

Supervised Similarity for Firm Linkages

Ryan Samson^{1, *},
Adrian Banner², Luca Candelori^{1, 3}, Sebastien Cottrell⁴, Tiziana Di Matteo^{5, 6, 7},
Paul Duchnowski², Vahagn Kirakosyan¹, Jose Marques², Kharen Musaelian^{1, 8},
Stefano Pasquali⁹, Ryan Stever², and Dario Villani^{1, 5, 8}

¹Qognitive, Inc., Miami Beach, FL

²Intech Investment Management LLC, West Palm Beach, FL

³Wayne State University, Department of Mathematics, Detroit, MI

⁴Deutsche Bank, New York, NY

⁵King's College London, Department of Mathematics, London, UK

⁶Complexity Science Hub Vienna, Vienna, Austria

⁷Centro Ricerche Enrico Fermi, Rome, Italy

⁸Duality Group, Miami Beach, FL

⁹Domyn, New York, NY

*Primary Investigator (ryan.samson@qognitive.io), remaining authors presented in alphabetical order.

The authors also wish to thank Hongli Lan for his contributions to this article.

The views expressed here are those of the authors alone and not of Deutsche Bank, Domyn, Duality Group, Intech Investment Management, or Qognitive.

The contents of this article do not constitute investment advice.

June 9, 2025

Abstract

We introduce a novel proxy for firm linkages, Characteristic Vector Linkages (CVLs). We use this concept to estimate firm linkages, first through Euclidean similarity, and then by applying Quantum Cognition Machine Learning (QCML) to similarity learning. We demonstrate that both methods can be used to construct profitable momentum spillover trading strategies, but QCML similarity outperforms the simpler Euclidean similarity.

1 Introduction

Prior literature has explored the use of fundamental information as a proxy for firm linkages. If investors have limited attention, then news impacting the price of a firm may only slowly be incorporated into prices of related firms, leading to return predictability across firms. Indeed, for many such firm linkages it has been shown that lagged returns of a firm are predictive of future returns for firms which are more similar to it. This effect is sometimes referred to as a momentum spillover effect, or a lead-lag strategy. Momentum spillover has been documented for similarities formed from a variety of fundamental information including industry [24], supply chain [12], analyst coverage [1], and geography [32], among others.

Unrelated literature explores the application of machine learning techniques to the learning of similarity relations between securities, often with the goal of clustering securities for risk management, signal generation, or portfolio construction. See e.g. the literature review in [37] for examples of classification and clustering techniques, [44] for a demonstration of how embeddings from Large Language Models can be used to extract company similarity relations, or [6] for a more general review of machine learning applications in finance.

More recent work has begun to explore the use of supervised learning techniques to extract similarity relationships. [21] proposed a supervised similarity framework based on extracting distance metrics from Random Forest, and demonstrated an application to clustering corporate bonds. [35] extended work on the corporate bond clustering problem, demonstrating the application of

Quantum Cognition Machine Learning [25, 38, 10], a novel paradigm for both supervised and unsupervised learning tasks rooted in the mathematical formalism of quantum theory, to distance metric learning.

Here, we introduce a novel proxy for firm linkages, which we name Characteristic Vector Linkages. Distance based on these linkages can most simply be derived from the Euclidean distance of a vector of firm characteristics, which can then be used to form a profitable momentum spillover trading strategy. Next, we demonstrate how an approach similar to that discussed in [35] can be used to learn distance relationships across equity securities, which can in turn be used to enhance the momentum spillover trading strategy derived from the simpler Euclidean distance.

2 Characteristic Vector Linkages and Euclidean Distance

Academic and practitioner literature has identified a large number of stock characteristics which are widely used as cross-sectional predictors of risk or returns¹. The values of these characteristics, or factors, give us small but important pieces of information about the underlying firms. We propose that, across a broad set of characteristics, the more similar the factor scores of any pair of firms, the more economically similar the underlying pair of firms truly is, and the more likely they are to display momentum spillover effects. In other words, similarity of sets of factor scores can be used to measure the strength of firm linkages. We name such purported factor / characteristic derived linkages Characteristic Vector Linkages (CVLs).

Define a characteristic observed at time t as $x_{t,j}^c$, where $c = 1, 2, \dots, C$ is the index of a characteristic and $j = 1, 2, \dots, J$ is the index of a firm. $\mathbf{x}_t^c \in \mathbb{R}^J$ is the vector over all firms for a single characteristic, $\mathbf{x}_{t,j} \in \mathbb{R}^C$ is the vector over all characteristics for a single firm, and $\mathbf{x}_t \in \mathbb{R}^{J \times C}$ is the matrix over firms and characteristics (we use similar index-omission notation throughout this article). Distance of factor scores could be defined in a variety of ways, but can simply and naturally be expressed as the Euclidean distance between two firm characteristic vectors. The Euclidean distance between firms i and j at time t is:

$$D_{Euclidean}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = \sqrt{(\mathbf{x}_{t,i} - \mathbf{x}_{t,j})^\top (\mathbf{x}_{t,i} - \mathbf{x}_{t,j})} \quad (1)$$

which we can then transform to a similarity measure on $[0, 1]$ as will be explained in section 4.

Euclidean distance is a natural choice in this context, as the magnitude of the factor scores and their differences carry meaning. By contrast, consider a normalized distance metric such as cosine distance:

$$D_{cosine}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = 1 - \frac{\mathbf{x}_{t,i} \cdot \mathbf{x}_{t,j}}{\|\mathbf{x}_{t,i}\| \|\mathbf{x}_{t,j}\|} \quad (2)$$

where input vectors are scaled by their norms. To understand why this is problematic, consider the extreme case of firm i with $\mathbf{x}_{t,i}^c = \epsilon \forall c$ and firm j with $\mathbf{x}_{t,j}^c = -\epsilon \forall c$. These firms have nearly identical factor scores and thus our distance metric should return a small value. In this case, equation 1 gives $D_{Euclidean}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = 2\epsilon\sqrt{C}$, a small value, while equation 2 gives $D_{cosine}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = 2$, the maximum distance possible.

3 Quantum Cognition Machine Learning Distance

3.1 Quantum Cognition Machine Learning Review

Quantum Cognition Machine Learning (QCML) [25, 38, 10] has recently been proposed as a new form of machine learning based on quantum cognition (see [33] for a recent survey of quantum cognition). QCML models learn a representation of the data into quantum states. Recall that 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 O (i.e. a quantum observable)

¹See e.g. [18] or [20], both of which evaluate the robustness of various features proposed in the financial literature, and also serve as excellent surveys of such features / characteristics

on a state $|\psi\rangle$ is denoted by $\langle\psi|O|\psi\rangle$, representing the expected outcome of the measurement corresponding to O on the state $|\psi\rangle$.

For QCML, we use $|\psi\rangle \in \mathbb{C}^N$, operators $O \in \mathbb{C}^{N \times N}$ and the standard Hermitian inner product, so kets correspond to column vectors, and bras to their conjugate transpose. Combinations of bras, kets, and operators are then interpreted using standard matrix multiplication rules. Hereafter, we omit the ket from ψ unless part of such an expression.

QCML starts by defining an error Hamiltonian over data and observables. We have flexibility in how we define the Hamiltonian, so long as the result is non-negative and Hermitian. In this article we have defined the error Hamiltonian as:

$$H(\mathbf{x}_{t,j}, \{A_c\}) = \sum_c (A_c - x_{t,j}^c \cdot I)^2 \quad (3)$$

where $\mathbf{x}_{t,j}$ is as defined in section 2, $\{A_c\} \in \mathbb{C}^{N \times N}$ is a set of C Hermitian observable operators which must be learned, and I denotes the $N \times N$ identity matrix.

Each of the C quantum observables $\{A_c\}$ can be viewed as a “quantization” of a corresponding feature of the original C -dimensional data set. The vector $\mathbf{x}_{t,j}$ is then mapped to the ground state $\psi_{t,j}$ of the error Hamiltonian,² giving a representation of data into quantum states (i.e., unit norm vectors in a complex Hilbert space). Conversely, given an arbitrary N -dimensional quantum state $\psi_{t,j}$, we can define its “position” to be the C -dimensional vector:

$$\hat{\mathbf{x}}_{t,j} = (\langle\psi_{t,j}|A_c|\psi_{t,j}\rangle)_c \in \mathbb{R}^C, \quad (4)$$

which in quantum theory represents the expected outcome of measuring the quantum observables A_c on $\psi_{t,j}$. In this way, given a set of quantum observables $\{A_c\}$, we have a way to map data into quantum states by sending $\mathbf{x}_{t,j}$ to its ground state $\psi_{t,j}$, and we can also retrieve information about a quantum state $\psi_{t,j}$ by taking its position $\hat{\mathbf{x}}_{t,j}$.

In an unsupervised setting, training a QCML model involves iterative updates to the estimated observables $\{\hat{A}_c\}$ so that the ground states $\psi_{t,j}$ of the error Hamiltonian “cohere” to the data. That is, the distance between $\mathbf{x}_{t,j}$ and its position $\hat{\mathbf{x}}_{t,j}$ is minimized, as well as the variance of the measurement. The overall ground state energy of the error Hamiltonian (with H defined in equation 3):

$$E_0(\mathbf{x}_{t,j}, \{\hat{A}_c\}) = \langle\psi_{t,j}|H|\psi_{t,j}\rangle \quad (5)$$

is equivalent to the sum of this distance and the variance of the measurement. See [10] for a detailed discussion of this decomposition. As discussed in [38], “targets” can be included in the set of C operators, and ground states and expectations can be taken over any subset of operators once the model is trained.

In the supervised setting, which is the primary focus of this article, the training process is slightly different. The target variable $y_{t,j} \in \mathbb{R}$ is assigned an N -dimensional quantum “forecast” observable B , and given a data vector $\mathbf{x}_{t,j}$ the corresponding forecast is given by:

$$\hat{y}_{t,j} = \langle\psi_{t,j}|B|\psi_{t,j}\rangle. \quad (6)$$

During the training process, the estimated quantum observables $\{\hat{A}_c\}$ and $\hat{B}$ are updated at each iteration so as to minimize the mean squared error $\sum_{t,j} (\hat{y}_{t,j} - y_{t,j})^2$. This can easily be extended to the case of multiple target variables, and we can also add to the loss any of the components of $E_0(\mathbf{x}_{t,j}, \{\hat{A}_c\})$ which we also wish to minimize. For the results in this article we use:

$$Loss_{t,j} = f(y_{t,j}, \mathbf{x}_{t,j}, \{\hat{A}_c\}, \hat{B}, w) = (\hat{y}_{t,j} - y_{t,j})^2 + w * \sum_c (\hat{x}_{t,j}^c - x_{t,j}^c)^2 \quad (7)$$

where w is a hyperparameter influencing how strongly the ground states cohere to the input data. This form of the loss is not essential - for example, our empirical results in subsequent sections were similar, albeit slightly weaker, with $w = 0$ in equation 7 or with the unsupervised approach (minimizing equation 5, with target included in $\{A_c\}$).

QCML univariate regression model training with input bias penalty

- Randomly initialize estimated feature operators $\{\hat{A}_c\}$ and target operator $\hat{B}$.
 - Iterate over training data until desired convergence:
 - 1: Generate error Hamiltonian $H(\mathbf{x}_{t,j}, \{\hat{A}_c\})$ (equation 3)
 - 2: Holding $\{\hat{A}_c\}$ constant, find the ground state $\psi_{t,j}$ of $H(\mathbf{x}_{t,j}, \{\hat{A}_c\})$
 - 3: Calculate gradients w.r.t $\{\hat{A}_c\}$ and $\hat{B}$ of the loss function (equation 7): $\sum_{t,j} Loss_{t,j}$
 - 4: Update $\{\hat{A}_c\}$ and $\hat{B}$ via gradient descent
-

The training algorithm can be summarized as follows:

The specifics of each of these steps will depend on the choice of parametrization for the operators $\{A_c\}$ and B . For results in this article, each operator is parameterized as a sum $O_1 + iO_2$, where O_1 (respectively, O_2) is a real symmetric (real antisymmetric) matrix. The dimension of the Hilbert space N is a hyperparameter of the algorithm. Larger values of N typically lower the loss, but could lead to overfitting and worse out-of-sample performance, while lower dimensions tend to have higher bias and lower variance [10]. Results in this article were obtained using $N = 12$, choices of $N \in [4, 32]$ have been seen to give optimal cross-validated accuracy for a wide variety of learning problems.

3.2 QCML Distance

As explained in section 3.1, QCML represents each observation as a quantum state. This allows one to have a natural notion of proximity between observations, since proximity between quantum states can be defined as *quantum fidelity* [27, III.9]

$$f(\psi_1, \psi_2) = |\langle \psi_1 | \psi_2 \rangle|^2, \quad (8)$$

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

Fidelity can be transformed to distance in a number of ways. Here, we choose to use the Bures distance defined in [42]:

$$D_{QCML}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = \sqrt{2 - 2 \cdot \sqrt{f(\psi_{t,i}, \psi_{t,j})}} = \sqrt{2 - 2 \cdot |\langle \psi_{t,i} | \psi_{t,j} \rangle|} \quad (9)$$

where vectors $\mathbf{x}_{t,i}, \mathbf{x}_{t,j}$ are mapped to ground states $\psi_{t,i}, \psi_{t,j}$ as explained in section 3.1. We chose the Bures distance for its similarity to Euclidean distance, noting that all ground states ψ are unit norm. Empirical results are not particularly sensitive to this choice of distance. For example, we obtained nearly identical results using the geodesic distance defined in [8] as: $\arccos |\langle \psi_{t,i} | \psi_{t,j} \rangle|$. We could also use fidelity directly as a similarity measure, rather than converting states to a distance and then converting distance to similarity. However, directly using these probabilities results in similarities which are too diffuse. Making these transformations allows us to form similarities with distributions comparable to those obtained using the Euclidean distance measure, allowing for a more fair comparison of the two.

4 Converting Distance to Similarity

Given a notion of distance, we form a similarity matrix following a similar approach to that outlined in [26] for affinity matrices:

$$S(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = \begin{cases} e^{-\gamma \cdot D(\mathbf{x}_{t,i}, \mathbf{x}_{t,j})^2} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

²The ground state $\psi_{t,j}$ is the eigenstate associated to the lowest eigenvalue of the error Hamiltonian, and is also called the quasi-coherent state

Using this definition and our distance definitions in equations 1 and 9, we can compute $J \times J$ similarity matrices by date, for all i, j firm pairs.

Note that, in contrast to the Euclidean similarity between two vectors, the QCML similarity is a learned measure, since the representation of the data in quantum states ψ_t has been optimized using the training targets. Thus, in the context of QCML, the similarity measure is estimated on the data via the following steps:

1. For C input characteristics, learn a set of operators $\{A_c\}$ on a set of training data as described in section 3.1
2. Compute $D_{QCML}(\mathbf{x}_{t,i}, \mathbf{x}_{t,j})$ as defined in equation 9. This requires first generating the error Hamiltonians $H(\mathbf{x}_{t,j})$ from equation 3 and finding the corresponding ground states $\psi_{t,j}$ as described in section 3.1
3. Compute the $J \times J$ similarity matrix $S_{QCML}(\mathbf{x}_t)$ with elements defined in equation 10

Our similarity matrix in equation 10 also requires a scaling parameter γ , which influences the concentration of the similarity values. For the results which follow in this paper, we use $\gamma_{Euclidean} = 1$ when computing Euclidean similarity, after verifying that results are not sensitive to this parameter and a value of 1 gives close to optimal results. When computing QCML similarity, we choose $\gamma_{QCML} = 16$, which results in $\gamma_{Euclidean} \cdot D_{Euclidean}(\mathbf{x}_t)^2$ and $\gamma_{QCML} \cdot D_{QCML}(\mathbf{x}_t)^2$ having approximately matching median values over the training data.

5 CVL Momentum Spillover Strategies

5.1 Overview

We now illustrate how to apply our concept of CVLs to form a profitable momentum spillover trading strategy, using an example set of characteristics. We first demonstrate the efficacy of the simple Euclidean similarity measure. We then show how QCML similarity can be used with the same input characteristics to learn similarity relationships across firms, which can in turn be used to enhance the momentum spillover trading strategy.

5.2 Data and Features

We collect daily data on a dynamic set of US firms screened for the largest size, liquidity and maturity at each point in time from October 2017 through June 2024. The average number of names available per trading day in this universe is 1,500, but the set of names changes over time to reflect the most mature and liquid names available. This universe of securities is used for both model training and subsequent evaluation.

We create a broad set of features to be used as inputs for our characteristic vectors. We do not take a view on what characteristics are best suited for this task, but here we focus on accounting and valuation ratios, as such characteristics are widely used, easy to compute, and fairly stable. Accounting values are averaged over the most recent 4 quarters, unless otherwise stated. All input characteristics are demeaned by GICS Industry Code to avoid characteristics linking firms based on industry biases. Characteristics are also cross-sectionally z-scored, and winsorized at the 1st/99th percentiles. We use market data sourced from Bloomberg and accounting data sourced from S&P Capital IQ:

Accounting Liquidity: $(0.5 * \text{Total Assets} + 0.25 * \text{Cash} + 0.25 * \text{Current Assets} - 0.5 * \text{Intangible Assets})$, scaled by Total Assets [28]

Accruals: Sign flipped 4 quarter change of $(\text{Total Assets} - \text{Working Capital} - \text{Total Liabilities} - \text{Long Term Investments} + \text{Long Term Debt})$, scaled by Total Assets [34, 40]

Book to Price: Common Equity, scaled by Market Cap [36]

Cash to Assets: Cash, scaled by Total Assets [29]

Earnings to Price: Net Income, scaled by Market Cap [7]

EBITDA to TEV: Earnings Before Interest, Taxes, Depreciation, and Amortization, scaled by Total Enterprise Value [23]

Implied Equity Duration: Defined in [13], a measure which adapts the concept of bond duration to equities. We use the forecasting parameters listed by the authors in Table 1, Panel B.

Leverage: Total Debt, scaled by Total Assets [9]

Net Leverage: Total Debt - Cash and Short Term Investments, scaled by Total Assets [31]

Net Operating Assets: Long-Term Debt + Current Portion of Long-Term Debt + Short-term Borrowings + Minority Interest in Equity + Preferred Equity + Total Common Equity - Non-Stock Preferred Equity - Total Cash And Short Term Investments - Total Other Investments, scaled by Total Assets [17]

Operating Cash-flows: Cash from Operations, scaled by Total Assets [14]

Operating Efficiency: Revenues, scaled by Total Assets [41]

Profit Margin: Net Income, scaled by Revenues [41]

Real Estate Ratio: Assets under Capital Lease + Assets under Operating Lease + Buildings, scaled by gross Property, Plant, and Equipment [43]

Revenue to Price: Revenues, scaled by Market Cap [5]

Tax Ratio: Earnings Before Taxes, including Unusual Items, scaled by Net Income to Company [22]

We also create the following set of controls for our portfolio construction process, using market data sourced from Bloomberg, and accounting and analyst ratings data sourced from S&P Capital IQ. All controls except GICS Dummies are winsorized at the 1st/99th percentiles:

Analyst Connected Stock Momentum: Momentum spillover signal formed on the basis of shared stock analyst coverage, following the approach in [1]. We create 5 versions of this control, each with input returns matched to the horizon of the input returns for the CVL signal being tested (see section 5.4 for details of our CVL signal construction). All versions are demeaned by GICS Industry Code and cross-sectionally z-scored

Analyst Coverage: Number of analysts covering a stock, cross-sectionally z-scored [30]³

Beta: Rolling 252-day time series estimated Beta to S&P 500 Index [39, 16, 15]

Momentum: Returns from 21 days ago to 252 days ago, cross-sectionally z-scored [19, 11]

Short Term Reversal: Returns from 1 day ago to 100 days ago, exponentially weighted with a 10 day half life, returns within 1 day of earnings announcements replaced by beta-adjusted market returns, demeaned by GICS Industry Code, cross-sectionally z-scored [36]

Size: Log of Market Cap, cross-sectionally z-scored [4, 15]

Subindustry Momentum: Log of market cap average weighted returns for the firm's GICS subindustry from 1 day ago to 200 days ago, exponentially weighted with a 20-day half-life, cross-sectionally z-scored [24]

Revenue to Price: As defined above

GICS Dummies: Dummy variables formed based on GICS Industry Group [3, 2]

³As detailed in [30], analyst coverage serves as an important proxy for investor attention, and some previously documented spillover signals are substantially weaker when controlling for analyst coverage

5.3 Similarity Construction

As explained in section 4, our Euclidean similarity measure is not a learned measure and requires no training. Given the inputs defined in section 5.2, we can directly compute $S_{Euclidean}(\mathbf{x}_t)$ on each day the measure is needed during the test period (January 2014 through June 2024). In contrast, our QCML similarity measure requires first training a QCML model, as explained in section 4.

QCML is trained using daily data from October 2007 through August 2013. For model training, we are required to choose a target variable. Note that the trained QCML model can be used to directly forecast the chosen target using equation 6. However, in this article we are not interested in the efficacy of the direct forecast, and thus we do not include herein any analysis of its performance. Instead, we will show that the ground states $\psi_{t,j}$ can be used to create a superior distance measure using the technique outlined in section 3.2.

It stands to reason that stocks which have similar future returns driven by observed characteristics are also likely to display stronger future co-movement due to similarity derived from those characteristics. Therefore, we choose as target variable 63-day forward returns, cross-sectionally z-scored. Other choices of target could be made. For example, future revenue / profitability growth or earnings surprises may more tightly link stocks for fundamental reasons.

Inputs are defined in section 5.2, and are the same inputs used for the Euclidean similarity measure. To create more robust similarity measures, we train 50 unique QCML models with varying seeds. Each seed randomly selects 10% of available training names to use. We form five distinct sub-groups of training dates (equally spaced), and seeds rotate through the sub-group selected for training. Signals produced from each seed are averaged with equal weight across all seeds. While QCML models can be updated online or through rolling / expanding retraining, for simplicity we keep the parameters static after initial training. Periodic retraining or online updating of QCML parameters could likely lead to stronger results than those we present herein.

Our QCML models herein used a Hilbert space dimensionality of 12. The training loop and loss function described in the algorithm in section 3.1 was implemented in PyTorch, and models were trained using the Adam optimizer. All the results and figures for this article were obtained on a laptop equipped with a 32-core 13th Gen Intel Core i9-13950HX CPU with 64GB of memory, supplemented by a NVIDIA RTX 5000 Ada Generation Laptop GPU.

Training a single QCML model on the laptop GPU takes approximately 15 seconds for the data and hyperparameters we selected. After we have successfully trained our QCML models on the training data, QCML parameters remain static over the course of the test period (January 2014 through June 2024). As explained in section 4, we can then compute $S_{QCML}(\mathbf{x}_t)$ on each day the measure is needed.

Computing $S_{QCML}(\mathbf{x}_t)$ for the full test period, also leveraging the laptop GPU, takes approximately 50 seconds per QCML model. This means, using only the laptop, all 50 seeds could be trained and final signals constructed in approximately 1 hour. A cloud computing approach could run all seeds in parallel, providing a final signal in less than 2 minutes. In this article, we use 50 seeds for our final results, but 20 seeds gives sufficient convergence for hyperparameter exploration, further reducing the computational burden.

5.4 Momentum Spillover Signal Construction

Once we have constructed a similarity matrix (denoted $S(\mathbf{x}_t)$ for the $J \times J$ similarity matrix on date t using a set of characteristics $\mathbf{x}_t$), we can create a simple momentum spillover signal. The signal uses $S(\mathbf{x}_t)$ to construct firm linkage weights, and the lagged returns of linked firms as additional inputs. Let $\mathbf{r}_{t-l:t-1,i}$ represent the returns for firm i from $t-l$ to $t-1$ for some prior number of days l , and let $w_{t,j,i}$ represent the signal input weight from firm i for the signal to be computed for firm j on t . We proceed as follows:

1. Define $w_{t,j,i} = S(\mathbf{x}_t)_{j,i} / \sum_i S(\mathbf{x}_t)_{j,i}$
2. The spillover signal using input return days l on trading date t for firm j is: $f_{l,t,j} = \sum_i (w_{t,j,i} * \mathbf{r}_{t-l:t-1,i})$

3. We construct $f_{l,t,j}$ using both Euclidean and QCML similarity, for l equal to 21 days, 63 days, 126 days, and 252 days. We also construct composite (over l) signals which average each of these 4 input returns and use the resultant $f_{\bar{l},t,j}$ as input to signal production
4. Each $f_{l,t,j}$ is individually demeaned by GICS Industry Code and cross-sectionally z-scored prior to signal evaluation

5.5 Signal Evaluation

To test performance, we use covariance estimated daily from daily returns⁴ to produce Markowitz optimal investment portfolios, where portfolio weight $w = V^{-1}Rf$ given covariance V , spillover signal forecast f , and projection⁵ operator R , to give risk optimal portfolio weights with no exposure to our controls.⁶

Investment portfolios are smoothed over 21 days to proxy for the fact that realistic investment portfolios must trade into new forecasts gradually. We compute daily returns to the smoothed portfolios, and analyze portfolio performance from January 2014 through June 2024. We do not remove transaction costs as these forecasts are not intended to represent a stand-alone investment strategy.

6 Results

6.1 Euclidean Similarity CVL Signals

We first examine the properties and performance of the Euclidean similarity CVL signals. In table 1 we show the average daily half-lives of the signals⁷. Relative to the horizon of the input returns, half-lives are faster than would be expected from a simple average of returns, particularly for longer dated input returns. This shows that linkages based on our accounting-data-focused inputs are fairly stable at a monthly horizon, but are an important source of variation over longer horizons.

In Table 2 we show the Sharpe ratios of portfolios formed from each of the signals. We plot the cumulative returns to the signals in Figure 1. Recall as explained in section 5.5 that our market neutral investment portfolios are formed to have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS Industry Groups, and thus have no linear contribution of those features to returns.

The Euclidean similarity signals all show positive performance on average, although performance significantly deteriorates in the final sub-period. We also note that the signals with longer input horizon returns perform substantially worse than the signals with short input horizon returns.

6.2 QCML Similarity CVL Signals

We next examine the properties and performance of the QCML similarity CVL signals. In table 3 we show the average daily half-lives of the signals. The signal half-lives here follow a similar pattern to those obtained with the simpler Euclidean similarity in table 1, but with the important distinction that the half-lives of the signals here are universally longer, and to a greater extent the longer the horizon of the input returns. In fact, the average half-life of the 252-day input return

⁴Covariance is estimated using a technique proprietary to Duality Group, the details of which are not relevant for this article

⁵If one wishes to solve for portfolio weights which maximize expected returns, with a penalty for expected portfolio variance, while maintaining zero exposure to a set of controls, then for weights w , spillover signal forecast f , asset variance V , risk aversion μ , and controls M , one needs to solve for w which minimizes $-w^\top f + 0.5\mu w^\top V w$ such that $w^\top M = 0$. The solution is $w = V^{-1}Rf$ where $R = I - M(M^\top V^{-1}M)^{-1}M^\top V^{-1}$. Thus R is a projection operator away from M , consistent with the desired investment process. Rf is also the residual from an inverse variance weighted regression of f on M

⁶Given a matrix of control features M , we have $f^\top M = 0$ and $w^\top M = 0$

⁷Define signal decay d as the average cross-sectional correlation of f_t with f_{t-1} . We approximate half-life as $\ln 0.5 / \ln d$

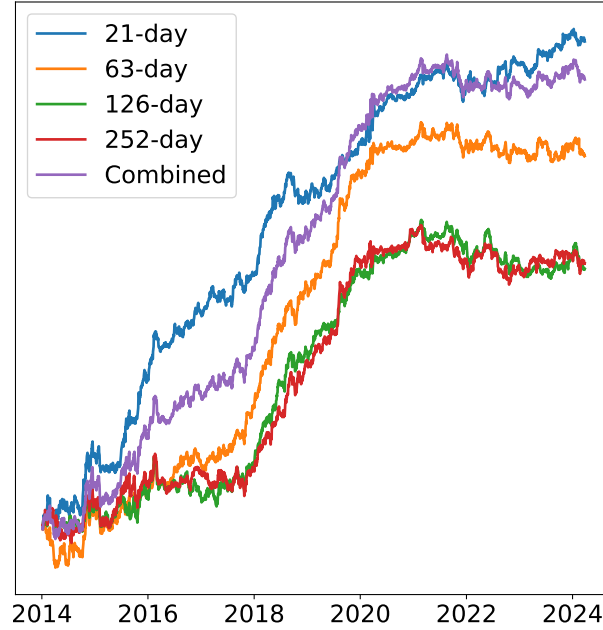
Table 1: Average daily half-lives of the Euclidean similarity CVL signals with varying input return horizons.

Period	21-day	63-day	126-day	252-day	Combined
Jan 2014 - Jun 2024	9.1	19.5	27.0	36.3	14.8
Jan 2014 - Jun 2017	9.2	19.7	26.9	35.6	14.9
Jul 2017 - Dec 2020	9.0	19.3	28.2	40.3	15.0
Jan 2021 - Jun 2024	9.0	19.4	25.9	33.5	14.5

Table 2: Sharpe ratios of returns to the Euclidean similarity CVL signals with varying input return horizons. Strategy portfolios are market neutral and have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS Industry Groups, and thus have no linear contribution of those features to returns.

Period	21-day	63-day	126-day	252-day	Combined
Jan 2014 - Jun 2024	1.35	1.03	0.71	0.73	1.24
Jan 2014 - Jun 2017	1.60	0.53	0.28	0.29	1.02
Jul 2017 - Dec 2020	1.63	2.50	2.00	2.00	2.41
Jan 2021 - Jun 2024	0.61	-0.13	-0.38	-0.34	0.00

Figure 1: Cumulative returns to the Euclidean similarity CVL signals with varying input return horizons. Strategy portfolios are market neutral and have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS Industry Groups, and thus have no linear contribution of those features to returns. Returns have been scaled by full sample realized volatility of each time series.



QCML similarity signal (90.9 days) is 2.5 times greater than the half-life of the 252-day input return Euclidean similarity signal (36.3 days). This implies that rather than adding instability from a highly parameterized learning approach, QCML has likely managed to focus on the more resilient relationships which can be inferred from the input data.

In Table 4 we show the Sharpe ratios of portfolios formed from each of the signals. We plot the cumulative returns to the signals in Figure 2. Recall as explained in section 5.5 that our market neutral investment portfolios (as for the Euclidean similarity investment portfolios) are formed to have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS

Table 3: Average daily half-lives of the QCML similarity CVL signals with varying input return horizons.

Period	21-day	63-day	126-day	252-day	Combined
Jan 2014 - Jun 2024	11.5	32.6	53.5	90.9	23.5
Jan 2014 - Jun 2017	12.5	34.9	54.6	90.2	25.0
Jul 2017 - Dec 2020	11.5	33.3	61.5	120.5	25.4
Jan 2021 - Jun 2024	10.6	29.7	46.0	72.2	20.5

Table 4: Sharpe ratios of returns to the QCML similarity CVL signals with varying input return horizons. Strategy portfolios are market neutral and have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS Industry Groups, and thus have no linear contribution of those features to returns.

Period	21-day	63-day	126-day	252-day	Combined
Jan 2014 - Jun 2024	1.38	1.14	1.04	1.10	1.42
Jan 2014 - Jun 2017	1.84	0.89	0.88	0.78	1.47
Jul 2017 - Dec 2020	1.68	2.14	2.14	2.48	2.31
Jan 2021 - Jun 2024	0.28	0.14	-0.25	-0.28	0.08

Industry Groups, and thus have no linear contribution of those features to returns.

Performance is similar to that obtained with the simpler Euclidean similarity in table 2 (the averages of the daily cross-sectional correlations of the underlying signals range from 0.74 to 0.78). However, the signals derived from QCML similarity all show out-performance, and to a greater extent as the horizon of the input returns increases. While the 21-day input return QCML only slightly outperforms, the 252-day input return QCML signal shows a Sharpe of 1.10 compared to 0.73 for the Euclidean signal (a greater than 50% improvement in Sharpe), despite the fact the QCML signal has a half-life (90.9 days) 2.5 times longer than the Euclidean signal (36.3 days). The QCML combined input return signal gives a Sharpe of 1.42, an improvement over the Euclidean combined input return signal, which has a Sharpe of 1.24, while the QCML signal has a half-life (23.5 days) nearly 1.6 times longer than the Euclidean signal (14.8 days).

Additionally, the returns to the majority of the QCML similarity signals show somewhat better sub-period consistency. These findings again imply that QCML has likely managed to learn more resilient cross-firm relationships than those computed using the simpler Euclidean similarity. Note that these results have also been obtained by training QCML a single time, on data prior to the test period. Periodic retraining or online updating of QCML parameters could likely lead to further improved results.

7 Conclusion

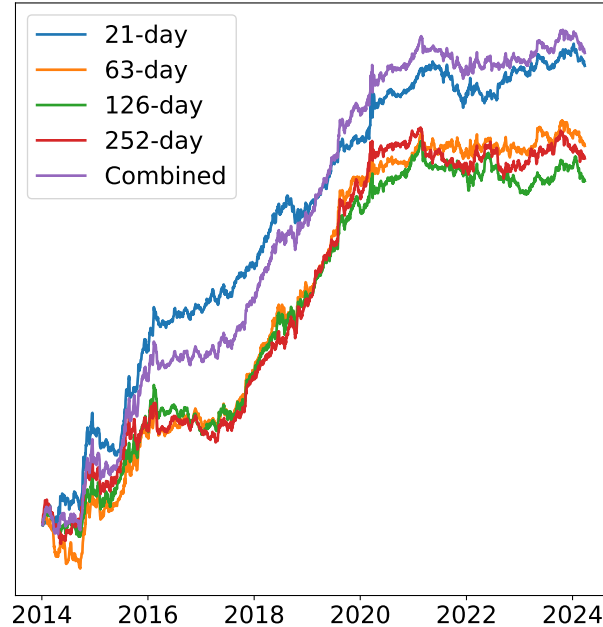
We introduced a novel form of firm linkages, which we call Characteristic Vector Linkages (CVLs). We then demonstrated how both simple Euclidean similarity CVLs and Quantum Cognition Machine Learning (QCML) similarity CVLs can be leveraged to capture momentum spillover effects in equity markets.

We find that CVLs constructed from vectors of firm characteristics provide a meaningful basis for identifying economically linked firms, evidenced by the positive Sharpe ratios achieved across the various input return horizons and similarity types, displayed in tables 2 and 4.

The Euclidean similarity approach, despite its simplicity, generates risk controlled portfolio returns with Sharpe ratios ranging from 0.73 to 1.37 depending on the input horizon, with shorter-horizon signals generally outperforming longer-horizon variants.

As discussed in detail in [10], QCML learns non-commutative quantum models for the data manifold. These quantum models have the ability to abstract out the fundamental features of the geometry of the data manifold, in a manner that is robust to noise. QCML therefore offers an elegant framework for representing complex, non-linear patterns in data, and subsequently

Figure 2: Cumulative returns to the QCML similarity CVL signals with varying input return horizons. Strategy portfolios are market neutral and have zero exposure to Analyst Connected Stock Momentum, Analyst Coverage, Beta, Momentum, Short Term Reversal, Size, Subindustry Momentum, Revenue to Price, and GICS Industry Groups, and thus have no linear contribution of those features to returns. Returns have been scaled by full sample realized volatility of each time series.



making more accurate measurements of distance between data points. In this paper, we take advantage of QCML’s abilities to ultimately learn relationships between firms that may not be captured by traditional distance metrics. By mapping firm characteristics into quantum states and leveraging quantum fidelity to derive similarity measures, QCML appears to uncover more accurate connections that enhance the predictive power of momentum spillover strategies.

Our supervised QCML similarity approach demonstrates meaningful improvements over the simple Euclidean method, particularly for longer-horizon signals. While the performance differences are modest for short-term signals, with 252-day input returns the QCML similarity measure achieves a Sharpe ratio of 1.12 compared to 0.76 for the corresponding Euclidean measure, while having a half-life (90.9 days) which is 2.5 times longer than the Euclidean signal’s half-life (36.3 days). The combined QCML signal delivers a Sharpe ratio of 1.43, an improvement over the Euclidean combined signal’s 1.26 Sharpe ratio, while having a half-life (23.5 days) which is nearly 1.6 times longer than the Euclidean signal’s half-life (14.8 days).

The QCML-based signals also exhibit somewhat superior consistency across sub-periods. While both approaches perform poorly during the January 2021 through June 2024 period, the QCML signals do modestly outperform the Euclidean signals. The QCML signals also derive relatively less of their overall performance from the strongest performing second sub-period (July 2017 through December 2020). This improved stability, as well as the considerably longer average half-lives observed for QCML signals, suggest that the supervised learning approach successfully identifies more robust and persistent firm relationships. Additionally, these results have been obtained by training QCML a single time, on data prior to the test period. Periodic retraining or online updating of QCML parameters could likely lead to stronger outperformance during the most recent sub-period. There is no such possible avenue to improve the performance of the Euclidean similarity signals.

Our work in this article could be extended in two obvious ways. First, as noted in section 5.3, our QCML models are only trained a single time, on data available prior to the test period. A more dynamic approach allowing for periodic retraining or online updating of parameters may give improved estimates of similarity. Second, as mentioned in section 5.2, we did not take a view regarding which sort of firm characteristics are most likely to be meaningful for capturing CVLs.

Future research could hypothesize economic mechanisms driving the ability of CVLs to capture firm linkages, and thereby refine the categories of input factors used or the target selected. The simplicity of the CVL approach makes testing such data variations trivial.

In conclusion, this work introduces Characteristic Vector Linkages as a promising new approach for capturing firm relationships and demonstrates the potential for Quantum Cognition Machine Learning to more accurately measure distance between data points and thus capture similarity relationships. As markets continue to evolve and traditional strategies face increasing competition, the ability to identify and exploit subtle firm linkages through advanced similarity learning may become increasingly valuable for quantitative investment strategies.

References

- [1] Usman Ali and David Hirshleifer. Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics*, 136(3):649–675, 2020.
- [2] Clifford S. Asness, Andrea Frazzini, and Lasse H. Pedersen. Low-risk investing without industry bets. *Financial Analysts Journal*, 70(4):24–41, 2014.
- [3] Clifford S. Asness, Roger B. Porter, and Ross L. Stevens. Predicting stock returns using industry-relative firm characteristics. *Capital Markets eJournal*, 2000.
- [4] Rolf W. Banz. The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18, 1981.
- [5] William C. Barbee, Sandip Mukherji, and Gary A. Raines. Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, 52:56–60, 1996.
- [6] Söhnke M. Bartram, Jürgen Branke, Giuliano De Rossi, and Mehrshad Motahari. Machine learning for active portfolio management. *The Journal of Financial Data Science*, 3(3):9–30, 2021.
- [7] Sanjoy Basu. The relationship between earnings’ yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics*, 12(1):129–156, 1983.
- [8] Ingemar Bengtsson and Karol Zyczkowski. *Geometry of Quantum States: An Introduction to Quantum Entanglement*, pages 187–191. Cambridge University Press, 2006.
- [9] Laxmi Chand Bhandari. Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43(2):507–528, 1988.
- [10] Luca Candelori, Alexander G. Abanov, Jeffrey Berger, Cameron J. Hogan, Vahagn Kirakosyan, Kharen Musaelian, Ryan Samson, James E.T. Smith, Dario Villani, Martin T. Wells, and Mengjia Xu. Robust estimation of the intrinsic dimension of data sets with quantum cognition machine learning. *Scientific Reports*, 15(1):6933, December 2025.
- [11] Mark Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.
- [12] Lauren Cohen and Andrea Frazzini. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011, 2008.
- [13] Patricia Dechow, Richard Sloan, and Mark Soliman. Implied equity duration: A new measure of equity risk. *Review of Accounting Studies*, 9(2):197–228, 06 2004.
- [14] Hemang Desai, Shivaram Rajgopal, and Mohan Venkatachalam. Value-glamour and accruals mispricing: One anomaly or two? *The Accounting Review*, 79(2):355–385, 2004.
- [15] Eugene Fama and Kenneth French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

- [16] Eugene F. Fama and James D. MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81:607 – 636, 1973.
- [17] David Hirshleifer, Kewei Hou, Siew Hong Teoh, and Yinglei Zhang. Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, 38:297–331, 2004. Conference Issue on Research on Market Efficiency, Valuation, and Mispricing: Risk, Behavioral, and Alternative Explanations.
- [18] Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133, 2020.
- [19] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, 1993.
- [20] Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.
- [21] Jerinsh Jeyapaulraj, Dhruv Desai, Peter Chu, Dhagash Mehta, Stefano Pasquali, and Philip Sommer. Supervised similarity learning for corporate bonds using random forest proximities, 2022.
- [22] Baruch Lev and Doron Nissim. Taxable income, future earnings, and equity values. *The Accounting Review*, 79(4):1039–1074, 10 2004.
- [23] Tim Loughran and Jay W. Wellman. New evidence on the relation between the enterprise multiple and average stock returns. *Journal of Financial and Quantitative Analysis*, 46(6):1629–1650, 2011.
- [24] Tobias J. Moskowitz and Mark Grinblatt. Do industries explain momentum? *The Journal of Finance*, 54(4):1249–1290, 1999.
- [25] Kharen Musaelian et al. Quantum cognition machine learning: AI needs quantum, 2024. <https://www.qognitive.io/papers/QCML%20-%20Qognitive,%20Inc.pdf>.
- [26] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.
- [27] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
- [28] Hernán Ortiz-Molina and Gordon M. Phillips. Real asset illiquidity and the cost of capital. *Journal of Financial and Quantitative Analysis*, 49(1):1–32, 2014.
- [29] Berardino Palazzo. Cash holdings, risk, and expected returns. *Journal of Financial Economics*, 104(1):162–185, 2012.
- [30] Christopher A Parsons, Riccardo Sabbatucci, and Sheridan Titman. Geographic lead-lag effects. *The Review of Financial Studies*, 33(10):4721–4770, 01 2020.
- [31] Stephen H. Penman, Scott A. Richardson, and İrem Tuna. The book-to-price effect in stock returns: Accounting for leverage. *Journal of Accounting Research*, 45(2):427–467, 2007.
- [32] Christo Pirinsky and Qinghai Wang. Does corporate headquarters location matter for stock returns? *The Journal of Finance*, 61(4):1991–2015, 2006.
- [33] Emmanuel M. Pothos and Jerome R. Busemeyer. Quantum cognition. *Annual Review of Psychology*, 73(1):749–778, 2022.
- [34] Scott A. Richardson, Richard G. Sloan, Mark T. Soliman, and İrem Tuna. Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics*, 39(3):437–485, 2005.

- [35] Joshua Rosaler, Luca Candelori, Vahagn Kirakosyan, Dhagash Mehta, Kharen Musaelian, Stefano Pasquali, Ryan Samson, and Martin T Wells. Supervised similarity for high-yield bonds, 2025. Risk.net.
- [36] Barr Rosenberg, Kenneth Reid, and Ronald Lanstein. Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*, 11(3):9–16, 1985.
- [37] Ayari Salah and Gatfaoui Hayette. A meta-analysis of supervised and unsupervised machine learning algorithms and their application to active portfolio management. *Expert Systems with Applications*, 271:126611, 2025.
- [38] Ryan Samson, Jeffrey Berger, Luca Candelori, Vahagn Kirakosyan, Kharen Musaelian, and Dario Villani. Quantum cognition machine learning: financial forecasting, 2024. Risk.net.
- [39] William Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.
- [40] Richard G. Sloan. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3):289–315, July 1996.
- [41] Mark T. Soliman. The use of dupont analysis by market participants. *Corporate Finance: Valuation*, 2007.
- [42] Dominique Spehner, Fabrizio Illuminati, Miguel Orszag, and Wojciech Roga. *Geometric Measures of Quantum Correlations with Bures and Hellinger Distances*, pages 105–157. Springer International Publishing, Cham, 2017.
- [43] Selale Tuzel. Corporate real estate holdings and the cross-section of stock returns. *The Review of Financial Studies*, 23(6):2268–2302, 03 2010.
- [44] Dimitrios Vamvourellis, Máté Toth, Snigdha Bhagat, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. Company similarity using large language models, 2023.