: 1) both data and metadata are registered with globally unique and persistent identifiers; 2) the association between metadata and the dataset itself is explicitly described in its metadata, and 3) the dataset is registered as a searchable resource and is searchable on a commonly used search engine. However, though searchable, the metadata fields are fairly sparse and information is lacking. Enriching the metadata with additional fields to include references that cite this dataset, or links to related or derived datasets, would make the data more readily available.

Accessible. To meet the FAIR accessible principle, data are required to be kept in a storage facility where they may be easily accessed or downloaded, with well-defined license and access conditions, which should be open access whenever possible, either at the level of metadata, or at the level of the actual data content.

The CMS $H(b\bar{b})$ dataset is retrievable using standard HTTP communication protocols, is open access, and is under the Creative Commons public domain dedication (license). Since the DOI has formal metadata, it satisfies the metadata longevity plan.

Interoperable. The interoperable principle requires that the data can be readily combined with other datasets by humans as well as by computer systems. For this dataset, (meta)data are represented using a formal and broadly applicable representation language. To improve the interoperability, the data descriptions were rewritten to be human-readable, removing jargon to make it accessible not only for domain experts, but also

Metric	Evaluation
F1. (Meta)data are assigned globally unique and persistent identifiers.	
Identifier Uniqueness: this metric measures whether there is a scheme to uniquely identify the digital resource.	**Pass**. The DOI for the data (which resolves to a URL[29]) follows a registered identifier scheme.
Identifier Persistence: this measures whether there is a policy that describes what the provider will do in the event an identifier scheme becomes deprecated.	**Pass**. The use of a DOI provide a persistent interoperable identifier.
F2. Data are described with rich metadata.	
Machine-readability of Metadata: to meet this metric, a URL to a document containing machine-readable metadata for the digital resource must be provided.	**Pass**. The URL for the metadata[57] in JSON Schema with REST API is available. The use of JSON Schema provides clear human and machine readable documentation. Also, running the URL through the Rich Result Test shows the data page contains rich results.
Richness of Metadata: data are described with rich metadata	**Partially pass**. Reviewing the DataCite metadata for the DOI shows a fairly sparse record. The metadata can be improved with richer fields.
F3. Metadata clearly and explicitly include the identifier of the data they describe.	
Resource Identifier in Metadata: this measures if the metadata document contains the identifier for the digital resource that meets F1 principle.	**Pass**. The association between the metadata and the dataset is made explicit because the dataset's globally unique and persistent identifier can be found in the metadata. Specifically, the DOI is a top-level and a mandatory field in the metadata record.
F4. (Meta)data are registered or indexed in a searchable resource	
Index in a searchable resource: this measures the degree to which the digital resource can be found using web-based search engines	**Pass**. The dataset is indexed by Google Dataset Search engine.
A1. (Meta)data are retrievable by their identifier using a standardized communications protocol	
A1.1: The protocol is open, free and universally implementable	
Access Protocol: it measures whether the URL is open access and free.	**Pass**. HTTP get on the identifier's URL returns a valid document
A1.2. The protocol allows for an authentication and authorization where necessary	
Access Authorization: it requires specification of a protocol to access restricted content.	**Pass**. This is an open dataset, accessible to everyone on the internet. The data is non-profit and privacy-unrelated, so no access authorization is needed.
A2. Metadata should be accessible even when the data is no longer available	
Metadata Longevity: it requires metadata to be present even in the absence of data	**Pass**. Metadata is stored separately in the CERN Open Data server. As per FAIR Principle F3, this metadata remains discoverable, even in the absence of the data, because it contains an explicit reference to the DOI of the data. Data and metadata will be retained for the lifetime of the repository. The host laboratory CERN, currently plans to support the repository for at least the next 20 years.

Table 1. Findable and Accessible principle assessment checks for the CMS H($b\bar{b}$) Open Dataset.

non-HEP researchers. The (meta)data use a set of FAIR vocabularies defined for both general purpose and HEP domain-related purpose. Although not all terms are findable in FAIR vocabularies, those that are not findable are well-defined and referenced. Lastly, the description of this dataset provides references to other datasets from which it is derived. However, a more extensive set of references that elaborate on the paper describing this dataset, and more information about the methods used to derive this dataset, could be added to aid in the comprehension of the problem to be addressed with this dataset.

Reusable. Reusability requires the data to be readily usable for future research and to be able to be processed further using different computational methods. We found that the metadata and data of this dataset are well-described with accurate and relevant attributes. Thus, we anticipate that the dataset will be reusable and can be integrated with additional data in future studies.

Methods

In this section we describe our approach to evaluate FAIRness of the CMS H($b\bar{b}$) dataset, and provide a human-readable description of the HEP dataset contents and its overall structure. These two complementary aspects of the dataset are critical elements in any pursuit of data FAIRification.

Dataset FAIRification. We have created a set of ready-to-use, domain-agnostic checks to facilitate the evaluation of how well a dataset meets the FAIR guiding principles[12], and applied them them to the H($b\bar{b}$) dataset. These checks provide researchers with a tool that can be used to assess the FAIRness of scientific datasets, and thus will streamline the use of such datasets for AI-driven analyses.

We have used the ARDC FAIR self assessment tools, developed by other researchers in the FAIR community, to validate our findings. We have also incorporated human-in-the-loop expertise in this process in the form of feedback from FAIR experts, who independently validated our results.

Dataset description. The CMS H($b\bar{b}$) Open Dataset consists of two data samples that have been a critical part of the understanding of physical phenomena associated with the Higgs boson. The Higgs boson, first observed at the LHC in 2012[10,11], is an elementary particle that is related to the Higgs mechanism for electroweak symmetry breaking, responsible generating the masses of the elementary particles.

Metric	Evaluation
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	
Use a Knowledge Representation (programming) Language: use a formal, accessible, shared, and broadly applicable language for knowledge representation	**Pass**. As described in Section 3, this dataset is represented based on the ROOT framework with Python interface. The notebook we release with this manuscript provides the required tools to handle this dataset using HDF5. The metadata is represented following the JSON Schema draft 4. Both are widely used formats in Physics.
Provide Human-readable descriptions	**Pass**. The description and data semantics of this dataset provides rich information on how to use the dataset.
I2. (Meta)data use vocabularies that follow FAIR principles.	
Use FAIR Vocabularies: it requires the metadata values and qualified relations should be FAIR themselves, that is, terms should be findable from open, community-accepted vocabularies.	**Partially pass**. I2 requires the controlled vocabulary used to describe datasets to be documented and resolvable using globally unique and persistent identifiers. For domain-specific terms, we leverage a vocabulary PhySH (Physics Subject Headings), a physics classification scheme developed by American Physical Society (APS). Some terms in dataset descriptions and semantics are registered in PhySH. However, since PhySH is still under development, there is not very good coverage of the narrower experimental concepts. For the terms not covered, references and hover definitions are provided. For general terms, the metadata follows the vocabulary from JSON Schema and a minimal set of FAIR terms are used.
I3. (Meta)data include qualified references to other (meta)data.	
Use Qualified References: The goal is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data.	**Partially pass**. There are connections with other datasets. A list of derived datasets is available at the dataset site [27]. Each referenced external piece of dataset is qualified by a resolvable URL and a unique CERN data identifier in metadata. To improve, the papers of these related data can be provided, from which more information about methods and workflow used to derive this dataset can be retrieved, and external datasets should be references by permanent identifiers rather than URLs.
R1.1. (Meta)data are released with a clear and accessible data usage license.	
Accessible Usage License: the existence of license document for (meta)data are being measured	**Pass**. This dataset is released under Creative Commons CC0 dedication. The license field is present in the metadata.
R1.2. (Meta)data are associated with detailed provenance.	
Detailed Provenance: Who / What / When produced the data? Why / How was the data produced?	**Pass**. The dataset is derived from other data, e.g.[58,59], using public software[60] that was made public to process and reduce it. We are able to track the original authors and data sources. But ideally, this workflow would be described in a machine-readable format.
R1.3. (Meta)data meet domain-relevant community standards.	
Meet Community Standards: it measures whether a certification of the resource meeting community standards exists.	**Pass**. Both metadata and data meet the CERN Open Data community standards and thus have been released on the CERN Open Data repository.

Table 2. Interoperable and Reusable principle assessment checks for CMS H($b\bar{b}$) Open Dataset.

One consequence of the Higgs mechanism is that the Higgs boson, which has a lifetime of only $\approx 10^{-22}$ seconds, couples to other particles in proportion to their mass and therefore will decay preferentially to elementary particles with comparatively higher masses. The H($b\bar{b}$) decay process is particularly important because the b quark is the most massive quark to which the Higgs boson can decay. By measuring precisely the rate of this decay process, the physics of the coupling between the Higgs boson and ordinary matter can be tested. Any significant deviations from the predicted values would be an indication of physics beyond the standard model of particle physics.

When a Higgs boson decays to b quarks, the quarks, which cannot be free in nature, are detected as clusters of particles moving away from the interaction vertex (jets) and recognized by a secondary decay vertex from a particle containing a b quark a short distance from the interaction. Collisions, or interactions between protons in the two circulating beams (events) occur at a rate of about 1 GHz, while the rate of production of Higgs bosons is only 0.001 Hz, about one every hour. The challenge of identifying Higgs bosons decaying to $b\bar{b}$ is to find them amid the much larger number of collisions (background) where a Higgs boson is not produced. In these background events, typically referred to as quantum chromodynamics (QCD) multijet events, a large number of particles are produced, which may include jets from b quarks, and can combine to resemble H($b\bar{b}$) events, which are the "signal" events.

To identify Higgs boson decays and separate them from the much larger QCD background, we use several key reconstructed components of proton-proton collisions. In particular, we reconstruct jets and analyze their characteristics which include tracks, secondary vertices (SVs), and substructure features. We also employ a particle-flow (PF) algorithm[27] to provide a comprehensive list of final-state particles that are identified and reconstructed via combination of information from multiple detector subsystems.

The following defines these elements:

- **Jets** are sprays of elementary particles in a cone-shaped pattern that radiate out from the collision vertex. They may be characterized by their *substructure*, including features like the jet mass, charge, and shape[28]. In total, the dataset contains 64 reconstructed jet features. These features are not necessarily independent from one another, and they may be derived from lower-level features related to the tracks, PF candidates, and secondary vertices.

- **Tracks** are the reconstructed helical paths of charged particles as they move away from the collision vertex in the magnetic field at the detector. Each charged particle leaves a characteristic set of hits in the tracking detector of CMS, which are used to reconstruct the track. In total, there are 45 track features.
- **Secondary vertices** are collections of tracks that originate from a particle decay that is not at the collision vertex. Secondary vertices are a interesting set of candidate features that discriminate between different classes of jets, because they are dominant signatures in bottom-quark decays. In total, there are 18 SV features.
- **Particle-flow candidates** are formed by combining tracks and clusters from other detectors outside of the tracking detector. The PF algorithm[27] is used to provide a complete event description through the generation of a comprehensive list of the particles produced in the collision. For each PF candidate, there are 24 features.

This dataset consists of particle jets extracted from simulated proton-proton collision events generated with a center-of-mass energy of 13 TeV. Each element of the dataset corresponds to a single jet, containing information about the jet, from jet-level features to track-level features (see later for the full details on the dataset).

The outcome of the default CMS reconstruction workflow is provided in the open simulation[29]. In particular, particle candidates are reconstructed using the particle-flow (PF) algorithm[27]. Particles produced nearly simultaneously with the events that leave extra hits in the detector (pileup) are removed with an algorithm developed for that purpose[30]. Jets are clustered from the remaining reconstructed particles[31,32] with a jet-size parameter $R = 0.8$ (AK8 jets). The standard CMS jet energy corrections are applied to the jets. In order to remove soft, wide-angle radiation from the jet, the soft-drop (SD) algorithm[33,34] is applied, with angular exponent $\beta = 0$, soft cutoff threshold $z_{cut} < 0.1$, and characteristic radius $R_0 = 0.8$[35]. The SD mass (m_{SD}) is then computed from the four-momenta of the remaining constituents.

The dataset is reduced by requiring the AK8 jets to have $300 < p_T < 2400$ GeV, $|\eta| < 2.4$, and $40 < m_{SD} < 200$ GeV. After this reduction, the dataset consists of 3.9 million H($b\bar{b}$) jets and 1.9 million QCD jets. Charged particles are required to have $p_T > 0.95$Ge and reconstructed secondary vertices (SVs) are associated with the AK8 jet using $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.8$. The dataset is divided into blocks of features, referring to different objects: tracks, secondary vertices, and particle candidates. See the CERN Open Data Portal for a complete list of features.

A typical use case for this dataset is the development of a machine-learning classifier to distinguish the H($b\bar{b}$) signal from the QCD background jets, which in ML terms, can be done via a binary-classification task. However, it is often useful to further classify the QCD jets, thereby, the task becomes multi-class classification with the following six jet classes: `H_bb`, `QCD_bb`, `QCD_cc`, `QCD_b`, `QCD_c`, and `QCD_others`. The labeling is performed sequentially. If a 'generator-level' Higgs boson is geometrically matched to the AK8 jet ($\Delta R < 0.8$) and the two bottom quark decay products are also matched to the jet, then it is labeled as `H_bb`. If instead, only two bottom (charm) quarks are found, the jet is labeled as `QCD_bb` (`QCD_cc`). If only a single bottom (charm) quark is found, it is labeled as `QCD_b` (`QCD_c`). Finally, if none of the above conditions are met, it is labeled as `QCD_others`. The distribution of labels is shown in Fig. 1. The large class imbalance is a common feature of classification problems in high energy physics: background jets occur at much larger rates than signal jets.

Specific signatures of b quark decays can be used in a ML algorithm to differentiate between H($b\bar{b}$) and QCD jets. For instance, one of the distinct signatures of b quarks is its long lifetime, which in a high energy collision translates to a particle that decays with a displacement with respect to the collision. The model can learn this information to improve the accuracy of the inference. An illustration of some of key features that can be used for H($b\bar{b}$) jet tagging are shown in Fig. 2. The distributions of some salient jet features are shown in Fig. 3.

Many different deep learning architectures have been developed and studied for the task of jet classification, such as: interaction networks (INs)[36], dynamic graph convolutional neural networks[37], and Lorentz-group equivariant networks[38]. The first was applied to this data[39] as a comparison with another ML model called the deep double-b (DDB) tagger created by the CMS Collaboration[40] that uses a smaller subset of the input features. In addition to jet classification[40–43], a further challenge within this dataset is a regression task, whereby one attempts to reconstruct the true energy of the Higgs boson. To perform this task, a regression loss needs to be constructed targeting the true Higgs boson energy. This promotes the exploration of physics-motivated loss functions, such as the earth (or energy) mover's distance (EMD)[18].

Dataset structure. Particle physics uses a variety of data formats (and analysis ecosystems), including the `ROOT` library[44]. `ROOT` is a framework for data processing, created at CERN that is widely used by the high-energy physics community. A `ROOT` file is a compressed binary file where objects of any type can be saved. There are Python bindings built into `ROOT`, which are called `PyROOT`. Recently, an additional library called uproot[45] has been developed that allows Python users to perform `ROOT` I/O directly. Unlike the standard `C++ ROOT` implementation, uproot is only an I/O library, primarily intended to stream data into machine learning libraries in Python. It can also make jagged or awkward arrays[46].

Trees are a data structure in `ROOT` that are tables of information. Trees are composed of *branches*, which are the columns of the table. In this dataset, each row represents a jet. Some branches contain only a single floating point number per entry (jet). Other branches contain a vector of floating point numbers, where the length of the vector varies for each entry. The former means there is only one number per jet (or event); the latter means there may be a variable number per jet.

In addition to the `ROOT` format, this dataset is also released in `HDF5` format. The `HDF5` files contain different arrays for each output variable, with only information for up to 100 particle candidates, 60 tracks, and 5 secondary vertices stored in zero-padded arrays.

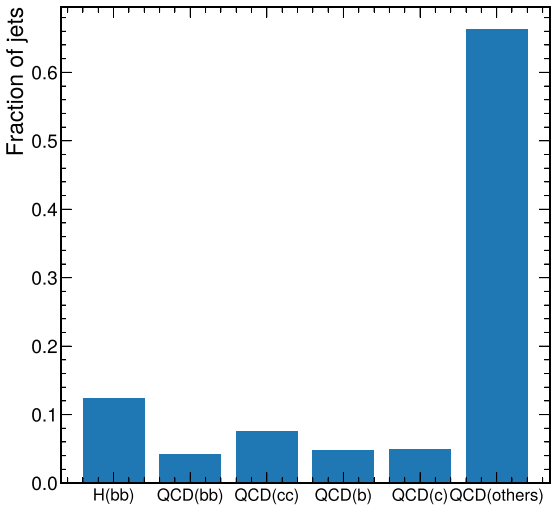

Fig. 1 The distribution of labels is shown for a representative file in the training dataset.

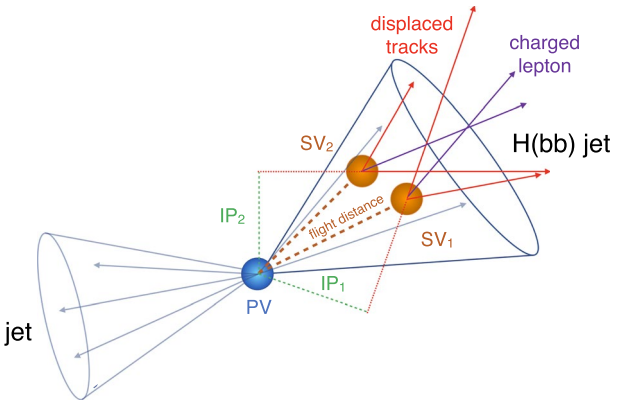

Fig. 2 Illustration of a $H(b\bar{b})$ jet with two secondary vertices (SVs) from the decay of two b hadrons resulting in charged-particle tracks (including a low-energy, or soft, lepton) that are displaced with respect to the primary collision vertex (PV), and hence with a large impact parameter (IP) value.

Discussion

The motivation of our work to adopt FAIR principles for the production, collection and curation of scientific datasets is to streamline and facilitate their use in the design, training, validation and testing of AI models. This approach is particularly relevant for ongoing efforts that aim to automate the inference of massive scientific datasets through the convergence of AI and modern computing environments[47,48]. It is often the case that AI models are trained with (abundant and easy to produce) synthetic data, large scale simulations, and first-principles mathematical models, although these may only provide an incomplete description of complex and highly nonlinear real-world phenomena. Thus, when AI models are used to extract new knowledge from realistic, experimental datasets, it is a common occurrence that AI predictions are off-target. However, once AI models are calibrated against experimental data, their predictions become increasingly accurate[49]. Given that this is a trend reported across many disciplines, it is useful to streamline the development of AI models with real, experimental datasets. This can be accomplished if synthetic and experimental datasets are produced, collected and curated following a common set of standards or, in this case, FAIR guiding principles.

Another motivation to understand and adopt FAIR principles to create AI-ready datasets is that some disciplines are subject to restrictive regulations that prevent data fusion and centralized analyses. This is a common issue in multi-modal biomedical datasets that are governed by federal regulations, consortium-specific data usage agreements, and institutional review boards. These restrictions have catalyzed the development of federated learning approaches, and the development of privacy-preserving methods and the use of secure data enclaves. It is clear that developing AI models by harnessing disparate data enclaves will only be feasible if datasets adhere to a common set of rules, or FAIR principles.

In this study we have shown that open source datasets may not be FAIR or AI-ready. The domain-agnostic checks that we provide in this article will provide researchers with a starting point, and guidance to FAIRify their datasets. The FAIR principles are comprehensive and can be used by AI and domain experts to enable the

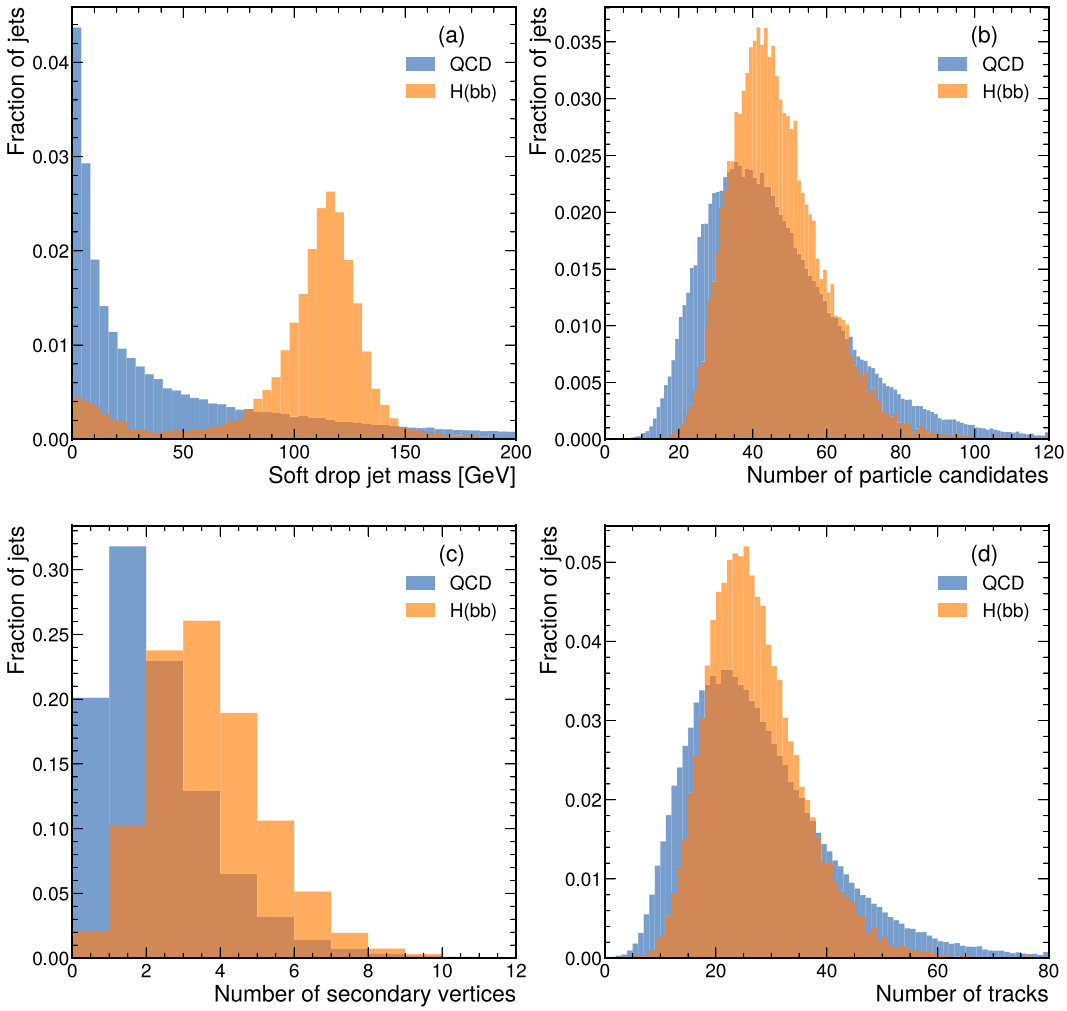

Fig. 3 The distributions of some salient jet features: **(a)** the soft-drop jet mass; **(b)** number of particle candidates; **(c)** number of secondary vertices; and **(d)** number of tracks, are shown for one file in the training dataset.

reusability of massive scientific datasets that will enable the creation of next-generation AI models leading to a digitally accurate, interpretable and reproducible description of natural phenomena.

Researchers who create FAIR datasets should keep in mind that this work aims to automate end-to-end AI studies, from data collection to inference. This will only be accomplished if datasets contain all the information needed to interpret, verify, and reproduce new findings. To accomplish this goal, we recommend that datasets are stored using formats that are widely available in modern computing environments, such as `HDF5` or `ROOT`. Using such data formats simplify the handling of large datasets, will allow experimental and synthetic datasets to be on the same footing, and will make accessible more datasets for widely used APIs for AI research, e.g., `TensorFlow` or `PyTorch`. In future work, we plan to introduce tools to automate the evaluation of FAIR metrics for datasets and to gain a better understanding of the relationship between data and AI models.

Data Formatting

The technical details of the simulation of the events and their selection for inclusion in the dataset are described in this section. The dataset consists of a signal model containing H($b\bar{b}$) jets available from simulated events containing the postulated Randall-Sundrum gravitons[50] that decay to two Higgs bosons, and thence to $b\bar{b}$ pairs.

The event generation was done by the CMS Collaboration with MADGRAPH5_aMCATNLO 2.2.2 at leading order, with graviton masses ranging between 0.6 and 4.5 TeV. Generation of this process enables better sampling of events where the Higgs boson is produced with a large lateral momentum component (p_T). The background dataset was generated with PYTHIA 8.205[51] in different bins of the average p_T of the final-state partons ($\hat{p}_T$). The parton showering and hadronization was also performed with PYTHIA 8.205, using the CMS underlying event tune CUETP8M1[52] and the NNPDF 2.3[53] parton distribution functions. Pileup interactions are modeled by overlaying each simulated event with additional minimum bias collisions, also generated with PYTHIA 8.205. The CMS detector response is modeled by GEANT4[54].

Data are composed of a set of characteristic variables relating to several broad types of objects consisting of event-level identifiers, features of charged particle tracks, secondary vertices, and particle-flow candidates, high level jet observables, and additional generator-level information, such as jet labels. There are 3 event-level features, 45 charged particle features, 18 secondary vertex features, 24 particle-flow candidate features, 64 high-level jet features, and 18 generator-level identifiers. For each of these variables, a detailed description is present on the CERN Open Data Portal. For the `HDF5` format, information is stored for up to 100 particle-flow candidates, with a maximum of 60 charged particle tracks, and up to 5 secondary vertices. In the instance where there are less candidates, inputs are zero-padded.

Data availability

The $H(b\bar{b})$ data for this work is available on the CERN Open Data Portal[29], in both `ROOT` and `HDF5` formats.

Code availability

To make the CMS $H(b\bar{b})$ Open Dataset more accessible, we provide notebooks[55] from the course "Particle Physics and Machine Learning" at University of California San Diego. The course notebooks provide a guide for the use of the `ROOT` format dataset. We also have released a second set of interactive Jupyter Notebooks on GitHub[56], where we visualize feature distributions and feature correlations, and provide machine learning examples on low-level features in this dataset. The Jupyter notebooks that we released show how the `HDF5`-formatted data can be accessed.

Received: 1 September 2021; Accepted: 15 December 2021;
Published online: 14 February 2022

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat* **521**, 436, https://doi.org/10.1038/nature14539 (2015).
2. Huerta, E. A. *et al.* Enabling real-time multi-messenger astrophysics discoveries with deep learning. *Nat Rev. Phys.* **1**, 600, https://doi.org/10.1038/s42254-019-0097-4 (2019).
3. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
4. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770, https://doi.org/10.1109/CVPR.2016.90 (2016).
5. van den Oord, A. *et al.* WaveNet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, 125 (2016).
6. Shamir, O. & Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Dasgupta, S. & McAllester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, 71–79 (PMLR, Atlanta, Georgia, USA, 2013).
7. Vázquez, F., Martínez, J. A. & Garzón, E. M. *GPU Computing*, 845–849 (Springer New York, New York, NY, 2013).
8. Wei, W. *et al.* Deep transfer learning for star cluster classification: I. application to the PHANGS-HST survey. *Mon. Not. R. Astron. Soc.* **493**, 3178–3193, https://doi.org/10.1093/mnras/staa325 (2020).
9. Whitmore, B. C. *et al.* Star cluster classification in the PHANGS-HST survey: Comparison between human and machine learning approaches. *Mon. Not. R. Astron. Soc.* **506**, 5294–5317, https://doi.org/10.1093/mnras/stab2087 (2021).
10. Aad, G. *et al.* Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716**, 1, https://doi.org/10.1016/j.physletb.2012.08.020 (2012).
11. Chatrchyan, S. *et al.* Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716**, 30, https://doi.org/10.1016/j.physletb.2012.08.021 (2012).
12. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
13. Wilkinson, M. D. *et al.* A design framework and exemplar metrics for FAIRness. *Sci Data*, https://doi.org/10.1038/sdata.2018.118 (2018).
14. Tripathee, A., Xue, W., Larkoski, A., Marzani, S. & Thaler, J. Jet Substructure Studies with CMS Open Data. *Phys. Rev. D* **96**, 074003, https://doi.org/10.1103/PhysRevD.96.074003 (2017).
15. Larkoski, A., Marzani, S., Thaler, J., Tripathee, A. & Xue, W. Exposing the QCD Splitting Function with CMS Open Data. *Phys. Rev. Lett.* **119**, 132003, https://doi.org/10.1103/PhysRevLett.119.132003 (2017).
16. Andrews, M., Paulini, M., Gleyzer, S. & Poczos, B. End-to-end physics event classification with CMS open data: Applying image-based deep learning to detector data for the direct classification of collision events at the LHC. *Comput. Softw. Big Sci.* **4**, 6, https://doi.org/10.1007/s41781-020-00038-8 (2020).
17. Andrews, M. *et al.* End-to-end jet classification of quarks and gluons with the CMS open data. *Nucl. Instrum. Meth. A* **977**, 164304 (2020).
18. Komiske, P. T., Metodiev, E. M. & Thaler, J. Metric space of collider events. *Phys. Rev. Lett.* **123**, 041801, https://doi.org/10.1103/PhysRevLett.123.041801 (2019).
19. Komiske, P. T., Mastandrea, R., Metodiev, E. M., Naik, P. & Thaler, J. Exploring the space of jets with CMS open data. *Phys. Rev. D* **101**, 034009, https://doi.org/10.1103/PhysRevD.101.034009 (2020).
20. Butter, A. *et al.* The Machine Learning landscape of top taggers. *SciPost Phys.* **7**, 014, https://doi.org/10.21468/SciPostPhys.7.1.014 (2019).
21. Kasieczka, G. *et al.* The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics. *Reports on Prog. Phys.* (2021).
22. Govorkova, E. *et al.* LHC physics dataset for unsupervised New Physics detection at 40 MHz. https://arxiv.org/abs/ (2021).
23. Amrouche, S. *et al. The Tracking Machine Learning challenge: Accuracy phase*. https://arxiv.org/abs/1904.06778 (2019).
24. Amrouche, S. *et al.* The Tracking Machine Learning challenge: Throughput phase. https://arxiv.org/abs/2105.01160 (2021).
25. Maguire, E., Heinrich, L. & Watt, G. HEPData: a repository for high energy physics data. *J. Phys. Conf. Ser.* **898**, 102006, https://doi.org/10.1088/1742-6596/898/10/102006 (2017).
26. CMS Collaboration. VBF1Parked primary dataset in AOD format from Run C of 2012 (/VBF1Parked/Run2012C-22Jan2013-v1/AOD). *CERN Open Data Portal* https://doi.org/10.7483/OPENDATA.CMS.4P88.F4RS (2012).
27. CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector. *JINST* **12**, P10003, https://doi.org/10.1088/1748-0221/12/10/P10003 (2017).
28. Thaler, J. & Van Tilburg, K. Identifying Boosted Objects with N-subjettiness. *JHEP* **03**, 015, https://doi.org/10.1007/JHEP03(2011)015 (2011).

29. CMS Collaboration, Duarte, J. Sample with jet, track and secondary vertex properties for Hbb tagging ML studies (HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC). *CERN Open Data Portal*. https://doi.org/10.7483/OPENDATA.CMS.JGJX.MS7Q (2019).
30. Sirunyan, A. M. *et al*. Pileup mitigation at CMS in 13 TeV data. *JINST* **15**, P09018, https://doi.org/10.1088/1748-0221/15/09/P09018 (2020).
31. Cacciari, M., Salam, G. P. & Soyez, G. The anti-k_T jet clustering algorithm. *JHEP* **04**, 063, https://doi.org/10.1088/1126-6708/2008/04/063 (2008).
32. Cacciari, M., Salam, G. P. & Soyez, G. FastJet user manual. *Eur. Phys. J. C* **72**, 1896, https://doi.org/10.1140/epjc/s10052-012-1896-2 (2012).
33. Dasgupta, M., Fregoso, A., Marzani, S. & Salam, G. P. Towards an understanding of jet substructure. *JHEP* **09**, 029, https://doi.org/10.1007/JHEP09(2013)029 (2013).
34. Butterworth, J. M., Davison, A. R., Rubin, M. & Salam, G. P. Jet substructure as a new Higgs search channel at the LHC. *Phys. Rev. Lett.* **100**, 242001, https://doi.org/10.1103/PhysRevLett.100.242001 (2008).
35. Larkoski, A. J., Marzani, S., Soyez, G. & Thaler, J. Soft drop. *JHEP* **05**, 146, https://doi.org/10.1007/JHEP05(2014)146 (2014).
36. Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D. & Kavukcuoglu, K. Interaction networks for learning about objects, relations and physics. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29 (Curran Associates, Inc., 2016).
37. Qu, H. & Gouskos, L. ParticleNet: Jet tagging via particle clouds. *Phys. Rev. D* **101**, 056019, https://doi.org/10.1103/PhysRevD.101.056019 (2020).
38. Bogatskiy, A. *et al*. Lorentz group equivariant neural network for particle physics. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 992 (PMLR, 2020).
39. Moreno, E. A. *et al*. Interaction networks for the identification of boosted $h \to b\bar{s}\bar{b}$ decays. *Phys. Rev. D* **102**, 012010, https://doi.org/10.1103/PhysRevD.102.012010 (2020).
40. CMS Collaboration. Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with phase 1 CMS detector. *CMS Detector Performance Note CMS-DP-2018-058, CERN* (2018).
41. Sirunyan, A. M. *et al*. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *JINST* **13**, P05011, https://doi.org/10.1088/1748-0221/13/05/P05011 (2018).
42. Chatrchyan, S. *et al*. Identification of b-quark jets with the CMS experiment. *JINST* **8**, P04013, https://doi.org/10.1088/1748-0221/8/04/P04013 (2013).
43. Bols, E., Kieseler, J., Verzetti, M., Stoye, M. & Stakia, A. Jet flavour classification using DeepJet. *JINST* **15**, P12012, https://doi.org/10.1088/1748-0221/15/12/P12012 (2020).
44. Brun, R. *et al*. Code for root-project/root. *Zenodo*. https://doi.org/10.5281/zenodo.3895860 (2019).
45. Pivarski, J. *et al*. Code for scikit-hep/uproot. *Zenodo*. https://doi.org/10.5281/zenodo.3952728 (2020).
46. Pivarski, J. *et al*. Code for scikit-hep/awkward-array. *Zenodo*. https://doi.org/10.5281/zenodo.3952674 (2020).
47. Huerta, E. A. *et al*. Accelerated, scalable and reproducible AI-driven gravitational wave detection. *Nature Astronomy* **5**, 1062–1068, https://doi.org/10.1038/s41550-021-01405-0 (2021).
48. Huerta, E. A. & Zhao, Z. *Advances in Machine and Deep Learning for Modeling and Real-Time Detection of Multi-messenger Sources*, 1–27, https://doi.org/10.1007/978-981-15-4702-7_47-1 (Springer Singapore, Singapore, 2020).
49. Lee, H. *et al*. DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding. In *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*, 12–19, https://doi.org/10.1109/DLS49591.2019.00007 (2019).
50. Randall, L. & Sundrum, R. Large mass hierarchy from a small extra dimension. *Phys. Rev. Lett.* **83**, 3370, https://doi.org/10.1103/PhysRevLett.83.3370 (1999).
51. Sjöstrand, T. *et al*. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.* **191**, 159, https://doi.org/10.1016/j.cpc.2015.01.024 (2015).
52. CMS Collaboration. Event generator tunes obtained from underlying event and multiparton scattering measurements. *Eur. Phys. J. C* **76**, 155, https://doi.org/10.1140/epjc/s10052-016-3988-x (2016).
53. Ball, R. D. *et al*. Parton distributions with LHC data. *Nucl. Phys. B* **867**, 244, https://doi.org/10.1016/j.nuclphysb.2012.10.003 (2013).
54. Agostinelli, S. *et al*. Geant4 —a simulation toolkit. *Nucl. Instrum. Meth. A* **506**, 250, https://doi.org/10.1016/S0168-9002(03)01368-8 (2003).
55. Duarte, J., Rao, A. & Würthwein, F. Code for jmduarte/capstone-particle-physics-domain. *Zenodo*. https://doi.org/10.5281/zenodo.5594610 (2021).
56. Chen, Y. & Duarte, J. Code for FAIR4HEP/FAIR4HEP-Toolkit. *Zenodo*, https://doi.org/10.5281/zenodo.5146623 (2021).
57. CMS Collaboration & Duarte, J. Record for the data set "Sample with jet, track and secondary vertex properties for Hbb tagging ML studies (HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC)". *CERN Open Data Portal*. http://opendata.cern.ch/api/records/12102 (2020).
58. CMS Collaboration. Simulated dataset BulkGravTohhTohbbhbb_narrow_M-600_13TeV-madgraph in MINIAODSIM format for 2016 collision data. *CERN Open Data Portal*., https://doi.org/10.7483/OPENDATA.CMS.R5U7.WV97 (2019).
59. CMS Collaboration. Simulated dataset QCD_Pt_300to470_TuneCUETP8M1_13TeV_pythia8 in MINIAODSIM format for 2016 collision data. *CERN Open Data Portal*. https://doi.org/10.7483/OPENDATA.CMS.DAY1.ZIQE (2019).
60. Duarte, J. *et al*. HiggsToBBNtupleProducerTool - ROOT ntuple producer for developing machine learning algorithms from CMS Run2 MiniAOD. *CERN Open Data Portal*. https://doi.org/10.7483/OPENDATA.CMS.MWG0.J8V6 (2019).

Acknowledgements

We thank Tom Honeyman from the Australian Research Data Commons (ARDC) and Chris Erdmann from the American Geophysical Union (AGU) for their help and advice on both FAIR data principles in general and on their application to our specific dataset, though any errors in interpretation of the principles are ours. We thank the CMS Collaboration for making the H(b$\bar{\text{b}}$) dataset publicly available and for helpful discussions in the preparation of this work. We also thank Tibor Simko, Kati Lassila-Perini, and the rest of CERN Open Data Portal Team. This work was performed as part of the FAIR Framework for Physics-Inspired Artificial Intelligence in High Energy Physics (FAIR4HEP) project (DE-SC0021258, DE-SC0021395, DE-SC0021225, and DE-SC0021396), support by the Office of Advanced Scientific Computing Research within U.S. Department of Energy Office of Science. FM was partially supported by an Halcoğlu Data Science Fellowship.

Author contributions

E.A.H. led this work and coordinated the writing of this manuscript. Y.C. and J.D. participated in the selection and FAIRification of our sample dataset. D.S.K. provided FAIR expertise to guide the initial FAIR assessment, and then worked with external FAIR experts to validate our results. V.K. contributed to the evaluation of the results. M.S.N. and P.H. contributed to the FAIR assessment of our sample dataset and as an internal editor of the

manuscript. Z.Z. contributed to the evaluation of the results and reviewed the manuscript. R.K., F.M. and D.D. provided feedback on the manuscript and contributed to the FAIRification of the dataset. S.E.P. provided feedback on the manuscript. R.R. acted as an internal editor of the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.A.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.