# Supplementary Information: Hybrid Quantum Learning Under Data Scarcity: A Protocol-Controlled Benchmark for Metal–Insulator Transition Classification

## Supplementary overview

This Supplementary Information provides the complete methodological and diagnostic evidence needed to reproduce and critically evaluate the primary conclusions of the study. It includes detailed dataset and split definitions, model and circuit specifications, complete $\rho$-resolved results, architecture- and bottleneck-matched classical controls, dependence-aware statistical analyses, calibration and screening diagnostics, and exploratory descriptor-based sensitivity analyses. These materials are necessary to verify that the reported comparisons are not driven by representation leakage, test-informed model selection, unmatched model capacity, dependence among repeated evaluations, or selective reporting of favorable data conditions.

The evidence is organized into three distinct evaluation protocols. The grouped, $\rho$-matched graph–quantum benchmark is the primary analysis: the CGCNN is retrained for every fold, seed, and retained-positive fraction, and all classifier heads receive the same fold-local representation. The fixed-encoder analyses are sensitivity diagnostics used to study architecture, imbalance, calibration, and top-k behavior; because their representation protocol differs, their values are not pooled with the primary benchmark. The six-descriptor benchmark removes the graph-representation interface and examines training-fold imbalance treatments, but it remains exploratory because configuration selection uses repeated test-split performance and the classical and quantum search spaces are not fully symmetric.

Dataset and split documentation establishes the evaluated population and possible selection effects. Model, circuit, and training specifications establish reproducibility and fairness of the comparison. Complete $\rho$-resolved results and matched controls test whether the conclusions persist across data-scarcity levels and progressively stricter classical comparisons. Calibration, screening, descriptor, and chemistry-family analyses determine whether ranking performance translates into practical and chemically uniform candidate-screening behavior. Collectively, these materials support conclusions about robustness, architecture sensitivity, and screening tradeoffs; they do not establish a quantum-specific advantage, a causal explanation for circuit-depth behavior, or readiness for deployment on unseen materials.

Dataset levels and split composition are summarized in Tables S1–S2 and Figure S1. Model, circuit, training, and experimental-design details are given in Tables S3–S9 and Figure S2. The aligned benchmark is reported in Tables S10–S11; matched and bottleneck controls are reported in Tables S12–S13; dependence-aware inference, complete inner-validation architecture selection, and the encoder-overlap analysis are reported in Tables S14–S16. Selected full-data metrics and depth frequencies are additionally reported in Tables S25–S26 and Figure S15. The fixed-encoder

protocol comparison and sensitivity controls are reported in Tables S17–S21 and Figures S3–S5. Calibration and screening diagnostics are reported in Tables S22–S23 and Figures S6–S7. The main manuscript presents a compact treatment-level descriptor comparison; the full per-treatment architecture landscapes, operating diagnostics, and metric profile are reported here in Figures S8–S13 and Table S24. The chemistry-family audit is shown in Figure S14.

# S1 Methods

## S1.1 Hybrid graph–quantum classification framework

The workflow is a post-hoc graph–quantum classification pipeline for binary metal–insulator transition (MIT) prediction. In the $\rho$-matched benchmark, MIT-positive training examples are subsampled first at fraction $\rho$, and the CGCNN is then retrained using that same fold-, seed-, and $\rho$-specific training subset. The fold-local CGCNN produces a 64-dimensional embedding vector and a CGCNN MIT probability. Its weights are frozen only during subsequent post-hoc head training. Thus, "frozen" refers to the head stage and does not mean that a single full-data encoder is reused across $\rho$.

The primary comparison consists of the $\rho$-matched CGCNN probability output, linear and MLP heads, and validation-selected VQC and REUP-family heads. Architecture selection is completed independently within every fold, seed, and $\rho$ using validation PR-AUC only. Fixed VQC (4q/2L) and fixed two-qubit/one-layer quantum heads are retained as controlled references; because $L = 1$ encodes the data once, the latter is the single-upload limit of the data re-uploading family. The fixed-encoder VQC, four-qubit/four-layer REUP, and associated controls are reported separately as sensitivity analyses. All quantum calculations use exact simulation. Resource counts for the fixed-encoder comparator circuits are given in Table S4.

Table S1: Summary of dataset levels used in the MIT classification benchmarks.

| **Dataset level** | **Total** | **MIT** | **non-MIT** | **Used for** |
|---|---|---|---|---|
| Curated descriptor dataset | 343 | 68 | 275 | Descriptor-based REUP benchmark |
| Structure-available subset with valid CIF/embedding | 316 | 59 | 257 | CGCNN embedding, VQC, REUP, and classical post-hoc benchmark |
| Excluded from structure-based benchmark | 27 | 9 | 18 | Missing/invalid CIF or failed embedding generation |

The two dataset sizes shown in Table S1 reflect different input requirements. Structure-based experiments require valid CIF files and successful graph embedding generation, whereas descriptor experiments require the six curated physical descriptors. The excluded set contains 9 MIT materials among 27 entries (33.3%), compared with an MIT prevalence of 19.8% in the full dataset. Structure availability is therefore associated with class composition and is treated as a selection effect rather than as label-neutral missingness.

## S1.2 Structure-available embedding subset and grouped nested cross-validation

We used two related but distinct evaluation datasets. The full curated descriptor dataset contains 343 materials, including 68 MIT and 275 non-MIT compounds. This dataset was used for

the descriptor-based re-uploading benchmark described in Section S1.7. For the structure-based graph–quantum benchmark, we retained only materials with valid CIF files and successfully generated CGCNN embeddings. This structure-available subset contains 316 compounds, including 59 MIT and 257 non-MIT materials. Thus, all CGCNN, linear-head, MLP-head, VQC, REUP, calibration, low-data, and top-k screening experiments based on frozen CGCNN embeddings were performed on the 316-compound structure-available subset, whereas the descriptor-only imbalance-treatment benchmark was performed on the full 343-material descriptor dataset.

The full structure-benchmark protocol is summarized in Table S2. To reduce leakage and obtain robust estimates, all structure-based experiments used grouped nested five-fold cross-validation. The grouping key is the material/composition identifier stored in the split file, so related records are kept in the same outer fold rather than being separated across training, validation, and held-out test partitions. In each outer fold, the test set was held out until final evaluation. The remaining compounds were split into training and validation subsets, where validation data were used only for early stopping and threshold tuning. The class distribution is visualized in Fig. S1.

Table S2: Structure-based graph–quantum benchmark dataset and evaluation protocol summary.

| **Item** | **Value** |
|---|---|
| Input representation | $\rho$-matched fold-local CGCNN embedding from the structure-available CIF subset |
| Embedding dimension | 64 |
| Valid compounds | 316 |
| MIT compounds | 59 |
| Non-MIT compounds | 257 |
| MIT prevalence | 0.187 (59 of 316) |
| Cross-validation | Grouped nested 5-fold CV |
| Random seeds | 42–46 |
| Primary metric | PR-AUC |
| Secondary metrics | ROC-AUC, F1, balanced accuracy, MCC, and calibration |
| Low-data fractions $\rho$ | 0.05, 0.10, 0.15, 0.20, 0.25, 0.333, 0.50, 0.75, and 1.00 |

### S1.3 Post-hoc classifier heads

The aligned $\rho$-matched benchmark evaluates five distinct outputs for each fold, seed, and $\rho$: (i) the fold-local CGCNN probability, (ii) a linear/logistic head, (iii) a standard MLP, (iv) a validation-selected VQC head, and (v) a validation-selected member of the REUP family. Fixed VQC (4q/2L) and two-qubit/one-layer single-upload heads are retained as controlled references. All post-hoc heads receive the same fold-, seed-, and $\rho$-specific CGCNN embeddings. The linear and logistic implementations are equivalent in this setup and produced identical predictions.

Additional controls reproduce model capacity or representation bottlenecks. Parameter-matched MLPs are sized to the trainable budgets of VQC and the single-upload head. Projection-bottleneck and latent-bottleneck classical heads reproduce the low-dimensional classical interfaces without a quantum circuit. A random-Fourier-feature (RFF) logistic model provides a nonlinear randomized-feature reference. The aligned model roles are summarized in Table S3; verified resource counts for the fixed-encoder comparator circuits are listed in Table S4.

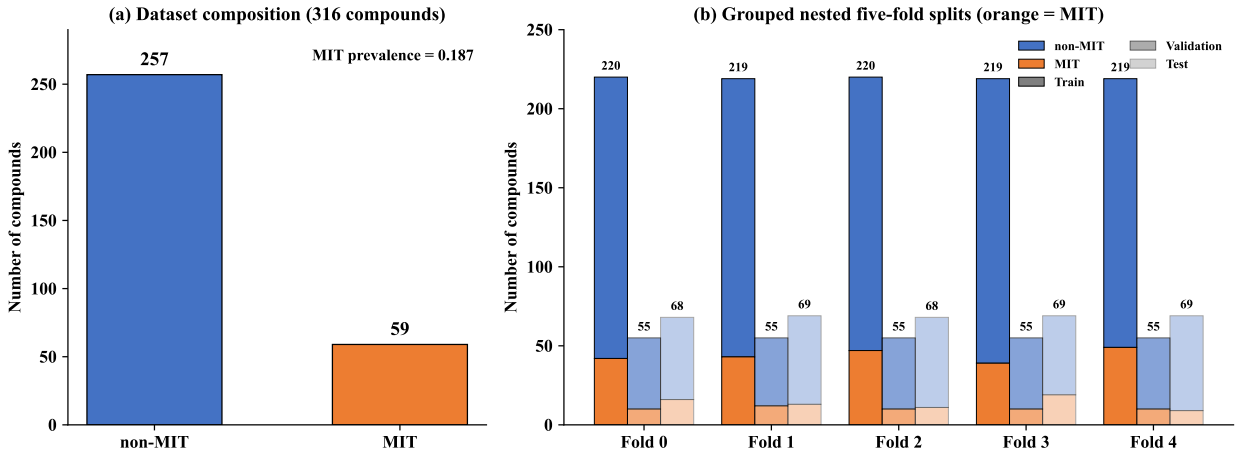


Figure S1: Dataset composition and grouped nested cross-validation splits. The figure shows the MIT/non-MIT class imbalance and fold-level train, validation, and test distributions.

Table S3: Classifier heads and their roles in the benchmark.

| **Model** | **Architecture** | **Role** |
|---|---|---|
| CGCNN ($\rho$-matched) | Fold-local graph model retrained separately at every $\rho$ | Representation source and direct probability reference |
| Linear/logistic | Linear classifier on the 64-dimensional embedding | Strong low-capacity classical baseline |
| Standard MLP | Two-layer nonlinear head | Strong nonlinear classical baseline |
| VQC | Learned projection, angle encoding, entangling circuit, all-$Z$ readout | Quantum reference head |
| Single-upload head (2q/1L) | Learned projection, two qubits, one data-encoding layer, quantum rotations/readout | $L = 1$ limit of the re-uploading family |
| Parameter-matched MLPs | Classical hidden width selected to match each quantum trainable budget | Capacity controls |
| Projection/latent bottlenecks | Classical heads constrained to the corresponding low-dimensional interfaces | Architecture-matched controls |
| RFF logistic | Random Fourier feature map followed by logistic regression | Nonlinear randomized-feature control |

Table S4: Verified trainable-parameter and circuit-resource summary for the fixed-encoder comparator configurations. All quantum results use exact simulation with `default.qubit`. Primary-run parameter counts and projection widths are preserved in the released run configurations.

| **Comparator model** | **Classical** | **Quantum** | **Total** | **Input dim** |
|---|---|---|---|---|
| Linear/logistic | 65 | 0 | 65 | 64 |
| Standard MLP | 2113 | 0 | 2113 | 64 |
| VQC (4q/2L fixed-encoder comparator) | 1113 | 8 | 1121 | 64 |
| REUP (4q/4L fixed-encoder comparator) | 1113 | 48 | 1161 | 64 |

### S1.4 Data re-uploading quantum classifier

Let $\mathbf{x} \in \mathbb{R}^{64}$ denote the frozen CGCNN embedding for a material. A classical projection network maps $\mathbf{x}$ to a low-dimensional latent vector,

$$\mathbf{z} = f_\theta(\mathbf{x}), \tag{1}$$

where $\mathbf{z}$ is converted to rotation angles for the quantum circuit. In the REUP classifier, the data-dependent angles are injected repeatedly across $L$ quantum layers. A single re-uploading layer can be written abstractly as

$$U_\ell(\mathbf{z}, \boldsymbol{\phi}_\ell) = U_{\text{ent}} U_{\text{train}}(\boldsymbol{\phi}_\ell) U_{\text{enc}}(\mathbf{z}), \tag{2}$$

where $U_{\text{enc}}$ encodes the latent data, $U_{\text{train}}$ contains trainable single-qubit rotations, and $U_{\text{ent}}$ applies the entangling operation. The full circuit is then

$$U_{\text{REUP}}(\mathbf{z}, \boldsymbol{\phi}) = \prod_{\ell=1}^{L} U_\ell(\mathbf{z}, \boldsymbol{\phi}_\ell). \tag{3}$$

The circuit output is obtained from Pauli-$Z$ expectation values and passed to a trainable linear readout to produce the MIT logit:

$$\hat{y} = \sigma\left(\mathbf{w}^\top \langle \mathbf{Z} \rangle + b\right), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function. For $L \geq 2$, the data are injected repeatedly and the circuit is a genuine re-uploading model. For $L = 1$, the data are encoded only once; this is the single-upload or no-re-upload limit of the architecture family. The primary structure benchmark uses two qubits and $L = 1$, whereas the fixed-encoder 4q/4L configuration uses four qubits and four re-uploading layers. Figure S2 shows the descriptor-based architecture, in which the physical descriptors are encoded directly as rotation angles.

### S1.5 Training protocol

The common optimization and evaluation settings are summarized in Table S5. Trainable neural heads use Adam [1] with binary cross-entropy loss. Positive-class weights are computed from the training split only. Early stopping uses validation PR-AUC, the best validation checkpoint is restored before test evaluation, and threshold-dependent metrics use a validation-tuned F1 threshold. Ranking metrics are computed directly from predicted probabilities.

For the $\rho$-matched setting, MIT-positive training samples are selected before CGCNN training. The same retained-positive identifiers are used for CGCNN, linear/logistic, MLP, VQC, single-upload, matched, bottleneck, and RFF controls. Validation and test sets remain unchanged. The CGCNN is retrained for every fold, seed, and $\rho$; its embedding is frozen only after that training step. Fixed-encoder controls, imbalance ablations, architecture sweeps, calibration, and top-$k$ analyses are reported separately.

### S1.6 Evaluation metrics

Because the MIT class is imbalanced, PR-AUC was used as the primary metric. ROC-AUC was included to measure global ranking quality, while F1, precision, recall, balanced accuracy, and

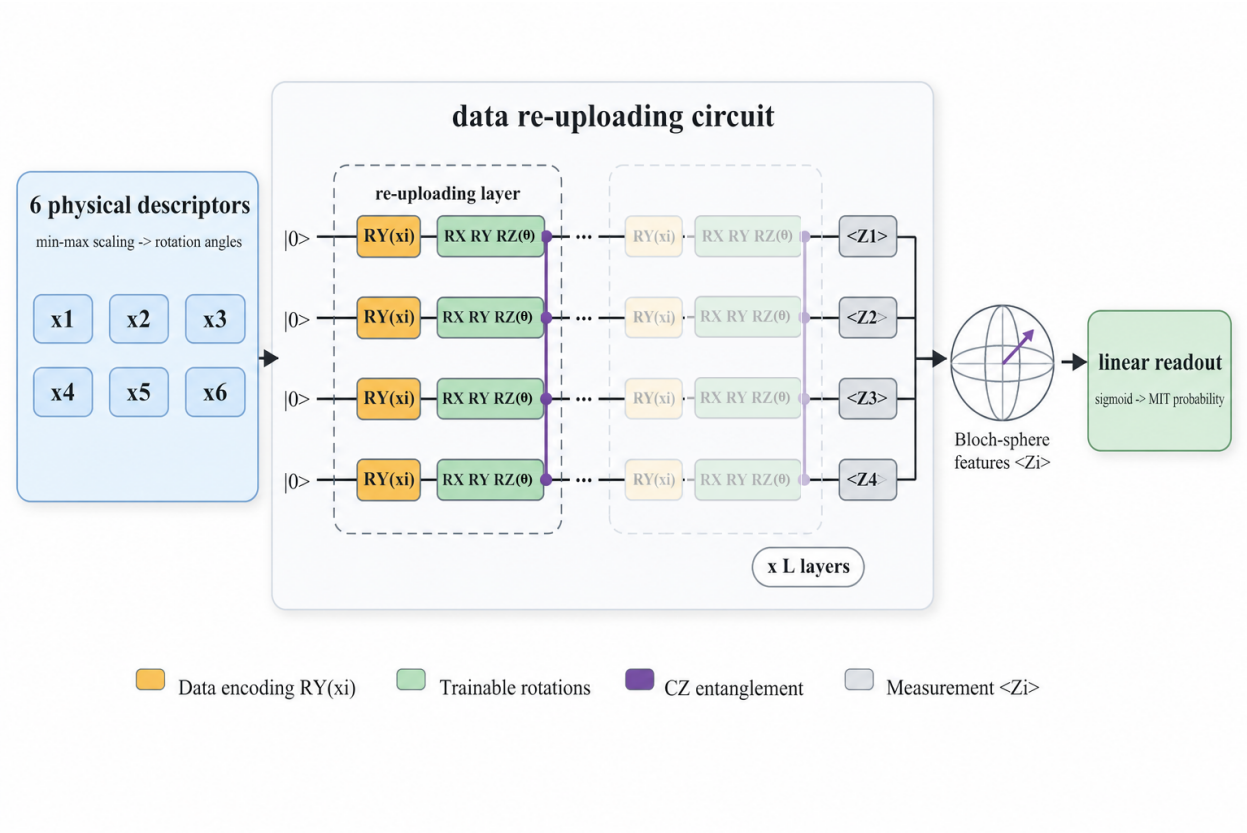


Figure S2: Data re-uploading quantum classifier, drawn for the descriptor-based benchmark. The six physical descriptors (Section S1.7) are min–max scaled into rotation angles $x_1, \ldots, x_6$ and injected through $R_Y$ data-encoding gates; each re-uploading layer re-injects the data, followed by trainable $R_X R_Y R_Z(\theta)$ rotations and an optional CZ entangling ring, and the layer is repeated $L$ times. Pauli-$Z$ expectation values $\langle Z_i \rangle$ form the Bloch-sphere features that a linear layer and sigmoid map to the predicted MIT probability. In the structure-based benchmark, the circuit input is instead the projected CGCNN embedding.

Table S5: Training and evaluation settings for post-hoc classifiers.

| **Setting** | **Value** |
|---|---|
| Neural optimizer | Adam |
| Loss | Weighted binary cross-entropy |
| Learning rate | $10^{-3}$ |
| Weight decay | $10^{-4}$ |
| Batch size | 32 |
| Maximum epochs | 100 |
| Early stopping | Validation PR-AUC |
| Patience | 12 epochs |
| Threshold tuning | Validation F1 |
| Quantum simulator | PennyLane `default.qubit` |
| Aggregation | Mean and 95% CI over fold × seed runs |

Matthews correlation coefficient (MCC) were used to assess thresholded classification [2]. Calibration was evaluated using Brier score [3] and expected calibration error (ECE) [4]. For the materials-discovery setting, top-$k$ precision, recall, and enrichment were also computed to measure how effectively each model ranks candidate MIT compounds near the top of the screening list. Because compounds recur across seeds, fold–seed confidence intervals are descriptive repeated-run summaries rather than intervals from independent experimental units.

### S1.7 Full 343-material descriptor benchmark with imbalance treatments

In addition to the structure-based graph–quantum benchmark, we report an exploratory descriptor-based re-uploading study that removes the learned graph-embedding interface. This second benchmark uses a curated MIT dataset of 343 materials, each described by six physical descriptors and a binary MIT/non-MIT label, with 68 MIT and 275 non-MIT compounds (19.8% MIT prevalence; descriptor row of Table S1). Because the inputs are six interpretable descriptors rather than 64-dimensional learned embeddings, the quantum circuit receives the complete classical representation directly, without an intermediate learned projection. The six descriptors are composition- and structure-derived features in the style of the curated MIT database literature [5]: the volume per atomic site, the composition-averaged covalent radius, the range of Mendeleev numbers of the constituent elements, the average deviation of the elemental ground-state band gap, the minimum elemental electronegativity, and the range of the neighbor-distance variation. The controlled variable in this benchmark is therefore not the representation but the treatment of class imbalance in the training data, a long-standing determinant of classifier behavior in imbalanced learning [6].

Table S6: Descriptor-benchmark split statistics. Stratified shuffle-split cross-validation with a 70/30 train–test ratio fixes the class composition of every split; the counts below therefore hold identically for all five splits of all five seeds (42–46).

| **Partition** | **Total** | **MIT** | **non-MIT** | **MIT prevalence** |
|---|---|---|---|---|
| Training fold | 240 | 48 | 192 | 0.200 |
| Test fold | 103 | 20 | 83 | 0.194 |

All descriptor-based experiments use stratified five-fold shuffle-split cross-validation with a 70/30 train–test ratio—each split is partitioned into a 70% training fold and a 30% test fold with no separate held-out validation fold, in contrast to the nested protocol of the structure-based benchmark—and five random seeds (42–46), matching the seeds used in the structure-based benchmark, so that the four imbalance settings are compared on identical splits and all descriptor metrics are reported as mean and 95% confidence interval across seeds. Stratification fixes the composition of every split: each training fold contains 240 materials (48 MIT) and each test fold 103 materials (20 MIT), identically across all splits and seeds (Table S6). The four settings differ only in how the training folds are treated after splitting: (i) the unbalanced distribution is kept unchanged; (ii) the minority MIT class is balanced by random oversampling; (iii) synthetic minority samples are generated by SMOTE; and (iv) SMOTE–Tomek combines synthetic oversampling with Tomek-link cleaning of ambiguous majority–minority pairs near the class boundary [7, 8, 9, 10]. In all four settings the test folds keep the original class distribution, so duplicated or synthetic samples can never leak into evaluation.

For each imbalance setting, a grid of 72 re-uploading configurations is trained with the derivative-free COBYLA optimizer [11] for 120 iterations, giving 288 configurations per seed. COBYLA is

used here, rather than the gradient-based Adam optimizer applied to the structure-based heads, because the descriptor circuits are optimized directly as parameterized quantum circuits without a classical automatic-differentiation front-end, and a derivative-free optimizer is a standard choice in that setting. The grid varies the number of qubits (1, 2, 4), the number of re-uploading layers (2, 4, 6, 8), the feature-scaling range ($[0, 1]$ or $[0, 2\pi]$, fitted on the training fold only), the re-uploading form (symmetrical or asymmetrical feature-to-qubit assignment), and the presence of ring entanglement; one-qubit circuits admit only the symmetrical, non-entangled form. Because the test folds remain imbalanced, the best configuration of each setting is selected by mean F1, with PR-AUC and balanced accuracy as complementary criteria [12, 13, 14]. Because the best configuration is selected by repeated test-split performance from a 72-point grid, the reported best-configuration metrics are optimistically biased. Averaging over seeds and reporting confidence intervals does not remove this selection bias. The descriptor results are therefore used only as exploratory sensitivity analyses; a confirmatory comparison requires nested model selection, a composition-relatedness audit or grouped splitting, optimizer-convergence evidence, and classical controls under matched resampling treatments.

To place the quantum results in context, three standard classical baselines—logistic regression, random forest, and XGBoost—are evaluated on the same descriptors, splits, seeds, and metrics, each tuned with a small per-model grid selected by the same mean-F1 criterion: logistic regression over regularization strength $C \in \{0.1, 1, 10\}$, class weighting $\{\text{none}, \text{balanced}\}$, feature scaling $\{\text{standard}, [0, 1]\}$, and an optional SMOTE training-fold resampling (24 configurations); random forest over $\{100, 300\}$ trees and maximum depth $\{\text{none}, 4\}$ with balanced class weights (8 configurations); and XGBoost over $\{75, 150\}$ estimators and maximum depth $\{2, 3\}$ with automatic positive-class weighting (8 configurations). Importantly, on the imbalance-treatment datasets the classical models handle class imbalance through class re-weighting (`class_weight` or `scale_pos_weight`) rather than through the fold-level random-oversampling, SMOTE, and SMOTE–Tomek resampling applied to the quantum circuits; in every setting the best classical model is a balanced-weight random forest, so the classical reference is effectively common across the four treatments and is reported with its complete metric profile in Table S24. For the operating-characteristic and calibration diagnostics of Section S3.8, the winning configuration of each treatment and the winning random forest were re-fitted once more on the identical splits with per-sample test scores retained; these re-runs reproduce the stored per-seed metrics exactly and add no new model selection. The results of this benchmark are reported in Sections S3.3–S3.8.

# S2 Experimental Design

## S2.1 Evaluation objective

The primary objective is to compare quantum and classical heads under the same low-data representation protocol. For every fold, seed, and retained MIT-positive fraction $\rho$, the CGCNN is retrained using the same retained-positive identifiers as the post-hoc heads. Strong linear and nonlinear classical baselines, capacity controls, bottleneck-matched controls, validation-only architecture selection, and dependence-aware paired inference are used to test whether any apparent quantum gain survives progressively stricter comparisons.

### S2.2 Primary $\rho$-matched low-data benchmark

Table S7 summarizes the aligned benchmark design. The benchmark evaluates

$$\rho \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.333, 0.50, 0.75, 1.00\}.$$

For each $\rho$, MIT-positive training samples are selected first. The CGCNN and all post-hoc heads then use that same training-label budget, while validation and test folds remain fixed.

Table S7: Summary of the $\rho$-matched primary benchmark.

| **Item** | **Setting** |
|---|---|
| Quantum models | VQC and a fixed 2-qubit/1-layer single-upload head |
| Classical baselines | CGCNN probability, linear/logistic, and standard MLP |
| Additional controls | Parameter-matched MLPs, projection/latent bottlenecks, and RFF logistic |
| Training fractions $\rho$ | 0.05–1.00 |
| Representation training | CGCNN retrained for each fold, seed, and $\rho$ |
| Cross-validation | Grouped nested 5-fold CV |
| Random seeds | 42–46 |
| Primary metric | PR-AUC |
| Threshold metric | Validation-tuned F1 |
| Simulation | Exact state vector |

### S2.3 Matched and bottleneck controls

Parameter-matched MLPs test whether differences can be explained by trainable-parameter budget. Projection-bottleneck and latent-bottleneck heads reproduce the low-dimensional classical interfaces without a quantum circuit, providing a stricter test of whether the circuit contributes beyond the learned projection. RFF logistic regression supplies a nonlinear randomized-feature comparison. Every control uses the same fold-local embedding and outer-test samples as the corresponding quantum head.

### S2.4 Class-imbalance ablation

The training-fold interventions are defined in Table S8. Because MIT-positive examples are scarce, additional runs evaluate whether REUP degradation is mainly caused by class imbalance. The following training variants are compared: original positive subsampling with all negatives retained, matched-negative sampling, preserved-ratio sampling, positive oversampling, and focal loss. Validation and test sets are never resampled. These experiments test whether imbalance-aware training can recover REUP performance or whether the limitation persists across sampling strategies.

### S2.5 Quantum architecture ablation and validation-only selection

The broader architecture variables are summarized in Table S9. The complete validation-selection experiment uses a controlled stage-1 subset consisting of qubit counts $\{2, 4, 6, 8\}$ and depths $\{1, 2, 4, 8\}$

Table S8: Class-imbalance ablation settings.

| Mode | Description |
|---|---|
| Unbalanced | Subsample MIT positives; retain all negatives |
| Matched negatives | Match negative count to selected positives |
| Preserved ratio | Subsample while preserving class ratio |
| Oversampling | Oversample selected MIT positives during training |
| Focal loss | Use focal loss instead of weighted BCE |

with ring entanglement, RY encoding, and all-$Z$ measurement fixed, giving 16 candidate architectures per head. The $L = 1$ configuration is treated explicitly as a single-upload control rather than as repeated re-uploading. For each head, $\rho$, seed, and fold, the architecture with maximum validation PR-AUC is selected without using the outer-test metric; the corresponding outer-test score is then reported once. Selection covers all 5 folds, 5 seeds, 9 values of $\rho$, and 2 heads, giving 450 selected cells.

Table S9: Quantum architecture variables explored in the QML ablation study.

| Variable | Values | Purpose |
|---|---|---|
| Number of qubits | 2, 4, 6, 8 | Tests quantum latent dimension |
| Circuit layers | 1, 2, 4, 8 | Tests circuit depth and re-uploading |
| Encoding gates | RY, RX+RY, RX+RY+RZ | Tests data-encoding richness |
| Entanglement | None, linear, ring, full | Tests correlation structure |
| Measurement | Single-$Z$, all-$Z$, readout layer | Tests output bottlenecks |

The ranges balance expressivity, trainability, and exact-simulation cost. The broader table documents the full sensitivity design, whereas the locked primary selection uses the controlled 4-by-4 qubit–depth grid described above. Architecture-selected and fixed-architecture estimates are reported separately because they answer different questions: fixed estimates characterize a locked model, whereas validation-selected estimates quantify model-selection performance without test-informed choice.

### S2.6 Embedding, calibration, and screening analyses

Additional analyses evaluate whether classifier behavior depends on representation compression, probability calibration, or screening utility. Representation-compression effects are probed by the complementary descriptor benchmark (Section S1.7), which removes the learned projection between the material representation and the circuit entirely. Calibration is evaluated using Brier score and expected calibration error. Screening utility is measured using top-$k$ precision, recall, and enrichment, which are relevant for ranking candidate MIT compounds for future validation.

### S2.7 Statistical analysis

PR-AUC is the primary endpoint because the MIT class is underrepresented. Per-model intervals summarize repeated fold–seed evaluations. Paired $\rho = 1$ comparisons use aligned predictions, a

fold-blocked cluster bootstrap for 95% confidence intervals, and fold-blocked sign-flip permutation tests. Blocking by outer fold preserves dependence among seeds that reuse the same held-out material groups. Thresholded metrics use validation-tuned predictions. Because only five outer-fold blocks are available, the blocked permutation test has limited resolution; confidence-interval exclusion of zero is therefore not treated as interchangeable with a two-sided $p < 0.05$ result.

# S3 Results

## S3.1 $\rho$-matched benchmark results

The complete aligned report contains 2475 model–$\rho$–fold–seed records across the main baselines and matched controls. The main benchmark contains 25 evaluations per model and $\rho$ (5 outer folds $\times$ 5 seeds). The independent validation-selection report adds 450 selected cells (5 folds $\times$ 5 seeds $\times$ 9 fractions $\times$ 2 quantum-head families).

Table S10: PR-AUC across retained MIT-positive training fractions for the aligned classical baselines and controlled fixed quantum architectures. Values are mean [95% CI] over 25 fold–seed evaluations. Linear and logistic heads are identical in this implementation.

| $\rho$ | CGCNN | Linear/logistic | MLP | VQC | Single-upload (2q/1L) |
|---|---|---|---|---|---|
| 0.05 | 0.409 [0.342,0.477] | 0.812 [0.749,0.875] | 0.847 [0.796,0.898] | 0.851 [0.808,0.894] | 0.753 [0.668,0.837] |
| 0.10 | 0.495 [0.437,0.552] | 0.827 [0.777,0.877] | 0.865 [0.821,0.909] | 0.876 [0.838,0.914] | 0.760 [0.686,0.834] |
| 0.15 | 0.525 [0.459,0.591] | 0.850 [0.799,0.902] | 0.860 [0.815,0.906] | 0.883 [0.846,0.920] | 0.815 [0.755,0.874] |
| 0.20 | 0.531 [0.464,0.598] | 0.801 [0.735,0.868] | 0.861 [0.815,0.907] | 0.897 [0.871,0.924] | 0.793 [0.726,0.859] |
| 0.25 | 0.532 [0.466,0.597] | 0.828 [0.775,0.882] | 0.854 [0.807,0.902] | 0.872 [0.828,0.915] | 0.787 [0.736,0.838] |
| 0.333 | 0.596 [0.544,0.648] | 0.839 [0.789,0.888] | 0.870 [0.828,0.911] | 0.863 [0.814,0.913] | 0.839 [0.782,0.895] |
| 0.50 | 0.583 [0.524,0.642] | 0.849 [0.798,0.899] | 0.868 [0.827,0.909] | 0.891 [0.851,0.931] | 0.801 [0.748,0.854] |
| 0.75 | 0.578 [0.517,0.639] | 0.851 [0.805,0.896] | 0.852 [0.800,0.904] | 0.903 [0.873,0.933] | 0.860 [0.809,0.910] |
| 1.00 | 0.571 [0.517,0.625] | 0.869 [0.820,0.918] | 0.861 [0.812,0.910] | 0.894 [0.856,0.931] | 0.860 [0.817,0.903] |

Table S11: Full metrics at $\rho = 1$ for the aligned classical baselines and controlled fixed quantum architectures. Values are mean [95% CI] over 25 fold–seed evaluations. Brier score and ECE are lower-is-better.

| Model | PR-AUC | ROC-AUC | Bal. acc. | F1 | Precision | Recall | MCC | Brier | ECE |
|---|---|---|---|---|---|---|---|---|---|
| CGCNN | 0.571 [0.517,0.625] | 0.822 [0.801,0.842] | 0.717 [0.681,0.754] | 0.503 [0.454,0.551] | 0.597 [0.501,0.693] | 0.551 [0.456,0.645] | 0.441 [0.391,0.490] | 0.166 [0.151,0.180] | 0.174 [0.151,0.198] |
| Linear/logistic | 0.869 [0.820,0.918] | 0.968 [0.958,0.977] | 0.888 [0.861,0.915] | 0.779 [0.738,0.820] | 0.768 [0.704,0.831] | 0.837 [0.777,0.898] | 0.745 [0.700,0.791] | 0.106 [0.094,0.119] | 0.244 [0.226,0.262] |
| MLP | 0.861 [0.812,0.910] | 0.970 [0.963,0.977] | 0.897 [0.875,0.919] | 0.808 [0.771,0.846] | 0.807 [0.746,0.869] | 0.848 [0.797,0.899] | 0.776 [0.736,0.817] | 0.104 [0.082,0.125] | 0.223 [0.182,0.264] |
| VQC | 0.894 [0.856,0.931] | 0.969 [0.960,0.979] | 0.892 [0.864,0.919] | 0.796 [0.754,0.839] | 0.798 [0.734,0.862] | 0.836 [0.774,0.897] | 0.766 [0.720,0.813] | 0.183 [0.167,0.200] | 0.352 [0.334,0.371] |
| Single-upload (2q/1L) | 0.860 [0.817,0.903] | 0.960 [0.951,0.969] | 0.857 [0.826,0.887] | 0.743 [0.697,0.789] | 0.762 [0.691,0.833] | 0.778 [0.708,0.847] | 0.705 [0.654,0.755] | 0.130 [0.115,0.145] | 0.257 [0.238,0.277] |

The aligned fixed-reference results show that VQC has the highest mean full-data PR-AUC, but linear, MLP, VQC, and the single-upload head have overlapping repeated-run intervals. Projection-bottleneck controls match the single-upload model, and fold-blocked paired tests do not support a quantum-specific improvement over the strong classical heads. Table S15 reports the complete validation-selected experiment. At full data, selected VQC reaches 0.870 and selected REUP reaches 0.811; across all fractions, one-layer circuits account for 82.7% of REUP selections. Complete selected metrics and depth-frequency summaries are given in Tables S25–S26 and Figure S15.

Table S12: Matched, bottleneck, and RFF controls across $\rho$. Values are PR-AUC mean [95% CI] over 25 fold–seed evaluations.

| $\rho$ | Single-upload | MLP (REUP-param.) | Projection bottleneck | Latent bottleneck | MLP (VQC-param.) | RFF logistic | VQC |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.753 [0.668,0.837] | 0.555 [0.424,0.685] | 0.818 [0.746,0.890] | 0.836 [0.788,0.885] | 0.588 [0.465,0.710] | 0.611 [0.553,0.669] | 0.851 [0.808,0.894] |
| 0.10 | 0.760 [0.686,0.834] | 0.603 [0.471,0.736] | 0.832 [0.778,0.885] | 0.854 [0.795,0.913] | 0.706 [0.597,0.816] | 0.654 [0.593,0.714] | 0.876 [0.838,0.914] |
| 0.15 | 0.815 [0.755,0.874] | 0.691 [0.582,0.799] | 0.836 [0.786,0.886] | 0.842 [0.790,0.894] | 0.662 [0.539,0.784] | 0.704 [0.660,0.748] | 0.883 [0.846,0.920] |
| 0.20 | 0.793 [0.726,0.859] | 0.616 [0.486,0.746] | 0.854 [0.808,0.901] | 0.872 [0.832,0.911] | 0.562 [0.423,0.701] | 0.703 [0.652,0.754] | 0.897 [0.871,0.924] |
| 0.25 | 0.787 [0.736,0.838] | 0.771 [0.677,0.864] | 0.870 [0.828,0.911] | 0.855 [0.812,0.898] | 0.706 [0.598,0.815] | 0.727 [0.675,0.779] | 0.872 [0.828,0.915] |
| 0.333 | 0.839 [0.782,0.895] | 0.687 [0.560,0.815] | 0.860 [0.813,0.908] | 0.839 [0.786,0.892] | 0.639 [0.514,0.763] | 0.775 [0.731,0.820] | 0.863 [0.814,0.913] |
| 0.50 | 0.801 [0.748,0.854] | 0.709 [0.596,0.821] | 0.842 [0.788,0.896] | 0.842 [0.793,0.892] | 0.684 [0.566,0.803] | 0.805 [0.773,0.837] | 0.891 [0.851,0.931] |
| 0.75 | 0.860 [0.809,0.910] | 0.601 [0.464,0.738] | 0.874 [0.830,0.918] | 0.849 [0.796,0.901] | 0.854 [0.787,0.922] | 0.833 [0.807,0.860] | 0.903 [0.873,0.933] |
| 1.00 | 0.860 [0.817,0.903] | 0.694 [0.576,0.811] | 0.861 [0.814,0.908] | 0.838 [0.784,0.892] | 0.723 [0.608,0.837] | 0.841 [0.813,0.868] | 0.894 [0.856,0.931] |

Table S13: Full metrics for matched and bottleneck controls at $\rho = 1$. Values are mean [95% CI] over 25 fold–seed evaluations.

| Model | PR-AUC | ROC-AUC | F1 | MCC | Brier |
|---|---|---|---|---|---|
| Single-upload (2q/1L) | 0.860 [0.817,0.903] | 0.960 [0.951,0.969] | 0.743 [0.697,0.789] | 0.705 [0.654,0.755] | 0.130 [0.115,0.145] |
| MLP (REUP-param.) | 0.694 [0.576,0.811] | 0.843 [0.753,0.933] | 0.688 [0.602,0.774] | 0.584 [0.449,0.720] | 0.207 [0.186,0.228] |
| Projection bottleneck | 0.861 [0.814,0.908] | 0.967 [0.958,0.976] | 0.778 [0.741,0.815] | 0.746 [0.708,0.785] | 0.068 [0.057,0.079] |
| Latent bottleneck | 0.838 [0.784,0.892] | 0.959 [0.946,0.973] | 0.795 [0.749,0.841] | 0.760 [0.707,0.813] | 0.156 [0.132,0.181] |
| MLP (VQC-param.) | 0.723 [0.608,0.837] | 0.869 [0.795,0.942] | 0.696 [0.602,0.790] | 0.611 [0.479,0.744] | 0.200 [0.179,0.221] |
| RFF logistic | 0.841 [0.813,0.868] | 0.924 [0.908,0.940] | 0.710 [0.670,0.750] | 0.674 [0.631,0.717] | 0.138 [0.126,0.150] |
| VQC | 0.894 [0.856,0.931] | 0.969 [0.960,0.979] | 0.796 [0.754,0.839] | 0.766 [0.720,0.813] | 0.183 [0.167,0.200] |

Table S14: Dependence-aware paired PR-AUC comparisons for the controlled fixed quantum architectures at $\rho = 1$. Differences are A minus B. Confidence intervals use a fold-blocked cluster bootstrap and $p$ values use fold-blocked sign-flip permutation tests.

| A | B | Mean difference | Fold-blocked 95% CI | $p$ |
|---|---|---|---|---|
| VQC | CGCNN | +0.322 | [0.252,0.367] | 0.058 |
| VQC | MLP | +0.033 | [−0.047,0.120] | 0.630 |
| VQC | Linear | +0.024 | [−0.047,0.088] | 0.572 |
| Single-upload | VQC | −0.034 | [−0.061,−0.011] | 0.058 |
| Single-upload | MLP | −0.001 | [−0.082,0.097] | 1.000 |
| Single-upload | CGCNN | +0.289 | [0.225,0.331] | 0.058 |
| Single-upload | REUP-param. MLP | +0.166 | [0.092,0.244] | 0.058 |
| Single-upload | Projection bottleneck | −0.001 | [−0.069,0.068] | 1.000 |
| Single-upload | Latent bottleneck | +0.022 | [−0.021,0.076] | 0.631 |
| VQC | VQC-param. MLP | +0.171 | [0.070,0.286] | 0.058 |
| VQC | Projection bottleneck | +0.033 | [−0.030,0.093] | 0.313 |
| VQC | Latent bottleneck | +0.056 | [0.005,0.104] | 0.186 |

Table S15: Complete inner-validation architecture selection. For each head, $\rho$, seed, and fold, the configuration with maximum validation PR-AUC is selected without test access. Test PR-AUC is reported only for that selected configuration. Each row contains 25 fold–seed cells. Fixed references are VQC 4q/2L and REUP 2q/1L.

| Head | $\rho$ | $n$ | Selected test PR-AUC | Fixed reference | $\Delta$ | Most selected architecture | Count | Unique |
|---|---|---|---|---|---|---|---|---|
| VQC | 0.05 | 25 | 0.841 [0.796,0.886] | 0.851 [0.808,0.894] | −0.010 | 2q/8L | 5/25 | 11 |
| VQC | 0.10 | 25 | 0.886 [0.856,0.915] | 0.876 [0.838,0.914] | +0.010 | 2q/2L | 5/25 | 11 |
| VQC | 0.15 | 25 | 0.868 [0.825,0.911] | 0.883 [0.846,0.920] | −0.015 | 2q/2L | 5/25 | 12 |
| VQC | 0.20 | 25 | 0.875 [0.845,0.905] | 0.897 [0.871,0.924] | −0.022 | 2q/1L | 7/25 | 11 |
| VQC | 0.25 | 25 | 0.883 [0.850,0.916] | 0.872 [0.828,0.915] | +0.011 | 2q/1L | 6/25 | 10 |
| VQC | 0.333 | 25 | 0.890 [0.862,0.919] | 0.863 [0.814,0.913] | +0.027 | 6q/1L | 6/25 | 9 |
| VQC | 0.50 | 25 | 0.878 [0.845,0.911] | 0.891 [0.851,0.931] | −0.013 | 2q/2L or 6q/1L | 4/25 | 11 |
| VQC | 0.75 | 25 | 0.875 [0.831,0.918] | 0.903 [0.873,0.933] | −0.028 | 2q/1L or 6q/1L | 5/25 | 9 |
| VQC | 1.00 | 25 | 0.870 [0.829,0.910] | 0.894 [0.856,0.931] | −0.024 | 2q/1L, 2q/2L, or 2q/4L | 4/25 | 12 |
| REUP | 0.05 | 25 | 0.837 [0.802,0.873] | 0.753 [0.668,0.837] | +0.085 | 2q/1L | 9/25 | 7 |
| REUP | 0.10 | 25 | 0.761 [0.699,0.823] | 0.760 [0.686,0.834] | +0.001 | 2q/1L | 9/25 | 6 |
| REUP | 0.15 | 25 | 0.834 [0.793,0.875] | 0.815 [0.755,0.874] | +0.019 | 2q/1L | 10/25 | 6 |
| REUP | 0.20 | 25 | 0.798 [0.739,0.858] | 0.793 [0.726,0.859] | +0.006 | 2q/1L, 4q/1L, or 8q/1L | 6/25 | 6 |
| REUP | 0.25 | 25 | 0.795 [0.750,0.840] | 0.787 [0.736,0.838] | +0.008 | 2q/1L | 8/25 | 6 |
| REUP | 0.333 | 25 | 0.832 [0.781,0.883] | 0.839 [0.782,0.895] | −0.007 | 2q/1L | 9/25 | 4 |
| REUP | 0.50 | 25 | 0.801 [0.748,0.855] | 0.801 [0.748,0.854] | −0.000 | 2q/1L | 12/25 | 7 |
| REUP | 0.75 | 25 | 0.830 [0.775,0.884] | 0.860 [0.809,0.910] | −0.030 | 2q/1L | 9/25 | 7 |
| REUP | 1.00 | 25 | 0.811 [0.765,0.858] | 0.860 [0.817,0.903] | −0.049 | 2q/1L | 8/25 | 7 |

Table S16: Encoder-protocol comparison for the CGCNN probability output. Overlap is measured against the upstream encoder's training collection.

| Quantity | Value |
|---|---|
| Upstream fixed-encoder PR-AUC ($\rho = 1$) | 0.876 |
| Nested fold-local CGCNN PR-AUC ($\rho = 1$) | 0.571 |
| Difference (nested − fixed) | −0.305 |
| Formula overlap with upstream training collection | 81.3% |
| Fingerprint overlap with upstream training collection | 53.5% |
| Missing or failed CIFs | 27 |
| Nested mean training MIT positives | 39 |
| Nested mean validation–test PR-AUC gap | 0.177 |

Table S17: Fixed-encoder and $\rho$-matched full-data PR-AUC values. Both the encoder protocol and quantum architecture differ, so the rows diagnose protocol sensitivity rather than a single controlled intervention.

| Model | Fixed-encoder | $\rho$-matched | Protocol distinction |
|---|---|---|---|
| Quantum head | 0.355 [0.294,0.416] | 0.860 [0.817,0.903] | 4q/4L repeated re-uploading versus fixed 2q/1L single-upload |
| CGCNN output | 0.876 [0.828,0.925] | 0.571 [0.517,0.625] | Fixed full-data probability reference versus $\rho$-matched retraining |

## S3.2 Fixed-encoder circuit-sensitivity benchmark

The following subsections report the fixed-encoder sensitivity results. They are not combined directly with Tables S10–S11 because the encoder-training protocol differs.

### S3.2.1 Overall performance across low-data regimes

Under the fixed-encoder protocol, the frozen CGCNN probability reference remains nearly constant across $\rho$ because the encoder is not retrained for the retained-positive fraction. The classical post-hoc heads remain competitive, VQC improves with more positive head-training examples, and the fixed 4q/4L REUP remains weak. These curves measure head-only sample efficiency on a representation trained with the full fold-level label budget.

At $\rho = 1.0$, VQC obtains the highest mean PR-AUC of 0.8987, compared with 0.8764 for CGCNN, 0.8692 for the linear head, and 0.8609 for the MLP head. However, CGCNN remains strongest in ROC-AUC, balanced accuracy, Brier score, and ECE, indicating better ranking stability and probability calibration. REUP reaches only 0.3553 PR-AUC and 0.2653 F1 at full data, showing that repeated data encoding does not automatically translate into effective MIT classification in the present graph-embedding setting.

### S3.2.2 Paired statistical comparisons

The paired outcomes are reported in Table S18. Paired tests were performed on aligned fixed-encoder fold–seed predictions. The VQC–CGCNN PR-AUC difference is +0.0223 with a 95% confidence interval of $[-0.0383, 0.0783]$ and permutation $p = 0.469$. VQC is therefore described as competitive rather than superior in the fixed-encoder protocol.

The REUP result is more decisive. At $\rho = 1.0$, REUP is significantly worse than VQC, MLP, and the REUP-matched classical baseline. The paired PR-AUC gap between REUP and VQC is -0.5434 with permutation $p = 0.0020$, confirming that the REUP underperformance is robust rather than a fold- or seed-specific artifact.

Table S18: **Paired statistical comparisons of PR-AUC differences at $\rho = 1$.** Mean difference (A − B) with 95% bootstrap CI and two-sided permutation $p$-value. Comparisons are matched over 25 fold–seed units. McNemar test uses tuned-threshold binary predictions where available. VQC should be interpreted as competitive rather than superior unless the paired test is significant ($p < 0.05$).

| Comparison | $N$ | Mean $\Delta$ | $CI_{2.5}$ | $CI_{97.5}$ | $p$ (perm.) | $p$ (McNemar) |
|---|---|---|---|---|---|---|
| VQC − CGCNN | 25 | +0.0223 | -0.0383 | +0.0783 | 0.469 | 1.000 |
| VQC − Linear | 25 | +0.0295 | -0.0325 | +0.0884 | 0.383 | 0.073 |
| VQC − MLP | 25 | +0.0378 | -0.0236 | +0.0935 | 0.242 | 0.762 |
| REUP − VQC | 25 | -0.5434 | -0.5993 | -0.4884 | 0.000* | 0.000 |
| REUP − MLP | 25 | -0.5056 | -0.5747 | -0.4389 | 0.000* | 0.000 |

* $p < 0.05$

### S3.2.3 Parameter-matched controls (fixed-encoder protocol)

The full-data matched-capacity results are summarized in Table S19. Parameter-matched classical controls were used to test whether the quantum heads behave differently from classical models of comparable trainable capacity. This comparison is especially important for REUP because a weak result could otherwise be attributed only to limited model size.

Within the fixed-encoder protocol, parameter count alone does not explain the poor 4q/4L result. At full data, the REUP classifier obtains a PR-AUC of 0.3553, while the parameter-matched MLP control reaches 0.8670. This gap narrows the plausible explanations but does not independently localize failure to compression, circuit optimization, or readout. In contrast, VQC remains stronger than its parameter-matched MLP control over most values of $\rho$, suggesting that the simpler quantum circuit provides a more effective inductive interface for the CGCNN embeddings than the present REUP design.

Table S19: **Parameter-matched classical controls at $\rho = 1$.** VQC and REUP are compared against latent MLPs with approximately the same total trainable parameters. All models use the same frozen CGCNN embeddings, splits, and training protocol. Means ± SD over 25 fold–seed units. PR-AUC is the primary metric.

| Model | PR-AUC | ROC-AUC | F1 | Bal. Acc. |
|---|---|---|---|---|
| *VQC group* | | | | |
| VQC (1121 params) | 0.899±0.101 | 0.964±0.029 | 0.799±0.097 | 0.887±0.063 |
| $\text{MLP}_{\text{VQC}}$ (matched) | 0.723±0.289 | 0.869±0.186 | 0.696±0.238 | 0.809±0.167 |
| *REUP group* | | | | |
| REUP (1161 params) | 0.355±0.156 | 0.603±0.147 | 0.265±0.152 | 0.557±0.081 |
| $\text{MLP}_{\text{REUP}}$ (matched) | 0.867±0.119 | 0.967±0.022 | 0.785±0.106 | 0.881±0.070 |

### S3.2.4 Effect of class imbalance (fixed-encoder protocol)

Because MIT-positive examples are scarce, class imbalance was evaluated as a possible cause of REUP degradation. Figure S3 compares the original training protocol with matched-negative sampling, positive oversampling, focal loss, and preserved-ratio sampling. These interventions do not rescue REUP performance. At $\rho = 1.0$, REUP F1 remains low across all imbalance treatments, ranging from 0.2653 under the original setting to 0.3260 with focal loss. This suggests that class imbalance contributes to the difficulty of the problem but is not the primary reason for REUP failure. The PR-AUC under five imbalanced treatments is summarized in Table S20.

### S3.2.5 Quantum architecture sensitivity

Quantum-circuit ablations were performed to determine whether the poor 4q/4L REUP result reflects an inherent limitation of data re-uploading or a sensitivity to circuit design. The default REUP configuration used four qubits and four re-uploading layers, giving a full-data PR-AUC of 0.3553 (Figure S4). However, the architecture sweep reveals that REUP performance is strongly dependent on circuit size and depth. The best configuration in the sweep used two qubits and one data-encoding layer and achieved a PR-AUC of 0.851. Because $L = 1$ contains no repeated data injection, it is the single-upload limit rather than a genuine re-uploading model.

Table S20: **PR-AUC under five imbalance treatments at $\rho = 1$ (316-compound structure-available benchmark).** Values are mean ± SD over 25 fold–seed units. Treatments are applied to the training fold only. Note: the imbalance treatments here apply to the CGCNN-embedding benchmark, not the 343-material descriptor dataset; see Figure S3 for the visualisation.

| Model | Original | Focal loss | Matched neg. | Oversample | Preserve ratio |
|---|---|---|---|---|---|
| Linear | 0.869±0.125 | 0.827±0.151 | 0.838±0.144 | 0.793±0.170 | 0.869±0.125 |
| MLP | 0.861±0.125 | 0.865±0.125 | 0.871±0.118 | 0.860±0.114 | 0.861±0.125 |
| VQC | 0.899±0.101 | 0.833±0.180 | 0.800±0.149 | 0.832±0.142 | 0.899±0.101 |
| REUP | 0.355±0.156 | 0.324±0.163 | 0.302±0.122 | 0.329±0.154 | 0.355±0.156 |

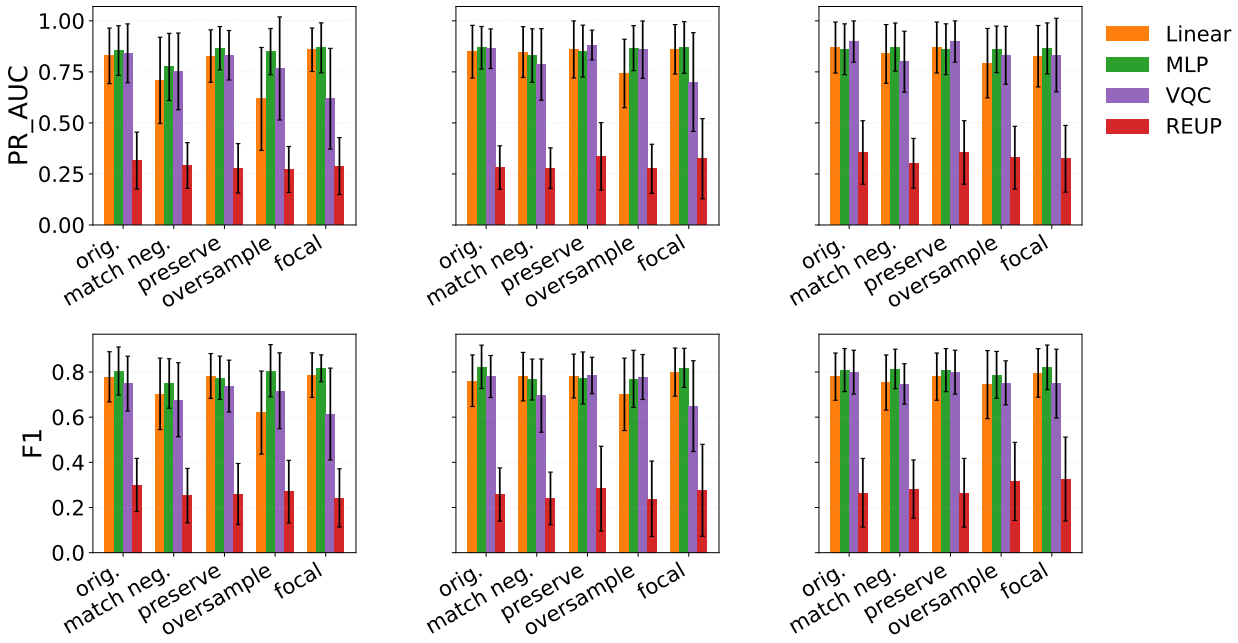


Figure S3: Class-imbalance ablation for classical and quantum heads. Imbalance-aware training changes REUP performance slightly but does not close the gap to VQC or classical baselines.

The sweep changes the interpretation of the fixed-encoder negative result. Performance is highly architecture-sensitive, and the strongest observed configuration removes repeated re-uploading entirely. Deeper circuits may be harder to optimize or generalize, but the present sweep does not independently identify the causal mechanism. Because the configuration was selected post hoc, the 0.851 value is exploratory. It motivated the fixed 2q/1L follow-up reported in Table S11, but it is not an independently selected architecture.

For comparison, the VQC architecture sweep shows several competitive configurations, with the best VQC reaching a PR-AUC of 0.924. Thus, the architecture ablations indicate that the limitation is not the use of quantum heads in general, but the sensitivity of the REUP design to depth, qubit count, and representation interface. The findings are summarized in Table S21.

Table S21: **Best architecture configurations from the quantum circuit sensitivity sweep.** PR-AUC values are means over 15 fold–seed units at $\rho = 1$, original imbalance. Results are *exploratory*: configurations were selected post-hoc and were not validated on a held-out split. Full encoding/entanglement labels use |-delimited keys in the saved CSV.

| Model | Architecture | Qubits | Layers | Entangle | Encoding | PR-AUC |
|---|---|---|---|---|---|---|
| VQC | best in grid | 8 | 8 | full | RX+RY+RZ | 0.9248±0.0424 |
| | default (main benchmark) | 4 | 2 | ring | RY | 0.9136±0.0772 |
| REUP | best in grid (single-upload limit) | 2 | 1 | ring | RY | 0.8513±0.1140 |

### S3.2.6 Calibration and screening behavior (fixed-encoder protocol)

Calibration analysis further separates the behavior of the models. CGCNN achieves the lowest Brier score and ECE at full data, indicating the most reliable probability estimates. VQC gives competitive PR-AUC but is less well calibrated, while REUP shows both weak classification and poor probability quality. This distinction is important for materials screening, where the ranked candidate list and confidence estimates both affect downstream validation decisions. The results are also shown in Table S22.

Table S22: **Calibration metrics at $\rho = 1$ on the 316-compound structure-available benchmark.** Brier score and ECE are computed on pooled out-of-fold test predictions (mean ± SD over 25 fold–seed units). Lower is better for both metrics. High PR-AUC does not imply reliable probabilities.

| Model | Brier score | ECE |
|---|---|---|
| CGCNN | 0.0416 ± 0.0098 | 0.0670 ± 0.0212 |
| Linear | 0.1064 ± 0.0319 | 0.2443 ± 0.0460 |
| Logistic | 0.1064 ± 0.0319 | 0.2443 ± 0.0460 |
| MLP | 0.1036 ± 0.0542 | 0.2228 ± 0.1040 |
| VQC | 0.1948 ± 0.0426 | 0.3467 ± 0.0493 |
| REUP | 0.2445 ± 0.0157 | 0.3085 ± 0.0586 |

Top-$k$ analysis confirms the same trend. At $\rho = 1.0$, VQC obtains the highest precision@5 of 0.9360, compared with 0.8800 for CGCNN, 0.8880 for the linear head, and 0.8720 for MLP. However, REUP

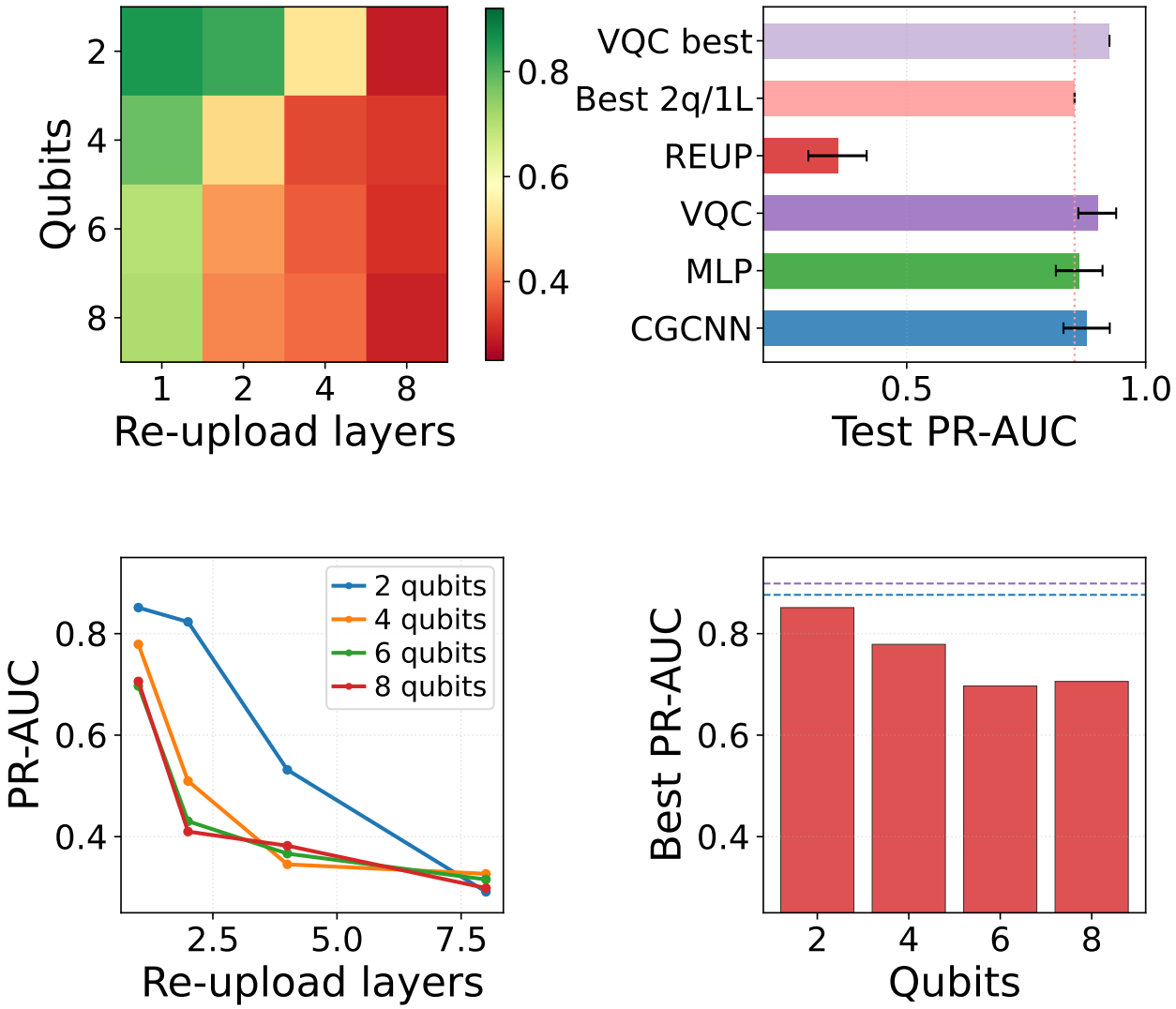


Figure S4: REUP architecture ablation at full data. The default REUP configuration uses four qubits and four re-uploading layers and gives weak performance, while the best sweep configuration uses two qubits and one data-encoding layer and reaches substantially higher PR-AUC. The result shows strong architecture sensitivity and that the best observed setting is the $L = 1$ no-re-upload limit.

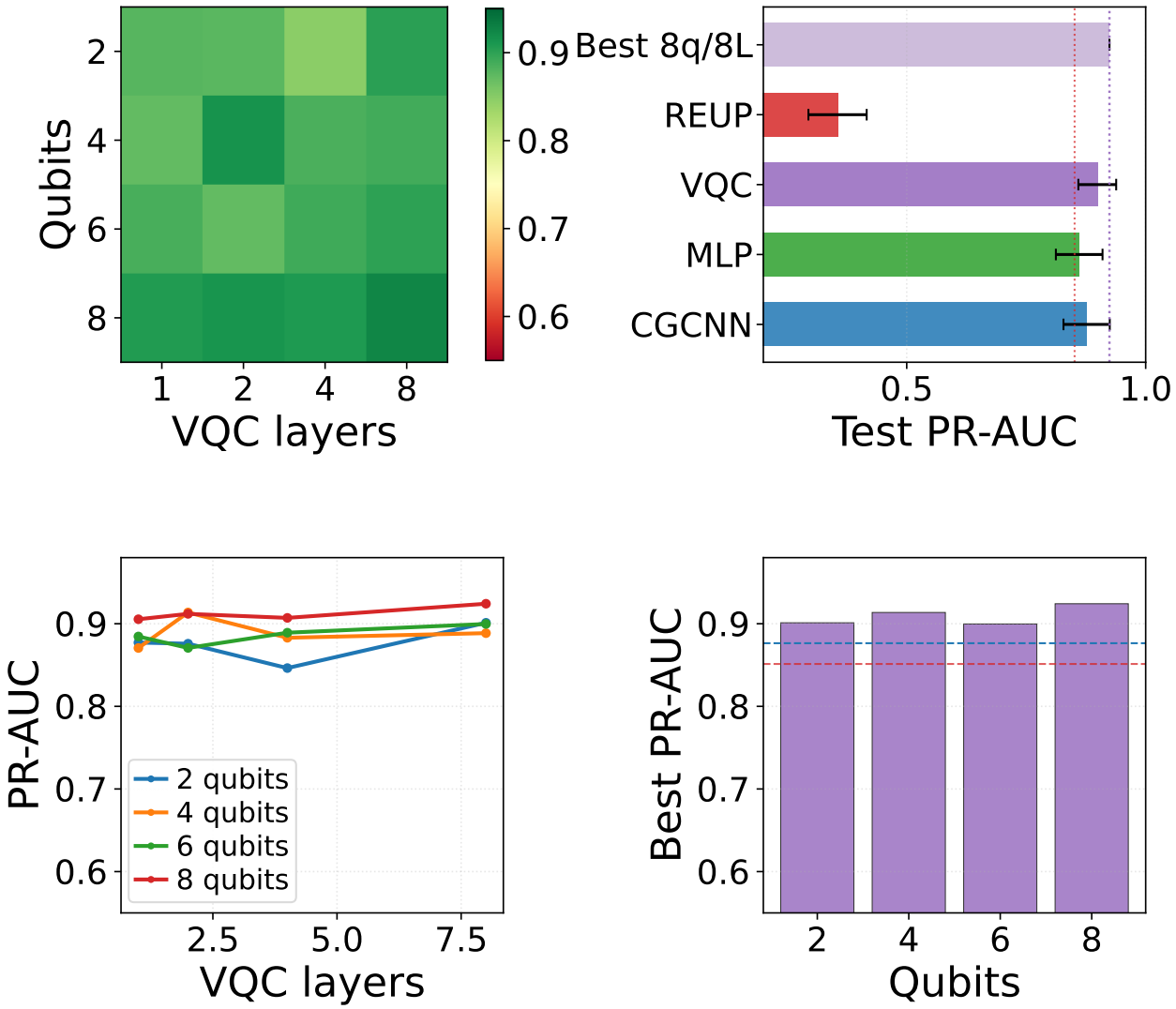


Figure S5: VQC architecture ablation at full data. Several VQC configurations remain competitive with the strongest baselines, indicating that the poor default REUP behavior is head-specific rather than a general failure of QML on CGCNN embeddings.

reaches only 0.3920 precision@5 and remains weak at larger $k$. The highest observed fixed-encoder precision@5 belongs to VQC, but no paired significance test for precision@5 is reported; the result is therefore descriptive. Precision@20 is retained only as a broad-retrieval diagnostic because $k = 20$ exceeds the approximate number of positives in a test fold.

Table S23: **Figure S7 visualizes the enrichment trends, and Table S23 reports the corresponding top-$k$ candidate screening metrics at $\rho = 1$.** Precision@$k$, Recall@$k$, and Enrichment@$k$ (relative to MIT prevalence in each test fold) for $k \in \{5, 10, 20\}$. Values are mean ± SD over 25 fold–seed units.

| Model | $k$ | Prec@$k$ | Recall@$k$ | Enrich@$k$ |
|---|---|---|---|---|
| CGCNN | 5 | 0.880±0.163 | 0.395±0.120 | 5.02±1.68 |
| | 10 | 0.860±0.122 | 0.759±0.114 | 4.80±0.75 |
| | 20 | 0.580±0.153 | 0.988±0.024 | 3.12±0.12 |
| Linear | 5 | 0.888±0.174 | 0.401±0.132 | 5.10±1.80 |
| | 10 | 0.848±0.123 | 0.750±0.127 | 4.74±0.83 |
| | 20 | 0.564±0.152 | 0.962±0.053 | 3.04±0.18 |
| Logistic | 5 | 0.888±0.174 | 0.401±0.132 | 5.10±1.80 |
| | 10 | 0.848±0.123 | 0.750±0.127 | 4.74±0.83 |
| | 20 | 0.564±0.152 | 0.962±0.053 | 3.04±0.18 |
| MLP | 5 | 0.872±0.172 | 0.391±0.122 | 4.97±1.71 |
| | 10 | 0.852±0.123 | 0.754±0.130 | 4.77±0.86 |
| | 20 | 0.562±0.148 | 0.960±0.049 | 3.03±0.18 |
| VQC | 5 | 0.936±0.138 | 0.420±0.110 | 5.31±1.43 |
| | 10 | 0.864±0.135 | 0.760±0.115 | 4.81±0.75 |
| | 20 | 0.560±0.148 | 0.956±0.056 | 3.02±0.19 |
| REUP | 5 | 0.392±0.248 | 0.177±0.134 | 2.24±1.71 |
| | 10 | 0.288±0.164 | 0.256±0.158 | 1.62±1.01 |
| | 20 | 0.252±0.118 | 0.426±0.167 | 1.34±0.52 |

### S3.3 Full descriptor benchmark: unbalanced training folds

The remaining results sections report the complementary benchmark on the full 343-material descriptor dataset of Section S1.7, in which the re-uploading classifier is trained directly on six physical descriptors and the controlled variable is the imbalance treatment of the training folds. All descriptor results are averaged over five seeds (42–46) and reported as mean ± 95% confidence interval. The original imbalanced setting serves as the reference point: it shows what the re-uploading family can extract from the raw 19.8% MIT distribution when no help is given. Figure S8 shows that a nontrivial MIT signal exists but remains fragile. The best configuration reached an F1 of $0.431 \pm 0.056$, a PR-AUC of $0.405 \pm 0.035$, a balanced accuracy of $0.654 \pm 0.041$, and a ROC AUC of $0.745 \pm 0.023$, with a precision of $0.443 \pm 0.021$ and a recall of $0.450 \pm 0.135$. The gap between ROC AUC and PR-AUC is the first diagnostic: the circuit ranks materials well above chance, but this ranking does not translate into reliable minority-class decisions, and the very wide recall interval (±0.135) shows that what little the untreated model detects is unstable from seed to seed—the untreated dataset supports detection, not yet decision-making.

Panel (b) of Figure S8 exposes the mechanism behind this fragility. The grid separates into two

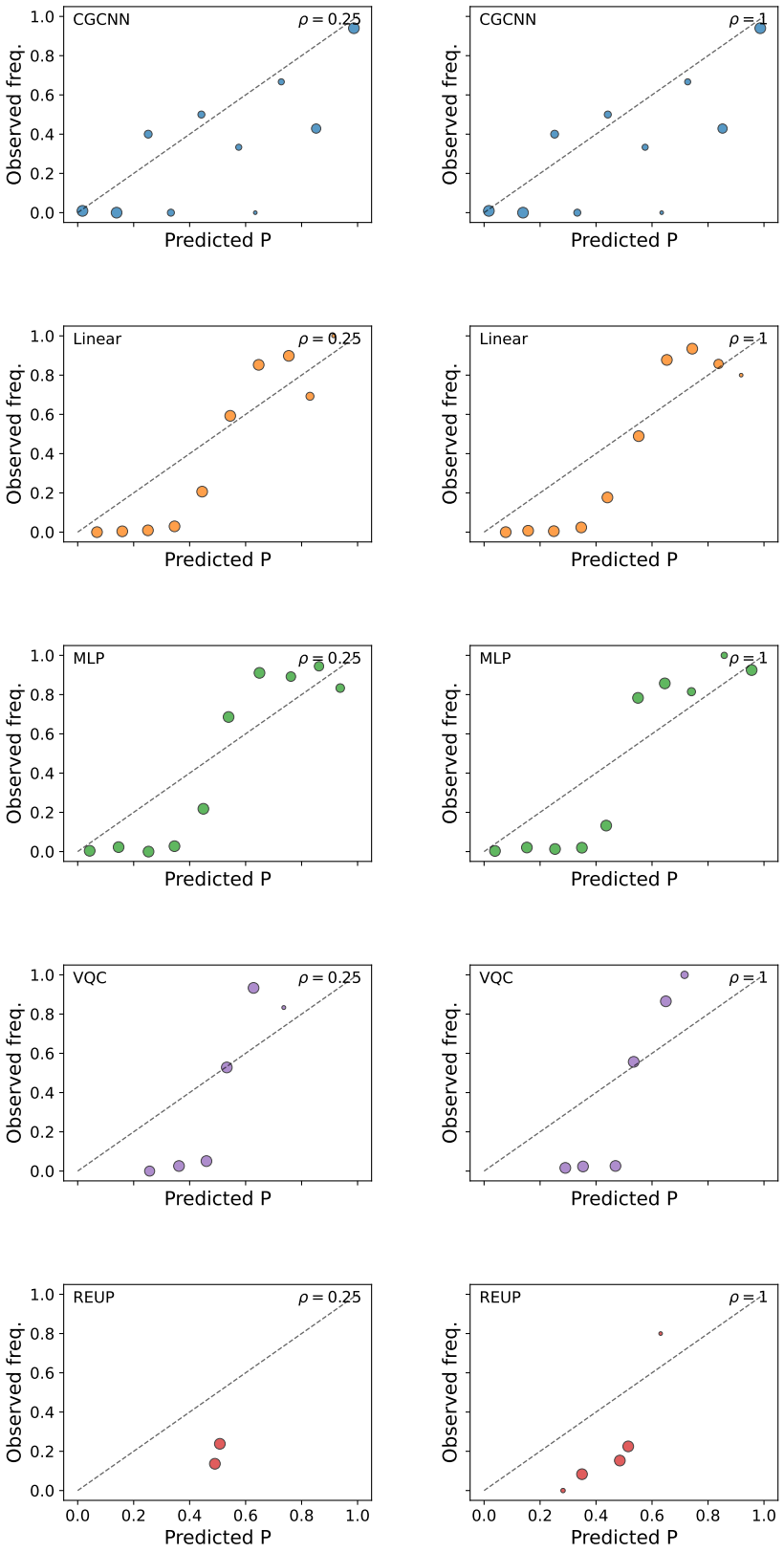


Figure S6: Calibration comparison at selected data fractions. CGCNN provides the best calibrated probabilities, while VQC and REUP require improved calibration before deployment in screening workflows.

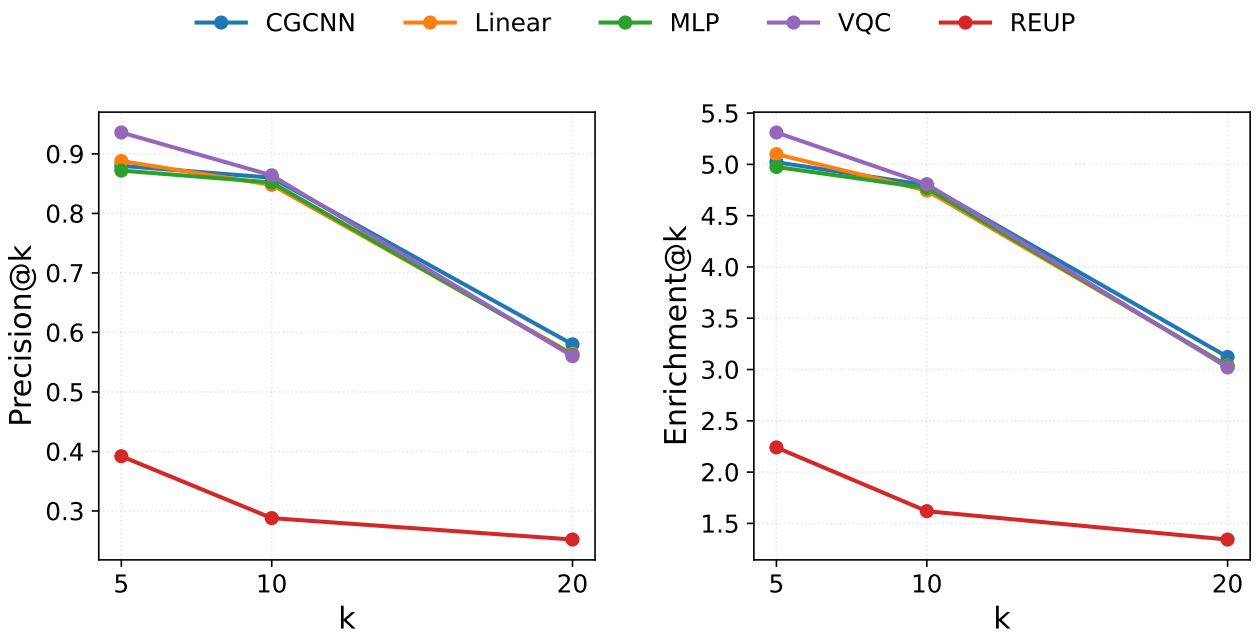


Figure S7: Top-$k$ screening performance at full data. VQC performs strongly among the highest-ranked candidates, whereas REUP provides limited enrichment compared with the other models.

clouds by feature scaler: configurations trained with $[0, 1]$ scaling cluster near a balanced accuracy of 0.5–0.55, the signature of a classifier that has collapsed onto the majority class, even though several retain a comparatively high PR-AUC. The $[0, 2\pi]$ configurations behave in the opposite way: they reach balanced accuracies up to about 0.66 but with systematically lower PR-AUC. In other words, under class imbalance the continuous output of the circuit can still rank MIT materials usefully while the induced decision behaviour fails, and the wider $[0, 2\pi]$ encoding range is what allows the decision boundary to fire on the minority class. This is why the best-F1 configuration in this setting—and only in this setting—uses the $[0, 2\pi]$ scaler; it is a one-qubit, two-layer, symmetrical, non-entangled circuit.

The hyperparameter landscape confirms that circuit size alone cannot solve the imbalance problem: panel (a) shows that the maximum F1 is low across the whole grid and not monotonic in either qubit count or depth, pointing to an optimization–generalization tradeoff rather than a simple expressivity limitation. What is missing is not variational capacity but minority-class support in the training folds, which motivates the three treatments analyzed next.

### S3.4 Full descriptor benchmark: random oversampling

Figure S9 shows a clear improvement once random minority oversampling is applied within each training fold. The best configuration improved to an F1 of $0.513 \pm 0.040$, a PR-AUC of $0.565 \pm 0.027$, a balanced accuracy of $0.733 \pm 0.024$, and a ROC AUC of $0.821 \pm 0.024$; recall rose from 0.450 to $0.776 \pm 0.053$ while precision held near 0.39. The simultaneous increase of F1 and PR-AUC is the important observation: repeated exposure to minority examples did not merely shift the decision threshold toward the MIT class, it produced circuit parameters whose continuous scores rank MIT materials better across thresholds. Panel (c) makes the same point at the level of the whole grid rather than the single winner: the PR-AUC distribution across all 72 configurations shifts upward,

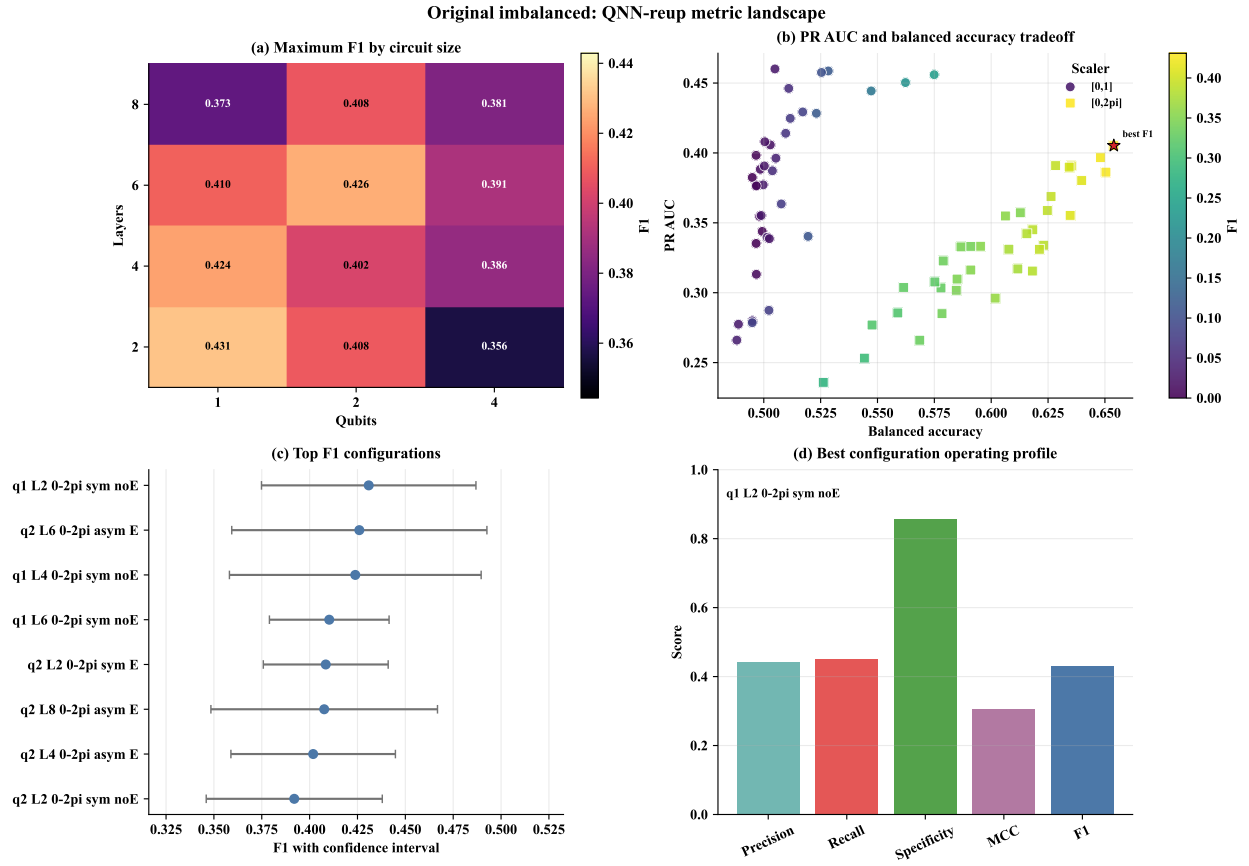


Figure S8: Descriptor-based re-uploading metric landscape for the unbalanced training-fold setting. The panels summarize the best F1 values by circuit size, the PR-AUC–balanced accuracy tradeoff, the top F1 configurations with confidence intervals, and the operating profile of the best configuration. All values are averaged over five seeds (42–46); intervals are 95% CIs across seeds.

into a range that only the very best configurations of the untreated setting could previously reach. Balancing therefore lifted the entire hyperparameter landscape, not just its optimum, which is stronger evidence that the treatment changed the effective learning problem.

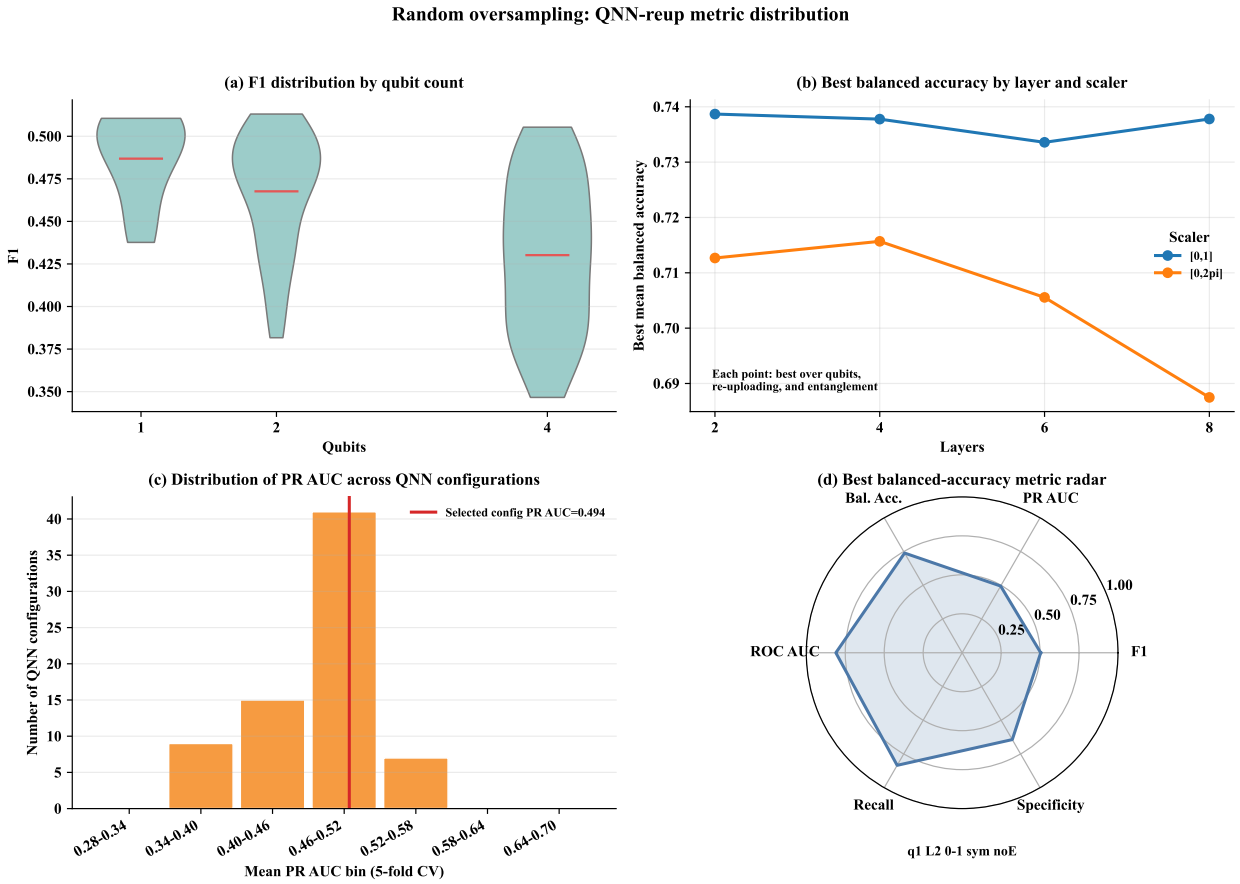


Figure S9: Descriptor-based re-uploading metric distribution for the random-oversampling setting. The panels show the F1 distribution by qubit count, the best balanced accuracy by layer and scaler, the PR-AUC distribution across all 72 configurations, and the metric radar of the best balanced-accuracy configuration.

The architecture response is more nuanced than a single circuit size. The best-F1 configuration here uses two qubits, eight layers, $[0, 1]$ scaling, asymmetrical re-uploading, and ring entanglement—not a minimal circuit—so the gain from oversampling is not tied to one compact architecture. The robust trend is in the input scaling: panel (b) shows that once the folds are balanced the $[0, 1]$ scaler dominates $[0, 2\pi]$ at every depth, with the $[0, 2\pi]$ curve collapsing at eight layers. This is the mirror image of the untreated setting: the classifier no longer needs the aggressive $[0, 2\pi]$ encoding to escape majority collapse, and the smoother $[0, 1]$ encoding generalizes better, consistent with the view that each re-uploading layer multiplies the frequency content of the learned function [15], so a restrained input range keeps deep re-uploading models from becoming too oscillatory for 343 samples. Panel (d) profiles the best balanced-accuracy configuration, whose radar is rounded across all six metrics. The structural limitation of duplication-based balancing remains visible in the modest precision ($\approx 0.39$): the gain is recall-driven class exposure, which motivates the interpolation-based treatments examined next.

## S3.5 Full descriptor benchmark: SMOTE

Figure S10 shows that SMOTE produced the strongest single-method shift toward minority-class sensitivity. The best configuration reached an F1 of $0.530 \pm 0.036$, a PR-AUC of $0.569 \pm 0.062$, a balanced accuracy of $0.750 \pm 0.030$, and a ROC AUC of $0.828 \pm 0.026$, with the highest recall of the descriptor benchmark ($0.806 \pm 0.050$) at a precision of $0.400 \pm 0.036$. Its central F1 estimate edges above that of random oversampling, although the two intervals overlap. Because SMOTE interpolates between neighbouring MIT materials, it populates the region between minority samples instead of re-weighting isolated points, and the classifier learns a decision region that covers the minority manifold more continuously. The result is that roughly four out of five MIT materials in the untouched test folds are recovered, a qualitatively different operating point from the 0.450 recall of the untreated setting.

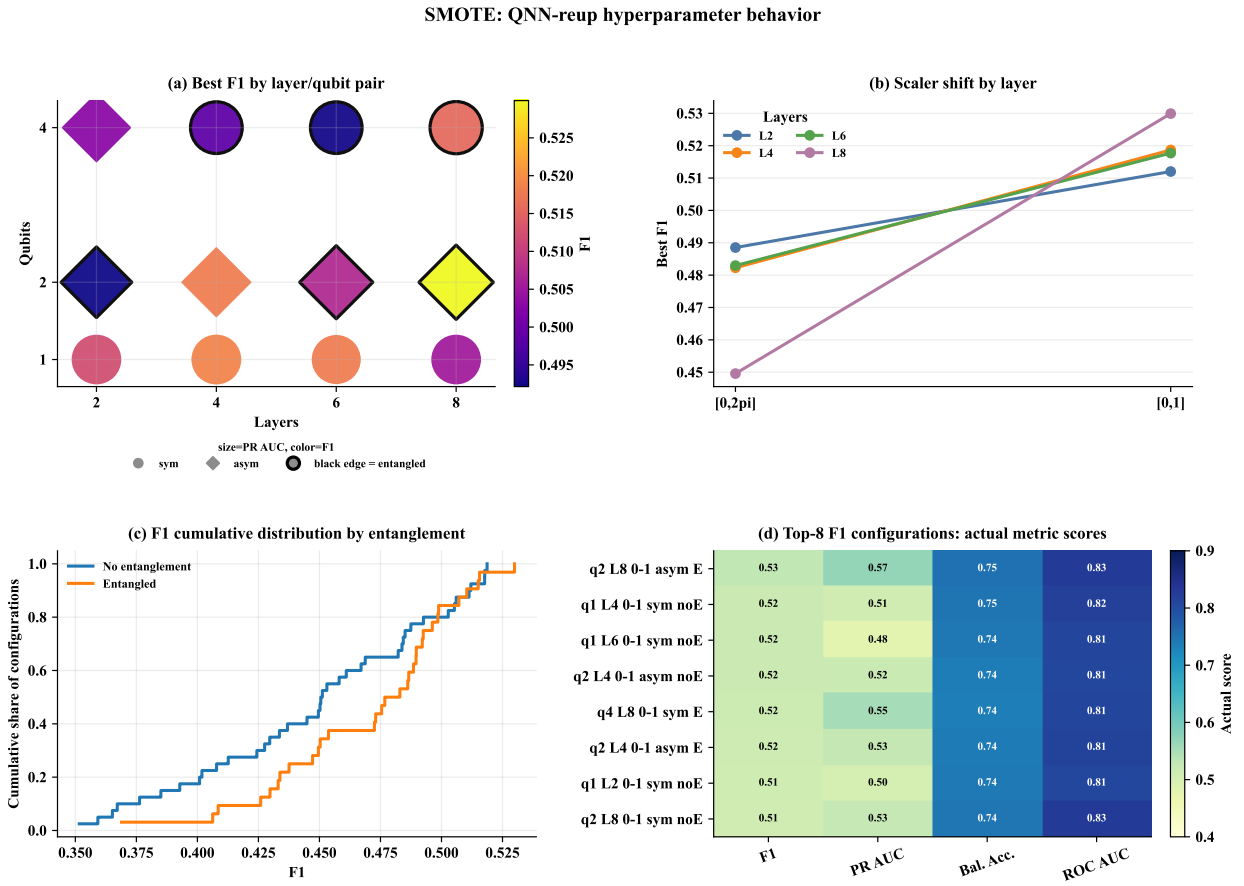


Figure S10: Descriptor-based re-uploading hyperparameter behavior for the SMOTE setting. The panels show the best F1 per layer–qubit pair (color = F1, size = PR-AUC, marker shape = re-uploading form, black edge = entangled), scaler-dependent F1 shifts across layers, the cumulative F1 distribution with and without entanglement, and the actual metric scores of the top-eight F1 configurations.

The best SMOTE configuration uses two qubits, eight layers, $[0, 1]$ scaling, asymmetrical re-uploading, and ring entanglement—the same architecture selected under random oversampling, and notably *not* a minimal circuit. Panel (b) shows the interaction that survives seed averaging: the advantage of $[0, 1]$ over $[0, 2\pi]$ widens steadily as the number of re-uploading layers grows,

so input range and circuit depth are not independent knobs—deep re-uploading is beneficial only when the encoding range is restrained enough to keep the learned function smooth. Panel (c) shows that the entangled and non-entangled cumulative F1 curves overlap across most of their range, and panel (d) that the top-eight configurations score within a narrow F1 band. The remaining weakness is precision (about 0.40), which caps PR-AUC: SMOTE can interpolate into ambiguous regions near the class boundary, where synthetic MIT samples overlap genuine non-MIT materials—the failure mode that Tomek-link cleaning targets.

### S3.6 Full descriptor benchmark: SMOTE–Tomek

Figure S11 shows that SMOTE–Tomek performs comparably to SMOTE rather than clearly surpassing it. The best configuration reached an F1 of $0.522 \pm 0.026$, a PR-AUC of $0.471 \pm 0.020$, a balanced accuracy of $0.746 \pm 0.023$, a ROC AUC of $0.801 \pm 0.021$, and an MCC of $0.403 \pm 0.040$, with a recall of $0.816 \pm 0.044$ at a precision of $0.391 \pm 0.028$. Its F1 interval overlaps almost entirely with those of SMOTE and random oversampling, so on this dataset the Tomek-link cleaning step does not yield a statistically distinguishable improvement; it does, however, produce the tightest F1 interval of the four settings, consistent with the intended effect of removing a few ambiguous majority–minority pairs near the class boundary. Panel (a) shows the best-F1 configuration at the corner of a short Pareto front, indicating little remaining F1–PR-AUC tradeoff among the leading models.

The selection diagnostics indicate which design choices actually matter once the data are well shaped. Panel (b) ranks the hyperparameters by the range of their group-mean F1: feature scaler (0.061) and qubit count (0.057) dominate, while layer count (0.023), entanglement (0.021), and re-uploading form (0.010) are comparatively inert—performance is governed by representation choices rather than by depth or entanglement. Panel (c) shows a shallow top-30 rank curve with broadly overlapping 95% confidence bands, a plateau of near-equivalent models instead of one sharp optimum, and panel (d) confirms that the leading models are overwhelmingly $[0, 1]$-scaled.

### S3.7 Full descriptor benchmark: cross-treatment synthesis

Reading the four settings as one controlled experiment, every minority-class metric improves substantially once any balancing is applied, but the three balancing treatments are not separable from one another. The best F1 rises from $0.431 \pm 0.056$ (untreated) to $0.513 \pm 0.040$ (random oversampling), $0.530 \pm 0.036$ (SMOTE), and $0.522 \pm 0.026$ (SMOTE–Tomek); the three balanced settings have heavily overlapping 95% confidence intervals, so no ordering among them is statistically supported. The same pattern holds for balanced accuracy (0.73–0.75), ROC AUC (0.80–0.83), and MCC (0.39–0.41), while recall jumps from 0.450 to 0.78–0.82. Specificity stays in a comparable range across settings, so the gains are minority-side rather than bought by sacrificing the majority class. These are real changes in the learned decision function, evaluated on test folds that always retain the original imbalance and averaged over five seeds.

The qualitative ordering of the treatments still matches the kind of minority-class information they inject—random oversampling adds *exposure*, SMOTE adds *coverage* of the minority manifold, and SMOTE–Tomek adds boundary cleaning—but on this dataset these mechanisms produce statistically indistinguishable end performance. One statistical caution applies to all four settings and is stated once here rather than repeated per treatment: each test fold contains only 20 MIT materials, so even after five-seed averaging the confidence intervals of leading configurations overlap heavily,

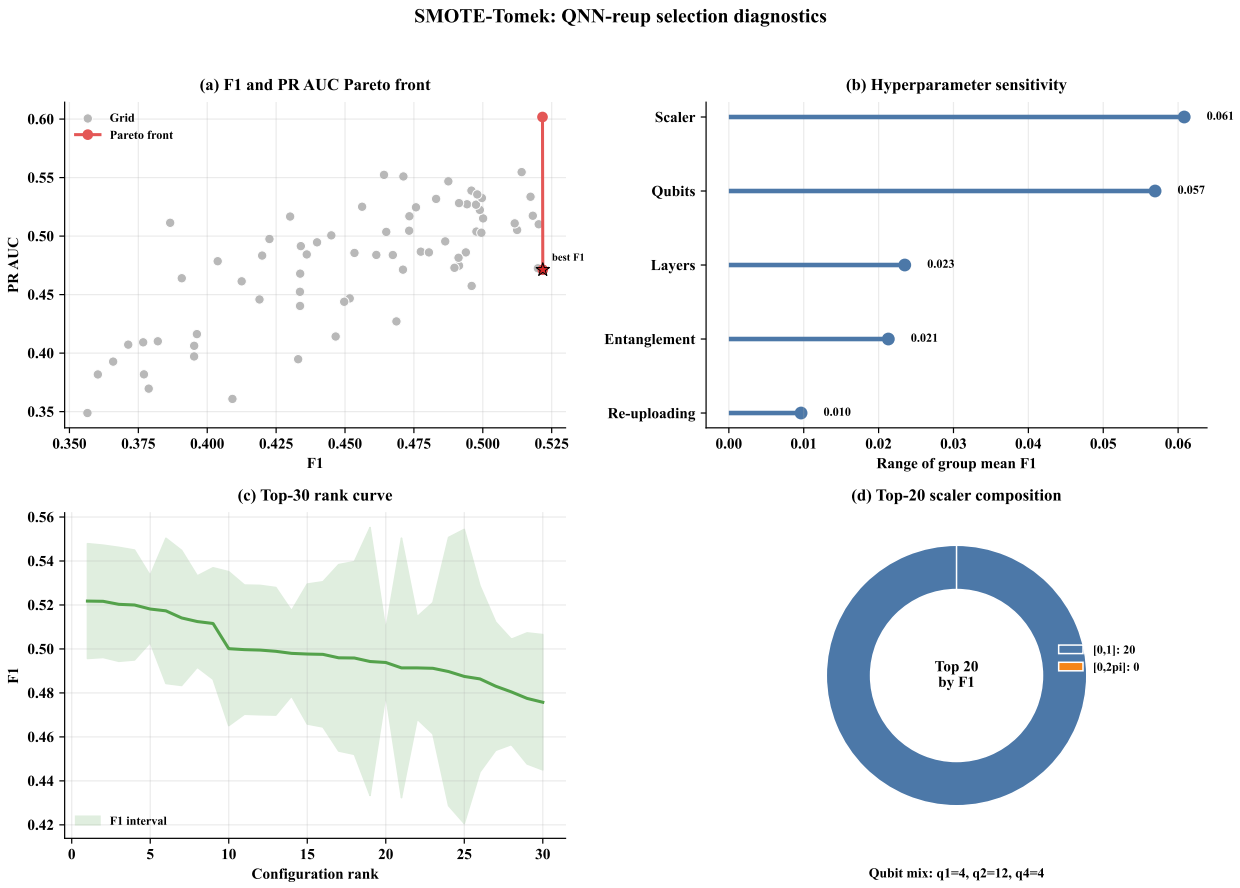


Figure S11: Descriptor-based re-uploading selection diagnostics for the SMOTE–Tomek setting. The panels show the F1–PR-AUC Pareto front, hyperparameter sensitivity, the top-30 F1 rank curve with confidence intervals, and the scaler composition of the top configurations. All values are averaged over five seeds (42–46); intervals are 95% CIs across seeds.

model selection is intrinsically noisy, and the per-treatment "best" circuits (which differ across treatments and seeds—e.g., the SMOTE–Tomek winner is a one-qubit, four-layer, non-entangled circuit while the SMOTE winner is a two-qubit, eight-layer, entangled one) should not be over-interpreted. Consistently with this, the hyperparameter-sensitivity analysis shows that input scaling and qubit count matter more than depth, entanglement, or re-uploading form. The practical lever is thus the training-fold treatment together with input scaling, not the fine circuit architecture.

Most important, the re-uploading classifier does not approach classical performance on the same descriptors. The compact comparison is retained in the main manuscript, and the complete metric profile is given below in Table S24. The best re-uploading model per treatment is compared against the best of three standard classical baselines (logistic regression, random forest, XGBoost) evaluated on the same descriptors, splits, and seeds, with class imbalance handled by class re-weighting (Section S1.7). A random forest reaches an F1 of 0.709 $[0.672, 0.747]$ and a balanced accuracy of 0.814, above every re-uploading variant, with the re-uploading F1 falling below the classical 95% confidence interval in all four settings. For practitioners, the recommendation is therefore clear in direction but bounded in degree: the original imbalanced data should not be used to train re-uploading classifiers on datasets of this size—some training-fold balancing is essential—but the choice among random oversampling, SMOTE, and SMOTE–Tomek is not critical, since all three perform equivalently within confidence intervals, and none reaches a well-tuned classical baseline. The descriptor benchmark therefore reinforces, rather than overturns, the central finding of this work: on this MIT task the re-uploading quantum classifier improves markedly with proper data treatment but remains below well-tuned classical baselines. This is consistent with the recognition in the QML literature that the structure and quality of the training data, rather than circuit expressivity alone, dominate practical performance [16, 17], and that supervised quantum advantages are structure-dependent rather than automatic [18].

### S3.8 Descriptor-space diagnostics

Two further diagnostics locate the descriptor-benchmark results in the materials-informatics context (Figure S12). First, the feature importances of the winning random-forest baseline (100 trees, maximum depth 4, balanced class weights; a refit over the same 25 splits reproduces the stored baseline, F1 $0.707 \pm 0.036$ versus 0.709) rank the volume per atomic site ($0.297 \pm 0.017$) and the mean covalent radius ($0.268 \pm 0.014$) as the dominant descriptors, followed by the range of the neighbor-distance variation ($0.175 \pm 0.011$); the remaining three descriptors contribute roughly 0.05–0.11 each. A model-agnostic permutation importance computed on the test folds confirms this ranking and sharpens it: permuting the volume per site or the mean covalent radius reduces test F1 by 0.29 each and the neighbor-distance range by 0.19, whereas the remaining three descriptors cost at most 0.03, with the average deviation of the ground-state band gap fully dispensable at test time ($-0.001 \pm 0.002$). The forest's decision structure therefore rests almost entirely on three size- and geometry-related descriptors. The prominence of covalent-radius and Mendeleev-number features is consistent with the curated MIT database literature, which identified exactly these families of size and chemical-identity descriptors as the strongest MIT discriminators [5]. Panel (b) shows why the task nonetheless remains hard for every model in this study: no single descriptor separates the classes on its own—the class-conditional distributions overlap heavily, with only modest median offsets—so successful classification must exploit descriptor interactions, which is precisely what a depth-4 forest captures cheaply and what a shallow re-uploading circuit must encode through repeated data injection.

Second, the aggregated test-fold confusion matrices in panel (c) quantify the mechanism behind the treatment gains of Sections S3.3–S3.7 at the level of raw counts (each test fold contains exactly 20 MIT and 83 non-MIT materials, so the fold-mean counts follow exactly from the stored recall and specificity). Untreated, the best re-uploading circuit recovers only 9.0 of 20 MIT materials per fold at 11.8 false positives. The three balancing treatments roughly double the recovered MIT count, to 15.5–16.3 of 20, but at the cost of 25–27 false positives per fold. The random forest occupies a different operating point: it recovers 13.8 of 20 MIT materials with only 5.1 false positives. After balancing, the re-uploading classifier is therefore *more* sensitive than the classical baseline (finding $\approx 2$ additional MIT materials per fold) but pays a roughly fivefold false-positive cost, which is what depresses its precision and F1. For a screening workflow in which every predicted MIT candidate is passed to DFT or experimental validation, this cost structure—not ranking ability alone—determines practical utility, and it identifies majority-class discrimination, rather than minority-class sensitivity, as the remaining deficit of the balanced re-uploading models.

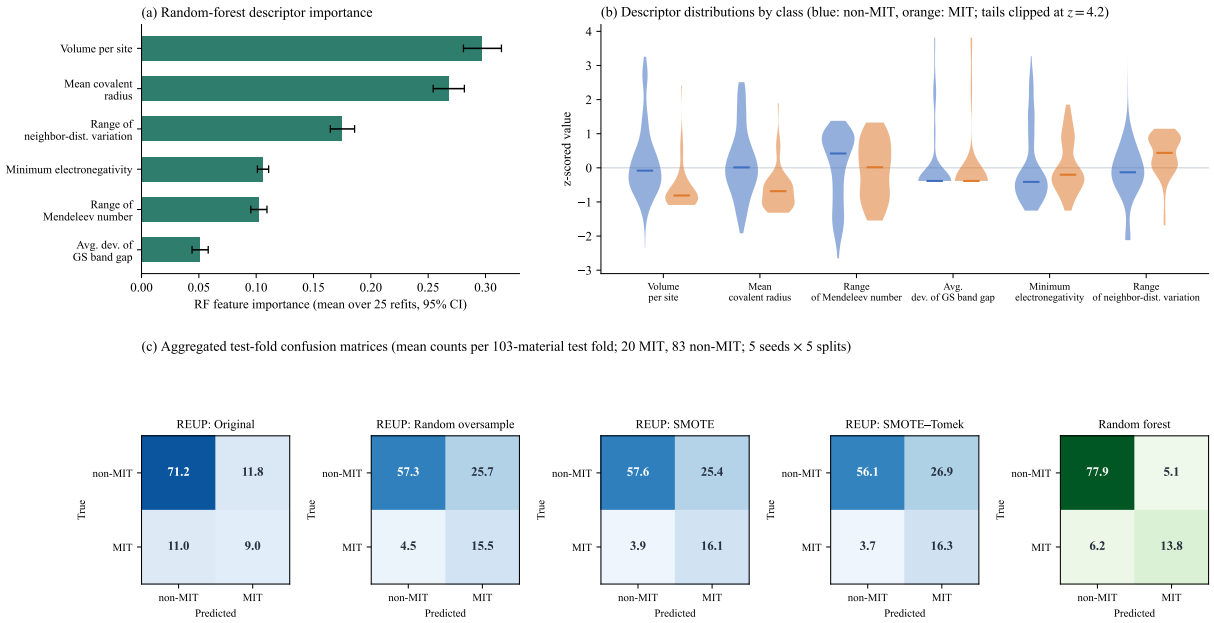


Figure S12: Descriptor-space diagnostics for the 343-material benchmark. (a) Feature importances of the winning random-forest baseline (100 trees, maximum depth 4, balanced class weights), averaged over 25 refits (5 seeds × 5 splits); error bars are 95% CIs. (b) z-scored descriptor distributions by class, showing that no single descriptor separates MIT from non-MIT materials. (c) Aggregated test-fold confusion matrices (mean counts per 103-material test fold; 20 MIT, 83 non-MIT) for the best re-uploading configuration of each imbalance treatment and for the random-forest baseline. The re-uploading matrices follow exactly from the stored mean recall and specificity because stratification fixes the fold composition.

Figure S13 extends the scalar metrics to full receiver-operating-characteristic, precision–recall, and reliability curves, computed from per-sample test scores obtained by re-fitting the winning configuration of each treatment and the winning random forest on the identical 25 splits (the re-runs reproduce the stored per-seed metrics exactly, so the curves describe the same models reported above). Two observations follow. First, the balancing treatments separate from the untreated

model across the entire operating range, not merely at the selected threshold: the treated ROC and precision–recall curves lie above the untreated curves at essentially every operating point, confirming that balancing reshapes the learned score function rather than shifting a threshold. Second, the random forest dominates every re-uploading variant wherever it matters for screening; most visibly in the precision–recall plane, it holds precision above 0.9 out to a recall of roughly 0.5, whereas the best re-uploading models fall below 0.8 precision almost immediately. Paired comparisons make this gap formal while respecting the dependence of the evaluation units: because the five splits of a seed are not mutually independent, we aggregate to per-seed means before testing, leaving five independent paired units. On these, the random-forest F1 exceeds that of the best re-uploading model by $0.276 \pm 0.062$ (original), $0.194 \pm 0.074$ (random oversampling), $0.177 \pm 0.035$ (SMOTE), and $0.186 \pm 0.037$ (SMOTE–Tomek; mean $\pm$ 95% CI over seeds), the difference is positive in every individual seed of every treatment, and a paired $t$-test on the seed-level means rejects equality in all four settings ($p \leq 0.002$, df $= 4$). Treating the 25 fold–seed units as independent instead (Wilcoxon signed-rank and sign-flip permutation tests, $p < 10^{-3}$) reaches the same conclusion and is reported only as a sensitivity check; the corresponding PR-AUC gaps are 0.18–0.34. The descriptor-side conclusion of Section S3.7 therefore holds not only within confidence intervals but under paired tests on matched splits, mirroring the paired-test discipline applied to the structure-based benchmark (Section S3.2.2).

The reliability panel adds a caveat that the threshold metrics conceal: training-fold balancing improves sensitivity but *degrades* probability calibration. The untreated model is the best-calibrated re-uploading variant (Brier $0.162 \pm 0.016$, ECE $0.146 \pm 0.036$), whereas all three balanced variants cluster near Brier 0.19 and ECE 0.24–0.25, systematically overconfident at intermediate probabilities (Figure S13c) because duplicated and interpolated minority samples inflate the apparent MIT density near the decision boundary. The random forest is both stronger and better calibrated (Brier $0.098 \pm 0.008$, ECE $0.109 \pm 0.009$). Table S24 collects the complete metric profile. For screening pipelines that consume probabilities rather than hard labels, balanced re-uploading models would therefore require post-hoc calibration before deployment, consistent with the calibration weakness of the embedding-based quantum heads (Section S3.2.6).

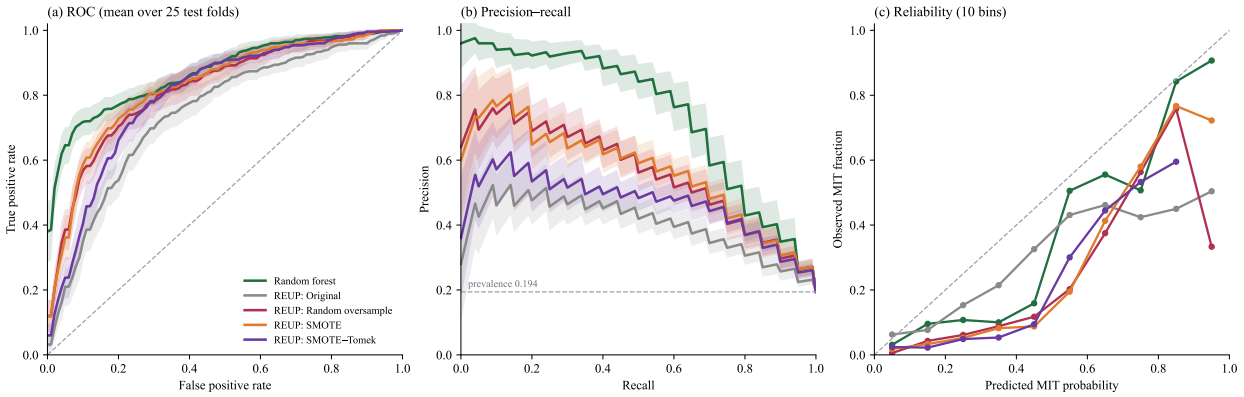


Figure S13: Operating-characteristic view of the descriptor benchmark, from per-sample test scores of the best re-uploading configuration per treatment and the random-forest baseline on the identical 25 splits. (a) ROC curves (mean over 25 test folds; bands are 95% CIs of the mean). (b) Precision–recall curves; the dashed line marks the test-fold MIT prevalence (0.194). (c) Reliability diagram (10 equal-width bins): all re-uploading variants are overconfident at intermediate predicted probabilities, and the balanced treatments more so than the untreated model.

Table S24: Complete metric profile of the best re-uploading (REUP) configuration per training-fold treatment and of the random-forest baseline, computed from per-sample test scores on the identical 25 splits (5 seeds × 5 splits; mean ± 95% CI across seeds). Brier score and expected calibration error (ECE) are lower-is-better.

| Model / treatment | F1 | PR-AUC | ROC-AUC | Bal. acc. | MCC | Precision | Recall | Specificity | Brier | ECE |
|---|---|---|---|---|---|---|---|---|---|---|
| Unbalanced | 0.431 ± 0.056 | 0.405 ± 0.035 | 0.745 ± 0.023 | 0.654 ± 0.041 | 0.307 ± 0.050 | 0.443 ± 0.021 | 0.450 ± 0.135 | 0.858 ± 0.054 | 0.162 ± 0.016 | 0.146 ± 0.036 |
| Random oversample | 0.513 ± 0.040 | 0.565 ± 0.026 | 0.821 ± 0.024 | 0.733 ± 0.024 | 0.385 ± 0.049 | 0.389 ± 0.052 | 0.776 ± 0.053 | 0.691 ± 0.079 | 0.187 ± 0.016 | 0.250 ± 0.028 |
| SMOTE | 0.530 ± 0.036 | 0.568 ± 0.062 | 0.828 ± 0.026 | 0.750 ± 0.030 | 0.411 ± 0.051 | 0.400 ± 0.036 | 0.806 ± 0.050 | 0.694 ± 0.046 | 0.185 ± 0.012 | 0.249 ± 0.024 |
| SMOTE–Tomek | 0.522 ± 0.026 | 0.471 ± 0.020 | 0.801 ± 0.021 | 0.746 ± 0.023 | 0.403 ± 0.040 | 0.390 ± 0.028 | 0.816 ± 0.044 | 0.676 ± 0.030 | 0.190 ± 0.009 | 0.239 ± 0.020 |
| Random forest | 0.707 ± 0.038 | 0.747 ± 0.056 | 0.871 ± 0.030 | 0.813 ± 0.024 | 0.642 ± 0.046 | 0.733 ± 0.036 | 0.688 ± 0.044 | 0.939 ± 0.010 | 0.098 ± 0.008 | 0.109 ± 0.009 |

## S3.9 Materials-family error analysis

Aggregate ranking and calibration metrics show how the classifiers behave on the full test distribution, but they do not reveal whether errors are uniformly distributed across chemistry space. This distinction is important for MIT materials, because different compound families can host different mixtures of structural, electronic, orbital, correlation-driven, and charge-ordering effects. We therefore performed a post-hoc materials-family error analysis on the fixed-encoder 316-compound graph–quantum predictions at $\rho = 1.0$. Because this analysis uses the fixed-encoder protocol, it is interpreted as a chemistry-dependent sensitivity diagnostic rather than part of the aligned $\rho$-matched comparison.

Because manually curated mechanism labels are not available for all entries, compounds were assigned to coarse chemistry families using formula-level composition rules. These family labels were used only for interpretation and were not provided to the classifiers during training. We therefore interpret the results as chemistry-dependent error diagnostics rather than as evidence that the model has learned specific Mott, Peierls, charge-density-wave, excitonic, or charge-order mechanisms.

Figure S14 summarizes the fixed-encoder family-level MIT recall of the CGCNN, VQC, and 4q/4L REUP heads across the formula-inferred families. The error structure is not uniform. Oxide-rich families such as vanadate, nickelate, cobaltate, and iron-oxide groups show comparatively high MIT recall, suggesting that the frozen CGCNN representation combined with the VQC head captures recurring structural or chemical motifs in these better represented regions of the dataset. In contrast, chalcogenides, manganite/oxide entries, and heterogeneous intermetallic or "other" compounds contain a larger fraction of missed MIT examples. These families are either sparsely represented, chemically diverse, or more likely to involve mechanisms that are not cleanly captured by the compressed graph embedding and post-hoc quantum readout.

This analysis supports two conclusions. First, the high aggregate PR-AUC of the VQC head does not imply uniform reliability across all MIT chemistries; family-dependent failure modes remain important for candidate screening. Second, the weak REUP result should not be interpreted only through global metrics, because representation compression and circuit design may affect sparse or chemically heterogeneous families more severely than well represented oxide families. A practical MIT-screening workflow should therefore combine aggregate ranking metrics with family-level error analysis before prioritizing candidates for experimental or first-principles validation.

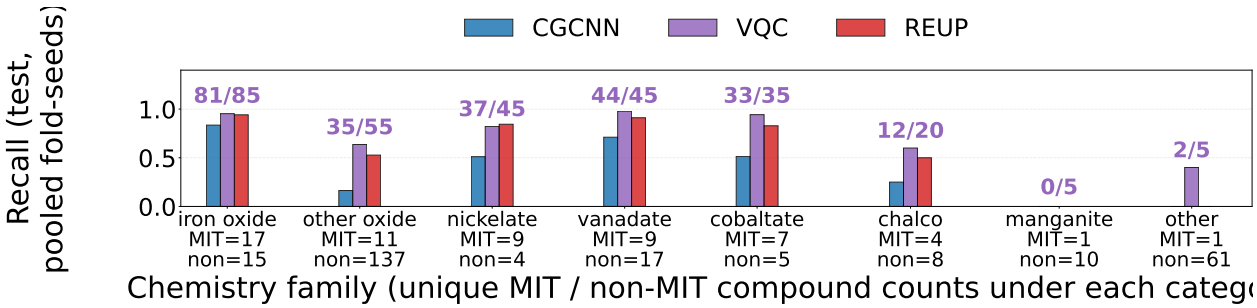


Figure S14: **Materials-family error analysis for the structure-based graph–quantum benchmark.** Post-hoc error analysis of CGCNN, VQC, and REUP predictions at $\rho = 1.0$ on the 316-compound structure-available benchmark. Compounds are grouped into coarse chemistry families inferred from formula-level composition rather than manually curated mechanism classes. Bars show family-level MIT recall for each fixed-encoder model, computed from pooled out-of-fold appearances across five seeds. The number of unique MIT compounds and evaluation records per family is required for interpreting uncertainty in sparse groups. The analysis shows that performance is family dependent: vanadate-, nickelate-, cobaltate-, and iron-oxide-rich groups are recovered more reliably, whereas chalcogenide, manganite/oxide, and heterogeneous intermetallic or "other" groups contain more missed MIT examples. Because MIT mechanisms can be coupled or contested, these formula-inferred families should be interpreted as diagnostic chemistry groupings rather than definitive physical mechanism labels.

## S4 Complete validation-selected architecture summaries

Tables S25–S26 and Figure S15 extend the primary architecture-selection analysis without changing the numbering of the preceding Supplementary items. The full per-architecture and per-$\rho$ selection records are retained in the machine-readable release.

Table S25: Full metrics at $\rho = 1$ for the validation-selected quantum-head families. Values are mean [95% CI] over 25 fold–seed evaluations.

| Model | PR-AUC | ROC-AUC | F1 | MCC |
|---|---|---|---|---|
| VQC, validation-selected | 0.870 [0.829,0.910] | 0.959 [0.944,0.974] | 0.788 [0.740,0.837] | 0.750 [0.693,0.808] |
| REUP, validation-selected | 0.811 [0.765,0.858] | 0.930 [0.907,0.953] | 0.726 [0.681,0.772] | 0.678 [0.625,0.732] |

Table S26: Overall architecture-selection frequency by circuit depth. Counts aggregate 225 cells per head (5 folds × 5 seeds × 9 fractions). For REUP, $L = 1$ is the single-upload limit.

| Head | $L = 1$ | $L = 2$ | $L = 4$ | $L = 8$ |
|---|---|---|---|---|
| VQC, count (rate) | 82 (36.4%) | 59 (26.2%) | 44 (19.6%) | 40 (17.8%) |
| REUP, count (rate) | 186 (82.7%) | 30 (13.3%) | 4 (1.8%) | 5 (2.2%) |

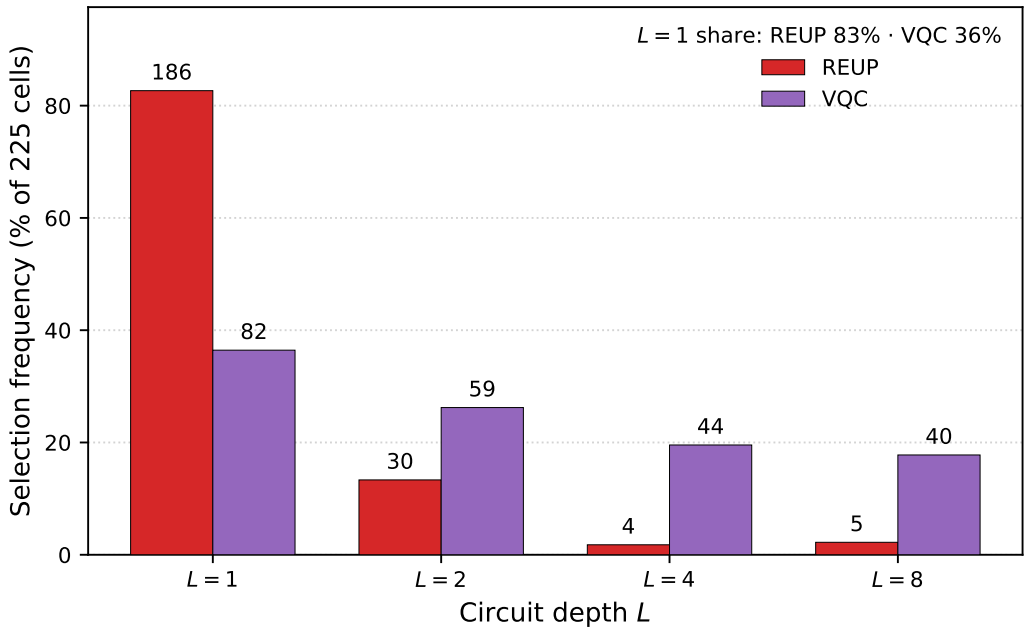


Figure S15: Validation-selected architecture frequency by depth over all 225 cells per quantum-head family. REUP selection is concentrated in the single-upload $L = 1$ limit, whereas VQC selection is distributed across depths.

# S5 Compound-resolved materials interpretation

## S5.1 Scientific objective and analysis scope

This analysis examines how the aggregate benchmarking results translate to individual materials. Its purpose is to identify which compounds are classified consistently, which known MIT materials are systematically missed, which curated non-MIT materials receive unexpectedly high MIT scores, and where quantum and classical classifier heads produce different decisions.

The analysis uses the full-data condition ($\rho = 1$) of the grouped nested structure benchmark. For every compound, only predictions obtained while that compound belonged to the locked outer test fold were retained. Each material therefore contributes one held-out prediction per random seed and model, and no compound-level result is based on a model trained on that compound. The analysis covers all 316 structure-available materials, comprising 59 MIT and 257 curated non-MIT compounds.

Six model outputs are included: the fold-local CGCNN probability, linear head, MLP head, validation-selected VQC, validation-selected REUP family, and projection-bottleneck classical head. For each compound and model, the five seed-level predictions were aggregated to obtain the mean MIT score, majority thresholded class, and cross-seed standard deviation. Thresholded predictions use the F1 threshold selected from the corresponding validation data rather than a universal probability threshold.

Because the models differ substantially in calibration, their raw scores should not be interpreted as directly comparable physical probabilities that a material undergoes an MIT. The compound-resolved analysis instead evaluates prediction consistency, model agreement, error persistence, and

sensitivity to the classifier readout.

## S5.2 Representative compound selection

Representative compounds were assigned to four diagnostic groups:

1. robustly identified MIT compounds that were correctly classified by most or all primary model heads;
2. persistent false-negative MIT compounds that were repeatedly classified as non-MIT;
3. high-scoring compounds carrying curated non-MIT labels; and
4. compounds showing disagreement between quantum and classical readouts.

The case-study selection was based on held-out prediction consistency, persistent error behavior, model disagreement, and chemistry-family coverage. It was not based solely on selecting the largest or smallest model scores. Supplementary Table S27 reports 23 representative compounds: five robust MIT compounds, five persistent false negatives, five high-scoring curated negatives, and eight quantum–classical disagreement cases.

The complete 316-compound aggregate table is supplied as Supplementary Data S1 so that readers can examine all materials rather than only the selected examples. Supplementary Data S2 contains every seed-, fold-, and model-level prediction underlying the aggregate analysis.

## S5.3 Robustly identified MIT compounds

Several known MIT materials are consistently identified across both quantum and classical readouts. Representative examples include $PrFe_4(CuO_4)_3$, $V_4O_7$, $EuNiO_3$, $NdCoO_3$, and $Ti_2O_3$. All five primary model heads assign the MIT class to these compounds by majority vote across seeds.

The robust group spans iron oxides, vanadate-derived oxides, nickelates, cobaltates, and titanium oxides. This family diversity indicates that successful classification is not limited to one narrow composition class. However, the compounds do not share a single universal descriptor signature. Some materials lie close to the MIT-class medians across most of the six physical descriptors, whereas others contain one or more values outside the central ranges of both classes. The consistent predictions therefore appear to arise from multivariate structure–composition patterns rather than from a single descriptor threshold.

For example, $V_4O_7$ is classified as MIT by all six evaluated heads, with mean held-out probabilities of 0.971 for CGCNN, 0.756 for the linear head, 0.854 for the MLP, 0.583 for the validation-selected VQC, 0.740 for the validation-selected REUP model, and 0.936 for the projection-bottleneck control. The lower absolute quantum scores should not be interpreted as weaker physical confidence because the model families are differently calibrated. The important result is the consistency of the validation-thresholded decision across models and seeds.

Similarly, $EuNiO_3$ is classified correctly by every model and exhibits low-to-moderate cross-seed variability. Its six-descriptor profile lies predominantly near the MIT-class medians, suggesting that it occupies a comparatively well-represented region of both the learned graph space and the interpretable descriptor space.

### S5.4 Persistent false-negative MIT compounds

A second group contains labeled MIT compounds that are systematically classified as non-MIT. Representative examples include $NbO_2$, $Pr(P_3Ru)_4$, $K_3(MoO_3)_{10}$, $Ca_2RuO_4$, and $Ca_4Mn_3O_{10}$.

For $NbO_2$, none of the five primary heads predicts MIT by majority vote. Its mean held-out probabilities range from 0.051 for the fold-local CGCNN to 0.466 for the validation-selected VQC. The prediction is also relatively stable across seeds, indicating that the error is not simply caused by one unfavorable random initialization.

Likewise, $Ca_2RuO_4$ is classified as non-MIT by all six model heads. Several of its physical descriptors, including volume per atomic site, mean covalent radius, Mendeleev-number range, and neighbor-distance variation, lie closer to the non-MIT class medians than to a clearly separated MIT region. This helps explain the statistical prediction but does not establish the microscopic origin of the transition.

These persistent false negatives expose a representation-level limitation. Static relaxed structures and broad composition–structure descriptors do not explicitly encode temperature-dependent structural evolution, strong electronic correlations, defects, stoichiometry, pressure, strain, magnetic order, or other external variables that may determine MIT behavior. The analysis therefore cannot identify which omitted physical factor causes an individual error. It can only show that replacing the final readout with a quantum circuit does not consistently recover information that is absent or weakly represented in the input features.

### S5.5 High-scoring curated non-MIT compounds

Several compounds assigned to the non-MIT category in the source dataset receive consistently high MIT scores. Representative examples include $IrO_2$, $V_7O_{13}$, $V_2O_5$, $TiO_2$, and TiO.

The clearest cases are $IrO_2$ and $V_7O_{13}$, which are classified as MIT by all five primary heads and by the projection-bottleneck control. Their mean scores across the evaluated models are approximately 0.816 and 0.811, respectively. For $V_7O_{13}$, the relatively small volume per atomic site, low mean covalent radius, and elevated neighbor-distance variation place it in a region occupied by several labeled MIT oxides.

These materials should be interpreted as high-scoring curated negatives or label-audit cases, not as newly discovered MIT compounds. The negative category in the source dataset combines curated metals and insulators and does not prove that no MIT can occur under every possible temperature, pressure, stoichiometry, doping, or structural condition. Conversely, a high model score is not physical verification of a transition. These cases indicate that the materials resemble labeled MIT compounds in the learned representation and may warrant closer literature or label review, but candidate-level claims require independent validation.

### S5.6 Quantum–classical disagreement cases

Some compounds lie close to the decision boundary and are classified differently by quantum and classical heads. These cases include $DyNiO_3$, $Eu(P_3Ru)_4$, $Ca(FeO_2)_2$, $CoP_3$, $KVO_3$, $LiTi_2O_4$, $Pr(P_3Ru)_4$, and $Ti_3O_5$.

For the labeled MIT compound $DyNiO_3$, the fold-local CGCNN and validation-selected VQC pre-

dict MIT, whereas the linear, MLP, validation-selected REUP, and projection-bottleneck heads predominantly predict non-MIT. Its six descriptor values lie close to the MIT-class medians, so the disagreement cannot be attributed to one obvious descriptor outlier. Instead, it reflects differences in how the classifier heads partition an ambiguous learned representation.

For $Ti_3O_5$, the linear, MLP, and REUP heads predict MIT, whereas the CGCNN probability, VQC, and projection-bottleneck control do not. Its elevated neighbor-distance variation provides one MIT-associated statistical signal, but this signal is not interpreted consistently across classifiers.

The curated non-MIT compound $Eu(P_3Ru)_4$ shows the opposite pattern: the validation-selected VQC and REUP heads predict MIT, whereas the CGCNN and most classical heads retain the non-MIT class. Such examples demonstrate that quantum heads can alter individual material decisions. They do not establish quantum-specific physical insight, because the aggregate benchmark and bottleneck-matched controls provide no evidence of systematic quantum superiority. These disagreements are therefore best interpreted as classifier uncertainty.

### S5.7 Descriptor context and interpretive limitations

The six physical descriptors reported for each selected compound are volume per atomic site, composition-averaged covalent radius, range of Mendeleev numbers, average deviation of elemental ground-state band gap, minimum electronegativity, and range of neighbor-distance variation. Their positions relative to the dataset distribution and class medians provide a statistical explanation for why a compound may lie closer to the learned MIT or non-MIT region.

The descriptor contribution values reported for the representative cases are obtained from a linear logistic attribution proxy. They are not SHAP values and should not be interpreted as model-independent feature importance. A positive contribution indicates that the corresponding descriptor pushes the logistic prediction toward the MIT class, while a negative contribution pushes it toward the non-MIT class.

These descriptor associations do not establish a microscopic MIT mechanism. Likewise, no atom-level or bond-level CGCNN attribution is used in the present analysis. Any future graph masking, integrated-gradient, or site-sensitivity analysis would describe sensitivity of the classifier output to structural inputs and should not be interpreted as proof that a particular atomic environment causes the transition.

### S5.8 Scientific conclusion

The compound-resolved analysis shows that the benchmark contains three distinct forms of materials-level information. First, several known MIT compounds are recognized consistently across chemistry families and classifier types, indicating that the fold-local embeddings contain transferable structure–composition signals associated with the MIT label. Second, persistent false negatives reveal known MIT materials whose behavior is not adequately represented by static structures and broad descriptors. Third, high-scoring curated negatives and classifier-disagreement cases identify regions in which dataset labels, representation limitations, and readout sensitivity require cautious interpretation.

These results strengthen the materials relevance of the benchmark by identifying which compounds are consistently easy, systematically difficult, or model dependent. They do not constitute new MIT discoveries, predictions of transition temperature or conductivity change, or validation of a

microscopic mechanism. Independent literature review, electronic-structure calculations, or experiment would be required before any compound-level screening hypothesis could be treated as a materials discovery.

Table S27: Compound-resolved case studies under the grouped full-data ($\rho = 1$) structure benchmark. Mean score is the average MIT probability across the six evaluated model outputs: the fold-local CGCNN, linear head, MLP head, validation-selected VQC, validation-selected REUP family, and projection-bottleneck classical head. "Correct heads" and "MIT votes" are calculated over the five primary heads: CGCNN, linear, MLP, validation-selected VQC, and validation-selected REUP. Curated non-MIT labels indicate the negative category in the source database and do not establish the absence of an MIT under all physical conditions. Detailed per-model probabilities, majority predictions, cross-seed standard deviations, and descriptor values are provided in the accompanying Supplementary Table S27 CSV.

| **Case-study group** | **Compound** | **Chemistry family** | **True label** | **Mean score** | **Correct heads** | **MIT votes** |
|---|---|---|---|---|---|---|
| *A. Robustly identified MIT materials* | | | | | | |
| Robust MIT | $PrFe_4(CuO_4)_3$ | Iron oxide | MIT | 0.809 | 5/5 | 5/5 |
| Robust MIT | $V_4O_7$ | Vanadate/oxide | MIT | 0.807 | 5/5 | 5/5 |
| Robust MIT | $EuNiO_3$ | Nickelate/oxide | MIT | 0.794 | 5/5 | 5/5 |
| Robust MIT | $NdCoO_3$ | Cobaltate/oxide | MIT | 0.777 | 5/5 | 5/5 |
| Robust MIT | $Ti_2O_3$ | Other oxide | MIT | 0.725 | 5/5 | 5/5 |
| *B. Persistent false-negative MIT materials* | | | | | | |
| Persistent false negative | $NbO_2$ | Other oxide | MIT | 0.251 | 0/5 | 0/5 |
| Persistent false negative | $Pr(P_3Ru)_4$ | Other/intermetallic | MIT | 0.275 | 1/5 | 1/5 |
| Persistent false negative | $K_3(MoO_3)_{10}$ | Other oxide | MIT | 0.358 | 0/5 | 0/5 |
| Persistent false negative | $Ca_2RuO_4$ | Other oxide | MIT | 0.366 | 0/5 | 0/5 |
| Persistent false negative | $Ca_4Mn_3O_{10}$ | Manganite/oxide | MIT | 0.411 | 0/5 | 0/5 |
| *C. High-scoring curated non-MIT materials* | | | | | | |
| High-scoring curated negative | $IrO_2$ | Other oxide | non-MIT | 0.816 | 0/5 | 5/5 |

Continued on next page

Table S27 continued

| Case-study group | Compound | Chemistry family | True label | Mean score | Correct heads | MIT votes |
|---|---|---|---|---|---|---|
| High-scoring curated negative | $V_7O_{13}$ | Vanadate/oxide | non-MIT | 0.811 | 0/5 | 5/5 |
| High-scoring curated negative | $V_2O_5$ | Vanadate/oxide | non-MIT | 0.696 | 0/5 | 5/5 |
| High-scoring curated negative | $TiO_2$ | Other oxide | non-MIT | 0.661 | 3/5 | 2/5 |
| High-scoring curated negative | TiO | Other oxide | non-MIT | 0.612 | 1/5 | 4/5 |
| *D. Quantum–classical disagreement cases* | | | | | | |
| Model disagreement | $DyNiO_3$ | Nickelate/oxide | MIT | 0.610 | 2/5 | 2/5 |
| Model disagreement | $Eu(P_3Ru)_4$ | Other/intermetallic | non-MIT | 0.366 | 3/5 | 2/5 |
| Model disagreement | $Ca(FeO_2)_2$ | Iron oxide | non-MIT | 0.497 | 1/5 | 4/5 |
| Model disagreement | $CoP_3$ | Other/intermetallic | non-MIT | 0.264 | 4/5 | 1/5 |
| Model disagreement | $KVO_3$ | Vanadate/oxide | non-MIT | 0.590 | 1/5 | 4/5 |
| Model disagreement | $LiTi_2O_4$ | Other oxide | non-MIT | 0.644 | 4/5 | 1/5 |
| Model disagreement | $Pr(P_3Ru)_4$ | Other/intermetallic | MIT | 0.275 | 1/5 | 1/5 |
| Model disagreement | $Ti_3O_5$ | Other oxide | MIT | 0.555 | 3/5 | 3/5 |

Table S27 shows that the material-level errors are not confined to a single chemistry family or classifier type. Robust MIT recognition occurs across several oxide families, whereas persistent false negatives such as $NbO_2$ and $Ca_2RuO_4$ are missed by nearly all readouts. Conversely, the unanimous MIT classification of curated negatives such as $IrO_2$ and $V_7O_{13}$ indicates either systematic representation-level false positives or cases requiring closer label and literature review. The disagreement group further shows that individual material decisions can depend strongly on the classifier readout even when aggregate performance is similar.

## References


[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6, 2020.

[3] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

[5] Alexandru B. Georgescu, Peiwen Ren, Aubrey R. Toland, Shengtong Zhang, Kyle D. Miller, Daniel W. Apley, Elsa A. Olivetti, Nicholas Wagner, and James M. Rondinelli. Database, features, and machine learning model to identify thermally driven metal–insulator transition compounds. *Chemistry of Materials*, 33(14):5591–5605, 2021.

[6] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[8] Ivan Tomek. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:769–772, 1976.

[9] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

[10] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18:1–5, 2017.

[11] M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*, pages 51–67. Springer, Dordrecht, 1994.

[12] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006.

[13] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10:e0118432, 2015.

[14] Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 3121–3124, 2010.

[15] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103:032430, 2021.

[16] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12:2631, 2021.

[17] Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13:4919, 2022.

[18] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021.