Neighboring Assets*

Sina Seyfi†

September, 2024

*I am thankful to Peter Nyberg, Lasse Heje Pedersen, Matthijs Lof, Matti Keloharju, Andreas Neuhierl, Tarun Chordia, Yong Chen, Petri Jylhä, Cristian Tiu (discussant), Theis Ingerslev Jensen, Anders Löflund (discussant), Mikko Leppämäki, Renato Lazo Paz (discussant), Markku Kaustia, Aleksi Pitkäjärvi, Ville Rantala, Farid AitSahlia (discussant), participants at the AFA annual meeting, FMA annual meeting, FMA European conference, Nordic Finance Network (NFN) workshop, Finance Brown Bag (Aalto University), and GSF Winter Workshop for their useful comments. I thank OP Group Research Foundation for the financial support. First version: December 2022

†Aalto University, School of Business, Department of Finance
Email: sina.seyfi@aalto.fi

Neighboring Assets

ABSTRACT

Firms with similar characteristics display similar expected returns. Defining neighboring assets as those with the most similar set of characteristics, I show that past returns of an asset's neighbors predict its future returns. If a majority of an asset's neighbors have performed poorly (well) in the past, it is likely that this asset also performs poorly (well) in the future. By classifying each asset into a decile portfolio based on the past performance of its neighbors, with 94 characteristics, a long-short portfolio generates an out-of-sample annualized Sharpe ratio of 1.40 with a monthly alpha of 1.75% (t = 9.45).

I. Introduction

If firm-level characteristics are determinants of expected stock returns, then firms with similar characteristics should have similar expected rates of returns. This suggests that past returns of firms which share similar characteristics should be similar to the future returns of a firm which is alike in characteristics. The empirical asset pricing literature, however, has discovered quite a large number of firm characteristics which relate to the expected stock returns.[1] With this zoo of characteristics, it is unclear how the similarity of firms should be measured and according to which characteristics assets should be sorted into portfolios (Cochrane, 2011). Also, the relationship between this large set of characteristics and expected returns is highly complex and the true functional form is unknown (Kelly, Malamud, and Zhou, 2024). In this paper, I tackle this large-dimensionality challenge and link similar firms through many characteristics.

For recognizing analogous firms, I start by finding *neighboring assets*, identified as those assets with the most similar set of characteristics. I measure this similarity between two assets as the distance of their characteristics and define an asset's *neighbors* as those whose characteristics have the closest distance to this very asset. By defining neighboring assets in this way, each asset has almost the same characteristics as its neighbors, and neighboring assets display fundamental linkages in many dimensions such as the similarity of accounting variables or financial ratios. If the cross-sectional variation in expected stock returns roots in the cross-sectional variation in asset-specific characteristics, naturally neighboring assets should have similar expected returns. This holds regardless of what is the true relationship between a large set of characteristics and expected returns.

[1] See, among many others, Green, Hand, and Soliman (2011), Harvey, Liu, and Zhu (2016), Hou, Xue, and Zhang (2017), Green, Hand, and Zhang (2017), Jensen, Kelly, and Pedersen (2021), and Chen and Zimmermann (2021).

Recognizing neighboring assets has several attractive applications in asset pricing. To begin with, grouping (separating) assets with similar (different) characteristics must generate cross-sectional dispersion in expected returns. This assumption provides a new method for cross-sectional return predictability through a large set of predictors in a flexible setting. In order to produce cross-sectional dispersion in average returns through portfolios of neighboring assets, I classify each asset into one of the decile portfolios based on the past performance of its neighbors, wherein decile 1 (10) is indicative of the loser (winner) portfolio. For example, if a majority of an asset's neighbors in the past, which have had the same characteristics as the most recent characteristics of the asset in question now, have belonged to decile portfolio 1, the loser portfolio, it is most likely that this asset itself also belongs to the loser portfolio, decile 1. Intuitively, if assets with a specific set of characteristics in the past have performed badly (well), it is likely that the assets in the future with the same set of characteristics will perform badly (well), as well. Considering the case with only two characteristics as an example, if assets with high momentum and low size have produced a high expected return in the in-sample data, all the assets with high momentum and low size in out-of-sample data are then grouped into a decile portfolio which is supposed to generate a high expected return. This portfolio reflects the behaviour of high momentum and low size stocks over time. Similarly, each decile portfolio is representative of a potentially large set of characteristics, generating the corresponding expected returns.

More formally, if a stock j at time t has a set of similar characteristics to the stock j_0 at time $t_0 < t$, then the expected return of stock j at time t should be similar to the expected return of stock j_0 at time t_0. In more detail, suppose that an asset j at time t has a set of most recent characteristics $\mathbf{x}_{j,t-1}$. All assets at time $t - t'$, the in-sample period, with a set of characteristics $\mathbf{x}_{t-t'-1}$, are assigned into one of

the decile portfolios based on their realized (or mean) returns at time $t - t'$, which is observable to the investor at time t. Hence, each asset in the in-sample data is assigned to a decile portfolio. For classifying asset j at time t, I first find the k nearest neighbors of this asset in the in-sample data, i.e. those k assets with the most similar set of characteristics in the in-sample data. Then I put asset j into the decile portfolio in which the majority of its k neighbors belonged to in the past. Inspired by Ali and Hirshleifer (2020), I also assume that *closer* neighbors, having a smaller distance of characteristics, must contribute more in decile prediction, so I weigh all of the k nearest neighboring assets by the inverse of their distance to the asset j.[2] The in-sample period is considered a rolling window consisting of $\tau = 120$ months before time t, which allows for the time-varying relationship between characteristics and expected returns. Figure 1 panel (a) visualizes this concept.

I find that this strategy generates a fairly large dispersion in the cross-section of stock returns. Recognizing neighboring assets according to 94 firm characteristics,[3] a value- and equally-weighted long-short strategy of extreme decile portfolios generate a monthly Fama-French five (FF5) alpha of 1.55% ($t = 8.17$) and 1.65% ($t = 11.52$), respectively. The corresponding annualized Sharpe ratios in the period 1980-2021 are 1.40 and 2.01 for value- and equally-weighted portfolios. Even after excluding micro caps (those which are below 20% NYSE percentile), the FF5 monthly alphas are 0.88% ($t = 5.67$) and 0.92% ($t = 7.64$), with Sharpe ratios of 0.91 and 1.24. The results are not derived by microcaps, not subsumed by industries,

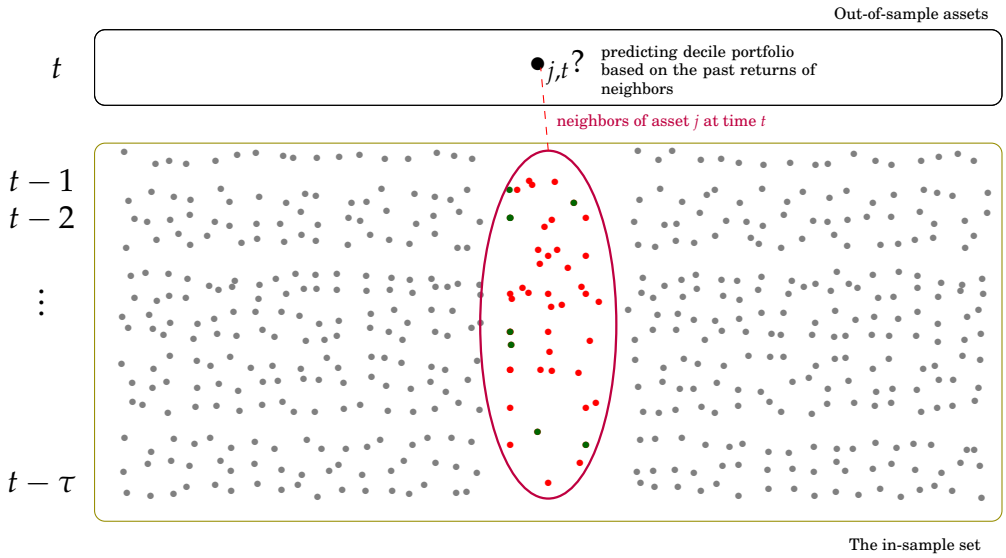
[2] Ali and Hirshleifer (2020) define the connection between two firms as the number of shared analyst coverage. For the return predictability, they assume that stocks which have more co-covered analysts are more likely to be similar, and hence they weigh stocks by the number of co-covered analysts (see equation 1 in their paper). In my analysis, I follow Ali and Hirshleifer (2020) by weighting neighboring assets with the inverse of their distance. However, the results are robust if I weigh all neighboring assets equally.

[3] These characteristics are listed in Table A1. I thank Dacheng Xiu for providing the characteristics data on his website.

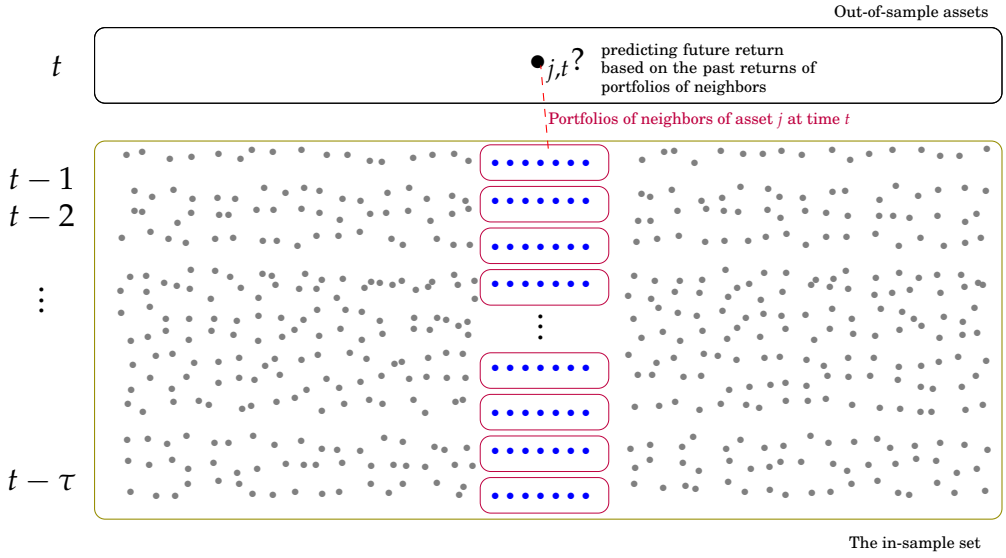
and remain unexplained by a large set of factor zoo. My results are robust to the choice of the number of neighbors, length of rolling windows, and different ways of distance measurement.

If past returns of an asset's neighbors at a stock level have predictive powers, then portfolios consisting of each asset's neighbors should also predict future returns of this asset. To test this, for each asset j at time t, I create a portfolio at each time $t-1$, $t-2$, $\cdots$, $t-\tau$ ($\tau = 120$ months) which contains the nearest neighbors of asset j at each month in the past. Then I predict the future return of asset j based on the average returns of its neighboring portfolios. Figure 1 panel (b) provides a visualisation. Forming decile portfolios based on this method, I find that an out-of-sample value-weighted long-short portfolio generates a monthly FF5 alpha of 1.38% ($t = 5.00$) with an annualized Sharpe ratio of 0.95. The monthly FF5 alpha for an equally-weighted counterpart long-short portfolio increases to 2.61% ($t = 12.44$) with an annualized Sharpe ratio of 2.12.

What mechanism drives the performance of neighboring assets strategy? If exposures to systematic risks are a function of firm characteristics—as well motivated by Kelly, Pruitt, and Su (2019) and Gu, Kelly, and Xiu (2021)—then stocks with similar characteristics have similar exposure to the systematic risk factors. Therefore, because of the similarity of their characteristics, neighboring assets should have the same exposure to the Stochastic Discount Factor (SDF). In turn, if the cross-sectional variation in expected stock returns is caused by exposure to systematic risk factors, then neighboring assets have similar expected returns. This holds regardless of what the true sources of risks are and what the functional form between characteristics and covariances is: as long as the compensation for exposure to risk factors—observed or latent—drives the relationship between characteristics and expected returns, neighboring assets display similar expected



(a) Individual neighbors



(b) Portfolios of neighboring assets

Figure 1. neighboring assets classification algorithm

Panel (a) shows the classification framework based on the neighboring assets. For an asset j at time t, I find its k neighbors in the in-sample data, from time $t-\tau$ to $t-1$, which have had the closest distance of characteristics to asset j at time t. If the majority of neighbors have performed poorly (well), it is likely that asset j at time t also performs poorly (well). Hence, I put asset j at time t into the decile portfolio to which the majority of its neighbors belonged in the past. The in-sample set is considered a rolling window. In panel (b), for each asset j at time t, I create a portfolio of neighboring assets in each month in the in-sample data. Blue dots show the assets which have had the most similar characteristics to asset j at time t. Then I predict the return of asset j at time t based on the average of in-sample neighboring portfolios.

returns and a long-short portfolio of neighboring stocks strategy is profitable.

The modern empirical asset pricing states that expected returns are a function of a large set of characteristics (Kozak and Nagel, 2023), and the relationship between these characteristics and expected returns is highly complex (Kelly et al. 2024, Bryzgalova, Pelger, and Zhu 2024). This is why, a new branch of literature has turned to machine learning methods to cope with this high-dimensionality challenge (Karolyi and Van Nieuwerburgh 2020, Gu, Kelly, and Xiu 2020, Chen, Pelger, and Zhu 2024). My neighboring asset algorithm finds the complex relationship between a large set of characteristics and expected returns without the need to estimate a large set of parameters.

As I show that an asset's neighbors' past returns predict its future returns, my paper contributes to the literature on finding peer firms by introducing a new way of defining related firms based on the similarity of many characteristics. The literature has defined various types of connections between firms, including same industry linkage (Moskowitz and Grinblatt, 1999), same principal customers (Cohen and Frazzini, 2008) and suppliers (Menzly and Ozbas, 2010), same technology (Lee, Ma, and Wang, 2016), common active mutual fund owners (Anton and Polk, 2014), and same analyst coverage (Ali and Hirshleifer, 2020), among others, all yielding to cross-sectional predictability of returns. In this regard, I introduce a new linkage between firms, that is, *closeness of many characteristics* at the same time.

Indeed, the economic link between firms with similar characteristics based on one or two characteristics has been studied in the literature. For instance, Fama and French (1995) document that earnings of firms with similar size and book-to-market are explained by common factors. Or, Hirshleifer, Hou, and Teoh (2012) argue that firms with similar accrual co-move. Considering seventeen characteristics one-by-one, Müller (2019) shows that firms which are in the same quantile

based on one characteristic have information spillovers and can be considered economically linked firms. He, Wang, and Yu (2021) argue that, because of investor attention, stocks with similar price, size, book-to-market, return on assets, and investment-to-assets have spillover effects. I generalize all these approaches by linking the firms based on many characteristics. If two firms are neighbors according to my definition, they share similar characteristics, and their values are sensitive to the same type of information. Therefore, I show that the similarity and dissimilarity of characteristics account for the cross-predictability of returns, and hence, using a neighborhood definition based on the characteristics is an effective way of recognizing related firms.

The neighborhood connection, importantly, becomes stronger when I use more characteristics for finding neighboring assets. This finding indicates that a large set of characteristics *jointly* predict future expected returns and the factor structure of expected returns is remarkably high-dimensional, which is also in line with the emerging literature on the high-dimensionality of the factor structure of expected returns.[4]

Return predictability through past returns, such as using a momentum strategy, has been well documented in the literature to be robust and pervasive (Jegadeesh and Titman, 2001; Goyal and Wahal, 2015). A momentum strategy uses each asset's own past history to predict its future returns. In contrast, and unlike Kelly, Malamud, and Pedersen (2020) who use all asset's signals to predict each individual returns, I use only each asset's neighbors past performance for return predictability of individual securities. I show that, while a momentum strategy fails to explain the performance of neighboring assets portfolios, the neighboring

[4]For instance, Kozak, Nagel, and Santosh (2020) and Jensen et al. (2021) show that there is little redundancy among a large set of documented anomalies, while Bryzgalova, Huang, and Julliard (2022) find that a large set of characteristics are needed to explain the variations in the cross-section of stock returns.

assets' long-short portfolio subsumes the momentum strategy and explains a wide range of momentum anomalies.

An asset can also be the neighbor of its own lags if its characteristics do not change dramatically over time. This raises the concern that an asset's own lags may drive the neighboring assets' strategy performance. To rule out any possibility that return prediction comes from an asset's own lag, I consider the following three tests. First, when making the prediction for each asset, I find up to 5000 of its neighbors in the past 120 months. Therefore, even if an asset is the neighbors of its own lags, an asset's lags constitute a small portion of the neighbors. Second, in a different scenario, when finding the neighbors of an asset, I remove its own lags from the in-sample set. Third and lastly, when considering a 10-year rolling window, I skip the most recent five years and find assets neighbors from five to ten years ago. In all cases, the performance of the neighboring assets strategy remains unchanged.

Inspired by Martin and Nagel (2022) who discriminate between in-sample and out-of-sample predictability, the focus of my paper is fully on out-of-sample prediction. My decile portfolios generate a wide range of average returns, namely from -0.40% in the lowest decile to 1.26% in the highest decile in excess of the risk-free rate. The large spread persists over time and does not disappear after excluding tiny stocks or when considering only large stocks. I track the pattern of characteristics in my decile portfolios in over 40 years of out-of-sample study (from 1980 to 2021), and find that most of these patterns are either a U-shape or an inverse of a U-shape or an M-shape. This finding suggests that the characteristics-returns relationship could be complex and interactions matter. These characteristics' patterns seem to be invariant over time and are stronger among tiny stocks.

Closely Related Literature. My paper contributes to the literature that attempts to model the future risk premium as a function of lagged stock-level char-

acteristics, such as Fama and French (2008). Traditional approaches for mapping characteristics to the expected returns include employing cross-sectional regressions used by Haugen and Baker (1996), Lewellen (2015) and Light, Maslov, and Rytchkov (2017) or portfolio sorts such as Daniel, Grinblatt, Titman, and Wermers (1997). Newer methods include machine learning methods surveyed by Gu et al. (2020). Gu et al. (2021), and Feng, He, Polson, and Xu (2018) use deep neural networks to map characteristics to the risk premiums. Kaniel, Lin, Pelger, and Van Nieuwerburgh (2023) and Li and Rossi (2020) show that machine learning methods can differentiate loser and winner funds based on their characteristics.

My paper also alleviates the curse-of-dimensionality problem in the cross-sections of returns. Among others, Kozak et al. (2020), Lettau and Pelger (2020) use PCA-based methods, and, Feng, Giglio, and Xiu (2020) and Freyberger, Neuhierl, and Weber (2020) opt LASSO-type approaches to pick the most important variables. I deviate from the literature by unifying the joint effect of characteristics zoo in the neighboring assets, a novel method for working with many characteristics simultaneously.

My paper also contributes to the literature of building basis assets, that is, how test portfolios should be created. Ahn, Conrad, and Dittmar (2009) use return correlations to sort assets into portfolios, while Moritz and Zimmermann (2016) and Bryzgalova et al. (2024) use Trees to group similar assets into portfolios. In Seyfi (2024), I suggest grouping stocks into portfolios such that stocks within each portfolio have the most similar characteristics. In Seyfi (2023), I propose creating test portfolios such that they are as similar as possible to the returns-sorted portfolios in the past. As stated by Nagel and Singleton (2011), managed portfolios are the optimal tools for hypothesis testing in conditional asset pricing. Inspired by Lewellen, Nagel, and Shanken (2010) who suggest that using double-

sort portfolios provides a low hurdle for asset pricing models, I use 94 characteristics to form portfolios of neighboring assets which is a new way of creating test assets based on many characteristics. The neighboring assets strategy sorts stocks into portfolios based on the expected returns and can be used for testing asset pricing models.

Lastly, my paper proposes a new way of defining connected firms through neighboring assets, complementary to the literature trying to find linked firms, such as Barberis and Shleifer (2003), Bouwman (2011), Shue (2013), Leary and Roberts (2014), and Kaustia and Rantala (2015), among others.

II. Neighboring Assets and Expected Returns

A. *Model*

In this section, I explain how to predict the future return premium of each asset given a large set of lagged characteristics through neighboring stocks. This return prediction requires first defining the target variable for both in-sample and out-of-sample sets. I define target variables in the in-sample set as the decile portfolio that each asset belonged to. To do so, I sort all assets at time $t_0 < t$ into one of the decile portfolios based on their realized returns at time t_0.[5]

I model $r_{j,t}$, the realized return for asset j at time t, as

$$r_{j,t} = \mathbb{E}(r_{j,t}) + \epsilon_{j,t}, \tag{1}$$

where the first part is the expected return at time t and the second part is the idiosyncratic noise.

[5]In an alternative approach, I also consider sorting stocks based on their 12-month average returns up to the end of the month t_0. The results are not sensitive to the choice of target variable.

Further, I assume that the expected return of asset j at time t is a function of the most recent characteristics $\boldsymbol{x_{j,t-1}}$. That is,

$$\mathbb{E}(r_{j,t}) = \mathcal{F}_{t-1}(\boldsymbol{x_{j,t-1}}), \tag{2}$$

where $\mathcal{F}_{t-1}(.)$ is a function that maps the most recent characteristics to the expected returns through neighboring assets and allows for a time-varying relationship between characteristics and expected returns. Note that the function $\mathcal{F}_{t-1}(.)$ is unknown and potentially highly complex (Kelly et al., 2024).

From equations 1 and 2, we get

$$r_{j,t} = \mathcal{F}_{t-1}(\boldsymbol{x_{j,t-1}}) + \epsilon_{j,t}. \tag{3}$$

That is, the realized returns at each time are a noisy function of the most recent characteristics. This implies that, regardless of the functional form, stocks with similar characteristics should have similar realized returns.

B. Methodology

In this section, I describe the framework of asset classification through neighboring assets. Forming decile portfolios that generate cross-sectional dispersion in expected returns is a classification problem. Instead of predicting the expected returns directly, I predicte the decile portfolio that each asset belongs to. I denote decile portfolio i by d_i, such that d_1 and d_{10} are loser and winner portfolios respectively. Suppose that asset j at time t_0, a_{j,t_0}, has the most recent characteristics $\mathbf{x}_{j,t_0-1}$, with a realized return of r_{j,t_0}. At each month in the in-sample period, I sort all assets based on their realized returns into one of the decile portfolios. I further denote the decile portfolio that asset a_{j,t_0} belongs to at time t_0 as c_{j,t_0}. Therefore, if

r_{j,t_0} is among the lowest 10 percentile of returns at time t_0, $c_{j,t_0} = d_1$, and similarly, if r_{j,t_0} is on the highest 10 percentile of returns $c_{j,t_0} = d_{10}$.

For classifying each asset at time t, I consider τ months for the length of the rolling window of in-sample data, that includes all available information up to time $t-1$. Let S_t be the in-sample set I use for time t, and the out-of-sample includes the whole cross-section of assets at time t. For predicting the decile portfolio that asset j belongs to at time t, I find the k-nearest neighbors of this asset in the S_t and put this stock to a decile portfolio that the majority of its neighbors belong to.

In the main analysis, I use a distance-weighted scheme for decile prediction. I assume that *closer* assets contribute more than further assets. I define the distance of asset j at time t (denoted as $a_{j,t-t'}$) to asset j' at time $t-t'$ (denoted as $a_{j',t-t'}$) as

$$l(a_{j,t}, a_{j',t-t'}) = \left[(\mathbf{x}_{j',t-t'-1} - \mathbf{x}_{j,t-1})'(\mathbf{x}_{j',t-t'-1} - \mathbf{x}_{j,t-1})\right]^{\frac{1}{2}}$$

and the distance weights become

$$\omega_{a_{j,t}}(a_{j',t-t'}) = \begin{cases} \frac{1}{l(a_{j,t}, a_{j',t-t'})} & \text{if asset } a_{j',t-t'} \in k \text{ neighbors of asset } a_{j,t} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

for some distance measure l. $\omega_{a_{j,t}}(a_{j',t-t'})$ is the weight with which asset $a_{j',t-t'}$ predicts asset $a_{j,t}$. In equation 4 assets which are among the k neighbors of $a_{j,t}$ get a weight proportional to the inverse of their distance, while other assets get a weight of zero, i.e., they are ignored for decile prediction. I show that the model performance is robust to different measures of l. I consider an Euclidean distance measure in this context and treat all the features equally, that is, I weigh characteristics evenly assuming that they contribute equally to separating assets with different expected returns.

Finally, for classifying asset $a_{j,t}$, I sum over the weights of each class and put $a_{j,t}$ into the decile portfolio which has the highest summation. Mathematically,

$$c_{j,t} = \underset{d \in \{d_1,\ldots,d_{10}\}}{\arg\max} \sum_{s \in S_t} \omega_{a_{j,t}}(a_{j',t-t'})\delta(d, c_{s,t-t'}), \tag{5}$$

where S_t is the set of in-sample data, and $\delta(.,.)$ is a Kronecker delta function:

$$\delta(z_1, z_2) = \begin{cases} 1, & \text{if } z_1 = z_2 \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Intuitively, asset $a_{j,t}$ goes to decile portfolio $d \in \{d_1, \ldots, d_{10}\}$ if the summation of weights of assets in portfolio d in the in-sample data is the highest. For an equally weighted (versus distance weighted) scheme, the equation 5 will simply become a mode function. Therefore equation 5 can be seen as a weighted mode function.

This framework has a simple probabilistic interpretation: if the majority of an asset's neighbors have been losers (winners) in the past, it is most likely that this asset will also be a loser (winner) in the future. I provide details of the Bayesian interpretation in the Internet Appendix .A.

C. Portfolios of Neighboring Assets

In this section, I consider another case wherein portfolios of neighboring assets would predict future returns of individual assets. Considering in-sample panel data from the past τ months, for each asset j at time t, I create a portfolio which contains asset j neighbors in each month. Formally, without loss of generality, assume that assets $a_{1,t-t'}, \cdots, a_{k,t-t'}$ are the k nearest neighbors of asset j at time $t - t'$. For this

month, I create a portfolio of these k assets:

$$r^*_{p_j,t-t'} = \frac{1}{k}\sum_{s=1}^{k} r_{s,t-t'}, \tag{7}$$

where $r^*_{p_j,t-t'}$ is the return of portfolio of neighboring assets of asset j at time $t-t'$, for $t' = 1, 2, \cdots, \tau$.[6] Then I predict a return for asset j at time t (out-of-sample) as:

$$\hat{r}_{j,t} = \frac{1}{\tau}\sum_{t'=1}^{\tau} r_{p_j,t-t'}. \tag{8}$$

Finally, I sort assets into decile portfolios based on the predicted returns and update portfolios monthly. Intuitively, when a set of assets with the same set of characteristics as asset j have generated an average return of $r^*_{p_j,t-t'}$ in the past, it is likely, then, that asset j at time t will have similar performance to the average performance of its neighbors in the past.

III. Empirical Analysis

In this section, I present the empirical analysis which contains 42 years of out-of-sample study.

A. Data

I obtain monthly individual stock returns from CRSP for all firms in the NYSE, AMEX, and NASDAQ with standard codes 10 and 11 for the period starting from January 1970 until the end of 2021. This universe of data contains more than 3 million individual observations.[7] In the beginning, I remove the smallest 5% stocks

[6] I find that the model is robust to distance-weighting or equally-weighting portfolios of neighbors.

[7] In an alternantive approach, I follow Gu et al. (2020) by including stocks with prices below \$5, share codes beyond 10 and 11, and financial firms, to make the data as large as possible. However,

based on their lagged market capitalization. These stocks are mostly illiquid and have many missing characteristics. In fact, because of their small sizes, including them does not distort the results as long as we use a value-weighting scheme. However, removing them assures that the results are not affected by very tiny stocks. After removing them still the data set includes more than 2.8 million observations with 24,586 unique assets. Following Fama and French (2008), I consider three sets of data for my analysis. First is the universe of all stocks except the 5% tiniest which I refer to "all stocks" hereafter. Second is the set of stocks with a market capitalization above 20% of NYSE, which I refer to as "all-but-tiny stocks". Third, I consider only "large stocks" which include only stocks with a market cap higher than the NYSE median.

For the out-of-sample study, I start the analysis in January 1980 and for each month I consider the data set in the past 10 years for the in-sample analysis. The in-sample set is a rolling window which moves before the out-of-sample set. For the out-of-sample study, the average number of stocks per month for all stocks exceeds 4800 assets, while more than 2387 and 1244 for all but tiny and large stocks, respectively.

For the characteristics, I use data provided by Green et al. (2017) and Gu et al. (2020) who build a large collection of 94 stock predictors.[8] I use the same acronyms used by Gu et al. (2020) and list them in Table A1 in the Appendix. One issue is the missing characteristics, which I follow Gu et al. (2020) by imputing them by the cross-sectional median. For finding neighboring assets, all the characteristics must have the same scale, otherwise, some features with larger scales dominate others. Therefore, I cross-sectionally demean all characteristics and make them

using or not using common filters, such as using only share codes 10 and 11, does not change the results.

[8] I download data from Dacheng Xiu website.

unite variance. In order to alleviate the effect of outliers, I also winsorize data in 0.01 and 0.99 levels. I obtain risk-free rate data from Fama and French library, as well as market returns, size, value, operating profitability, and investment factors. I download q and augmented q model (q^5) from the work based on Hou, Mo, Xue, and Zhang (2019) and Hou, Mo, Xue, and Zhang (2021). I obtain 2-digit SIC codes for controlling for industries from the Gu et al. (2020) dataset.

B. Model Parameters

The neighboring assets framework requires only one hyper-parameter to be fixed before the analysis starts. This one parameter is an extremely small number compared to other machine-learning methods. This parameter is k, the number of neighbors for each asset. As I use a distance-weighted scheme, the role of k becomes less important, as the further neighbors receive lower weights. I show that results in the distance-weighted scheme are robust for large enough ks. I choose $k = 1000$ for the main analysis and later consider different k in the robustness checks (varying k from 100 to 5000) and show that the model is not sensitive to the value of k.

However, there is one more parameter for defining the in-sample set at each month t, S_t. That is, τ which measures the length of the in-sample rolling window. For the main analysis, I set $\tau = 120$, which means I include 10 previous years in S_t that can be used in the out-of-sample predictions. I show that model performance is robust over the various amounts of τ: I vary τ from 24 months to 168 months, the results remain unchanged.

Lastly, I consider different distance metrics rather than Euclidean distance and show the results do not change with respect to the distance measure.

C. Out of Sample Study of Cross-Sections

I begin the out-of-sample study from January 1980 with a 10-years in-sample rolling window ($\tau = 120$). I consider the whole 94 characteristics at once for finding neighboring assets. Based on the neighboring assets, I predict the decile portfolio for each asset in each period, and update portfolios month by month.

C.1. Out-of-sample Performance of Neighboring Assets Strategy

Table 1 shows the results for out-of-sample portfolios when 94 characteristics are considered for finding the neighboring assets. The first four columns show value-weighted portfolios, while the second four columns indicate equally-weighted portfolios. The mean columns show the average of monthly excess realized returns of decile portfolios in the period 1980-2021 in percentage, while t-stat tests if the realized mean excess returns are significantly different from zero. The columns SR report annualized Sharpe ratios. Panel (a) shows the results when the universe of stocks includes all stocks, whereas panels (b) and (c) show the results when considering all but tiny and large stocks, respectively. The "LS" shows the performance of a long-short portfolio which goes long in decile 10 and short in decile 1. On the left side, in panel (a), the average returns vary between -0.40% ($t = -1.00$) and 1.26% ($t = 3.30$) in the lowest and highest decile portfolio, with a 1.66% ($t = 9.10$) of average returns for the long-short portfolio. The annualized Sharpe ratio in this case is 1.40. The long-short equally weighted portfolio even has a higher Sharpe ratio of 2.01 as shown in the right side of Table 1. In the right side of Table 1, the decile 1 average is -0.23%, while decile 10 has a mean average of 1.58%, yielding to a performance of 1.81% for a long-short portfolio with $t = 13.00$. The dispersion for panel (b) is lower, 0.89 ($t = 5.90$) on the left side, and 1.05% ($t = 8.06$) on the right side. When I consider only large stocks this dispersion drops to 0.69 ($t = 4.39$) and

	Value weighted				Equally weighted			
	mean	std	*t*-stat	SR	mean	std	*t*-stat	SR
Panel (a): *All data*								
1	-0.40	9.10	-1.00	-0.15	-0.23	9.32	-0.55	-0.09
2	0.29	8.00	0.82	0.13	0.47	7.33	1.43	0.22
3	0.65	6.33	2.31	0.36	0.72	6.02	2.68	0.41
4	0.65	5.15	2.83	0.44	0.73	4.80	3.42	0.53
5	0.75	4.36	3.85	0.59	0.91	4.03	5.04	0.78
6	0.67	4.07	3.71	0.57	0.88	3.89	5.10	0.79
7	0.82	4.43	4.18	0.64	1.02	4.53	5.07	0.78
8	0.83	5.41	3.46	0.53	1.10	5.57	4.45	0.69
9	1.07	7.66	3.13	0.48	1.17	6.96	3.79	0.58
10	1.26	8.56	3.30	0.51	1.58	8.39	4.22	0.65
LS	1.66	4.10	9.10	1.40	1.81	3.12	13.00	2.01
Panel (b): *All but tiny stocks*								
1	0.24	8.40	0.63	0.10	0.19	8.65	0.49	0.08
2	0.60	6.62	2.03	0.31	0.67	6.61	2.28	0.35
3	0.88	5.49	3.62	0.56	0.84	5.51	3.44	0.53
4	0.79	4.63	3.85	0.59	0.89	4.75	4.20	0.65
5	0.70	4.09	3.83	0.59	0.89	4.30	4.63	0.71
6	0.81	4.26	4.27	0.66	0.90	4.32	4.66	0.72
7	0.81	4.55	4.01	0.62	0.97	4.57	4.78	0.74
8	0.76	5.11	3.35	0.52	1.01	5.23	4.33	0.67
9	1.00	6.34	3.55	0.55	1.12	6.05	4.15	0.64
10	1.12	7.69	3.28	0.51	1.23	7.29	3.80	0.59
LS	0.89	3.37	5.90	0.91	1.05	2.91	8.06	1.24
Panel (c): *Large stocks*								
1	0.33	7.90	0.93	0.14	0.31	8.01	0.86	0.13
2	0.59	6.01	2.20	0.34	0.71	5.86	2.70	0.42
3	0.84	5.29	3.58	0.55	0.88	5.19	3.79	0.58
4	0.72	4.50	3.60	0.56	0.87	4.66	4.18	0.64
5	0.69	4.10	3.76	0.58	0.83	4.24	4.42	0.68
6	0.81	4.18	4.36	0.67	0.87	4.33	4.52	0.70
7	0.82	4.51	4.10	0.63	0.89	4.44	4.50	0.69
8	0.86	4.77	4.07	0.63	0.92	4.91	4.21	0.65
9	0.84	5.72	3.31	0.51	0.98	5.51	3.98	0.61
10	1.01	7.29	3.12	0.48	0.98	7.02	3.12	0.48
LS	0.69	3.51	4.39	0.68	0.67	3.23	4.66	0.72

Table 1- The out-of-sample performance of neighboring assets strategy

Continued from previous page - This table reports the out-of-sample performance of neighboring assets' strategy. I put each asset into one of the portfolios based on the performance of the majority of its neighbors in the past ten years. I recognize each asset's neighbors based on the 94 characteristics listed in table A1. The four left columns show the results of value-weighted portfolios, while the four right columns show the equally-weighted counterpart. Columns titled "mean" show the average monthly excess returns of portfolios from Jan 1980 to Dec 2021 in percentage. std is the monthly standard deviation of portfolios. $t-$stat shows if the risk premiums are significantly different from zero and SR demonstrates the annualized Sharpe ratio. Panel (a) considers all data except the 5% tiniest for creating portfolios, while panels (b) and (c) include all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median), respectively. LS shows the performance of a long-short portfolio which goes long (short) in decile 10 (1). The number of neighbors is considered $k = 1000$.

0.67% ($t = 4.66$) in the left and right side of the Table 1 respectively. Decile 1 (10) has the lowest (highest) average realized returns and all decile portfolios fairly line up monotonically.

Next, I evaluate my long-short portfolios with common factor models. I consider six different factor models, namely CAPM, Fama French three-factor model (FF3), Carhart four-factor model (Carhart), Fama French five-factor model (FF5), q model and augmented q model with expected growth (q^5). I show alphas of a value-weighted long-short portfolio, their $t-$statistics, as well as adjusted R^2 for all of the six different factor models in the panel (a) of Table 2 in time-series regressions. The first three columns are related to value-weighted portfolios while the second three columns show the equally-weighted portfolios. For the case of all sotcks in the left side of Table 2, the long-short portfolio generates a monthly alpha of 1.75% with $t = 9.45$ with respect to FF3. In the left side, the results are even stronger: FF3 monthly alpha 1.84% with $t = 13.00$. None of the other factor models help explain alphas.

These portfolios are well diversified. Table A2 in the Appendix shows the average number of assets in each portfolio in the out-of-sample set. In the case of "all stocks", all decile portfolios contain several hundreds of stocks at each month.

	Value weighted			Equally weighted		
	all stocks	All but tiny stocks	Large stocks	all stocks	All but tiny stocks	Large stocks
$\alpha_{\mathbf{CAPM}}$	1.72	0.99	0.79	1.83	1.22	0.80
t-stat	9.34	6.64	5.03	13.03	9.87	5.68
adj R^2	0.01	0.04	0.03	0.00	0.12	0.06
$\alpha_{\mathbf{FF3}}$	1.75	1.02	0.81	1.84	1.19	0.79
t-stat	9.45	6.76	5.16	13.00	9.69	5.54
adj R^2	0.01	0.04	0.03	-0.00	0.14	0.06
$\alpha_{\mathbf{Carhart}}$	1.36	0.65	0.44	1.54	0.88	0.45
t-stat	8.41	5.28	3.37	12.50	8.93	3.80
adj R^2	0.27	0.38	0.35	0.27	0.46	0.38
$\alpha_{\mathbf{FF5}}$	1.55	0.88	0.63	1.65	0.92	0.53
t-stat	8.17	5.67	3.88	11.52	7.64	3.71
adj R^2	0.04	0.05	0.06	0.06	0.24	0.13
α_q	1.36	0.67	0.44	1.45	0.76	0.36
t-stat	7.40	4.54	2.81	10.81	6.74	2.69
adj R^2	0.12	0.16	0.15	0.19	0.35	0.24
α_{q^5}	1.19	0.50	0.29	1.30	0.65	0.21
t-stat	6.10	3.21	1.74	9.11	5.46	1.48
adj R^2	0.13	0.17	0.16	0.20	0.36	0.25

Table 2- Risk-adjusted returns for long-short portfolios of neighboring assets

This table reports monthly alphas, $t-$values and adjusted R^2 for the out-of-sample performance of long-short portfolios of neighboring assets strategy. I put each asset into one of the portfolios based on the performance of the majority of its neighbors in the past ten years. I recognize each asset's neighbors based on the 94 characteristics listed in table A1. The three left columns show the results for the value-weighted portfolios, while the three right columns show the equally-weighted counterparts. all stocks include the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six different factor models and report alpha and their $t-$stats based on them. These six models include CAPM, the Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French five-factor model (FF5), the q model, and the augmented q model with expected growth (q^5). The number of neighbors is considered $k = 1000$ and the time period is 1980-2021.

The minimum number of stocks in a portfolio in all months is 60. In the case of "all but tiny stocks", the average number of assets in each portfolio varies between 80 and 416. This range for large stocks is between 43 and 196.

C.2. Portfolios of Neighboring Assets

Now, I create portfolios of neighboring assets according to section II.C. I use all but the 5% tiniest stocks for creating portfolios. For each asset at each month between Jan 1980 and Dec 2021, I create $\tau = 120$ months past portfolios, which contain $k = 50$ nearest neighbors at each month.[9] By monthly updating portfolios, I present the out-of-sample performance of decile portfolios in Table 3 panel (a). On average there are 643 assets in each portfolio. The left side shows the results for a value-weighted portfolio while the right side presents the equally-weighted counterparts. Again, the average excess returns line up monotonically, and a value-weighted long-short portfolio generates a monthly average return of 1.66% ($t = 6.17$), with an annualized Sharpe ratio of 0.95. The equally-weighted long-short portfolio has a higher average monthly return of 2.79% ($t = 13.74$) with a Sharpe ratio of 2.12. Panel (b) shows the risk-adjusted returns for long-short portfolios. The CAPM monthly alpha for a value-weighted long-short portfolio is 1.70% ($t = 6.23$), while for an equally-weighted counterpart, it increases to 2.83% ($t = 13.72$).

The results stay the same when I create a distance-weighting portfolio of neighbors, that is, weighting each asset based on the inverse of their distance according to equation 4. The results also are robust to changing the number of neighbors in each portfolio in the in-sample data. The results are weaker when I remove tiny stocks. For example, a value-weighted long-short portfolio from all but tiny stocks generates a monthly FF3 alpha of 0.89% ($t = 3.84$). This alpha drops to 0.46% ($t = 2.21$) when considering only large stocks.

[9] I find that the results are robust to different values of k.

	value-weighted			equally-weighted		
	mean	$t-$stat	SR	mean	t-stat	SR
Panel (a): *Decile portfolios*						
1	-0.37	-1.08	-0.17	-0.61	-1.68	-0.26
2	0.35	1.36	0.21	0.45	1.75	0.27
3	0.47	2.21	0.34	0.57	2.60	0.40
4	0.61	2.84	0.44	0.60	2.73	0.42
5	0.64	3.02	0.47	0.75	3.42	0.53
6	0.72	3.31	0.51	0.85	3.81	0.59
7	0.77	3.56	0.55	0.93	4.05	0.63
8	0.77	3.44	0.53	1.05	4.42	0.68
9	0.92	3.64	0.56	1.26	4.89	0.75
10	1.29	4.20	0.65	2.18	6.44	0.99
LS	1.66	6.17	0.95	2.79	13.74	2.12
Panel (b): *Risk adjusted returns of long-short portfolios*						
	alpha	$t-$stat	adj R^2	alpha	$t-$stat	adj R^2
CAPM	1.70	6.23	0.00	2.83	13.72	0.00
FF3	1.78	6.56	0.02	2.87	13.98	0.02
Carhart	1.05	5.04	0.45	2.38	14.03	0.35
FF5	1.38	5.00	0.06	2.61	12.44	0.05
q	1.03	3.91	0.17	2.36	11.64	0.13
q^5	0.82	2.91	0.18	2.19	10.12	0.14

Table 3- Out-of-sample performance of decile portfolios formed by the portfolios of neighboring assets

This table reports the out-of-sample performance of decile portfolios predicted based on the portfolios of neighboring assets. For each asset, I create a portfolio containing 50 assets of its nearest neighbors for each month in all past 120 months. Then I predict future return of this asset based on the monthly average of its 120 past neighboring portfolios. Finally, I sort assets into decile portfolios based on the predicted returns. I update decile portfolios monthly from Jan 1980 till Dec 2021. The average number of assets in each decile portfolio is 643 assets. Panel (a) shows the out-of-sample performance of decile portfolios. Column "mean" shows the monthly average returns of each portfolio, and SR shows the annualized Sharpe ratio. LS is a long-short portfolio. In panel (b), I show monthly risk-adjusted returns of long-short portfolios with respect to six factor models. adj R^2 shows the adjusted R^2 of time-series regressions. The left (right) side presents value-weighted (equally-weighted) portfolios.

C.3. The Dispersion over Time

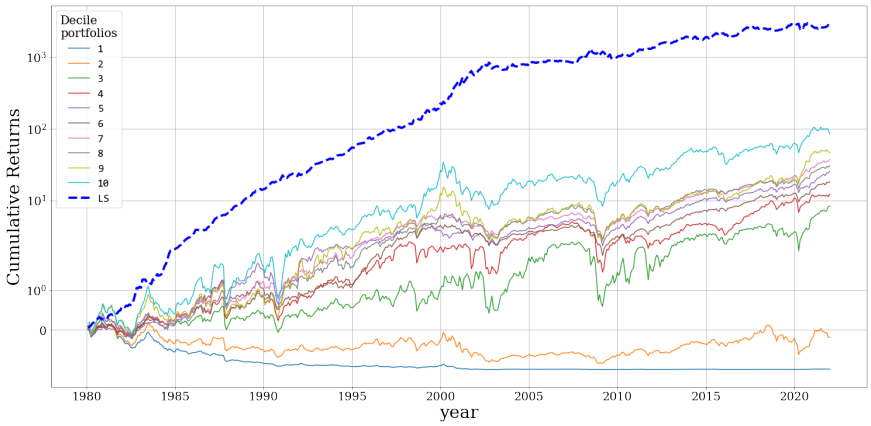
In the long run, my long-short portfolios generate large cumulative returns. The cumulative returns of a long-short portfolio have ascending trends in almost over 40 years. Figure 2 shows the cumulative returns of the decile portfolios from neighboring assets strategy. Panel (a), (b) and (c) illustrate cumulative returns when all, all but tiny, and large stocks are considered for creating the portfolios. The cumulative returns are shown in the vertical axis on a logarithmic scale.

The dispersion generated by decile portfolios persists over time. In Figure 3, I plot the rolling average of realized returns for deciles 1 and 10. The left column shows the 240 months (20 years) of rolling averages, and the right column shows the 10-year moving average of realized returns. Panel (a), (b), and (c) show portfolios consisting of all, all but tiny, and large stocks, respectively. For three datasets, the gap between portfolio 10 and 1 seems to persist over time, although this gap shrinks between the period 2010-2020.

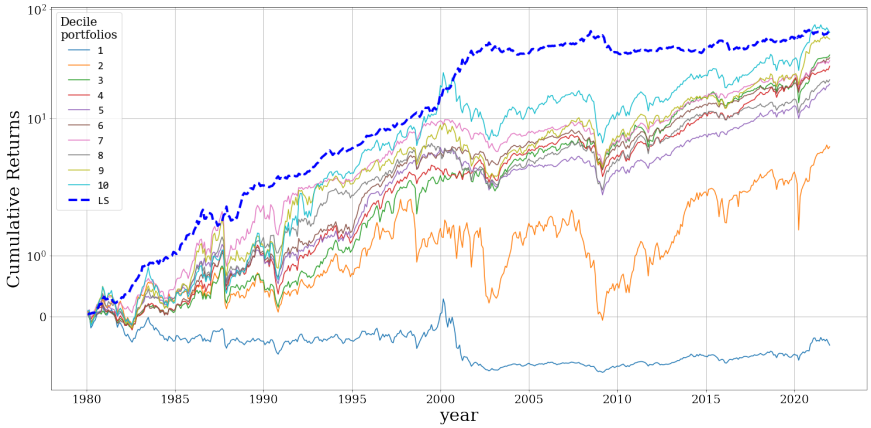
I also show the behavior of the annualized Sharpe ratio for deciles 1 and 10 as well as a long-short portfolio, all value-weighted, in Figure 4.

C.4. Subsample Analysis

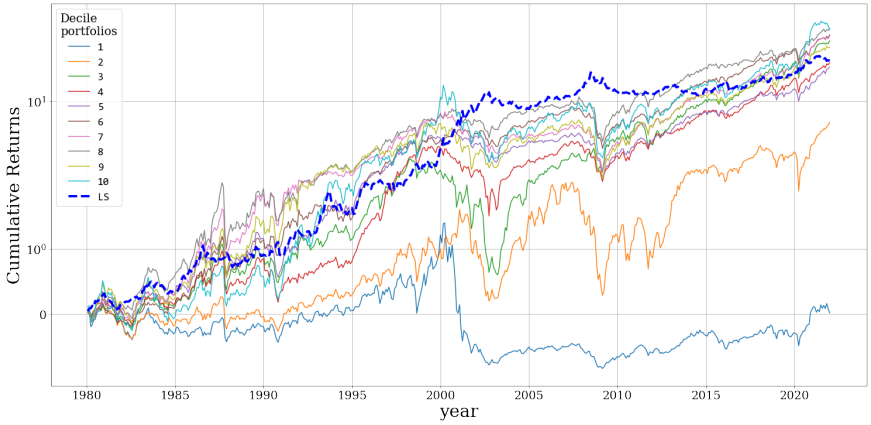
Now I consider three different time horizons and study the performance of value-weighted long-short portfolios in these horizons. First, I consider the first 20 years, i.e. from the beginning of 1980 until the end of 1999 (1980-2000). Second I consider the time after 2000 till the end of 2021 (2000-2021). Lastly, I focus on the performances of the last 12 years (2010-2021). Table 4 shows the results. The first three left columns show the results for the long-short value-weighted portfolios, while the second three columns show the counterpart results for the equally-weighted portfolios. Panel (a) shows the results for the case of all stocks. In



(a) All stocks



(b) All but tiny stocks



(c) Large stocks

Figure 2. Cumulative returns of decile portfolios for the period 1980-2021

This figure shows the cumulative returns of portfolios based on neighboring assets strategy from 1980-2021. "LS" indicates the long-short portfolio. Panel (a) shows portfolios consisting of all but 5% tiniest ones. Panel (b) and (c) show portfolios containing all but tiny (above 20% NSYE percentile) and large stocks (above NYSE median). The y-axis is in a logarithmic scale.

panel (a), all of the alphas are economically meaningful and statistically significant. The Sharpe ratios for the period between 1980-2000 are strongest. Panels (b) and (c) show the results for all but tiny and large stocks. In most cases, the mean returns and alphas are statistically significant and economically large. Overall the results seem robust in different time horizons, but stronger in the beginning of the sample.

D. Importance of Using a Large Set of Characteristics

In this section, I show that the neighboring assets' performance is strongest when a large set of characteristics are used.

I consider two models wherein I find the neighboring assets with fewer characteristics. One is the case that I find neighboring assets with three well-known characteristics: size, book-to-market, and 12-month momentum. The second model adds 9 more characteristics to the previous ones. These 9 characteristics are documented to be among the most important features by Gu et al. (2020) among others. The second model, therefore, contains 12 characteristics including 1-month momentum (Jegadeesh and Titman, 1993), change in momentum (Gettleman and Marks, 2006), maximum daily return (Bali, Cakici, and Whitelaw, 2011), industry momentum (Moskowitz and Grinblatt, 1999), return volatility (Ang, Hodrick, Xing, and Zhang, 2006), dollar trading volume (Chordia, Subrahmanyam, and Anshuman, 2001), sales to price (Barbee Jr, Mukherji, and Raines, 1996), share turnover (Datar, Naik, and Radcliffe, 1998), and asset growth (Datar et al., 1998) as well as the 3 characteristics in the first model. The neighboring assets will be found only using these characteristics and the portfolios are updated monthly.

Table 5 shows the results for value-weighted portfolios. Mean (1), (10), and (LS) show the average excess returns for the lowest and highest deciles, and a long-short portfolio, respectively. SR shows the annualized Sharpe ratio of the

	Value weighted			Equally weighted		
	1980-2000	2000-2021	2010-2021	1980-2000	2000-2021	2010-2021
Panel (a): *All stocks*						
mean	2.52	1.05	0.74	2.41	1.37	1.27
t-stat	11.21	3.69	2.59	13.93	6.38	5.31
SR	2.45	0.79	0.75	3.04	1.36	1.53
$\alpha_{\textbf{CAPM}}$	2.42	1.23	0.94	2.31	1.49	1.40
t-stat	10.72	4.50	3.19	13.43	7.09	5.69
$\alpha_{\textbf{FF3}}$	2.54	1.20	0.91	2.40	1.47	1.42
t-stat	11.00	4.38	3.03	13.65	7.03	5.72
$\alpha_{\textbf{Carhart}}$	2.22	0.95	0.76	2.15	1.27	1.35
t-stat	9.56	4.24	2.75	12.18	7.54	5.55
$\alpha_{\textbf{FF5}}$	2.53	0.85	0.93	2.35	1.18	1.34
t-stat	10.53	3.02	3.03	12.86	5.53	5.43
α_{q}	2.40	0.79	0.76	2.16	1.08	1.22
t-stat	9.51	3.04	2.61	11.54	5.60	5.00
α_{q^5}	2.42	0.59	0.79	2.09	0.92	1.13
t-stat	8.64	2.18	2.64	10.06	4.65	4.56
Panel (b): *All but tiny stocks*						
mean	1.40	0.56	0.35	1.48	0.83	0.67
t-stat	7.21	2.40	1.47	10.29	3.78	3.12
SR	1.57	0.51	0.43	2.25	0.81	0.90
$\alpha_{\textbf{CAPM}}$	1.42	0.72	0.43	1.59	1.04	0.84
t-stat	7.20	3.22	1.74	11.32	5.17	3.83
$\alpha_{\textbf{FF3}}$	1.45	0.71	0.32	1.50	1.02	0.81
t-stat	7.08	3.17	1.31	10.45	5.15	3.73
$\alpha_{\textbf{Carhart}}$	0.95	0.49	0.21	1.16	0.83	0.73
t-stat	5.11	2.78	0.90	8.81	5.30	3.50
$\alpha_{\textbf{FF5}}$	1.45	0.41	0.28	1.35	0.62	0.70
t-stat	6.87	1.80	1.13	9.32	3.19	3.23
α_{q}	1.23	0.30	0.23	1.15	0.58	0.61
t-stat	5.50	1.52	1.01	7.78	3.33	2.92
α_{q^5}	1.21	0.14	0.16	1.11	0.45	0.55
t-stat	4.90	0.67	0.67	6.83	2.49	2.56

continued

Continued from previous page

	Value weighted			Equally weighted		
	1980-2000	2000-2021	2010-2021	1980-2000	2000-2021	2010-2021
Panel (c): *Large stocks*						
mean	0.87	0.66	0.45	0.87	0.69	0.51
t-stat	4.40	2.68	1.74	5.16	2.85	2.32
SR	0.96	0.57	0.50	1.13	0.61	0.67
$\alpha_{\textbf{CAPM}}$	0.86	0.83	0.52	0.89	0.91	0.68
t-stat	4.28	3.54	1.94	5.21	4.07	3.06
$\alpha_{\textbf{FF3}}$	0.89	0.81	0.39	0.84	0.87	0.62
t-stat	4.32	3.46	1.50	4.84	3.95	2.78
$\alpha_{\textbf{Carhart}}$	0.36	0.60	0.30	0.46	0.67	0.52
t-stat	1.96	3.11	1.19	2.81	3.71	2.49
$\alpha_{\textbf{FF5}}$	0.85	0.45	0.37	0.71	0.49	0.60
t-stat	4.12	1.91	1.39	4.02	2.21	2.67
α_q	0.54	0.45	0.39	0.46	0.47	0.52
t-stat	2.44	2.03	1.46	2.53	2.30	2.35
α_{q^5}	0.54	0.27	0.34	0.41	0.28	0.44
t-stat	2.19	1.19	1.26	2.03	1.32	1.94

Table 4- Subsample analysis

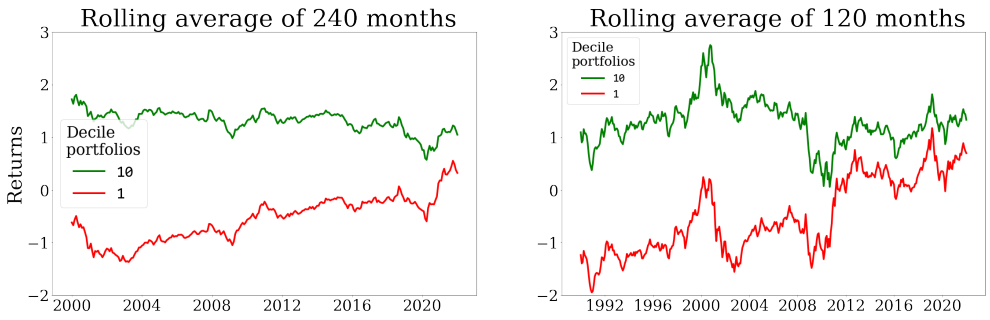
This table reports the performance of value-weighted long-short portfolios based on the neighboring assets strategy in different subsamples. The first three left columns show the value-weighted portfolios, while the second three right columns show the equally-weighted counterpart. The results show different time horizons: Jan 1980 till Dec 1999, Jan 2000 till Dec 2021, and Jan 2010 till Dec 2021. Panel (a), (b) and (c) show the results for all, all but tiny and large stocks, respectively. The results for the full sample are shown in Tables 1 and 2.

long-short portfolio. In all data sets the Sharpe ratios and t-values increase by increasing the number of characteristics, suggesting that the predictive power of neighboring assets increases when more characteristics are taken into account for finding the neighbors. Comparing these results to the previous case where I use 94 characteristics, we observe that the neighboring assets' predictive power is strongest when a large set of characteristics are used.

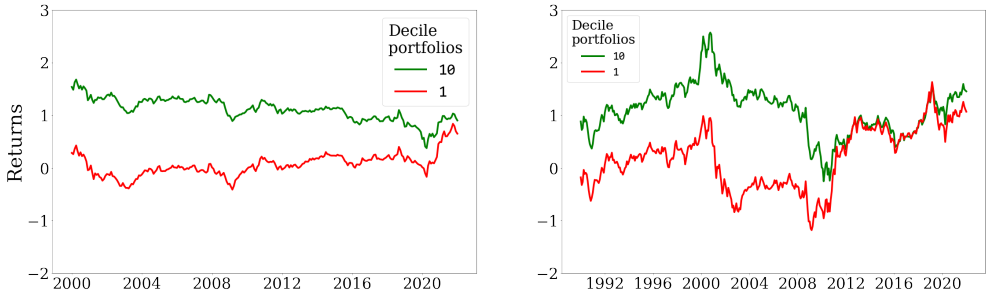
	All data		All but tiny		Large stocks	
	3 char	12 char	3 char	12 char	3 char	12 char
mean (1)	0.23	-0.23	0.36	0.10	0.47	0.25
t-stat	0.61	-0.57	1.11	0.26	1.58	0.74
mean (10)	1.25	1.37	1.17	1.28	1.16	1.17
t-stat	3.63	3.72	3.83	3.85	4.00	3.69
mean (LS)	1.02	1.60	0.82	1.18	0.68	0.92
t-stat	4.29	8.22	3.92	6.79	3.30	5.68
SR	0.66	1.27	0.61	1.05	0.51	0.88
$\alpha_{\mathbf{CAPM}}$	1.16	1.69	0.91	1.30	0.76	1.00
t-stat	4.88	8.64	4.34	7.52	3.66	6.14
$\alpha_{\mathbf{FF3}}$	1.28	1.75	1.03	1.33	0.91	1.04
t-stat	5.52	8.93	5.09	7.67	4.59	6.44
$\alpha_{\mathbf{Carhart}}$	0.60	1.29	0.36	0.87	0.24	0.62
t-stat	3.64	7.87	2.97	6.47	2.10	4.88
$\alpha_{\mathbf{FF5}}$	1.03	1.41	0.84	1.10	0.80	0.85
t-stat	4.30	7.17	4.00	6.19	3.89	5.11
α_q	0.60	1.19	0.47	0.88	0.44	0.64
t-stat	2.73	6.37	2.39	5.20	2.24	4.07
α_{q^5}	0.46	1.05	0.28	0.69	0.29	0.44
t-stat	1.96	5.30	1.35	3.82	1.36	2.66

Table 5- The performance of neighboring assets strategy based on 3 and 12 characteristics

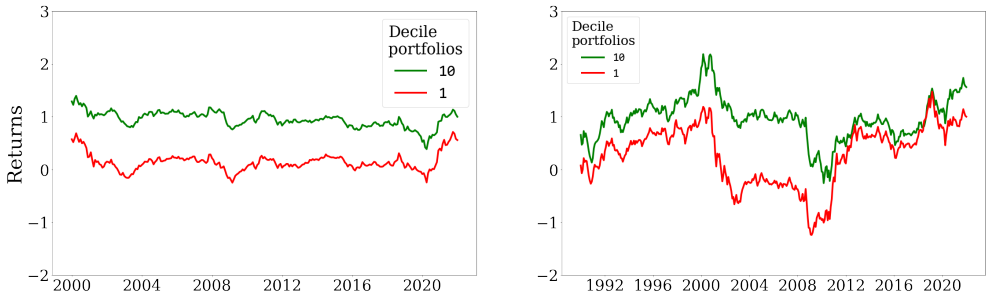
This table reports the average monthly out-of-sample performance of value-weighted portfolios based on 3 and 12 characteristics. I put each asset into one of the portfolios based on the performance of the majority of its neighbors in the past ten years. In the columns “3 char” I recognize each asset’s neighbors based on the size, book-to-market, and momentum. In the columns “12 char”, I use size, book-to-market, momentum, 1-month momentum, change in momentum, maximum daily return, industry momentum, return volatility, dollar trading volume, sales to price, share turnover, and asset growth. Rows titled "mean (1), (10) and (LS)" show the average monthly excess returns of portfolios 1, 10 and long-short portfolio from Jan 1980 to Dec 2021 in percentage and SR demonstrates the annualized Sharpe ratio of the long-short portfolio. I also show the monthly alphas of the long-short portfolios with respect to different factor models.



(a) All stocks



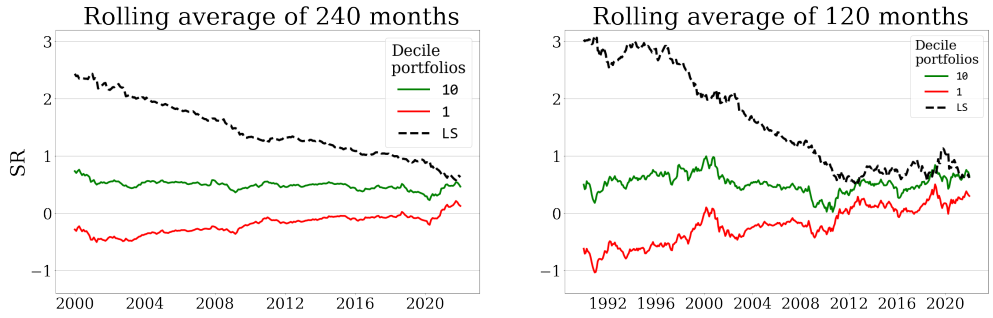
(b) All but tiny stocks



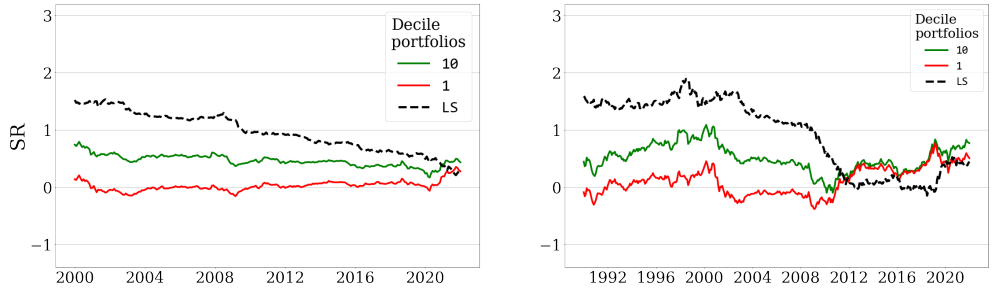
(c) Large stocks

Figure 3. Rolling average of returns in decile 1 and 10

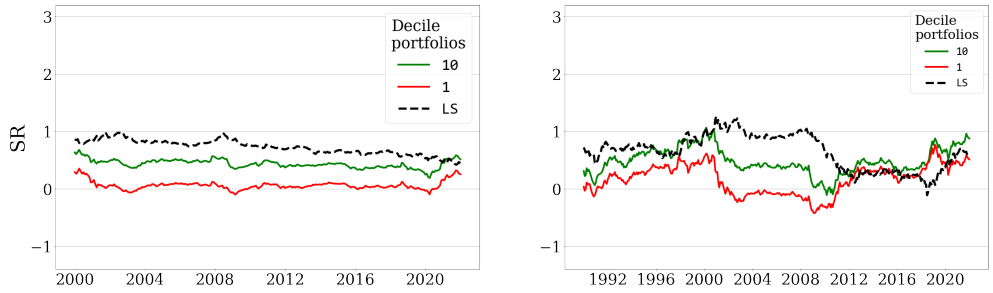
This figure shows the rolling average of returns from decile portfolios 1 (red) and 10 (green), where stocks are sorted into value-weighted deciles according to neighboring assets strategy. The left column shows the rolling average when the rolling window is 240 months, while the right column shows the 120-month rolling window. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks.



(a) All stocks



(b) All but tiny stocks



(c) Large stocks

Figure 4. Rolling average of Sharpe Ratio for decile 1, 10 and the long-short portfolio

This figure shows the rolling average of Sharpe ratios from decile portfolios 1 (red) and 10 (green), as well as for the long-short portfolio (black). Stocks are sorted to value-weighted deciles according to neighboring assets strategy. The left column shows the rolling average when the rolling window is 240 months, while the right column shows the 120-month rolling window. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks.

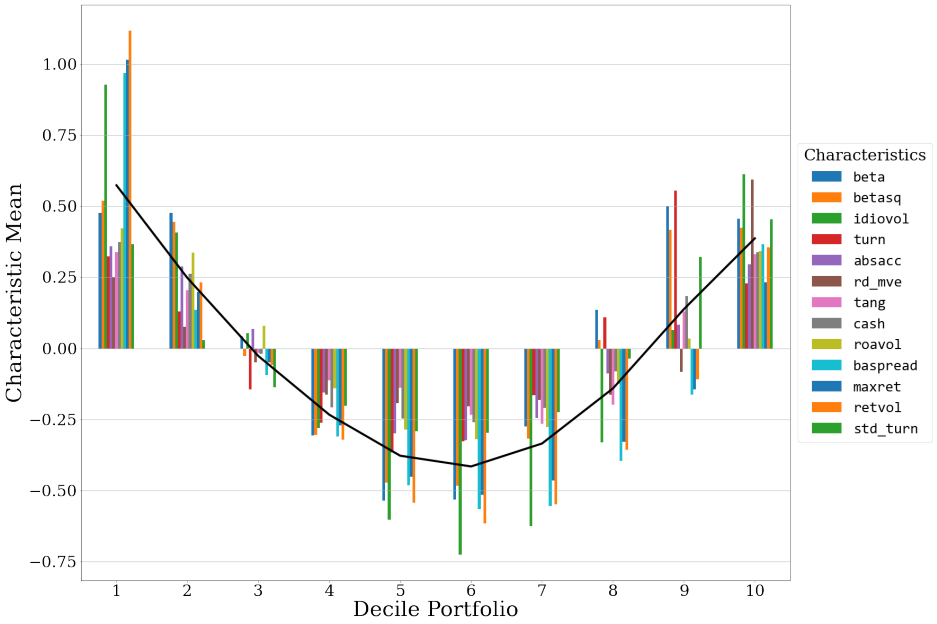
E. Characteristics Patterns

As the neighboring assets strategy maps a large set of characteristics to the expected returns, studying the patterns of characteristics in each portfolio sheds light on the relationship between characteristics and expected returns.

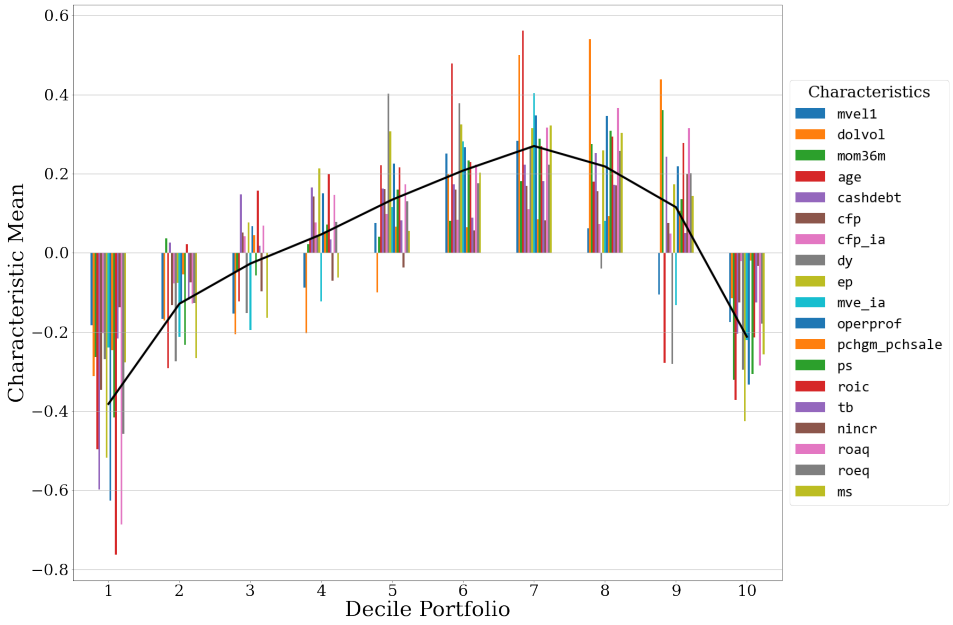
To begin with, considering all 94 characteristics, I find the average of characteristics in each portfolio. Then I find the time-series average of all characteristics from 1980-2021 for each decile portfolio. I observe that most characteristics do *not* linearly monotonically line up with portfolios of neighboring assets. Instead, they have some of the following patterns: First, some characteristics show a U-shape relationship with expected returns—they are high in portfolios 1 and 10, and low in other portfolios. Examples of these characteristics include market beta and idiosyncratic volatility. Figure 5a shows these characteristics. Another group of characteristics (such as firm size and firm age) show an inverse of a U-shape with expected returns. That is, stocks within portfolio 1 and 10 have lowest values of these characteristics. Figure 5b shows these characteristics. The third group of characteristics show, on average, an M-shape relationship with expected returns. Figure 5c shows this group of characteristics. The results suggest that the relationship between the joint distribution of characteristics and expected returns can be quite complex and non-linear.

F. Neighboring Assets versus Momentum Strategy

A momentum strategy predicts an asset's future returns based on its own past, while the neighboring assets strategy uses the past performance of neighboring assets to predict future expected returns. One important consideration is to contrast the momentum strategy with neighboring assets strategy. I show that neighboring assets strategy is distinct from momentum. Also, I show that the neighboring

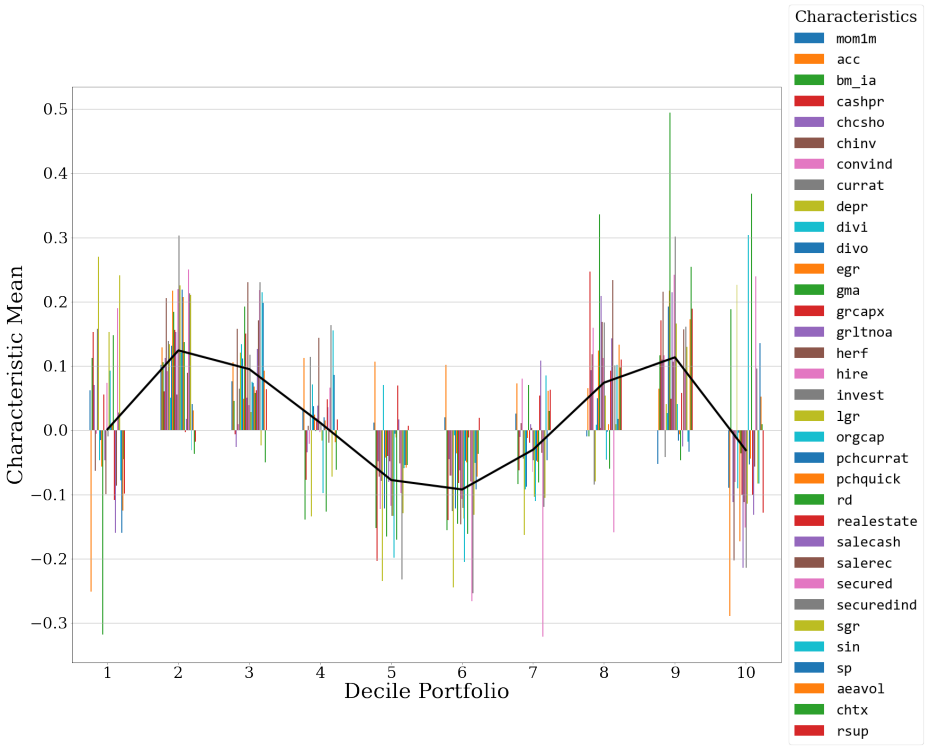


(a) Characteristics with a U-shape relationship



(b) Characteristics with an inverse U-shape relationship

continued



(c) Characteristics with an M shape relationship

Figure 5. Average of characteristics in each of the decile portfolios

This figure shows the pattern of characteristics among decile portfolios formed based on the neighboring assets strategy. I define the characteristics of each decile portfolio as the average characteristics of its stocks. The mean of characteristics shows the average of characteristics in each portfolio over time. Panel (a) shows characteristics which first decrease and then increase with the decile portfolio, yielding to have a U-shape relationship with portfolios of neighboring assets. Panel (b) illustrates characteristics with an inverse-U shape with mean returns. Panel (c) shows the characteristics which first increase, then decrease, and then again increase and decrease, having an M-shape pattern. The black line shows the average of bars in a decile portfolio. The x-axis shows the decile portfolios, while the y-axis shows the mean of characteristics. The characteristics are from portfolios containing all but the tiniest 5% stocks. The list of characteristics is defined in Table A1. Before creating portfolios, all characteristics are cross-sectionally standardized and winsorized at 0.01 level. All portfolios are value-weighted. The time period is 1980-2021.

assets strategy explains a large set of momentum anomalies (including the momentum factor), while the momentum factor fails to explain the performance of the neighboring strategy.

F.1. Dropping an Assets' Own Lag

The first test I consider to rule out the possibility that the neighboring assets strategy is merely a variant of momentum involves excluding an asset's own lags from the in-sample set when identifying its neighbors. This ensures that the return predictability does not stem from the asset's own past performance. The performance of neighboring assets strategy does not change when I drop an asset's own lags from the in-sample set. Table 6 panel (a) shows the out-of-sample alphas of a long-short portfolio when excluding each assets' own lags: The performance of the neighboring assets strategy remains the same. Carhart's four-factor alpha is 1.68% with $t = 7.09$ in a value-weighted portfolio. This test controls not just for momentum, but also short- and long-term reversals.

F.2. Skipping the Five Most Recent Years

The second test I perform is that, in the neighboring assets strategy, where I find each asset's neighbors in the past, I skip the past five recent years from the in-sample set. Thus, I only find each asset's neighbors from the past five years to the past ten years. This method also controls for variations of momentum. Table 6 panel (b) shows the out-of-sample performance of the long-short portfolio, confirming that the results stay robust. Carhart's four-factor alpha is 1.51% with $t = 6.48$ in a value-weighted portfolio.

	value-weighted			equally-weighted		
	alpha	$t-$stat	adj R^2	alpha	t-stat	adj R^2
Panel (a): *Dropping an asset own lags*						
$\alpha_{\textbf{CAPM}}$	2.35	7.37	0.00	1.24	5.04	0.01
$\alpha_{\textbf{FF3}}$	2.53	8.11	0.06	1.33	5.47	0.03
$\alpha_{\textbf{Carhart}}$	1.68	7.09	0.48	0.60	3.57	0.56
$\alpha_{\textbf{FF5}}$	2.29	7.09	0.07	1.12	4.44	0.05
α_q	1.77	5.67	0.15	0.58	2.51	0.22
α_{q^5}	1.45	4.35	0.16	0.41	1.68	0.22
Panel (b): *Dropping the five recent years*						
$\alpha_{\textbf{CAPM}}$	2.22	7.01	0.00	1.13	4.79	0.01
$\alpha_{\textbf{FF3}}$	2.38	7.63	0.05	1.20	5.12	0.03
$\alpha_{\textbf{Carhart}}$	1.51	6.48	0.48	0.54	3.11	0.49
$\alpha_{\textbf{FF5}}$	2.23	6.87	0.05	1.00	4.11	0.04
α_q	1.66	5.33	0.14	0.51	2.28	0.20
α_{q^5}	1.35	4.06	0.15	0.35	1.48	0.21

Table 6- Out-of-sample performance of neighboring assets strategy when excluding each asset's own lags or skipping the five recent years

This table reports the out-of-sample performance of neighboring assets' strategy when controlling for each asset's own lags. In panel (a), when finding an assets neighbors in the past, I remove each asset's own lags. In panel (b), when finding each asset's neighbors in the past, I skip the five most recent years and find an asset's neighbors between the past five to ten years. Then I put each asset into one of the portfolios based on the performance of the majority of its neighbors. I show monthly risk-adjusted returns of long-short portfolios with respect to six factor models. adj R^2 shows the adjusted R^2 of time-series regressions. The left (right) side presents value-weighted (equally-weighted) portfolios.

F.3. Momentum Anomalies

A momentum strategy contains buying (selling) high- (low-) mean stocks repeatedly (Conrad and Kaul, 1998). The neighboring asset strategy also buys (sells) high- (low-) mean stocks, but based on a large set of characteristics rather than the past returns of the asset itself. Thus, it is important to contrast the two strategies. In this section, I show that while the momentum strategy fails to explain the

neighboring assets strategy, the latter explains the former. Moreover, I show that the neighboring assets strategy explains a wide range of momentum anomalies.

To begin, I consider a list of 41 momentum anomalies provided by Hou, Xue, and Zhang (2020). Table A3 in the Appendix describes these anomalies. I make long-short value-weighted portfolios from these momentum anomalies and regress them on my value-weighted long-short portfolio from all but tiny stocks. I treat this long-short portfolio as a neighboring factor (`Neighb`). With a critical value of $|t| \geq 3$, only 1 anomaly portfolio can generate a significant alpha when exposed to my `Neighb` factor. As shown in Table 7 panel (a), other factor models generate more significant alphas. For $|t| \geq 1.96$ only 7 alphas stand significant with my factor, while even a Carhart model cannot explain 14 of these momentum anomalies. The average of absolute values of alphas for 41 momentum anomaly is 0.20 with an average absolute value of $t = 1.04$. `Neighb` factor has a low average absolute t values compared to other factor models in explaining the anomalies. A Carhart model generates an average absolute t value of 1.53 and CAPM, FF3, and FF5 produce average absolute values of t greater than 2.62.

Next, I regress my `Neighb` factor on the momentum factor. I consider three long-short portfolios as a proxy for the momentum factor. First, I collect the up-minus-down factor from Fama and French Data Library which is the momentum factor. Second, I sort assets into 10 decile portfolios based on their 12- and 6-month momentum and create long-short portfolios from extreme deciles. Table 7 panel (b) shows the regression result. In fact, the momentum factor, as a tradable portfolio, no longer generates a profit when exposed to the `Neighb` portfolio. The intercept simply is -0.40% with $t = -2.17$, with a negative sign. The `Neighb` factor coefficient, in this case, is 0.79 ($t = 16.57$), meaning that the `Neighb` factor strongly explains the momentum factor. When using a long-short 12-month momentum portfolio

	avg $\lvert\alpha\rvert$	avg $\lvert t\rvert$	num $\lvert t\rvert \geq 1.96$	num $\lvert t\rvert \geq 2.58$	num $\lvert t\rvert \geq 3$	avg $\lvert\beta_{\text{Neighb.}}\rvert$	avg $\lvert t(\beta_{\text{Neighb.}})\rvert$
Panel (a): *Regressing momentum anomalies on different models*							
Neighb.	0.20	1.04	7	2	1	0.43	8.71
Neighb. + Market	0.19	1.04	6	2	2	0.41	8.40
CAPM	0.49	2.78	33	23	16		
FF3	0.57	3.35	39	30	25		
Carhart	0.22	1.53	14	7	4		
FF5	0.46	2.64	32	20	11		
q **model**	0.19	1.17	8	5	4		
q^5 **model**	0.18	1.03	6	2	1		

Panel (b): *Regressing momentum factors on neighboring Assets long-short factor*

Regressand	alpha	$t-$stat	$\beta_{\text{Neighb.}}$	$t(\beta_{\text{Neighb.}})$
Mom factor	-0.40	-2.17	0.56	13.51
LS mom 12m	-0.68	-1.98	1.19	15.25
LS mom 6m	-0.43	-1.31	1.00	13.59

Panel (c): *Regressing neighboring Assets long-short factor on momentum factors*

Regressor	alpha	$t-$stat	β_{MOM}	$t(\beta_{\text{MOM}})$
Mom factor	1.41	8.94	0.47	13.51
LS mom 12m	1.32	8.62	0.27	15.25
LS mom 6m	1.33	8.41	0.27	13.59

Table 7- Performance of neighboring assets long-short portfolio as a factor on momentum anomalies

This table shows the regression results when a long-short portfolio formed from neighboring assets is used as a tradeable factor (Neighb factor). I collect 41 momentum anomalies provided by Hou et al. (2020) and check if my Neighb factor can explain a long-short strategy from those anomalies. Panel (a) shows the results of time-series regression of Neighb factor as well as other factor models on momentum anomalies. In panel (b), I regress momentum factors on my Neighb factor. I consider three different portfolios as a proxy for momentum anomalies: a momentum factor from Fama and French data library, and two long-short portfolios from decile portfolios sorted on the 12- and 6-momentum. Panel (c) shows the regression results when these momentum factors are regressed on the Neighb Portfolio. I use a value-weighted long-short portfolio from all but tiny stocks as Neighb factor. num $\lvert t\rvert$ shows the number of momentum anomalies that generate an absolute t-value of alpha greater than 1.96, 2.58, and 3 from the given factor. avg $\lvert\beta\rvert$ shows the average values of beta. avg $\lvert t(\beta)\rvert$ shows the average values of the t-values of the coeffieicnts.

as a proxy for momentum factor, the alpha is still negative, namely -0.68 with $t = -1.98$. For a long-short 6-month momentum portfolio, the alpha magnitude is -0.43 ($t = -1.31$). Next, I flip the regressions: I regress momentum factors on my long-short portfolio. Panel (c) shows the results. With momentum factors, 12- and 6-month long-short momentum portfolios, `Neighb` portfolio generates a statistically significant positive alpha of 1.41, 1.32, and 1.33 with t-statistics of 8.94, 8.62, and 8.41, respectively. The result suggests that the neighboring asset strategy subsumes the momentum factor and momentum anomalies, whereas the momentum factor does not explain the neighboring asset strategy.

G. Neighboring Assets and the Factor Zoo

Now I contrast the neighboring assets' long-short portfolio against a large set of long-short factors. More specifically, from each characteristic, I create a value-weighted long-short portfolio by sorting stocks into deciles on that characteristic. This leads to 94 long-short portfolios. Then I regress my neighboring assets' long-short portfolio on each of the long-short portfolios on characteristics. The first column of Table 8 shows the summary statistics of 94 t-statistics for alphas, that the neighboring assets strategy generates. The average t-values of alpha for neighboring assets strategy is 9.02. No alpha t-statistics is lower than 8.39. This suggests that none of the factors in the set of the 94 factor zoo can explain the neighboring assets' performance.

In the second column of Table 8, I reverse the analysis: I regress each of the 94 factors on the neighboring assets long-short portfolio. The average absolute t-values for alpha that all of the factors generate is 1.08. Only 12 factors generate an alpha with absolute t-value greater than 1.96, while only 3 factors have a $|t| > 3$. This shows that the neighboring asset long-short portfolio does a decent job in explaining

	Neighboring assets $\|t\|$-stats against factor zoo	Factor zoo $\|t\|$-stats against neighboring assets	Factor zoo $\|t\|$-stats against CAPM
mean	9.02	1.08	2.13
min	8.39	0.06	0.03
25%	8.93	0.53	1.09
50%	9.07	0.87	2.00
75%	9.12	1.52	3.00
max	9.65	3.34	6.20
count	94	94	94

Table 8- Neighboring assets versus the factor zoo

This table contrasts the neighboring assets portfolio against the factor zoo. For each of the 94 factors, I regress the value-weighted long-short portfolio of the neighboring assets on the factor, and I obtain the alpha and its t-value. The first column shows the descriptive statistics of $|t|$-value of alphas from time-series regression of neighboring assets' strategy against the factor zoo. The second column reverses the regression: I regress each of the 94 factors on the neighboring assets long-short portfolio and report the descriptive statistics of the $|t|$-values of alphas. The third column shows the descriptive statistics of the $|t|$-values of alphas from regressing each of the 94 factors on the value-weighted market factor. All factors are a value-weighted long-short portfolio from sorting stocks into deciles based on one characteristic. The period is 1980-2021.

a large set of factors.

To compare these results with a benchmark, the third column of Table 8 shows the summary statistics of absolute t-stats of alphas of the factor zoo when regressed on the market factor. The average absolute t-stats is 2.13, while at least half of the factors show a $|t|$-value of greater than 2 when regressed on the value-weighted market factor.

Lastly, I regress the long-short portfolio of the neighboring assets on all 94 factors simultaneously. The neighboring assets portfolio strategy still generates an out-of-sample monthly alpha of 0.92% ($t = 5.19$).

H. Principle Portfolios of Neighboring Assets

To understand the factor structure of deciles portfolios of neighboring assets, I perform a Principal Component (PC) Analysis (Kozak, Nagel, and Santosh, 2018). In a similar vein to Clarke (2022), I find the PC portfolios of neighboring assets' deciles and find the correlation of the first three PC portfolios with other factors. Figure 6 shows the absolute correlation of the most correlated factors with the first three PC portfolios of neighboring assets. The top figure shows that the market factor has the highest correlation of 0.93% with the first PC portfolio. The next most correlated factors with the first PC portfolio are long-short portfolios of stocks sorted on beta, beta squared, zerotrade, and idiosyncratic volatility. The middle figure shows factors with the highest absolute correlation with the second PC. Long-short portfolios based on operating profitability, bid-ask spread, and RMW (robust-minus-weak) show the highest correlation with the second PC portfolio.

The last figure shows the absolute correlations of factors with the third PC portfolio. Portfolios sorted on momentum display the highest correlation with the third PC portfolio. The results suggest that while a large set of one-sorted factors are related to the portfolios of neighboring assets, they show little correlation with the PC portfolios and are distinct from neighboring asset deciles.

I. Robustness Checks

In this section, I show that the results are not sensitive to the neighborhood definition, nor to the distance measurement. I show that the results are robust to different rolling years embedded in the in-sample set. I show that the results stay the same if we randomly sample from the in-sample set and if in the in-sample set, we use average returns instead of realized returns for sorting stocks to deciles. I show that the industries do not explain the results.

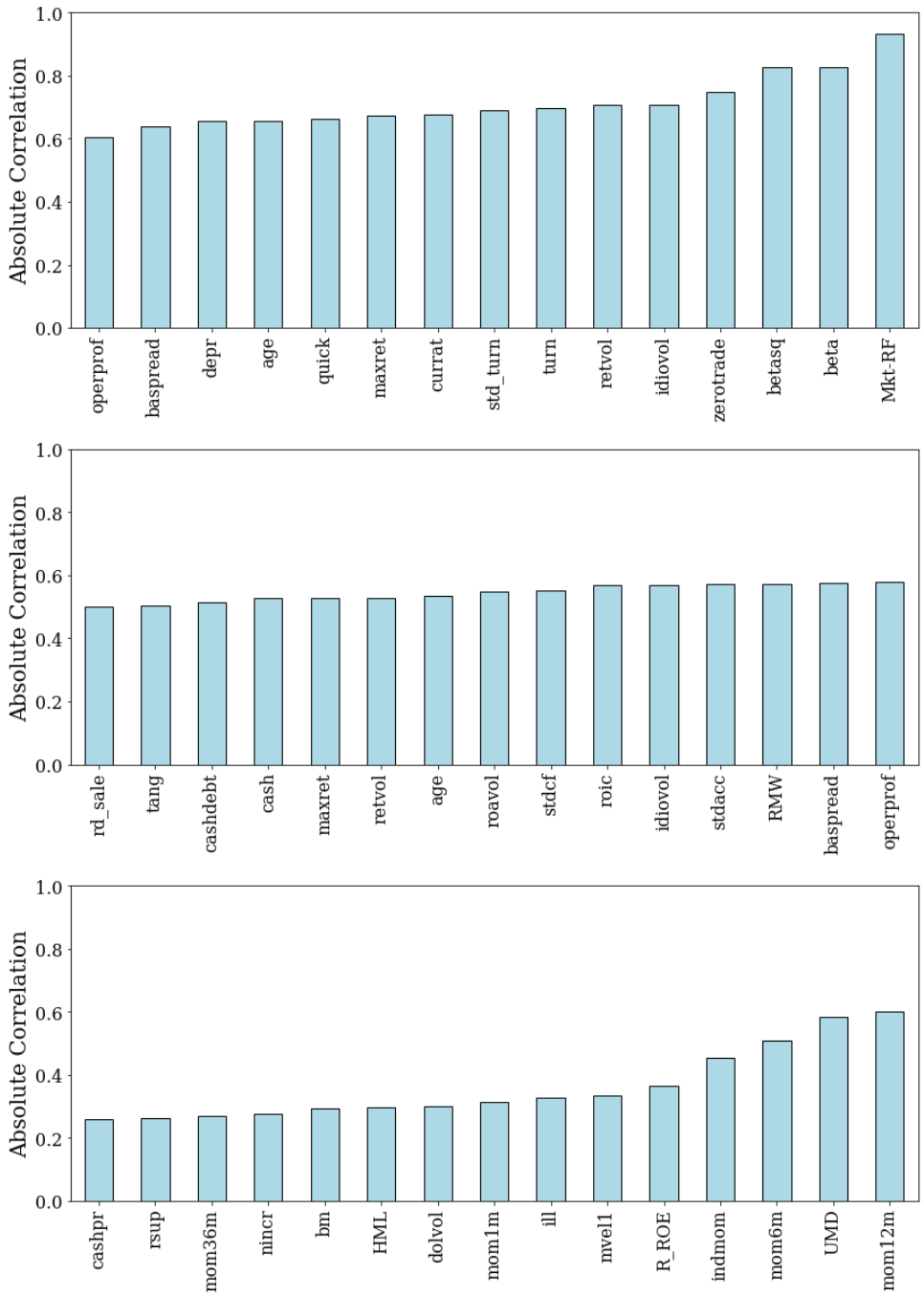


Figure 6. Absolute correlation of top factors with the PC portfolios of neighboring assets

This figure shows the absolute correlation of top factors with the first three Principle Component (PC) portfolios of the neighboring assets. The top photo shows the top 15 factors which have the highest correlation with the first PC portfolio. The middle and bottom photos show the top 15 factors that have the highest correlation with the second and third PC portfolios, respectively. Mkt-RF is the value-weighted market factor. RMW, HML, and UMD are robust-minus-weak, high-minus-low, and up-minus-down factors. The characteristics are defined in Appendix Table A1.

I.1. Controlling for Industries

I start the robustness checks from the industries. Based on the SIC 2-digit codes, I identify the industries. Then I create industry-adjusted returns of the neighboring asset deciles. That is, in each month, I subtract the return of each stock from its average industry. Table 9 shows the value-weighted and equally-weighted out-of-sample performance of a long-short portfolio of neighboring assets when creating industry-adjusted returns.

As Table 9 shows, the average excess returns of a long-short portfolio is 1.40% ($t = 7.21$) with a Sharpe ratio of 1.57. The FF5 monthly alpha is 1.45% ($t = 6.87$). Overall, the results are not weaker when creating industry-adjusted returns.

I.2. Different Number of Neighbors

Now I check how the results are sensitive to the number of neighbors considered for classifying assets into portfolios. For robustness checks, I hold everything else constant and change k, and report the dispersion produced by value-weighted decile portfolios 1 and 10. I repeat the analysis for three sets of data in Table 10. I change k from 100 to 5000 and report the performance of a value-weighted long-short portfolio. When considering all stocks, the average realized return of a long-short portfolio does not go below 1.08%. The performance of the model starts increasing by increasing k, and it reaches its maximum around between $k = 900$ and $k = 1200$. Panels (b) and (c) show the performance of long-short portfolios from all but tiny and large stocks. Overall the results are suggestive that the model performance is not sensitive to the neighborhood definition and persists over all three data sets.

	Value weighted			Equally weighted		
	All stocks	All but tiny	Large stocks	All stocks	All but tiny	Large stocks
mean	1.40	0.56	0.35	1.48	0.83	0.67
t-stat	7.21	2.40	1.47	10.29	3.78	3.12
SR	1.57	0.51	0.43	2.25	0.81	0.90
$\alpha_{\textbf{CAPM}}$	1.42	0.72	0.43	1.59	1.04	0.84
t-stat	7.20	3.22	1.74	11.32	5.17	3.83
$\alpha_{\textbf{FF3}}$	1.45	0.71	0.32	1.50	1.02	0.81
t-stat	7.08	3.17	1.31	10.45	5.15	3.73
$\alpha_{\textbf{Carhart}}$	0.95	0.49	0.21	1.16	0.83	0.73
t-stat	5.11	2.78	0.90	8.81	5.30	3.50
$\alpha_{\textbf{FF5}}$	1.45	0.41	0.28	1.35	0.62	0.70
t-stat	6.87	1.80	1.13	9.32	3.19	3.23
α_{q}	1.23	0.30	0.23	1.15	0.58	0.61
t-stat	5.50	1.52	1.01	7.78	3.33	2.92
α_{q^5}	1.21	0.14	0.16	1.11	0.45	0.55
t-stat	4.90	0.67	0.67	6.83	2.49	2.56

Table 9- Industry adjusted portfolios

This table shows the out-of-sample performance of the long-short portfolio of neighboring assets strategy where the portfolio returns are industry-adjusted. I identify the industries based on the 2-digit SIC code. When creating portfolios, I subtract the realized returns of each assets from its average industry at each month. The three left columns show the results for the value-weighted portfolios, while the three right columns show the equally-weighted counterparts. all stocks include the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six different factor models and report alpha and their $t-$stats based on them. These six models include CAPM, the Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French 5 factor model (FF5), the q model, and augmented q model with expected growth (q^5). The number of neighbors is considered $k = 1000$ and the time period is 1980-2021.

I.3. Different Rolling Windows in the In-sample Set

In my framework, I assume that observations from several years ago are relevant for classifying a new asset. The main analysis is based on 10 years ($\tau = 120$ months) rolling window. In this section, I repeat the analysis by varying τ—the number of months in the sample period—from 2 to 14 years of rolling windows and show that

k	100	300	700	900	1200	1500	2000	2500	5000
Panel (a): *All data*									
mean	1.08	1.60	1.61	1.56	1.61	1.69	1.68	1.63	1.58
t-stat	5.84	8.69	8.99	8.47	8.77	9.02	8.98	8.60	8.19
SR	0.90	1.34	1.39	1.31	1.35	1.39	1.39	1.33	1.26
$\alpha_{\textbf{CAPM}}$	1.16	1.67	1.67	1.63	1.66	1.76	1.73	1.68	1.66
t-stat	6.24	8.96	9.22	8.78	8.93	9.27	9.10	8.76	8.53
$\alpha_{\textbf{FF3}}$	1.21	1.73	1.71	1.66	1.68	1.78	1.75	1.70	1.69
t-stat	6.52	9.30	9.41	8.91	9.02	9.33	9.16	8.87	8.66
$\alpha_{\textbf{Carhart}}$	0.81	1.35	1.33	1.27	1.30	1.39	1.38	1.31	1.25
t-stat	5.04	8.23	8.37	7.79	7.92	8.26	8.06	7.75	7.53
$\alpha_{\textbf{FF5}}$	0.99	1.54	1.52	1.45	1.46	1.57	1.59	1.56	1.47
t-stat	5.23	8.07	8.13	7.61	7.67	8.03	8.04	7.83	7.33
α_q	0.77	1.35	1.33	1.25	1.25	1.34	1.39	1.32	1.18
t-stat	4.25	7.27	7.36	6.81	6.81	7.10	7.23	6.86	6.21
α_{q^5}	0.63	1.21	1.17	1.08	1.13	1.18	1.20	1.23	1.10
t-stat	3.24	6.12	6.07	5.50	5.72	5.86	5.89	5.98	5.40
Panel (b): *All but tiny stocks*									
mean	0.68	0.74	0.92	0.90	0.92	0.96	0.96	0.90	1.04
t-stat	3.88	4.92	5.85	5.90	5.80	6.22	5.95	5.52	6.34
SR	0.60	0.76	0.90	0.91	0.89	0.96	0.92	0.85	0.98
$\alpha_{\textbf{CAPM}}$	0.80	0.85	1.02	1.01	1.04	1.08	1.09	1.05	1.22
t-stat	4.58	5.62	6.53	6.62	6.62	7.13	6.88	6.59	7.75
$\alpha_{\textbf{FF3}}$	0.84	0.86	1.04	1.03	1.05	1.08	1.08	1.04	1.18
t-stat	4.82	5.71	6.66	6.72	6.65	7.04	6.76	6.47	7.53
$\alpha_{\textbf{Carhart}}$	0.44	0.47	0.64	0.65	0.66	0.69	0.67	0.63	0.79
t-stat	3.00	3.95	5.16	5.23	5.15	5.63	5.28	4.92	6.23
$\alpha_{\textbf{FF5}}$	0.64	0.72	0.94	0.90	0.90	0.90	0.89	0.79	0.85
t-stat	3.54	4.60	5.77	5.67	5.50	5.73	5.43	4.86	5.50
α_q	0.43	0.46	0.72	0.68	0.69	0.70	0.67	0.57	0.64
t-stat	2.45	3.17	4.63	4.52	4.43	4.67	4.32	3.73	4.41
α_{q^5}	0.34	0.33	0.54	0.52	0.52	0.53	0.54	0.47	0.56
t-stat	1.80	2.13	3.26	3.24	3.17	3.35	3.26	2.89	3.57

continued

Continued from previous page

k	100	300	700	900	1200	1500	2000	2500	5000
Panel (c): *Large stocks*									
mean	0.45	0.57	0.61	0.64	0.73	0.71	0.76	0.79	0.69
t-stat	2.54	3.70	3.83	4.03	4.70	4.49	4.73	4.83	4.18
SR	0.39	0.57	0.59	0.62	0.73	0.69	0.73	0.75	0.64
$\alpha_{\mathbf{CAPM}}$	0.54	0.65	0.71	0.75	0.85	0.85	0.89	0.94	0.87
t-stat	3.08	4.19	4.48	4.77	5.52	5.47	5.61	5.82	5.41
$\alpha_{\mathbf{FF3}}$	0.63	0.70	0.75	0.77	0.87	0.86	0.90	0.95	0.86
t-stat	3.66	4.54	4.69	4.89	5.63	5.48	5.61	5.87	5.29
$\alpha_{\mathbf{Carhart}}$	0.22	0.35	0.35	0.38	0.51	0.49	0.52	0.57	0.48
t-stat	1.52	2.68	2.72	2.98	3.94	3.76	3.90	4.23	3.53
$\alpha_{\mathbf{FF5}}$	0.53	0.59	0.59	0.57	0.64	0.60	0.67	0.72	0.60
t-stat	2.94	3.64	3.58	3.52	4.07	3.80	4.11	4.38	3.67
α_q	0.33	0.42	0.40	0.39	0.48	0.43	0.49	0.54	0.40
t-stat	1.89	2.70	2.54	2.50	3.17	2.88	3.19	3.45	2.59
α_{q^5}	0.20	0.23	0.24	0.21	0.31	0.24	0.33	0.38	0.21
t-stat	1.06	1.36	1.44	1.27	1.90	1.48	1.98	2.27	1.29

Table 10- Performance of a value-weighted long-short portfolio for different values of k

This table reports the performance of a value-weighted long-short portfolio formed based on a different number of neighbors k. I consider six factor models and report alpha and their $t-$stats based on them. These six models include CAPM, Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French 5 factor model (FF5), the q model, and augmented q model with expected growth (q^5). The regression periods capture realized returns in percentage from 1980 to 2021. Panel (a) reports the results for all but 5% tiniest stocks, panel (b) shows the counterpart results for all but tiny stocks, and panel (c) represents the results for large stocks.

the results remain the same.

Table 11 shows the performance of value-weighted long-short portfolios for different τ from 1980 to 2021. To keep everything else constant, I set $k = 1000$ as the main analysis. In panel (a) which contains all but 5% tiniest stocks, for all columns, the mean of a long-short portfolio is over 1.27%, and the Sharpe ratios are

over 1. All models produce a significant alpha. Panel (b) shows the counterpart results for all but tiny stocks. Large stocks are shown in panel (c). The results from the whole table suggest that the model performance is not sensitive to the rolling window considered in the in-sample set.

I.4. Stochastically shrinking the In-sample Set

The in-sample set includes 120 months of data, and in each month there are several thousands of assets. Therefore, in-sample data contains several hundred thousand assets. Finding $k = 1000$ nearest neighbors through several hundred thousand is computationally time-consuming. Now from the in-sample set, I randomly select a subset of assets and use these assets as the in-sample set. By doing so, outliers are likely to be removed from the in-sample set, and the in-sample set will have a smaller number of stocks.[10] With a stochastic selection of assets in the in-sample data, it turns out that the performance of the model does not come from a specific group of stocks. This suggests that the factor structure embedded in cross-sections of returns exists even in a smaller group of randomly selected stocks. Here I show that the dispersion does not disappear when the size of in-sample data is stochastically shrunk.

In the in-sample set, I randomly drop 3 out of 4 stocks, leading to 75% reduction in the sample size. For classifying each stock in out-of-sample portfolios, among the remaining 25% stocks in the in-sample set, I find the neighboring assets and make the prediction.

Table 12 shows the out-of-sample performance of the long-short portfolios. In both value-weighted and equally-weighted case and for all three sets of datasets

[10] Hinton and Roweis (2002) and Tarlow, Swersky, Charlin, Sutskever, and Zemel (2013) propose stochastic selections for a k-nearest neighbors classifier in a machine learning setting. Kusner, Tyree, Weinberger, and Agrawal (2014) point out that using the stochastic selection in the in-sample data makes the kNN classifier substantially more robust in noisy environments.

τ	24	36	48	96	132	144	156	168
Panel (a): *All stocks*								
mean	1.27	1.38	1.38	1.38	1.62	1.55	1.61	1.61
t-stat	6.53	7.07	6.88	7.58	9.10	8.63	9.18	8.76
SR	1.01	1.09	1.06	1.17	1.40	1.33	1.42	1.35
$\alpha_{\textbf{CAPM}}$	1.35	1.44	1.45	1.44	1.67	1.61	1.65	1.68
t-stat	6.88	7.31	7.13	7.86	9.25	8.91	9.34	9.07
$\alpha_{\textbf{FF3}}$	1.40	1.49	1.49	1.47	1.69	1.64	1.68	1.71
t-stat	7.16	7.50	7.35	8.01	9.33	9.00	9.43	9.15
$\alpha_{\textbf{Carhart}}$	1.04	1.10	1.05	1.07	1.32	1.27	1.33	1.33
t-stat	5.85	6.20	6.00	6.77	8.26	7.88	8.36	8.05
$\alpha_{\textbf{FF5}}$	1.26	1.39	1.34	1.29	1.49	1.45	1.48	1.50
t-stat	6.23	6.72	6.36	6.81	8.02	7.79	8.13	7.86
α_q	1.02	1.13	1.06	1.06	1.29	1.23	1.30	1.32
t-stat	5.19	5.64	5.22	5.86	7.20	6.87	7.35	7.09
α_{q^5}	1.03	1.14	0.97	0.95	1.12	1.10	1.15	1.17
t-stat	4.87	5.34	4.50	4.92	5.87	5.76	6.12	5.91
Panel (b): *All but tiny stocks*								
mean	0.66	0.85	0.83	0.83	0.96	1.00	0.99	0.97
t-stat	3.79	5.22	5.28	5.38	6.32	6.81	6.62	6.52
SR	0.58	0.81	0.81	0.83	0.97	1.05	1.02	1.01
$\alpha_{\textbf{CAPM}}$	0.79	0.98	0.95	0.95	1.07	1.10	1.09	1.06
t-stat	4.52	6.15	6.01	6.25	7.03	7.47	7.30	7.19
$\alpha_{\textbf{FF3}}$	0.81	0.99	0.96	0.95	1.08	1.11	1.11	1.07
t-stat	4.64	6.16	6.09	6.24	7.08	7.53	7.38	7.21
$\alpha_{\textbf{Carhart}}$	0.39	0.62	0.57	0.57	0.71	0.76	0.75	0.71
t-stat	2.72	4.58	4.45	4.64	5.68	6.21	6.03	5.84
$\alpha_{\textbf{FF5}}$	0.69	0.85	0.83	0.80	0.93	1.00	0.99	0.95
t-stat	3.79	5.09	5.05	5.09	5.91	6.53	6.35	6.15
α_q	0.43	0.63	0.61	0.58	0.73	0.81	0.79	0.74
t-stat	2.49	3.97	3.88	3.86	4.85	5.47	5.30	4.98
α_{q^5}	0.36	0.60	0.49	0.39	0.58	0.70	0.65	0.55
t-stat	1.91	3.54	2.92	2.48	3.59	4.42	4.09	3.53

continued

Continued from previous page

τ	24	36	48	96	132	144	156	168
Panel (c): *Large stocks*								
mean	0.56	0.52	0.53	0.67	0.65	0.72	0.75	0.72
t-stat	3.00	2.98	3.17	4.21	4.32	4.83	4.88	4.79
SR	0.46	0.46	0.49	0.65	0.67	0.75	0.75	0.74
$\alpha_{\textbf{CAPM}}$	0.63	0.60	0.64	0.79	0.76	0.81	0.84	0.82
t-stat	3.38	3.44	3.82	5.02	5.00	5.42	5.48	5.44
$\alpha_{\textbf{FF3}}$	0.71	0.66	0.67	0.82	0.77	0.83	0.86	0.84
t-stat	3.86	3.78	4.00	5.21	5.09	5.51	5.59	5.57
$\alpha_{\textbf{Carhart}}$	0.31	0.27	0.29	0.43	0.40	0.46	0.47	0.48
t-stat	1.93	1.80	2.05	3.38	3.22	3.77	3.84	3.85
$\alpha_{\textbf{FF5}}$	0.60	0.53	0.50	0.59	0.60	0.66	0.70	0.69
t-stat	3.13	2.93	2.87	3.67	3.82	4.24	4.43	4.40
α_q	0.41	0.33	0.29	0.40	0.42	0.49	0.53	0.53
t-stat	2.17	1.89	1.76	2.64	2.79	3.28	3.45	3.50
α_{q^5}	0.34	0.19	0.18	0.23	0.25	0.34	0.38	0.39
t-stat	1.69	1.01	1.03	1.44	1.56	2.15	2.33	2.41

Table 11- Performance of a value-weighted long-short portfolio for different values of τ

This table reports the performance of a value-weighted long-short portfolio formed based on different lengths of the rolling window, τ. That is, for finding each asset neighbors, I search the neighbors through the past τ months. All 94 characteristics are taken into account while creating portfolios. I consider six factor models and report alpha and their $t-$stats based on them. These six models include CAPM, the Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French 5 factor model (FF5), q model, and augmented q model with expected growth (q^5). The regression periods capture realized returns in percentage from 1980 to 2021. Panel (a) reports the results for all but 5% tiniest stocks, panel (b) shows the counterpart results for all but tiny stocks, and panel (c) represents the results for large stocks.

the results do not disappear and all of the alphas are economically large and statistically significant.

	Value weighted			Equally weighted		
	All data	All but tiny	Large stocks	All data	All but tiny	Large stocks
mean	1.51	0.77	0.71	1.58	0.92	0.66
std	4.14	3.31	3.65	3.03	2.69	3.27
t-stat	8.19	5.23	4.34	11.71	7.68	4.54
SR	1.26	0.81	0.67	1.81	1.18	0.70
$\alpha_{\textbf{CAPM}}$	1.50	0.87	0.77	1.59	1.06	0.78
t-stat	8.02	5.93	4.70	11.64	9.17	5.40
adj R^2	-0.00	0.03	0.01	-0.00	0.10	0.05
$\alpha_{\textbf{FF3}}$	1.55	0.90	0.81	1.62	1.04	0.80
t-stat	8.32	6.06	4.97	11.74	9.01	5.52
adj R^2	0.01	0.03	0.02	-0.00	0.10	0.05
$\alpha_{\textbf{Carhart}}$	1.20	0.56	0.54	1.33	0.75	0.52
t-stat	7.13	4.48	3.54	11.03	8.09	4.01
adj R^2	0.22	0.32	0.19	0.26	0.44	0.26
$\alpha_{\textbf{FF5}}$	1.39	0.74	0.79	1.39	0.79	0.74
t-stat	7.19	4.84	4.61	10.08	6.91	4.89
adj R^2	0.03	0.06	0.02	0.07	0.20	0.05
α_q	1.19	0.57	0.61	1.20	0.63	0.54
t-stat	6.32	3.89	3.64	9.39	6.00	3.75
adj R^2	0.10	0.14	0.07	0.22	0.33	0.16
α_{q^5}	1.11	0.43	0.52	1.05	0.54	0.44
t-stat	5.53	2.73	2.91	7.71	4.83	2.87
adj R^2	0.10	0.15	0.08	0.24	0.33	0.17

Table 12- Performance of a value-weighted long-short portfolio with randomly shrinking the in-sample set

This table reports monthly alphas, $t-$values and adjusted R^2 for out-of-sample performance of value-weighted long-short portfolios where there is a stochastic selection in the training sample. Specifically, from the in-sample set I randomly eliminate 75% of the stocks, thereby using only 25% of the original dataset to find neighboring assets. The three left columns show the results for the value-weighted portfolios, while the three right columns show the equally-weighted counterparts. all stocks include the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six different factor models and report alpha and their $t-$stats based on them. These six models include CAPM, the Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French five-factor model (FF5), q model and augmented q model with expected growth (q^5). The number of neighbors is considered $k = 1000$ and the time period is 1980-2021.

	All stocks		All but tiny stocks		Large stocks	
	uniform	l-norm	uniform	l-norm	uniform	l-norm
mean	1.73	1.56	0.93	0.87	0.63	0.73
t-stat	9.08	7.84	6.38	5.41	4.28	4.30
SR	1.40	1.21	0.98	0.84	0.66	0.66
$\alpha_{\textbf{CAPM}}$	1.72	1.58	0.98	0.96	0.72	0.82
t-stat	8.89	7.83	6.63	5.93	4.85	4.81
$\alpha_{\textbf{FF3}}$	1.78	1.64	1.00	1.00	0.76	0.90
t-stat	9.21	8.15	6.74	6.19	5.15	5.35
$\alpha_{\textbf{Carhart}}$	1.37	1.22	0.66	0.58	0.41	0.48
t-stat	8.13	6.92	5.27	4.56	3.36	3.54
$\alpha_{\textbf{FF5}}$	1.56	1.43	0.86	0.80	0.63	0.77
t-stat	7.89	6.89	5.61	4.84	4.14	4.42
α_q	1.34	1.23	0.70	0.60	0.48	0.59
t-stat	7.05	6.12	4.74	3.78	3.27	3.50
α_{q^5}	1.21	1.19	0.57	0.43	0.32	0.30
t-stat	5.94	5.52	3.60	2.54	2.04	1.71

Table 13- Performance of a value-weighted long-short portfolio for different distance measures

This table reports the performance of value-weighted long-short portfolios created with uniform weighting scheme and distance weighting with l-norm distance metrics. The regression period is from Jan 1980 till Dec 2021. The uniform column shows the case when I use equal weights for neighboring assets when making the prediction. (All the portfolios are still value-weighted with market-cap). $l-$norm shows the case when I use absolute-value norm (instead of Euclidean norm) to find k closest assets. I consider six factor models and report alpha and their $t-$stats based on them. These six models include CAPM, Fama French three-factor model (FF3), the Carhart four-factor model (Carhart), the Fama French 5 factor model (FF5), the q model, and the augmented q model with expected growth (q^5). The regression periods capture realized returns in percentage from 1980 to 2021. Panel (a) reports the results for all but 5% tiniest stocks, panel (b) shows the counterpart results for all but tiny stocks, and panel (c) represents the results for large stocks.

I.5. Different Distance Measures

Now I show that the results are not sensitive to distance measurement. The distance measurement I use for the whole analysis is based on a Euclidean distance, which is l^2-norm, and then I use a distance weighting scheme. For a robustness

check, I consider a uniform distance weighting such that all neighboring assets receive equal weights. In other words, $\omega_{a_{j,t}}(a_{j',t-t'})$ in equation 4 will be the same across all assets (for $j' = 1, \ldots, k$) in the neighborhood of asset $a_{j,t}$. Next, I deviate from a l^2-norm measurement to a l-norm. I show that the patterns persist among different definitions of distance and uniformly weighting.

Table 13 shows the performance of value-weighted long-short portfolios based on both uniform weighting and l-norm distance. For all three sets of data, changing the distance measure does not change the performance of the neighboring assets portfolios.

IV. Summary and Conclusion

If two firms are alike in terms of characteristics, they display similar expected rates of returns. The similarity of firms in the characteristics is an indicator of fundamental linkages. By considering a large set of characteristics, for each firm, I recognize the firms that most resemble it. Identifying them as neighboring firms (assets), I find that neighboring assets tend to produce similar expected returns over time. This simply is implied by assuming that firm-level characteristics determine expected returns. Hence, because each asset has a fundamental connection with its neighbors in many aspects, each asset should have similar expected returns as its neighbors.

My results show that past returns of each asset's neighbors are a strong predictor of its future returns. I borrow the so-called k-nearest neighbors classifier from machine learning to find assets with the most similarity. Grouping assets to decile portfolios based on the past performance of their neighbors generates a large dispersion in the cross-sections of returns. A large set of factor zoo fail to explain

the long-short portfolio of neighbouring assets.

While the true relationship between a large set of characteristics and expected returns is unknown, through neighboring assets I map a large set of firm characteristics to their expected returns in a non-linear setting. This sheds light on the relationship between firm characteristics and expected returns. My method also alleviates the curse of dimensionality implied by the zoo of characteristics. I propose a novel way of forming test portfolios by grouping neighboring assets into decile portfolios. Finally, I develop a method to unify the joint effect of characteristics between connected firms.

References

Ahn, Dong-Hyun, Jennifer Conrad, and Robert F Dittmar, 2009, Basis assets, *The Review of Financial Studies* 22, 5133–5174.

Ali, Usman, and David Hirshleifer, 2020, Shared analyst coverage: Unifying momentum spillover effects, *Journal of Financial Economics* 136, 649–675.

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *The journal of finance* 61, 259–299.

Anton, Miguel, and Christopher Polk, 2014, Connected stocks, *The Journal of Finance* 69, 1099–1127.

Bali, Turan G, Nusret Cakici, and Robert F Whitelaw, 2011, Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of financial economics* 99, 427–446.

Barbee Jr, William C, Sandip Mukherji, and Gary A Raines, 1996, Do sales–price

and debt–equity explain stock returns better than book–market and firm size?, *Financial Analysts Journal* 52, 56–60.

Barberis, Nicholas, and Andrei Shleifer, 2003, Style investing, *Journal of financial Economics* 68, 161–199.

Bouwman, Christa HS, 2011, Corporate governance propagation through overlapping directors, *The Review of Financial Studies* 24, 2358–2394.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2022, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *The Journal of Finance* .

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, 2024, Forest through the trees: Building cross-sections of stock returns, *The Journal of Finance* .

Chen, Andrew Y, and Tom Zimmermann, 2021, Open source cross-sectional asset pricing, *Critical Finance Review, Forthcoming* .

Chen, Luyang, Markus Pelger, and Jason Zhu, 2024, Deep learning in asset pricing, *Management Science* 70, 714–750.

Chordia, Tarun, Avanidhar Subrahmanyam, and V Ravi Anshuman, 2001, Trading activity and expected stock returns, *Journal of financial Economics* 59, 3–32.

Clarke, Charles, 2022, The level, slope, and curve factor model for stocks, *Journal of Financial Economics* 143, 159–187.

Cochrane, John H, 2011, Presidential address: Discount rates, *The Journal of finance* 66, 1047–1108.

Cohen, Lauren, and Andrea Frazzini, 2008, Economic links and predictable returns, *The Journal of Finance* 63, 1977–2011.

Conrad, Jennifer, and Gautam Kaul, 1998, An anatomy of trading strategies, *The Review of Financial Studies* 11, 489–519.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *The Journal of finance* 52, 1035–1058.

Datar, Vinay T, Narayan Y Naik, and Robert Radcliffe, 1998, Liquidity and stock returns: An alternative test, *Journal of financial markets* 1, 203–219.

Fama, Eugene F, and Kenneth R French, 1995, Size and book-to-market factors in earnings and returns, *The journal of finance* 50, 131–155.

Fama, Eugene F, and Kenneth R French, 2008, Dissecting anomalies, *The Journal of Finance* 63, 1653–1678.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *The Journal of Finance* 75, 1327–1370.

Feng, Guanhao, Jingyu He, Nicholas G Polson, and Jianeng Xu, 2018, Deep learning in characteristics-sorted factor models, *Journal of Financial and Quantitative Analysis* 1–36.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting characteristics nonparametrically, *The Review of Financial Studies* 33, 2326–2377.

Gettleman, Eric, and Joseph M Marks, 2006, Acceleration strategies, *SSRN Electronic Journal* .

Goyal, Amit, and Sunil Wahal, 2015, Is momentum an echo?, *Journal of Financial and Quantitative Analysis* 50, 1237–1267.

Green, Jeremiah, John RM Hand, and Mark T Soliman, 2011, Going, going, gone? the apparent demise of the accruals anomaly, *Management Science* 57, 797–816.

Green, Jeremiah, John RM Hand, and X Frank Zhang, 2017, The characteristics that provide independent information about average us monthly stock returns, *The Review of Financial Studies* 30, 4389–4436.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2021, Autoencoder asset pricing models, *Journal of Econometrics* 222, 429–450.

Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, . . . and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.

Haugen, Robert A, and Nardin L Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of financial economics* 41, 401–439.

He, Wei, Yuehan Wang, and Jianfeng Yu, 2021, Similar stocks, *Available at SSRN 3815595* .

Hinton, Geoffrey E, and Sam Roweis, 2002, Stochastic neighbor embedding, *Advances in neural information processing systems* 15.

Hirshleifer, David, Kewei Hou, and Siew Hong Teoh, 2012, The accrual anomaly: risk or mispricing?, *Management Science* 58, 320–335.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2019, Which factors?, *Review of Finance* 23, 1–35.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2021, An augmented q-factor model with expected growth, *Review of Finance* 25, 1–41.

Hou, Kewei, Chen Xue, and Lu Zhang, 2017, A comparison of new factor models, *Fisher college of business working paper* 05.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *The Review of Financial Studies* 33, 2019–2133.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of finance* 48, 65–91.

Jegadeesh, Narasimhan, and Sheridan Titman, 2001, Profitability of momentum strategies: An evaluation of alternative explanations, *The Journal of finance* 56, 699–720.

Jensen, Theis Ingerslev, Bryan T Kelly, and Lasse Heje Pedersen, 2021, Is there a replication crisis in finance?, *The Journal of Finance, Forthcoming* .

Kaniel, Ron, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh, 2023, Machine-learning the skill of mutual fund managers, *Journal of Financial Economics* 150, 94–138.

Karolyi, G Andrew, and Stijn Van Nieuwerburgh, 2020, New methods for the cross-section of returns, *The Review of Financial Studies* 33, 1879–1890.

Kaustia, Markku, and Ville Rantala, 2015, Social learning and corporate peer effects, *Journal of Financial Economics* 117, 653–669.

Kelly, Bryan, Semyon Malamud, and Lasse Heje Pedersen, 2020, Principal portfolios, *The Journal of Finance* .

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, The virtue of complexity in return prediction, *The Journal of Finance* 79, 459–503.

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Kozak, Serhiy, and Stefan Nagel, 2023, When do cross-sectional asset pricing factors span the stochastic discount factor?, Technical report, National Bureau of Economic Research.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Kusner, Matt, Stephen Tyree, Kilian Weinberger, and Kunal Agrawal, 2014, Stochastic neighbor compression, in *International conference on machine learning*, 622–630, PMLR.

Leary, Mark T, and Michael R Roberts, 2014, Do peer firms affect corporate financial policy?, *The Journal of Finance* 69, 139–178.

Lee, Charles, Paul Ma, and Charles CY Wang, 2016, The search for peer firms: When do crowds provide wisdom?, *Harvard Business School Accounting & Management Unit Working Paper* 14–46.

Lettau, Martin, and Markus Pelger, 2020, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.

Lewellen, Jonathan, 2015, The cross section of expected stock returns, *Critical Finance Review* 4, 1–44.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial economics* 96, 175–194.

Li, Bin, and Alberto G Rossi, 2020, Selecting mutual funds from the stocks they hold: A machine learning approach, *Available at SSRN 3737667* .

Light, Nathaniel, Denys Maslov, and Oleg Rytchkov, 2017, Aggregation of information about the cross section of stock returns: A latent variable approach, *The Review of Financial Studies* 30, 1339–1381.

Martin, Ian WR, and Stefan Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.

Menzly, Lior, and Oguzhan Ozbas, 2010, Market segmentation and cross-predictability of returns, *The Journal of Finance* 65, 1555–1580.

Moritz, Benjamin, and Tom Zimmermann, 2016, Tree-based conditional portfolio sorts: The relation between past and future stock returns, *Available at SSRN 2740751* .

Moskowitz, Tobias J, and Mark Grinblatt, 1999, Do industries explain momentum?, *The Journal of finance* 54, 1249–1290.

Müller, Sebastian, 2019, Economic links and cross-predictability of stock returns: Evidence from characteristic-based "styles", *Review of Finance* 23, 363–395.

Nagel, Stefan, and Kenneth J Singleton, 2011, Estimation and evaluation of conditional asset pricing models, *The Journal of Finance* 66, 873–909.

Seyfi, Sina, 2023, Essence of the cross section, *Available at SSRN 4466972* .

Seyfi, Sina, 2024, Basis portfolios, *Available at SSRN 4687058* .

Shue, Kelly, 2013, Executive networks and firm policies: Evidence from the random assignment of mba peers, *The Review of Financial Studies* 26, 1401–1442.

Tarlow, Daniel, Kevin Swersky, Laurent Charlin, Ilya Sutskever, and Rich Zemel, 2013, Stochastic k-neighborhood selection for supervised and unsupervised learning, in *International Conference on Machine Learning*, 199–207, PMLR.

Appendices

Acronym	Firm characteristic	Acronym	Firm characteristic
absacc	Absolute accruals	mom36m	36-month momentum
acc	Working capital accruals	mom6m	6-month momentum
aeavol	Abnormal earnings announcement	ms	Financial statement score
age	# years since first Compustat coverage	mvel1	Size
agr	Asset growth	mve	ia Industry-adjusted size
baspread	Bid-ask spread	nincr	Number of earnings increases
beta	Beta	operprof	Operating profitability
betasq	Beta squared	orgcap	Organizational capital
bm	Book-to-market	pchcapx_ia	ia Industry adjusted % change in capital expenditures
bm _ia	ia Industry-adjusted book to market	pchcurrat	% change in current ratio
cash	Cash holdings	pchdepr	% change in depreciation
cashdebt	Cash flow to debt	pchgm_pchsale	% change in gross margin - % change in sales
cashpr	Cash productivity	pchquick	% change in quick ratio
cfp	Cash flow to price ratio	pchsale_pchinvt	% change in sales - % change in inventory
cfp _ia	Industry-adjusted cash flow to price ratio Asness	pchsale_pchrect	% change in sales - % change in A/R
chatoia	Industry-adjusted change in asset turnover	pchsale_pchxsga	% change in sales - % change in SG&A
chcsho	Change in shares outstanding	pchsaleinv	% change sales-to-inventory
chempia	Industry-adjusted change in employees	pctacc	Percent accruals
chinv	Change in inventory	pricedelay	Price delay
chmom	Change in 6-month momentum	ps	Financial statements score
chpmia	Industry-adjusted change in profit margin	quick	Quick ratio
chtx	Change in tax expense	rd	R&D increase
cinvest	Corporate investment	rd _mve	R&D to market capitalization
convind	Convertible debt indicator	rd_sale	R&D to sales
currat	Current ratio	realestate	Real estate holdings
depr	Depreciation / PP&E	retvol	Return volatility
divi	Dividend initiation	roaq	Return on assets
divo	Dividend omission	roavol	Earnings volatility
dolvol	Dollar trading volume	roeq	Return on equity
dy	Dividend to price	roic	Return on invested capital
ear	Earnings announcement return	rsup	Revenue surprise
egr	Growth in common shareholder equity	salecash	Sales to cash
ep	Earnings to price	saleinv	Sales to inventory
gma	Gross profitability	salerec	Sales to receivables
grCAPX	Growth in capital expenditures	secured	Secured debt
grltnoa	Growth in long term net operating assets	securedind	Secured debt indicator
herf	Industry sales concentration	sgr	Sales growth
hire	Employee growth rate	sin	Sin stocks
idiovol	Idiosyncratic return volatility	sp	Sales to price
ill	Illiquidity	std_dolvol	Volatility of liquidity (dollar trading volume)
indmom	Industry momentum	std_turn	Volatility of liquidity (share turnover)
invest	Capital expenditures and inventory	stdacc	Accrual volatility
lev	Leverage	stdcf	Cash flow volatility
lgr	Growth in long-term debt	tang	Debt capacity/firm tangibility
maxret	Maximum daily return	tb	Tax income to book income
mom12m	12-month momentum	turn	Share turnover
mom1m	1-month momentum	zerotrade	Zero trading days

Table A1- List of acronyms and characteristics provided by Gu et al. (2020)

Portfolio	All stocks	All but tiny	Large stocks
1	823	416	163
2	335	80	43
3	243	106	67
4	410	176	110
5	385	308	167
6	862	398	196
7	557	255	163
8	376	159	79
9	456	165	81
10	354	325	176

Table A2- Average number of assets in each portfolio

This table shows the average number of stocks in each of the decile portfolios based on the neighboring assets strategy. I put each asset into one of the portfolios based on the performance of the majority of its neighbors in the past ten years. I recognize each asset's neighbors based on the 94 characteristics listed in table A1. The first column considers all stocks except the 5% tiniest for creating portfolios, while the second and third columns include all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median), respectively. The number of neighbors is considered $k = 1000$.

Anomaly
1- cumulative abnormal returns around earnings announcement dates, 1-month holding period
2- cumulative abnormal returns around earnings announcement dates, 6-month holding period
3- cumulative abnormal returns around earnings announcement dates, 12-month holding period
4- customer industries momentum, 1-month holding period
5- customer industries momentum, 6-month holding period
6- customer industries momentum, 12-month holding period
7- customer momentum, 1-month holding period
8- customer momentum, 12-month holding period
9- changes in analyst earnings forecasts, 1-month holding period
10- changes in analyst earnings forecasts, 6-month holding period
11- changes in analyst earnings forecasts, 12-month holding period
12- industry lead-lag effect in earnings surprises, 1-month holding period
13- industry lead-lag effect in prior returns, 1-month holding period
14- industry lead-lag effect in prior returns, 6-month holding period
15- industry lead-lag effect in prior returns, 12-month holding period
16- industry momentum, 1-month holding period
17- industry momentum, 6-month holding period
18- industry momentum, 12-month holding period
19- the number of quarters with consecutive earnings increase, 1-month holding period
20- 52-week high, 6-month holding period
21- 52-week high, 12-month holding period
22- prior 6-month returns, 1-month holding period
23- prior 6-month returns, 6-month holding period
24- prior 6-month returns, 12-month holding period
25- prior 11-month returns, 1-month holding period
26- prior 11-month returns, 6-month holding period
27- prior 11-month returns, 12-month holding period
28- revisions in analyst earnings forecasts, 1-month holding period
29- revisions in analyst earnings forecasts, 6-month holding period
30- 6-month residual momentum, 6-month holding period
31- 6-month residual momentum, 12-month holding period
32- 11-month residual momentum, 1-month holding period
33- 11-month residual momentum, 6-month holding period
34- 11-month residual momentum, 12-month holding period
35- revenue surprises, 1-month holding period
36- supplier industries momentum, 1-month holding period
37- supplier industries momentum, 12-month holding period
38- segment momentum, 1-month holding period
39- segment momentum, 12-month holding period
40- standard unexpected earnings, 1-month holding period
41- standard unexpected earnings, 6-month holding period

Table A3- List of momentum anomalies provided by Hou et al. (2020)

Internet Appendix

Appendix A. Bayesian Perspective of Neighboring Assets Algorithm

In this internet appendix, I estimate the posterior probability that asset $a_{j,t}$ belongs to portfolio d_i for $i = 1, \ldots, 10$. For the sake of simplicity, I first consider the case that each neighboring asset in the in-sample data has equal weight, that is, uniform weighting. For each asset $a_{j,t}$ with characteristics $\mathbf{x}_{j,t-1}$, consider a region $\mathcal{R}$ in characteristics hyperspace with the center $\mathbf{x}_{j,t-1}$ such that $\mathcal{R}$ captures a set of k assets from set S_t, namely $a_{s',t-t'-1}$ for $t' = 1, \ldots \tau$, with characteristics $\mathbf{x}_{s',t-t'-1}$, and $s' = 1, \ldots, k$ (without loss of generality). It is clear that

$$l(\mathbf{x}_{s',t-t'-1}, \mathbf{x}_{j,t-1}) \leq l(\mathbf{x}_{s,t-t'-1}, \mathbf{x}_{j,t-1})$$

for all s which is not among the set $s' \in \{1, \ldots, k\}$, and for some distance measure l. I denote the volume of $\mathcal{R}$ in characteristic hyperspace as V and suppose that it contains n_i assets which belong to decile portfolio d_i. Moreover, I denote the whole number of assets belonging to portfolio d_i in the entire set S_t by N_i, so that

$$\sum_i N_i = |S_t|.$$

Then the conditional decile portfolio estimation for asset $a_{j,t}$ with set of characteristics $\mathbf{x}_{j,t-1}$ is:

$$p(a_{j,t}|d_i; \mathbf{x}_{j,t-1}) = \frac{n_i}{N_i V}. \tag{A1}$$

Likewise, the unconditional probability would be

$$p(a_{j,t}|\mathbf{x}_{j,t-1}) = \frac{\sum_i n_i}{|S_t|V} = \frac{k}{|S_t|V}. \tag{A2}$$

And the prior labels of portfolios are

$$p(d_i) = \frac{n_i}{|S_t|}. \tag{A3}$$

Now using Bayes' theorem we observe that the posterior probability that asset $a_{j,t}$ belongs to portfolio i at time t is given by:

$$p(d_i|a_{j,t}; \mathbf{x}_{j,t-1}) = \frac{p(a_{j,t}|d_i; \mathbf{x}_{j,t-1})p(d_i)}{p(a_{j,t}|\mathbf{x}_{j,t-1})} = \frac{n_i}{k}. \tag{A4}$$

The interpretation is straightforward: the probability that an asset belongs to decile d_i is proportional to the number of its neighbors belonging to decile d_i in the past.

Next, I weigh all n_i points in the $\mathcal{R}$ by the inverse of their distance from asset $a_{j,t}$, that is,

$$\omega_{a_{j,t}}(a_{s',t-t'}) = \frac{1}{l(\mathbf{x}_{s',t-t'-1}, \mathbf{x}_{j,t-1})}$$

for all assets s' in the region $\mathcal{R}$ centered by lagged-characteristics of asset $a_{j,t}$. In this case, the posterior probability of asset classification is

$$p(d_i|a_{j,t}; \mathbf{x}_{j,t-1}) = \frac{\sum_{s'=1}^{k} \omega_{a_{j,t}}(a_{s',t-t'})\delta(d, c_{s',t-t'})}{\sum_{s'=1}^{k} \omega_{a_{j,t}}(a_{s',t-t'})}, \tag{A5}$$

where $\delta(.,.)$ is defined according to equation 6. Equation A5 means that the posterior probability of asset $a_{j,t}$ belongs to portfolio d_i given a set of characteristics is proportional to the sum of the weights of its neighbors which belonged to portfolio d_i. Therefore, in order to maximize the probability of true classifications, one should put $a_{j,t}$ to portfolio d_i that has the highest summation of weights,

$\sum_{s'=1}^{k} \omega(a_{s',t-t'}|a_{j,t})\delta(d_i, c_{s',t-t'})$:

$$c_{j,t} = \underset{d \in \{d_1,\ldots,d_{10}\}}{\arg\max} \quad p(d|a_{j,t}; \mathbf{x}_{j,t-1}). \tag{A6}$$

In a case where all weights are equal, equation A5 becomes as equation A4. Intuitively, it is most likely that asset $a_{j,t}$ belongs to portfolio d_i if the majority of its *closest* neighbors have belonged to this portfolio in the past.