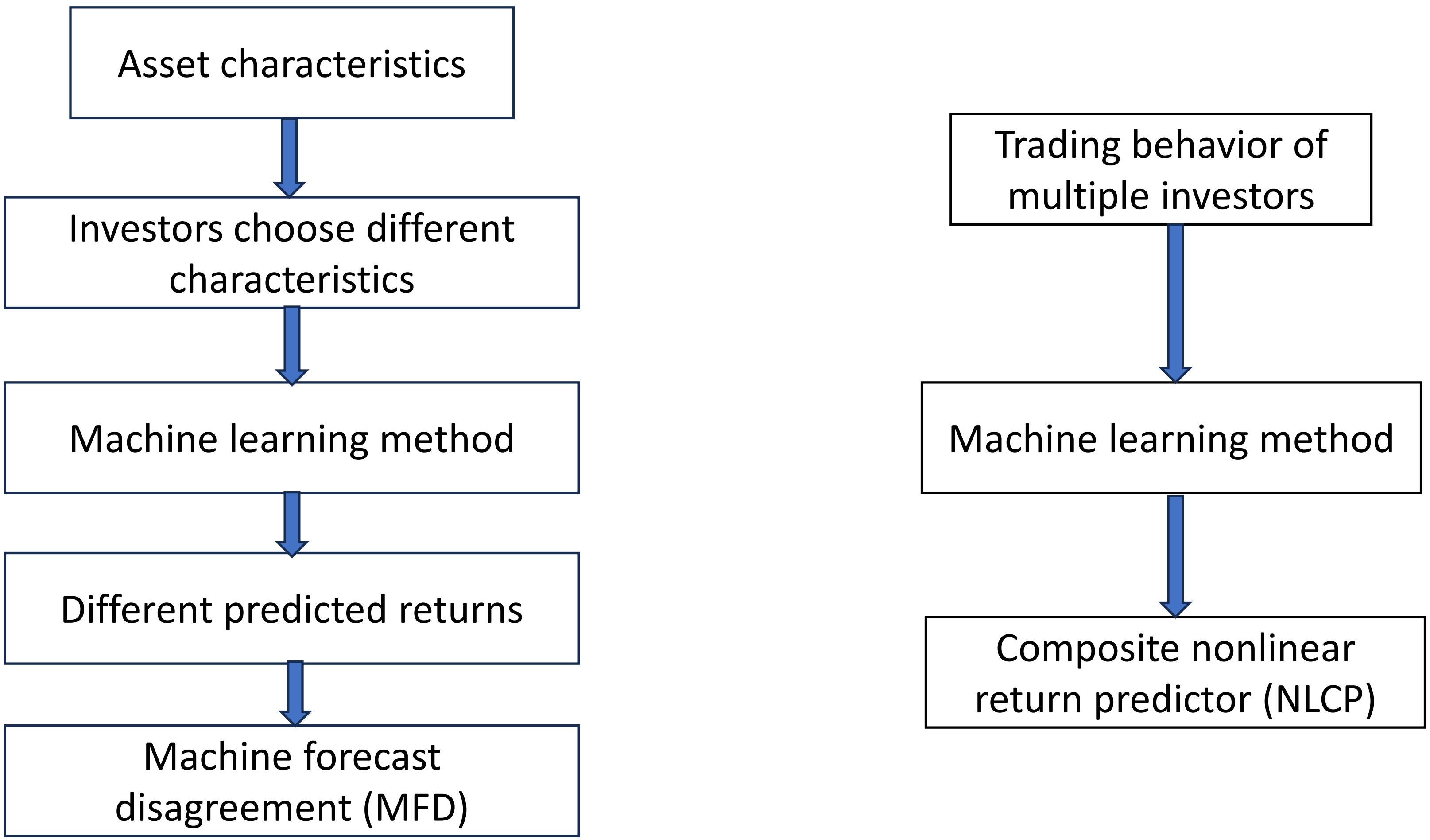
Machine learning, heterogeneous investors, and stock returns



prediction

simulation

Machine Forecast Disagreement

Turan G. Bali, Bryan T. Kelly, Mathis M¨orke, and Jamil Rahman

Working paper, 2023

1. Introduction-- Motivation

- Belief disagreement is a primary motivation for trade
- A **theoretical** literature seeks to understand how differences of opinion among investors impact market prices and volumes.
  - Miller (1977) predicts that stock prices are upward biased due to short-sale constraints
- **Empirical** work on disagreement is more limited due to the difficulty in measuring investor beliefs.
  - Diether et al. (2002) proxies for belief heterogeneity using analyst earnings forecasts: analyst forecast dispersion (AFD) ↑ → future return ↓
  - Johnson (2004) questions the interpretation of Diether et al. (2002) and argues that AFD proxies for firm-specific risk

➢ We propose a new measure of investor disagreement.

1. Introduction-- Challenge

- We propose a statistical surrogate for belief disagreement
  - not directly observable
- Each investor is a **prediction model** from which beliefs about future returns are formed
  - have access to a common set of predictive information, but use available information in **different** ways
  - endowing each investor with a machine learning model but introduce **random variation** in model specification
- Whether the distribution from which we simulate model specifications is plausible
  - the **calibration** of our simulation distribution is reasonable
  - our results are **robust** to a range of distributions for simulating investors' models.

1. Introduction-- Contents

Collection of investors: different information

↓

Random Forest

↓

Different returns

↓

Machine forecast disagreement (MFD)

↓

Explanatory power of MFD for return

↓

Economic underpinnings of MFD alpha

每只股票存在多个预测值

1. Introduction-- Contribution

- Proposing a **new measure** of belief disagreement at the asset level (MFD)
  - more data coverage, more objective than AFD (Dugar and Nathan, 1995; Michaely and Womack, 1999; Chan et al., 2007)
- Documenting the strong explanatory power of MFD for the cross-sectional pricing of individual stocks.
  - 23% correlation with AFD, more explanatory power than AFD
- Investigating the economic underpinnings of MFD alpha
  - More overpricing with higher short-sale costs and higher retail ownership (Miller,1977)
  - The MFD premium is associated with high-MFD stocks being mispriced (Stambaugh et al.,2015)

2. An empirical model of disagreement

- Gu et al. (2020) consider a general conditional risk premium formulation

$$E_t[r_{i,t+1}] = g(z_{i,t}),$$

- We consider a collection of investors $k = 1, \ldots, K$. Each investor k differs in her information set $z_{k,i,t}$, an investor k forms beliefs according to

$$E_{k,t}[r_{i,t+1}] = g_k(z_{k,i,t}).$$

$$g_k(z_{i,t}) = RF_k(z_{k,i,t}),$$

  - Each investor is endowed with an incomplete information set $z_{k,i,t} \in \mathbb{R}^{d_k}$, $1 \leq d_k \leq d$. $z_{i,t} \in \mathbb{R}^d$.
  - Investors estimate $g_k(\cdot)$ using a random forest regression (bootstrapping and dropout)
- Machine forecast disagreement (MFD), for stock i as the standard deviation of $E_{k,t}\ [r_{i,t+1}]$ across investors.

3. Data and variables

- Use the dataset from Jensen et al. (2022b)
- Stock returns and characteristics(153)
- Cross-sectionally rank into [-1,1]
- 1966.07-2022.12
- 10-year rolling window
- MFD construction
  - the number of investors $K = 100$
  - the dimension of the incomplete information set $d_k = 76$

4. Univariate Sorts on MFD

Panel A: Equal-Weighted Portfolios				
	Excess Return	t-stat	CAPM	t-stat
Low	1.14***	(5.09)	0.50***	(3.72)
2	1.06***	(4.85)	0.42***	(3.30)
3	1.06***	(4.62)	0.38***	(2.97)
4	1.00***	(4.28)	0.29**	(2.43)
5	0.97***	(4.02)	0.23**	(1.96)
6	0.87***	(3.44)	0.11	(0.92)
7	0.77***	(2.87)	-0.03	(-0.22)
8	0.61**	(2.08)	-0.22	(-1.51)
9	0.44	(1.44)	-0.42***	(-2.63)
High	-0.18	(-0.52)	-1.12***	(-5.68)
H-L	-1.32***	(-5.61)	-1.62***	(-7.15)

- The negative cross-sectional relation between MFD and future stock returns

4. Analyst Forecast Dispersion and MFD

Panel A: Average AFD in MFD Decile Portfolio

	Low	High	H-L	t-stat
AFD	0.08	0.26	0.18***	18.48

Panel B: Bivariate Portfolio Sort on AFD

	Low	High	H-L	t-stat	FF6	t-stat
AFD Low	1.18	0.66	-0.51**	-2.23	-0.39	-1.52
AFD 2	0.92	0.07	-0.85***	-3.22	-0.83***	-2.79
AFD 3	0.96	-0.02	-0.98***	-3.43	-0.73**	-2.38
AFD 4	0.87	-0.09	-0.96***	-2.79	-0.79**	-2.18
AFD High	0.61	-0.59	-1.20***	-3.41	-1.11***	-2.92
AFD H-L	-0.57	-1.25	-0.68**	-2.34	-0.72**	-2.19

- The relatively higher predictive power of MFD with respect to AFD
- MFD is much stronger for equities with high AFD.

4. Average Stock Characteristics of MFD-sorted Portfolios

	Low	2	3	4	5	6	7	8	9	High	H-L	t-stat
MFD	1.21	1.39	1.52	1.64	1.84	1.97	2.10	2.27	2.49	2.87	1.66***	(19.21)
SUE	0.04	0.02	0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.03	-0.07**	(-2.14)
AG	0.06	0.07	0.08	0.09	0.09	0.11	0.12	0.13	0.16	0.23	0.17***	(7.32)
MOM	0.16	0.13	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.09	-0.07**	(-2.11)
ILLIQ	0.06	0.08	0.07	0.07	0.08	0.09	0.09	0.10	0.11	0.17	0.11***	(4.97)
OP	0.29	0.27	0.26	0.25	0.24	0.23	0.21	0.19	0.15	0.03	-0.26***	(-11.36)
IVOL	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.02***	(21.58)
BETA	0.92	0.98	1.03	1.08	1.11	1.15	1.19	1.26	1.33	1.46	0.54***	(16.51)
SIZE ($\times 10^{-9}$)	1.35	1.01	0.87	0.73	0.62	0.54	0.45	0.37	0.31	0.24	-1.11***	(-6.77)
BM	0.51	0.53	0.53	0.54	0.54	0.53	0.53	0.53	0.50	0.45	-0.06**	(-2.37)
MAX	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.03***	(18.29)
TURN ($\times 10^{3}$)	4.13	4.27	4.47	4.73	4.85	5.08	5.29	5.58	5.89	7.29	3.16***	(7.98)
STR ($\times 10^{3}$)	8.33	8.31	9.15	8.52	8.51	8.74	9.16	8.29	10.32	23.05	14.72***	(4.30)

- Other firm characteristics can explain the negative relation between MFD and future stock returns.
- The stocks with higher MFD are indeed smaller, less liquid, and have higher idiosyncratic volatility and stronger lottery features.

4. Average Stock Characteristics of MFD-sorted Portfolios

Panel A: Equal-Weighted Portfolios												
	SUE	AG	MOM	ILLIQ	OP	IVOL	BETA	SIZE	BM	MAX	TURN	STR
1	1.12***	1.08***	1.06***	1.13***	1.11***	1.09***	1.13***	1.12***	1.13***	1.11***	1.12***	1.12***
High	-0.00	0.02	-0.00	-0.17	0.07	0.17	0.14	-0.03	-0.08	0.13	-0.09	-0.11
	(-0.01)	(0.07)	(-0.01)	(-0.49)	(0.23)	(0.63)	(0.47)	(-0.10)	(-0.26)	(0.44)	(-0.28)	(-0.32)
H-L	-1.12***	-1.06***	-1.06***	-1.30***	-1.04***	-0.92***	-0.99***	-1.15***	-1.21***	-0.99***	-1.21***	-1.22***
	(-4.99)	(-5.61)	(-5.18)	(-5.71)	(-6.48)	(-7.32)	(-5.61)	(-5.01)	(-6.01)	(-6.94)	(-6.21)	(-6.11)
FF6	-0.77***	-0.77***	-0.79***	-0.88***	-0.87***	-0.72***	-0.80***	-0.73***	-0.85***	-0.64***	-0.81***	-0.84***
	(-6.90)	(-7.04)	(-7.51)	(-7.83)	(-7.43)	(-6.55)	(-6.85)	(-5.82)	(-7.73)	(-6.38)	(-7.23)	(-7.36)

- The negative association between MFD and future stock returns exists while controlling for the established equity return predictors

5. Sources of return predictability--**Mispricing** versus risk

Mispricing and MFD (Stambaugh et al. (2015))

Panel A: Average MISP in MFD Decile Portfolio							
	Low	2	3	4	High	H-L	t-stat
MISP	44.07	46.78	48.98	51.10	55.05	10.97***	16.68

Panel B: Bivariate Portfolio Sort on MISP									
	Low	2	3	4	High	H-L	t-stat	FF6	t-stat
MISP Low	1.32	1.36	1.29	1.16	0.92	-0.40***	-2.66	-0.28*	-1.83
MISP 2	1.11	1.20	1.08	1.00	0.60	-0.51***	-3.00	-0.37**	-2.13
MISP High	0.78	0.56	0.29	0.19	-0.40	-1.18***	-5.33	-0.97***	-4.10
MISP H-L	-0.54	-0.80	-1.00	-0.98	-1.32	-0.79***	-4.75	-0.69***	-3.77

- High MFD stocks indeed have a higher average mispricing score than the low MFD stocks

5. Sources of return predictability--**Mispricing** versus risk

Earnings Announcement Returns Prediction

	Panel A: One-day Window		Panel B: Three-day Window	
Dep. variable	Ret_t^d	Ret_t^d	Ret_t^d	Ret_t^d
MFD	-0.26***	-0.32***	-0.25***	-0.31***
	(-6.31)	(-6.84)	(-6.16)	(-6.67)
MFD × EDAY	-0.50***	-0.50***	-0.36***	-0.36***
	(-3.43)	(-3.42)	(-5.18)	(-5.13)
EDAY	0.25***	0.26***	0.15***	0.15***
	(9.28)	(9.44)	(11.60)	(11.78)
Lagged Controls?	No	Yes	No	Yes
Day Fixed Effects?	Yes	Yes	Yes	Yes

- An earnings announcement window dummy variable (EDAY)
- Mispricing explanation: negative cross-sectional relation is stronger on earnings announcement days(the interaction term)

5. Sources of return predictability--**Short-selling costs**

the indicative borrowing fee provided by HIS Markit, and institutional ownership

Panel A: Average BORROWFEE in MFD Decile Portfolio

	Low	2	3	4	High	H-L	t-stat
BORROWFEE	0.67	0.68	0.88	1.27	3.82	3.15***	8.85

Panel B: Bivariate Portfolio Sort on BORROWFEE

	Low	2	3	4	High	H-L	t-stat	FF6	t-stat
BORROWFEE Low	0.92	1.00	0.97	0.89	0.83	-0.09	-0.45	-0.07	-0.38
BORROWFEE 2	1.00	0.98	0.80	0.66	0.28	-0.72**	-2.38	-0.77**	-2.51
BORROWFEE High	0.76	0.33	-0.38	-1.08	-1.94	-2.69***	-5.45	-2.30***	-4.92
BORROWFEE H-L	-0.16	-0.67	-1.35	-1.97	-2.76	-2.60***	-5.89	-2.23***	-5.27

- The MFD premium is stronger among stocks with more severe short sale costs

5. Sources of return predictability--**Limits to arbitrage**

ARB is composed of IVOL, illiquidity, and size

Panel A: Average ARB in MFD Decile Portfolio

	Low	2	3	4	High	H-L	t-stat
ARB	13.73	14.69	15.75	16.80	18.66	$\boxed{4.93^{***}}$	8.45

Panel B: Bivariate Portfolio Sort on ARB

	Low	2	3	4	High	H-L	t-stat	FF6	t-stat
ARB Low	1.00	0.94	0.93	0.81	0.68	-0.32***	-3.24	-0.28***	-2.73
ARB 2	1.10	1.08	1.01	0.87	0.48	-0.62***	-3.54	-0.38**	-2.16
ARB High	1.16	0.79	0.45	0.28	-0.50	-1.66***	-6.70	-1.37***	-5.51
ARB H-L	0.16	-0.15	-0.47	-0.54	-1.18	$\boxed{-1.34^{***}}$	-6.02	-1.09***	-4.98

- Slow diffusion of information into stock prices due to limits-to-arbitrage provides a complementary explanation to the predictive power of MFD

6. Conclusion

- This paper introduces a statistical model of investor beliefs from which we build a novel measure of investor belief disagreement.

- We find a significantly negative and highly cross-sectional relation between this proposed measure, MFD, and future stock returns.

- We investigate the source of the MFD spread portfolio’s alpha.

Do the Collective Trades of Market Participants Contain Information about Stocks? A Machine Learning Approach

Victor DeMiguel, Li Guo, Bo Sang, Zhe Zhang

Working paper, 2023

1. Introduction-- Motivation

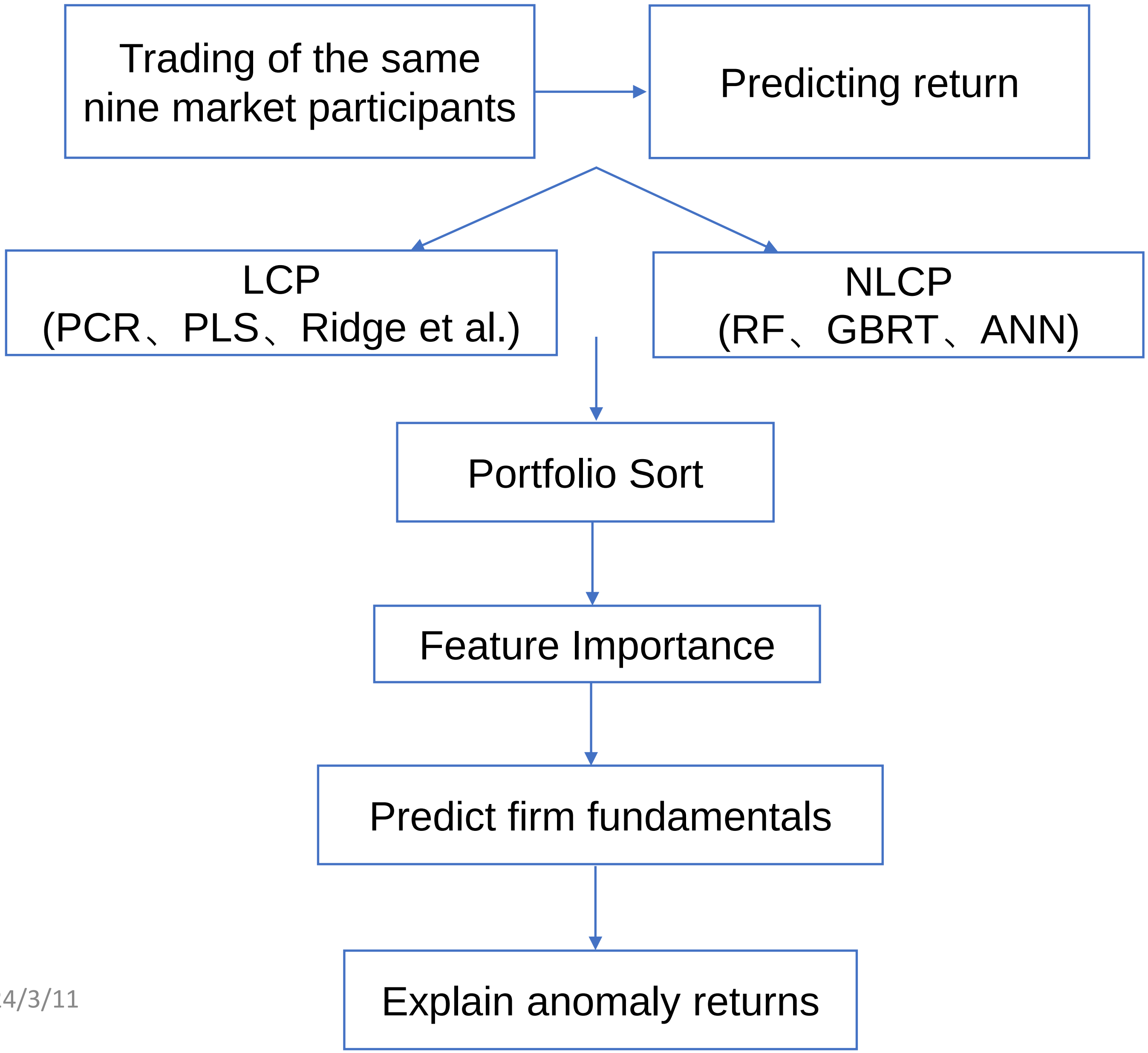
- How information is impounded into asset prices through trading is a central theme

- Most studies tend to focus on the trades of **one particular type** of investors
  - mutual funds, hedge funds, short sellers, or retail investors

- McLean, Pontiff, and Reilly (2022) conduct a comprehensive analysis on the trades of nine market participants
  - focusing on the **marginal effect** of each type of participant

- Studying the **interactions** between investors is important (Diamond and Verrecchia, 1981; Goldstein and Yang, 2015)

1. Introduction-- Contents



1. Introduction-- Question

- Do the composite signals predict returns?

Yes

- The return predictability comes from the (nonlinear) interaction or the trading of a few types?

Both

- The predicted returns contain information about firm fundamentals?

Yes

- The predicted returns is on the right side of anomaly returns?

Yes

1. Introduction-- Contribution

- Contribute to the literature whether market participant trading contains information about future stock returns, using **machine learning** method.
  - (Diether, Lee, and Werner, 2009; Aggarwal and Jorion, 2010; Boehmer, Huszar, and Jordan, 2010; Baker et al., 2010; Kaniel et al., 2012; Kelley and Tetlock, 2013; Cao et al., 2018; Boehmer et al., 2021; McLean, Pontiff, and Reilly, 2022)
- Contribute to the literature that uses machine learning models in asset pricing and investment, based on the trading signals of **multiple market participants**
  - (Gu, Kelly, and Xiu, 2020; Kozak, Nagel, and Santosh, 2020; Bryzgalova, Pelger, and Zhu, 2021; Chatigny, Goyenko, and Zhang, 2022; Leippold, Wang, and Zhou, 2022).

2. Data

- **Institutional trading signals**: institutional holdings data from Thomson/Refinitiv S12 and 13F
  - **mutual funds, insurance companies, banks, hedge funds, wealth management firms, and other institutions**.
  - Changes in 13F institutional holdings
  - The level of holdings of each type
- **Retail trading signals**: compute retail order imbalance on a monthly basis from TAQ trade dataset, following Boehmer et al. (2021)
- **Short seller trading signals**: monthly short interest data scaled by the number of shares outstanding, from Compustat
- **Firm trading signals**: calculate changes in shares, that is, share issues minus share repurchases, divided by shares outstanding from CRSP.

2. Data

- 2008.01-2020.12
- Normalizing trading signals to the (-1,1) interval
- Rolling window: five-year length
- **linear combination**
  - OLS、Alasso,、Ridge、Enet、PCR、PLS
- **Nonlinear combination**
  - GBRT、RF、ANN1、 ANN2、 ANN3、ANN4

3. FF5 alpha of univariate trading signals

	Low	2	3	8	9	High	**High - Low**
Bank Trading	0.01%	-0.08%	-0.05%	0.07%	-0.36%*	-0.11%	-0.12%
	(0.04)	(-0.47)	(-0.29)	(0.42)	(-1.99)	(-0.76)	(-0.49)
Firm Trading	-0.28%	0.10%	0.19%	-0.09%	0.04%	-0.05%	0.22%
	(-0.94)	(0.50)	(1.31)	(-0.70)	(0.61)	(-0.43)	(0.66)
Hedge Fund Trading	0.00%	-0.10%	-0.05%	0.06%	-0.13%	0.07%	0.08%
	(-0.02)	(-0.69)	(-0.38)	(0.39)	(-1.03)	(0.46)	(0.29)
Insurance Company Trading	0.10%	-0.03%	-0.16%	-0.10%	-0.27%	-0.16%	-0.26%
	(0.49)	(-0.19)	(-0.85)	(-0.42)	(-1.40)	(-1.20)	(-0.98)
Mutual Fund Trading	-0.02%	0.04%	0.08%	0.02%	0.26%*	0.00%	0.02%
	(-0.15)	(0.33)	(0.27)	(0.06)	(1.82)	(0.04)	(0.12)
Other Institutional Trading	0.00%	-0.15%	0.01%	0.21%	-0.02%	-0.10%	-0.10%
	(0.01)	(-1.18)	(0.08)	(1.57)	(-0.14)	(-0.70)	(-0.38)
Short Seller Trading	0.08%	-0.16%	-0.06%	-0.16%	0.19%**	0.07%	-0.01%
	(0.52)	(-1.45)	(-0.53)	(-1.51)	(2.19)	(0.67)	(-0.04)
Wealth Management Trading	0.06%	-0.13%	-0.16%	-0.11%	0.10%	0.07%	0.01%
	(0.29)	(-1.50)	(-0.90)	(-0.71)	(0.61)	(0.44)	(0.03)
Retail Trading_MPR	0.12%	0.22%*	0.15%	0.01%	-0.14%	-0.45%	-0.57%**
	(0.62)	(1.84)	(0.92)	(0.08)	(-0.48)	(-1.22)	(-2.09)
Retail Trading_BJZZ	0.16%	0.02%	-0.18%**	-0.09%	0.02%	0.43%**	0.27%
	(0.93)	(0.19)	(-2.44)	(-0.52)	(0.14)	(2.02)	(1.20)

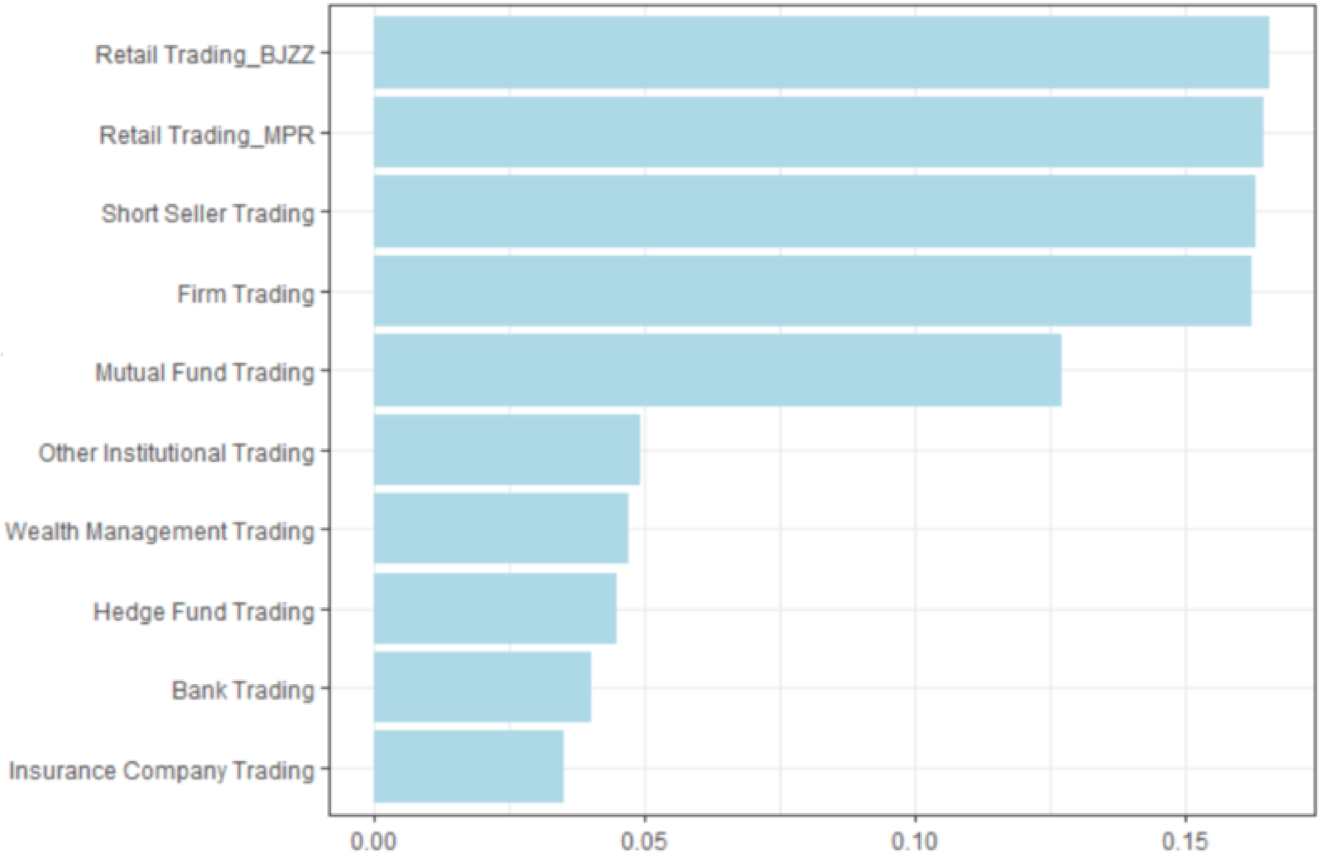
- Retail trading as measured by Boehmer et al. (2021) shows strong return predictability

3. Portfolio performance from composite return predictors

	LCP			NLCP		
	FF5	q5	MISP	FF5	q5	MISP
Low	-0.53%	-0.15%	-0.61%**	-1.04%***	-0.75%***	-0.95%***
	(-1.57)	(-0.54)	(-2.53)	(-4.69)	(-3.15)	(-5.04)
High	0.32%**	0.19%	0.32%**	0.35%**	0.36%**	0.35%**
	(2.26)	(1.23)	(2.56)	(2.15)	(2.19)	(2.35)
H - L	0.85%**	0.34%	0.93%***	1.39%***	1.11%***	1.30%***
	(2.19)	(1.02)	(3.58)	(4.88)	(3.70)	(5.73)

- The nonlinear composite return predictor (NLCP) preforms better

3. Importance of each trading signal



- These trading signals are individually insignificant in predicting returns, **collectively** they provide significant contribution to the return predictor, possibly from their **nonlinear interactions**.

3. Predicting profitability and stock fundamentals

Variable	ΔROA	ΔCF	ΔSUE	ROA	CF	SUE
NLCP	0.0597**	0.0392***	0.0062**	0.2692***	0.2517***	0.0156***
	(2.23)	(2.68)	(2.62)	(6.95)	(7.94)	(7.70)
SIZE	-0.0007	0.0085	-0.0004	0.1056***	0.1098***	-0.0002
	(-0.12)	(1.61)	(-0.46)	(18.27)	(19.42)	(-0.28)
BM	-0.0463***	-0.0289***	-0.0014	0.0380***	0.0493***	0.0002
	(-4.44)	(-4.09)	(-1.34)	(3.48)	(6.92)	(0.23)
MOM	0.0166***	0.0066*	-0.0026**	0.0338***	0.0242***	0.0005
	(3.11)	(1.98)	(-2.13)	(6.08)	(5.25)	(0.58)
STR	0.0047	0.0046*	-0.0259***	-0.0004	-0.0003	0.0008*
	(1.18)	(1.81)	(-26.17)	(-0.09)	(-0.10)	(1.78)
AG	0.0428***	0.0164**	-0.0002	-0.0371***	-0.0340***	-0.0008
	(3.57)	(2.16)	(-0.20)	(-4.45)	(-5.79)	(-1.41)
GP	-0.0924***	-0.0521***	0.0000	0.0463*	0.0599***	0.0006
	(-3.78)	(-3.26)	(0.02)	(1.73)	(3.22)	(0.76)
Obs.	226,329	222,214	286,434	230,541	223,447	288,550
R-squared	0.04	0.03	0.04	0.17	0.27	0.01

- NLCP contains information related to firm fundamentals

3. Predicting profitability and anomalies

	Low	2	3	4	High	High - Low	t-stat
Market capitalization	12.427	13.585	13.937	14.017	13.664	1.24***	10.37
Book-to-market	0.554	0.567	0.561	0.554	0.596	0.04***	2.68
Gross margin	0.187	0.291	0.298	0.301	0.293	0.11***	18.80
Illiquidity	0.810	0.874	0.864	0.765	1.154	0.34**	2.56
Idiosyncratic volatility	0.078	0.052	0.046	0.045	0.047	-0.03***	-16.27
Momentum_12m	-0.010	0.084	0.103	0.120	0.113	0.12***	5.43
Momentum_1m	-0.002	0.009	0.012	0.013	0.015	0.02*	1.97
Asset growth	0.137	0.118	0.107	0.104	0.093	-0.04***	-5.75
Dividend yield	0.011	0.013	0.014	0.014	0.014	0.00***	4.41
Analyst coverage	5.168	7.350	8.069	8.080	7.096	1.93***	6.16
Price delay	0.089	0.064	0.060	0.057	0.077	-0.01*	-1.74
Combined fundamental	3.572	4.163	4.326	4.361	4.237	0.67***	13.77

	FF5	q5	MISP	vw NLCP + market factor
Average $\|\alpha\|$	0.64%	0.52%	0.58%	0.40%
Average $\|t\|$	1.9	1.5	1.9	1.2
Delta	0.44	0.38	0.72	0.34
F-statistic	1.6*	1.3	2.8***	1.2
p-value	0.073	0.234	0.001	0.256

- The predicted returns is on the right side of most anomaly returns
- Two-factor model suggests their ability to explain the stock return anomalies

4. Conclusion

- We use machine learning to study whether the joint trading behavior of multiple market participants contains information about future stock returns

- A long-short portfolio based on the nonlinear composite predictor (NLCP) generates monthly alphas from various factor models exceeding 1%