Machine-Learning the Skill of Mutual Fund Managers *

Ron Kaniel† Zihan Lin‡ Markus Pelger§ Stijn Van Nieuwerburgh¶

August 7, 2023

Abstract

We show, using machine learning, that fund characteristics can consistently differentiate high from low-performing mutual funds, before and after fees. The outperformance persists for more than three years. Fund momentum and fund flow are the most important predictors of future risk-adjusted fund performance, while characteristics of the stocks that funds hold are not predictive. Returns of predictive long-short portfolios are higher following a period of high sentiment. Our estimation with neural networks enables us to uncover novel and substantial interaction effects between sentiment and both fund flow and fund momentum.

JEL: G11, G12, G17, G23, C45

Key words: Mutual fund performance, fund flow, momentum, machine learning, sentiment, big data, neural networks

*First draft: January 14, 2021.

We thank Will Cong, Kay Giesecke, Stefano Giglio, Markus Ibert (discussant), Allan Timmermann (discussant), Alberto Rossi (discussant), and seminar and conference participants at Imperial College London, Oxford, Stanford, the Annual Society for Financial Econometrics Conference, the BI-SHoF Conference, the Conference on Emerging Technologies in Accounting and Financial Economies, and the NBER Summer Institute in Big Data for helpful comments.

†Department of Finance, Simon School of Business, Rochester University, ron.kaniel@simon.rochester.edu.

‡Institute for Computational and Mathematical Engineering, Stanford University, zihanl@stanford.edu.

§Department of Management Science and Engineering, Stanford University, mpelger@stanford.edu.

¶Department of Finance, Columbia Business School, svnieuwe@gsb.columbia.edu.

1 Introduction

The asset management industry is enormous and growing rapidly. U.S. mutual funds had $24 trillion in assets under management at the end of 2020, more than half of which were in equity mutual funds. Over 100 million Americans rely on such funds to save for retirement and meet other financial objectives. Many of these mutual funds actively trade stocks in an effort to out-perform their benchmarks and create value for their investors. The literature has found mixed results in terms of the investment performance of actively-traded equity mutual funds. We revisit the evidence using modern techniques, and ask which–if any–characteristics of mutual funds and of the stocks they hold can help separate the corn from the chaff. We uncover new evidence of economically substantial and long-lasting abnormal return predictability. Fund flows and fund return momentum are the main two characteristics that can meaningfully and robustly help distinguish funds that outperform from those that under-perform, particularly in times of high investor sentiment.

We study the universe of actively-traded U.S. equity mutual funds between 1980 and 2019 and the stocks that they hold. The object we predict is the abnormal fund return, defined as the locally-estimated four-factor alpha of the mutual fund. The predictors are a long list of 46 stock characteristics weighted by the funds' holdings of each stock and 13 fund and fund-family characteristics. We also include a variable that summarizes the overall state of the market, either proxied by investor sentiment or by a comprehensive measure of macro-economic activity. Our main method is a feedforward neural network, which can reliably estimate a complex functional relationship among a large set of predictors. It is trained and tuned on one subset of the data and evaluated on another subset of the data. Hence, all of our predictions are out-of-sample. Our method identifies fund characteristic information, and specifically fund flow and fund momentum, as the key predictors of mutual fund out-performance. Moreover, there is an important interaction effect between these two fund characteristics and sentiment, which linear models fail to pick up.

The model generates large differences in out-of-sample performance. Buying the ten percent of mutual funds with the best predicted performance each month, and using the model not only to select but also to weight the funds within the top decile, generates a cumulative abnormal return of 72%. Buying the ten percent of mutual funds with the worst predicted performance each month produces a cumulative abnormal return of -119%. The 191% difference in out-of-sample performance based on the model's predictions is economically large and statistically significant. It translates into a monthly out-performance of 15 basis points for the 10% best funds and 25 basis points per month under-performance for the 10% worst funds. Since the best and the worst funds have similar fees, the same result holds for after-fee abnormal returns. The performance improves further when we directly predict net-of-fees performance.

The performance differential is nearly identical if we constrain the model by removing all stock characteristic information. In fact, we can also remove most fund and fund family characteristics.

The predictions of a model that are only given data on fund flow, fund momentum, and sentiment are nearly as good as those of the full model. They deliver out-performance of 48 basis points per month for the top relative to the bottom deciles of predicted performers. The Sharpe ratio on this strategy is 0.24 per month.

The predictability we uncover is surprisingly persistent. Even though investments are made based on one-month ahead prediction, the best decile of funds significantly outperforms the worst decile for three years. Even after 36 months, the monthly Sharpe ratio on the long-short portfolio is still 0.20, compared to a 0.30 Sharpe ratio 3-months ahead. This result is remarkable in light of the literature's difficulty in finding evidence for persistence in abnormal fund returns.

We decompose the fund abnormal return into a between-disclosure component, which holds fixed the funds' stock holdings at their previous quarter-end values, and a within-disclosure component, which accounts for mutual fund trades during the quarter. The latter is the sum of the return gap and a risk exposure differential.[1] We find that about half of the outperformance comes from the model's ability to predict between-disclosure abnormal returns and the other half from predicting within-disclosure abnormal returns. Both fund flow and fund momentum predict the return gap and the risk exposure differential, while most stock characteristics that predict the return gap do so by taking on more systematic risk resulting in little within-disclosure abnormal return. These results shed additional light on the sources and persistence of out-performance.

The salience of flow and fund return momentum as the key predictors suggests that some investors can detect skill and (re)allocate their investment towards such skilled managers. This reallocation of investment flows is not as strong as the frictionless model of Berk and Green (2004) predicts. Skill leaves a trail in the form of fund return momentum for investors to exploit in the next period. Put differently, the flows are gradual and small enough that it takes several periods until the fund runs into zero marginal abnormal returns.

The results are potentially also consistent with funds and fund families attracting flows through marketing rather than—or in addition to—through investment skill (Gallaher, Kaniel, and Starks, 2009; Ibert, Kaniel, Van Nieuwerburgh, and Vestman, 2018; Roussanov, Ruan, and Wei, 2021). Marketing-induced inflows create buying pressure for stocks that the fund typically invests in. In a world with downward-sloping demand curves (Coval and Stafford, 2007; Koijen and Yogo, 2019; Gabaix and Koijen, 2021), this raises prices and lifts fund returns. Through the flow-performance relationship, as well as through persistence in marketing-driven flows, the out-performance creates more inflows in the next period. The demand pressure increases prices further, generating momentum in fund returns. The fact that flows and fund momentum have a much stronger association with fund performance in high-sentiment periods lends further credence to this marketing-driven channel.

[1] The return gap is the difference between the fund's actual returns over the period and the hypothetical returns generated by keeping the fund's portfolio holdings constant.

Our paper makes several methodological contributions adding to the protocol of how to use machine learning models for asset pricing. First, we contribute to relative performance prediction. We show that abnormal returns, obtained as local residuals to a factor model, are not only an economically motivated, but also the statistically better target for prediction. In contrast, the level of fund (and stock) returns is extremely hard to predict. Abnormal returns remove the level effect of market and other risk factors, which makes the prediction of abnormal returns a relative objective. The commonly-used machine learning prediction of total returns can be dominated by the prediction error in the common component in return levels, resulting in suboptimal use of cross-sectional information relevant to relative performance. Indeed, we show that using the same flexible methods for predicting abnormal returns instead of total returns results in higher accuracy and better portfolio performance.

Second, we quantify the economic benefit of different information sets. We suggest to compare the prediction and trading benefits by varying the information set available to the same flexible machine-learning algorithm. The focus is the comparison of information sets instead of a horse race of model specifications.

Third, we show how to measure the dependencies on macroeconomic states. Specifically, we propose a cross-out-of-sample evaluation of conditional models using the full time-series. Importantly, the data points for estimating and evaluating the model have to be sampled such that all relevant economic conditions are represented in all subsamples, which can be achieved by random sampling over time. This is particularly important for measuring the dependencies on macroeconomic states which are only available in a subset of the data and might be neglected in the estimation or evaluation with a conventional chronological data split. Our evaluation approach allows us to take advantage of all sample periods for the out-of-sample analysis, diminishing the effect of particular subperiods. That said, our main results are robust to using a chronological cross-validation as well as an expanding-window sampling approach.

Fourth, in order to better assess the investment benefits of prediction, we suggest prediction-weighted portfolios. These portfolios result in the largest return spread as they take advantage not only of the ranking but also of the relative strength of the prediction signal. The prediction-weighted portfolios dominate the widely-used equally-weighted portfolios based on prediction quantiles.

Last but not least, we propose a new measure for interaction effects in machine learning algorithms, which does not only measure a local slope, but a more informative global slope. For this interpretable measure, we provide a formal statistical significance test based on functional central limit theorems for neural networks.

Related Literature An enormous literature in empirical asset pricing studies whether mutual fund managers outperform their benchmarks through stock picking and market timing. The sem-

inal paper of Berk and Green (2004) suggests that a large fraction of fund managers out-performs before fees while Fama and French (2010) find no out-performance before fees. Kacperczyk, Van Nieuwerburgh, and Veldkamp (2014; 2016) find that a modest fraction of managers displays enough skill to persistently outperform, through a strategy that switches between market timing in recessions and stock picking in expansions. The presence of uninformed mutual fund managers and retail traders makes this possible as an equilibrium phenomenon (Stambaugh, 2014).

While investors direct flows to funds that out-perform, at least as measured by the CAPM alpha (Berk and Van Binsbergen, 2016; Barber, Huang, and Odean, 2016), there is mounting evidence that other factors besides fees and before-fee performance determine fund flows. Gallaher, Kaniel, and Starks (2009) shows advertising impacts flows at the industry, family and fund level. Roussanov, Ruan, and Wei (2021) argues that marketing is an important determinant of flows, necessary to understand the empirical joint distribution of fund size and performance. Consistent with this, Ibert, Kaniel, Van Nieuwerburgh, and Vestman (2018) shows that fund manager compensation is tied to the component of assets-under-management that is orthogonal to current and past fund performance.

The predictive role of flows to fund performance was first uncovered by Gruber (1996) and Zheng (1999), who identified a positive, but fairly short-lived and weak relationship. The "smart money" relation they found exists for small but not for large funds. Risk adjustment in these papers did not include momentum. Sapp and Tiwari (2004) shows the smart money effect disappears once a stock return momentum factor is considered (Carhart, 1997). Prior work identified different directional effects for different components of flow and fund returns. Lou (2012) shows that the expected part of flow-induced trading positively forecasts mutual fund returns in the following year. Song (2020) finds that fund flows associated with positive factor related returns lead to negative future fund performance. Our machine learning approach revives the predictive role of flows, with a 4-factor risk-adjustment, and shows that fund flow predicts performance positively.

Carhart (1997) finds that persistence in fund net performance essentially disappears once a stock momentum factor is added, apart for the worst performing funds where it arises from persistently high expenses. With the aid of machine learning, we identify an important predictive role for fund past performance both gross and net of fees, even after controlling for stock momentum. Furthermore, the predictive power of our method is long lived. Bollen and Jeffrey (2005) argue that part of the reason for the lack of performance persistence in Carhart (1997) is that he forms decile portfolios and considers the time series of performance of these decile portfolios, instead of computing an abnormal return at the stock level and averaging that across stocks in each subsequent period. Our predictive results, which find an important role for fund past performance, hold for long-short portfolios as well, highlighting that including fund past performance as part of a neural network prediction model is important.

Our paper more broadly relates to the fund return predictability literature. Cremers and

Petajisto (2009)'s Active Share—funds with holdings that differ greatly from their benchmarks—predicts benchmark index-adjusted Carhart alphas. Kacperczyk, Sialm, and Zheng (2008)'s Return Gap predicts 4-factor monthly alphas. The monthly abnormal return we identify is about twice as large as theirs. While they finds significance for the short but not the long leg, we find significance for both.[2] More importantly, we show that the predictive power of fund momentum and fund flows are substantially amplified when investor sentiment is high at the time of forming portfolios.

There is fairly little evidence on the impact of macro-economic conditions on performance. Moskowitz (2000) and Kosowski (2011) find that risk-adjusted performance of mutual funds is better in recessions than in booms. Massa and Yadav (2015) show that a fund's level of exposure to high-sentiment beta stocks predicts lower future returns. Sentiment affects the out-performance of fund managers differently than long-short anomaly strategies. Similar to the findings in Stambaugh, Yu, and Yuan (2012) for stock return anomalies, we find that high sentiment periods coincide with more fund return predictability. While the effect for equity anomalies comes primarily from its short leg, the out- respectively under-performance of the best and worst fund managers in high sentiment periods is symmetric, suggesting a different economic channel. In contrast to the novel interaction effects between sentiment and fund characteristics, we find no equivalent interaction effects with the state of the macro-economy as proxied by CFNAI.

Our work connects to the growing Machine Learning (ML) literature in finance (see Karolyi and Van Nieuwerburgh, 2020, for a summary). This literature has focused on analyzing the cross-section of stock returns using a plethora of return predictors.[3] Similar techniques are beginning to be used in other asset classes.[4] Independent work to ours by Li and Rossi (2021) and DeMiguel, Gil-Bazo, Nogales, and Santos (2023) study mutual fund performance with ML techniques, providing a comparison study of predicting fund returns or abnormal returns, respectively, with machine learning methods similar to Gu, Kelly, and Xiu (2020). In addition to our methodological innovations, we use a richer information set, which allows us to disentangle the relative value of holdings-based, fund-specific, and macroeconomic information. In addition, Li and Rossi (2021) predict fund total returns using holding-based stock characteristics. We confirm and refine their analysis by showing that holdings-based stock characteristics only predict the systematic component of fund returns. Our main object of interest is the fund abnormal return, which is orthogonal

[2]Some other predictive variables identified in the literature include, for example, Industry Concentration of holdings (Kacperczyk, Sialm, and Zheng (2005)) and fund R^2 (Amihud and Ruslan (2013)).

[3]Recent contributions include among others: return prediction with flexible and regularized models in Freyberger, Neuhierl, and Weber (2020) and Gu, Kelly, and Xiu (2020), robust stochastic discount factor construction with many characteristics in Kozak, Nagel, and Santosh (2020), Chen, Pelger, and Zhu (2023), Bryzgalova, Pelger, and Zhu (2021) and Cong, Feng, He, and He (2022) and estimation and evaluation of risk factors in Lettau and Pelger (2020), Kelly, Pruitt, and Su (2019), and Feng, Giglio, and Xiu (2020).

[4]Bianchi, Büchner, and Tamoni (2021); Bianchi, Büchner, Hoogteijling, and Tamoni (2021) study bonds, Filippou, Rapach, Taylor, and Zhou (2022) currencies, and Wu, Chen, Yang, and Tindall (2021) hedge fund strategies

to the systematic component of fund returns.[5] These abnormal returns are only predicted by fund-specific characteristics and sentiment, and not by stock-specific characteristics. In contemporaneous work, DeMiguel, Gil-Bazo, Nogales, and Santos (2023) also predicts abnormal returns, but uses only fund-specific characteristics without macroeconomic information. In order to capture how a model can change depending on macroeconomic conditions, it is necessary to also include the macroeconomic variables as predictors.[6] Furthermore, neither paper includes price trends of fund returns as predictors, which we find to be the most relevant. In summary, our results emphasize the role of fund-specific characteristics and the interaction with the state of the economy, thereby making progress on understanding the economic mechanism.

The rest of the paper is organized as follows. Section 2 describes our data. Section 3 describes our neural network model and our main results. Section 4 analyzes the main results in depth. Section 5 concludes. The appendix provides additional empirical results (A), implementation details (B), and statistical significance tests (C). An Internet Appendix contains auxiliary results.

2 Data

2.1 Mutual Funds

As is customary, we focus on actively-managed mutual funds holding mostly domestic equities. The mutual fund returns, expenses, total net assets (TNA), investment objectives and other fund characteristics are from the Center for Research in Security Prices (CRSP) Survivor Bias-Free Mutual Fund Database. Our analysis requires fund holdings, which we obtain by linking the database to the Thomson Financial Mutual Fund Holdings. Our cleaned data set includes 407,158 (mutual fund by month) observations for 3,275 mutual funds spanning the period from January 1980 until January 2019. We restrict our study to mutual funds with raw returns observed at time t and holdings data and total net assets observed at $t-1$, which guarantees that holding-based abnormal return returns at time t, as defined below, are observed. At each time t, mutual funds are also required to have at least 30 non-missing return observations in the last 36 months, which guarantees that the regression-based abnormal returns are well-defined. Internet Appendix IA.1 contains more details and summary statistics.

[5] A two-step procedure that predicts total, rather than abnormal, fund returns in a first step and then estimates a four-factor model on ex-post prediction-based return portfolios to form abnormal returns in the second step is fundamentally different from our procedure which directly predicts abnormal returns.

[6] An ex-post regression on sentiment or sentiment indicators is not sufficient to detect the interaction effects with sentiment our ML method uncovers. We show that the overall model changes depending on interactions of fund-level variables with sentiment and not just an additive sentiment component.

2.2 Abnormal Fund Returns

Our main object of interest is the abnormal mutual fund return. It measures fund performance after subtracting compensation for systematic risk factor exposure. We construct the abnormal return for each fund-month observation relative to the Carhart (1997) model, following a similar procedure. First, factor loadings are estimated over the prior 36 months:

$$R_{i,t-36:t-1} = \alpha_i + F_{t-36:t-1}\hat{\beta}_{i,t-1} + \eta_{i,-36:t-1}, \tag{1}$$

where $R_{i,t}$ is the gross (before-fee) return of fund i in month t in excess of a one-month T-bill yield. The rolling window regressions allow for time-varying factor exposures. Second, abnormal returns ($R^{abn}_{i,t}$) are computed:

$$R^{abn}_{i,t} = R_{i,t} - F_t\hat{\beta}_{i,t-1}. \tag{2}$$

Abnormal returns are not guaranteed have a mean of zero. Their mean and median is -0.03% per month in our sample with a standard deviation of 2.00%. Hence, mutual funds earn returns commensurate with the predictions of the Carhart model on average, but with substantial cross-sectional dispersion. While there is some controversy over which return model actual mutual fund investors use (Berk and Van Binsbergen, 2015; Barber, Huang, and Odean, 2016; Jegadeesh and Mangipudi, 2021), the Carhart model arguably remains the main factor model in the mutual fund literature and hence a natural benchmark for our purposes. The main results are robust to using abnormal returns with respect to an eight-factor model.[7]

We will use machine-learning techniques to connect abnormal fund returns to the characteristics of mutual funds, including the characteristics of the stocks they hold, and to variables that capture the state of the economy.

2.3 Holdings-based Characteristics

Mutual funds hold stocks. The stock characteristics are from Chen, Pelger, and Zhu (2023) and cover 46 characteristics that have been shown to have predictive power for the cross-section of expected returns. They are listed in Table 1 in six subgroups.

There are 332,294 fund-by-time observations with fully observed fund characteristics. We impute the missing fund characteristics with a latent factor model in the characteristics space as described in IA.1. Hence, we have a complete set of fund characteristics for all 407,158 fund-by-time observations. Our results are robust to the data imputation and are essentially identical on the subset of funds with fully observed data.

[7] That model includes market, size, value, momentum, investment, profitability, short-term reversal, and long-term reversal factors. The results are available upon request.

All stock characteristics are cross-sectionally normalized to range from -0.5 to 0.5 based on stocks' rankings on that characteristic. We normalize the sign of the characteristic ranking of stocks such that the corresponding long-short factor has a positive risk premium. For example for size (LME), the largest stocks have negative rankings while small stocks have positive rankings. The stock-specific characteristics of each fund are weighted by the fund's holdings.

Table 1: Fund-specific and stock-specific characteristics by category

	Past Returns	
(1)	r2_1	Short-term momentum
(2)	r12_2	Momentum
(3)	r12_7	Intermediate momentum
(4)	r36_13	Long-term momentum
(5)	ST_Rev	Short-term reversal
(6)	LT_Rev	Long-term reversal
	Investment	
(7)	Investment	Investment
(8)	NOA	Net operating assets
(9)	DPI2A	Change in property, plants, and equipment
(10)	NI	Net Share Issues
	Profitability	
(11)	PROF	Profitability
(12)	ATO	Net sales over lagged net operating assets
(13)	CTO	Capital turnover
(14)	FC2Y	Fixed costs to sales
(15)	OP	Operating profitability
(16)	PM	Profit margin
(17)	RNA	Return on net operating assets
(18)	ROA	Return on assets
(19)	ROE	Return on equity
(20)	SGA2S	Selling, general and administrative expenses to sales
(21)	D2A	Capital intensity
	Intangibles	
(22)	AC	Accrual
(23)	OA	Operating accruals
(24)	OL	Operating leverage
(25)	PCM	Price to cost margin
	Value	
(26)	A2ME	Assets to market cap
(27)	BEME	Book to Market Ratio
(28)	C	Ratio of cash and short-term
(29)	CF	Free Cash Flow to Book Value
(30)	CF2P	Cashflow to price
(31)	D2P	Dividend Yield
(32)	E2P	Earnings to price
(33)	Q	Tobin's Q
(34)	S2P	Sales to price
(35)	Lev	Leverage
	Trading Frictions	
(36)	AT	Total Assets
(37)	Beta	CAPM Beta
(38)	IdioVol	Idiosyncratic volatility
(39)	LME	Size
(40)	LTurnover	Turnover
(41)	MktBeta	Market Beta
(42)	Rel2High	Closeness to past year high
(43)	Resid_Var	Residual Variance
(44)	Spread	Bid-ask spread
(45)	SUV	Standard unexplained volume
(46)	Variance	Variance
	Fund Momentum	
(47)	F_ST_Rev	Fund short-term reversal
(48)	F_r2_1	Fund short-term momentum
(49)	F_r12_2	Fund momentum
	Fund Characteristics	
(50)	age	Fund age
(51)	tna	Fund tna
(52)	flow	Fund flow
(53)	exp_ratio	fund expense ratio
(54)	turnover ratio	turnover ratio
	Fund Family Characteristics	
(55)	family_tna	family tna
(56)	fund_no	number of funds in family
(57)	Family_r12_2	family momentum
(58)	Family_age	Family age
(59)	Family_flow	Family flow

This table shows all 59 characteristics sorted into nine categories. The first six categories represent stock-specific characteristics and the last three characteristic groups are fund-specific characteristics.

Table 2: Fund momentum characteristics

Acronym	Name	Definition	Reference
F_r2_1	Short-term momentum	Lagged one-month abnormal return	Jegadeesh and Titman (1993)
F_r12_2	Momentum	Mean abnormal return from past 12 months before the abnormal return prediction to two months before. Need at least 8 non-missing samples to be included.	Fama and French (1996)
F_ST_Rev	Short-term reversal	Prior month abnormal return	Jegadeesh and Titman (1993)

This table summarizes the fund momentum characteristics. We use a 'F' prefix to denote that these characteristics are based on mutual funds. It includes their acronym, name, definition and reference of their stock based counterpart. The fund momentum characteristics follow the same definition as their stock counterpart in Chen, Pelger, and Zhu (2023).

2.4 Fund and Family Characteristics

In addition to the 46 stock characteristics, we also form 13 fund characteristics sorted in the last three subgroups shown in Table 1: fund momentum, fund characteristics, and fund family characteristics. The three fund momentum characteristics are computed from fund abnormal returns as defined in Table 2. Fund momentum is different from holdings-based stock momentum. First, portfolio holdings information is only available quarterly, while funds also trade within the quarters. Hence, holdings-weighted averages of stock momentum, which use quarterly-updated weights for monthly-updated stock quantiles, can only provide an approximation. Second, fund momentum is based on the time series of residuals after removing the correlation with a stock market-based momentum factor.

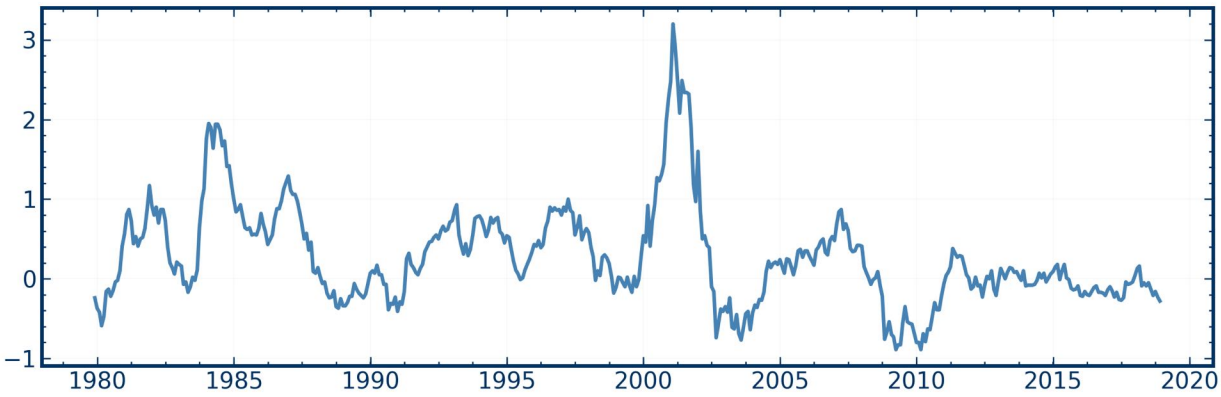
Following Brown and Wu (2016), fund family is identified by the management company code. The variables "Family_r12_2" and "Family flow" are the average of the fund-level counterparts, "F_r12_2" and "flow," weighted by TNA of all funds in the family, excluding the fund itself. "Family age" is the age of the oldest fund in the family, excluding the fund itself. "Fund no" is the number of funds in the family and "Family tna" is the sum of TNAs of all funds in the family excluding the fund itself. The fund and family characteristics are similarly normalized.

On average, mutual funds in our sample are 13.7 years old, manage $1,153 million dollars in assets, and charge a monthly expense ratio of around 0.1%. The fund's flow is defined as $flow_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1}(1+R_{i,t})}{TNA_{i,t-1}}$. Throughout the sample period the mutual fund industry is growing; on average funds enjoy a 1.6% monthly inflow.

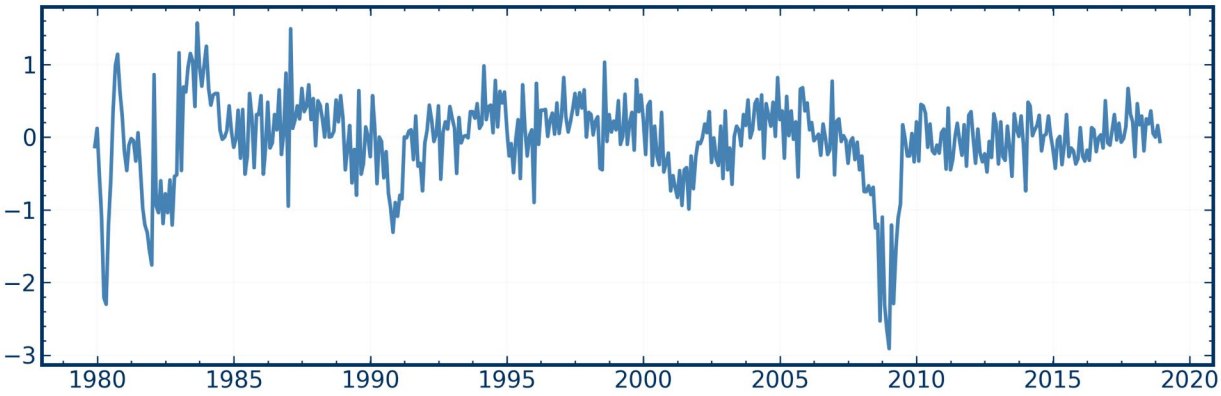
2.5 Macro-economic Information

To study whether fund performance can be linked to the state of the economy, we include investor sentiment (Baker and Wurgler (2006)) and the Chicago Fed National Activity Index (CFNAI), a series which captures the state of the macro economy and is itself an index of many macro time

Figure 1: Macroeconomic time series plots



(a) Sentiment



(b) CFNAI

These figures show the macroeconomic time series plot. Panel (a) plots the sentiment time series and panel (b) plots the CFNAI time series.

series.[8] Figure 1 plots the time series plots of both macro variables. Kacperczyk, Van Nieuwerburgh, and Veldkamp (2014) shows that mutual fund performance depends on CFNAI.

3 Main Analysis

Our main analysis aims to predict mutual fund abnormal returns. This analysis is an out-of-sample prediction analysis with many conditioning variables. It allows for interactions of characteristics (the 59 characteristics in Table 1 plus sentiment/CFNAI), as well as for non-linearities in the relationship between characteristics and future fund outperformance. To that end, we use

[8]CFNAI is the first principal component of 85 economic indicators from four broad categories: production and income; employment, unemployment, and hours; personal consumption and housing; and sales, orders, and inventories. Sentiment also uses principal component analysis to combine the information from multiple economic indicators, which include the closed-end fund discount, NYSE share turnover, number of IPOs, average first-day returns on IPOs, equity share in new shares and the dividend premium.

an artificial neural network, similar to Gu, Kelly, and Xiu (2020). In their extensive comparison study, they show that this method dominates other ML techniques for predicting stock returns. We predict fund abnormal returns with a neural network of lagged predictors:

$$R^{abn}_{i,t+1} = g(z_{it}, z_t) + \epsilon_{i,t+1} \tag{3}$$

The structure of the neural network $g(\cdot)$ is selected based on a validation sample. It uses as its inputs the characteristics $z_{i,t}$ specific to mutual funds, and macro-economic variables z_t to build the best predictors of fund abnormal returns. We focus on sentiment as our main macroeconomic variable, and discuss the results with CFNAI as a robustness check.

3.1 Sampling Scheme

We use a cross-out-of-sample analysis to evaluate the performance of the neural network model. We split the full sample into three folds, periods of the same length. We use two of the folds to estimate the model and select the tuning parameters, and evaluate the prediction out-of-sample on the remaining fold. Following Kozak, Nagel, and Santosh (2020), Lettau and Pelger (2020) and Bryzgalova, Pelger, and Zhu (2021), we cross-validate the estimation on three different combinations of the three folds, thereby obtaining an out-of-sample prediction for each data point in our sample. This cross-out-of-sample evaluation diminishes the effect of particular subperiods in the out-of-sample analysis. The estimation and validation time period (2/3 of the sample) is split into 3/4 used for estimation (training) and 1/4 used for validation (to select the tuning parameters).

Our baseline results select the dates that go into each of the three folds randomly. The top panel of Figure 2 shows this random sampling scheme, where different colors denote the three folds. The bottom panel shows the more traditional chronological sampling. A second alternative is an expanding-window chronological estimation and evaluation without cross-validation. We analyze the two alternative sampling schemes below, and show that our benchmark predictability results are robust to the sampling scheme.[9]

Each sampling scheme has benefits and drawbacks. The random sampling approach has the important advantage of having a more equal distribution of high- and low-sentiment observations in each fold, as illustrated in Figure 2. In contrast, the chronological sampling approach may have no high-sentiment periods in the evaluation fold (in the validation where the green fold is used for out-of-sample evaluation) or the estimation fold (in the validation where the green fold is used for estimation). If sentiment is an important conditioning variable in the prediction problem, as we will show is the case, the chronological results will fail to accurately capture the dependencies

[9] In Internet Appendix IA.3, we also investigate the impact of overlap between data points that are both in the 36-month window used for the estimation of the abnormal return (equations 1 and 2) and in the testing fold in the random cross-out-of-sample validation. We find that eliminating this overlap has minimal impact on the results, but reduces estimation precision.

on the underlying economic state. By using the full time span of the data, the random sampling scheme allows the model to better learn the non-linear function $g(z_{it}, z_t)$.

The random sampling scheme does not create look-ahead bias. The economic object of interest is the conditional abnormal return $g(z_{it}, z_t)$. A prediction can be interpreted as a cross-sectional non-parametric regression, where the time-series order of the cross-sectionally stacked triplets $\{R^{abn}_{i,t+1}, z_{it}, z_t\}_{i,t}$ is not explicitly taken into account. This regression only uses variables known at time t to predict abnormal returns at time $t+1$. Information in the error term ϵ_{t+1} of the prediction is never used in the prediction.

What is true is that both the random and chronological sampling schemes with cross-validation are not available to investors in real-time. In contrast, the expanding-window chronological sampling scheme does represent a feasible investment strategy at each date. However, the latter estimates a new model at each point in time, making the interpretation more problematic. It also uses less data for evaluation and estimation, sacrificing precision. Because it uses less data and because is not a cross-validation, it has an even less representative distribution of economic states to learn from at each date.

3.2 Neural Network

A feedforward network (FFN) is a flexible non-parametric estimator that can learn any functional relationship $y = f(x)$ between an input x and output variable y with sufficient data.[10] A deep neural network combines several layers by using the output of one hidden layer as an input to the next hidden layer. Our best model structure is a one-layer neural network.

It combines the raw predictor variables (or features) $z = z^{(0)} \in \mathbb{R}^{K^{(0)}}$ linearly and applies a non-linear transformation. This non-linear transformation is based on an element-wise operating activation function. We choose the popular rectified linear unit activation function (ReLU),[11] which component-wise thresholds the inputs and is defined as:

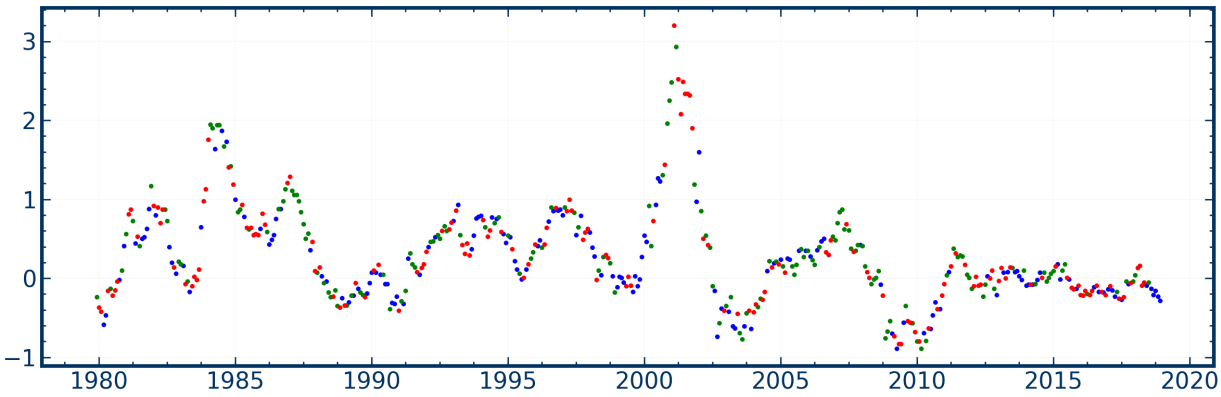
$$\text{ReLU}(z_k) = \max(z_k, 0).$$

The result is the hidden layer $z^{(1)} = (z^{(1)}_1, ..., z^{(1)}_{K^{(1)}})$ of dimension $K^{(1)}$ which depends on the parameters $W^{(0)} = (w^{(0)}_1, ..., w^{(0)}_{K^{(0)}})$ and the bias term $w^{(0)}_0$. The output layer is simply a linear trans-

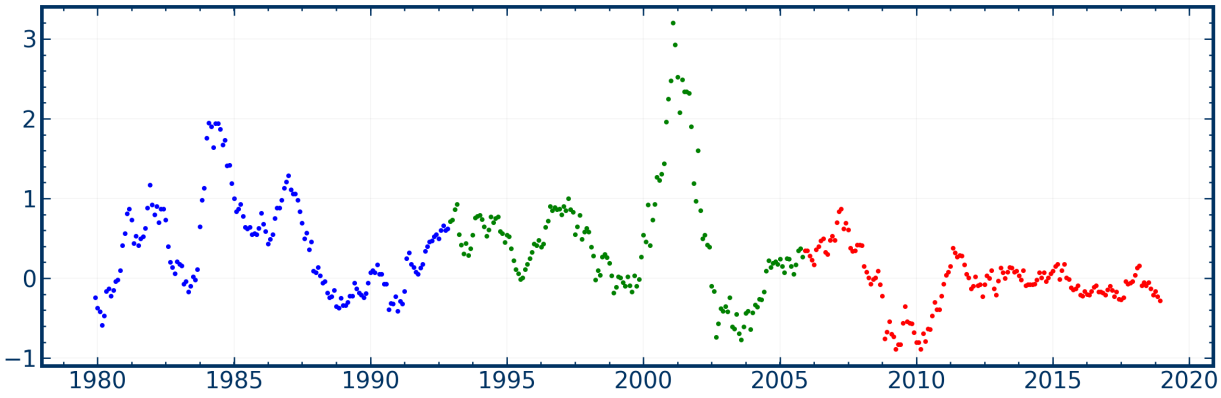
[10] FFN are among the simplest neural networks and treated in detail in standard machine learning textbooks, e.g. Goodfellow, Bengio, and Courville (2016).

[11] ReLU activation functions have a number of advantages including the non-saturation of its gradient, which greatly accelerates the convergence of stochastic gradient descent compared to the sigmoid/hyperbolic functions and fast calculations of expensive operations.

Figure 2: Sentiment time series for the different cross-out-of-sample folds



(a) Random sampling



(b) Chronological sampling

This figure plots the Baker and Wurgler (2006) sentiment measure from 1979/12 to 2018/12. Different colors denote the three different cross-out-of-sample folds, which we use throughout the paper. The top figure shows the random sampling into three folds, while the bottom figure shows chronological sampling.

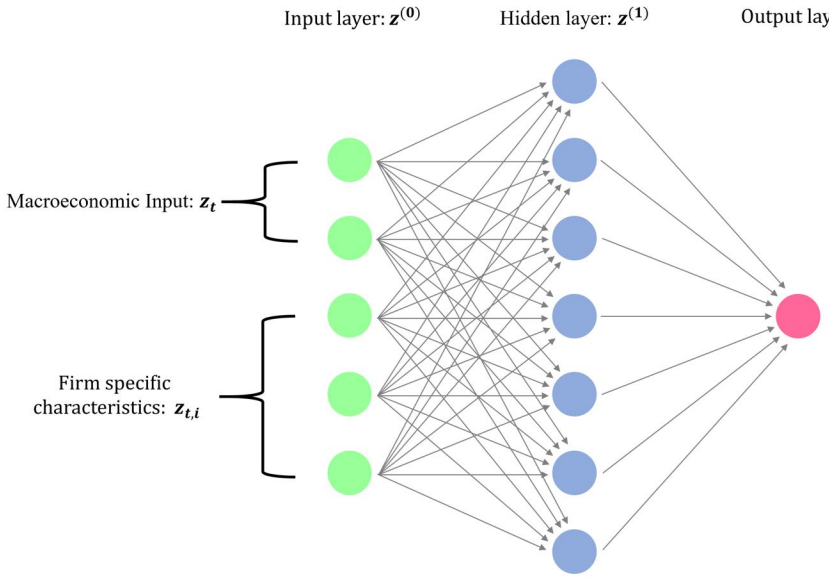
formation of the output from the hidden layer.

$$z^{(1)} = \text{ReLU}(W^{(0)\top} z^{(0)} + w_0^{(0)}) = \text{ReLU}\left(w_0^{(0)} + \sum_{k=1}^{K^{(0)}} w_k^{(0)} z_k^{(0)}\right)$$
$$R^{abn} = W^{(1)\top} z^{(1)} + w_0^{(1)} \qquad \text{with } z^{(1)} \in \mathbb{R}^{K^{(1)}}, W^{(0)} \in \mathbb{R}^{K^{(1)} \times K^{(0)}}, W^{(1)} \in \mathbb{R}^{K^{(1)}}.$$

Note that without the non-linearity in the hidden layer, the one-layer network would reduce to a generalized linear model. Our optimal network has 64 nodes in the hidden layer, which can be interpreted as representing the information set with 64 basis functions which are non-linear transformations of the original characteristics and macroeconomic variables and which are linearly combined to predict the abnormal return.

Our results are extremely robust to the choice of tuning parameters. Networks with more lay-

Figure 3: Illustration of Feedforward Network with Single Hidden Layer



ers and nodes result in a very similar performance and estimated functional form as our optimal network. This is consistent with the findings in Chen, Pelger, and Zhu (2023) and Gu, Kelly, and Xiu (2020). Hence, it matters primarily to allow for the flexible functional form and interaction effects, which can be achieved with many model specifications. The details of the hyperparameter tuning are in Appendix B, where we also show the robustness to the number of layers in Section IA.4 of the Internet Appendix.

We quantify the economic benefit of different information sets by comparing the prediction and trading benefits of each information set available to the neural network. As appropriately tuned neural networks can approximate any functional relationship, they allow us to understand what the best possible prediction is for a given information set.

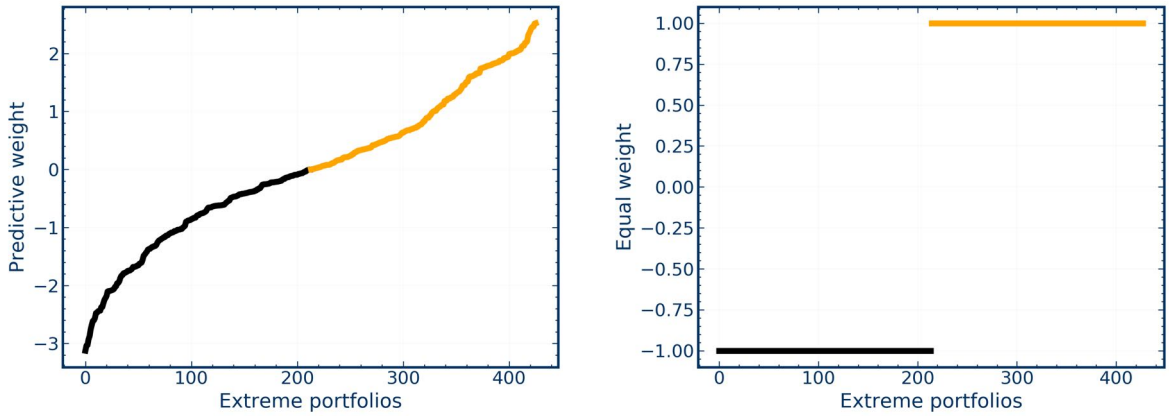
As a robustness check, we explore Gradient Boosted Trees as an alternative machine learning approach. We obtain broadly similar conclusions, and relegate an in-depth comparison to Internet Appendix IA.5. Neural networks have the advantage that we can provide valid confidence intervals, similar to regression analysis, which is not always the case for alternative machine learning predictors.

3.3 Optimal Prediction

Having estimated the neural network model, we form the model's prediction of fund abnormal returns for each fund-month using all 59 characteristics listed in Table 1 and investor sentiment. We sort funds in deciles based on their predicted abnormal return for the next month.

Within the deciles we weight the funds either by their predicted value or equally. Figure 4 illustrates the two weighting schemes in the extreme deciles for a representative month. The

Figure 4: Equally and prediction-weighted portfolio weights within deciles



(a) Prediction weights **(b)** Equal weights

These figures show the equal and prediction-weighted portfolio weights for the first and tenth decile in a long-short portfolio for the representative example month 2000/01. The x-axis refers to the sorted firms in the bottom 10% and in the top 10% of the predicted abnormal return distribution

prediction weights in the left panel exploit the heterogeneity in the prediction and assign a higher relative weight to predictions that deviate more from the center of the decile.[12]

Figure 5 plots the cumulative abnormal return from investing in each of these 10% of funds. The right panel equally-weights the abnormal returns of the funds within each decile, using the neural network model only to sort funds into deciles. The left panel additionally uses the neural network model prediction to form portfolio weights; we refer to this as the prediction-weighted return. A prediction-weighted approach uses more information resulting in a larger spread in the prediction portfolios. The baseline model for the rest of the paper is prediction-weighted.[13] An investor who had followed an investment strategy that invests in the 10% best mutual funds based on the neural network model's predictions would have earned a cumulative abnormal return of

[12]The prediction based weights are defined as the following shifted and scaled weights:

$$\text{For top portfolio: } \tilde{\mu}_{i,t} = \hat{\mu}_{i,t} - \min_{i \in \text{Top}} (\hat{\mu}_{i,t}) \tag{4}$$

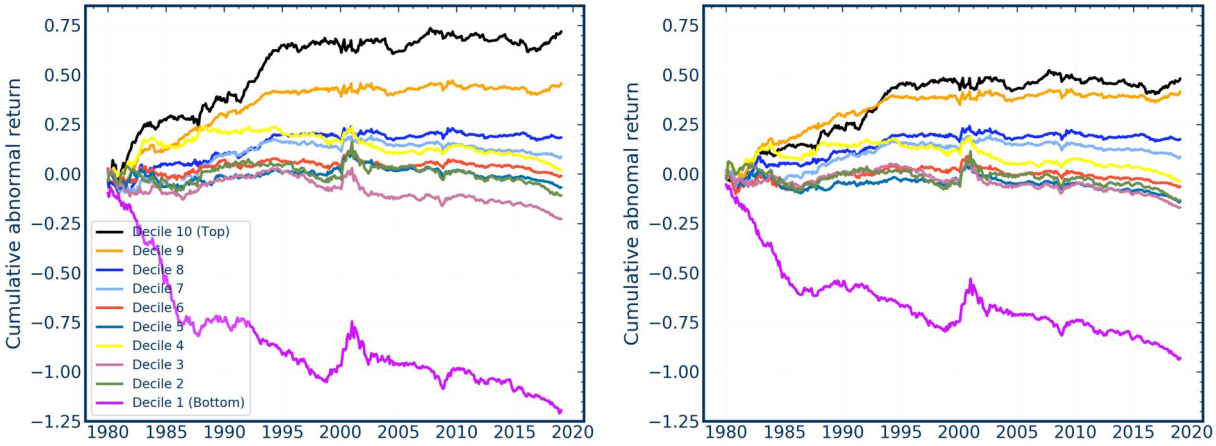
$$\text{For bottom portfolio: } \tilde{\mu}_{i,t} = \hat{\mu}_{i,t} - \max_{i \in \text{Bottom}} (\hat{\mu}_{i,t}) \tag{5}$$

$$w_{i,t}^{\text{pred}} = \frac{\tilde{\mu}_{i,t}}{\sum_{i=1}^{N} \tilde{\mu}_{i,t}} \tag{6}$$

where $\hat{\mu}_{i,t}$ are the predictions of neural network models. For top-performing funds, we subtract the smallest model prediction within the group in equation (4) to ensure that the top portfolio is a long-only portfolio. For bottom-performing funds, we subtract the largest model prediction within the group in equation (5) to ensure that the bottom portfolio is a short-only portfolio. We then standardize the normalized predictions to sum up to 1 per equation (6). The prediction weights are similar for the other deciles. The results for quintiles and 20 quantiles are very similar and available upon request. An alternative to the prediction weights within the quantiles are rank weights. The results with rank weights are very similar to those with prediction weights and are omitted for brevity.

[13]The results for the equally-weighted approach are similar and presented in Appendix A.1.3. We explore a third weighting scheme, which is value-weighted returns in Appendix A.1.8.

Figure 5: Cumulative abnormal returns of prediction deciles for all characteristics.



(a) Prediction-weighted abnormal returns.

(b) Equally-weighted abnormal returns.

These figures show the cumulative abnormal returns sorted into prediction deciles when considering the complete information set (fund-specific and stock-specific characteristics + sentiment) to predict abnormal returns.

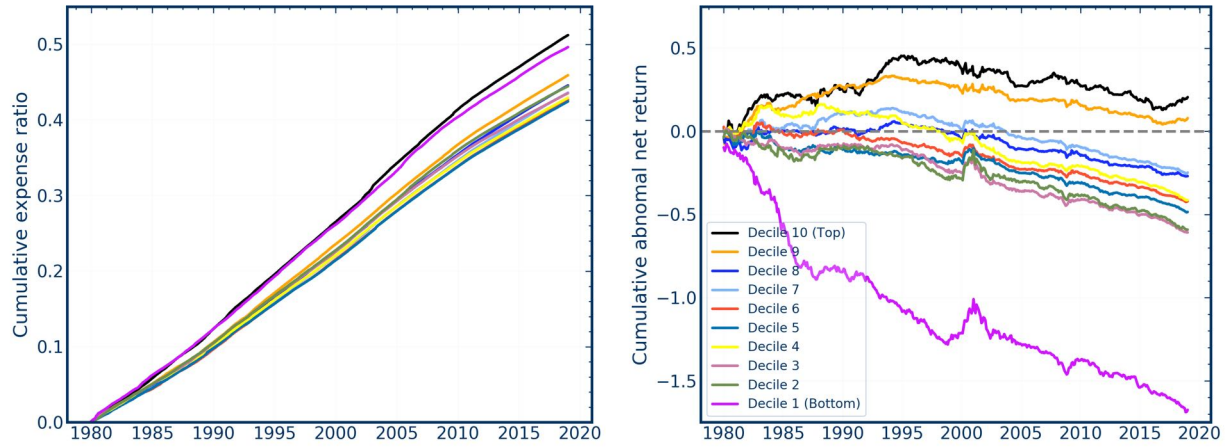
72% prediction-weighted and 48% equally-weighted. The difference between these two numbers shows that the neural net is not only good at predicting which funds are likely to be in the top performance decile, but also at how good some of the funds in the top decile are relative to other top-performing funds.

At the other end of the spectrum, the 10% worst funds according to the out-of-sample prediction of the neural network model generate cumulative abnormal returns of -119% prediction-weighted and -93% equally-weighted. Hence, avoiding the 10% worst mutual funds is even more valuable than investing in the best 10% of mutual funds.

The main conclusion is that abnormal mutual fund returns are predictable and the extent of predictability over the past 40 years is economically large. The cumulative abnormal return translates into a monthly out-performance of 15 basis points for the 10% best funds prediction-weighted (10 basis points equally-weighted). The 10% worst funds see 25 basis point per month underperformance prediction-weighted (20 basis points equally-weighted).

Panel (a) of Figure 6 shows the average fees for the different prediction-based decile portfolios. While those funds with higher predicted and realized abnormal returns charge a higher fee, the spread in fees does not explain the spread in expected returns. In fact, the worst and best 10% of funds both have a cumulative expense ratio of around 50%, which are higher than the expense ratios of the funds in the middle of the predicted performance distribution. Given that the 10% best and 10% worst funds have the same fees, fees can explain nothing of their relative performance. The 10% best funds earn cumulative abnormal gross returns of 72%, exceeding cumulative fees. Indeed, Panel (b) of Figure 6 shows that the funds in the top two prediction deciles still earn a positive abnormal return after fees. Note that these are lower bounds as predicting abnormal re-

Figure 6: Cumulative expense ratios and abnormal net return.



(a) Cumulative prediction-weighted expense ratios

(b) Cumulative abnormal net returns

The left figure shows the cumulative expense ratios of prediction-weighted deciles based on the full information set (fund-specific and stock-specific characteristics + sentiment). The right figures shows the abnormal net returns for the prediction-weighted deciles, that is, the abnormal returns minus the fees.

turns after fees (rather than before fees) improves the predicted after-fee performance. Figure A.1 shows the performance for predicting abnormal after-fee returns. The 10% best funds achieve a cumulative abnormal return after fees of 37%, which substantially exceeds the 72% gross outperformance minus 50% in fees. On the other hand, the 10% worst funds earn cumulative abnormal returns after fees of around -170%, further highlighting the usefulness of the neural network in identifying which funds to avoid.

In the frictionless Berk and Green (2004) model, all the out-performance should go to the managers in the form of higher fee revenues, resulting in zero abnormal returns after fees. Rather, we find that about 20% of funds outperform after fees, while the remaining 80% have negative after-fee performance. The outperformance suggests the presence of frictions, while the underperformance is consistent with the presence of unsophisticated investors that do not account properly for risk after fees and neglect to withdraw their investments (Ben-David, Li, Rossi, and Song, 2022). The asymmetry seems to support a separation of investors into "smart" and "unsophisticated" investors. The best funds charge among the highest fees, in line with the predictions of Berk and Green (2004). However, unsophisticated investors do not properly measure skill and end up investing into funds that earn negative abnormal returns after fees.

The magnitude of the predictability deserves some discussion. First, the measured performance is not the alpha of a trading strategy, but the realized performance of mutual funds. Second, it refers to an abnormal return beyond the compensation for exposure to the four factors. Most, importantly, it represents an out-of-sample performance. Related work by Roussanov, Ruan, and Wei (2021, 2022) uses Bayesian forecasts of after-fee outperformance that are based on deviations

of fund size relative to what the Berk and Green (2004) model implies. They obtain magnitudes in performance comparable to ours, but their results are not entirely out-of-sample. Hence, our out-of-sample predictability of 15 basis points monthly for the top funds and -25 basis points for the worst funds over a 40-year sample is substantial.[14]

3.4 Which Information Most Useful When Predicting Fund Abnormal Returns?

To assess the importance of stock-specific characteristics (labeled 1-46 in Table 1), fund momentum (47-49), fund characteristics (50-54), family characteristics (55-59), and sentiment for the prediction of fund abnormal performance, we estimate neural network models that are given subsets of predictors. Our main finding is that the combination of fund-level variables and sentiment results in the best performance. Stock-specific characteristics of the stocks held by funds do not help predict fund abnormal return.

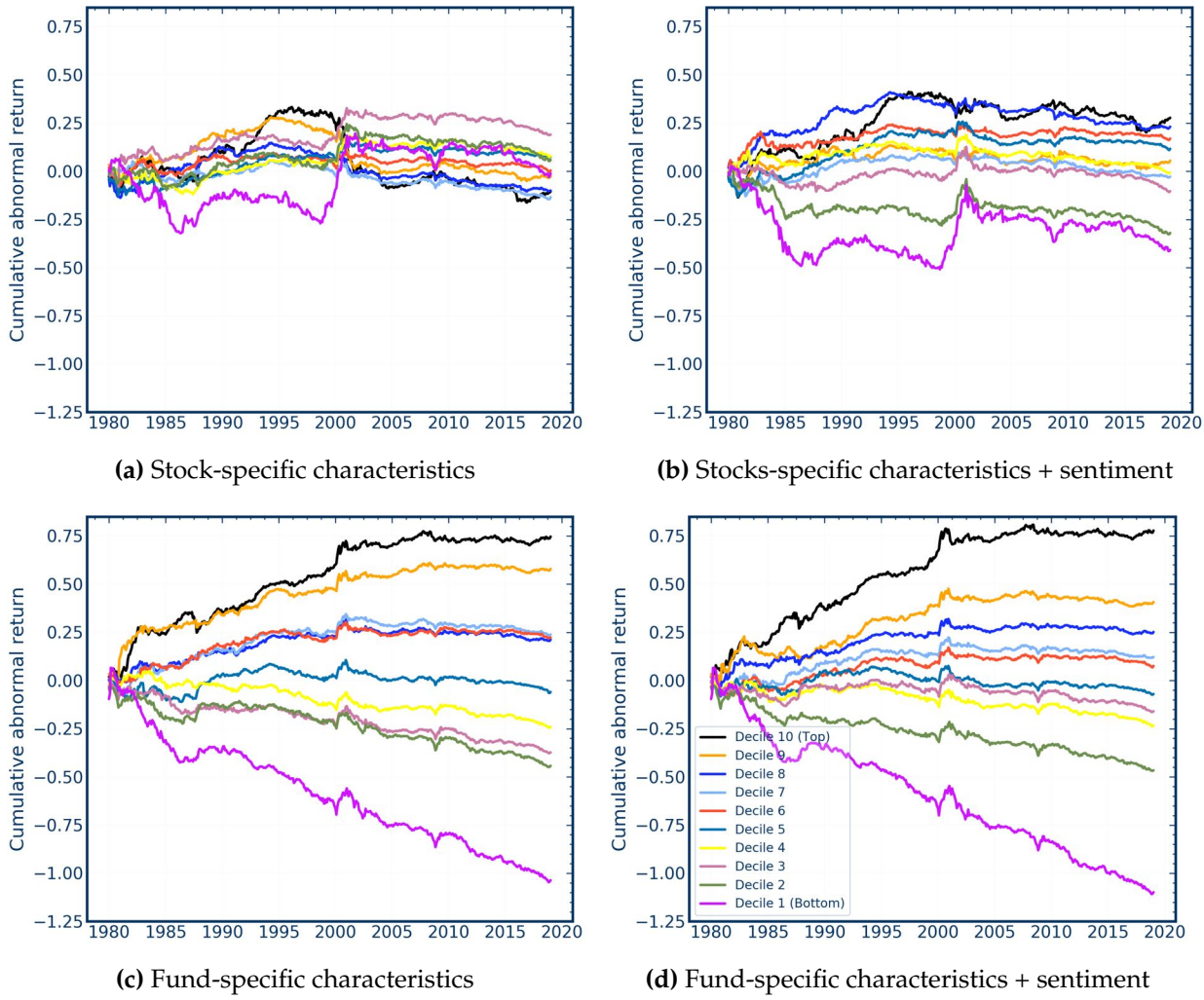
Figure 7 shows the cumulative abnormal returns for the fund deciles when using stock characteristics 1-46 only (Panel A), stock characteristics and sentiment (Panel B), fund characteristics 47-59 (Panel C), and fund characteristics and sentiment (Panel D). Fund abnormal returns within each decile are prediction-weighted. The best model for predicting fund abnormal returns ignores stock characteristics entirely. Fund characteristics, in sharp contrast to stock characteristics, are extremely useful for prediction, as is sentiment. We note the monotone pattern in Panel D. As we will see shortly, fund characteristics interact with sentiment in important ways.

As most of the predictability is in the extreme deciles, we propose a long-short prediction portfolio of the top and bottom decile as a measure for the spread in skill. This is a convenient economic measure of the spread, not a tradable investment strategy. Figure 8 shows that the portfolio that goes long in the (predicted) best 10% of funds and short in the (predicted) worst 10% of funds earns cumulative returns of -9% out-of-sample when only stock information is used, 69% when using stock plus sentiment information, 178% when using fund information, 188% when using fund plus sentiment information, and 191% when using stock plus fund plus sentiment information. The results are qualitatively the same for equally-weighted portfolios as shown in Appendix A.1.3.

To assess whether these different long-short investment strategies incur different amounts of risk, we compute the Sharpe Ratio on the long-short decile portfolio, which Table 3 reports alongside the mean return. The highest Sharpe-ratio strategy ignores stock-specific information. Using fund information and sentiment to select the best and worst 10% of funds results in a monthly long-short return of 40 basis points with a monthly Sharpe ratio of 0.25, which translates into a

[14] Roussanov, Ruan, and Wei (2021) show in their Figure 2 that the top predicted-skill decile outperforms by 5.8 basis points monthly and the bottom decile underperforms by 23 basis points. Roussanov, Ruan, and Wei (2022) show in their Table 4 that the top (bottom) decile has a performance of 11 basis points (-26 basis points) value weighted. Similar to our results, only the top decile has a significantly positive net-of-fee alpha. Overall, their results provide a useful benchmark to compare to for our out-of-sample analysis.

Figure 7: Cumulative abnormal returns for different information sets.



(a) Stock-specific characteristics

(b) Stocks-specific characteristics + sentiment

(c) Fund-specific characteristics

(d) Fund-specific characteristics + sentiment

These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment, stock-specific characteristics+ sentiment, fund-specific characteristics or stock-specific characteristics to predict abnormal returns.

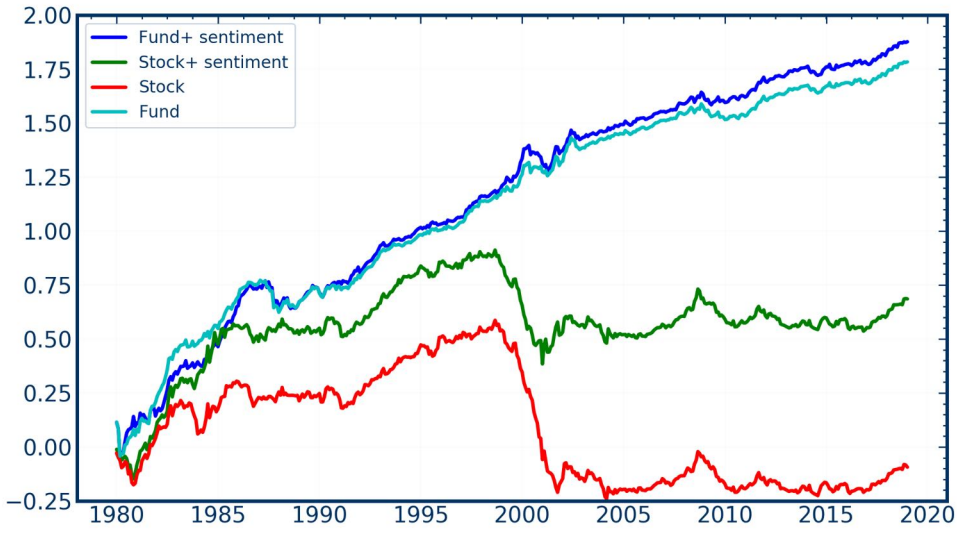
0.87 annual Sharpe ratio.

The last three rows show that when one only uses four out of the 13 fund characteristics, namely the fund momentum group (F_r12_2, F_r2_1 and F_ST_Rev) and flow, combined with sentiment, the resulting long-short portfolio has a similar mean return and Sharpe ratio to the portfolio based on all fund information. The last row shows that using all fund characteristics except for these two results in substantially worse performance. In summary, fund momentum and flow, interacted with sentiment, are the key variables for predicting fund abnormal returns.

The last column of Table 3 reports the R^2_F statistic, which measures how well the realized long-short portfolio return is predicted by the neural network model.[15] If the realized long-short abnor-

[15]We denote by F_t the realized return and by $\hat{F}_t$ the predicted return of the long-short portfolio based on prediction-

Figure 8: Cumulative abnormal returns of long-short prediction portfolios



This figure plots the cumulative abnormal returns of prediction-weighted long-short decile portfolios that use different information sets for prediction. We consider fund-specific and stock-specific characteristics combined with sentiment.

mal return factor is predicted more accurately, then an investor knows better by how much funds in the top decile will outperform funds in the bottom decile in the next period.[16] The highest R^2_F of 5.00% monthly, which is substantial, is obtained for the full model with fund, sentiment, and stock information. Dropping sentiment information results in a large decline in R^2_F, which suggests that sentiment is important for predicting the high-minus-low abnormal fund return. As we will see below, the conditional mean of the long-short portfolio is higher in high sentiment periods. Replacing fund- with stock-level information also results in a large drop in R^2_F. Note that the R^2_F measures accuracy of both the relative cross-sectional prediction of funds (the fund ranking) and the level of the abnormal returns. The prediction based only on fund information correctly predicts the ranking of future abnormal fund returns, but not their magnitude as suggested by the low R^2_F. Including sentiment slightly improves the relative prediction, due to the interaction effects studied in Section 4.1, but substantially improves the level prediction of abnormal returns.[17] Table A.1 reports the results for the top and bottom prediction deciles separately with the same findings.

sorted deciles. The normalized time-series prediction error is measured by $R^2_F = 1 - \frac{\sum_{t=1}^{T}(\hat{F}_t - F_t)^2}{\sum_{t=1}^{T} F_t^2}$.

[16] This information could be used for timing and sizing the portfolio investment in the spirit of Haddad, Kozak, and Santosh (2020).

[17] The results for the different measures also illustrate the challenges of using them to select the best model. The measures are tools for understanding what economic structure is captured. The models do not need to improve uniformly along all measures.

Table 3: Performance of long-short abnormal return portfolios for different information sets.

Information set	mean(%)	t-stat	SR	R^2_F (%)
Stock+ fund+ sentiment	0.41	4.5***	0.21	5.00
Fund+ sentiment	0.40	5.4***	0.25	2.73
Fund	0.38	5.5***	0.25	0.19
Stock+ fund	0.28	3.3***	0.15	2.30
Stock+ sentiment	0.15	1.6	0.07	1.27
Stock	-0.02	-0.2	-0.01	-1.60
Fund momentum + sentiment	0.35	4.4***	0.21	0.29
Fund momentum + Flow + sentiment	0.48	5.2***	0.24	0.92
Fund excl. fund momentum and flow	0.06	1.5	0.07	0.12

This table reports the Sharpe ratio, mean and factor R^2 of long-short prediction-weighted decile portfolios that use different information sets for the prediction. We consider nine different information sets which combine fund-specific and stock-specific characteristics and sentiment. We also include flow and fund momentum (F_r12_2, F_r2_1 and F_ST_Rev) individually.

Spanning Do the long-short portfolio returns formed from the neural network model's prediction reflect true abnormal return or compensation for risk? To explore this question, we estimate a multivariate regression of the long-short portfolio return on several sets of return factors from the literature. The results for the main information set is reported in the first row of Table 4. The second row reports results when using fund characteristics and sentiment in the predictor set. In both cases, we find large and highly significant intercepts α, relative to the mean return μ (reported in the last column), and low R^2. Hence it is not the case that our approach results in a mutual fund portfolio return that inadvertently captures compensation for systematic factor exposure.

Robustness to Fund Size The predictability is robust to excluding or down-weighting small mutual funds. First, we exclude mutual funds with less than 15 million asset under management (TNA), which is an often used cutoff in the literature (see for example Doshi, Elkamhi, and Simutin (2015)). Figure 9(b) shows the result is essentially unaffected by dropping the smaller funds. Second, we combine the prediction classification with the value of the assets under management of the funds to form value-weighted prediction portfolios. Figure 9(a) shows that the magnitude of the outperformance of the best over the worst decile of funds for value-weighted portfolios is very similar to prediction-weighted portfolios. Interestingly, the top decile performs better under value-weighting, suggesting that some of the best funds have relatively high AUM. Appendix A.1.8 provides further details.

3.5 Longer Holding Periods

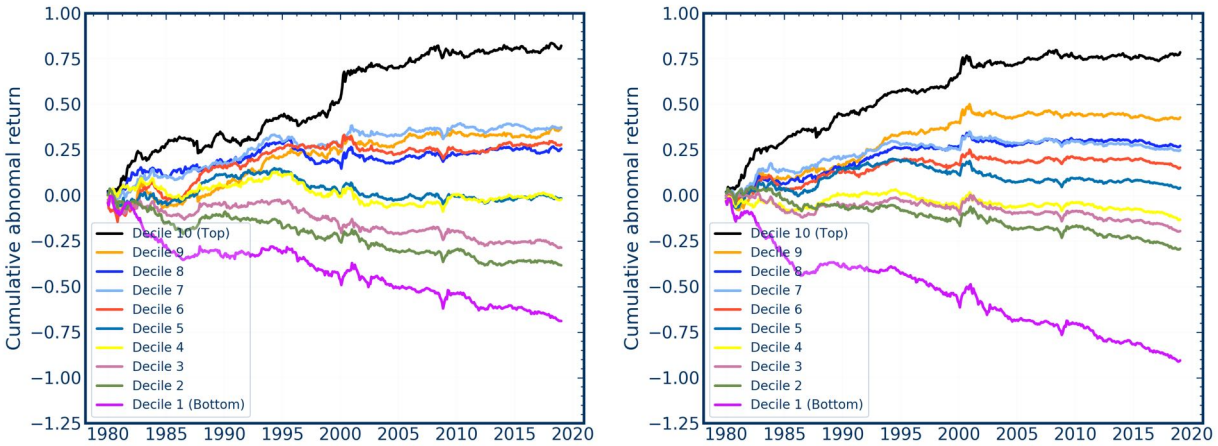
The monthly rebalancing of mutual funds is not crucial to earning the high abnormal returns associated with the relative performance of funds. Figure 10 shows the abnormal returns on a

Table 4: Spanning of long-short abnormal prediction portfolios with different factor models.

	FF 4 factors		FF 5 factors		FF 6 factors		FF 8 factors		
	α	R^2	α	R^2	α	R^2	α	R^2	mean μ
Stock+ fund+ sentiment	0.15***	0.27	0.12***	0.31	0.10**	0.35	0.09**	0.35	0.15***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Fund+ sentiment	0.16***	0.12	0.21***	0.03	0.17***	0.12	0.19***	0.14	0.18***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Fund	0.15***	0.16	0.21***	0.05	0.16***	0.17	0.17***	0.18	0.19***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Stock+ fund	0.09*	0.14	0.10**	0.13	0.08*	0.16	0.06	0.19	0.06
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Stock+ sentiment	0.10**	0.28	0.05	0.36	0.03	0.38	0.03	0.38	0.09*
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Stock	0.04	0.15	0.03	0.17	0.02	0.18	0.00	0.22	0.01
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Flow+ fund momentum+ sentiment	0.13***	0.26	0.23***	0.12	0.17***	0.29	0.20***	0.35	0.17***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
F_r12_2+ sentiment	0.08*	0.12	0.18***	0.04	0.13***	0.17	0.13***	0.17	0.12***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)

This table reports the time-series regression results of long-short prediction-weighted decile portfolios for different factor models. We compare different information sets to predict abnormal returns. We consider the 4-factor Fama-French-Carhart model (market, size, value and momentum), the 5-factor Fama-French model (market, size, value, profitability and investment), a 6-factor model which adds the momentum factor to the Fama-French 5 factors, and an 8-factor model which adds the momentum, short-term reversal and long-term reversal factors to the Fama-French 5 factors. The α column reports the time-series pricing error and R^2 is the explained variation of the regression. Both the long-short abnormal return portfolios and the factor models are normalized to have a standard deviation of 1. Standard errors are in brackets and stars denote the significance levels.

Figure 9: Cumulative abnormal returns of prediction deciles for all characteristics.



(a) Prediction-value-weighted

(b) Prediction-weighted with size cutoff.

These figures show the out-of-sample cumulative abnormal returns sorted into prediction deciles. We predict abnormal returns with fund-specific characteristics and sentiment. The left subfigure uses prediction-value weighted portfolios, while the right subfigure shows the cumulative abnormal returns for prediction-sorted portfolios for mutual funds with at least 15 million $ assets under management

long-short prediction portfolio for holding periods up to 3 years. Fund investments are made

every month based on the one-month ahead prediction, but investments are held for a longer holding period, ranging from 1 to 36 months (with overlapping holding periods). As expected, the mean return decreases over time but it stays significant for all holding periods (top left panel). At the same time, the longer holding periods decrease the standard deviation of the return (top right panel). A 3-month holding period reduces the standard deviation more than the mean, and hence results in a higher Sharpe ratio than the one-month holding period (bottom left). Even after 36 months, the monthly Sharpe ratio is still 0.20, compared to 0.30 after 3 months. The outperformance remains statistically significant even after 36 months (bottom right panel). This occurs even though the model only attempts to predict abnormal returns one-month ahead. In light of the literature's difficulty in finding persistence in abnormal returns, this result is remarkable.

Removing stock characteristics results in similar long-horizon Sharpe Ratios as with the full information set. Trading based on the most important predictors (flow, fund momentum, and sentiment) results in similar holding-period patterns as using the full information set.

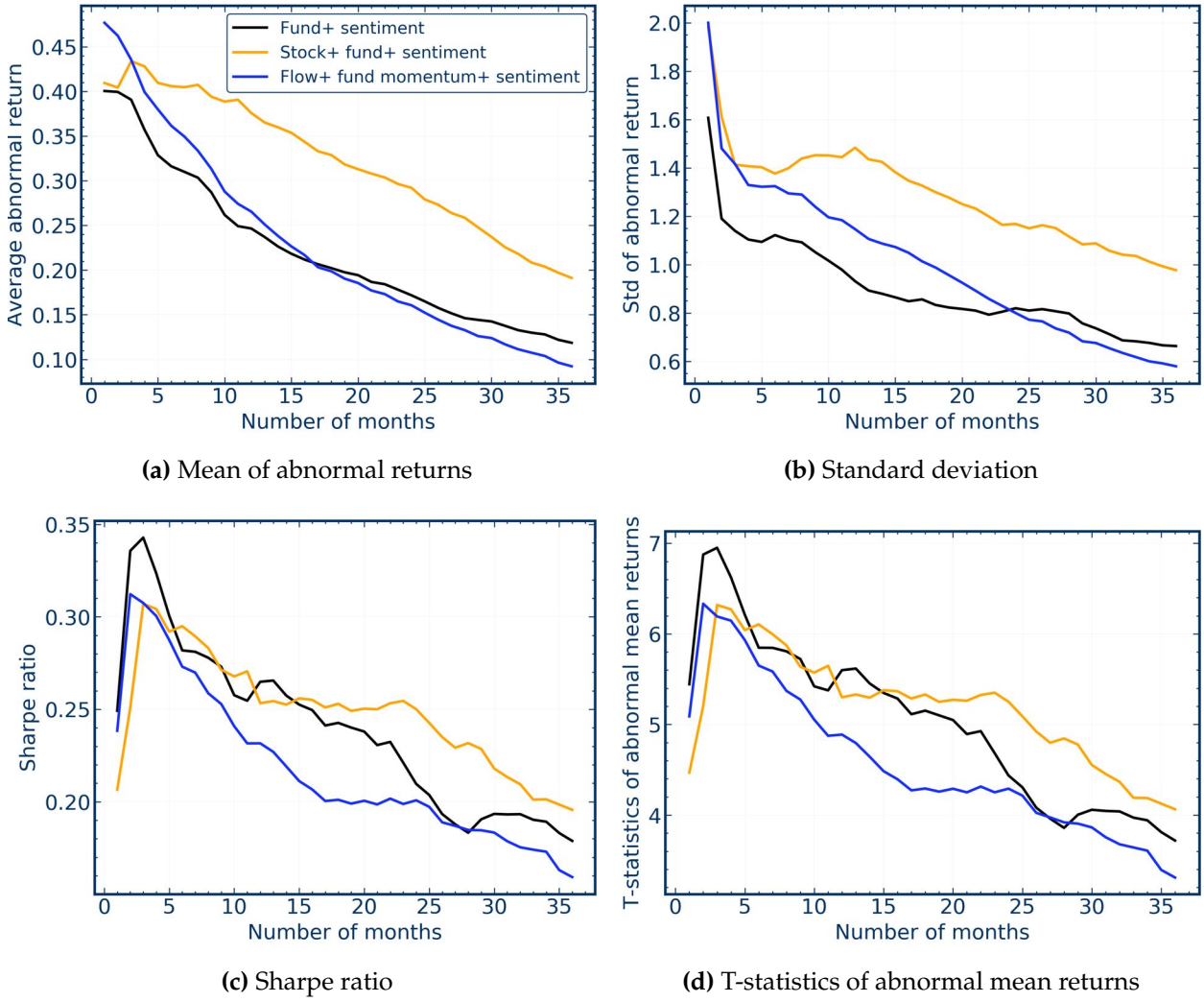
We obtain similar results for longer holding periods when using the longer-horizon abnormal return as the prediction objective. Appendix A.1.4 shows the strong predictability for an annual abnormal return prediction. While 12-month-horizon prediction lowers the mean return, it also reduces the variance and, hence, results in similar monthly Sharpe ratios as the one-month ahead prediction.

The predictability lasts for longer because many fund-specific characteristics contain predictive information that remains relevant for a longer horizon. Figure A.9 shows the autocorrelation of fund-specific characteristics. Except for short-term momentum (F_r2_1) and short-term reversal (F_ST_Rev), the fund characteristics are persistent. This is also reflected in the persistence of the classification of funds. Figure A.10 shows the transition matrix between the different prediction quantiles for each month. Over 60% of the top-20% and bottom-20% funds stay in the same prediction quantile in the next month. The classification remains stable for longer time periods.

3.6 Sampling Scheme

Results are not sensitive to the random sampling scheme with three-fold cross-validation. Appendix A.2 revisits all results using chronological sampling with cross-validation, while Appendix A.3 does the same for a chronological expanding-window estimation. We highlight the main finding here. Panel A of Figure 11 shows a very similar out-performance of the predicted top-decile and under-performance of the predicted bottom decile with chronological sampling with three-fold cross-validation. Panel B shows again similar performance with chronological expanding-window sampling (without cross-validation). Table 5 shows that, if anything, the mean return and Sharpe ratio of the long-short portfolio based on the full information set are higher for the two alternative sampling schemes than for the benchmark random sampling scheme.

Figure 10: Performance for Different Holding Periods



(a) Mean of abnormal returns

(b) Standard deviation

(c) Sharpe ratio

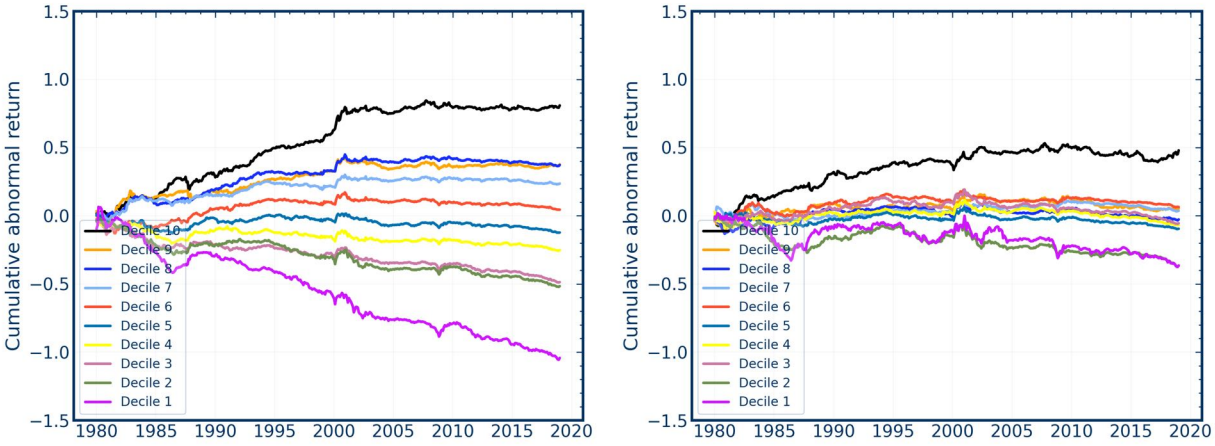
(d) T-statistics of abnormal mean returns

This figure shows the results for long-short prediction-weighted portfolios for different holding periods. At each time t, we sort funds based on the one-month prediction into deciles and hold the long-short prediction portfolio for s months with overlapping returns. We calculate the mean, Sharpe ratio, standard deviation, and t-statistics of the overlapping abnormal returns. The one-month prediction uses either fund+sentiment, stock+fund+sentiment or flow+fund+sentiment.

The second main result, that stock characteristics are less useful for predicting out- and under-performance of mutual funds than fund characteristics is also robust to the sampling scheme. This is clearly visible in the right panels of Figure 11 and confirmed in the row "Stock" of Table 5. While there is some statistical evidence for predictability of stock characteristics in the chronological scheme, the mean return and Sharpe ratio are less than half as large as when fund and sentiment information are added. The results in the expanding-window sampling are weaker still.

Figure 11: Cumulative abnormal returns for different sampling schemes.

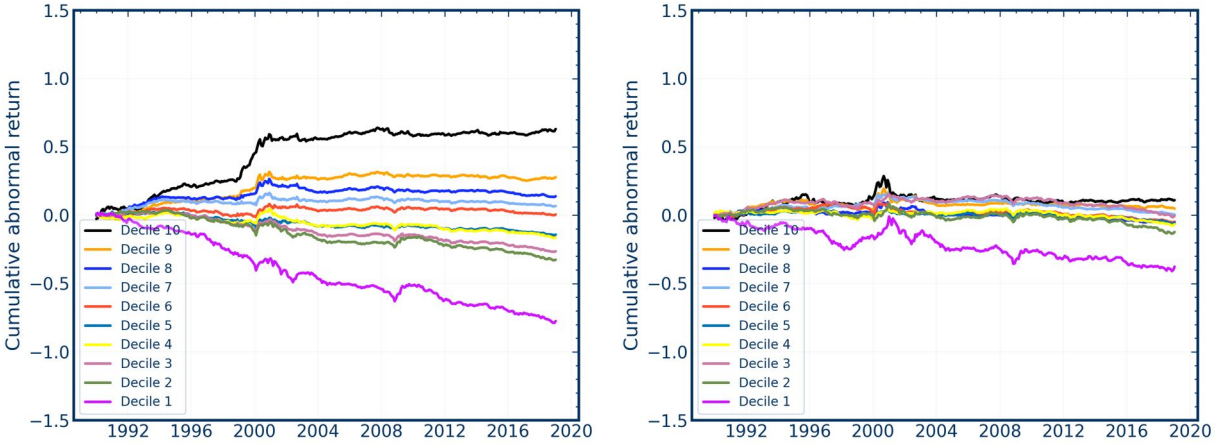
Panel A: Chronological sampling



(a) Fund-specific characteristics + sentiment

(b) Stock-specific characteristics

Panel B: Expanding window



(c) Fund-specific characteristics + sentiment

(d) Stock-specific characteristics

These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The abnormal returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment and stock-specific characteristics. In Panel A, the three cross-out-of-sample folds keep the chronological order. In Panel B, we estimate prediction models on an expanding window, that is, for each t we use all available data to estimate new predictions for year $t + 1$. As we need an initial training data set, the out-of-sample analysis for the expanding window starts in 1990.

Table 5: Performance of abnormal return portfolios: Sampling Schemes.

Sampling	Information set	mean(%)	t-stat	SR	R^2_F (%)
		Long-short			
Random	Stock+ fund+ sentiment	0.41	4.5***	0.21	5.00
	Fund+ sentiment	0.40	5.4***	0.25	2.73
	Stock	-0.02	-0.2	-0.01	-1.60
Chronological	Stock+ fund+ sentiment	0.52	5.7***	0.26	2.66
	Fund+ sentiment	0.39	5.2***	0.24	1.49
	Stock	0.18	2.2**	0.10	-1.38
Expanding	Stock+ fund+ sentiment	0.45	5.4***	0.29	8.92
	Fund+ sentiment	0.40	5.8***	0.31	5.71
	Stock	0.14	1.7*	0.09	-0.47

This table reports the Sharpe ratio, mean and factor R^2 of the top-minus-bottom prediction-weighted decile portfolios that use different information sets for the prediction. We consider three different information sets. We compare three different sampling schemes: three random cross-out-of-sample folds, three cross-out-of-sample folds that keep the chronological order, and estimation on an expanding window. The out-of-sample analysis for the expanding window starts in 1990.

4 Understanding the Results

4.1 Variable Importance and Interaction Effects

Importance of Predictors In order to visualize which variables are the most important for prediction, we construct a metric based on the average squared gradient of the abnormal return prediction for each characteristic, following Sadhwani, Giesecke, and Sirignano (2020) and Horel and Giesecke (2020):

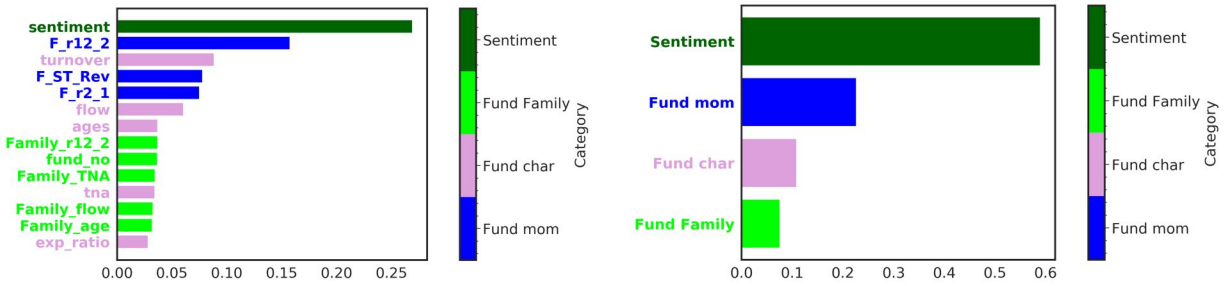
$$\text{Sensitivity}(z_k) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(\frac{\partial \hat{R}^{abn}_{i,t+1}}{\partial z_{i,k,t}}\right)^2} \tag{7}$$

The partial derivatives are evaluated at the observed characteristics and are approximated with numerical derivatives. T is the number of periods and N_t is the number of funds available at time t. The $Sensitivity(z_k)$ is then averaged over the three cross-out-of-sample folds and normalized to sum up to 1. The partial derivative simplifies to the standard slope coefficient in the special case of a linear regression framework. A larger sensitivity means that a variable has a larger effect on the neural network prediction.[18] In addition to the point estimates, we can provide formal statistical tests for the significance of $Sensitivity(z_k)$. We apply the theory developed by Horel and Giesecke (2020) to test the null hypothesis that the sensitivity measure is statistically different from zero.

[18] An alternative sensitivity measure is based on the average absolute values of the partial derivatives. The relative importance results are very similar to the average squared gradients, but the measure in Equation 7 has the advantage that we can provide a valid asymptotic inference.

This corresponds to a generalization of t-statistics in linear regressions.[19] The test enables model-free inference for the importance of variables. Appendix C provides the technical details on how we extend their model.

Figure 12: Top variable importance for explaining abnormal returns.



(a) Top variable importance.

(b) Top variable group importance.

This figure shows the importance ranking for individual variables and variable groups. The ranking is square root of the average of the squared gradient for the eight ensemble fits. The variable importance measures are evaluated on the test data and averaged across three cross-out-of-sample folds. Fund-specific characteristics and sentiment are used as network input.

The left panel of Figure 12 shows the sensitivities for the neural network model with fund-level information and sentiment. Sentiment is the most important variable, followed by fund momentum, turnover, fund reversal, and flow. In the right panel, we define the variable importance measure of a group by taking the average of the sensitivity measures within that group. The most important fund-specific characteristics group is fund momentum characteristics.

Table 6 reports the level and statistical significance of the measure for $Sensitivity(z_k)$ in the first column of each panel. We do not normalize the measures to add up to one. Hence, the values represent the predicted average change in returns. We confirm that sentiment and most fund-level characteristics are highly statistically significant on a 1% level. This is true both for our benchmark sampling scheme in the left panel and chronological sampling in the right panel. One difference is that sentiment is quantitatively somewhat less important in the chronological sampling. The random sampling approach allows the model to better learn the non-linear interaction between sentiment and fund-level variables, as we discuss next.

Interaction Effects We now analyze the interactions between sentiment and fund characteristics that are implied by the neural network model. Figure 13 plots the predicted abnormal fund return (on the y-axis) as a function of one fund-level variable (on the x-axis), keeping all the other variables at their median level. The function is averaged over three cross-out-of-sample folds. In order

[19] Horel and Giesecke (2020) study the large sample asymptotic behavior of gradient based variable importance measures. They view the neural network estimator as a nonparametric sieve estimator and show under mild assumptions that it converges to the argmax of a Gaussian process with mean zero and a specific covariance function. Then, they use a functional delta method to derive the distribution of gradient based test statistics.

Table 6: Statistical significance of variable importance and interaction effects

Random sampling			Chronological sampling		
Fund characteristics	Sensitivity	Interaction	Fund characteristics	Sensitivity	Interaction
sentiment	0.14***		F_r12_2	0.09***	0.06***
F_r12_2	0.08***	0.09***	sentiment	0.08***	
turnover	0.05***	0.06***	F_ST_Rev	0.06***	0.02***
F_ST_Rev	0.04***	0.04***	F_r2_1	0.03***	-0.00
F_r2_1	0.04***	-0.03***	flow	0.03***	0.02***
flow	0.03***	0.03***	turnover	0.03***	0.01
ages	0.02***	0.02***	ages	0.02***	0.01
fund_no	0.02***	-0.01**	Family_r12_2	0.02**	0.01
tna	0.02***	0.01**	tna	0.02***	0.01
Family_r12_2	0.02***	0.01	fund_no	0.02***	0.00
Family_flow	0.02**	0.01***	exp_ratio	0.02***	0.00
Family_TNA	0.02**	0.00	Family_age	0.01**	0.00
Family_age	0.02*	-0.01	Family_flow	0.01	0.01
exp_ratio	0.01	0.01	Family_TNA	0.01	0.00

This table reports the magnitude and significance of the measures for $Sensitivity(z_k)$ and $Interaction(z_k, \text{sentiment})$. Both measures are reported in percentages. Fund-specific characteristics and sentiment are used as network input. The significance levels are given by $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. The left panel reports the results for random sampling, while the right panel shows the corresponding results for chronological sampling.

to study the interaction effects with sentiment, we plot this one-dimensional function for different quantiles of the sentiment distribution. Hence, the plots show the mean of abnormal fund returns conditional on the values of one fund variable and sentiment.

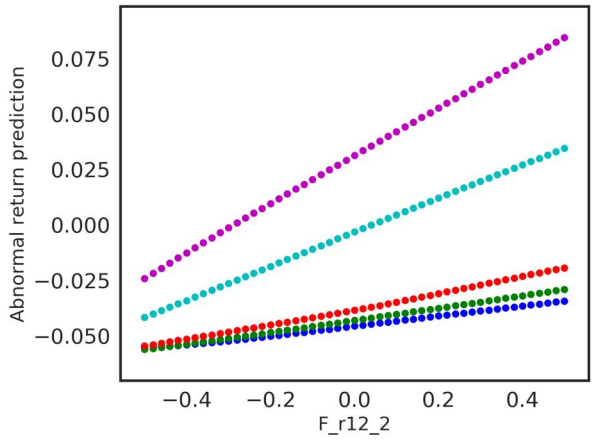
There are clear interaction effects between sentiment and fund-level variables. Predicted abnormal returns are almost linear in fund-specific variables, but the slope of that relationship is substantially higher in times of high sentiment. Note that without interaction effects between sentiment and the flow variable, the different curves in each panel would be parallel shifts. They clearly are not. The interaction effect with sentiment is particularly strong for fund momentum in panel (a). In contrast, there is no interaction effect for family momentum.

It turns out the interaction effects of sentiment with fund-level variables is monotonic for all our variables. In order to assess the economic magnitude and the relative importance, we introduce the following new interaction measure, which measures the differences in slopes for high and low macroeconomic states. While we consider only sentiment as macroeconomic variable in this section, we extend this measure to other macroeconomic states in the next section.

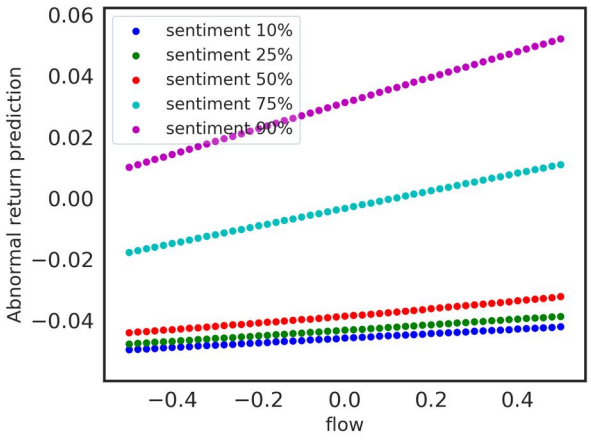
$$\begin{aligned}\text{Interaction}(z, \text{macro}) = &\left(\hat{R}^{abn}(\text{high } z, \text{high macro}) - \hat{R}^{abn}(\text{low } z, \text{high macro})\right) \\ &- \left(\hat{R}^{abn}(\text{high } z, \text{low macro}) - \hat{R}^{abn}(\text{low } z, \text{low macro})\right).\end{aligned}$$

We evaluate the predicted abnormal return $\hat{R}^{abn}$ for the highest and lowest value of the fund

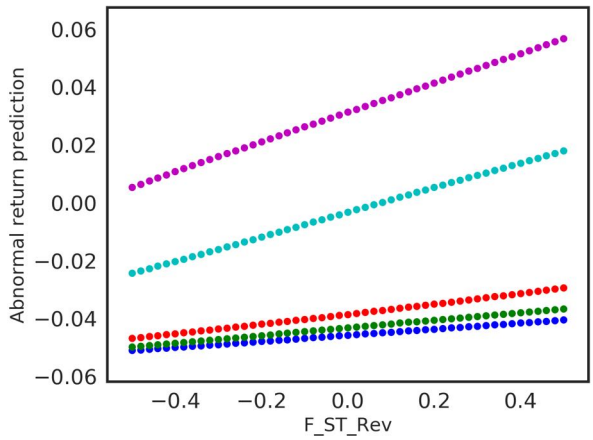
Figure 13: Conditional mean as a function of fund characteristics and sentiment.



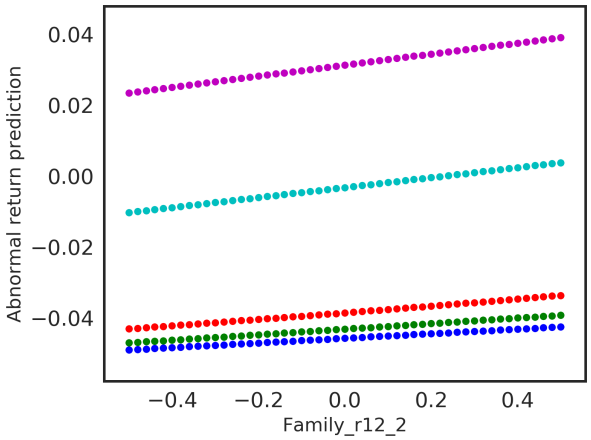
(a) Predicted abnormal returns as function of *F_r12_2*



(b) Predicted abnormal returns as function of *flow*



(c) Predicted abnormal returns as function of *F_ST_Rev*



(d) Predicted abnormal returns as function of *Family_r12_2*

This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic conditional on different sentiment quantiles. The other variables are set to their median. The neural network model is estimated with fund-specific characteristics and sentiment. The interaction effects are evaluated on the test data and averaged across three cross-out-of-sample folds. The high-minus-low portfolios have a higher mean conditional on high past sentiment. This is a non-linear interaction effect.

variable z and the high (90% quantile) and the low (10% quantile) macroeconomic state. The other variables are set to their median values. A high absolute value in this measure indicates a strong interaction effect and measures the difference in the return spread of characteristic z in high and low sentiment states. In addition to introducing this new measure, Appendix C extends the statistical distribution results of Horel and Giesecke (2020) to test the statistical significance of interaction effects.

Table 6 reports the interaction measure for sentiment with the fund-level characteristics. Return spreads due to fund momentum, turnover, flow and reversal are the most affected by sentiment. The table shows that the predicted monthly spread in fund momentum is nine basis points higher in high sentiment states compared to low sentiment states. The large interaction effects

that we observe are statistically significant.

Why is our neural network structure able to generate such an interaction effect between sentiment and fund-level characteristics? The hidden layer of the neural network performs a nonlinear transformation of original characteristics into new characteristics: $z(1) = ReLU(\sum_{k=1}^{K} w_k^{(0)} z_k^{(0)} + w_0^{(0)})$. There are some hidden-layer neurons that get activated only when sentiment is high (z_t is large), which changes the dependency of the abnormal return prediction on fund-level characteristics. When the neuron gets activated, the slope of this dependency gets higher, which is exactly what we see in Figure 13. While the interaction effects with sentiment are the strongest, there are also interaction effects between the fund specific variables as shown in Appendix A.1.7.

The interaction effects are weaker for chronological than random sampling because chronological sampling hurts the model's ability to learn the non-linear relationship between fund variables and sentiment, given that high-sentiment periods are clustered chronologically. Appendices A.2 and A.3 discuss sensitivity and interaction effects for chronological and expanding-window sampling.

4.2 Which Macro-economic Variable?

Having shown the importance of sentiment and its interaction effects with fund characteristics, it is reasonable to ask whether other variables like CFNAI might play a similarly important role in predicting mutual fund out-performance. Or maybe they add a useful piece of macro-economic information that is not contained in sentiment? Appendix A.1.6 answers these questions in great detail. The main finding is that replacing sentiment with CFNAI results does not affect much which mutual funds are sorted bottom and top predicted performance deciles. However, the relative performance prediction within the top and bottom deciles is weaker with CFNAI than with sentiment. The reason can be traced back to the much weaker interaction effects between fund variables and CFNAI.

4.3 A Parsimonious Model

In order to illustrate a more interpretable model, we estimate a simplified model that uses only flow, F_r12_2, and sentiment as inputs and present its functional form. Previously, we showed the results for the information set flow + fund momentum + sentiment, where fund momentum refers to the fund momentum group consisting of F_r2_1, F_r12_2, F_ST_Rev. Table 7 shows that these three variables already capture a large fraction of the predictability. However, the predictability is weaker than for a model that includes all fund momentum characteristics, which illustrates the benefit of F_r2_1 and F_ST_Rev.

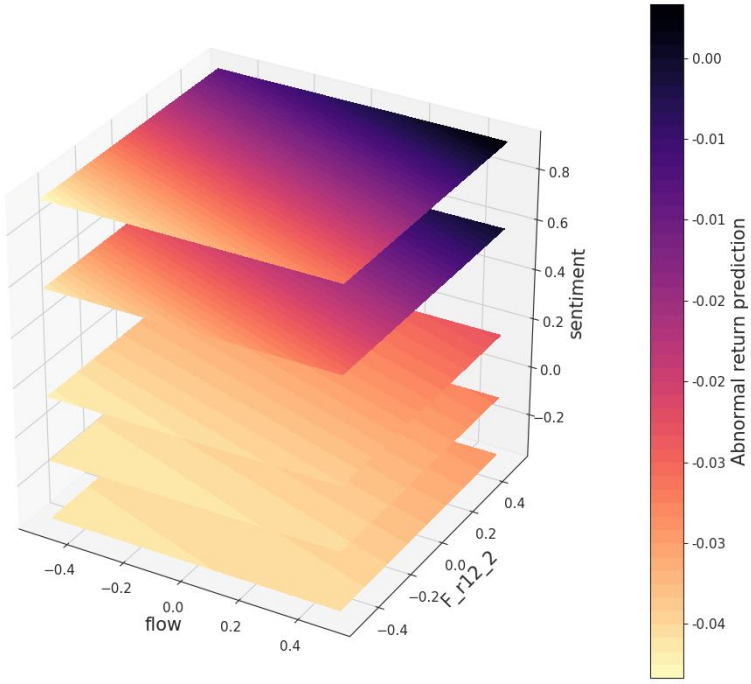
This simple model with only three variables can be easily visualized and interpreted. Figure 14 shows non-trivial interaction effects between the three variables. The interaction effects with

Table 7: Performance of abnormal return portfolios conditioned on flow + F_r12_2 + sentiment.

Decile	mean(%)	t-stat	SR	R^2_F (%)
Long-short	0.40	5.4***	0.25	0.70
Top	0.17	3.4***	0.16	-0.73
Bottom	-0.23	-3.6***	-0.21	0.82

This table reports the Sharpe ratio, mean and factor R^2 of long-short prediction-weighted, and top and bottom decile portfolios that use only flow, F_r12_2, and sentiment as information set.

Figure 14: Conditional abnormal returns as a function of F_r12_2, flow and sentiment



This figure shows the conditional abnormal return as a function of F_r12_2, flow and sentiment. The neural network only uses F_r12_2, flow, and sentiment as input. The results are shown in percentages.

sentiment are particularly strong. The highest conditional abnormal return occurs during high sentiment periods for funds with high momentum and high flow. The lowest abnormal returns are predicted for low sentiment periods and funds with low momentum and low flow. These non-linear interaction effects cannot be captured by a linear model. A simple ex-post regression of abnormal returns on sentiment or high-sentiment indicators is also not sufficient to detect these interaction effects, since the key point is that sentiment must be included in the prediction model itself.

Gruber (1996) and Zheng (1999) identified a positive, but fairly short-lived and weak relationship between flows and subsequent fund abnormal returns for small but not for large funds. Importantly, Sapp and Tiwari (2004) show that this "smart money" effect is explained away once the risk adjustment controls for stock return momentum. Lou (2012) shows that the expected part

of flow-induced trading positively forecasts mutual fund returns, while Song (2020) finds that fund flows associated with positive factor returns lead to negative future fund performance. Our machine learning approach revives the predictive role of flows, with a 4-factor risk-adjustment, and shows that fund flow predicts performance positively and persistently. It also uncovers that fund abnormal return momentum strongly and positively predicts fund abnormal return. Both predictive relationships are stronger in high-sentiment periods.

These results are consistent with theories where at least some managers are skilled and at least some investors can detect skill and (re)allocate their investment towards skilled managers. There are reasons to believe that this reallocation of investment flows may not be as strong and quick as the frictionless model of Berk and Green (2004) predicts because of transaction or broader search costs (Roussanov, Ruan, and Wei, 2021), investor inertia, or weak transmission from investor beliefs to allocation decisions (Giglio, Maggiori, Stroebel, and Utkus, 2021). With imperfect reallocation, skill leaves a trail in the form of fund return momentum for investors to exploit in the next period. Put differently, the flows are gradual and small enough that it takes several periods until the fund runs into zero marginal abnormal returns. The results are potentially also consistent with funds and fund families attracting flows through marketing rather than—or in addition to—through investment skill (Ibert, Kaniel, Van Nieuwerburgh, and Vestman, 2018; Roussanov, Ruan, and Wei, 2021). Marketing-induced inflows create buying pressure for stocks that the fund typically invests in. In a world with downward-sloping demand curves (Coval and Stafford, 2007; Koijen and Yogo, 2019; Gabaix and Koijen, 2021), this raises prices and lifts fund returns. Through the flow-performance relationship, as well as through persistence in marketing-driven flows, the out-performance creates more inflows in the next period. The demand pressure increases prices further, generating momentum in fund returns. The fact that flows and fund momentum have a much stronger association with fund performance in high-sentiment periods lends further credence to this marketing-driven channel.

4.4 Decomposing Abnormal Returns

Having established that fund characteristics, and in particular fund momentum and flow, and their interaction with sentiment are key inputs for predicting mutual fund abnormal returns, we now try to understand in more detail the mechanisms behind this prediction. For simplicity, we do so in a univariate setting.[20] We decompose the mutual funds' abnormal return into a component that reflects trading between disclosure dates (between quarter ends) and a component that reflects trading within a disclosure period (within quarter).

[20] For a full set of univariate predictability results, see Appendix IA.2.

Table 8: Decomposition of univariate long-short abnormal return factors

	Total		Between-disclosure		Within-disclosure		Risk difference		Return gap	
	SR	mean	SR	mean	SR	mean	SR	mean	SR	mean
F_r12_2	0.28	0.36***	0.14	0.20***	0.20	0.17***	0.14	0.11***	0.10	0.06***
F_ST_Rev	0.20	0.30***	0.14	0.16***	0.15	0.15***	0.12	0.08**	0.12	0.06***
Family_r12_2	0.19	0.13***	0.10	0.09***	0.09	0.04**	0.12	0.07**	-0.06	-0.03
Beta	0.15	0.18***	0.12	0.16***	0.03	0.03	-0.01	-0.00	0.05	0.03
Rel2High	0.14	0.20***	0.13	0.25***	-0.05	-0.05	-0.03	-0.03	-0.03	-0.03
RNA	0.13	0.13***	0.11	0.12***	0.01	0.01	-0.03	-0.02	0.04	0.02
Family_TNA	0.13	0.09***	0.09	0.07	0.05	0.03	-0.12	-0.06**	0.16	0.08***
fund_no	0.13	0.10***	0.10	0.07**	0.06	0.03	-0.12	-0.05**	0.14	0.07***
flow	0.12	0.11**	0.08	0.08**	0.06	0.03	-0.00	-0.00	0.08	0.03**
Family_age	0.11	0.09**	0.08	0.07	0.03	0.02	-0.13	-0.06**	0.13	0.08***
ROA	0.10	0.10**	0.11	0.13***	-0.03	-0.03	-0.05	-0.03	0.01	0.01
PM	0.10	0.10**	0.10	0.11**	-0.01	-0.01	-0.03	-0.02	0.02	0.01
ROE	0.10	0.11**	0.09	0.12**	-0.01	-0.01	-0.02	-0.01	0.00	0.00
ST_Rev	0.09	0.13**	0.06	0.11	0.02	0.02	0.02	0.02	0.01	0.01
CF	0.09	0.09**	0.11	0.16**	-0.07	-0.06**	-0.06	-0.04	-0.04	-0.03

This table reports mean and Sharpe ratio of the decomposition of univariate long-short abnormal return factors. Means of abnormal returns are reported in percentages. The results are sorted according to the Sharpe ratio of the long-short factors and show the first 15 factors. The full results are in Table IA.11. For each of the 59 characteristics and each abnormal return, we construct decile-sorted portfolios. The long-short factors are the differences between the top decile and the bottom decile. Stars denote the significance levels.

$$R^{abn}_{i,t} = \underbrace{\tilde{R}_{i,t} - f_t\tilde{\beta}_i}_{\text{Between-disclosure abnormal return}} + \underbrace{R_{i,t} - f_t\beta_i - (\tilde{R}_{i,t} - f_t\tilde{\beta}_i)}_{\text{Within-disclosure abnormal return}} \tag{8}$$

$$= \underbrace{\tilde{R}_{i,t} - f_t\tilde{\beta}_i}_{\text{Between-disclosure abnormal return}} + \underbrace{R_{i,t} - \tilde{R}_{i,t}}_{\text{Return gap}} + \underbrace{f_t(\tilde{\beta}_i - \beta_i)}_{\text{Risk exposure difference}} \tag{9}$$

The between-disclosure abnormal return is the abnormal return of an investor who invests in the most recently disclosed stock positions of a fund and holds that portfolio until next fund disclosure. In the equation above, $\tilde{R}_{i,t}$ is the hypothetical return of a mutual fund i that keeps its portfolio weights fixed at their last-disclosed levels (at $t-1$), f_t is the contemporaneous return vector on the Carhart factors, and $\tilde{\beta}_i$ is the vector of exposures to the Carhart factors associated with this hypothetical fund return $\tilde{R}_{i,t}$.[21] A positive average between-disclosure abnormal return means that the mutual fund can pick stocks with positive alpha at quarterly frequency.

A high value of within-disclosure abnormal returns indicates that the fund is adding value by actively trading between adjacent disclosure dates. The within-disclosure abnormal return can be decomposed further into two parts: the return gap, as defined in Kacperczyk, Sialm, and Zheng

[21] The exposures $\tilde{\beta}_i$ are estimated from a regression of $\tilde{R}_{i,t-h}$ on the Carhart factors in the previous 36 months ($h = 1, \cdots, 36$), where $\tilde{R}_{i,t-h} = \sum_j w_{i,j,t} R_{j,t-h}$. That is, $\tilde{R}_{i,t-h}$ is the return on a portfolio that holds the identity of the stocks j and their portfolio weights $w_{i,j,t}$ fixed at the last-disclosed period t for every h.

Table 9: Decomposition of prediction long-short abnormal return portfolios

	Total		Between-disclosure		Within-disclosure		Risk difference		Return gap	
	SR	mean	SR	mean	SR	mean	SR	mean	SR	mean
Stock+ fund+ sentiment	0.21	0.41***	0.10	0.28**	0.13	0.13***	0.07	0.06	0.09	0.06**
Fund+ sentiment	0.25	0.40***	0.15	0.24***	0.16	0.16***	0.16	0.13***	0.03	0.03
Fund	0.25	0.38***	0.15	0.20***	0.17	0.18***	0.15	0.12***	0.08	0.06**
Stock+ fund	0.15	0.28***	0.05	0.13	0.14	0.15***	0.06	0.06	0.11	0.09***
Stock+ sentiment	0.07	0.15	0.04	0.12	0.02	0.02	0.00	0.00	0.02	0.02
Stock	-0.01	-0.02	-0.01	-0.03	0.01	0.01	-0.01	-0.01	0.03	0.02
Flow+fund momentum+ sentiment	0.24	0.48***	0.14	0.26***	0.17	0.22***	0.13	0.12***	0.11	0.10***

This table reports the mean and Sharpe ratio for the decomposition of the prediction-weighted long-short decile portfolios. We use different information sets to predict abnormal returns. Means of abnormal returns are reported in percentages. Stars denote the significance levels.

(2008), and the risk exposure differential, which is the difference between the risk exposure of the hypothetical fixed portfolio from the latest stock holding disclosure and the real risk exposure from the current (since rebalanced) portfolio.

We ask which characteristics best predict each of these three components of the fund abnormal return. Table 8 shows the results for the three-way decomposition. Columns 2 and 3 report the Sharpe Ratio and mean return associated with an investment that goes long the 10% best funds and short the 10% worst funds based on a univariate prediction using the variable listed in the first column.[22] The next four sets of two columns predict one of the components of the abnormal fund return.

Momentum characteristics in the first three rows of the table are the most important characteristics for both between-disclosure and within-disclosure abnormal returns, with each component of returns accounting for about half of the return. Flow, the number of funds in the family as well as a few stock-specific characteristics are also useful for predicting between-disclosure returns, while these momentum characteristics are the only significant predictors of within-disclosure abnormal returns.[23] When it comes to understanding within-disclosure returns, we find that fund momentum and reversal are the only characteristics that predict both the return gap and the risk difference with the same sign. Flows predict the return gap as well. Other fund and fund-family variables predict the return gap significantly, but this effect is offset by an opposite-sign prediction for the risk difference. That is, while funds with these characteristics are trading in a way that increases the fund's return, they do so by taking on more systematic risk. Funds with high fund momentum and reversal characteristics, in contrast, trade within the quarter in ways that both increase the return gap and reduce the systematic risk of the portfolio significantly.

These univariate insights carry over to the neural network prediction model as shown in Table 9. The decomposition is a complex average of the univariate results. A prediction model that is

[22]The rows are ranked by Sharpe Ratio in the column, from highest abnormal return predictive SR to lowest. For brevity, we only report the first 15 rows of this table; the full set of results appears in Table IA.11.

[23]An exception being CF that is a significant negative predictor of within and positive predictor of between.

only based on fund momentum, flow, and sentiment (third row of the table) has the strongest within-disclosure effects, which is driven in equal part by the significant positive risk difference and return gap. Adding more fund characteristics, lowers both the within-disclosure and between-disclosure mean returns.

4.5 Abnormal versus Total Return Prediction

One of our key findings is that stock characteristics contribute little to the prediction of best and worst funds. This may be a surprising result, and it may appear to contradict the findings in Li and Rossi (2021) who emphasize that one can predict the best and worst funds based on the stocks that they hold. We explain why there is no contradiction. Our paper predicts fund abnormal returns, $R^{abn}_{i,t}$. Fund total returns $R_{i,t}$ have a strong common component, due to fund exposures to common return factors F_t. Figure A.15 and Table A.10 in the Appendix report the results for predicting the *total* returns of mutual funds rather than *abnormal* returns. First, we find that stock characteristics are substantially more predictive for total fund returns than for their abnormal returns. In other words, the stock characteristics seem to be able to predict the systematic factor component in fund returns, consistent with Li and Rossi (2021).[24] However, as we have established above, once this factor component is taken out, stock characteristics lose most of their predictability. Second, the Sharpe ratio of long-short portfolios based on total return prediction is lower than from predicting abnormal returns. This points to an important methodological contribution of this paper. The level of fund returns (and also stock returns) is extremely hard to predict, while the relative performance is more predictable. Abnormal returns are a relative prediction objective as they remove the level effect arising from compensation for systematic risk factor exposures. Third, the comparison between returns and abnormal returns also illustrates the difference between conditional and unconditional factor models. Predicting total returns with a ML model and subsequently estimating an unconditional Carhart 4-factor model on the prediction portfolio returns is fundamentally different from first constructing abnormal returns from a conditional Carhart 4-factor model and subsequently predicting abnormal returns with a ML model. Appendix A.1.9 provides a detailed discussion.

4.6 Time-Variation in Performance

The predictability of the performance of mutual fund managers seems to be time-varying. First, the predictability with stock characteristics only sharply declines after the year 2000 as shown in Figure 8. In contrast, the performance of the strategy that only uses fund-specific information continues to do reasonably well post-2000. Second, Figures 5 and 7 show that the performance of the top deciles based on fund-specific information and sentiment declines post 2000. A large part of

[24] As an aside, this result also suggests that the list of stock-specific characteristics we use is not driving our results.

the performance of the long-short strategy can be attributed to predicting the bottom decile. Third, Figure 9 indicates that performance for the top-decile deteriorates less post-2000 when funds are value-weighted. This suggests that the outperformance of the most skilled fund managers at the largest funds is more consistent over time, and stronger than for the most skilled managers at the smaller funds.

This time-variation in performance is not explained by turnover and expense ratios. Figure A.16 shows that turnover with consequently higher transactions costs does not systematically increase after 2000. Other changes might provide an explanation. Between late 2000 and early 2003 a set of regulations were enacted that had important implications for the information collection environment, transparency in securities markets, timely disclosure of information by public firms, and reduction in trading frictions.[25] These changes likely negatively impacted mutual funds' ability to generate abnormal returns both in terms of reducing information collection advantages, and in facilitating easier and cheaper entry of arbitrageurs competing with mutual funds in taking advantage of potential price anomalies. It seems possible that the largest mutual funds have better ways of managing compliance with all these new rules, which could at least partly help to explain the value-weighted results.

Relatedly, Hanson and Sunderam (2014) show that the amount of arbitrage capital devoted to familiar quantitative equity strategies, such as value and momentum, have grown dramatically since the early 2000s. This influx of capital resulted in lower strategy returns, whose signals decay more rapidly following portfolio formation. Akbas, Ay, and Koch (2023) document that markets are more efficient in responding to both firm-specific and market-wide news from the early 2000s in comparison to the period between 1980 and 2000, and that the increased efficiency is related to the capacity of arbitrageurs to update prices, as proxied by the size and skill level of the finance industry. Finally, Green, Hand, and Zhang (2017) find a marked shift in characteristics-based predictability in 2003 or slightly earlier, which is consistent with the collapse in predictability of stock-specific characteristics for abnormal returns.

5 Conclusion

In this paper, we revisit the question of predicting actively-managed mutual fund performance. While predictability has been difficult to establish thus far, using modern neural network techniques we find strong evidence of abnormal return predictability. An important advantage of

[25]The new regulations included: Reg FD requiring that when a firm discloses material nonpublic information it must also make public disclosure of that information, decimalization of quotes, Sarbanes-Oxley enacting enhanced transparency and accountability in financial reporting and internal auditing mechanisms, acceleration of deadlines for filing with the SEC annual and quarterly reports to make 10-Q and 10-K filings more timely, and NYSE's introduction of its auto-quoting software that led to dramatic reductions in trading frictions and costs and an equally dramatic increase in algorithmic trading by hedge funds.

non-linear neural network methods is that they can reliably estimate a complex functional relationship among a large set of variables. This turns out especially advantageous in predicting abnormal returns of mutual funds. The predictability we identify is out-of-sample, long-lived, and economically meaningful. It holds both before and after fees. Most of the benefits accrue from avoiding funds that the model predicts to be the worst performers. However, the prediction model is also able to identify about 10-20% of funds that generate positive abnormal returns even after fees. The predictability lasts for at least 36 months.

We identify two fund characteristics, fund flow and fund momentum, as key predictors of mutual fund out-performance. Characteristics of the stocks that funds hold do not play a significant role in predicting abnormal returns. Moreover, these two fund characteristics matter much more when investor sentiment is high. A linear model would fail to pick up this important interaction effect. While including CFNAI, a proxy for macro-economic activity, also improves predictability, there are no discernible interaction effects associated with CFNAI unlike with sentiment. These results should prove useful for improving theories of delegation in the mutual fund market.

Methodologically, we show that abnormal returns, obtained as local residuals to a factor model, are not only an economically motivated, but also the statistically better target for prediction. We demonstrate how to measure dependencies on macro-economic states. We suggest that instead of the typical horse race of model specifications it may be better to compare the prediction and trading benefits by varying the information set available to the same flexible machine learning algorithm. Finally, we introduce a novel measure for interaction effects in machine learning algorithms, which does not only measure a local slope, but a more informative global slope. For this interpretable measure, we provide a formal statistical significance test. These methodological contributions will help advance future asset pricing and investment research using machine learning, a growing area of research.

This paper focused on actively-managed equity mutual funds. Natural next steps are to study bond mutual funds, as well as portfolios managed by hedge funds, pension funds, and endowments, in an effort to uncover both the presence and the drivers of skill in other asset classes and types of institutions.

References

AKBAS, F., L. AY, AND P. KOCH (2023): "The Evolution of Market Efficiency Over the Past Century," *Available at SSRN 4373735*.

AMIHUD, Y., AND G. RUSLAN (2013): "Mutual Fund's R2 as Predictor of Performance," *Review of Financial Studies*, 26(3), 667–694.

BAKER, M., AND J. WURGLER (2006): "Investor sentiment and the cross-section of stock returns," *Journal of Finance*, 61(4), 1645–1680.

BARBER, B. M., X. HUANG, AND T. ODEAN (2016): "Which factors matter to investors? Evidence from mutual fund flows," *Review of Financial Studies*, 29(10), 2600–2642.

BEN-DAVID, I., J. LI, A. ROSSI, AND Y. SONG (2022): "What do mutual fund investors really care about?," *The Review of Financial Studies*, 35(4), 1723–1774.

BERK, J. B., AND R. C. GREEN (2004): "Mutual fund flows and performance in rational markets," *Journal of Political Economy*, 112(6), 1269–1295.

BERK, J. B., AND J. H. VAN BINSBERGEN (2015): "Measuring skill in the mutual fund industry," *Journal of Financial Economics*, 118(1), 1–20.

——— (2016): "Assessing asset pricing models using revealed preference," *Journal of Financial Economics*, 119(1), 1–23.

BIANCHI, D., M. BÜCHNER, T. HOOGTEIJLING, AND A. TAMONI (2021): "Corrigendum: Bond risk premiums with machine learning," *Review of Financial Studies*, 34(2), 1090–1103.

BIANCHI, D., M. BÜCHNER, AND A. TAMONI (2021): "Bond risk premiums with machine learning," *Review of Financial Studies*, 34(2), 1046–1089.

BOLLEN, N., AND B. JEFFREY (2005): "Short-Term Persistence in Mutual Fund Performance," *Review of Financial Studies*, 18(2), 569–597.

BROWN, D. P., AND Y. WU (2016): "Mutual fund flows and cross-fund learning within families," *Journal of Finance*, 71(1), 383–424.

BRYZGALOVA, S., M. PELGER, AND J. ZHU (2021): "Forest through the Trees: Building Cross-Sections of Stock Returns," *Working paper*.

CARHART, M. M. (1997): "On persistence in mutual fund performance.," *Journal of Finance*, 52(1), 57–82.

CHEN, L., M. PELGER, AND J. ZHU (2023): "Deep learning in asset pricing," *Management Science*.

CONG, L. W., G. FENG, J. HE, AND X. HE (2022): "Asset pricing with panel trees under global split criteria," *Working paper*.

COVAL, J., AND E. STAFFORD (2007): "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 86(2), 479–512.

CREMERS, M., AND A. PETAJISTO (2009): "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *Review of Financial Studies*, 22(9), 3329–3365.

DEMIGUEL, V., J. GIL-BAZO, F. J. NOGALES, AND A. A. P. SANTOS (2023): "Machine Learning and Fund Characteristics Help to Select Mutual Funds with Positive Alpha," *Working paper*.

DOSHI, H., R. ELKAMHI, AND M. SIMUTIN (2015): "Managerial activeness and mutual fund performance," *Review of Asset Pricing Studies*, 5(2), 156–184.

FAMA, E. F., AND K. R. FRENCH (1996): "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, 51(1), 55–84.

——— (2010): "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *Journal of Finance*, 65(5), 1915–1947.

FENG, G., S. GIGLIO, AND D. XIU (2020): "Taming the factor zoo: A test of new factors," *The Journal of Finance*, 75(3), 1327–1370.

FILIPPOU, I., D. RAPACH, M. P. TAYLOR, AND G. ZHOU (2022): "Exchange Rate Prediction with Machine Learning and a Smart Carry Portfolio," *Working Paper*.

FREYBERGER, J., A. NEUHIERL, AND M. WEBER (2020): "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, 33(5), 2326–2377.

GABAIX, X., AND R. S. KOIJEN (2021): "In search of the origins of financial fluctuations: The inelastic markets hypothesis," *Working Paper*.

GALLAHER, S. T., R. KANIEL, AND L. T. STARKS (2009): "Advertising and Mutual Funds: From Families to Individual Funds," *Working Paper*.

GIGLIO, S., M. MAGGIORI, J. STROEBEL, AND S. UTKUS (2021): "Five facts about beliefs and portfolios," *American Economic Review*, 111(5), 1481–1522.

GOODFELLOW, I., Y. BENGIO, AND A. COURVILLE (2016): *Deep Learning*. MIT Press.

GREEN, J., J. R. HAND, AND X. F. ZHANG (2017): "The characteristics that provide independent information about average US monthly stock returns," *The Review of Financial Studies*, 30(12), 4389–4436.

GRUBER, M. (1996): "Another Puzzle: The Growth of Actively Managed Mutual Funds," *Journal of Finance*, 51(3), 783–810.

GU, S., B. T. KELLY, AND D. XIU (2020): "Empirical Asset Pricing Via Machine Learning," *Review of Financial Studies*, 33(5), 2223–2273.

HADDAD, V., S. KOZAK, AND S. SANTOSH (2020): "Factor timing," *Review of Financial Studies*, 33(5), 1980–2018.

HANSON, S. G., AND A. SUNDERAM (2014): "The growth and limits of arbitrage: Evidence from short interest," *The Review of Financial Studies*, 27(4), 1238–1286.

HOREL, E., AND K. GIESECKE (2020): "Towards Explainable AI: Significance Tests for Neural Networks," *Journal of Machine Learning Research, forthcoming*.

IBERT, M., R. KANIEL, S. VAN NIEUWERBURGH, AND R. VESTMAN (2018): "Are mutual fund managers paid for investment skill?," *Review of Financial Studies*, 31(2), 715–772.

JEGADEESH, N., AND C. S. MANGIPUDI (2021): "What do fund flows reveal about asset pricing models and investor sophistication?," *Review of Financial Studies*, 34(1), 108–148.

JEGADEESH, N., AND S. TITMAN (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 48(1), 65–91.

KACPERCZYK, M., C. SIALM, AND L. ZHENG (2005): "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 60(4), 1983–2011.

——— (2008): "Unobserved actions of mutual funds," *Review of Financial Studies*, 21(6), 2379–2416.

KACPERCZYK, M., S. VAN NIEUWERBURGH, AND L. VELDKAMP (2014): "Time-varying fund manager skill," *Journal of Finance*, 69(4), 1455–1484.

——— (2016): "A rational theory of mutual funds' attention allocation," *Econometrica*, 84(2), 571–626.

KAROLYI, G. A., AND S. VAN NIEUWERBURGH (2020): "New methods for the cross-section of returns," *Review of Financial Studies*, 33(5), 1879–1890.

KELLY, B. T., S. PRUITT, AND Y. SU (2019): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134(3), 501–524.

KOIJEN, R. S., AND M. YOGO (2019): "A demand system approach to asset pricing," *Journal of Political Economy*, 127(4), 1475–1515.

KOSOWSKI, R. (2011): "Do Mutual Funds Perform When It Matters Most to Investors? US Mutual Fund Performance and Risk in Recessions and Expansions," *Quarterly Journal of Finance*, 1(3), 607–664.

KOZAK, S., S. NAGEL, AND S. SANTOSH (2020): "Shrinking the Cross Section," *Journal of Financial Economics*, 135(2), 271–292.

LETTAU, M., AND M. PELGER (2020): "Factors that Fit the Time-Series and Cross-Section of Stock Returns," *Review of Financial Studies*, 33(5), 2274–2325.

LI, B., AND A. G. ROSSI (2021): "Selecting mutual funds from the stocks they hold: A machine learning approach," *Working Paper*.

LOU, D. (2012): "A Flow-Based Explanation for Return Predictability,," *Review of Financial Studies*, 25(6), 3457–3489.

MASSA, M., AND V. YADAV (2015): "Investor Sentiment and Mutual Fund Strategies," *Journal of Financial and Quantitative Analysis*, 50(4), 699–727.

MOSKOWITZ, T. J. (2000): "Mutual fund performance: an empirical decomposition into stock-picking talent, style, transactions costs, and expenses. Discussion," *Journal of Finance*, 55, 1695–1703.

ROUSSANOV, N., H. RUAN, AND Y. WEI (2021): "Marketing mutual funds," *Review of Financial Studies*, 34(6), 3045–3094.

ROUSSANOV, N. L., H. RUAN, AND Y. WEI (2022): "Mutual Fund Flows and Performance in (Imperfectly) Rational Markets?," *Working Paper*.

SADHWANI, A., K. GIESECKE, AND J. SIRIGNANO (2020): "Deep Learning for Mortgage Risk*," *Journal of Financial Econometrics*, 19(2), 313–368.

SAPP, T., AND A. TIWARI (2004): "Does Stock Return Momentum Explain the "Smart Money" Effect?," *Journal of Finance*, 59(6), 2605–2622.

SONG, Y. (2020): "The Mismatch Between Mutual Fund Scale and Skill," *The Journal of Finance*, 75(5), 2555–2589.

STAMBAUGH, R. F. (2014): "Presidential address: Investment noise and trends," *Journal of Finance*, 69(4), 1415–1453.

STAMBAUGH, R. F., J. YU, AND Y. YUAN (2012): "The short of it: Investor sentiment and anomalies," *Journal of Financial Economics*, 104(2), 288–302.

WU, W., J. CHEN, Z. YANG, AND M. L. TINDALL (2021): "A cross-sectional machine learning approach for hedge fund return prediction and selection," *Management Science*, 67(7), 4577–4601.

ZHENG, L. (1999): "Is Money Smart? A Study of Mutual Fund investors' Fund Selection Ability," *Journal of Finance*, 54, 901–933.

A Additional Empirical Results

A.1 Random Sampling Method: Additional Results

The results in the main text use the random sample split scheme. This appendix contains additional results referenced in the main text.

A.1.1 Predicted Top and Bottom Decile Returns

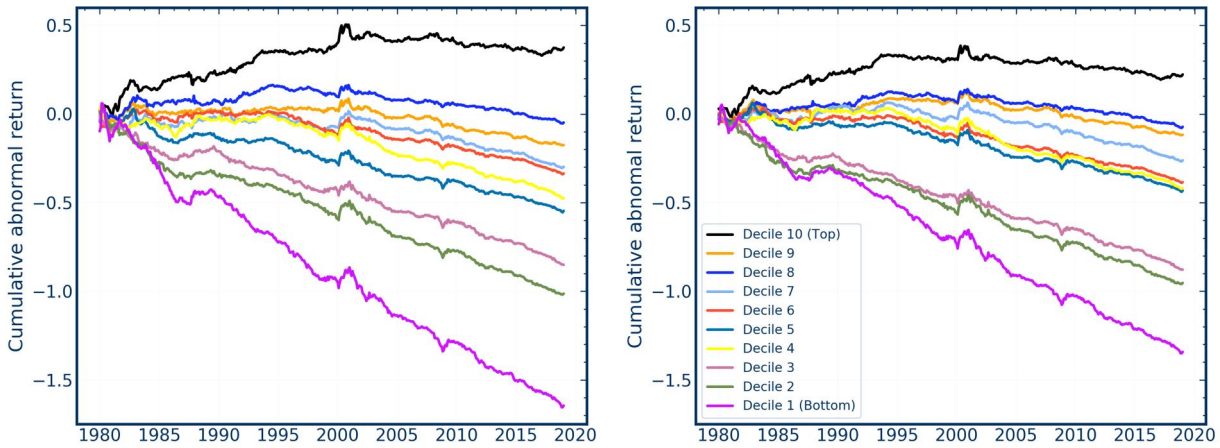
Table A.1: Performance of extreme abnormal return decile portfolios.

	Information set	mean(%)	t-stat	SR	R^2_F (%)
Top decile	Stock+ fund+ sentiment	0.15	2.9***	0.13	1.87
	Fund+ sentiment	0.17	3.5***	0.16	1.46
	Fund	0.16	3.7***	0.17	-1.20
	Stock+ fund	0.10	1.7*	0.08	-0.52
	Stock+ sentiment	0.06	1.2	0.06	0.61
	Stock	-0.02	-0.4	-0.02	-2.52
	Flow+ fund momentum+ sentiment	0.19	3.2***	0.15	-0.15
	Fund exclude momentum and flow	-0.01	-0.2	-0.01	-0.17
	F_r12_2+ sentiment	0.12	2.0**	0.09	-0.58
Bottom decile	Stock+ fund+ sentiment	-0.25	-3.5***	-0.22	1.99
	Fund+ sentiment	-0.23	-3.8***	-0.23	1.38
	Fund	-0.22	-3.7***	-0.23	0.74
	Stock+ fund	-0.19	-2.6***	-0.15	1.33
	Stock+ sentiment	-0.09	-1.2	-0.08	-0.03
	Stock	-0.00	-0.0	-0.00	-0.82
	Flow+ fund momentum+ sentiment	-0.29	-4.2***	-0.23	1.05
	Fund exclude momentum and flow	-0.07	-1.8*	-0.09	-0.32
	F_r12_2+ sentiment	-0.23	-3.8***	-0.18	0.88

This table reports the Sharpe ratio, mean and factor R^2 of prediction-weighted of the first and tenth decile portfolios that use different information sets for the prediction. We consider nine different information sets which combine fund-specific and stock-specific characteristics and sentiment. We also include flow and fund momentum individually.

A.1.2 After-Fee Abnormal Returns

Figure A.1: Predicting After-Fee Abnormal Returns



(a) Prediction-weighted abnormal net return

(b) Equally-weighted abnormal net return

These figures show the cumulative abnormal after-fee returns for prediction-sorted decile portfolios. We use fund-specific characteristics and sentiment to predict abnormal after-fee returns. The left panel weights funds based on their prediction, while the right panel equally weights funds within the prediction deciles.

A.1.3 Equally-Weighted Prediction Portfolios

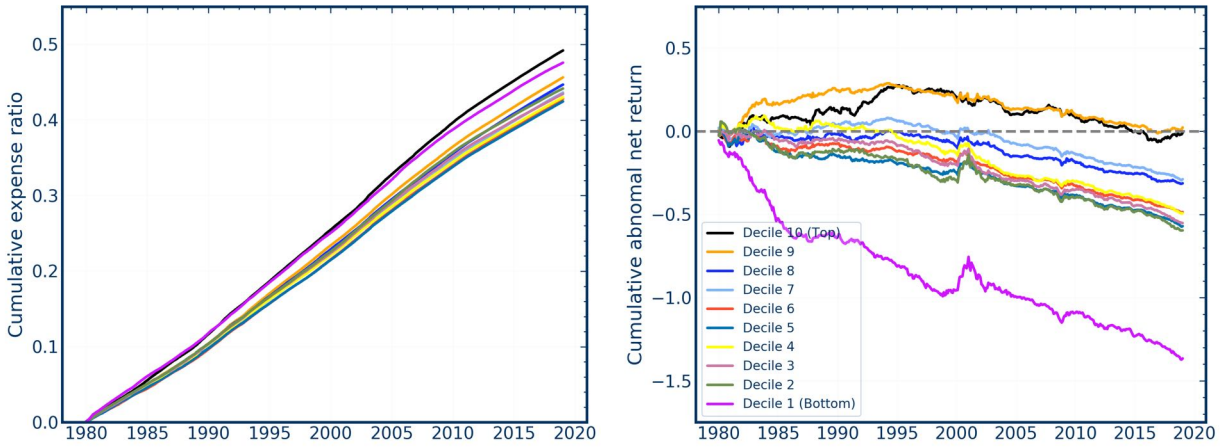
Unless explicitly mentioned, all results in the main text are for prediction-weighted returns, where mutual funds' abnormal returns in each prediction decile are weighted using the ML algorithm. Here we explore the alternative of equally-weighting within each prediction decile.

Table A.2: Performance of equally-weighted long-short abnormal return portfolios for different information sets.

Information set	mean (%)	t-stat	SR	R^2_F (%)
Stock+ fund+ sentiment	0.30	4.3***	0.20	4.47
Fund+ sentiment	0.33	5.9***	0.27	3.50
Fund	0.31	5.8***	0.27	0.24
Stock+ fund	0.21	3.1***	0.14	2.03
Stock+ sentiment	0.13	1.9*	0.09	0.18
Stock	0.01	0.1	0.01	-1.29
Flow+ fund momentum+ sentiment	0.38	5.7***	0.26	1.26
F_r12_2+ sentiment	0.38	6.0***	0.28	0.43

This table reports the Sharpe ratio, mean and R^2_F of long-short equally-weighted decile portfolios that use different information sets to predict abnormal returns. We consider eight different information sets which combine fund-specific and stock-specific characteristics and sentiment. We also include flow and fund momentum individually.

Figure A.2: Cumulative expense ratios and abnormal net return for equally-weighted deciles.



(a) Cumulative equally-weighted expense ratios

(b) Cumulative abnormal net returns

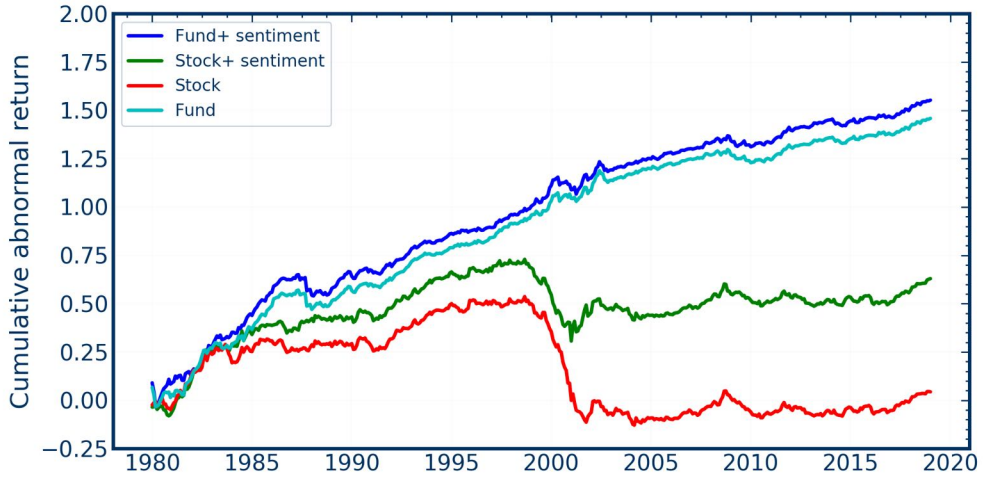
The left figure shows the cumulative expense ratios of equally-weighted prediction deciles. We use the full information set (fund-specific and stock-specific characteristics + sentiment) to predict abnormal returns before fees. The right figures shows the abnormal returns for the equally-weighted deciles after fees, that is, the abnormal returns before fees minus the fees.

Table A.3: Decomposition of equally-weighted prediction long-short abnormal return portfolios

	Total		Between-disclosure		Within-disclosure		Risk difference		Return gap	
	SR	mean	SR	mean	SR	mean	SR	mean	SR	mean
Stock+ fund+ sentiment	0.20	0.30***	0.08	0.18**	0.16	0.12***	0.10	0.08**	0.09	0.05**
Fund+ sentiment	0.27	0.33***	0.16	0.18***	0.20	0.15***	0.16	0.11***	0.07	0.04
Fund	0.27	0.31***	0.15	0.17***	0.18	0.14***	0.14	0.10***	0.07	0.04**
Stock+ fund	0.14	0.21***	0.04	0.09	0.14	0.11***	0.08	0.05	0.10	0.06**
Stock+ sentiment	0.09	0.13**	0.05	0.11	0.03	0.03	0.03	0.03	0.00	0.00
Stock	0.01	0.01	-0.01	-0.01	0.03	0.02	0.01	0.01	0.03	0.02
Flow+ fund momentum+ sentiment	0.26	0.38***	0.13	0.20***	0.19	0.18***	0.16	0.11***	0.11	0.07***

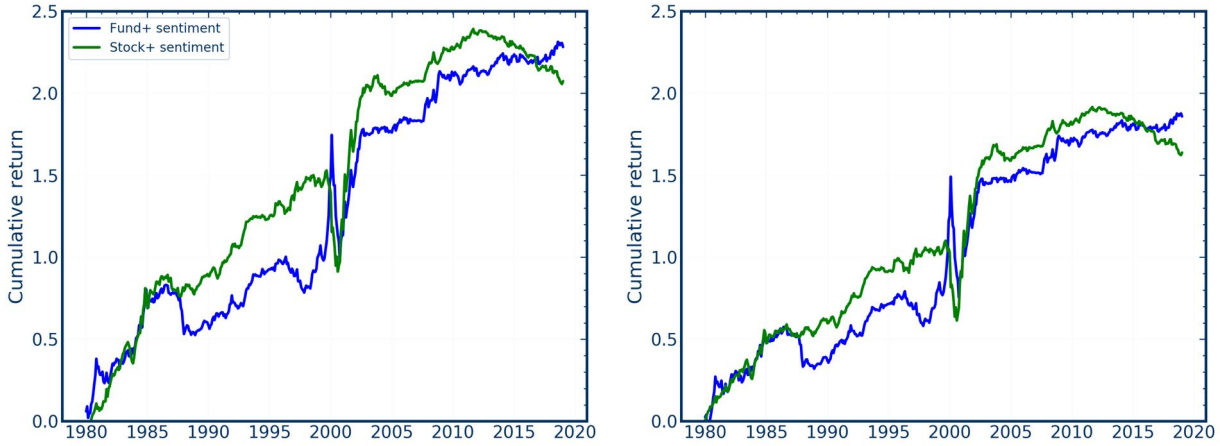
This table reports the mean and Sharpe ratio for the decomposition of equally-weighted long-short abnormal return portfolios. We use different information sets to predict abnormal returns. Means of abnormal returns are reported in percentages. The long-short portfolios are the differences between the top decile and the bottom decile. Stars denote the significance level.

Figure A.3: Cumulative abnormal returns of equally-weighted long-short prediction portfolios.



This figure plots the cumulative abnormal returns of equally-weighted long-short decile portfolios that use different information sets to predict abnormal returns. We consider fund-specific and stock-specific characteristics combined with sentiment.

Figure A.6: Cumulative total return of long-short portfolio



(a) Prediction-weighted **(b)** Equally-weighted

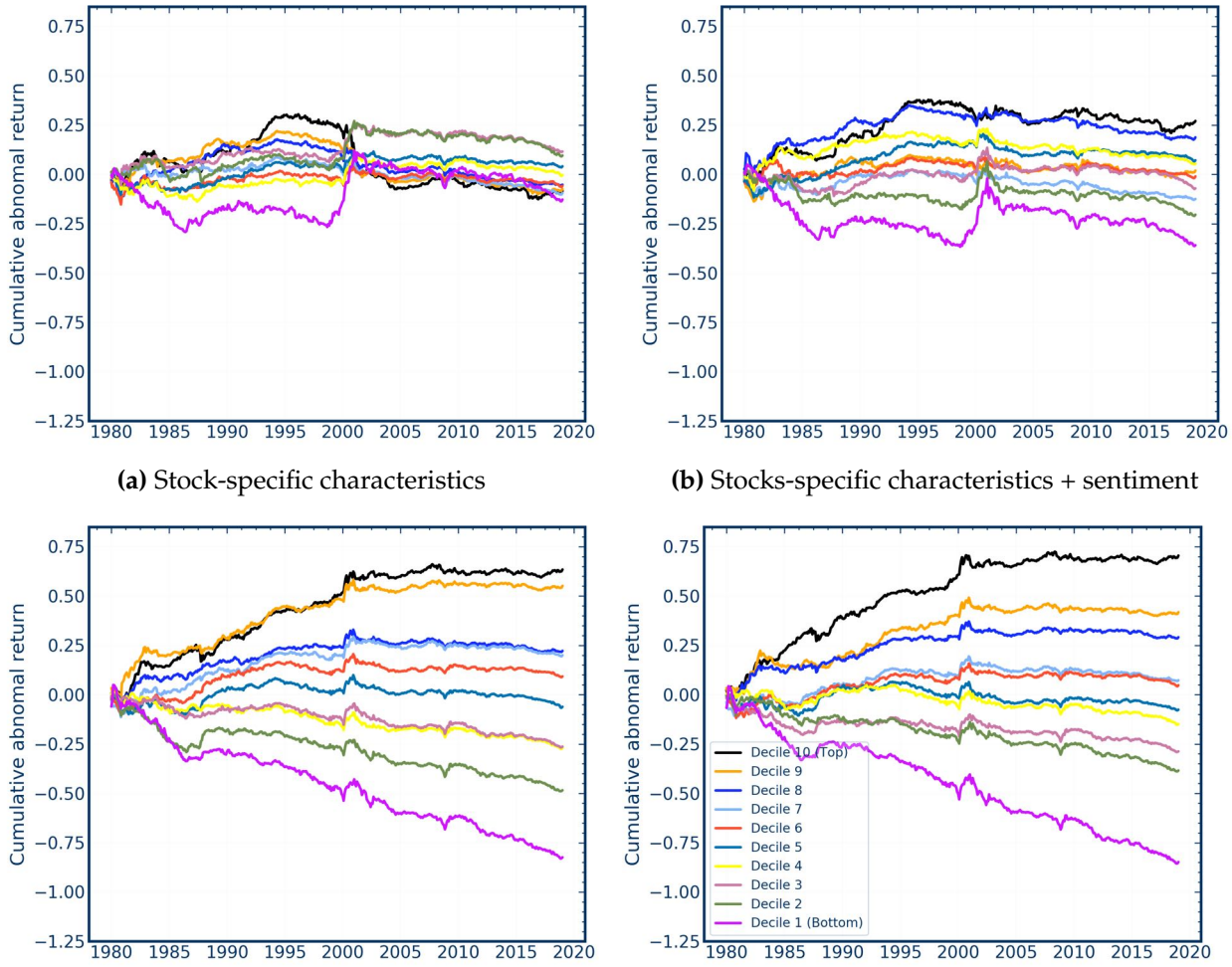
This figure plots the cumulative return of long-short decile portfolios that use fund- or stock-specific characteristics and sentiment to predict total returns (not abnormal returns).

Table A.4: Performance of equally-weighted long-short total return portfolios for different information sets.

Information set	mean (%)	t-stat	SR	R^2_F (%)
Stock+ fund+ sentiment	0.35	2.9***	0.14	-25.20
Fund+ sentiment	0.40	2.7***	0.13	0.56
Fund	0.44	3.2***	0.15	0.84
Stock+ sentiment	0.35	2.9***	0.14	-18.98
Stock	0.06	0.8	0.04	-55.93

This table reports the Sharpe ratio, mean and R^2_F of long-short equally-weighted decile portfolios based on predicting *returns* instead of abnormal returns with different information sets. We consider five different information sets which combine fund-specific and stock-specific characteristics and sentiment.

Figure A.4: Cumulative abnormal returns of equally-weighted deciles: different information sets.



(a) Stock-specific characteristics

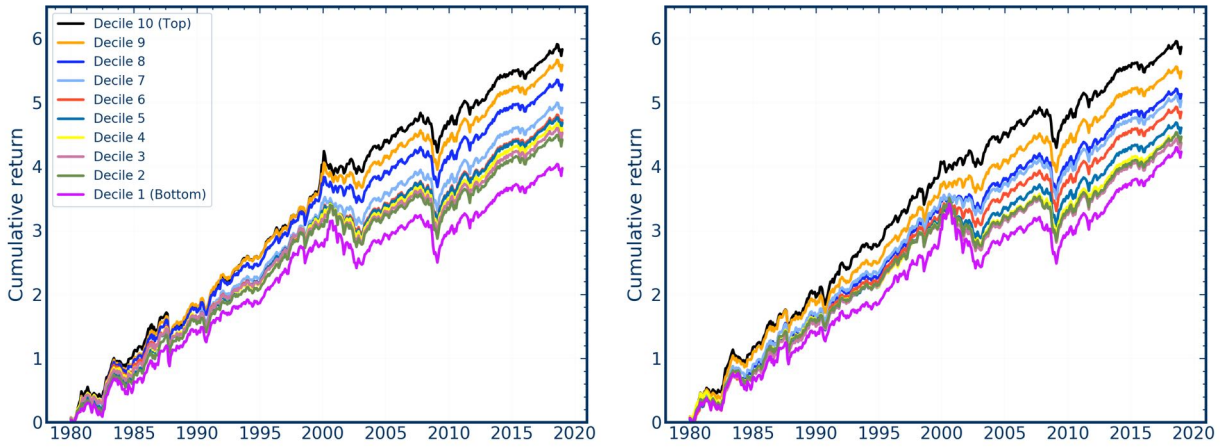
(b) Stocks-specific characteristics + sentiment

(c) Fund-specific characteristics

(d) Fund-specific characteristics + sentiment

This figure shows the cumulative abnormal returns sorted into prediction deciles for different information sets. The returns are equally-weighted within deciles. We consider fund-specific characteristics + sentiment, stock-specific characteristics+ sentiment, fund-specific characteristics or stock-specific characteristics to predict abnormal returns.

Figure A.5: Cumulative total returns of equally-weighted deciles: different information sets.



(a) Fund-specific characteristics+ sentiment

(b) Stock-specific characteristics+ sentiment

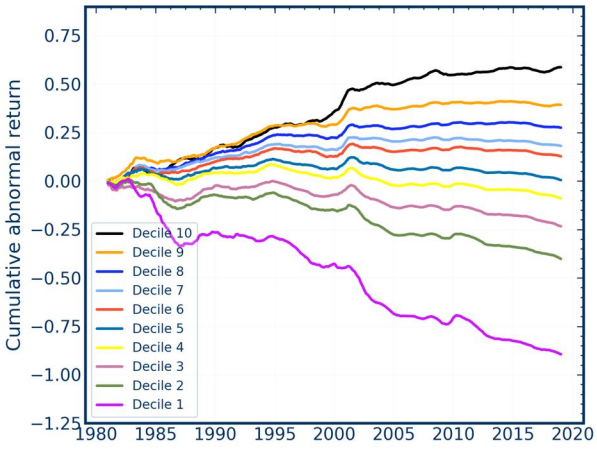
This figure plots the cumulative total (not abnormal) returns sorted into prediction deciles for different information sets. The returns are equally-weighted within deciles. We consider information sets which combine fund-specific and stock-specific characteristics and sentiment to predict returns instead of abnormal returns.

A.1.4 One-Year Holding Period

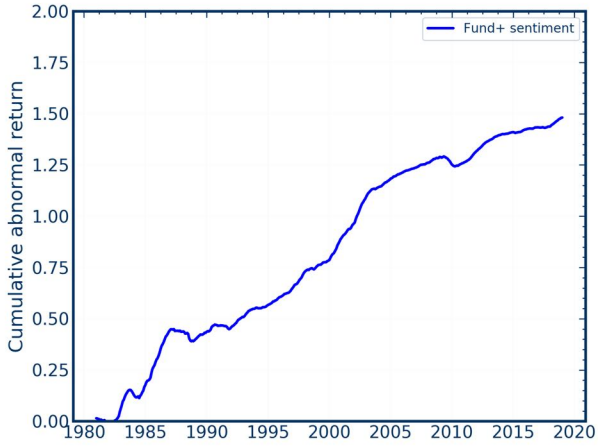
The results in main text are based on one-month ahead prediction. We obtain better results for longer holding periods when we estimate our model with a longer horizon prediction objective.

In this section, we use the same network structure and cross-out-of-sample evaluation method as in the main text, but predict one-year ahead abnormal returns. At each time t, we use characteristics to predict the average abnormal return from time $t+1$ to $t+12$. Figure A.7 reports the performance of annual abnormal return decile portfolios based on the annual prediction target. As expected, the performance improves relative to holding the one-month prediction portfolio for 12 months. We confirm that the predictability lasts over longer horizons.

Figure A.7: Results for annual overlapping abnormal return prediction portfolios



(a) Cumulative abnormal returns of deciles



(b) Cumulative abnormal returns of long-short portfolio

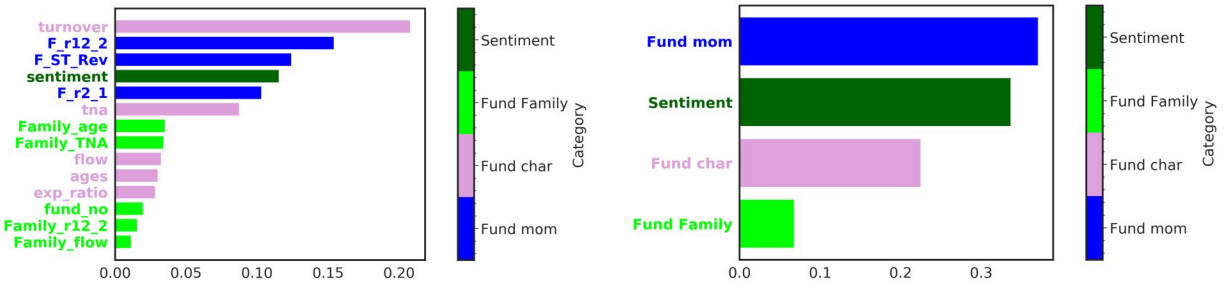
Information set	SR	mean(%)	t-stat
Fund+ sentiment	0.27	0.31	6.6***

(c) Performance of long-short abnormal return portfolio for fund + sentiment information

These panels show the out-of-sample results for prediction deciles for one-year holding periods. Each month, we predict cumulative abnormal returns over the next 12 months with fund-specific characteristics and sentiment. Panels (a) and (b) show the overlapping cumulative annual abnormal returns of prediction-weighted deciles and the long-short portfolio of the top decile minus the bottom decile. Panel (c) reports the Sharpe ratio, mean, and t-statistics of the annual overlappping abnormal return long-short portfolio. The variances and standard errors are adjusted with the Newey-West estimator with twelve lags. The performance is scaled to monthly abnormal returns.

Figure A.8 plots the most important variables for predicting annual abnormal returns. The variable importance changes, and shifts to more persistent fund characteristics.

Figure A.8: Top variable importance for explaining annual overlapping abnormal returns.



(a) Top variable importance.

(b) Top variable group importance.

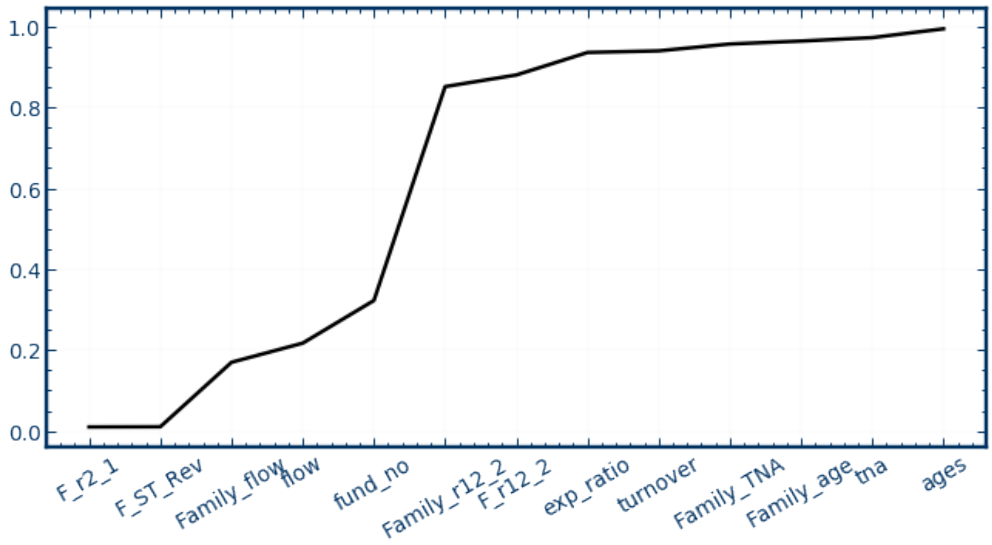
This figure shows the importance ranking for individual variables and variable groups. The ranking is the square root of average of the squared gradient for the eight ensemble fits as in equation 7. The variable importance measures are evaluated on the test data and averaged across three cross-out-of-sample folds. Fund-specific characteristics and sentiment are used as network input. The model's prediction target is annually averaged and overlapping abnormal returns

A.1.5 Persistence of Signals and Classification

The prediction of returns and investment strategies for longer horizons is directly related to the persistence of the fund characteristics. The robustness to longer holding periods requires at least some of the characteristics to be persistent. Indeed, we find that many fund characteristics are stable over short time horizons.

Figure A.9 shows the autocorrelation of the fund characteristics averaged over time and funds. Several fund characteristics are very persistent, in particular turnover and F_r12_2, which are two of the most important model predictors.

Figure A.9: Persistence of fund characteristics



This figure shows the persistence of fund characteristics. We estimate the autocorrelation for each time series of fund and characteristics. We exclude time series with less than 50 observations. The figure reports the persistence as the average autocorrelation for each characteristics sorted in increasing order.

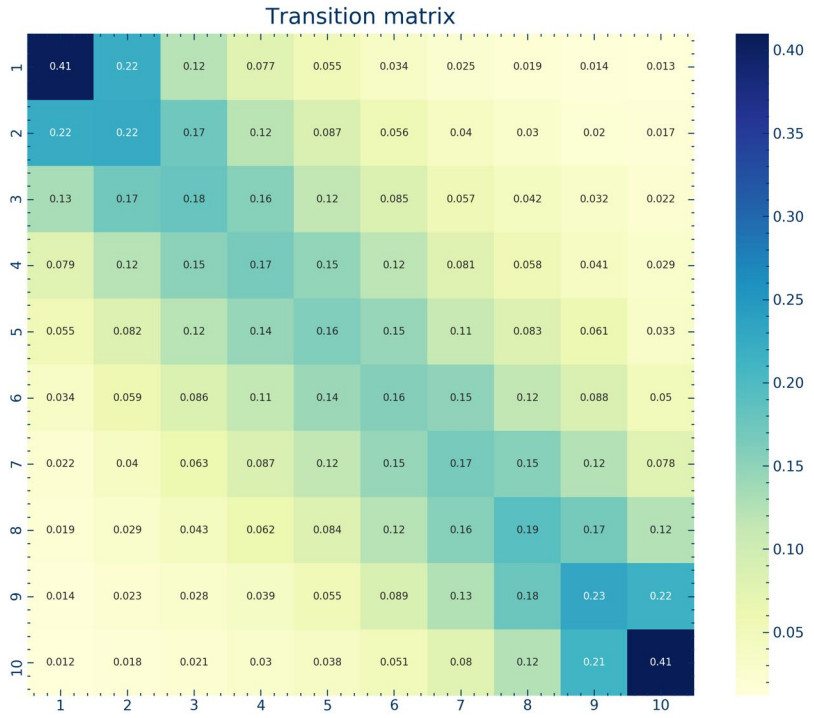
A closely related aspect is the persistence of the classification of funds. For this purpose we estimate the transition matrix between prediction quantiles based on the one-month prediction. In more detail, we count the percentage of funds that move from prediction decile i to decile j in the next month. This transition probability is calculated as

$$P_{i,j} = \frac{1}{T} \sum_t \frac{\sum_k 1^i_{k,t} 1^j_{k,t+1}}{\sum_k 1^i_{k,t}},$$

where $1^i_{k,t}$ is an indicator variable if a fund k is in decile i at time t and decile j at time $t+1$.

The result is shown in Figure A.10. Rows correspond to i and column to j. For example, the value of 0.22 of the $(1,2)$ element implies that on average 22% of the funds in the bottom decile at time t move to the second decile at time $t+1$. We observe that the top 20 and bottom 20 percent of fund classifications are very persistent, which explains why long-short portfolios

Figure A.10: Transition matrix of machine learning prediction



This matrix shows the transition probabilities of the prediction decile classifications between time t and time $t+1$. The transition probability is defined as $P_{i,j} = \frac{1}{T}\sum_t \frac{\sum_k 1^i_{k,t} 1^j_{k,t+1}}{\sum_k 1^i_{k,t}}$, where $1^i_{k,t}$ is an indicator variable if a fund k is in decile i at time t and decile j at time $t+1$.

remain profitable over longer horizons.

A.1.6 Macroeconomic Conditioning Variables

The main text shows the importance of sentiment and its interaction effects with fund characteristics. It is reasonable to ask whether other variables like CFNAI might play a similarly important role in predicting mutual fund out-performance. Or maybe they add a useful piece of macroeconomic information that is not contained in sentiment? To answer these questions, we estimate several additional neural network models which combine fund-level information with the following macro variables: sentiment (benchmark), CFNAI, sentiment orthogonalized to CFNAI, CFNAI orthogonal to sentiment, and sentiment plus CFNAI.[26] Table A.5 shows the results.

[26]We use a least-squares orthogonalization. The results are similar for a least absolute deviation orthogonalization. Sentiment and both orthogonalized sentiment series are all very similar because sentiment has a low 10% correlation with CFNAI. The prediction of abnormal fund returns can only use a relatively small number of macroeconomic predictors as macroeconomic variables are different from cross-sectional characteristics. In our sample we have a large cross-sectional dimension, but only a comparatively small time dimension. The effect of macroeconomic variables is estimated from the time-series, while cross-sectional variables can take advantage from the variation in the large cross-sectional dimension. Therefore, there is simply to little time-series data to estimate a model with a large number of macroeconomic variables.

Table A.5: Long-short abnormal return portfolios for different macro-economic information.

Weighting method	Information set	mean (%)	t-stat	SR	R^2_F (%)
Prediction-weighted	Fund+sentiment	0.40	5.4***	0.25	2.73
	Fund+CFNAI	0.39	6.0***	0.28	0.72
	Fund+sentiment+CFNAI	0.42	6.3***	0.29	2.48
	Fund+sentiment_orth	0.43	6.4***	0.29	1.22
	Fund+CFNAI_orth	0.38	5.4***	0.25	0.92
	Fund	0.38	5.5***	0.25	0.19
Equally-weighted	Fund+sentiment	0.33	5.9***	0.27	3.50
	Fund+CFNAI	0.33	6.5***	0.30	0.85
	Fund+sentiment+CFNAI	0.32	6.0***	0.28	2.71
	Fund+sentiment_orth	0.34	6.8***	0.31	1.58
	Fund+CFNAI_orth	0.31	5.8***	0.27	1.11
	Fund	0.31	5.8***	0.27	0.24

This table reports the Sharpe ratio, mean and R^2_F of long-short prediction-weighted and equally-weighted decile portfolios that use different macro-economic information sets for the prediction. We consider six different macro-economic information sets: none, sentiment, CFNAI, sentiment orthogonal to CFNAI, CFNAI orthogonal to sentiment and sentiment+CFNAI. We use least-squares orthogonalization.

In terms of out-of-sample mutual fund return predictability, using CFNAI leads to equally strong results in terms of the mean and Sharpe Ratio of the long-short portfolio. This result is surprising at first in light of the low correlation between sentiment and CFNAI. However, the low (linear) correlation is misleading. Sorting the respective time series of sentiment, CFNAI, and orthogonalized sentiment into high, medium, and low states (terciles) results in large overlap between the states. High-sentiment and high-CFNAI periods are often the same periods. What matters for mutual fund abnormal return predictability is to distinguish between the good and the bad states. This can be done equally effectively with sentiment and CFNAI. Put differently, different neural network models place very similar mutual funds into the same deciles. We calculate that 78% of mutual funds that are put in the bottom decile by the model that uses sentiment are also put in the bottom decile by the model that uses CFNAI. The corresponding number for the top decile is 74%. These numbers are even higher for the orthogonalized sentiment measure and when using both sentiment and CFNAI as shown in Table A.6.

Figure A.11 shows cumulative abnormal returns on the long-short portfolio, using prediction-weighting and equally-weighted funds within the top and bottom deciles. It reinforces the result that the four different macro-economic information sets result in similarly strong out-performance. Consistent with our previous results, the prediction-weighting results in larger return spreads between the extreme deciles and hence a larger mean return.

Does that mean that the predictions with sentiment and CFNAI are equally good? No. The R^2_F statistic is substantially higher when sentiment is used than when CFNAI is used. In other words, the actual outperformance of the best relative to the worst funds aligns better with the predicted outperformance when sentiment is used. The reason is that the model with sentiment

Table A.6: Prediction based classification relative to fund+ sentiment information.

	# bin = 2		# bin = 5		# bin = 10		# bin = 20	
Bin	1st	2nd	1st	5th	1st	10th	1st	20th
Fund+ CFNAI	0.91	0.91	0.84	0.82	0.78	0.74	0.70	0.61
Fund+ sentiment_orth	0.95	0.95	0.89	0.89	0.85	0.81	0.79	0.70
Fund+CFNAI_orth	0.91	0.91	0.83	0.83	0.78	0.74	0.70	0.61
Fund+ sentiment and CFNAI	0.94	0.94	0.87	0.89	0.83	0.82	0.75	0.73
Fund	0.93	0.93	0.87	0.86	0.81	0.78	0.73	0.67
Stock+ sentiment	0.55	0.55	0.27	0.27	0.17	0.15	0.11	0.08
Stock	0.55	0.55	0.27	0.27	0.17	0.16	0.11	0.08

This figure shows the percentage of funds that overlap with the prediction quantiles based on fund+ sentiment information. We consider two, five, 10 or 20 quantiles and six different information sets for predicting abnormal returns. The reference classification is fund + sentiment.

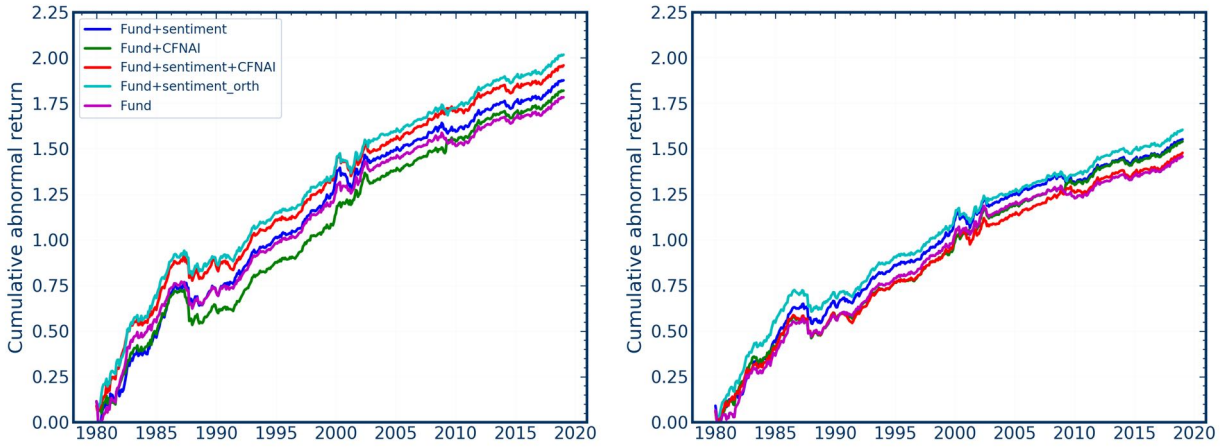
does a better job predicting the actual (the cardinal and not just the ordinal) abnormal return of the funds in the top and bottom deciles than the model with CFNAI. In other words, while sentiment and CFNAI result in similar decile rankings of funds, the model with sentiment is substantial better in predicting the level of the performance. This can be exploited for timing the investments. Table A.7 shows the out-of-sample performance based on sentiment terciles. The predictability of abnormal returns is the highest for medium and high sentiment states. An investor, who only invests into the long-short strategy during periods of high predictability, earns more than twice the expected return compared to the low predictability state. The table also shows that even investors who can only take long positions can still achieve an average monthly abnormal return of 0.27% by investing into the top funds in high sentiment periods. Note that these results represent a valid out-of-sample performance as the strategy uses lagged sentiment and estimates the terciles cut-offs without the use of out-of-sample data.

Table A.7: Long-short abnormal return portfolios in different sentiment terciles.

	T^L				T^M				T^H			
Portfolio	SR	mean	t-stat	R^2_F	SR	mean	t-stat	R^2_F	SR	mean	t-stat	R^2_F
D10-D1	0.12	0.23	1.6	0.50	0.37	0.42	4.6***	3.39	0.32	0.55	4.0***	4.83
D1	-0.11	-0.18	-1.4	0.71	-0.25	-0.23	-3.1***	3.65	-0.23	-0.29	-2.9***	1.35
D2	-0.19	-0.16	-2.4**	1.77	-0.15	-0.10	-1.8*	1.77	-0.05	-0.04	-0.6	-3.43
D3	-0.12	-0.09	-1.5	0.80	-0.05	-0.04	-0.7	1.14	0.03	0.02	0.4	-7.41
D4	-0.15	-0.10	-1.9*	1.10	0.01	0.01	0.1	-0.53	-0.07	-0.06	-0.9	-4.49
D5	-0.09	-0.06	-1.1	0.44	0.01	0.01	0.2	0.49	0.01	0.00	0.1	-2.90
D6	-0.09	-0.06	-1.2	0.67	0.12	0.07	1.5	-0.85	0.06	0.04	0.7	-3.20
D7	-0.15	-0.09	-1.8*	1.05	0.10	0.06	1.2	-0.12	0.15	0.11	1.9*	-0.43
D8	-0.10	-0.06	-1.2	-0.55	0.15	0.10	1.8*	-1.01	0.14	0.12	1.7*	-1.69
D9	-0.06	-0.05	-0.8	-0.80	0.19	0.24	2.3**	-0.32	0.07	0.08	0.9	-0.12
D10	0.05	0.04	0.6	-0.86	0.22	0.19	2.7***	1.00	0.21	0.27	2.6**	2.68

This table reports the Sharpe ratio, mean, t-statistic of mean and R^2_F of prediction-weighted decile portfolios evaluated in different sentiment terciles. The mean and R^2_F are reported in percentages. The low, medium and high tercile (T^L, T^M and T^H) splits for sentiment are based on the in-sample data of each of the three folds, and represent a valid out-of-sample performance. We use fund-specific characteristics and sentiment to predict abnormal returns.

Figure A.11: Cumulative abnormal returns of long-short portfolios for different macroeconomic information sets



(a) Prediction-weighted long-short portfolio

(b) Equally-weighted long-short portfolio

This figure plots the cumulative abnormal return returns of long-short prediction-weighted and equally-weighted decile portfolios that use different macroeconomic information and fund-specific characteristics to predict abnormal returns. We consider the following macro-economic information sets: none, sentiment, CFNAI, sentiment orthogonal to CFNAI and sentiment+CFNAI. We use least-squares orthogonalization.

A.1.7 Interaction Effects

To help understand the origin of the weaker performance of the model with CFNAI, Figure A.12 revisits the interaction effects of fund-level characteristics with CFNAI. It shows no interaction effects: the predictive effect of a fund characteristic on abnormal returns in high CFNAI periods is a parallel shift from the relationship in low-CFNAI periods. This is in contrast to the strong interaction effects for sentiment.

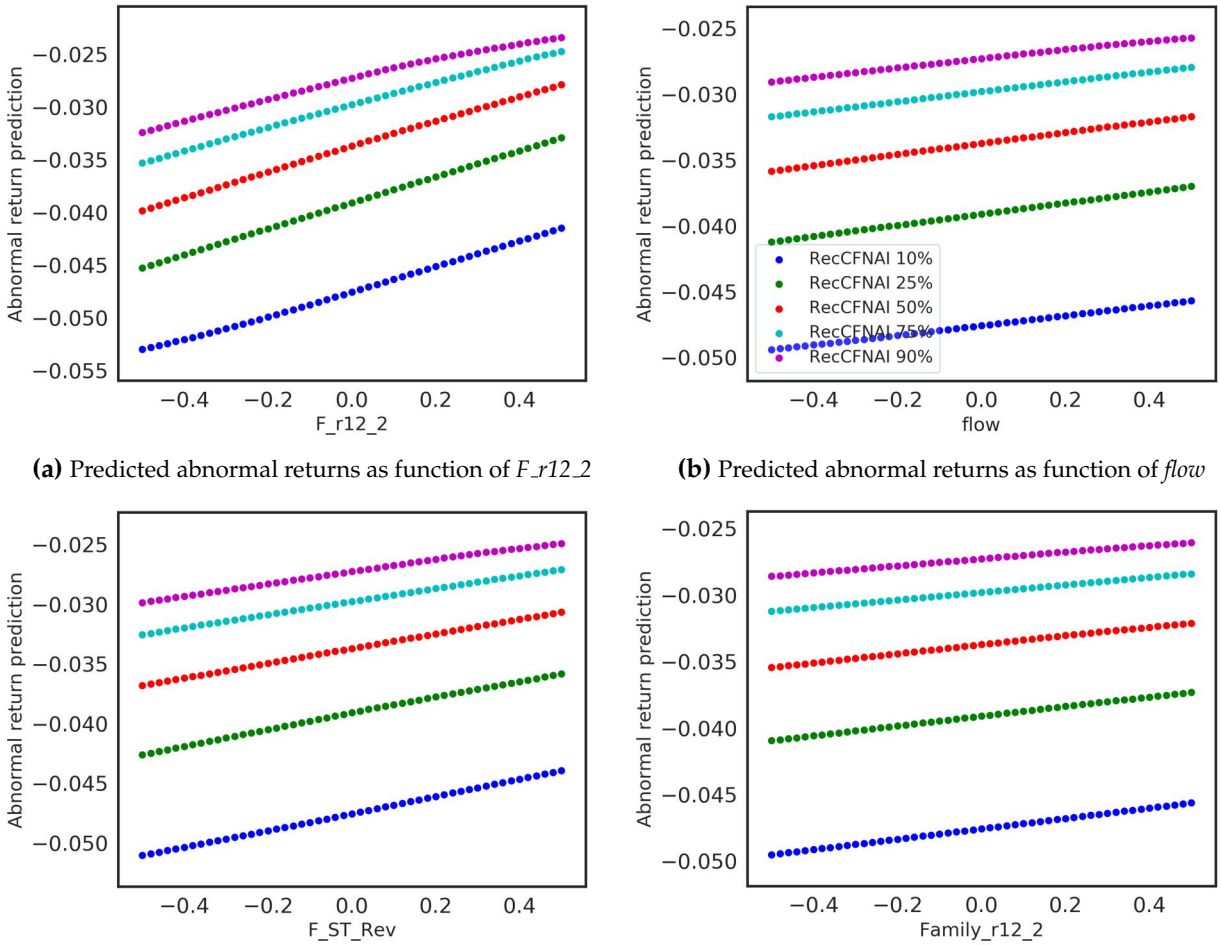
Figure A.13 compares the interaction measure for sentiment (orange bars) and CFNAI (black bars) with the fund characteristics. As noted in the main text, return spreads due to fund momentum, turnover, flow and reversal are the most affected by sentiment. In contrast, CFNAI has virtually no interaction with the fund characteristics.

We assess more general interactions between different fund characteristics. Here, we calculate the interactions between fund-specific characteristics. The interaction measure is defined similar to the main text as:

$$\begin{aligned}\text{Interaction}(z_i, z_j) = &\left(\hat{R}^{abn}(\text{high } z_i, \text{high } z_j) - \hat{R}^{abn}(\text{low } z_i, \text{high } z_j)\right) \\ &- \left(\hat{R}^{abn}(\text{high } z_i, \text{low } z_j) - \hat{R}^{abn}(\text{low } z_i, \text{low } z_j)\right).\end{aligned}$$

We set the high values of z_i and z_j to 0.5, and the low value to -0.5, which makes the measure symmetric with respect to z_i and z_j, that is, $\text{Interaction}(z_i, z_j) = \text{Interaction}(z_j, z_i)$. The other

Figure A.12: Conditional mean as a function of fund characteristics and CFNAI



(a) Predicted abnormal returns as function of *F_r12_2*

(b) Predicted abnormal returns as function of *flow*

(c) Predicted abnormal returns as function of *F_ST_Rev*

(d) Predicted abnormal returns as function of *Family_r*12_2

This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic and different CFNAI quantiles. The other variables are set to their median values. The neural network model is estimated with fund-specific characteristics and CFNAI. The interaction effects are evaluated on test data and averaged across three cross-out-of-sample folds. The high-minus-low factors have almost the same mean conditional on different past CFNAI. There is essentially no interaction effect.

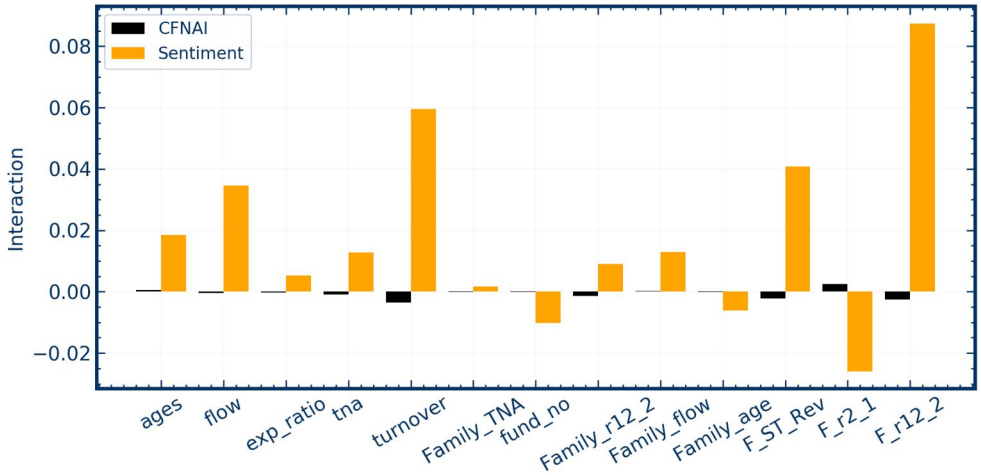
variables are set to their median values. The results are presented in Table A.8.

Table A.8: Interaction for fund-specific characteristics in the neural network model

	flow	F_r12_2	F_ST_Rev	Family_r12_2	turnover
F_r12_2	0.251				
F_ST_Rev	0.068	0.336			
Family_r12_2	0.037	0.147	0.018		
turnover	0.004	0.070	0.050	-0.020	
F_r2_1	0.080	0.284	0.049	0.048	-0.007

This table shows the interaction measure between the fund-specific characteristics. The results are presented in basis points. The neural networks use fund-specific characteristics and sentiment as input.

Figure A.13: Interaction of fund variables with sentiment and CFNAI



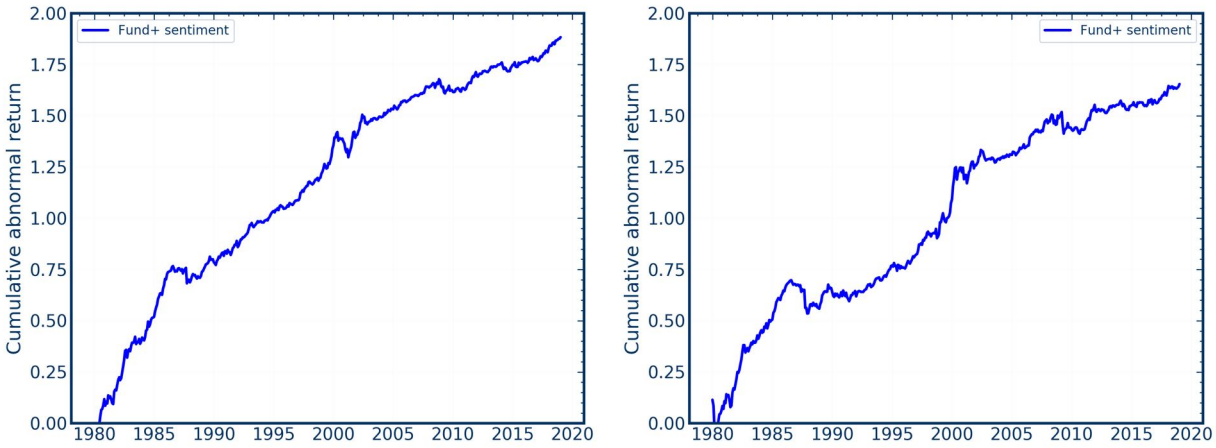
This figure reports the measure *Interaction(z,macro)* for fund characteristics and sentiment and CFNAI as macro-economic variables. We evaluate the predicted abnormal return $\hat{R}^{abn}$ for the highest and lowest value of the fund variable z and the high (90% quantile) and the low (10% quantile) macro-economic state. The other variables are set to their median values. The measure is reported in percentages.

A.1.8 Robustness to Size of Funds

The predictability is robust to excluding or down-weighting small mutual funds. We first present the results after excluding small mutual funds and second for value-weighted prediction portfolios.

We exclude mutual funds with less than 15 million asset under management (TNA), which is an often used cutoff in the literature. Figures 9 and A.14 and Table A.9 show the results for prediction weighted portfolios. The predictability is essentially not affected by dropping the smaller funds and our results are not driven by small funds.

Figure A.14: Results for abnormal return prediction long-short portfolios for mutual funds larger than 15 million and for value weights



(a) Prediction-weighted with size cutoff

(b) Prediction-value-weigthed

These figures show the cumulative abnormal returns of the long-short portfolios with prediction-weights for mutual funds with at least 15 million asset under management and for prediction-value-weights. We predict abnormal returns with fund-specific characteristics and sentiment. The left subfigure shows the results for mutual funds with at least 15 million asset under management, while the right subfigure uses prediction-value weighted portfolios.

Table A.9: Results for abnormal return prediction portfolios for mutual funds larger than 15 million and for value weights

Decile	mean(%)	t-stat	SR	R^2_F (%)
Prediction-weighted with size cutoff				
Long-short	0.40	6.5***	0.30	4.07
Top	0.17	3.9***	0.18	2.03
Bottom	-0.23	-4.5***	-0.23	2.05
Prediction-value-weighted				
Long-short	0.35	4.3***	0.20	2.30
Top	0.18	3.0***	0.14	1.12
Bottom	-0.18	-2.8***	-0.14	0.76

This table shows the out-of-sample results for prediction-sorted portfolios for mutual funds with at least 15 million asset under management and for prediction-value-weights using all funds. We predict abnormal returns with fund-specific characteristics and sentiment. We report the out-of-sample Sharpe ratios, mean returns and t-statistics of the long-short portfolio.The top subtable shows the results for mutual funds with at least 15 million asset under management, while the bottom table uses prediction-value weighted portfolios with all funds.

As shown in the main text, prediction-weighted portfolios use normalized model predictions as portfolio weights and thus take advantage of both the ranking and relative magnitude information. The prediction based weights are defined in equation (6), where $\tilde{\mu}_{i,t}$ are the normalized prediction weights. Value-weighted portfolios assign proportionally larger weights to funds with more assets under management (aum). We combine the prediction with the value of funds to form predication-value weights

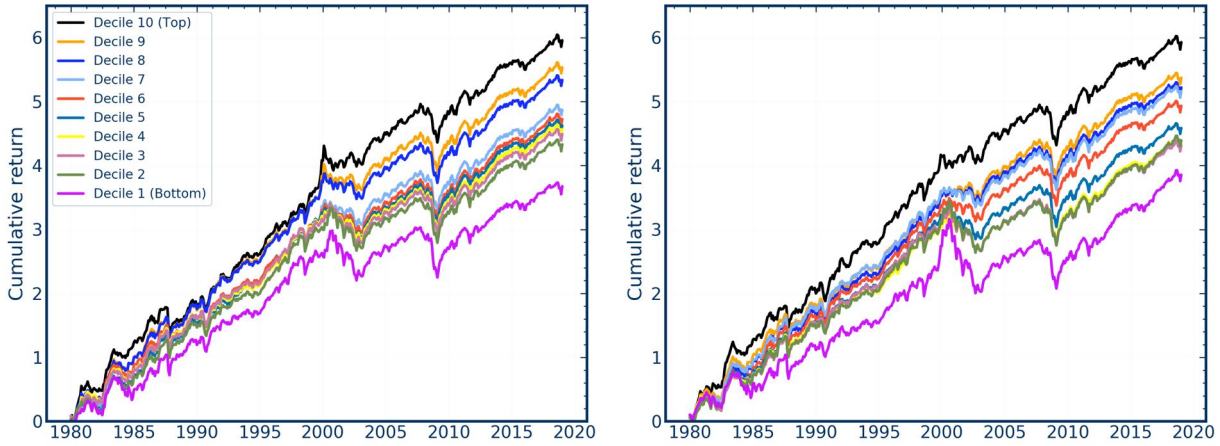
$$w_{i,t}^{\text{pred,val}} = \frac{\tilde{\mu}_{i,t} \times aum_{i,t}}{\sum_{i=1}^{N} \tilde{\mu}_{i,t} \times aum_{i,t}}$$

Figures 9 and A.14 and Table A.9 summarize the results for prediction-value weighted portfolios and show that the predictability is overall robust to value weighting.

A.1.9 Abnormal versus Total Return Prediction

Li and Rossi (2021) predict total mutual fund returns, $R_{i,t}$, while our paper predicts fund abnormal returns, $R_{i,t}^{abn}$. Fund total returns have a strong common component, due to fund exposures to common return factors F_t. Figure A.15 and Table A.10 report the results for predicting the *total* returns of mutual funds rather than *abnormal* returns. First, stock characteristics are substantially more predictive for total fund returns than for their abnormal returns. In other words, the stock characteristics seem to be able to predict the systematic factor component in fund returns, consistent with Li and Rossi (2021). However, as we have established above, once this factor component is taken out, stock characteristics lose most of their predictability. Indeed, our object of interest, abnormal returns, is orthogonal to the systematic component of fund returns by construction. It is

Figure A.15: Predicting Total Fund Returns



(a) Fund-specific characteristics+sentiment **(b)** Stock-specific characteristics+ sentiment

This figure plots the cumulative *returns* sorted into prediction deciles for different information sets. The returns are prediction-weighted within deciles. We consider information sets which combine fund-specific and stock-specific characteristics and sentiment to predict returns instead of abnormal returns.

this systematic component of returns that is predicted by stock characteristics. Second, the Sharpe ratio of long-short portfolios based on total return prediction is lower than from predicting abnormal returns. The Sharpe ratio in Table A.10 are only about 0.15. This points to an important methodological contribution of this paper. The level of fund returns (and also stock returns) is extremely hard to predict, while the relative performance is more predictable. Abnormal returns remove the level effect arising from compensation for systematic risk factor exposures. Hence, an abnormal return prediction objective is mainly a relative objective. In contrast, a machine learning prediction for returns might not select a model with a correct relative cross-sectional ranking of funds if it has a high prediction error in the level, which is largely irrelevant for relative ranking. In summary, abnormal returns are not only the object of interest to us, since we want to measure the returns managers earn in excess of systematic risk compensation, but they may also be the better objective for machine learning prediction in a statistical sense.

Table A.10: Performance of long-short return portfolios for different information sets.

Data	mean(%)	t-stat	SR	R^2_F (%)
Stock+ fund+ sentiment	0.45	3.1***	0.14	-26.54
Fund+ sentiment	0.49	3.0***	0.14	0.97
Fund	0.53	3.5***	0.16	0.97
Stock+ sentiment	0.44	3.1***	0.14	-20.03
Stock	0.11	1.1	0.05	-53.21

This table reports the Sharpe ratio, mean and factor R^2 of long-short prediction-weighted decile portfolios based on predicting *total returns* instead of abnormal returns with different information sets. We consider five different information sets which combine fund-specific and stock-specific characteristics and sentiment to predict returns instead of abnormal returns.

The comparison between returns and abnormal returns also illustrates the difference between conditional and unconditional factor models. The abnormal returns are estimated from a factor model using rolling-window regressions and hence reflect time-varying risk exposure. Holding-based stock characteristics seem to predict the time-varying risk exposure of the factors. The residuals of these local regressions take out the component that is predictable with holding-based stock characteristics.

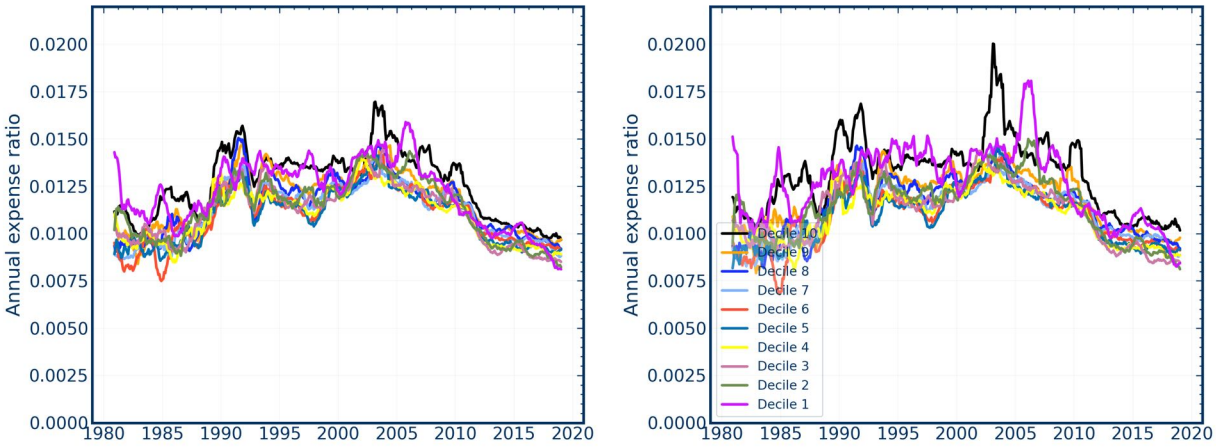
As a coda to this discussion, predicting total returns with a ML model and subsequently estimating an unconditional Carhart 4-factor model on the prediction portfolio returns is fundamentally different from first constructing abnormal returns from a conditional Carhart 4-factor model and subsequently predicting abnormal returns with a ML model. It is possible to find significant unconditional alphas for mutual fund portfolio returns formed from total return predictions using only stock characteristics, as shown in Table A.13. However, once we obtain local residuals with respect to Carhart factors, the stock characteristics lose most of their predictability. As we showed, these residuals depend almost entirely on fund-specific characteristics and sentiment.

A.1.10 Turnover and Expense Ratios Over Time

Figure A.16 plots turnover and expense ratios among predicted performance deciles. There are no systematic changes over time that differ between the top- and bottom-predicted performance portfolios in either the expense ratio or the turnover ratio.

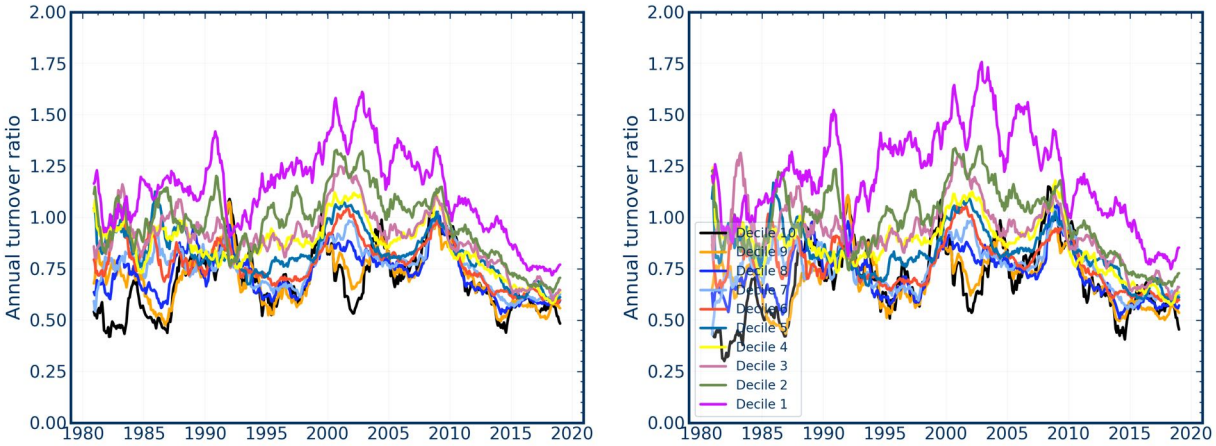
Figure A.16: Moving average expense and turnover ratios

Panel A: Expense ratio



(a) Equally-weighted **(b)** Prediction-weighted

Panel B: Turnover ratio



(c) Equally-weighted **(d)** Prediction-weighted

These figures show the 12-months moving average of expense and turnover ratios for the equally- and prediction-weighted deciles. The information set for the prediction are all characteristics and sentiment. We use the benchmark random sampling.

A.2 Chronological Sampling Method

The results in the main text apply a random cross-out-of-sample analysis, where the time periods for the three folds are sampled randomly to ensure that low and high sentiment states are represented in all three folds. The predictability results are robust to a cross-out-of-sample analysis with chronological sampling. This shows that the predictability in the main text is not driven by our sample selection.

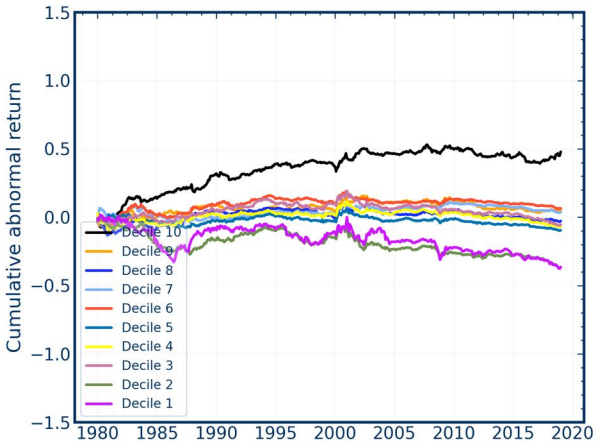
The chronological cross-out-of-sample analysis keeps the chronological order of observations within the folds. Exactly as in the main text, we use two of the folds to estimate the model and select the tuning parameters (two-thirds of the sample), and evaluate the prediction out-of-sample on the remaining fold (one-third of the sample). We repeat the estimation for three different combinations to obtain an out-of-sample prediction for each observation in the sample.

A.2.1 Long-short portfolio results

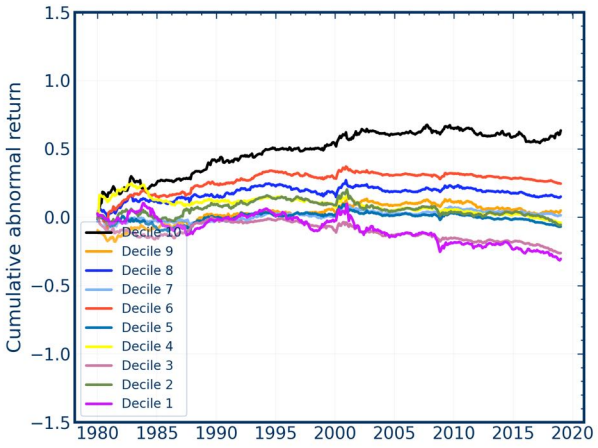
This subsection confirms that our main findings on abnormal returns are robust to chronological sampling. We can still strongly predict out-of-sample the abnormal returns of mutual fund managers, and the mean spread in skill between the top and bottom fund managers is similar to random sampling. Fund characteristics still have the strongest predictive power, but now the predictability with stock characteristics is stronger than before. However, the economic and statistical significance of stock characteristics is still much weaker compared to fund characteristics. When we zoom in on the time series of long-short portfolios, stock characteristics lose their predictive power after the year 2000. Fund momentum, flow, and sentiment still emerge as the most important variables for predicting abnormal returns. Consistent with our intuition, the role of investor sentiment is now weaker: it shows up as the second most important variable in the variable importance ranking and the interaction effects between fund variables and sentiment is weaker.

The cumulative abnormal returns for different information sets are in Figure A.17, the cumulative abnormal returns of long-short prediction portfolios are in Figure A.18. The Sharpe ratio, mean, and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios are in Table A.11.

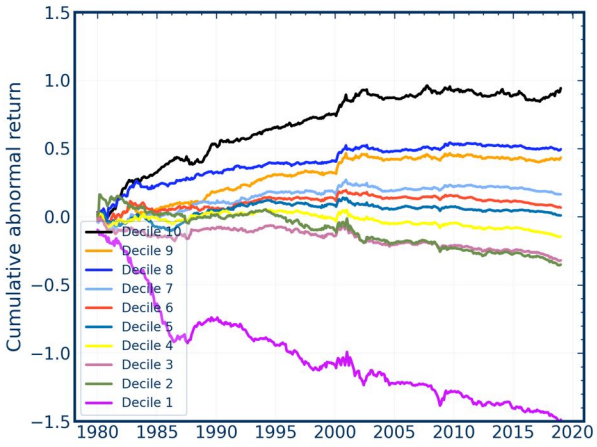
Figure A.17: Cumulative abnormal returns for different information sets with chronological data split.



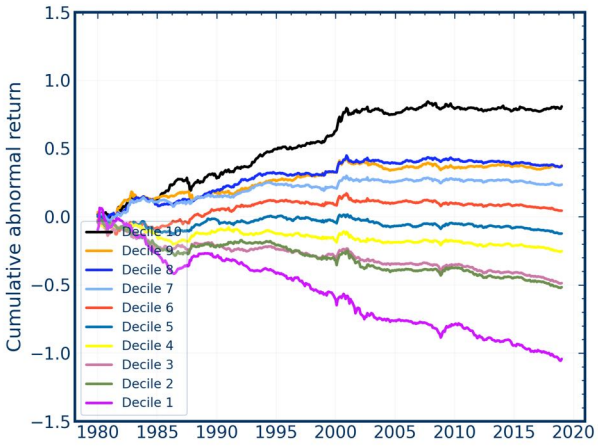
(a) Stock-specific characteristics



(b) Stock-specific characteristics + sentiment



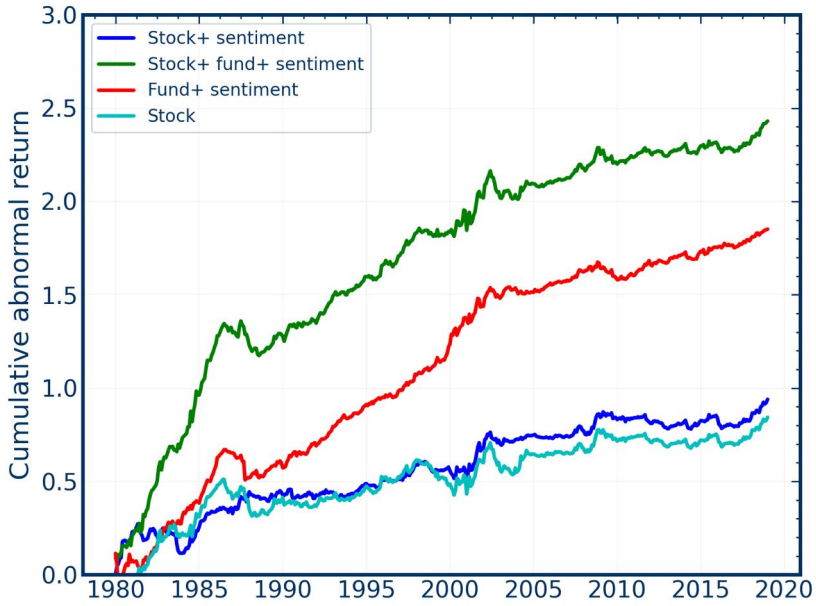
(c) All characteristics



(d) Fund-specific characteristics + sentiment

These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The abnormal returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment, stock-specific characteristics+ sentiment, stock-specific characteristics or all characteristics to predict abnormal returns. Three cross-out-of-sample folds keep the chronological order.

Figure A.18: Cumulative abnormal returns of long-short prediction portfolios with chronological data split.



This figure plots the cumulative abnormal returns of prediction-weighted long-short decile portfolios that use different information sets for prediction. We consider fund-specific and stock-specific characteristics combined with sentiment. Three cross-out-of-sample folds keep the chronological order.

Table A.11: Performance of abnormal return portfolios with chronological data split.

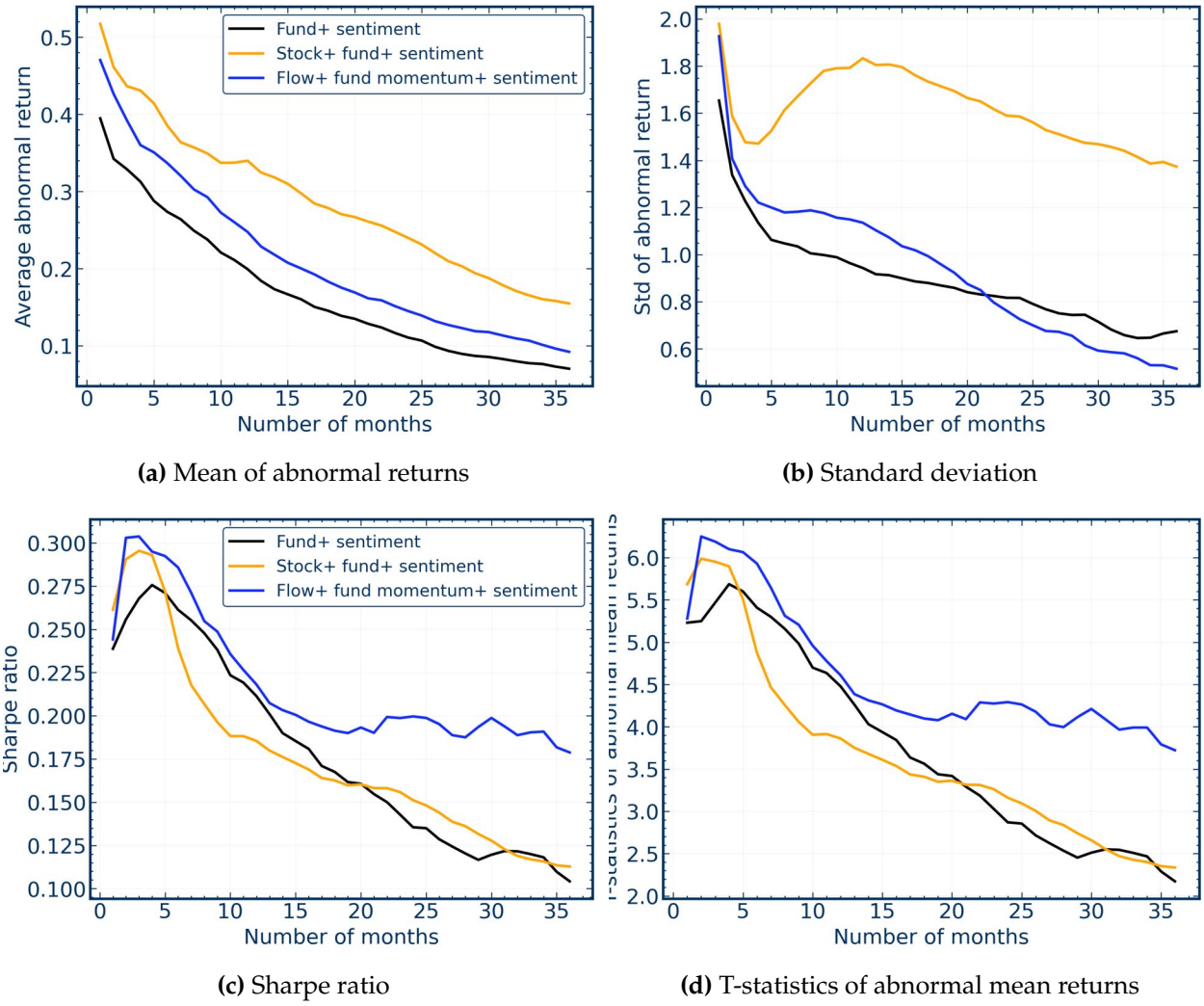
Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ fund+ sentiment	0.52	5.7***	0.26	2.66
	Fund+ sentiment	0.39	5.2***	0.24	1.49
	Fund	0.41	5.6***	0.26	0.52
	Stock+ fund	0.40	4.9***	0.23	-0.78
	Stock+ sentiment	0.20	2.9***	0.13	-13.48
	Stock	0.18	2.2**	0.10	-1.38
	Fund + CFNAI	0.36	5.2***	0.24	1.15
	Flow+ fund momentum+ sentiment	0.47	5.3***	0.24	1.12
	Fund exclude momentum and flow	0.00	0.1	0.00	0.10
	F_r12_2+ sentiment	0.32	4.0***	0.19	0.88
Top Decile	Stock+ fund+ sentiment	0.20	4.5***	0.21	-3.65
	Fund+ sentiment	0.17	3.6***	0.17	-0.44
	Fund	0.18	3.9***	0.18	-1.63
	Stock+ fund	0.16	3.3***	0.15	-7.07
	Stock+ sentiment	0.14	2.7***	0.12	-7.98
	Stock	0.10	2.4**	0.11	-6.38
	Fund + CFNAI	0.14	3.1***	0.14	-1.72
	Flow+ fund momentum+ sentiment	0.19	3.6***	0.17	-0.05
	Fund exclude momentum and flow	-0.03	-0.8	-0.03	-0.52
	F_r12_2+ sentiment	0.10	1.6	0.08	-0.05
Bottom Decile	Stock+ fund+ sentiment	-0.32	-4.1***	-0.33	2.93
	Fund+ sentiment	-0.22	-3.6***	-0.22	1.03
	Fund	-0.23	-3.8***	-0.23	0.30
	Stock+ fund	-0.23	-3.5***	-0.22	1.95
	Stock+ sentiment	-0.07	-1.2	-0.06	-5.46
	Stock	-0.08	-1.1	-0.08	0.31
	Fund + CFNAI	-0.22	-3.9***	-0.22	0.64
	Flow+ fund momentum+ sentiment	-0.27	-4.0***	-0.23	0.76
	Fund exclude momentum and flow	-0.03	-1.0	-0.04	-0.04
	F_r12_2+ sentiment	-0.22	-3.6***	-0.16	0.56

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios that use different information sets for the prediction. We consider nine different information sets which combine fund-specific and stock-specific characteristics and sentiment. We also include flow and fund momentum (F_r12_2, F_r2_1 and F_ST_Rev) individually. The cross-out-of-sample folds keep the chronological order in each fold.

A.2.2 Holding period

This subsection shows that the persistence of performance in our long-short portfolio is robust to chronological sampling. Figure A.19 shows the abnormal returns on a long-short prediction portfolio for holding periods ranging from 1 month to 36 months under chronological sampling.

Figure A.19: Performance for Different Holding Periods



(a) Mean of abnormal returns

(b) Standard deviation

(c) Sharpe ratio

(d) T-statistics of abnormal mean returns

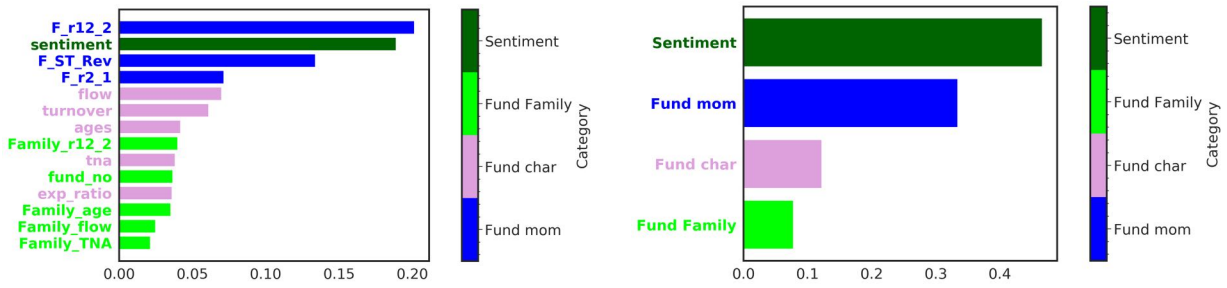
This figure shows the results for long-short prediction-weighted portfolios for different holding periods. The time periods are sampled chronologically. At each time t, we sort funds based on the one-month prediction into deciles and hold the long-short prediction portfolio for s months with overlapping returns. We calculate the mean, Sharpe ratio, standard deviation, and t-statistics of the overlapping abnormal returns. The one-month prediction uses either fund+sentiment, stock+fund+sentiment or flow+fund+sentiment.

A.2.3 Interaction effects

This subsection studies the interaction effects between fund characteristics and sentiment under chronological sampling. The interaction effects between sentiment and fund variables are less pronounced than with random sampling, but the interaction effects between sentiment and fund momentum, fund short term reversal, and flow are still significant, as in the main text.

The top variable importance for explaining abnormal returns and interaction effects between sentiment and fund characteristics are in Figures A.20 and A.21. Since the third chronological fold does not include any high sentiment states, there is mechanically no interaction effect with sentiment for this specific fold. As a result, Figure A.20 shows that the importance of sentiment declines. A model evaluation that would only use the last fold to assess the conditional abnormal return would not detect the strong interaction effects with sentiment.

Figure A.20: Top variable importance for explaining abnormal returns with chronological sampling.



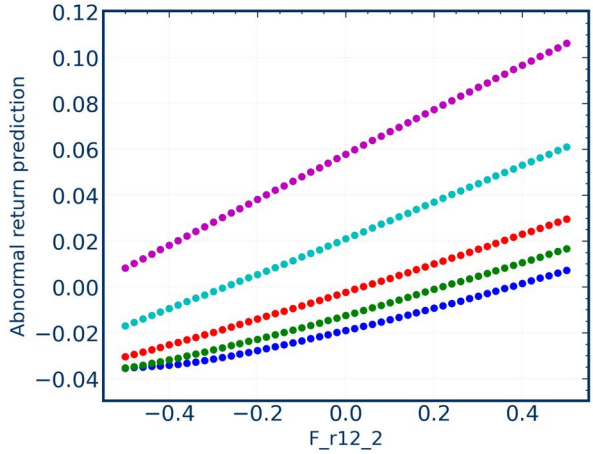
(a) Top variable importance.

(b) Top variable group importance.

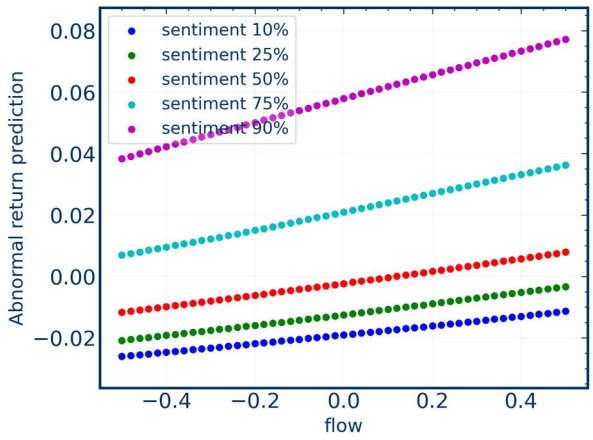
This figure shows the importance ranking for individual variables and variable groups. The ranking is the square root of average of the squared gradient for the eight ensemble fits as in equation 7. The variable importance measures are evaluated on the test data and averaged across three cross-out-of-sample folds. Fund-specific characteristics and sentiment are used as network input. Three cross-out-of-sample folds keep the chronological order.

Figure A.21 plots the mean of abnormal fund returns conditional on the values of one fund variable and sentiment. It shows that the interaction effects between sentiment and fund variables are less pronounced compared to the interaction effects at random sampling.

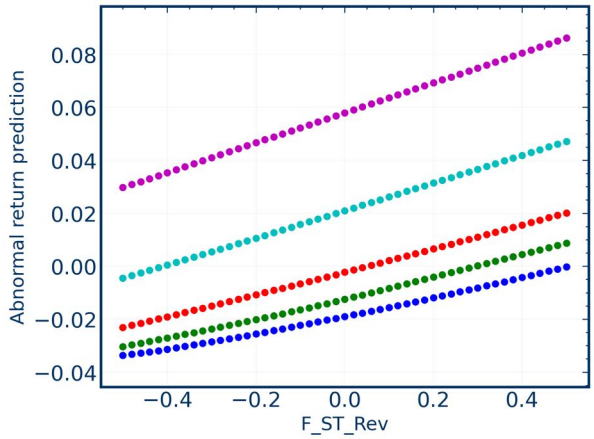
Figure A.21: Conditional mean as a function of fund characteristics and sentiment with chronological data split.



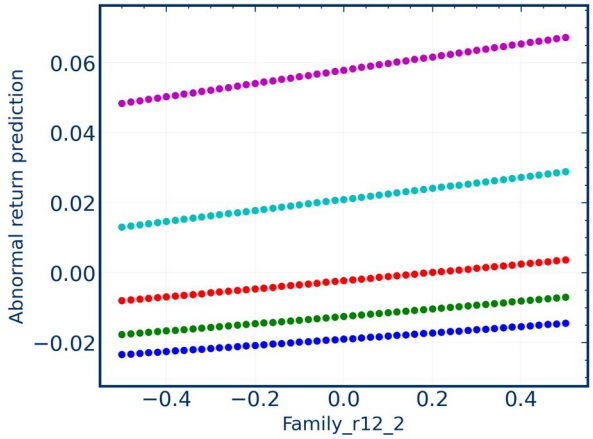
(a) Predicted abnormal returns as function of *F_r12_2*



(b) Predicted abnormal returns as function of *flow*



(c) Predicted abnormal returns as function of *F_ST_Rev*

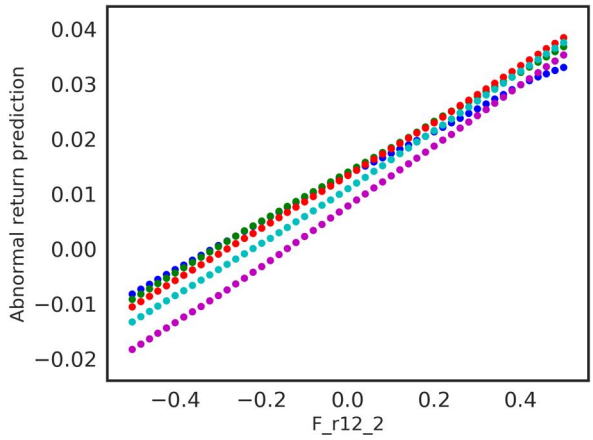


(d) Predicted abnormal returns as function of *Family_r12_2*

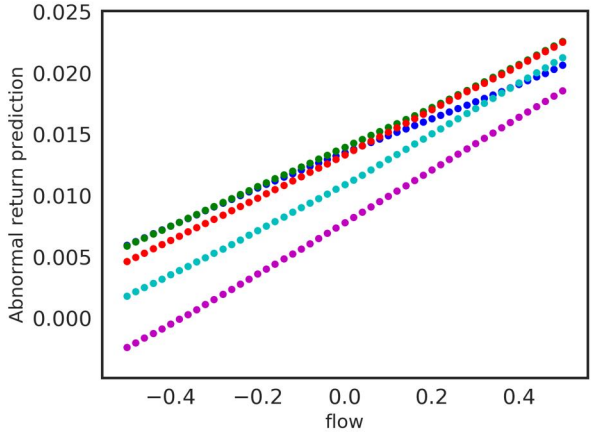
This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic conditional on different sentiment quantiles. The other variables are set to their median. The neural network model is estimated with fund-specific characteristics and sentiment. The interaction effects are evaluated on the test data and averaged across three cross-out-of-sample folds. The cross-out-of-sample folds keep the chronological order in each fold. The high-minus-low portfolios have a higher mean conditional on high past sentiment. This is a non-linear interaction effect.

Figure A.22 plots the mean of abnormal fund returns conditional on CFNAI. It shows weak interaction effects between CFNAI and fund characteristics, just like in the random sampling approach.

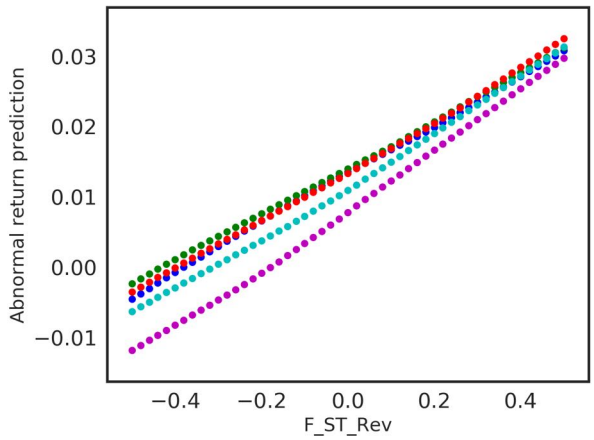
Figure A.22: Conditional mean as a function of fund characteristics and CFNAI with chronological data split.



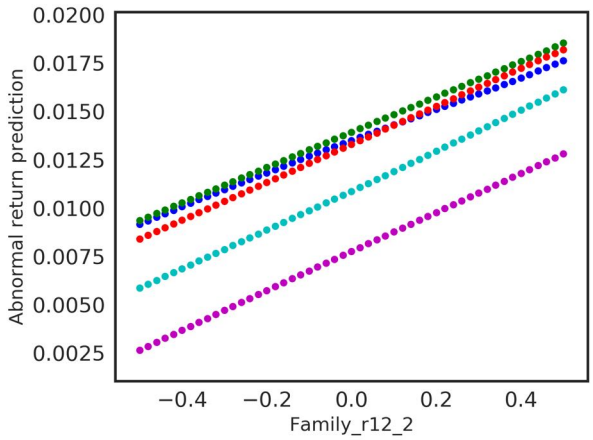
(a) Predicted abnormal returns as function of *F_r12_2*



(b) Predicted abnormal returns as function of *flow*



(c) Predicted abnormal returns as function of *F_ST_Rev*



(d) Predicted abnormal returns as function of *Family_r*12_2

This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic conditional on different CFNAI quantiles. The other variables are set to their median. The neural network model is estimated with fund-specific characteristics and CFNAI. The interaction effects are evaluated on the test data and averaged across three cross-out-of-sample folds. The cross-out-of-sample folds keep the chronological order in each fold.

A.2.4 Predicting Total Returns

Table A.12 shows the results for predicting *total* (as opposed to abnormal) return with various information sets under chronological sampling. Table A.13 shows exp-post unconditional factor regressions on total return prediction long-short portfolios. The alphas are significant and the R^2 modest.

Table A.12: Performance of long-short total return portfolios for different information sets with chronological data split.

Information set	mean(%)	t-stat	SR	R^2_F (%)
Stock+ fund+ sentiment	0.63	4.8***	0.22	-1.68
Fund+ sentiment	0.65	3.6***	0.17	0.56
Fund	0.61	3.6***	0.16	0.22
Stock+ fund	0.47	4.1***	0.19	-5.48
Stock+ sentiment	0.62	5.0***	0.23	0.62
Stock	0.42	3.8***	0.17	-3.19

This table reports the Sharpe ratio, mean and factor R^2 of long-short prediction-weighted decile portfolios based on total return prediction with different information sets. We consider six different information sets which combine fund-specific and stock-specific characteristics and sentiment. Three cross-out-of-sample folds keep the chronological order and the network structure is the same as the benchmark setup for predicting abnormal returns other than l1 penalty = $1e-5$.

Table A.13: Spanning of long-short return prediction portfolios with different factor models and chronological data split.

	FF 4 factors		FF 5 factors		FF 6 factors		FF 8 factors		
	α	R^2	α	R^2	α	R^2	α	R^2	mean μ
Stock+ fund+ sentiment	0.19***	0.08	0.19***	0.05	0.16***	0.09	0.20***	0.25	0.22***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Fund+ sentiment	0.14***	0.15	0.18***	0.07	0.14***	0.16	0.20***	0.51	0.17***
	(0.03)		(0.03)		(0.03)		(0.03)		(0.05)
Fund	0.15***	0.09	0.17***	0.04	0.14***	0.09	0.22***	0.58	0.16***
	(0.03)		(0.03)		(0.03)		(0.03)		(0.05)
Stock+ fund	0.16***	0.16	0.23***	0.05	0.19***	0.18	0.23***	0.36	0.19***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Stock+ sentiment	0.20***	0.11	0.20***	0.07	0.17***	0.12	0.21***	0.23	0.23***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
Stock	0.14***	0.24	0.24***	0.09	0.19***	0.27	0.22***	0.36	0.17***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)

This table reports the time-series regression results of long-short prediction-weighted decile portfolios for different factor models. The model predictions are based on machine learning predictions on fund returns with different information sets. Three cross-out-of-sample folds keep the chronological order and the network structure is the same as the benchmark setup for predicting abnormal returns other than l1 penalty = $1e-5$. We consider the 4-factor Fama-French-Carhart model (market, size, value and momentum), the 5-factor Fama-French model (market, size, value, profitability and investment), a 6-factor model which adds the momentum factor to the Fama-French 5 factors, and an 8-factor model which adds the momentum, short-term reversal and long-term reversal factors to the Fama-French 5 factors. The α column reports the time-series pricing error and R^2 is the explained variation of the regression. Both the long-short abnormal return portfolios and the factor models are normalized to have a standard deviation of 1. Standard errors are in brackets and stars denote the significance levels.

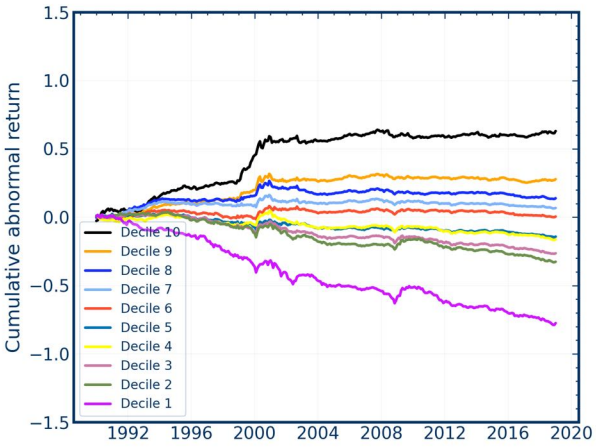
A.3 Expanding-Window Sampling Method

This section studies how our results change under an expanding window estimation of the neural network. This model assumes a time-varying conditional abnormal return function, in contrast with the constant abnormal return function assumed in the random and chronological cross-validation exercises. We predict abnormal returns with an expanding window estimation, updating the model every year. Since we need to "warm start" the model, the analysis now starts in the year 1990.

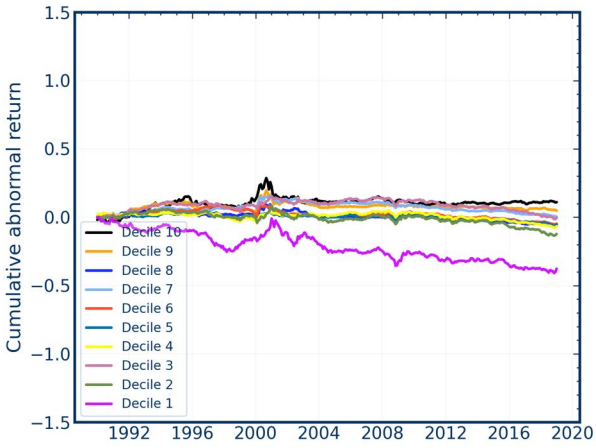
The cumulative abnormal returns for different information sets are in Figure A.23. The cumulative abnormal returns of long-short prediction portfolios are in Figure A.24. The Sharpe ratio, mean, and factor R^2 of the long-short, the first, and the tenth prediction-weighted decile portfolios are in table A.14. These results establish the strong predictability of fund manager skill and the important role of fund characteristics for prediction. The predictive power of stock characteristics also remains weaker compared to information sets that include fund information.

Figures A.25 and A.26 show that the important predictive role of fund momentum and fund flow and the interaction effects between fund variables and sentiment remain robust.

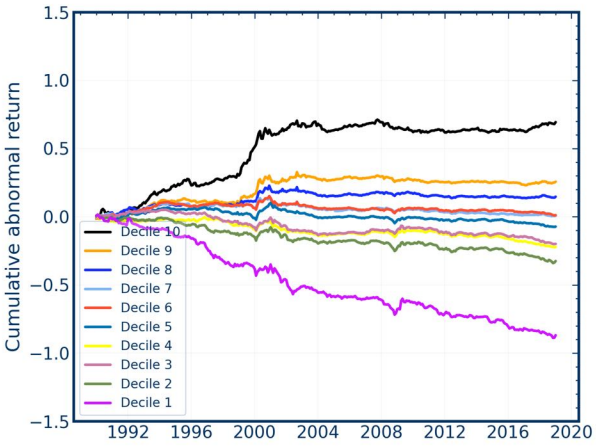
Figure A.23: Cumulative abnormal returns from rolling-window predictions for different information sets.



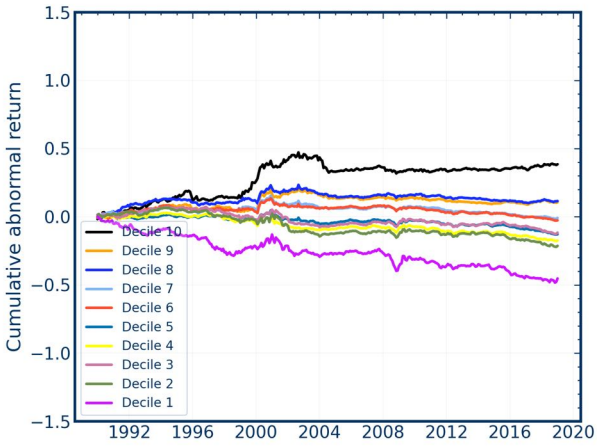
(a) Fund-specific characteristics + sentiment



(b) Stock-specific characteristics



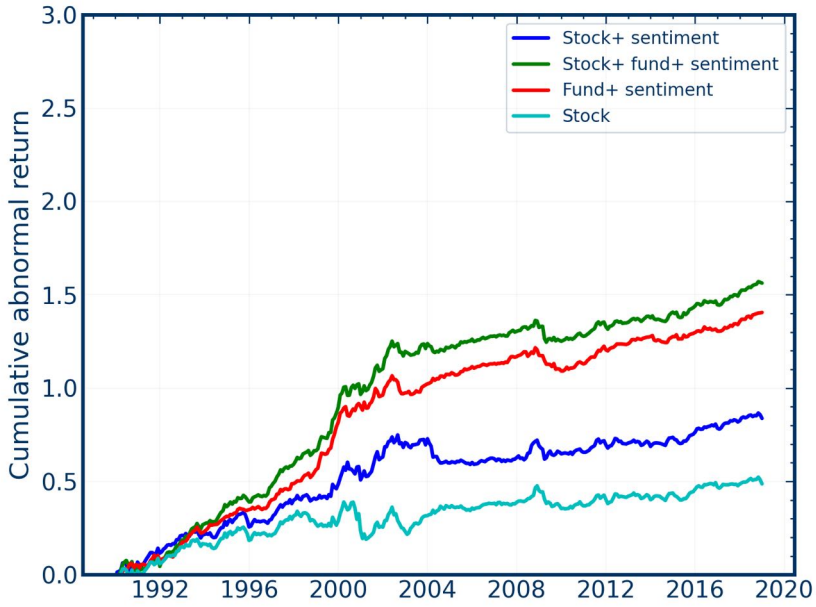
(c) All characteristics



(d) Stock-specific characteristics + sentiment

These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment, stock-specific characteristics+ sentiment, stock-specific characteristics or all characteristics to predict abnormal returns. The model predictions are generated in a rolling way. That is, we use data until year to generate predictions for year t+1, with t varying. To make sure that we have enough training data for out-of-sample evaluations, we start the out-of-sample analysis on year 1990.

Figure A.24: Cumulative abnormal returns of long-short rolling-window prediction portfolios



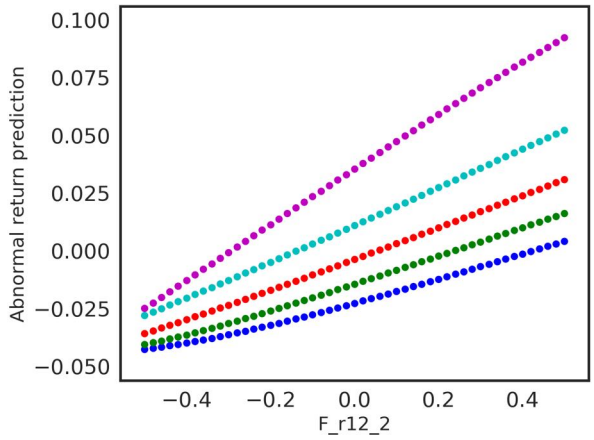
This figure plots the cumulative abnormal returns of prediction-weighted long-short decile portfolios that use different information sets for prediction. We consider fund-specific and stock-specific characteristics combined with sentiment. The model predictions are generated in a rolling way. That is, we use data until year to generate predictions for year t+1, with t varying. To make sure that we have enough training data for out-of-sample evaluations, we start the out-of-sample analysis on year 1990.

Table A.14: Performance of abnormal return portfolios with rolling-window predictions.

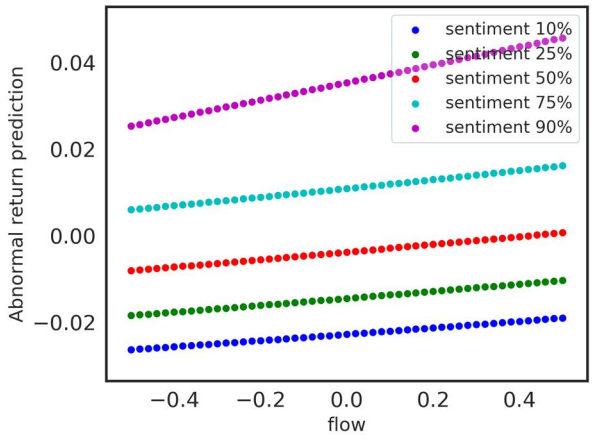
Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ fund+ sentiment	0.45	5.4***	0.29	8.92
	Fund+ sentiment	0.40	5.8***	0.31	5.71
	Stock+ sentiment	0.24	2.8***	0.15	2.78
	Stock	0.14	1.7*	0.09	-0.47
Top decile	Stock+ fund+ sentiment	0.20	3.2***	0.17	0.13
	Fund+ sentiment	0.18	3.4***	0.18	-1.37
	Stock+ sentiment	0.11	1.9*	0.10	-1.29
	Stock	0.03	0.5	0.03	-2.75
Bottom decile	Stock+ fund+ sentiment	-0.25	-4.0***	-0.21	3.31
	Fund+ sentiment	-0.22	-4.1***	-0.23	-1.44
	Stock+ sentiment	-0.13	-2.1**	-0.12	0.42
	Stock	-0.11	-1.6	-0.10	0.57

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios that use different information sets for the prediction. We consider four different information sets which combine fund-specific and stock-specific characteristics and sentiment. The model predictions are generated in a rolling way. That is, we use data until year to generate predictions for year t+1, with t varying.

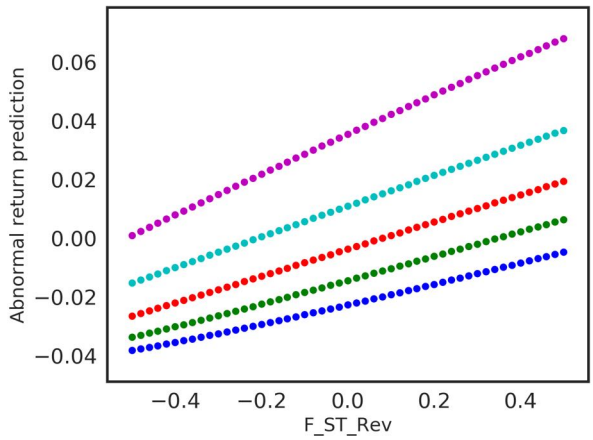
Figure A.25: Conditional mean as a function of fund characteristics and sentiment with rolling window predictions.



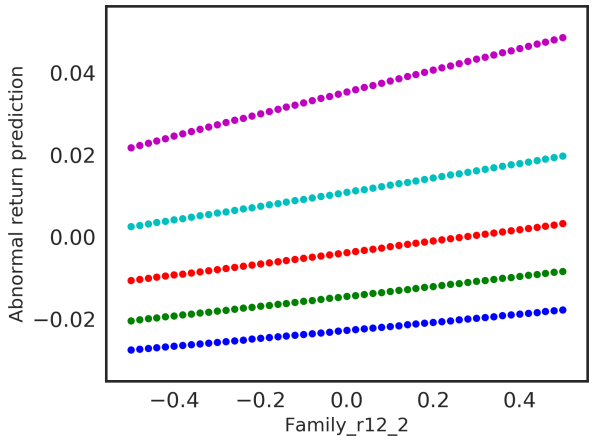
(a) Predicted abnormal returns as function of *F_r12_2*



(b) Predicted abnormal returns as function of *flow*



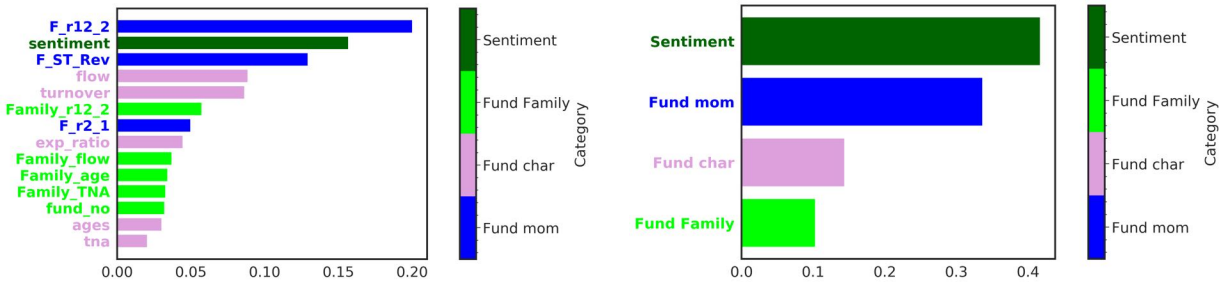
(c) Predicted abnormal returns as function of *F_ST_Rev*



(d) Predicted abnormal returns as function of *Family_r*12_2

This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic conditional on different sentiment quantiles. The other variables are set to their median. The neural network model is estimated with fund-specific characteristics and sentiment. When the training data is until time t, the interaction effects are evaluated on year t+1. The final interaction measure is averaged across all out-of-sample years. The high-minus-low portfolios have a higher mean conditional on high past sentiment. This is a non-linear interaction effect.

Figure A.26: Top variable importance for explaining abnormal returns with rolling window prediction.



(a) Top variable importance.

(b) Top variable group importance.

This figure shows the importance ranking for individual variables and variable groups. The ranking is the square root of average of the squared gradient for the eight ensemble fits as in equation 7. When the training data is until time t, the importance measure is on year t+1, and the variable importance measure is averaged across all out-of-sample years. Fund-specific characteristics and sentiment are used as network input.

B Implementation: Tuning Parameters

Table B.1 summarizes the tuning parameters for the possible network structures. HU, the number of hidden units in each layer, deserves more explanation. The nodes of first layer are 64 or 32 and the number of nodes in each layer is half of the previous layer. For example, for a neural network with 3 layers the number of nodes are $32, 16, 8$ and $64, 32, 16$. In mathematical terms the number of nodes in the ith layer is 2^{7-i} or 2^{6-i}.

Table B.1: Selection of tuning parameters

Notation	Tuning Parameters	Candidates	Optimal
HL	Number of layers in Neural Network	1, 2, 3	1
HU	Number of hidden units in each layer	2^{6-i} or 2^{7-i}, i = 1 to HL	64
DR	Dropout	0.90, 0.95	0.95
LR	Learning rate	0.001 , 0.01	0.01
L1	l1 regularization	0, 1e-5	0
L2	l2 regularization	0, 1e-2, 1e-3	1e-3

This table shows the set of tuning parameters, which result in 144 candidate models. The optimal parameters are selected on the validation data.

We obtain robust and stable fits by ensemble averaging over several fits of the models. A distinguishing feature of neural networks is that the estimation results can depend on the starting value used in the optimization. The standard practice which has also been used by Chen, Pelger, and Zhu (2023) is to train the models separately with different initial values chosen from an optimal distribution. Averaging over multiple fits achieves two goals: First, it diminishes the effect of a local suboptimal fit. Second, it reduces the estimation variance of the estimated model. All our neural networks are averaged over eight model fits.

We split the full time-series sample into three periods of the same length but select the dates randomly for each fold as shown in Figure 2. We keep the same three randomly selected folds throughout our analysis. We use two of the periods to estimate the model and select the tuning parameters, and evaluate the prediction out-of-sample on the remaining third of the sample. We repeat the estimation on three different combinations of the three time periods and report the average results. The estimation and validation time period is split into 3/4 used for training and 1/4 used for validation to select the optimal tuning parameters from the candidate set in Table B.1. For each combination of candidate tuning parameters we train the network for 512 epochs. Our results are robust to the choice of tuning parameters as demonstrated in Section IA.4 of the Internet Appendix. In particular, our results do not depend on the structure of the network and all models with good performance on the validation data provide essentially an identical model with the same relative performance on the test data.

The number of layers of the network is a tuning parameter selected on the validation data. The-

oretically a shallow network with few layers but with more nodes can be equivalent to a deeper network with fewer nodes. Hence, a discussion about the number of layers needs to be related to the number of nodes used in each layer. More layers obviously means more parameters to be estimated and therefore requires either more data or a stronger signal-to-noise ratio to be useful. A panel of individual stock returns is a larger data set and returns–instead of abnormal returns–seem to have a stronger structure to detect. The data on abnormal fund returns is comparatively smaller and seems to have a less complex structure than the data on individual stock returns. Therefore, the structure that can be estimated robustly in our data seems to be simpler. In conclusion, a smaller number of layers provides a parsimonious and robust model for our data.

C Variable Importance and Interaction Effects: Statistical Significance Test

We use the large-sample asymptotic theory of Horel and Giesecke (2020) to study the statistical significance of the measures for Sensitivity(z_k) and Interaction(z_k, macro). Horel and Giesecke (2020) develop a pivotal test to assess the statistical significance of the feature variables in a single-layer feedforward neural network regression model. They study the asymptotics of gradient-based test statistics using nonparametric techniques. Using an empirical-process approach, they show that the large-sample asymptotic distribution of the rescaled neural network sieve estimator is given by the argmax of a Gaussian process. A second-order functional delta method is then used to obtain the asymptotic distribution of the test statistic as the weighted average of the squared partial derivative of the argmax of the Gaussian process with respect to the variable of interest. They show that a test statistics based on the squared partial derivatives can be asymptotically represented by a mixture of chi-squared distributions.

The asymptotic theory in Horel and Giesecke (2020) is developed for a one-layer neural network, which is what we use as the benchmark model in this paper. They apply their method to univariate partial derivatives, while we extend it to also measure interaction effects. The functional delta method arguments directly carry over to our interaction measure, which is simply a linear function of the estimated network. We use the same sensitivity measure as in their paper to study the univariate effects. Their results are derived for sigmoid basis activation functions, while we use ReLU activation function, which can be viewed as an approximation. This approximation is not expected to affect the empirical distribution results. The functional central limit theorems are derived under the assumption that the samples $\{R_{i,t}^{abn}, z_{i,t}, z_t\}_{i,t}$ are i.i.d., which essentially imposes that the error terms from the neural network prediction are independent and identically distributed. Obtaining a distribution theory for a general estimation approach like neural networks requires to impose some assumptions of this form.

The difficulty of the procedure lies in obtaining the asymptotic distribution of the neural network. Our implementation follows the "discretization approach" of Horel and Giesecke (2020), which obtains a cover of the argmax of the Gaussian process by randomly sampling neural networks which approximate the function space. More specifically, we generate random neural networks with the same network structure as the benchmark model, but sample the network parameters randomly.

In more detail, we randomly sample $M = 1,000$ functions f_l with the same network structure as the benchmark setup. The functions are scaled such that the standard deviation of their predictions is the same as for the estimated model $g(z_{i,t})$. In order to generate a sample from the asymptotic distribution, we first generate a random sample from a multivariate normal distribution of dimension $M = 1,000$ with mean 0 and covariance matrix approximated by a diagonal matrix with diagonal elements $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t} f_l(z_{i,t})^2, l = 1 \cdots M$. T is the number of periods and N_t is the number of funds available at time t. We extract the maximum index from this multivariate normal, l^*, and the argmax function h is approximated by f_{l^*}. Given the approximate argmax function f_{l^*}, we generate an approximate sample from the asymptotic distribution of the test statistics: $[\text{Sensitivity}(z_k)]^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}(\frac{\partial f_{l^*}(z_{i,t})}{\partial z_{i,k,t}})^2$ and Interaction(z, macro) = $(f_{l^*}(\text{high z, high macro}) - f_{l^*}(\text{low z, high macro})) - (f_{l^*}(\text{high z, low macro}) - f_{l^*}(\text{low z, low macro}))$. We repeat the process for 1,000 times to estimate the quantile of the test statistics at 1%, 5%, and 10% levels. We obtain the statistical significance levels by comparing the realized test statistics with the properly scaled quantiles. The scaling for Interaction(z, macro) is $r_n = (\frac{n}{log\, n})^{\frac{d+1}{2(2d+1)}}$ while the scaling for $[\text{Sensitivity}(z_k)]^2$ equals r_n^2 , where n is the number of samples and d the number of features in the model. The scaling constants are different because $[\text{Sensitivity}(z_k)]^2$ is a squared function, while Interaction(z, macro) is linear.

This approach is computationally efficient and less expensive than a bootstrap approach. The partial derivative underpinning the test statistic is basically a byproduct of the widely used gradient-based fitting algorithms and is provided by standard software packages used for fitting neural networks. The test procedure does not require re-fitting the neural network. Furthermore, the test is not susceptible to the non-identifiability of neural networks.

Internet Appendix to
Machine-Learning the Skill of Mutual Fund Managers

Ron Kaniel* Zihan Lin† Markus Pelger‡ Stijn Van Nieuwerburgh§

June 16, 2023

*Department of Finance, Simon School of Business, Rochester University, ron.kaniel@simon.rochester.edu.
†Institute for Computational and Mathematical Engineering, Stanford University, zihanl@stanford.edu.
‡Department of Management Science and Engineering, Stanford University, mpelger@stanford.edu.
§Department of Finance, Columbia Business School, svnieuwe@gsb.columbia.edu.

Contents

IA.1 Data Cleaning and Imputation

IA.1.1 Data Cleaning

We make use of the code of Doshi, Elkamhi, and Simutin (2015) for processing mutual fund data. The fund returns, expenses, total net assets (TNA), investment objectives and other fund characteristics are from the Center for Research in Security Prices (CRSP) Survivor Bias-Free Mutual Fund Database. Our analysis requires fund holdings, which we obtain by linking the database to the Thomson Financial Mutual Fund Holdings. The stock characteristics are from Chen, Pelger, and Zhu (2023) and cover 46 characteristics that have been shown to have predictive power for the cross section of expected returns.

We restrict the analysis to diversified domestic active managed equity mutual funds. We use the newly introduced CRSP funds' investment objectives, crsp_obj_cd to define our sample and funds' style categories. The final sample selected with crsp_obj_cd is nearly identical to the one obtained when including funds with AGG, GMC, GRI, GRO, ING, and SCG Strategic Insight codes, EIEI, G, LCCE, LCGE, LCVE, MCCE, MCGE, MCVE, MLCE, MLGE, MLVE, SCCE, SCGE, and SCVE Lipper codes, and G, G-I, AGG, GCI, GRI, GRO, LTG, MCG, and SCG Wiesenberger codes. We screen styles and fund names to exclude international, balanced, sector, bond, money market, and index funds. There are eleven (mutual fund, time) observations for which the raw return is larger than 1, which we also remove.

A fund's total net asset value (TNA) is summed across share classes, and its return, expense ratio, turnover and flow are averages weighted by the lagged asset value of each share class. Fund age is defined as the number of years since the inception of the oldest share class. The fund momentum characteristics are constructed as defined in Table 2.

Following Brown and Wu (2016), fund family is identified by the management company code. For the quarters with a missing company code, we use the mapping between the company name and company code identified in other quarters. For a given fund and month, Family_r12_2 and Family_flow are the averages of F_r12_2 and flow weighted by tna of all funds in the family, excluding the fund itself. Family_age is the age of the oldest fund in the family, excluding the fund itself. Fund_no is the number of funds in the family and Family_tna is the sum of TNAs of all funds in the family excluding the fund itself.

Table IA.1: Summary statistics of fund characteristics

Statistic	N	Mean	St. Dev.	Median
turnover	358,303	0.826	1.015	0.620
ages	407,139	13.669	10.200	11.000
flow (%)	406,661	1.601	419.975	−0.392
r12_2	407,158	0.108	0.173	0.107
LME	407,158	−0.385	0.108	−0.424
BEME	407,158	−0.153	0.376	−0.161
abnormal return (%)	407,158	−0.028	2.000	−0.028
exp_ratio (%)	407,043	0.097	0.086	0.095
TNA	406,802	1,153.180	4,833.920	214.700
Family_TNA	398,655	19,834.410	60,108.400	1,584.700
Family_age	399,011	22.657	19.512	18.000
fund_no	399,011	13.017	18.165	6.000

This table reports summary statistics of the fund characteristics. The sample period is from 1980/01 to 2019/01.

IA.1.2 Data Imputation

We use a panel of monthly firm returns and characteristics from 1963/01 to 2019/12. The data set has in total 30,000 different stocks with around 5,000-9,000 stocks available in each month. However, in each month only around 2,000-3,000 stocks have all 46 characteristics available. In order to use almost all stocks in our sample, we impute the missing characteristic information. This is important as a large number of funds' holdings include stocks that have missing characteristics. We follow the insights of Bryzgalova, Lettau, Lerner, and Pelger (2021) to impute the missing characteristics. More specifically, we apply the cross-sectional factor model advocated by Bryzgalova, Lettau, Lerner, and Pelger (2021), which is a modification of the method developed in Xiong and Pelger (2021) applied to characteristics. Intuitively, each month we estimate a latent factor model with principal components analysis (PCA) in the characteristic space and impute the missing observations as the common components of the factor model. This approach has the advantage that it takes into account the dependency between characteristics, which is not the case with a simple mean or median imputation. Importantly, the method of Xiong and Pelger (2021) allows the missing pattern to depend on the latent factor model or characteristic specific features, which is crucial as the data is not missing at random as shown by Bryzgalova, Lettau, Lerner, and Pelger (2021).

We have a three dimensional array of firm-characteristics $C_{t,l,i}$ which denotes the characteristics l of firm i at time t. We have in total L firm characteristics which we report as cross-sectional quantiles from -0.5 to 0.5. The number of total time periods is T and at time t there are N_t stocks available. We denote by C^t the $L \times N_t$ matrix of L characteristics for the N_t stocks in month t. We assume that the characteristics can be modeled by an approximate K-factor model as in Xiong and

Pelger (2021):

$$C_{i,l}^t = F_i^t {\Lambda_l^t}^\top + e_{i,l}^t$$

Without missing values the latent factors and loadings can be estimated by PCA from the "characteristics covariance matrix" $\Sigma^t = \frac{1}{N_t} C^t {C^t}^\top$. In the presence of missing values, we use the method of Xiong and Pelger (2021) to estimate the latent factors. More specifically, we estimate the $L \times L$ matrix Σ_t as

$$\Sigma_{l,r}^t = \frac{1}{|Q_{l,r}|} \sum_{i \in Q_{l,r}} C_{l,i}^t C_{r,i}^t, \tag{IA.1}$$

where $Q_{l,r}$ are the indices of all the stocks that have characteristics l and r in common at time t and $|Q_{l,r}|$ is the cardinality of this set. The Λ^t are estimated with PCA as the normalized eigenvectors of Σ^t. Last but not least, we estimate the "characteristic factors" with a weighted regression

$$\underbrace{\hat{F}_i^t}_{K \times 1} = \left(\sum_{l \in Q_{i,t}} \hat{\Lambda}_l^t \hat{\Lambda}_l^{t\top} \right)^{-1} \sum_{l \in Q_{i,t}} \hat{\Lambda}_l^t C_{l,i}^t.$$

Given the estimated factors and loadings, the missing values are imputed with $\hat{C}_{l,i}^t = \hat{F}_i^t \hat{\Lambda}_l^{t\top}$.

The factor imputation depends on two parameters. First, we select the number $L_{\min}$ of characteristics that a stock needs to have in order to be included in the sample. By construction this is an upper bound on the number of latent factors. Second, we select the number of latent factors K. Based on an extensive analysis we have set $L_{min} = 10$ and $K = 10$ as the benchmark model. The results are robust to this choice, while it allows us to include almost all stocks in the sample.

We assess the accuracy of the data imputation method based on how well the characteristic factor model approximates the observed characteristic entries. We measure the amount of explained variation by the following R^2 using only the observed entries:

$$R^2 = 1 - \frac{\sum_{t,l,i} e_{t,l,i}^2}{\sum_{t,l,i} C_{t,l,i}^2}.$$

Table IA.2 shows that a model with $K = 10$ factors explains around 75% of the cross-sectional variation in characteristics. Note that the first latent factor is very close to a cross-sectional average for each characteristic. Hence, a one-factor model is essentially imputation with a cross-sectional median, which is strongly suboptimal.

Table IA.3 shows that after the data imputation we have full characteristic information for almost 99% of the stocks held by mutual funds. Without the data imputation, we could only observe the characteristics for around 57% of the stocks held by mutual funds. The choice of

L_{min} trades off the accuracy of the imputation and the goal of keeping as many stocks as possible. The extensive empirical study in Bryzgalova, Lettau, Lerner, and Pelger (2021) provides further support for this imputation model.

Table IA.2: R^2 of Factor Model

K	2	4	6	8	10
R^2	34.7%	50.7%	60.6%	68.4%	74.5%

This table shows the time average of R^2 of the characteristics imputation for different number of factors K. We require at least $L_{min} = 10$ observed characteristics for each stock.

Table IA.3: Proportion of Missing Characteristics for Different L_{min}

L_{min}	Missing (i, j, t) No.	Proportion of total observations
8	565,663	0.98%
10	744,087	1.29%
12	1,794,131	3.12%
14	2,025,437	3.52%
46, original characteristic data	24,596,754	42.80%

This table shows the number of missing characteristics observations and the corresponding proportion relative to the total number of entries. The results are summed over mutual funds, stocks and time.

IA.2 In-Sample Fund Performance By Univariate Characteristic

We explore which fund characteristics are associated with strong mutual fund performance in a univariate analysis. This analysis is not a substitute for our full machine-learning analysis since it is (i) in-sample rather than predictive, (ii) it ignores the possibility of important non-linearities in the relationship between fund characteristics and fund abnormal returns, and (iii) it ignores the possibility of important interaction effects between multiple characteristics or between characteristics and macro variables.

For each of the 59 characteristics, we sort fund abnormal returns into deciles based on the value of the characteristic. Then, we construct long-short portfolios as the difference between the top and bottom deciles. The first two columns of Table IA.4 report the mean and Sharpe ratio of these long-short portfolio returns, ranked from highest to lowest Sharpe ratio. The stars report the significance of a test that the mean of the long-short portfolio return is different from zero. The main finding, which foreshadows the results in our main analysis, is that portfolios based on *fund characteristics*, and in particular fund momentum and flow, are associated with the highest Sharpe ratios as well as a large and statistically significant mean abnormal fund return. Of the ten characteristics that are associated with a monthly Sharpe ratio above 0.10, seven are fund-level variables and only three are stock characteristics. Most *stock-specific* characteristics cannot

systematically differentiate between the performance of mutual funds. Put simply, little can be learned about fund abnormal returns from the stocks that they hold.

IA.2.1 Univariate Groups

In Table IA.5, we show that the pre-eminence of fund characteristics as (univariate) drivers of fund performance also emerges when we collapse characteristics by group (using the nine groups in Table 1). Group characteristics are formed as the equally-weighted average of the characteristics within each category, and then long-short portfolios of funds are formed based on the deciles of the group characteristics. Fund momentum rises to the top as the group characteristic that is associated most strongly with fund performance (Sharpe ratio of 0.22), followed by Family characteristics (SR of 0.16), and Fund characteristics (0.13). The holdings-based characteristics Value (0.16) and Profitability (0.11) are in third and fifth place. Thus, the last five places out of nine are reserved for stock characteristics.

Table IA.5: Group characteristic long-short factors from abnormal returns

	mean (%)	std (%)	SR	t-stat
fund momentum	0.33	1.45	0.22	4.9***
family	0.12	0.74	0.16	3.5***
value	0.18	1.11	0.16	3.4***
fund	0.09	0.71	0.13	2.8***
profitability	0.13	1.02	0.13	2.7***
friction	0.17	1.72	0.10	2.1**
intangile	0.08	0.88	0.09	1.9*
past	0.11	1.39	0.08	1.7*
investment	0.06	0.93	0.06	1.3

This table reports the summary statistics for univariate long-short factors based on group-averaged abnormal returns and sorted according to their Sharpe ratios. For each of the 9 categories, we construct group characteristics as the equally-weighted average of the characteristics within each category. The long-short factors are the differences between the top decile and the bottom decile. "Mean", "std" and "Sharpe ratio" report the mean, standard deviation and Sharpe ratio of the factors and the fourth column, "t-stat" denotes the t-statistics for a test that the factor mean is different from 0.

IA.2.2 Interaction Effects with Macro Variables

The univariate analysis confirms a second main result of the paper, which is that fund characteristics exercise a different influence on fund performance depending on the state of the economy. The last six columns of Table IA.4 report Sharpe ratios and mean returns of the same univariate long-short portfolios, but conditional on the level of investor sentiment. The sample is split into terciles based on the value of the sentiment index in the prior month. We find that the strong association between abnormal performance and fund characteristics such as fund momentum, fund short-run reversal, and flow is driven by above-average sentiment periods. This hints at impor-

Table IA.4: Univariate long-short portfolios from mutual fund abnormal returns

	Full sample		Low sentiment		Medium sentiment		High sentiment	
	SR	mean (%)	SR	mean (%)	SR	mean (%)	SR	mean (%)
F_r12_2	0.28	0.36***	0.16	0.19*	0.46	0.50***	0.25	0.40***
F_ST_Rev	0.20	0.30***	0.11	0.15	0.22	0.28***	0.26	0.49***
Family_r12_2	0.19	0.13***	0.24	0.13***	0.31	0.21***	0.06	0.05
Beta	0.15	0.18***	0.17	0.19**	0.06	0.06	0.23	0.31***
Rel2High	0.14	0.20***	0.06	0.08	0.20	0.24**	0.18	0.31**
RNA	0.13	0.13***	0.16	0.14*	0.14	0.11*	0.10	0.13
Family_TNA	0.13	0.09***	0.01	0.01	0.24	0.15***	0.13	0.12
fund_no	0.13	0.10***	0.03	0.01	0.24	0.17***	0.11	0.11
flow	0.12	0.11**	0.21	0.15**	0.11	0.09	0.08	0.09
Family_age	0.11	0.09**	0.01	0.01	0.27	0.17***	0.08	0.08
ROA	0.10	0.10**	0.16	0.15*	0.05	0.04	0.11	0.13
PM	0.10	0.10**	0.11	0.10	0.19	0.13**	0.07	0.09
ROE	0.10	0.11**	0.13	0.12	0.04	0.04	0.13	0.18
ST_Rev	0.09	0.13**	0.04	0.06	0.09	0.11	0.15	0.24*
CF	0.09	0.09**	0.08	0.07	0.02	0.01	0.15	0.21*
Resid_Var	0.09	0.14**	0.11	0.15	0.10	0.13	0.09	0.18
ages	0.09	0.05**	0.10	0.05	0.10	0.04	0.09	0.07
MktBeta	0.08	0.14**	0.09	0.16	0.13	0.16	0.06	0.12
r12_2	0.08	0.11**	0.02	0.02	0.17	0.18**	0.06	0.11
Spread	0.08	0.13**	0.11	0.16	0.06	0.07	0.10	0.21
D2P	0.08	0.12**	0.04	0.06	-0.04	-0.05	0.19	0.32**
r12_7	0.08	0.11**	0.08	0.11	0.10	0.10	0.05	0.10
F_r2_1	0.08	0.11	0.02	0.02	0.23	0.27***	0.02	0.04
LTurnover	0.07	0.13	0.13	0.24	-0.00	-0.00	0.09	0.20
Variance	0.07	0.13	0.11	0.17	0.03	0.04	0.10	0.21
IdioVol	0.07	0.12	0.09	0.13	0.09	0.11	0.07	0.15
C	0.07	0.09	0.03	0.04	0.11	0.11	0.06	0.08
Lev	0.07	0.08	0.08	0.09	0.05	0.05	0.05	0.07
Family_flow	0.07	0.04	0.09	0.04	0.06	0.03	0.07	0.06
ATO	0.07	0.07	0.02	0.01	0.01	0.01	0.12	0.16
exp_ratio	0.07	0.04	0.12	0.06	0.01	0.01	0.07	0.06
CTO	0.06	0.07	0.03	0.02	0.05	0.04	0.08	0.11
tna	-0.06	-0.04	-0.03	-0.02	-0.19	-0.11**	-0.00	-0.00
SUV	0.06	0.06	0.15	0.14*	-0.06	-0.05	0.08	0.10
SGA2S	0.05	0.06	0.11	0.12	0.04	0.04	-0.01	-0.01
OL	0.05	0.05	0.02	0.01	0.02	0.02	0.07	0.08
PCM	0.05	0.05	0.15	0.14*	0.05	0.04	-0.04	-0.05
r2_1	0.05	0.06	0.02	0.02	0.06	0.07	0.07	0.13
CF2P	0.04	0.06	0.03	0.05	0.09	0.10	-0.00	-0.01
NI	0.04	0.05	0.12	0.17	-0.06	-0.05	0.05	0.08
Q	0.04	0.05	0.05	0.07	0.01	0.01	0.03	0.04
FC2Y	0.04	0.05	0.10	0.12	0.04	0.03	-0.02	-0.03
PROF	0.04	0.04	0.11	0.08	0.01	0.01	0.01	0.02
LME	0.04	0.03	0.01	0.01	-0.03	-0.02	0.11	0.11
D2A	0.03	0.03	-0.05	-0.04	0.03	0.02	0.10	0.10
turnover	0.03	0.03	0.03	0.02	-0.03	-0.02	0.08	0.08
AT	0.03	0.04	0.02	0.02	-0.04	-0.03	0.06	0.09
OA	0.03	0.03	-0.02	-0.01	0.08	0.06	0.05	0.06
r36_13	-0.03	-0.04	0.02	0.03	0.01	0.01	-0.09	-0.13
AC	0.03	0.02	-0.01	-0.01	0.08	0.06	0.04	0.04
E2P	-0.03	-0.04	0.05	0.06	-0.05	-0.06	-0.04	-0.07
OP	0.02	0.03	0.12	0.10	-0.02	-0.02	0.01	0.02
NOA	0.02	0.02	-0.03	-0.02	0.18	0.13**	-0.03	-0.04
DPI2A	0.01	0.01	0.00	0.00	0.19	0.14**	-0.08	-0.09
A2ME	0.01	0.02	0.07	0.11	0.03	0.03	-0.07	-0.12
S2P	-0.01	-0.02	-0.07	-0.10	-0.07	-0.07	0.10	0.16
Investment	0.01	0.01	-0.01	-0.01	0.12	0.09	-0.05	-0.06
LT_Rev	0.01	0.01	0.08	0.08	0.03	0.03	-0.04	-0.06
BEME	-0.00	-0.00	-0.07	-0.09	-0.03	-0.02	0.09	0.13

This table reports summary statistics for univariate long-short portfolios in the full sample and in different sentiment terciles. The results are sorted according to the Sharpe ratio of univariate long-short factors. For each of the 59 characteristics, we construct ten sorted decile portfolios. The long-short factors are the differences between the top decile and the bottom decile. We split the time-series into T^H, T^M and T^L, which denotes the high, medium and low sentiment terciles based on the previous months. For each of these time periods we separately estimate "Mean" and "Sharpe ratio", which report the mean and Sharpe ratio of this factor conditional on a sentiment tercile. The stars are the signifance of t-statistics for the test that the factor mean is different from 0. The sample period is 1980/01 to 2019/01 with monthly rebalancing.

tant interaction effects between macro variables and (fund) characteristics. Table IA.6 shows that this result also applies to the first principal component of each group.

Table IA.6: First Principal Component within each Characteristic Group in Different Sentiment Terciles

	T^L		T^M		T^H	
Name	SR	t-stat	SR	t-stat	SR	t-stat
fund momentum	0.02	0.3	0.30	5.0***	0.23	4.6***
family	0.08	1.0	0.15	1.8*	0.16	2.5**
profitability	0.01	0.1	0.20	1.9**	0.17	2.5**
fund	0.10	0.9	0.04	0.4	0.16	2.1**
past	-0.07	-0.6	0.24	1.7***	0.12	1.3
friction	0.02	0.2	0.01	0.1	0.18	2.1**
value	0.04	0.1	0.19	0.2**	-0.12	-0.2
intangile	-0.00	-0.0	-0.08	-0.1	0.12	0.2
investment	-0.02	-0.0	-0.10	-0.1	0.12	0.1

This table reports the mean, standard deviation and Sharpe ratio for different sentiment terciles. For each of the 9 categories, we apply PCA to the abnormal returns of all long-short factors within a group to obtain a group factor. We split the time-series into T^H, T^M and T^L, which denotes the high, medium and low sentiment terciles based on the previous months. For each of these time periods we separately estimate the mean, standard deviation and Sharpe ratio of the factors conditional on a sentiment tercile. The column, "t-stat" denotes the t-statistics for factor mean different from 0 in each sentiment tercile.

Table IA.7 shows the results when when using the economic activity variable CFNAI as the conditioning variable.

Table IA.7: Univariate long-short portfolios in different CFNAI terciles

	Low CFNAI		Medium CFNAI		High CFNAI	
	SR	mean (%)	SR	mean (%)	SR	mean (%)
F_r12_2	0.26	0.39***	0.34	0.40***	0.25	0.30***
F_ST_Rev	0.17	0.31**	0.19	0.26**	0.25	0.34***
Family_r12_2	0.28	0.23***	0.13	0.08*	0.13	0.08
Beta	0.17	0.24**	0.12	0.13	0.17	0.18**
Rel2High	0.07	0.14	0.22	0.25***	0.20	0.21**
RNA	0.18	0.21**	0.17	0.12**	0.05	0.05
Family_TNA	0.04	0.04	0.26	0.16***	0.11	0.08
fund_no	0.02	0.02	0.25	0.15***	0.16	0.12*
flow	0.03	0.04	0.15	0.13*	0.18	0.15**
Family_age	0.04	0.04	0.27	0.15***	0.10	0.07
ROA	0.17	0.20**	0.09	0.07	0.04	0.04
PM	0.21	0.24**	0.13	0.11*	-0.04	-0.04
ROE	0.19	0.23**	0.08	0.07	0.02	0.02
ST_Rev	0.09	0.16	0.15	0.18*	0.05	0.06
CF	0.03	0.04	0.19	0.16**	0.08	0.08
Resid_Var	0.11	0.23	0.06	0.08	0.08	0.10
ages	-0.00	-0.00	0.06	0.04	0.19	0.12**
MktBeta	0.12	0.25	0.06	0.09	0.06	0.09
r12_2	-0.03	-0.05	0.25	0.27***	0.09	0.12
Spread	0.13	0.26	0.06	0.08	0.04	0.07
D2P	0.12	0.22	0.03	0.04	0.08	0.10
r12_7	-0.05	-0.10	0.20	0.23**	0.14	0.19*
F_r2_1	-0.01	-0.03	0.19	0.23**	0.09	0.12
LTurnover	0.12	0.25	-0.01	-0.01	0.10	0.17
Variance	0.10	0.20	0.03	0.05	0.09	0.13
IdioVol	0.11	0.22	0.04	0.05	0.06	0.08
C	0.02	0.04	0.24	0.27***	-0.04	-0.04
Lev	0.07	0.11	0.03	0.03	0.11	0.11
Family_flow	0.09	0.05	0.06	0.03	0.06	0.04
ATO	-0.01	-0.02	0.07	0.08	0.17	0.14**
exp_ratio	0.06	0.04	0.16	0.09**	-0.00	-0.00
CTO	-0.01	-0.02	0.07	0.07	0.16	0.14*
tna	-0.06	-0.05	-0.06	-0.03	-0.06	-0.05
SUV	0.01	0.01	0.12	0.12	0.04	0.04
SGA2S	0.01	0.02	0.18	0.18**	-0.02	-0.02
OL	0.02	0.02	0.08	0.07	0.05	0.04
PCM	-0.02	-0.02	0.15	0.13*	0.03	0.03
r2_1	0.02	0.03	0.12	0.13	0.03	0.03
CF2P	0.05	0.08	0.06	0.08	0.02	0.02
NI	0.09	0.13	-0.09	-0.11	0.14	0.15*
Q	0.04	0.07	0.04	0.04	0.05	0.05
FC2Y	0.01	0.01	0.14	0.15*	-0.01	-0.02
PROF	0.00	0.00	0.09	0.06	0.06	0.05
LME	0.11	0.12	0.00	0.00	-0.03	-0.02
D2A	0.06	0.05	-0.00	-0.00	0.05	0.04
turnover	0.03	0.03	0.04	0.02	0.03	0.03
AT	0.06	0.09	0.02	0.01	0.01	0.01
OA	0.14	0.14*	0.11	0.08	-0.16	-0.13**
r36_13	0.08	0.12	-0.18	-0.19**	-0.04	-0.04
AC	0.12	0.12	0.12	0.08	-0.16	-0.12**
E2P	-0.03	-0.04	-0.04	-0.05	-0.01	-0.02
OP	-0.01	-0.01	-0.02	-0.02	0.13	0.11
NOA	-0.02	-0.02	0.12	0.09	-0.02	-0.01
DPI2A	-0.01	-0.01	0.13	0.11*	-0.06	-0.06
A2ME	-0.07	-0.12	0.09	0.10	0.05	0.06
S2P	0.05	0.09	-0.08	-0.10	-0.02	-0.03
Investment	-0.05	-0.07	0.10	0.09	0.00	0.01
LT_Rev	0.09	0.13	-0.13	-0.11	0.02	0.02
BEME	0.11	0.14	-0.10	-0.11	-0.03	-0.04

This table reports summary statistics for univariate long-short factors in different CFNAI terciles. The results are sorted according to the Sharpe ratio of univariate long-short factors. For each of the 59 characteristics, we construct ten sorted decile portfolios. The long-short factors are the differences between the top decile and the bottom decile. We split the time-series into T^H, T^M and T^L, which denotes the high, medium and low CFNAI terciles based on the previous months. For each of these time periods we separately estimate "Mean" and "Sharpe ratio", which report the mean and Sharpe ratio of this factor conditional on a CFNAI tercile. The stars are the signifance of t-statistics for the test that factor mean is different from 0.

IA.2.3 Spanning

Next, we show that the long-short portfolios that are most strongly associated with fund outperformance do not simply reflect compensation for exposure to standard risk factors. To that end, we estimate multivariate regressions of the long-short portfolios based on fund characteristics on the four Carhart factors and an intercept. The results are in Table IA.8. All R^2 are small (below 10%) except for fund momentum (23%). Most of the factor loadings are also insignificant. Importantly, a vast majority of alphas are highly significant, meaning that the returns on long-short portfolios are not spanned by the equity asset pricing factors. We also report the mean return on the long-short portfolios. In the few cases where the alpha is not significant, the risk premium of the portfolios is typically not significant. For most fund variables, the mean return and the mean intercept are similar in magnitude, which implies that the Carhart factor exposure explains little of the abnormal returns of fund characteristic portfolios.

Table IA.8: Spanning of univariate long-short portfolios with FFC-4 factors.

	Mkr	SMB	HML	Mom	α	Factor mean	R^2
F_ST_Rev	-0.09*	0.07	0.08*	0.13***	0.18***	0.20***	0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
F_r2_1	0.10**	-0.05	0.05	0.34***	0.01	0.08	0.11
	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	
F_r12_2	0.29***	0.04	0.11**	0.44***	0.17***	0.28***	0.23
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	
ages	-0.00	-0.11**	0.07	0.06	0.08	0.09*	0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
flow	0.12**	-0.10**	0.03	0.03	0.10**	0.12**	0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
exp_ratio	-0.07	0.02	-0.08*	-0.20***	0.11**	0.07	0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
tna	-0.09*	-0.03	0.03	0.05	-0.05	-0.06	0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
turnover	-0.00	0.04	0.03	0.06	0.02	0.03	0.01
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
Family_TNA	0.11**	-0.03	-0.06	-0.07	0.13***	0.13***	0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
fund_no	0.17***	0.02	0.03	0.08	0.09*	0.13***	0.03
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
Family_r12_2	0.10*	0.04	0.10**	0.21***	0.14***	0.19***	0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
Family_age	0.18***	0.01	0.00	0.07	0.08	0.11**	0.03
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	
Family_flow	-0.04	0.04	-0.01	-0.06	0.08*	0.07	0.01
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	

This table reports the multivariate time-series regression results of univariate long-short abnormal return portfolios on the four Fama-French-Carhart factors. The first four columns report the slope coefficient on the FFC-4 factors in the regression and the fifth column, α reports the time-series pricing error of the regression. The last two columns report the mean of the univariate long-short decile portfolios and the R^2 of the regression. Both, the univariate long-short abnormal return portfolios and FFC-4 factors, are normalized to have a standard deviation of 1.

Table IA.9 shows that these results are robust to using alternative factor models: the Fama-French three-factor model, the Fama-French five-factor model (with investment and profitability factors), a six-factor model that adds momentum to the Fama-French five factors, and an eight-factor model that adds short-term reversal, long-term reversal, and momentum to the Fama-French five factors.

Table IA.9: Spanning of univariate long-short portfolios with different factor models.

	4 factors		5 factors		6 factors		8 factors		
	α	R^2	α	R^2	α	R^2	α	R^2	mean
F_ST_Rev	0.18***	0.04	0.19***	0.02	0.18***	0.04	0.23***	0.31	0.20***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
F_r2_1	0.01	0.11	0.03	0.03	-0.01	0.12	0.01	0.17	0.08
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
F_r12_2	0.17***	0.23	0.22***	0.06	0.16***	0.23	0.16***	0.23	0.28***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.05)
ages	0.08	0.02	0.06	0.03	0.05	0.03	0.06	0.03	0.09*
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
flow	0.10**	0.02	0.08	0.03	0.08	0.03	0.07	0.03	0.12**
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
exp_ratio	0.11**	0.04	0.14***	0.07	0.16***	0.09	0.16***	0.11	0.07
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
tna	-0.05	0.02	-0.05	0.02	-0.06	0.02	-0.07	0.04	-0.06
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
turnover	0.02	0.01	-0.05	0.08	-0.05	0.08	-0.07	0.12	0.03
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Family_TNA	0.13***	0.02	0.16***	0.05	0.16***	0.05	0.18***	0.07	0.13***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
fund_no	0.09*	0.03	0.12**	0.03	0.11**	0.04	0.13***	0.07	0.13***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Family_r12_2	0.14***	0.04	0.17***	0.01	0.15***	0.05	0.14***	0.05	0.19***
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Family_age	0.08	0.03	0.11**	0.04	0.10**	0.04	0.11**	0.07	0.11**
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)
Family_flow	0.08*	0.01	0.07	0.00	0.08	0.01	0.07	0.02	0.07
	(0.05)		(0.05)		(0.05)		(0.05)		(0.05)

This table reports the multivariate time-series regression results of univariate long-short abnormal return portfolios for different factor models. We consider the 4-factor Fama-French-Carhart model (market, size, value and momentum), the 5-factor Fama-French model (market, size, value, profitability and investment), a 6-factor model which adds the momentum factor to the Fama-French 5 factors, and an 8-factor model which adds the momentum, short-term reversal and long-term reversal factors to the Fama-French 5 factors. The α column reports the time-series pricing error and R^2 is the explained variation of the regression. Both the univariate long-short abnormal return portfolios and the factor models are normalized to have a standard deviation of 1. Standard errors are in brackets and stars denote the significance levels.

IA.2.4 Holding Period

The univariate results are robust to alternative holding-period assumptions. This is germane, as mutual funds tend to be held by investors for longer periods than one month. We consider three robustness tests: (i) we update characteristics annually, (ii) we hold each fund investment for one

year with overlapping returns, and (iii) we consider quarterly returns of quarterly updated positions. For all three, as for monthly returns, portfolios sorted on fund-specific characteristics—in particular flow and fund momentum—have the highest Sharpe ratio and a statistically significant mean, while most stock-specific characteristics cannot systematically differentiate the performance of funds. These results show that monthly rebalancing of mutual funds is not crucial to earning the high abnormal returns associated with strongly-performing funds. We study even longer holding periods in our main analysis below.

First, we construct annually updated monthly abnormal return portfolios. Specifically, we sort funds into portfolios only once a year. We keep the weights constant in the sorting portfolios for 12 months. We use the December characteristics to obtain the decile sorting allocations and update these allocations each December. Then we compute the monthly abnormal returns as before. We report the univariate sorting results in Panel (a) of Table IA.10. Second, we consider overlapping abnormal returns with one-year holding periods. At each time $t-1$, investors use the current month characteristics to form portfolios and hold the portfolio for one year. The overlapping holding results are in Panel (b). Last but not least, we consider the abnormal return of investors who rebalance their portfolios every quarter. They rebalance fund portfolios using the fund characteristics at the end of each quarter and hold the same portfolio for one quarter. These returns are not overlapping. They are in Panel (c). For all three specifications, we observe that the Sharpe ratios and average abnormal returns are comparable to the results in Table IA.4. The fund-specific characteristics, in particular flow and fund momentum, have the highest Sharpe ratio and a statistically significant mean, while most stock-specific characteristics are not significant.

Table IA.10: Univariate long-short portfolios from abnormal returns at different holding frequency

	(a) Annually updated characteristics				(b) Overlapping annual holdings				(c) Quarterly			
	mean	std	SR	t-stat	mean	std	SR	t-stat	mean	std	SR	t-stat
F_r12_2	0.23	1.16	0.20	4.3***	0.32	2.35	0.24	5.1***	0.20	12.82	0.19	4.0***
F_ST_Rev	0.18	1.18	0.16	3.4***	0.19	2.71	0.12	2.7***	0.17	8.86	0.23	4.9***
Family_r12_2	0.06	0.84	0.07	1.5	0.10	1.39	0.13	2.8***	0.05	8.29	0.07	1.4
Beta	0.24	1.13	0.21	4.6***	0.20	1.98	0.18	3.8***	0.18	9.71	0.23	4.8***
Rel2High	0.19	1.29	0.15	3.2***	0.26	2.29	0.20	4.2***	0.20	13.55	0.18	3.8***
RNA	0.09	0.88	0.10	2.2**	0.13	1.74	0.13	2.8***	0.07	7.36	0.12	2.6**
Family_TNA	0.08	0.73	0.12	2.5**	0.10	1.23	0.14	3.0***	0.09	8.20	0.13	2.6***
fund_no	0.07	0.73	0.10	2.2**	0.10	1.24	0.14	3.0***	0.09	7.54	0.14	3.0***
flow	0.16	0.93	0.17	3.6***	0.16	1.33	0.20	4.4***	0.05	4.71	0.12	2.6***
Family_age	0.09	0.76	0.12	2.6***	0.10	1.26	0.14	3.0***	0.08	7.54	0.12	2.5**
ROA	0.07	0.92	0.08	1.7*	0.10	1.79	0.10	2.2**	0.07	6.85	0.12	2.5**
PM	0.06	0.94	0.06	1.3	0.14	1.64	0.14	3.1***	0.08	6.48	0.14	2.9***
ROE	0.06	0.97	0.06	1.4	0.09	1.73	0.09	2.0**	0.07	7.56	0.11	2.3**
ST_Rev	0.08	1.31	0.06	1.3	0.11	2.59	0.07	1.6	0.09	7.46	0.15	3.1***
CF	0.02	0.97	0.02	0.5	0.05	1.92	0.04	0.9	0.09	10.98	0.10	2.1**
Resid_Var	0.15	1.57	0.09	2.0**	0.20	2.67	0.13	2.8***	0.18	18.18	0.12	2.4**
ages	0.05	0.60	0.08	1.7*	0.05	1.07	0.08	1.6	0.05	6.35	0.10	2.1**
MktBeta	0.12	1.65	0.07	1.5	0.13	2.79	0.08	1.7*	0.18	17.37	0.13	2.7***
r12_2	0.04	1.30	0.03	0.7	0.10	2.77	0.06	1.4	0.03	13.73	0.02	0.5
Spread	0.14	1.58	0.09	1.9*	0.18	2.79	0.11	2.5**	0.16	17.85	0.11	2.3**
D2P	0.10	1.46	0.07	1.5	0.14	2.41	0.10	2.2**	0.15	15.23	0.12	2.4**
r12_7	0.04	1.40	0.03	0.6	0.09	2.64	0.06	1.3	-0.01	12.64	-0.01	-0.2
F_r2_1	0.14	1.27	0.11	2.5**	0.16	2.38	0.12	2.6**	0.14	8.61	0.20	4.2***
LTurnover	0.17	1.73	0.10	2.1**	0.18	2.88	0.11	2.3**	0.22	19.92	0.13	2.8***
Variance	0.16	1.62	0.10	2.1**	0.20	2.83	0.12	2.6**	0.20	19.48	0.12	2.6**
IdioVol	0.13	1.59	0.08	1.7*	0.19	2.66	0.13	2.7***	0.17	18.21	0.11	2.4**
C	-0.01	1.23	-0.01	-0.2	0.05	2.37	0.03	0.7	-0.02	12.20	-0.02	-0.4
Lev	0.10	1.20	0.08	1.7*	0.06	2.18	0.05	1.1	0.06	10.74	0.07	1.4
Family_flow	0.04	0.66	0.07	1.4	0.01	1.10	0.02	0.4	0.01	3.54	0.04	0.8
ATO	0.03	1.04	0.03	0.6	0.11	1.84	0.10	2.1**	0.07	9.14	0.09	1.9*
exp_ratio	0.04	0.73	0.05	1.1	-0.01	1.17	-0.01	-0.2	-0.04	6.50	-0.07	-1.5
CTO	0.05	1.02	0.05	1.0	0.09	1.74	0.09	1.9*	0.07	7.71	0.11	2.4**
tna	-0.02	0.77	-0.03	-0.7	-0.06	1.14	-0.09	-2.0*	-0.01	6.70	-0.01	-0.3
SUV	0.07	1.02	0.07	1.5	0.07	1.74	0.07	1.5	0.00	5.76	0.00	0.1
SGA2S	0.04	0.95	0.04	0.9	0.06	2.13	0.05	1.0	0.04	11.44	0.05	1.0
OL	0.05	0.97	0.05	1.1	0.09	1.71	0.09	1.9*	0.08	8.77	0.11	2.2**
PCM	-0.01	0.90	-0.01	-0.2	0.01	1.88	0.01	0.3	0.02	7.48	0.03	0.6
r2_1	0.11	1.41	0.07	1.6	0.11	2.42	0.08	1.6	0.07	7.68	0.12	2.4**
CF2P	0.03	1.34	0.03	0.6	0.03	2.60	0.02	0.4	-0.01	13.92	-0.01	-0.2
NI	0.02	1.10	0.02	0.3	0.04	2.10	0.03	0.6	0.08	11.23	0.09	1.8*
Q	0.06	1.22	0.05	1.0	0.07	2.42	0.05	1.0	-0.01	11.26	-0.01	-0.1
FC2Y	0.04	0.98	0.04	0.8	0.06	2.24	0.05	1.0	0.04	11.67	0.04	0.8
PROF	0.07	0.86	0.08	1.7*	0.04	1.90	0.04	0.8	0.11	8.71	0.15	3.2***
LME	0.05	0.87	0.06	1.2	-0.01	1.21	-0.01	-0.2	0.01	5.89	0.01	0.2
D2A	-0.03	0.82	-0.03	-0.7	-0.01	1.46	-0.01	-0.1	-0.02	6.23	-0.04	-0.9
turnover	0.05	0.92	0.05	1.2	0.03	1.73	0.03	0.6	0.07	12.73	0.07	1.4
AT	0.07	1.11	0.06	1.4	0.00	1.82	0.00	0.1	-0.01	8.89	-0.02	-0.4
OA	0.00	0.73	0.00	0.1	0.02	1.48	0.02	0.5	0.01	5.30	0.02	0.5
r36_13	-0.02	1.04	-0.02	-0.4	-0.02	2.03	-0.02	-0.3	0.02	11.73	0.02	0.4
AC	0.01	0.71	0.01	0.3	0.03	1.43	0.04	0.8	0.01	5.02	0.02	0.4
E2P	-0.03	1.33	-0.02	-0.5	-0.00	2.47	-0.00	-0.0	0.06	14.70	0.05	0.9
OP	0.04	0.96	0.04	0.8	0.03	1.91	0.03	0.6	0.08	9.40	0.10	2.2**
NOA	-0.02	0.92	-0.02	-0.5	-0.03	1.39	-0.04	-0.8	-0.02	6.17	-0.04	-0.8

	mean	std	SR	t-stat	mean	std	SR	t-stat	mean	std	SR	t-stat
DPI2A	0.04	0.93	0.04	0.8	0.08	1.83	0.07	1.6	-0.01	6.27	-0.02	-0.4
A2ME	-0.03	1.13	-0.02	-0.5	0.07	2.38	0.05	1.1	0.01	11.54	0.01	0.3
S2P	0.01	1.27	0.01	0.2	-0.02	2.48	-0.02	-0.4	0.02	13.55	0.02	0.4
Investment	0.08	1.01	0.08	1.6	0.07	2.00	0.06	1.3	0.02	7.40	0.03	0.6
LT_Rev	0.02	0.98	0.02	0.5	0.04	1.71	0.04	0.9	0.04	9.41	0.05	1.1
BEME	0.05	1.11	0.05	1.1	-0.03	2.34	-0.02	-0.5	0.00	11.50	0.00	0.1

This table reports summary statistics for univariate long-short factors based on fund abnormal returns. Mean and std of abnormal returns are reported in percentages. The results are sorted according to the Sharpe ratio of univariate long-short factors based on abnormal returns, in the same order as in Table IA.4. In panel (a), the characteristics of mutual funds are updated in December of each year and kept constant throughout the year. Panel (b) reports the results for overlapping annual returns. The standard deviation and t-statistics are reported with Newey-West correction for time-series correlation with 12 lags. Panel (c) displays the results for non-overlapping quarterly abnormal return. For each of the 59 characteristics, we construct ten sorted decile portfolios based on fund abnormal returns. The long-short factors are the differences between the top decile and the bottom decile. The fourth column, "t" reports the t-statistics for a test that the factor mean is different from 0 and stars denote the significance levels.

IA.2.5 Decomposing Abnormal Returns

Table IA.11: Decomposition of univariate long-short portfolios from mutual fund abnormal returns

	Total		Between-disclosure		Within-disclosure		Risk difference		Return gap	
	SR	mean	SR	mean	SR	mean	SR	mean	SR	mean
F_r12_2	0.28	0.36***	0.14	0.20***	0.20	0.17***	0.14	0.11***	0.10	0.06***
F_ST_Rev	0.20	0.30***	0.14	0.16***	0.15	0.15***	0.12	0.08**	0.12	0.06***
Family_r12_2	0.19	0.13***	0.10	0.09***	0.09	0.04**	0.12	0.07**	-0.06	-0.03
Beta	0.15	0.18***	0.12	0.16***	0.03	0.03	-0.01	-0.00	0.05	0.03
Rel2High	0.14	0.20***	0.13	0.25***	-0.05	-0.05	-0.03	-0.03	-0.03	-0.03
RNA	0.13	0.13***	0.11	0.12***	0.01	0.01	-0.03	-0.02	0.04	0.02
Family_TNA	0.13	0.09***	0.09	0.07	0.05	0.03	-0.12	-0.06**	0.16	0.08***
fund_no	0.13	0.10***	0.10	0.07**	0.06	0.03	-0.12	-0.05**	0.14	0.07***
flow	0.12	0.11**	0.08	0.08**	0.06	0.03	-0.00	-0.00	0.08	0.03**
Family_age	0.11	0.09**	0.08	0.07	0.03	0.02	-0.13	-0.06**	0.13	0.08***
ROA	0.10	0.10**	0.11	0.13***	-0.03	-0.03	-0.05	-0.03	0.01	0.01
PM	0.10	0.10**	0.10	0.11**	-0.01	-0.01	-0.03	-0.02	0.02	0.01
ROE	0.10	0.11**	0.09	0.12**	-0.01	-0.01	-0.02	-0.01	0.00	0.00
ST_Rev	0.09	0.13**	0.06	0.11	0.02	0.02	0.02	0.02	0.01	0.01
CF	0.09	0.09**	0.11	0.16**	-0.07	-0.06**	-0.06	-0.04	-0.04	-0.03
Resid_Var	0.09	0.14**	0.08	0.17**	-0.03	-0.03	0.03	0.03	-0.06	-0.06
ages	0.09	0.05**	0.01	0.01	0.06	0.04	0.07	0.04	0.02	0.01
MktBeta	0.08	0.14**	0.07	0.14	0.00	0.00	0.05	0.05	-0.07	-0.05**
r12_2	0.08	0.11**	0.08	0.19**	-0.06	-0.08**	-0.08	-0.09**	0.01	0.01
Spread	0.08	0.13**	0.07	0.14	-0.01	-0.01	0.04	0.04	-0.04	-0.04
D2P	0.08	0.12**	0.08	0.13**	-0.01	-0.01	0.09	0.08***	-0.16	-0.09***
r12_7	0.08	0.11**	0.10	0.19***	-0.07	-0.08**	-0.09	-0.09**	0.02	0.02
F_r2_1	0.08	0.11	-0.00	-0.00	0.13	0.11***	0.12	0.08**	0.07	0.04**
LTurnover	0.07	0.13	0.06	0.14	-0.01	-0.01	0.08	0.08**	-0.12	-0.09***
Variance	0.07	0.13	0.07	0.17	-0.04	-0.04	0.04	0.04	-0.09	-0.08**
IdioVol	0.07	0.12	0.07	0.15	-0.04	-0.04	0.03	0.03	-0.07	-0.06
C	0.07	0.09	0.08	0.12**	-0.04	-0.03	-0.10	-0.08***	0.11	0.05***
Lev	0.07	0.08	0.05	0.07	0.02	0.01	-0.02	-0.01	0.04	0.02
Family_flow	0.07	0.04	0.05	0.03	0.02	0.01	0.04	0.02	-0.03	-0.01
ATO	0.07	0.07	0.05	0.05	0.02	0.02	-0.03	-0.02	0.09	0.04**
exp_ratio	0.07	0.04	-0.03	-0.02	0.13	0.06***	0.08	0.04**	0.05	0.03
CTO	0.06	0.07	0.06	0.07	-0.01	-0.01	-0.07	-0.04	0.07	0.03
tna	-0.06	-0.04	-0.04	-0.03	-0.03	-0.02	0.09	0.04**	-0.11	-0.06***
SUV	0.06	0.06	0.06	0.07	-0.03	-0.02	-0.04	-0.02	0.01	0.01
SGA2S	0.05	0.06	0.08	0.10**	-0.06	-0.04	-0.07	-0.05**	0.02	0.01
OL	0.05	0.05	0.03	0.04	0.01	0.01	-0.01	-0.00	0.03	0.01
PCM	0.05	0.05	0.10	0.12**	-0.09	-0.07**	-0.06	-0.04	-0.06	-0.03
r2_1	0.05	0.06	0.00	0.00	0.06	0.06	0.05	0.04	0.03	0.02
CF2P	0.04	0.06	0.04	0.08	-0.02	-0.02	-0.08	-0.06**	0.08	0.04**
NI	0.04	0.05	0.04	0.05	0.01	0.01	0.12	0.09**	-0.16	-0.08***
Q	0.04	0.05	0.02	0.04	0.01	0.01	-0.07	-0.05**	0.10	0.07**
FC2Y	0.04	0.05	0.07	0.09	-0.06	-0.05	-0.06	-0.04**	-0.00	-0.00
PROF	0.04	0.04	0.05	0.08	-0.06	-0.05	-0.06	-0.03	-0.02	-0.01
LME	0.04	0.03	0.04	0.04	-0.01	-0.00	0.03	0.03	-0.03	-0.03
D2A	0.03	0.03	0.03	0.03	0.00	0.00	-0.02	-0.01	0.01	0.01
turnover	0.03	0.03	0.04	0.04	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01
AT	0.03	0.04	0.01	0.02	0.02	0.02	-0.04	-0.02	0.05	0.04
OA	0.03	0.03	-0.01	-0.00	0.05	0.03	0.07	0.04	-0.01	-0.00
r36_13	-0.03	-0.04	-0.10	-0.13**	0.12	0.09***	0.12	0.08***	0.02	0.01
AC	0.03	0.02	-0.00	-0.00	0.04	0.02	0.06	0.03	-0.02	-0.01
E2P	-0.03	-0.04	-0.01	-0.02	-0.03	-0.02	0.06	0.03	-0.10	-0.05**
OP	0.02	0.03	0.02	0.03	-0.00	-0.00	-0.02	-0.01	0.01	0.01
NOA	0.02	0.02	0.05	0.05	-0.06	-0.03	-0.11	-0.06**	0.05	0.03
DPI2A	0.01	0.01	0.08	0.08**	-0.11	-0.07***	-0.19	-0.09***	0.05	0.02
A2ME	0.01	0.02	0.05	0.07	-0.07	-0.06**	-0.15	-0.12***	0.12	0.06***
S2P	-0.01	-0.02	-0.02	-0.03	0.02	0.02	0.13	0.11***	-0.18	-0.09***
Investment	0.01	0.01	0.08	0.08**	-0.11	-0.07***	-0.14	-0.09***	0.04	0.02
LT_Rev	0.01	0.01	-0.01	-0.01	0.02	0.02	0.02	0.01	0.01	0.01
BEME	-0.00	-0.00	-0.03	-0.04	0.05	0.04	0.13	0.11***	-0.15	-0.07***

This table reports the mean and Sharpe ratio for the decomposition of univariate long-short abnormal return factors. Means of abnormal returns are reported in percentages. The results are sorted according to the Sharpe ratio of univariate long-short factors. For each of the 59 characteristics and each abnormal return, we construct ten sorted decile portfolios. The long-short factors are the differences between the top decile and the bottom decile.

IA.3 Abnormal Return Construction on Out-Of-Sample-Folds

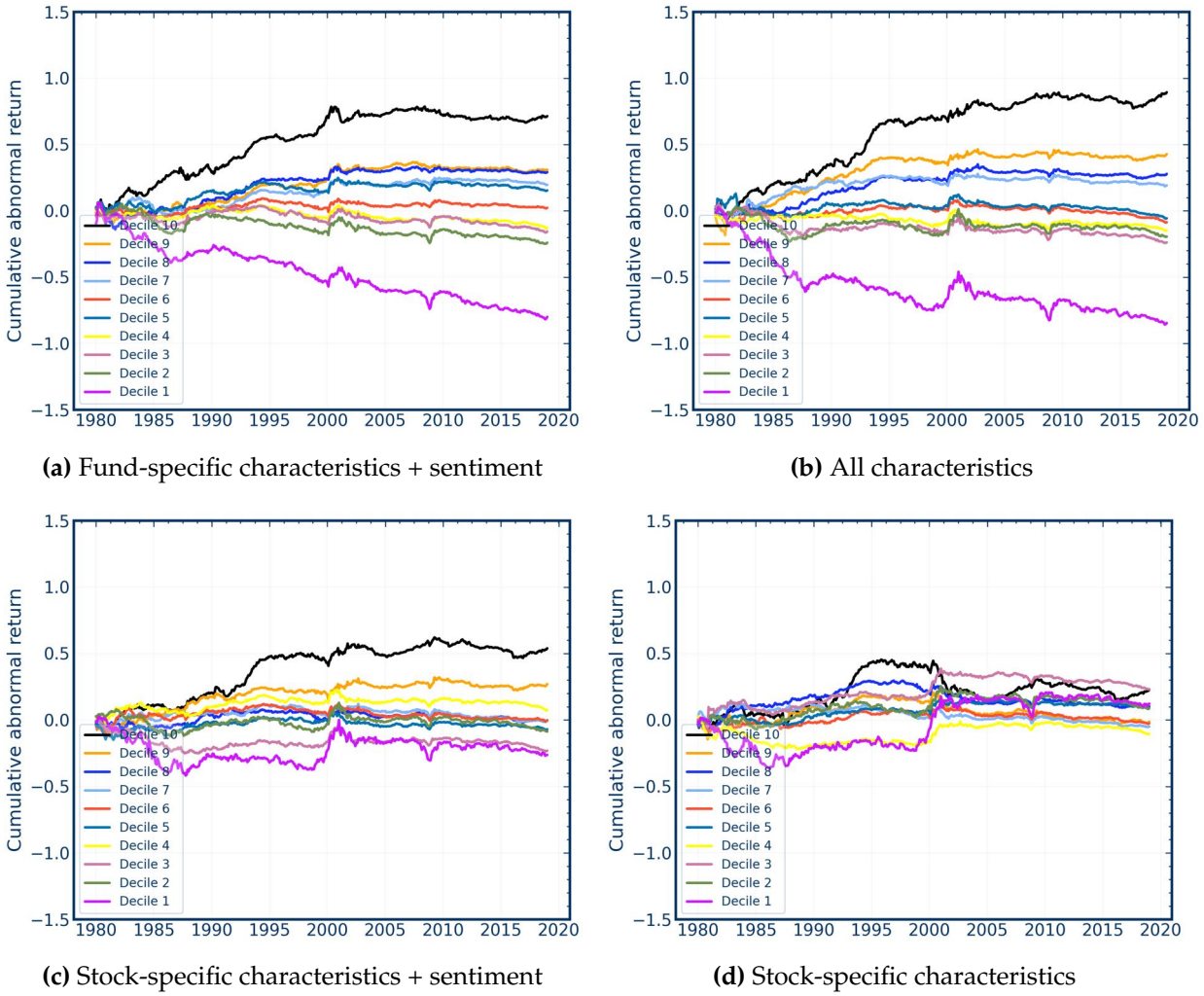
The construction of abnormal returns (equations 1 and 2 in the main text) in our main analysis respects the chronological order of the data. The main neural network analysis then randomly samples dates for the cross-out-of-sample analysis based on those abnormal returns. This means that we use information from time $t-36$ to $t-1$ to construct the abnormal return at date t in the first step, while in the cross-out-of-sample analysis in the second step, some dates from time $t-36$ to $t-1$ are used for training and validation. In this appendix, we investigate the impact of this data structure by studying an alternative approach that avoids this overlap. Specifically, for each test-sample in one of the test folds, we still look back three years, but only using data points that are not within this test fold. Since in our cross-out-of-sample evaluations, each data point is in one of the test folds, this guarantees that all abnormal returns are defined out-of-sample but also respect the chronological order locally. The downside of this sampling scheme is that we have fewer samples in the local regressions and thus this abnormal return definition will be noisier.

IA.3.1 Prediction Results

The cumulative abnormal returns for different information sets are in Figure IA.1 and the cumulative abnormal returns of long-short prediction portfolios are in Figure IA.2. The Sharpe ratio, mean, and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios are in table IA.12. The results are very similar to the analysis in the main text, though the economic magnitudes are a few basis points smaller.

The top variable importance for explaining abnormal returns and interaction effects between sentiment and fund characteristics are in figures IA.3 and IA.4.

Figure IA.1: Cumulative completely out-of-sample abnormal returns for different information sets with benchmark random data split.



(a) Fund-specific characteristics + sentiment

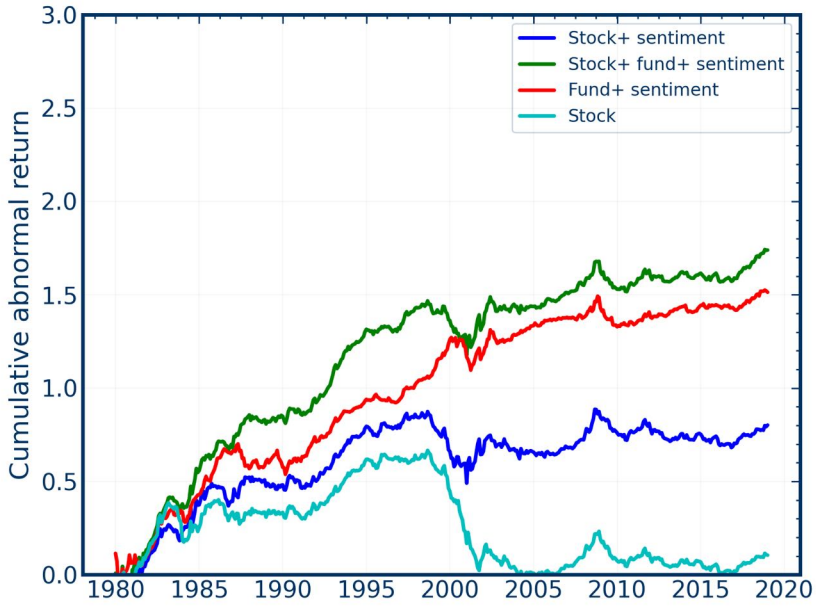
(b) All characteristics

(c) Stock-specific characteristics + sentiment

(d) Stock-specific characteristics

These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment, stock-specific characteristics+ sentiment, stock-specific characteristics or all characteristics to predict abnormal returns. Fund abnormal returns are defined completely out-of-sample and the data split is random cross-out-of-sample as in our benchmark setup.

Figure IA.2: Cumulative completely out-of-sample abnormal returns of long-short prediction portfolios with benchmark random data split.



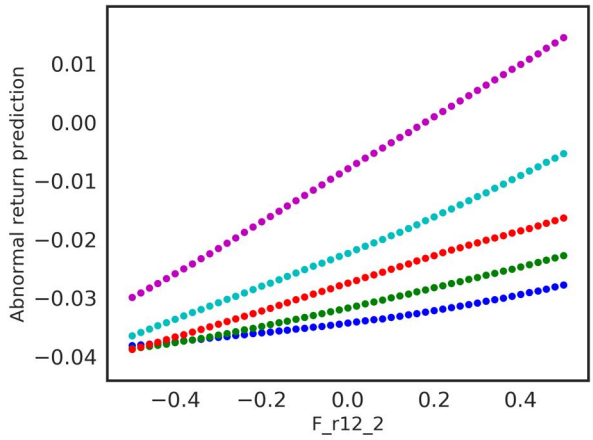
This figure plots the cumulative abnormal returns of prediction-weighted long-short decile portfolios that use different information sets for prediction. We consider fund-specific and stock-specific characteristics combined with sentiment. Fund abnormal returns are defined completely out-of-sample and the data split is random cross-out-of-sample as in our benchmark setup.

Table IA.12: Performance of completely out-of-sample abnormal return portfolios with random sampling.

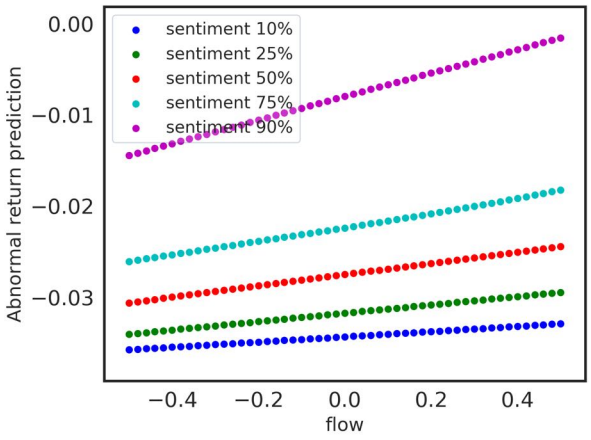
Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ sentiment	0.17	1.8*	0.08	1.78
	Stock+ fund+ sentiment	0.37	4.2***	0.19	4.41
	Fund+ sentiment	0.32	4.1***	0.19	1.49
	Stock	0.02	0.3	0.01	-1.11
Top decile	Stock+ sentiment	0.11	2.3**	0.11	1.89
	Stock+ fund+ sentiment	0.19	3.4***	0.16	3.32
	Fund+ sentiment	0.15	2.9***	0.13	0.57
	Stock	0.05	0.8	0.04	-0.51
Bottom decile	Stock+ sentiment	-0.06	-0.7	-0.05	-0.51
	Stock+ fund+ sentiment	-0.18	-2.5**	-0.15	-0.27
	Fund+ sentiment	-0.17	-2.7***	-0.15	0.18
	Stock	0.03	0.3	0.02	-1.13

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios that use different information sets for the prediction. We consider four different information sets which combine fund-specific and stock-specific characteristics and sentiment. The abnormal returns are defined completely out-of-sample and the data split is random cross-out-of-sample as in our benchmark setup.

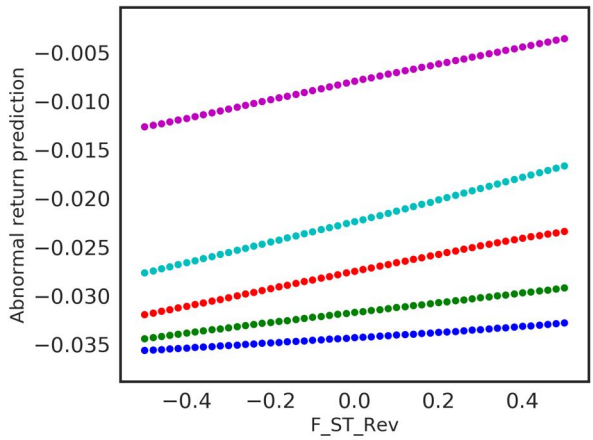
Figure IA.3: Conditional mean as a function of fund characteristics and sentiment with completely out-of-sample abnormal returns and random sampling.



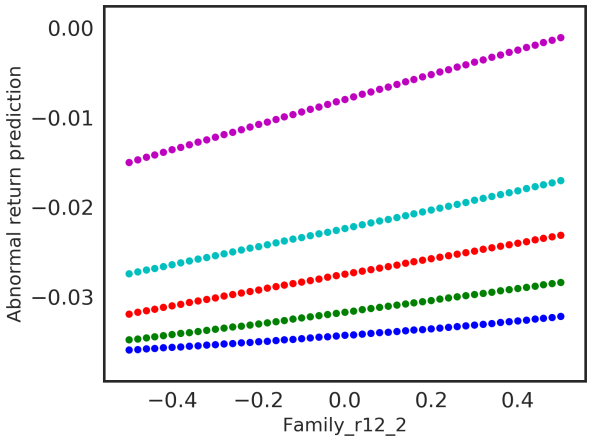
(a) Predicted abnormal returns as function of *F_r12_2*



(b) Predicted abnormal returns as function of *flow*



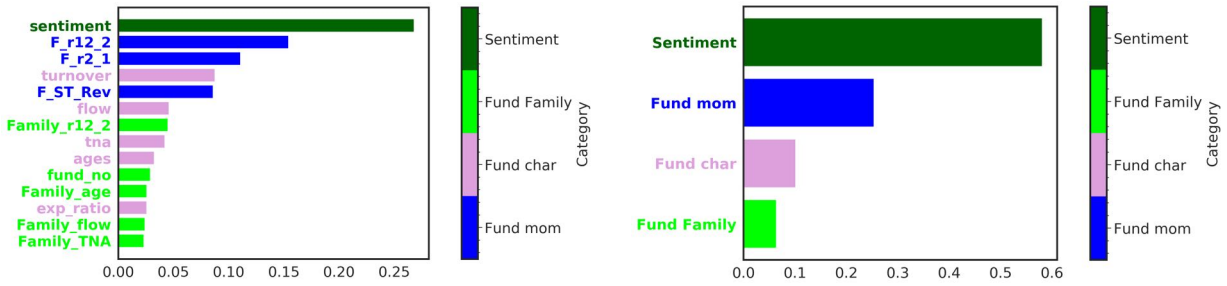
(c) Predicted abnormal returns as function of *F_ST_Rev*



(d) Predicted abnormal returns as function of *Family_r*12_2

This figure shows the predicted abnormal returns (in percentages) as a function of one fund characteristic conditional on different sentiment quantiles. The other variables are set to their median. The neural network model is estimated with fund-specific characteristics and sentiment. The interaction effects are evaluated on the test data and averaged across three cross-out-of-sample folds. The high-minus-low portfolios have a higher mean conditional on high past sentiment. This is a non-linear interaction effect. Fund abnormal returns are defined completely out-of-sample and the data split is random cross-out-of-sample as in our benchmark setup.

Figure IA.4: Top variable importance for explaining abnormal returns with completely out-of-sample abnormal returns and random sampling.



(a) Top variable importance.

(b) Top variable group importance.

This figure shows the importance ranking for individual variables and variable groups. The ranking is the square root of average of the squared gradient for the eight ensemble fits as in equation 7. The variable importance measures are evaluated on the test data and averaged across three cross-out-of-sample folds. Fund-specific characteristics and sentiment are used as network input. Fund abnormal returns are defined completely out-of-sample and the data split is random cross-out-of-sample as in our benchmark setup.

IA.4 Neural Networks with More Layers

Our findings are robust to the structure of the neural networks. Our main analysis uses the parameters of the network structure that is selected on the validation data. This optimal network has one hidden layer. In this section, we show that the results are robust to using more layers.

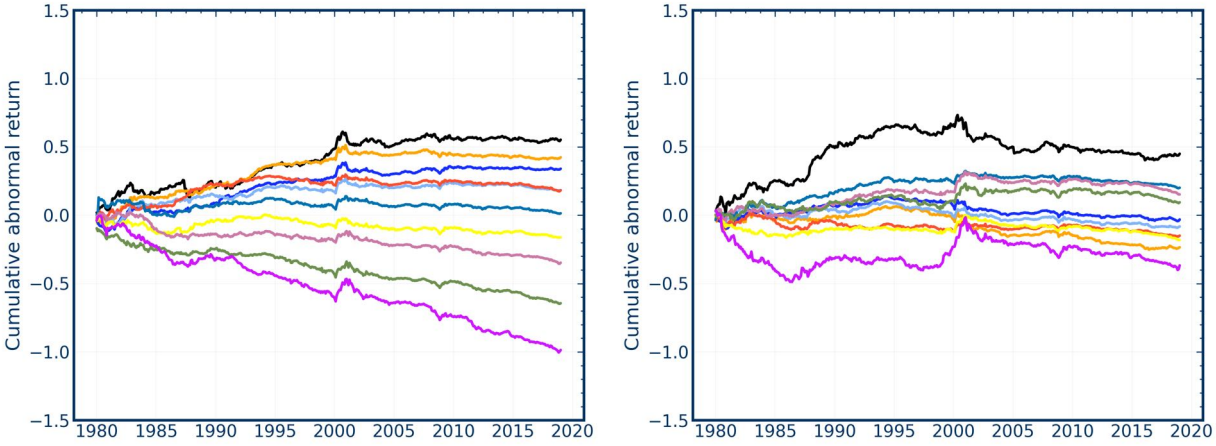
This analysis is motivated by the results in Gu, Kelly, and Xiu (2020), who document that a neural network with three layers can predict stock returns better than a neural network with one layer. It is helpful to provide some context to the discussion about the number of layers in a neural network.

First, our prediction target are not stock returns, but residuals of mutual fund returns. Predicting fund residuals seem to be a "simpler" problem in terms of functional complexity than predicting stock returns, which essentially requires to model the systematic component as well. Hence an optimal network for predicting stock returns does not need to be the optimal network for abnormal fund returns. More layers obviously also mean more parameters and therefore require either more data or a stronger signal to noise ratio to be useful. A panel of individual stock returns is a larger data set and returns (instead of abnormal returns) have a stronger structure to detect. The data of abnormal fund returns is comparatively smaller and seems to have a less complex structure than the panel of individual stock returns. Therefore, the structure that can be estimated robustly in our data is simpler.

Second, the complexity of the network is a combination of the number of layers and nodes. Theoretically a shallow network with few layers but with more nodes can be equivalent to a deeper network with less nodes. Hence, a discussion about the number of layers needs to be related to the number of nodes used in each layer. Our network with one layer has more nodes

Figure IA.5: Cumulative abnormal returns for 2 and 3 layers

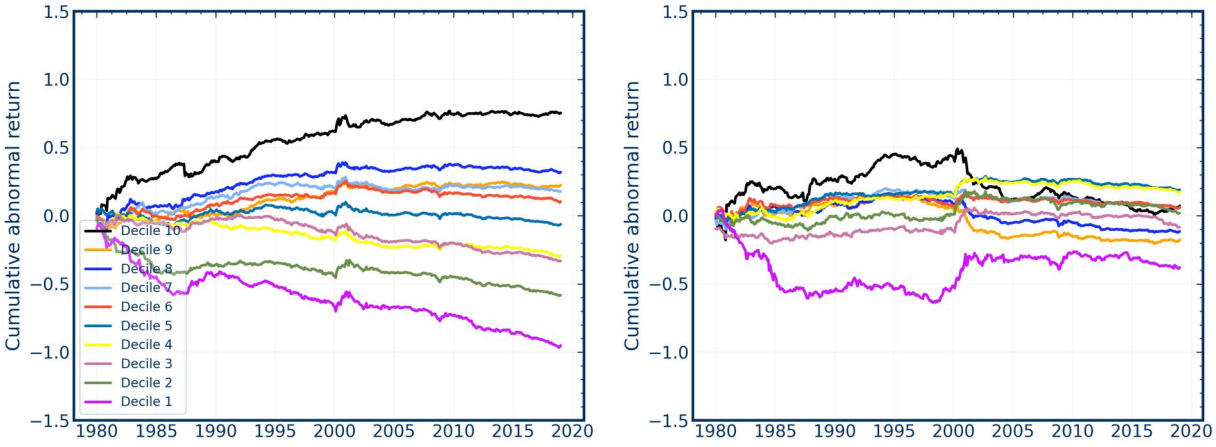
Panel A: Optimal neural network with 2 layers



(a) Fund-specific characteristics + sentiment

(b) Stock-specific characteristics

Panel B: Optimal neural network with 3 layers



(c) Fund-specific characteristics + sentiment

(d) Stock-specific characteristics

These figures show the cumulative abnormal returns sorted into prediction deciles for different number of layers of the neural networks. The abnormal returns are prediction-weighted within deciles. We consider fund-specific characteristics + sentiment and stock-specific characteristics. In Panel A, we show the results for an optimal neural network with 2 layers, and in panel B we show the results for the optimal neural network with 3 layers. The data split is random cross-out-of-sample as in our benchmark setup.

than, for example, the corresponding network in Gu, Kelly, and Xiu (2020). In summary, a comparison of different models would need to take into account all parameters of a network (including regularization parameters) and not only the number of layers.

Third, a direct comparison with Gu, Kelly, and Xiu (2020) could also be misleading in two

other aspects. We select the number of layers on the validation data. Gu, Kelly, and Xiu (2020) do not select the number of layers on the validation data, but show the out-of-sample results for different number of layers. Furthermore, the comparison is in terms of different metrics.

We show that our main results are robust to the number of layers. For this purpose, we report the results of multi-layer neural networks under the random sampling scheme. The optimal tuning parameters for both, the 2-layer and 3-layer neural networks, are similar.[1] For the 2-layer neural network, the optimal number of hidden nodes are 64 and 32, and for the 3-layer neural network, the optimal number of hidden nodes are 64, 32, and 16. The analysis follows the same procedure as for our main model.

Figure IA.5 shows the cumulative abnormal returns for the 2- and 3-layer neural networks. Our main finding are still valid, that is, fund characteristics are predictive for fund abnormal return but not stock characteristics. Table IA.13 reports the Sharpe ratio, t-statistics, Sharpe ratio, and R^2_F of the extreme prediction deciles for the multi-layer neural networks. The qualitative results confirm our main findings. As expected, the metrics for 2 and 3 layers are lower than for one layer, which is reported in Table 3. This confirms that our optimal network selected on the validation data is indeed a one-layer neural network.

Table IA.13: Performance of abnormal return portfolios with multi-layer neural networks.

Portfolio	Layer	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	2	Fund+ sentiment	0.33	5.5***	0.26	0.23
		Stock	0.17	2.6**	0.12	0.61
	3	Fund+ sentiment	0.36	5.1***	0.24	2.71
		Stock	0.10	0.9	0.04	-0.02
Top decile	2	Fund+ sentiment	0.12	2.5**	0.11	1.32
		Stock	0.10	1.6	0.07	-1.10
	3	Fund+ sentiment	0.16	2.9***	0.14	1.13
		Stock	0.02	0.2	0.01	-0.68
Bottom decile	2	Fund+ sentiment	-0.21	-4.5***	-0.20	-2.10
		Stock	-0.08	-1.3	-0.06	-0.94
	3	Fund+ sentiment	-0.20	-3.6***	-0.17	1.31
		Stock	-0.08	-1.1	-0.05	0.06

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios that use different information sets for the prediction. We consider fund-specific characteristics + sentiment and stock-specific characteristics. We show the results for 2 and 3 layers with optimally selected tuning parameters. The data split is random cross-out-of-sample as in our benchmark setup.

[1] The optimal dropout rate is 0.95, l1 and l2 penalty are 0, and the learning rate is 0.01.

IA.5 Gradient Boosted Tree Prediction Results

Our findings are robust to the machine learning method that is used. Based on the discussion of Li and Rossi (2021), we compare the predictability of gradient boosted trees (GBT) with our neural networks. We use the same setup as for the neural network prediction and use the default parameters of the GBT package except for the maximum depth, which is selected optimally from the validation data.

IA.5.1 Random Sampling

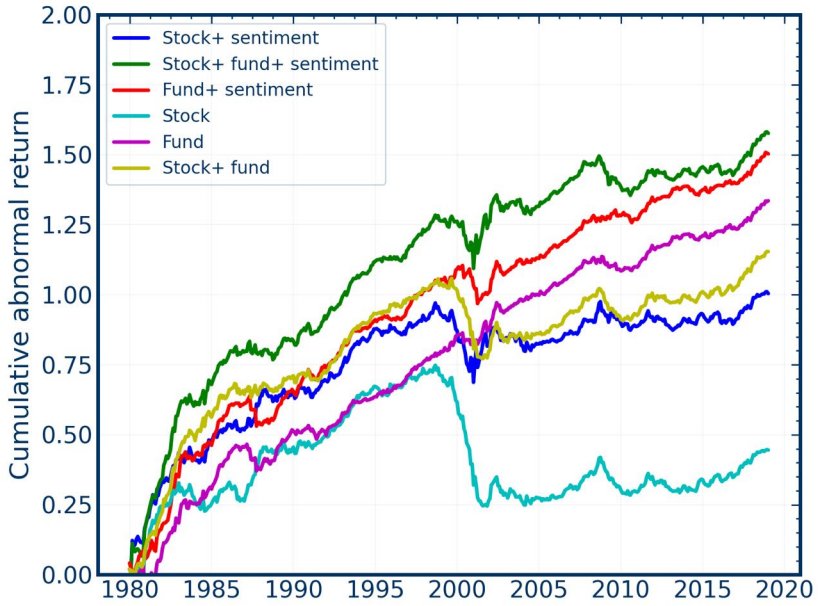
We use the same random sample for the three folds as in the main text for the neural network analysis. Subsection IA.5.3 shows that there are a few extreme outlier GBT predictions pre-1990, which makes long-short prediction-weighted portfolio results unreliable. Thus we focus on the equally-weighted portfolio results for the main analysis. The results for this equally-weighted portfolio are in Table IA.14 and a comparison between different information sets is given in Figure IA.6. In Subsection IA.5.3, we also show the prediction-weighted portfolio results post 1990 and find that the qualitative results are similar. Analogously to the main results with neural networks, we find that under our benchmark random sampling method, stock characteristics have little predictive power for abnormal returns of mutual funds.

Table IA.14: Performance of equally-weighted abnormal return portfolios of GBT with random sampling.

Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ sentiment	0.21	3.1***	0.14	-14.46
	Stock+ fund+ sentiment	0.34	4.7***	0.22	-8.31
	Fund+ sentiment	0.32	6.3***	0.29	-7.97
	Stock	0.09	1.5	0.07	-5.07
	Fund	0.28	5.7***	0.26	4.48
	Stock+ fund	0.25	4.4***	0.20	1.65
Tenth decile	Stock+ sentiment	0.11	2.3**	0.11	-4.52
	Stock+ fund+ sentiment	0.17	3.1***	0.14	-0.90
	Fund+ sentiment	0.13	3.2***	0.15	-0.74
	Stock	0.05	1.1	0.05	-3.19
	Fund	0.12	2.9***	0.13	0.21
	Stock+ fund	0.10	2.3**	0.11	-2.14
First decile	Stock+ sentiment	-0.10	-1.9*	-0.10	-23.90
	Stock+ fund+ sentiment	-0.17	-3.2***	-0.15	-21.58
	Fund+ sentiment	-0.19	-4.6***	-0.20	-22.62
	Stock	-0.05	-0.9	-0.05	-4.14
	Fund	-0.17	-3.6***	-0.19	0.17
	Stock+ fund	-0.14	-2.9***	-0.15	0.10

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth equally-weighted decile portfolios that use different information sets for the prediction. We consider six different information sets which combine fund-specific and stock-specific characteristics and sentiment. The three cross-out-of-sample folds use the benchmark random sampling.

Figure IA.6: Cumulative abnormal returns of equally-weighted long-short prediction portfolios with random sampling.



These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The abnormal returns are equally-weighted within deciles. The three cross-out-of-sample folds use the benchmark random sampling.

IA.5.2 Chronological Sampling

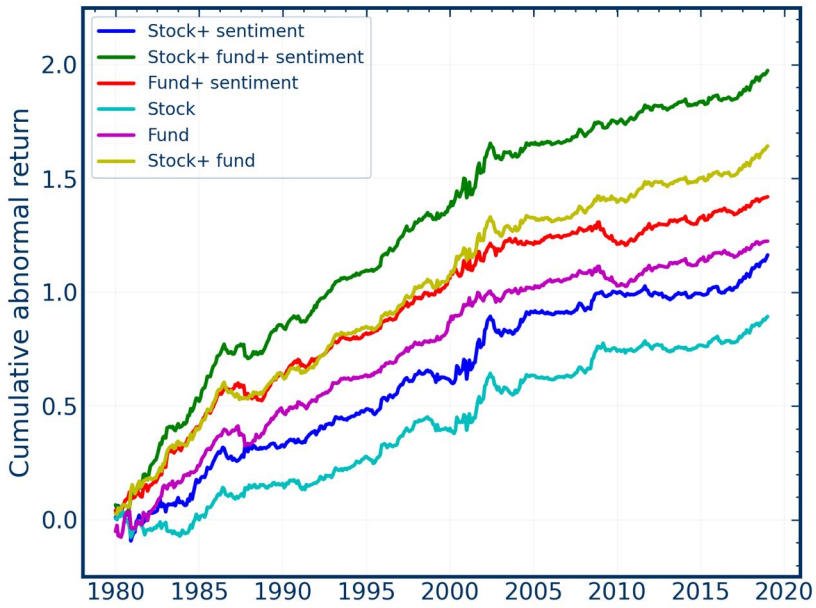
Because of the effect explained in Subsection IA.5.3, we focus on the results for equally-weighted portfolios. The results are shown in Table IA.15 and Figure IA.7. Similar to the chronological split results from neural network predictions, the predictability of stock characteristics is stronger than with random sampling but weaker compared to information sets that contain fund characteristics. The stock characteristics lose their predictive power after the year 2000.

Table IA.15: Performance of abnormal return portfolios with GBT and chronological sampling.

Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ sentiment	0.25	3.8***	0.17	2.30
	Stock+ fund+ sentiment	0.42	6.7***	0.31	6.04
	Fund+ sentiment	0.30	5.8***	0.27	1.68
	Stock	0.19	3.4***	0.16	-1.08
	Fund	0.26	5.1***	0.23	-3.15
	Stock+ fund	0.35	6.5***	0.30	6.00
Tenth decile	Stock+ sentiment	0.13	3.2***	0.15	-0.29
	Stock+ fund+ sentiment	0.18	4.2***	0.20	-1.21
	Fund+ sentiment	0.14	3.5***	0.16	-7.00
	Stock	0.12	3.3***	0.15	0.55
	Fund	0.12	2.7***	0.12	-3.75
	Stock+ fund	0.19	4.5***	0.21	2.37
First decile	Stock+ sentiment	-0.12	-2.2**	-0.15	-1.66
	Stock+ fund+ sentiment	-0.25	-5.0***	-0.27	0.95
	Fund+ sentiment	-0.17	-3.8***	-0.20	-1.54
	Stock	-0.07	-1.3	-0.09	-1.36
	Fund	-0.15	-3.2***	-0.16	-2.93
	Stock+ fund	-0.16	-3.7***	-0.18	1.63

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth equally-weighted decile portfolios that use different information sets for the prediction. We consider six different information sets which combine fund-specific and stock-specific characteristics and sentiment. Three cross-out-of-sample folds keep the chronological order.

Figure IA.7: Cumulative abnormal returns of equally-weighted long-short prediction portfolios with GBT and chronological sampling.



These figures show the cumulative abnormal returns sorted into prediction deciles for different information sets. The abnormal returns are equally-weighted within deciles. The three cross-out-of-sample folds keep the chronological order.

IA.5.2.1 Predicting Total Returns

When predicting total (as opposed to abnormal)returns, we are not affected by the outlier issue described in Subsection IA.5.3. Hence, we return to the benchmark setup and report results for prediction-weighted portfolios. The results for equally-weighted portfolios are comparable.

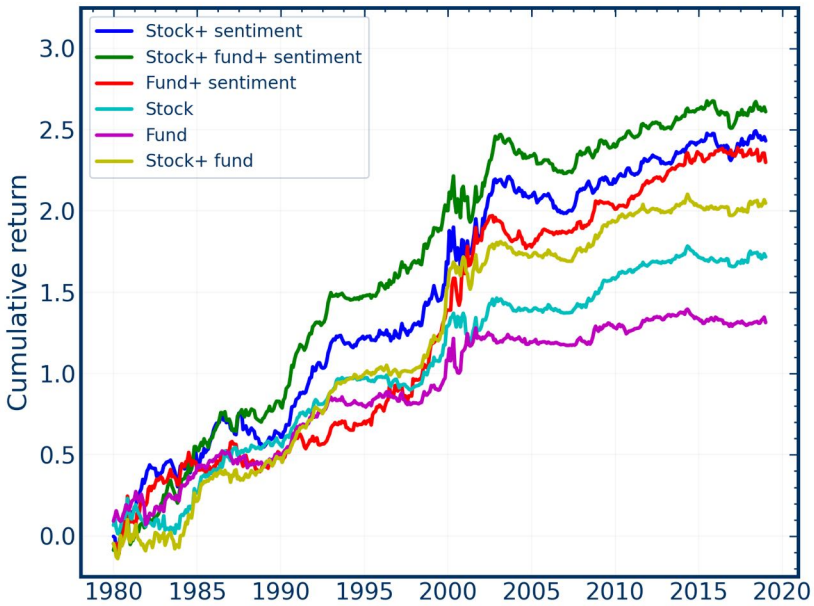
Table IA.16 and Figure IA.8 show the results for predicting total returns with GBT and chronological sampling. Both fund characteristics and stock characteristics strongly predict out-of-sample mutual fund returns, and the economic magnitudes are similar to the results in section A.2.4. That is, we do not find evidence that GBT has stronger predictive power for fund total return than our benchmark neural networks.

Table IA.16: Performance of return portfolios with GBT and chronological sampling.

Portfolio	Information set	mean(%)	t-stat	SR	R^2_F (%)
Long-short	Stock+ sentiment	0.52	3.6***	0.17	-13.40
	Stock+ fund+ sentiment	0.56	3.9***	0.18	-14.90
	Fund+ sentiment	0.49	3.6***	0.17	-12.34
	Stock	0.37	3.6***	0.17	-218.36
	Fund	0.28	2.8***	0.13	-119.70
	Stock+ fund	0.44	3.8***	0.18	-152.90
Top decile	Stock+ sentiment	0.13	3.2***	0.15	-0.29
	Stock+ fund+ sentiment	0.18	4.2***	0.20	-1.21
	Fund+ sentiment	0.14	3.5***	0.16	-7.00
	Stock	0.12	3.3***	0.15	0.55
	Fund	0.12	2.7***	0.12	-3.75
	Stock+ fund	0.19	4.5***	0.21	2.37
Bottom decile	Stock+ sentiment	-0.12	-2.2**	-0.15	-1.66
	Stock+ fund+ sentiment	-0.25	-5.0***	-0.27	0.95
	Fund+ sentiment	-0.17	-3.8***	-0.20	-1.54
	Stock	-0.07	-1.3	-0.09	-1.36
	Fund	-0.15	-3.2***	-0.16	-2.93
	Stock+ fund	-0.16	-3.7***	-0.18	1.63

This table reports the Sharpe ratio, mean and factor R^2 of long-short, the first, and the tenth prediction-weighted decile portfolios that use different information sets for the return prediction. We consider six different information sets which combine fund-specific and stock-specific characteristics and sentiment. The three cross-out-of-sample folds keep the chronological order.

Figure IA.8: Cumulative returns of prediction-weighted long-short prediction portfolios with GBT and chronological sampling.

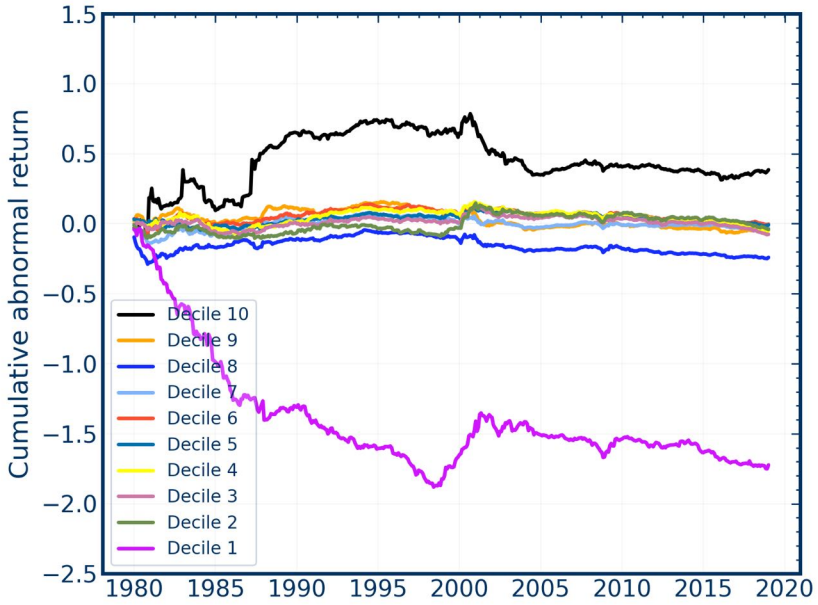


These figures show the cumulative returns sorted into prediction deciles for different information sets. The returns are prediction-weighted within deciles. The three cross-out-of-sample folds keep the chronological order.

IA.5.3 Instability of GBT Prediction Pre-1990

The GBT prediction in the early part of the sample is more extreme that that of the neural networks. This makes the long-short abnormal return prediction portfolio look extremely good in the early part of the sample. An illustration of this effect is shown in Figure IA.9.

Figure IA.9: Cumulative abnormal returns of long-short prediction-weighted prediction portfolios of GBT with random sampling



These figures show the cumulative abnormal returns sorted into prediction deciles for stock-specific characteristics. The abnormal returns are prediction-weighted within deciles.

However, if we analyze equally-weighted portfolios instead, we arrive at a different conclusion. The results are shown in Figure IA.10. Hence, the good results in the early part of the sample arise not because we are systematically able to differentiate good from bad mutual fund managers but because we can predict some extreme outliers correctly with GBT. This is the reason why in the prior subsections of GBT prediction, we focused on the results with equally-weighted portfolios.