



# City Research Online

## City St George's, University of London

**Citation:** Araz, J. Y., Buckley, A., Kasieczka, G., Kieseler, J., Kraml, S., Kvellestad, A., Lessa, A., Procter, T., Raklev, A., Reyes-González, H., et al (2024). Les Houches guide to reusable ML models in LHC analyses. SciPost Physics Community Reports, 3. doi: 10.21468/scipostphyscommrep.3

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/36214/>

**Link to published version:** <https://doi.org/10.21468/scipostphyscommrep.3>

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

# Les Houches guide to reusable ML models in LHC analyses

Jack Y. Araz<sup>1</sup>, Andy Buckley<sup>2</sup>, Gregor Kasieczka<sup>3</sup>, Jan Kieseler<sup>4</sup>,  
Sabine Kraml<sup>5</sup>, Anders Kvellestad<sup>6</sup>, Andre Lessa<sup>7</sup>, Tomasz Procter<sup>2</sup>,  
Are Raklev<sup>6</sup>, Humberto Reyes-Gonzalez<sup>8,9,10</sup>, Krzysztof Rolbiecki<sup>11</sup>,  
Sezen Sekmen<sup>12</sup> and Gokhan Unel<sup>13</sup>

<sup>1</sup> Jefferson Lab, Newport News, VA 23606, USA

<sup>2</sup> University of Glasgow, Glasgow, UK

<sup>3</sup> Univ. Hamburg, Germany

<sup>4</sup> Karlsruhe Institute for Technology, Karlsruhe, Germany

<sup>5</sup> Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 38000 Grenoble, France

<sup>6</sup> University of Oslo, 0316 Oslo, Norway

<sup>7</sup> Universidade Federal do ABC, Santo André, 09210-580 SP, Brazil

<sup>8</sup> Department of Physics, University of Genova, Via Dodecaneso 33, 16146 Genova, Italy

<sup>9</sup> INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy

<sup>10</sup> Institut für Theoretische Teilchenphysik und Kosmologie,  
RWTH Aachen, 52074 Aachen, Germany

<sup>11</sup> Faculty of Physics, University of Warsaw, 02-093 Warsaw, Poland

<sup>12</sup> Department of Physics, Kyungpook National University, Daegu, South Korea

<sup>13</sup> U.C. Irvine, Physics & Astronomy Dept., Irvine, CA, USA

## Abstract

With the increasing usage of machine-learning in high-energy physics analyses, the publication of the trained models in a reusable form has become a crucial question for analysis preservation and reuse. The complexity of these models creates practical issues for both reporting them accurately and for ensuring the stability of their behaviours in different environments and over extended timescales. In this note we discuss the current state of affairs, highlighting specific practical issues and focusing on the most promising technical and strategic approaches to ensure trustworthy analysis-preservation. This material originated from discussions in the LHC Reinterpretation Forum and the 2023 PhysTeV workshop at Les Houches.



Copyright J. Y. Araz *et al.*

This work is licensed under the Creative Commons  
[Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Published by the SciPost Foundation.

Received 11-01-2024

Accepted 08-10-2024

Published 06-11-2024

doi:[10.21468/SciPostPhysCommRep.3](https://doi.org/10.21468/SciPostPhysCommRep.3)



Check for  
updates

## Contents

|          |                                                   |          |
|----------|---------------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                               | <b>2</b> |
| <b>2</b> | <b>Mechanisms and examples</b>                    | <b>3</b> |
| <b>3</b> | <b>Analysis design</b>                            | <b>5</b> |
| <b>4</b> | <b>Material for implementation and validation</b> | <b>8</b> |

|                                  |           |
|----------------------------------|-----------|
| <b>5 Surrogate models</b>        | <b>9</b>  |
| <b>6 Summary and conclusions</b> | <b>10</b> |
| <b>References</b>                | <b>11</b> |

---

## 1 Introduction

While the preservation and reuse of traditional “cut-and-count” collider analyses is increasingly standard [1, 2], the preservation and reuse of analyses that employ machine-learning (ML) techniques is still in its relative infancy. Despite the growing reliance of experiments on ML-based analyses, there still remain serious obstacles to their practical reuse and long-term stability.

Machine-learning models are used in many different contexts within an experiment [3–7]. Some of these uses, such as  $b$ -tagging or electron/photon identification, are attempting to reconstruct a well-defined property of the physical event and can be relatively well-modelled using efficiency functions. In other cases, however, the ML model defines the analysis output, for example a signal-vs-background score, so that reinterpretability of the analysis directly depends on the reusability of the ML model.<sup>1</sup> There also exist cases in between (e.g. an analysis-specific top-tagger trained specifically on signal samples, where the efficiencies may differ significantly for different signal models), and all these use cases can take a wide variety of different inputs requiring different amounts of detector simulation/emulation.

Accessible and reusable ML information is thus a necessity that is increasingly recognised in the HEP community. For example, the Snowmass 2021 white paper on *Data and Analysis Preservation, Recasting, and Reinterpretation* [8] explicitly recommends:

“... that reinterpretability and reuse should be kept in mind early on in the analysis design. This concerns, for instance, the choice of input parameters in ML models, the full specification of the fiducial phase space of measurement in terms of the final state, including any vetos applied, and generally the choice of non-overlapping regions and standard naming of shared nuisances to facilitate the combination of analyses.”

A few LHC analyses have recently published ML models on HEPData [9, 10]. This is a major step in the right direction. In this paper, we review the present status and discuss what is needed to make ML models, as often employed in LHC analyses, usable in reinterpretation studies. Our aim is to provide concrete guidelines for appropriate ML design, the serialisation and communication of the learned model, and the provision of validation material needed for the correct usage of reinterpretation tools. This concerns measurements and searches alike.

Following the guidelines of this paper will benefit not only the phenomenology community but also the analysis preservation efforts within the experimental collaborations by extending the usefulness of the experimental work. Furthermore, it can benefit the education of the next generation of physicists.

---

<sup>1</sup>Throughout this document by reinterpretability we mean reusability of the ML model or the analysis to infer results for scenarios not originally considered by the experimental analysis.

## 2 Mechanisms and examples

Standards exist for a degree of ML interoperability. For neural networks (NNs), the most common format is Open Neural Network Exchange (ONNX) [11], though, within ATLAS in particular, the Light-Weight Neural Network (LWTNN) [12] format is also used. For boosted decision trees (BDTs), possible output formats include dependency-free serialisation through petrify-bdt [13] and treelite [14], custom formats like XML files from ROOT TMVA<sup>2</sup> [15, 16], and to some extent also ONNX. Which output formats are available is likely to depend on the training environment, meaning that consideration to the preservation strategy is needed from an early stage of the ML-analysis design. There also exist ML architectures for which there is no current framework-independent representation, such as elements of generative networks used in normalizing-flow methods: this challenge is hence one that can be expected to continue, and the practicalities of preservation and reuse need to be considered when choosing an ML architecture.

Successful publication of an ML model requires not only sharing the model itself (including architectures, weights, and complete specification of any software dependencies) but also the detailed specification of the input data and any preprocessing performed (which to an extent can be encoded in ONNX), along with material to validate the model. Moreover, the long-term stability of the preservation format needs to be addressed – i.e. that model files saved using, for example, a version of ONNX from 2023 will still produce identical results in 2050. For this reason, it is useful to advertise the exact software versions used to produce, save and run the neural network. However, even with exact versioning, the long-term future ability to run such legacy code is not guaranteed, in particularly within frameworks that may or may not be maintained in the long term. Nonetheless, detailed documentation of the model used as well as validation material can greatly increase the future re-usability of analyses. A few initial encouraging steps are already being taken by the experimental collaborations in this direction.

Below we give some examples of analyses that have published ML information. The list is dominated by ATLAS analyses due to the current volumes of public information, but CMS is expected to follow suit soon.

- The ATLAS search for squarks and gluinos in final states with jets and missing transverse momentum, ATLAS-SUSY-2018-22 [17], has published BDT weights in XML format for use in ROOT TMVA on HEPData [18]. This is accompanied by a code snippet [19] with the implementation of the analysis selection at the truth level (also included in the SimpleAnalysis framework [20, 21]). Observed BDT score distributions in control and signal regions are provided for validation. This is an example of good practice which can readily be reused.
- The ATLAS search for  $R$ -parity-violating supersymmetry, ATLAS-SUSY-2019-04 [22] published trained networks in ONNX format on HEPData [23]. This analysis uses 64 high-level variables plus a continuous  $b$ -tag score from a ML-based tagger. Also in this case, a truth-level implementation is available [24], which proved essential for understanding the input variables. However, it is unclear how to reproduce the  $b$ -tag score, and how to verify that physics objects from any given fast-simulation package produce the intended ML responses within acceptable uncertainties.
- Another analysis for which the trained NN has been published in the ONNX format is the ATLAS search for supersymmetry in final states with many  $b$ -jets and missing transverse energy, ATLAS-SUSY-2018-30 [25] (targeting gluino decays to 3rd generation quarks

---

<sup>2</sup>Toolkit for Multivariate Data Analysis with ROOT.

and electroweakinos). Here, the ONNX files are contained in the SimpleAnalysis distribution [21]. The SimpleAnalysis code also details input variables and their normalisation. For  $b$ -tagging, a binary variable is used. The NN returns scores for evaluation of signal and control regions, however the latter are not included in SimpleAnalysis code and therefore cannot be used in recasting tools, e.g. for validation or for reuse with alternative signal models. Publishing also distributions of NN score for benchmark signal events would assist validation.

- The ATLAS search for neutral long-lived particles that decay into displaced hadronic jets in the calorimeter (“CalRatio LLP search”) ATLAS-EXOT-2019-23 [26] published ONNX records of the NNs [27] as well as pure-C++ and pure-Python standalone executables of the BDTs [28] (preserved with petrify-bdt) on HEPData. These employ a large amount of low-level detector information and are not intended to be used outside the collaboration; the publication on HEPData is purely for the purpose of preservation. However, for public use, the analysis also provides 6-dimensional efficiency maps [29] parametrising the BDT+NN selection (plus example code to read and use them [30]). We will return to this kind of “model surrogates” in Section 5.
- The ATLAS anomaly-detection search in hadronic final states ATLAS-HDBS-2019-23 [31], which targets new resonances decaying into a Higgs boson and a generic new particle  $X$ , published the post-training weights from their variational recurrent NN (VRNN) as a PyTorch .pth file [32] on HEPData. Also provided is the VRNN Python code [33], but no description of the input/output variables, nor any description of how to use the code.

We also note that three of the above analyses have been presented in RAMP (‘Reinterpretation: Auxiliary Material Presentation’) seminars: slides and recordings are available online [34–36].

On the reinterpretation tools side, all the major frameworks (CheckMATE [37], GAMBIT’s ColliderBit [38], MadAnalysis 5 [39], Rivet [40], and ADL/CutLang [41, 42]) have developed interfaces for using published ML models. This was extensively discussed at the last two workshops of the Reinterpretation Forum: in December 2022 at CERN [43] and in August 2023 at Durham University [44].

Three of the above-mentioned ML-based searches have been implemented in recasting tools, with various degree of success. For models published in the ONNX format, the ONNX interface has been realised in all cases using the C++ ONNXRuntime library [45]. In particular, the ATLAS multi- $b$  jet-plus-MET analysis, ATLAS-SUSY-2018-30 [25], has been implemented in CheckMATE, Rivet, GAMBIT and ADL. The validation in Rivet, GAMBIT and CheckMATE showed generally good agreement with the cut-flows provided by ATLAS, though CheckMATE found up to 30% lower efficiencies than reported by ATLAS (the discrepancy is due to  $b$ -tagging efficiency); and the ADL/CutLang validation is ongoing.

These same four tools, as well as MadAnalysis 5, have also implemented ATLAS-SUSY-2019-04 [22]. All these implementations, however, have noted apparent tensions with the published cut-flows even before applying the NN score, with similar issues encountered using ATLAS’ own public SimpleAnalysis code. The NN score distribution for a sample of signal events revealed significant discrepancies to the corresponding ATLAS distribution, and high sensitivity to the truth-level schemes used for emulation of the  $b$ -tag score input feature. More information is definitely needed for a reliable implementation of this analysis. Finally, CheckMATE has also implemented ATLAS-SUSY-2018-22 [17] using ROOT TMVA for BDT classification, and found good agreement with the ATLAS cut-flows.

These concrete implementations in public reinterpretation frameworks all relied heavily on the ATLAS SimpleAnalysis codes available from HEPData, effectively as pseudo-code documentation as the descriptions in the respective paper texts were usually insufficient on their own. Although it does not seem necessary to provide such a detailed description in publications, a short text document detailing shape of input tensors, variables, output tensors (e.g. the NN output in ATLAS-SUSY-2018-30 is non-trivial), etc. would be a further great assistance. Integration of the corresponding naming conventions of inputs and outputs in the ONNX file may also be useful. These hands-on experiences with the currently available material also inform the guidelines elaborated below.

### 3 Analysis design

In the design of an ML-based analysis, a number of choices have to be made, which will significantly influence the ease of publication and subsequent reuse. It is thus worthwhile that reinterpretability and reuse be kept in mind early on in the analysis design.

**Choice of framework:** The first choice is likely to be which ML package to implement the network or tree in. It is essential that this package be open source, e.g. Tensorflow, Keras, PyTorch, scikit-learn, JAX, or even ROOT TMVA, or that, at the minimum, the selected package can save the ML model for execution in an open-source or framework-independent form.

Proprietary packages, such as NeuroBayes [46] – used in several public analyses e.g. Refs. [47, 48] – or MATLAB-based packages are effectively impossible to reuse in any reinterpretation tool other than collaboration-internal containerized preservations such as RECAST [49], which are not generally available to the phenomenology community.

Changes to formats or default settings, even in minor framework-version updates, can fundamentally change ML model behaviours, so the more complex an execution framework and its dependency tree, the larger the risk of algorithm behaviour becoming dependent on framework versioning.

**Choice of preservation format:** The typical complexity of machine-learning models, particularly neural-network architectures, makes the choice of ML preservation format an important question. This is largely independent of the programming language from which the ML-training framework is used, but it is crucial that the chosen format be readable and executable from compiled languages, particularly C++. The majority of event-loop reinterpretation tools are implemented in C++, and these must be able to execute ML models directly rather than e.g. inefficiently embedding a Python interpreter. Other non-C++ language options may also arise and become dominant in future, e.g. the Julia language is seeing gradual uptake in use for certain ML tasks.

This frames the central importance of flexible data formats. Although the overwhelming majority of ML training takes place in Python, the Python “pickle” format is not suitable as a long-term storage format because of a) its specificity to Python, b) versioning sensitivities within pickle itself, and c) its embedding of dependencies on (specific versions of) dependency packages. Fortunately, it has become common for ML frameworks to support export to compiled form for production deployments, and this has motivated industry development of interchange formats which are also natural candidates for preservation use. We note, however, that the industry focus is on short-timescale shifts between platforms rather than long-term preservation. Most major ML packages from outside HEP, e.g. TensorFlow and PyTorch, now support preservation and reloading via the ONNX format, including in compiled frameworks.

LWTNN is a lightweight C++ neural network framework that originated in the HEP community and runs the familiar risk of dependence on a few time-poor individuals. While it is currently supported, is used in ATLAS for efficient deployment of several neural-net applications, and has the benefit of having a far smaller code dependency than ONNXRuntime, the absence of long-term commitment is a concern for preservation. This is an issue that could benefit from discussion within and between the experimental collaborations. Conversion to LWTNN's own JSON from ONNX is possible, while the inverse conversion is currently not. **ONNX is hence the current format with most institutional/industrial support and the strongest case for use as the primary preservation format for HEP neural nets.**

We note, however, that newer techniques such as normalizing flows (NFs) are gaining significant uptake for publishing likelihoods [50] and many other applications [51]. However, currently, NFs are not supported by ONNX, and active involvement from the HEP community might be necessary to achieve this. This reflects the constantly changing nature of this technology and its format ecosystem. This will likely evolve further, adding a requirement that HEP archival systems perform periodic re-exporting, versioning, and regression testing of analysis ONNX data files to maintain usability with the evolving tool set. (As with many issues in ML reinterpretation, this concern also applies to other archive formats such as HistFactory [52], HS3<sup>3</sup> [53], or ROOT, but the relative complexity and novelty of ML methods make the issue particularly acute.)

We also note that ROOT TMVA has recently developed an experimental system, SOFIE [54], for converting ONNX-based network preservations to C++ for use with TMVA. This has yet to be evaluated for reinterpretation purposes: a further decoupling from the large dependency on ROOT itself would be of great interest to public reinterpretation frameworks, most of which are not based on ROOT.

While neural nets have been the main focus of recent reinterpretation issues, it is important also to note the usage of other methods, most significantly BDTs, which have, in fact, been used more than NNs at the analysis level. The HEP-oriented petrify-bdt package can convert BDTs from TMVA and sci-kit-learn into C++ or Python code, using only conditional branching and no dependencies to make a very stable read/execute-only record of a trained BDT.<sup>4</sup> In addition, the treelite library and format purport to be a universal exchange mechanism for BDTs, similar to the role played by ONNX for NNs; while this, like ONNX, does add a library dependency and potential long-term stability issues, it has support for the XGBoost and LightGBM BDT libraries (but not currently TMVA).

For both NNs and BDTs and beyond, preserved ML files should be tracked and stored as a full part of the analysis to ensure they cannot accidentally be lost or overwritten. In case files are larger than can be supported by HEPData, **the CERN-based Zenodo data archive is the natural platform for hosting and preservation of trained ML models** – stable links should be provided from the HEPData archive to Zenodo or similar records via DOI codes.

**Choice of network architecture:** After choosing the package, thoughts turn to network architecture and choice of inputs. Some very advanced architectures may not yet be convertible to ONNX format (and LWTNN is even more restricted), so this choice can impact preservability. Similarly, heavily customised layers or activation functions (e.g. custom

<sup>3</sup>High Energy Physics Statistics Serialization Standard.

<sup>4</sup>This simpler representation has been seen to also greatly improve the speed and memory footprint with respect to the original.

“lambda” layers in TensorFlow) may not be well preserved. Testing whether porting the network to a format such as ONNX is possible at an early stage and should be included in continuous-integration (CI) testing of analysis-code development. This will minimise the amount of work needed later to ensure the network is preservable.

The size of ML models may also become a potential stumbling block. In global fits, typically, tens if not hundreds of analyses are used at the same time. The combination of many very large models, in terms of disk or memory footprint, may make them unsuitable for reinterpretation. If designing lighter models is not possible, techniques such as neural-network pruning could be considered.

**Choice of input features:** The choice of input features leads to a parting of the ways between how best to reuse networks. In the extreme case, analyses may design taggers that are based largely or entirely on detector-level data that reinterpretation tools cannot hope to reproduce accurately – reinterpretation of this class of ML tool is discussed in Section 5.

However, there also exist many taggers that mainly or entirely use reco-level data that can be reproduced fairly well by the simple forms of detector simulation/emulation used by reinterpretation tools (either Delphes [55] or some variant of 4-vector smearing as in Rivet [56], MadAnalysis SFS [57], or GAMBIT ColliderBit [38]). Examples of such event features include the kinematics of reconstruction-level jets or leptons and missing energy, which admit fairly simple response-function parametrisations and already form the basis of most “cut-and-count” analyses’ reinterpretation.

In these cases, reinterpretation would be greatly aided by avoiding inclusion of just one or two low-level variables in an otherwise reproducible ML model. As discussed above, several current analyses include among their input features a continuous ML-score variable produced by a complex and closed-source *b*-tagging algorithm operating on detector-level quantities such as hits in layers of silicon tracker. Using such unreproducible input variables renders the entirety of the preserved analysis-level ML discriminant unreproducible, even if those features only play a subleading role in its behaviour.

This problem can be ameliorated by a small tweak to the analysis design: although this score (and other such variables) is not easy to emulate in reinterpretation frameworks, replacing the continuous score with a tagged/not-tagged boolean value for ML purposes can ensure the network input can be well-approximated using efficiency-based techniques.

### ML model design checklist:

- Use machine-learning software that can be easily converted to a stable interchange format supported by open-source tools. The ONNX and LWTNN JSON formats are the current most stable options for neural networks.
- Alternatively, if possible, export the ML model to executable code without dependencies beyond standard libraries.
- Preserved networks should be runnable with as few dependencies as possible from an API to a compiled language (e.g. C++), not just from Python.
- Avoid over-complexity in network design, e.g. not using customised layers or custom activation functions if the application does not require them. Ensure the chosen architecture has sufficient preservation-format support, particularly with ONNX.
- Where possible, and especially if the model is dominated by simple kinematic inputs, avoid input features that are heavily dependent on detector and reconstruction details.

- Where inputs are heavily detector-based, in addition to preserving the ML model itself, provide detailed efficiency maps (including mistag rates) or an equivalent surrogate network using less detector-sensitive input features (see Section 5).

## 4 Material for implementation and validation

Input features for ML algorithms should be described with sufficient detail to unambiguously recreate them, just like any other variable that is used directly to define a cut in an analysis. In addition, the exact format of the input features must be provided: an ML algorithm whose inputs are not well-defined is not usable! Specific issues include the parameter ordering (if the input to the network is a vector) or their exact keys (if it is a map), along with the units, conventions used (for example for angles), and any normalisation or padding information.

Possible formats to convey this information include auxiliary tables, an accompanying documentation note, README files, etc. Code snippets, like SimpleAnalysis implementations, are particularly helpful, as they automatically provide the needed information. A welcome example was the SimpleAnalysis routines provided in Ref. [22, 24] – we strongly recommend that such routines are saved to HEPData.

For validation, a plain data file of representative feature vectors and their corresponding ML output values would serve as a purely technical validation dataset (nb. this will also help ensuring version stability). Moreover, detailed cut-flow tables or histograms for the relevant signal model(s), including immediately before and after any ML-based cuts, can help validate both the network inputs and outputs. Plots of the model's output(s) assist in checking that the network has been implemented correctly, as the pass/fail in a cut flow may not be sufficiently informative. For example, as mentioned above, the BDT score distributions provided by ATLAS-SUSY-2018-22 proved very helpful. Similarly, plots of kinematic distributions *after* ML-based cuts can also be important, especially if the analysis goes on to make further kinematic cuts. As in other recasting situations, explicit configuration details for MC event-generator configurations allow direct comparison with plots in the paper or supporting documentation.

Some understanding of feature importance would also be very helpful, especially if accompanied by plots illustrating the distributions of the most important inputs and the NN output. This will allow the reinterpretation effort to focus on the most central variables for increased fidelity. As an example, see the discussion in Ref. [58], even though this network has not been made public. However, for this type of plot, it is very important to be very explicit about the validation sample being used in the plot – is it from the samples used in training (and if so, how do we re-generate them), or is it representative of the events reaching that stage in the analysis?

While this requested information is extensive, we note that it is not a great extension of the information already regarded as most important for “traditional” cut-and-count analysis validation and that the extras are motivated by the “black box” inscrutability of ML observables offering relatively few physics sanity-checks, as well as the increased risk of corruption by framework translation or evolution. The information is available internally, but not always in a convenient form: we urge experiments to consider the pipeline by which it can be efficiently made available in a way usable outside the experiment code-base.

Finally, we emphasise that for much validation information, specific formats and polished presentation are not essential: it is much more important that it be available at all, than to erect barriers to publication which in practise only ensure that they remain publicly unavailable.

**Documentation/validation checklist:**

- Provide exact definitions (including units, ordering and conventions) of network input features, either as code examples or documentation.
- Provide a sample of input features and output values for technical validation.
- Provide plots of input and output variables for validation samples, if possible, with some indication of feature importance, as analysis supporting data.
- Provide cut-flow information, both before and after any ML-based selections.
- Validated and runnable published analysis code can be the clearest expression of both the general analysis logic and the specific interfacing with the ML functions.
- Full descriptions of the physics models used to generate the information above, e.g. SLHA files and generator run cards, are essential inputs to validating any serious reinterpretation.

**5 Surrogate models**

In some cases, the inputs to a trained ML algorithm will not be realistically accessible to those outside the experiment – examples include detector-level inputs, such as hits, or variables which are heavily dependent on detector geometry, such as soft jet-substructure variables. Examples of this sort of network include the  $b$ -jet taggers in both ATLAS and CMS [59, 60].

To first order, this can be approximated using truth-level efficiencies – this approach has been used by reinterpretation tools for both  $b$ -tagging and tagging for SM resonances. If using this approach, providing information for type-II errors, i.e. mistag rates, such as light or charm jets in  $b$ -tagging, is very important. Similarly, if the efficiency has any strong kinematic dependence, providing the efficiency in a set of kinematic bins – for example,  $p_T$  and  $\eta$  – would be necessary. An outstanding example are the 6D(!) efficiency maps parametrising the BDT+NN selection of the ATLAS CalRatio LLP search [26] mentioned in section 2.

It would be worthwhile to explore if the accuracy of modelling could be improved using surrogate ML models, or surrogates – neural networks trained to replicate the output of the original tagger but using input events with a simpler set of attributes, typical as used in the reinterpretation frameworks. For all surrogates it would be crucial to provide a quantification of the surrogate’s uncertainty.

What inputs the surrogate should take will depend on the inputs of the original network, i.e., the physics the network is trying to capture, and the level of accuracy required of the surrogate. Broadly speaking, we can split the different sets of inputs a surrogate may take into three classes depending on the level of information required as input:

- **Reco-level surrogates:** features of reconstructed (clustered) objects after showering, hadronization and detector emulation, such as jets, dressed leptons and missing energy. In most cases (“simple reco-level”) these features will include the object kinematics; which are available at some level of precision, from public detector-emulation and clustering tools like FastJet, Delphes or the Rivet/MadAnalysis 5 convolution-based “smearing” systems. Adding in more complex reco-level variables, such as jet substructure, is also possible, but may stretch the capabilities of fast detector emulation.
- **Particle-level surrogates:** 4-vectors of all particle-level objects after showering and hadronisation (with or without detector smearing), i.e. hadrons, leptons, photons or long-lived Beyond-the-Standard Model (BSM) particles. This level may also include information about the positions and timings of decay or production vertices.

- **Parton-level surrogates:** 4-vectors of parton-level particles produced by the MC generator prior to hadronisation or detector emulation. This level can be particularly useful when building surrogate models involving long-lived BSM particles. Whether or not showering needs to be included depends on the physics case and on whether initial-state radiation is included in the hard scattering process.

Which of these to use will depend heavily on the use case. For example, “reco-level” surrogates clearly present the easiest task to neural networks, but this may rely on detector emulation providing reliable results. At present, this is not always the case, e.g. for soft jet-substructure features. Methods that require extensive jet reclustering may also be unsuited to reinterpretation tools that prioritise performance.

A clear distinction should also be drawn between surrogates that provide the true label (e.g. a boolean variable expressing if a jet actually contains a  $b$ -hadron), and those that do not. The surrogates that do, effectively become very well-parameterised efficiency maps, and will likely perform better. However, there may be valid reasons not to provide this information to the surrogate: for example, if a tagger is being used to tag BSM particles directly, then directly checking the event record for a particular BSM particle-ID code becomes dangerously model-dependent and hence cannot be reinterpreted beyond the original model. Such implications again hint at questions to be asked at the analysis-design stage, about how to maximise the general-purpose usefulness and hence the long-term impact of a physics study.

The field of preserving analyses using surrogates is even younger than the other content of this note. In the future we hope to see more examples – among them several projects started at the 2023 Les Houches PhysTeV workshop – that can shape best practices for analyses going forward. That being said, analyses should not shy away from making their own surrogates immediately! Of course the guidelines for design, documentation and validation of ML models elaborated above also apply to surrogates.

### Surrogates checklist:

- If the ML model in an analysis is not suitable for sharing and reuse in its original form, detailed efficiency map or a trained surrogate model are necessary to enable reinterpretation.
- For providing a surrogate model, follow the analysis design and documentation/validation guidelines above.

## 6 Summary and conclusions

Machine-learning methods have great potential to boost the scientific sensitivity of physics experiments, and are undoubtedly here to stay as a major part of how we do physics. But when applying ML-based methods we should bear in mind that the long-term impact of an analysis depends not only on its sensitivity to the initial physics model, but also on its future application to other models. This, combined with the large investments in major collider facilities and data-taking, demand the possibility for data reuse rather than publish-and-forget, and so priority should be given to the long-term and framework-independent re-usability of ML methods. In this note we have highlighted practical issues (and currently preferred partial solutions) to be actively considered as early as possible in ML-based analysis design and implementation. We anticipate future updates of these recommendations as technological capabilities and physics challenges evolve, but an early adoption of the present recommendations will go a long way towards ensuring long-lasting scientific impact of current and near-future ML-based analyses.

## Acknowledgments

**Funding information** AB acknowledges support by the UK STFC Consolidated Grant scheme (ST/W000520/1) and TP an STFC doctoral training studentship. The work of JK is supported by the Alexander von Humboldt-Stiftung. The work of SK is supported in part by the IN2P3 master project “Théorie – DATAMatter”; financial support by the CNRS Formation Permanente for attending the Les Houches workshop is also gratefully acknowledged. AK and AR are supported by the Research Council of Norway (RCN) through the FRIPRO grant 323985 PLUMBIN’. AL is supported by FAPESP grants no. 2018/25225-9 and 2021/01089-1. HRG is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 396021762 – TRR 257: Particle Physics Phenomenology after the Higgs Discovery. HRG also acknowledges the support from the Italian PRIN grant 20172LNEEZ. KR is supported by the Norwegian Financial Mechanism 2014-2021, grant no. 2019/34/H/ST2/00707; and by the National Science Centre, Poland grant 2019/35/B/ST2/02008. The work of SS is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under contract NRF-2021R1I1A3048138.

## References

- [1] S. Kraml et al., *Searches for new physics: Les Houches recommendations for the presentation of LHC results*, Eur. Phys. J. C **72**, 1976 (2012), doi:[10.1140/epjc/s10052-012-1976-3](https://doi.org/10.1140/epjc/s10052-012-1976-3).
- [2] W. Abdallah et al., *Reinterpretation of LHC results for new physics: Status and recommendations after Run 2*, SciPost Phys. **9**, 022 (2020), doi:[10.21468/SciPostPhys.9.2.022](https://doi.org/10.21468/SciPostPhys.9.2.022).
- [3] D. Guest, K. Cranmer and D. Whiteson, *Deep learning and its application to LHC physics*, Annu. Rev. Nucl. Part. Sci. **68**, 161 (2018), doi:[10.1146/annurev-nucl-101917-021019](https://doi.org/10.1146/annurev-nucl-101917-021019).
- [4] K. Albertsson et al., *Machine learning in high energy physics community white paper*, J. Phys.: Conf. Ser. **1085**, 022008 (2018), doi:[10.1088/1742-6596/1085/2/022008](https://doi.org/10.1088/1742-6596/1085/2/022008).
- [5] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao and T. Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics*, Nature **560**, 41 (2018), doi:[10.1038/s41586-018-0361-2](https://doi.org/10.1038/s41586-018-0361-2).
- [6] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman and D. Shih, *Machine learning in the search for new fundamental physics*, Nat. Rev. Phys. **4**, 399 (2022), doi:[10.1038/s42254-022-00455-1](https://doi.org/10.1038/s42254-022-00455-1).
- [7] P. Shanahan, K. Terao and D. Whiteson, *Snowmass 2021 computational frontier CompF03 topical group report: Machine learning*, (arXiv preprint) doi:[10.48550/arXiv.2209.07559](https://doi.org/10.48550/arXiv.2209.07559).
- [8] S. Bailey et al., *Data and analysis preservation, recasting, and reinterpretation*, (arXiv preprint) doi:[10.48550/arXiv.2203.10057](https://doi.org/10.48550/arXiv.2203.10057).
- [9] A. Buckley, M. Whalley and J. W. Monk, *HepData - The HEP data archive reloaded*, Proc. Sci. **093**, 067 (2011), doi:[10.22323/1.093.0067](https://doi.org/10.22323/1.093.0067).
- [10] E. Maguire, L. Heinrich and G. Watt, *HEPData: A repository for high energy physics data*, J. Phys.: Conf. Ser. **898**, 102006 (2017), doi:[10.1088/1742-6596/898/10/102006](https://doi.org/10.1088/1742-6596/898/10/102006).
- [11] Open neural network exchange, <https://onnx.ai>.

- [12] D. H. Guest et al., *lwttn/lwttn: Version 2.13*, Zenodo (2022), doi:[10.5281/zenodo.6467676](https://doi.org/10.5281/zenodo.6467676).
- [13] *petrify-bdt*, <https://gitlab.com/hepcedar/petrify-bdt>.
- [14] P. H. Cho et al., *treelite*, <https://github.com/dmlc/treelite>.
- [15] R. Brun et al., *root-project/root: v6.18/02*, Zenodo (2019), doi:[10.5281/zenodo.3895860](https://doi.org/10.5281/zenodo.3895860)
- [16] *Machine learning with ROOT* (2024), <https://root.cern/manual/tmva>.
- [17] G. Aad et al., *Search for squarks and gluinos in final states with jets and missing transverse momentum using 139 fb<sup>-1</sup> of  $\sqrt{s} = 13$  TeV pp collision data with the ATLAS detector*, J. High Energy Phys. **02**, 143 (2021), doi:[10.1007/JHEP02\(2021\)143](https://doi.org/10.1007/JHEP02(2021)143).
- [18] G. Aad et al., *ZeroLepton2018-SRBDT-weight.tar.gz*, HEPData (2021), doi:[10.17182/hepdata.95664.v2/r8](https://doi.org/10.17182/hepdata.95664.v2/r8).
- [19] G. Aad et al., *ZeroLeptonBDT2018.cxx*, HEPData (2021), doi:[10.17182/hepdata.95664.v2/r6](https://doi.org/10.17182/hepdata.95664.v2/r6).
- [20] G. Aad et al., *SimpleAnalysis: Truth-level analysis framework*, Tech. Rep. ATL-PHYS-PUB-2022-017, CERN, Geneva, Switzerland (2022), doi:[10.17181/CERN.R6S3.0QKV](https://doi.org/10.17181/CERN.R6S3.0QKV).
- [21] B. Petersen and G. Stark, *SimpleAnalysis, v1.1.0*, <https://gitlab.cern.ch/atlas-sa/simple-analysis>.
- [22] G. Aad et al., *Search for R-parity-violating supersymmetry in a final state containing leptons and many jets with the ATLAS experiment using  $\sqrt{s} = 13$  TeV proton-proton collision data*, Eur. Phys. J. C **81**, 1023 (2021), doi:[10.1140/epjc/s10052-021-09761-x](https://doi.org/10.1140/epjc/s10052-021-09761-x).
- [23] G. Aad et al., *SUSY-2019-04-ONNX.tar.gz*, HEPData (2021), doi:[10.17182/hepdata.104860.v1/r3](https://doi.org/10.17182/hepdata.104860.v1/r3).
- [24] G. Aad et al., *SUSY-2019-04.cxx*, HEPData (2021), doi:[10.17182/hepdata.104860.v1/r1](https://doi.org/10.17182/hepdata.104860.v1/r1).
- [25] G. Aad et al., *Search for supersymmetry in final states with missing transverse momentum and three or more b-jets in 139 fb<sup>-1</sup> of proton-proton collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector*, Eur. Phys. J. C **83**, 561 (2023), doi:[10.1140/epjc/s10052-023-11543-6](https://doi.org/10.1140/epjc/s10052-023-11543-6).
- [26] G. Aad et al., *Search for neutral long-lived particles in pp collisions at  $\sqrt{s} = 13$  TeV that decay into displaced hadronic jets in the ATLAS calorimeter*, J. High Energy Phys. **06**, 005 (2022), doi:[10.1007/JHEP06\(2022\)005](https://doi.org/10.1007/JHEP06(2022)005).
- [27] G. Aad et al., *ONNX.tar.gz*, HEPData (2023), doi:[10.17182/hepdata.115578.v3/r1](https://doi.org/10.17182/hepdata.115578.v3/r1).
- [28] G. Aad et al., *BDTS.tar.gz*, HEPData (2023), doi:[10.17182/hepdata.115578.v3/r2](https://doi.org/10.17182/hepdata.115578.v3/r2).
- [29] G. Aad et al., *Data from figures 11 and 12 of aux. mat. of Ref. [26] (version 3)*, HEPData (2023), doi:[10.17182/hepdata.115578.v3/t48](https://doi.org/10.17182/hepdata.115578.v3/t48), doi:[10.17182/hepdata.115578.v3/t49](https://doi.org/10.17182/hepdata.115578.v3/t49).
- [30] G. Aad et al., *readMap.py*, HEPData (2023), doi:[10.17182/hepdata.115578.v3/r4](https://doi.org/10.17182/hepdata.115578.v3/r4).
- [31] G. Aad et al., *Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using  $\sqrt{s} = 13$  TeV pp collisions with the ATLAS detector*, Phys. Rev. D **108**, 052009 (2023), doi:[10.1103/PhysRevD.108.052009](https://doi.org/10.1103/PhysRevD.108.052009).

- [32] G. Aad et al., *VRNN\_weights.pth*, HEPData (2023), doi:[10.17182/hepdata.135828.v2/r2](https://doi.org/10.17182/hepdata.135828.v2/r2).
- [33] G. Aad et al., *VRNN\_model.py*, HEPData (2023), doi:[10.17182/hepdata.135828.v2/r1](https://doi.org/10.17182/hepdata.135828.v2/r1).
- [34] K. Uno, *Material for the ATLAS 0-lepton gluino/squark search with 139/fb (incl. BDT weights!)*, CERN, Geneva, Switzerland (2021), <https://cds.cern.ch/record/2763449>.
- [35] J. M. Berlingen, *Material from the ATLAS SUSY RPV lepton(s) + jets analysis (ATLAS-SUSY-2019-04)*, CERN, Geneva, Switzerland (2021), <https://indico.cern.ch/event/1083851>.
- [36] L. Corpe, *Material from the ATLAS EXOT-2019-23 (neutral LLP to displaced jets in calo)*, CERN, Geneva, Switzerland (2023), <https://indico.cern.ch/event/1233294>.
- [37] D. Dercks, N. Desai, J. S. Kim, K. Rolbiecki, J. Tattersall and T. Weber, *Check-MATE 2: From the model to the limit*, Comput. Phys. Commun. **221**, 383 (2017), doi:[10.1016/j.cpc.2017.08.021](https://doi.org/10.1016/j.cpc.2017.08.021).
- [38] C. Balázs et al., *ColliderBit: A GAMBIT module for the calculation of high-energy collider observables and likelihoods*, Eur. Phys. J. C **77**, 795 (2017), doi:[10.1140/epjc/s10052-017-5285-8](https://doi.org/10.1140/epjc/s10052-017-5285-8).
- [39] E. Conte and B. Fuks, *Confronting new physics theories to LHC data with MADANALYSIS 5*, Int. J. Mod. Phys. A **33**, 1830027 (2018), doi:[10.1142/S0217751X18300272](https://doi.org/10.1142/S0217751X18300272).
- [40] C. Bierlich et al., *Robust independent validation of experiment and theory: Rivet version 3*, SciPost Phys. **8**, 026 (2020), doi:[10.21468/SciPostPhys.8.2.026](https://doi.org/10.21468/SciPostPhys.8.2.026).
- [41] G. Unel, S. Sekmen, A. M. Toon, B. Gokturk, B. Orgen, A. Paul, N. Ravel and J. Setpal, *CutLang v2: Advances in a runtime-interpreted analysis description language for HEP data*, Front. Big Data **4**, 659986 (2021), doi:[10.3389/fdata.2021.659986](https://doi.org/10.3389/fdata.2021.659986).
- [42] H. B. Prosper, S. Sekmen and G. Unel, *Analysis description language: A DSL for HEP analysis*, (arXiv preprint) doi:[10.48550/arXiv.2203.09886](https://doi.org/10.48550/arXiv.2203.09886).
- [43] A. Buckley, M. Feickert, N. Desai, S. Kraml and S. Sekmen, *7th workshop of the “Forum on the interpretation of the LHC results for BSM studies”*, CERN, Geneva, Switzerland (2022), <https://indico.cern.ch/event/1197680>.
- [44] J. Araz, J. Butterworth, M. Spannowsky, N. Wardle and S. Kraml, *8th workshop of the “Forum on the interpretation of the LHC results for BSM studies”*, Institute for Particle Physics Phenomenology Durham, Durham, UK (2023), <https://conference.ippp.dur.ac.uk/event/1178>.
- [45] ONNX Runtime (2024), <https://onnxruntime.ai>.
- [46] M. Feindt, *A neural Bayesian estimator for conditional probability densities*, (arXiv preprint) doi:[10.48550/arXiv.physics/0402093](https://doi.org/10.48550/arXiv.physics/0402093).
- [47] G. Aad et al., *Search for flavour-changing neutral-current interactions of a top quark and a gluon in pp collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector*, Eur. Phys. J. C **82**, 334 (2022), doi:[10.1140/epjc/s10052-022-10182-7](https://doi.org/10.1140/epjc/s10052-022-10182-7).
- [48] G. Aad et al., *Search for pair-produced scalar and vector leptoquarks decaying into third-generation quarks and first- or second-generation leptons in pp collisions with the ATLAS detector*, J. High Energy Phys. **06**, 188 (2023), doi:[10.1007/JHEP06\(2023\)188](https://doi.org/10.1007/JHEP06(2023)188).

- [49] K. Cranmer and I. Yavin, *RECAST - Extending the impact of existing analyses*, J. High Energy Phys. **04**, 038 (2011), doi:[10.1007/JHEP04\(2011\)038](https://doi.org/10.1007/JHEP04(2011)038).
- [50] H. Reyes-Gonzalez and R. Torre, *The NFLikelihood: An unsupervised DNN-Likelihood from normalizing flows*, SciPost Phys. Core **7**, 048 (2024), doi:[10.21468/SciPostPhysCore.7.3.048](https://doi.org/10.21468/SciPostPhysCore.7.3.048).
- [51] *A living review of machine learning for particle physics*, <https://iml-wg.github.io/HEPML-LivingReview>.
- [52] K. Cranmer, G. Lewis, L. Moneta, A. Shibata and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, Tech. Rep. CERN-OPEN-2012-016, CERN, Geneva, Switzerland (2012), doi:[10.17181/CERN-OPEN-2012-016](https://doi.org/10.17181/CERN-OPEN-2012-016).
- [53] C. Burgard, *HEP statistics serialization standard*, in *8th workshop of the “Forum on the interpretation of the LHC results for BSM studies”*, Institute for Particle Physics Phenomenology Durham, Durham, UK (2023).
- [54] S. An, L. Moneta, S. Sengupta, A. Hamdan, F. Sossai and A. Saxena, *C++ code generation for fast inference of deep learning models in ROOT/TMVA*, J. Phys.: Conf. Ser. **2438**, 012013 (2023), doi:[10.1088/1742-6596/2438/1/012013](https://doi.org/10.1088/1742-6596/2438/1/012013).
- [55] J. de Favereau et al., *DELPHES 3: A modular framework for fast simulation of a generic collider experiment*, J. High Energy Phys. **02**, 057 (2014), doi:[10.1007/JHEP02\(2014\)057](https://doi.org/10.1007/JHEP02(2014)057).
- [56] A. Buckley, D. Kar and K. Nordström, *Fast simulation of detector effects in Rivet*, SciPost Phys. **8**, 025 (2020), doi:[10.21468/SciPostPhys.8.2.025](https://doi.org/10.21468/SciPostPhys.8.2.025).
- [57] J. Y. Araz, B. Fuks and G. Polykratis, *Simplified fast detector simulation in MadAnalysis 5*, Eur. Phys. J. C **81**, 329 (2021), doi:[10.1140/epjc/s10052-021-09052-5](https://doi.org/10.1140/epjc/s10052-021-09052-5).
- [58] G. Aad et al., *Search for pair production of heavy vectorlike quarks decaying into hadronic final states in pp at  $\sqrt{s} = 13$  TeV with the ATLAS detector*, Phys. Rev. D **98**, 092005 (2018), doi:[10.1103/PhysRevD.98.092005](https://doi.org/10.1103/PhysRevD.98.092005).
- [59] G. Aad et al., *Performance of b-jet identification in the ATLAS experiment*, J. Instrum. **11**, P04008 (2016), doi:[10.1088/1748-0221/11/04/P04008](https://doi.org/10.1088/1748-0221/11/04/P04008).
- [60] A. M. Sirunyan et al., *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, J. Instrum. **13**, P05011 (2018), doi:[10.1088/1748-0221/13/05/P05011](https://doi.org/10.1088/1748-0221/13/05/P05011).