nature portfolio

Peer Review File

Quantum anomaly detection in the latent space of proton collision events at the LHC

This manuscript has been previously reviewed at another *Nature Portfolio* journal. This document only contains reviewer comments and rebuttal letters for versions considered at *Communications Physics*.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This is an interesting paper that I am not sure is of sufficient impact to warrant publication in Nature Communications Physics, but perhaps I can be persuaded otherwise. The authors have an extremely intriguing result about which they make a sensational claim that, if taken seriously, will invite a tremendous amount of scrutiny. If they can really defend their claim, this paper should clearly be accepted. If they choose to back off and moderate their claims, the paper may not be of sufficiently high impact for this journal.

The paper is nominally about anomaly detection at the LHC using quantum classifiers to probe a compressed version of a "realistic" dataset. We have to make a concession about the plausibility of this exercise - a real LHC experiment would never compress the data and throw away information in this way unless it were part of the trigger system with a hardware accelerated auto encoder and classifier system such that analysis of the event was faster than some critical threshold. Furthermore, the layout of the data, especially post-compression, means there is nothing "quantum" about the data or any reason to expect a classifier leveraging quantum features should perform better than a purely classical approach. We are just mapping classical data into a high-dimensional space and hoping to get lucky with a mapping that produces an especially good hyper-plane (hyper-sphere in this work, really). In many senses studies like this would benefit from using much simpler benchmark datasets where it is easier to compare the performance of the classical benchmark. Because they data analysis task they've chosen is esoteric and unfamiliar to the quantum computing and machine learning communities, it is hard to form an intuition about classifier performance. By using a "particle physics" dataset, the authors connect their quantum machine learning (QML) work to the domain science goals at the LHC, but there is no real "physics" work here and in general attempting to make this sort of connection just makes this work harder to evaluate. Is the absolute value of the performance metric good here? It is hard to say. I of course trust the authors to have chosen useful metrics (in terms of true and false positive rates), appropriate for their domain science and am fully willing to believe all the details about the "new physics" scenarios are fine.

Once we have accepted this difficulty in evaluating the machine learning (ML) task at hand, the paper compares the performance of a family of "quantum" anomaly detection (AD) algorithms to a set of purely classical benchmarks. There are three algorithms - an unsupervised kernel machine, a quantum K-means clustering algorithm, and a quantum K-median clustering algorithm. All of the excitement will be focused on the unsupervised kernel machine where the authors make the very strong and bold claim of a "quantum performance advantage," and so that is what this review will focus on.

The other algorithms see performance similar to the purely classical benchmark. While the paper itself is a bit confusing and often brief where it should have been more details (for example, in explaining the mathematics of their classifiers), and long-winded where brevity would have made sense (for example, while discussing things like “jet invariant mass cuts” while constructing the dataset), the authors have done a truly superb job in releasing a well-documented software package that I was able to read in order to figure out what they were actually doing. This reviewer hopes the entire community can live up to the example they are setting here for reproducibility. The clustering algorithms are technically sound and I don’t have much further comment on those portions of the paper, honestly mostly because the results are as expected - essentially comparable behavior between the “quantum” algorithms on the purely classical benchmarks.

I believe more work is necessary to really support the main, sensational claim. First, I refer to the algorithm as “quantum” in quotes because it is not really a quantum algorithm. It is a fully classical simulation of a quantum algorithm in a regime where classical simulation is extremely easy to perform. It is a quantum-inspired algorithm. To actually claim this algorithm is quantum, I believe the authors need to do two things. First, they should rigorously establish the computational hardness of their kernel. The authors make no claims about whether their kernel is classically hard. I am not an expert on whether it is or isn’t - it passes the obvious simple test of not being fully composed of Clifford gates, but if they want to make the claim their algorithm offers a real “quantum advantage,” we need to unambiguously establish the circuit is classically hard. Next, we need to establish that the algorithm’s performance scales with qubit count - or better would be quantum volume. We need to establish that there is strong reason to believe the kernel is not one that may be implemented easily on classical resources. If it is classically easy, it will essentially always be expedient to do so and the claim needs to be modified that the authors have identified a quantum-inspired, but classical, kernel. The authors do make a small piece of the analysis about scaling qubit count and I commend them for it - but it isn’t enough to cover these concerns.

To be clear, I am not asking for a beyond-classical simulation of their algorithm on classical resources, or demanding to see better performance on hardware using 50+ qubits or something like that. But I do think that to make the claim they are making they need to provide a theoretical proof of the difficulty of computing their kernel, and they need to demonstrate a performance trend that increases with quantum volume all the way up to the largest reasonable size that may simulated on classical resources. You can argue about “reasonable,” but it should be well beyond laptop-scale even if it doesn’t test the limits of the classical simulation of quantum circuits. The important point is a trend supporting value from more quantum resources being available.

Note - I do understand the authors made an evaluation of whether the kernel was feasible on NISQ hardware, but their central, controversial, claim is separate from that. It would be somewhat

interesting to understand what they did in terms of error mitigation to get a good result on hardware, but that isn't the main point of the paper.

I think it would also be interesting to compare the performance of the exact quantum simulation of their algorithm to a "fuzzy" method of circuit simulation like a tensor network. The paper shows some evidence that increasing entanglement in the simulation helps, but only up to a point after which it actually begins to penalize performance of the kernel. This suggests to me a tensor network simulation of their circuit may offer all the performance with scaling costs that keep the overall accounting in favor of a classical ML algorithm for problems of arbitrary size. I am not requiring the authors to do this work here - that is an unreasonable scope increase. However, they should moderate their claims until this sort of study is performed.

We also need to understand a lot more about the classical benchmark. We don't know much about how many approaches they tried or how much optimization was done on the purely classical side - we just know there is at least one purely classical algorithm their quantum-inspired algorithm significantly outperforms. The authors have chosen a good classical algorithm, but we need more insurance against potential claims of cherry-picking. Because they have chosen an esoteric task, it is hard to have an intuition about performance. At best we can say we have a state-of-the-art (SOTA) measurement using a quantum-inspired algorithm and invite the community to try to best it. It isn't clear there is motivation to do so, but the claim will hinge on beating all comers. If the authors had chosen a well-studied dataset, their new SOTA would be meaningful, but as it is we have to hope they are able to build a community willing to compete on this task. The authors should, of course, be commended again for publishing the dataset they used and their code - this makes such a competition possible in the future.

Note that I am not asking the authors to beat every classical algorithm in existence before attempting to publish - that isn't possible. But the claim needs to be adjusted - they found a quantum-inspired algorithm that beats at least one other classical algorithm, and they should explain in some depth why that benchmark was chosen and defend their benchmark as at least a plausible competitor for the best available option.

The authors face one other problem in the defense of this work, and it is difficulty they alluded to in the paper - namely, this is a hybrid algorithm at best with a substantial amount of the "work" done by the auto-encoder. This is relevant because their circuit ansatz ties the number of qubits to the feature size of the auto-encoder. This makes disentangling, if you will, the qubit scaling effects from the feature richness of the latent space as the dimension changes. This is balanced somewhat by the classical benchmark, but it remains an onerous detail to wrestle with because the classical benchmark needs to be re-optimized as you go along.

A smaller detail, but one thing that is unclear is how the authors deal with computing expectation values for the kernel matrix. For the classical simulation, was the exact value used? For real hardware this is never possible. Of course you can make the statistical uncertainty arbitrarily small in the limit of infinite shots, but shots are not free and the price of growing a kernel matrix grows like the dataset size squared, so it is not practical to assume for a very long time that we will have the ability to take an “arbitrary” number of shots for every element of a large kernel matrix. Furthermore, since readout rates are a few kHz, collecting shots adds some significant burden to the overall cost of running the algorithm if you really need your statistical error on the kernel matrix elements to be below a tight threshold. And if you don’t need the statistical error below a tight threshold, their “quantum” algorithm will likely always be outperformed by a tensor-network simulation of their circuit. What is the required statistical uncertainty on the elements of the kernel matrix to achieve the observed performance?

One other note - the authors call their kernel “novel” but it is very similar to the kernel in https://www.nature.com/articles/s41534-021-00498-9 from 2 years ago. The major difference is the repetition of the ansatz - which is a big difference to be sure! - but the structure of the kernels is otherwise quite similar (aside from the use of a CNOT for an entangling gate as opposed to sqrt(iSWAP) - a choice owing to the native 2-qubit entangling option on IBM vs Google hardware).

Response to Referee

We thank Referee for their exceptionally thorough and detailed review. Overall, the Referee has found our results to be of interest, and we are pleased to note that all of the Referee's comments have contributed to the enhancement of our substantially revised manuscript.

The feedback from the Referee has helped us in providing a clearer context for our contributions within the existing literature. Additionally, the Referee has identified a limitation in our previous results, specifically the necessity to extend our investigation into the quantum resources utilised by the quantum kernel model. While we have implemented numerous changes in the revised manuscript, we would especially like to highlight the new Fig. 4 and the corresponding discussion regarding the quantum resources utilised by the quantum kernel model. We have significantly expanded our numerical experiments, observing a trend of enhanced performance in the quantum kernel machine corresponding to increased allocation of quantum resources, namely, the number of qubits and circuit depth. We are confident that the revised manuscript achieves enhanced clarity and effectively addresses the concerns raised by the Referee, meeting the high standards for publication in Nature Communications Physics.

In the following, we proceed to address the specific comments of the Referee individually and describe the corresponding modifications made to the manuscript. The Referee's comments will be presented in italic, followed by our responses in plain text.

Replies and comments to Referee

R: *This is an interesting paper that I am not sure is of sufficient impact to warrant publication in Nature Communications Physics, but perhaps I can be persuaded otherwise. The authors have an extremely intriguing result about which they make a sensational claim that, if taken seriously, will invite a tremendous amount of scrutiny. If they can really defend their claim, this paper should clearly be accepted. If they choose to back off and moderate their claims, the paper may not be of sufficiently high impact for this journal.*

A: We thank the reviewer for carefully reading our manuscript. We appreciate the Referee's guidance, which has allowed us to enhance our previous findings and underscore the connections between our work and relevant prior studies. Through the Referee's constructive feedback, we have enhanced the clarity and precision of our manuscript, aligning it more closely with the objectives and primary outcomes of our research. We are confident that these revisions have significantly improved the accessibility and presentation of our results for the interdisciplinary readership of this journal.

R: *The paper is nominally about anomaly detection at the LHC using quantum classifiers to probe a compressed version of a "realistic" dataset. We have to make a*

concession about the plausibility of this exercise - a real LHC experiment would never compress the data and throw away information in this way unless it were part of the trigger system with a hardware accelerated auto encoder and classifier system such that analysis of the event was faster than some critical threshold.

A: We acknowledge the Referee's concern on the plausibility of our approach in a real LHC setting. To provide more context to our proposed methodology and enhance clarity, we have added a paragraph that provides further information and references to other anomaly detection studies at the LHC that employ models which involve data compression. The ATLAS and CMS experiments at the LHC frequently compress data as part of their searches for new-physics signatures [1–8] and have already produced such anomaly detection studies [9, 10]. Typically, this is accomplished via autoencoder models, where one uses the difference between the encoder and decoder networks of the model as a metric to identify anomalies. In our work, we apply instead the quantum anomaly detection algorithms on the latent space produced by the encoder network of our autoencoder model (cf. Fig. 1) to identify anomalies. As highlighted by the referee, models that involve data compression have been also proposed for the real-time event selection system (trigger) of the LHC experiments [11, 12].

Furthermore, the CMS experiment is currently evaluating algorithms and dedicated hardware which directly process and store collision data in a compressed format to tackle the immense requirements regarding the data volume which will drastically increase after the High Luminosity upgrade of the LHC [13].

R: *Furthermore, the layout of the data, especially post-compression, means there is nothing "quantum" about the data or any reason to expect a classifier leveraging quantum features should perform better than a purely classical approach. We are just mapping classical data into a high-dimensional space and hoping to get lucky with a mapping that produces an especially good hyper-plane (hyper-sphere in this work, really).*

A: We acknowledge the Referee's concern regarding the "quantumness" of HEP datasets. There is a fundamental difference between conventional benchmark datasets, e.g., the widely used MNIST handwritten digits dataset, and HEP data: HEP data is generated by a fundamentally quantum process. As the Referee accurately states, data recorded from particle collisions is initially processed, frequently compressed, and subsequently stored classically. Such data processing methodology is used in our work and in all HEP analyses at the LHC. However, even after classical processing, quantum effects stemming from the initial quantum mechanical process –particle interaction– are still present. Specifically, quantum effects in HEP can be observed through entanglement between particles produced in proton collisions [14–17], spin correlation measurements [18], and violation of Bell inequalities [19–21].

To the best of our knowledge, our study is the first to provide numerical evidence showing that the amount entanglement and the number of qubits play a crucial role

to the quantum model's performance. Specifically, increasing the quantum resources utilised by the model significantly enhances its ability to differentiate between new-physics anomalies and the Standard Model in the high-dimensional Hilbert space. Thanks to the Referee's comments, discussed further below, we have substantially extended our analysis in terms of the depth and the number of qubits of the studied circuits (cf. new Figure 4).

Furthermore, the Referee's remark points towards an open question in the field of QML applications on classical datasets. Namely, the existence of a universal metric of "quantumness" that measures the suitability of a classical dataset for QML is yet unknown. We believe that the completeness and clarity of our revised manuscript has been improved by providing more information.

R: *In many senses studies like this would benefit from using much simpler benchmark datasets where it is easier to compare the performance of the classical benchmark. Because the data analysis task they've chosen is esoteric and unfamiliar to the quantum computing and machine learning communities, it is hard to form an intuition about classifier performance. By using a "particle physics" dataset, the authors connect their quantum machine learning (QML) work to the domain science goals at the LHC, but there is no real "physics" work here and in general attempting to make this sort of connection just makes this work harder to evaluate.*

A: We thank the Referee for their comment and for encouraging us to provide more context for our dataset and underlying motivation. Exploring datasets and applications for which quantum models could serve as natural candidates is a goal of paramount importance for the QML community; this is why we provide our dataset and developed code publicly. Recent studies stress the importance and need of investigating datasets that are generated by fundamentally quantum processes [22–25]. Furthermore, these studies also discourage the use of simple benchmark ML datasets, in which classical models are known to excel saturating the dataset by reaching frequently perfect accuracy. The authors of Refs. [24, 26] demonstrate that entanglement in the QML models worsens the performance on these classical benchmark datasets, implying that no "quantumness" is needed to perform well in such datasets. These results are in stark contrast to our findings, in which entanglement plays a crucial role for the performance of our quantum model; as seen in the updated Fig. 4. For these reasons, we refrain from using standardised benchmark datasets. We strongly believe that our work is both timely and offers important perspectives as a physics methodology utilising quantum machine learning to uncover new physics at the LHC, providing valuable insights not only to the QML and HEP communities but also extending beyond. We agree with the Referee that it is important to contextualise our work within the existing literature and enhance clarity for the broad readership of this journal. In the revised manuscript, we highlight the above and include the relevant references.

R: *Is the absolute value of the performance metric good here? It is hard to say. I of course trust the authors to have chosen useful metrics (in terms of true and false positive rates), appropriate for their domain science and am fully willing to believe all the details about the "new physics" scenarios are fine.*

A: The absolute values that we obtain for the area under the ROC curves (AUC), e.g., at the order of 0.9, presented in Fig. 3, as well as the values for the background rejection rates at the order of $10^1 - 10^2$, presented in Table 1, are comparable to other state-of-the-art results in anomaly detection studies at the LHC [3, 5–7, 11]. Of course, the exact values of these metrics depend on the considered signal (anomalous) dataset.

R: *The other algorithms see performance similar to the purely classical benchmark. While the paper itself is a bit confusing and often brief where it should have been more details (for example, in explaining the mathematics of their classifiers), and long-winded where brevity would have made sense (for example, while discussing things like "jet invariant mass cuts" while constructing the dataset), the authors have done a truly superb job in releasing a well-documented software package that I was able to read in order to figure out what they were actually doing. This reviewer hopes the entire community can live up to the example they are setting here for reproducibility. The clustering algorithms are technically sound and I don't have much further comment on those portions of the paper, honestly mostly because the results are as expected - essentially comparable behavior between the "quantum" algorithms on the purely classical benchmarks.*

A: We thank the Referee for acknowledging our efforts and the importance of providing the software we developed publicly to the community. We strongly believe in open science, reproducibility, and the practical significance of open source code.
Furthermore, we provide the mathematical definitions of the different models, information on how they are trained, and references for further reading in the Methods section of the manuscript. Following the Referee's comments, we added more details for the QK-medians model, and moved parts of the data preprocessing information related to the jet cuts from the main text to the Methods section. Our aim is to respect the formatting and length requirements of this journal. We are open to relocating information from the Methods section to the main text or providing additional details, either in the Methods section or the main text, if deemed appropriate by the Referee or the Editor.

R: *I believe more work is necessary to really support the main, sensational claim. First, I refer to the algorithm as "quantum" in quotes because it is not really a quantum algorithm. It is a fully classical simulation of a quantum algorithm in a regime where*

classical simulation is extremely easy to perform. It is a quantum-inspired algorithm.

A: We understand the Referee's sentiment regarding the ability to simulate and test our models in a low qubit number regime. Directly designing, investigating, optimising, and executing quantum algorithms on real hardware without relying on classical simulations remains a challenge. Currently, this is primarily due to limited access to quantum computers, lengthy queuing times, noise effects, frequent hardware re-calibrations, and other factors that affect the reliability of results, hindering comprehensive studies.

For these reasons, quantum simulations on classical computers are indispensable for all studies on quantum algorithms and their applications [25, 27]. Such simulations enable us to reliably investigate the proposed quantum models in an ideal environment, their scaling properties and performance with respect to the circuit depth and number of qubits; as for instance in the revised Fig. 4. These studies are crucial to establish the feasibility of the designed quantum algorithms. Subsequently, we demonstrate the execution of our model on real quantum hardware.

R: *To actually claim this algorithm is quantum, I believe the authors need to do two things. First, they should rigorously establish the computational hardness of their kernel. The authors make no claims about whether their kernel is classically hard. I am not an expert on whether it is or isn't - it passes the obvious simple test of not being fully composed of Clifford gates, but if they want to make the claim their algorithm offers a real "quantum advantage," we need to unambiguously establish the circuit is classically hard. [...] We need to establish that there is strong reason to believe the kernel is not one that may be implemented easily on classical resources. If it is classically easy, it will essentially always be expedient to do so and the claim needs to be modified that the authors have identified a quantum-inspired, but classical, kernel.*

A: Our quantum kernel circuit is composed of universal single-qubit gates G and two-qubit CNOT gates, forming a universal set of gates for quantum computation. Specifically, each universal G gate needs at least one T gate when decomposed in the gate set of {Clifford, T}. The number of universal G gates scales linearly with the number of qubits in our circuit. In general, such circuits are expected to be hard to simulate classically. One single T gate is sufficient to make the output probability distributions, such as the ones extracted from the quantum kernel, of quantum circuits classically hard [28]. Additionally, in Ref. [29] it is shown that computing output probabilities of most such quantum circuits is classically hard, using techniques based on random quantum circuit and their universality properties. Apart from results on average-case complexity, techniques to prove classical hardness for specific instances of circuits and input data still do not exist, and remain an open problem in the field.

We study the expressibility of the designed circuit in Supplementary Fig. 1b, showing numerically that our circuit approximates a Haar random distribution. That is, it uniformly explores the available Hilbert space, approaching a universal unitary opera-

tion over all qubits, as L increases.

It is important to note, that even if the specific quantum circuit is expected to be classically hard to simulate this does not preclude the existence of other classical algorithms capable of achieving comparable or superior performance on the dataset. For instance, algorithms based on deep learning or tensor networks, as discussed in a subsequent comment. In the revised manuscript we acknowledge this point.

In our numerical experiments, we progressed from the original study involving 8 qubits, which utilized 96 CPU cores, taking hours to complete to simulations of up to 24 qubits in the revised manuscript, employing 384 CPU cores and memory usage at the order terrabytes (TB), with runtimes spanning several days. The exact runtimes, up to 8 days for 24 qubits, vary depending on the depth L of our kernel circuit. The aforementioned runtimes and memory usage correspond to a single training and evaluation of the model. For the full numerical study presented in the updated Fig. 4, multiple training and evaluation runs where conducted.

In our study we focus on our quantum model's ability to utilise quantum resources, such as entanglement and number of qubits, to solve a fundamental physics problem; discovering new-physics signatures in particle physics data. This is reflected more clearly in the revised manuscript and is discussed in the following comment.

R: *Next, we need to establish that the algorithm's performance scales with qubit count - or better would be quantum volume. [...] The authors do make a small piece of the analysis about scaling qubit count and I commend them for it - but it isn't enough to cover these concerns. To be clear, I am not asking for a beyond-classical simulation of their algorithm on classical resources, or demanding to see better performance on hardware using 50+ qubits or something like that. But I do think that to make the claim they are making they need to provide a theoretical proof of the difficulty of computing their kernel, and they need to demonstrate a performance trend that increases with quantum volume all the way up to the largest reasonable size that may simulated on classical resources. You can argue about "reasonable", but it should be well beyond laptop-scale even if it doesn't test the limits of the classical simulation of quantum circuits. The important point is a trend supporting value from more quantum resources being available.*

A: We greatly appreciate the Referee's feedback and thank the Referee for acknowledging our initial analysis on the qubit count scaling. This feedback had crucial impact on the improvement of the manuscript since it encouraged us to greatly expand our investigation. We are confident that our expanded numerical experiments adequately address the important concerns raised by the Referee. Previously, we studied the impact of entanglement and expressibility of the circuit, only for the case of 8 qubits and only up to $L = 6$ repetitions of our data encoding circuit; as seen in the previous Fig. 4. In the revised manuscript, we substantially extend our analysis. Specifically, we extend the studied circuit depths, from no entanglement (NE) to $L = 13$, and we

perform numerical experiments for up to 24 qubits. As also discussed in the previous comment, the computing resources needed to perform our study are well beyond laptop-scale. Specifically, our simulations were conducted on distributed computing clusters at ETH and dedicated computing nodes at CERN equipped with 6 TB of memory. The runtime and memory consumption increase drastically when going from 4 to 24 qubits, from the order of minutes to the order of days, respectively.

The revised Fig. 4 illustrates the trend of improving model performance as quantum resources, both the number of qubits n_q and circuit depth L, are increased. We refrain from using quantum volume to quantify the quantum resources utilised by our algorithm since quantum volume is dependent on the hardware characteristics. Specifically, it depends on the qubit connectivity, the noise levels, the calibration techniques, and the algorithm used to transpile the circuit to the native gates of the hardware. We are in favour of presenting our results in a manner that is not device-specific. For this reason, we quantify our results as function of the quantum circuit "surface", given by the depth L and the number of qubits n_q (width).

Specifically, in Fig. 4a the performance of our algorithm is quantified by Δ_{QC} as a function of the depth of the circuit L, for all the investigated qubit numbers. Fig. 4b summarises the increase of performance of the quantum model as a function of the qubits and Fig. 4c presents the entanglement as a function of L. We observe that we reach best performance when entanglement is close to maximal. Conversely, when entanglement is absent (NE) the performance of the quantum model suffers for all values of n_q.

R: *Note - I do understand the authors made an evaluation of whether the kernel was feasible on NISQ hardware, but their central, controversial, claim is separate from that. It would be somewhat interesting to understand what they did in terms of error mitigation to get a good result on hardware, but that isn't the main point of the paper.*

A: As the Referee correctly identified, our objective is to demonstrate the effectiveness of our kernel in operating on current noisy devices. The presented hardware results do not have any error mitigation, demonstrating the robustness of our architecture. We experimented with out-of-the-box measurement error mitigation, however, it did not show a noticeable improvement in the model performance. We agree with the Referee that an interesting future study would be to systematically investigate different and specific error mitigation techniques required to run on larger system sizes.

R: *I think it would also be interesting to compare the performance of the exact quantum simulation of their algorithm to a "fuzzy" method of circuit simulation like a tensor network. The paper shows some evidence that increasing entanglement in the simulation helps, but only up to a point after which it actually begins to penalize performance of the kernel. This suggests to me a tensor network simulation of their circuit may*

offer all the performance with scaling costs that keep the overall accounting in favor of a classical ML algorithm for problems of arbitrary size. I am not requiring the authors to do this work here - that is an unreasonable scope increase. However, they should moderate their claims until this sort of study is performed.

A: This comment further encouraged us to substantially expand our analysis in terms of circuit depth and entanglement, addressing limitations of the original study. In the revised manuscript, we demonstrate that for each n_q the performance of the quantum models stabilizes at an optimal value as the entanglement approaches its maximum with increasing L, as evidenced by the updated Fig. 4a and Fig. 4c. The Referee's remark, *"The paper shows some evidence that increasing entanglement in the simulation helps, but only up to a point after which it actually begins to penalize performance of the kernel."*, refers to the previous version of Fig. 4, in which the performance of the quantum kernel seemingly drops for $L = 6$ and FE. In our extended numerical experiments we resolve this issue, explained in the following.

The slight drop in performance observed in the previous version of Fig. 4 is a statistical fluctuation, as showcased in our extended numerical studies presented in the revised Fig. 4. Furthermore, FE in the previous version of Fig. 4 refers to a circuit with all-to-all CNOT connections and $L = 3$. Both the expressibility and entanglement of the FE circuit is lower than that of all circuits with $L \in [4, 5, 6]$ with nearest neighbour CNOT connections, as seen from the calculations presented in Supplementary Fig. 1a and Fig. 1b of the original manuscript. All the above lead to a potentially misleading and not adequately informative previous version of Fig. 4. For these reasons, we exclude FE in the revised manuscript and provide the updated Fig. 4 instead to enhance clarity of our results.

Additionally, we concur with the Referee's assessment that a tensor network approach for our problem would be interesting and could offer further insights into our findings. Such study will be conducted in forthcoming work. We have adjusted our conclusions to account for future tensor network investigations.

R: *We also need to understand a lot more about the classical benchmark. We don't know much about how many approaches they tried or how much optimization was done on the purely classical side - we just know there is at least one purely classical algorithm their quantum-inspired algorithm significantly outperforms. The authors have chosen a good classical algorithm, but we need more insurance against potential claims of cherry-picking. Because they have chosen an esoteric task, it is hard to have an intuition about performance. At best we can say we have a state-of-the-art (SOTA) measurement using a quantum-inspired algorithm and invite the community to try to best it. It isn't clear there is motivation to do so, but the claim will hinge on beating all comers. If the authors had chosen a well-studied dataset, their new SOTA would be meaningful, but as it is we have to hope they are able to build a community willing to compete on this task. The authors should, of course, be commended again for pub-*

lishing the dataset they used and their code - this makes such a competition possible in the future. Note that I am not asking the authors to beat every classical algorithm in existence before attempting to publish - that isn't possible. But the claim needs to be adjusted - they found a quantum-inspired algorithm that beats at least one other classical algorithm, and they should explain in some depth why that benchmark was chosen and defend their benchmark as at least a plausible competitor for the best available option.

A: We thank the Referee for recognizing our dedication to publicly sharing both our dataset and code. We are confident that our results valuable insights to the community. Many prior studies focusing on classical benchmark datasets fail to demonstrate any improvement in performance by quantum models, nor do they provide numerical evidence supporting the utilisation of entanglement to enhance quantum model performance [24]. This underscores the significance of our study's contributions as discussed in previous comments.

Our objective is to fully characterise the behaviour of our quantum model and compare it to classical models with similar structures and complexities, using the same training data. We assessed the performance of various classical kernel-based models, each equipped with linear, polynomial, sigmoid, and Radial Basis Function (RBF), or Gaussian, kernels. We further evaluated the performance of classical clustering algorithms, namely the K-means and K-medians as seen in Fig. 3. Additionally, we explored different values of the hyperparameter ν (refer to the Methods section for a detailed discussion). We chose the RBF classical kernel as a benchmark because it consistently exhibited superior performance among all tested classical models.

Moreover, we observed that the autoencoder alone, as discussed in the subsequent section and illustrated in the attached Figure 1, demonstrated significantly poorer anomaly detection performance without the inclusion of quantum models in the latent space. This underscores the importance of integrating quantum models for enhanced performance. Notably, the autoencoder's performance can be interpreted as a further classical benchmark against which the quantum models showcase improvements. In the revised manuscript, we clarify that our results do not obexclude every possible classical model.

R: *The authors face one other problem in the defense of this work, and it is difficulty they alluded to in the paper - namely, this is a hybrid algorithm at best with a substantial amount of the "work" done by the auto-encoder. This is relevant because their circuit ansatz ties the number of qubits to the feature size of the auto-encoder. This makes disentangling, if you will, the qubit scaling effects from the feature richness of the latent space as the dimension changes. This is balanced somewhat by the classical benchmark, but it remains an onerous detail to wrestle with because the classical benchmark needs to be re-optimized as you go along.*

A: We would like to address the Referee's concern by presenting the anomaly detection performance using only autoencoder. The autoencoder achieves an average of 75%

AUC, across the different latent space dimensions, as seen by the ROC curve in the attached Figure 1). Hence, the quantum algorithms provide a substantial enhancement of performance with respect to this baseline. We added this important information in the revised manuscript.

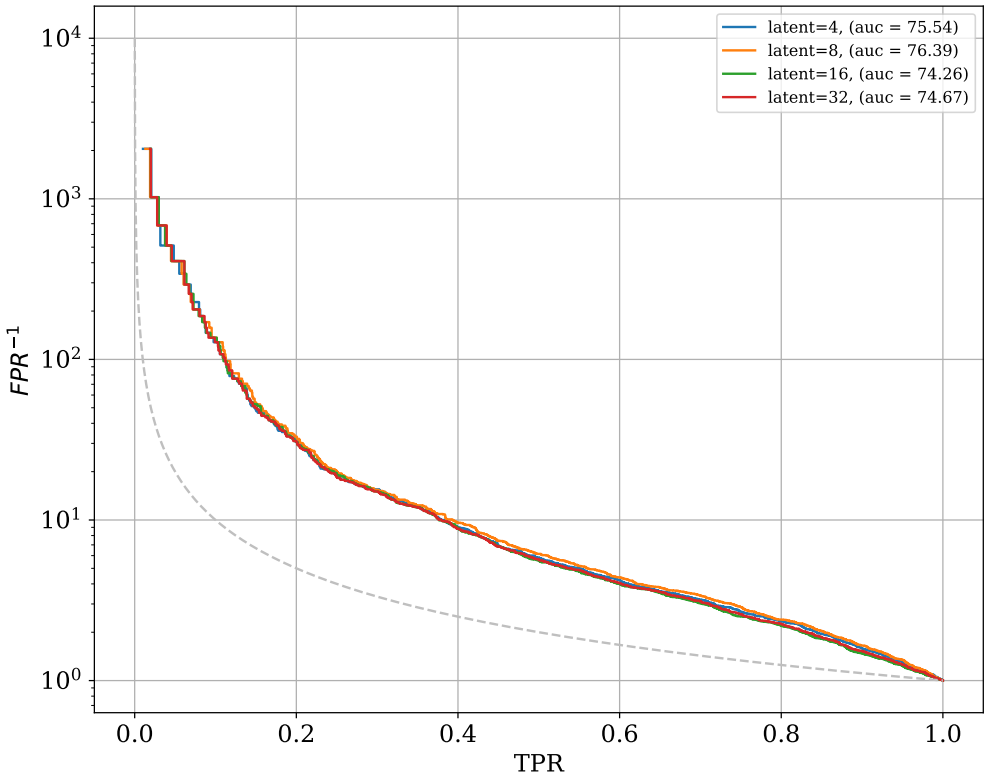

Figure 1: Anomaly detection performance of the autoencoder for different latent space dimensions.

R: *A smaller detail, but one thing that is unclear is how the authors deal with computing expectation values for the kernel matrix. For the classical simulation, was the exact value used? For real hardware this is never possible. Of course you can make the statistical uncertainty arbitrarily small in the limit of infinite shots, but shots are not free and the price of growing a kernel matrix grows like the dataset size squared, so it is not practical to assume for a very long time that we will have the ability to take an "arbitrary" number of shots for every element of a large kernel matrix. Furthermore, since readout rates are a few kHz, collecting shots adds some significant burden to the overall cost of running the algorithm if you really need your statistical error on the kernel matrix elements to be below a tight threshold. And if you don't need the statistical error below a tight threshold, their "quantum" algorithm will likely always be outperformed by a tensor-network simulation of their circuit. What is the required statistical*

uncertainty on the elements of the kernel matrix to achieve the observed performance?

A: For the quantum hardware experiments, we empirically found that 10^4 shots were sufficient. This threshold can be different for larger numbers of qubits and noise profiles of different quantum computers. This investigation is an interesting topic for future hardware studies.

Concerning the measurement costs, we agree with the Referee that the standard measurement protocol will prevent the scaling of the method towards larger data sets. However, recent developments in the domain of classical shadows [30] and informationally-complete positive operator-valued measurements (IC-POVM) [31] make possible the evaluation of a set of generic operators, using only a logarithmic number of measurements, dramatically reducing the costs for the evaluation of the quantum kernel. In addition, the quality of the expectation values (i.e., the standard deviation) can be improved by means of a further optimization of the IC-POVM effects [31] or duals [32, 33]. The last two sentences have been added to the 'Results' section of the main text.

R: *One other note - the authors call their kernel "novel" but it is very similar to the kernel in https://www.nature.com/articles/s41534-021-00498-9 from 2 years ago. The major difference is the repetition of the ansatz - which is a big difference to be sure! - but the structure of the kernels is otherwise quite similar (aside from the use of a CNOT for an entangling gate as opposed to sqrt(iSWAP) - a choice owing to the native 2-qubit entangling option on IBM vs Google hardware).*

A: We thank the Referee for bringing our attention to this interesting study. We have added the reference in the revised manuscript. Our quantum circuit is substantially different in two key aspects: firstly, the input data features are permuted at the second block of rotation gates (cf. Fig. 2a), and secondly, the whole data encoding circuit, as pointed out by the Referee, is repeated L times to achieve high expressibility and entanglement capability (demonstrated in Supplementary Fig. 1a and 2b).

References

[1] Georges Aad *et al.* (ATLAS), "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), arXiv:2306.03637 [hep-ex] .

[2] M. Farina *et al.*, "Searching for New Physics with Deep Autoencoders," Phys. Rev. D **101**, 075021 (2020), arXiv:1808.08992 [hep-ph] .

[3] Tuhin S. Roy and Aravind H. Vijay, "A robust anomaly finder based on autoencoders," (2019), arXiv:1903.02032 [hep-ph] .

[4] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer M. Thompson, "QCD or What?" SciPost Phys. **6**, 030 (2019), arXiv:1808.08979 [hep-ph] .

[5] O. Knapp *et al.*, "Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark," Eur. Phys. J. Plus **136**, 236 (2021), arXiv:2005.01598 [hep-ex] .

[6] Taoli Cheng, Jean-Fran çois Arguin, Julien Leissner-Martin, Jacinthe Pilette, and Tobias Golling, "Variational autoencoders for anomalous jet tagging," Phys. Rev. D **107**, 016002 (2023).

[7] Andrew Blance, Michael Spannowsky, and Philip Waite, "Adversarially-trained autoencoders for robust unsupervised new physics searches," JHEP **10**, 047 (2019), arXiv:1905.10384 [hep-ph] .

[8] Blaž Bortolato, Aleks Smolkovič, Barry M. Dillon, and Jernej F. Kamenik, "Bump hunting in latent space," Phys. Rev. D **105**, 115009 (2022), arXiv:2103.06595 [hep-ph] .

[9] Georges Aad *et al.* (ATLAS), "Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ TeV with the ATLAS detector," (2023), arXiv:2307.01612 [hep-ex] .

[10] *Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at* $\sqrt{s} = 13\ TeV$, Tech. Rep. (CERN, Geneva, 2024).

[11] Ekaterina Govorkova *et al.*, "Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider," Nature Mach. Intell. **4**, 154–161 (2022), arXiv:2108.03986 [physics.ins-det] .

[12] O. Cerri *et al.*, "Variational Autoencoders for New Physics Mining at the Large Hadron Collider," JHEP **05**, 036 (2019), arXiv:1811.10276 [hep-ex] .

[13] "Anomaly Detection in the CMS Global Trigger Test Crate for Run 3," (2023).

[14] Georges Aad *et al.* (ATLAS), "Observation of quantum entanglement in top-quark pairs using the ATLAS detector," (2023), arXiv:2311.07288 [hep-ex] .

[15] Alba Cervera-Lierta, Jose Ignacio Latorre, Juan Rojo, and Luca Rottoli, "Maximal entanglement in high energy physics," SciPost Physics **3**, 036 (2017).

[16] Claudio Severi, Cristian Degli Esposti Boschi, Fabio Maltoni, and Maximiliano Sioli, "Quantum tops at the lhc: from entanglement to bell inequalities," The European Physical Journal C **82**, 285 (2022).

[17] M. Fabbrichesi, R. Floreanini, E. Gabrielli, and L. Marzola, "Bell inequalities and quantum entanglement in weak gauge bosons production at the lhc and future colliders," (2023), arXiv:2302.00683 [hep-ex, physics:hep-ph, physics:quant-ph].

[18] CMS Collaboration, "Measurement of the top quark polarization and $t\bar{t}$ spin correlations using dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ tev," Physical Review D **100**, 072002 (2019), arXiv:1907.03729 [hep-ex].

[19] M. Fabbrichesi, R. Floreanini, and G. Panizzo, "Testing bell inequalities at the lhc with top-quark pairs," Physical Review Letters **127**, 161801 (2021).

[20] Yoav Afik and Juan Ramón Muñoz de Nova, "Quantum information with top quarks in qcd," Quantum **6**, 820 (2022).

[21] Diptimoy Ghosh and Rajat Sharma, "Bell violation in $2 \to 2$ scattering in photon, gluon and graviton efts," (2023), arXiv:2303.03375 [hep-ph, physics:hep-th, physics:quant-ph].

[22] J. Kübler *et al.*, "The inductive bias of quantum kernels," in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates, Inc., 2021) p. 12661–12673.

[23] Casper Gyurik and Vedran Dunjko, "Exponential separations between classical and quantum learners," (2023), arXiv:2306.16028 [quant-ph] .

[24] Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld, "Better than classical? The subtle art of benchmarking quantum machine learning models," (2024), arXiv:2403.07059 [quant-ph] .

[25] H.Y. Huang *et al.*, "Power of data in quantum machine learning," Nature Communications **12**, 2631 (2021).

[26] Sergio Altares-López, Angela Ribeiro, and Juan José García-Ripoll, "Automatic design of quantum feature maps," Quantum Science and Technology **6**, 045015 (2021), 2105.12626 .

[27] Evan Peters *et al.*, "Machine learning of high dimensional data on a noisy quantum processor," npj Quantum Information **7**, 161 (2021), arXiv:2101.09581 [quant-ph] .

[28] M. Hinsche, M. Ioannou, A. Nietner, J. Haferkamp, Y. Quek, D. Hangleiter, J.-P. Seifert, J. Eisert, and R. Sweke, "One T gate makes distribution learning hard," Physical Review Letters **130**, 240602 (2023), 2207.03140 .

[29] Ramis Movassagh, "The hardness of random quantum circuits," Nature Physics **19**, 1719–1724 (2023).

[30] Hsin-Yuan Huang, Richard Kueng, and John Preskill, "Predicting many properties of a quantum system from very few measurements," Nature Physics **16**, 1050–1057 (2020).

[31] Guillermo García-Pérez, Matteo A.C. Rossi, Boris Sokolov, Francesco Tacchino, Panagiotis Kl. Barkoutsos, Guglielmo Mazzola, Ivano Tavernelli, and Sabrina Maniscalco, "Learning to measure: Adaptive informationally complete generalized measurements for quantum algorithms," PRX Quantum **2**, 040342 (2021).

[32] Joonas Malmi, Keijo Korhonen, Daniel Cavalcanti, and Guillermo García-Pérez, "Enhanced observable estimation through classical optimization of informationally over-complete measurement data – beyond classical shadows," (2024), arXiv:2401.18049 [quant-ph] .

[33] Laurin E. Fischer, Timothée Dao, Ivano Tavernelli, and Francesco Tacchino, "Dual frame optimization for informationally complete quantum measurements," (2024), arXiv:2401.18071 [quant-ph] .

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I appreciate the detailed and thoughtfully-written rebuttal from the authors, and the many domain-specific clarifications they included. They went to great lengths to address my many concerns and I am very appreciative of their patience and their dedication to high-quality scholarship. I will, in this review, focus on the paper and the changes in the text of the paper (as opposed to going though the reubuttal with my thoughts) since it is the text that matters for readers.

First, the authors addressed the single most important issue, which was how the model performs with increasing quantum resources in terms of entanglement and number of qubits deployed. Further, they took appropriate steps to scale the computation up far enough in quantum resources to reasonably begin to look at associated trends. Additionally, they moderated the language around "quantum advantage" in a way that is more consistent with the observed effects and less sensationalist. As such I believe this paper is nearly ready to be accepted.

I believe there is one "big" issue that should be addressed, and I have a couple of other very minor points. However, I do not think it is appropriate to require another round of review on my part. Instead, I will make some suggestions that are relatively brief and trust the authors to take them seriously. They have shown so far they are serious about the quality of this work, so I feel confident they will address these relatively minor points effectively. Hopefully we can avoid the associated delay of another round of referee review (although I am happy to look at the paper again if the Editors would like me to do so).

In particular, the "big" (although really not so large) issue is with uncertainties on the kernel matrix elements. The authors explain the number of shots used in experiments, and discuss techniques for evaluating operators in a shot-efficient manner, but I cannot determine what the uncertainties on the kernel matrix elements were. In Figure 3 they show uncertainty bands on the various algorithms in the FPR^-1 vs TPR plots. How are these uncertainties computed? I suppose they assumed zero error on the kernel matrix elements and used statistical spread in repeated experiments. It isn't clear.

Put another way, the important issue to address is — what uncertainty is required on the kernel matrix elements to see significant quantum performance advantages over their classical benchmark? We need a clear answer to that question because otherwise we will remain in a situation where it is unclear whether one could get the same (or better) performance with a fuzzy

circuit estimation method like tensor networks that admits some small uncertainty in the kernel elements in exchange for vastly faster computations.

To be clear, I am not asking the authors to show that their technique is outperforming a tensor network implementation. I only believe it is important to know clearly what the uncertainty they assume on the the kernel matrix elements in their simulated results is (it isn't important for the hardware demonstrator) and how large it could be while still admiting a performance advantage over their classical benchmarks. This should not be much work, and should only change a couple of sentences in the paper.

A couple of other small points that confused me slightly, and may confuse a reader, that the authors should briefly address in the text: (1) At the conclusion of Table 1, the authors mention they do 5-fold testing with N-test = 10^5. But they have already defined a training sample of size 600... so what does the 10^5 refer to? (2) They mention the performance of the autoencoder in a standalone anomaly detection algorithm mode, but don't discuss how it is used. I assume they do some sort of cut on the reconstructed event after it is passed through the autoencoder, but it isn't clear what that cut is or how it was optimized.

Response to Referee

We thank Referee for carefully reviewing our revised manuscript. Here, we proceed to address the comments of the Referee and describe the corresponding modifications made to the manuscript.

Replies and comments to Referee

R: *I appreciate the detailed and thoughtfully-written rebuttal from the authors, and the many domain-specific clarifications they included. They went to great lengths to address my many concerns and I am very appreciative of their patience and their dedication to high-quality scholarship. I will, in this review, focus on the paper and the changes in the text of the paper (as opposed to going though the rebuttal with my thoughts) since it is the text that matters for readers.*

First, the authors addressed the single most important issue, which was how the model performs with increasing quantum resources in terms of entanglement and number of qubits deployed. Further, they took appropriate steps to scale the computation up far enough in quantum resources to reasonably begin to look at associated trends. Additionally, they moderated the language around "quantum advantage" in a way that is more consistent with the observed effects and less sensationalist. As such I believe this paper is nearly ready to be accepted.

A: We appreciate the Referee's positive feedback and acknowledgment of the improvements in our manuscript following the thorough revision. In the following, we address the comments and provide clarifications in the revised manuscript.

R: *I believe there is one "big" issue that should be addressed, and I have a couple of other very minor points. However, I do not think it is appropriate to require another round of review on my part. Instead, I will make some suggestions that are relatively brief and trust the authors to take them seriously. They have shown so far they are serious about the quality of this work, so I feel confident they will address these relatively minor points effectively. Hopefully we can avoid the associated delay of another round of referee review (although I am happy to look at the paper again if the Editors would like me to do so).*

In particular, the "big" (although really not so large) issue is with uncertainties on the kernel matrix elements. The authors explain the number of shots used in experiments, and discuss techniques for evaluating operators in a shot-efficient manner, but I cannot determine what the uncertainties on the kernel matrix elements were. In Figure 3 they show uncertainty bands on the various algorithms in the FPR^{-1} vs TPR plots. How are these uncertainties computed? I suppose they assumed zero error on the kernel matrix elements and used statistical spread in repeated experiments. It isn't clear.

A: The Referee is correct that the uncertainty bands in Figure 3 represent the statistical fluctuations from the repeated testing of our models during training and k-fold testing. In our simulations, we find that these fluctuations are the largest source of uncertainty in the anomaly detection performance of the models. For this reason, in Figure 3 the uncertainty on the kernel matrix elements is assumed to be zero. We investigate and discuss the uncertainty on the kernel matrix elements further in the following comments and provide further details in the revised manuscript.

R: *Put another way, the important issue to address is — what uncertainty is required on the kernel matrix elements to see significant quantum performance advantages over their classical benchmark? We need a clear answer to that question because otherwise we will remain in a situation where it is unclear whether one could get the same (or better) performance with a fuzzy circuit estimation method like tensor networks that admits some small uncertainty in the kernel elements in exchange for vastly faster computations.*

To be clear, I am not asking the authors to show that their technique is outperforming a tensor network implementation. I only believe it is important to know clearly what the uncertainty they assume on the the kernel matrix elements in their simulated results is (it isn't important for the hardware demonstrator) and how large it could be while still admitting a performance advantage over their classical benchmarks. This should not be much work, and should only change a couple of sentences in the paper.

A: We conducted additional numerical experiments to assess the impact of uncertainty in the kernel matrix elements on model performance. Specifically, we investigated the amount of uncertainty that could be allowed in the kernel values while still achieving $\Delta_{\mathrm{QC}} > 1$, i.e., a performance advantage over the classical benchmarks. We find that an uncertainty of order 10^{-3} still allows us to achieve $\Delta_{\mathrm{QC}} > 1$, and has a negligible effect on the uncertainty bands of Figure 3 and the Δ_{QC} values presented in Figure 4. Conversely, when the uncertainty reaches order of 10^{-2} or higher, our quantum model performs similarly to the classical benchmarks, resulting in $\Delta_{\mathrm{QC}} \sim 1$. Following the Referee's recommendation, we have added two sentences providing this information in the revised manuscript.

R: *A couple of other small points that confused me slightly, and may confuse a reader, that the authors should briefly address in the text: (1) At the conclusion of Table 1, the authors mention they do 5-fold testing with N-test = 10^5. But they have already defined a training sample of size 600... so what does the 10^5 refer to?*

A: Indeed, the training sample size is 600, while the number of testing samples is 10^5. Specifically, the models are trained using $N_{\mathrm{train}} = 600$ data samples and their performance is tested using $N_{\mathrm{test}} = 10^5$ independent test data samples. This test dataset is

divided into $k = 5$ subsets (folds), and the trained models are evaluated on each subset. From these k model evaluations we compute the mean and standard deviation of the different performance metrics we define in the manuscript. This way, we compute the uncertainties in Figure 3 and in Table I. We have clarified the above in the revised caption of Table I.

R: *(2) They mention the performance of the autoencoder in a standalone anomaly detection algorithm mode, but don't discuss how it is used. I assume they do some sort of cut on the reconstructed event after it is passed through the autoencoder, but it isn't clear what that cut is or how it was optimized.*

A: Indeed, for the standalone autoencoder anomaly detection the data are passed through the autoencoder and the anomaly metric is the reconstruction loss function. In the revised manuscript, we a more clear explanation in the Autoencoder section of the Methods. As for all studied models, we vary the cut on the reconstruction loss to measure the performance of the autoencoder by constructing the ROC curves. The intuition as correctly implied by the Referee is that anomalous data have a higher reconstruction loss since the autoencoder is optimised during the training to minimise the reconstruction loss for Standard Model (normal) collision events.