Quantity, Risk, and Return*

Yu An† Yinan Su‡ Chen Wang§

October 19, 2024

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 trading flows, and hence term the model *beta times quantity* (BTQ). The rationale is that sophisticated investors should demand a higher factor premium when they have absorbed noise trading flows of stocks with high loadings 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 Federico Bandi, Hank Bessembinder, Andrew Chen, Zhi Da, Darrell Duffie, Robin Greenwood, Zhiguo He, Ben Hébert, Shiyang Huang, Bryan Kelly, Ralph Koijen, Serhiy Kozak, Jiacui Li, Dong Lou, Paolo Pasquariello, Nagpurnanand Prabhala, Seth Pruitt, Alessandro Rebucci, Nikolai Roussanov, Paul Schultz, Dongho Song, Yang Song, Zhaogang Song, Semih Üslü, Wei Wu, Jun Yu; conference discussant, Raymond Kan; and the participants at the NFA conference 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 with 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 endeavors 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] We show 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 these trades, which causes fluctuations in the quantities of their holdings of the underlying

[1] See papers that report elusive risk-return association in Footnote 4 and those that predict stock returns in Footnote 6.

[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.

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, i.e., they have less demand for that risk. This gives rise 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 (short for "quantity" or variable q throughout the paper), which we term the *beta times quantity* (BTQ) model.

This framework, though still abstracted away from many details of the market microstructure, captures a significant economic force central to risk aversion that, nonetheless, has long been missing in empirical studies of risk and return. The new mechanism considered here is not new to the literature that studies the price impacts of noise trading flows for individual assets, factor portfolios, or asset classes.[3] Our contribution is integrating quantities into the factor pricing framework, which enables smooth upgrades of workhorse methods in cross-sectional asset pricing research. We demonstrate that incorporating quantity information leads to important empirical discoveries in the following 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-$\mathbb{E}r$ relation emerges *conditional on* high levels of market factor q. That is, the risk-aversion implied high-risk-high-return relation holds

[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 (2020) question whether common factor exposure (β) really explains the cross-sectional variation in average returns.

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] Additional results that support this view are obtained with similar SMLs for other factors, and with Fama-MacBeth regressions properly upgraded with 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, and statistically predicting individual stock returns serves as a touchstone for proposed explanations. This task is empirically difficult, and researchers have only recently made significant progress by resorting to unstructured machine learning models designed for forecasting and using a large number of firm characteristics, which inevitably sacrifice interpretability. The state-of-the-art methods can reliably predict stock returns at the monthly horizon, even though the explained variation is small given the low signal-to-noise nature of market prices.[6] We build an economically grounded predictor that interacts stock-level factor exposures (β) with factor-level quantity fluctuations (q). Beta times quantity (BTQ) reliably predicts the panel of monthly individual stock returns with an OOS R^2 of around 1% in various robustness settings, a level comparable to high-dimensional machine learning models. Without quantity, the "β-only" model has almost no predictive power, consistent with previously reported null results (Lopez-Lira and Roussanov, 2020).

Third, quantity information offers a new and better perspective to address the factor zoo problem and provide new factor selection results. The proliferation of proposed factors challenges the asset pricing literature regarding which factors are important for expected returns and fundamental to investors' pricing decisions. The existing tests focus on the *existence* of factor premium: essentially, they ask whether there is a positive spread in

[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 (2014). 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).

expected returns between stocks with high and low factor exposures in the cross section.[7] The new test asks an upgraded question on the quantity-driven *changes* of factor premium: whether the expected return spread *widens* when the sophisticated investors' factor quantity (q) is high (and vice versa). For one, using quantity as an instrument for factor premium provides 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 in various specifications. A few other factors are also selected, including betting-against-beta, volatility, idiosyncratic risk, and value. However, size is dismissed in various settings, challenging its perceived importance as a fundamental risk factor. These results are obtained with a variable selection method (Lasso) that allows for the inclusion of 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 results highlight the importance of incorporating quantity into the factor pricing framework to empirically establish a risk-based explanation of expected returns. We emphasize the joint economic roles of quantity and risk in determining expected returns. To sharpen this argument, we compare the BTQ model with two alternative baseline models that contain only quantity or only risk. The "quantity-only" alternative disregards the factor structure and arbitrage pricing condition, while the "β-only" model represents the

[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.

traditional factor pricing framework that considers risk but not quantity. We find neither alternative explains the cross section of expected returns.

First, the emphasis on risk (and the comparison with the "quantity-only" alternative) 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" used in risk management.[8] For example, if noise investors sell a large quantity of value stocks with high HML loadings, then, from sophisticated investors' perspective, q of HML should increase accordingly. This construction reflects the economic mechanism that investors are averse to systematic risk and the degree of aversion responds to the quantity of systematic risk they hold.

This setup is contrasted with the "quantity-only" model, in which stock-level flows and quantity variations directly affect stocks' expected returns, short-circuiting the factor structure (see Figure 6 for the architecture comparison). This alternative model does not observe the cross-sectional no-(statistical) arbitrage condition, and implies investors are seemingly averse to the physical quantity of stocks rather than the systematic risk they represent. We hardly find any predictive power for stock returns in various implementations of this "quantity-only" model. This comparison highlights risk's role 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 contrast of macro vs. micro elasticities: stocks with comparable risk loadings are close substitutes, while the demand for systematic risks is more inelastic to price (Gabaix and Koijen, 2022; Li and Lin, 2022).

Second, the BTQ model is a direct upgrade from the canonical factor pricing framework,

[8]Stock-level noise trading flows from retail investors are built from mutual fund holding 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 constructing procedure of q.

in which the expected return is determined by a "β-only" baseline. The upgrade 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. As discussed before, this upgrade brings greater identification power as well as economic relevance in selecting the factors fundamental to investors and asset pricing. Future empirical studies can easily subject a newly proposed factor to BTQ factor pricing tests, given that the factor's BTQ term can be easily constructed from factor returns. Similarly, quantity is smoothly integrated into several workhorse methods in asset pricing research, including the security market line (SML), Fama-MacBeth regressions, stock return prediction, and latent factor models. We show significant improvements in empirical performances across these settings. These properties highlight the advantages and broad applicability for future research to incorporate quantity into factor pricing.

Two related frameworks in the literature have differences with our research 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.[9] Instead, we still use previously proposed factors, and the newly proposed factor-level quantity variables work together with risks in the form of "β times quantity".

Second, this paper belongs to the burgeoning literature of demand-based asset pricing, which argues 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, 2023). We focus on the empirical study of the cross-section of expected stock returns, with return prediction accuracy as the central criterion for empirical success.[10] For this purpose, our approach

[9]The approach that treats flow or quantity information as a source of risk is related to De Long, Shleifer, Summers, and Waldmann (1990), Shleifer and Vishny (1997), Adrian, Etula, and Muir (2014), He, Kelly, and Manela (2017), and Dou, Kogan, and Wu (2022).

[10]Koijen and Yogo (2019) demonstrate the mean reversion in latent demand introduces a new source of predictability for the cross-sectional variation in stock returns.

aligns more closely with the traditional factor pricing framework: we explicitly model returns' factor structure; maintain the associated factor pricing condition; and contribute to solving the factor zoo problem, rather than using the factor model as a micro-foundation for the characteristic-based demand system. In terms of the core economic mechanism, we specify that a factor's premium varies with the factor's quantity fluctuations induced by trading flows (q). This mechanism closely guides the construction of the quantity variable (q) and its incorporation into the factor pricing framework in the form of BTQ. 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. In this sense, our channel aligns more closely with the "macro" elasticities emphasized by Gabaix and Koijen (2022) because it is at the factor level.

Some existing papers have reported that trading flow or financial intermediaries' holding quantities are relevant for future returns in various settings. Examples include Teo and Woo (2004), Ben-David, Li, Rossi, and Song (2022a), Kang, Rouwenhorst, and Tang (2022), Li (2022), Li and Lin (2022), and Huang, Song, and Xiang (2024) in stock markets, Greenwood and Vayanos (2014), Vayanos and Vila (2021), Bretscher, Schmid, Sen, and Sharma (2022), and Jansen, Li, and Schmid (2024) in bond markets, Garleanu, Pedersen, and Poteshman (2008) in option markets, and Moskowitz, Ross, Ross, and Vasudevan (2024) for covered-interest parity (CIP) deviations. Unlike these papers, which focus on establishing the pricing effects for individual assets, factor portfolios, or asset classes, our primary emphasis is on integrating quantity into the factor pricing framework to investigate cross-sectional risk-return tradeoffs.[11]

[11] Additionally, 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. However, they do not focus on the asset pricing properties of the selected factors.

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 empirical results for the BTQ model; Section 5 contrasts the BTQ with the alternative "quantity-only" model; Section 6 concludes.

2 Theoretical motivation, empirical model, and methods

2.1 Theoretical motivation

The theoretical reason why quantity information should be integrated into factor pricing is that market trading activities matter for sophisticated investors' risk holdings and, in turn, their required compensation for bearing risks. We focus on a prominent channel where a significant aspect of trading activities, the noise trading flows, matters for the central element of asset pricing, the factor premium, although there can be many other market microstructure mechanisms in which trading activities have price impacts. We outline this theoretical channel below.

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, which can induce significant fluctuations when aggregated to the factor level.[12]

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 ag-

[12]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, they tend to sell growth stocks and in the next, they buy small (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.

gregation 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 (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.[13] 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. Meanwhile, sophisticated investors enforce no-arbitrage pricing across stocks, so the canonical factor pricing condition still holds.[14] 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.

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,

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

[13] Limited risk-bearing capacity can arise from liquidity constraints or misallocation of risk, e.g., Adrian, Etula, and Muir (2014); Gabaix and Maggiori (2015); He, Kelly, and Manela (2017); Kondor and Vayanos (2019); Haddad and Muir (2021); Eisfeldt, Herskovic, and Liu (2024).

[14] This 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.

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 equation is a general specification in which μ_k is an unspecified function of $q_{k,t}$. In most empirical settings, we implement a linear specification as in the second equation.[15] The first 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 the factor's quantity fluctuations.

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) to the extent that having it in the BTQ model even hurts the empirical fit. Therefore, we typically omit the

[15] The linear specification can be microfounded using the standard theoretical framework with mean-variance utility and normally distributed payoffs (see Rostek and Yoon, 2023). In the upgraded Fama-MacBeth regressions of Section 4.2, we implement a non-parametric estimation of $\mu_k(\,\cdot\,)$. See Section 2.3 for an overview of the various parametric and non-parametric empirical methods.

term in parentheses and only include the BTQ 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 q variation tracks sophisticated investors' holding condition. Hence, we are no longer inferring investors' risk pricing process from asset and asset price information alone. Therefore, the new framework can lead 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.[16]

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

[16]The model specifies that that factor k's premium $\mu_{k,t}$ is affected only by its own quantity $q_{k,t}$, not by the quantities $q_{j,t}$ of other factors (in Eq. 3). Theoretically, this specification is justifiable when the factor risks are non-redundant (which is particularly guaranteed for the selected factors and the orthogonalized PC factors). Empirically, allowing for cross-factor interactions would complicate the model, increasing the number of parameters from K to K^2, which becomes particularly challenging when K is large.

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.

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 return. 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 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 averaging the returns of the Fama-MacBeth factor mimicking portfolio (FMP, which are coefficients of 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 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 setup of the stock return panel and evaluate our model with the same metric of empirical success, the out-of-sample (OOS) prediction fit (R^2) besides the in-sample (IS) R^2.

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 explain the most time-series return variation of the many candidate factors, which according to existing literature, can be more reliable candidate factors for explaining expected returns.

In summary, we put forward the message that integrating quantitative information into various empirical methods can lead to significant empirical discoveries. We implement the methods outlined above to support this message, although the methods here are far from exhaustive, given the vast asset pricing literature. We believe the quantity variables can similarly interact with many other existing methods and lead to a broad avenue of potential 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 $r_{i,t+1}$ and a panel of $\beta_{i,k,t}$ and a time series of $q_{k,t}$, for each factor k, for the right-hand side predictors. Among them, $\beta_{i,k,t}$ 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, with which one can calculate both β and q of 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 with 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

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

where $\widehat{\text{cov}}_t$ and $\widehat{\text{var}}_t$ are realized covariance and variance estimated with daily returns in a 12-month rolling window up to month t.[17]

We construct the stock-level dollar flow $flow_{i,t}^{\text{stock}}$ 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.[18] 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 A.1.

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

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

[17] Notice $\widehat{\beta}_{i,k,t}$ corresponds to the regression coefficient of a single-factor model. This is different 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, 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.

[18] The mutual fund industry witnessed significant growth and sustained inflows throughout the 1990s (Lou, 2012; Ben-David, Li, Rossi, and Song, 2022a). In the post-2000 era, the monthly flows of mutual funds have generally maintained relative stability, prompting us to start our sample period from 2000, aligning with Gabaix and Koijen (2022).

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,

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

We use the negative sign to switch the perspective from retail investors to sophisticated investors in accounting the flow. In particular, a positive $flow^{\text{stock}}_{i,t}$ dollar number indicates that retail investors are selling stock i in month t, and sophisticated investors are buying. Moreover, we use the two-quarter-lagged mutual fund holding weight, denoted as $weight^{\text{fund}}_{i,m,\text{quarter}(t)-2}$. For instance, quarter(July) $-$ 2 = Q1.[19]

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 in Section 2.1 and has 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:

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

[19] 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 $flow^{\text{stock}}_{i,t}$ 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.

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, the sophisticated investors add a marginal portfolio to their existing holdings in response to retail flows, and $flow_{i,t}^{\text{stock}}$ is the dollar weights of this portfolio. The portfolio's risk characteristics are determined by its composition (portfolio weights $flow_{i,t}^{\text{stock}}$), 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.[20] In this sense, we are indeed tracking the quantity of factor *risk*, not the physical quantity of securities or portfolios.

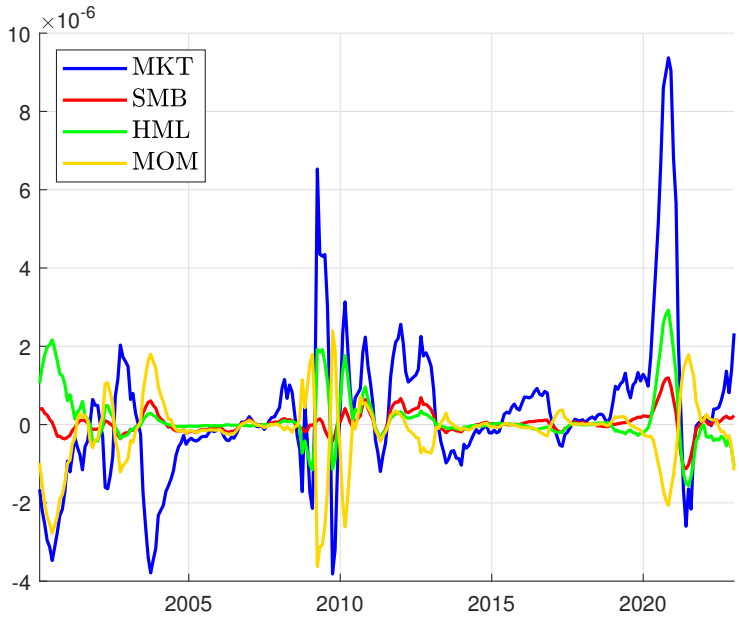
Second, these flow shocks are normalized by the lagged total US stock market capitalization and accumulated in a six-month lookback window,

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

The normalization accounts for the upward trend in dollar flows that aligns with the overall growth of the equity market as well as the growing capacities of sophisticated investors to absorb these flows. Accumulating $flow_{k,t}^{\text{factor}}$ 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, $flow_{k,t}^{\text{factor}}, flow_{k,t-1}^{\text{factor}}, flow_{k,t-2}^{\text{factor}} \ldots$ 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).

[20] 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 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.

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.

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.

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 central 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.[21] 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

[21]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.1). 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.

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-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 B.1 reports 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 series have a multi-factor structure with independent variation along various dimensions and substantial idiosyncratic variation. This result suggests each long-short factor's quantity series offers valuable pricing information beyond that of the market. It also means the results of BTQ's predictive power further below attained with different factors are not mechanically driven by repeated q variations and suggests 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 economic 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 Empirical results

4.1 Security market line (SML) depends on quantity

The security market line is a simple and familiar tool to visualize the relationship between systematic risk exposure and expected return (β-$\mathbb{E}r$) in the cross section of stocks without resorting to parametric modeling. We construct the empirical SML and the upgraded versions conditional on factor q. We show 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 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: $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$, and the purpose is to show 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 uncondi-

tional 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 $\widehat{\beta}_{i,k,t}$.[22]

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

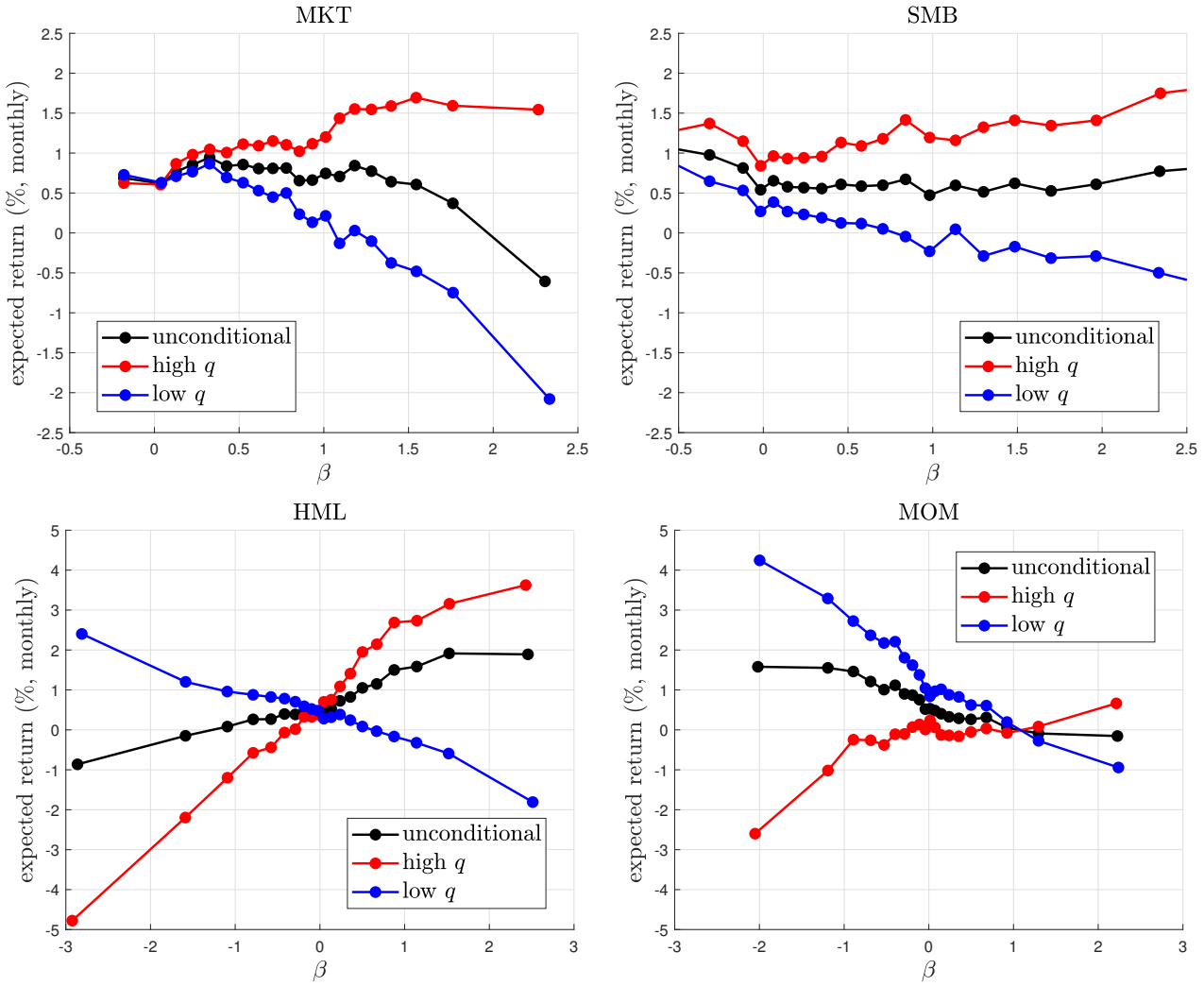
We find that the unconditional SML is nearly flat for the market factor, with a slight downward slope in the higher beta range. This implies that the market beta *alone* cannot explain the cross-sectional variation in expected returns, which is consistent with similar reports in the existing literature. 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 two conditional SMLs have distinct slopes, and the unconditional SML (black) lies in between them as the mixed average. The positive high-q slope means the cross-sectional risk-return tradeoff is strong and positive, suggesting that sophisticated investors demand higher additional compensation for bearing high systematic risk investments in high-q environments. The negative low-q slope indicates a negative risk-return tradeoff, suggesting that 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.[23] The gaps in the slopes suggest that sophisticated investors' risk-holding conditions matter for their demand for risk, which 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, suggesting the quantity's effects

[22]Formally, an unconditional bin is defined as $\{(i,t) \text{ s.t. } \widehat{\beta}_{i,k,t} \in [a,b)\}$, 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 $\{(i,t) \text{ s.t. } \widehat{\beta}_{i,k,t} \in [a,b) \text{ and } q_{k,t} \geq \text{median}(q_{k,t})\}$, 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 "$<$".

[23]The negative low-q slope is puzzling in the sense that it suggests a risk preference (rather than aversion) in low-q months. The frictions regarding sophisticated investors' risk management as described in Frazzini and Pedersen (2014) can be a potential explanation, which is an interesting direction for future research.

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 x- and y-axis scales are two times larger in the bottom two panels than in the top two to accommodate the greater ranges of HML and MOM β's.

on factor premiums is general and the underlying economic mechanism is robust.

The magnitude of q's effects is economically massive. For instance, a market beta-neutral stock has an expected return of around 0.75% per month unconditional on q. 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 0.25% in low-q months, with the average being still around 0.75%. The high v.s. low-q gap is around 1% per month or more than 10% annualized. Such a gap is even greater for higher

market β stocks. For HML, the gap for a $\beta_{\text{HML}} = 1$ stock is around 30% annualized, while the HML-neutral stock's expected return does not depend on q, shown 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 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.[24]

All the SMLs 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), although the slope (risk premium) can vary significantly with the q condition. Next, we specify the linearity of expected returns in β, while still leaving the effect of q non-parametric in an upgraded Fama-MacBeth regression framework.

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:

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

[24] It is also interesting to note that the crossings of the high/low-q and unconditional SMLs are at around $\beta = 0$ for MKT and HML, but not for SMB and MOM. Crossing at $\beta = 0$ is consistent with the parametric BTQ model and simpler to understand with the theoretical motivation: the expected return of a factor risk-neutral stock should not be affected by that factor's quantity fluctuations. Not crossing at $\beta = 0$ warrants further investigation.

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. It is often cited as evidence against factor pricing because the unconditional Fama-MacBeth factor premium is close to zero (Lopez-Lira and Roussanov, 2020).

The innovation is that we estimate the mean of $\gamma_{k,t+1}$ conditional on $q_{k,t}$. To this end, we form four unit bins of $q_{k,t}$ (which is already standardized) and calculate the average of $\gamma_{k,t+1}$ in each bin. Figure 3 presents the conditional (solid lines) and the unconditional (dashed lines) factor premiums for the four single-factor specifications.

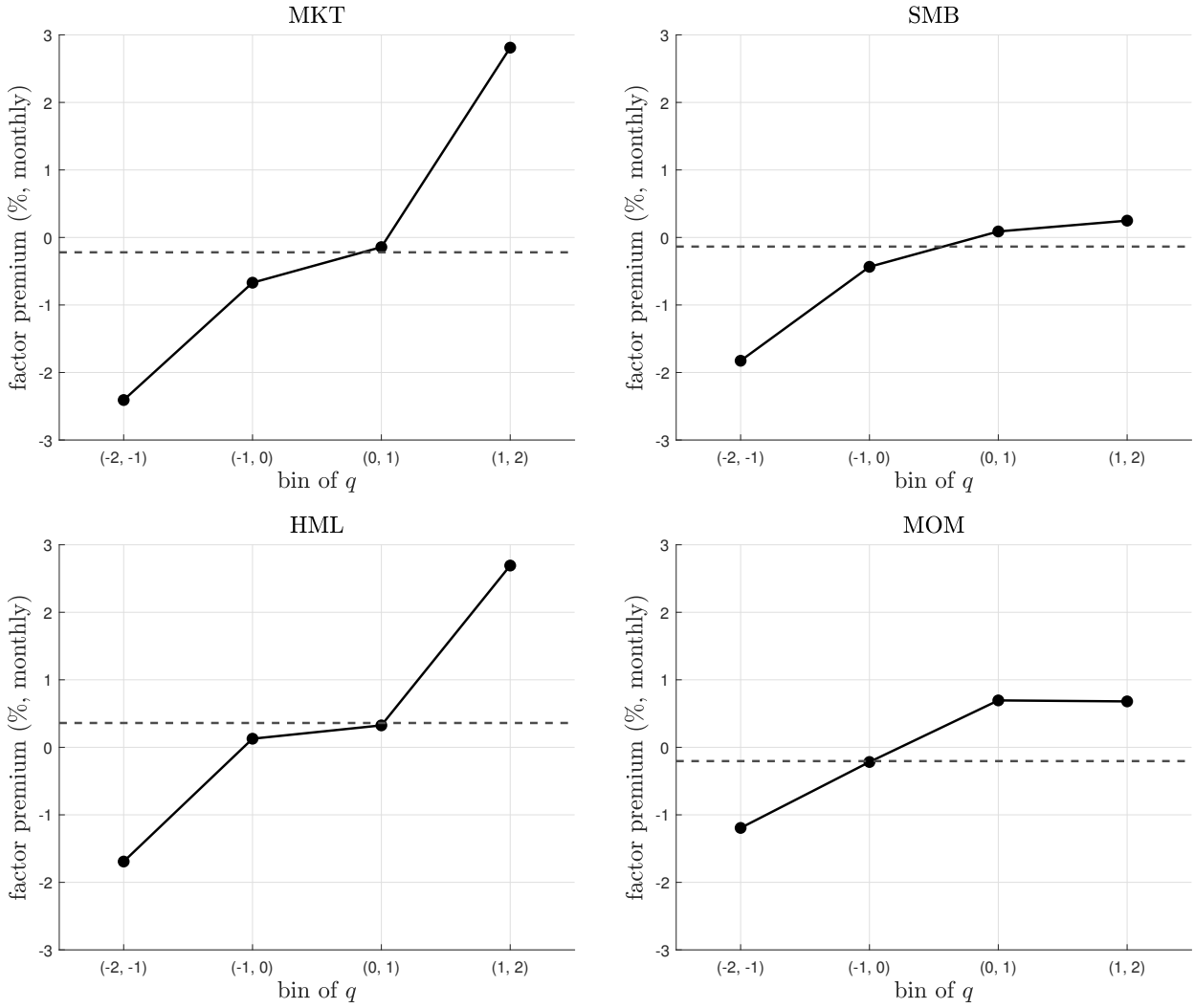
The plot shows strong and consistent evidence that the Fama-MacBeth factor premium is not zero but increasing in factor quantity $q_{k,t}$. Specifically, the cross-sectional risk-return relationship is strong and positive when quantity $q_{k,t}$ is high. And the factor premium is negative when $q_{k,t}$ 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 reflects only a small part of the interesting big picture that unfolds only when we condition on quantity.

The increasing relationship in $\mu_k(q_{k,t})$ is consistent across the four factors, although the market factor exhibits 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 percentage points.

4.3 Beta times quantity (BTQ) forecasts individual stock returns

The empirical results so far with non-parametric plots show the quantity information significantly affects the cross-sectional risk-return relationship. Next, we 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 the BTQ

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).

model provides a compelling explanation for the expected return of individual stocks.

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

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

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, and 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 ($K = 1$) with $k =$ 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.[25] 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 ones with a high model fit in 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 have an OOS R^2 above 0. This level of predictability is economically significant and comparable to unstructured machine learning models that use a large number of firm characteristics to predict stock returns. The state-of-the-art machine learning models typically achieve an OOS R^2 of around 1% to 2% in the literature. In contrast, the β-only models have a low R^2 around 0, and 50 out of the 153 JKP factors have a negative OOS R^2.

Turning to the coefficients estimates, the BTQ model's λ_k are significantly positive for all four Fama-French-Carhart factors and most of the 153 JKP factors. The economic magnitude of the λ_k estimates is substantial. For example, $\lambda_{\text{MKT}} = 1.8\%$, meaning 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, so on and so forth. In contrast, the β-only model's μ_k coefficients are mostly statistically insignificant from zero, and 90 out of the 153 JKP factors even have negative coefficient point estimates.

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

In addition, Appendix Table A.1 presents an incidental empirical finding: 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. Additionally, 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

[25] 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.

the BTQ model's cross-sectional return predictability, it is a much weaker result to argue for quantity's pricing power and peripheral to our research focus (more discussion in Appendix B.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-Five-Carhart factors. The BTQ model still significantly outperforms the β-only model in all multi-factor specifications. Allowing multiple factors further boosts BTQ's predictive accuracy with the best OOS R^2 values reaching above 1%. In contrast, the β-only model still barely predicts stock returns with low R^2 values even in sample.

In terms of factor importance, controlling for other factors' contributions, MKT is the most prominent with the highest and most statistically significant coefficients across all multi-factor models, even though λ_{MKT} attenuates when more factors are included. HML and MOM also have positive coefficients but are statistically insignificant. When these factors are added to the model, both IS and OOS R^2 increase, indicating that their BTQ terms provide additional predictive power, and that they are priced factors. SMB, CMA, and RMW's coefficients are near zero or negative, indicating they are not priced factors according to the BTQ model. This is also consistent with the fact that the OOS R^2 drops when adding these factors to the model. The β-only model's μ coefficients are all insignificant from zero or negative. (These numbers are relegated to Appendix Table A.2.)

Comparing BTQ's IS v.s. OOS model fits, we see slight reductions in R^2 when moving from IS to OOS for CAPM, FF3, and FF3C. This indicates mild overfitting or parameter instability issues. It 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. When moving to FF5 and FF5C, the IS R^2 keeps increasing 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

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.

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 it, 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. Therefore, 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:

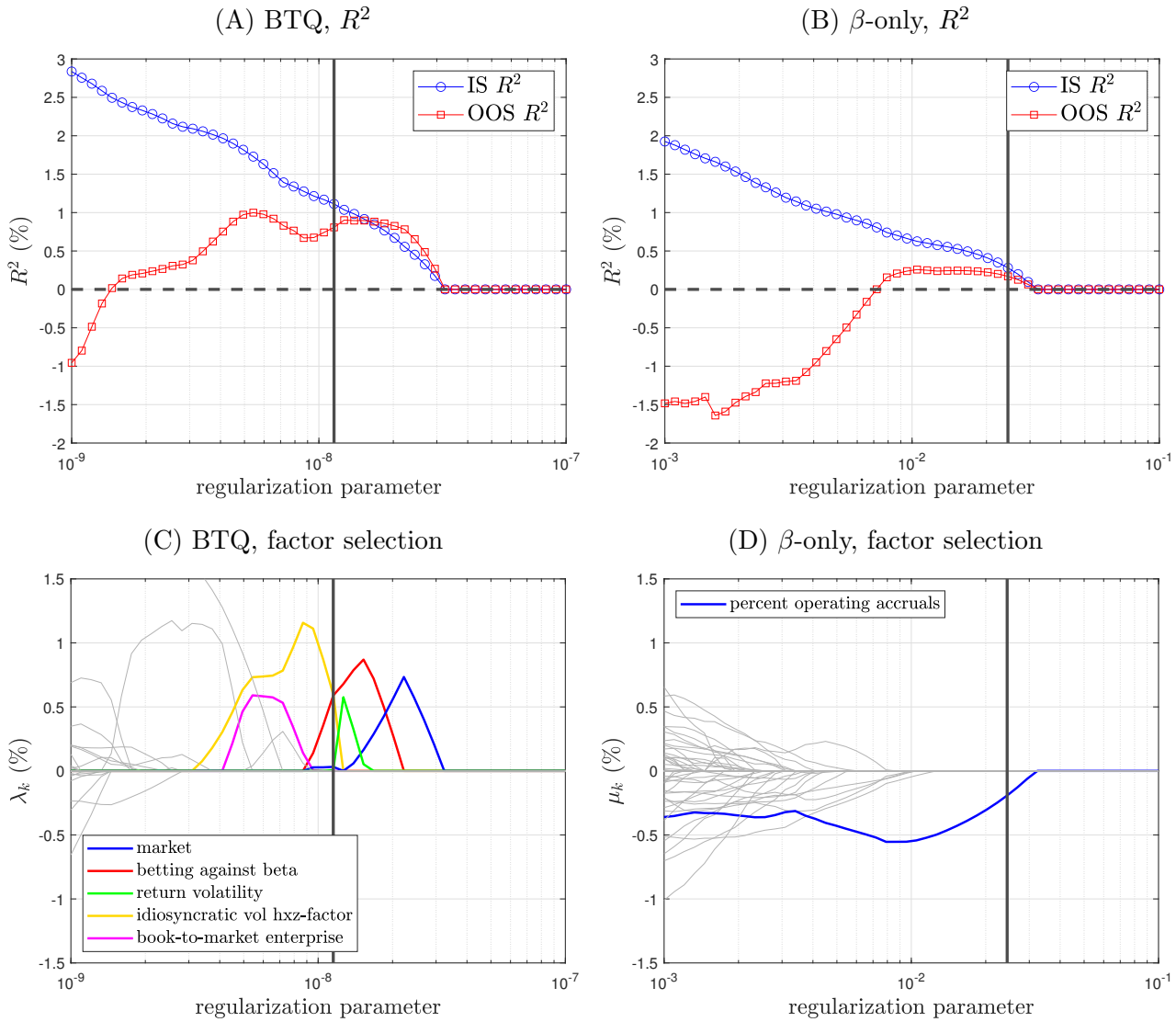
$$\min_{\lambda_1 \dots \lambda_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \lambda_k \widehat{\beta}_{i,k,t} q_{k,t} \right)^2 + \omega \sum_{k=1}^{K} \frac{1}{\sigma(\widetilde{q}_{k,t})} |\lambda_k| , \tag{13}$$

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.[26]

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 A.2.) As ω increases, the fitted BTQ model 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 is the expected behavior of the Lasso method. 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%.

[26] The penalty on λ_k is normalized by the standard deviation of $\widetilde{q}_{k,t}$, so that the regularization is "fair" across different factors whose quantity $\widetilde{q}_{k,t}$ have different scales of fluctuations before standardization ($q_{k,t} = \widetilde{q}_{k,t}/\sigma(\widetilde{q}_{k,t})$). See technical details in Appendix A.2.

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 B.4 with factor importance measures. The vertical black line indicates the tuned ω based on cross-validation; see Appendix A.2 for details.

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 use 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 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 theory such as Merton's (1973) ICAPM model, and is historically the most important factor in workhorse empirical models such as the CAPM and Fama French. However, 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 is still very effective for explaining expected stock returns. However, this conclusion cannot be achieved 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. On the other hand, SMB and other factors related to size are dismissed by the Lasso selection. The unselected factors are in gray, reported in Appendix B.4 with factor importance measures.

Moreover, notice that the λ estimates of these selected factors from the BTQ model are all positive, which is consistent with the risk-based explanation as discussed in Section 2.2.

The β-only model only selects one factor, percent operating accruals (Panel D). It has a negative coefficient, which is inconsistent with the risk-based explanation. We believe this is not a reliable result in a misspecified model, as indicated by β-only's 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 overcome the problem, we provide the tuned ω using only IS information via ten-fold cross-validation, shown as the vertical black line (see technical details in Appendix A.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.[27] We demonstrate 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.[28] Then, we construct $\widehat{\beta}$ and, in turn, quantity q with respect to 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.

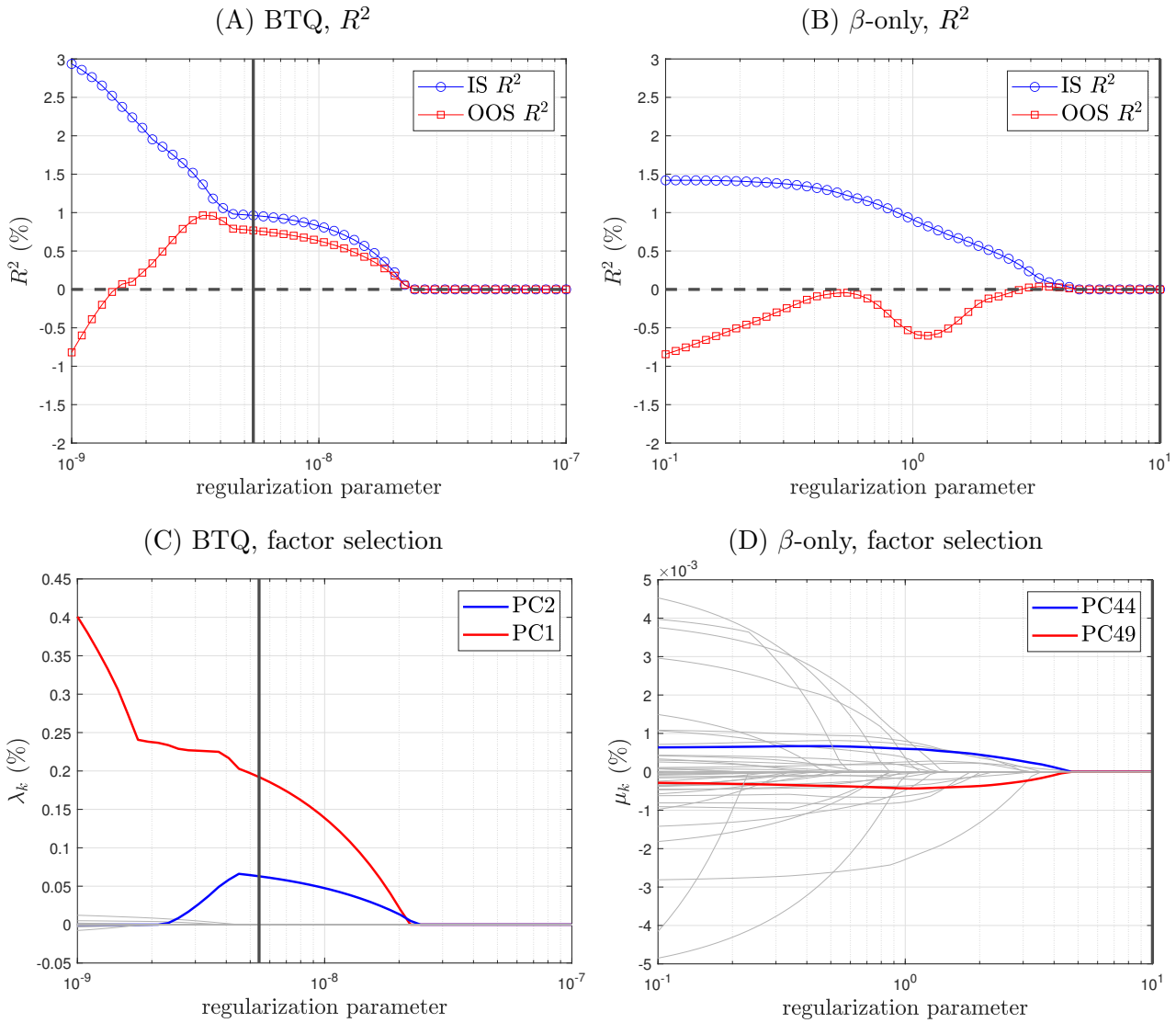
Figure 5 shows that the BTQ model with PC factors has strong predictive power for stock returns, with the OOS R^2 peaking at around 1.0%, similar to the previous Figure 4 with the original factors. The high OOS R^2 is, once again, robust to the choice of ω, shown with the broad peak of the OOS R^2 hump-shaped curve. In contrast, the β-only model with PC factors hardly delivers any predictive power, with the OOS R^2 less than 0 for almost all ω values.

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

[27] See, e.g., Kelly, Pruitt, and Su (2019), Kozak, Nagel, and Santosh (2020), and Lettau and Pelger (2020).

[28] 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 models 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 A.2 for tuning details.

selected as the most important factors for predicting stock returns. The magnitude of their λ estimates dominates the subsequent PC factors (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).

Notice, once again, the signs of the λ estimates for the selected factors, PC1 and PC2, are both positive. This is required by the risk-based theory no matter how the signs of the PCs are specified. In contrast, the β-only model's selection and parameter estimates do not show a discernible pattern, which we believe are mostly estimation noise given that the β-only model is misspecified.

4.6 Robustness

This section reports robustness checks that validate the predictive power of the BTQ model 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.

First, we 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).[29] Appendix B.5 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 is usually the most challenging section for stock return prediction. Characteristics-

[29]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.

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.[30]

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.[31]

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

[31] 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

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).

Second, 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 $h = 5$ 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 of the speed and persistence with which

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.

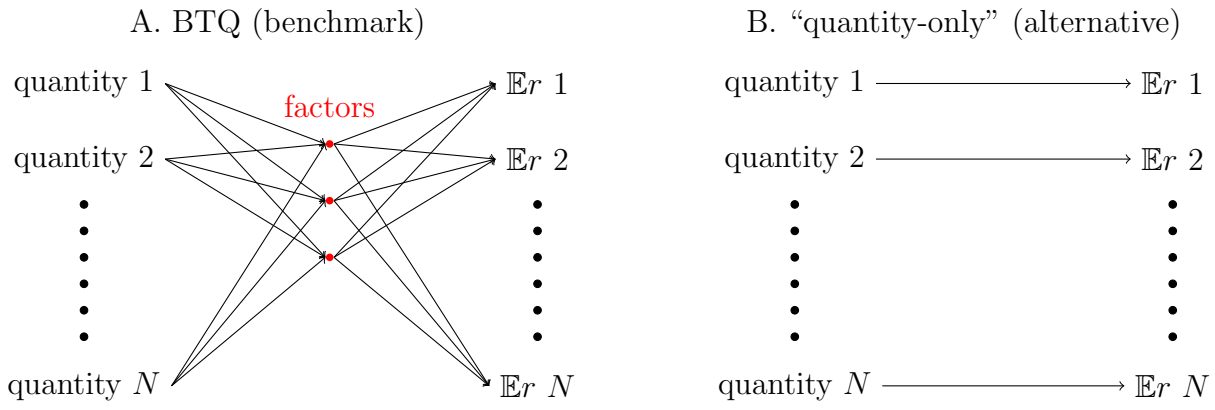
factor flows influence sophisticated investors' pricing of factor risks. Flow shocks likely have an immediate impact on the factor premium next month, given that $h = 1$ already has some predictive power. The R^2 is higher with an intermediate window ($h = 5$, 6, or 7), suggesting the lagged flows in 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-only" models do not explain expected returns either

This paper emphasizes a *risk-based* explanation of expected stock returns that incorporates quantity information. Is the risk modeling essential, or can quantity information alone explain expected stock returns? We now examine an alternative economic model in which stock-level flow and quantity variations directly affect the expected stock returns without considering the risk factors and the arbitrage pricing condition. This exercise is important for understanding the joint economic roles of quantity and risk in asset pricing. So far, the analyses above have focused on comparing benchmark model BTQ against the "β-only" baseline that accounts for risk but not quantity. This exercise introduces an alternative baseline: solely using quantity without considering risk also falls short by far 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 premium, and then feed back to stock-level expected returns. The "quantity-only" alternative model specifies that stock-level flows and quantity variations *directly* affect expected stock returns, short-circuiting the factor premium

Figure 6: Comparison of predictive architectures of the two models



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

mechanism (see the comparison in Figure 6). Specifically, the alternative model is

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

where $q_{i,t}^{\text{stock}}$ is a stock-specific flow or quantity measure, and λ^{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. μ_i^{stock} is specified as 0 and omitted since we focus on the dynamic effects of quantity.)

This "quantity-only" model implies a very different economic mechanism, although (14) is similar in form to the main model's factor premium specification (3). In the main model, factor premiums dynamically vary, 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. Suppose two stocks have the same risk exposure but have received different noise flow shocks, the alternative model would imply an immediate cross-sectional arbitrage opportunity. It implies an inability of cross-sectional substitution, such that each stock is independently priced regardless of their factor exposures. This might be the case if rigid frictions prevent cross-sectional arbitrage, or if the stocks' idiosyncratic risks are not diversifiable and treated as individually priced risks.

Another way to look at the specification (14) is a quantity-driven *alpha* model when viewed in conjunction with the BTQ model: $\mathbb{E}_t r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock}} + \sum_k \lambda_k \beta_{i,k,t} q_{k,t}$. In this view, the term $\lambda_i^{\text{stock}} q_{i,t}^{\text{stock}}$ captures dynamic *alpha*, the part of quantity-driven expected stock return that is not explained by the risk channel (BTQ, or the second term).

To empirically implement the alternative model, we experiment with various specifications of (14) and find none of them even come close to the BTQ model's explanatory power for expected stock returns. Specifically, we normalize the dollar stock-level mutual fund flow ($flow_{i,t}^{\text{stock}}$, see Eq. 7) by the stock's one-month-lagged market capitalization, so that the 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:

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

The construction is similar to that of the factor-level quantity variables in Eq. 9 but at the stock level.

We explore the predictive power of $q_{i,t}^{\text{stock},h}$ for stock returns, and compare it with the BTQ model's predictive power. We experiment with different specifications of the sensitivity coefficient λ^{stock}, varying the degrees of parameter freedom. First, we specify it as a constant for all stocks: $r_{i,t+1} = \lambda^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$ (results reported in Table 6 Panel A). Second, we allow a size-dependent sensitivity coefficient such that λ^{stock} is indexed by the NYSE size quintile of the stock: $r_{i,t+1} = \lambda_{\text{size\_quintile}(i,t)}^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$ (in Panel B). Lastly, we allow stock-specific λ_i^{stock}, which is the most flexible specification: $r_{i,t+1} = \lambda_i^{\text{stock}} q_{i,t}^{\text{stock},h} + error_{i,t+1}$ (in Panel C).[32]

Consistent across all specifications, the "quantity-only" models are very weak and un-

[32]The second specification (size-dependent λ^{stock}) effectively runs five separate univariate predictive regressions, one for each size bin. The third specification (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.

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
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 6 means 0.005%, a very small value.

reliable in predicting stock returns. In Table 6 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 equally small, but also negative in some specifications of the lookback length (h). Allowing size-dependent λ^{stock} brings a very slight improvement in these predictive power evaluations, but no qualitative changes (Panel B). The low R^2 values suggest that these two specifications are too restrictive, and 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 v.s. one or five parameters). The in-sample R^2 mechanically increases but is still smaller than the BTQ model's. 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 constant predictive coefficient λ^{stock} estimates are mostly positive, but none is statistically significant. The positive sign is consistent with our expectations based on the existing literature. Recall that a positive $flow_{i,t}^{\text{stock}}$ means net outflow from noise traders, so the positive sign of λ^{stock} indicates a positive expected return response after sophisticated investors absorb flows (or, equivalently, a negative concurrent price impact). Since we have normalized $flow_{i,t}^{\text{stock}}$ in constructing the $q_{i,t}^{\text{stock}}$ predictors, the λ^{stock} coefficient is comparable to the (inverse) demand elasticities of stocks to flows, typically defined as $(\Delta P/P)/(\Delta Q/Q)$, where $\Delta Q/Q$ is the percentage change in the security's quantity. The magnitude of λ^{stock} stabilizes between 0.5 and 0.8 for larger h values, which is close to the elasticity estimates in the literature.[33] Nonetheless, the λ^{stock} estimates are not statistically significant, and the predictive power is too weak to offer a meaningful explanation of stock returns.

The poor performance of this alternative model underscores the predictive power of the BTQ model, which is greater by orders of magnitude and much more robust to specification perturbations. It shows the empirical success of the BTQ model is not driven by quantity per se, once again highlighting the paper's core argument that quantity and risk should work together to explain expected stock returns. In particular, the comparison implies that the factor structure is still essential in modeling expected stock returns. This 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 noise traders (Kozak, Nagel, and Santosh, 2018).

The different performance of BTQ vs. "quantity-only" is also consistent with the contrast of macro vs. micro elasticities in Gabaix and Koijen (2022). At the stock level, securities are highly elastic substitutes, especially if their factor loadings are similar. The quantity's effect on price is more salient at the factor level, where the demand is more inelastic.

From a statistical/machine learning perspective, we can view both the BTQ and the

[33]The literature typically runs a concurrent regression of price on quantity using appropriate demand instruments, whereas we run a predictive regression. Despite the many differences, our estimated λ^{stock} is comparable in magnitude to the literature's micro-elasticity around 1 to 2 (e.g., Da, Larrain, Sialm, and Tessada, 2018; Hartzmark and Solomon, 2022; Li, Pearson, and Zhang, 2024).

"quantity-only" alternative as prediction models of the cross section of 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 conducts a dimension reduction on the predictors and, in turn, uses the reduced predictors to predict the cross section. This is an encoder-decoder architecture in machine learning terms, where the low-dimensional "code" is the factor-level quantities (see Figure 6 Panel A for encoder-decoder illustration). In this 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. On the other hand, the "quantity-only" model is limited by the noisy predictor inputs at the stock level. See Gu, Kelly, and Xiu (2021) and Kelly, Malamud, and Pedersen (2023) for applications of encoder-decoder structures in asset pricing. However, BTQ specifies both the encoding and decoding weights as the factor risk exposure (β) according to the economic theory, rather than estimating them solely based on statistics.

6 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 in three main aspects of new findings. First, quantity variation elicits risk-return tradeoff relationships, which have been hard to capture with β only and cast doubt on whether risk is the main driver of 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.

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 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.

This paper fits into the general agenda of using market trading and investor holding quantity data to better model investors' *demand* for risk and the resulting implication for asset prices. We provide a simple and actionable way to incorporate quantity into workhorse asset pricing models, including the security market line, Fama-MacBeth regressions, predicting stock returns, and factor selection. 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 quantity variables. A concurrent paper (Gabaix, Koijen, Richmond, and Yogo, 2023) is highly relevant to this purpose; the method proposed therein can help avoid a potential "quantity zoo" problem.

References

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

Barber, Brad M, Xing Huang, and Terrance Odean, 2016, Which factors matter to investors? evidence from mutual fund flows, *The 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.

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, *Unpublished working paper* .

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, 2022, Institutional corporate bond pricing, Working paper, USC.

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

Choi, Darwin, Wenxi Jiang, and Chao Zhang, 2023, Alpha go everywhere: Machine learning and international stock returns, *Available at SSRN 3489679* .

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, *The 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.

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.

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, *The 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, *arXiv preprint arXiv:1804.09314* .

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, *The 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, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, 2023, Asset embeddings, *Available at SSRN 4507511* .

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

Garleanu, Nicolae, Lasse Heje Pedersen, and Allen M Poteshman, 2008, Demand-based option pricing, *The Review of Financial Studies* 22, 4259–4299.

Giglio, Stefano, Yuan Liao, and Dacheng Xiu, 2021, Thousands of alpha tests, *The 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, *Available at SSRN 4802345* .

Greenwood, Robin, and Dimitri Vayanos, 2014, Bond supply and excess bond returns, *The Review of Financial Studies* 27, 663–713.

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

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

Haddad, Valentin, and Tyler Muir, 2021, Do intermediaries matter for aggregate asset prices? *The 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.

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, *The Journal of Finance* 71, 2095–2144.

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

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

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

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, *Swiss Finance Institute Research Paper* .

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

Kang, Wenjin, K Geert Rouwenhorst, and Ke Tang, 2022, Crowding and factor returns, Working paper, Yale University.

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, *The Journal of Finance* 79, 459–503.

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross-section of present values, *The 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.

Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo, 2023, Which investors matter for equity valuations and expected returns? *Review of Economic Studies* Forthcoming.

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.

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.

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

Lewellen, Jonathan, 2014, The cross section of expected stock returns, *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper* .

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.

Lopez-Lira, Alejandro, and Nikolai L Roussanov, 2020, Do common factors really explain the cross-section of stock returns?, *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper* .

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

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

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

Moskowitz, Tobias J, Chase P Ross, Sharon Y Ross, and Kaushik Vasudevan, 2024, Quantities and covered-interest parity, *Available at SSRN 4820243* .

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.

Rostek, Marzena J, and Ji Hee Yoon, 2023, Imperfect competition in financial markets: Recent developments, Working paper, University of Wisconsin - Madison.

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

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, *The Review of Financial Studies* 21, 1455–1508.

Appendix

A Technical details

A.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.

A.1.1 Return cleaning

We first correct data errors in monthly 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.[34] 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 negative returns before a fund's closure. Thus, we manually check only the subsample of

[34] 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 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.

A.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.[35] 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.[36] Morningstar's TNA data (`Net_Assets_ShareClass_Monthly`) suffer to a lesser extent from these issues than

[35] 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.

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

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.[37] 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.

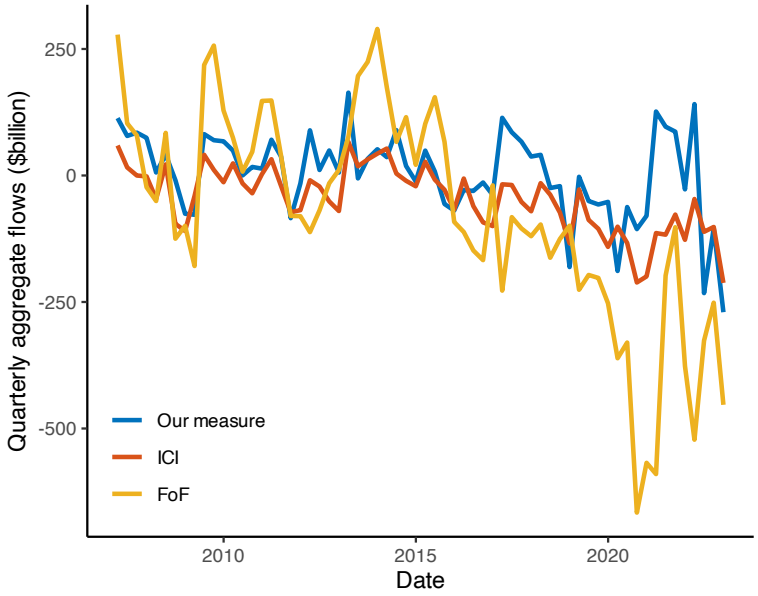
A.1.3 Cross-validating the data-cleaning procedure

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

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

Figure A.1: 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.

Figure A.1 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.1 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.[38] 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.[39] In comparison, our mutual fund sample contains U.S. mutual funds covered by CRSP, which collects historical data from various sources.[40] Due to the nature of the data collection process, CRSP's

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

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

[40]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,

coverage is smaller than ICI's coverage.

A.2 Technical details of Lasso implementation

In optimization (13), adding "$1/\sigma(\widetilde{q}_{k,t})$" 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) is equivalent to running the vanilla Lasso on the pre-standardized BTQ $(\widehat{\beta} \times \widetilde{q})$

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

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. A factor with greater variation in $\widetilde{q}_{k,t}$ will have an inflated λ_k after standardizing to $q_{k,t}$, but we do not want to penalize it more for that reason. Standard Lasso implementation where the economic interpretation is not a concern would recommend standardizing the predictor (BTQ together) across the $\{i, t\}$ panel. We are effectively creating a customized standardization based on the required economic interpretation.

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

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

We perform ten-fold cross-validation to tune hyperparameter ω based on only in-sample information (from 2000 to 2009). For each fold, we exclude one year of observations and solve the Lasso problem (A.1) using the remaining nine years of in-sample data. The model is evaluated in the left-out year to form predicted returns $\widehat{r}_{i,t+1}^{[\text{cv}]}$. After enumerating all folds and forming predicted returns for all in-sample observations, we calculate the cross-validated

and the Wiesenberger Investment Companies Annual Volumes.

(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.

B Additional empirical results

B.1 Additional properties of the $q_{k,t}$ time series variables

Figure A.2 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. It shows there is significant comovement (both positive and negative) among the four variables, which is also evident in the time series plot (Figure 1 in the main text). HML-MOM has the greatest (in absolute value) correlation of -0.75. All pairwise correlations are far from ± 1, indicating that the $q_{k,t}$ variables are far from collinear, and that they each capture unique information about the underlying quantity variations.

Panel B shows the similar situation of limited comovement for 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 shows that there is a factor structure among the 153 factor-level $q_{k,t}$ variables, but there is significant unique information across the different dimensions of the $q_{k,t}$ variables. 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 the these in-sample PC statistics are likely exaggerated due to overfitting.

In summary, the $q_{k,t}$ variables are not collinear, and they each capture unique information about the underlying quantity variations. It 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 predic-

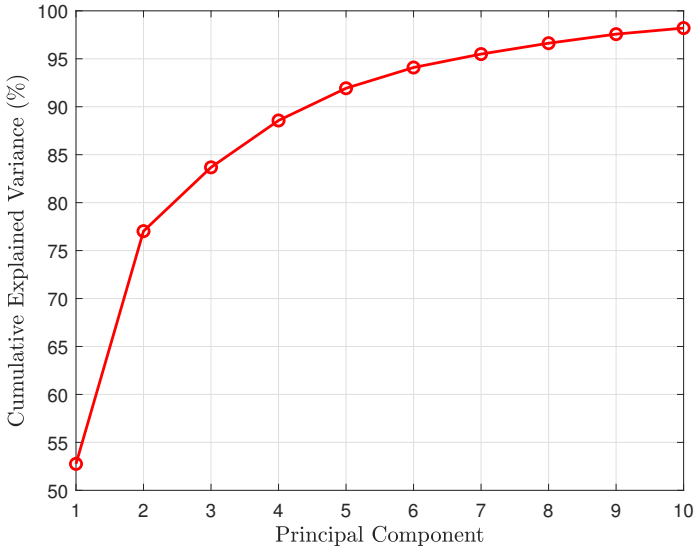
tive power is consistent across various factor specifications speaks to the robustness of the underlying economic mechanism.

Figure A.2: 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: cummulative explained variances by principal components of $q_{k,t}$ series of the 153 JKP factors.

B.2 Predicting factor returns with factor quantity

This appendix subsection presents the results of predicting factor returns with factor quantity. We notice the time-series predictability is not the main focus of this paper, although it is implied by a part of the paper's theoretical motivation. In particular, it is related to the factor premium modeling in Eq. 3. Empirically, we successfully detect the predictability to a certain extent, consistent with the theoretical motivation. However, we 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 is predominantly positive and statistically significant for all Fama-French-Carhart factors and most JKP factors. This indicates that each factor's expected return is positively related to its quantity, consistent with the theoretical motivation. The full-sample R^2 values are around 5%, which is relatively high for factor return

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 $R^2 = 1 - \sum_t (f_{k,t+1} - \widehat{f}_{k,t+1})^2 / \sum_t f_{k,t+1}^2$.

prediction (see Welch and Goyal, 2008)

However, the OOS R^2 values are mostly negative. The exception is the market factor whose R^2 is even higher at 6.8%. However, we consider this high R^2 non-robust possibly due to the significant noise in the time-series R^2 metric. We acknowledge that simple univariate time-series regression has apparent limitations in predicting aggregate factor returns. Our 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).

B.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.

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.

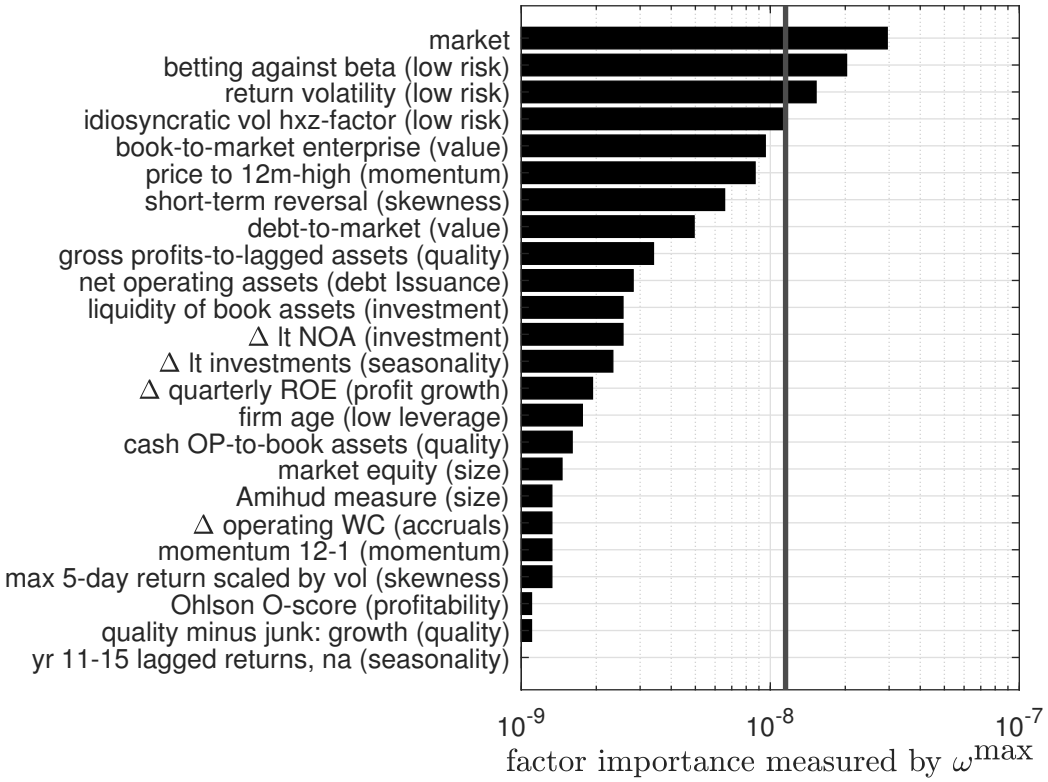
The main text Table 3 already shows that the β-only model has weak predictive power, with low and even negative R^2 values in some OOS cases. Table A.2 shows that the μ coefficients in the β-only model are either statistically insignificant or negative in various factor specifications. This, once again,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

B.4 Additional results on factor selection

We present a more formal factor importance analysis and report 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.3 reports $\omega_k^{\max}$ for the top 24 factors in the JKP factor zoo,

omitting the rest with $\omega_k^{\max} < 10^{-9}$.

Figure A.3: 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.

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 style investment clusters such as the value, quality, investment, seasonality, etc. Specifically, the 24 factors' full names, the factor clusters they belong to, and code names (as in JKP data) are:

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.

B.5 Additional robustness results

Table A.3 evaluates the robustness of the BTQ model's predictive power in 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 according to concurrent NYSE market capitalization quintiles. The same OOS predicted returns ($\widehat{r}_{i,t+1}$) are respectively evaluated in each size group. Panel B breaks downs 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 already reported in the main text Table 3 Panel B Line "BTQ" for ease of reference.

Table A.3 shows the BTQ model's predictive power as reported in the main text 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. In terms of 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

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 reports the original joint OOS (2010-2022) evaluation for reference.

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.