VI. SUPPLEMENTARY INFORMATION

A. Energy analysis

To establish the quality of the QUBO construction, we evaluate the Ising model energy E (Eq. 9) for a 500-track sector and compare the minimum-energy solution to the ground truth. Since the quadratic unconstrained binary optimization (QUBO) formulation does not allow the track efficiency and purity to be directly optimized, the QUBO construction provides an indirect objective function that should be highly correlated with precision and recall.

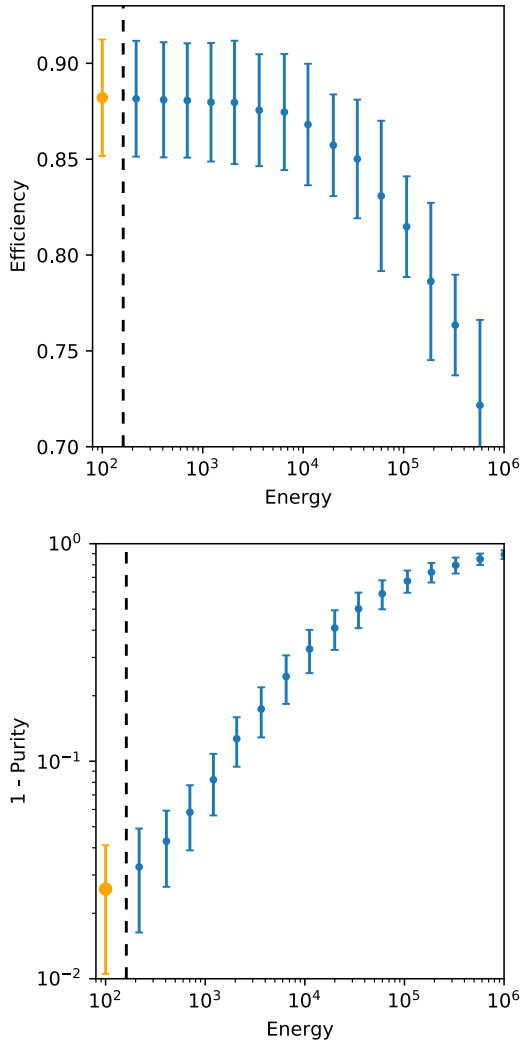

FIG. 8. Correlation of Ising model energy with track efficiency and purity for a 500-track sector. The Hamiltonian has been offset to normalize the lowest energy (orange) to 1000. Ground truth energy (100% purity and 93% efficiency) is shown by the black dashed line.

As seen in Figure 8, the ground truth energy is in the neighborhood of the minimum energy relative to the purity and efficiency metrics, providing further confirmation that the QUBO construction accurately represents the tracking problem.

$$
\begin{aligned}
E = &- \sum_{a,b,c} \left(\frac{\cos^{\lambda}(\theta_{abc}) + \rho \cos^{\lambda}(\phi_{abc})}{r_{ab} + r_{bc}} \right) s_{ab} s_{bc} \\
&+ \eta \sum_{a,bc} \left(z_c - \frac{z_c - z_a}{r_c - r_a} r_c \right)^{\zeta} s_{ab} s_{bc} \\
&+ \alpha \left(\sum_{b \neq c} s_{ab} s_{ac} + \sum_{a \neq c} s_{ab} s_{cb} \right) \\
&+ \sum_{a,b} \left(\beta P(s_{ab}) - \gamma \right) s_{ab}
\end{aligned}
\tag{9}
$$

B. Sparse QUBO construction

Due to the lack of full connectivity on current programmable quantum annealers, we propose additional methodology to increase sparsity in the quadratic unconstrained binary optimization (QUBO) formulation of the tracking problem. Three general stages are used to iteratively simplify the problem, with the quantum annealer solving the final nontrivial component (as shown in the main Results section). The impact of each stage of preprocessing is visualized in terms of purity and efficiency in Figure 10.

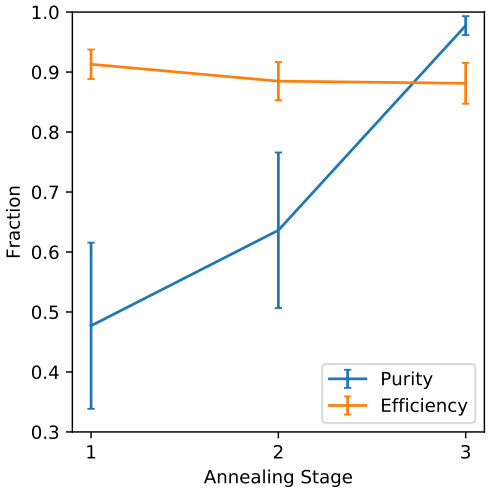

FIG. 10. Purity and efficiency dependence on annealing stage in the sparse QUBO construction. Purity is significantly improved without a large impact on overall tracking efficiency (maximum 93% from the Gaussian KDE cutoff).

The total procedure is outlined in detail in Figure 3. In the first stage, a Gaussian kernel density estimator

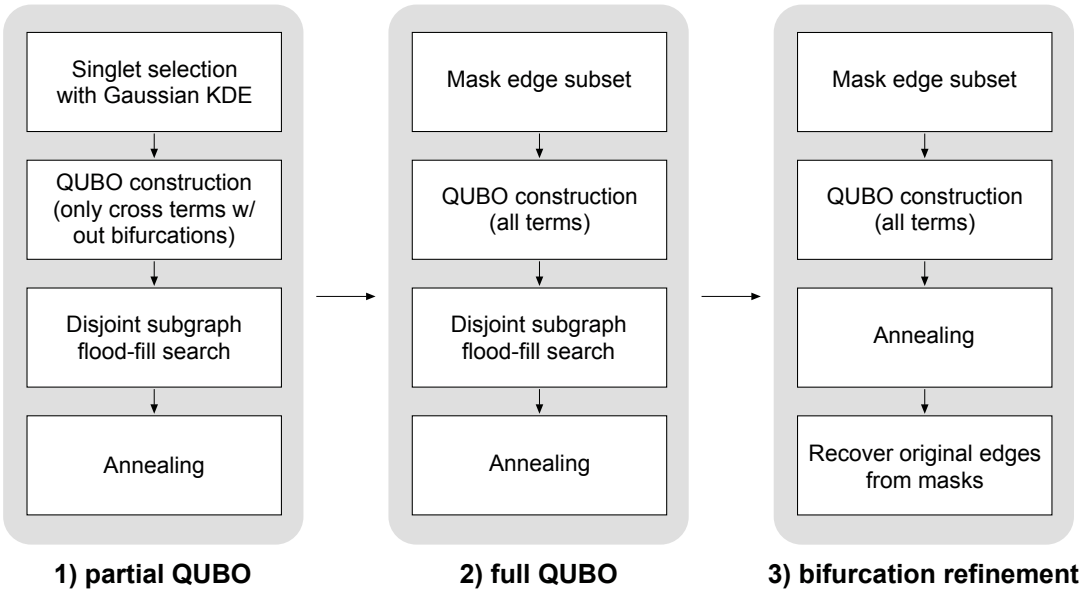

FIG. 9. Summary of dimension reduction methodology. After solving a partial QUBO with no bifurcation or single-spin terms, quantum annealing may be performed on the full QUBO using a partially connected and then fully connected graph. By tuning the cutoff parameters at each stage (i.e. constraining the sizes of edge subsets), the problem can be scaled to quantum hardware while attempting to preserve the quality of the solution.

(KDE) is used to approximate a prior probability on a given edge being on or off based on the angle in the rz-plane and z-intercept of a line segment between two hits (Figure 11). Although geometric approaches may be taken in place of a Gaussian KDE, the complex geometry of the detector end-caps (right and left regions of Figure 12) as well as the versatility in extending the methodology to new detector geometries provide justification for using a generic prior probability estimation methodology.

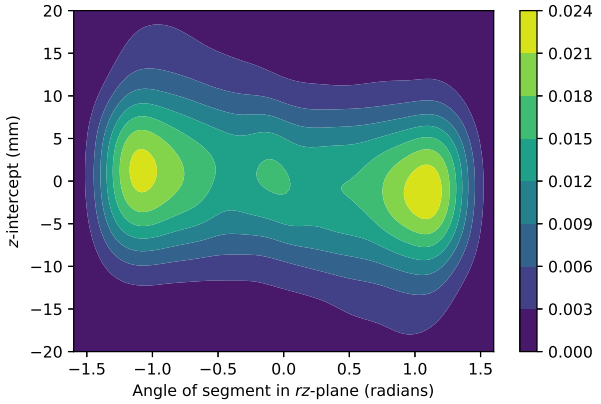

FIG. 11. Gaussian kernel density estimation of a prior probability for two hits to be connected by an edge, allowing information from the interaction points and detector geometry to be introduced into the QUBO.

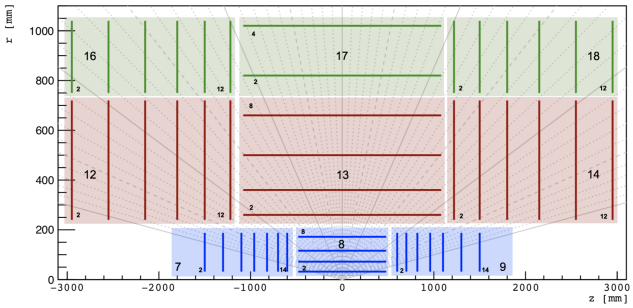

FIG. 12. TrackML detector geometry in the rz-plane [29].

By placing a threshold on the minimum prior probability, the candidate set of edges can be classically reduced in size. A QUBO is then constructed from edge compatibility scores without bifurcations or bias terms:

$$E = -\sum_{a,b,c}\left(\frac{\cos^{\lambda}(\theta_{abc}) + \rho\cos^{\lambda}(\phi_{abc})}{r_{ab}+r_{bc}}\right)s_{ab}s_{bc} + \eta\sum_{a,bc}\left(z_c - \frac{z_c - z_a}{r_c - r_a}r_c\right)^{\zeta} s_{ab}s_{bc} \tag{10}$$

Since the remaining terms in the QUBO are negative, solving this QUBO yields high-efficiency but low-purity tracks. To reduce connectivity, we perform a disjoint sub-graph flood-fill search on the space of edges, labeling up to the 5 best edge connections of a starting edge and spreading the label of each of the children edges to their neighbors. Specifically, for each sub-graph, we begin at the vertex closest to the origin and select the 5

best edge connections; each connected vertex is recursively processed in the same way, incurring a time complexity linear in the number of vertices (i.e. total number of hits). After the search, each edge is connected to at most 5 other edges, allowing the problem to be more easily embedded in the Chimera graph architecture of the D-Wave annealer since each edge is encoded as a qubit. Moreover, the event is divided into sub-graphs that are disjoint by edges, allowing it to be annealed in multiple QUBO problems.

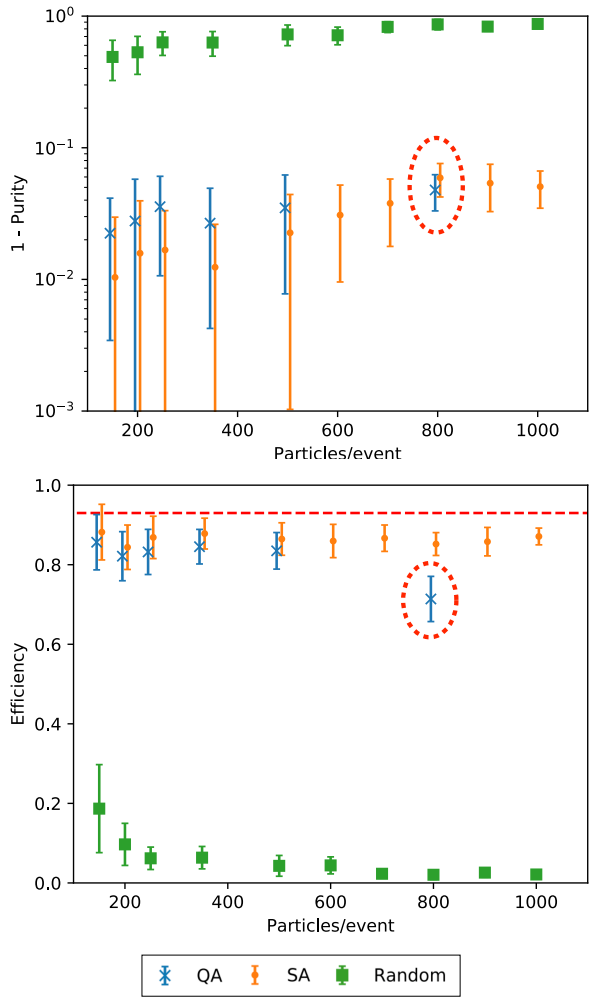

FIG. 13. Purity and efficiency for with QUBO construction parameters tuned to process larger events with high purity but lower efficiency. The QA point for 800 tracks (circled) is shown with the new parameters, while the remaining data uses the parameters reported in the main text.

Since the first stage retains efficiency, we may use the subset of edges it requires to be on in the second stage while discarding all edges classified as off. The full QUBO (Eq. 9) is then annealed on a sub-graphed event, following the same procedure as outlined above. A final annealing on the new edge subset is applied to the full QUBO without any sub-graph constraints. This ensures that all possible bifurcations are included in the annealing process, yielding a QUBO that approximates the full event by only considering high-quality edges determined by the preceding annealings. Each annealing is preceded by a classical heuristic to fix QUBO variables in polynomial time, following standard D-Wave API usage to permit the largest possible QUBO to fit on the quantum processor.

For the parameters reported in the main text, this pre-processing procedure with sub-graphs enables events of up to 500 tracks to be annealed on the D-Wave 2X, corresponding to encoding 60,000 QUBO variables in a fully-connected graph on an annealer with only 33 fully-connected logical qubits. By changing the parameters of the QUBO construction, larger events may be annealed with further cost to performance. For instance, we may modify the threshold of the edge affinity term, which sets the reward weight to 0 if $\cos^{\lambda}(\theta_{abc}) < \tau$, i.e. if two adjacent edge segments do not sufficiently line up to a helix (shown in Figure 2). As τ increases, the cross-terms are increasingly sparse, allowing the event to be broken into a larger number of sub-graphs. By increasing $\tau = 0.996$ (used in the main text) to $\tau = 0.9997$, we may fit events as large as 800 tracks on the D-Wave 2X, corresponding to encoding 160,000 QUBO variables in a fully-connected graph. However, this decreases overall performance (see Figure 13).

C. Simulated annealing

In simulated annealing (SA) [52], we expect the time scaling after pre-processing to increase exponentially as $O\left(\sum_{i=1}^{K} \exp\left(cm_i\right)\right)$ for $c > 0$ where the i^{th} sub-graph has m_i edges. Thus, the scaling is a function of the distribution over sub-graph sizes (Figure 14).

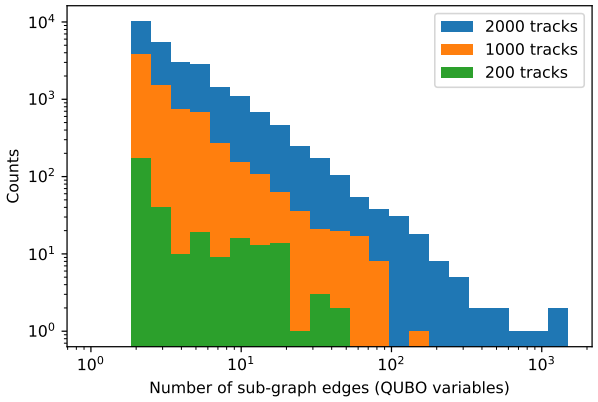

FIG. 14. Histogram of sub-graph sizes m_i summed over the 5 largest event sectors (1/16 of an event) for different track densities. Each bin corresponds to the number of edges in a sub-graph, which is equivalent to the number of variables in a QUBO.

This exponential-time fit with respect to the number of sub-graph edges (or equivalently, the number of QUBO variables) appropriately models the computational time of SA with $c = (8.52 \pm 0.076) \times 10^{-3}$, as seen in Figure 15 where the quality of the fit is shown as a function of track number.

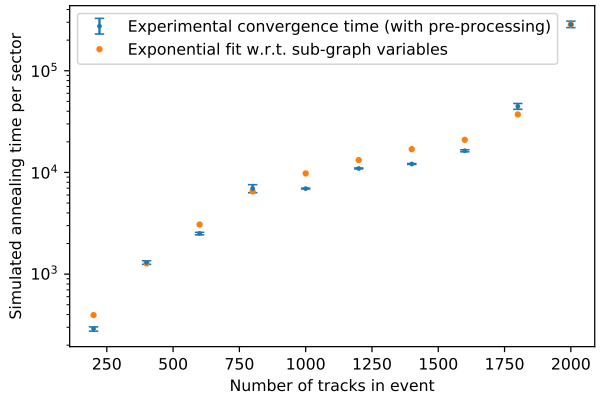

FIG. 15. SA convergence time with a fit $O\left(\sum_i^K \exp\left(cm_i\right)\right)$. By assuming the logarithm of convergence time is proportional to the distribution of sub-graph sizes $\{m_i\}$, we find a reasonable model for pre-processed scaling with a proportionality constant of $c = (8.52 \pm 0.076) \times 10^{-3}$. Error bars show the 1σ variation in annealing time, not including variation across events.

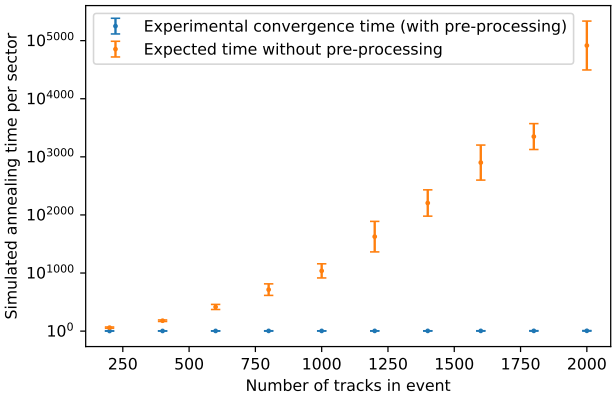

FIG. 16. SA convergence time bounded from above by $O\left(\exp\left(ch^2\right)\right)$, with $c = (8.52 \pm 0.076) \times 10^{-3}$.

Note that although the convergence time is exponential after pre-processing, this is bounded from above by $O\left(\exp\left(ch^2\right)\right)$ obtained if there were no pre-processing (see Figure 16).

To perform SA, we start with a randomly initialized vector of spins $\vec{s}$ of length N. In each sweep, we flip N variables one by one. If the flip decreases the QUBO energy, $E(\vec{s}') < E(\vec{s})$, the new state is accepted. However, if the energy is increased, we accept the flip with probability $\exp\left(-\beta\left(E(\vec{s}') - E(\vec{s})\right)\right)$ where β is the inverse temperature at that time. After each sweep, the inverse temperature is increased. We use $\beta_{\text{init}} = 0.1$ and $\beta_{\text{fin}} = 10$ with a linear annealing schedule (at each sweep, β is increased by $(\beta_{\text{fin}} - \beta_{\text{init}})/S$ where S is the total number of sweeps performed). For the time complexity analysis above, we only seek to find a lower bound on the runtime of SA, and we thus select parameters to better estimate time to convergence at the cost of minimizing energy. The convergence time for each QUBO is computed after performing $S = 15000$ sweeps and 5 runs of SA to obtain the lowest energy state for these schedule parameters. We then begin with $S = 1$ and increment S until we reach the previously found lowest energy, thus obtaining the number of sweeps required to converge. This process is carried out 5 times to obtain 5 samples for convergence times for each QUBO. We then use bootstrapping to compute mean convergence time and confidence intervals, generating 250 pseudo-samples for every tracking event and sampling uniformly from the 5 convergence time samples for each sub-QUBO.

Note that for large QUBOs in the 2000-track data point, we weakened our criteria for convergence by accepting convergence if the found state was ϵ-close to the converged low-energy state. However, this strictly reduces the number of sweeps required for the 2000-tracks data point and still provides exponential scaling with respect to number of tracks. Weakening the lower bound through this heuristic still yielded exponential scaling, confirming the time complexity of SA.

In measuring convergence time, it is important to realize that the converged low-energy state might not be the global ground state. However, the time to ground-state solution must be bounded from below by the convergence time. Hence, given that convergence times grow exponentially with number of tracks as in Figure 15, we expect time to solution to grow exponentially as well, indicating that the pre-processing methodology did not reduce the QUBO to a polynomial-time problem.

To compare quantum and simulated annealing performance as reported in Results, we use 1000 reads and 1000 sweeps to ensure a lower energy is reached than in the timing analysis, since we wish to compare purity and efficiency rather than gain an exact estimate of time to convergence. No significant difference was observed when increasing the number of reads or sweeps in simulated annealing over the range of track densities we report.

D. Quantum annealing

While numerous classical methods exist to solve QUBO problems, a quantum annealer is in general expected to yield a different path to the solution and may consequently achieve a different exponential scaling (or possibly even polynomial scaling) in comparison to classical approaches. As discussed in the main text, the limitations of the D-Wave hardware make it difficult to characterize the scaling of the proposed quantum algorithm when evaluated on a physical quantum annealer. How-

ever, there is evidence in related work of favorable performance of the quantum annealer beyond the ground state. In particular, the excited states in preparing a quantum machine learning classifier have been seen to offer better generalization than the excited states of simulated annealing [53]. Motivated by such results, future work given access to faster quantum anneal times may compare the performance of quantum and classical methods (in terms of efficiency and purity) when a particular runtime is enforced, either in terms of wall-clock time or the runtime scaling with larger problems.