

Data-mined Anomalies and the Expected Market Return*

Peng Li[†], Emmanouil Platanakis[‡], Xiaoxia Ye[§], Guofu Zhou[¶]

May 2026

Abstract

Building on a theoretical framework of mispricing correction persistence (MCP), we develop a two-stage anomaly selection procedure to predict market returns. In the first stage, we screen anomalies based on the t-statistics of their long–short returns. In the second stage, we estimate the slope coefficients from predictive regressions to identify the most promising signals for forecasting the market. The anomalies selected from a universe of several thousand data-mined signals exhibit strong MCP. We show that the aggregate returns of long–short portfolios constructed from these data-mined anomalies are strongly associated with the predictability of aggregate excess market returns and deliver statistically significant out-of-sample forecasts. Applying the same two-stage procedure, we further demonstrate that data-mined anomalies outperform published anomalies, suggesting that MCP is more pronounced among the former, which are less exploited by investors.

Keywords: Data-mined anomalies, published anomalies, mispricing correction persistence, time-series predictability, stock market return.

JEL Classification: G11, G12, G17.

*We are grateful to conference participants at the Frontiers of Factor Investing Conference 2026 (Lancaster) and the European Financial Management Association 2026 Annual Meeting (Norway), and seminar participants at Olin Business School - Washington University in St. Louis and University of York for their useful comments and suggestions.

[†]Peng Li is at School of Management, University of Bath, UK. Email: pl750@bath.ac.uk

[‡]Emmanouil Platanakis is at School of Management, University of Bath, UK. Email: ep643@bath.ac.uk

[§]Xiaoxia Ye is at Business School, University of Nottingham, UK. Email: xiaoxia.ye@nottingham.ac.uk

[¶]Guofu Zhou is at Olin Business School, Washington University in St. Louis, USA. Email: zhou@wustl.edu

1 Introduction

Both cross-sectional (anomalies) and time-series (market premium) return predictability are fundamental topics in asset pricing. Motivated by the asymmetric persistence of mispricing corrections (Shleifer and Vishny, 1997), where mispricing creates autocorrelation if it requires a long period to be corrected and the positive serial correlation is stronger for overvalued assets, Dong et al. (2022) demonstrate that cross-sectional anomaly long-short returns, used as proxies for mispricing correction persistence (MCP), can help forecast market returns. In this study, we construct data-mined anomalies and, by aggregating them, show that they improve out-of-sample market return forecasts compared to published anomalies. Furthermore, the out-of-sample performance, controlled for overfitting in-sample data using data-mined anomalies, outperforms published anomalies across various aggregation techniques, including the predictor average as our base case and popular machine learning methods. Overall, the out-of-sample predictability of market returns using data-mined anomalies is statistically and economically significant. Our empirical findings provide support for the persistence of mispricing corrections.

Building on the mispricing narrative and the connection between cross-sectional and time-series predictability, we investigate the predictive power of data-mined anomalies for market returns. Following Yan and Zheng (2017), data-mined anomalies refer to long-short portfolios constructed by searching numerous accounting or return variables without clear return-characteristic relation guidance. McLean and Pontiff (2016) find that the long-short returns of cross-sectional anomalies decrease by more than 50% after publication. First, this aligns with the mispricing hypothesis, as mispricing should be arbitrated away when more investors become aware of the anomalies. Second, it implies that undiscovered or unpublished anomalies may contain more information about mispricing. These anomalies are less likely to be exploited as investment strategies, making their associated mispricing more persistent. To this end, data-mined anomalies lacking strong theoretical economic or financial backing are less likely to be published. Thus, they should capture stronger mispricing correction persistence than published anomalies. If the predictability of long-short anomaly returns for market returns stems from persistent mispricing, we expect data-mined

anomalies to outperform published anomalies in predicting market returns.

According to the principle of MCP, both the long and short legs of an anomaly exhibit positive serial correlation. Intuitively, for the long leg of an anomaly, if it is undervalued at time $t - 1$, it will yield higher returns at time t . Due to the persistence of mispricing correction, it will continue to generate higher returns at $t + 1$ and beyond as the mispricing correction process unfolds. Similarly, the short leg of an overvalued anomaly at $t - 1$ will generate lower returns at t and remain low at $t + 1$ and onward until the mispricing correction process concludes. Since the returns of the long and short legs contribute to the market returns, both legs should positively predict the market returns. Additionally and importantly, there are asymmetric limits to arbitrage, meaning that the short leg is more difficult to arbitrage, which makes the MCP stronger for the short leg. Consequently, one should expect: first, the short leg primarily drives the predictability of long-short returns for market returns. Second, since the long-short return reflects an adverse effect of the short leg, it should predict future market returns negatively.

Building on this intuition, we extend [Dong et al. \(2022\)](#) to develop a theoretical framework in which individual anomalies load on market-wide mispricing dynamics. The framework implies that an anomaly's predictive power depends on its exposure to MCP, which is shaped by both the strength of its long-short return (captured by its t-statistic) and its loading on the persistent component of market-wide overpricing. Therefore, it suggests a two-stage selection procedure to identify the anomalies most informative about future market returns. Accordingly, the first stage of our procedure screens for anomalies based on the significance of their long-short returns, while the second stage estimates the predictive regression slope coefficients that empirically reflect their MCP exposure. Combining these two criteria allows us to select anomalies that, in theory and in practice, are most likely to forecast aggregate market excess returns.

The construction of our data-mined anomalies follows the methodology of [Yan and Zheng \(2017\)](#). We generate combinations of accounting variables, including ratios and growth rates. For example, for any accounting variable Z , its growth rate is calculated as the difference between Z in year t and Z in year $t - 1$. We first select significant anomalies from these data-mined

anomalies whose absolute t-values are greater than 3. Next, we narrow the selection to the top 120 anomalies with the most negative beta coefficients, estimated from predictive regressions of individual long-short anomaly returns on subsequent market returns, which matches the number of published anomalies we consider in this study.

Consistent with our theoretical framework, the rationale for this two-step screening of data-mined anomalies lies in two primary considerations. First, anomalies with significant returns reflect the MCP, whereas insignificant anomalies are merely noise and cannot serve as proxies for the MCP. Second, the sign of the slope coefficients is determined by the difference between the covariance of the long-leg return with the market return and the covariance of the short-leg return with the market return. Due to the asymmetric nature of the mispricing correction process, where overpricing tends to dominate underpricing, the covariance of the short-leg return with the market return typically exceeds that of the long-leg return with the market return. This asymmetry in mispricing leads to a negative beta. In the second step, our selection procedure prioritizes data-mined anomalies that align with this mispricing asymmetry. Ultimately, we select 120 data-mined anomalies to match the number of published anomalies, allowing for a fair comparison between the two groups.

Goyal and Welch (2008) and Martin and Nagel (2022), among others, argue that the out-of-sample analysis is critical to assessing predictability in practice. Hence, we focus on out-of-sample analysis. We use both the R_{OS}^2 metric of Campbell and Thompson (2008) and the mean-squared forecasting errors (MSFE)-adjusted statistic of Clark and West (2007). Furthermore, in line with theory, we set the cross-sectional average (predictor average) as our base case for $N = 100, 120, 140, 160$, and 180 data-mined anomalies selected with the proposed two-stage selection procedure ($DM^{2-stage \sim N}$), and the 120 published anomalies ($ANOM$).

We show that $DM^{2-stage \sim N=120}$ generates an economically significant R_{OS}^2 of 0.91%, statistically significant at the 5% level. Similarly, for alternative values of N , $DM^{2-stage \sim N}$ also generates economically significant R_{OS}^2 's of 0.83% ($N=100$), 1.09% ($N=140$), 1.13% ($N=160$), and 1.19% ($N=180$), all statistically significant at the 5% level. In contrast, $ANOM$ generates a negative R_{OS}^2 of -0.12% . Further, $DM^{2-stage \sim N}$ outperforms a wide variety of popular predic-

tors from the literature. From an asset allocation perspective, the cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 7.42, outperforming $ANOM$ (2.87). The outperformance of the data-mined anomalies in asset allocation persists when examining the 1st and 2nd half of the out-of-sample period separately.

We examine the recursive slope coefficients for $DM^{2-stage \sim N}$ ($N = 100, 120, 140, 160,$ and 180) and $ANOM$, to compare their magnitudes of MCP for overpricing vis-à-vis underpricing. We show that the recursive estimates for the selected data-mined anomalies are always greater in magnitude than published anomalies across the entire evaluation period.

We evaluate the predictability performance of the data-mined anomalies using the combination ENet (C-ENet) of [Rapach and Zhou \(2020\)](#) and [Dong et al. \(2022\)](#), an efficient framework when the number of predictors is large. We show that $DM^{2-stage \sim N=120}$ generates an economically significant R_{OS}^2 of 2.68% using C-ENet, statistically significant at the 1% level, which outperforms $ANOM$ (0.80%). In terms of asset allocation, the cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 6.63 that outperforms $ANOM$ (3.72). The outperformance is maintained when examining the 1st and 2nd half of the evaluation period separately.

We apply our proposed two-stage selection framework to the 120 published anomalies to examine its generalization and further practical applicability, selecting 10 of them. For the cross-sectional average, $ANOM^{2-stage}$ yields an economically significant R_{OS}^2 of 0.68% and outperforms $ANOM$ (-0.12%), as is the case with C-ENet. Further, the cumulative wealth generated by $ANOM^{2-stage}$ is 5.37, almost double that of $ANOM$ (2.87).

Motivated by [Kelly et al. \(2024\)](#) who theoretically and empirically demonstrate that model complexity can be a virtue rather than a weakness in market return prediction, we extend their framework to the domain of anomaly-based predictors in forecasting market returns using the cross-sectional average and C-ENet to aggregate $p = 500$ and 1000 random Fourier features generated by the 120 data-mined anomalies selected with the proposed two-stage procedure, and the 120 published anomalies. We observe that $DM^{2-stage \sim p=500}$ generates an economically significant R_{OS}^2 of 1.20%, statistically significant at the 5% level, outperforming $ANOM^{p=500}$

(0.22%). Using C-ENet, $DM^{2-stage \sim p=500}$ generates an economically significant R_{OS}^2 of 3.36%, statistically significant at the 1% level and outperforms $ANOM^{p=500}$ (2.72%). The outperformance persists when using $p = 1000$ features. We also note that the forecasting performance of the 120 published anomalies improves from the random Fourier features generating process.

Our main conclusions remain unchanged after conducting a series of additional tests to ensure robustness of our core empirical findings, including alternative values of risk aversion, short sales constraints and transaction costs for asset allocation, as well as using C-LASSO and C-Ridge, instead of C-ENet, to examine the predictability performance of the data-mined anomalies selected with the two-stage procedure and compare them against the published anomalies.

We contribute to several areas of the existing literature. First, our study adds to research on time-series market return predictability. Previous studies have primarily relied on economic and financial indicators for forecasting. [Campbell \(1987\)](#) documents that the term structure of interest rates can predict market returns. [Cochrane \(2008\)](#) and [Cochrane \(2011\)](#) show supporting evidence that dividend yields can be used to predict market returns. [Goyal and Welch \(2008\)](#) and [Goyal et al. \(2024\)](#) find that many existing predictors of equity premium struggle to make accurate predictions in-sample and out-of-sample. Our findings extend this stream of literature by showing that data-mined anomalies can also be utilized to forecast market returns.

Second, we contribute to the vast literature on market return predictability through cross-sectional anomalies ([Dong et al., 2022](#); [Engelberg et al., 2023](#); [Wang et al., 2025](#); [Dong et al., 2025](#)). Our study provides new empirical evidence based on data-mined anomalies in this field.

Third, we also add to the debate on whether cross-sectional anomalies effectively forecast market returns. For example, [Cakici et al. \(2024\)](#) question the predictive power of anomalies for market returns by analyzing alternative sets of published anomalies. [Dong et al. \(2025\)](#) find a weak link between the returns of published anomalies and market returns in countries outside the US, but the predictability is strong when aggregating across nations. Our study, which relies on data-mined anomalies, provides supportive evidence for [Dong et al. \(2022\)](#) by showing that anomalies have predictive power for the expected market return. Finally, our results suggest that most data-mined

anomalies can be attributed to mispricing, aligning with the mispricing explanation proposed in many cross-sectional anomaly studies. Our findings highlight the importance of data-mined anomalies in enhancing our understanding of market return predictability.

We organize the remainder of our study as follows. [Section 2](#) presents the theoretical motivations and our extension of the [Dong et al. \(2022\)](#) framework to multiple observable anomalies with an idiosyncratic component and anomaly-specific loadings. [Section 3](#) describes the data for published and data-mined anomalies. [Section 4](#) presents our main out-of-sample analysis, and [Section 5](#) includes the additional analysis. Finally, [Section 5](#) concludes. Technical details and supplementary results are available in the online supplementary appendix.

2 Theoretical motivations

We follow [Dong et al. \(2022\)](#) and extend their framework to multiple observable anomalies with an idiosyncratic component and anomaly-specific loadings on the changes of market-wide mispricing. Specifically, the market return has three components:

$$R_{M,t} = f_t + \frac{\Delta u_{L,t} + \Delta u_{S,t}}{2}, \quad (1)$$

where f_t is the period t increment of a common martingale component, $\Delta u_{L,t}$ ($\Delta u_{S,t}$) is the change of market-wide underpricing (overpricing), $\text{Cov}(f_t, \Delta u_{L,t}) = \text{Cov}(f_t, \Delta u_{S,t}) = 0$; we model the i^{th} anomaly's long-short portfolio return as:

$$R_{LS,t}^i = \gamma^i \Delta u_{L,t} - \theta^i \Delta u_{S,t} + \omega_t^i, \quad (2)$$

where $\gamma^i > 0$ and $\theta^i > 0$ are the anomaly-specific loadings to $\Delta u_{L,t}$ and $\Delta u_{S,t}$, respectively, and ω_t^i is a serially uncorrelated idiosyncratic shock.

Now, consider the predictive regression relating the i^{th} anomaly's long-short portfolio return to

next period's market return:

$$R_{M,t+1} = \eta^i + \zeta^i R_{LS,t}^i + \varepsilon_{t+1}^i, \quad (3)$$

where $R_{M,t}$ is the market return at period t and ε_t^i is the regression error term at period t for the i^{th} anomaly. The standardized slope coefficients of [eq. \(3\)](#) is given by:

$$\tilde{\zeta}^i = \frac{0.5 [\gamma^i \text{Cov}(\Delta u_{L,t+1}, \Delta u_{L,t}) - \theta^i \text{Cov}(\Delta u_{S,t+1}, \Delta u_{S,t})]}{\sqrt{(\gamma^i)^2 \text{Var}(\Delta u_{L,t}) + (\theta^i)^2 \text{Var}(\Delta u_{S,t}) + \text{Var}(\omega_t^i)}}. \quad (4)$$

Following [Dong et al. \(2022\)](#), we assume $\text{Cov}(\Delta u_{L,t+1}, \Delta u_{L,t}) = 0$, which effectively means mispricing in the short leg is more persistent than that in the long leg, consistent with the evidence of limit to arbitrage being more severe in the short leg due to short-sell constraint. Therefore, $\tilde{\zeta}^i$ in [eq. \(4\)](#) can be written as:

$$\tilde{\zeta}^i = \lambda^i \text{Cov}(\Delta u_{S,t+1}, \Delta u_{S,t}), \quad (5)$$

where

$$\lambda^i = -\frac{\theta^i}{2\sqrt{(\gamma^i)^2 \text{Var}(\Delta u_{L,t}) + (\theta^i)^2 \text{Var}(\Delta u_{S,t}) + \text{Var}(\omega_t^i)}}. \quad (6)$$

[Dong et al. \(2022\)](#) use their framework to provide a theoretical reason why a market-wide anomaly's long-short portfolio has predictive power. This is because the long-short portfolio is associated with the MCP correlated with future market returns. We extend [Dong et al. \(2022\)](#)'s simple framework to account for individual anomalies, thereby enabling us to develop criteria for comparing different anomalies in terms of their predictive power for market returns. [Equation \(5\)](#) provides us with these criteria: essentially, it says that an anomaly's market return predictive power can be measured by the exposure (λ^i) to the market-wide MCP, i.e., $\text{Cov}(\Delta u_{S,t+1}, \Delta u_{S,t})$. [Equation \(6\)](#) says that the lower the standard deviation of an anomaly's long-short portfolio returns,

the larger the exposure’s magnitude. Therefore, anomalies with more significant long-short returns, that is, with higher t -statistics, tend to be better candidates for predicting the market. However, the t -statistic is not the only criterion: the numerator θ^i of [eq. \(6\)](#) also plays an important role in affecting the magnitude of the exposure, λ^i . If the assumption of [\(2\)](#) is reasonable, an empirical estimate of $\tilde{\zeta}^i$ should be proportional to λ^i to a good degree. Given these insights, we propose combining both the t -statistic and the estimate of the slope coefficient $\tilde{\zeta}^i$ in [eq. \(3\)](#) as criteria to select the best candidate anomalies for predicting the market.

3 Cross-sectional anomalies

3.1 Published anomalies

We utilize CRSP and Compustat data spanning the earliest available dates in these databases through Dec 2024 to construct 138 anomalies. Our sample encompasses a broader range of anomalies than that of [Dong et al. \(2022\)](#), who included 100 published anomalies. The scope of our analysis is comparable to that of other meta-studies. For instance, [McLean and Pontiff \(2016\)](#) examine 97 published anomalies, while [Green et al. \(2017\)](#) analyze 94 firm characteristics. [Chen and Zimmermann \(2022\)](#) replicate 161 anomalies, and [Jensen et al. \(2023\)](#) construct 153 anomalies. Overall, our study covers a comprehensive set of published anomalies.

To construct the 138 anomalies, we follow the anomaly definitions provided by [Chen and Zimmermann \(2022\)](#), as their calculations closely align with the methodologies described in the original papers. To compute the long-short returns, we construct decile portfolios, assigning the returns based on the relationship between firm characteristics and stock returns.

For anomalies with a negative relationship, such as the asset growth anomaly — which indicates a negative correlation between total asset growth and stock returns — we calculate the long-short return by going long the lowest asset growth decile (decile 1) and shorting the highest (decile 10). Conversely, for anomalies exhibiting a positive relationship, such as momentum, we go long on decile 10 and short on decile 1.

We remove anomalies with more missing values to obtain a sample with a good cross-sectional number of anomalies per month and sufficient time periods. We aim to balance the trade-off between the number of anomalies and the length of the time horizon. Our sample includes 120 published anomalies and spans from Jan 1969 to Dec 2024. Panel A of [Table 1](#) shows that the average t -statistic of all 138 anomalies is 2.33, and 53.6% of them are significant at the t threshold of 2 for the period up to December 2003. After 2003, the average t -statistic of all 138 anomalies is 0.37, and the percentage of the anomalies that are significant at the t threshold of 2 decreases to 12.3%. The 120 anomalies used in our main analysis have an average t -value of 2.05, and 48.3% of them have a t -statistic greater than 2 for the period up to December 2003. After 2003, the average t -statistic of the 120 anomalies is 0.38, and the percentage of the anomalies that are significant at the t threshold of 2 decreases to 5.8%.

[Insert [Table 1](#) about here]

3.2 Construction and selection of data-mined anomalies

We construct data-mined cross-sectional anomalies following [Yan and Zheng \(2017\)](#) and [Chen et al. \(2025\)](#). These anomalies are computed using data from CRSP and Compustat (we use the earliest available date for each database through Dec 2024). The data-mining strategies are structured as follows. For return-based anomalies, we compute cumulative returns from $t - k - j$ to $t - k - 1$, where k ranges from 0 to 12, 24, 36, 48, and 60 months, and j spans from 1 to 240. Each k - j pair serves as an anomaly-sorting variable.

For accounting-based anomalies, we apply six methods to define anomaly variables. These methods include the ratio of Z to Y (Z/Y), the change in Z/Y , the percent change in Z/Y , the change in Z scaled by lagged Y , the percent change in Z minus the percent change in Y , and the percent change in Z . In these definitions, Z represents an accounting variable, and Y serves as the base variable for scaling. For accounting variables Z , we require at least 1,000 firms per year and at least 20 years of data for each anomaly variable. This filtering process identifies 245 accounting-related variables from Compustat, of which we select 15 as base variables Y . The

final scaling variable is market value, which is sourced from CRSP. When constructing anomalies that involve one of the base variables (Y), we exclude it from being used as a Z -variable because changes or ratios involving the same variable lack cross-sectional variation. For instance, the total asset is one of the Y -variables. If total assets were also used as Z , all firms would have a Z/Y ratio of 1, resulting in no cross-sectional variation.

Next, we divide all stocks into deciles each month based on every anomaly variable. We introduce a six-month lag for accounting-based anomalies to account for the time lag between the fiscal year-end and the public release of financial reports. This ensures that the information is available when portfolios are formed. After forming portfolios, we compute each anomaly variable's return spread between decile 10 and decile 1.

Since only significant anomalies can be used to measure MCP, the first step in our data-mined anomaly selection is to retain only significant anomalies. To achieve this, we calculate t -statistics for anomalies if the long-short return data spans at least 120 months before the end of 2003. This results in 18,022 data-mined anomalies. We set 2003 as the end when computing t -values to avoid look-forward bias when examining out-of-sample performance, since we use post-2003 data as the test period.

The second step of anomaly selection examines the coefficient of the long-short return from the regression of subsequent market returns on the long-short return of individual anomalies. The long-short return is composed of a long leg and a short leg. The regression coefficient is proportional to the difference between the covariance of the long leg return with the market return and the covariance of the short leg return with the market return. The short leg, which represents overvalued assets, faces higher limits to arbitrage, including liquidity risk ([Pastor and Stambaugh, 2003](#)), idiosyncratic risk ([Ang et al., 2006](#)), and short-selling costs ([D'Avolio, 2002](#); [Asness et al., 2018](#)). Consequently, MCP tends to be stronger for the short leg. Given this, the covariance of the short leg return with the market return should be larger. A negative coefficient, therefore, captures this asymmetric MCP. Based on this insight, we include anomalies with more negative coefficients from the pool of significant anomalies.

Specifically, to make results comparable with our published anomalies, we aim to keep the exact number of data-mined anomalies equal to the number of published anomalies. To this end, we rank the t -statistic values from high to low for data-mined anomalies and retain $N = 120$ of them with the most negative coefficients, excluding those with missing data from Jan 1969. In our out-of-sample analysis, for robustness, we also consider retaining an alternative number (N) of significant data-mined anomalies with the most negative coefficients, for example, $N = 100, 140, 160$, and 180 . This two-step selection procedure ensures that the data-mined anomalies included in our analysis are reliable proxies for asymmetric MCP.

Panel B of [Table 1](#) shows that the average t -statistic of all 18,022 data-mined anomalies is 1.30, and 22% of them are significant at the t threshold of 2 for the period up to December 2003. After 2003, the average t -statistic of all the data-mined anomalies is 0.01, and the percentage of these anomalies that are significant at the t threshold of 2 decreases to 3.4%. The 120 double-screened data-mined anomalies used in our main analysis have an average t -value of 3.84, and 99.2% of them have a t -statistic greater than 2 for the period up to December 2003. After 2003, the average t -statistic for the 120 double-screened data-mined anomalies is 1.14, and 28.3% are significant at the t threshold of 2.

4 Out-of-sample analysis

We follow the standard practice of the market prediction literature and start with the following predictive regression:

$$R_{M,t+1} = \alpha + \beta X_t + \varepsilon_t, \quad (7)$$

where $R_{M,t+1}$ is the excess stock market return at $t + 1$, and X_t is the predictor at t . We compute the out-of-sample forecasts as follows:

$$\hat{R}_{M,t+1} = \hat{\alpha} + \hat{\beta}X_t, \quad (8)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimates of the predictive regression (7). As usual, we use an expanding estimation window and recursive estimation to generate out-of-sample forecasts.

Although an in-sample analysis provides more accurate estimates when all data are known with certainty, Goyal and Welch (2008), among others, argue that out-of-sample tests are more relevant for assessing genuine return predictability in practice. Our forecast will be absolutely out-of-sample, with no look-ahead bias: no future information will be used at any out-of-sample time.

We use Campbell and Thompson (2008)'s R_{OS}^2 and Clark and West (2007)'s MSFE-adjusted statistic to evaluate the out-of-sample forecasting performance. The R_{OS}^2 measures the proportional reduction in the mean squared forecast error (MSFE) of the predictive regression relative to the benchmark forecast (the historical average excess market return from the beginning of the sample period up to time t).

In particular, the errors for the mean benchmark and a predictive forecast are defined as follows:

$$\hat{e}_{0,t|t-1} = R_{M,t} - \hat{R}_{t|t-1}^{MB}, \quad (9)$$

$$\hat{e}_{1,t|t-1} = R_{M,t} - \hat{R}_{M,t|t-1}, \quad (10)$$

where $R_{M,t}$ is the market excess return, $\hat{R}_{t|t-1}^{MB}$ denotes the mean benchmark forecast and $\hat{R}_{M,t|t-1}$ represents the predictive forecast. The sample MSFE is given by

$$\widehat{MSFE}_j = \frac{1}{T} \sum_{t=1}^T \hat{e}_{j,t|t-1}^2, \quad for \ j = 0, \dots, 1, \quad (11)$$

where T is the number of out-of-sample observations.

The R_{OS}^2 statistic, which is very common to report when comparing MSFEs between the mean benchmark forecast and a predictive forecast is computed as:

$$R_{OS}^2 = 1 - \frac{\widehat{MSFE}_1}{\widehat{MSFE}_0}. \quad (12)$$

When $R_{OS}^2 > 0$, the predictive regression forecast outperforms the benchmark forecast in terms of MSFE. [Campbell and Thompson \(2008\)](#) suggest that a monthly R_{OS}^2 greater than 0.5% can signal economic significance. To uncover whether the predictive forecasts generate a statistically significant improvement in MSFE, we use [Clark and West \(2007\)](#)'s MSFE-adjusted statistic to test the hypothesis $H_0 : R_{OS}^2 \leq 0$ against $H_0 : R_{OS}^2 > 0$.

We use the first 35 years (1969:01 to 2003:12) of the entire sample period as the initial in-sample estimation period. 2004:01 to 2025:01 (253 observations) constitutes the out-of-sample period for forecast evaluation. We set 2003 as the split date to be conservative - many studies show that anomalies have weaker cross-sectional predictability after 2003 or 2004, for example, [Chordia et al. \(2014\)](#), [Green et al. \(2017\)](#), [Chen and Velikov \(2023\)](#), [Chen and Dim \(2024\)](#), and [Chen et al. \(2025\)](#), among others. Therefore, if anomalies can still capture information to forecast market returns in the post-2003 test period, the results are more credible and less likely to be due to overfitting.

We set the cross-sectional average (predictor average) as our base case for two reasons. First, the simple predictor average is model-free and requires no subjective parameter selection. Second, considering the predictor average allows us to directly compare the (negative) slopes of the predictive regressions across different sets of anomalies and to extract useful information about the pattern of stronger MCP for overpricing vis-à-vis underpricing. Although the predictor average, which is our base case, can accommodate anomalies with missing data, we exclude anomalies with missing returns from our sample, since the ML aggregation methods we will employ later in the paper also require non-missing predictor data.

4.1 Forecast evaluation

[Table 2](#) reports the out-of-sample R_{OS}^2 's and MSFE-adjusted statistics for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies selected with the proposed two-stage selection procedure ($DM^{2-stage \sim N}$) from a universe of several thousand signals, and the cross-sectional average of the

120 published anomalies (*ANOM*). We observe that $DM^{2-stage \sim N=120}$ yields an economically significant R_{OS}^2 of 0.91%, which is statistically significant at the 5% level according to the MSFE-adjusted statistic. Similarly, $DM^{2-stage \sim N}$ generates economically significant R_{OS}^2 's of 0.83% (N = 100), 1.09% (N = 140), 1.13% (N = 160), and 1.19% (N = 180), which are statistically significant at the 5% level. In contrast, *ANOM* generates a negative R_{OS}^2 of -0.12%.

[Insert *Table 2* about here]

As discussed by [Gârleanu and Pedersen \(2013\)](#) and others, market frictions such as short-sale restrictions and hard-to-borrow fees can cause a very slow response to mispricing. Since the selected data-mined anomalies using our proposed two-stage procedure exhibit strong MCP, we expect they should also contain more relevant predictive information than published anomalies during times of high market frictions. To test this hypothesis, we use the financial uncertainty index of [Jurado et al. \(2015\)](#), which measures the aggregate unpredictability and uncertainty of several financial variables, as a proxy for market frictions. *Table 3* reports the differences in R_{OS}^2 's between high- and low-friction periods, defined using the sample median of the uncertainty index. In contrast to the published anomalies, we show that the differences in R_{OS}^2 's between high- and low-friction regimes are always positive for the N = 100 (4.30%), 120 (3.09%), 140 (2.21%), 160 (1.75%), and 180 (1.31%) data-mined anomalies selected with the two-stage procedure.

[Insert *Table 3* about here]

Further, to alleviate concerns that the outperformance of data-mined anomalies may stem from a small number of anomalies, we randomly draw two subsets of 100 anomalies from the 120 data-mined anomalies selected with the 2-stage procedure and the 120 published anomalies, respectively. We repeat the process 1,000 times and, each time, use the cross-sectional average to compute out-of-sample market excess return forecasts and the corresponding R_{OS}^2 . *Figure 1* presents histograms for the R_{OS}^2 's for the two subsets. Data-mined anomalies generate an R_{OS}^2 greater than 0.5% in 99.8% of the cases, outperforming published anomalies, which have no cases with an R_{OS}^2 greater than 0.5%.

[Insert *Figure 1* about here]

4.2 Slope coefficients and MCP

Table 2 reports the mean standardized recursive slope of the predictive regressions, and **Figure 2** plots the recursive slopes across the out-of-sample period. This is a convenient and efficient way to compare $DM^{2-stage \sim N}$, $N = 120, 140, 160$, and 180 , against $ANOM$, in terms of the magnitude of the MCP for overpricing vis-à-vis underpricing, since aggregate information in all the long-short portfolio returns is captured in a single slope coefficient.

[Insert *Figure 2* about here]

The predictive regressions underlying the predictor average forecasts for $DM^{2-stage \sim N}$ and $ANOM$ are given as follows:

$$R_{M,t+1} = \alpha + \beta \bar{R}_{LS,t}^{DM^{2-stage \sim N}} + \varepsilon_t, \quad (13)$$

$$R_{M,t+1} = \alpha + \beta \bar{R}_{LS,t}^{ANOM} + \varepsilon_t. \quad (14)$$

For $i = 1, \dots, N$, $N = 100, 120, 140, 160$ and 180 , $\bar{R}_{LS,t}^{DM^{2-stage \sim N}}$ and $\bar{R}_{LS,t}^{ANOM}$ are computed as:

$$\bar{R}_{LS,t}^{DM^{2-stage \sim N}} = \frac{1}{N} \sum_{i=1}^N R_{LS,t}^{DM^{2-stage \sim N},i}, \quad (15)$$

$$\bar{R}_{LS,t}^{ANOM} = \frac{1}{120} \sum_{i=1}^{120} R_{LS,t}^{ANOM,i}, \quad (16)$$

where $R_{LS,t}^{DM^{2-stage \sim N},i}$ and $R_{LS,t}^{ANOM,i}$ represent the long-short portfolio returns of the i^{th} data-mined and published anomaly, respectively, at time t .

As expected, the recursive slope coefficients are always negative for $DM^{2-stage \sim N}$, $N = 100, 120, 140, 160$, and 180 , and $ANOM$, highlighting the stronger MCP for overpricing vis-à-vis underpricing. However, the recursive estimates for the data-mined anomalies are always larger in

magnitude than those of published anomalies across the entire evaluation period. For example, the absolute mean standardized recursive slope for $DM^{2-stage \sim N=120}$ is 0.81, which is more than double that of *ANOM*, suggesting that the pattern of stronger MCP for overpricing vis-à-vis underpricing is more pronounced in the long-short portfolio returns of the data-mined anomalies than in published anomalies, which in turn explains their stronger predictability power for the expected market return.

4.3 Economic value

A monthly out-of-sample R_{OS}^2 of 0.5% or greater can potentially generate substantial economic benefits (Campbell and Thompson, 2008). Hence, in this section, we conduct an asset allocation analysis to examine potential gains for investors. We follow Kandel and Stambaugh (1996), Ferreira and Santa-Clara (2011), Neely et al. (2014), Huang et al. (2015), Han et al. (2016), Rapach et al. (2016), Jiang et al. (2019), Chen et al. (2022), Chen et al. (2023), and Chen et al. (2025), among others, and examine a mean-variance investor's portfolio that is optimally allocating across equities and the risk-free asset by using the out-of-sample predictive regression forecasts each month.

Suppose a mean-variance investor allocates a proportion of w_t to the market portfolio and $1 - w_t$ to the risk-free asset by maximizing the following objective function at the end of each month:

$$U(R_{p,t+1}) = E(R_{p,t+1}) - \frac{\lambda}{2} Var(R_{p,t+1}), \quad (17)$$

where $E(R_{p,t+1})$ and $Var(R_{p,t+1})$ denote the mean and variance of the portfolio returns at $t + 1$, and λ is the investor's risk aversion. The investor's portfolio return at the end of each month ($t+1$) is given by:

$$R_{p,t+1} = w_t R_{M,t+1} + R_{f,t+1}. \quad (18)$$

$R_{M,t+1}$ and $R_{f,t+1}$ denote the excess stock market return and the risk-free rate, respectively. The

optimal portfolio weight allocated to the market portfolio is computed as follows:

$$w_t = \frac{1}{\lambda} \frac{\hat{R}_{M,t+1}}{\hat{\sigma}_{M,t+1}^2}, \quad (19)$$

where λ is the risk aversion coefficient, $\hat{R}_{M,t+1}$ is the out-of-sample forecast of stock excess return, and $\hat{\sigma}_{M,t+1}^2$ is its variance. Following [Campbell and Thompson \(2008\)](#) and many others, we use a 60-month rolling window to estimate the variance of market excess returns. Further, we allow for short sales and leverage for our base-case results. We define the cumulative wealth (CW) as follows:

$$CW = \prod_{t=1}^T (1 + R_{p,t}) \quad (20)$$

Table 4 reports the cumulative wealth of a mean-variance investor, with a risk aversion coefficient of five, using the cross-sectional average of the $N = 100, 120, 140, 160,$ and 180 data-mined anomalies selected with the two-stage procedure ($DM^{2-stage \sim N}$), and the cross-sectional average of the 120 published anomalies ($ANOM$), to predict future excess market returns. We also allow for short sales and leverage. In addition, **Figure 3** plots the cumulative wealth across the entire out-of-sample period for all cases we considered. We observe that investment portfolios based on data-mined anomalies outperform those of the published anomalies. When examining the entire evaluation period, the cumulative wealth of $DM^{2-stage \sim N=120}$ is 7.42, more than double that of $ANOM$ (2.87). Similarly, the cumulative wealth generated by $DM^{2-stage \sim N}$ is 6.83 ($N=100$), 7.68 ($N=140$), 7.87 ($N=160$), and 7.92 ($N=180$), outperforming $ANOM$. The outperformance of the data-mined anomalies in asset allocation is maintained when examining the 1st and 2nd half of the out-of-sample period separately. In particular, the cumulative wealth generated by $DM^{2-stage \sim N=120}$ in the 1st (2nd) half is 1.68 (4.42).

[Insert **Table 4** and **Figure 3** about here]

4.4 Benchmarking against popular predictors and a single-stage selection procedure

In this section, we show that data-mined anomaly portfolio returns exhibit stronger out-of-sample predictive power than 14 popular economic variables from [Goyal and Welch \(2008\)](#). They are the log dividend-price ratio, log dividend yield ratio, log earnings-price ratio, log dividend-payout ratio, stock return variance, book-to-market ratio, net equity expansion, treasury bill rate (3-month Treasury bill interest rate), long-term yield (long-term government bond yield), long-term return (return on long-term government bonds), term spread (long-term yield minus the Treasury bill rate), default yield spread (difference between BAA- and AAA-rated corporate bond yields), default return spread (long-term corporate bond return minus the long-term government bond return) and inflation (CPI: 2-month lagged inflation in the regression to account for the delay in CPI releases). [Table 5](#) reports the R_{OS}^2 statistics for univariate predictive regression forecasts based on the 14 economic predictors. Among the predictors, we observe that two produce positive R_{OS}^2 statistics: the treasury bill rate (0.25%) and the long-term yield (0.28%), both of which are economically and statistically insignificant.

[Insert [Table 5](#) about here]

Further, [Table 6](#) reports the R_{OS} statistics for univariate predictive regression forecasts based on nine popular predictors from the literature. They are the average correlation of daily stock returns of [Pollet and Wilson \(2010\)](#), nearness to Dow 52-week high: the current difference to the lifetime high and nearness to Dow 52-week high: the scaled current difference to the 52-week high of [Li and Yu \(2012\)](#), aggregation of 9 illiquidity measures of [Chen et al. \(2018\)](#), ratio of new orders to shipments of durable goods of [Jones and Tuzel \(2013\)](#), cross-sectional standard deviation on the set of 100 size and book-to-market portfolios of [Maio \(2016\)](#), average of monthly skewness values across firms of [Jondeau et al. \(2019\)](#), tail risk from cross-section of [Kelly and Jiang \(2014\)](#), and stock-bond yield gap of [Maio \(2013\)](#). We observe that data-mined anomaly portfolio returns outperform all nine predictors since, among them, two produce a positive R_{OS}^2 statistic: the nearness

to Dow 52-week high:the current difference to the lifetime high (0.26%), and the tail risk from cross-section (0.20%), which are neither economically nor statistically significant.

[Insert *Table 6* about here]

We also show that the data-mined anomalies selected with the two-stage procedure outperform data-mined anomalies selected using only the t-statistic ($DM^{1-stage \sim N}$) for predicting the average monthly excess stock market return. *Table 7* reports the corresponding out-of-sample R_{OS}^2 's and MSFE-adjusted statistics. We observe that ($DM^{1-stage \sim N}$) generates R_{OS}^2 's of 0.63% (N =100), 0.73% (N=120), 0.76% (N =140), 0.50% (N =160), and 0.35% (N=180), which are all statistically insignificant.

[Insert *Table 7* about here]

4.5 Combination elastic net (C-ENet)

In this section, we examine whether our conclusions remain unchanged when using the combination ENet (C-ENet). Inspired by [Diebold and Shin \(2019\)](#), [Rapach and Zhou \(2020\)](#), [Dong et al. \(2022\)](#), and [Han et al. \(2024\)](#), among others, the C-ENet forecasts take the average of the individual forecasts selected by the ENet in a [Granger and Ramanathan \(1984\)](#) multiple regression relating the actual market excess return to the individual univariate forecasts, which is particularly efficient when the number of predictors is large. We use the first 20 years (1969:01 to 1988:12) for estimation, and the subsequent 15 years (1989:01 to 2003:12) as the initial holdout out-of-sample period for computing the forecasts. Further, we consider three values of the parameter that blends the l_1 and l_2 components in the penalty term for C-ENet ($\alpha = 0.25, 0.50$, and 0.75).

Table 8 reports the out-of-sample R_{OS}^2 's and MSFE-adjusted statistics for predicting the average monthly excess stock market return using C-ENet ($\alpha = 0.25, 0.50$, and 0.75) to aggregate the data-mined anomalies selected with the proposed two-stage procedure, and the 120 published anomalies. For $\alpha = 0.5$, we observe that $DM^{2-stage \sim N=120}$ generates an economically significant R_{OS}^2 of 2.68%, which is statistically significant at the 1% level and outperforms *ANOM* (0.80%). The

outperformance of the data-mined anomalies is maintained when using C-ENet with $\alpha = 0.25$ and 0.50, and across alternative numbers of data-mined anomalies selected in the second stage of the procedure.

[Insert *Table 8* about here]

By employing C-ENet ($\alpha = 0.5$), *Table 9* reports the cumulative wealth of a mean-variance investor with a risk aversion coefficient of five that allows for short sales and leverage, using the data-mined anomalies selected with the two-stage procedure and the 120 published anomalies to predict future excess market returns. *Figure 4* plots the cumulative wealth across the entire out-of-sample period for all the cases we considered. We observe that investment portfolios based on data-mined anomalies outperform those based on published anomalies. The cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 6.63, outperforming *ANOM* (3.72). Similarly, the cumulative wealth generated by $DM^{2-stage \sim N}$ is 6.57 (N=100), 6.49 (N=140), 6.54 (N=160), and 6.55 (N=180), outperforming *ANOM*. The outperformance persists when examining the 1st and 2nd half of the evaluation period separately.

[Insert *Table 9* and *Figure 4* about here]

4.6 Applying the two-stage selection procedure to published anomalies

So far, in the previous sections, we have shown that our proposed two-stage theoretical framework selects a subset of data-mined anomalies that exhibit strong and persistent mispricing correction dynamics and deliver economically and statistically significant out-of-sample predictions of market excess returns, outperforming the predictive power of published anomalies. However, it would be of great interest to examine the generalization of the two-stage framework and investigate whether its application could also improve the predictive performance of the published anomalies by selecting a subset of anomalies that are significantly related to the predictability of aggregate excess market returns.

We apply our two-stage framework to the 120 published anomalies, selecting 20 anomalies in the first step based on the t-statistic and 10 anomalies (out of 20) in the second step based on the slope coefficient of predictive regressions. [Table 10](#) reports the out-of-sample R_{OS}^2 's and MSFE-adjusted statistics for predicting the average monthly excess stock market return using a subset of published anomalies selected with the two-stage framework ($ANOM^{2-stage}$) and the 120 published anomalies ($ANOM$). For the cross-sectional average, $ANOM^{2-stage}$ generates an economically significant R_{OS}^2 of 0.68%, outperforming $ANOM$ (-0.12%), as is the case with C-ENet at $\alpha = 0.5$.

[Insert [Table 10](#) about here]

[Table 11](#) reports the cumulative wealth of a mean-variance investor with a risk aversion coefficient of five that allows for short sales and leverage to examine whether using the $ANOM^{2-stage}$ to predict future excess market returns could be helpful for asset allocation. When using the cross-sectional average, the cumulative wealth generated by $ANOM^{2-stage}$ over the entire evaluation period is 5.37, almost double that of $ANOM$ (2.87). For C-ENet ($\alpha = 0.5$), $ANOM^{2-stage}$ generates a cumulative wealth of 4.30, outperforming $ANOM$ (3.72). The outperformance of $ANOM^{2-stage}$ persists when examining the 1st and 2nd half of the evaluation period separately.

[Insert [Table 11](#) about here]

4.7 Virtue of complexity and anomalies

[Kelly et al. \(2024\)](#) demonstrate, both theoretically and empirically, that model complexity can be a virtue rather than a weakness in predicting market returns. Contrary to the traditional emphasis on parsimony, they show that when appropriate shrinkage or regularization is applied, high-dimensional predictive models (i.e., those with more parameters than observations) can deliver superior out-of-sample performance and economically meaningful improvements in market return forecasts. Their analysis reveals that complex models are better equipped to approximate the unknown, nonlinear relationship between predictive signals and expected market returns.

Motivated by the virtue of complexity, our study extends their framework to the domain of anomaly-based predictors in forecasting market returns. Specifically, we follow the random Fourier feature approach to systematically expand the set of predictors in our machine learning models, applying it separately to both data-mined anomalies and published anomalies. This allows our models to incorporate a substantially larger number of features than observations, enabling an empirical investigation of whether model complexity can similarly enhance the predictive content of anomaly returns for aggregate market returns. If complexity indeed contributes to predictability, we expect to observe stronger forecasting power in both sets of anomalies. Moreover, because data-mined anomalies reflect less arbitrated and potentially stronger mispricing than published anomalies, we further anticipate that complex models based on data-mined anomalies will exhibit superior predictive performance relative to those based on published anomalies.

Table 12 reports the out-of-sample R_{OS}^2 's and MSFE-adjusted statistics for predicting the average monthly excess stock market return using the cross-sectional average and C-ENet ($\alpha = 0.5$) to aggregate $p = 500$ and 1000 random Fourier features generated by the 120 data-mined anomalies selected with the proposed two-stage procedure, and the 120 published anomalies. Using the cross-sectional average, we observe that $DM^{2-stage \sim p=500}$ generates an economically significant R_{OS}^2 of 1.20%, which is statistically significant at the 5% level, outperforming $ANOM^{p=500}$ (0.22%). Using C-ENet, $DM^{2-stage \sim p=500}$ generates an economically significant R_{OS}^2 of 3.36%, which is statistically significant at the 1% level and outperforms $ANOM^{p=500}$ (2.72%). The outperformance of the data-mined anomalies is maintained when considering $p = 1000$. It is worth noting that generating many random Fourier features improves the forecasting performance of the 120 published anomalies, both with the cross-sectional average and C-ENet.

[Insert **Table 12** about here]

Table 13 reports the cumulative wealth of a mean-variance investor, with a risk aversion coefficient of five that allows for short sales and leverage, using the cross-sectional average and C-ENet ($\alpha = 0.5$) to aggregate $p = 500$ and 1000 random Fourier features generated by the 120 data-mined anomalies selected with the proposed two-stage procedure, and the 120 published anomalies.

When examining the entire evaluation period, the cumulative wealth generated by $DM^{2-stage \sim p=500}$ for the case of cross-sectional average (C-ENet) is 5.95 (9.35), outperforming $ANOM^{p=500}$. The outperformance of the data-mined anomalies persists when considering $p = 1000$ random Fourier features.

[Insert *Table 13* about here]

4.8 Additional analysis

In this section, we conduct a series of additional tests, reported in the Online Appendix, to ensure robustness of our core empirical findings. First, we evaluate the economic value of forecasting stock market returns in asset allocation for alternative reasonable values of risk aversion (Tables A1-A4). Second, we examine asset allocation when prohibiting short sales and allowing leverage of up to 50% (Tables A5-A9). Third, we evaluate asset allocation when considering transaction costs of 50 bps for all the cases (Tables A10-A19). Fourth, we evaluate asset allocation using C-ENet with alternative reasonable risk aversion coefficients and when prohibiting short sales (Tables A20-A24), and with transaction costs of 50 bps (Tables A25-A26). Fifth, we evaluate the forecasting performance of the data-mined anomalies when using C-LASSO and C-Ridge (Table A27), as well as the economic gains in asset allocation (Tables A28-A35). Sixth, we evaluate the predictive performance of the data-mined anomalies and the published anomalies using both the cross-sectional average and C-ENet over the 1st and 2nd half of the out-of-sample period separately (Tables A36-A39). Overall, we show that the data-mined anomalies selected with the proposed two-stage theoretical framework outperform the published anomalies both in predicting excess market returns and in asset allocation.

Tables A1-A4 report the cumulative wealth of a mean-variance investor with alternative values of risk aversion ($\lambda = 3, 4, 6$, and 7). For the entire evaluation period, we observe that the cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 12.87 for $\lambda = 3$, 9.68 for $\lambda = 4$, 6.00 for $\lambda = 6$, and 5.07 for $\lambda = 7$, outperforming ANOM in all cases. Similar conclusions hold when selecting $N = 100$, 140, 160, and 180 data-mined anomalies in the second stage of the selection procedure.

Tables A5-A9 examine the cumulative wealth of a mean-variance investor when prohibiting short sales and allowing leverage of up to 50%. In this case, we observe that $DM^{2-stage \sim N=120}$ generates a cumulative wealth of 6.90 for $\lambda = 3$, 5.91 for $\lambda = 4$, 4.88 for $\lambda = 5$, 4.10 for $\lambda = 6$, and 3.70 for $\lambda = 7$, outperforming ANOM. The conclusions also hold when selecting an alternative number of data-mined anomalies in the second stage.

Tables A10-A19 examine asset allocation of a mean-variance investor when considering transaction costs of 50 bps. When allowing for short selling and leverage, we observe that the cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 2.27 for $\lambda = 3$, 2.66 for $\lambda = 4$, 2.65 for $\lambda = 5$, 2.54 for $\lambda = 6$, and 2.43 for $\lambda = 7$. When prohibiting short sales and allowing leverage of up to 50%, $DM^{2-stage \sim N=120}$ generates a cumulative wealth of 3.68 for $\lambda = 3$, 3.17 for $\lambda = 4$, 2.69 for $\lambda = 5$, 2.35 for $\lambda = 6$, and 2.21 for $\lambda = 7$, outperforming ANOM in all cases.

Tables A20-A21 evaluate the asset allocation of a mean-variance investor using C-ENet with alternative reasonable risk aversion coefficients when allowing for short selling and leverage. The cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 14.62 for $\lambda = 3$ and 4.42 for $\lambda = 7$, outperforming ANOM. Tables A22-A24 examine asset allocation using C-ENet when prohibiting short sales and allowing leverage of up to 50%. In this case, $DM^{2-stage \sim N=120}$ generates a cumulative wealth of 11.46 for $\lambda = 3$, 6.24 for $\lambda = 5$, and 4.26 for $\lambda = 7$, outperforming ANOM. Tables A25-A26 also report the asset allocation using C-ENet, and when considering transaction costs of 50 bps, the main conclusions remain unchanged.

Table A27 reports the forecasting performance of data-mined anomalies and published anomalies when using C-LASSO and C-Ridge. For C-LASSO, $DM^{2-stage \sim N=120}$ yields an economically significant R_{OS}^2 of 2.36%, which is statistically significant at the 1% level and outperforms ANOM (1.30%). Similarly, $DM^{2-stage \sim N}$ generates economically significant R_{OS}^2 's of 2.65% (N = 100), 2.40% (N = 140), 2.83% (N = 160), and 2.83% (N = 180), which are statistically significant at the 1% level. For C-Ridge, $DM^{2-stage \sim N=120}$ generates an economically significant R_{OS}^2 of 1.68%, statistically significant at the 1% level, outperforming ANOM. The main conclusions remain unchanged when considering N = 100, 140, 160, and 180 data-mined anomalies in the second stage

of the selection procedure.

Tables A28-A35 report the asset allocation when using C-LASSO and C-Ridge. For the C-LASSO case, the cumulative wealth generated by $DM^{2-stage \sim N=120}$ is 6.00 when allowing short selling and leverage, and 5.34 when prohibiting short selling and allowing leverage up to 50%. For the case of C-Ridge, $DM^{2-stage \sim N=120}$ generates a cumulative wealth of 4.91 when allowing for short selling and leverage, and 4.73 when prohibiting short selling and allowing leverage of up to 50%. Further, the main conclusions remain unchanged for $N = 100, 140, 160$, and 180 , and for transaction costs of 50 bps.

Further, Tables A36-A39 report the predictive performance of the data-mined and published anomalies over the 1st and 2nd half of the out-of-sample period using both the cross-sectional average and C-ENet, confirming the outperformance of the data-mined anomalies in most cases.

5 Conclusions

Following the framework of mispricing correction persistence proposed by [Dong et al. \(2022\)](#), we revisit the predictability of long-short anomaly returns for aggregate market returns using a large set of data-mined anomalies. Unlike well-documented published anomalies, data-mined anomalies are less likely to be known and exploited by investors, potentially allowing for stronger, more persistent mispricing corrections. As a result, long-short portfolios based on these less known signals are expected to exhibit greater predictive power for market excess returns.

We construct nearly 20,000 data-mined cross-sectional signals and identify 120 anomalies that meet two key screening criteria: (1) long-short returns must be statistically significant ($|t| \geq 2$), and (2) anomalies must exhibit asymmetric mispricing consistent with the correction persistence mechanism. These filters select signals that are more likely to reflect persistent mispricing corrections.

Consistent with model predictions, we find that the selected 120 long-short data-mined anomaly portfolios deliver stronger out-of-sample predictive performance for market excess returns when

evaluated using machine learning techniques that mitigate overfitting. Furthermore, investment strategies based on these data-mined anomalies achieve higher Sharpe ratios and higher certainty equivalent returns (CER) than those based on published anomalies. A series of robustness checks confirms that our main findings remain intact under various alternative specifications.

Our results suggest that data-mined anomalies contain valuable and previously untapped information for forecasting aggregate market excess returns, offering both academic insight and practical relevance for asset pricing and investment strategies.

References

- Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang.** 2006. “The cross-section of volatility and expected returns.” *The Journal of Finance* 61 259–299.
- Asness, Clifford, Andrea Frazzini, Ronen Israel, Tobias J. Moskowitz, and Lasse H. Pedersen.** 2018. “Size matters, if you control your junk.” *Journal of Financial Economics* 129 479–509.
- Cakici, Nusret, Christian Fieberg, Daniel Metko, and Adam Zaremba.** 2024. “Do Anomalies Really Predict Market Returns? New Data and New Evidence.” *Review of Finance* 28 1–44.
- Campbell, John Y.** 1987. “Stock returns and the term structure.” *Journal of Financial Economics* 18 373–399.
- Campbell, John Y, and Samuel B Thompson.** 2008. “Predicting excess stock returns out of sample: Can anything beat the historical average?” *The Review of Financial Studies* 21 (4): 1509–1531.
- Chen, Andrew Y., and Chukwuma Dim.** 2024. “High-throughput asset pricing or: how I learned to stop worrying and love data mining.” *Working paper*.
- Chen, Andrew Y., Alejandro Lopez-Lira, and Tom Zimmermann.** 2025. “Does peer-reviewed research help predict stock returns?” *Working paper*.
- Chen, Andrew Y., and Mihail Velikov.** 2023. “Zeroing in on the expected returns of anomalies.” *Journal of Quantitative and Financial Analysis* 58 (3): 968–1004.
- Chen, Andrew Y., and Tom Zimmermann.** 2022. “Open Source Cross-Sectional Asset Pricing.” *Critical Finance Review* 11 207–264.
- Chen, Jian, Guohao Tang, Jiaquan Yao, and Guofu Zhou.** 2022. “Investor attention and stock returns.” *Journal of Financial and Quantitative Analysis* 57 (2): 455–484.
- Chen, Jian, Guohao Tang, Jiaquan Yao, and Guofu Zhou.** 2023. “Employee sentiment and stock returns.” *Journal of Economic Dynamics and Control* 149 104636.
- Chen, Jian, Guohao Tang, Guofu Zhou, and Wu Zhu.** 2025. “ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?” *Available at SSRN 4660148*.
- Chen, Yong, Gregory W Eaton, and Bradley S Paye.** 2018. “Micro (structure) before macro? The predictive power of aggregate illiquidity for stock returns and economic activity.” *Journal of Financial Economics* 130 (1): 48–73.
- Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong.** 2014. “Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?” *Journal of Accounting and Economics* 58 (1): 41–58.

- Clark, Todd E, and Kenneth D West.** 2007. “Approximately normal tests for equal predictive accuracy in nested models.” *Journal of Econometrics* 138 (1): 291–311.
- Cochrane, John.** 2008. “The dog that did not bark: a defense of return predictability.” *The Review of Financial Studies* 21 1533–1575.
- Cochrane, John.** 2011. “Discount rates.” *The Journal of Finance* 66 1047–1108.
- Diebold, Francis X, and Minchul Shin.** 2019. “Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives.” *International Journal of Forecasting* 35 (4): 1679–1691.
- Dong, Xi, Yan Li, Yanran Li, David Rapach, and Guofu Zhou.** 2025. “Anomalies and Links to Market Return Predictability: Supranational Evidence Based on New Market Efficiency Measures.” *Available at SSRN*.
- Dong, Xi, Yan Li, David E. Rapach, and Guofu Zhou.** 2022. “Anomalies and the expected market return.” *The Journal of Finance* 77 639–681.
- D’Avolio, Gene.** 2002. “The market for borrowing stock.” *Journal of Financial Economics* 66 271–306.
- Engelberg, Joseph, David McLean, Jeffrey Pontiff, and Matthew C. Ringgenberg.** 2023. “Do cross-sectional predictors contain systematic information?” *Journal of Financial and Quantitative Analysis* 58 1172–1201.
- Ferreira, Miguel A, and Pedro Santa-Clara.** 2011. “Forecasting stock market returns: The sum of the parts is more than the whole.” *Journal of Financial Economics* 100 (3): 514–537.
- Gârleanu, Nicolae, and Lasse Heje Pedersen.** 2013. “Dynamic trading with predictable returns and transaction costs.” *The Journal of Finance* 68 (6): 2309–2340.
- Goyal, Amit, and Ivo Welch.** 2008. “A comprehensive look at the empirical performance of equity premium prediction.” *The Review of Financial Studies* 21 (4): 1455–1508.
- Goyal, Amit, Ivo Welch, and Athanasse Zafirov.** 2024. “A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction.” *The Review of Financial Studies* 37 3490–3557.
- Granger, Clive WJ, and Ramu Ramanathan.** 1984. “Improved methods of combining forecasts.” *Journal of Forecasting* 3 (2): 197–204.
- Green, Jeremiah, John Hand, and Frank Zhang.** 2017. “The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns.” *The Review of Financial Studies* 30 4389–4436.

- Han, Yufeng, Ai He, David E Rapach, and Guofu Zhou.** 2024. “Cross-sectional expected returns: new Fama–MacBeth regressions in the era of machine learning.” *Review of Finance* 28 (6): 1807–1831.
- Han, Yufeng, Guofu Zhou, and Yingzi Zhu.** 2016. “A trend factor: Any economic gains from using information over investment horizons?” *Journal of Financial Economics* 122 (2): 352–375.
- Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou.** 2015. “Investor sentiment aligned: A powerful predictor of stock returns.” *The Review of Financial Studies* 28 (3): 791–837.
- Jensen, Theis, Bryan Kelly, and Lasse Pedersen.** 2023. “Is there a replication crisis in finance?” *The Journal of Finance* 78 2465–2518.
- Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou.** 2019. “Manager sentiment and stock returns.” *Journal of Financial Economics* 132 (1): 126–149.
- Jondeau, Eric, Qunzi Zhang, and Xiaoneng Zhu.** 2019. “Average skewness matters.” *Journal of Financial Economics* 134 (1): 29–47.
- Jones, Christopher S, and Selale Tuzel.** 2013. “New orders and asset prices.” *The Review of Financial Studies* 26 (1): 115–157.
- Jurado, Kyle, Sydney C Ludvigson, and Serena Ng.** 2015. “Measuring uncertainty.” *American Economic Review* 105 (3): 1177–1216.
- Kandel, Shmuel, and Robert F Stambaugh.** 1996. “On the predictability of stock returns: an asset-allocation perspective.” *The Journal of Finance* 51 (2): 385–424.
- Kelly, Bryan, and Hao Jiang.** 2014. “Tail risk and asset prices.” *The Review of Financial Studies* 27 (10): 2841–2871.
- Kelly, Bryan, Semyon Malamud, and Kangying Zhou.** 2024. “The virtue of complexity in return prediction.” *The Journal of Finance* 79 (1): 459–503.
- Li, Jun, and Jianfeng Yu.** 2012. “Investor attention, psychological anchors, and stock return predictability.” *Journal of Financial Economics* 104 (2): 401–419.
- Maio, Paulo.** 2013. “The “Fed model” and the predictability of stock returns.” *Review of Finance* 17 (4): 1489–1533.
- Maio, Paulo.** 2016. “Cross-sectional return dispersion and the equity premium.” *Journal of Financial Markets* 29 87–109.
- Martin, Ian WR, and Stefan Nagel.** 2022. “Market efficiency in the age of big data.” *Journal of Financial Economics* 145 (1): 154–177.

- McLean, David, and Jeffrey Pontiff.** 2016. “Does academic research destroy return predictability?” *The Journal of Finance* 71 5–32.
- Neely, Christopher J, David E Rapach, Jun Tu, and Guofu Zhou.** 2014. “Forecasting the equity risk premium: the role of technical indicators.” *Management Science* 60 (7): 1772–1791.
- Pastor, Lubos, and Robert F. Stambaugh.** 2003. “Liquidity risk and expected stock returns.” *Journal of Political Economy* 111 642–685.
- Pollet, Joshua M, and Mungo Wilson.** 2010. “Average correlation and stock market returns.” *Journal of Financial Economics* 96 (3): 364–380.
- Rapach, David E, Matthew C Ringgenberg, and Guofu Zhou.** 2016. “Short interest and aggregate stock returns.” *Journal of Financial Economics* 121 (1): 46–65.
- Rapach, David E, and Guofu Zhou.** 2020. “Time-series and cross-sectional stock return forecasting: New machine learning methods.” *Machine learning for asset management: New developments and financial applications* 1–33.
- Shleifer, Andrei, and Robert Vishny.** 1997. “The limits of arbitrage.” *The Journal of Finance* 52 35–55.
- Wang, Jianqiu, Zhuo Wang, and Ke Wu.** 2025. “Forecasting stock market return with anomalies: Evidence from China.” *International Journal of Forecasting*.
- Yan, Xuemin, and Lingling Zheng.** 2017. “Fundamental analysis and the cross-section of stock returns: a data-mining approach.” *The Review of Financial Studies* 30 1382–1423.

Figure 1: Histograms for out-of-sample R^2_{OS} (%) - 1,000 randomly selected groups of 100 anomalies

Histograms for the out-of-sample R^2_{OS} statistics. This figure shows histograms for R^2_{OS} (%) for forecasting market excess returns based on the cross-sectional average. The forecasts are based on 1,000 randomly selected groups of 100 long-short portfolio returns from the 120 data-mined anomalies (DM^{2-stage}) selected with the two-stage procedure (Panel A) and the 120 published anomalies (ANOM) (Panel B).

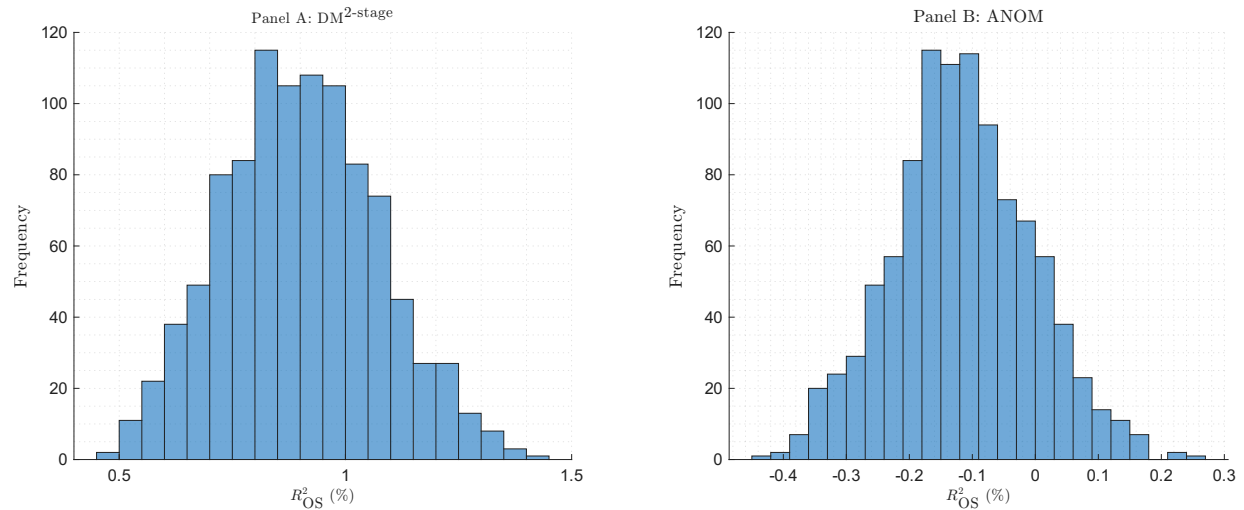


Figure 2: Slope coefficient dynamics comparisons

The panels in this figure compare the standardized recursive slope - simple average - coefficient estimates used to compute the $DM^{2\text{-stage}}$ and ANOM forecasts of the market excess return. The solid (dashed) lines are the slope estimates for the $DM^{2\text{-stage}}$ (ANOM). The five panels show the comparisons for $N = 100, 120, 140, 160$, and 180 . N is the number of data-mined anomalies with the most negative slope coefficient that are chosen in the 2nd stage.

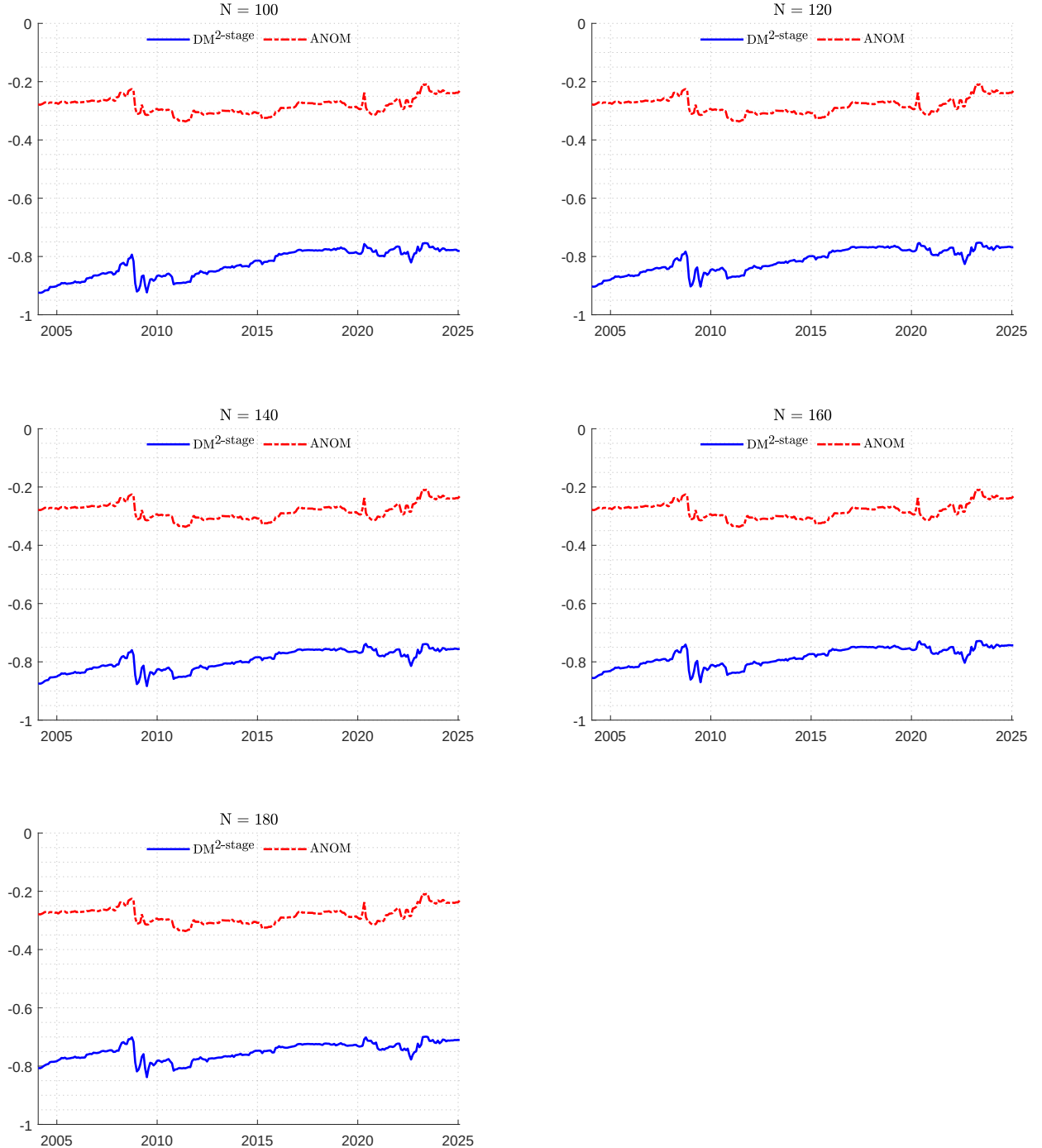


Figure 3: Cumulative wealth dynamics comparisons: cross-sectional average

The panels in this figure compare the cumulative wealth generated by strategies based on the cross-sectional average of the $DM^{2\text{-stage}}$ and the ANOM forecasts of market excess returns. The solid lines depict the cumulative wealth associated with the $DM^{2\text{-stage}}$ strategy, while the dashed lines represent the ANOM-based strategy. The five panels show the comparisons for $N = 100, 120, 140, 160$, and 180 . N is the number of data-mined anomalies with the most negative slope coefficient that are chosen in the 2nd stage.

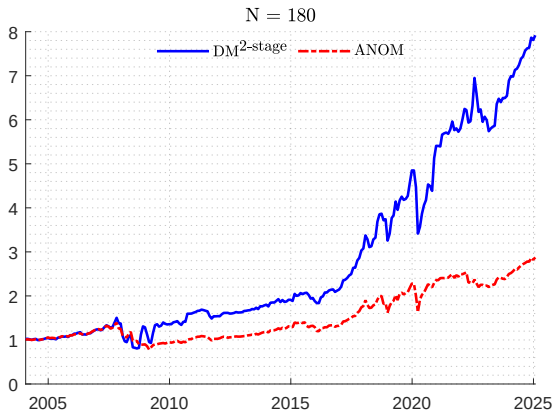
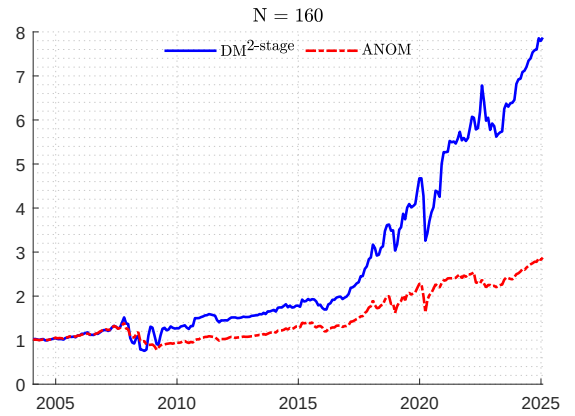
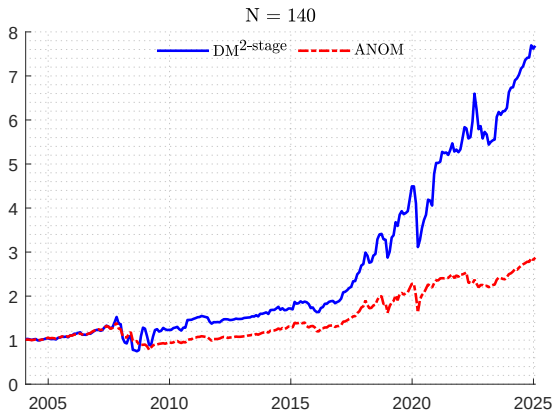
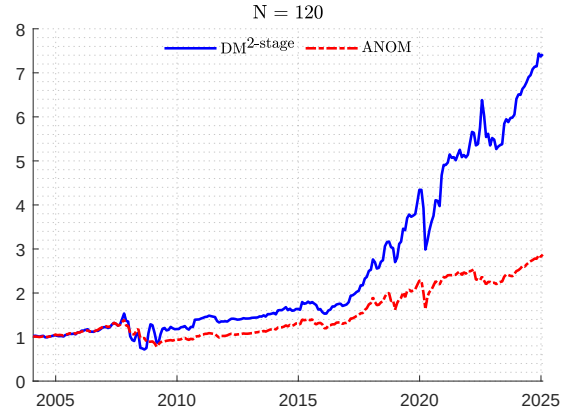
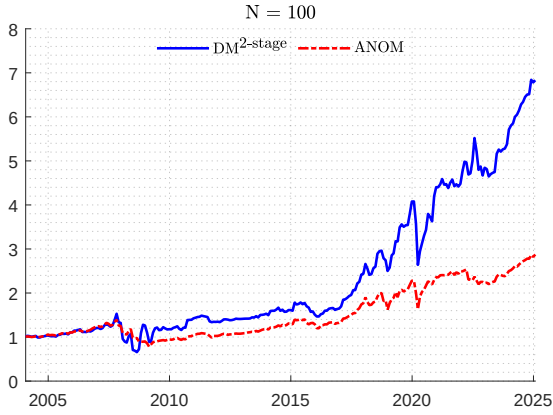


Figure 4: Cumulative wealth dynamics comparisons: C-ENet

The panels in this figure compare the cumulative wealth generated by strategies using the C-ENet ($\alpha = 0.5$) aggregation of the DM^{2-stage} and the ANOM forecasts of market excess returns. The solid lines depict the cumulative wealth associated with the DM^{2-stage} strategy, while the dashed lines represent the ANOM-based strategy. The five panels show the comparisons for $N = 100, 120, 140, 160$, and 180 . N is the number of data-mined anomalies with the most negative slope coefficient that are chosen in the 2nd stage.

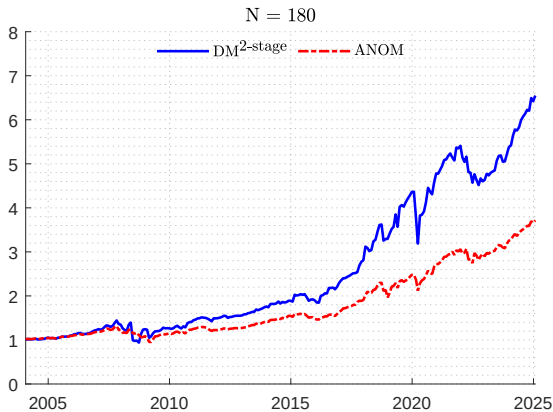
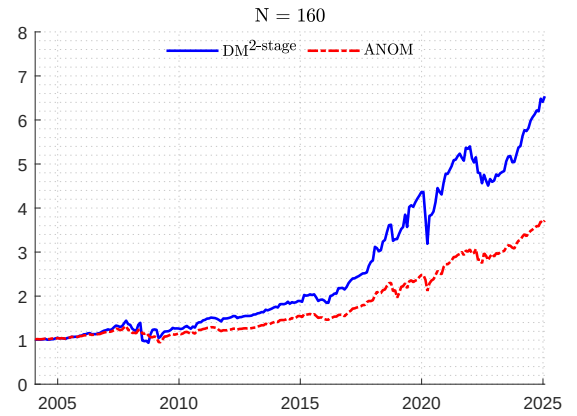
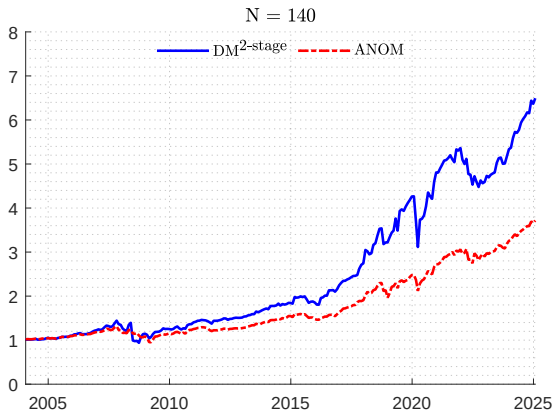
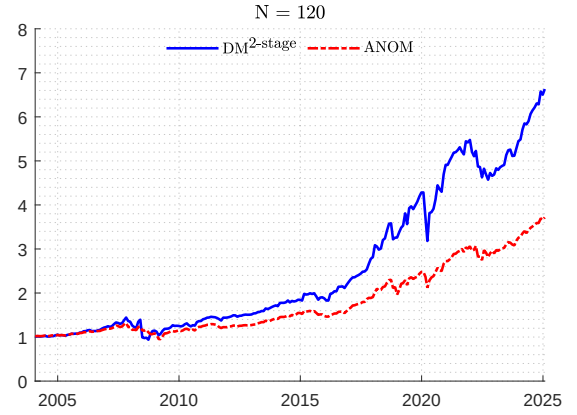
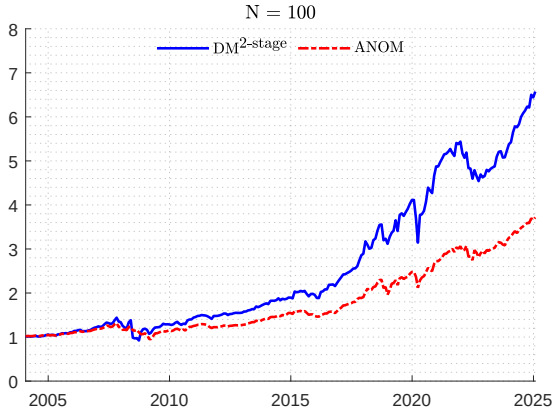


Table 1: Summary statistics

This table reports the distribution statistics of published anomalies and data-mined anomalies. Panel A reports the results of published anomalies. The full sample includes 138 anomalies. 120 anomalies remain after anomalies with missing values are removed in the early period. Panel B reports the results of data-mined anomalies. The first stage of double screening for data-mined anomalies uses the t-statistic of returns, and the second stage estimates the slope coefficient of predictive regression to select the most promising anomalies for predicting the market.

Panel A: Published anomalies

	Pre-2003		Post-2003	
	All	Valid	All	Valid
Number of anomalies	138	120	138	120
% of t-stat ≥ 2	53.6%	48.3%	12.3%	5.8%
% of t-stat ≥ 3	32.6%	26.7%	2.9%	0.0%
Mean of t-stat	2.33	2.05	0.37	0.38
Minimum of t-stat	-2.17	-2.62	-2.42	-1.91
Maximum of t-stat	11.32	7.59	4.73	2.96

Panel B: Data-mined anomalies

	Pre-2003		Post-2003	
	All	Double screened	All	Double screened
Number of anomalies	18022	120	18022	120
% of t-stat ≥ 2	22.0%	99.2%	3.4%	28.3%
% of t-stat ≥ 3	5.0%	86.7%	0.4%	5.8%
Mean of t-stat	1.30	3.84	0.01	1.14
Minimum of t-stat	0.00	1.90	-3.94	-1.89
Maximum of t-stat	7.19	5.84	3.69	3.64

Table 2: Out-of-sample R_{OS}^2 's and slope coefficients: cross-sectional average

This table reports the out-of-sample R_{OS}^2 (%), the MSFE-adjusted statistic, and the mean standardized recursive slope (Coef) of the predictive regressions across the out-of-sample period for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average		
	R_{OS}^2 (%)	MSFE-adjusted	Coef
$DM^{2\text{-stage}}, N = 100$	0.83**	1.80	-0.83
$DM^{2\text{-stage}}, N = 120$	0.91**	1.92	-0.81
$DM^{2\text{-stage}}, N = 140$	1.09**	1.98	-0.80
$DM^{2\text{-stage}}, N = 160$	1.13**	1.98	-0.78
$DM^{2\text{-stage}}, N = 180$	1.19**	1.93	-0.75
ANOM	-0.12	0.36	-0.28

Table 3: Differences in out-of-sample R_{OS}^2 's between high- and low-friction periods

This table reports the differences in out-of-sample R_{OS}^2 's between high- and low-friction periods, defined using the sample median of the financial uncertainty index of [Jurado et al. \(2015\)](#). OOS period: Jan 2004 to Jan 2025.

	Difference in R_{OS}^2 (%)
DM ^{2-stage} , N = 100	4.30
DM ^{2-stage} , N = 120	3.09
DM ^{2-stage} , N = 140	2.21
DM ^{2-stage} , N = 160	1.75
DM ^{2-stage} , N = 180	1.31
ANOM	-0.56

Table 4: Cumulative wealth: cross-sectional average

This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	6.83	1.67	4.10
$DM^{2\text{-stage}}, N = 120$	7.42	1.68	4.42
$DM^{2\text{-stage}}, N = 140$	7.68	1.76	4.37
$DM^{2\text{-stage}}, N = 160$	7.87	1.82	4.33
$DM^{2\text{-stage}}, N = 180$	7.92	1.93	4.11
ANOM	2.87	1.27	2.27

Table 5: Out-of-sample R_{OS}^2 's for 14 popular economic variables from [Goyal and Welch \(2008\)](#)

This table reports the out-of-sample R_{OS}^2 (%), and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using univariate predictive regression forecasts based on 14 economic predictors from [Goyal and Welch \(2008\)](#): the log dividend-price ratio (d/p), log dividend yield ratio (d/y), log earnings-price ratio (e/p), log dividend-payout ratio (d/e), stock return variance (svar), book-to-market ratio (b/m), net equity expansion (ntis), treasury bill rate (tbl), long-term yield (lty), long-term return (ltr), term spread (tms), default yield spread (dfy), default return spread (dfr), and inflation (Infl). Statistical significance for the R_{OS}^2 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Predictors	R_{OS}^2 (%)	MSFE-adjusted
d/p	-0.65	-0.93
d/y	-0.69	-0.75
e/p	-0.49	-0.52
d/e	-3.41	-0.35
svar	-3.23	-0.21
b/m	-0.15	-0.70
ntis	-1.31	-0.66
tbl	0.25	0.82
lty	0.28	0.77
ltr	-0.22	0.90
tms	-1.70	-0.85
dfy	-1.60	-0.75
dfr	-1.19	-0.08
Infl	-0.43	-0.76

Table 6: Out-of-sample R^2_{OS} 's for various popular predictors

This table reports the out-of-sample R^2_{OS} (%), and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using univariate predictive regression forecasts based on several popular predictors: the average correlation of daily stock returns (avgcor), nearness to Dow 52-week high: the current difference to the lifetime high (dtoat), nearness to Dow 52-week high: the scaled current difference to the 52-week high (dtoy), aggregation of 9 illiquidity measures (lzrt), ratio of new orders to shipments of durable goods (ndrbl), cross-sectional standard deviation on the set of 100 size and book-to-market portfolios (rdsp), average of monthly skewness values across firms (skvw), tail risk from cross-section (tail), and stock-bond yield gap (ygap). Statistical significance for the R^2_{OS} 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Predictors	R^2_{OS} (%)	MSFE-adjusted
avgcor	-0.85	0.36
dtoat	0.26	0.67
dtoy	-0.24	-0.30
lzrt	-1.97	0.17
ndrbl	-0.38	0.30
rdsp	-0.72	-0.44
skvw	-2.88	-0.31
tail	0.20	0.98
ygap	-0.98	-0.52

Table 7: Out-of-sample R_{OS}^2 's: two-stage vs. single-stage selection procedures

This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies selected with the two-stage procedure ($DM^{2\text{-stage}}, N$), the single-stage procedure ($DM^{1\text{-stage}}, N$), and the cross-sectional average of the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average	
	R_{OS}^2 (%)	MSFE-adjusted
$DM^{2\text{-stage}}, N = 100$	0.83**	1.80
$DM^{1\text{-stage}}, N = 100$	0.63	1.09
$DM^{2\text{-stage}}, N = 120$	0.91**	1.92
$DM^{1\text{-stage}}, N = 120$	0.73	1.14
$DM^{2\text{-stage}}, N = 140$	1.09**	1.98
$DM^{1\text{-stage}}, N = 140$	0.76	1.15
$DM^{2\text{-stage}}, N = 160$	1.13**	1.98
$DM^{1\text{-stage}}, N = 160$	0.50	0.98
$DM^{2\text{-stage}}, N = 180$	1.19**	1.93
$DM^{1\text{-stage}}, N = 180$	0.35	0.87
ANOM	-0.12	0.36

Table 8: Out-of-sample R_{OS}^2 's: C-ENet

This table reports the out-of-sample R_{OS}^2 (%), and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	C-ENet ($\alpha = 0.25$)		C-ENet ($\alpha = 0.5$)		C-ENet ($\alpha = 0.75$)	
	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj
$DM^{2\text{-stage}}, N = 100$	2.84***	3.38	2.73***	3.01	2.73***	3.22
$DM^{2\text{-stage}}, N = 120$	2.83***	3.48	2.68***	3.20	2.64***	3.37
$DM^{2\text{-stage}}, N = 140$	2.84***	3.55	2.63***	3.17	2.65***	3.38
$DM^{2\text{-stage}}, N = 160$	2.66***	3.22	2.58***	2.99	3.05***	3.10
$DM^{2\text{-stage}}, N = 180$	2.66***	3.22	2.59***	3.01	3.10***	3.17
ANOM	1.23**	1.98	0.80**	1.74	1.31***	2.34

Table 9: Cumulative wealth: C-ENet

This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	6.57	1.88	3.50
$DM^{2\text{-stage}}, N = 120$	6.63	1.83	3.62
$DM^{2\text{-stage}}, N = 140$	6.49	1.83	3.55
$DM^{2\text{-stage}}, N = 160$	6.54	1.87	3.49
$DM^{2\text{-stage}}, N = 180$	6.55	1.87	3.50
ANOM	3.72	1.49	2.49

Table 10: Out-of-sample R_{OS}^2 's and slope coefficients for ANOM

This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using the cross-sectional average and C-ENet to aggregate 10 published anomalies (ANOM^{2-stage}) selected with the two-stage procedure, and the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average		
	R_{OS}^2 (%)	MSFE-adj	Coef
ANOM ^{2-stage}	0.68	1.18	-0.40
ANOM	-0.12	0.36	-0.28
	C-ENet ($\alpha = 0.5$)		
	R_{OS}^2 (%)	MSFE-adj	Coef
ANOM ^{2-stage}	1.23*	1.64	--
ANOM	0.80**	1.74	--

Table 11: Cumulative wealth for ANOM

This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average and C-ENet to aggregate 10 published anomalies (ANOM^{2-stage}) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
ANOM ^{2-stage}	5.37	1.62	3.32
ANOM	2.87	1.27	2.27
	C-ENet ($\alpha = 0.5$), $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
ANOM ^{2-stage}	4.30	1.68	2.56
ANOM	3.72	1.49	2.49

Table 12: Out-of-sample R_{OS}^2 's with RFF

This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using the cross-sectional average and C-ENet to aggregate $p = 500$ and $p = 1,000$ random Fourier features (RFF) generated by the 120 data-mined anomalies (DM^{2-stage}, $p = 500$ & DM^{2-stage}, $p = 1000$) selected with the two-stage procedure, and by the 120 published anomalies (ANOM $p = 500$ & ANOM $p = 1000$). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average		C-ENet ($\alpha = 0.5$)	
	p = 500			
	R^2_{OS} (%)	MSFE-adj	R^2_{OS} (%)	MSFE-adj
DM ^{2-stage}	1.20**	1.94	3.36***	2.81
ANOM	0.22	1.02	2.72**	1.94
	p = 1000			
	R^2_{OS} (%)	MSFE-adj	R^2_{OS} (%)	MSFE-adj
DM ^{2-stage}	1.69**	2.13	3.50***	2.58
ANOM	0.19	0.63	2.83*	1.30

Table 13: Cumulative wealth with RFF

This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average and C-ENet to aggregate $p = 500$ and $p = 1,000$ random Fourier features (RFF) generated by the 120 data-mined anomalies ($DM^{2\text{-stage}}$, $p = 500$ & $DM^{2\text{-stage}}$, $p = 1000$) selected with the two-stage procedure, and by the 120 published anomalies (ANOM $p = 500$ & ANOM $p = 1000$). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$			C-ENet ($\alpha = 0.5$), $\lambda = 5$		
	$p = 500$			$p = 500$		
	full	1st half	2nd half	full	1st half	2nd half
$DM^{2\text{-stage}}$	5.95	1.83	3.24	9.35	2.32	4.04
ANOM	2.92	1.09	2.68	8.13	2.44	3.33
	$p = 1000$			$p = 1000$		
	full	1st half	2nd half	full	1st half	2nd half
$DM^{2\text{-stage}}$	5.34	1.88	2.84	9.68	2.44	3.98
ANOM	3.01	1.12	2.70	7.93	3.42	2.32

Online Appendix to
‘Data-mined Anomalies and the Expected Market Return’

A	Technical Details for ML Techniques	1
B	Supplementary Empirical Results	3
C	Anomalies	44
C.1	List of published anomalies	45
C.2	Data-mined anomalies	49

A Technical Details for ML Techniques

C-ENet, C-LASSO & C-Ridge

Consider the following set of univariate predictive regressions based on the individual predictors $(X_{i,t})$,

$$R_{M,t+1} = n_i + \theta_i X_{i,t} + \varepsilon_{i,t+1}, \quad \text{for } i = 1, \dots, N. \quad (21)$$

For each individual predictor, the corresponding forecast for the excess stock market return $(\hat{R}_{M,i,t+1|t})$ based on (25) is as follows:

$$\hat{R}_{M,i,t+1|t} = \hat{n}_{i,(1:t)} + \hat{\theta}_{i,(1:t)} X_{i,t}, \quad (22)$$

where $\hat{n}_{i,(1:t)}$ and $\hat{\theta}_{i,(1:t)}$ are the OLS estimates of n_i and θ_i by using data up to time t .

The Granger and Ramanathan (1984) multiple regression that relates the actual return $(R_{M,t})$ to the individual univariate predictive regression forecasts $(\hat{R}_{M,i,t|t-1})$ is given as follows:

$$R_t = \varphi + \sum_{i=1}^n f_i \hat{R}_{i,t|t-1} + \varepsilon_t. \quad (23)$$

The C-ENet forecasts are computed in three steps.

Step-1: For each predictor variable, compute recursive univariate predictive regression forecasts based on equation (26) over the holdout out-of-sample period,

$$\hat{R}_{M,i,s|s-1} = \hat{n}_{i,(1:s-1)} + \hat{\theta}_{i,(1:s-1)} X_{i,s-1}, \quad (24)$$

for $s = t_1, \dots, t$ and $i = 1, \dots, N$, where t_1 is the length of the initial in-sample estimation period.

Step-2: Estimate the Granger and Ramanathan (1984) regression in equation (27) via the ENet over the holdout out-of-sample period,

$$R_s = \varphi + \sum_{i=1}^N f_i \hat{R}_{M,i,s|s-1} + \varepsilon_s, \quad (25)$$

for $s = t_1, \dots, t$, and $I_t \subseteq 1, \dots, N$ denotes the set of individual univariate predictive regression forecasts selected by the ENet in (29).

Step-3: Compute the C-ENet forecast as follows:

$$\hat{R}_{M,t+1|t}^{C-ENet} = \frac{1}{|I_t|} \sum_{i \in I_t} \hat{R}_{i,t+1|t}, \quad (26)$$

where $|I_t|$ is the cardinality of I_t , and $\hat{R}_{M,i,t+1|t}$ is given by (26) for $i = 1, \dots, N$.

We follow the same process for generating C-LASSO forecasts, but instead of applying ENet over the holdout out-of-sample period in step-2, we use LASSO.

B Supplementary Empirical Results

Table A1: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 3$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	10.21	1.34	7.59
$DM^{2\text{-stage}}, N = 120$	12.87	1.45	8.85
$DM^{2\text{-stage}}, N = 140$	14.29	1.64	8.70
$DM^{2\text{-stage}}, N = 160$	15.06	1.75	8.60
$DM^{2\text{-stage}}, N = 180$	16.21	2.04	7.93
ANOM	3.81	1.24	3.08

Table A2: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of four, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 4$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	8.53	1.62	5.27
$DM^{2\text{-stage}}, N = 120$	9.68	1.66	5.84
$DM^{2\text{-stage}}, N = 140$	10.23	1.78	5.76
$DM^{2\text{-stage}}, N = 160$	10.58	1.85	5.70
$DM^{2\text{-stage}}, N = 180$	10.83	2.03	5.35
ANOM	3.26	1.26	2.58

Table A3: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of six, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 6$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	5.65	1.65	3.42
$DM^{2\text{-stage}}, N = 120$	6.00	1.65	3.63
$DM^{2\text{-stage}}, N = 140$	6.14	1.71	3.59
$DM^{2\text{-stage}}, N = 160$	6.26	1.75	3.57
$DM^{2\text{-stage}}, N = 180$	6.25	1.83	3.41
ANOM	2.61	1.26	2.07

Table A4: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 7$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.84	1.62	2.98
$DM^{2\text{-stage}}, N = 120$	5.07	1.62	3.14
$DM^{2\text{-stage}}, N = 140$	5.15	1.66	3.11
$DM^{2\text{-stage}}, N = 160$	5.24	1.70	3.09
$DM^{2\text{-stage}}, N = 180$	5.21	1.75	2.97
ANOM	2.42	1.25	1.93

Table A5: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 3$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	6.52	1.76	3.71
$DM^{2\text{-stage}}, N = 120$	6.90	1.79	3.85
$DM^{2\text{-stage}}, N = 140$	7.25	1.83	3.96
$DM^{2\text{-stage}}, N = 160$	7.24	1.85	3.92
$DM^{2\text{-stage}}, N = 180$	6.99	1.88	3.72
ANOM	4.81	1.71	2.81

Table A6: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of four, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 4$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	5.56	1.59	3.49
$DM^{2\text{-stage}}, N = 120$	5.91	1.64	3.60
$DM^{2\text{-stage}}, N = 140$	6.10	1.67	3.66
$DM^{2\text{-stage}}, N = 160$	6.17	1.68	3.68
$DM^{2\text{-stage}}, N = 180$	5.97	1.68	3.56
ANOM	3.61	1.50	2.40

Table A7: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.56	1.48	3.09
$DM^{2\text{-stage}}, N = 120$	4.88	1.50	3.26
$DM^{2\text{-stage}}, N = 140$	5.05	1.52	3.32
$DM^{2\text{-stage}}, N = 160$	5.05	1.52	3.32
$DM^{2\text{-stage}}, N = 180$	4.86	1.52	3.21
ANOM	3.05	1.39	2.19

Table A8: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of six, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 6$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	3.90	1.38	2.83
$DM^{2\text{-stage}}, N = 120$	4.10	1.39	2.94
$DM^{2\text{-stage}}, N = 140$	4.19	1.41	2.97
$DM^{2\text{-stage}}, N = 160$	4.20	1.42	2.96
$DM^{2\text{-stage}}, N = 180$	4.22	1.46	2.88
ANOM	2.69	1.31	2.06

Table A9: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 7$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	3.49	1.33	2.62
$DM^{2\text{-stage}}, N = 120$	3.70	1.35	2.74
$DM^{2\text{-stage}}, N = 140$	3.83	1.38	2.77
$DM^{2\text{-stage}}, N = 160$	3.86	1.39	2.77
$DM^{2\text{-stage}}, N = 180$	3.86	1.44	2.68
ANOM	2.40	1.25	1.92

Table A10: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 3$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	1.63	0.59	2.76
$DM^{2\text{-stage}}, N = 120$	2.27	0.70	3.26
$DM^{2\text{-stage}}, N = 140$	2.71	0.82	3.31
$DM^{2\text{-stage}}, N = 160$	2.96	0.89	3.34
$DM^{2\text{-stage}}, N = 180$	3.47	1.09	3.19
ANOM	2.09	1.00	2.10

Table A11: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of four, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 4$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.19	0.89	2.46
$DM^{2\text{-stage}}, N = 120$	2.66	0.96	2.76
$DM^{2\text{-stage}}, N = 140$	2.97	1.06	2.79
$DM^{2\text{-stage}}, N = 160$	3.14	1.12	2.81
$DM^{2\text{-stage}}, N = 180$	3.42	1.27	2.70
ANOM	2.08	1.07	1.93

Table A12: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.31	1.04	2.23
$DM^{2\text{-stage}}, N = 120$	2.65	1.09	2.42
$DM^{2\text{-stage}}, N = 140$	2.86	1.17	2.44
$DM^{2\text{-stage}}, N = 160$	2.99	1.22	2.46
$DM^{2\text{-stage}}, N = 180$	3.15	1.33	2.37
ANOM	2.01	1.11	1.80

Table A13: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of six, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 6$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.29	1.12	2.05
$DM^{2\text{-stage}}, N = 120$	2.54	1.16	2.20
$DM^{2\text{-stage}}, N = 140$	2.70	1.22	2.21
$DM^{2\text{-stage}}, N = 160$	2.80	1.26	2.22
$DM^{2\text{-stage}}, N = 180$	2.90	1.35	2.16
ANOM	1.93	1.13	1.71

Table A14: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 7$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.23	1.16	1.93
$DM^{2\text{-stage}}, N = 120$	2.43	1.19	2.04
$DM^{2\text{-stage}}, N = 140$	2.55	1.24	2.05
$DM^{2\text{-stage}}, N = 160$	2.62	1.28	2.06
$DM^{2\text{-stage}}, N = 180$	2.70	1.35	2.01
ANOM	1.87	1.14	1.64

Table A15: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 3$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	3.44	1.30	2.65
$DM^{2\text{-stage}}, N = 120$	3.68	1.34	2.75
$DM^{2\text{-stage}}, N = 140$	3.90	1.37	2.85
$DM^{2\text{-stage}}, N = 160$	3.93	1.39	2.83
$DM^{2\text{-stage}}, N = 180$	3.80	1.41	2.69
ANOM	3.32	1.45	2.28

Table A16: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of four, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 4$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.98	1.20	2.48
$DM^{2\text{-stage}}, N = 120$	3.17	1.25	2.54
$DM^{2\text{-stage}}, N = 140$	3.30	1.27	2.59
$DM^{2\text{-stage}}, N = 160$	3.36	1.28	2.62
$DM^{2\text{-stage}}, N = 180$	3.26	1.29	2.53
ANOM	2.58	1.32	1.96

Table A17: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 5$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.52	1.15	2.19
$DM^{2\text{-stage}}, N = 120$	2.69	1.18	2.29
$DM^{2\text{-stage}}, N = 140$	2.79	1.19	2.34
$DM^{2\text{-stage}}, N = 160$	2.81	1.20	2.34
$DM^{2\text{-stage}}, N = 180$	2.74	1.20	2.29
ANOM	2.24	1.24	1.80

Table A18: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of six, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 6, TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.25	1.10	2.03
$DM^{2\text{-stage}}, N = 120$	2.35	1.12	2.09
$DM^{2\text{-stage}}, N = 140$	2.41	1.14	2.11
$DM^{2\text{-stage}}, N = 160$	2.43	1.15	2.11
$DM^{2\text{-stage}}, N = 180$	2.47	1.19	2.08
ANOM	2.03	1.18	1.72

Table A19: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	Cross-sectional Average, $\lambda = 7$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	2.08	1.09	1.90
$DM^{2\text{-stage}}, N = 120$	2.21	1.12	1.98
$DM^{2\text{-stage}}, N = 140$	2.30	1.16	2.01
$DM^{2\text{-stage}}, N = 160$	2.33	1.16	2.01
$DM^{2\text{-stage}}, N = 180$	2.36	1.20	1.97
ANOM	1.87	1.14	1.64

Table A20: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 3$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	14.33	2.23	6.42
$DM^{2\text{-stage}}, N = 120$	14.62	2.16	6.78
$DM^{2\text{-stage}}, N = 140$	14.00	2.15	6.50
$DM^{2\text{-stage}}, N = 160$	13.85	2.18	6.34
$DM^{2\text{-stage}}, N = 180$	13.88	2.18	6.35
ANOM	6.53	1.70	3.84

Table A21: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 7$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.40	1.68	2.62
$DM^{2\text{-stage}}, N = 120$	4.42	1.64	2.69
$DM^{2\text{-stage}}, N = 140$	4.36	1.64	2.65
$DM^{2\text{-stage}}, N = 160$	4.40	1.68	2.62
$DM^{2\text{-stage}}, N = 180$	4.41	1.68	2.63
ANOM	2.85	1.40	2.04

Table A22: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of three, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 3$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	10.29	2.29	4.49
$DM^{2\text{-stage}}, N = 120$	11.46	2.38	4.82
$DM^{2\text{-stage}}, N = 140$	10.59	2.38	4.45
$DM^{2\text{-stage}}, N = 160$	10.49	2.40	4.36
$DM^{2\text{-stage}}, N = 180$	10.52	2.40	4.37
ANOM	5.39	1.68	3.21

Table A23: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	5.87	1.73	3.40
$DM^{2\text{-stage}}, N = 120$	6.24	1.78	3.51
$DM^{2\text{-stage}}, N = 140$	5.78	1.77	3.26
$DM^{2\text{-stage}}, N = 160$	5.36	1.67	3.20
$DM^{2\text{-stage}}, N = 180$	5.37	1.67	3.21
ANOM	3.52	1.50	2.34

Table A24: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of seven, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100$, 120, 140, 160, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet ($\alpha = 0.5$), $\lambda = 7$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.09	1.53	2.68
$DM^{2\text{-stage}}, N = 120$	4.26	1.56	2.73
$DM^{2\text{-stage}}, N = 140$	4.17	1.56	2.68
$DM^{2\text{-stage}}, N = 160$	3.96	1.50	2.64
$DM^{2\text{-stage}}, N = 180$	3.97	1.50	2.65
ANOM	2.78	1.41	1.98

Table A25: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet, $\lambda = 5$, $TCs = 50$ bps		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.31	1.57	2.74
$DM^{2\text{-stage}}, N = 120$	4.17	1.52	2.75
$DM^{2\text{-stage}}, N = 140$	3.94	1.51	2.60
$DM^{2\text{-stage}}, N = 160$	3.92	1.52	2.57
$DM^{2\text{-stage}}, N = 180$	3.92	1.52	2.57
ANOM	2.67	1.27	2.09

Table A26: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-ENet, $\lambda = 5$, $TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.17	1.51	2.76
$DM^{2\text{-stage}}, N = 120$	4.30	1.54	2.80
$DM^{2\text{-stage}}, N = 140$	3.91	1.54	2.54
$DM^{2\text{-stage}}, N = 160$	3.59	1.43	2.51
$DM^{2\text{-stage}}, N = 180$	3.59	1.43	2.51
ANOM	2.60	1.29	2.02

Table A27: This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using C-LASSO and C-Ridge to aggregate $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. The sample spans the period from January 1969 to January 2025. OOS period: January 2004 to January 2025. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	C-LASSO		C-Ridge	
	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj
$DM^{2\text{-stage}}, N = 100$	2.65***	3.36	1.65***	2.91
$DM^{2\text{-stage}}, N = 120$	2.36***	3.35	1.68***	3.10
$DM^{2\text{-stage}}, N = 140$	2.40***	3.44	1.76***	3.23
$DM^{2\text{-stage}}, N = 160$	2.83***	3.02	1.64***	3.14
$DM^{2\text{-stage}}, N = 180$	2.83***	3.02	1.56***	3.05
ANOM	1.30**	2.32	0.72***	1.92

Table A28: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-LASSO to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-LASSO, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	6.48	1.87	3.47
$DM^{2\text{-stage}}, N = 120$	6.00	1.71	3.50
$DM^{2\text{-stage}}, N = 140$	6.07	1.71	3.54
$DM^{2\text{-stage}}, N = 160$	7.30	2.01	3.64
$DM^{2\text{-stage}}, N = 180$	7.30	2.01	3.64
ANOM	4.11	1.53	2.69

Table A29: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-LASSO to aggregate $N = 100$, 120, 140, 160, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-LASSO, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	6.05	1.81	3.35
$DM^{2\text{-stage}}, N = 120$	5.34	1.66	3.21
$DM^{2\text{-stage}}, N = 140$	5.41	1.66	3.25
$DM^{2\text{-stage}}, N = 160$	5.53	1.67	3.30
$DM^{2\text{-stage}}, N = 180$	5.53	1.67	3.30
ANOM	3.80	1.54	2.46

Table A30: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-Ridge to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-Ridge, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	5.00	1.58	3.17
$DM^{2\text{-stage}}, N = 120$	4.91	1.53	3.21
$DM^{2\text{-stage}}, N = 140$	5.03	1.59	3.16
$DM^{2\text{-stage}}, N = 160$	4.81	1.57	3.06
$DM^{2\text{-stage}}, N = 180$	4.71	1.57	3.00
ANOM	3.35	1.31	2.56

Table A31: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-Ridge to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-Ridge, $\lambda = 5$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.61	1.50	3.07
$DM^{2\text{-stage}}, N = 120$	4.73	1.51	3.13
$DM^{2\text{-stage}}, N = 140$	4.82	1.54	3.13
$DM^{2\text{-stage}}, N = 160$	4.68	1.55	3.02
$DM^{2\text{-stage}}, N = 180$	4.60	1.54	2.99
ANOM	3.33	1.30	2.56

Table A32: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-LASSO to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-LASSO, $\lambda = 5$, $TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.10	1.54	2.66
$DM^{2\text{-stage}}, N = 120$	3.60	1.40	2.57
$DM^{2\text{-stage}}, N = 140$	3.63	1.40	2.60
$DM^{2\text{-stage}}, N = 160$	4.25	1.62	2.63
$DM^{2\text{-stage}}, N = 180$	4.25	1.62	2.63
ANOM	2.91	1.29	2.25

Table A33: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-LASSO to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-LASSO, $\lambda = 5$, $TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	4.17	1.56	2.67
$DM^{2\text{-stage}}, N = 120$	3.56	1.42	2.52
$DM^{2\text{-stage}}, N = 140$	3.60	1.42	2.54
$DM^{2\text{-stage}}, N = 160$	3.62	1.42	2.56
$DM^{2\text{-stage}}, N = 180$	3.62	1.42	2.55
ANOM	2.77	1.31	2.12

Table A34: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which allows for short selling and leverage, for predicting the average monthly excess stock market return using C-Ridge to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-Ridge, $\lambda = 5$, $TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	3.82	1.41	2.70
$DM^{2\text{-stage}}, N = 120$	3.82	1.38	2.76
$DM^{2\text{-stage}}, N = 140$	3.92	1.44	2.73
$DM^{2\text{-stage}}, N = 160$	3.80	1.42	2.68
$DM^{2\text{-stage}}, N = 180$	3.76	1.42	2.65
ANOM	2.74	1.20	2.28

Table A35: This table reports the cumulative wealth (CW) for a mean-variance investor with a risk-aversion coefficient of five, which prohibits short selling and allows leverage of up to 50%, for predicting the average monthly excess stock market return using C-Ridge to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). We consider transaction costs (TCs) of 50 bps. The sample spans the period from January 1969 to January 2025. OOS period: Jan 2004 to Jan 2025.

	C-Ridge, $\lambda = 5$, $TCs = 50 \text{ bps}$		
	CW (full)	CW (1st half)	CW (2nd half)
$DM^{2\text{-stage}}, N = 100$	3.64	1.37	2.65
$DM^{2\text{-stage}}, N = 120$	3.78	1.39	2.72
$DM^{2\text{-stage}}, N = 140$	3.86	1.42	2.72
$DM^{2\text{-stage}}, N = 160$	3.78	1.42	2.66
$DM^{2\text{-stage}}, N = 180$	3.75	1.42	2.65
ANOM	2.74	1.20	2.29

Table A36: This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. OOS period: Jan 2004 to Jun 2014 (1st half). *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average (1st half)	
	R_{OS}^2 (%)	MSFE-adjusted
$DM^{2\text{-stage}}, N = 100$	0.33	0.87
$DM^{2\text{-stage}}, N = 120$	0.41	0.94
$DM^{2\text{-stage}}, N = 140$	0.79	1.03
$DM^{2\text{-stage}}, N = 160$	0.95	1.06
$DM^{2\text{-stage}}, N = 180$	1.44	1.14
ANOM	0.68	1.04

Table A37: This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using the cross-sectional average of $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the cross-sectional average of the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of Clark and West (2007)'s MSFE-adjusted statistic. OOS period: Jul 2014 to Jan 2025 (2nd half). *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Cross-sectional Average (2nd half)	
	R_{OS}^2 (%)	MSFE-adjusted
$DM^{2\text{-stage}}, N = 100$	0.97**	1.88
$DM^{2\text{-stage}}, N = 120$	1.07**	1.94
$DM^{2\text{-stage}}, N = 140$	1.11**	1.94
$DM^{2\text{-stage}}, N = 160$	1.04**	1.94
$DM^{2\text{-stage}}, N = 180$	0.76**	1.78
ANOM	-0.88	-0.43

Table A38: This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160,$ and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. OOS period: Jan 2004 to Jun 2014 (1st half). *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	C-ENet ($\alpha = 0.25$)		C-ENet ($\alpha = 0.5$)		C-ENet ($\alpha = 0.75$)	
	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj
$DM^{2\text{-stage}}, N = 100$	3.58**	2.19	3.44**	2.10	3.32**	2.17
$DM^{2\text{-stage}}, N = 120$	3.44**	2.21	3.30**	2.25	3.01**	2.17
$DM^{2\text{-stage}}, N = 140$	3.58***	2.33	3.30**	2.25	3.01**	2.16
$DM^{2\text{-stage}}, N = 160$	3.28**	2.04	3.34**	2.08	3.80**	2.00
$DM^{2\text{-stage}}, N = 180$	3.28**	2.04	3.34**	2.08	3.80**	2.00
ANOM	1.89*	1.62	1.24*	1.35	1.44*	1.59

Table A39: This table reports the out-of-sample R_{OS}^2 (%) and the MSFE-adjusted statistic for predicting the average monthly excess stock market return using C-ENet to aggregate $N = 100, 120, 140, 160$, and 180 data-mined anomalies ($DM^{2\text{-stage}}, N$) selected with the two-stage procedure, and the 120 published anomalies (ANOM). Statistical significance for the R_{OS}^2 's is based on the p-value of [Clark and West \(2007\)](#)'s MSFE-adjusted statistic. OOS period: Jul 2014 to Jan 2025 (2nd half). *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	C-ENet ($\alpha = 0.25$)		C-ENet ($\alpha = 0.5$)		C-ENet ($\alpha = 0.75$)	
	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj	R_{OS}^2 (%)	MSFE-adj
$DM^{2\text{-stage}}, N = 100$	2.08***	3.17	1.99**	2.17	2.09***	2.42
$DM^{2\text{-stage}}, N = 120$	2.16***	3.21	2.02**	2.26	2.20***	2.64
$DM^{2\text{-stage}}, N = 140$	2.08***	3.03	1.91**	2.21	2.22***	2.65
$DM^{2\text{-stage}}, N = 160$	2.00***	2.93	1.79**	2.24	2.27***	2.66
$DM^{2\text{-stage}}, N = 180$	1.99***	2.92	1.80**	2.27	2.37***	2.82
ANOM	0.63	1.20	0.41	1.18	1.18**	1.74

Figure A1: Histograms for the out-of-sample R_{OS}^2 statistics. This figure shows histograms for R_{OS}^2 (%) for forecasting market excess returns based on the cross-sectional average. The forecasts are based on 1,000 randomly selected groups of 80 long-short portfolio returns from the 120 data-mined anomalies (DM^{2-stage}) selected with the two-stage procedure (Panel A) and the 120 published anomalies (ANOM) (Panel B).

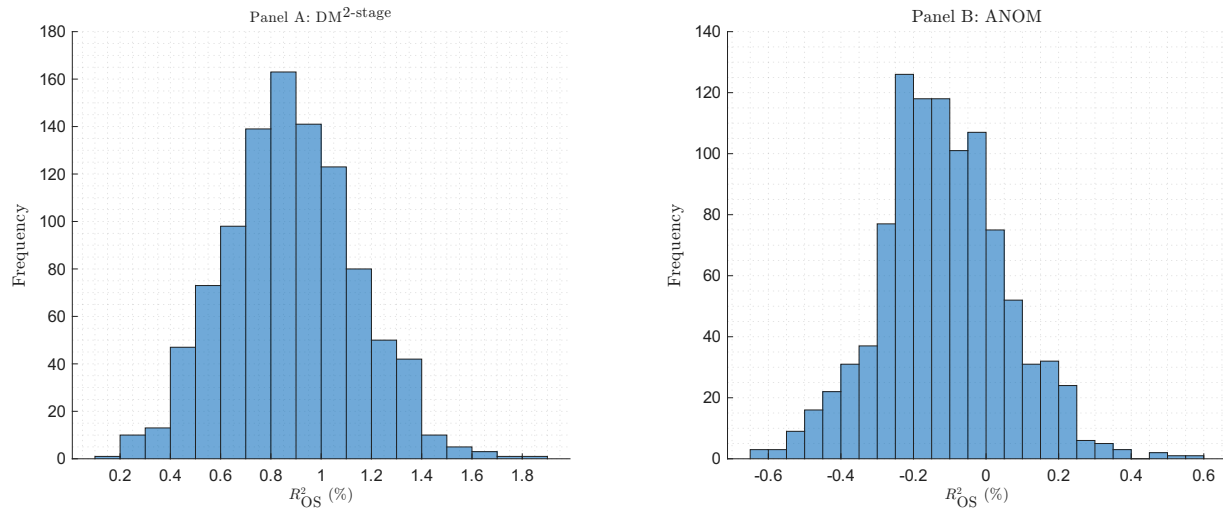
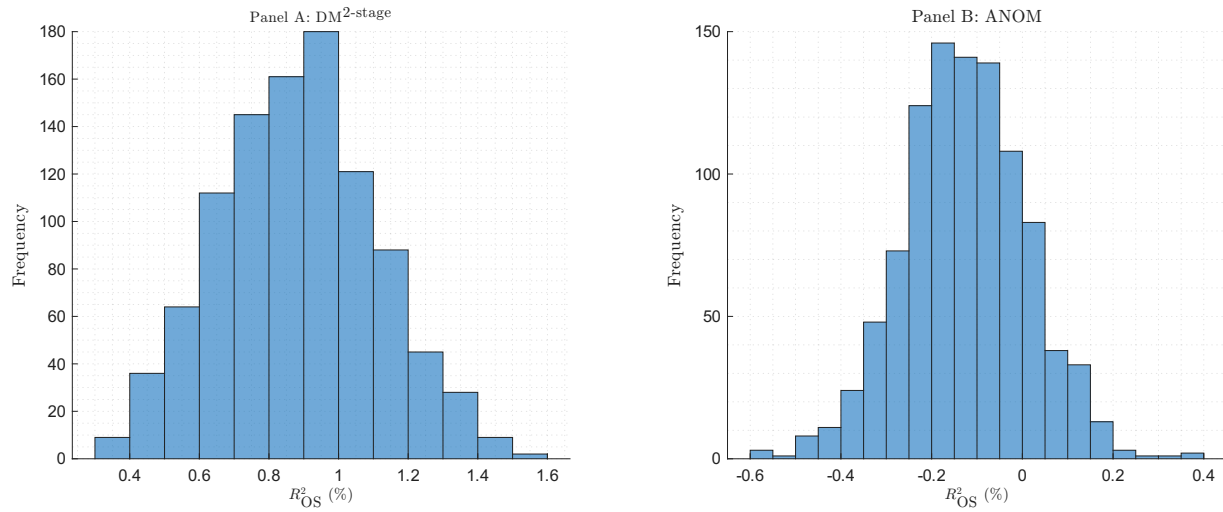


Figure A2: Histograms for the out-of-sample R_{OS}^2 statistics. This figure shows histograms for R_{OS}^2 (%) for forecasting market excess returns based on the cross-sectional average. The forecasts are based on 1,000 randomly selected groups of 90 long-short portfolio returns from the 120 data-mined anomalies (DM^{2-stage}) selected with the two-stage procedure (Panel A) and the 120 published anomalies (ANOM) (Panel B).



C Anomalies

C.1 List of published anomalies

The table in this section shows the 138 published anomalies used in our study and their descriptions. We follow [Chen and Zimmermann \(2022\)](#) to construct the anomalies and their Table 1 of online appendix provided the detailed definitions (which is available via the link: <https://drive.google.com/file/d/1dqqQD7fjMuy0XN9aDyEHxCxVJIGonXQF/view>). The code of our anomaly construction is public available from this GitHub repo: <https://github.com/mk0417/anomalies-replication-python>.

Anomaly	Description	Anomaly	Description
abnormal_accruals	Abnormal accruals	abnormal_accruals_pct	Percent abnormal accruals
accruals	Accruals	am	Total assets to market
ann_ret	Earnings announcement return	asset_growth	Asset growth
beta_liq_ps	Pastor-Stambaugh liquidity beta	beta_tail_risk	Tail risk beta
beta_vix	Systematic volatility	bm	Book to market, original (Stattman 1980)
bm_rrl	Book to market using most recent ME	bm_dec	Book to market using December ME
book_leverage	Book leverage (annual)	bpebm	Leverage component of BM
cash	Cash to assets	cash_prod	Cash Productivity
cb_oper_prof	Cash-based operating profitability	cf	Cash flow to market
cfp	Operating Cash flows to price	ch_asset_turnover	Change in Asset Turnover
ch_eq	Growth in book equity	ch_inv	Inventory Growth
ch_inv_ia	Change in capital inv (ind adj)	ch_nncoa	Change in Net Noncurrent Op Assets
ch_nwc	Change in Net Working Capital	ch_tax	Change in Taxes
comp_equ_iss	Composite equity issuance	comp_debt_iss	Composite debt issuance
coskew_acx	Coskewness using daily returns	del_coa	Change in current operating assets
del_col	Change in current operating liabilities	del_equ	Change in equity to assets
del_finl	Change in financial liabilities	del_lti	Change in long-term investment
del_net_fin	Change in net financial assets	dnoa	change in net operating assets
dol_vol	Past trading volume	earnings_consis	Earnings consistency
earnings_surp	Earnings Surprise	ebm	Enterprise component of BM
ent_mult	Enterprise Multiple	ep	Earnings-to-Price Ratio
equity_duration	Equity Duration	firm_age_mom	Firm Age - Momentum
frontier	Efficient frontier index	gp	gross profits / total assets
gr_ad_exp	Growth in advertising expenses	gr_capx	Change in capex (two years)
gr_capx_1yr	Change in capex (two years)	gr_capx_3yr	Change in capex (three years)
herf	Industry concentration (sales)	herf_be	Industry concentration (equity)

Anomaly	Description	Anomaly	Description
hire	Employment growth	idio_vol_ff3	Idiosyncratic risk (3 factor)
idio_vol_aht	Idiosyncratic risk (AHT)	illiquidity	Amihud's illiquidity
intan_bm	Intangible return using BM	intan_cfp	Intangible return using CFtoP
intan_ep	Intangible return using EP	intan_sp	Intangible return using Sale2P
int_mom	Intermediate Momentum	investment	Investment to revenue
invest_ppe_inv	change in ppe and inv/assets	inv_growth	Inventory Growth
leverage	Market leverage	lt_reversal	Long-run reversal
max_ret	Maximum return over month	mean_rank_rev_growth	Revenue Growth Rank
mom12m	Momentum (12 month)	mom12m_off_season	Momentum without the seasonal part
mom6m	Momentum (6 month)	mom6m_junk	Junk Stock Momentum
mom_off_season	Off season long-term reversal	mom_off_season_6yr_plus	Off season reversal years 6 to 10
mom_off_season_16yr_plus	Off season reversal years 16 to 20	mom_season	Return seasonality years 2 to 5
mom_season_6yr_plus	Return seasonality years 6 to 10	mom_season_11yr_plus	Return seasonality years 11 to 15
mom_season_16yr_plus	Return seasonality years 16 to 20	mom_season_short	Return seasonality last year
net_debt_fin	Net debt financing	net_debt_price	Net debt to price
net_equity_fin	Net equity financing	net_payout_yield	Net Payout Yield
noa	Net Operating Assets	op_leverage	Operating leverage
order_backlog	Order backlog	order_backlog_chg	Change in order backlog
org_cap	Organizational capital	org_cap_noadj	Organizational capital
payout_yield	Payout Yield	pct_acc	Percent Operating Accruals
pct_tot_acc	Percent Total Accruals	price	Price
rd	R&D over market cap	rd_ability	R&D ability
rd_cap	R&D capital-to-assets	realized_vol	Realized (Total) Volatility
resid_mom6m_ff3	Momentum based on FF3 residuals	resid_mom12m_ff3	Momentum based on FF3 residuals

Anomaly	Description	Anomaly	Description
return_skew	Return skewness	return_skew_ff3	Idiosyncratic skewness (3F model)
revenue_surp	Revenue Surprise	roa_q	Return on assets (qtrly)
share_iss_1yr	Share issuance (1 year)	share_iss_5yr	Share issuance (5 year)
size	Size	std_turn	Share turnover volatility
st_reversal	Short term reversal	tax	Taxable income to income
total_accruals	Total accruals	trend_factor	Trend Factor
vol_std	Volume Variance	x_fin	Net external financing
zero_trade6m	Days with zero trades	zero_trade1m	Days with zero trades
zero_trade12m	Days with zero trades	ad_exp	Advertising Expense
beta	CAPM beta	beta_fp	Frazzini-Pedersen Beta
coskew	Coskewness	del_drc	Deferred Revenue
firm_age	Firm age based on CRSP	gr_lt_noa	Growth in long term operating assets
gr_sale_to_gr_inv	Sales growth over inventory growth	gr_sale_to_gr_overhead	Sales growth over overhead growth
herf_asset	Industry concentration (assets)	high52	52 week high
mom_off_season_11yr_plus	Off season reversal years 11 to 15	mom_reversal	Medium-run reversal
oper_prof	operating profits / book equity	oper_prof_rd	Operating profitability R&D adjusted
real_estate	Real estate holdings	roe	net income / book equity
sp	Sales-to-price	var_cf	Cash-flow to price variance
vol_mkt	Volume to market equity	vol_trend	Volume Trend

C.2 Data-mined anomalies

This section provided more details of the accounting variables that are used to construct the data-mined anomalies. The data-mined anomalies are constructed using the following methods: ratio of Z to Y (Z/Y), the change in Z/Y , the percent change in Z/Y , the change in Z scaled by lagged Y , the percent change in Z minus the percent change in Y , and the percent change in Z . As there are more than 20,000 data-mined anomalies, we do not list them one by one. Instead, we list the Compustat items for both Z and Y .

The Z variables include: acchg, acdo, aco, acodo, acominc, acox, act, aldo, am, ao, aocidergl, aociother, aocipen, aocisecgl, aodo, aoloch, aox, ap, apalch, aqc, at, bkvlps, caps, capx, capxv, ceq, ceql, ceqt, ch, che, chech, cld2, cld3, cld4, cld5, cogs, cstc, cstkev, cstke, dc, dclo, dcom, dcvsr, dcvsb, dcvt, dd, dd1, dd2, dd3, dd4, dd5, diladj, dilavx, dlc, dlch, dltis, dlto, dltp, dltr, dlth, dm, dn, do, donr, dp, dpact, dpc, dpvieb, drc, drlt, ds, dudd, dv, dvc, dvp, dvpa, dvt, dxd2, dxd3, dxd4, dxd5, ebit, ebitda, epsfi, epsfx, epspi, epspx, esopct, esopnr, esopr, esopt, esub, esubc, exre, fatb, fatc, fate, fatl, fatn, fato, fatp, fiaf, fincf, fopo, fopox, gdw, gp, ib, ibadj, ibc, ibcom, icapt, idit, intan, intano, intc, intpn, invch, invfg, invo, invrm, invt, invwip, itcb, itci, ivaco, ivaeq, ivao, ivch, ivncf, ivst, ivstch, lco, lcox, lcoxdr, lct, lifr, lo, loxdr, lse, lt, mib, mii, mrc1, mrc2, mrc3, mrc4, mrc5, mrct, mrcta, msa, ni, niadj, nopi, nopio, np, oancf, oiadp, oibdp, opeps, oprepsx, optprcey, pi, pidom, pifo, ppegt, ppent, ppeveb, prstkc, pstk, pstkc, pstkl, pstkn, pstkr, pstkrv, rca, rcd, rceps, rcp, rdip, rdipa, rdipd, rdipeps, re, rea, reajo, recch, recco, recd, rect, recta, rectr, reuna, revt, sale, seq, seqo, siv, spi, sppe, sppiv, sstk, stkco, tlc, tsk, tskc, tskp, txach, txbco, txc, txdb, txdba, txdc, txdfed, txdf, txdi, txditc, txds, txfed, txfo, txndb, txndba, txndbl, txndbr, txo, txp, txpd, txr, txs, txt, txw, weap, xacc, xad, xi, xido, xidoc, xint, xintd, xopr, xpp, xpr, xrd, xrent, xsga

The Y variables include: at, act, invt, ppent, lt, lct, dlth, ceq, seq, icapt, sale, cogs, xsga, emp, me

Table C2: This table reports the distribution statistics of data-mined anomalies using two-stage screening. The first stage of double screening for the data-mined anomalies uses the t-statistic of returns, and the second stage estimates the slope coefficient of predictive regression to select the most promising anomalies for predicting the market.

	Pre-2003				Post-2003			
Num of selected anomalies	100	140	160	180	100	140	160	180
% of t-stat ≥ 2	99.0%	99.3%	99.4%	99.4%	30.0%	25.0%	21.9%	20.6%
% of t-stat ≥ 3	86.0%	85.7%	84.4%	84.4%	6.0%	5.0%	4.4%	3.9%
Mean of t-stat	3.85	3.81	3.81	3.80	1.16	1.08	1.00	0.97
Minimum of t-stat	1.90	1.90	1.90	1.89	-1.89	-1.89	-1.89	-1.89
Maximum of t-stat	5.84	5.84	6.84	6.84	3.64	3.64	3.64	3.64

