Quantity, Risk, and Return*

Yu An† Yinan Su‡ Chen Wang§

February 17, 2025

Abstract

We propose a new model of expected stock returns that incorporates quantity information from market trading activities into the factor pricing framework. We posit that the expected return of a stock is determined by not only its factor risk exposures (β) but also the factor's quantity fluctuations (q) induced by noise trading flows, and hence term the model *beta times quantity* (BTQ). The rationale is that a factor's premium should be higher when sophisticated investors have absorbed flows of stocks with high exposure to that factor. The BTQ model provides a compelling risk-based explanation for stock returns, which is otherwise obscured without considering the quantity information. The cross-sectional risk-return association, which is nearly flat unconditionally, strongly depends on the quantity variable. The structured BTQ model reliably predicts monthly stock returns out of sample, and addresses the factor zoo problem by selecting a small number of factors.

Keywords: quantity, flow, noise trader, risk and return, cross section of return, return prediction, factor zoo, Lasso, PCA, BTQ

JEL Codes: G11, G12

*We thank Caio Almeida, Federico Bandi, Hank Bessembinder, Andrew Chen, Hui Chen, Zhi Da, Darrell Duffie, Zhiyu Fu, Robin Greenwood, Zhiguo He, Ben Hébert, Shiyang Huang, Bryan Kelly, Ralph Koijen, Nikolai Roussanov, Serhiy Kozak, Martin Lettau, Jiacui Li, Dong Lou, Sydney Ludvigson, Jun Pan, Paolo Pasquariello, Nagpurnanand Prabhala, Seth Pruitt, Alessandro Rebucci, Paul Schultz, Dongho Song, Yang Song, Zhaogang Song, Semih Üslü, Wei Wu, Jun Yu; conference discussants, Andrea Eisfeldt, Raymond Kan; and participants at the NBER Asset Pricing Program Meeting and NFA for valuable comments and suggestions. The previous version of the paper, titled "A Factor Framework for Cross-Sectional Price Impacts," was presented at the Fed Board, Wolfe, MFA, Southern Methodist University, FMCG, DC Junior, SoFiE, CICF, U. of Macau, CityU HK, CUHK, SAFE Asset Pricing Workshop, UT Dallas Finance Conference, World Symposium on Investment Research, FMA Applied Finance, Michigan Mitsui Symposium, with discussants Aref Bolandnazar, Aditya Chaudhry, Thummim Cho, Fotis Grigoris, Badrinath Kottimukkalur, Xin Liu, Marcel Müller, Andrey Pankratov, Oleg Rytchkov, Andrea Vedolin. We also thank them for valuable feedback and suggestions.

†Carey Business School, Johns Hopkins University; yua@jhu.edu.

‡Carey Business School, Johns Hopkins University; ys@jhu.edu.

§Mendoza College of Business, University of Notre Dame; chen.wang@nd.edu.

1 Introduction

Explaining the expected returns of different stocks is a central question in asset pricing. The theoretical answer is clear—risk—investors are averse to risk and require compensation for bearing risk. Therefore, riskier investments should earn higher expected returns in equilibrium. However, the empirical answer has proven more complicated: evidence of the risk-return tradeoff, such as their positive association in the cross section, is elusive in data; and risk-based models hardly predict individual stock returns, in contrast to unstructured predictions using firm characteristics and machine learning models.[1] A revamped model is critically needed for the risk-based approach to expected returns.

This paper makes headway in this important area by incorporating a new aspect of risk's economic role in determining asset prices—the *quantity* variation in investors' risk holdings induced by trading flows. Many existing efforts focus on the statistical aspects of risk, such as identifying the common factors and estimating factor premiums, and on the properties of the securities per se, such as risk exposures and firm characteristics.[2] In contrast, we show that the canonical risk framework equipped with the quantity variables, which are constructed from market trading activities and are about sophisticated investors' risk-holding conditions, yields a compelling explanation for the cross section of expected returns.

We integrate quantity into factor pricing by considering market trading activity's effect on sophisticated investors' risk holdings and, in turn, their required compensation for bearing risk. First, we acknowledge that the market is not populated with representative agents but is modeled with two groups of investors: noise investors (such as retail investors) and sophisticated investors (such as hedge funds and market makers). Noise investors generate large and correlated flows in individual stocks. Sophisticated investors take the other side of

[1] Lopez-Lira and Roussanov (2023), for example, question whether factor exposure can explain the cross section of expected stock returns. See Footnote 4 for other papers reporting an elusive risk-return relationship and Footnote 6 for those focused on predicting stock returns.

[2] These related topics constitute a large and growing body of literature. We contribute to three sub-areas with references listed in Footnotes 4, 6, and 7, respectively.

these trades, which causes fluctuations in the quantities of their holdings of the underlying systematic risks. For example, if noise investors sell a large quantity of value stocks with high HML (high-minus-low) loadings, sophisticated investors' holdings of the HML risk will increase. The sophisticated investors are the marginal investors whose demand determines asset prices. We posit that they require greater compensation for a systematic risk factor when they hold more of it. This leads to a key innovation in factor model specification: a factor's premium varies with the factor's quantity fluctuations induced by trading flows. Meanwhile, sophisticated investors enforce no-arbitrage pricing across stocks, so the canonical factor pricing condition still holds. These two forces combined give rise to our main empirical model, in which the expected return of a stock is determined by the interactions of its factor risk exposures (β) and the factors' quantity fluctuations induced by trading flows (shortened to "quantity" or q throughout the paper), which we term the *beta times quantity* (BTQ) model.

This framework, though abstracted from many details of the market microstructure, captures a significant economic force central to risk aversion that has long been missing in empirical studies of risk and return. Our approach draws from the literature that studies the price impacts of noise trading flows.[3] The novelty lies in integrating quantities into the factor pricing framework and adapting the "price impacts" mechanism to explain expected future stock returns. This advance enhances the empirical power of workhorse methods in cross-sectional asset pricing, addresses previous limitations, and leads to important empirical discoveries in three aspects.

First, quantity information elicits risk-return tradeoff relationships that would otherwise be obscured. Previous studies report a flat security market line (SML, which plots expected return $\mathbb{E}r$ against market β), inconsistent with the theoretical premise of high-risk-high-return.[4] However, a significant positive β-$\mathbb{E}r$ association emerges *conditional on* high levels

[3] See Gabaix and Koijen (2022) for a review. We discuss related papers in detail further below.

[4] Black (1972), Black, Jensen, and Scholes (1972), and Frazzini and Pedersen (2014) report a flat SML. Along this direction but with more involved investigations, Lopez-Lira and Roussanov (2023) question whether common factor exposure (β) really explains the cross-sectional variation in expected returns.

of market factor q. That is, the risk-aversion implied high-risk-high-return association holds when sophisticated investors have absorbed more market factor quantity. In this view, the previously reported flat SML is an unconditional average when the quantity information is ignored.[5] This positive association between the factor-level q and the factor's risk-return tradeoff (i.e., factor premium) holds across SMLs of other factors and in Fama-MacBeth regressions conditional on quantity information.

Second, quantity information enables a risk-based model that predicts individual stock returns. A central goal of asset pricing is to explain conditional expected returns, with the statistical prediction of individual stock returns serving as a touchstone for proposed explanations. This task is empirically difficult, and researchers have only recently achieved significant progress by using unstructured machine learning models designed for forecasting and a large number of firm characteristics, which inevitably sacrifice interpretability. The state-of-the-art machine learning methods can reliably predict stock returns at the monthly horizon, although the explained variation is small given the low signal-to-noise nature of market prices.[6] In contrast, we build an economically grounded predictor that interacts stock-level factor exposures (β) with factor-level quantity fluctuations (q). The resulting beta times quantity (BTQ) model reliably predicts the panel of monthly individual stock returns with an out-of-sample (OOS) R^2 of around 1% in various robustness settings, a level comparable to high-dimensional machine learning models. The predictability is robust to different sample periods, firm size groups, and model specifications. Without quantity, the "β-only" model exhibits no predictive power, aligning with the reported null result that using risk alone hardly explains expected stock returns (Lopez-Lira and Roussanov, 2023).

Third, quantity offers a new perspective for addressing the factor zoo problem and pro-

[5]Relatedly, Hong and Sraer (2016), Jylhä (2018), and Hendershott, Livdan, and Rösch (2020) find varying slopes of the SML conditional on investor disagreement, margin requirements, and whether returns occur during the day or night.

[6]Studies on stock (and equity portfolio) return forecasting include Fama and French (2008), Welch and Goyal (2008), Koijen and Van Nieuwerburgh (2011), Rapach and Zhou (2013), and Lewellen (2015). More recent advances with machine learning methods include Gu, Kelly, and Xiu (2020), Feng, He, and Polson (2018), Freyberger, Neuhierl, and Weber (2020), Choi, Jiang, and Zhang (2023), and Kelly, Malamud, and Zhou (2024).

viding new results on factor selection. The proliferation of proposed factors challenges the asset pricing literature in identifying factors that are important for expected returns and investors' pricing decisions. The traditional tests focus on the *existence* of factor premium: essentially, whether there is a positive spread in expected returns between stocks with high and low factor exposures in the cross section.[7] Our new test asks an upgraded question about *changes* in factor premium driven by quantity: whether the expected return spread *widens* when the sophisticated investors' factor quantity (q) is high (and vice versa).[8] For one, using quantity as an instrument for factor premium should provide more variation and, hence, greater identification power. More importantly, this upgrade is more informative of the economic mechanism through which risk aversion takes place and, therefore, should lead us closer to identifying the fundamental risks to investors. We find the market factor is the most prominent across various specifications, while other selected factors include betting-against-beta, volatility, idiosyncratic risk, and value. These results are obtained by conducting variable selection (Lasso) from a BTQ configuration that includes a large number of candidate factors (including 153 factors from Jensen, Kelly, and Pedersen, 2023, henceforth JKP). Alternatively, pre-processing the candidate factors with principal component analysis (PCA) to "shrink the cross section" (Kozak, Nagel, and Santosh, 2020) leads to a similar but even more parsimonious result in which only the first two principal components are selected, and the return predictive power is equally strong.

In summary, these three key results highlight the importance of incorporating quantity into the factor pricing framework to empirically establish a risk-based explanation of expected

[7]The proliferation of proposed factors to explain the cross section of expected stock returns (a.k.a. the factor zoo problem) is noted by Cochrane (2011), Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016), and Hou, Xue, and Zhang (2017). Existing studies address the problem by selecting or "shrinking" the factors (broadly speaking, estimating a low-dimensional factor space), including Feng, Giglio, and Xiu (2020), Lettau and Pelger (2020), Kozak, Nagel, and Santosh (2020), Giglio, Liao, and Xiu (2021), and Giglio and Xiu (2021). Essentially, they discipline a factor by whether its factor premium is positive (i.e., positive cross-sectional risk-return association). In this sense, these are developments of the more traditional Fama and MacBeth (1973) method.

[8]The new test is analogous to the difference-in-differences (DID) analysis commonly used in applied microeconomics: β captures the cross-sectional variation while q provides the time-series variation in expected returns. In this analogy, the "β-only" model has only one dimension of "difference" and assumes constant factor premiums.

returns. To sharpen this argument, we compare the BTQ model with two alternative baseline models that contain only risk or only quantity, respectively.

The first alternative, the "β-only" model, represents the traditional factor pricing framework where risk (specifically, β) is the sole determinant of differences in expected stock returns. Our main results benchmark BTQ against the "β-only" baseline, and show compelling empirical improvements in various familiar workhorse asset pricing settings: the security market line (SML), Fama-MacBeth regressions, and stock return prediction. Moreover, as discussed earlier, incorporating quantity provides an additional perspective for selecting factors from the "zoo" based on their economic relevance. Future studies can easily test newly proposed factors, as a factor's BTQ term can be easily constructed from the factor's return series. These properties highlight the advantages and broad applicability of incorporating quantity into factor pricing for future research.

Second, relative to the "quantity-only" alternative, the emphasis on risk is embedded in our construction of the q variables. They track the fluctuations of sophisticated investors' factor risk holdings induced by retail trading flows. This is achieved by aggregating stock-level flows to the factor level according to each stock's factor exposure (β), in a way similar to "portfolio beta" in risk management.[9] For example, if noise investors sell a large quantity of value stocks with high HML loadings, then from the perspective of sophisticated investors, the q of HML increases accordingly. This construction underscores the economic mechanism in which investors are averse to systematic risk, with their degree of aversion adjusting based on the amount of systematic risk they bear.[10]

This setup is contrasted with the "quantity-only" model, where stock-level flows and quantity variations directly affect stocks' expected returns, bypassing the factor structure (see Figure 6 for a comparison of the architectures). This alternative model does not adhere to the cross-sectional no-(statistical)-arbitrage condition and implies that investors are averse

[9]Stock-level noise trading flows from retail investors are constructed using mutual fund holdings and flow data, following standard procedures in the literature (Coval and Stafford, 2007; Froot and Ramadorai, 2008; Lou, 2012). See Section 3.2 for the complete construction procedure of q.

[10]Appendix A provides the theoretical foundation that formalizes this statement.

to the physical quantity of stocks rather than the systematic risk they represent. Empirically, we find little to no predictive power for stock returns in various implementations of the "quantity-only" model. This comparison highlights the critical role of risk in the BTQ model. It is consistent with the view that statistical arbitrage activities by some sophisticated investors are effective in determining the cross section of expected returns, even in the presence of significant impacts of noise trading flows on prices (Kozak, Nagel, and Santosh, 2018). It is also related to the distinction between micro and macro elasticities: stocks with similar risk loadings are close substitutes, whereas the demand for systematic risks is more inelastic to price (Gabaix and Koijen, 2022; Li and Lin, 2022).

We provide further evidence to support the economic interpretation that quantity explains expected stock returns through factor risk. We find that different factors' q variables provide distinct pricing information along their respective risk dimensions, and that BTQ variants crossing one factor's q with other factors' β terms fail to predict future stock returns. This result highlights that the quantity-risk premium association is independently robust across factors. We also find that no other conditioning variables—such as factor momentum signals and a comprehensive set of macroeconomic variables (dividend yield, default spread, income growth, etc.)—can substitute for q in reproducing BTQ's predictive power. This result rejects the idea that the factor premium variation we report is driven by other underlying economic forces and that the quantity variable is merely a facade.

In summary, the core message is that both quantity and risk matter for expected stock returns. At a high level, this naturally stems from the interaction between sophisticated investors and noise traders (Shleifer and Summers, 1990). Considering the interaction allows us to bridge factor pricing (which emphasizes rational agents' aversion to risk) and the price impact of noise flows (which emphasizes noise traders' non-fundamental flows cause price dislocation). The contribution of this empirical paper is providing a playbook for integrating quantity information into the canonical factor framework and showing its significant improvement to factor pricing models' empirical relevance.

Literature. This paper is related to two frameworks in the literature but has differences in its objective and approach. First, we do not treat flow or quantity fluctuations as a source of risk, and the constructed quantity time-series variables are not new risk factors, as in a recent paper by Dou, Kogan, and Wu (2022).[11] Instead, we still use previously proposed factors, and the newly proposed factor-level quantity variables work together with risks in the form of "beta times quantity."

Second, this paper belongs to the growing literature on demand-based asset pricing, which shows that investor demand plays a critical role in determining asset prices and that incorporating flow and quantity data can improve empirical asset pricing research (Koijen and Yogo, 2019; Gabaix and Koijen, 2022; Koijen, Richmond, and Yogo, 2024; Haddad, Huebner, and Loualiche, 2024, etc.). In particular, a strand of the literature estimates factor-level price multipliers, including Teo and Woo (2004), Peng and Wang (2021), Ben-David, Li, Rossi, and Song (2022a), Li (2022), Li and Lin (2022), and Huang, Song, and Xiang (2024). We focus on the empirical study of expected future stock returns rather than impacts on contemporaneous prices.[12] In this regard, our goal and approach align more closely with the factor pricing literature: we explicitly model the factor structure of returns; maintain the associated factor pricing condition; and take return prediction accuracy as the central criterion of empirical success.[13]

Additionally, our use of the no-arbitrage factor pricing (APT) condition to link the cross-sectional quantity-return relationship also differs from existing approaches, such as nested

[11] Other papers that treat flow or quantity information as sources of risk include De Long, Shleifer, Summers, and Waldmann (1990), Shleifer and Vishny (1997), Lo and Wang (2000), Hasbrouck and Seppi (2001), Adrian, Etula, and Muir (2014), and He, Kelly, and Manela (2017).

[12] Appendix A discusses the connection between flow/quantity's impact on contemporaneous prices and expected future returns (risk premiums) with a formal theoretical microfoundation.

[13] Koijen and Yogo's (2019) demand system models a stock's demand elasticities with respect to a) the stock's price (or the market capitalization) and b) the stock's factor risk exposures (proxied by the stock's characteristics). Neither is exactly our channel: a) operates at the stock level, rather than the factor level, and b) is about the cross-sectional demand variation related to a stock's factor loadings or characteristics, rather than time-series demand variation driven by aggregated factor risk quantity. They use the factor framework as a microfoundation for the characteristic-based demand system. Related to our research objective, one of their applications shows that mean reversion in latent demand introduces a new source of predictability for cross-sectional variation in stock returns.

logit demand systems in Koijen and Yogo (2020), Bretscher, Schmid, Sen, and Sharma (2024), and Jiang, Richmond, and Zhang (2024), controlling for close substitutes as in Chaudhary, Fu, and Li (2023), and mean-variance optimization as in Vayanos and Vila (2021), Davis, Kargar, and Li (2024), and Jansen, Li, and Schmid (2024).[14]

In the remainder of the paper, Section 2 provides the theoretical motivation, empirical model, and methods; Section 3 constructs the quantity and other empirical measures; Section 4 presents the main empirical results; Section 5 shows that quantity must be combined with risk to forecast returns; Section 6 investigates alternative economic channels; Section 7 concludes.

2 Theoretical motivation, empirical model, and methods

2.1 Theoretical motivation

The theoretical rationale for integrating quantity information into factor pricing to explain *expected* stock return is that market trading activities affect sophisticated investors' risk holdings and, in turn, their required compensation for bearing risks. We focus on a prominent channel where the noise trading flows—a significant type of trading activities—matter for the central element of asset pricing, namely the factor premium, although there can be many other market microstructure mechanisms in which trading activities generate contemporaneous price impacts. We outline this theoretical channel below, and Appendix A provides the formal microfoundation.

Suppose the market is populated with two groups of investors: noise investors and sophisticated investors. Noise investors, such as retail traders, generate uninformed flows in and out of individual stocks over time. The noise flows are large and correlated across stocks,

[14]Relatedly, Berk and Van Binsbergen (2016), Barber, Huang, and Odean (2016), and Ben-David, Li, Rossi, and Song (2022b) use a revealed preference approach to determine which factors investors care about, and Bretscher, Lewis, and Santosh (2023) show that betas measured relative to institutional investor portfolios explain stock returns.

which can induce significant fluctuations when aggregated to the factor level.[15]

Sophisticated investors, such as hedge funds and market makers, take the other side of the retail trades by absorbing the noise flows and supplying liquidity. Therefore, noise flows induce fluctuations in the sophisticated investors' holding quantities of the underlying systematic risks. For example, if retail investors sell lots of value stocks with high HML exposures, then sophisticated investors will accumulate more HML risk holdings. The aggregation from stock-level flows to factor-level quantities accounts for each stock's factor exposure (β) in the fashion of "portfolio beta" commonly used in risk management practice (see Section 3.2 for aggregation details). The sophisticated investors are the marginal investors whose risk-holding conditions drive asset prices. They have limited capacity to bear risk and absorb flows, and require greater compensation for a systematic risk factor when they hold more of it.[16] This gives rise to the key model specification that a factor's premium varies with the factor's quantity fluctuations induced by trading flows, and we hypothesize that the relationship is positive.[17] Meanwhile, sophisticated investors enforce no-arbitrage pricing across stocks, so the canonical factor pricing condition still holds.[18] These two forces combined imply the main empirical model specified below, in which both the stock's factor risk exposures (β) and factor quantity (q) determine its expected return.

[15]Previous studies report (which we also confirm empirically) that the retail flows are not only significant in magnitude but also correlated across stocks due to the commonality in retail investors' trading behaviors. The correlation aligns with investment styles, such that, say in one period, retails tend to sell growth stocks and in the next, they buy small stocks (Li, 2022; Huang, Song, and Xiang, 2024). This fact supports that retail flows can induce significant fluctuations in the quantity of risk when aggregated to the factor level.

[16]Limited risk-bearing capacity can stem from liquidity or balance-sheet constraints (e.g., Adrian, Etula, and Muir, 2014; Gabaix and Maggiori, 2015; He, Kelly, and Manela, 2017; Kondor and Vayanos, 2019; Haddad and Muir, 2021). In particular, Eisfeldt, Herskovic, and Liu (2024) and Kargar (2021) emphasize that heterogeneity within the intermediary sector can further lead to risk misallocation, offering a novel explanation for why liquidity is priced.

[17]Appendix A provides the formal theoretical model to microfound the quantity-factor premium association (Eq. 3 specified further below). This specification is related to the demand-based literature, which emphasizes the "price multiplier" is high, or, in other words, the demand is inelastic to price (Gabaix and Koijen, 2022). The empirical distinction is that our goal is explaining the expected future returns, rather than the contemporaneous price impact (although the two are theoretically connected as high expected returns imply low current prices).

[18]Enforcing the cross-sectional APT condition is consistent with Kozak, Nagel, and Santosh (2018), who argue that cross-sectional no-arbitrage conditions are still valid in the presence of noise traders as long as there exist some sophisticated investors. This is in contrast to those models that directly link each individual stock's flow to its price. See Section 5.1 for the comparison against this benchmark.

2.2 Empirical model

The empirical model starts with the canonical factor pricing framework, in which the cross section of stock returns follows a factor structure

$$r_{i,t+1} = \sum_{k=1}^{K} \beta_{i,k,t} f_{k,t+1} + \epsilon_{i,t+1}, \qquad \forall i, t, \tag{1}$$

where $r_{i,t+1}$ is the excess return of stock i in month $t+1$, k indexes factors, f is factor return (zero-cost or excess return), and β is the stock's factor exposure, which is subsequently estimated using realized daily returns. According to the APT (Ross, 1976), the cross section of expected return follows the factor pricing condition,

$$\mathbb{E}_t[r_{i,t+1}] = \sum_{k=1}^{K} \beta_{i,k,t} \mu_{k,t}, \qquad \forall i, t, \tag{2}$$

where $\mathbb{E}_t[r_{i,t+1}]$ is the conditional expected stock return, our research object, and $\mu_{k,t}$ is the factor premium conditional on time-t information.

The departure from the canonical framework lies in the modeling of the factor premium. According to the theoretical motivation above, we specify that the factor premium is not a constant but varies with the factor's quantity fluctuations induced by trading flows.

$$\mu_{k,t} = \mu_k(q_{k,t}) = \mu_k + \lambda_k q_{k,t}, \qquad \forall k, t, \tag{3}$$

where the first is a general non-parametric form in which $\mu_k(\,\cdot\,)$ is an unspecified function of $q_{k,t}$, while the second is the parametric linear specification, which is implemented in most empirical settings.[19] Parameter μ_k corresponds to the constant factor premium, which is the key interest of estimation in traditional factor pricing tests. The linear coefficient λ_k is the new central parameter of interest, which measures the sensitivity of the factor premium to

[19] The various parametric and non-parametric empirical methods are detailed further below in Section 2.3.

the factor's quantity fluctuations.[20]

Plugging the factor premium specification into the factor pricing condition (Eq. 3 into Eq. 2), we arrive at the main empirical model, the beta times quantity (BTQ) model of expected stock returns:

$$\mathbb{E}_t[r_{i,t+1}] = \left(\sum_{k=1}^{K} \mu_k \beta_{i,k,t}\right) + \sum_{k=1}^{K} \lambda_k \beta_{i,k,t} q_{k,t}, \qquad \forall i, t. \tag{4}$$

The first summation term is the traditional factor pricing model, which we refer to as the "β-only" model, serving as the baseline in empirical comparisons. The second is the new beta times quantity (BTQ) term. In empirical implementation, we often find the β-only term is so close to zero (and so noisy for explaining expected returns) that including it in the BTQ model can even hurt the empirical fit. Therefore, the BTQ model typically omits the β-only term in parentheses and only includes the beta times quantity term.

The key hypothesis implied by the theoretical motivation is that, for a "true" fundamental risk factor k, $\lambda_k > 0$. The hypothesis means that the cross-sectional return dispersion between high and low β stocks widens when the factor's quantity is high. This is similar to the difference-in-differences (DID) analysis: β captures the cross-sectional variation in expected returns while q provides the time-series variation. In other words, the observed factor risk aversion is stronger when q is high. This offers a new perspective compared to the traditional hypothesis $\mu_k > 0$, which asks whether higher exposure to that factor is associated with higher *average* returns, i.e., only the first "difference" in the DID analysis. The new test has more identification power provided by the time-series variation in q. More importantly, this test has more economic relevance since the q variation tracks sophisticated investors' risk-holding conditions. Hence, we are no longer inferring investors' risk pricing process from asset and asset price information alone. Therefore, the new framework can lead

[20] Appendix A.3 provides a microfoundation for the linear specification and the economic interpretation of its parameters. The parameter λ_k reflects the inelasticity of sophisticated investors' demand for factor risk. This inelasticity is further attributed to two primitives: high risk aversion and the limited capital of sophisticated investors relative to the aggregate stock market.

us closer to identifying the fundamental risks that investors care about.

The model allows for multiple factors and allows each to have a different λ_k coefficient. This is useful for testing each factor's marginal importance in a joint setting, controlling for other factors' contribution to expected returns.[21]

An important property of the sign of λ_k is noted. Regardless of the sign of the factor (e.g., small-minus-big or big-minus-small), the sign of λ_k should, theoretically speaking, always be positive. This is because when factor return f flips its sign, both β and q flip their signs, and β times q remains unchanged. A positive λ_k estimate, nonetheless, is not empirically guaranteed. Thus, it provides another layer of testing for the risk-based theory, regardless of the specification of the factor's sign. A negative λ_k estimate would be an unambiguous rejection of the risk-based theory, and the empiricist could not blame the "wrong" sign of the factor as an excuse. Notice that μ_k in the traditional β-only model does not have this property: big-minus-small would have a negative μ_k.

We focus on testing the hypothesis "$\lambda_k > 0$" in the cross-sectional setting of the BTQ model (Eq. 4), not in the time series context of predicting factor returns $f_{k,t+1}$ with $q_{k,t}$. Although the BTQ model is theoretically motivated by the time-series specification of factor premium (Eq. 3), empirically, a positive time-series predictive coefficient between $q_{k,t}$ and $f_{k,t+1}$ is far from implying the cross-sectional hypothesis of $\lambda_k > 0$. The gap between the two is the cross-sectional variation of the risk exposures (β), which is not present in the time series setting. A similar gap is familiar in the traditional factor pricing framework: a long-short portfolio with a high average return does not guarantee that it is a priced factor in cross-sectional tests, such as the Fama-MacBeth regressions.

[21] The model (Eq. 3) specifies that factor k's premium $\mu_{k,t}$ is affected only by its own quantity $q_{k,t}$, not by the quantities of other factors $q_{j,t}$. Allowing for cross-factor impacts would complicate the model, increasing the number of parameters from K to K^2, which becomes impractical for large K. Our most salient empirical results are attained with single-factor settings, where cross-factor impacts are irrelevant.

2.3 Empirical methods

We use a series of empirical methods to estimate and test the BTQ model. The methods are presented as upgrades of familiar procedures in asset pricing, such as the security market line, Fama-MacBeth factor premium estimates, and return prediction exercises, for ease of comparison and to demonstrate the value of incorporating quantity information into the factor model. We present an overview of the methods here, while the details are provided when presenting the empirical results in Section 4.

From the methodological perspective, the progression of the methods can be seen as gradually adding parameterization to the model of expected stock returns. To start with, the familiar security market line (SML) can be seen as a simple non-parametric model, $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t})$, where $Er(\,\cdot\,)$ is an unspecified function. (The SML is typically estimated with the market beta, i.e., $k = \text{MKT}$, but we implement it with other factors as well.) The conditional SML (Section 4.1) upgrades it to a bi-variate non-parametric model that includes q, $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$. We estimate this non-parametric model with a simple kernel method by binning observations of β and q. This method is easy to interpret via the familiar SML plot, and clearly shows that q is a highly relevant variable in the expected return function (Er) with significant effects on the risk-return (β-$\mathbb{E}r$) relation.

The second method, the quantity upgraded Fama-MacBeth factor premium estimates, is semi-parametric (Section 4.2). It imposes a linear relationship between risk (β) and expected return according to APT, but is still non-parametric about q's effect: $Er(\beta_{i,k,t}, q_{k,t}) = \beta_{i,k,t}\mu_k(q_{k,t})$, where the factor premium function $\mu_k(\,\cdot\,)$ is left unspecified. It is still estimated non-parametrically by binning q and then averaging the returns of the Fama-MacBeth factor mimicking portfolio (FMP, which are coefficients from the cross-sectional regression $r_{i,t+1}$ on $\beta_{i,k,t}$) within each bin.

Third, once the $\mu_k(\,\cdot\,)$ function is also specified as linear, we arrive at the parametric BTQ model $Er(\beta_{i,k,t}, q_{k,t}) = \lambda_k\beta_{i,k,t}q_{k,t}$. The parametric setting easily accommodates multiple factors, and is estimated with a linear predictive regression on the panel of monthly stock

returns $r_{i,t+1} = \sum_{k=1}^{K} \lambda_k \beta_{i,k,t} q_{k,t} + error_{i,t+1}$ (Section 4.3). Notice that each factor's beta times quantity (BTQ) term together serves as a predictor, and the BTQ terms of different factors serve as multivariate predictors. Predicting stock returns has experienced significant progress with firm characteristics and machine learning models. We follow the literature's stock-month panel setup and use the same measure of empirical success: the monthly stock return predictive R^2 evaluated out-of-sample (OOS). This is our (and also the literature's) key evaluation metric for "explaining expected stock returns."

Lastly, in response to the factor zoo problem, when the number of candidate factors (K) is large, the number of BTQ predictors grows accordingly to more than 100. In such a setting, we use machine learning methods designed for high-dimensional prediction, such as Lasso, to select a small number of priced factors (Section 4.4). By inducing sparsity in the λ_k coefficients, Lasso allows us to select a small number of BTQ terms and reveal which factors are priced in a joint setting, controlling for other factors. Additionally, we follow Kozak, Nagel, and Santosh (2020) and pre-process the candidate factors with principal component analysis (PCA). Then, we supply the principal component factors to the same BTQ construction and Lasso prediction exercise (Section 4.5). The potential benefit of this method is to "shrink the cross section" of factors and elicit latent factors that capture most of the time-series return variation among the many candidates. According to existing literature, such latent factors are often more reliable for explaining expected returns.

In summary, we put forward the message that integrating quantity information into various empirical methods can lead to significant empirical discoveries. We implement the methods outlined above to support this message, but they are far from exhaustive given the vast asset pricing literature. We believe these quantity variables can similarly interact with many other existing methods, opening a broad avenue for further empirical discoveries.

3 Constructing quantity (q) and other variables

The data to run a BTQ predictive regression include the (unbalanced) panel of monthly excess stock returns $r_{i,t+1}$, along with a panel of $\beta_{i,k,t}$ and a time series of $q_{k,t}$ for each factor k, which serve as right-hand side predictors. Among these, $\beta_{i,k,t}$ is constructed from the time series of factor return $f_{k,t}$ as in the first stage of the Fama-MacBeth procedure. The construction of $q_{k,t}$ is new. It requires the stock-level retail flow in the same unbalanced panel structure as the returns, which is then aggregated to the factor level according to each stock's factor exposure measures. In summary, the source data are only the panel of returns and the panel of flows at the stock level, from which one can calculate both β and q for any factor, given the time series of factor returns $f_{k,t}$.

3.1 Return, risk, and flow variables constructed with standard procedures

The factor and stock return, risk exposure, and stock-level dollar flow variables are all constructed using data sources and procedures standard in the literature.

We use delisting-adjusted stock returns from CRSP. The six Fama-French-Carhart (i.e., Fama and French, 1993, 2015; Carhart, 1997) factors are from Kenneth French's website, and the 153 Jensen, Kelly, and Pedersen (2023, JKP) factors are from the authors' website. All returns are obtained in both daily and monthly frequencies in excess of the risk-free rate.

Each stock's exposure to factor k in month t is

$$\widehat{\beta}_{i,k,t} := \frac{\widehat{\mathrm{cov}}_t(r_{i,t}, f_{k,t})}{\widehat{\mathrm{var}}_t(f_{k,t})}, \qquad \forall i, t, k, \tag{5}$$

where $\widehat{\mathrm{cov}}_t$ and $\widehat{\mathrm{var}}_t$ are realized covariance and variance estimated with daily returns in a 12-month rolling window up to month t.[22]

[22]Notice $\widehat{\beta}_{i,k,t}$ corresponds to the regression coefficient of a single-factor model. This differs from the original Fama-MacBeth procedure, where the first stage is a multi-factor regression. A single-factor beta is simply the realized covariance normalized by scalar variance and offers two advantages. First, multi-factor regressions can be unreliable even with a moderately high number of factors. Second, a single-factor beta, and consequently each factor's BTQ term, can be constructed independently of other factors in the model,

We construct the stock-level dollar flow $\$\text{flow}^{\text{stock}}_{i,t}$ panel using the mutual fund flow-induced trading (FIT) metric, proposed by Coval and Stafford (2007), Froot and Ramadorai (2008), and Lou (2012). We use the standard mutual fund data source but carefully clean data errors by cross-validating several sources. In particular, we obtain monthly mutual fund returns and characteristics from the CRSP Survivorship-Bias-Free Mutual Fund database and quarterly holdings data from the Thomson/Refinitiv Mutual Fund Holdings Data (S12). Our sample period spans from January 2000 through December 2022.[23] The mutual fund sample comprises both active and passive mutual funds. To ensure accuracy in our flow measure, we cross-validate mutual funds' monthly returns and total net assets (TNA) obtained from the CRSP database with corresponding data from Morningstar and Thomson/Refinitiv. In the process, we manually correct several data input inaccuracies. Details regarding this process are in Appendix B.1.

The standard $\$\text{flow}^{\text{stock}}_{i,t}$ construction procedure has three steps. First, dollar mutual fund flows are

$$\$\text{flow}^{\text{fund}}_{m,t} := \text{TNA}_{m,t} - \text{TNA}_{m,t-1}(1 + r^{\text{fund}}_{m,t}), \tag{6}$$

where $\text{TNA}_{m,t}$ is the total net assets of mutual fund m at the end of month t, and $r^{\text{fund}}_{m,t}$ is mutual fund m's net-of-fee return in month t.

Second, we allocate mutual fund flows to dollar stock-level flows, based on the established assumption in the literature that mutual funds buy or sell stocks in proportion to their prior holdings,

$$\$\text{flow}^{\text{stock}}_{i,t} := -\sum_{\text{fund } m} \$\text{flow}^{\text{fund}}_{m,t} \text{weight}^{\text{fund}}_{i,m,\text{quarter}(t)-2} \ . \tag{7}$$

The negative sign is used to shift the perspective from retail investors to sophisticated in-

allowing for a more convenient empirical procedure. See Feng, Giglio, and Xiu (2020) for a related discussion, who also use covariances rather than multi-variate betas.

[23] We start the sample period in 2000, following the convention in the literature. The mutual fund industry experienced significant growth and sustained inflows throughout the 1990s (Lou, 2012; Ben-David, Li, Rossi, and Song, 2022a). Since 2000, the size of the mutual fund sector has remained stable relative to the total equity market, resulting in stationary monthly flow shocks.

vestors when accounting for the flow. Specifically, a positive $\$\text{flow}_{i,t}^{\text{stock}}$ dollar value indicates that retail investors are selling stock i in month t, while sophisticated investors are buying. Moreover, we use the two-quarter-lagged mutual fund holding weight, denoted as $\text{weight}_{i,m,\text{quarter}(t)-2}^{\text{fund}}$. For instance, $\text{quarter}(\text{July}) - 2 = \text{Q1}$.[24]

In total, we have around 1,644,000 stock-month observations in a full sample of 276 months from January 2000 to December 2022, or on average around 6,000 stock-month observations per month.

3.2 Constructing quantity variables

The construction of $q_{k,t}$ is guided by the theoretical motivation outlined in Section 2.1 and the microfoundation detailed in Appendix A.3. It involves two steps. First, we aggregate stock-level flows to the factor level, using the same risk measures, $\widehat{\text{cov}}_t(r_{i,t}, f_{k,t})$, from Eq. 5:

$$\text{flow}_{k,t}^{\text{factor}} := \sum_i \$\text{flow}_{i,t}^{\text{stock}} \widehat{\text{cov}}_t(r_{i,t}, f_{k,t}) = \sum_i \$\text{flow}_{i,t}^{\text{stock}} \widehat{\beta}_{i,k,t} \widehat{\text{var}}_t(f_{k,t}), \qquad \forall k, t. \tag{8}$$

The aggregation accounts for each stock's factor exposure, in a similar spirit to calculating the portfolio beta commonly used in risk management. The second expression in Eq. 8 is for explaining the intuition: every month, sophisticated investors add a marginal portfolio to their existing holdings in response to retail flows, and $\$\text{flow}_{i,t}^{\text{stock}}$ is the dollar weights of this portfolio. The portfolio's risk characteristics are determined by its composition (portfolio weights $\$\text{flow}_{i,t}^{\text{stock}}$), as well as each constituent stock's factor exposures ($\widehat{\beta}_{i,k,t}$). For example, if retail investors sell a large quantity of value stocks with high HML loadings, the sophisticated investors' HML quantity would experience a positive flow shock.[25] Moreover, multiplying

[24]The use of a two-quarter lag deviates from the conventional one-quarter lag (Lou, 2012) to be more conservative and ensures that the constructed $\$\text{flow}_{i,t}^{\text{stock}}$ is observable with information up to month t. In particular, mutual fund holding is reported with a maximum statutory delay of 45 days (Christoffersen, Danesh, and Musto, 2015), which means the end of Q2 holdings may not be observable in July. By using a two-quarter lag, July relies on the end of Q1 holdings, which are guaranteed to be available. Our results remain robust when we apply the one-quarter lag commonly used in the literature. These results are available upon request.

[25]Notice we aggregate flow to the factor level (HML in this example) based on each stock's HML exposure (β), not on the stock's characteristics (the book-to-market ratio) or its weight in the HML portfolio. This

by $\widehat{\text{var}}_t(f_{k,t})$ modulates the portfolio's risk by the time-series fluctuation in factor return variance.[26] In this sense, we are indeed tracking the quantity of factor *risk*, not the physical quantity of securities or portfolios.[27]

Second, the flow shocks $\text{flow}_{k,t}^{\text{factor}}$ are normalized by the lagged total US stock market capitalization and accumulated in a six-month lookback window,

$$\widetilde{q}_{k,t} := \frac{1}{h}\sum_{h'=0}^{h-1} \frac{\text{flow}_{k,t-h'}^{\text{factor}}}{\text{total stock market cap}_{t-h'-1}}, \qquad \forall k,t, \qquad \text{with } h=6. \tag{9}$$

This normalization accounts for the upward trend in dollar flows, which reflects the overall growth of the equity market, as well as the increasing capacity of sophisticated investors to absorb these flows.[28] Accumulating $\text{flow}_{k,t}^{\text{factor}}$ over time accounts for the persistent effects of older flows on future returns. What matters for the expected return in month $t+1$ is the factor quantity held at the end of month t, which is impacted by flow shocks in all previous periods, $\text{flow}_{k,t}^{\text{factor}}, \text{flow}_{k,t-1}^{\text{factor}}, \text{flow}_{k,t-2}^{\text{factor}} \ldots$ The speed at which sophisticated investors can absorb these shocks and eliminate their effect on risk premiums is not our research focus. We accumulate past flows in a 6-month lookback window for simplicity and transparency to avoid a more involved study of the speed. The empirical results are robust to alternative specifications (see Section 4.6).

In many empirical exercises, we standardize the raw $\widetilde{q}_{k,t}$ time series as $q_{k,t} := \widetilde{q}_{k,t}/\sigma(\widetilde{q}_{k,t})$, where $\sigma(\widetilde{q}_{k,t})$ is the full-sample time-series standard deviation, for ease of interpreting the regression coefficients.

---

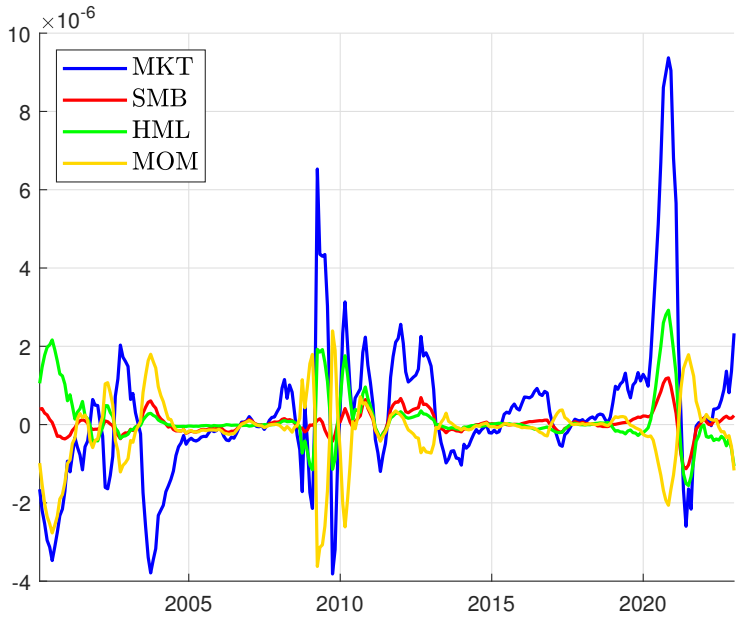
choice is based on the theoretical motivation that sophisticated investors are *averse to factor risk*, not the factor portfolio itself. The goal is to measure the quantity variation in each factor's risk, not the factor portfolio itself. Li (2022) aggregates using portfolio weights, which can be reconciled with our framework if characteristics are viewed as proxies for factor exposures.

[26] More specifically, the variance term arises in the theoretical model that assumes CARA utility for sophisticated investors (see Appendix A.3).

[27] Appendix B.1.4 discusses an alternative method that directly constructs factor-level flows from mutual fund flows.

[28] Appendix A.3 provides a theoretical justification for normalizing by the total stock market capitalization under the assumption that the fraction π of sophisticated investors relative to the total stock market remains constant over time. The smaller this fraction (π), the more sensitive the risk premium (the higher λ).

Figure 1: Quantity ($\widetilde{q}_{k,t}$) time series plot



Note: Time series of the constructed quantity ($\widetilde{q}_{k,t}$) variables for the Fama-French-Carhart factors. The monthly observations span from January 2000 to December 2022.

3.3 Basic properties of the constructed quantity variables

Next, we present the summary statistics of the flow-induced quantity, $\widetilde{q}_{k,t}$, the key new variable introduced in this paper. Figure 1 shows the time-series plots of $\widetilde{q}_{k,t}$ for the Fama-French-Carhart (FF3C) factors. We plot the pre-standardized series $\widetilde{q}$ to show magnitudes.[29] Table 1 presents the full-sample statistics of FF3C's $\widetilde{q}$ and summaries of these statistics across the 153 JKP factors.

Examining the basic time series properties of $\widetilde{q}_{k,t}$, we find that variation dominates its trend, making quantity fluctuation the primary feature compared to the secular trend in retail flows. The series also exhibits dynamic volatility clustering, similar to that seen in more familiar factor return time series.

Among the four factors plotted in Figure 1, MKT's quantity (in blue) has the most time-

[29]The magnitudes of $\widetilde{q}$ are in the unit of 10^{-6}. The absolute level is irrelevant for empirical analysis, as the variables are standardized in regressions. To understand this magnitude, we know the monthly mutual fund flows are in the order of tens of billions of dollars, and the total market capitalization is in the order of tens of trillions of dollars (see Appendix Figure A.2). So the first term in Eq. 8 is in the order of 10^{-3} (given market β around 1). The last term, monthly $\widehat{\text{var}}_t(f_{k,t})$ is in the order of 10^{-3}, so $\widetilde{q}$ is in the order of 10^{-6}.

Table 1: Summary statistics of quantity $\widetilde{q}_{k,t}$ (unit: 10^{-6})

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
Mean	0.29	0.04	0.13	-0.15	-0.05	-0.01	0.03
Std	1.88	0.29	0.65	0.82	0.23	0.39	0.76

Note: The mean and standard deviation of the constructed quantity time series $\widetilde{q}_{k,t}$ for the Fama-French-Carhart factors and JKP factors.

series variation. The reason is that most stocks have positive market beta centered around one, so $\widetilde{q}_{\text{MKT},t}$ roughly aggregates the *overall* retail flows into (and out of) the entire mutual fund sector. In contrast, the three long-short factors have stock betas that are more evenly distributed around zero, so their $\widetilde{q}_{k,t}$ series reflect the *net* retail flows into (and out of) stocks of particular investment styles. Therefore, these series are not mechanically correlated, even though they are all constructed from the same retail flow panel data.

Appendix C.1 reports that the pairwise correlations of the four $q_{k,t}$ series are far from ± 1, indicating that series are not collinear. It also reports a principal component analysis (PCA) on the $q_{k,t}$ series for the 153 JKP factors. These q series have a multi-factor structure with independent variation along various principal dimensions as well as substantial idiosyncratic variation. Section 5.2 further shows each factor's q provides distinct and independent pricing information along its respective risk dimension. These results suggest BTQ's consistent predictive power across different factors is not mechanically driven by one (or a few) special "secrete sauce" q series, highlighting the robustness of the underlying economic mechanism.

Turning to notable spikes in the plot, we note $\widetilde{q}_{\text{MKT},t}$ experiences significant increases during the Global Financial Crisis and the COVID-19 pandemic in the spring of 2020. These spikes are attributed to significant outflows from mutual fund investors during these periods. As a result, the sophisticated investors' risk holding quantity increases, making them more "averse" to the market risk, which can be related to market crashes and subsequent rebounds. However, this is a highly simplified and anecdotal explanation of the main eco-

nomic mechanism, as it does not consider cross-sectional variation in factor exposures, more nuanced fluctuations, or factors beyond MKT. Next, we turn to formal empirical analysis.

4 Main empirical results

4.1 Security market line (SML) depends on quantity

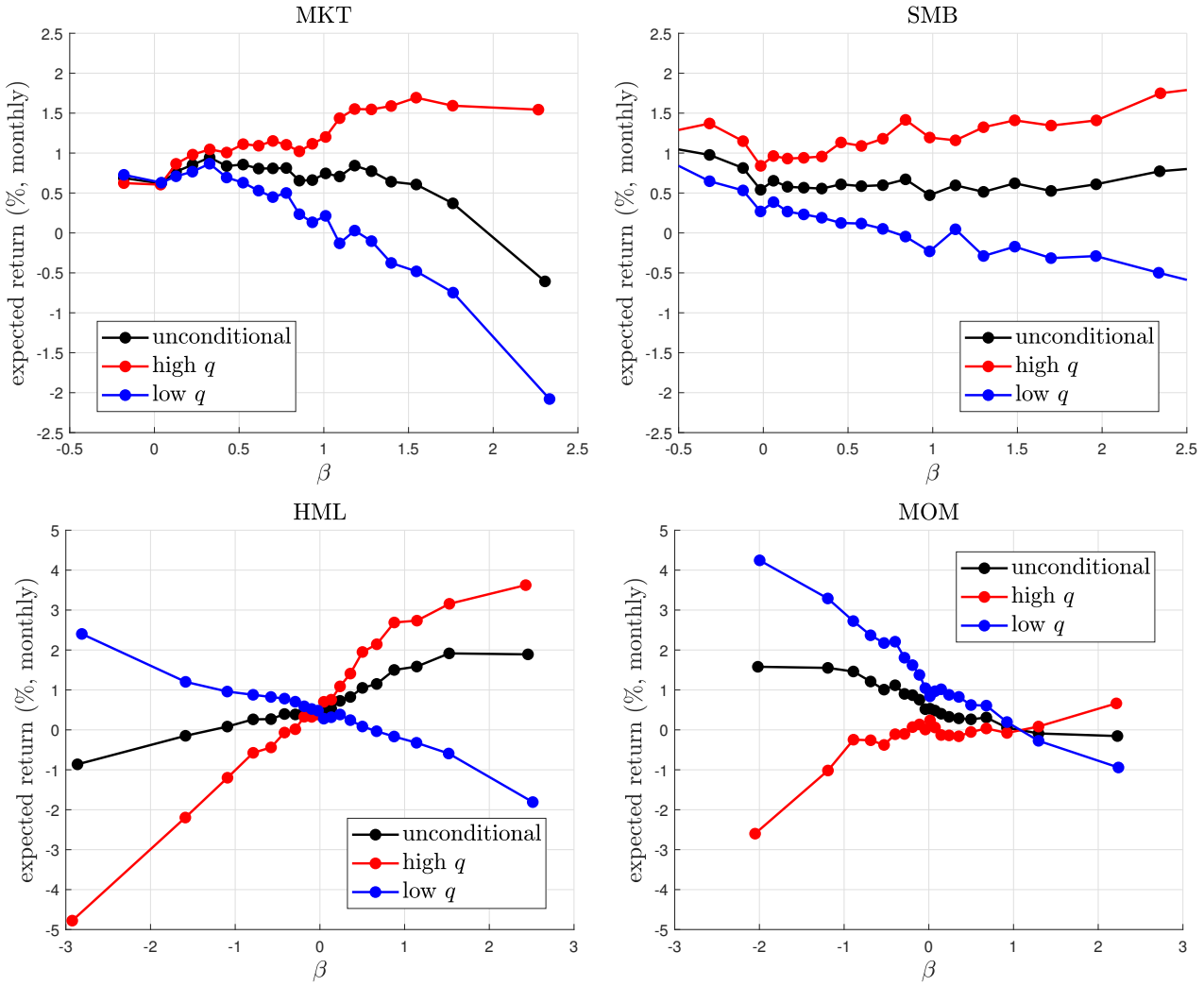
The security market line is a simple and commonly used tool to visualize the relationship between systematic risk exposure and expected return (β-$\mathbb{E}r$) in the cross section of stocks, without relying on parametric modeling. We construct the empirical SML and its conditional versions based on factor q. We show that the β-$\mathbb{E}r$ relationship is nearly flat unconditionally, which is consistent with the existing empirical results that factor exposure alone cannot adequately explain the cross-sectional variation in stock returns. However, once conditional on quantity information, the SML reveals interesting risk-return patterns that strongly support a risk-based explanation.

The unconditional SML displays the β-$\mathbb{E}r$ relationship in the non-parametric regression model: $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t})$. We estimate it with a simple kernel method by sorting stock-month observations into twenty quantile bins by $\widehat{\beta}_{i,k,t}$, and then plotting the average of $r_{i,t+1}$ against the average $\widehat{\beta}_{i,k,t}$ within each bin. Notice that return $r_{i,t+1}$ leads $\widehat{\beta}_{i,k,t}$ by one month, so that it estimates conditional expected returns.

The upgraded SML *conditional on quantity* estimates the bi-variate non-parametric model: $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$. Our purpose is to show that the second entry, q, matters for the risk-return relationship. Again, we conduct a simple non-parametric estimation for transparency and intuitiveness. The estimation procedure is the same as the unconditional SML, but we further split each bin of stock-month observations into two sub-bins by the time-series median of $q_{k,t}$, and plot sub-bin average $r_{i,t+1}$ against average $\widehat{\beta}_{i,k,t}$.[30]

[30]Formally, an unconditional bin is defined as $\{(i,t)$ s.t. $\widehat{\beta}_{i,k,t} \in [a,b)\}$, where a and b are boundaries of the 20 quantiles of $\widehat{\beta}_{i,k,t}$, for example, the first pair is $\big[\text{quantile}(\widehat{\beta}_{\cdot,k,\cdot}, 0\%), \text{quantile}(\widehat{\beta}_{\cdot,k,\cdot}, 5\%)\big)$. A "high q" bin is defined as $\{(i,t)$ s.t. $\widehat{\beta}_{i,k,t} \in [a,b)$ and $q_{k,t} \geq \text{median}(q_{k,t})\}$, where $\text{median}(q_{k,t})$ is the time-series median of $q_{k,t}$. And, "low q" is the same as "high q" but with "$\geq$" replaced by "$<$".

Figure 2: Security market line (SML) conditioning on quantity: $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$



Note: Security market line (SML) plots expected stock returns against β. The unconditional SML (black): sorts the stock-month observations into twenty quantile bins of $\widehat{\beta}_{i,k,t}$ and plots the average return $r_{i,t+1}$ against average $\widehat{\beta}_{i,k,t}$ within each bin. The conditional SMLs (red for high q, blue for low q): the same process but split bins by the time-series median of $q_{k,t}$. Notice the scales of the x- and y-axes in the bottom two panels are zoomed out by a factor of two to accommodate the larger ranges of HML and MOM β's.

Figure 2 presents single-factor models using the Fama-French-Carhart factors (MKT, SMB, HML, MOM). The black curves represent the unconditional SMLs, while the red and blue curves correspond to conditional SMLs for high and low $q_{k,t}$, respectively.

We find that the unconditional SML is nearly flat for the market factor, with a slight downward slope in the higher beta range. This implies that the market beta *alone* cannot explain the cross-sectional variation in expected returns, which is consistent with similar

reports in the existing literature. Similar null results for unconditional SMLs are observed for SMB and MOM, while HML's SML is slightly upward-sloping.

In contrast, the conditional SMLs show interesting risk-return patterns that are not observable without conditioning on q. The high-q SMLs (red) exhibit a strong positive slope, while the low-q (blue) SMLs are downward sloping. The unconditional SML (black) lies in between these conditional SMLs as the mixed average. The gaps in the slopes suggest that sophisticated investors' risk-holding conditions matter for their demand for risk, which in turn significantly impacts the pricing of factor risks in the cross section. Notice the four plots are produced with different $q_{k,t}$ time series and $\widehat{\beta}_{i,k,t}$ panels, yet the slope patterns are consistent across factors. This consistency suggests that quantity's effects on factor premiums are general and robust, reflecting a stable underlying economic mechanism.

The positive high-q slope suggests that sophisticated investors demand higher additional compensation for high systematic risk in high-q environments. Conversely, the negative low-q slope indicates high-risk investments have low expected returns (or high concurrent prices). This is likely because sophisticated investors are more willing to hold high-risk investments when they are required to sell lots of such stocks to retail traders in low-q months, i.e., when they are in a relatively short position of the factor.[31]

The magnitude of q's effects is economically large. For instance, a market beta-neutral stock has an unconditional expected return of around 0.75% per month. In contrast, for a stock with a market beta of 1, the expected return is as high as 1.25% in high-q months or as low as 0.25% in low-q months, with the average still around 0.75%. The high vs. low-q gap is around 1% per month or more than 10% annualized. This gap is even greater for stocks with higher market betas. For HML, the gap is even more pronounced: a $\beta_{\text{HML}} = 1$ stock is expected to earn around 30% annually, while an HML-neutral stock's expected return remains unaffected by q, as evidenced by the crossing of the three curves at $\beta_{\text{HML}} = 0$. This result reveals that HML is a salient fundamental factor for sophisticated investors, as both

[31] The sophisticated investors' risk management mechanisms as described in Frazzini and Pedersen (2014) can provide a potential explanation for the pricing behavior in the low-q months.

high β exposure and high quantity holdings are compensated by significantly higher risk premiums. For the SMB factor, while the general patterns of SML slopes remain consistent, the effects of both β and q are smaller in magnitude compared to the other factors. We provide additional support for these findings and present more precise point estimates using parametric estimations further below.[32]

All SMLs, regardless of their slopes, are approximately straight lines, regardless of their slopes, particularly around the central range of β, where most stocks are concentrated, and sampling noise is less pronounced. This linearity in β is consistent with the cross-sectional law of one price (LOOP), even as the slope (risk premium) varies significantly with q. Next, we specify the linearity of expected returns in β, while still leaving the effect of q non-parametric.

4.2 Fama-MacBeth factor premium increases with quantity

We specify a linear relationship between factor exposure (β) and expected return, where the linear coefficient (factor premium) is allowed to vary with quantity: $Er(\beta_{i,k,t}, q_{k,t}) = \beta_{i,k,t}\mu_k(q_{k,t})$.

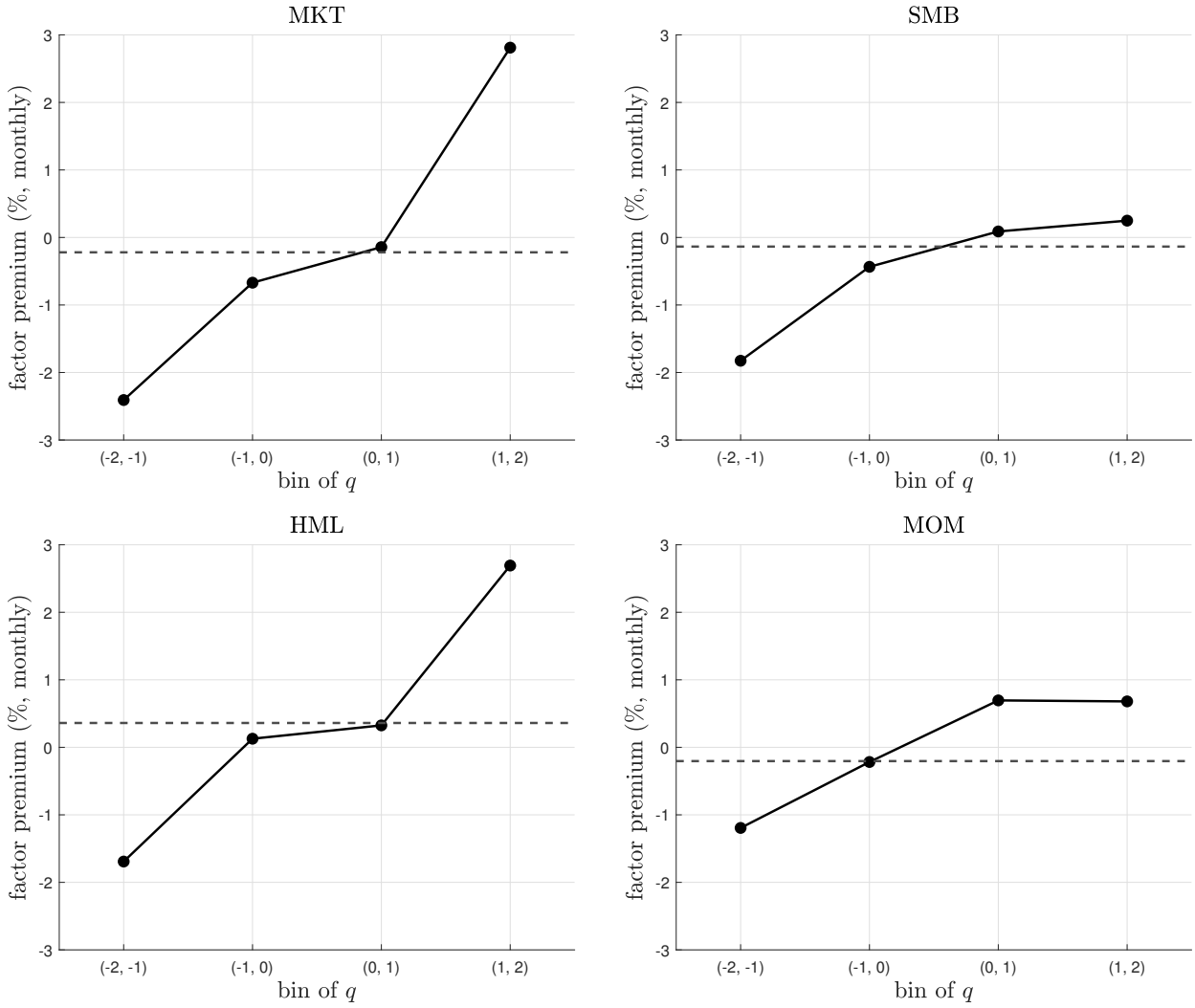
To estimate this model, the first stage of the Fama-MacBeth regressions provides factor risk exposures $\widehat{\beta}_{k,i,t}$ from time-series regression (already detailed in Section 3.1). The second stage of the Fama-MacBeth regressions runs cross-sectional regression for each t:

$$r_{i,t+1} = \gamma_{0,t+1} + \gamma_{k,t+1}\widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, \tag{10}$$

where $\gamma_{k,t+1}$ is the Fama-MacBeth factor mimicking portfolio (FMP) return. Canonically, the factor premium is estimated as the time-series average of $\gamma_{k,t+1}$. It measures the average cross-sectional association between factor loading and stock return. The average factor

[32]It is also interesting to note that the crossings of the high/low-q and unconditional SMLs are almost exactly at $\beta = 0$ for MKT and HML, and somewhat near zero for SMB and MOM. Crossing at $\beta = 0$ is consistent with the parametric BTQ model and the theoretical motivation: the expected return of a factor risk-neutral stock should not be affected by that factor's quantity fluctuations.

Figure 3: Fama-MacBeth factor premium conditioning on quantity, $\mu_k(q_{k,t})$



Note: Fama-MacBeth factor mimicking portfolio returns (FMP, $\gamma_{k,t+1}$) averaged unconditionally (dashed line) and averaged within unit bins of $q_{k,t}$ (solid line).

premiums are often found to be close to zero, challenging factor pricing (Lopez-Lira and Roussanov, 2023).

The innovation of our approach is to estimate the mean of $\gamma_{k,t+1}$ conditional on $q_{k,t}$. To achieve this, we form four unit bins of $q_{k,t}$ (which is already standardized) and calculate the average of $\gamma_{k,t+1}$ within each bin. Figure 3 presents the conditional (solid lines) and the unconditional (dashed lines) factor premiums for each of the four FF3C factors.

The plot shows strong and consistent evidence that the Fama-MacBeth factor premium

is not zero but increasing in factor quantity $q_{k,t}$. Specifically, the cross-sectional risk-return relationship is strong and positive when quantity $q_{k,t}$ is high, while the factor premium becomes negative when $q_{k,t}$ is low, suggesting that the risk-return tradeoff is reversed in low-q environments. On average, the unconditional premium is close to zero, but this masks the significant dynamics that only unfold when we condition on quantity information.

The increasing relationship in $\mu_k(q_{k,t})$ is consistent across the four factors, with the market factor exhibiting the most substantial variation. The market factor premium varies from less than -2% per month when market $q_{k,t}$ is in the lowest $(-2,-1)$ standard deviation range to nearly $+3\%$ per month when market q is in the $(1,2)$ range. Consistent with the SML results, the magnitude of factor premium fluctuation driven by $q_{k,t}$ can reach double-digit annualized percentages, highlighting the economic relevance of quantity in driving factor premiums.

4.3 Beta times quantity (BTQ) forecasts individual stock returns

The empirical results so far from non-parametric plots show that the quantity information significantly impacts the cross-sectional risk-return relationship. We now turn to the parametric BTQ model, which allows us to include multiple factors, provide more formal point estimates, and conduct OOS model fit evaluation and factor selection tests. We show that the BTQ model provides a compelling explanation for the expected return of individual stocks.

Once the factor premium function $\mu_k(q_{k,t})$ is specified in linear form as $\mu_k(q_{k,t}) = \lambda_k q_{k,t}$, we arrive at the parametric BTQ model, which is estimated using the following return predictive regression with a panel of individual stocks:

$$r_{i,t+1} = \sum_{k=1}^{K} \lambda_k q_{k,t} \widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, t. \tag{11}$$

Table 2: Predicting stock returns with and without quantity, single factor

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
Panel A: IS R^2 comparison, full sample 2000-2022 (%)							
BTQ	1.01	0.30	1.00	0.91	0.39	0.62	0.95
β-only	0.05	0.05	0.12	0.06	0.02	0.06	0.10
Panel B: OOS R^2 comparison, evaluation window 2010-2022 (%)							
BTQ	0.75	0.60	0.84	0.65	0.20	0.38	0.67
β-only	0.05	-0.10	0.15	0.02	-0.03	0.04	0.11
Panel C: full-sample coefficient comparison: 2000-2022							
BTQ							
λ_k (%)	1.80	0.72	1.48	1.77	0.62	0.99	1.48
t-stat	(4.18)	(2.76)	(3.52)	(3.38)	(2.24)	(2.96)	(3.69)
β-only							
μ_k (%)	0.38	0.31	0.56	-0.50	-0.33	-0.14	0.22
t-stat	(1.07)	(1.25)	(1.71)	(-1.23)	(-1.52)	(-0.71)	(1.11)

Note: BTQ and β-only return predictions (Eq. 11 and 12), single-factor models ($K = 1$). The first four columns repeat the same prediction exercises with k = MKT, SMB, HML, MOM, respectively. The last three columns report the summary statistics across the 153 JKP factors. The t-statistics (in parentheses) are calculated using standard errors clustered by month. Return prediction R^2 is calculated without demeaning ($R^2 := 1 - \sum_{i,t} (r_{i,t+1} - \widehat{r}_{i,t+1})^2 / \sum_{i,t} r_{i,t+1}^2$, where $\widehat{r}_{i,t+1}$ is predicted return) throughout the paper following Gu, Kelly, and Xiu (2020).

We compare it with the "β-only" model, which is implied by a constant factor premium μ_k:

$$r_{i,t+1} = \sum_{k=1}^{K} \mu_k \widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, t. \tag{12}$$

We first present the results of the single-factor predictive regressions ($K = 1$), using each of the four Fama-French-Carhart factors (MKT, SMB, HML, MOM) and the 153 JKP factors (Table 2).

The key finding is that the BTQ model significantly outperforms the β-only model in

predicting stock returns, with substantial R^2 improvements across different factor choices and in both in-sample and out-of-sample evaluations.[33] Even with only one factor, the BTQ model's OOS return predictive R^2's are around 0.8% for MKT and HML, which are among the highest fit within the 153 JKP factors. The median OOS R^2 across the 153 JKP factors is around 0.4%, and 139 out of the 153 factors yield a positive OOS R^2.[34] This return predictability at the individual stock level is economically significant and comparable to unstructured state-of-the-art machine learning models that use a large number of firm characteristics to predict stock returns, which typically achieve an OOS R^2 of 1% to 2%. In contrast, the β-only models have a low R^2 close to zero, with the median OOS R^2 across the 153 JKP factors at 0.04% and even the 75th percentile reaching only 0.11%.

Turning to the coefficients estimates, the BTQ model's λ_k are significantly positive for all four Fama-French-Carhart factors and for most of the 153 JKP factors. The economic magnitude of the λ_k estimates is substantial. For example, $\lambda_{\text{MKT}} = 1.8\%$, meaning that for one standard deviation increase in market factor q, the expected return of a stock with a market beta of 1 increases by 1.8% per month, or $1.8\% \times 2 = 3.6\%$ per month for a stock with a market beta of 2, and so on.[35] In contrast, the β-only model's μ_k coefficients are mostly statistically insignificant, with 90 out of the 153 JKP factors even exhibiting negative coefficient point estimates.

In summary, the single-factor results show that the BTQ model reliably predicts stock returns, with coefficients consistent with the risk-based explanation, while the β-only model fails in both model fit and coefficient estimates.

[33] For OOS evaluations, we estimate the model parameters (λ_k and μ_k) using the sample period from 2000 to 2009 and apply these estimates to calculate the OOS R^2 for the period from 2010 to 2022. Return prediction $R^2 := 1 - \sum_{i,t} \left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2 / \sum_{i,t} r_{i,t+1}^2$, where $\widehat{r}_{i,t+1}$ is the predicted return. These standards are maintained throughout the paper.

[34] Appendix C.5 provides further interpretation of the economic magnitude of these R^2 values. Roughly speaking, with various simplifications, a one-monthly standard deviation shock in quantity corresponds to 1% of the mutual fund sector's market capitalization or about 0.2% of the total U.S. stock market capitalization. Assuming a price multiplier of 5 (Gabaix and Koijen, 2022), this translates to approximately $1\% = 5 \times 0.2\%$ of expected return fluctuation, which fits about $R^2 = 1\%$ of the monthly variation of realized stock returns.

[35] See Appendix C.5 for additional details showing that the magnitude of the λ_{MKT} estimates is comparable to those reported in the literature.

In addition, Appendix Table A.1 presents an incidental empirical finding regarding factor returns: each factor's return $f_{k,t+1}$ is predictable by its quantity $q_{k,t}$, with the predictive coefficients predominantly positive and statistically significant. However, the OOS R^2's are unstable and mostly negative, due to the limited statistical power of the simple time-series prediction of factor returns. As discussed in Section 2.2, while this time-series predictability is consistent with the BTQ model's cross-sectional return predictability, it is a much weaker argument for the pricing power of quantity and is peripheral to our primary research focus (see further discussion in Appendix C.2).

Moving onto multi-factor models, Table 3 presents the results for these models while maintaining a relatively low dimensionality with $K \leq 6$. This is achieved by using various combinations of the Fama-French-Carhart (FF5C) factors. The BTQ model continues to significantly outperform the β-only model across all multi-factor specifications. Allowing for multiple factors further boosts BTQ's predictive accuracy, with the best OOS R^2 values exceeding 1%. In contrast, the β-only model still struggles to predict stock returns, with low R^2 values even within the sample.

Regarding factor importance, MKT stands out as the most prominent after controlling for the contributions of other factors. It has the highest and most statistically significant coefficients across all multi-factor models, despite an attenuation in λ_{MKT} as more factors are included. HML and MOM also have positive coefficients but lack statistical significance. The inclusion of these factors in the BTQ model increases both IS and OOS R^2, indicating that their BTQ terms provide additional predictive power and that they are priced factors. Conversely, the coefficients for SMB, CMA, and RMW are either near zero or negative, indicating they are not priced factors according to the BTQ model. This is also evidenced in the fact that the OOS R^2 drops when these factors are added to the model. The β-only model's μ_k coefficients are all insignificant or negative. (These numbers are relegated to Appendix Table A.2.)

Comparing BTQ's IS vs. OOS model fits, we observe slight reductions in R^2 when mov-

Table 3: Predicting stock returns with and without quantity: multi-factor models

	CAPM	FF3	FF3C	FF5	FF5C
	$K = 1$	3	4	5	6
Panel A: IS R^2 comparison, full sample 2000-2022 (%)					
BTQ	1.01	1.17	1.19	1.17	1.21
β-only	0.05	0.17	0.21	0.18	0.22
Panel B: OOS R^2 comparison, evaluation window 2010-2022 (%)					
BTQ	0.75	1.03	1.07	0.44	0.65
β-only	0.05	0.15	0.22	-0.26	-0.05
Panel C: coefficients, full sample 2000-2022					
BTQ, λ_k (%) and t-statistics in parentheses					
MKT	1.80	1.27	1.15	1.28	1.16
	(4.18)	(2.08)	(1.96)	(2.00)	(1.98)
SMB		-0.23	-0.16	-0.20	-0.10
		(-0.77)	(-0.59)	(-0.69)	(-0.38)
HML		0.82	0.50	0.80	0.50
		(1.43)	(0.70)	(1.55)	(0.73)
MOM			0.53		0.74
			(0.71)		(0.93)
CMA				0.10	0.08
				(0.35)	(0.28)
RMW				-0.09	-0.25
				(-0.28)	(-0.68)
β-only					
— see Appendix Table A.2 —					

Note: BTQ and β-only return predictions (Eq. 11 and 12). Same as Table 2 but with multi-factor models ($K \geq 1$). The coefficients of the β-only model are relegated to Appendix Table A.2.

ing from IS to OOS for CAPM, FF3, and FF3C models, indicative of mild overfitting or parameter instability. This underscores the robustness of the BTQ model's predictive power, especially considering the inherent difficulty of forecasting monthly stock returns due to the low signal-to-noise ratio in stock prices. For FF5 and FF5C, the IS R^2 continues to increase slightly, while the OOS R^2 reverses to lower values of 0.5% and 0.7%. These levels of prediction accuracy are still economically significant, but the gap between IS and OOS R^2 indicates an overfitting issue. It suggests the ordinary least squares (OLS) estimation

method has limitations for moderately higher-dimensional BTQ models. The additional factors might be noisy or redundant and introduce sample estimation errors. Next, we adopt a regularization method to select factors from a much greater number of candidates.

4.4 Taming the factor zoo with BTQ

The proliferation of proposed factors challenges the asset pricing literature, and the BTQ model offers a new method to select factors. This method has stronger identification power and economic relevance than traditional factor premium tests.

To implement this approach, we use the same return prediction framework (Eq. 11) but overload it with a large number of proposed factors ($K = 159$, including six from FF5C and 153 from JKP). It is well expected that many of these factors are noisy or redundant when controlling for other factors for pricing stock returns. To address this, we use the Lasso method to induce sparsity in the predictive model and filter out the factors that are not priced according to the BTQ model.

Lasso is a regularization method that adds a penalty term to the OLS objective function to shrink and threshold the coefficients towards zero. Specifically, the parameter estimates solve the following optimization problem:

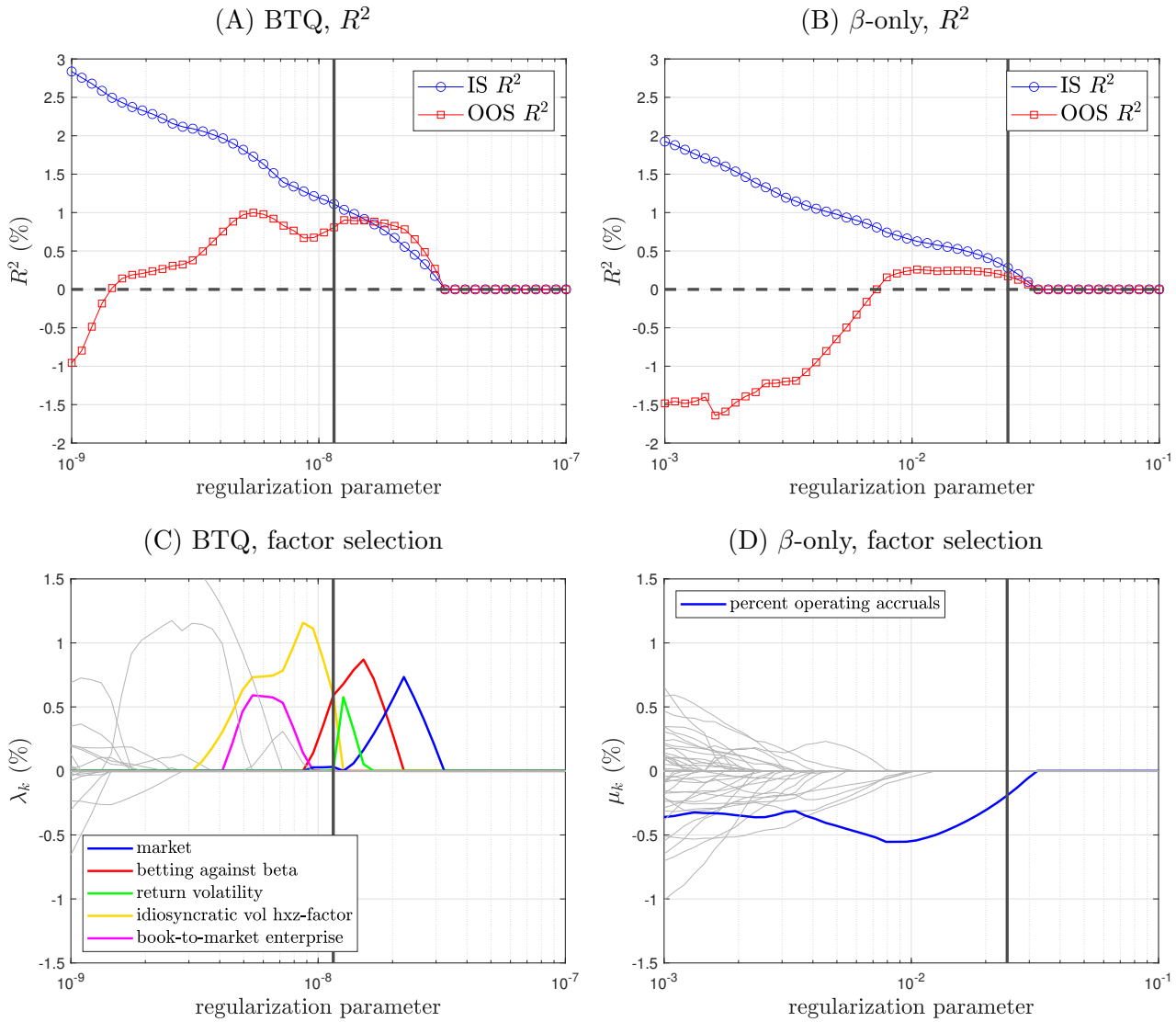
$$\min_{\lambda_1 \dots \lambda_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \lambda_k \widehat{\beta}_{i,k,t} q_{k,t} \right)^2 + \omega \sum_{k=1}^{K} \frac{1}{\sigma(\widetilde{q}_{k,t})} |\lambda_k| , \tag{13}$$

where $|\text{IS}|$ is the number of stock-month observations in the training sample, and ω is the regularization parameter that controls the strength of the penalty term.[36]

Figure 4 plots the model fit and factor selection results for the BTQ and β-only models as the regularization parameter (ω) varies. (The β-only model's Lasso implementation is similar; see technical details in Appendix B.2.) As ω increases, the fitted BTQ model

[36]The penalty on λ_k is normalized by the standard deviation of $\widetilde{q}_{k,t}$ for technical reasons. It allows the economic interpretation of λ_k with respect to the standardized $q_{k,t}$ as used throughout the paper. See technical details in Appendix B.2, where the Lasso essentially is conducted with the pre-standardized $\widetilde{q}_{k,t}$, and these two forms are mathematically equivalent.

Figure 4: Return prediction with factor selection from the factor zoo



Note: Model fit and parameter estimates as the regularization parameter (ω, horizontal axis) varies. In Panels A and B, the IS R^2 is evaluated in the training window (2000-2009), and the OOS R^2 is the same model evaluated in the testing window (2010-2022). Panels C and D plot the parameter estimates from the training window, which are also brought out of the sample for evaluating the OOS R^2 in Panels A and B. The selected factors (colored curves) are, for BTQ: market (`mkt`), betting against beta (`betabab_126d`), return volatility (`rvol_21d`), idiosyncratic volatility from HXZ q-factor model (`ivol_hxz4_21d`), and book-to-market enterprise value (`bev_mev`); and for β-only, percent operating accruals (`oaccruals_ni`). The unselected factors are in gray, reported in Appendix C.4 with factor importance measures. The vertical black line indicates the tuned ω based on cross-validation; see Appendix B.2 for details.

becomes more parsimonious, as shown by the decreasing IS R^2 (Panel A, blue curve) and the decreasing number of selected factors (those with non-zero λ_k in Panel C). This behavior is expected from Lasso. More importantly, the OOS R^2 (Panel A, red curve) displays a hump

shape, with a broad and relatively stable peak that reaches around 1.0%. This suggests that the BTQ model's predictive power is strong and robust to the choice of ω. In contrast, the β-only model's OOS R^2 never exceeds 0.3% and is only positive in a smaller range of ω values. This comparison once again highlights that quantity is essential for a risk-based explanation of expected stock returns.

The most important application of the BTQ + Lasso setup is a new way to investigate which factors are important for pricing stock returns. We find that only a few factors out of the factor zoo are sufficient for the models' high predictive power. The selected factors (those with non-zero λ_k when OOS R^2 peaks) are colored in Panel C. We find that MKT is the first and most important factor, consistent with the observations in previous sections (4.1 to 4.3). The MKT factor is central to multi-factor pricing theories such as Merton's (1973) ICAPM model, and has historically been the most important factor in workhorse empirical models such as the CAPM and Fama-French models. Nevertheless, some research casts doubt on whether market beta is indeed related to expected returns (Black, 1972; Black, Jensen, and Scholes, 1972; Frazzini and Pedersen, 2014). Our results show that the market factor *equipped with quantity* variation remains highly effective in explaining expected stock returns. However, this conclusion cannot be reached with β-only models.

The other selected factors include three based on technical information, betting against beta, return volatility, and idiosyncratic volatility from Hou, Xue, and Zhang's (2015) q-factor model, and one based on fundamental information, book-to-market enterprise value (which is a variant of the HML factor). These are among the usual suspects in the literature, while our results reinforce their importance when considering quantity. Moreover, it is worth noting that the λ estimates of these selected factors from the BTQ model are all positive, which is consistent with the risk-based explanation discussed in Section 2.2. On the other hand, SMB and other size-related factors are excluded by the Lasso selection process. The unselected factors are in gray and can be found in Appendix C.4 with factor importance measures.

The β-only model only selects one factor, percent operating accruals (Panel D), with a negative coefficient. This result is inconsistent with the risk-based explanation and likely reflects the model's misspecification, as indicated by its low model fit.

Additionally, choosing ω based on the OOS R^2 peak is sufficient for the purpose of interpreting the BTQ model's factor selection. However, for the purpose of forecasting stock returns, it has a look-ahead bias. To address the problem, we provide the tuned ω using only IS information via ten-fold cross-validation, as shown by the vertical black lines in Figure 4 (see technical details in Appendix B.2). The IS tuned ω is close to the OOS R^2 peak, suggesting the robustness of prediction and selection results.

4.5 BTQ with latent factors

Latent factors estimated using statistical methods to fit the realized time-series variation of returns have shown superior explanatory power for expected returns.[37] We demonstrate that the BTQ framework can be applied to latent factors as well, and it leads to a strong two-factor structure with high predictive power for stock returns that is unattainable with the β-only counterpart.

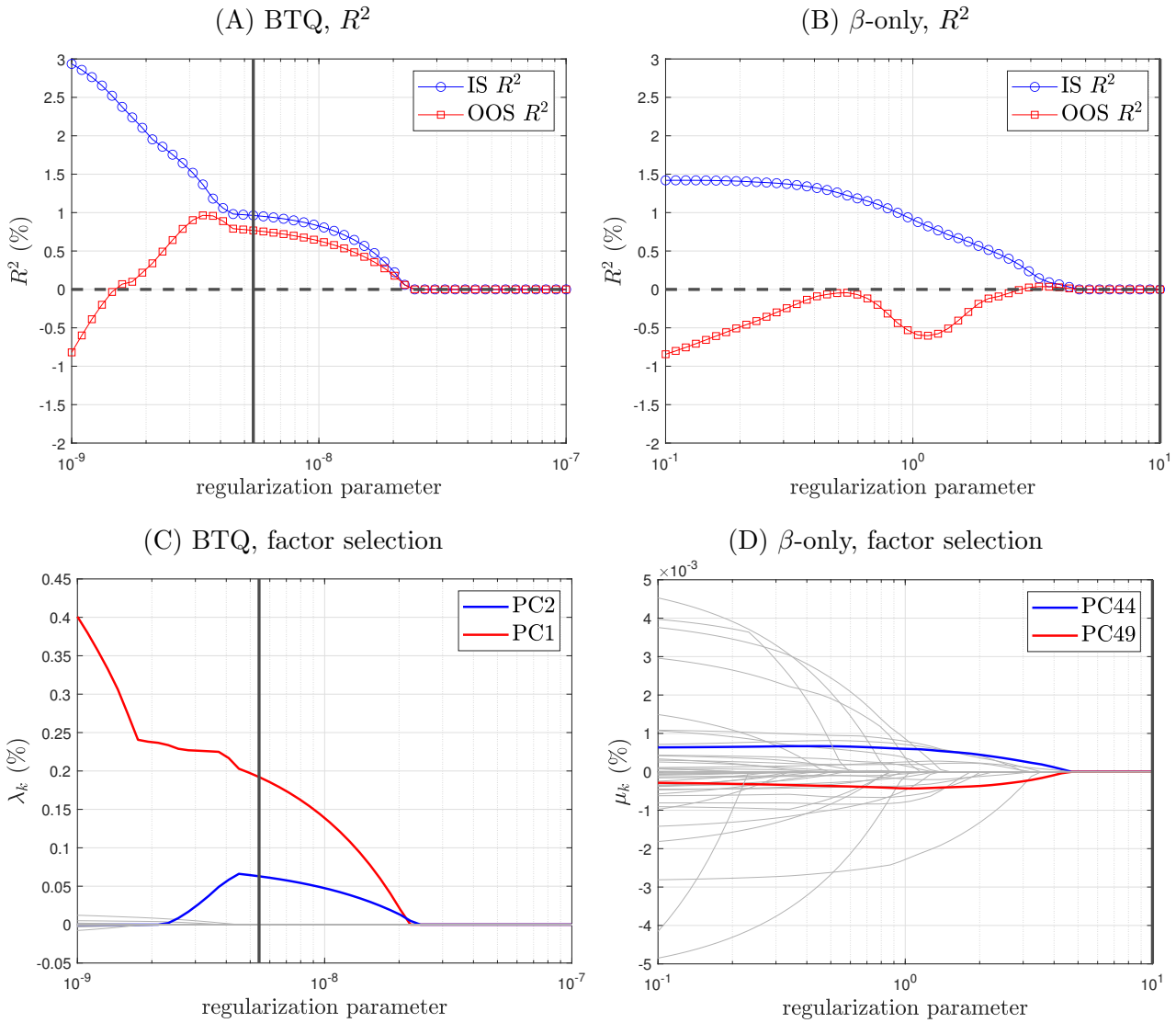
We extract the principal components (PC) of the factor zoo portfolio returns, which are the linear combinations of the factor returns that capture the most time-series variation.[38] Then, we construct $\widehat{\beta}$ and, in turn, the quantity q for each of these PC factors from scratch, following the same procedure reported in Section 3. Based on these variables, we conduct the same BTQ predictive regression with Lasso as in the previous section. The new set of $\widehat{\beta}$ and q variables provides some external validation of our method's robustness and generalizability.

Figure 5 shows that the BTQ model with PC factors has strong predictive power for stock returns, with the OOS R^2 peaking at around 1.0%, similar to the previous Figure 4 using original factors. The high OOS R^2 is, once again, robust to the choice of ω, as evidenced by

[37] See, e.g., Kelly, Pruitt, and Su (2019, 2020), Kozak, Nagel, and Santosh (2020), Lettau and Pelger (2020), Chen, Roussanov, and Wang (2023), and Chen, Roussanov, Wang, and Zou (2024).

[38] Specifically, we use the first 50 principal components estimated from the monthly returns of the FF5C and 153 JKP factors from 1970 to 2009.

Figure 5: Return prediction with PC and factor selection



Note: Model fit and parameter estimates as the regularization parameter (ω, horizontal axis) varies. In Panels A and B, the IS R^2 is evaluated in the training window (2000-2009), and the OOS R^2 is the same model evaluated in the testing window (2010-2022). Panels C and D plot the parameter estimates from the training window, which are also brought out of the sample for evaluating the OOS R^2 in Panels A and B. We perform Lasso regression using the first 50 principal components derived from the monthly returns of the FF5C and JKP factors from 1970 to 2009. The unselected factors are in gray. The vertical black line indicates the tuned ω based on ten-fold cross-validation; see Appendix B.2 for tuning details.

the broad peak of the OOS R^2 hump-shaped curve. In contrast, the β-only model with PC factors hardly delivers any predictive power, with OOS R^2 remaining below zero for almost all ω values.

More importantly, Panel C reveals a strong two-factor structure, with PC1 and PC2

emerging as the most important factors for predicting stock returns. The magnitude of their λ estimates dominates those of subsequent PC factors (depicted as gray curves). This parsimonious structure attained with the BTQ model with latent factors can explain expected stock returns well with high OOS R^2. This is consistent with the literature that suggests latent factors are helpful in "shrinking the cross section" and reducing the dimensionality of the factor zoo (Kozak, Nagel, and Santosh, 2020).

Notably, the signs of the λ estimates for the selected factors, PC1 and PC2, are both positive. This is required by the risk-based theory, regardless of how the signs of the PCs are specified, and further reinforces the validity of the BTQ model. In contrast, the β-only model's selection and parameter estimates exhibit no discernible pattern, which likely stems from estimation noise, as the β-only model is misspecified.

4.6 Robustness of the main predictive results above

This subsection reports robustness checks that validate the BTQ model's predictive power reported above. We have already shown the BTQ model is robust to different factor specifications, including single-factor, multi-factor, selected factors, and latent factors extracted from the factor zoo. We further change the specifications in different dimensions, including various sub-sample evaluations and alternative constructions of the quantity variable.

Subsamples. We first evaluate the forecasts of the BTQ models reported above in different size and time sub-samples. Table 4 Panel A breaks down the OOS panel into five size groups according to NYSE market capitalization quintiles and reports the OOS R^2 in each size group. Panel B similarly breaks down the OOS evaluation into three sub-periods: 2010-2014, 2015-2018, and 2019-2022. Panel C reports the original joint OOS (2010-2022) evaluation for reference. This table evaluates the BTQ models with factors selected from the factor zoo (initially reported in Section 4.4) and with selected PC factors (in Section 4.5).[39]

[39]Table 4 evaluates the OOS forecasts ($\widehat{r}_{i,t+1}$) produced with the in-sample cross-validated hyperparameter ω. That is, Panel C reports the same OOS R^2 values at the vertical black line in Figures 4 and 5 Panel A.

Table 4: BTQ OOS prediction accuracy (R^2 in %) in size and time sub-samples

evaluation sample	# of obs.	selection	PC+selection
Panel A: size group evaluation			
1 (small)	323,617	0.46	0.44
2	165,059	1.12	1.07
3	141,153	1.48	1.40
4	115,763	2.02	1.91
5 (big)	103,927	2.16	2.09
Panel B: sub-period evaluation			
2010-2014	321,425	1.16	1.18
2015-2018	255,959	0.15	0.14
2019-2022	272,135	1.00	0.92
Panel C: original benchmark OOS evaluation			
OOS (2010-2022)	849,519	0.81	0.77

Note: OOS R^2 evaluated in different size and time sub-samples for the BTQ models with factors selected from the factor zoo (in Section 4.4) and with selected PC factors (in Section 4.5). Panel A breaks down the OOS panel into five size groups according to NYSE market capitalization quintiles and reports the OOS R^2 in each size group. Panel B breaks down the OOS evaluation into three sub-periods: 2010-2014, 2015-2018, and 2019-2022. Panel C reports the original joint OOS (2010-2022) evaluation for reference.

Appendix C.6 contains the same sub-sample robustness evaluations for the Fama-French-Carhart factors (in Section 4.3), and the results are mostly the same.

Table 4 shows the BTQ model's predictive results reported above are consistent in most size and time sub-samples. In particular, Panel A shows the accuracy is higher in large stocks, which are usually the most challenging section for stock return prediction. Characteristics-based anomalies and machine learning models typically find stronger predictive power in the small groups due to stronger limits to arbitrage in small stocks, including illiquidity and information asymmetry. This result indicates that BTQ's predictive power can be more reliably implemented in investment strategies in practice, given liquidity costs and trading constraints are typically weaker for larger stocks.[40]

[40]Cf. Jensen, Kelly, Malamud, and Pedersen (2024); Goyenko, Kelly, Moskowitz, Su, and Zhang (2024).

Table 5: BTQ OOS R^2 (%) robustness to lookback window length in $q_{k,t}$ construction

lookback (h)	selection	PC+selection	lookback (h)	selection	PC+selection
1	0.36	0	7	0.78	0.88
2	0.41	0.42	8	0.62	0.70
3	0.65	-0.15	9	0.62	0.10
4	0.49	0.38	10	0.33	0.11
5	0.85	0.82	11	0.48	0.19
6 (benchmark)	0.81	0.77	12	0.48	0.25

Note: OOS R^2 evaluated for BTQ models with $q_{k,t}$ constructed with alternative lookback window lengths (h) using factors selected from the factor zoo (in Section 4.4) and selected PC factors (in Section 4.5).

Regarding sub-periods, the BTQ model's predictive power is mostly stable over time. The first and the last sub-periods (2010-2014 and 2019-2022) have higher R^2 values than the middle sub-period (2015-2018) in both model specifications. We attribute this to the fact that quantity fluctuations in the middle sub-period are less volatile, as shown in Figure 1.[41]

Alternative constructions of the quantity variable. Next, we evaluate the robustness of the BTQ model to alternative specifications in constructing the quantity time series $q_{k,t}$. In particular, there is no explicit theoretical guidance on whether factor-level flows have immediate or lagged effects on factor premium, or how fast past flows' effects decay. The benchmark specification of the quantity variables (in Section 3) accumulates past flows in a six-month lookback window, which aligns with the common expectation. We now change the specification by constructing the quantity variables using lookback windows ranging from 1 to 12 months. That is, in Eq. 9, $h = 6$ is replaced by $h = 1$ to $h = 12$. The same empirical analyses from the previous sections are re-run with these alternative quantity variables.

Table 5 shows the BTQ model's predictive accuracy is robust to alternative lookback window lengths in constructing the quantity variables. Certain perturbations (such as $h = 5$

[41]Notice for each model specification, the predictive model is trained once with the 2000-2009 training sample (IS). Repeated size group-specific training (a.k.a. expert models) and rolling-window training have the potential to further improve the R^2 in OOS sub-samples above. We leave these extensions for future research due to their focus on forecasting engineering.

or 7) can even improve the R^2, meaning the benchmark results are not sensitive to the exact specification of the quantity variable. Having an h that is too short or too long will hurt the predictive performance, but the OOS R^2 values are mostly significant and positive, especially for the method that directly selects factors from the factor zoo.

Additionally, Table 5 offers suggestive evidence regarding the speed and persistence with which factor flows influence sophisticated investors' pricing of factor risks. Flow shocks likely have an immediate impact on the factor premium next month, given that $h = 1$ already has some predictive power. The R^2 is higher with an intermediate window ($h =$ 5, 6, or 7), suggesting the lagged flows in the recent few months also have impacts on factor premium, and that accumulating flows in a lookback window has, at least, statistical benefits in smoothing the predictors. On the other hand, longer windows near one year suppress prediction accuracy, suggesting that flows older than seven months have attenuated impacts on factor premiums. The attenuation is likely related to mechanisms through which sophisticated investors can gradually unwind their absorbed positions and adjust their risk holdings over time. A more detailed investigation of the dynamics between factor flows and factor premiums is left for future research, which likely requires models and data more focused on investor holdings.

5 Quantity must be combined with risk to forecast returns

This paper emphasizes a *risk-based* explanation of expected stock returns that incorporates quantity information. But is the risk modeling essential? Can quantity information alone explain expected stock returns? This section makes the case that risk and quantity must operate in tandem to forecast returns effectively. First, stock-level quantity information must be aggregated at the factor level to predict stock returns. Second, almost all factor-level q variables in the "factor zoo" exhibit predictive power; however, each q is effective only in explaining the cross-sectional return dispersion along its corresponding factor's own risk dimension—any mismatch between β and q significantly diminishes prediction accuracy.

These results show that previously reported empirical success is unique to the factor risk structure, supporting our economic interpretation.

5.1 "Quantity-only" models do not explain expected returns

We examine an alternative economic model in which stock-level flow and quantity variations directly affect the expected stock returns without considering the factor structure and the arbitrage pricing condition. This exercise is important for understanding the joint economic role of quantity and risk in asset pricing. The main results presented earlier compare the benchmark BTQ model against the "β-only" baseline that accounts for risk without quantity; here, we demonstrate that an alternative "quantity-only" baseline—relying solely on quantity while disregarding risk—also falls far short in explaining expected stock returns.

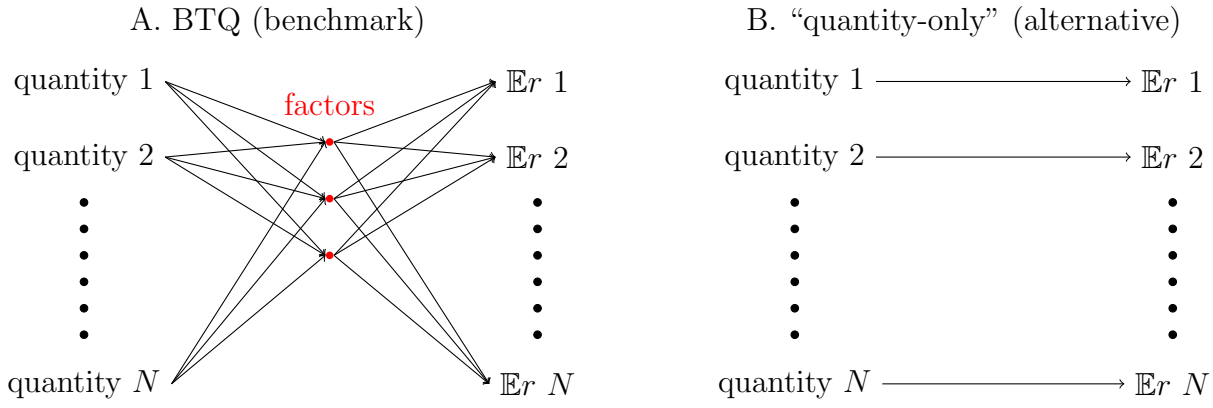
In the benchmark model (BTQ), stock-level quantity variations are first aggregated to the factor-level quantities, which affect factor premiums, and then feed back to stock-level expected returns. In contrast, the "quantity-only" model specifies that stock-level flow and quantity variations *directly* affect expected stock returns, short-circuiting the factor premium adjustment mechanism (see the contrast in Figure 6). Specifically, the alternative model is:

$$\mathbb{E}_t r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock}}, \qquad \forall i, t, \tag{14}$$

where $q_{i,t}^{\text{stock}}$ is a stock-specific flow or quantity measure, and λ_i^{stock} is the sensitivity coefficient of stock returns to $q_{i,t}^{\text{stock}}$. (λ_i^{stock} may or may not vary across stocks, to be specified below.)

This "quantity-only" model implies a fundamentally different economic mechanism, although Eq. 14 is similar in form to the main model's factor premium specification in Eq. 3. In the main model, factor premiums vary dynamically, yet the cross-sectional no-(statistical) arbitrage pricing condition (with respect to the factors) holds each period. In contrast, the alternative model dispenses with the APT condition. For instance, two stocks with identical risk exposures but differing noise flow shocks are priced differently by the alternative model,

Figure 6: Comparison of predictive architectures of the two models



Note: A. BTQ model: stock-level quantity variations affect expected stock returns *via* quantities of factor risks and factor premiums. B. "quantity-only" alternative model: stock-level quantity *directly* affects expected stock returns, short-circuiting the factor premium mechanism.

creating an immediate arbitrage opportunity. The "quantity-only" model implies a lack of cross-sectional substitution, such that each stock is priced independently of its factor exposures. This might be the case if rigid frictions prevent cross-sectional arbitrage; or if stocks' idiosyncratic risks are not diversifiable and individually priced.

We experiment with various specifications of Eq. 14 and find that none come close to the BTQ model's explanatory power for expected stock returns. Specifically, stock-level $q_{i,t}^{\text{stock},h}$ is constructed similarly to that of the factor-level in Eq. 9:

$$q_{i,t}^{\text{stock},h} := \frac{1}{h}\sum_{h'=0}^{h-1} \frac{\$\text{flow}_{i,t-h'}^{\text{stock}}}{\text{market\_cap}_{i,t-1-h'}}, \qquad \forall i,t, \text{ and } h = 1,\ldots,12.^{42} \tag{15}$$

We explore different specifications of the sensitivity coefficient λ^{stock}, with varying degrees of parameter freedom. Table 6 Panel A specifies λ^{stock} as a constant for all stocks: $r_{i,t+1} = \lambda^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. Panel B allows a size-dependent sensitivity coefficient such that λ^{stock} is indexed by the NYSE size quintile of the stock: $r_{i,t+1} = \lambda^{\text{stock}}_{\text{size\_quintile}(i,t)} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. Panel C allows stock-specific λ_i^{stock}, which is the most flexible specification:

[42] We normalize the dollar stock-level mutual fund flow ($\$\text{flow}_{i,t}^{\text{stock}}$, see Eq. 7) by the stock's one-month-lagged market capitalization, so that the sensitivity coefficients are more interpretable. We accumulate past flows over various lookback windows, since we are agnostic about whether flow shocks have immediate or lagged effects on expected returns.

Table 6: "Quantity-only" alternative model does not forecast stock returns

	A. constant λ^{stock}				B. λ^{stock} by size quintile		C. λ_i^{stock} by stock	
h	IS R^2(%)	OOS R^2(%)	λ^{stock}	t-stat	IS R^2(%)	OOS R^2(%)	IS R^2(%)	OOS R^2(%)
1	0.000	0.000	-0.10	-0.27	0.003	-0.001	0.47	-232
2	0.000	-0.001	0.05	0.12	0.002	-0.003	0.44	-215
3	0.000	0.000	0.17	0.29	0.003	0.000	0.39	-155
6	0.005	0.006	0.75	1.01	0.007	0.006	0.41	-107
9	0.004	0.004	0.75	0.99	0.006	0.006	0.38	-96
12	0.003	-0.007	0.77	1.01	0.006	-0.004	0.38	-81

Note: Panel A: univariate predictive regression, $r_{i,t+1} = \lambda^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. B: size-dependent predictive regression, $r_{i,t+1} = \lambda^{\text{stock}}_{\text{size\_quintile}(i,t)} q_{i,t}^{\text{stock},h} + error_{i,t+1}$, where $\lambda^{\text{stock}}_{\text{size\_quintile}(i,t)}$ is indexed by the NYSE size quintile of the stock. C: stock-specific predictive regression, $r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. The R^2 values are in percentages, e.g., 0.005 in row $h = 6$ means 0.005%, a very small value. The table skips some rows to save space, see complete results ($h = 1 \sim 12$) in Appendix Table A.4.

$r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$.[43]

The "quantity-only" models are too weak and unreliable to predict stock returns in any alternative specification, as shown in Table 6. In Panel A, the constant λ^{stock} specification's in-sample R^2 values are about 100 times smaller than the BTQ model's. The out-of-sample R^2 values are not only small in magnitude, but also negative for some lookback lengths (h). The λ^{stock} estimates are mostly positive, consistent with the existing literature—outflows from noise traders have negative concurrent price impacts and positively predict future returns.[44] However, the estimates are statistically insignificant, and too weak to offer a meaningful R^2 in predicting stock returns. Allowing size-dependent λ^{stock} slightly improves these predictive power evaluations, but no qualitative changes (Panel B). The low R^2 values

[43]The specification in Panel B (size-dependent λ^{stock}) effectively runs five separate univariate predictive regressions, one for each size bin ("mixture of experts" in machine learning terms). The specification in Panel C (stock-specific λ_i^{stock}) effectively runs stock-by-stock time-series predictive regressions. To address the unbalanced panel, we restrict the analysis to stocks with more than 80% of monthly observations available in both the in-sample and out-of-sample windows. Stocks with fewer observations would be even more challenging to forecast.

[44]See Appendix C.5 for additional details showing that the magnitude of these λ^{stock} estimates is comparable to those reported in the literature.

suggest that these two specifications are too restrictive, and that the model is underfitting the data. In Panel C, stock-specific λ_i^{stock} allows a much greater degree of freedom in parameterization (thousands of stocks vs. one or five parameters). The in-sample R^2 mechanically increases but is still smaller than the BTQ model. More importantly, the out-of-sample R^2 values are extremely negative, suggesting the in-sample R^2 values are greatly exaggerated by overfitting.

The poor performance of the "quantity-only" alternative model underscores that the empirical success of the BTQ model is not driven by quantity alone, highlighting the necessity of combining quantity and risk to explain expected stock returns. In particular, the comparison implies that the factor structure is still essential in modeling expected stock returns, and that cross-sectional quantity-driven mispricing (alpha) is too weak to detect.[45] This is consistent with the view that statistical arbitrage activities by some sophisticated investors are effective in enforcing the cross-sectional APT condition, even in the presence of noise traders (Kozak, Nagel, and Santosh, 2018). The differing performance of BTQ and "quantity-only" models mirrors the contrast of macro vs. micro elasticities in Gabaix and Koijen (2022). At the stock level, securities are highly substitutable, whereas quantity's effect on prices is more salient at the factor level due to greater inelasticity of factor demand (e.g., Peng and Wang, 2021; Li and Lin, 2022).

From a statistical and machine learning perspective, we can view both the BTQ and the "quantity-only" alternative as prediction models of stock returns based on the stock-level quantity information, i.e., they use the same predictors to predict the same targets. The difference is that the BTQ model employs a dimension reduction, aggregating predictors to the factor level, which, in turn, are used to predict the whole cross section. This is an encoder-decoder architecture in machine learning terms, where the low-dimensional "code" is the factor-level q (see Figure 6 Panel A for the encoder-decoder illustration). In this

[45] The "quantity-only" model (Eq. 14) can be viewed as a quantity-driven alpha, especially when viewed in conjunction with the BTQ predictors: $\mathbb{E}_t r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock}} + \sum_k \lambda_k \beta_{i,k,t} q_{k,t}$. The term $\lambda_i^{\text{stock}} q_{i,t}^{\text{stock}}$ captures dynamic *alpha*, the part of the quantity-driven expected stock return that is not explained by the risk channel (BTQ, or the second term).

perspective, BTQ performs well in forecasting because encoding reduces the noise in the predictors and captures the economically meaningful quantity variation aggregated at the factor level. In contrast, the "quantity-only" model is limited by the noisy inputs at the stock level.[46]

5.2 Different factors' q's explain risk-return tradeoffs along different dimensions

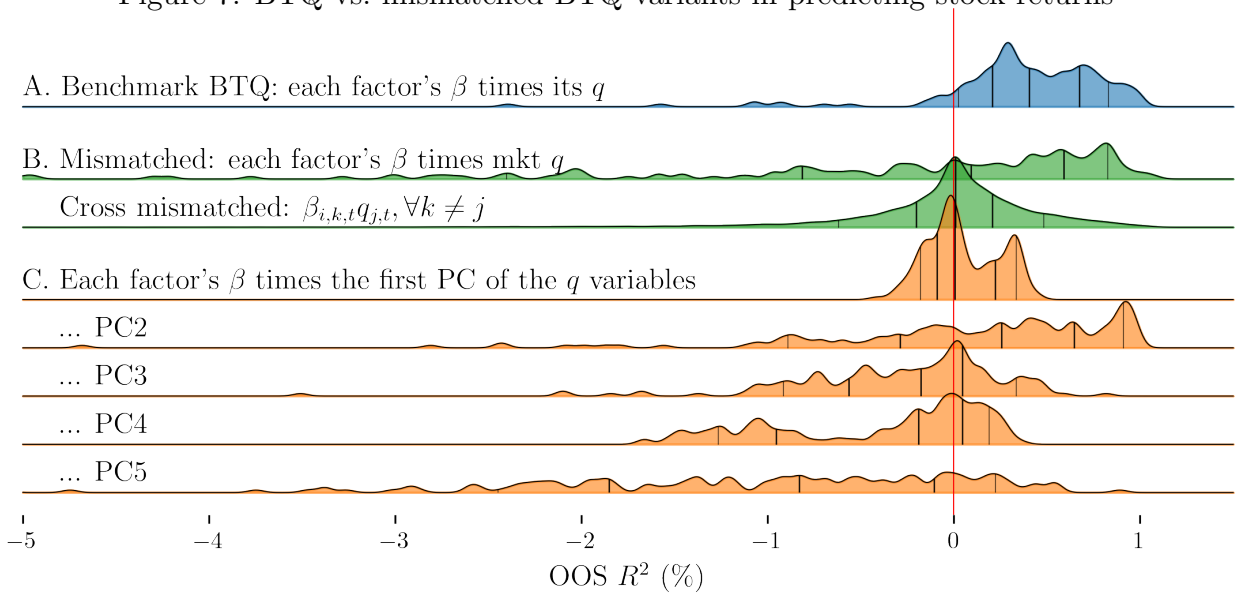
The analysis above highlights the need to aggregate the quantity information at the factor level to effectively predict stock returns, consistent with our risk-based explanation. In addition, the theory also imposes a corresponding structure between a factor's exposure and the factor's quantity—BTQ must be built with each factor's own β times its own q. We show that dispensing with this corresponding structure does not unleash more statistical power but, to the contrary, reduces it significantly, further validating the risk-return tradeoff channel of quantity's role in expected stock returns.

We know that stock returns exhibit a multi-factor risk structure. A remarkable aspect of BTQ is that its pricing power holds independently across a diverse array of factors in the "factor zoo" (see Table 2 and repeated in Figure 7 Panel A). We now show that each factor's q provides distinct pricing information along its respective risk dimension. Mismatching one factor's β to another factor's q significantly reduces the prediction accuracy, meaning that a factor's q pertains to explaining the cross-sectional return dispersion *only* along its own risk direction. Each dimension of risk provides independent evidence on the q-μ association, supporting the robustness of the economic interpretation. This result refutes the idea that the main results are driven by one (or a few) special "secret sauce" q series.

In detail, Figure 7 Panel A plots the benchmark distributions of the 159 OOS R^2 values for the single-factor BTQ models, with each factor's β times its own q (repeating the result in Table 2). More than 90% of these factors yield positive OOS R^2 values, demonstrating

[46]See Gu, Kelly, and Xiu (2021) and Kelly, Malamud, and Pedersen (2023) for applications of encoder-decoder structures in asset pricing. Notice that BTQ specifies both the encoding and decoding weights as β according to economic theory, rather than solely relying on statistical estimation.

Figure 7: BTQ vs. mismatched BTQ variants in predicting stock returns



Note: Each distribution represents the outcomes (OOS R^2 values) of (A) 159 single-factor BTQ models, or (B, C) 159 mismatched BTQ variants, except Line "Cross mismatched" has $159 \times 159 - 159 = 25{,}122$ OOS R^2 values. Kernel density estimates' (KDE) areas under the curve are standardized across all the 8 distributions. Vertical lines, thin: 10, 90 quantiles; thick: 25, 50, 75 quantiles.

the effectiveness of the BTQ model when correctly specified.

The subsequent panels are BTQ variants where β and q are mismatched in various ways. The first variant in Panel B pairs each factor's $\beta_{i,k,t}$ with the market factor's quantity $q_{\text{mkt},t}$. The second distribution is for the "cross mismatched" BTQ variants: each factor's β is paired with the q of every other factor ($\beta_{i,k,t} \times q_{j,t}, \forall j \neq k$), resulting in $159 \times 159 - 159 = 25{,}122$ mismatched models. Panel C examines any underlying common signal among the 159 q series, by pairing each factor's β with a principal component of the q series.[47]

The mismatched variants perform significantly worse than the benchmark BTQ model. When correctly specified, more than 90% of the factors in the "zoo" yield BTQ models with positive OOS R^2 values. However, the mismatched variants have approximately half of their density below zero, often with long left tails of highly negative R^2 values. In particular, the market factor's q does not generalize to other factors. This means that the market factor's q is not a special series that gives rise to other factors' BTQ results, despite its dominant

[47]Notice, we conduct PCA of the 159 q series here, rather than calculating the quantities for the PCs of the factor returns as in Section 4.5. Appendix Figure A.3 Panel B reports the properties of the PC q series.

time-series variation (Figure 1) and leading prominence in factor selection (Figure 4). Other factors' q's are essential to explain the risk-return tradeoff along their respective risk dimensions. The same conclusion holds for each principal component q variant.[48] These results highlight the independence of each factor's risk-return tradeoff and its unique association with its own q, providing robust evidence against the idea that our main results are driven by some particular signals in the q series.

6 Quantity or alternative channels driving factor premiums?

In the economic framework of the BTQ model, we argue that quantity is associated with factor premiums, representing the degree of cross-sectional risk-return trade-offs. Can alternative economic channels explain the observed empirical success of the BTQ model? What if some other underlying state variables drive the reported variation in factor premiums and, with quantity variables merely reflecting those state variables? If that is the case, the BTQ's reported empirical performance is merely a facade, and quantity would not be the direct driver of factor premiums.

To investigate this possibility, we examine alternative variables, including factor momentum signals and a large set of macroeconomic variables, both of which are well-documented in previous studies as predictors of factor returns. We find that their empirical performance is far behind the BTQ model across a wide range of empirical specifications. These findings suggest factor quantity variation is directly associated with factor premiums and further validate the BTQ model's economic channel.

The first exercise considers factor momentum and retail flow's "performance chasing" behavior. Factor momentum implies past factor performances are positively associated with future factor returns. Meanwhile, retail mutual fund flows also positively respond to past performances.[49] Can these two forces combined explain BTQ's reported empirical perfor-

[48] In terms of the right tails, the market q and the second principal component q retain some factors' predictive power from the benchmark BTQ model, likely due to the two's commonality with those factors.

[49] Factor momentum is well-documented in the literature (e.g., Moskowitz, Ooi, and Pedersen, 2012; Gupta

Table 7: BTQ vs. "beta times momentum"

CAPM	FF3	FF3C	FF5	FF5C	selction	PC+selection
BTQ, benchmark results reported in Section 4						
0.75	1.03	1.07	0.44	0.65	0.81	0.77
"beta times momentum", best of 1- to 12-month momentum signal formation periods						
0.01	-0.05	-0.04	0.12	0.09	0	0

Note: All numbers are OOS predictive R^2 in %. To save space, we only report the best R^2 values across the lookback window lengths. The complete results are available upon request, where most of the R^2 values are negative and with no discernible pattern across varying formation length $h = 1 \sim 12$. $R^2 = 0$ is because Lasso selects no predictors, and the model forecasts zero return for all stock-months.

mance?

The answer is no. First, this channel would imply a negative q-μ association, which is opposite to the BTQ model's prediction and our empirical findings. Performance-chasing behavior of retail investors would reduce sophisticated investors' holdings (q) when past factor returns are high. Under factor momentum, this reduction would be associated with high future factor returns—contradicting the observed positive association. Regardless of the sign restriction, the data show that replacing $q_{k,t}$ in the BTQ model with a variety of factor-level momentum/reversal signals fails to yield reliable predictions of stock returns. Even the best "beta times momentum" predictor specifications ($\beta_{i,k,t} \times \text{mom}_{k,t}$, where $\text{mom}_{k,t}$ is the past return of factor k in 1- to 12-month windows) have only slightly positive OOS R^2 values, with most cases yielding negative values (see Table 7). Our interpretation is that while factor momentum and performance-chasing are respectively valid phenomena supported in the literature, they are insufficient to reproduce the explanatory power of BTQ, and that our q variable captures quantity variation beyond performance-chasing alone.

The second exercise examines whether macroeconomic variables or aggregate financial metrics, used individually or combined linearly, can replace the q variables and provide pre-

and Kelly, 2018; Ehsani and Linnainmaa, 2022; Arnott, Kalesnik, and Linnainmaa, 2023). Relatedly, Huebner (2023) demonstrates the generation of price momentum in equilibrium driven by investor demands. Mutual fund performance chasing has also been extensively studied (e.g., Lou, 2012; Ben-David, Li, Rossi, and Song, 2024).

Table 8: BTQ vs. "beta times macro variables"

A. BTQ (benchmark: market beta times market q)

0.75

B. Market beta times each macro variable (126 FRED-MD variables one at a time)

best of the 126	95th percentile	75th percentile	50th percentile	25th percentile
0.28	0.18	0.00	-0.12	-0.63

C. Market beta times the principal component of the macro variables

PC1	PC2	PC3	PC4	PC5
-0.17	0.01	-1.05	-0.09	-0.46

D. Market beta times the best linear combination of the macro variables

OLS (best in-sample fit, no regularization)	LASSO (best among reg. param.)
< -100	0

Note: Predicting stock returns with conditional CAPM predictors in the form of "market beta times ...". All numbers are OOS R^2 in percentage. Multi-factor results in Appendix Table A.5 are similar. Details of Panel D results across regularization parameters are in Appendix Figure A.5.

dictive power comparable to the BTQ model. We explore the FRED-MD dataset (McCracken and Ng, 2016), which contains a comprehensive set of 126 monthly macroeconomic variables, including dividend yield, default spread, and personal income growth. These are frequently used as conditioning variables in conditional factor models (Jagannathan and Wang, 1996; Lettau and Ludvigson, 2001; Petkova and Zhang, 2005; Daniel and Titman, 2011).[50]

The results show that neither these variables nor their linear combinations come close to q's predictive power. When tested individually in the form of "market beta times a macro variable," even the best-performing macroeconomic variable among the 126 series achieves an OOS predictive R^2 that is only a fraction of the "market beta times market q" specification. Most macroeconomic variables yield negative OOS R^2 (Table 8 Panels A vs. B).[51] For linear combinations of the macroeconomic variables, neither unsupervised principal components (Panel C) nor supervised combinations fitted on stock returns (Panel D, fitting details in

[50]We pre-process the Fred-MD series following standard procedures, detailed in Appendix C.7.

[51]The best macro variable, with $R^2 = 0.28\%$, is unemployment insurance initial claims. The second and third are housing starts in the northeast and personal consumption expenditures on services, respectively. These variables offer no coherent economic explanation, as they were chosen ex post.

Appendix Figure A.5) yield any predictive power.[52] Multi-factor specifications report similar results in Appendix Table A.5.

7 Conclusion

This paper considers a new but important aspect of risk's economic role in determining asset prices—the *quantity* variation in investors' risk holdings induced by trading flows. The economic rationale is simple: when sophisticated investors hold more of a systematic risk factor, they require greater compensation for bearing that risk, which in turn drives the expected return of every stock exposed to the factor. Yet the empirical model yields a compelling risk-based explanation for expected stock returns.

We show that incorporating quantity into canonical factor pricing has important implications for asset pricing studies with three new findings. First, quantity variation elicits risk-return tradeoff relationships, which have been hard to capture with β only and thereby cast doubt on whether risk explains expected returns. We find the cross-sectional relationship between factor exposures and expected returns (β-$\mathbb{E}r$ relationship) strongly depends on factor quantity variation, and the previous null result is a mixed average unconditional on quantity. Second, quantity enables a risk-based predictive model (termed beta times quantity, BTQ) for monthly stock returns. The model delivers high prediction accuracy in this hard empirical task dominated by unstructured machine learning models and firm characteristics. Third, incorporating quantity provides a new way for factor selection and, thereby, new answers to the factor zoo problem. Instrumenting factor premiums with quantity variation has not only greater identification power but also more economic relevance than traditional factor premium tests. We find that a few factors out of the factor zoo are selected for the model's high predictive power, and in a latent factor setting, the first two principal components overwhelmingly dominate the remaining components.

[52] Principal components of the macroeconomic series are used for forecasting in, for example, Stock and Watson (2002) and Ludvigson and Ng (2009) for the bond market.

Besides showing the improvements against the β-only baseline, we also implement various versions of the "quantity-only" model, which directly relates stock-level quantity to expected stock returns. We find this alternative baseline also falls short by far in explaining expected stock returns. This result implies that the stock returns' factor structure and the no-arbitrage pricing condition are important for modeling expected returns, even in the presence of significant price impacts from noise flows.

In summary, we show both quantity and risk should work together for modeling expected stock returns. At a high level, this is a natural result given the interaction between sophisticated investors and noise traders. It bridges factor pricing (which emphasizes rational agents' aversion to risk) and the price impact of noise flows (which emphasizes the price dislocation effects of non-fundamental flows and inelastic demand). The contribution of this empirical paper is providing a simple and actionable way to integrate quantity information into canonical factor models and showing its significant improvement to the factor model's empirical relevance.

We are confident that future research can similarly incorporate quantity information into other existing asset pricing methods to yield new insights for various research questions. Another interesting direction for further research is to explore a richer set of asset holdings information to construct other quantity variables; a concurrent paper (Gabaix, Koijen, Richmond, and Yogo, 2023) is highly relevant for this potential direction.

References

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial intermediaries and the cross-section of asset returns, *Journal of Finance* 69, 2557–2596.

Arnott, Robert D, Vitali Kalesnik, and Juhani T Linnainmaa, 2023, Factor momentum, *Review of Financial Studies* 36, 3034–3070.

Barber, Brad M, Xing Huang, and Terrance Odean, 2016, Which factors matter to investors? Evidence from mutual fund flows, *Review of Financial Studies* 29, 2600–2642.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022a, Ratings-driven demand and systematic price fluctuations, *Review of Financial Studies* 35, 2790–2838.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022b, What do mutual fund investors really care about? *Review of Financial Studies* 35, 1723–1774.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2024, Discontinued positive feedback trading and the decline of return predictability, *Journal of Financial and Quantitative Analysis* 59, 3062–3100.

Berk, Jonathan B, and Jules H Van Binsbergen, 2016, Assessing asset pricing models using revealed preference, *Journal of Financial Economics* 119, 1–23.

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, *Journal of business* 45, 444–455.

Black, Fischer, Michael C. Jensen, and Myron Scholes, 1972, The capital asset pricing model: Some empirical tests, in Michael C. Jensen, ed., *Studies in the Theory of Capital Markets*, 79–121 (Praeger Publishers).

Boehmer, Ekkehart, Charles M Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking retail investor activity, *Journal of Finance* 76, 2249–2305.

Bretscher, Lorenzo, Ryan Lewis, and Shrihari Santosh, 2023, Investor betas, Working paper, University of Colorado, Boulder.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, 2024, Institutional corporate bond pricing, *Review of Financial Studies* Forthcoming.

Carhart, Mark M, 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, 2015, Regression discontinuity and the price effects of stock market indexing, *Review of Financial Studies* 28, 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, 2023, Corporate bond multipliers: Substitutes matter, Working paper, Columbia Business School.

Chen, Qihui, Nikolai Roussanov, and Xiaoliang Wang, 2023, Semiparametric conditional factor models: Estimation and inference, Working paper, University of Pennsylvania, Wharton.

Chen, Zhongtian, Nikolai Roussanov, Xiaoliang Wang, and Dongchen Zou, 2024, Common risk factors in the returns on stocks, bonds (and options), redux, Working paper, University of Pennsylvania, Wharton.

Choi, Darwin, Wenxi Jiang, and Chao Zhang, 2023, Alpha go everywhere: Machine learning and international stock returns, Working paper, CUHK.

Christoffersen, Susan Kerr, Erfan Danesh, and David K Musto, 2015, Why do institutions delay reporting their shareholdings? Evidence from form 13F, Working paper, University of Toronto.

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

Coval, Joshua, and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479–512.

Da, Zhi, Borja Larrain, Clemens Sialm, and José Tessada, 2018, Destabilizing financial advice: Evidence from pension fund reallocations, *Review of Financial Studies* 31, 3720–3755.

Daniel, Kent, and Sheridan Titman, 2011, Testing factor-model explanations of market anomalies, *Critical Finance Review* 1, 103–139.

Davis, Carter, Mahyar Kargar, and Jiacui Li, 2024, Why is stock-level demand inelastic? A portfolio choice approach, Working paper, Indiana University.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Noise trader risk in financial markets, *Journal of Political Economy* 98, 703–738.

Dou, Winston, Leonid Kogan, and Wei Wu, 2022, Common fund flows: Flow hedging and factor pricing, *Journal of Finance* Forthcoming.

Ehsani, Sina, and Juhani T Linnainmaa, 2022, Factor momentum and the momentum factor, *Journal of Finance* 77, 1877–1919.

Eisfeldt, Andrea L, Bernard Herskovic, and Shuo Liu, 2024, Interdealer price dispersion and intermediary capacity, Working paper, UCLA.

Fama, Eugene F, and Kenneth R French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

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

Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

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

Feng, Guanhao, Jingyu He, and Nicholas G Polson, 2018, Deep learning for predicting asset returns, Working paper, University of Chicago.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

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

Froot, Kenneth A, and Tarun Ramadorai, 2008, Institutional portfolio flows and international investments, *Review of Financial Studies* 21, 937–971.

Gabaix, Xavier, and Ralph SJ Koijen, 2022, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper, Harvard University.

Gabaix, Xavier, and Ralph SJ Koijen, 2024, Granular instrumental variables, *Journal of Political Economy* 132, 2274–2303.

Gabaix, Xavier, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, 2023, Asset embeddings, Working paper, Harvard University.

Gabaix, Xavier, and Matteo Maggiori, 2015, International liquidity and exchange rate dynamics, *Quarterly Journal of Economics* 130, 1369–1420.

Gao, Ming, and Cong Zhang, 2023, Optimizing return forecasts: A bayesian intermediary asset pricing approach, Working paper, University of Chicago.

Giglio, Stefano, Yuan Liao, and Dacheng Xiu, 2021, Thousands of alpha tests, *Review of Financial Studies* 34, 3456–3496.

Giglio, Stefano, and Dacheng Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.

Goyenko, Ruslan, Bryan T Kelly, Tobias J Moskowitz, Yinan Su, and Chao Zhang, 2024, Trading volume alpha, Working paper, Yale University.

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

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

Gupta, Tarun, and Bryan T Kelly, 2018, Factor momentum everywhere, Working paper, Yale University.

Haddad, Valentin, Paul Huebner, and Erik Loualiche, 2024, How competitive is the stock market? Theory, evidence from portfolios, and implications for the rise of passive investing, *American Economic Review* Forthcoming.

Haddad, Valentin, and Tyler Muir, 2021, Do intermediaries matter for aggregate asset prices? *Journal of Finance* 76, 2719–2761.

Hartzmark, Samuel M, and David H Solomon, 2022, Predictable price pressure, Working paper, Boston College.

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

Hasbrouck, Joel, and Duane J Seppi, 2001, Common factors in prices, order flows, and liquidity, *Journal of Financial Economics* 59, 383–411.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

Hendershott, Terrence, Dmitry Livdan, and Dominik Rösch, 2020, Asset pricing: A tale of night and day, *Journal of Financial Economics* 138, 635–662.

Hong, Harrison, and David A Sraer, 2016, Speculative betas, *Journal of Finance* 71, 2095–2144.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2017, A comparison of new factor models, Working paper, OSU.

Huang, Shiyang, Yang Song, and Hong Xiang, 2024, Noise trading and asset pricing factors, *Management Science* Forthcoming.

Huebner, Paul, 2023, The making of momentum: A demand-system perspective, Working paper, Stockholm School of Economics.

Jagannathan, Ravi, and Zhenyu Wang, 1996, The conditional CAPM and the cross-section of expected returns, *Journal of Finance* 51, 3–53.

Jansen, Kristy AE, Wenhao Li, and Lukas Schmid, 2024, Granular treasury demand with arbitrageurs, Working paper, USC.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, 2023, Is there a replication crisis in finance? *Journal of Finance* 78, 2465–2518.

Jensen, Theis Ingerslev, Bryan T Kelly, Semyon Malamud, and Lasse Heje Pedersen, 2024, Machine learning and the implementable efficient frontier, Working paper, Yale University.

Jiang, Zhengyang, Robert J Richmond, and Tony Zhang, 2024, A portfolio approach to global imbalances, *Journal of Finance* 79, 2025–2076.

Jylhä, Petri, 2018, Margin requirements and the security market line, *Journal of Finance* 73, 1281–1321.

Kargar, Mahyar, 2021, Heterogeneous intermediary asset pricing, *Journal of Financial Economics* 141, 505–532.

Kelly, Bryan, Semyon Malamud, and Lasse Heje Pedersen, 2023, Principal portfolios, *Journal of Finance* 78, 347–387.

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

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross-section of present values, *Journal of Finance* 68, 1721–1756.

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

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2020, Instrumented principal component analysis, Working paper, Yale University.

Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo, 2024, Which investors matter for equity valuations and expected returns? *Review of Economic Studies* 91, 2387–2424.

Koijen, Ralph SJ, and Stijn Van Nieuwerburgh, 2011, Predictability of returns and cash flows, *Annu. Rev. Financ. Econ.* 3, 467–491.

Koijen, Ralph SJ, and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.

Koijen, Ralph SJ, and Motohiro Yogo, 2020, Exchange rates and asset prices in a global demand system, Working paper, University of Chicago.

Kondor, Péter, and Dimitri Vayanos, 2019, Liquidity risk and the dynamics of arbitrage capital, *Journal of Finance* 74, 1139–73.

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

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

Lee, Charles MC, and Mark J Ready, 1991, Inferring trade direction from intraday data, *The Journal of Finance* 46, 733–746.

Lettau, Martin, and Sydney Ludvigson, 2001, Resurrecting the (C) CAPM: A cross-sectional test when risk premia are time-varying, *Journal of Political Economy* 109, 1238–1287.

Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *Review of Financial Studies* 33, 2274–2325.

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

Li, Jennifer Jie, Neil D Pearson, and Qi Zhang, 2024, Impact of demand shocks on the stock market: Evidence from Chinese IPOs, Working paper, INSEAD.

Li, Jiacui, 2022, What drives the size and value factors? *Review of Asset Pricing Studies* 12, 845–885.

Li, Jiacui, and Zihan Lin, 2022, Prices are less elastic at more aggregate levels, Working paper, University of Utah.

Lo, Andrew W, and Jiang Wang, 2000, Trading volume: definitions, data analysis, and implications of portfolio theory, *Review of Financial Studies* 13, 257–300.

Lopez-Lira, Alejandro, and Nikolai L Roussanov, 2023, Do common factors really explain the cross-section of stock returns? Working paper, Wharton.

Lou, Dong, 2012, A flow-based explanation for return predictability, *Review of Financial Studies* 25, 3457–3489.

Ludvigson, Sydney C., and Serena Ng, 2009, Macro factors in bond risk premia, *Review of Financial Studies* 22, 5027–5067.

McCracken, Michael W, and Serena Ng, 2016, Fred-md: A monthly database for macroeconomic research, *Journal of Business & Economic Statistics* 34, 574–589.

McLean, R David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.

Merton, Robert C, 1973, An intertemporal capital asset pricing model, *Econometrica* 867–887.

Moskowitz, Tobias J, Yao Hua Ooi, and Lasse Heje Pedersen, 2012, Time series momentum, *Journal of Financial Economics* 104, 228–250.

Pavlova, Anna, and Taisiya Sikorskaya, 2023, Benchmarking intensity, *Review of Financial Studies* 36, 859–903.

Peng, Cameron, and Chen Wang, 2021, Factor demand and factor returns, Working paper, LSE.

Petkova, Ralitsa, and Lu Zhang, 2005, Is value riskier than growth? *Journal of Financial Economics* 78, 187–202.

Rapach, David, and Guofu Zhou, 2013, Forecasting stock returns, in *Handbook of Economic Forecasting*, volume 2, 328–383 (Elsevier).

Ross, Stephen A, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–60.

Shleifer, Andrei, and Lawrence H Summers, 1990, The noise trader approach to finance, *Journal of Economic Perspectives* 4, 19–33.

Shleifer, Andrei, and Robert W. Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.

Smith, Simon C, George Bulkley, and David S Leslie, 2020, Equity premium forecasts with an unknown number of structural breaks, *Journal of Financial Econometrics* 18, 59–94.

Smith, Simon C, and Allan Timmermann, 2021, Break risk, *The Review of Financial Studies* 34, 2045–2100.

Stock, James H, and Mark W Watson, 2002, Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association* 97, 1167–1179.

Teo, Melvyn, and Sung-Jun Woo, 2004, Style effects in the cross-section of stock returns, *Journal of Financial Economics* 74, 367–398.

Vayanos, Dimitri, and Jean-Luc Vila, 2021, A preferred-habitat model of the term structure of interest rates, *Econometrica* 89, 77–112.

Warther, Vincent A., 1995, Aggregate mutual fund flows and security returns, *Journal of Financial Economics* 39, 209–235.

Welch, Ivo, and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* 21, 1455–1508.

Zhou, Kangying, 2024, Active mutual funds and media narratives, Working paper, Yale University.

Online Appendix of "Quantity, Risk, and Return"

A Equilibrium theory and microfoundation of quantity-factor premium relationship

In this section, we develop a theoretical model to provide a formal and explicit interpretation of the empirical framework and findings. This model formalizes key concepts such as sophisticated investors' inelastic demand, noise traders' flows, and how their interaction in equilibrium determines factor premiums. The microfoundation delivers the main empirical specification that observed factor premiums (which are the equilibrium outcomes) are positively related to factor quantities (Eq. 3). The model highlights two theoretical underpinnings that support the strong explanatory power of BTQ for expected stock returns (i.e., high R^2): 1) sophisticated investors exhibit sufficiently inelastic demand, driven by two primitive conditions: high factor-level risk aversion and limited risk-bearing capacity relative to the aggregate stock market; and 2) the noise traders' flows are indeed noisy and exhibit sufficient variation.

A.1 Factor pricing identities when factor premium is an equilibrium outcome

We first provide basic factor pricing identities when the factor premium is allowed to be an endogenous equilibrium outcome. We show that a high factor premium and a low price of the factor risk are two sides of the same coin. The model has two periods t and $t+1$, and the risk-free rate is r_{rf}.

Random state variable $\widetilde{f}_{t+1}$, with $\mathbb{E}_t[\widetilde{f}_{t+1}] = 0$, represents the physical systematic risk of a factor, independent of time-t equilibrium trading outcomes. For simplicity and without loss of generality, we omit subscript k for variables $\widetilde{f}$ as well as q, μ etc. below, as the theory applies generically to any factor.

Let $f_{t+1}(q_t)$ be the payoff of a zero-cost factor portfolio that has a unit exposure to $\widetilde{f}_{t+1}$.

This variable represents the observed long-short portfolio *return* (e.g., the SMB series from French's website). With "(q_t)," we allow for a general form where the payoff random variable depends on the equilibrium quantity variable q_t. By construction,

$$f_{t+1}(q_t) = \widetilde{f}_{t+1} + \mu_t(q_t), \qquad \forall q_t, \tag{A.1}$$

where $\mu_t(q_t) = \mathbb{E}_t[f_{t+1}(q_t)]$ is the factor premium, the focus of this section. In the cross section, by the APT, a stock with exposure β_t to systematic risk $\widetilde{f}_{t+1}$ has expected excess return $\beta_t \mu_t(q_t)$ (assuming no exposure to other systematic risks).

Turning to time-t prices, we define $P_t^{\widetilde{f}}(q_t) := \mathbb{P}\text{rice}_t[\widetilde{f}_{t+1}](q_t)$ as the price of the state contingent payoff $\widetilde{f}_{t+1}$, where $\mathbb{P}\text{rice}_t[\,\cdot\,](q_t) := \mathbb{E}_t[M_{t+1}(q_t)\cdot\,]$ is the payoff pricing operator with the stochastic discount factor (SDF) $M_{t+1}(q_t)$. Once again, "(q_t)" indicates that the prices are endogenous to q_t.

By construction, the price of the factor portfolio payoff is 0: $\mathbb{P}\text{rice}_t[f_{t+1}(q_t)](q_t) = 0, \forall q_t$. This leads to the intuitive identity between equilibrium factor premium and factor risk price. Applying the pricing operator to both sides of the factor return decomposition (A.1), by the law of one price, we obtain $0 = \mathbb{P}\text{rice}_t\left[f_{t+1}(q_t)\right] = P_t^{\widetilde{f}}(q_t) + \mu_t(q_t)/(1 + r_{\text{rf}})$. Therefore,

$$\mu_t(q_t) = -(1 + r_{\text{rf}}) P_t^{\widetilde{f}}(q_t), \qquad \forall q_t. \tag{A.2}$$

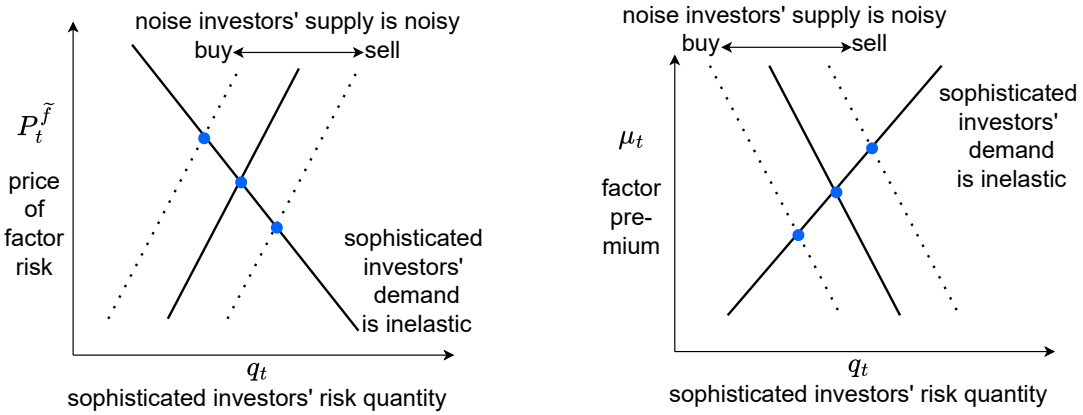
This expression captures the canonical inverse relationship between the zero-cost factor premium and the price of zero-mean factor risk. Our model extends the canonical relationship by allowing both sides of the equation to be endogenous equilibrium outcomes.

A.2 Demand functions and the equilibrium

Next, we introduce the demand and supply functions of the sophisticated and noise investors, along with their equilibrium interactions, as illustrated in Figure A.1.

Sophisticated investors obey factor pricing and enforce the law of one price. The afore-

Figure A.1: Demand functions and the equilibrium



Note: Two forces for strong quantity-factor premium empirical relationship are 1) inelastic sophisticated investors' demand (sloped, not flat), and 2) significant fluctuations in noise traders' supply (shifts left and right). The two panels are equivalent illustrations with price or factor premium—just flip everything upside down according to Eq. A.2.

mentioned SDF, $M_{t+1}(q_t)$, and the pricing operator are interpreted as theirs, which are influenced by their quantity holdings: q_t^{sophi}. Hence, their demand function is: $P_t^{\widetilde{f}}(q_t^{\text{sophi}}) = \mathbb{P}\text{rice}_t[\widetilde{f}_{t+1}](q_t^{\text{sophi}}) = \mathbb{E}_t[M_{t+1}(q_t^{\text{sophi}})\widetilde{f}_{t+1}]$.

Both $P_t^{\widetilde{f}}(\,\cdot\,)$ and $\mu_t(\,\cdot\,)$ are equivalent representations of demand, connected by Eq. A.2. The theory can be equivalently stated with either one. This equivalence indicates that the realized zero-cost returns $f_{t+1}(q_t)$, such as those from Fama-French, should be modeled as the equilibrium outcome of time-t trading activities, rather than exogenous variables.

Empirically, typical demand-based asset pricing studies target *concurrent* price impacts. They model investors' demand curves as $\mathbb{P}\text{rice}_t[\widetilde{X}_{t+1}](q_t)$, where a payoff $\widetilde{X}_{t+1}$ (in our case $\widetilde{f}_{t+1}$) is independent of the time-t equilibrium quantity. In this paper, we shift the focus from concurrent price impact to risk premiums, aligning more closely with factor pricing literature's emphasis on expected future returns. With the identity in Eq. A.2, we explicitly show how our and the literature's focuses are related. Shifting from concurrent price impact to future returns also dispenses with the endogeneity issue between q_t and $P_t^{\widetilde{f}}$ (such as flow chasing concurrent return), which necessitates instrumental variable methods in Gabaix and

Koijen (2024). Nonetheless, our model still requires the exogeneity of time-t trading to future risk $\widetilde{f}_{t+1}$ realization, which is easier to justify.

Inelastic demand by sophisticated investors is key to the reported empirical relationships. That is, $P^{\widetilde{f}}(\,\cdot\,)$ is downward sloping (or $\mu(\,\cdot\,)$ is upward sloping). Perfectly elastic demand (a flat line where price remains constant regardless of quantity) would imply that the sophisticated investors have unlimited risk-bearing capacity or "deep pockets," inconsistent with real-world observations. The microfoundation in the next subsection connects inelasticity to sophisticated investors' risk aversion and their capital share in the market.

Since sophisticated investors trade with noise traders, the "supply" function of noise traders (how much q they sell) is $q_t^{\text{noise}}(P_t^{\widetilde{f}}) = \text{DeterministicSupply}(P_t^{\widetilde{f}}) + \eta_t$, where η_t is the noise supply component and unspecified function $\text{DeterministicSupply}(P_t^{\widetilde{f}})$ is the deterministic component of supply that can respond elastically to price.

We are agnostic about most aspects of the noise traders' supply, but require that the noise component η_t exhibits significant variation and is indeed noisy in the sense of not predicting the future factor risk $\widetilde{f}_{t+1}$. It does not affect the result whether $\text{DeterministicSupply}(\cdot)$ is perfectly inelastic (meaning noise traders are completely insensitive to price, vertical in Figure A.1) or somewhat elastic (meaning they partially adjust supply to price). This is because the observed equilibrium quantity is q_t, and that η_t is unobserved and unmodelled anyway. It is implausible that $\text{DeterministicSupply}(\cdot)$ is or near perfectly elastic (horizontal in Figure A.1), as it would imply that the noise traders are not only "sophisticated" about price but also "deep-pocketed." We are also agnostic about the sources of the "noise," which could be driven by various factors such as investor sentiment, beliefs, or media narratives (e.g., Zhou, 2024).

The equilibrium is obtained by market clearing: $q_t = q_t^{\text{sophi}} = q_t^{\text{noise}}$, given the two demand (supply) functions. Equilibrium q_t is determined by noise trading shock η_t (with an unspecified function).

The empirical model additionally assumes that sophisticated investors' demand curve

is time-invariant: $P_t^{\widetilde{f}}(q_t) = P^{\widetilde{f}}(q_t)$, or equivalently $\mu_t(q_t) = \mu(q_t)$ for any t and q_t. This is largely an empirical restriction stating that, besides q_t^{sophi}, the demand function is not affected by any other time-varying variables. Section 6 explores alternative variables, such as factor momentum and macro variables, but finds none with reliable empirical power for pricing the cross section of stock returns. The next subsection provides a theoretical demand function with this property, under appropriate assumptions.

With all these setups, we conclude that the observed equilibrium outcomes, $\{q_t, P_t^{\widetilde{f}}\}$, lie on the sophisticated investors' demand curve $P^{\widetilde{f}}(\,\cdot\,)$, and equivalently, the observed $\{q_t, \mu_t\}$ on $\mu(\,\cdot\,)$. This establishes that, across periods, the factor premium μ_t is positively related to the factor quantity q_t.

A.3 A microfoundation of the inelastic demand function

In this subsection, we provide a specific microfoundation for sophisticated investors' pricing kernel, which results in an inelastic, downward-sloping demand function with a closed-form expression, $\mu(q_t) = \mu + \lambda q_t$, consistent with the empirical model in Eq. 3. The inelasticity arises from two key factors: 1) the risk aversion of sophisticated investors and 2) their limited total capital for absorbing the factor risk. Here, we derive an analytical expression for these two forces, providing one simple mechanism for inelastic demand, though other mechanisms, such as investment mandates (Gabaix and Koijen, 2022), are also possible.

Suppose sophisticated investors' total wealth (AUM) is $\$W_t$ and their existing portfolio has a random payoff $\$W_{t+1}$, with a return $r_{t+1}^{\text{W}} := W_{t+1}/W_t - 1$. A representative sophisticated investor with \$1 AUM has CARA utility with risk aversion γ, $\mathbb{E}_t[-\exp(-\gamma r_{t+1}^{\text{W}})]$. The utility and demand have identical functional forms across sophisticated investors and scale proportionally with their individual AUMs, under the standard assumption that CARA risk aversion scales inversely with the AUM.

Taking zero-cost factor payoff $f_{t+1}(q_t)$ as given, the representative sophisticated investor

optimally allocates additional exposure to systematic risk factors b by solving:

$$b_t = \arg\max_b \mathbb{E}_t[-\exp(-\gamma(r^{\mathrm{W}}_{t+1} + b f_{t+1}(q_t)).] \tag{A.3}$$

Assuming that the wealth return r^{W}_{t+1} and factor return $f_{t+1}(q_t)$ are jointly normally distributed, the first-order condition of Eq. A.3 implies:

$$\begin{aligned}\mu_t(q_t) = \mathbb{E}_t[f_{t+1}(q_t)] &= \gamma \mathrm{cov}_t(r^{\mathrm{W}}_{t+1}, f_{t+1}(q_t)) + \gamma b_t \mathrm{var}_t(f_{t+1}(q_t)) \\ &= \gamma \mathrm{cov}_t(r^{\mathrm{W}}_{t+1}, \widetilde{f}_{t+1}) + \gamma b_t \mathrm{var}_t(\widetilde{f}_{t+1}),\end{aligned} \tag{A.4}$$

where the last equality uses Eq. A.1. Given per AUM demand b_t, the aggregated demand of additional factor exposure is $b_t W_t$.

Up to this point, we have established the demand function connecting the factor premium to the sophisticated investors' risk holdings measured in terms of their factor exposure: $b_t W_t$. The empirical counterpart of this measure is the flow-induced factor beta: $\sum_i \$\mathrm{flow}^{\mathrm{stock}}_{i,t} \widehat{\beta}_{i,k,t}$, the intermediate term in constructing the quantity variable in Eq. 8. To connect the demand function in this measure to the quantity variable q_t in the empirical model, we assume that the aggregate stock market capitalization at time t is AGG_t, and sophisticated investors' total wealth W_t is a constant fraction π of AGG_t, such that $W_t = \pi \mathrm{AGG}_t$. Therefore, the factor quantity defined in Section 3.2 can be simplified as

$$\widetilde{q}_t = b_t W_t \mathrm{var}_t[f_{t+1}(q_t)]/\mathrm{AGG}_t = b_t \mathrm{var}_t[f_{t+1}(q_t)]\pi. \tag{A.5}$$

Substituting into Eq. A.4, we have the demand function in terms of $\widetilde{q}_t$:

$$\mu_t(q_t) = \gamma \mathrm{cov}_t(r^{\mathrm{W}}_{t+1}, \widetilde{f}_{t+1}) + \frac{\gamma}{\pi}\widetilde{q}_t. \tag{A.6}$$

In the empirical implementation, we also standardize $\widetilde{q}_t$ as $q_t = \widetilde{q}_t/\sigma(\widetilde{q}_t)$, so that the estimated

linear coefficient λ can be more intuitively interpreted as the price effect per standard-deviation shock in $\widetilde{q}_t$. (The magnitude of $\widetilde{q}_t$ tends to be very small, because this raw measure is normalized by the aggregate stock market capitalization.)

Finally, the demand function matches the empirical specification in Eq. 3:

$$\mu(q_t) = \gamma \text{cov}_t(r^{\text{W}}_{t+1}, \widetilde{f}_{t+1}) + \frac{\gamma}{\pi}\sigma(\widetilde{q}_t)q_t \tag{A.7}$$

$$= \mu + \lambda q_t, \tag{A.8}$$

with $\mu := \gamma \text{cov}_t(r^{\text{W}}_{t+1}, \widetilde{f}_{t+1})$ and $\lambda := (\gamma/\pi)\sigma(\widetilde{q}_t)$.

Both terms in Eq. A.7 have clear interpretations. In the first term, we assume $\gamma \text{cov}_t(r^{\text{W}}_{t+1}, \widetilde{f}_{t+1}) = \mu$. That is, the background factor risk premium (the risk premium when no additional factor exposure is taken, $b_t = q_t = 0$) is constant. Empirically, it implies that no conditioning variables other than q_t affect the factor premium. Section 6 supports this assumption and finds no reliable alternative conditioning variables for pricing the cross section of stock returns. In fact, the main empirical results show, via the "beta-only" benchmark, that μ is indistinguishable from zero, suggesting that other conditioning variables have effects too weak to be empirically detected.

The second term, $\lambda := (\gamma/\pi)\sigma(\widetilde{q}_t)$, shows the conditions underlying the strong empirical relationship between quantity and factor premium. The first condition requires that sophisticated investors' demand is sufficiently inelastic. This inelasticity is governed by two primitive parameters: 1) high risk-aversion (high γ) and 2) a small share of sophisticated investors relative to the stock market (low π). Specifically, the factor premium multiplier is γ/π; and the price multiplier, according to Eq. A.2, is $\gamma/(\pi(1 + r_{\text{rf}}))$, inversely related to the elasticity of the demand function. This aligns with Gabaix and Koijen (2022), who also highlight the limited share of hedge funds capital (around 5% of total investors) as a key driver of the inelastic demand. The second condition requires noise traders' supply to exhibit significant variation (large $\sigma(\widetilde{q}_t)$). Greater variation in the noise trading—reflected in large

left-right shifts in the supply curve—translates to greater variation in the factor premium attributed to the quantity channel, which this paper captures empirically.

B Technical details

B.1 Construction and cleaning of mutual fund flows

In this appendix, we present details related to constructing and cleaning mutual fund flows. Our primary data source is the CRSP Survivorship-Bias-Free Mutual Fund database. We start with all funds' return and total net assets (TNA) data at the share-class level. A mutual fund may include multiple share classes. We first drop observations with no valid CRSP identifier, `crsp_fundno`. A few fund-share classes report multiple TNAs in a given month. These are likely data duplicates, so we keep only the first observation of the month. In what follows, we discuss the cleaning steps for returns and TNA at the share-class level. After cleaning, we aggregate the share-class level data to the fund level.

B.1.1 Return cleaning

We first correct data errors in monthly mutual fund net returns, `mret`.

First, we address extremely positive returns. We study the case in which a particular fund share has returns greater than 100% and has existed for more than one year.[1] We manually check the entire time series of each share class in this subsample. Most of these extreme returns reflect misplaced decimal points, which confound returns in decimal and percentage formats. For these cases, we divide the faulty returns by 100.

Second, we address extreme negative returns. Similarly, we study the case in which a particular fund share has existed for more than one year and has returns lower than -50%. With extremely negative returns, we need to distinguish data errors from significantly

[1] We use the one-year threshold because mutual fund return and TNA during the first year are sometimes inaccurate in the CRSP database. For example, return and TNA can be stale or reported using a placeholder number such as 0.1. We address these issues by cross-checking with the alternative database.

negative returns before a fund's closure. Thus, we manually check only the subsample of negative returns that occur at least one year prior to the last observation of a closed fund. We manually check whether these extreme returns reflect data-input errors for this subsample. For the cases with misplaced decimal points, we divide the faulty returns by 100.

B.1.2 TNA cleaning

Unlike many prior studies that construct percentage mutual fund flows, we study dollar-value flows to preserve the cross-sectional relative magnitudes. The mutual fund size distribution features a very long right tail. Winsorizing the extreme dollar-value TNA likely removes both valid large values and input errors, generating significant bias. We devise an algorithm to identify and correct erroneous observations of TNA:

1. Using the sample with corrected returns, we calculate dollar flows as in Eq. 6 at the share-class level.

2. We study the top and bottom 0.5% of all dollar flows.[2] We manually check this subsample's TNA time series of all share classes. We identify several common errors:

   - Misplaced decimal points (usually by hundredths or thousandths).
   - Stale TNA observations from CRSP, typically when a fund reorganizes its share class offering (e.g., adding a new share class and moving assets into a single share class).
   - CRSP sometimes sets TNA = 0.1 for the first few months of a new fund or a new share class.

   We correct the misplaced decimal issue. For funds suffering from the latter two problems, we obtain their TNA from Morningstar.[3] Morningstar's TNA data

[2]The choice of the top and bottom 0.5% is motivated by the distribution of dollar flows, where most extreme values tend to occur at these tails.

[3]We merge the CRSP and Morningstar databases using a fund's ticker.

(`Net_Assets_ShareClass_Monthly`) suffer to a lesser extent from these issues than CRSP's TNA data. We conclude by further cross-checking other third-party vendors (e.g., Yahoo Finance and Bloomberg Terminal). Hence, whenever a fund's CRSP TNA deviates more than 50% from its Morningstar TNA, we use the Morningstar TNA.

3. We repeat the previous steps one more time to ensure that we have accounted for most, if not all, extreme errors.

4. We compare our cleaned TNA with total assets (`assets`) from Thomson/Refinitiv Holdings data.[4] Following Coval and Stafford (2007) and Lou (2012), we drop observations whenever our cleaned TNA deviate more than 50% from `assets` from Thomson/Refinitiv.

Using cleaned return and TNA data, we calculate dollar flows at the share-class level using equation (6). We obtain a fund's flows by adding up the flows of all share classes in the same fund. The final sample contains 1,707,742 fund×month observations.

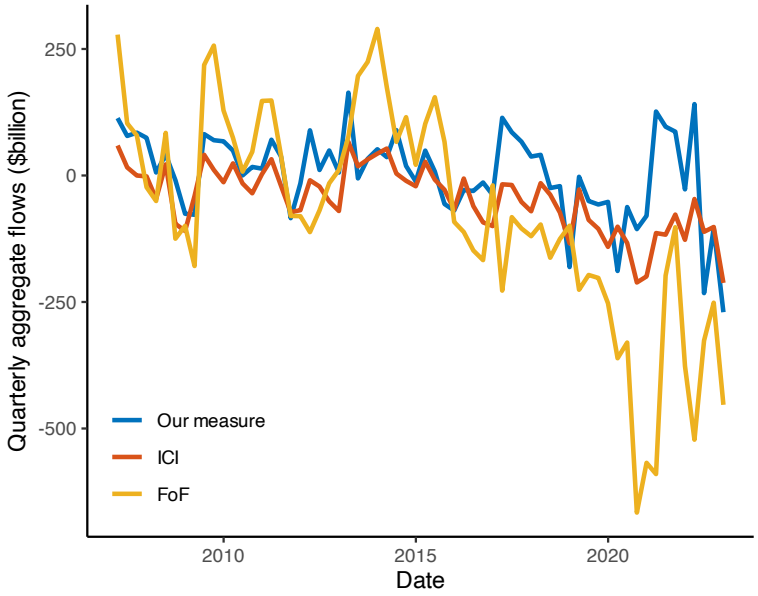
B.1.3 Cross-validating the data-cleaning procedure

We cross-validate our data-cleaning procedure by comparing our aggregated mutual fund flows with alternative sources. We compute the quarterly aggregate flows in dollar amounts from our main sample and compare them with data from the Investment Company Institute (ICI) and the Flow of Funds (FoF).

The ICI publishes aggregate monthly mutual fund flows, from which we extract quarterly data spanning from 2007 to 2022. Specifically, we use the ICI's Total Equity mutual fund flows, which align closely with the coverage of mutual funds in our sample. Additionally, we draw on data from the Federal Reserve Board's Financial Accounts of the United States – Z.1 (formerly known as the Flow of Funds or FoF) from the same time period, providing quarterly observations. For our analysis, we focus on mutual fund flows (Line 28) within

[4] We merge the two databases via the linking table MFLINKS, which WRDS provides.

Figure A.2: Time series of aggregate mutual fund flows from various sources



Note: The figure plots the quarterly time series of our measure, ICI flows, and Flow of Funds (FoF) flows.

Corporate Equities (Table 223) and use unadjusted flows (FU).

Figure A.2 plots the quarterly time series of aggregate mutual fund flows from all three sources. Our measure of aggregate mutual fund flows is broadly consistent with the other two sources. The correlation between our aggregate flow measure and ICI flow is 0.63, while the correlation between our measure and FoF flow is 0.47.

The differences observed in Figure A.2 among the three measures likely reflect variations in mutual fund coverage. Specifically, the ICI flow tracks virtually all U.S. equity mutual funds that invest in both domestic and world equity markets.[5] The FoF flow, sourced from unpublished ICI data, aggregates unadjusted flows into and out of all U.S. mutual funds (including variable annuity long-term mutual funds). It is calculated based on mutual fund assets in common stock, preferred stock, and rights and warrants.[6] In comparison, our mutual fund sample contains U.S. mutual funds covered by CRSP, which collects historical

[5]The ICI is a trade association for the mutual fund industry, and virtually all U.S. mutual funds are ICI members (Warther, 1995).

[6]See https://www.federalreserve.gov/apps/fof/SeriesAnalyzer.aspx?s=FA653064100&t=F.223&suf=Q.

data from various sources.[7] Due to the nature of the data collection process, CRSP's coverage is smaller than ICI's coverage.

B.1.4 Alternative method to construct factor-level flow directly from mutual fund flow

In an alternative method, flows into mutual funds, $\$\text{flow}_{m,t}^{\text{fund}}$, can be directly aggregated into flows into factors, $\text{flow}_{k,t}^{\text{factor}}$, by substituting Eq. 7 into Eq. 8. This approach approximately aggregates $\$\text{flow}_{m,t}^{\text{fund}}$ based on each fund's beta to each factor. Specifically, we have

$$\text{flow}_{k,t}^{\text{factor}} = -\sum_{i}\sum_{\text{fund } m} \$\text{flow}_{m,t}^{\text{fund}} \text{weight}_{i,m,\text{quarter}(t)-2}^{\text{fund}} \widehat{\beta}_{i,k,t} \widehat{\text{var}}_t(f_{k,t}). \tag{A.9}$$

The alternative is

$$-\sum_{\text{fund } m} \$\text{flow}_{m,t}^{\text{fund}} \widehat{\beta}_{m,k,t} \widehat{\text{var}}_t(f_{k,t}), \tag{A.10}$$

where $\widehat{\beta}_{m,k,t}$ is the beta of fund m to factor k in month t. These two are approximately the same because $\sum_i \text{weight}_{i,m,\text{quarter}(t)-2}^{\text{fund}} \widehat{\beta}_{i,k,t} \approx \sum_i \text{weight}_{i,m,t}^{\text{fund}} \widehat{\beta}_{i,k,t} = \widehat{\beta}_{m,k,t}$.

We do not use the alternative method for two reasons. First, our method follows the literature and can help exclude potential informed trading by mutual fund managers. The direct approach is only an approximation because we use lagged, not current, mutual fund holdings to construct $\$\text{flow}_{i,t}^{\text{stock}}$, an important detail to exclude potential informed trading by mutual fund managers (Lou, 2012). Second, starting from stock-level flows and building upward (Eq. 8) is more general and extends beyond mutual fund flow-induced trading. For example, retail investor order flow imbalance, a widely used measure of noise trading in the literature (Lee and Ready, 1991; Boehmer, Jones, Zhang, and Zhang, 2021; Li and Lin, 2022), is constructed directly at the stock level.

[7] The sources include the Fund Scope Monthly Investment Company Magazine, the Investment Dealers Digest Mutual Fund Guide, Investor's Mutual Fund Guide, the United and Babson Mutual Fund Selector, and the Wiesenberger Investment Companies Annual Volumes.

B.2 Technical details of Lasso implementation

In optimization (13), adding the term "$1/\sigma(\widetilde{q}_{k,t})$" is technically necessary because we have already standardized $\widetilde{q}_{k,t}$ to $q_{k,t} = \widetilde{q}_{k,t}/\sigma(\widetilde{q}_{k,t})$ (see Section 3.2). Optimization (13) therefore is equivalent to running the standard Lasso on the pre-standardized BTQ ($\widehat{\beta} \times \widetilde{q}$)

$$\min_{\widetilde{\lambda}_1 \ldots \widetilde{\lambda}_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \widetilde{\lambda}_k \widehat{\beta}_{i,k,t} \widetilde{q}_{k,t} \right)^2 + \omega \sum_{k=1}^{K} \left| \widetilde{\lambda}_k \right|, \tag{A.11}$$

and then standardizing the coefficients for economic interpretation: $\lambda_k = \widetilde{\lambda}_k \sigma(\widetilde{q}_{k,t})$. Although we standardize $\widetilde{q}_{k,t}$ for interpretability, we do not want to lose the information contained in the original quantity $\widetilde{q}_{k,t}$ during the Lasso selection process. A factor with greater variation in $\widetilde{q}_{k,t}$ will have an inflated λ_k after being standardized to $q_{k,t}$, but it should not be additionally penalized for this reason. Standard Lasso implementation where the economic interpretation is not a priority would typically standardize the predictor (BTQ together) across the $\{i, t\}$ panel. Here, we customize the standardization based on the required economic interpretation.

Similarly, for the β-only model, the Lasso implementation is

$$\min_{\mu_1 \ldots \mu_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \mu_k \widehat{\beta}_{i,k,t} \right)^2 + \omega \sum_{k=1}^{K} |\mu_k|. \tag{A.12}$$

We perform ten-fold cross-validation to tune hyperparameter ω based on only in-sample information from 2000 to 2009. In each fold, we exclude one year of observations and solve the Lasso problem (A.11) using the remaining nine years of in-sample data. The model is evaluated in the left-out year to form predicted returns $\widehat{r}^{[\text{cv}]}_{i,t+1}$. After enumerating all folds and forming predicted returns for all in-sample observations, we calculate the cross-validated (CV) in-sample mean squared errors (MSE) as $\sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \widehat{r}^{[\text{cv}]}_{i,t+1} \right)^2$. Hyperparameter ω is tuned by choosing the one with the minimum CV MSE.

C Additional empirical results

C.1 Additional properties of the quantity variable $q_{k,t}$

Figure A.3 reports various statistics of the constructed quantity variables $q_{k,t}$ to show the extent to which these time-series variables comove. Panel A shows the pairwise correlation matrix of the four Fama-French-Carhart factors. The results reveal some comovement (both positive and negative) among the four variables, which is also evidenced in the time series plot in Figure 1 in the main text. HML-MOM has the greatest (in absolute value) correlation of -0.75. Nonetheless, all pairwise correlations are far from ± 1, indicating that the $q_{k,t}$ variables are far from collinear, and each captures unique information about the underlying quantity variations.

Panel B shows a similar pattern of limited comovement among the 153 JKP factors. Instead of reporting pairwise correlations, we conduct a principal component analysis (PCA) on the $q_{k,t}$ variables of the 153 JKP factors and report the cumulative explained variances by principal components. The plot reveals a clear factor structure among the 153 factor-level $q_{k,t}$ variables, yet also indicates significant unique information across different dimensions of $q_{k,t}$. The first principal component explains around half of the total variance, and the first two principal components in total explain around 77% of the total variance. It requires five principal components to explain 90% of the total variance, and seven to explain 95% of the total variance. We also note that these in-sample PC statistics are likely exaggerated due to overfitting.

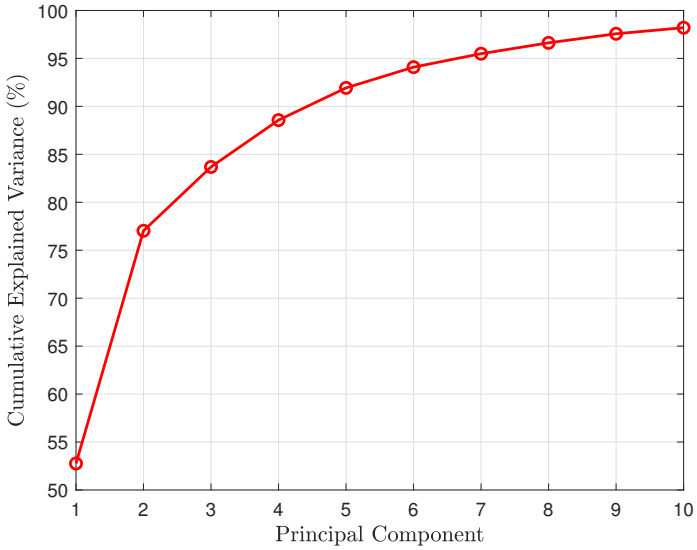
In summary, the $q_{k,t}$ variables are not collinear and capture unique information about the underlying quantity variations. This feature indicates the paper's main result on BTQ's predictive power is not driven by a few special $q_{k,t}$ variables. The fact that BTQ's predictive power is consistent across various factor specifications speaks to the robustness of the underlying economic mechanism.

Figure A.3: Degree of comovement among the quantity variables (q) of different factors

A. Correlation matrix

	MKT	SMB	HML	MOM
MKT	1			
SMB	0.55	1		
HML	0.47	0.57	1	
MOM	-0.47	-0.23	-0.75	1

B. PC variances for q's of 153 JKP factors



Note: Panel A: pairwise correlation for $q_{k,t}$ of the four Fama-French-Carhart factors. Panel B: cumulative explained variances by principal components of $q_{k,t}$ series of the 153 JKP factors.

C.2 Predicting factor returns with factor quantity

This appendix subsection presents the results of using factor quantities to predict factor returns. While time-series predictability is not the main focus of this paper, it is naturally implied by the paper's theoretical framework, particularly the factor premium modeling in Eq. 3. Empirically, we successfully detect the predictability to a certain extent, consistent with the theoretical motivation. However, we also note the apparent methodological limitations of predicting factor returns with simple time-series regressions.

Table A.1 presents the results of the time-series regression $f_{k,t+1} = \mu_k + \lambda_k q_{k,t} + error_{k,t+1}$ for various factors. The estimated λ_k coefficients are predominantly positive and statistically significant for all Fama-French-Carhart factors and the majority of JKP factors. This indicates that each factor's expected return is positively related to its quantity, consistent with our theoretical motivation. The full-sample R^2 values are around 5%, which is relatively high for factor return prediction at the monthly frequency (see Welch and Goyal, 2008).

However, the OOS R^2 values are mostly negative, with the exception of the market fac-

Table A.1: Predicting factor return $f_{k,t+1}$ using quantity $q_{k,t}$

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
λ_k (%)	1.04	0.49	0.82	1.10	0.25	0.66	1.00
t-stat	(3.25)	(2.45)	(2.89)	(1.76)	(1.41)	(2.00)	(2.64)
μ_k (%)	0.38	0.19	0.08	0.36	-0.20	-0.01	0.27
t-stat	(1.39)	(1.16)	(0.38)	(1.41)	(-1.41)	(-0.10)	(1.63)
R^2 (%)	5.05	2.48	5.59	4.35	1.45	4.74	7.36
OOS R^2 (%)	6.74	-1.05	-14.70	-1.29	-7.08	-0.95	2.07

Note: Factor return predictive regressions ($f_{k,t+1} = \mu_k + \lambda_k q_{k,t} + error_{k,t+1}$) for $k =$ each of the Fama-French-Carhart factors and JKP factors. The point estimates are in percentage terms. That is, the first cell indicates a one standard deviation increase in $q_{k,t}$ predicts a 1.04% increase in market return in the following month. The t-statistics are based on Newey-West standard errors. The first five rows are full-sample regressions (2000-2022) with R^2 evaluated in the same full sample. The ordinary IS R^2 with a constant term is reported: $R^2 = 1 - \sum_t (f_{k,t+1} - \widehat{f}_{k,t+1})^2 / \sum_t (f_{k,t+1} - \widehat{\mu}_k)^2$. The last row "OOS R^2" is with the regressions estimated in 2000-2009 and evaluated in 2010-2022, and we benchmark the R^2 against predicting zero: OOS $R^2 = 1 - \sum_t (f_{k,t+1} - \widehat{f}_{k,t+1})^2 / \sum_t f_{k,t+1}^2$.

tor, which achieves a higher R^2 of 6.8%. Nonetheless, we should interpret this high R^2 with caution as the time-series R^2 metric may contain significant noise. These results underscore the apparent limitations of using simple univariate time-series regression for predicting aggregate factor returns. The construction of the factor quantity series is not designed for time-series return prediction, which is understood to be a challenging task that requires more sophisticated methods and richer predictor data (Kelly and Pruitt, 2013; Kelly, Malamud, and Zhou, 2024).[8]

C.3 Additional results on stock return forecasting

This appendix subsection contains additional empirical results on stock return forecasting that are omitted in the main text Subsection 4.3. Table A.2 completes Table 3 by providing the full-sample coefficient estimates for the β-only model.

[8]Dynamic regime shift models can further enhance prediction accuracy by accounting for structural breaks (Smith, Bulkley, and Leslie, 2020; Smith and Timmermann, 2021; Gao and Zhang, 2023).

Table A.2: Table 3 continued, β-only model's coefficient estimates

	CAPM	FF3	FF3C	FF5	FF5C
β-only model: μ_k (%, monthly), t-statistics in parentheses					
MKT	0.38	0.45	0.35	0.55	0.50
	(1.07)	(0.90)	(0.75)	(1.21)	(1.15)
SMB		-0.05	0.06	-0.04	0.09
		(-0.15)	(0.19)	(-0.11)	(0.27)
HML		0.58	0.51	0.56	0.44
		(1.74)	(1.59)	(1.49)	(1.23)
MOM			-0.41		-0.48
			(-1.09)		(-1.26)
CMA				0.04	0.10
				(0.17)	(0.51)
RMW				0.09	0.13
				(0.34)	(0.52)

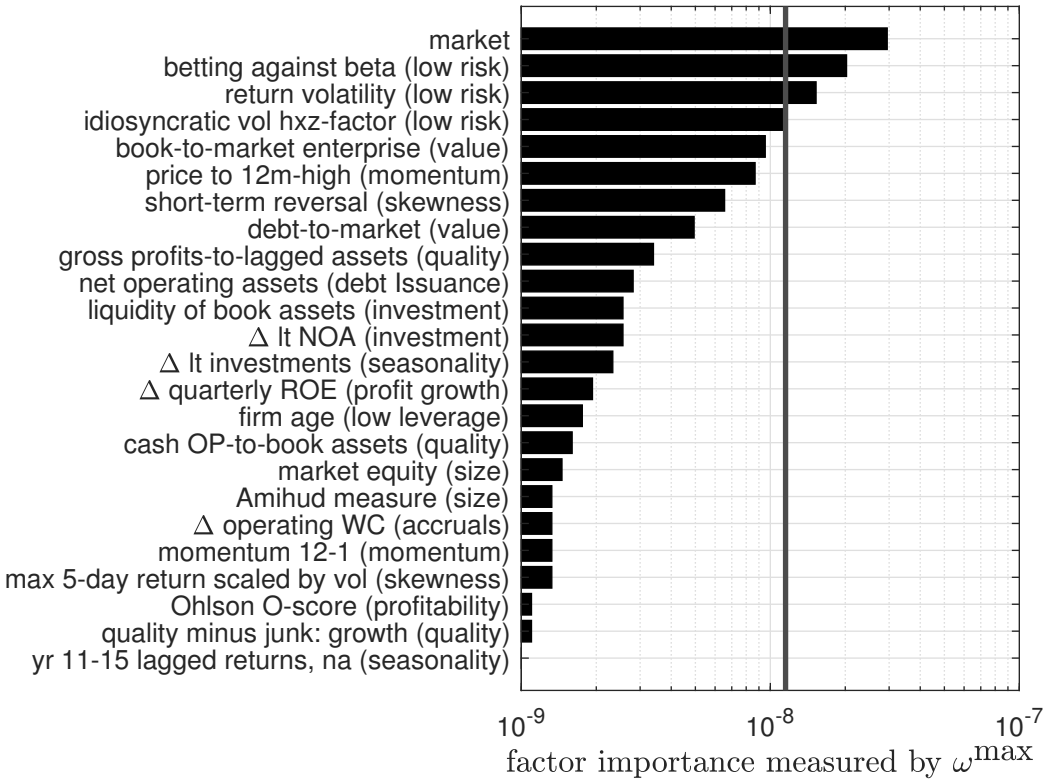
Note: Table 3 in the main text reports the R^2 values for the BTQ and the β-only models, as well as the coefficients for the BTQ model. This table reports the β-only model's coefficient estimates. Same as the main text table, the μ_k coefficients are in percentage terms, and the t-statistics are based on standard errors clustered by month.

Table 3 in the main text already shows that the β-only model has weak predictive power, with low and even negative R^2 values in some OOS cases. Table A.2 further shows that the μ coefficients in the β-only model are either statistically insignificant or negative in various factor specifications. This, once again, shows the empirical difficulty in establishing a positive risk-return association using β only without quantity information.

C.4 Additional results on factor selection

This appendix subsection presents a more formal factor importance analysis and reports other important factors besides the top five reported in Section 4.4 for the Lasso estimation (Eq. 13). Factor importance is measured using a feature selection metric from the Lasso regressions. In particular, we measure the importance of factor k by $\omega_k^{\max}$, the largest ω value at which factor k is still selected. Specifically, $\omega_k^{\max} := \sup\{\omega : \widehat{\lambda}_k(\omega) \neq 0\}$, where $\widehat{\lambda}_k(\omega)$ is the Lasso estimate of λ_k at hyperparameter ω. Figure A.4 reports $\omega_k^{\max}$ for the top

Figure A.4: Factor importance in Lasso factor selection



Note: We measure the importance of factor k by $\omega_k^{\max}$, the largest ω value at which factor k is still selected. Specifically, $\omega_k^{\max} := \sup\{\omega : \widehat{\lambda}_k(\omega) \neq 0\}$, where $\widehat{\lambda}_k(\omega)$ is the Lasso estimate of λ_k at hyperparameter ω. This figure reports $\omega_k^{\max}$ for the top 24 factors in the JKP factor zoo, omitting the rest with $\omega_k^{\max} < 10^{-9}$. The vertical black line indicates the tuned ω based on ten-fold cross-validation.

24 factors in the JKP factor zoo, omitting the rest with $\omega_k^{\max} < 10^{-9}$.

The most significant factor is unambiguously the market factor, followed by the low-risk factors constructed with technical (past return) information, the value factor constructed with fundamental information, and a version of the momentum factors. The remaining less important factors are related to investment style clusters such as the value, quality, investment, seasonality, etc. Specifically, the top 24 factors' full names, the factor clusters they belong to, and code names (as in JKP data) are detailed below:

market	–	`mkt`,
betting against beta	low risk	`betabab_1260d`,
return volatility	low risk	`rvol_21d`,
idiosyncratic volatility q-factor	low risk	`ivol_hxz4_21d`,
book-to-market enterprise	value	`bev_mev`,
current price to high price over last year	momentum	`prc_highprc_252d`,

short-term reversal	skewness	`ret_1_0`,
debt-to-market	value	`debt_me`,
gross profits-to-lagged assets	quality	`gp_atl1`,
net operating assets	debt issuance	`noa_at`,
liquidity of book assets	investment	`aliq_at`,
change in long-term net operating assets	investment	`lnoa_gr1a`,
change in long-term investments	seasonality	`lti_gr1a`,
change in quarterly return on equity	profit growth	`niq_be_chg1`,
firm age	low leverage	`age`,
cash-based operating profits-to-book assets	quality	`cop_at`,
market equity	size	`market_equity`,
Amihud measure	size	`ami_126d`,
change in current operating working capital	accruals	`cowc_gr1a`,
price momentum t-12 to t-1	momentum	`ret_12_1`,
highest 5 days of return scaled by volatility	skewness	`rmax5_rvol_21d`,
Ohlson O-score	profitability	`o_score`,
quality minus junk: growth	quality	`qmj_growth`,
years 11-15 lagged returns, nonannual	seasonality	`seas_11_15na`.

C.5 Interpreting the magnitude of λ estimates and connection to the literature

This appendix shows that the magnitude of the λ estimates—reported at the factor level in Tables 2 and 3 and at the stock level in Table 6—is consistent with the market reality and recent estimates from the literature. This alignment further supports our interpretation of the economic channel, although identifying the coefficient per se is not our goal; rather, the focus is on demonstrating the predictive power.

The estimated market-level λ_{MKT}, presented in Tables 2 and 3, ranges from 1.2 to 1.8 across various univariate and multivariate settings. This implies that for a one standard deviation increase in the quantity (q) of the market factor, the expected return of a stock with a market beta of 1 increases by 1.2% to 1.8% per month.

Our λ coefficient can be converted to the "price multiplier" in the demand-based asset pricing literature. The λ coefficient is the sensitivity between risk premium (μ) and the quantity (q, constructed in Section 3). In the literature, the price multiplier is typically defined as $(\Delta P/P)/(\Delta Q/Q)$, where $\Delta Q/Q$ is the percentage change in the security's quantity (Gabaix and Koijen, 2022).

To relate the magnitude of our λ coefficient to that of the price multiplier, consider the following conversions. The monthly standard deviation of q is 1.88×10^{-6} in terms of the pre-standardized $\widetilde{q}$ measure (see Table 1). According to Eq. 8, this is roughly 0.22% of (β-aggregated) monthly dollar flow shock as a fraction of the total U.S. stock market capitalization, after adjusting for the market factor's monthly variance: $0.22\% = \sqrt{6} \times 1.88 \times 10^{-6}/(16\%^2 \div 12)$.[9] In terms of the price multiplier measure, let the numerator be the flow shock, $dQ/Q = 0.22\%$, then the price multiplier itself is between $1.2\%/0.22\% = 5.5$ and $1.8\%/0.22\% = 8.2$, which matches in magnitude with the estimate of 5 reported in Gabaix and Koijen (2022).

The magnitude of the q variation is also consistent with market realities.[10] As shown above, a one-standard-deviation increase in market q corresponds to a 0.22% increase in dollar flow as a fraction of the total stock market capitalization. Considering that the mutual fund sector holds about $20 \sim 30\%$ of the total U.S. stock market capitalization (Ben-David, Li, Rossi, and Song, 2022a), this one standard deviation flow shock corresponds to about $0.7\% = 0.22/30 \sim 1.1\% = 0.22/20$ of the mutual fund sector's total AUM, which is a reasonable level.

From the opposite direction, we can also justify the magnitude of our predictability using the literature's price multiplier estimates. Assuming that a one-standard-deviation change in q captures a flow shock of about 1% of the total mutual fund sector's AUM, which multiplies to about $0.2\% = 1\% \times 20\%$ of the total stock market capitalization. With a market-level price multiplier of 5 (Gabaix and Koijen, 2022), this implies a monthly market factor premium variation of about $1\% = 0.2\% \times 5$. Under a one-factor model, this factor premium variation matches the observed total stock return variation, with a monthly standard deviation of 10%, at an $R^2 \approx 1\% = (1\%)^2/(10\%)^2$.

[9] The annualized standard deviation of the market factor return during our sample period is 16%. We multiply by $\sqrt{6}$ because our pre-standardized $\widetilde{q}$ measure in Eq. 9 uses a 6-month lagged average of flow shocks, and we assume these monthly shocks are independent.

[10] Significant noise in q fluctuation is also one of the two requirements for our channel to have a significant empirical impact, as shown in Appendix A.

At the stock level, the coefficients in Table 6 column λ^{stock}, are also comparable to the price multipliers of stocks to flows reported in the literature. This is because we have normalized $\$\text{flow}_{i,t}^{\text{stock}}$ by the stock's own market cap in constructing the $q_{i,t}^{\text{stock}}$ predictors (see Eq. 15). The magnitude of our λ^{stock} stabilizes between 0.5 and 0.8 for larger h values, which is close to the multiplier estimates in the literature, generally ranging from 0.5 to 1 (e.g., Lou, 2012; Chang, Hong, and Liskovich, 2015; Da, Larrain, Sialm, and Tessada, 2018; Hartzmark and Solomon, 2022; Pavlova and Sikorskaya, 2023; Li, Pearson, and Zhang, 2024).

C.6 Additional robustness checks on return predictability

Table A.3 evaluates the robustness of the BTQ model's predictive power across different size and time sub-samples for the Fama-French-Carhart factors. Panel A breaks down the stock-month observations in the OOS evaluation panel into five size groups using concurrent NYSE market capitalization breakpoints. The same OOS predicted returns ($\widehat{r}_{i,t+1}$) are respectively evaluated in each size group. Panel B breaks down the OOS panel by time into three sub-periods: 2010-2014, 2015-2018, and 2019-2022, and reports sub-sample R^2 similarly. Panel C repeats the original joint OOS (2010-2022) evaluation reported in the main text Table 3 Panel B Line "BTQ" for ease of reference.

Table A.3 shows that the BTQ model's predictive power reported in Table 3 is robust in most size and time sub-samples. In particular, the FF3 and FF3C specifications perform even better in large stocks, which are usually the most challenging group for stock return prediction. Across sub-periods, the BTQ model's predictive power is relatively stable over time. The first and the last sub-periods (2010-2014 and 2019-2022) have higher R^2 values than the middle sub-period (2015-2018) across various model specifications. These size and time sub-sample results for the Fama-French-Carhart factors are very similar to those reported for the factors selected from the factor zoo and the selected PC factors in main text Table 4.

Table A.4 extends the main text Table 6 to all h from 1 to 12 and confirms that the

Table A.3: BTQ OOS prediction accuracy (R^2 in %) in size and time sub-samples

evaluation sample	# of obs.	CAPM	FF3	FF3C	FF5	FF5C
		$K = 1$	3	4	5	6
Panel A: size group evaluation						
1 (small)	323,617	0.69	0.72	0.72	0.58	0.64
2	165,059	0.99	1.37	1.44	0.51	0.79
3	141,153	1.16	1.74	1.83	0.42	0.82
4	115,763	0.76	1.97	2.20	-0.33	0.46
5 (big)	103,927	-0.56	1.66	2.00	-1.18	-0.17
Panel B: sub-period evaluation						
2010-2014	321,425	1.10	1.33	1.34	1.03	0.98
2015-2018	255,959	0.17	0.11	0.11	0.07	0.07
2019-2022	272,135	0.90	1.38	1.47	0.37	0.81
Panel C: original OOS evaluation (in Table 3 Panel B)						
OOS (2010-2022)	849,519	0.75	1.03	1.07	0.44	0.65

Note: OOS R^2 evaluated in different size and time sub-samples for the Fama-French-Carhart factors. Panel A breaks down the OOS panel into five size groups according to NYSE market capitalization quintiles and reports the OOS R^2 in each size group. Panel B breaks down the OOS evaluation into three sub-periods: 2010-2014, 2015-2018, and 2019-2022. Panel C repeats the original joint OOS (2010-2022) evaluation reported in the main text Table 3 Panel B Line "BTQ" for ease of reference.

"quantity-only" alternative model does not forecast stock returns.

C.7 Details for macroeconomic variables as alternatives to factor quantity

The second exercise in Section 6 examines the use of macroeconomic variables as alternatives to factor quantity (q) in the BTQ model. This subsection reports the accompanying technical details and additional results.

We preprocess the 126 FRED-MD macro series following standard procedures. We first transform each raw series to a stationary process according to the transformation code provided by FRED-MD. We then standardize these macro variables by demeaning them and dividing by their standard deviations, using the mean and standard deviation estimated

Table A.4: "Quantity-only" alternative model does not forecast stock returns: expanding Table 6 to all h from 1 to 12

	A. constant λ^{stock}				B. λ^{stock} by size quintile		C. λ_i^{stock} by stock	
h	IS R^2(%)	OOS R^2(%)	λ^{stock}	t-stat	IS R^2(%)	OOS R^2(%)	IS R^2(%)	OOS R^2(%)
1	0.000	0.000	-0.10	-0.27	0.003	-0.001	0.47	-232
2	0.000	-0.001	0.05	0.12	0.002	-0.003	0.44	-215
3	0.000	0.000	0.17	0.29	0.003	0.000	0.39	-155
4	0.000	-0.001	0.20	0.33	0.004	-0.002	0.39	-156
5	0.003	0.001	0.51	0.75	0.006	0.002	0.41	-150
6	0.005	0.006	0.75	1.01	0.007	0.006	0.41	-107
7	0.006	0.008	0.82	1.12	0.008	0.008	0.41	-107
8	0.003	0.003	0.66	0.87	0.006	0.005	0.37	-142
9	0.004	0.004	0.75	0.99	0.006	0.006	0.38	-96
10	0.003	0.000	0.69	0.96	0.006	0.003	0.38	-101
11	0.003	-0.003	0.73	0.95	0.006	0.000	0.38	-98
12	0.003	-0.007	0.77	1.01	0.006	-0.004	0.38	-81

Note: Panel A: univariate predictive regression, $r_{i,t+1} = \lambda^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. B: size-dependent predictive regression, $r_{i,t+1} = \lambda^{\text{stock}}_{\text{size\_quintile}(i,t)} q_{i,t}^{\text{stock},h} + error_{i,t+1}$, where $\lambda^{\text{stock}}_{\text{size\_quintile}(i,t)}$ is indexed by the NYSE size quintile of the stock. C: stock-specific predictive regression, $r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$. The R^2 values are expressed as percentages, e.g., 0.005 in row $h = 6$ means 0.005%, a very small value.

during the in-sample period (2000-2009).

Table 8 in the main text reports the comparisons in the single-factor setting by predicting stock returns with predictors in the form of market beta times a macro variable (or linear combinations of the macro variables). In this appendix, Table A.5 expands the analysis to the multi-factor settings. The predictors are $\left[\widehat{\beta}_{i,\text{MKT},t} x_{j,t}, \widehat{\beta}_{i,\text{SMB},t} x_{j,t}, \widehat{\beta}_{i,\text{HML},t} x_{j,t}, \ldots\right]$, where $x_{j,t}$ is the j-th macroeconomic variable, or a full-sample principal component (PC) of the 126 macroeconomic variables. We observe that beta times macro variables in multi-factor settings also do not predict stock returns.

Table 8 Panel D searches for the optimal linear combinations of the 126 FRED-MD macro variables that can substitute for the $q_{\text{mkt},t}$ variables in the single-factor CAPM BTQ model.

Table A.5: "Beta times macro variables", expanding Table 8 to multi-factor settings

	CAPM	FF3	FF3C	FF5	FF5C
A. Benchmark					
BTQ	0.75	1.03	1.07	0.44	0.65
B. Multi-factor beta times each macro-variable					
Mean	-0.63	-1.85	-2.11	-8.69	-8.60
Q10	-1.88	-6.08	-6.81	-19.43	-18.27
Q25	-0.63	-1.67	-2.65	-4.04	-5.92
Q50	-0.12	-0.47	-0.57	-1.25	-1.28
Q75	0.00	-0.08	-0.12	-0.31	-0.38
Q90	0.13	0.02	0.01	-0.07	-0.12
Q95	0.18	0.11	0.12	0.01	0.00
Max	0.28	0.38	0.37	0.41	0.39
C. Multi-factor beta times PCs of macro-variables					
PC1	-0.17	-0.67	-2.82	-20.07	-15.37
PC2	0.01	0.02	-0.06	-0.26	-0.26
PC3	-1.05	-5.93	-5.90	-5.70	-4.43
PC4	-0.09	-1.88	-1.87	-6.41	-9.73
PC5	-0.46	-0.82	-0.81	0.04	-0.24

Note: Predicting stock returns using conditional multi-factor predictors in the form of "market beta times ...", "SMB beta times ...", "HML beta times ...", and so on. All numbers are OOS R^2 in percentage.

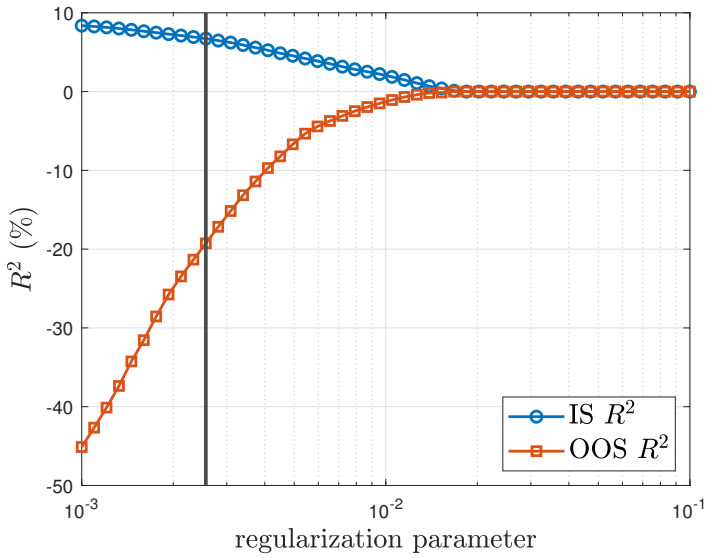
In particular, we minimize the following predictive least squares loss:

$$\min_{\{b_j\}} \sum_{i,t\in \text{IS}} \left(r_{i,t+1} - \widehat{\beta}_{i,\text{mkt},t} \left(\sum_{j\in \text{FRED-MD}} b_j x_{j,t} \right) \right)^2, \tag{A.13}$$

where $x_{j,t}$ is the j-th macro variable at time t, and b_j is the coefficient to be estimated. This search is equivalent to an OLS regression of $r_{i,t+1}$ on 126 predictors $\widehat{\beta}_{i,\text{mkt},t} x_{j,t}$.

This OLS setup drastically overfits with an extremely negative OOS predictive R^2, as shown in Table 8 Panel D, which is expected given the large number of predictors. To address the overfitting issue, we apply regularization to the linear combination search by

Figure A.5: Market beta times the best linear combination of the macro variables, prediction accuracy along regularization path



Note: Model fit and parameter estimates as the regularization parameter (ω, horizontal axis) varies. The IS R^2 is evaluated in the training window (2000-2009), and the OOS R^2 is the same model evaluated in the testing window (2010-2022). The vertical black line indicates the tuned ω based on ten-fold cross-validation.

regressing $r_{i,t+1}$ on $\widehat{\beta}_{i,\text{mkt},t}x_{j,t}$ with LASSO penalty on the coefficients. Note that $x_{j,t}$ has already been demeaned and standardized using the in-sample period estimates. In detail, the Lasso optimization problem is:

$$\min_{\{b_j\}} \frac{1}{2|\text{IS}|} \sum_{i,t\in\text{IS}} \left(r_{i,t+1} - \widehat{\beta}_{i,\text{mkt},t} \left(\sum_{j\in\text{FRED-MD}} b_j x_{j,t} \right) \right)^2 + \omega \sum_{j\in\text{FRED-MD}} |b_j|, \tag{A.14}$$

where ω is the regularization parameter.

Figure A.5 plots both the IS and OOS R^2 values across different regularization parameters. The OOS R^2 values remain below 0 across all ω values and only approach 0 at large ω values, where all coefficients are shrunk to zero, resulting in a model that predicts zero returns for all stock-months. This result suggests the fitted dimension reduction combinations of macro variables do not have predictive power either.