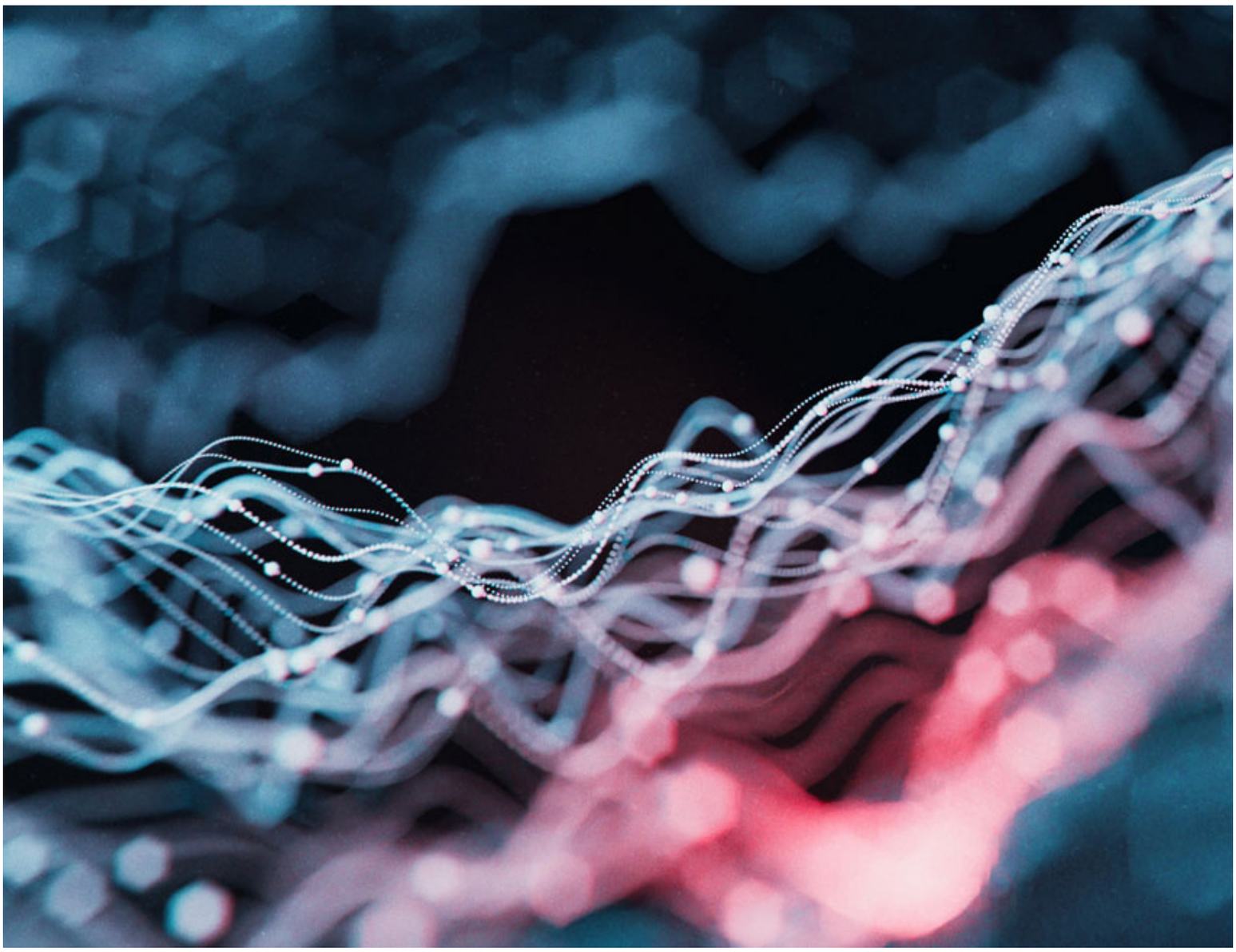
Risk

REPRINTED FROM

Risk.net

Risk.net May 2025



Supervised similarity for high-yield bonds

Supervised similarity for high-yield bonds

Joshua Rosaler, Luca Candelori, Vahagn Kirakosyan, Kharen Musaelian, Ryan Samson, Martin T. Wells, Dhagash Mehta and Stefano Pasquali

apply quantum cognition machine learning (QCML) to distance metric learning for corporate bonds. A measure of similarity is useful for the trading of illiquid bonds, identification of similar tradable alternatives and pricing securities with few recent quotes or trades. QCML for supervised distance metric learning outperforms tree-based models in high-yield markets while giving comparable performance in investment-grade markets

Learning about similarity relationships between securities is central in the application of machine learning to financial contexts. For example, a portfolio manager may wish to include a particular security in a portfolio but then find that there is insufficient liquidity in the market and instead try to find a tradable substitute that is similar to the desired security. Alternatively, similarity can also be used to price securities for which there are no recent trades or quotes in the market by looking at the prices of similar securities with more up-to-date market data. In addition, a measure of similarity between securities can be used for hedging risk. Yet another application is anomaly detection, which may involve identifying atypical securities within a particular cohort as being those that are most dissimilar from the others. Beyond the individual security level, distance metric learning can also be used to assess the similarity of different portfolios; for example, basket trading may require assessing the similarity of an entire basket of securities to a given exchange-traded fund (ETF).

Although similarity or distance metric learning is typically unsupervised, supervised approaches, in which similarity is defined with respect to a target variable, are often more effective in financial applications. For example, a portfolio manager might substitute a hard-to-trade bond with one of a similar spread and risk profile, based on features such as the coupon, rating, sector and maturity. Recent research demonstrates that tree-based models, such as random forests (RFs) or gradient-boosted machines (GBMs), can be trained on a target variable to yield a feature-space distance metric, with proximity implying similar target values (Breiman & Cutler 2001). Financial applications of supervised similarity based on RFs and GBMs have recently been explored, for example by Jeyapaulraj *et al* (2022), Rosaler *et al* (2024) and Li *et al* (2024). Currently, supervised similarity based on RFs represents the state-of-the-art for supervised similarity (Desai *et al* 2024).

A supervised similarity metric can be evaluated in terms of the performance of a k-nearest-neighbours (k-NN) regressor, with the notion of 'nearest' taken from the learned metric. The core idea is that a distance metric is more effective when proximity in feature space aligns with similarity in the target variable. In such cases, averaging the target values of nearest neighbours yields more accurate predictions than using a generic, unsupervised metric.

The emerging field of quantum cognition machine learning (QCML) (Candelori *et al* 2025; Musaelian *et al* 2024; Samson *et al* 2024) offers a new class of machine learning models based on the mathematical formalism of quantum theory. In QCML, individual data points are represented by quantum states, and features and target variables are represented by Hermitian operators or 'observables' that are learned by optimising a particular objective function. Although the objective functions and evaluation metrics of QCML models are largely the same as those used in the training and evaluation of classical machine learning models, the way in which data is represented deviates in fundamental ways. In particular, in this article we introduce a new notion of proximity between data points that arises naturally in QCML by taking the quantum fidelity (ie, the absolute value of the inner product) of two quantum states, each representing a data point.

A QCML model can naturally handle both numerical and categorical data as well as missing and/or noisy data. It creates a global quantum manifold model – in the sense of quantum geometry (Steinacker 2024) – of the original data manifold that is robust to noise and able to effectively generalise beyond training samples (Candelori *et al* 2025; Samson *et al* 2024). Its ability to control variance comes partly from the fact that the number of parameters in a QCML model scales linearly with the number of features, thus achieving logarithmic economy of representation. By way of comparison, in an RF it is the depth of trees that typically scales linearly with the number of features, resulting in an exponential growth in parameters. These different scaling regimes result in profound differences between RF proximity and QCML proximity. For example, we show that, in the presence of a large number of features and a large number of outliers in the data, RF proximity tends to inflate sparsity in the data by placing points as far apart as possible. This is a consequence of the RF model attempting to grow individual tree branches for each outlier, thus overparameterising the data. QCML proximity instead creates a compact representation of the data even in the presence of a large number of outliers. These effects are shown to be particularly dramatic when analysing a cohort of high-yield bonds, which exhibit not only a large number of features, arising from one-hot encoding of categorical variables (bond rating, country and industry), but also a large number of outliers, arising from bonds that are near default. Within this cohort, we demonstrate a significant advantage of QCML proximity over RF proximity, as measured by a k-NN regressor. For control, we also analyse a data set consisting of investment-grade bonds and show that the performance of QCML is similar to that of RF. This is as expected, since investment-grade bonds tend to have a smaller number of outliers than high-yield bonds.

Supervised metric learning with random forests

In tree-based regression models such as RFs and GBMs, the similarity between points can be quantified by the proportion of trees in which they share a leaf. We focus on RFs, given their more established literature and tool support.

Breiman's original definition. Breiman originally proposed a definition of supervised similarity for RFs as the proportion of trees in the RF ensemble for which two points fall in the same leaf node:

$$\mathrm{Prox}(i,j) = \frac{1}{M}\sum_{T=1}^{M}\frac{1}{N_i^T} I[j \in \mathcal{L}_i^T] \tag{1}$$

where I is the indicator function, M is the number of trees in the forest and N_i^T is the number of training points in the leaf $\mathcal{L}_i^T$ of the tree T into which both i and j fall (Brieman & Cutler 2001).

■ **Geometry-and-accuracy-preserving proximities.** Lin & Jeon (2006) showed that any tree-based ensemble model, such as an RF or a GBM, can be viewed as an adaptive weighted k-NN model. That is, the prediction of the RF or GBM for any test point can be expressed exactly as a weighted average of the target labels of points in the training set, with the weights of the training points varying depending on the location of the test point in the feature space. In the context of regression, this result entails that the prediction $\hat{y}_i$ of the model can be expanded locally as a linear combination of target labels in the training dataset:

$$\hat{y}_i = \boldsymbol{k}_i(x) \cdot \boldsymbol{y}_{\text{train}} = k_{i,1}(x) y_{\text{train},1} + \cdots + k_{i,N}(x) y_{\text{train},N} \quad (2)$$

where $y_{\text{train},j}$ is the ground-truth target label for the jth training example, and $k_{i,j}(x)$ is the input-dependent weight of observation j in the expansion for observation i. This weight corresponds to yet another notion of proximity or similarity between points in the feature space, and like the other proximities we have discussed, it depends on the number of trees in the RF ensemble for which two points fall in the same leaf node (Rhodes *et al* 2023).

Rhodes *et al* (2023) showed that for RFs the correct form of the expansion coefficients $k_{ij}(x)$ appearing in (2), which they call the geometry-and-accuracy-preserving (GAP) RF proximity, is:

$$\text{Prox}_{\text{GAP}}(i,j) = \frac{1}{|S_i|} \sum_{t \in S_i} \frac{c_j(t) I[j \in J_i(t)]}{|M_i(t)|} \quad (3)$$

where S_i is the set of trees in the RF for which observation i is out-of-bag, $M_i(t)$ is the multiset of bagged points in the same leaf as i in tree t, $J_i(t)$ is the corresponding set (ie, without repetitions) of bagged points in the same leaf as i in tree t and $c_j(t)$ is the multiplicity of the index j in the bootstrap sample.

Quantum cognition machine learning

QCML (Candelori *et al* 2025; Musaelian *et al* 2024; Samson *et al* 2024) has recently been proposed as a new framework for machine learning models based on quantum cognition (see Pothos & Busemeyer (2022) for a recent survey). QCML models learn a representation of the data as quantum states. In quantum mechanics a 'state' is a vector of unit norm in a Hilbert space and is represented in bra-ket notation by a ket, $|\psi\rangle$. The inner product of two states $|\psi_1\rangle$, $|\psi_2\rangle$ is represented by a bra-ket $\langle\psi_1|\psi_2\rangle$. The expectation value of a Hermitian operator M (ie, a quantum observable) on a state $|\psi\rangle$ is denoted by $\langle\psi|M|\psi\rangle$, representing the expected outcome of the measurement corresponding to M on the state $|\psi\rangle$. In this interpretation the expression $\langle\psi|M|\psi\rangle$ corresponds exactly to the expected value of the discrete random variable given by measuring M on ψ (Nielsen & Chuang 2000, section I.2.2).

In QCML, for each vector $\boldsymbol{x}_t \in \mathbb{R}^K$ belonging to a data set consisting of $t = 1, \ldots, T$ K-dimensional observations, define an error Hamiltonian:

$$H(\boldsymbol{x}_t) = \frac{1}{2} \sum_k (A_k - \boldsymbol{x}_{t,k} \cdot I)^2 \quad (4)$$

depending on a set of N-dimensional quantum observables $\{A_k\}$. In this formula, I denotes the $N \times N$ identity matrix. Each of these K quantum observables can be viewed as a 'quantisation' of a corresponding feature of the original K-dimensional data set. The vector $\boldsymbol{x}_t$ is mapped to the ground state $|\psi_t\rangle$ of the error Hamiltonian (ie, the eigenstate associated with the lowest eigenvalue), giving a representation of data as quantum states. Conversely, given an arbitrary N-dimensional quantum state $|\psi\rangle$, we can define its 'position' to be the following K-dimensional vector:

$$\boldsymbol{x}(\psi) = (\langle\psi|A_k|\psi\rangle)_k \in \mathbb{R}^K$$

which in quantum theory represents the expected outcome of measuring the quantum observables A_k on ψ. In this way, given a set of quantum observables $\{A_k\}$, we have a way to map data into quantum states by sending $\boldsymbol{x}_t$ to its ground state $|\psi_t\rangle$, and we can also retrieve information about a quantum state $|\psi\rangle$ by taking its position $\boldsymbol{x}(\psi)$. In an unsupervised setting, training a QCML model involves iterative updates to the observables $\{A_k\}$ so that the ground states $|\psi_t\rangle$ 'cohere' with the data; that is, the distance between $\boldsymbol{x}_t$ and its position $\boldsymbol{x}(\psi_t)$ is minimised, as well as the variance of the measurement. Coherence can be achieved by minimising this distance directly, or by minimising the overall energy of the error Hamiltonian, as discussed in detail by Candelori *et al* (2025).

In a supervised setting, which is the primary focus of this article, the training process is slightly different (Samson *et al* 2024). The target variable $y \in \mathbb{R}$ is assigned an N-dimensional quantum 'forecast' observable B, and given a data point $\boldsymbol{x}_t$, the corresponding forecast is given by:

$$\hat{y}_t = \langle\psi_t|B|\psi_t\rangle$$

During the training process, the quantum observables $\{A_k\}$ and B are updated at each iteration to minimise the mean absolute error $\sum_t |\hat{y}_t - y_t|$. The nondifferentiability of the loss function here does not add further complexity to the algorithm, since the mapping from $\boldsymbol{x}_t$ to its ground state ψ_t is already nondifferentiable, with the nondifferentiable points corresponding to the locus of degeneracy of the error Hamiltonian (4). This setup can easily be extended to the case of multiple target variables. The training algorithm can be summarised as follows.

Algorithm 1 QCML univariate regression model training

- Randomly initialise feature operators $\{A_k\}$ and target operator B
- Iterate over training data and operators until the desired convergence:
  1: generate error Hamiltonian $H(\boldsymbol{x}_t)$
  2: holding A_k constant, find the ground state $|\psi_t\rangle$ of $H(\boldsymbol{x}_t)$
  3: calculate the gradients of the loss function $\sum_t |\hat{y}_t - y_t|$ with respect to A_k and B
  4: update A_k and B via gradient descent

The specifics of each of these steps will depend on the choice of parameterisation for the operators A_k and B. For this article, each operator is parameterised as a sum $M_1 + \mathrm{i}M_2$, where M_1 (respectively, M_2) is a real symmetric (real antisymmetric) matrix. The dimension of the Hilbert space, N, is a hyperparameter of the algorithm and can be optimised using cross-validation. Larger values of N typically reduce loss but could lead to overfitting and worse performance outside of the sample, while lower dimensions tend to have higher bias and lower variance (Candelori *et al* 2025). A typical choice for many applications would be $N \approx 8$.

Note that the Hamiltonian (4) and its ground states originally appeared in the field of theoretical physics known as quantum geometry (Steinacker 2024).

■ **QCML proximity.** There is a natural notion of proximity for quantum states given by 'quantum fidelity' (Nielsen & Chuang 2000, section III.9):

$$f(\psi_1, \psi_2) = |\langle\psi_1|\psi_2\rangle|^2$$

which can be interpreted as the probability of identifying the state ψ_1 (resp. ψ_2) with the state ψ_2 (resp. ψ_1) when performing a quantum measurement designed to test whether a given quantum state is equal to ψ_2 (resp. ψ_1).

In the context of QCML this type of proximity can be used to define a similarity measure for the data. Suppose that a (supervised) QCML model has been trained with feature observables A_k and target observables B. We then have a representation of the data in quantum states given by $\boldsymbol{x}_t \to |\psi_t\rangle$. Given two data points $\boldsymbol{x}_t, \boldsymbol{x}_{t'}$, define their 'QCML distance' by:

$$d_Q(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) = 1 - f(\psi_t, \psi_{t'}) = 1 - |\langle\psi_t|\psi_{t'}\rangle|^2 \qquad (5)$$

Note that the square root of this function satisfies all the axioms required by a 'distance' in the strict mathematical sense and corresponds to the trace distance when the quantum states ψ are expressed in terms of their density matrix $|\psi\rangle\langle\psi|$. For Hilbert spaces of very high dimension N, this distance tends to be infinitesimally small, since any two states will tend to be orthogonal. But in practical QCML applications, as noted above, the Hilbert space dimension tends to be very small, so this is not an issue.

Unlike standard Euclidean distance, QCML proximity defines a supervised similarity measure, with quantum states ψ_t optimised using training targets. Like RF proximity, QCML proximity can handle both numerical and categorical features.

Data description

We selected two data sets of corporate bonds, consisting of the holdings of
(1) the iShares iBoxx $ High Yield ETF (ticker: HYG) and
(2) the iShares 1-5 Year Investment Grade ETF (ticker: IGSB).
The data features selected are cross-sectional, and they were downloaded at end of day on September 18, 2024. They consist of seven numerical variables (coupon, coupon frequency, days to maturity, duration, age, amount issued and amount outstanding) and three categorical variables (rating, country and industry), with the target variable being the bond yield for HYG and the credit spread for IGSB. We use one-hot encoding for the categorical variables, obtaining a total of $K = 99$ features.

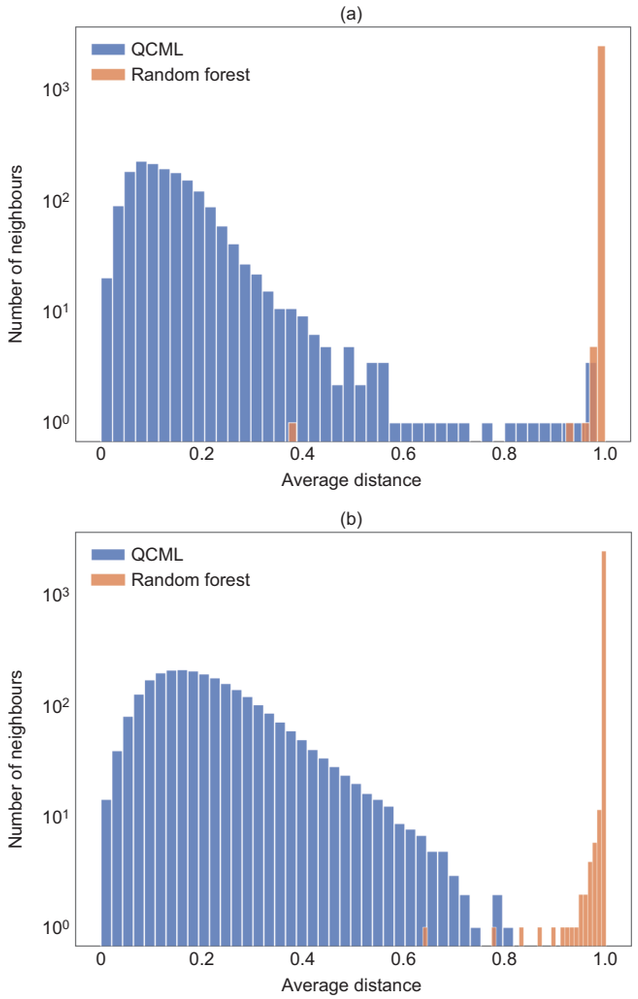
We select the credit spread as the prediction target for IGSB since investment-grade bonds are typically priced and traded based on the spread rather than the absolute yield. In contrast, for HYG we use the yield of the bond as the prediction target. Given the elevated risk and volatility associated with high-yield bonds, it is less meaningful to compare their yields to those of Treasuries.

There are significant differences in distribution between the HYG and IGSB data. The distributions of bond yield and spread tend to have wider support and a larger number of outliers for high-yield bonds compared with investment grade.

Methods

We now describe how to train and evaluate the QCML model on the regression tasks described above (yield prediction for HYG and spread prediction for IGSB), and how to calculate and evaluate the supervised QCML distance between data points.

1 Average distance from a reference bond to its neighbours for (a) HYG and (b) IGSB



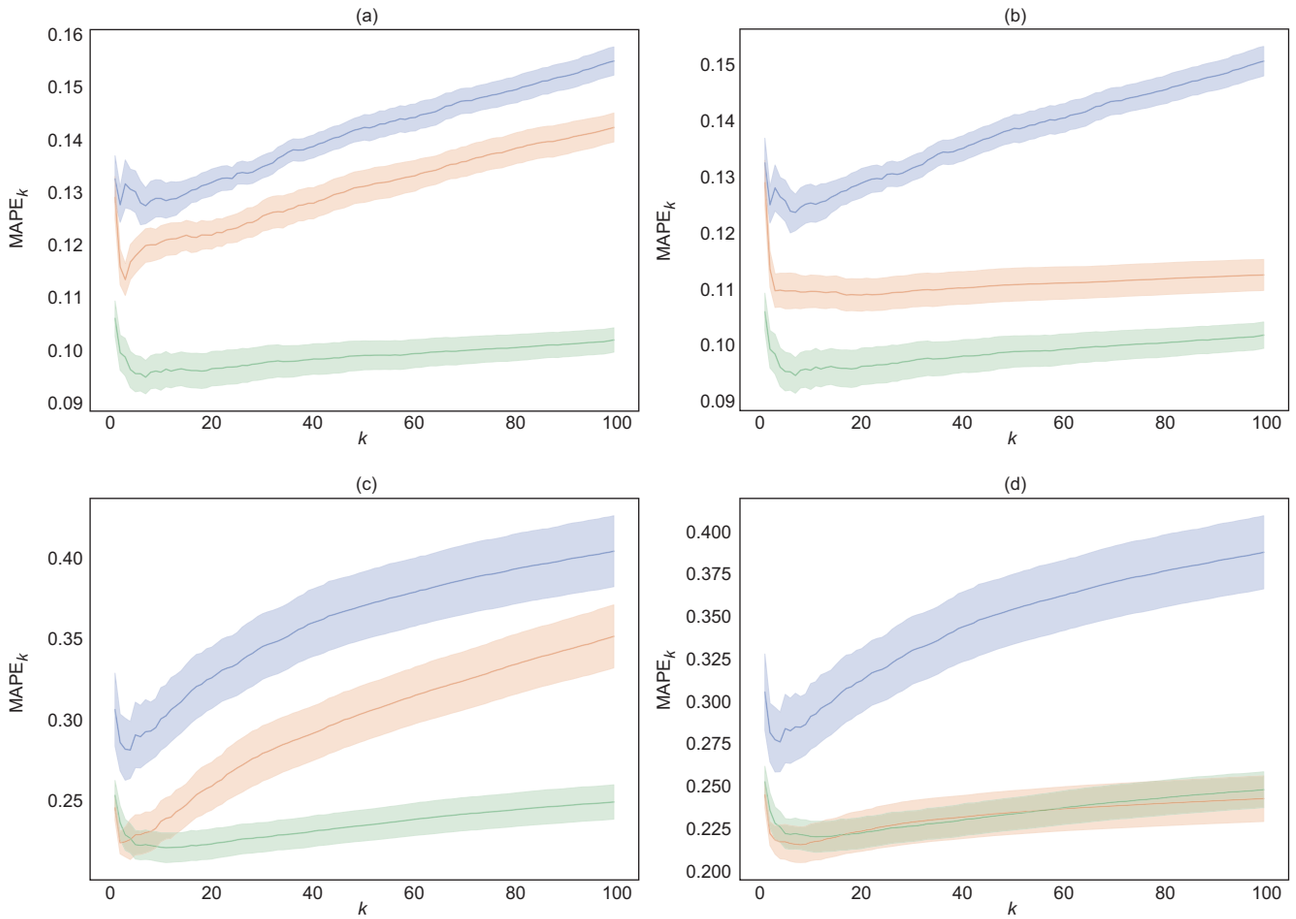
The average is taken over all bonds in the data set. For RF (GAP proximity), almost all neighbours are at the maximum distance of 1.0 from the reference bond

■ **Train-test split and hyperparameter optimisation.** We first performed a random 80/20 train/test split of the data and used the training data for hyperparameter optimisation. For the QCML model, we optimised the Hilbert space dimension using a threefold cross-validation procedure, minimising the mean squared error (MSE). We obtained an optimal Hilbert space dimension of $N = 7$ for HYG and $N = 12$ for IGSB.

For comparison, we also did a hyperparameter search for RF, varying the number of trees, maximum depth, minimum number of samples per leaf, maximum number of features considered in each split and training objective function. The optimal choice of parameters for HYG was 1000 trees, a maximum depth of 50, a minimum of one sample per leaf, a maximum number of features considered in each split equal to the square root of the total number of features and a training objective equal to the mean absolute error (MAE). For IGSB, the results were the same, except for a choice of 200 trees and a training objective equal to the MSE.

■ **Similarity matrix evaluation.** To evaluate the quality of the QCML similarity measure, we ran the k-NN regression method using the QCML

2 Average MAPE for k-NN regression over 10 train/test splits of the data, for HYG (parts (a) and (b)) and IGSB (parts (c) and (d))



The bands represent the standard error in estimation of the mean for each k. In (a) and (c), the k-NN prediction is computed as a simple average of the training targets for the k nearest neighbours. In (b) and (d), the k-NN prediction is computed as a proximity-weighted average of the k nearest neighbours. In all figure parts, the colours are as follows: blue, Euclidean; orange, RF GAP; and green, QCML

A. Average test set metrics and standard deviations for target variable predictions for the HYG and IGSB data sets

(a) HYG: yield prediction				
	MAPE	**MAE**	**RMSE**	$\boldsymbol{R^2}$
Linear regression	0.13 ±0.007	1.06 ±0.08	1.86 ±0.19	0.59 ±0.06
RF	0.11 ±0.007	0.93 ±0.08	1.80 ±0.19	0.62 ±0.04
QCML	**0.09** ±0.01	**0.79** ±0.06	**1.73** ±0.17	**0.65** ±0.07
(b) IGSB: credit spread prediction				
	MAPE	**MAE**	**RMSE**	$\boldsymbol{R^2}$
Linear regression	0.26 ±0.03	15.90 ±0.48	22.83 ±1.40	0.69 ±0.01
RF	0.25 ±0.04	**13.74** ±0.47	**20.30** ±1.28	**0.76** ±0.02
QCML	**0.23** ±0.05	13.87 ±0.49	20.78 ±1.08	0.74 ±0.02

The average and standard deviations are taken over 10 different 80/20 train/test splits of the data

distance, the RF GAP distance and the standard Euclidean distance. The idea is that a good measure of distance should bring bonds of similar yield or spread closer, and it should separate bonds with differing yields or spreads, which translates into better performance under the k-NN method.

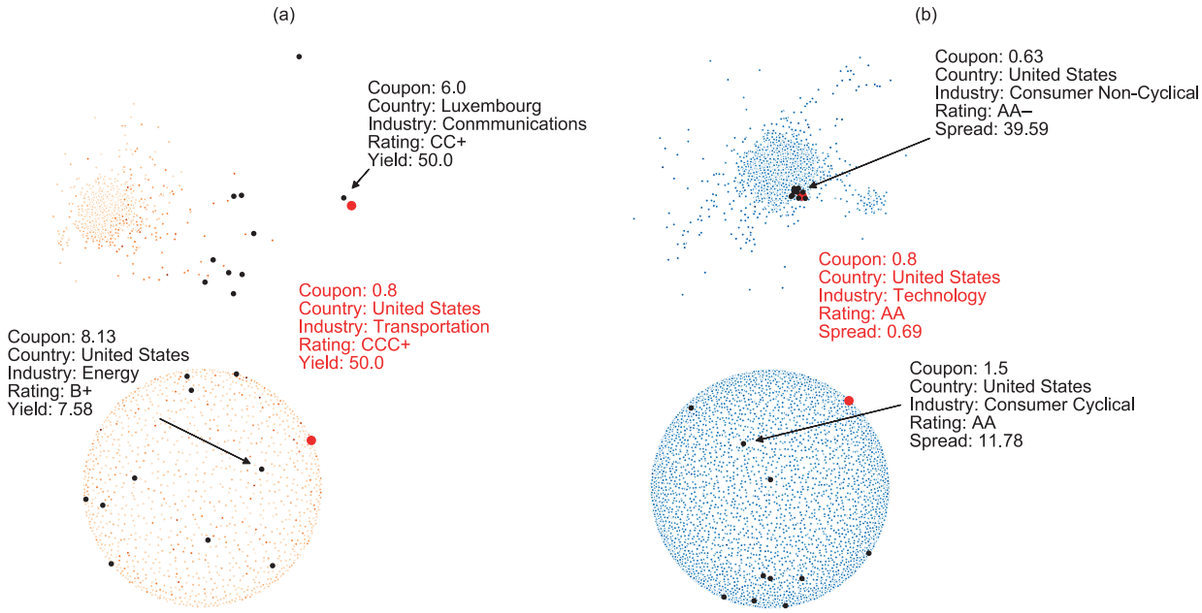
For the QCML metric, we ensembled three different distance matrixes, corresponding to QCML models with different initialisation weights, by taking the average of each entry, which is given by (5). For RF-based metrics, the proximities are given directly by (2). For the Euclidean metric, the proximity is defined as $1 - d$, where d is the Euclidean distance normalised by its maximum training value.

There is a subtlety in how to weigh the contributions of the k nearest neighbours when computing the k-NN prediction. For the use case of identifying tradable substitutes, we should place greater weight on bonds with higher proximity, which will have more similar values for the most important input features.

For the RF GAP and QCML proximities, it is most straightforward to associate this weight with the proximity itself. In the specific case of the RF GAP proximities, these weights have the further compelling interpretation that they are exactly equal to the proportional contribution of each bond in the training set to the RF prediction for the bond of interest.

However, this interpretation of proximity is not available for any of the other metrics, which obstructs a like-for-like comparison between the various proximity-weighted k-NN predictions for different metrics. In fact, basing the weights in the k-NN average on model proximity makes them strongly model-dependent. This point is reinforced by the graphs in figure 1, which show the distributions of the average distance for both QCML and GAP

3 MDS visualisation of QCML proximities (top) and RF GAP proximities (bottom), for (a) HYG and (b) IGSB



Darker colours represent higher yields. In all plots, a reference bond is displayed in red and the 10 closest neighbours are displayed in black

proximities. With the QCML metric, distances tend to be less than 0.5, while for RF, they tend to be greater than 0.9, which means that for RF only a relatively small number of bonds contribute substantially to the proximity-weighted k-NN prediction, while for QCML, a substantially broader region of neighbours contributes.

Given the element of arbitrariness that enters into the choice of a weighting scheme to calculate the k-NN prediction, in figure 2 we plot the performance as a function of k both without weighting by proximity (to show what the results look like in the absence of a weighting scheme) and with weighting.

Results

Our main goal in this section is to evaluate the quality of the similarity matrixes extracted from QCML compared with those extracted from the standard Euclidean metric and from state-of-the-art supervised similarity learning with RFs.

■ **Target variable prediction.** We first test the QCML model on the regression task to verify its prediction performance. We repeat the test over 10 different 80/20 train/test splits of the data, with the data randomly shuffled before each split. The regression metrics tracked are the mean absolute percentage error (MAPE), MAE, root mean squared error (RMSE) and R^2. The regression results for QCML are compared with those of linear regression (as a baseline) and RF. The results are shown in table A.

Our focus is on evaluating the similarity matrixes derived from models such as QCML or RF, not their predictive performance. However, as shown in the examples below, stronger regression performance directly improves the quality of the extracted similarity metric, as measured by the k-NN method.

For example, in the case of HYG, we see that the superior performance of QCML as a yield predictor, shown in part (a) of table A, also translates into better results for the associated similarity metric, shown in parts (a) and (b) of figure 2. In contrast, in the case of IGSB, the fact that the performance of QCML as a predictor of spread is closer to that of RF translates into a comparable performance for the associated similarity metrics, particularly in the case of proximity weighting, shown in figure 2(d); for the unweighted k-NN curve in figure 2(c), QCML still performs better for the most part.

■ **Similarity measure evaluation.** Next, we compute the similarity matrixes and apply the k-NN method to evaluate each metric. Three distance metrics are evaluated: standard (unsupervised) Euclidean, RF GAP and QCML. For each distance metric we report the test MAPE of a k-NN regressor for a range of neighbours $k = 1, \ldots, 100$, first in the case where the k-NN average is unweighted, and then in the case where the k-NN average is weighted by proximity. The k-NN method for evaluating similarity is performed for bonds in the HYG and IGSB indexes.

For the corporate bonds dataset we see in figure 2 that, for HYG, the QCML metric outperforms other metrics for the unweighted and proximity-weighted methods of k-NN regression. For IGSB, we see that the QCML metric again outperforms other metrics for the unweighted k-NN regression and performs comparably, if slightly worse, for the proximity-weighted k-NN regression. As discussed, RF-based proximities tend to be at or near the minimum value of 0 for all but a relatively small proportion of training points; as a result, these proximities tend to place more weight on a smaller number of nearest neighbours than QCML proximities, which have support over a larger proportion of the training set. In this respect, proximity weighting inherently favours RF-based proximities, as it counts only a small number of nearest neighbours that are more likely to have yields close to those of the test point.

The most significant performance difference between RF and QCML is found for the high-yield bonds (parts (a) and (b) of figure 2). In RF regression, imbalanced data can cause issues; since RF regression takes the average of predictions from multiple decision trees, if one part of the dataset dominates, then the model's predictions may be biased towards the majority region

of the target variable. The model may struggle to predict extreme values if the target variable is highly skewed, because few trees are trained on those cases. If most trees are trained on similar value ranges (due to imbalanced data), the overall variance of predictions is reduced, making the model less flexible in capturing minority patterns. Consequently, RF regression struggles with imbalanced target distributions because it averages predictions, leading to poor performance on rare values. In contrast, as we illustrate visually in the next section, QCML is well suited to handling imbalanced target distributions, since its predictions are based on a faithful representation of the underlying data manifold, including possibly sparse regions of the data.

Visualising bond similarities

We next visualise both QCML and RF GAP proximity metrics using multi-dimensional scaling (MDS). This technique can be applied using an arbitrary distance matrix, and it is therefore suitable for visualising both the QCML and RF GAP proximities (figure 3).

The two-dimensional MDS plots help illustrate qualitatively the differences between QCML and RF GAP proximities. For example, the MDS plots for RF resemble a disc, with a higher density of points near the boundary, a consequence of RF GAP proximity placing most points at a maximum distance from each other (already noted in figure 1).

MDS can also be used to visualise and compare the rankings of the top k neighbours of an individual bond, for both QCML and RF proximity. In figure 3(a), a HYG reference bond is plotted along with its top 10 neighbours according to QCML and RF proximity. This bond has a 50% yield, making it an outlier that lies in a particularly sparse region of the data set. This sparsity is clear in the MDS plot of the QCML proximities (top of figure 3(a)). However, among the top 10 nearest neighbours, QCML also identifies a similar bond with a 50% yield. In contrast, most of the 10 neighbours for RF are relatively far from the reference bond, resulting in higher yield-prediction error (RF predicts a yield of 18.17%, versus 33.04% for QCML). In figure 3(b), the same experiment is repeated for IGSB. In this case, the reference bond chosen has very low spread and is situated in the central core of the data. For this example, the spread approximation for RF (11.2 basis points) is better than that of QCML (24.3 basis points).

In general, we can expect QCML to outperform within a cohort of bonds that lie in sparse patches of the data. This is the source of the advantage of QCML proximity that was noted earlier for high-yield bonds.

Conclusions

This study has shown that QCML, a novel paradigm for machine learning, performs comparably to, and in some cases better than, traditional machine learning methods in the context of supervised similarity. This was illustrated by identifying tradable alternatives for high-yield corporate bonds in the HYG index, with the alternative bonds indicated by the QCML metric being substantially closer in yield to the desired bond than those indicated by distance metrics based on RFs or Euclidean distance. ■

Joshua Rosaler is a quantitative researcher at BlackRock, Inc. in New York. Luca Candelori is the director of research at Qognitive, Inc. and an associate professor in the department of mathematics at Wayne State University. He is based in Miami Beach, FL. Vahagn Kirakosyan is the director of quantitative analytics, Kharen Musaelian is the co-founder and chief science officer and Ryan Samson is the director of research, all at Qognitive, Inc. in Miami Beach, FL. Martin T. Wells is the Charles A. Alexander Professor of Statistical Sciences in the department of statistics and data science at Cornell University in Ithaca, NY. Dhagash Mehta is the head of applied AI for investment management and Stefano Pasquali is the managing director and head of investment AI modeling and research, both at at Blackrock, Inc. in New York. The views expressed here are those of the authors alone and not of BlackRock, Inc. The contents of this article do not constitute investment advice.
Email: joshua.rosaler@blackrock.com, luca.candelori@qognitive.io, vahagn.kirakosyan@qognitive.io, kharen@qognitive.io, ryan.samson@qognitive.io, mtw1@cornell.edu, dhagash.mehta@blackrock.com, pacca.pasquali@gmail.com.

REFERENCES

Brieman L and A Cutler, 2001
Random forests
https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htm, 2001

Candelori L et al, 2025
Robust estimation of the intrinsic dimension of data sets with quantum cognition machine learning
Scientific Reports 15(1):6933, 2025

Desai D, D Mehta and J Urquidi, 2024
rfproximity: a Python package for rf proximity analysis, 2024
https://pypi.org/project/rfproximity

Jeyapaulraj J, D Desai, D Mehta, P Chu, S Pasquali and P Sommer, 2022
Supervised similarity learning for corporate bonds using random forest proximities
In Proceedings of the Third ACM International Conference on AI in Finance (pages 411--419)

Li M, B Sarmah, D Desai, J Rosaler, S Bhagat, P Sommer and D Mehta, 2024
Quantile regression using random forest proximities
arXiv preprint arXiv:2408.02355

Lin Y and Y Jeon, 2006
Random forests and adaptive nearest neighbors
Journal of the American Statistical Association 101(474):578--590

Musaelian K et al, 2024
Quantum cognition machine learning: AI needs quantum
https://www.qognitive.io/QCML%20-%20Qognitive,%20Inc.pdf

Nielsen MA and IL Chuang, 2000
Quantum Computation and Quantum Information
Cambridge University Press

Pothos EM and JR Busemeyer, 2022
Quantum cognition
Annual Review of Psychology 73(1):749--778

Rhodes JS, A Cutler and KR Moon, 2023
Geometry- and accuracy-preserving random forest proximities

Rosaler J, D Desai, B Sarmah, D Vamvourellis, D Onay, S Pasquali and D Mehta, 2024
Enhanced local explainability and trust scores with random forest proximities
In Proceedings of the 5th ACM International Conference on AI in Finance (pages 521--529)

Samson R et al, 2024
Quantum cognition machine learning: financial forecasting
Risk.net, https://www.risk.net/7960053

Steinacker HC, 2024
Quantum Geometry, Matrix Theory, and Gravity
Cambridge University Press