

Towards replacing detector simulations with heterogeneous graph neural networks in flavor physics analyses

G. Hijano^{1,3}, D. Lancierini², A. Marshall^{1,*}, A. Mauri², P. Owen³, M. Patel², K. Petridis¹,
S. R. Qasim³, N. Serra³, W. Sutcliffe³ and H. Tilquin²

¹*University of Bristol, Bristol, United Kingdom*

²*Imperial College London, London, United Kingdom*

³*University of Zürich, Zurich, Switzerland*



(Received 9 July 2025; accepted 4 December 2025; published 13 January 2026)

Driven by the increasing volume of recorded data, the demand for simulation from experiments based at the Large Hadron Collider will rise sharply in the coming years. Addressing this demand solely with existing computationally intensive workflows is not feasible. This paper introduces a new fast simulation tool designed to address this demand at the LHCb experiment. This tool emulates the detector response to arbitrary multibody decay topologies at the LHCb. Rather than memorizing specific decay channels, the model learns generalizable patterns within the response, allowing it to interpolate to channels not present in the training data. Novel heterogeneous graph neural network architectures are employed that are designed to embed the physical characteristics of the task directly into the network structure. We demonstrate the performance of the tool across a range of decay topologies, showing that the networks can correctly model the relationships between complex variables. The architectures and methods presented are generic and could readily be adapted to emulate workflows at other simulation-intensive particle physics experiments.

DOI: [10.1103/k9ls-9pwc](https://doi.org/10.1103/k9ls-9pwc)

I. INTRODUCTION

Experiments at the Large Hadron Collider [1] (LHC) are now collecting data at an unprecedented rate, and with the planned upgrades, this rate will only increase [2]. To maximize the scientific output from the data taken, one requires a detailed understanding of detector efficiencies for a wide range of decay topologies. This is achieved by generating vast quantities of simulated data, which currently requires repeatedly running the same computationally expensive simulation software for millions of CPU hours annually. This forms a large fraction of each experiment's computing budget [3–5]. Despite the significant allocation of resources, the limited size of available simulation samples remains a major source of uncertainty in high-profile published analyses [6–9]. As the volume of recorded data grows, the amount of simulated data needed to maintain a comparable level of understanding of both signal and background processes increases proportionally. The development of fast simulation tools is therefore

essential to ensure the precision of future analyses is not constrained by this limit [10].

The tools being developed across the experiments at the LHC to simulate detector responses broadly fall into three categories. The first category involves tools which make simplifications to existing full simulation frameworks; these often include parametrizing or entirely avoiding expensive elements of the simulation. Examples include ReDecay [11] and FastSim [12]. These strategies have been in use for years and are well-established. However, their continued reliance on full simulation frameworks limits the overall speed-up possible, which generally proves modest relative to future requirements. The second category encompasses fully parametric tools such as DELPHES [13] and RapidSim [14] that offer extremely fast performance by approximating the detector response using simplified parametrizations. Although highly efficient, these tools are limited in accuracy and lack the level of detail required for precision studies. Finally, the third category covers a growing class of tools that leverage machine learning techniques to attempt to reproduce the accuracy of the full simulation at the speeds of the fully parametric tools. These tools either model parts of the detector response to overcome specific bottlenecks or, in some cases, even attempt to model the entirety of the detector response. The full detector response is often modeled by generating the low-level energy deposits or hits, and then passing this information to the full

*Contact author: alex.marshall@cern.ch

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

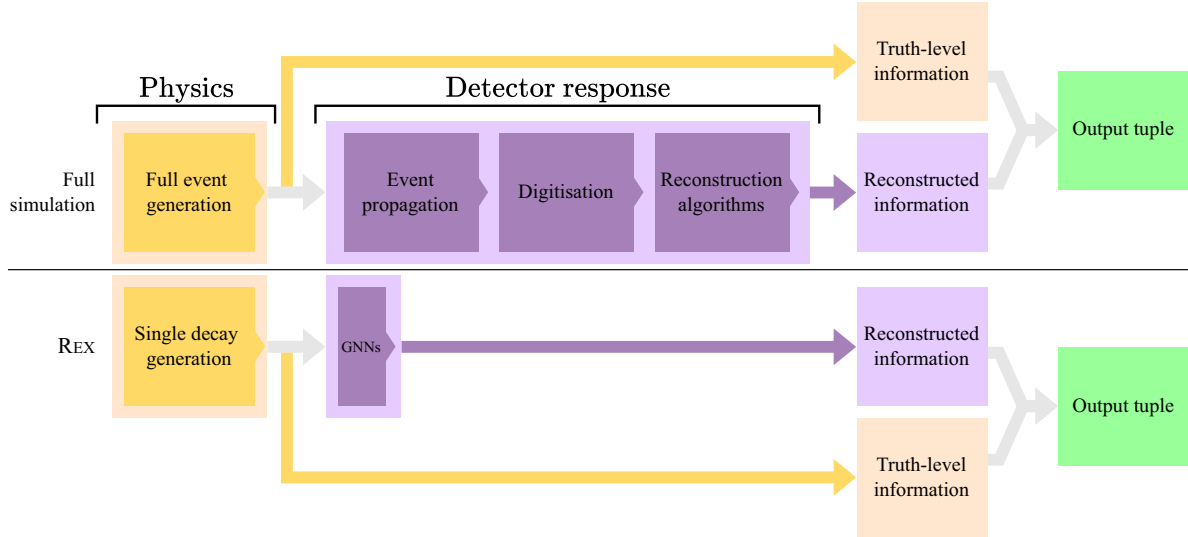


FIG. 1. Overview of the stages of the full simulation and those of the approach of Rex.

reconstruction algorithms to produce analysis-level variables. A diverse range of such efforts, in various stages of development, exists across the LHC experiments; some examples are available in Refs. [15–21]. Of particular interest is the approach presented by the CMS Collaboration in Ref. [19], which offers the high-fidelity generation of analysis-level variables in a single step. This so-called end-to-end approach has the most potential for speed, albeit with some trade-off in flexibility and interpretability. There is significant demand for such a tool to be developed for the requirements of the LHCb experiment. A well-designed tool would prove invaluable in reducing the load on existing simulation frameworks and reducing the existing dependence on parametric tools.¹ Additionally, it would enable novel studies that are currently impossible due to the computational demands of full simulations, such as exhaustive grid searches for relevant background processes.

This paper introduces *Rex*, a graph neural network (GNN) based tool that generates realistic sets of analysis-level variables on an event-by-event basis. These variables include reconstructed kinematic quantities, particle identification (PID) variables, and variables associated with track reconstruction and vertex quality. This generation is conditioned on the kinematics of each decay, which are directly available from existing event generators. These networks learn stochastic mappings with which they can generate a realistic detector output from these conditions, along with vectors of randomly generated noise. At no point does the model replicate the internal steps of the LHCb reconstruction algorithms; instead, it learns to

directly produce the required high-level output. Therefore, in deployment, this tool is isolated from the LHCb simulation and reconstruction software. The contrast between this approach and the full simulation is illustrated in Fig. 1.

Trained in a manner agnostic to any specific topologies or final states, *Rex* is designed to be as generalized as possible and is capable of modeling a broad range of decay modes commonly encountered in analyses of heavy flavor decays. This includes fully reconstructed modes, partially reconstructed decays with missing particles, modes involving misidentified tracks, and those affected by significant radiative losses. Whilst the model is trained on a sample of heavy flavor decays, *Rex* is able to produce representative samples for any multibody decay² at the LHCb.

The approach taken is founded on the premise that the detector response across a wide range of decay topologies is already well-represented in existing fully simulated samples. While the set of explicitly simulated modes within these samples is limited, missing modes will typically differ only by minor structural variations, such as an extra particle in the final state or an alternate resonance structure. Crucially, as the characteristics of the target high-level variables are governed by the same principles across all examples, by learning these principles, a model can accurately interpolate between examples. Whilst there is a stochastic nature to the full simulation, over large samples, the distributions of the target high-level variables follow reproducible, learnable patterns.

This paper is structured as follows. The description of the training sample is provided in Sec. II. An outline of *Rex* is given in Secs. III and IV describes the various network

¹Note that despite its approximate nature, RapidSim is still widely used within LHCb analyses in cases where selection effects can be estimated or are less important, see Refs. [22–27].

²For this initial implementation, reconstructed topologies are limited to having four, or fewer, final-state tracks.

architectures employed. This is followed by examples of performance in Sec. V, and then, finally, conclusions are presented in Sec. VI.

II. LHCb SIMULATION FRAMEWORK

This tool has been developed for the requirements of the LHCb experiment; therefore, a brief outline of the standard LHCb simulation framework is provided below.

In the LHCb simulation, pp collisions are generated using Pythia [28] with a specific LHCb configuration [29]. In a typical simulation workflow, for example, for that of $B^+ \rightarrow K^+ e^+ e^-$ decays, pp collisions are simulated until a B^+ meson is generated. In this case, the B^+ particle, the head of the decay, is then forced to decay into the target signal final state, $K^+ e^+ e^-$, using a predefined decay model. The kinematics of particle decays are described by EvtGen [30]. Final-state radiation is generated using Photos [31,32]. Each generated so-called event is made up of the B^+ meson decay products of interest and any other particles generated in the collision, which are all then passed to the detector response stage. The kinematics and identification of the involved particles are stored at this point; this is the so-called truth-level information.

The propagation of generated particles through the detector material and any interaction with the detector is then computed using the Geant4 toolkit [33], as described in Ref. [34]. This produces a simulated response in all detector subsystems. This is then digitized and passed through the same trigger, particle identification, and reconstruction algorithms as used during online data-taking; here, reconstructed track objects are combined to form reconstructed candidates. It is the reconstruction-level output quantities from these algorithms that form the final detector response that can be used to compute the derived quantities, such as efficiencies, required for physics analyses. It is these high-level quantities that are extracted from simulated events to create training samples for this tool.

A. Training data preparation and preprocessing

Training samples are extracted from the extensive archive of simulated data produced for LHCb analyses, covering the full range of previous simulation requests. For each simulated event, all final-state tracks are truth-matched [35] to ensure they are associated with single particles and originate from heavy meson decays; ghosts, tracks containing hits from multiple particles, and clones, multiple tracks matched to the same particle, are excluded. Valid tracks, and various subsets of these tracks, are systematically combined to form candidates representing every possible decay topology. These topologies include both direct multibody decays and configurations with reconstructed intermediate resonances. Tracks are reused as appropriate to cover all topologies and unique combinations. At no point does the original true topology of a given simulated event affect the candidates constructed. Such a

process creates an exhaustive sample of fully- and partially-reconstructed candidates. For the studies presented in this paper, only charged tracks from pions, kaons, muons, and electrons are included in the sample; neutral particles and protons are therefore excluded from the first iteration of Rex. Additionally, for this iteration, only final states of four tracks or fewer are considered; however, by employing graph representations for decays, this tool can naturally accommodate higher multiplicity final states with no change to the network architecture. For this study, only a small subset of the available simulated data is employed. Events are processed until there are 1.5 million training samples for each option of 2, 3, and 4 final-state particles; these are evenly balanced across the topologies for each option.

The first prototype of Rex produces the following variables for each track:

- (i) $\text{prob}_{\text{ghost}}$: The probability that a track is a fake track that does not correspond to the trajectory of a true particle, but originates from a mismatch of hits from separate particles or from detector noise, a so-called ghost [36];
- (ii) $\chi^2_{\text{TrackFit}}/\text{ndof}$: The fit quality of the Kalman filter-based track fit [36–38];
- (iii) χ^2_{IP} : A measure of how likely it is that a given track originated from the primary vertex (PV), computed as the difference between the χ^2 of the PV fit when including and excluding the track in question;
- (iv) PID_i : A set of particle identification variables computed by taking the difference in the log likelihood of a track being of a specific type (e.g., K , e , μ , p) relative to the π hypothesis. These variables incorporate information from the calorimeter, RICH, and muon systems [39];
- (v) ProbNN_i : A set of optimized PID variables obtained using multivariate techniques that combine tracking and PID information from each subsystem into a single probability value for each particle hypothesis, i is one of π , K , e , μ , p .

Then, for each reconstructed candidate, whether that is an intermediate resonance or the parent candidate at the head of the decay chain, the following variables are generated:

- (i) χ^2_{toPV} : A measure of how well the decay vertex of the reconstructed particle in question can be separated from the primary vertex;
- (ii) $\cos(\text{DIRA})$: The cosine of the direction angle (DIRA), the angle between the momentum vector of the reconstructed particle in question and the vector connecting the associated PV and the reconstructed decay vertex.
- (iii) χ^2_{IP} : A measure of how likely it is that the reconstructed particle in question originated from the PV, computed as the difference between the χ^2 of the PV fit when including and excluding the particle in question.
- (iv) $\chi^2_{\text{DV}}/\text{ndof}$: A measure of the fit quality of the secondary vertex fit [37,38].

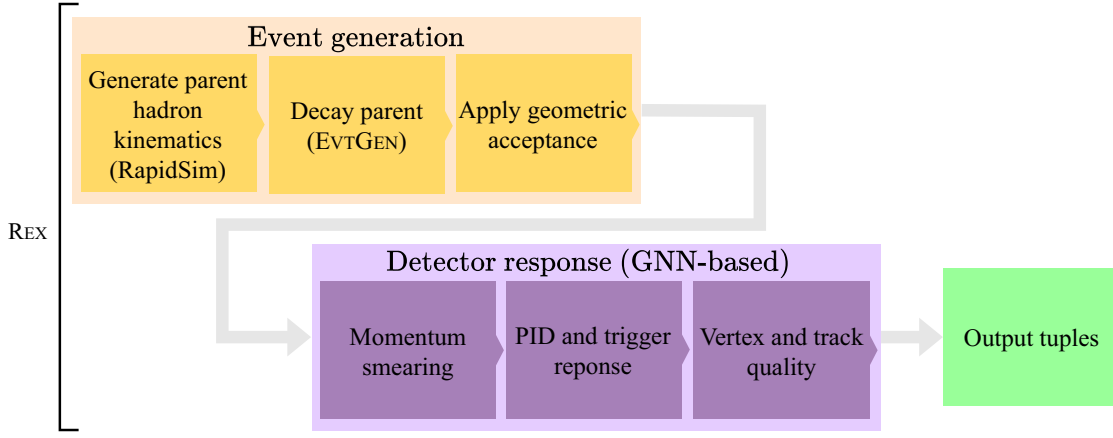


FIG. 2. Overview of the structure of Rex

Finally, for the parent candidate only, there are the vertex isolation variables ($VtxIso$), which employ event-wide information to determine whether or not any unaccounted for charged tracks also originate from the vertex in question. This set of target variables includes the majority of the high-level variables used in most of the physics analysis at LHCb. For any analysis pipeline requiring the generation of bespoke variables, retraining the network will be necessary; however, the software is designed to make this process as straightforward as possible.

Additionally, a set of conditional variables is extracted from each event, corresponding to a set of properties that guide the response of each network within Rex, as discussed in Sec. III. The conditional inputs include:

- (i) *Per-track variables* such as the true identification of the particle that formed each track, the true and reconstructed momenta and pseudorapidity of each track;
- (ii) *Track-pair variables* describing relationships between tracks; for example, the angle between their momentum vectors, and the angles between their reconstructed momentum vectors.
- (iii) *Head candidate-level variables* computed for the reconstructed head particle. These include quantities such as the DIRA, calculated from both true and reconstructed momenta, the total true and reconstructed momenta of the candidate, and the difference between the true momentum of the head particle and the sum of the true momenta of its daughter particles.
- (iv) *Intermediate candidate-level variables* are constructed similarly to those of the head candidate. However, since an intermediate particle in the reconstruction topology may not correspond to a physical particle, no truth-level quantities are available, and only reconstructed information is used.

The chosen set of conditional variables is designed to capture as much information as possible upon which the target high-level variables may depend.

The raw values of both the target and conditional variables in the training samples are preprocessed such that the distributions are easier for the network to handle. This constitutes mapping distributions into smoother representations, normalizing them to align with the chosen activation space, and reducing the density of events near the activation space boundaries. Quantile transformations are used to achieve processed distributions with these properties [40]. The distribution of each variable is mapped onto a truncated normal distribution that is subsequently normalized to the range $[-1, 1]$ to align with the $\tanh$ activation function. This is a monotonic and reversible transformation that preserves correlations between variables while being robust to outliers and large differences in scale between the core and the tails of distributions. This preprocessing is found to improve training stability and facilitate faster convergence.

III. FAST SIMULATION FRAMEWORK

An overview of the structure of the tool is provided in Fig. 2, and described in the following paragraphs.

Event generation This approach entirely bypasses Pythia and instead directly samples the kinematics of single heavy-flavor hadrons from fixed-order next-to-leading-order predictions of charm and bottom production at the LHC [41,42]. In collisions at the LHC, heavy-flavor hadrons are often produced in pairs. Any influence of the corresponding partner hadron or of particles from the rest of the event is implicitly included on average by training on fully simulated events, which were generated with Pythia and within which event-wide effects are present. This is demonstrated in Sec. VC, where the generated reconstruction performance is shown to closely match that of the full simulation, rather than resembling an idealized case. The location of the PV is generated using a dedicated network trained on fully simulated events, which is introduced in Sec. IVA, and the lifetimes of any particles,

along with any flight distances, are modeled within the RapidSim package.

Particle decays are handled with the same generator as in the full simulation, EvtGen. This ensures a physical description of decay amplitudes and allows users to generate events with standardized physics models. Final-state radiation is also included with the same library as the full simulation, Photos.

In pp collisions at LHCb, heavy-flavor hadrons and their decay products are produced in 4π steradians. In the full simulation, the geometrical acceptance is naturally captured. However, in this fast approach, the geometrical acceptance must be approximated. This is achieved by requiring that all final-state particles fall within a specified pseudorapidity range. Events are only retained in cases where all final-state particles are within this range. A more accurate modeling of this acceptance will be included in a later version *Rex*.

Detector response The modeling of the detector response is decomposed into three interdependent components: momentum smearing, generation of particle identification and trigger responses, and vertexing. While each component is interdependent during inference, during training, each network is trained independently. Each component is handled with a GNN, the architecture of each is described in Sec. IV. This modular decomposition is motivated by considerations of both training efficiency and flexibility during inference. With this treatment, it becomes possible to update the network weights of specific stages as needed. For instance, one could substitute a different model for momentum smearing without retraining the entire system. The networks are implemented using PyTorch Geometric [43].

Inference and output *Rex* is available as a lightweight Python package with minimal dependencies and is isolated from the extensive LHCb software stack. The package interfaces with RapidSim automatically, executing queries in the background before running the required networks and then finally organizing the output. The generated output is formatted such that it mirrors that of the full simulation and can therefore act as a drop-in replacement. The package handles the computation of the reconstructed masses of all intermediate and parent particles, automatically accounting for different mass hypotheses in the case of misidentified tracks.

To provide a sense of timing performance, generating analysis-ready output for a sample of ten million $B^+ \rightarrow K^+ \mu^+ \mu^-$ decays takes approximately one hour on a 64-core AMD EPYC 7702P CPU. This time includes the RapidSim and EvtGen-based event generation, which accounts for a similar portion of the total time taken as the detector response emulation. This is a speed-up of $\mathcal{O}(10^5)$ over the full simulation using a typical approximation that each event takes one minute [44]. Such a significant speed-up enables on-demand generation, which has the potential to reduce the pressure of long-term storage

of simulated samples significantly. Furthermore, such a workflow of generating on demand could enable a partial decentralization of the simulation process, helping to reduce administrative overhead for the Collaboration.

IV. NETWORK ARCHITECTURE

During inference, the networks used within this tool are the generators of conditional generative adversarial networks (GANs) [45]. In the standard GAN framework, a generator network is trained to produce synthetic samples that are indistinguishable from the training samples. To enable this, during training, updates to the weights of the generator network are driven by a second network, the discriminator network. The goal of the discriminator network is to distinguish between real and generated samples. The two networks are trained in tandem in a technique known as adversarial training. Updates to each network are performed iteratively via backpropagation, with losses being computed over minibatches of training data. In the simplest GAN setup, the generator learns to produce samples using only random noise vectors as inputs. However, the GANs employed here are all conditional, that is, they have additional input vectors containing values that represent physical characteristics, as described in Sec. II A, that can be used to guide the generation. These variables are also shown to the discriminator so that it can flag generated samples with characteristics that are incompatible with the conditional inputs.

To emulate the detector response, these networks are structured as heterogeneous graph neural networks [46] (HGNNs). GNNs are a broad class of models designed to operate on graph-structured data, where nodes represent entities (such as particle tracks) and edges represent interactions between those entities. Information is propagated across a graph via message-passing operations, in which each node exchanges messages with its connected neighbors, then updates its internal representation based on any received information. These updates often include pooling operations, aggregating incoming information at nodes connected to multiple neighbors. Details of the message-passing operations employed in these networks are provided in the following sections. HGNNs extend the framework of GNNs, adding flexibility that allows for multiple types of nodes and edges, each with its own learnable update mechanism. For instance, different node types may represent tracks formed by different species of particles. In this case, edge types would encode directed, typed interactions between those tracks; for example, an edge from a kaon track to an electron track may be treated differently from one in the reverse direction. This increased flexibility of HGNNs enables them to capture complex dependencies for a wide variety of graph structures, making them well-suited to the task at hand. These qualities of HGNNs have been exploited in other particle physics applications, see Refs. [47–49]. The structure of each

network and the construction of each graph are detailed in the following sections.

Training was performed on NVIDIA Tesla T4 (16 GB) GPUs. The HGNN-based networks were each trained for approximately 48 hours.

A. Primary vertex smearing

The process outlined in Sec. III, with which the kinematics of the head particle of each decay are generated, does not produce a generated location of the PV. As the reconstruction quality has a dependence on the location of the PV, a realistic distribution of PVs is required in the generated sample. We train a very simple GAN constructed of fully connected layers to generate the true location of PVs using a training sample extracted from fully simulated events. This GAN is conditioned on the true momenta of the decaying meson to account for boosting effects. When queried, the generated smearing is passed on to all downstream vertices such that any separation between vertices is maintained.

B. Momentum smearing

Momentum smearing constitutes the first stage of the detector response model. The structure of this network is designed to handle two effects. Firstly, the scale of resolution effects varies across particle species, most notably for electrons, whose tendency to emit bremsstrahlung radiation significantly degrades their reconstruction quality at the LHCb. Secondly, global properties of an event can impact the momentum resolution. For example, in the case that two particles are emitted with very small angles between them, they can influence each other's reconstruction. Consequently, to treat each track in isolation with independent queries of a network would be an incomplete solution.

To address both these challenges, we employ a heterogeneous graph-based approach using multiple node types. Each event is structured into an all-to-all connected graph within which each node represents a reconstructed track and is assigned a type corresponding to its particle species,

independently of charge. The graph incorporates multiple edge types that facilitate track-to-track communication. In this setup, a distinct edge type is defined for each ordered pair of node types. For example, *kaon-to-electron* and *electron-to-kaon*, and these are used to connect each node to every other. This structure gives the network flexibility to capture any differences that may exist in the various species-specific interactions. Each graph is initialized with vectors of conditions and vectors of uniquely generated random noise at each node; additionally, the angles between particles are provided as conditional edge features in the initialization. Each graph structure is then passed through the network, resulting in vectors of output at each node.

During a pass of the network, the hidden representations of each graph are updated multiple times by a sequence of self-loop layers and communication layers. Illustrations of the structure of the nodes and edges in these layers are presented in Fig. 3. In the self-loop layers, each node type is processed by a distinct and isolated linear transformation. In the communication layers, information is exchanged between nodes along the edges via graph attention convolution layers [50] (GATs) in the process of message-passing. In a standard graph convolution, for each target node, the hidden states of its neighbors (and optionally the node itself) are first linearly transformed to create a collection of incoming messages; these are then aggregated by pooling to update the state of the target node. GATs extend this operation to include a learnable attention mechanism that assigns a weight to each message to scale its relative importance before pooling. Since intertrack correlations are expected to introduce only subdominant modifications to the smearing, we restrict this message passing so that only a fixed subset of the hidden representation of each node is shared with its neighbors. The rest of each hidden representation is not updated by the communication layers. The subset of nodes used in the communication is defined before the training process.

This heterogeneous node-based approach removes the need for the particle species labels to be included in the conditional

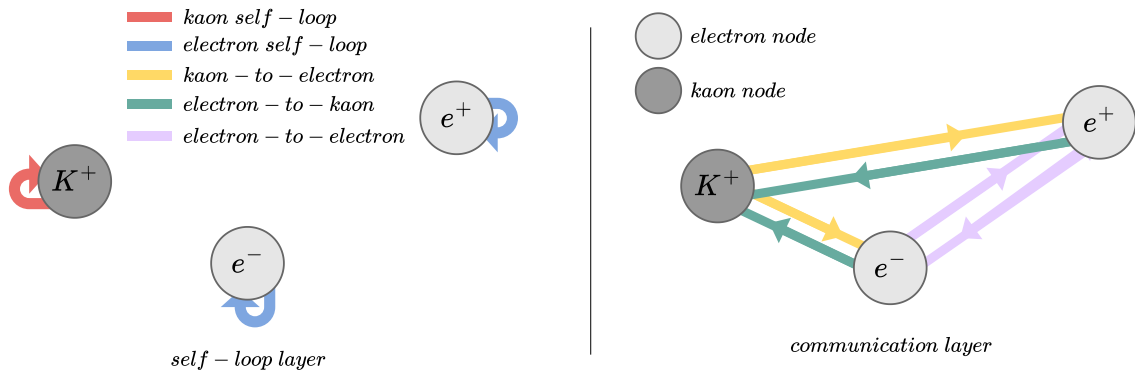


FIG. 3. Example of the edge connection structure of a graph employed during a communication layer of the smearing and PID networks.

input and instead encodes this information into the graph structure itself. This approach automatically handles imbalances in the training sample, with the weights of particle-specific nodes and edges never being updated unnecessarily.

Additionally, this heterogeneous graph formulation enables species-specific preprocessing. This is helpful when the scale of input variable distributions differs significantly across particle types. This is true for the distributions of momentum residuals, where electrons, being more poorly reconstructed, exhibit a much broader spread than other species. An illustration of this effect and both a common and a species-specific preprocessing strategy is shown in Fig. 4. In the case of a single global preprocessing, the majority of the electron distribution exists in the tail regions of the processed distribution, which is undesirable. The species-specific preprocessing avoids this, significantly improving the stability of the network during training and speeding up convergence.

C. PID

The network architecture employed here is the same as that of the momentum smearing, with a unique node type per particle species. The target variables here are the PID variables PID_i for each of K , e , μ , and p , and ProbNN $_i$ for each of π , K , e , μ , and p . The conditional input variables used to query this network are the same as those of the smearing network; however, they now also include the generated reconstructed momenta, such that the generated PID information is correctly correlated to the momentum

reconstruction quality. Again, unique preprocessing per node type is advantageous, as the distribution of each target variable is very different, being variables designed to separate different particle species.

D. Vertexing

The architecture of the vertexing network is significantly different. Here, a graph represents the reconstruction of a decay tree from track objects. Within each graph, there are now only three possible node types; track nodes, nodes representing reconstructed intermediate candidates and nodes representing parent candidates. Each graph is initialized with vectors of conditions and of uniquely generated random noise at each node. Within each graph, nodes and their connecting edges are organized in a hierarchical arrangement in accordance with the requested reconstructed topology. Some examples are provided in Fig. 5; these include cases of partially reconstructed decays in which one or more final state particles are not reconstructed; missing particles are indicated in the graphs with faded nodes. There is an edge type for all the possible connections, such as track-to-intermediate or parent-to-track, and each edge type has a unique set of weights. All edges in the network are directional, and information is propagated within a graph sequentially, as illustrated in Fig. 6. Within each layer of the generator network, two passes are made; first, information flows *down* the graph from the reconstructed parent candidate, through any reconstructed intermediate objects, to the track nodes. This *downward* pass allows the network to

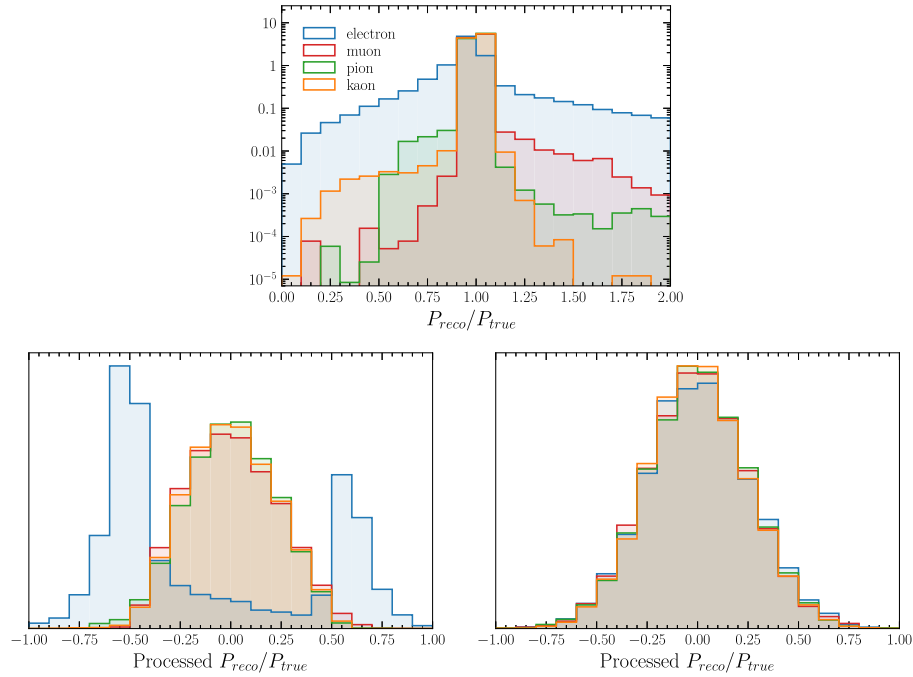


FIG. 4. Examples of pre-processing for a variable describing the level of momenta smearing, the ratio of the reconstructed momenta P_{reco} to the true momenta P_{true} . The physical distribution of the variable (top), the distribution using shared preprocessing (left), and unique per-particle-type preprocessing (right) are all provided.

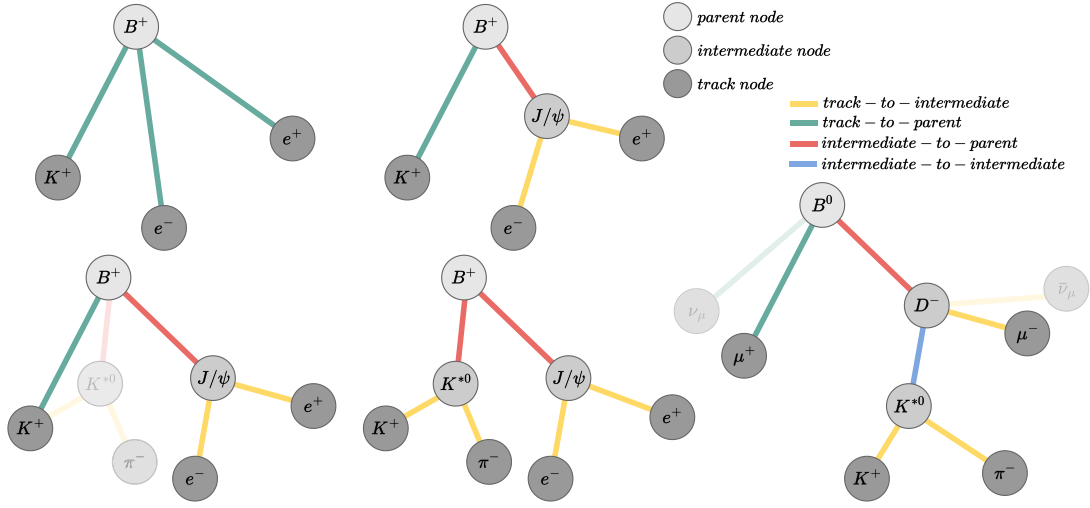


FIG. 5. Example edge structures within graphs passed through the vertexing network, the examples cover various 3- and 4-body reconstruction topologies. Faded nodes indicate particles missed; these are not included in the graph structures that encode the reconstruction topology.

correlate track-level features with properties of the reconstructed objects. This is particularly important for variables such as the track χ^2_{IP} , which encode how the parent candidate's χ^2_{IP} changes when a track is excluded. Then, a second pass proceeds *up* the graph. This *upward* pass is motivated by the existing LHCb reconstruction algorithms, which build candidates having access solely to reconstructed track momenta. The final pass being *upward* is stylistically important, as it ensures that any downstream nodes remain unchanged and the representations left behind are the exact representations used in any pooling operations. For the discriminator network, the order of these passes is reversed.

On top of the primary edges connecting nodes in accordance with the reconstructed hierarchy, additional skip connections were trialed. These were added to allow direct communication between otherwise disconnected nodes, for example, the K^+ and the J/ψ nodes of the graph in Fig. 6. There was no significant improvement in the performance, so these are avoided for simplicity. The hierarchical structure without skip connections is also philosophically closer to how the existing LHCb reconstruction algorithms build up reconstructed candidates.

For this network, we implemented a customized version of the PyTorch Geometric heterogeneous graph class that allows for node-independent pooling during the propagation up each graph, as illustrated in Fig. 7. In the standard heterogeneous graph class, pooling at nodes occurs in two stages; firstly, hidden representations from any edges of each edge type are pooled, then secondly, these resulting representations are pooled across edge types. This approach is problematic for this application as it would change the relative importance of each edge across different cases. We avoid this by combining representations in a single step, agnostic of edge type. Three pooling operations are used simultaneously; min, max, and

sum. A learnable attention mechanism then allows the network to adapt the importance of each operation for each dimension of the hidden representation.

This architecture incorporates minibatch discrimination (MBD) layers [51], which enables the discriminator to employ information from across a small batch of samples rather than just assessing each example independently. This was found to improve the diversity of the generated outputs. This is particularly valuable when modeling variables that are only weakly correlated with the conditioning inputs, such as $\chi^2_{\text{TrackFit}}/\text{ndof}$ and $\text{prob}_{\text{ghost}}$. MBD layers are applied both at the level of individual track nodes and on the graph-level pooled representation before its final compression into the discriminator output.

The conditional inputs at each track node type include information about the true kinematics, the generated reconstructed kinematics, and trigger and PID information. As described in Sec. II A, different conditions are given as inputs to nodes of each type. Importantly, all conditional variables used are derived exclusively from lab-frame observables; quantities such as reconstructed invariant masses are deliberately excluded. This design choice aims to improve the network's ability to generalize across different decay modes and more easily cover modes not shown during training. This avoids learning about specific resonant structures present in the training data. For example, in decays such as $B^+ \rightarrow K^+ J/\psi (\rightarrow \mu^+ \mu^-)$, providing the invariant mass of the dimuon system as a condition would implicitly encode knowledge of the J/ψ resonance. The network may then associate large deviations from the J/ψ mass with poor reconstruction quality. If the network were subsequently queried on nonresonant $B^+ \rightarrow K^+ \mu^+ \mu^-$ decays, it would incorrectly generate variables indicating poor reconstruction quality for events with dilepton invariant mass far from the J/ψ resonance. This approach also

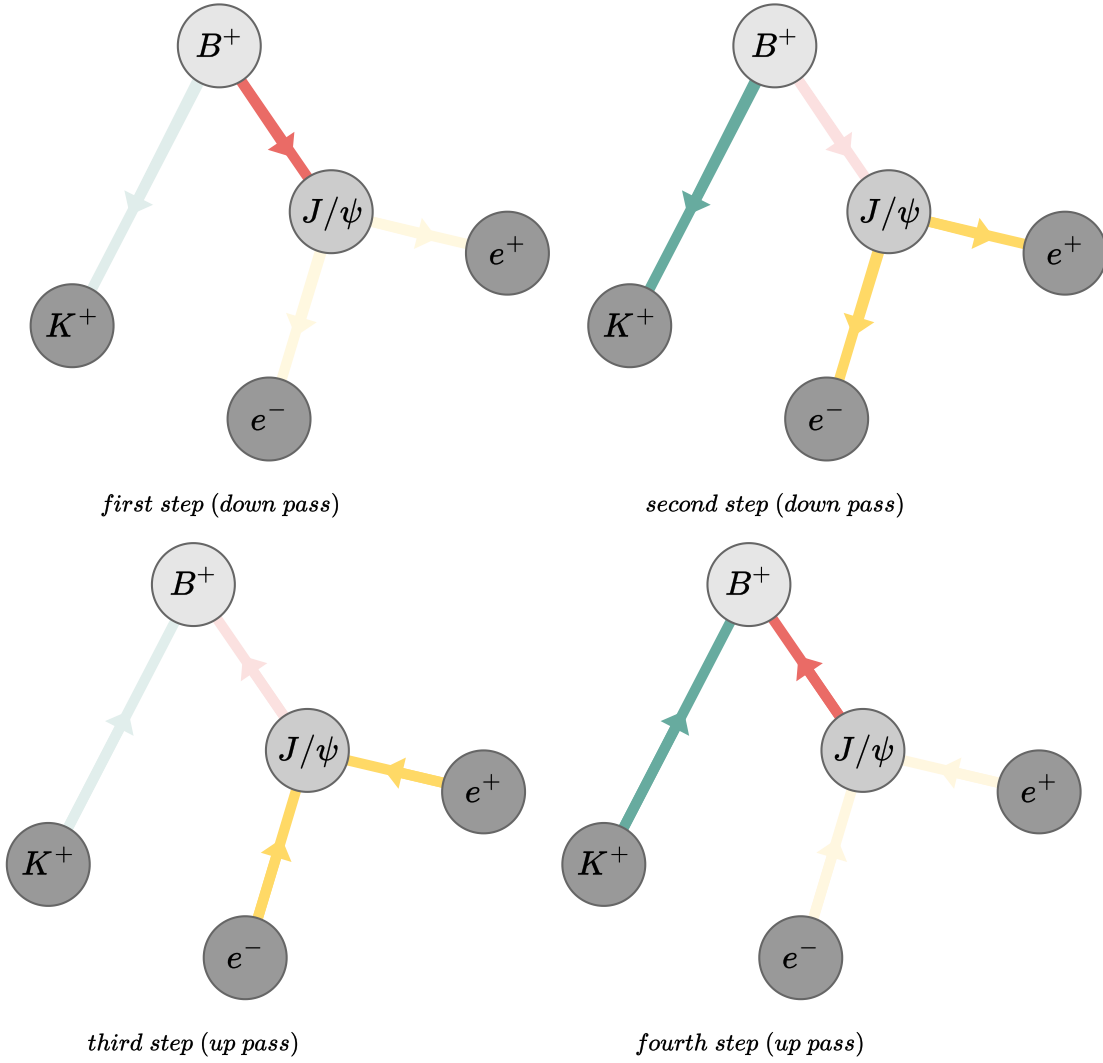


FIG. 6. The stages of the hierarchical propagation of information during a single layer of the generator of the vertexing network, for the example of $B^+ \rightarrow K^+ J/\psi (\rightarrow e^+ e^-)$. Faded edge connections indicate inactive edges in a given step.

mirrors the LHCb vertexing algorithms, which only have access to lab-frame variables when constructing candidates.

One condition worth highlighting is that which encodes the physical separation between the origin vertices of final

state particles and the decay of the parent particle. Separation is possible in cases where a decay has occurred via a long-lived intermediate state, such as a D^0 . In the example of $B^+ \rightarrow \bar{D}^0 (\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$, the $\bar{D}^0$ meson has a long enough lifetime to create a physical separation between the B^+ decay vertex and the origin vertex of the K^+ and e^- .

Regarding training convergence, no formal stopping criterion has yet been implemented—a limitation acknowledged in Sec. VI as a target for future development. Training is currently stopped based on a qualitative assessment of when the performance on an unseen validation subsample of the training data ceases to improve.

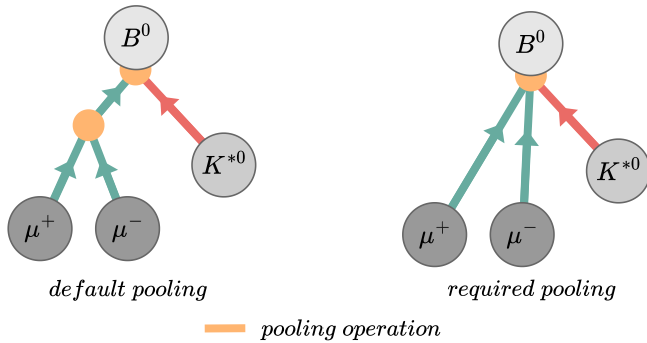


FIG. 7. Example of a single vertex within a wider graph indicating two possible approaches to pooling hidden representations.

V. PHYSICS PERFORMANCE

Within this section, the performance of each component of Rex is presented, followed by an example analysis case study and a discussion on uncertainties.

A. Momentum smearing

The true momenta of each particle, as generated by EvtGen, are smeared to account for detector resolution by querying the smearing network with node types assigned according to true particle species. Figure 8 shows a comparison of distributions of $\Delta E/E$ between generated and simulated events, for different final state particles. Note the drastically different scale of the smearing for the electrons. Also provided in Fig. 8 are distributions of the reconstructed B -mass for the examples of $B^+ \rightarrow K^+ e^+ e^-$ and $B^+ \rightarrow K^+ \mu^+ \mu^-$. The individual distributions of $\Delta E/E$ appear well-modeled, and this translates into the full reconstructed masses, which show the behavior is correctly captured out into the tails of the distribution. In each case, the network correctly picks up the required asymmetry in the smearing.

As described in Sec. IV B, the advantage of smearing the momenta of particles on an event-by-event basis rather than independently is that the smearing can capture relationships between individual track smearings. Figure 9 shows the $\Delta E/E$ distributions for the ℓ^+ in two examples of $B^+ \rightarrow K^+ \ell^+ \ell^-$ decays, where $\ell = e, \mu$. Whilst the overall agreement is imperfect, for cases in which the angle between the leptons is small, the network correctly degrades the resolution. This is only possible because of the communication within the graph structure.

B. PID

The efficiencies of an example cut of $\text{PIDK} > 0$, as a function of $\log(p_T)$ and η for supported particle types, are presented in Fig. 10. The network correctly models the

shape and scale of these efficiencies. A full set of PID variables and similar efficiency maps for a variety of example cuts are presented in the Appendix (see Figs. 20 and 21). In general, the model performs well, capturing the efficiency dependence across kinematic variables. Some variables, though, appear systematically mismodeled. In particular, the distribution of PIDE for electrons. We observe that this can be corrected by weighting events such that the distribution of B kinematics more closely aligns with Pythia. Improving the generation of B kinematics is listed as an area for future development in Sec. VI. It is worth noting here that PID variables are one of the most poorly modeled parts of the LHCb full simulation, and it is therefore essential for analysts to weight even fully-simulated events to correct for this. Comprehensive data-driven tools have been developed within the Collaboration to standardize this procedure [52]. One is able to apply the same tools to weight samples generated with Rex, where any generated correlations with other variables will be included.

C. Vertexing and reconstruction

Distributions of vertexing and reconstruction quality variables are presented in Fig. 11 for two example channels, $B^+ \rightarrow K^+ \mu^+ \mu^-$ and $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$. Despite both having three final-state charged tracks, these channels have significantly different decay topologies, being fully- and partially-reconstructed decays, respectively. The final state particles in the $B^+ \rightarrow K^+ \mu^+ \mu^-$ decays all originate from the B^+ decay vertex; this is not the case for the $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ decay in which the intermediate state

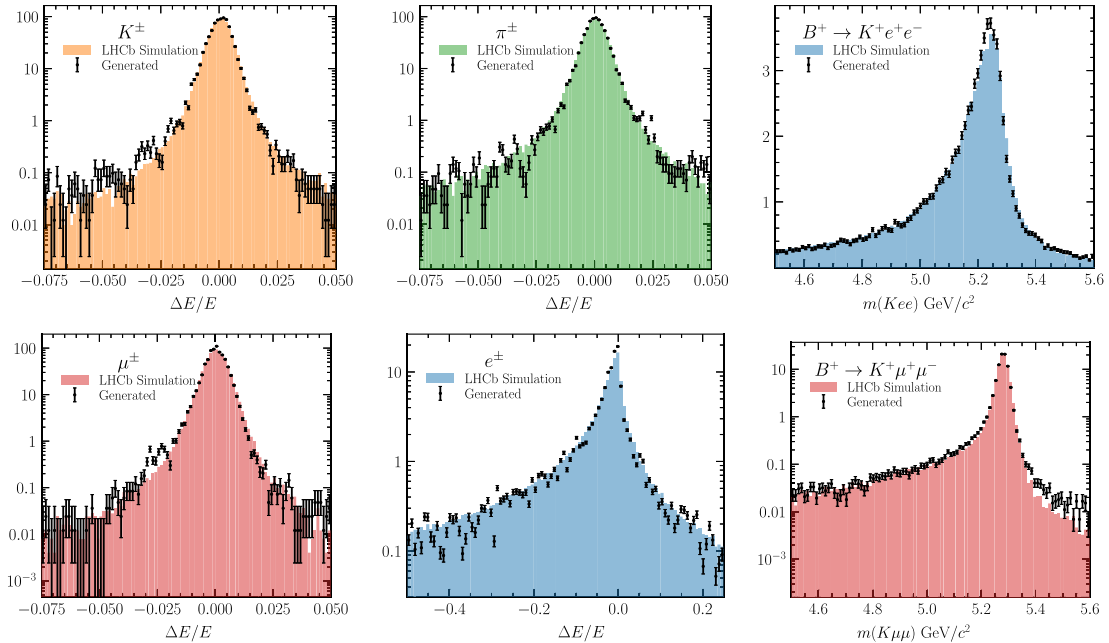


FIG. 8. Distributions (left and centre) of $\Delta E/E$ for each particle type, and (right) reconstructed- B mass distributions for $B^+ \rightarrow K^+ e^+ e^-$ and $B^+ \rightarrow K^+ \mu^+ \mu^-$. Both simulated and generated distributions are provided in each case.

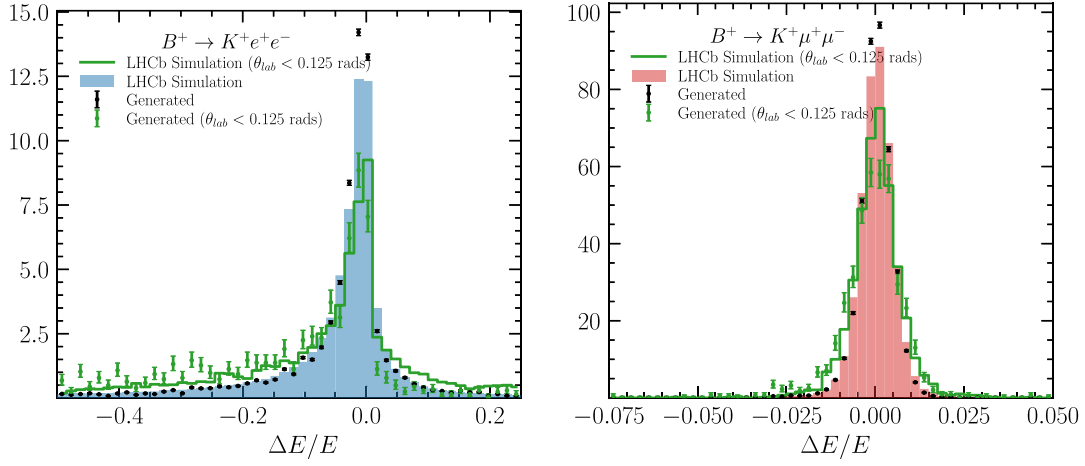


FIG. 9. Distributions of $\Delta E/E$ for the ℓ^+ in $B^+ \rightarrow K^+ \ell^+ \ell^-$, where ℓ is e (left) and μ (right). Both distributions are shown twice, firstly for all events, and then just for events where the angle between the leptons in the lab frame was smaller than 0.125 rads.

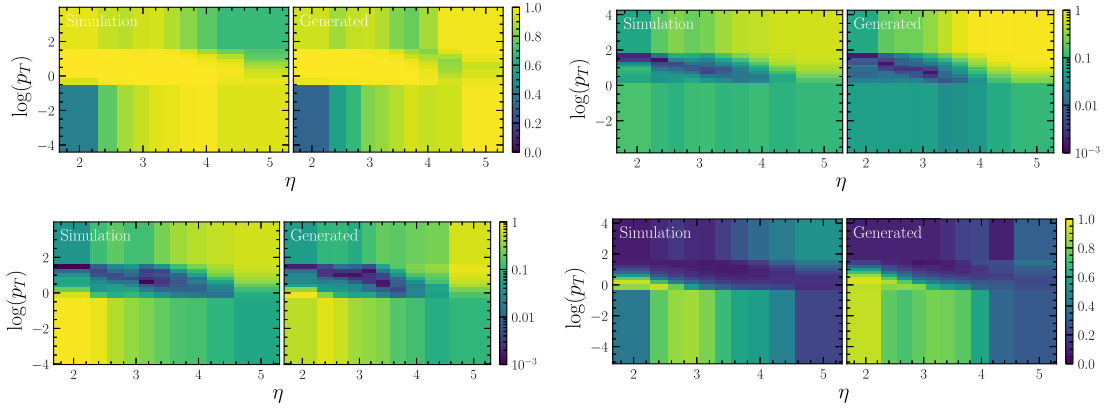


FIG. 10. A comparison between simulated and generated maps of the efficiency of the cut “PIDK > 0” across $\log(p_T)$ and η for each true particle type. The binning scheme is adapted with quantiles to keep the number of events in each bin similar.

$\bar{D}^0$ has a lifetime long enough to create some physical separation, meaning the K^+ and e^- originate from a vertex displaced from that of the B^+ . This is accounted for in the graph structure parsed within which the $\bar{D}^0$ intermediate is reconstructed. Consequently, the variables associated with this intermediate state are additionally presented for this channel in Fig. 11. Illustrations of the generated and reconstructed topologies of $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ are shown in Fig. 12, where, as the neutrino is not detected, it does not appear in the reconstructed topology. Note that the variables presented in Fig. 11 only depend on the choice of reconstructed topology; they are unaffected by any potential track misidentifications and incorrect mass hypotheses. Several variables appear imperfectly modeled, which is likely due to instabilities during the convergence of the network. As mentioned in Secs. V E and VI, this is an area for future development.

An accurate description of these variables is achieved, with the model successfully capturing differences in kinematic and topological features between the two

example decay modes. The clearest examples of differences are in the distributions of the vertex fit quality χ^2_{DV}/ndof and impact parameter significance χ^2_{IP} of the parent particle, where the distributions in the partially reconstructed case are significantly broader. This broadening is attributed to degraded vertex resolution caused by momentum lost in the undetected neutrino and the spatial displacement due to the flight distance of the intermediate $\bar{D}^0$.

To further explore the origin of the degradation in the $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ case, we partition the data into subsets containing the most extreme examples from the sample. In Fig. 13, distributions of the same χ^2_{DV}/ndof and χ^2_{IP} variables are shown for events with the hardest (2.5th percentile) and softest neutrinos. Additionally, the sample is divided according to the largest and smallest $\bar{D}^0$ flight distances, which is also shown in Fig. 13. In each case, the network correctly models the impact on these variables.

In the examples shown up to this point, the $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ decays are correctly reconstructed with the K^+ and e^- tracks forming the intermediate candidate. We can

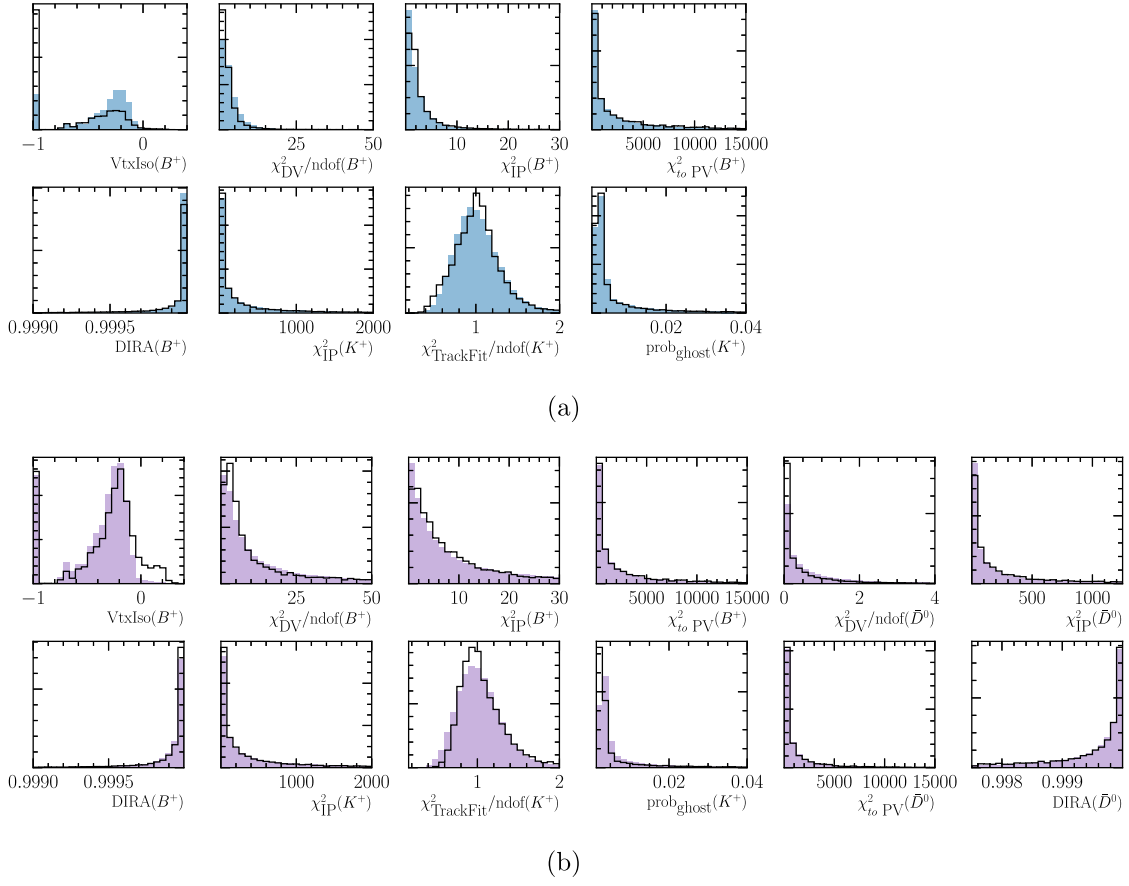


FIG. 11. Distributions of generated vertexing and reconstruction quality variables for the channels $B^+ \rightarrow K^+ \mu^+ \mu^-$ and $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$. The full simulation is shown with the filled histograms, and the empty stepped histograms are derived from the output of Rex.

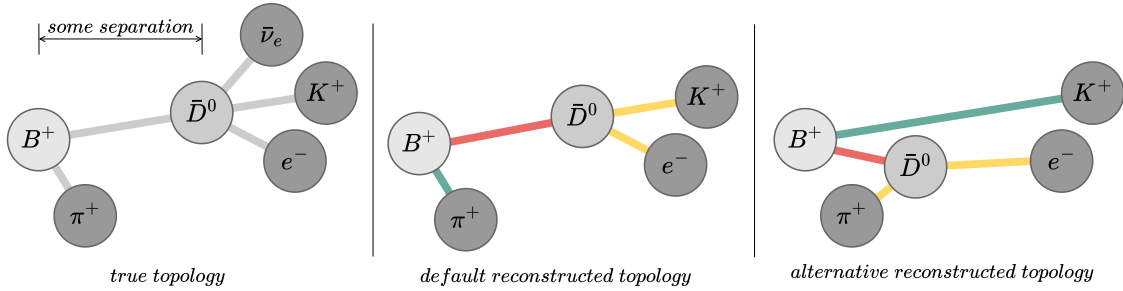


FIG. 12. Illustrations of the generated and reconstructed topologies referred to in the text.

look at how the generated variables change in the case that one changes the reconstruction hypothesis and reconstructs the same events with the e^- and π^+ tracks combined in the reconstruction to form the intermediate candidate, illustrated as the alternative reconstruction topology in Fig. 12. In Fig. 14, we see that the effect on the distributions of these variables is again modeled correctly.

As discussed in Sec. III, we avoid the generation of the rest of the event in the interest of speed and instead include the impact from the surrounding event activity by training

on the full simulation in which the effect is present. With this approach, the model effectively samples the size of the effect to apply such that it is consistent with the simulated conditions of the decay under study. There is, though, no explicit conditional control of the size of the effect, but it is nevertheless included in the output. While the impact from the rest of the event varies from event to event, it is generally correlated with event occupancy—a proxy for which is the number of reconstructed tracks, with vertexing quality improving in events with fewer reconstructed

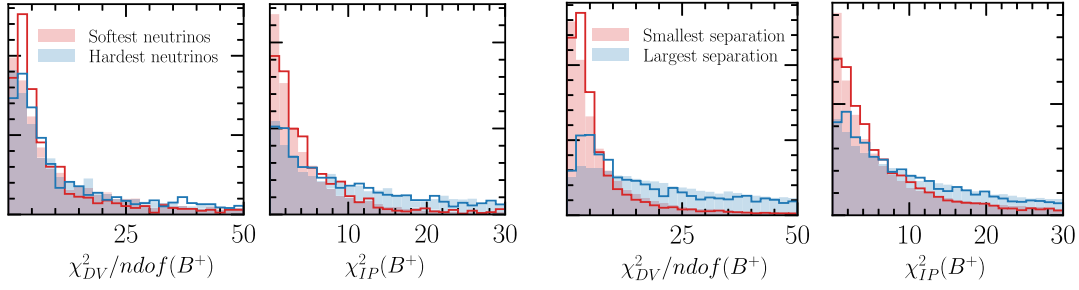


FIG. 13. Distributions of variables for the reconstructed B^+ in $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ events in extreme cases of $\bar{\nu}_e$ momentum and of $\bar{D}^0$ flight distance. The full simulation is shown with the filled histograms, and the empty stepped histograms are derived from the output of Rex.

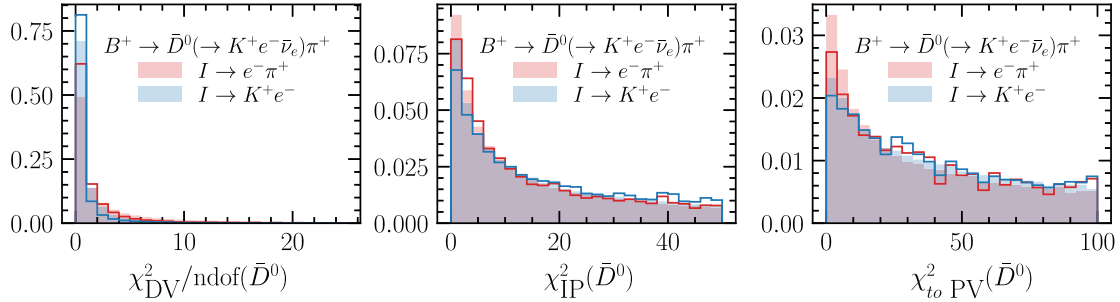


FIG. 14. Distributions of generated vertexing quality variables assigned to the reconstructed intermediate of $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+$ in the correctly reconstructed case (blue) and swapped case (red). The full simulation is shown with the filled histograms, and the empty stepped histograms are derived from the output of Rex.

tracks. We demonstrate in Fig. 15 that the generated distributions follow the full distributions rather than the idealized low-occupancy distributions. In future versions, the network should include occupancy variables in the output so that any correlations with other variables are modeled and in the case that an analyst wishes to include cuts on occupancy in a selection, the events that pass have the correct properties.

D. Example analysis selection chain

To assess the performance of Rex in a realistic analysis context, we emulate the full sequence of an analysis selection criteria on both signal and a range of background channels. For this study, we look at the analysis of the $B^+ \rightarrow K^+ e^+ e^-$ decay. This channel was selected as its associated backgrounds represent a broad range of topologies commonly encountered within other LHCb analyses; these backgrounds also include tracks from all particle species supported within this version of Rex. Channels with electrons are particularly useful test cases, as electrons are both underrepresented in the training sample and exhibit the most distinctive reconstruction characteristics.

All signal and background candidates are reconstructed under the same decay hypothesis of $B^+ \rightarrow K^+ I(\rightarrow e^+ e^-)$, where I denotes a fictitious intermediate state representing the dilepton system. The squared invariant mass of the

intermediate state I defines the variable q^2 . The following background processes are considered:

- (i) $B^+ \rightarrow K^+ J/\psi(\rightarrow e^+ e^-)$: This channel closely resembles the signal but is confined to a narrow q^2 region corresponding to the invariant mass of the J/ψ ;
- (ii) $B^0 \rightarrow K^{*0}(\rightarrow K^+ \pi^-) e^+ e^-$: This channel contributes as a partially reconstructed background in cases where the π^- is not detected;
- (iii) $B^+ \rightarrow \bar{D}^0(\rightarrow K^+ e^- \bar{\nu}_e) \pi^+_{e^+}$: This is another partially reconstructed channel, due to the undetected neutrino; however, this channel presents several additional challenges for this tool. As discussed in Sec. V C, the finite lifetime of the $\bar{D}^0$ meson results in a spatial separation between the decay vertices of the final-state particles, which can distort vertex-related observables. Furthermore, the true intermediate particle is the $\bar{D}^0$, formed from the K^+ and e^- , whereas the reconstruction assumes the intermediate state I is composed of the e^- and π^+ , introducing additional effects which must be accounted for in order to correctly reproduce the characteristics of fully-simulated events. Finally, this background includes a particle misidentification, where the π^+ is incorrectly reconstructed as an e^+ ;
- (iv) $B^+ \rightarrow K^+ \mu^+_{e^+} \mu^-_{e^-}$: This background shares the same topology as the signal but involves two misidentified

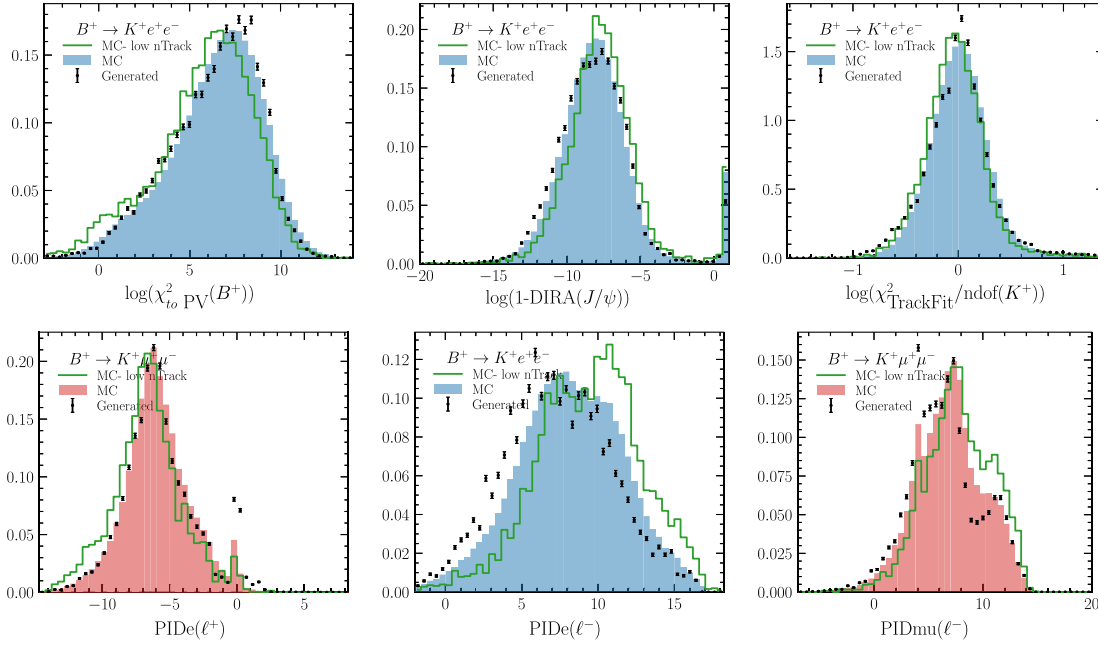


FIG. 15. A comparison between distributions of various reconstruction variables in simulated and generated $B^+ \rightarrow K^+e^+e^-$ and $B^+ \rightarrow K^+\mu^+\mu^-$ samples reconstructed with the hypothesis of $B^+ \rightarrow K^+J/\psi(\rightarrow \ell^+\ell^-)$. The simulated sample is additionally shown with a cut on the occupancy of the event.

muons reconstructed as electrons. This background is well controlled by PID requirements; however, the performance of the tool can still be studied here before these selection requirements are made.

A typical analysis selection includes cuts on reconstruction quality variables, PID requirements, a boosted decision tree (BDT) to suppress combinatorial background, and finally, cuts on reconstructed invariant mass variables. The cuts used for this example are taken from Ref. [53] and are detailed in Table I. These cuts serve to ensure that basic reconstruction quality requirements are met. Following this, the BDT of Ref. [53] is employed to reduce contamination from combinatorial background candidates formed in random combinations of unrelated particles from the same event. The BDT is trained to separate fully simulated signal candidates, labeled as 1, from candidates in the data selected from a region known to be dominated by combinatorial background, labeled as 0. At no point is the BDT trained on candidates generated using Rex. This BDT uses the variables listed in Table I, this list does not include PID variables nor vertex isolation variables, which are the least well-modeled variables in Rex. However, this is a typical list of the most important features used across LHCb to suppress combinatorial backgrounds. Accurately reproducing the efficiency of this BDT requires the model to capture both the individual variable distributions and the correlations between them correctly. Figure 16 shows the output of the BDT for all events (top panels) and for those that pass the cut-based selection (bottom panels), for each channel. The distributions are well-reproduced in all cases, both before and after the

cut-based selection is applied. Note that the distributions are normalized before selection, meaning that both the shape and scale must be correctly modeled for good agreement. For the rest of the examples in this paper, events are selected with a threshold of > 0.9 on the BDT output. The efficiencies of each individual cut, the BDT, some PID cuts (also from Ref. [53]) and the kinematic B -mass cuts as used in the analysis, along with the running efficiency, are shown in Fig. 17.

The top panels of Fig. 18 show the distribution of reconstructed B -mass in the hypothesis of $B^+ \rightarrow K^+e^+e^-$ after both the cut-based and the BDT selection are made. The bottom panels show the efficiency of the

TABLE I. List of selection requirements and variables which are included in the combinatorial BDT.

Selection cuts				Included in BDT	
B^+	$\chi^2_{to PV}$	> 100	B^+	$\chi^2_{to PV}$	✓
	$\cos(\text{DIRA})$	> 0.995	B^+	$\cos(\text{DIRA})$	✓
	χ^2_{IP}	< 25		χ^2_{IP}	✓
	χ^2_{DV}/ndof	< 9		χ^2_{DV}/ndof	✓
I	$\chi^2_{to PV}$	> 16	I	$\chi^2_{to PV}$	✓
	$\cos(\text{DIRA})$			$\cos(\text{DIRA})$	✓
	χ^2_{IP}			χ^2_{IP}	✓
	χ^2_{DV}/ndof	< 9		χ^2_{DV}/ndof	
All tracks	$\text{prob}_{\text{ghost}}$	< 0.3	All tracks	$\text{prob}_{\text{ghost}}$	
	χ^2_{TrackFit}	< 3		χ^2_{TrackFit}	✓
	χ^2_{IP}	> 9		χ^2_{IP}	✓

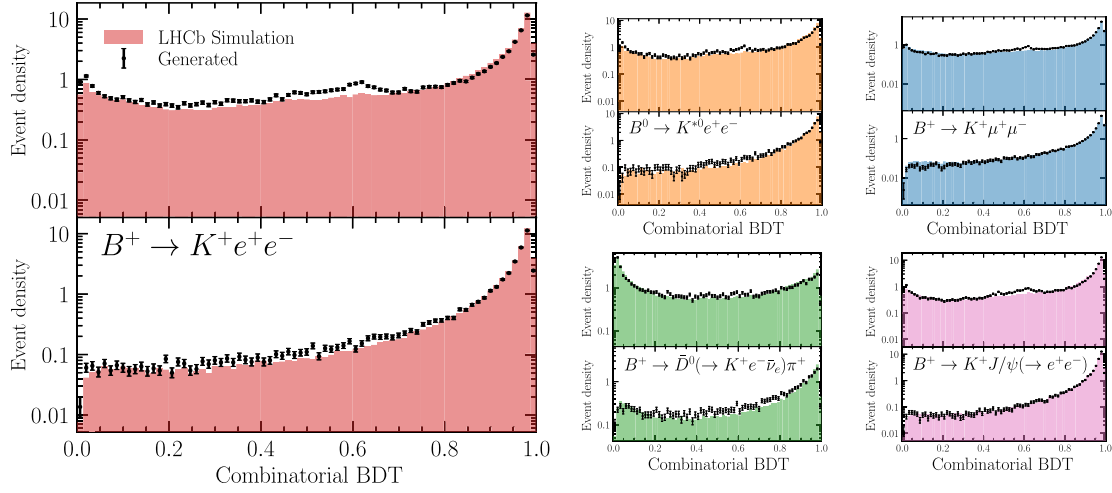


FIG. 16. Combinatorial BDT output of all events (top panels), and events passing cut-based selection (bottom panels) for each example in the case study.

selection as a function of reconstructed B -mass. Similar distributions are produced for other kinematic variables and are available in the Appendix (see Fig. 22). To reiterate, these invariant kinematic variables are not given to the network in the conditions, yet the model correctly captures the dependence of the efficiency as a function of them. Resolution effects also appear well-modeled in each case.

Following the procedure outlined in Ref. [54] describing the measurement of R_K in $B^+ \rightarrow K^+ \ell^+ \ell^-$ transitions, we find that the replacement of the fully simulated samples with ones produced through R_{ex} would lead to a systematic shift of 0.5% in R_K . This shift would constitute a subleading source of systematic uncertainty for this measurement. Further details regarding this study are given in Appendix A. A more stringent validation of the performance of R_{ex} is provided by the measurement of $r_{J/\psi}$, also described in Ref. [54], which probes lepton universality in J/ψ decays from $B^+ \rightarrow K^+ J/\psi$ transitions. Unlike R_K , which makes use of a double ratio to cancel systematic uncertainties related to efficiency determination and to reduce the impact of mismodeling differences between electrons and muons, $r_{J/\psi}$ is based on a single efficiency ratio and is therefore directly sensitive to such mismodeling effects. In this case, using R_{ex} to produce the simulated samples, the shift to $r_{J/\psi}$ is 2%, which is of similar order to existing systematic effects. Updates to come in future releases of R_{ex} , as listed in the conclusions of this paper, will improve on this.

E. Uncertainties

Inaccuracies in the output of R_{ex} may be ascribed to four principal sources:

Convergence-related uncertainties The network's performance is sensitive to instabilities in the training process. These may arise from premature or late stopping—a consequence of the difficulty in defining a convergence

criterion that generalizes well. The model may also become saturated, lacking the capacity to capture the required behavior across phase space; this may manifest in continual trade-offs between different regions. Instabilities may also emerge in localized parts of condition space, latent space, or output space due to the stochastic nature of batch ordering during training. The GAN loss itself is infamously unstable, often requiring significant tuning of hyperparameters and of training schedules to maintain the balance between the discriminator and generator required to achieve convergence. These effects are more likely to manifest in sparsely populated or high-variance regions of phase space. Repeating training or resampling with different random seeds can help in assessing these uncertainties.

Model-related limitations Systematic biases may arise from limitations in the model architecture. For example, it may be the case that the network or graph structures do not sufficiently allow or unknowingly prohibit certain relationships between objects. In complex topologies, skip connections or other architectural components may be required to accurately model correlations between variables. A more likely cause of biases within this category is insufficient conditioning: we are inherently restricted in the set of variables we can use as conditions. For any effects that depend on something not parsed, any behavior will only be captured on average. This could result in biases when the network is queried for specific channels. Studying the performance across a range of decay topologies may help mitigate these issues, though edge cases may remain vulnerable.

Inaccuracies in conditions Even with a perfectly trained network, any discrepancies between the training conditions (as derived from simulation) and those generated for inference will cause systematic bias. Known mismodelings include differences in the kinematics of hadrons originating from the primary vertex and a simplified geometric acceptance. These effects appear to be small, and future

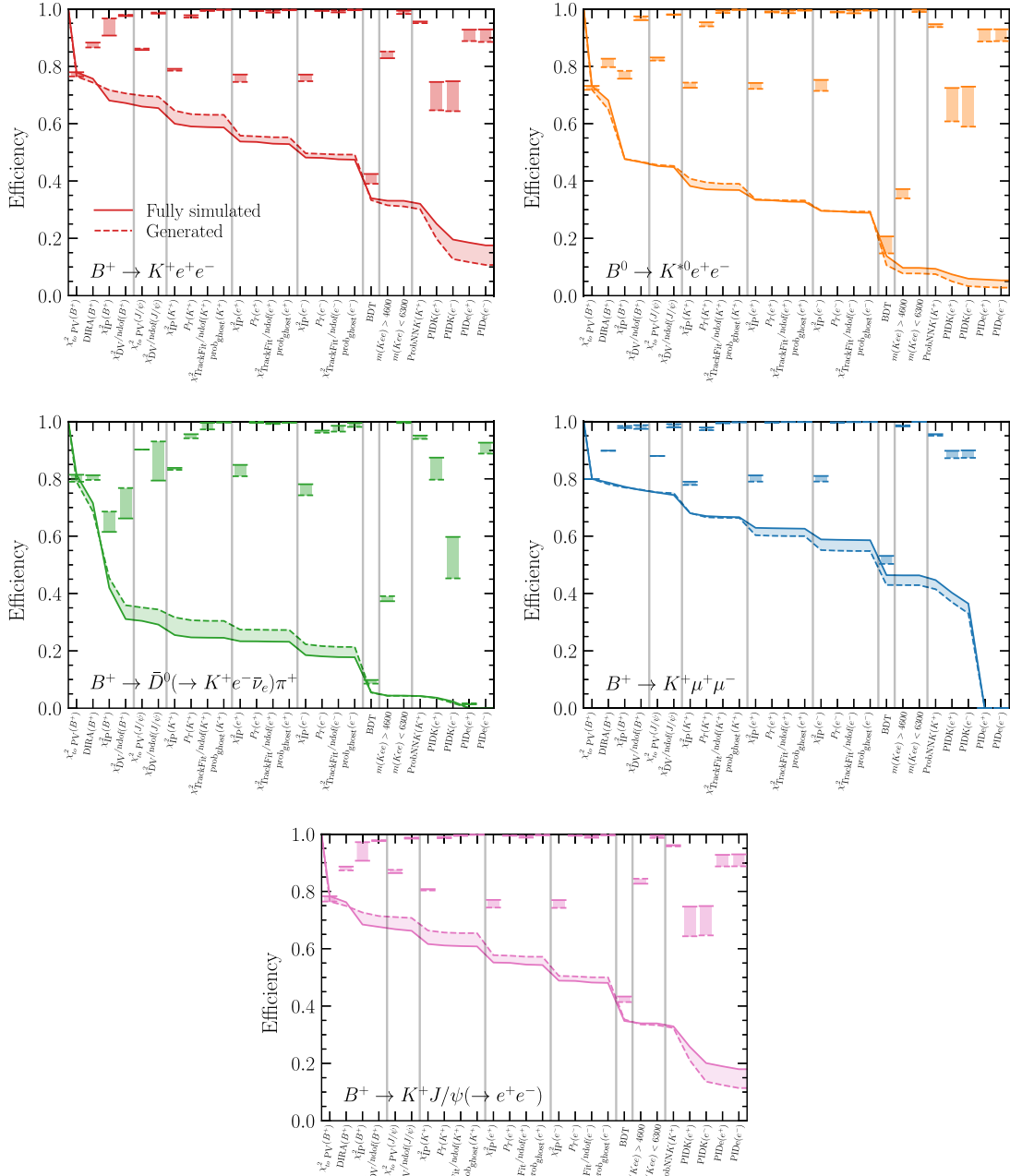


FIG. 17. A comparison of the efficiencies of the applied selection criteria, as computed using full simulation (solid) and Rex (dashed). Individual cut efficiencies are presented alongside the cumulative (running) efficiency.

versions of Rex will aim to address them. Physically motivated perturbations to the conditions can be used to probe the sensitivity of the outputs to such differences. With the simplified geometrical acceptance, the tool is currently unable to account for particles in acceptance that may not have been reconstructed. This will be accounted for in later versions by employing an improved probabilistic geometrical acceptance map.

Simulation inaccuracies The full simulation is known to imperfectly describe data. Analyses at the LHCb routinely apply significant data-driven weighting procedures to

improve the fidelity of simulated distributions. This tool inherits any inaccuracies present in the underlying simulation, forming a source of irreducible uncertainty.

Beyond identifying the sources of uncertainty, we also must consider how best to present them to end users. Given the generalized nature Rex, with outputs in the form of raw variables rather than derived quantities, the most appropriate approach is likely to provide multiple alternative sets of output variables. Users can then propagate each set through their analysis workflows (e.g., selection criteria) and study the resulting changes in efficiency or yield. While this tool

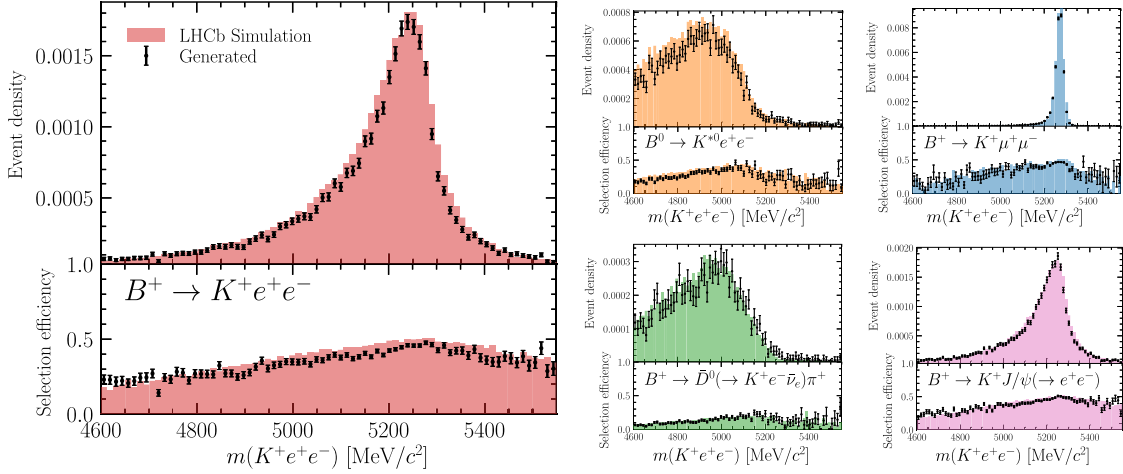


FIG. 18. Reconstructed B -mass distributions (top panels) using the signal hypothesis of $B^+ \rightarrow K^+ e^+ e^-$ for the signal channel and the various backgrounds, and efficiency of combinatorial BDT and cut-based selection excluding PID cuts (bottom panels) as a function of reconstructed mass.

could incorporate some built-in functionality to support that workflow, the diversity of use cases suggests that the final implementation details are best left to the user.

VI. CONCLUSION

This paper presents a set of novel heterogeneous graph-based GAN network architectures designed to emulate the LHCb simulation pipeline in a manner that is as general as possible. The tool is trained in such a way as to be agnostic to any specific final states, to support any physical decay topology and any arbitrary analysis selection criteria. Where possible, the architectures employed have been designed to embed the logic of each interdependent stage of modeling into the structure of the networks themselves.

Using a representative case study, we have demonstrated that the tool is capable of modeling selection efficiencies and invariant mass distributions of a broad spectrum of topologies, including fully reconstructed decays, partially reconstructed decays, final states involving long-lived intermediates, and various types of misidentification with excellent fidelity. Furthermore, we have demonstrated that second-order effects, such as interactions between tracks and interference from the rest of the event, are also correctly included.

This work represents the first iteration of Rex and is available as a user-friendly Python package capable of modeling the sculpting of a decay phase space by arbitrary selection criteria. The tool reads simple configuration files and outputs in the same format as the full LHCb simulation, allowing for seamless integration with existing analysis workflows. Estimates of absolute efficiencies of any selection chain, including tracking, vertexing, impact parameter, and flight distance variables, can be produced across a wide range of decay topologies. Developments to come in future versions of Rex will reduce these discrepancies.

On the inference side, the next steps are to address the limitations mentioned in Sec. III in the modeling of the

detector's geometrical acceptance and the generation of kinematics of heavy-flavor hadrons originating from the PV, as mentioned in Sec. VB this is likely to improve the modeling of the PID variables. Additionally, there are a number of commonly used variables which are not currently supported by the model, most notably trigger decisions and calorimetric variables. Note that extending the list of variables is possible without a change to the architecture. Finally, updated versions of Rex should also include extensions to support protons and higher multiplicity final states. Then, regarding the performance, the next steps are to begin some trials within ongoing analyses and to begin quantifying the agreement with the full simulation and the size of the various sources of uncertainty. Here, improvements in the training stability and convergence criteria may be required; this could be driven by a move away from GANs to diffusion models, to avoid the complexities of adversarial training. Beyond this, it should be possible to implement inference mechanisms to estimate new background types, such as combinatorial background or backgrounds involving particles from the primary vertex, using the existing network architecture. Finally, looking further ahead, it may even be possible to develop methods that enable training Rex directly on data, potentially avoiding some of the accuracy limitations in the full simulation.

The speed-ups achieved by the methods presented in this paper have the potential to fundamentally reshape the simulation workflow at the LHCb. Trained models can be stored, one set for each configuration of data-taking conditions, and used repeatedly for on-demand generation of analysis-ready samples, reducing the existing need for long-term storage of large samples of full detector readout. This workflow would offer significant savings in both compute and storage, and opens the door to a more decentralized workload that could offer more flexibility to analysts.

The architectures and methods presented in this paper can easily be employed to emulate reconstruction processes at

other particle physics experiments, many of which are seeing similar pressures relating to growing demands for simulation.

ACKNOWLEDGMENTS

We are extremely grateful to colleagues from the LHCb Simulation Project, the quality of this tool and the ability to train it at all rely on many years of development of the LHCb software. Additionally, we would like to thank colleagues from across the Data Processing and Analysis (DPA) project who have provided feedback during the development of this tool. Finally, this paper was significantly improved based on comments from Mark Whitehead and Mark Smith. The University of Bristol group acknowledges support from the STFC under Grant No. ST/W000490/1. The UZH group acknowledges support from the Swiss National Science Foundation (SNSF) under Grants No. 209263 and 204238, and from Forschungskredit under the UZH Candoc Grant No. FK-25-086. The Imperial College group acknowledges support from the STFC (United Kingdom) and The Leverhulme Trust (United Kingdom).

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they are owned by a third party and the terms of use prevent public distribution. The data are available from the authors upon reasonable request.

APPENDIX A: REMEASURING R_K WITH *Rex*

We can employ fully simulated tuples and tuples generated by *Rex* to get a handle on how different R_K would be if one were to perform the measurement using only *Rex*.

First, we look at the fit to the reconstructed B -mass, $m(K^+e^+e^-)$, in the electron mode. We generate toy event sets using distributions from selected fully simulated tuples and using yields in accordance with the fits of Ref. [54]. An exponential component is used to represent the combinatorial background. We fit each event set with the same set of fit components as obtained from the full simulation and additionally a set obtained using *Rex*, then compare the extracted signal yield. In each case, the signal yield is obtained via a binned likelihood template fit in which the only parameters not fixed are the yields of the components. An example toy, and the fitted components are presented in the left panel of Fig. 19. In the right panel is a distribution of the difference in the extracted signal yield expressed as a percentage. For the vast majority of cases, this difference is $< 1\%$, and the most likely difference is a $\sim 0.5\%$ shift.

In a measurement of a branching fraction or of a ratio of branching fractions, such as R_K , as well as the measured yield, one also requires integrated efficiencies. These appear as the ϵ quantities, and the yields appear as the N quantities, in the following expression for R_K :

$$R_K = \frac{N_{K\mu\mu}}{N_{Kee}} \cdot \frac{\epsilon_{Kee}}{\epsilon_{K\mu\mu}} \cdot \frac{N_{KJ/\psi(ee)}}{N_{KJ/\psi(\mu\mu)}} \cdot \frac{\epsilon_{KJ/\psi(\mu\mu)}}{\epsilon_{KJ/\psi(ee)}}. \quad (\text{A1})$$

The ratios of the efficiencies as produced by *Rex* and by the full simulation for the channels involved in R_K are presented in Table II. Here, we can see that efficiencies are mismodeled by *Rex*, in its current form. For example, *Rex* underestimates the efficiency of $B^+ \rightarrow K^+e^+e^-$ by $\sim 7\%$. However, it is important to stress that measurements at the LHCb are almost always conducted with respect to control channels, which, by design, have final states and decay topologies as similar as possible to the channels of interest, such that in the measurement, mismodelings cancel. Indeed, it is the case too that in *Rex* the mismodelings of control channels, in this case

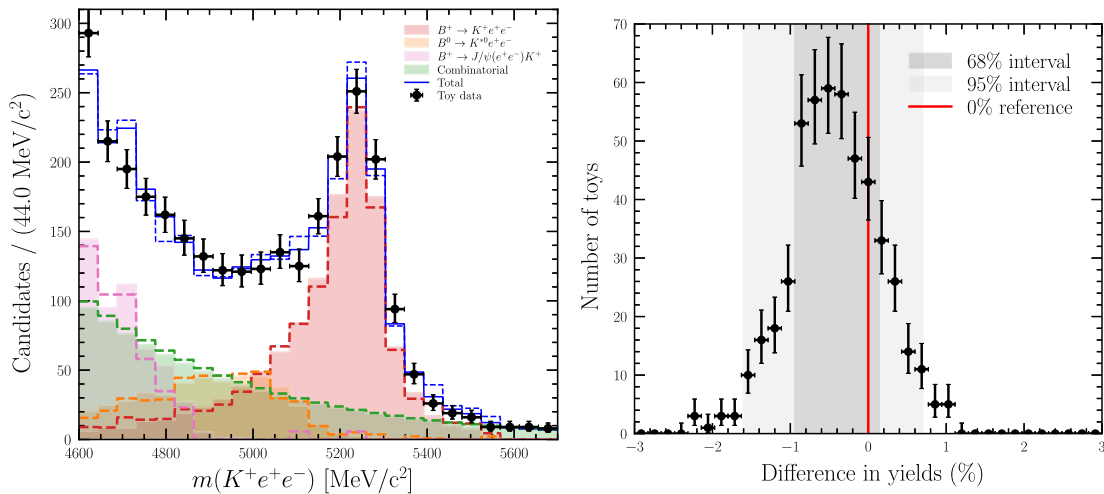


FIG. 19. An example pseudodataset for the measurement of $B^+ \rightarrow K^+e^+e^-$ and fitted components, with those from *Rex* being dashed (left), and the observed differences obtained from fitting pseudodatasets with *Rex* relative to the full simulation (right).

TABLE II. Ratios of efficiencies for channels involved in the extraction of R_K .

	$B^+ \rightarrow K^+ e^+ e^-$	$B^+ \rightarrow J/\psi(e^+ e^-) K^+$	$B^+ \rightarrow K^+ \mu^+ \mu^-$	$B^+ \rightarrow J/\psi(\mu^+ \mu^-) K^+$
$\epsilon_{\text{REX}}/\epsilon_{\text{LHCb}}$	0.93	0.92	0.95	0.94

the J/ψ channels, are similar to their corresponding signal channels. Therefore, in the measurement scenario of R_K , effects from mismodelings in R_{ex} are suppressed to a $< 0.1\%$ effect. Leaving the full shift expected if one were to measure R_K with R_{ex} to be $\sim 0.5\%$, with this being dominated by the difference in the extraction of the signal yield as observed in Fig. 19. Another related quantity is the single ratio $r_{J/\psi}$, which is employed as a validation channel in R_K measurements, it is defined as

$$r_{J/\psi} = \frac{N_{KJ/\psi(\mu\mu)}}{N_{KJ/\psi(ee)}} \cdot \frac{\epsilon_{KJ/\psi(ee)}}{\epsilon_{KJ/\psi(\mu\mu)}}. \quad (\text{A2})$$

Studying any systematic shift here is a much more stringent test of the performance of R_{ex} as cancellations within the

quantity $r_{J/\psi}$ are not as powerful as in R_K , and the ratio remains vulnerable to differences in the extent to which the detector response to electrons and to muons is mismodeled. In this case, $r_{J/\psi}$ shifts by 2% when using simulated samples from R_{ex} , which is commensurate with the systematic uncertainty on $r_{J/\psi}$ reported in Ref. [55]. The shifts presented here are computed using tuples before any data-simulation correction weights that are applied to the full simulation tuples in any analysis at the LHCb. Applying such weights would more closely align R_{ex} and the full simulation and reduce systematic shifts reported here.

APPENDIX B: ADDITIONAL PID PLOTS

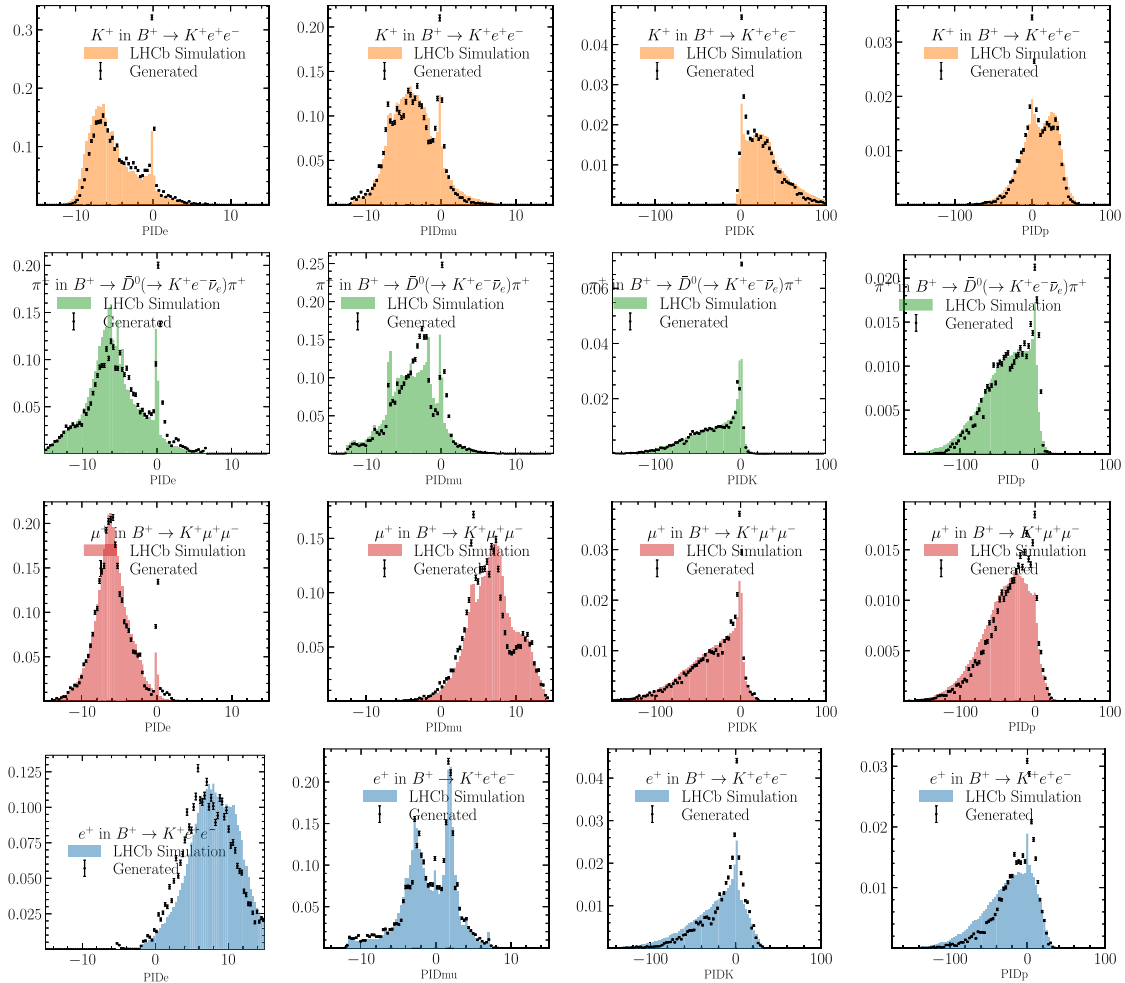
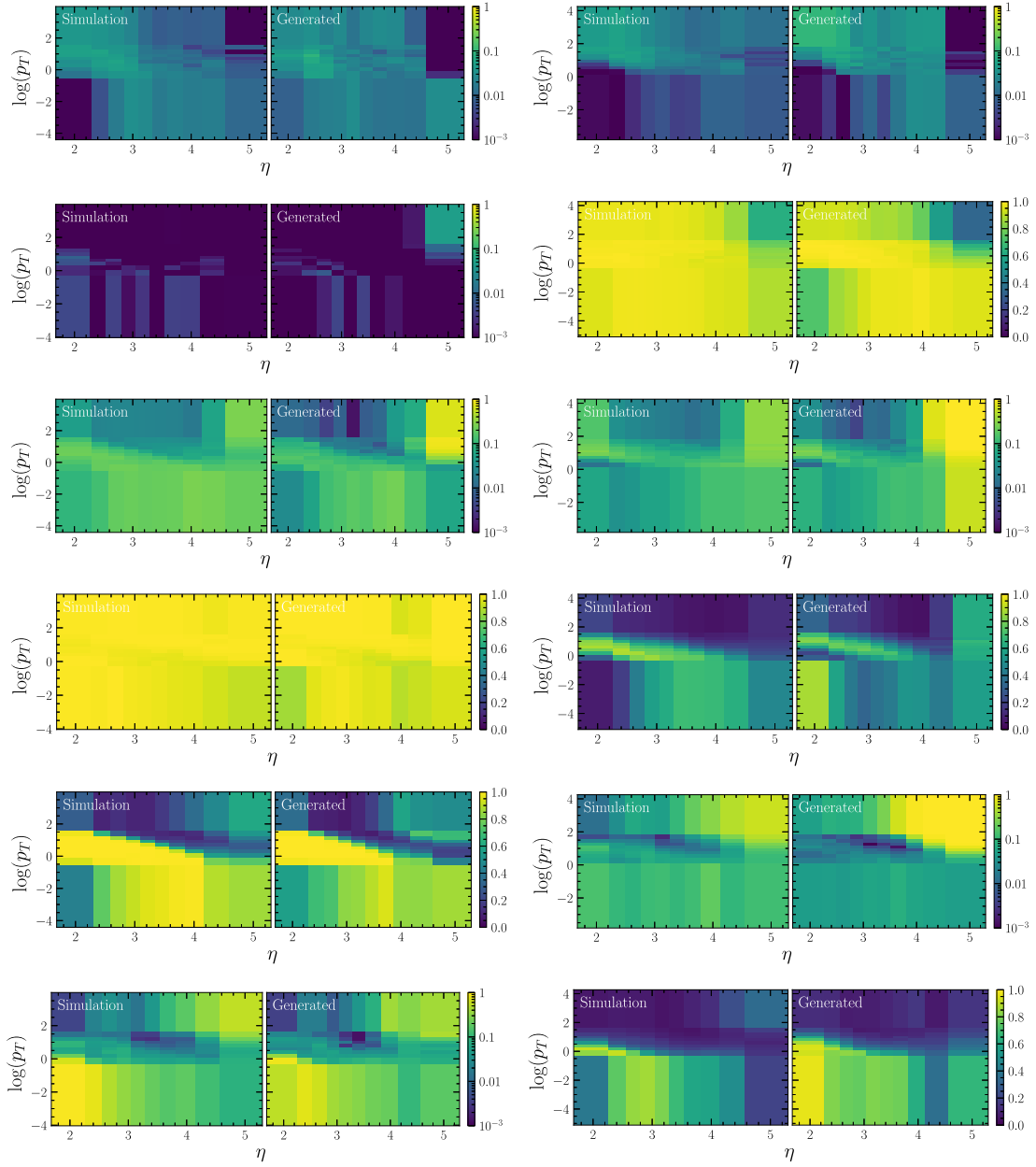


FIG. 20. Example distributions of generated PID variables.

FIG. 21. Efficiency maps [in $\log(p_T)$ and η] for various PID cuts.

APPENDIX C: ADDITIONAL VERTEXING PLOTS

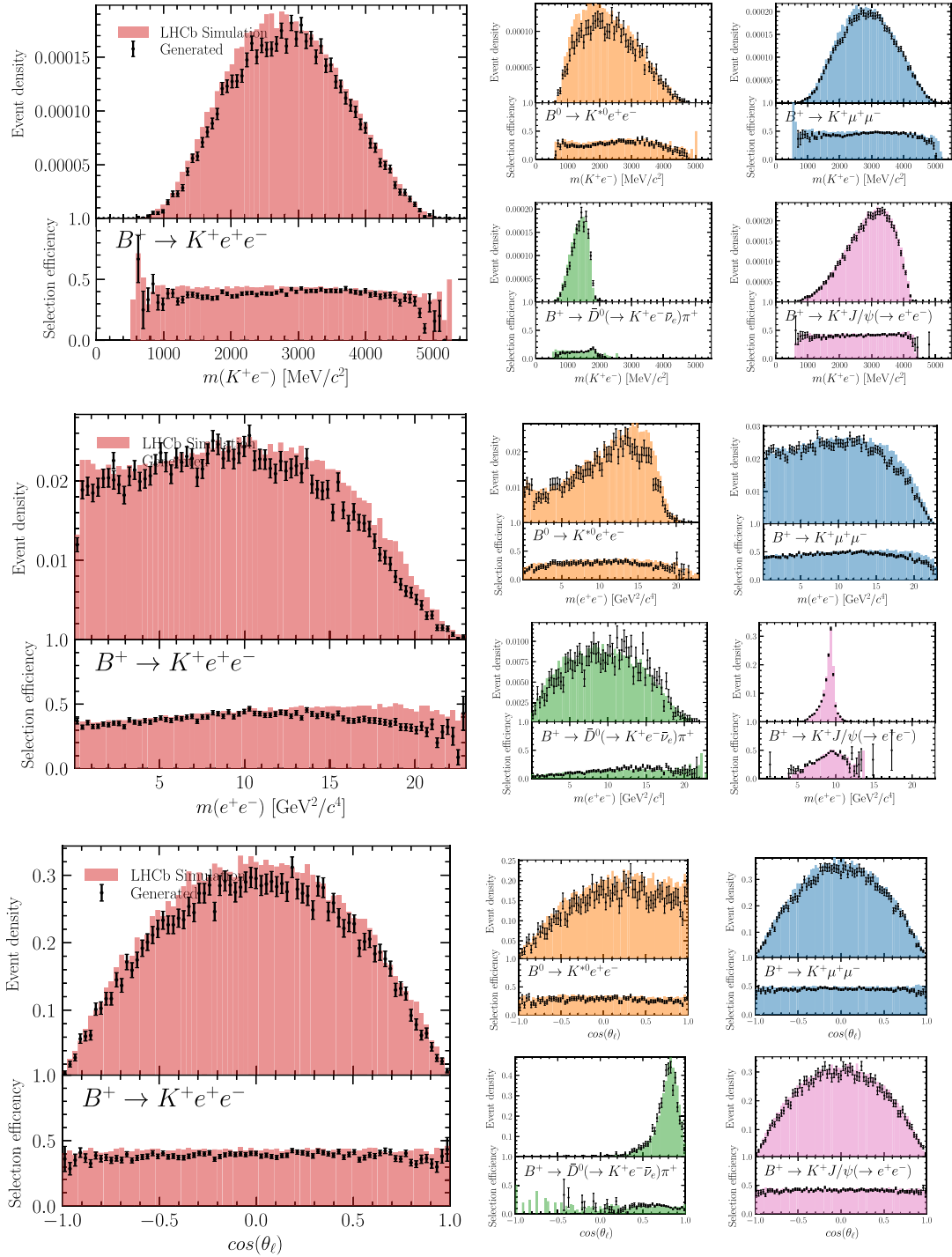


FIG. 22. Example response to combinatorial BDT and stripping as a function of various reconstructed kinematic variables.

- [1] L. Evans and P. Bryant, eds., LHC Machine, *J. Instrum.* **3**, S08001 (2008).
- [2] LHCb Collaboration, Framework TDR for the LHCb Upgrade II: Opportunities in flavour physics, and beyond, in the HL-LHC era, Technical Report No. CERN-LHCC-2021-012, LHCb-TDR-023, CERN, Geneva, 2021, <https://cds.cern.ch/record/2776420>.
- [3] C. Bozzi, LHCb computing resource usage in 2023, Report No. LHCb-PUB-2024-003, CERN-LHCb-PUB-2024-003, CERN, Geneva, 2024, <https://cds.cern.ch/record/2888940>.
- [4] Atlas Collaboration, Atlas software and computing HL-LHC roadmap, Technical Report No. CERN-LHCC-2022-005, LHCC-G-182, CERN, Geneva, 2022, <https://cds.cern.ch/record/2802918>.
- [5] C.O. Software and Computing, CMS Phase-2 computing model: Update document, Technical Report, CERN, Geneva, 2022.
- [6] LHCb Collaboration, Measurement of the branching fraction ratios $R(D^+)$ and $R(D^{*+})$ using muonic τ decays, *Phys. Rev. Lett.* **134**, 061801 (2025).
- [7] LHCb Collaboration, Measurement of the ratios of branching fractions $\mathcal{R}(D^*)$ and $\mathcal{R}(D^0)$, *Phys. Rev. Lett.* **131**, 111802 (2023).
- [8] LHCb Collaboration, Test of lepton flavor universality using $B^0 \rightarrow D^{*-}\tau + \nu\tau$ decays with hadronic τ channels, *Phys. Rev. D* **108**, 012018 (2023).
- [9] LHCb Collaboration, Branching fraction measurements of the rare $B_s^0 \rightarrow \phi\mu^+\mu^-$ and $B_s^0 \rightarrow f_2'(1525)\mu^+\mu^-$ decays, *Phys. Rev. Lett.* **127**, 151801 (2021).
- [10] HEP Software Foundation Collaboration, A roadmap for HEP software and computing R&D for the 2020s, *Comput. Software Big Sci.* **3**, 7 (2019).
- [11] D. Müller, M. Clemencic, G. Corti, and M. Gersabeck, ReDecay: A novel approach to speed up the simulation at LHCb, *Eur. Phys. J. C* **78**, 1009 (2018).
- [12] S. Abdullin, P. Azzi, F. Beaudette, P. Janot, A. Perrotta (on behalf of the CMS Collaboration), The fast simulation of the CMS detector at LHC, *J. Phys. Conf. Ser.* **331**, 032049 (2011).
- [13] DELPHES 3 Collaboration, DELPHES 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [14] G. A. Cowan, D. C. Craik, and M. D. Needham, RapidSim: An application for the fast simulation of heavy-quark hadron decays, *Comput. Phys. Commun.* **214**, 239 (2017).
- [15] LHCb Collaboration, Fast data-driven simulation of Cherenkov detectors using generative adversarial networks, *J. Phys. Conf. Ser.* **1525**, 012097 (2020).
- [16] A. Rogachev and F. Ratnikov, GAN with an auxiliary regressor for the fast simulation of the electromagnetic calorimeter response, *J. Phys. Conf. Ser.* **2438**, 012086 (2023).
- [17] M. Barbetti, Lamarr: LHCb ultra-fast simulation based on machine learning models deployed within Gauss, in *21th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI meets Reality* (2023).
- [18] CMS Collaboration, Reweighting simulated events using machine-learning techniques in the CMS experiment, *Eur. Phys. J. C* **85**, 495 (2025).
- [19] CMS Collaboration, FlashSim: Accelerating HEP simulation with an end-to-end Machine Learning framework, *EPJ Web Conf.* **295**, 09020 (2024).
- [20] ATLAS Collaboration, AtlFast3: The next generation of fast simulation in ATLAS, *Comput. Software Big Sci.* **6**, 7 (2022).
- [21] Atlas Collaboration, Deep generative models for fast photon shower simulation in atlas, *Comput. Software Big Sci.* **8**, 7 (2024).
- [22] LHCb Collaboration, Observation of the doubly-charmed-baryon decay $\Xi_{cc}^{++} \rightarrow \Xi_c^0\pi^+\pi^+$, *J. High Energy Phys.* **10** (2025) 136.
- [23] LHCb Collaboration, Measurement of the CKM angle γ in $B^\pm \rightarrow DK^*(892)^\pm$ decays, *J. High Energy Phys.* **02** (2025) 113.
- [24] LHCb Collaboration, Measurement of CP violation observables in $D^+ \rightarrow K^-K + \pi^+$ decays, *Phys. Rev. Lett.* **133**, 251801 (2024).
- [25] LHCb Collaboration, Search for the lepton-flavor violating decay $B_s^0 \rightarrow \phi\mu \pm \tau \mp$, *Phys. Rev. D* **110**, 072014 (2024).
- [26] LHCb Collaboration, First observation of the $\Lambda_b^0 \rightarrow D^+D^-\Lambda$ decay, *J. High Energy Phys.* **07** (2024) 140.
- [27] LHCb Collaboration, Study of CP violation in $B^0 \rightarrow DK(892)^0$ decays with $D \rightarrow K\pi(\pi\pi), \pi\pi(\pi\pi)$, and KK final states, *J. High Energy Phys.* **05** (2024) 025.
- [28] T. Sjöstrand, S. Mrenna, and P. Skands, Pythia 6.4 physics and manual, *J. High Energy Phys.* **05** (2006) 026.
- [29] I. Belyaev *et al.*, Handling of the generation of primary events in Gauss, the LHCb simulation framework, *J. Phys. Conf. Ser.* **331**, 032047 (2011).
- [30] D. J. Lange, The EvtGen particle decay simulation package, *Nucl. Instrum. Methods Phys. Res., Sect. A* **462**, 152 (2001).
- [31] P. Golonka and Z. Was, Photos Monte Carlo: A precision tool for QED corrections in Z and W decays, *Eur. Phys. J. C* **45**, 97 (2006).
- [32] N. Davidson, T. Przedzinski, and Z. Was, Photos interface in C++: Technical and physics documentation, *Comput. Phys. Commun.* **199**, 86 (2016).
- [33] Geant4 Collaboration, Geant4 developments and applications, *IEEE Trans. Nucl. Sci.* **53**, 270 (2006).
- [34] M. Clemencic, G. Corti, S. Easo, C. R. Jones, S. Miglieranzi, M. Pappagallo, and P. Robbe, The LHCb simulation application, Gauss: Design, evolution and experience, *J. Phys. Conf. Ser.* **331**, 032023 (2011).
- [35] S. Tolk, J. Albrecht, F. Dettori, and A. Pellegrino, Data driven trigger efficiency determination at LHCb.
- [36] M. De Cian, S. Farry, P. Seyfert, and S. Stahl, Fast neural-net based fake track rejection in the LHCb reconstruction.
- [37] LHCb Collaboration, Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC, *J. Instrum.* **14**, P04013 (2019).
- [38] P. Billoir, M. De Cian, P. A. Günther, and S. Stemmle, A parametrized Kalman filter for fast track fitting at LHCb, *Comput. Phys. Commun.* **265**, 108026 (2021).
- [39] LHCb Collaboration, LHCb detector performance, *Int. J. Mod. Phys. A* **30**, 1530022 (2015).
- [40] J. Gavranovič and B. P. Kerševan, Systematic evaluation of generative machine learning capability to simulate distri-

- butions of observables at the large hadron collider, *Eur. Phys. J. C* **84**, 911 (2024).
- [41] M. Cacciari, M. Greco, and P. Nason, The p_T spectrum in heavy-flavour hadroproduction, *J. High Energy Phys.* **05** (1998) 007.
- [42] M. Cacciari, S. Frixione, N. Houdeau, M. L. Mangano, P. Nason, and G. Ridolfi, Theoretical predictions for charm and bottom production at the LHC, *J. High Energy Phys.* **10** (2012) 137.
- [43] M. Fey and J. E. Lenssen, Fast graph representation learning with PyTorch geometric, *arXiv:1903.02428*.
- [44] LHCb Collaboration, Adopting new technologies in the LHCb Gauss simulation framework, *EPJ Web Conf.* **214**, 02004 (2019).
- [45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, *Commun. ACM* **63**, 139 (2020).
- [46] X. Yang, M. Yan, S. Pan, X. Ye, and D. Fan, Simple and efficient heterogeneous graph neural network, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37 (2023), pp. 10816–10824.
- [47] W. Sutcliffe, M. Calvi, S. Capelli, J. Eschle, J. García Pardiñas, A. Mathad *et al.*, Scalable multi-task learning for particle collision event reconstruction with heterogeneous graph neural networks, *Mach. Learn. Sci. Tech.* **6**, 045060 (2025).
- [48] A. Huang, X. Ju, J. Lyons, D. Murnane, M. Pettee, and L. Reed, Heterogeneous graph neural network for identifying hadronically decayed tau leptons at the High Luminosity LHC, *J. Instrum.* **18**, P07001 (2023).
- [49] S. Caillou, C. Collard, C. Rougier, J. Stark, H. Torres, and A. Vallier, Novel fully-heterogeneous GNN designs for track reconstruction at the HL-LHC, *EPJ Web Conf.* **295**, 09028 (2024).
- [50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, *arXiv:1710.10903*.
- [51] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, Improved techniques for training GANs, *arXiv:1606.03498*.
- [52] L. Anderlini, A. Contu, C. R. Jones, S. S. Malde, D. Muller, S. Ogilvy *et al.*, The PIDCalib package, 2016.
- [53] LHCb Collaboration, Measurement of the branching fraction ratio R_K at large dilepton invariant mass, *J. High Energy Phys.* **07** (2025) 198.
- [54] LHCb Collaboration, Test of lepton universality in $b \rightarrow s \ell^+ \ell^-$ decays, *Phys. Rev. Lett.* **131**, 051803 (2023).
- [55] LHCb Collaboration, Test of lepton universality in beauty-quark decays, *Nat. Phys.* **18**, 277 (2022).