Review of Finance, 2024, 1–44
https://doi.org/10.1093/rof/rfad025
Advance Access Publication Date: 9 August 2023

Do Anomalies Really Predict Market Returns? New Data and New Evidence

Nusret Cakici[1], Christian Fieberg[2,3,4], Daniel Metko[5], and Adam Zaremba[6,7,8]

[1]Gabelli School of Business, Fordham University, New York, NY, USA, [2]School of International Business, City University of Applied Sciences, Bremen, Germany, [3]Department of Economics and Management, University of Luxembourg, Luxembourg, Luxembourg, [4]Department of Finance, Concordia University, Montreal, Canada, [5]Faculty of Business Studies and Economics, University of Bremen, Bremen, Germany, [6]Montpellier Business School, Montpellier, France, [7]Department of Investment and Financial Markets, Institute of Finance, Poznan University of Economics and Business, Poznań, Poland and [8]Department of Finance and Tax, Faculty of Commerce, University of Cape Town, South Africa

Abstract

Using new data from US and global markets, we revisit market risk premium predictability by equity anomalies. We apply a repertoire of machine-learning methods to forty-two countries to reach a simple conclusion: anomalies, as such, cannot predict aggregate market returns. Any ostensible evidence from the USA lacks external validity in two ways: it cannot be extended internationally and does not hold for alternative anomaly sets—regardless of the selection and design of factor strategies. The predictability—if any—originates from a handful of specific anomalies and depends heavily on seemingly minor methodological choices. Overall, our results challenge the view that anomalies as a group contain helpful information for forecasting market risk premia.

Keywords: Equity anomalies, Return predictability, Machine learning, International stock markets, Equity premium

JEL classification: G11, G12, G14

Received April 23, 2023; accepted August 8, 2023 by Editor Marcin Kacperczyk.

1. Introduction

Asset pricing literature typically views return predictability from two perspectives. The first strain examines *cross-sectional* predictability, exploring whether stock characteristics assist in answering why some stocks outperform others.[1] The other focuses on the *time-series* dynamics of the market equity premium.[2] A few recent studies have attempted to link these two seemingly disjointed lines of research. Engelberg *et al.* (2023) scrutinize whether aggregate versions of cross-sectional stock characteristics make good time-series predictors. They find little evidence to support this thesis. In turn, Dong *et al.* (2022) investigate the information content of cross-sectional anomaly strategies. Using a large pool of factor returns and various forecasting models, they document that past returns on long–short anomaly portfolios help to predict the market risk premium within the US market.

In this article, we comprehensively reexamine market return predictability by equity anomalies. Using new data from both US and international markets, we analyze hundreds of anomalies in forty-two countries around the world. Our baseline sample contains more than 80,000 stocks, 10 million return observations, and 1 billion monthly stock characteristics. With these data at hand, we take advantage of machine-learning models and forecast market equity premia based on anomaly portfolio returns. With a focus on variable selection and dimension reduction techniques, machine learning is well equipped to tackle challenging prediction problems by reducing degrees of freedom and condensing redundant variation among predictors (Gu, Kelly, and Xiu, 2020). Finally, we employ various evaluation methods: out-of-sample R^2 (R^2_{OS}), utility gains, and Sharpe ratio comparisons.

Our findings yield a simple, yet unequivocal conclusion: *equity anomalies*, *as such*, *cannot predict market returns*. Any apparent predictability lacks external validity in two critical aspects: stock market selection and anomaly sample. While some evidence may be spotted in individual markets—such as the USA—it originates from a handful of specific anomalies and depends heavily on seemingly unimportant methodological choices.

First, the return predictability does not extend to international stock markets. To demonstrate this, we apply the models from Dong *et al.* (2022): ordinary least squares (OLS) regressions, elastic net (ENet), simple combination (Comb), combination ENet (C-ENet), predictor average (Avg), principal component (PC) regression, and partial least squares (PLS). Using data from 1990 to 2021, we feed these models with up to 153 anomaly portfolios from Jensen, Kelly, and Pedersen (2023). We aim to verify whether any useful information can be extracted from factor returns.

We find no robust evidence of market risk premium predictability in global markets. The average R^2_{OS}s across all forty-two countries in our sample are *negative* or indistinguishable from zero. The few significant R^2_{OS} coefficients seem to be almost randomly scattered across countries and models. In other words, there is no systematic international evidence that machine-learning models based on equity anomalies reliably forecast market portfolio returns. These negative conclusions are supported by alternative measures of prediction performance. An overview of strategies' utility gains and Sharpe ratios confirms that the equity anomalies have negligible practical value for international investors. Pursuing anomaly-based forecasts fails to generate measurable utility gains within global markets. Furthermore, it does not lead to reliable improvement in the risk-adjusted portfolio performance. The Sharpe ratios on portfolios that are formed using anomaly-based predictions

[1] See, for example, Fama and French (1993, 2015, 2018); Harvey, Liu, and Zhu (2016); Green, Hand, and Zhang (2017); Hou, Xue, and Zhang (2020); and Chen and Zimmermann (2022).

[2] See, for example, Rapach, Strauss, and Zhou (2010, 2013); Rapach, Wohar, and Rangvid (2005); Campbell and Thomson (2008); Welch and Goyal (2008); Hollstein *et al.* (2020); and Goyal, Welch, and Zafirov (2021).

fail to beat simple naïve strategies, such as buy-and-hold or allocations based on predictions from a historical mean model.

Our findings are robust to various methodological modifications. They hold not only for long–short portfolios but also for long and short legs separately. They do not depend on specific portfolio breakpoints used to form anomaly strategies—surviving both in tercile and quintile portfolios. Finally, they are qualitatively influenced neither by choice of a recursive or rolling training window nor by the precise length of these windows.

The second major weakness of the market return predictability concerns the anomaly selection: it fails to work for alternative anomaly sets. To illustrate this, we revisit the playing field from the seminal paper by Dong *et al.* (2022): the US stock market from 1970 to 2017. Besides the anomalies from Dong *et al.* (2022), we scrutinize the performance of the most well-known anomaly sets from the asset pricing literature: 153 tercile and decile portfolios based on data from Jensen, Kelly, and Pedersen (2023), 207 anomalies from Chen and Zimmermann (2022), and 188 factors from Hou, Xue, and Zhang (2020). Among all the tested samples, the return predictability holds only for one in four: the original sample of Dong *et al.* (2022). No other anomaly set generates any evidence of a similar pattern.

Though the predictability cannot be detected for full alternative anomaly sets, perhaps it holds for some of its parts. Imaginably, some combinations of individual factors could generate reliable forecasts. To scrutinize this conjecture, we randomly select samples of 100 anomalies from various sets. We find that the selection dimension indeed affects return predictability. Modifications of the anomaly sets can lead to considerable differences in forecasting performance. For certain models, such as ENet or C-ENet, the R^2_{OS} can span from deeply negative to slightly positive—depending on which anomalies are supplied to the model. Nevertheless, despite this apparent dispersion, our primary conclusion still prevails: virtually no combination of factors can produce reliable forecasts. Typically, even the best percentile of the random draws fails to generate significantly positive R^2_{OS}.

Importantly, the prediction performance may depend not only on *which* anomaly portfolios are picked but also on *how* they are formed. Arbitrary decisions concerning portfolio construction may markedly influence their return distributions (Walter, Weber, and Weiss, 2022; Menkveld *et al.*, 2023; Soebhag, van Vliet, and Verwijmeren, 2023). Do they also matter for market return predictability? Perhaps, some specific construction of factor portfolios sets the stage for reliable forecasts. To investigate this possibility, we consider various methodological choices that are commonly applied to anomaly portfolio design in the asset pricing literature. Using the 153 signals from Jensen, Kelly, and Pedersen (2023) and 207 from Chen and Zimmermann (2022), we investigate, for example, different weighting schemes, stock price limits, winsorization rules, or cut-off points. Rather than checking individual specifications, we take a holistic approach—amassing 2,592 distinct implementations of factor strategies.

The outcomes of this experiment confirm that the portfolio design affects the prediction performance. The same forecasting model, supplied with an identical selection of anomalies, may have R^2_{OS} differing by even several percentage points. Nonetheless—again—our baseline inference remains intact: anomalies are not reliable predictors of market portfolio returns. Hardly any implementation of the hundreds of factor portfolios from Jensen, Kelly, and Pedersen (2023) and Chen and Zimmermann (2022) can generate reliable forecasts. Moreover, only a handful can compare with the predictive abilities of the 100 portfolios from Dong *et al.* (2022). Last, we observe no systematic difference between the predictability of significant and insignificant anomalies, refuting the mispricing correction mechanism, which serves as the theoretical foundation for market return predictability by anomaly portfolios.

Our study is most closely related to Dong *et al.* (2022). Our findings challenge their conclusions that factor returns contain information about future market portfolio performance. On the other hand, our results align with Engelberg *et al.* (2023), who argue that information from the cross-section, that is, stock characteristics, has little value for forecasting aggregate market risk premia.

From a broader perspective, our article bridges the two strains of anomaly research: on cross-sectional predictability of stock returns (e.g., Fama and French, 1993, 2015, 2018; Bali, Engle, and Murray, 2016; Harvey, Liu, and Zhu, 2016; Green, Hand, and Zhang, 2017; Hou, Xue, and Zhang, 2020; Baltussen, Swinkels, and Van Vliet, 2021; Chen and Zimmermann, 2022) and forecasting market risk premium in the time series (e.g., Rapach, Wohar, and Rangvid, 2005; Rapach, Strauss, and Zhou, 2010, 2013; Campbell and Thompson, 2008; Welch and Goyal, 2008; Hollstein *et al.*, 2020; Goyal, Welch, and Zafirov, 2021; Engelberg *et al.*, 2023). From the perspective of our findings, these two worlds remain disconnected; the anomaly returns do not help predict aggregate market returns.

Last, our article contributes to the long-standing discussion on the replication crisis in finance and data snooping within anomaly research (Lo and MacKinlay, 1990; Fama, 1998; Schwert, 2003). In this context, Jensen, Kelly, and Pedersen (2023) distinguish between internal and external validity. Recent work has brought mixed evidence. Harvey, Liu, and Zhu (2016), as well as Linnainmaa and Roberts (2018), argue that numerous anomalies are likely to be false. Many studies document that return predictability may not survive real-life trading constraints (Hou, Xue, and Zhang, 2020; Chen and Velikov, 2023), hangs on minor methodological choices (Bali and Cakici, 2008; Cakici and Zaremba, 2022), and fails to extend to international markets (Goyal and Wahal, 2015). On the other hand, Jensen, Kelly, and Pedersen (2023), Chen and Zimmermann (2022), and Jacobs and Müller (2020) conclude that most anomalies are likely to be true. While the replication debate has mainly focused on *cross-sectional return predictability*, the more limited *time-series* literature has recently gained momentum. Fresh evidence casts doubt on the validity of such well-established predictors, such as dividend growth (Hjalmarsson and Kiss, 2021) or Treasury rates (Gray and Huynh, 2021). Goyal, Welch, and Zafirov (2021) collect forty-six variables from twenty-seven papers to show that only a handful of them perform decently out of sample. Hollstein *et al.* (2020) extend a similar analysis internationally to conclude that return predictability may depend on market development. In line with these views, our study highlights the dangers of generalizing conclusions from a specific market, study period, or empirical design to a broader setting.

The remainder of this article proceeds as follows. Section 2 summarizes our empirical strategy: research sample, anomaly selection, prediction models, and performance evaluation methods. Section 3 presents the results for international markets. Section 4 reevaluates the US evidence. Section 5 explores why our results differ from earlier evidence. Finally, Section 6 concludes the article.

2. Data and Methods

2.1 Data Sources and Sample Preparation

Our international sample covers forty-two equity markets from around the world. The overall study period is from January 1990 to December 2021. However, the precise start

Table I. Research sample for international tests

The table presents the forty-two markets covered in our sample for international examinations. Start date indicates the first month when a country is included in the sample; the end date for all markets is December 2021. The table also presents the total (Total # firms) and average (Average # firms) number of companies in the market.

Country	Start date	Total	Average	Country	Start date	Total	Average
		# firms	# firms			# firms	# firms
Argentina	September 1991	164	61	Kuwait	September 2001	237	138
Australia	January 1990	3,783	1,158	Malaysia	January 1990	1,431	726
Austria	February 1990	203	72	Mexico	May 1992	280	84
Belgium	February 1990	335	127	Netherlands	January 1990	421	160
Brazil	May 1998	396	98	New Zealand	January 1990	322	105
Canada	January 1990	3,095	1,031	Norway	January 1990	733	171
Chile	December 1993	277	112	Philippines	June 1994	336	167
China	August 1996	4,750	1,496	Poland	July 1996	1,232	377
Denmark	January 1990	428	147	Portugal	January 1995	146	52
Finland	May 1990	303	109	Russia	October 1996	681	143
France	January 1990	1,974	636	Saudi Arabia	December 2001	238	123
Germany	January 1990	1,857	670	Singapore	January 1990	1,148	471
Greece	December 1994	432	176	South Africa	January 1990	1,064	308
Hong Kong	January 1990	2,891	1,072	Spain	January 1990	462	145
India	February 1990	5,110	1,563	Sweden	January 1990	1,505	345
Indonesia	May 1992	893	313	Switzerland	January 1990	510	208
Ireland	September 1990	125	42	Taiwan	January 1990	2,629	1,059
Israel	February 1996	873	292	Thailand	January 1990	1,120	469
Italy	January 1990	838	258	Turkey	April 1991	592	260
Japan	January 1990	5,894	3,302	UK	January 1990	6,070	1,788
Korea	January 1990	3,498	1,249	USA	January 1990	21,305	5,565

dates differ across countries; we accept a country for our analysis when at least twenty-five factors are available. Table I illustrates the composition of our sample.

Market data for the USA are obtained from CRSP. Market data for other countries, as well as all accounting data, come from Compustat.[3] We measure the portfolio performance with monthly returns; however, the stock characteristics are based on monthly, weekly, or daily observations. Following the typical approach in asset pricing research, we express all market data in US dollars (e.g., Fama and French, 2012, 2017; Hanauer, 2020; Baltussen, Swinkels, and Van Vliet, 2021; Hollstein, 2022; Windmüller, 2022) using the exchange rates from Compustat.[4] Consistently, the risk-free return is captured by the US 1-month Treasury bill rate. Furthermore, the market portfolio returns for each country are calculated as the value-weighted average return of all stocks available.

[3] We collect all data using the publicly available code from Jensen, Kelly, and Pedersen (2023). See https://github.com/bkelly-lab/ReplicationCrisis.

[4] In an additional analysis, we verify that our results also hold for when local currency returns are used. The change of the currency convention does not materially affect the overall findings. Further details are available in Supplementary Appendix Table A17.

We prepare and clean the data using standard methods found within asset pricing literature. In particular, we reproduce the procedures from Jensen, Kelly, and Pedersen (2023). Our sample contains only common stocks, which Compustat recognizes as primary securities of the underlying firms. We categorize them into countries based on the location of their exchange. Finally, following Jensen, Kelly, and Pedersen (2023), we winsorize the international stock returns each month at 0.1% and 99.9% using CRSP breakpoints.[5]

Once we apply all the filters, our international sample contains 80,581 unique companies and 10,216,980 monthly stock return observations.

2.2 International Anomaly Portfolios

Our international tests rely on factor returns from Jensen, Kelly, and Pedersen (2023). To avoid any arbitrariness in anomaly selection, we use all 153 stock characteristics available therein. This comprehensive sample captures the universe of the most prominent equity anomalies documented within asset pricing literature. Supplementary Appendix Table A1 lists all the characteristics that are used in the dataset. We calculate all the variables following the procedures in Jensen, Kelly, and Pedersen (2023) based on their original replication code.[6] Given over three decades of data on over 80,000 firms across forty-two markets, our sample contains 1,121,104,394 monthly characteristic observations.

To avoid arbitrariness in factor portfolio formation, we reproduce procedures from Jensen, Kelly, and Pedersen (2023). Each month, we rank all stocks on characteristics and group them into terciles. A factor strategy is then represented by a zero-investment long–short portfolio that buys (or sells) a value-weighted tercile of stocks with the strongest (or weakest) anomaly characteristic. Notably, to minimize the influence of small and thinly traded companies, we determine the breakpoints using non-microcap stocks. We classify stocks as microcaps if their market capitalization falls below the 20th percentile in a month. Next, we distribute the microcaps into the terciles based on their characteristic.

In addition to the above, we apply several restrictions associated with data availability. As seen in Jensen, Kelly, and Pedersen (2023), we require at least five firms per long and short leg. If the number of companies on either of the sides falls below 5, we set the return to missing. Furthermore, we include only the factors with at least 75% of non-missing returns throughout the study period. If this condition cannot be met, we exclude the factor entirely from the sample.

Table II presents the summary statistics for the anomaly portfolios that are used in our sample. The total number of factor strategies pooled across all markets is 5,058. The aggregate number of monthly factor return observations pooled across all factors, markets, and total study period equals 1,960,394.

2.3 Prediction Models

Our approach assumes employing machine-learning models to predict market portfolio returns. With their ability to digest vast amounts of data, distill information from multiple signals, and avoid overfitting, machine-learning routines are well-fitted to extract return predictability from the anomaly zoo. Regarding the choice of specific algorithms, we use seven

[5] In an unpublished analysis, we also experimented with non-winsorized data and observed that this operation had no measurable effect on our key findings.

[6] We are grateful for making this available at https://github.com/bkelly-lab/ReplicationCrisis.

Table II. Anomaly portfolios for international markets

The table presents the summary statistics for the monthly returns on anomaly portfolios used in international market tests. For each anomaly, we sort stocks based on an underlying return-predicting variable. The long–short anomaly strategies buy (or sell) value-weighted terciles of stocks with the strongest (or weakest) stock characteristics. The set of anomalies in each market comprises up to 153 stock characteristics from JKP3. The study periods for different markets are provided in Table I. #factors is the total number of anomalies available in a given market. The *t*-statistics in the middle section pertain to the alphas from the three-factor model of Fama and French (1993). The sample comprises forty-two markets. The means and standard deviations of returns (the rightmost section) are reported in percentage terms. The last column contains the average pairwise Pearson product–momentum correlation coefficients between anomaly portfolio returns.

	#factors	Three-factor model alphas				Anomaly portfolio returns		
		#|*t*-stat| >1.645	#|*t*-stat| >1.96	#|*t*-stat| >2.58	#|*t*-stat| >3	Average of sample means	Average of sample std. dev.	Average correlation
Argentina	67	14	9	4	2	0.33	9.19	0.03
Australia	135	65	52	31	19	0.14	3.23	0.05
Austria	108	22	17	9	4	0.11	4.77	0.04
Belgium	125	42	34	20	11	0.10	4.73	0.04
Brazil	121	38	32	14	7	0.23	5.80	0.08
Canada	151	76	68	50	37	0.25	4.48	0.06
Chile	128	25	20	8	2	0.10	3.84	0.03
China	135	67	61	43	29	0.16	4.04	0.05
Denmark	130	52	44	21	11	0.22	5.11	0.05
Finland	127	43	32	12	7	0.18	7.27	0.04
France	138	61	54	32	22	0.17	3.55	0.06
Germany	136	68	55	35	22	0.24	4.12	0.05
Greece	102	51	45	27	16	0.53	7.16	0.09
Hong Kong	130	50	37	22	12	0.13	4.74	0.06
India	45	32	26	14	9	0.18	5.89	0.09
Indonesia	125	57	49	33	26	0.29	5.99	0.04
Ireland	98	25	19	7	5	0.18	9.70	0.05
Israel	117	40	31	20	14	0.18	5.26	0.03
Italy	130	36	27	10	2	0.16	4.41	0.04
Japan	136	49	39	23	16	0.10	3.13	0.04
Korea	120	56	40	25	16	0.29	5.97	0.04
Kuwait	105	15	7	2	1	0.04	4.55	0.03
Malaysia	123	42	32	24	14	0.09	4.24	0.07
Mexico	128	18	11	3	3	0.06	4.56	0.04
Netherlands	135	45	35	15	6	0.10	4.78	0.05
New Zealand	111	39	26	12	5	0.15	4.29	0.04
Norway	125	53	32	18	8	0.22	5.00	0.05
Philippines	84	25	15	6	1	0.15	5.16	0.07
Poland	82	19	12	8	5	0.13	5.10	0.03
Portugal	120	31	21	7	3	0.22	5.91	0.03
Russia	109	36	28	15	4	0.24	6.95	0.04
Saudi Arabia	135	31	17	2	1	0.04	5.51	0.08

(continued)

Table II. Continued

	#factors	Three-factor model alphas				Anomaly portfolio returns		
		#|t-stat| >1.645	#|t-stat| >1.96	#|t-stat| >2.58	#|t-stat| >3	Average of sample means	Average of sample std. dev.	Average correlation
Singapore	128	46	35	16	7	0.13	4.59	0.06
South Africa	133	39	30	14	8	0.23	4.98	0.04
Spain	131	52	44	24	10	0.21	4.53	0.05
Sweden	133	49	39	17	8	0.11	5.17	0.04
Switzerland	135	51	39	16	8	0.14	4.36	0.05
Taiwan	112	53	39	20	13	0.12	5.14	0.03
Thailand	117	46	38	22	12	0.14	5.61	0.07
Turkey	108	17	9	4	2	0.12	6.84	0.03
UK	147	64	50	25	18	0.15	3.36	0.07
USA	153	107	96	71	53	0.15	3.05	0.08

prediction models employed originally by Dong *et al.* (2022).[7] All the models aim to forecast the month t market excess return ($r_{M,t}$) with long–short anomaly portfolio returns available through month $t-1$. For all the models, we define the market M excess return at time t as:

$$r_{M,t} = E_{t-1}(r_{M,t}) + \varepsilon_{M,t}, \tag{1}$$

where $t = 1, \ldots, T$. We estimate the expected excess return as a function of monthly returns on long–short anomaly portfolios a available in the market M at time $t-1$:

$$E_{t-1}(r_{M,t}) = g(a_{M,t-1}). \tag{2}$$

The function $g(.)$ is flexible and depends on the forecasting algorithm. The monthly anomaly returns from time $t-1$, which serve as inputs to the prediction model, are represented by the P-dimensional vector $a_{M,t-1}$. If a given factor return is missing, we replace it with the cross-sectional median of all factor returns available within the respective market. For conciseness, we describe the forecasting models only briefly; in addition, a comprehensive explanation is available from Dong *et al.* (2022) and the Internet Appendix therein.

2.3.a. Conventional OLS regression

The linear regression via the OLS assumes fitting multiple predictive regressions that use all long–short anomaly returns as model inputs. The algorithm is relatively straightforward; it requires no hyperparameters or a split into training and validation samples. For high-dimensional models, however, it may be prone to overfitting—resulting in poor out-of-sample accuracy (Gu, Kelly, and Xiu, 2020).

2.3.b. ENet

ENet attempts to reduce overfitting by shrinking the slope coefficients with a penalty term (Zou and Hastie, 2005). Specifically, the penalty function combines the components from

[7] See Dong *et al.* (2022), along with its Internet Appendix entry, for a detailed description of the models. To assure a possibly close replication, we build on the original replication code available along with the published article.

the least absolute shrinkage and selection operator and the ridge regression of Hoerl and Kennard (1970). We tune the regularization hyperparameters by using the corrected Akaike (1973) criterion of Hurvich and Tsai (1989).

2.3.c. Simple combination (Comb)

The Comb method, advocated by Rapach, Strauss, and Zhou (2010), replaces the multivariate predictive regressions from OLS with a series of univariate regressions. Concretely, in the first step, we estimate the univariate regressions of future market excess returns on each of the long–short anomaly portfolio returns separately. Then, we compute the arithmetic mean of the individual univariate predictions.

2.3.d. C-ENet

Rapach and Zhou (2020) and Han *et al.* (2023) propose the C-ENet to cope with the Comb tendency to shrink the forecasts excessively toward the prevailing mean. While Comb averages the univariate regression forecasts based on all variables, C-ENet forms averages of forecasts that ENet selects.

2.3.e. Predictor average (Avg)

The Avg method aims to cope with overfitting in another way. First, it combines the available return predictors in a single signal. Specifically, Dong *et al.* (2022) perform this operation by calculating the average return on the long–short anomaly portfolio returns. Subsequently, this consolidated predictor is employed in a univariate OLS regression to forecast the market excess returns.

2.3.f. Principal component

The PC method is implemented in two steps. To begin with, we derive the first PC from the full set of cross-sectional anomaly returns. Then, we use this component to predict the market excess return via a simple univariate OLS regression.

2.3.g. Partial least squares

While PC effectively reduce the number of dimensions, the extraction procedure is disconnected from the returns themselves. Kelly and Pruitt (2013, 2015) offer an alternative three-pass regression that concentrates on the predictors with the strongest links with stock returns. The method assumes forming a new factor from a set of covariates that is maximally correlated with future market returns. Next, we use this lagged factor as an independent variable in an OLS regression in order to predict the market portfolio excess return.

We implement and estimate all the above models using the standard procedures from the literature. The return predictions are always out of sample because we only use data available through month $t-1$ to predict month t. We split the study period into subperiods, as required by a specific model—training period, validation period, and testing period—which all maintain the temporal order. The first 10 years are used as a training sample and the subsequent 5 as the validation period. If a given method does not require cross-validation (e.g., OLS), we use the entire 15 years as the training period. The country-specific start dates may vary depending on the length of the actual study period (we always begin the study period whenever at least twenty-five different factors are available). The testing period begins no earlier than January 2005.

Regardless of the differences between countries and forecasting models, our baseline approach assumes a recursive training window. This means that we re-calculate the models each month, extending the training sample by 1 month at each re-estimation. Importantly, we estimate the models for each country separately.

2.4 Evaluation Methods

In order to assure the robustness of our findings, we use three different evaluation methods—all of which build on Dong *et al.* (2022): out-of-sample predictive R^2 coefficient (R^2_{OS}), annualized utility gains, and Sharpe ratio comparisons. The three methods complement each other and provide different perspectives on return predictability. Out-of-sample R^2_{OS} assesses the predictive accuracy of our models by comparing their mean-squared forecast errors (MSFEs) with a prevailing mean benchmark model. It allows us to directly compare our predictive models with established benchmarks, giving us an insight into how much our models improve (or worsen) their predictive ability. On the other hand, utility gains capture the practical economic impact of the model's forecast. By considering the asset allocation of a risk-averse investor, they help to quantify the economic benefits of using a particular model's forecast rather than a standard forecast. This provides a more concrete economic perspective on the model's value, which is of direct interest to many readers. Finally, the improvement in the Sharpe ratio provides a practitioner's perspective. It assesses whether the use of a particular model's predictions incrementally and significantly improves the risk-return profile of an asset allocation strategy. This metric adds a risk management dimension to the evaluation of our models, which is essential in real-world applications.

2.4.a. Out-of-sample R^2 measure

R^2_{OS} evaluates the prediction accuracy by comparing the models' MSFEs with those of a prevailing mean benchmark model. To obtain it, we begin by calculating the monthly forecast errors:

$$\widehat{e}_{0,t|t-1} = r_{M,t} - \widehat{r}^{\mathrm{PM}}_{M,t|t-1}, \tag{3}$$

$$\widehat{e}_{1,t|t-1} = r_{M,t} - \widehat{r}_{M,t|t-1}, \tag{4}$$

where $\widehat{e}_{0,t|t-1}$ and $\widehat{e}_{1,t|t-1}$ denote the monthly forecast errors of the prevailing mean benchmark prediction and the competing model, respectively, and the $\widehat{r}^{\mathrm{PM}}_{M,t|t-1}$ and $\widehat{r}_{M,t|t-1}$ are the two models' predictions. $r_{M,t}$ indicates the actual realized market return. The prevailing mean forecast is calculated as the average market excess return from the beginning of the study period through $t-1$; this utilizes all information available at the time of portfolio formation. Welch and Goyal (2008) argue that the prevailing mean forecast is difficult to beat due to the inherently low predictability of market returns.

Next, we compute the MSFE as:

$$\widehat{\mathrm{MSFE}}_j = \frac{1}{T}\sum_{t=1}^{T} \widehat{e}^2_{j,t|t-1} \text{ for } j = 0,\ 1, \tag{5}$$

where T indicates the total number of out-of-sample observations. Finally, we obtain the Campbell and Thompson (2008) R^2_{OS} measure by comparing the MSFE of the competing model forecast with the prevailing mean benchmark forecast:

$$R^2_{\mathrm{OS}} = 1 - \frac{\widehat{\mathrm{MSFE}}_1}{\widehat{\mathrm{MSFE}}_0}. \tag{6}$$

We calculate the statistical significance of R^2_{OS} estimates using the Clark and West (2007) statistic for MSFE comparisons. Using this test to evaluate:

$$\text{H}_0:\ \text{MSFE}_0 \leq \text{MSFE}_1 \text{ versus } \text{H}_\text{A}:\ \text{MSFE}_0 > \text{MSFE}_1, \tag{7}$$

is equivalent to testing

$$\text{H}_0:\ R^2_{\text{OS}} \leq 0 \text{ versus } \text{H}_\text{A}:\ R^2_{\text{OS}} > 0. \tag{8}$$

Specifically, we run a one-sided test to verify whether R^2_{OS} significantly exceeds zero. Since we study seven different models, we consider the statistical significance in two ways. First, we treat them separately—ignoring that several competing models are treated simultaneously. Second, as we consider several prediction models simultaneously, this increases the probability of type I errors. Hence, we address the multiple testing framework by using the Bonferroni adjustment. This divides the significance threshold by the number of simultaneously considered comparisons. For example, for the one-sided 5% significance level, the relevant *t*-statistics are equal to 1.645 and 2.45. Admittedly, the Bonferroni correction may seem relatively strict as it does not account for the correlation between individual tests. Therefore, we report both adjusted and unadjusted significance and leave the evaluation to the reader.

Last, to provide a global assessment of return predictability, we supplement the country-specific R^2_{OS} measures with two further statistics. First, we compute cross-sectional averages across countries. Second, we also report the R^2_{OS} values for the pooled international sample calculated as in Han *et al.* (2023).

2.4.b. Utility gain

To illustrate the economic gains from pursuing the model's forecast, we consider a risk-averse investor who allocates their assets across equities (market portfolio) and Treasury bills each month. The assumed objective function faced by the investor is

$$\text{argmax} w_{t|t-1}\widehat{r}_{M,t|t-1} - 0.5\gamma w^2_{t|t-1}\widehat{\sigma}^2_{t|t-1}, \tag{9}$$

where $w_{t|t-1}$ is the share of the portfolio allocated to the market portfolio in period t, γ is the relative risk-aversion coefficient, and $\widehat{r}_{M,t|t-1}$ and $\widehat{\sigma}^2_{t|t-1}$ denote the forecasts of market excess returns for month t and its variance. The optimal equity allocation in Equation (9) can be solved by

$$w^*_{t|t-1} = \left(\frac{1}{\gamma}\right)\left(\frac{\widehat{r}_{M,t|t-1}}{\widehat{\sigma}^2_{t|t-1}}\right). \tag{10}$$

The sample variance in Equation (10) is calculated using the 60-month trailing estimation window. Furthermore, for practical purposes, we limit the equity allocation to the range from −100% to 200%. Next, the average realized utility for the portfolios that are formed using the competing model ($\overline{U}_1$) and the prevailing mean benchmark forecast ($\overline{U}_0$) is estimated using their mean-realized portfolio returns ($\overline{r}_1$, $\overline{r}_0$) and variances ($\widehat{\sigma}^2_1$, $\widehat{\sigma}^2_0$) over the testing periods as:

$$\overline{U}_j = \overline{r}_j - 0.5\gamma\widehat{\sigma}^2_j \text{ for } j = 0,\ 1. \tag{11}$$

In Equation (11), we assume that $\gamma = 3$. We then calculate the average annualized utility gain (UG_1) of the investor that pursues a competing model's forecast instead of the prevailing mean forecast:

$$\text{UG}_1 = 12\left(\overline{U}_1 - \overline{U}_0\right). \tag{12}$$

UG_1 has an intuitive interpretation as the maximum annual management fee (expressed in percentage terms) that a rational investor would accept for accessing the information to competing forecasts in place of the prevailing mean prediction.

2.4.c. Sharpe ratio improvement

Finally, to provide a practitioner perspective, we calculate the Sharpe ratios on the portfolios formed using the competing model forecasts and the prevailing mean benchmark model. The portfolios allocate actively between the stock market and Treasury bills in line with Equation (1). Hence, the Sharpe ratios are given by:

$$\text{SR}_j = \frac{\overline{r}_j^{\text{ex}}}{\left(\widehat{\sigma}_0^{\text{ex}}\right)^2} \text{ for } j = 0,\ 1. \tag{13}$$

where $\overline{r}_1^{\text{ex}}$ ($\overline{r}_0^{\text{ex}}$) and $\widehat{\sigma}_1^{\text{ex}}$ ($\widehat{\sigma}_0^{\text{ex}}$) are the mean and standard deviation, respectively, of the portfolio excess return in the out-of-sample period for the investor that follows the competing model (prevailing mean) forecast. Subsequently, we compare the Sharpe ratios using the well-known statistics of Ledoit and Wolf (2008). This allows us to ascertain whether using a given model's predictions incrementally and significantly improves the risk-return profile of an asset allocation strategy.

3. Empirical Findings: International Evidence

We begin the discussion of our results with the findings from the international markets. Next, we zoom closer to the US market to perform additional tests. The analyses rely on three testing methods: out-of-sample R^2, utility gains and Sharpe ratios. We comment on each of them separately below.

Table III presents the prediction performance of different models across forty-two markets. The general picture emerging from the table seems disappointing: the anomalies cannot predict market excess returns. The positive and significant R^2_{OS} values can be observed in a handful of cases. Several high R^2_{OS} coefficients are seen in Israel, Japan, and Norway; however, overall, the significant values make an impression of being almost randomly scattered across the different models and countries. The quantity of insignificant R^2_{OS} measures prevails. Furthermore, the R^2_{OS} values for the pooled international sample fail to pass any significance threshold and are, in fact, all slightly negative.

Admittedly, the international data offer shorter time series than the US market. Consequently, one might expect lower statistical significance. Nevertheless, as can be seen in the bottom section of Table III, the R^2_{OS} are not only typically insignificant but also negative on average. Essentially, all models also fail to produce convincing results. The cross-country average R^2_{OS} coefficients range from –5.08% to 0.12%. The only positive mean is recorded for the Comb model (0.12%); however, even in this case, the statistical significance is too low to pass the 5% threshold. Overall, the R^2_{OS} coefficients are statistically significant at the 5% level in between three and eleven countries—depending on the model.

Table III. Prediction accuracy in international markets

The table presents the out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008) for the market excess returns forecast in forty-two countries based on long–short anomaly portfolio returns. There are seven forecasting models considered: conventional OLS, ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The predictions are based on up to 153 long–short value-weighted tercile anomaly portfolios from JKP3. The total study period is from January 1990 to December 2021; the testing period starts in January 2005. The models are estimated using a recursive training window. The R^2_{OS} are expressed in percentage terms. The numbers in parentheses are Clark and West's (2007) *t*-statistics. The values in bold are significant at the 5% level in standalone tests ($t\text{-stat} > 1.645$). The underline font indicates the 5% significance after the Bonferroni adjustment for multiple testing framework ($t\text{-stat} > 2.45$). The bottom section presents the summary statistics for international markets: the cross-country averages with corresponding bootstrap *t*-statistics, pooled R^2_{OS} coefficients of Han *et al.* (2023) for international sample, and the numbers of countries with *t*-statistics above the significance thresholds.

	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Argentina	−91.89 (−0.36)	−0.67 (0.34)	−0.51 (−1.22)	−3.59 (−0.03)	−0.39 (0.01)	−2.37 (−1.94)	−17.51 (−1.34)
Australia	−111.38 (−0.05)	−1.02 (0.70)	0.22 (0.71)	−5.42 (−1.47)	−0.29 (0.27)	−0.49 (0.13)	−6.61 (0.78)
Austria	−51.13 (1.63)	−3.04 (−0.25)	0.34 (0.76)	−2.11 (0.07)	−1.83 (−0.78)	0.24 (0.82)	−3.67 (1.02)
Belgium	**−93.29 (2.07)**	−1.86 (1.49)	1.24 (1.48)	2.65 (1.44)	−0.28 (0.55)	3.71 (1.58)	**5.11 (1.70)**
Brazil	−170.06 (−0.41)	−4.75 (0.08)	−0.75 (−0.85)	−1.39 (−0.73)	−5.91 (−1.17)	−5.12 (−0.80)	−7.54 (−0.90)
Canada	−145.44 (0.97)	−2.34 (0.16)	−0.01 (0.14)	−1.44 (0.29)	−0.22 (0.38)	−0.13 (−0.02)	−1.10 (0.06)
Chile	−111.48 (−0.04)	−1.09 (−0.81)	0.09 (0.44)	−0.42 (−0.82)	−0.15 (0.60)	0.17 (0.82)	−4.72 (0.31)
China	−163.20 (0.47)	−2.76 (0.84)	0.07 (0.30)	−2.71 (−0.06)	−0.74 (−0.80)	−0.57 (−1.78)	−22.63 (0.79)
Denmark	−202.35 (−1.80)	−5.91 (−1.92)	0.00 (0.10)	−2.47 (−1.31)	−0.01 (0.44)	−0.98 (0.03)	−4.60 (0.31)
Finland	−237.95 (1.00)	−0.26 (0.61)	0.35 (0.48)	−2.66 (−0.78)	0.61 (1.31)	0.53 (0.60)	0.11 (0.46)
France	−96.19 (0.31)	0.02 (0.37)	−0.12 (−0.50)	−0.99 (−1.60)	−0.08 (−0.05)	−0.56 (−0.28)	−5.42 (−0.27)
Germany	−245.10 (−1.34)	0.42 (0.83)	0.10 (0.44)	1.74 (1.27)	0.34 (0.59)	−0.07 (0.30)	−1.35 (0.49)
Greece	−171.53 (0.90)	−5.36 (0.09)	−0.18 (0.05)	1.16 (1.28)	−4.70 (0.43)	−3.08 (0.23)	−8.16 (−0.02)
Hong Kong	−134.21 (0.68)	−5.22 (0.51)	0.39 (1.00)	−1.90 (−0.33)	−0.56 (0.75)	−0.14 (−0.25)	−2.12 (1.07)
India	−14.23 (1.35)	**−1.27 (1.81)**	0.06 (0.43)	−0.11 (1.13)	−0.37 (−0.08)	−0.21 (−0.08)	−5.63 (0.82)
Indonesia	−185.57 (1.09)	−0.53 (1.32)	0.57 (0.91)	−0.12 (0.39)	0.75 (0.68)	−0.20 (0.56)	−2.55 (1.05)
Ireland	−112.29 (0.16)	−2.07 (0.65)	−0.75 (−1.44)	−0.77 (−0.61)	−1.50 (−0.78)	−2.61 (−1.02)	−10.78 (−1.22)
Israel	−182.19 (−0.34)	**3.62 (2.24)**	**1.27 (2.16)**	**3.07 (2.02)**	−1.30 (−0.18)	1.34 (1.31)	**3.82 (2.06)**
Italy	−141.61 (−1.05)	−3.63 (−0.27)	−0.14 (−0.93)	1.54 (1.16)	−0.34 (−1.05)	−0.64 (−1.35)	−8.90 (−0.65)
Japan	−281.37 (0.62)	−0.04 (0.24)	**0.59 (2.18)**	−0.33 (0.38)	**4.31 (2.69)**	0.06 (0.41)	**1.60 (2.02)**
Korea	−144.13 (−0.02)	−7.55 (−1.18)	−0.21 (−0.51)	−3.51 (−1.69)	−0.45 (−0.53)	−0.54 (−0.09)	−7.96 (−0.56)

(continued)

Table III. Continued

	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Kuwait	−172.77 (0.44)	**3.83 (2.01)**	0.27 (0.75)	−0.98 (−0.65)	−0.45 (0.00)	−0.68 (−0.70)	−7.11 (0.14)
Malaysia	−308.76 (−0.54)	−29.10 (0.68)	**0.75 (1.82)**	0.49 (1.15)	1.22 (1.30)	0.28 (0.71)	**0.73 (1.82)**
Mexico	−216.86 (1.41)	−1.89 (0.40)	**0.57 (2.04)**	−1.19 (−0.30)	2.03 (1.58)	−1.27 (−0.45)	**0.52 (2.01)**
Netherlands	−123.08 (0.78)	−10.38 (−0.13)	0.34 (0.84)	−0.42 (0.32)	1.15 (1.06)	1.01 (1.02)	−1.81 (1.24)
New Zealand	−77.70 (−0.66)	−3.18 (−2.24)	−0.37 (−2.12)	−1.25 (−0.64)	−1.91 (−1.69)	−0.69 (−0.75)	−5.82 (−2.21)
Norway	−87.14 (0.47)	0.69 (1.37)	**0.77 (1.99)**	0.73 (1.45)	1.17 (1.19)	**2.50 (1.80)**	**3.20 (2.11)**
Philippines	**−60.47 (1.86)**	−5.89 (−0.02)	0.47 (0.83)	−0.86 (−1.68)	−1.16 (−0.18)	0.67 (0.77)	−1.87 (0.71)
Poland	−109.47 (0.13)	−4.79 (−0.84)	−0.03 (0.03)	−0.84 (0.17)	−0.21 (−1.17)	−0.72 (−1.15)	−8.55 (0.17)
Portugal	−345.95 (−0.14)	−1.33 (0.60)	0.35 (0.75)	−0.91 (−0.82)	−1.00 (−0.14)	1.27 (1.26)	−7.89 (1.04)
Russia	−136.73 (0.40)	0.17 (1.26)	0.11 (0.37)	−0.07 (−0.18)	−0.34 (−0.16)	−0.57 (−0.17)	−3.80 (−0.06)
Saudi Arabia	−536.89 (0.97)	−46.81 (−1.05)	−1.10 (−2.07)	0.00 (0.00)	−2.02 (−1.48)	−3.41 (−1.17)	−17.19 (−2.11)
Singapore	−158.83 (0.37)	2.67 (1.52)	0.41 (1.24)	**2.51 (1.65)**	0.06 (0.30)	0.70 (0.90)	2.22 (1.39)
South Africa	−128.74 (1.43)	−2.39 (−0.54)	−0.25 (−1.35)	−1.12 (−0.80)	−1.11 (−0.38)	−0.84 (−1.80)	−23.38 (−1.06)
Spain	−126.54 (0.57)	−0.34 (−1.23)	−0.41 (−2.05)	−2.99 (−1.19)	−0.51 (−1.39)	−0.45 (−0.61)	−12.63 (−2.00)
Sweden	−121.33 (0.84)	1.35 (1.53)	0.15 (0.80)	−0.71 (−1.57)	1.08 (1.03)	−0.52 (−0.76)	0.24 (0.76)
Switzerland	−164.74 (−1.49)	−0.19 (0.64)	0.05 (0.28)	−0.27 (0.06)	−0.22 (−0.37)	−0.78 (−1.69)	−3.45 (0.28)
Taiwan	−159.02 (0.38)	−1.00 (0.18)	−0.59 (−1.19)	−1.47 (−0.10)	0.03 (0.42)	−1.68 (−1.14)	−2.56 (−0.39)
Thailand	−125.03 (−0.78)	−0.34 (−0.17)	0.01 (0.15)	−1.94 (−0.61)	0.22 (0.62)	0.47 (1.25)	−3.45 (0.01)
Turkey	−177.00 (−0.55)	−11.30 (−0.42)	−0.33 (−1.11)	−2.48 (−0.75)	−2.01 (−1.78)	−0.59 (−1.42)	−14.10 (−0.61)
UK	**−132.22 (1.78)**	**6.56 (1.87)**	**1.34 (1.68)**	2.30 (1.42)	3.12 (1.54)	**3.77 (1.81)**	**6.00 (1.76)**
USA	−185.10 (−0.88)	−3.14 (−0.88)	−0.21 (−0.47)	−0.57 (−1.05)	0.64 (0.81)	−0.56 (−0.46)	−2.05 (−0.64)
Summary statistics							
Average	−160.50 (−11.58)	−3.62 (−2.74)	0.12 (1.49)	−0.76 (−2.78)	−0.32 (−1.24)	−0.33 (−1.33)	−5.08 (−4.95)
Pooled R^2_{OS}	−187.32 (−0.55)	−3.00 (−0.55)	−0.11 (−0.23)	−0.82 (−0.82)	0.78 (1.30)	−0.37 (−0.24)	−3.02 (0.11)
# *t*-stat >1.645	3	4	6	2	1	2	7
# *t*-stat >2.45	0	0	0	0	1	0	0

When we additionally account for the multiple model comparisons, almost none of the models prove significant in more than one country. The only exception is the Avg model applied in Japan, where the R^2_{OS} coefficient amounts to 4.31% (*t*-stat 2.69). To sum up, Table III fails to provide evidence that the anomalies can predict market portfolio returns within international markets.

Table IV sheds further light on the benefits of using anomaly returns to predict market performance by presenting the utility gains. The conclusions seem equally disappointing: we hardly find evidence of anomaly-based forecasts generating utility gains. The average utility gains are negative in almost all cases—except for Comb, which does not pass the 5% significance threshold anyway. Admittedly, the utility gains for certain models and countries seem high. For example, in Israel, the gains for ENet and C-ENet reach 9.64% and 7.18%, respectively. Nevertheless, these few instances cannot change the broader image—which provides little support for the predictability.

Next, we supplement the results from Section 3 with a further practical perspective. Table V shows the annualized Sharpe ratios on the strategies based on competing models' forecasts around the world. The numbers in parentheses are Ledoit and Wolf's (2008) statistics for the ratio comparison. They indicate whether a strategy can beat the portfolio formed using the prevailing mean benchmark forecast. The performance measure of the benchmark strategy itself is reported in the leftmost column.

The conclusions from Table V are consistent with our earlier findings: the anomaly-based forecasts do not provide a measurable improvement in investment performance. Although we can observe certain variability in the risk-adjusted performance across the models and countries, the overall picture leaves little room for optimism: international market performance is typically subdued. The average annualized Sharpe ratio across countries' prevailing mean benchmark strategy is 0.15. The equivalent Sharpe ratios for the competing models vary from 0.16 (OLS) to 0.22 (ENet). Only in the case of Comb, the average Sharpe ratio—equaling 0.22 (*t*-stat = 2.08)—significantly beats the strategy based on a benchmark forecast. To sum up, from a practical perspective, we see little—if any—value added in pursuing the market return forecasts based on the cross-sectional anomaly returns.

While the comparisons in Table V rely on the prevailing mean benchmark forecast, Löffler (2022) argues that using a strategy based on average historical returns may be inappropriate. In practice, the simple buy-and-hold benchmark is much harder to beat and does not require any allocation or market timing decision. Hence, we also reproduce the analysis from Table V by using the market portfolios as the benchmark strategy. The results are summarized in Supplementary Appendix Table A2. Indeed, the market portfolio displays a higher Sharpe ratio (0.37 on an annualized basis) than the prevailing mean strategy (0.15)—and, indeed, it is harder to beat. To be precise, according to Ledoit and Wolf's (2008) tests, no forecast model in any country outperforms the alternative benchmark strategy.

Finally, to assure the validity of our findings, we supplement our examinations with several additional robustness checks. We want to ensure that our findings do not hang on some arbitrary methodological choices. Specifically, we modify several assumptions that underlie the implementation of our models.

Thus far, our international tests have focused on the models trained based on long–short tercile anomaly strategies using a 15-year recursive in-sample window. Now, we relax four essential assumptions that are embedded in this framework. First, alongside the long–short portfolios, we check the role of long and short legs separately. Second, we modify the breakpoints used to construct the strategies. While terciles ensure diversified portfolios

Table IV. Utility gains in international markets

The table presents the annualized average utility gains of an investor that uses anomaly portfolio returns to predict market excess returns in forty-two countries instead of the prevailing mean benchmark forecast. There are seven forecasting models considered: conventional OLS, ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The predictions are based on up to 153 long–short value-weighted tercile anomaly portfolios from JKP3. The total study period is from January 1990 to December 2021; the testing period starts in January 2005. The models are estimated using a recursive training window. The utility gains are expressed in percentage terms. The bottom section presents the summary statistics for international markets: the cross-country averages with corresponding *t*-statistics.

	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Argentina	−27.09	−2.39	−1.57	−13.02	−0.02	−7.07	−24.20
Australia	−16.08	−3.32	0.71	−7.33	−1.73	−1.77	−4.84
Austria	−4.27	−8.12	−0.23	−2.13	−10.66	−7.96	−14.00
Belgium	7.80	7.22	2.18	4.21	1.01	6.26	7.54
Brazil	−14.81	−9.05	−1.93	−1.80	−13.56	−11.23	−16.60
Canada	−7.39	−1.26	0.00	0.36	0.65	−2.10	0.71
Chile	−10.54	−2.63	0.26	−0.57	−1.43	0.32	−6.46
China	−21.90	−4.87	−0.25	−2.06	−1.79	−0.90	−13.93
Denmark	−21.53	−8.79	−0.50	−2.75	−2.32	−1.72	−7.96
Finland	−5.24	−0.41	1.27	−2.65	−0.22	0.28	−0.21
France	−8.39	−0.19	−0.84	−1.20	−0.48	−2.45	−7.69
Germany	−5.95	−3.37	−0.07	1.25	−0.23	0.19	−6.40
Greece	−26.85	−6.09	0.11	1.58	−3.02	−2.00	−8.23
Hong Kong	−6.31	−5.35	1.15	0.07	−0.38	−0.56	0.66
India	−3.92	2.83	0.79	2.38	1.87	1.74	2.08
Indonesia	0.17	5.16	1.96	−0.41	3.18	1.44	0.14
Ireland	−21.41	−7.39	−1.90	−0.20	−3.27	−3.68	−12.52
Israel	−5.57	9.64	2.13	7.18	−1.69	−1.33	4.85
Italy	−21.52	−3.88	−0.28	3.08	−0.63	−1.05	−12.43
Japan	1.56	−0.36	0.90	0.27	7.97	0.02	6.72
Korea	−13.74	−9.27	0.28	−5.98	−1.80	0.46	−7.14
Kuwait	1.63	4.00	0.27	−2.21	1.48	−1.67	−5.21
Malaysia	−9.91	−0.67	1.77	1.33	3.66	1.08	5.08
Mexico	−9.73	−3.90	1.54	−1.58	6.93	−0.12	4.36
Netherlands	−2.33	−5.41	0.01	−1.14	1.39	−0.15	0.35
New Zealand	−12.38	−7.82	−0.83	−0.96	−3.47	−2.51	−8.89
Norway	−21.17	2.19	1.41	2.94	−0.97	3.45	6.14
Philippines	6.92	0.40	0.97	−1.54	−1.90	1.30	4.50
Poland	−14.13	−5.57	0.19	−1.08	−0.28	−2.20	−7.44
Portugal	−11.72	0.84	0.38	−1.01	−3.02	2.88	−1.65
Russia	−13.97	0.25	0.57	−1.17	0.70	0.64	−3.85
Saudi Arabia	1.13	−5.10	−1.20	0.00	−2.07	−2.35	−14.22
Singapore	−12.22	3.07	1.44	5.96	−0.09	2.45	4.46
South Africa	−7.67	−2.58	−0.57	−3.91	−2.88	−1.89	−18.25
Spain	−9.67	−0.73	−0.72	−3.25	−2.01	−0.41	−10.45
Sweden	−16.58	5.67	0.31	−1.96	−0.59	−0.63	1.07
Switzerland	−9.79	2.21	0.24	−0.38	−0.01	−1.28	2.73

(continued)

Table IV. Continued

	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Taiwan	−8.27	−0.41	−1.21	−0.18	−0.42	−3.15	−2.22
Thailand	−23.40	0.00	0.00	−3.78	0.61	1.05	−6.68
Turkey	−43.28	−8.27	−0.48	−4.18	−4.43	−1.24	−14.00
UK	4.03	4.52	3.22	1.58	5.91	7.52	7.28
USA	−9.29	−0.31	−0.08	−0.02	2.28	0.82	−1.22
Average	−10.83	−1.66	0.27	−0.86	−0.66	−0.70	−4.24
	(−6.95)	(−2.29)	(1.56)	(−1.65)	(−1.15)	(−1.41)	(−3.49)

even in relatively small markets, they may result in a relatively narrow spread between top and bottom characteristics, that is, a weaker signal of mispricing. Hence, we replace the terciles with quintiles.[8] Third, we consider an alternative in-sample period of 10 years. For the models requiring validation, it comprises 7-year training and 3-year validation windows; this is opposed to the previously used 10- and 5-year frames. Whereas such an approach may potentially lead to less "tuned" models, it enables the extension of the testing period from 17 to 22 years.[9] Fourth, in addition to the recursive training window, we investigate the rolling (or sliding) variant. The rolling-window approach emphasizes the use of the most recent data. Therefore, it may theoretically downplay the impact of older or obsolete predictors that could have already lost their information content. As in our earlier tests, we stick to the 15-year length of the in-sample period (10-year training and 5-year validation).

For brevity, we limit the presentation of the results of these robustness checks in the main manuscript to aggregate summary statistics only. For each of the tested variants, Table VI presents the R^2_{OS} coefficient for the pooled global sample, the average R^2_{OS} value across individual markets, as well as the number of significant single-country R^2_{OS} measures in each case. The detailed outcomes of these analyses for particular countries are available in the Supplementary Appendix (see Supplementary Appendix Tables A4–A6).

Our robustness checks' results fully confirm the earlier findings. The anomaly portfolios fail to reliably predict aggregate market returns. The pooled global R^2_{OS} are predominantly slightly negative and fail to pass any commonly accepted significance thresholds. The average country-specific R^2_{OS} values also typically fall below zero. The few sole exceptions, typically for the Comb strategy, are still hardly distinguishable from zero and insignificant. For each of the strategies, the R^2_{OS} are significantly positive only in a handful of countries—from zero to ten. Nonetheless, if we control for the multiple hypothesis testing using the Bonferroni correction, the strategies successfully forecast returns in one or two countries per specification at best. In short, our tests do not support the view that anomalies assist in predicting future market returns.

[8] Our tests of quintile portfolios cover only thirty-nine countries instead of forty-two. This is because three markets—Brazil, Kuwait, and Saudi Arabia—have an insufficient number of stocks to run the tests. The descriptive statistics of these portfolios are provided in Supplementary Appendix Table A3.

[9] Importantly, in the case of one method (OLS), we continue to use the 15-year in-sample window. This exception is dictated by the considerations of the degrees of freedom in the regressions. With a 10-year window, the number of observations might be lower than the number of estimated parameters for the first 3 years of our sample.

Table V. Sharpe ratios for international markets

The table presents the annualized Sharpe ratios for an investor who allocates between the local market portfolio and risk-free Treasury bills by using the different models indicated in the first column to predict the market portfolio excess returns in forty-two countries. There are seven forecasting models considered: conventional OLS, ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The predictions are based on up to 153 long–short value-weighted tercile anomaly portfolios from JKP3. The total study period is from January 1990 to December 2021; the testing period starts in January 2005. The models are estimated using a recursive training window. The numbers in parentheses are *t*-statistics from the test of Ledoit and Wolf (2008), which compares a competing model with a prevailing mean benchmark forecast (as is indicated in the *Bench* column). The values in bold are significant at the 5% level in stand-alone tests (|*t*-stat|> 1.96). The underline font indicates the 5% significance after the Bonferroni adjustment for multiple testing framework (|*t*-stat| > 2.69). The bottom section presents the cross-country averages with corresponding *t*-statistics.

	Bench	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Argentina	−0.18	0.10 (0.59)	0.12 (0.92)	−0.31 (−0.67)	0.11 (0.83)	0.10 (1.08)	−0.42 (−0.79)	−0.39 (−0.64)
Australia	0.23	0.07 (−0.57)	0.21 (−0.08)	0.28 (0.79)	0.01 (−1.80)	0.20 (−0.23)	0.21 (−0.11)	0.30 (0.24)
Austria	0.11	0.36 (0.74)	−0.15 (−1.10)	0.08 (−0.25)	0.14 (0.08)	−0.26 (−1.88)	−0.06 (−0.61)	−0.04 (−0.48)
Belgium	−0.05	**0.56 (2.00)**	**0.50 (2.22)**	0.12 (1.63)	0.31 (1.67)	0.14 (0.71)	0.44 (1.78)	0.51 (1.94)
Brazil	−0.21	0.30 (1.25)	0.09 (0.70)	−0.23 (−0.10)	−0.34 (−0.44)	−0.36 (−0.39)	−0.25 (−0.08)	−0.29 (−0.20)
Canada	0.23	0.25 (0.07)	0.30 (0.36)	0.25 (0.56)	0.30 (0.32)	0.32 (0.57)	0.20 (−0.26)	0.32 (0.54)
Chile	0.23	0.06 (−0.48)	0.07 (−1.44)	0.23 (−0.11)	0.20 (−0.37)	0.22 (−0.03)	0.23 (0.11)	0.04 (−0.65)
China	0.45	−0.07 (−1.38)	0.22 (−0.70)	0.43 (−0.33)	0.33 (−1.01)	0.30 (−1.23)	0.38 (−1.01)	0.03 (−0.81)
Denmark	0.49	**−0.09 (−2.69)**	0.14 (−2.53)	0.47 (−0.78)	0.36 (−0.74)	0.41 (−0.74)	0.43 (−0.46)	0.27 (−1.36)
Finland	0.21	0.22 (0.03)	0.32 (0.43)	0.29 (0.72)	0.10 (−0.57)	0.21 (0.04)	0.31 (0.44)	0.30 (0.38)
France	0.14	0.11 (−0.07)	0.14 (−0.06)	0.09 (−1.22)	0.05 (−1.61)	0.13 (−0.22)	0.00 (−1.25)	−0.17 (−0.99)
Germany	0.13	0.25 (0.49)	0.16 (0.13)	0.13 (0.10)	0.33 (1.09)	0.16 (0.18)	0.18 (0.26)	0.04 (−0.43)
Greece	−0.65	**0.05 (2.05)**	−0.05 (1.43)	−0.43 (0.76)	**0.14 (2.70)**	0.01 (1.45)	−0.02 (1.51)	−0.10 (1.30)
Hong Kong	0.37	0.29 (−0.28)	0.22 (−1.01)	0.43 (1.32)	0.40 (0.27)	0.43 (0.24)	0.36 (−0.53)	0.48 (0.42)
India	0.36	0.38 (0.06)	0.50 (0.56)	0.40 (0.95)	0.46 (0.62)	0.44 (1.52)	0.43 (0.88)	0.55 (0.88)
Indonesia	0.10	0.46 (0.83)	0.51 (1.06)	0.20 (1.01)	0.11 (0.01)	0.34 (0.85)	0.23 (0.51)	0.34 (0.59)
Ireland	−0.22	−0.27 (−0.12)	−0.11 (0.54)	−0.26 (−1.22)	−0.20 (0.25)	−0.28 (−0.63)	−0.25 (−0.19)	−0.36 (−0.65)

(continued)

Table V. Continued

	Bench	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Israel	0.26	0.21 (−0.13)	0.75 (1.92)	0.37 (1.45)	0.64 (1.65)	0.24 (−0.12)	0.26 (−0.02)	0.53 (1.01)
Italy	−0.13	−0.21 (−0.25)	0.00 (0.49)	−0.15 (−0.56)	0.28 (1.34)	−0.20 (−1.13)	−0.18 (−0.36)	−0.14 (−0.01)
Japan	−0.17	0.35 (1.06)	−0.09 (0.31)	0.01 (1.75)	0.03 (0.88)	**0.73 (2.30)**	−0.14 (0.58)	0.59 (1.89)
Korea	0.09	0.15 (0.21)	−0.11 (−0.60)	0.14 (0.28)	−0.23 (−1.17)	−0.10 (−1.36)	0.21 (0.39)	−0.06 (−0.39)
Kuwait	0.71	0.78 (0.13)	0.87 (1.35)	0.72 (0.48)	0.61 (−0.78)	0.77 (0.58)	0.64 (−1.09)	0.47 (−0.90)
Malaysia	0.24	0.00 (−0.79)	0.33 (0.40)	0.35 (1.57)	0.34 (1.07)	0.48 (1.07)	0.30 (0.70)	0.56 (1.05)
Mexico	−0.05	0.10 (0.44)	0.07 (0.72)	**0.03 (2.31)**	−0.02 (0.26)	0.40 (1.38)	−0.06 (−0.06)	0.39 (1.37)
Netherlands	0.20	0.23 (0.08)	0.03 (−0.76)	0.22 (0.41)	0.19 (−0.12)	0.34 (0.65)	0.25 (0.28)	0.33 (0.53)
New Zealand	0.31	0.04 (−0.95)	**−0.05 (−2.45)**	**0.26 (−2.05)**	0.26 (−0.37)	0.10 (−1.77)	0.20 (−1.29)	−0.10 (−1.79)
Norway	0.26	0.09 (−0.59)	0.44 (1.02)	0.35 (1.81)	0.45 (1.68)	0.33 (0.39)	0.52 (1.21)	0.65 (1.53)
Philippines	−0.12	0.62 (1.77)	0.20 (0.85)	0.10 (1.34)	−0.35 (−1.77)	−0.09 (0.08)	0.22 (1.01)	0.49 (1.55)
Poland	−0.06	−0.07 (−0.02)	−0.24 (−0.60)	−0.02 (0.61)	0.06 (0.42)	−0.09 (−0.52)	−0.24 (−1.64)	−0.03 (0.08)
Portugal	0.12	0.05 (−0.16)	0.26 (0.54)	0.18 (0.48)	0.00 (−0.90)	−0.01 (−0.61)	0.43 (1.06)	0.26 (0.41)
Russia	0.22	0.05 (−0.41)	0.22 (−0.43)	0.22 (0.02)	0.19 (−0.99)	0.24 (0.17)	0.23 (0.07)	0.03 (−0.54)
Saudi Arabia	0.66	0.69 (0.06)	0.38 (−0.80)	**0.59 (−2.11)**	0.66 (0.00)	0.54 (−1.05)	0.53 (−0.76)	**−0.14 (−2.43)**
Singapore	0.16	0.00 (−0.46)	0.41 (1.13)	0.22 (1.47)	0.50 (1.41)	0.17 (0.08)	0.28 (0.83)	0.45 (1.20)
South Africa	0.15	0.34 (0.68)	0.03 (−0.58)	0.09 (−1.40)	0.00 (−1.09)	0.01 (−0.60)	**−0.02 (−2.20)**	−0.15 (−1.09)
Spain	−0.03	0.07 (0.27)	−0.05 (−1.06)	−0.07 (−1.56)	−0.16 (−1.05)	−0.09 (−1.42)	−0.04 (−0.26)	−0.38 (−1.09)
Sweden	0.29	0.04 (−0.94)	0.59 (1.61)	0.31 (0.86)	0.21 (−1.58)	0.34 (0.43)	0.26 (−0.54)	0.39 (0.75)
Switzerland	0.40	0.03 (−1.17)	0.52 (0.68)	0.41 (0.65)	0.39 (−0.19)	0.40 (0.01)	0.33 (−1.59)	0.54 (0.55)
Taiwan	0.34	0.22 (−0.32)	0.26 (−0.31)	0.10 (−1.08)	0.27 (−0.20)	0.24 (−0.35)	−0.08 (−1.47)	0.15 (−0.60)
Thailand	0.20	−0.12 (−0.99)	0.20 (0.03)	0.20 (0.03)	−0.01 (−0.95)	0.25 (0.58)	0.29 (1.06)	0.03 (−0.62)
Turkey	0.04	−0.36 (−0.90)	0.09 (0.19)	0.03 (−0.12)	−0.05 (−0.56)	−0.09 (−1.61)	0.01 (−1.00)	−0.12 (−0.50)
UK	−0.04	0.41 (1.72)	**0.39 (1.97)**	0.20 (1.62)	0.18 (1.14)	0.41 (1.65)	0.49 (1.87)	**0.50 (2.10)**
USA	0.38	0.04 (−1.12)	0.38 (−0.11)	0.38 (−0.05)	0.38 (−0.02)	0.50 (0.57)	0.43 (0.29)	0.33 (−0.25)
Average	0.15	0.16 (0.30)	0.22 (1.89)	0.18 (2.08)	0.18 (1.07)	0.20 (1.44)	0.18 (1.04)	0.18 (0.59)

Table VI. Prediction performance in international markets: alternative test specifications

The table presents the summary statistics for the prediction performance of different models across forty-two global stock markets. Seven different models are employed to forecast market excess returns using anomaly portfolio returns: conventional OLS, ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The total study period is from January 1990 to December 2021 and the baseline testing period starts in January 2005. The predictions are based on up to 153 value-weighted anomaly portfolios from JKP3; we consider three types of inputs: long–short portfolio returns (left-most section), long leg portfolio excess returns (middle section), and short leg portfolio excess returns (right-most section). *Average* R^2_{OS} is the cross-country average of the single-market out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008), along with the corresponding bootstrap *t*-statistic (in parentheses). *Pooled* R^2_{OS} is the pooled R^2_{OS} coefficients of Han *et al.* (2023) for the pooled international sample. *Average* R^2_{OS} and *Pooled* R^2_{OS} are expressed in percentage terms. The table also reports the number of markets with Clark and West's (2007) *t*-statistics that are significant at the 5% level in standalone tests *(#t* >1.645) and significant after the Bonferroni adjustment for multiple testing framework *(#t* >2.45). Panel A reports the results of the baseline approach. Panels B–D consider various modifications of the methodological assumptions: replacing terciles with quintiles in portfolio construction (Panel B), shortening the in-sample period to 10 years (Panel C), and training the models using rolling—rather than extending—training windows (Panel D).

	Long–short portfolio returns				Long leg portfolio excess returns				Short leg portfolio excess returns			
	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45
Panel A: Baseline approach												
OLS	−160.50 (−11.58)	−187.32 (−0.55)	3	0	−157.63 (−13.26)	−176.10 (−0.40)	2	0	−169.46 (−8.74)	−162.10 (−0.59)	2	1
ENet	−2.46 (−3.34)	−3.00 (−0.55)	4	0	−2.34 (−3.06)	−1.95 (−1.07)	0	0	−2.34 (−3.06)	−1.95 (−0.54)	4	0
Comb	−0.12 (−0.44)	−0.11 (−0.23)	6	0	−0.10 (−0.37)	−0.20 (0.35)	2	0	−0.10 (−0.37)	0.07 (0.64)	1	0
C-ENet	−0.92 (−3.01)	−0.82 (−0.82)	2	0	−1.02 (−3.84)	−0.21 (0.85)	0	0	−1.02 (−3.84)	−0.28 (0.17)	0	0
Avg	−0.39 (−1.24)	0.78 (1.30)	1	1	−0.36 (−1.12)	−0.21 (0.43)	2	0	−0.36 (−1.12)	0.07 (0.69)	1	0
PC	−0.40 (−1.28)	−0.37 (−0.24)	2	0	−0.36 (−1.13)	−0.24 (0.41)	2	0	−0.36 (−1.13)	0.05 (0.68)	1	0
PLS	−0.52 (−1.49)	−3.02 (0.11)	7	0	−0.52 (−1.42)	−0.16 (0.53)	2	0	−0.52 (−1.42)	0.12 (0.75)	1	0
Panel B: Terciles replaced with quintiles												
OLS	−147.81 (−14.70)	−191.40 (−1.09)	1	0	−138.52 (−13.79)	−168.97 (0.24)	3	0	−161.13 (−10.63)	−158.71 (−0.32)	1	0
ENet	−1.90 (−3.53)	−4.54 (−1.05)	3	0	−1.97 (−3.73)	−0.78 (−0.03)	2	0	−1.97 (−3.73)	−1.07 (0.12)	3	0
Comb	−0.04 (−0.11)	−0.11 (−0.20)	4	0	0.03 (0.14)	−0.19 (0.33)	2	0	0.03 (0.14)	0.14 (0.69)	2	0

(continued)

Table VI. Continued

	Long–short portfolio returns				Long leg portfolio excess returns				Short leg portfolio excess returns			
	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45	Average R^2_{OS}	Pooled R^2_{OS}	#t > 1.65	#t > 2.45
C-ENet	−0.92 (−2.61)	−1.62 (−1.48)	2	0	−0.68 (−1.73)	−1.29 (−0.56)	1	0	−0.68 (−1.73)	−0.43 (0.42)	0	0
Avg	−0.34 (−0.93)	0.77 (1.32)	2	0	−0.29 (−0.70)	−0.20 (0.46)	2	0	−0.29 (−0.70)	0.13 (0.77)	2	0
PC	−0.36 (−1.00)	−0.36 (−0.10)	3	0	−0.28 (−0.70)	−0.25 (0.42)	2	0	−0.28 (−0.70)	0.10 (0.74)	2	0
PLS	−0.67 (−1.61)	−2.91 (0.19)	6	0	−0.29 (−0.61)	−0.11 (0.59)	2	0	−0.29 (−0.61)	0.19 (0.83)	3	0
Panel C: 10-year instead of 15-year in-sample period												
OLS	−160.50 (−11.58)	−2.17e5 (−0.96)	3	0	−157.63 (−13.26)	−8.05e6 (1.52)	2	0	−169.46 (−8.74)	−6.57e5 (0.83)	2	1
ENet	−6.99 (−3.76)	−6.04 (0.24)	2	1	−5.92 (−4.85)	−3.78 (−0.91)	4	0	−5.92 (−4.85)	−4.14 (−1.09)	4	0
Comb	0.23 (0.81)	−0.36 (−0.81)	8	1	0.26 (0.88)	−0.69 (−0.35)	5	0	0.26 (0.88)	−0.58 (−0.05)	6	0
C-ENet	−0.66 (−2.26)	−1.80 (−1.85)	2	1	−0.47 (−1.19)	−0.15 (0.91)	4	0	−0.47 (−1.19)	−0.10 (0.88)	3	0
Avg	−0.02 (−0.05)	−0.45 (0.57)	4	0	0.03 (0.09)	−0.76 (−0.23)	5	0	0.03 (0.09)	−0.64 (0.04)	5	0
PC	−0.03 (−0.06)	−0.59 (−0.18)	4	1	0.03 (0.10)	−0.78 (−0.26)	5	0	0.03 (0.10)	−0.66 (0.01)	5	0
PLS	−0.13 (−0.34)	−4.14 (0.22)	10	1	−0.18 (−0.47)	−1.47 (−0.58)	7	0	−0.18 (−0.47)	−0.58 (0.27)	6	0
Panel D: Rolling instead of recursive training window												
OLS	−365.39 (−13.90)	−639.64 (−1.12)	2	0	−353.15 (−14.77)	−779.14 (0.82)	2	0	−383.62 (−10.39)	−620.32 (−1.36)	5	2
ENet	−2.31 (−5.53)	−6.09 (−1.33)	3	0	−2.05 (−3.31)	−1.82 (−0.51)	2	0	−2.05 (−3.31)	−5.81 (−1.25)	2	0
Comb	−0.35 (−1.42)	−0.31 (−0.43)	4	0	−0.29 (−1.10)	−0.83 (0.22)	0	0	−0.29 (−1.10)	−0.56 (0.45)	1	0
C-ENet	−0.51 (−1.83)	0.63 (1.15)	0	0	−0.66 (−2.48)	0.09 (0.65)	2	0	−0.66 (−2.48)	0.70 (1.08)	0	0
Avg	−0.73 (−2.59)	−0.26 (0.65)	1	1	−0.65 (−2.19)	−0.88 (0.34)	0	0	−0.65 (−2.19)	−0.61 (0.53)	1	0
PC	−0.74 (−2.67)	−1.00 (−0.17)	3	0	−0.64 (−2.21)	−0.92 (0.32)	0	0	−0.64 (−2.21)	−0.63 (0.52)	1	0
PLS	−1.12 (−3.06)	−3.90 (−0.03)	5	0	−0.65 (−1.91)	−0.84 (0.43)	0	0	−0.65 (−1.91)	−0.62 (0.57)	1	0

4. USA and Further Evidence

The analysis of the international markets fails to provide compelling evidence for market return predictability with equity anomalies. The patterns in global markets seem to differ vividly from the USA, where the phenomenon was initially documented.

To assure robustness, we explore the US market using five different sets of anomalies. We focus on the most respectable anomaly sets in literature, which are commonly employed in asset pricing research. Besides our 153 tercile portfolios from Jensen, Kelly, and Pedersen (2023) (henceforth: JKP3), we also use the original anomaly set of Dong *et al.* (2022)—comprising 100 representative anomalies (henceforth: DLRZ).[10] Next, we use the same data and procedures as in JKP3 and form analogous decile portfolios (JKP10). The only difference between JKP3 and JKP10 lies in the choice of the breakpoint. The use of deciles aims at aligning the portfolio construction closer with DLRZ, where this cut-off point is employed. Notably, the wider spread in the decile portfolios is associated with stronger mispricing and—consequently—with a sharper predictive signal. The fourth portfolio set is the 207 anomalies from Chen and Zimmermann (2022) (CZ).[11] Finally, the last group contains the 188 testing portfolios based on Hou, Xue, and Zhang (2020) (HXZ).[12] The detailed composition of the DLRZ, CZ, and HXZ anomaly sets is provided in Supplementary Appendix Tables A7–A9.

Notably, the factor sets differ substantially in terms of data source, stock filtering procedures, portfolio construction methods, as well as size and composition of the anomaly universe. For example, considering the choice of breakpoints: JKP3 relies on tercile portfolios, DLRZ uses deciles, HXZ deciles or quintiles, and CZ employs different rules carefully following the original research design. While at this point we want to have a general overview of the return predictability in the USA, we explore these methodological differences closer in further analyses. To ensure comparability with the seminal study of DLRZ, we follow their original study period choice in all the tests of the US market. Hence, our sample, in this case, runs from January 1970 to December 2017. As in our earlier tests, the in-sample period equals 15 years, comprising 10-year training and 5-year validation windows.[13]

Table VII summarizes the different anomaly samples employed in the tests of the US market. Most sets display comparable performance statistics, such as average returns or volatility. However, the exceptions are the CZ and HXZ factors—which exhibit noticeably higher profitability than DLRZ, JKP3, and JKP10.

[10] Notably, the replication package for the Dong *et al.* (2022) paper, which is available at https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13099, contains only the anomaly returns and lacks the code to compute them independently. The code to reproduce the anomaly portfolios is also not available from the authors.

[11] We thank the authors for making this data available at https://www.openassetpricing.com/.

[12] The 188 factors are a subset of those anomalies explored by Hou, Xue, and Zhang (2020) that proved significant in their tests. For details, see Xue (2022). We are grateful to the authors for providing this dataset at: https://global-q.org/testingportfolios.html.

[13] For the OLS method applied to the CZ and HXZ anomaly sets, we exceptionally extend the in-sample period to 20 years. This departure from baseline methodology is dictated by the otherwise insufficient number of degrees of freedom. With the 15-year window, the number of anomalies could exceed the number of monthly observations in the initial years of the sample. Importantly, we experiment with alternative training windows—both shorter and longer—and their choice have no qualitative impact on the findings.

Table VII. Anomaly portfolios for the US market

The table presents the summary statistics for monthly returns on anomaly portfolios used for the US market tests. For each anomaly, we sort stocks based on an underlying return-predicting variable. The respective long–short strategies buy (or sell) value-weighted portfolios of stocks with the strongest (or weakest) anomaly characteristics. The return statistics are reported separately for the long–short differential returns and the long and short legs. The underlying sample comprises NYSE, AMEX, and NASDAQ stocks. The study period runs from January 1970 to December 2017. We consider five different sets of anomalies: the original anomaly set used in DLRZ; 153 tercile and decile portfolios based on JKP3 (JPK10); 207 portfolios from CZ; and 188 anomalies from HXZ. The means and standard deviations of returns are reported in percentage terms.

	DLRZ	JKP3	JKP10	CZ	HXZ
Total number of factors	100	153	153	207	188
Average number of factors available	98	153	153	191	179
Fama and French (1993) three-factor model alphas					
Number of factors with $\|t\text{-stat}\| > 1.645$	58	69	78	127	83
Number of factors with $\|t\text{-stat}\| > 1.96$	48	49	60	113	60
Number of factors with $\|t\text{-stat}\| > 2.58$	28	20	26	76	28
Number of factors with $\|t\text{-stat}\| > 3$	16	6	17	62	21
Average correlation across anomaly excess returns					
Long leg	0.76	0.90	0.78	0.84	0.79
Short leg	0.82	0.90	0.81	0.85	0.84
Long–short	0.08	0.06	0.07	0.05	0.08
Long-leg anomaly portfolio excess returns					
Average of sample means	0.71	0.64	0.69	1.32	1.20
Average of sample standard deviations	5.16	4.65	5.37	6.14	5.43
Short-leg anomaly portfolio excess returns					
Average of sample means	0.33	0.46	0.34	0.78	0.72
Average of sample standard deviations	6.20	5.23	6.40	6.60	5.87
Long–short anomaly portfolio excess returns					
Average of sample means	0.38	0.18	0.34	0.55	0.47
Average of sample standard deviations	4.37	3.02	4.86	3.56	4.06

4.1 Preliminary Evidence

Table VIII summarizes the prediction performance of different models for the US market, reporting the R^2_{OS} values, utility gains, and Sharpe ratios. The results seem critically sensitive to the choice of the anomaly sample: they hold for DLRZ but not others. Observe, for example, the R^2_{OS} coefficients in Panel A. The return predictability for the original DLRZ sample is stellar. The R^2_{OS} scores are mostly positive and statistically significant; furthermore, Comb, C-ENet, and PLS even pass the Bonferroni-adjusted 5% threshold. Except for OLS, all the models produce sizeable positive R^2_{OS} ranging from 0.89 to 2.81. The OLS, which suffers from an overfitting problem, clearly underperforms—falling behind the other models. Overall, our results perfectly match the original results of DLRZ.

The picture for other samples, however, is remarkably different: the R^2_{OS} values are prevailingly negative. Literally, in none of the cases do they significantly exceed zero. The highest R^2_{OS} score, which is recorded for the Avg method fed with the JKP10 anomalies, reaches

Table VIII. Market return predictability in the US market

The table presents the measures of prediction performance for the US market excess returns forecast based on various sets of long–short anomaly portfolio returns: the out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008) (Panel A), the annualized average utility gains (Panel B), and the annualized Sharpe ratios for an investor who allocates between the US market portfolio and risk-free Treasury bills (Panel C). We consider five different sets of long–short anomaly portfolios: the original set used in DLRZ; 153 tercile and decile portfolios based on JPK3 (JPK10); 207 portfolios from CZ; and 188 portfolios from HXZ. The total study period is from January 1970 to December 2017; the testing period starts in January 1985. The underlying sample comprises US stocks. The R^2_{OS} and utility gains are expressed in percentage terms. The numbers in parentheses are: in Panel A, Clark and West's (2007); and in Panel B, *t*-statistics from the test of Ledoit and Wolf (2008), which compares a given Sharpe ratio with the Sharpe ratio of the portfolio formed using the prevailing mean benchmark forecast (=0.42). The values in bold are significant at the 5% level in standalone tests (*t*-stat >1.645 in Panel and |*t*-stat| >1.96 in Panel C); furthermore, the underline font indicates the 5% significance after the Bonferroni adjustment for multiple testing framework (*t*-stat >2.45 in Panel A and |*t*-stat| >2.69 in Panel C).

	DLRZ	JKP3	JKP10	CZ	HXZ
Panel A: Predictive R^2 coefficients					
OLS	−2513.86 (0.53)	−186.46 (−1.37)	−167.93 (0.89)	−27,288.47 (1.27)	−61,224.58 (0.71)
ENet	**2.03 (2.26)**	−0.51 (0.95)	−0.35 (0.97)	−1.54 (−0.98)	−4.27 (−1.31)
Comb	$\underline{\textbf{0.89 (2.50)}}$	0.31 (1.60)	0.18 (1.03)	−0.01 (0.07)	0.16 (1.10)
C-ENet	$\underline{\textbf{2.81 (2.49)}}$	−0.91 (−0.92)	−0.36 (−0.18)	−0.82 (−1.52)	−1.57 (−0.97)
Avg	**1.89 (2.13)**	0.78 (1.48)	0.85 (1.57)	0.09 (0.58)	0.65 (1.41)
PC	**1.25 (1.84)**	0.11 (0.77)	0.46 (1.21)	−0.01 (0.43)	−0.16 (−0.54)
PLS	$\underline{\textbf{2.06 (2.65)}}$	−0.33 (1.36)	−1.11 (0.95)	−4.72 (0.00)	−−2.47 (1.17)
Panel B: Utility gains					
OLS	−4.97	−9.85	−5.11	−11.86	−19.29
ENet	6.26	0.90	1.05	−1.17	−1.96
Comb	2.59	0.76	0.72	0.03	0.05
C-ENet	6.06	−2.09	−0.71	−1.14	−2.12
Avg	3.74	2.54	2.88	1.03	2.14
PC	3.28	1.37	2.35	0.74	−0.25
PLS	6.38	3.30	2.00	0.54	1.48
Panel C: Sharpe ratios					
OLS	0.23 (−0.82)	0.08 (−1.64)	0.26 (−0.72)	**−0.02 (−1.99)**	**−**$\underline{\textbf{0.37 (−3.33)}}$
ENet	**0.81 (1.99)**	0.48 (0.38)	0.49 (0.47)	0.37 (−0.50)	0.32 (−1.76)
Comb	**0.59 (1.96)**	0.47 (1.65)	0.47 (1.42)	0.43 (0.13)	0.43 (0.19)
C-ENet	0.79 (1.79)	0.29 (−1.70)	0.39 (−0.63)	0.35 (−1.70)	0.31 (−0.78)
Avg	0.65 (1.18)	0.58 (1.01)	0.61 (1.13)	0.49 (0.94)	0.56 (1.13)
PC	0.62 (1.13)	0.51 (0.72)	0.57 (1.07)	0.47 (0.47)	0.41 (−0.57)
PLS	0.80 (1.55)	0.62 (0.88)	0.55 (0.50)	0.46 (0.19)	0.52 (0.53)

only 0.85% and remains insignificant at the 5% level with a *t*-value of 1.57. The analysis of other measures of prediction performance leads to similar conclusions. The utility gains (seen in Table VIII, Panel B) are impressively high for the DLRZ dataset. Leaving the OLS aside, they range from 2.59% for Comb to 6.38% for PLS. For other samples, the gains are

substantially lower—and sometimes even negative. For example, in the case of our baseline anomaly sample (JKP3), they range from −2.09% for C-ENet to 3.30% for PLS. The results for the remaining prediction algorithms are qualitatively similar.

Finally, Panel C concentrates on the Sharpe ratio comparisons. The annualized Sharpe ratio of the portfolio based on the prevailing mean forecast equals 0.42. Each strategy based on DLRZ in Panel A, except for OLS, beats this benchmark. Their Sharpe ratios range from 0.59 (Comb) to 0.81 (ENet). However, the statistical significance of this outperformance is relatively low. Only two strategies (Comb and C-ENet) display Ledoit and Wolf's (2008) statistics marking significance at the 5% level on a standalone basis; furthermore, none pass this threshold in the multiple hypotheses framework.

As seen in Panels B and C, relaxing the assumptions concerning the anomaly set detrimentally affects the portfolio performance. The Sharpe ratios are no different in this regard: for JKP3, JKP10, CZ, and HXZ, they are typically substantially lower than for DLRZ. Specifically, they fail to significantly beat the prevailing mean forecast benchmark in any of the considered specifications.[14]

To wrap up our discussion of the preliminary findings for the US market, the market return predictability by anomalies is not robust to using alternative anomaly strategies as model inputs. The forecasting models can produce substantial economic gains when fed with DLRZ while failing for others. The magnitude of the market return predictability by equity anomalies may hinge on at least two dimensions. First, it may be affected by the composition of the anomaly set. Additionally, the anomaly portfolio's structure may also play a role. We now look closer into these aspects one by one.

4.2 Anomaly Selection

The general assertion that *anomalies predict market returns* yields a vital question: which anomalies? The finance literature has documented hundreds of cross-sectional predictors in stock returns (Harvey, Liu, and Zhu, 2016; Hou *et al.*, 2020; Chen and Zimmermann, 2022). Their role and information content about future market returns may be uneven. Given the substantial differences in the predictive power of the various anomaly sets seen in Section 4.1, the selection of signals may be an important factor. Hence, we are interested in knowing how sensitive the return predictability is to the choice of anomalies that are included in the set of predictors.

To explore this issue, we run a bootstrap simulation of alternative samples of anomalies. We randomly select 100 anomalies from different datasets. Specifically, we use the four samples of long–short portfolios for the US market that were described in Section 4.1 (JKP3, JKP10, CZ, and HXZ) and run 1,000 bootstrap simulations for each of them. During each draw, we randomly chose 100 anomalies—the number that was originally employed by DLRZ. Next, for each selection, we employ all seven models described in Section 2.3 to predict market returns and use the full battery of prediction performance evaluation measures: predictive R^2_{OS}, utility gains, and Sharpe ratios. We want to see the potential dispersion in results that would stem from the selection of factors.

[14] As we already noted when discussing Table V, while our comparisons rely on the prevailing mean forecast, the buy-and-hold benchmark offers a viable alternative—which tends to be harder to beat. In an unreported analysis, we find that its annualized Sharpe ratio equals 0.57. In consequence, none of the competing models can outperform it; even the strategies based on the DLRZ dataset fail to succeed.

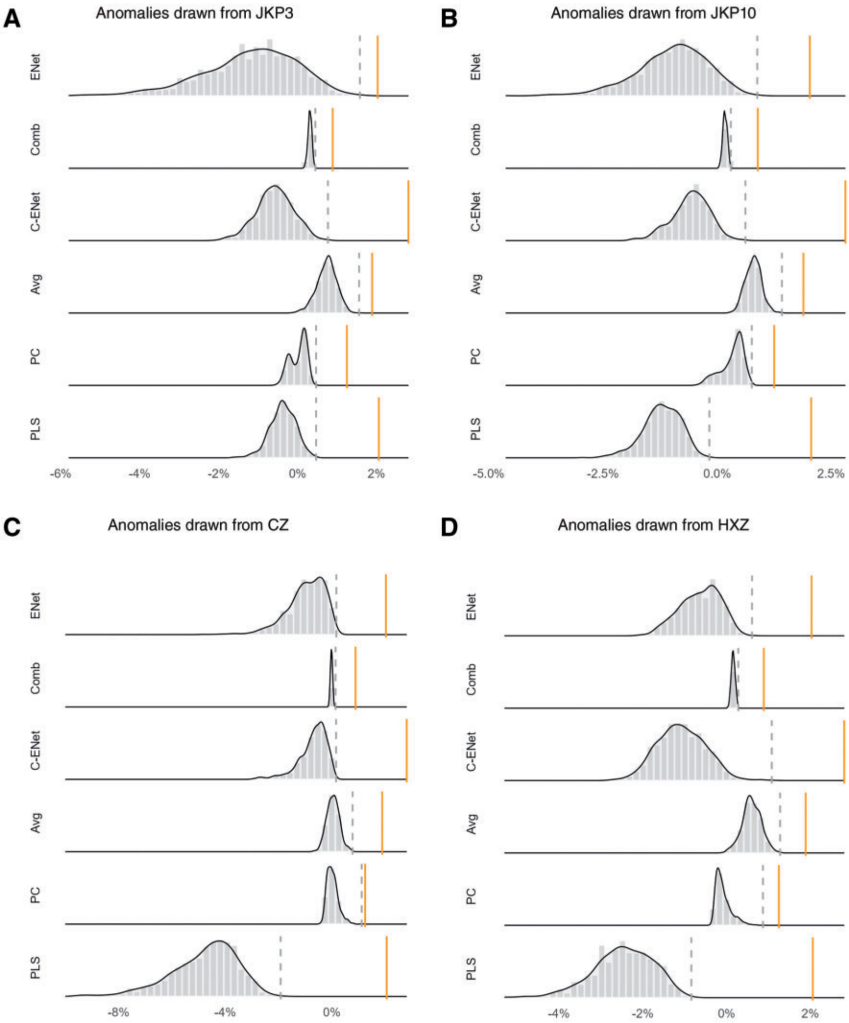

Figure 1. Predictive R^2 distributions for random anomaly sets. The figure illustrates the predictive R^2_{OS} coefficient by Campbell and Thompson (2008) for random selections of different long–short anomaly portfolios. We consider six different prediction models: the ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. These are supplied with 100 long–short anomaly portfolios randomly selected from one of four anomaly sets: 153 tercile and decile portfolios based on JKP3 (JKP10); 207 portfolios from CZ; and 188 anomalies from HXZ. Panels A, B, C, and D report the results for 1,000 random draws from the JKP3, JKP10, CZ, and HXZ sets, respectively. The underlying sample comprises US stocks, and the study period is from January 1970 to December 2017. The R^2_{OS} are reported in percentage terms. The histograms are supplied by the Gaussian kernel density plots. The gray dashed vertical lines represent the draw with the highest R^2_{OS}, and the orange vertical lines indicate the results of DLRZ.

Figure 1 displays the distributions of predictive R^2_{OS} measures calculated based on the random samples of anomalies. For comparison, we also illustrate the initial results based on the DLRZ sample. Supplementary Appendix Table A10 provides more detailed statistics on the R^2_{OS} distributions, along with the associated Clark and West's (2007) *t*-statistics. Notably, we limit the presentation of the OLS results to the Supplementary Appendix; this is because both DLRZ and our analyses generate incomparably large negative R^2_{OS}.

The random draws reveal a substantial variation in the prediction performance associated with various anomaly sets. The dispersion is particularly vivid for certain prediction models—such as ENet, C-ENet, or PLS. Notably, these models emphasize a small set of selected variables. Observe, for example, the JKP3 bootstrap results. For ENet, the 1st and 99th distribution percentiles of R^2_{OS} equal roughly −4.5% and 1%, respectively. Depending on the anomaly choice, it implies that the prediction accuracy could be either substantially negative or slightly positive. The patterns are qualitatively similar for all anomaly sets, marking a consistent magnitude of dispersion for the same methods. The only exception is the Comb model, in which the variation in results is substantially lower.

Most importantly, regardless of the model and anomaly set, the bootstrapped results fall visibly behind the original forecasting accuracy seen in DLRZ. The DLRZ anomaly sample generates incomparably higher R^2_{OS} than our bootstrap experiment, consistently beating the best single random draw. Let us reiterate it: literally, no combination of anomalies from any of our samples produces results comparably strong as the original DLRZ sample. Supplementary Appendix Table A10, Panel B, confirms the same observation at the *t*-statistics level. Even the top percentile yields insignificant findings for most anomaly sets and prediction models. In other words, not only is it unfeasible to match the actual results from DLRZ with any combination of established factors, but it is also hardly possible to generate any significant predictability at all.

While Figure 1 focused on the R^2_{OS} coefficient, Supplementary Appendix Figures A1 and A2 extend this analysis to the other prediction performance measures: utility gains and Sharpe ratios. The results are consistent, confirming the key conclusions from the R^2_{OS} analysis. The utility gains and Sharpe ratios exhibit notable dispersion originating from anomaly selection—with the highest variation observable for ENet, C-ENet, and PLS, and the lowest for Comb. Again, the DLRZ sample performs remarkably better than all simulated anomaly sets.

4.3 Anomaly Portfolio Construction

Our considerations in Sections 4.1 and 4.2 reveal that the established sets of anomalies from the US market fail to generate any evidence of return predictability. However, each of these sets relies on some specific portfolio construction rules. Recent studies documented that the design of factor portfolios may critically affect their properties, and consequently, their expected returns (Walter, Weber, and Weiss, 2022; Menkveld *et al.*, 2023; Soebhag, van Vliet, and Verwijmeren 2023). Most importantly, the large number of arbitrary construction choices paves the way for many different factor portfolio designs. Not only is there no broadly acclaimed way to form the factor strategies, but these discretionary decisions may also materially affect the market return predictability.

Can some specific portfolio designs improve the performance of our prediction models? To scrutinize this issue, we reproduce our calculations with alternative implementations of factor strategies. To explore this issue, we consider eight common methodological decisions on sample preparation and portfolio construction. They pertain to weighting scheme, anomaly significance, microcaps inclusion, minimum share price, selection of share codes, industry exclusions, return winsorization, and breakpoint selection. A detailed description of the data and methods used in this experiment is available in Supplementary Appendix Section B. In total, the combination of all possible decisions yields 2,592 variants of portfolio construction. Notably, these also include all methodological approaches employed by DLRZ, including the specific combination of sample selection and portfolio construction choices made in their study. For each of these variants, we run our standard tests to see whether any of these combinations lead to measurable improvements in return predictability.

Figure 2 presents the distributions of prediction performance measures for all 2,592 portfolio designs.[15] Clearly, the research design choices affect the effectiveness of market return forecasts. The dispersion of results is substantial, in particular ENet and C-ENet. Observe, for instance, ENet results for the JKP dataset. The worst 1% of portfolio specifications produces R^2_{OS} lower than −6%, while the best top percentile exceeds 0.07% (see Supplementary Appendix Table A11 for details). In other words, one could conclude that the anomaly-based forecasts either destroy or generate value for investors—depending on the assumed portfolio construction. The market return predictability seems strongly reliant on the research design.

Importantly, as seen in Figure 2, nearly all construction variants in Figure 2 fare considerably worse than the models based on the original DRLZ anomaly sample. For most models, the prediction performance of the DLRZ sets vastly outperforms even the best possible combination of all methodological choices—for both JKP and CZ datasets. Moreover, a further look at the t-statistic distributions in Supplementary Appendix Table A11 suggests that very few factor implementations yield significant results. The t-statistics exceed the 5% significance threshold only in between 1% and 10% of the factor designs—at best. For some methods, such as ENet, even the top percentile fails to generate any significant predictability. Finally, if we additionally account for the Bonferroni adjustment for multiple model testing, the predictability cannot be confirmed even for the top percentile of the portfolio variants.

Last, we are also interested in knowing how particular choices affect the market return predictability with anomalies. For example, what is the marginal benefit of utilizing a specific weighting scheme or share price filter? To shed light on this issue, for each methodological choice, we calculate the average R^2_{OS} scores across all the portfolio specifications that it includes. For example, for value-weighted portfolios, we calculate the average R^2_{OS} across the 864 specifications among all 2,592 that involve this security weighting scheme.

The results in Table IX indicate that the forecasting model performance might be sensitive to several categories of portfolio design choices. Overall, the return predictability seems to benefit from emphasizing the role of large and liquid companies in anomaly portfolios. For example, the R^2_{OS} coefficients are noticeably better for value-weighted portfolios than for equal-weighted ones. Analogously, excluding microcaps and filtering out penny stocks boost prediction accuracy marginally. Nevertheless, one conclusion remains unchanged regardless of this relative variation in results: the evidence for return predictability is unclear at best.

A separate remarkable insight from Table IX is the role of anomaly significance. Specifically, we do not observe any evident link between the prediction performance and the significance of anomalies. Specific models fare even marginally better on average when fed with portfolios with insignificant mean returns rather than significant ones. Notably,

[15] We provide further details of the distributions—as well as the OLS results—in Supplementary Appendix Table A11.

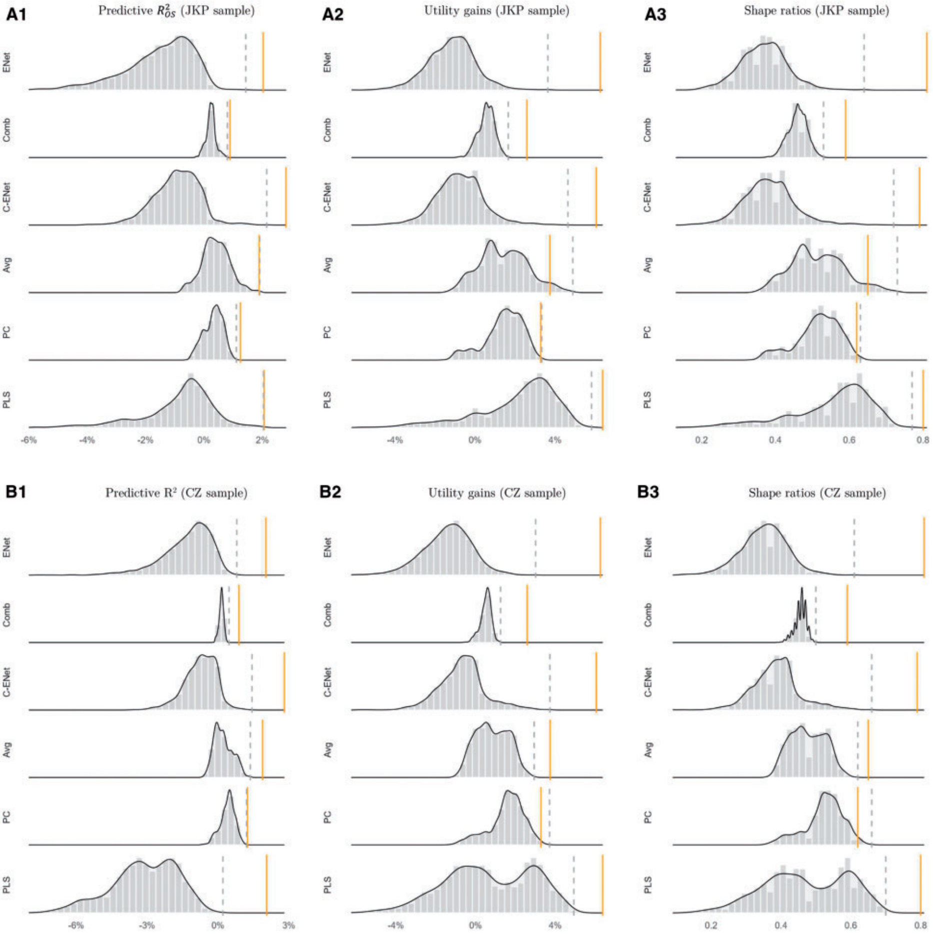

Figure 2. Prediction performance distributions for alternative anomaly portfolio designs. The figure illustrates the distributions of prediction performance measures for various methodological choices to predict market returns with anomaly portfolios. The considered prediction models include: the ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The models in Panels A and B use anomalies from the sample of JKP3 and CZ, respectively. The equity universe comprises US stocks. The study period is from January 1970 to December 2017, and the testing period starts in January 1985. The distributions are based on 2,592 variants of anomaly implementations, accounting for different weighting schemes (equal-, value-, or capped-value weighting); minimum share price (1$, 5$, or none); treatment of financial companies (included or excluded), utility companies (included or excluded), and micro-cap stocks (included or excluded); inclusion of different CRSP share price codes (all or 10 and 11 only); winsorization of stock returns (at 99%, 99.9%, or none); equity universes used to determine portfolio breakpoints (CRSP or NYSE), and anomaly significance (all, significant only, insignificant only). The reported performance measures are the out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008), the annualized average utility gains, and annualized Sharpe ratios for an investor who allocates between the US market portfolio and risk-free Treasury bills. The histograms are supplied by the Gaussian kernel density plots. The gray dashed vertical lines represent the best replication among all 2,592 variants, and the orange lines indicate the original results from DLRZ. (A.1) Predictive R^2_{OS} (JKP sample). (A.2) Utility gains (JKP sample). (A.3) Sharpe ratio ratios (JKP sample). (B.1) Predictive R^2 (CZ sample). (B.2) Utility gains (CZ sample). (B.3) Sharpe ratio ratios (CZ sample).

Table IX. Average predictive R^2 coefficients for different anomaly portfolio designs

The table illustrates the impact of various methodological choices on the predictive performance of long–short anomaly portfolio returns. Seven different models are employed to forecast market excess returns using anomaly portfolio returns: conventional OLS, ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The models in Panels A and B use anomalies from the sample of JKP3 and CZ, respectively. The equity univers comprises US stocks. The study period is from January 1970 to December 2017 and the testing period starts in January 1985. The distributions are based on 2,592 variants of anomaly implementations, accounting for different weighting schemes (equal-, value-, or capped-value weighting); minimum share price (1$, 5$, or none); treatment of financial companies (included or excluded), utility companies (included or excluded), and micro-cap stocks (included or excluded); inclusion of different CRSP share price codes (all or 10 and 11 only); winsorization of stock returns (at 99%, 99.9%, or none); equity universes used to determine portfolio breakpoints (CRSP or NYSE), and anomaly significance (all, significant only, insignificant only). For each methodological choice, we calculate the average out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008) across all specifications calculated using the given approach. The R^2_{OS} values are reported in percentage terms. The last two columns represent means and medians across all seven considered models, with the OLS model excluded. The underlying sample comprises US stocks and the study period runs from January 1970 to December 2017; the testing period starts in January 1985.

Panel A: JKP sample

Choice	Option	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Weighting scheme	Equal-weighted	−89.06	−1.80	0.13	−1.06	0.02	0.27	−1.19
Weighting scheme	Capped value-weighted	−73.46	−1.25	0.27	−0.91	0.51	0.56	−0.44
Weighting scheme	Value-weighted	−52.21	−0.95	0.28	−0.74	0.66	0.22	−0.39
Anomalies	All	−171.23	−1.30	0.25	−1.20	0.65	0.42	−0.58
Anomalies	Insignificant	−11.66	−1.27	0.27	−0.52	0.40	0.24	−0.88
Anomalies	Significant	−72.75	−1.38	0.21	−1.00	0.20	0.36	−0.49
Minimum share price	None	−73.45	−1.24	0.22	−0.90	0.35	0.30	−0.73
Minimum share price	1$	−73.51	−1.25	0.23	−0.93	0.39	0.31	−0.67
Minimum share price	5$	−71.01	−1.43	0.28	−0.84	0.45	0.44	−0.40
Financial companies	Included	−72.26	−1.70	0.25	−0.85	0.45	0.32	−0.69
Financial companies	Excluded	−72.80	−0.97	0.23	−0.93	0.34	0.40	−0.51
Utility companies	Included	−73.21	−1.07	0.22	−0.79	0.35	0.35	−0.63
Utility companies	Excluded	−72.25	−1.48	0.27	−0.97	0.42	0.37	−0.53
Micro-cap stocks	Included	−73.81	−1.10	0.18	−0.97	0.35	0.22	−1.17
Micro-cap stocks	Excluded	−71.88	−1.50	0.30	−0.84	0.42	0.47	−0.33
Share codes included	All	−73.37	−1.29	0.21	−0.92	0.36	0.32	−0.77
Share codes included	10 and 11	−72.54	−1.34	0.28	−0.87	0.43	0.40	−0.39
Winsorization	None	−72.26	−1.32	0.24	−0.89	0.35	0.35	−0.60
Winsorization	At 99.9%	−73.18	−1.30	0.23	−0.90	0.38	0.35	−0.60
Winsorization	At 99%	−72.87	−1.32	0.24	−0.89	0.44	0.37	−0.57
Factor breakpoints	Based on CSRP	−78.11	−1.29	0.22	−0.94	0.36	0.34	−0.82
Factor breakpoints	Based on NYSE	−65.78	−1.34	0.26	−0.84	0.42	0.38	−0.45

(continued)

Table IX. Continued

Panel B: CZ sample

Choice	Option	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Weighting scheme	Equal-weighted	−1874.60	−1.35	0.11	−0.74	0.08	0.43	−3.62
Weighting scheme	Capped value-weighted	−167.22	−1.28	0.20	−0.50	0.25	0.60	−2.57
Weighting scheme	Value-weighted	−195.93	−0.76	0.13	−0.77	0.22	0.31	−2.90
Anomalies	All	−17,558.23	−1.14	0.16	−0.86	0.65	0.53	−3.16
Anomalies	Insignificant	−62.66	−1.08	0.13	−0.59	−0.04	0.26	−3.21
Anomalies	Significant	−162.24	−1.04	0.15	−0.62	0.05	0.52	−2.71
Minimum share price	None	−509.43	−0.92	0.15	−0.72	0.16	0.40	−3.11
Minimum share price	1$	−504.95	−1.10	0.14	−0.62	0.11	0.43	−3.14
Minimum share price	5$	−455.14	−1.26	0.16	−0.72	0.23	0.49	−2.86
Financial companies	Included	−491.79	−1.24	0.14	−0.71	0.19	0.40	−3.53
Financial companies	Excluded	−457.90	−0.93	0.16	−0.66	0.17	0.50	−2.17
Utility companies	Included	−514.62	−1.28	0.13	−0.65	0.19	0.42	−3.25
Utility companies	Excluded	−428.55	−0.93	0.17	−0.73	0.16	0.47	−2.76
Micro-cap stocks	Included	−430.37	−0.98	0.11	−0.58	0.10	0.33	−3.29
Micro-cap stocks	Excluded	−546.57	−1.26	0.18	−0.78	0.22	0.51	−2.67
Share codes included	All	−509.43	−1.05	0.14	−0.67	0.12	0.42	−3.18
Share codes included	10 and 11	−445.14	−1.12	0.16	−0.71	0.21	0.47	−2.85
Winsorization	None	−471.28	−1.06	0.15	−0.71	0.17	0.45	−3.12
Winsorization	At 99.9%	−479.81	−1.11	0.15	−0.70	0.18	0.45	−3.08
Winsorization	At 99%	−476.33	−1.10	0.15	−0.66	0.17	0.45	−2.91
Factor breakpoints	Based on CSRP	−450.82	−1.07	0.13	−0.75	0.19	0.45	−3.41
Factor breakpoints	Based on NYSE	−519.73	−1.13	0.17	−0.62	0.17	0.45	−2.62

this observation is inconsistent with the "mispricing correction mechanism," which serves as the theoretical foundation for return predictability by anomaly returns in DLRZ. Consequently, our findings not only fail to align with the earlier empirical evidence but also contradict the proposed theoretical mechanism.

To conclude, the prediction performance of equity anomalies partly depends on anomaly portfolio design. Specific construction methods can improve forecasting performance. Nonetheless, the overall picture is clear: hardly any specifications produce significant predictability.

5. Why Do Our Results Differ?

Our findings thus far visibly depart from the seminal study of DLRZ. While they document robust market return predictability by equity anomalies, we can hardly detect any. Since our methodological approach is identical—and the tests in Section 4.3 encompass all possible choices taken and not taken by DLRZ—the discrepancy must stem from the sample of equity anomalies and the implementation of the anomalies. Hence, we now look closer at the contribution of individual anomalies to return predictability, as well as the role of various implementation choices.

5.1 Return Predictability by Individual Anomalies

We begin with an overview of the predictability of market returns by various anomalies. To this end, we run univariate predictive regressions of market portfolio excess returns by individual factors. We examine five sets from our previous analyses: DLRZ, JKP10, JKP3, CZ, and HXZ. Table X tabulates the results of this exercise, with Panel A reporting the R^2_{OS} values for the top fifteen anomalies in each dataset, and Panel B providing additional summary statistics. The full R^2_{OS} statistics for all anomalies can be found in Supplementary Appendix Table A12.

A quick glance reveals remarkable differences between the different anomaly sets. The DLRZ selection stands out, both in terms of the number of significant predictors and their sheer predictive power. As seen in Panel B, as many as eighteen anomalies, representing 18% of the dataset, are significant return predictors at the 5% level (on a standalone basis). On the other hand, in other datasets, the analogous proportion does not exceed 7%. In particular, in the selection of 207 anomalies by CZ, not a single portfolio helps to predict market returns reliably. Notably, among all the anomaly samples we consider, CZ come closest to replicating the original studies by strictly following their portfolio construction procedures.

Finally, it is worth noting that when we adjust the test statistics for the multiple hypotheses framework—using either the Bonferroni or Romano and Wolf (2016) approach—no single anomaly in any dataset passes the 5% significance threshold. Admittedly, both adjustments are relatively strict. Nevertheless, it is a refreshing reminder that, given a sufficiently large number of anomalies, return predictability can occur by chance.

In addition to the larger number of significant predictors, the DLRZ anomalies also have a higher average R^2_{OS} values. For example, the average R^2_{OS} for the top fifteen strategies shown in Panel A is 1.94% for the DLRZ and ranges from 0.39% (CZ) to 0.95% (JKP10) for the other datasets. When all anomalies from DLRZ are considered, the average R^2_{OS} for the 100 strategies is positive, equal to 0.27%, while for other datasets, such averages are negative, ranging from −0.43% (CZ) to −0.20% (JKP3). In other words, while a typical anomaly in DLRZ helps predict market portfolio returns, a typical anomaly in other recognized datasets has no predictive ability at all. The anomalies used by DLRZ clearly outperform other samples.

Which anomalies are the best predictors of future returns? Panel A of Table X reveals a surprising truth. First, consider the DLRZ sample. The leaders contain almost exclusively two types of predictors: issuance effects and anomalies related to earnings surprises. Specifically, the top fifteen variables include three net share issuance anomalies that differ only in rebalancing frequency (NSIMO, NSIANN, and NSIFY), two composite equity issuance variables (CEIANN and CEIMO), and two share issuance anomalies (SHR5MO and SHR5ANN). The second notable category is anomalies related to earnings surprises and beating analysts' forecasts: FERR, CSUE, and SUE.

Interestingly, the predictors related to the issuance and earnings surprise often top the list in other samples as well: JKP10, JKP3, CZ, and HXZ. Nonetheless, neither their R^2_{OS} are as high as in DLRZ, nor are they comparably abundant. In fact, there are more

Table X. Prediction performance of individual factor returns

The table illustrates the out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008) for the US market excess returns forecast based on returns on individual asset pricing factors. We consider five different sets of long–short anomaly portfolios: the original set used in DLRZ, 153 tercile and decile portfolios based on JKP3 (JPK10); 207 portfolios from CZ; and 188 portfolios from HXZ. All the anomalies are derived from the US market and the prediction models rely on recursive training windows. The underlying sample comprises US stocks. The total study period is from January 1970 to December 2017; the testing period starts in January 1985. The R^2_{OS} are expressed in percentage terms. The numbers in parentheses are Clark and West's (2007) *t*-statistics. Panel A reports the fifteen variables with the highest R^2_{OS} in each anomaly set. Panel B displays the summary statistics: averages for all and top-fifteen variables (Average all, Average top 15), the total number of factors (# factors), the number and percentage of R^2 coefficients significant at the 5% level at a standalone basis (# signif. (raw), % signif. (raw)), and after adjustment for multiple testing framework using the Bonferroni (# signif. (Bonf), % signif. (Bonf)) and Romano and Wolf (2016) (# signif. (RW), % signif. (RW)) methods. The symbols of anomalies are explained in Supplementary Appendix Tables A1, A7, A8, and A9.

DLRZ			JKP10			JKP3			CZ			HXZ		
Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW
Panel A: Top 15 anomalies														
NSIMO	4.29	3.19	ncoa_gr1a	2.05	2.64	dbnetis_at	1.79	2.47	CashProd	0.82	1.54	NSI	1.67	1.87
FERR	3.14	2.61	chcsho_12m	1.92	2.30	coa_gr1a	1.47	2.32	PS	0.65	1.13	EM	1.42	2.46
CEIANN	2.60	2.70	mispricing_mgmt	1.52	2.16	cop_at	1.25	2.01	EarningsSurprise	0.62	1.38	NOP	1.16	1.86
NSIANN	2.55	2.78	nncoa_gr1a	1.11	1.92	mispricing_mgmt	1.06	1.90	CompEquIss	0.49	1.08	CEI	1.10	1.82
CEIMO	2.40	2.53	netis_at	1.08	1.86	ocf_at	1.01	1.76	DebtIssuance	0.44	1.48	TBIq_12	0.98	2.03
CSUE	1.82	2.50	rskew_21d	0.98	1.64	ni_ar1	0.92	1.71	hire	0.39	1.28	IVG	0.91	1.85
SUE	1.77	2.42	dbnetis_at	0.78	1.47	dsale_dinv	0.86	2.07	IdioVolAHT	0.37	1.04	OA	0.90	1.73
CHATOIA	1.61	2.06	eqnpo_12m	0.72	1.55	ncoa_gr1a	0.81	1.89	ChNNCOA	0.33	1.26	CP	0.55	1.42
ROEQ	1.53	1.78	ivol_capm_252d	0.70	1.37	ppeinv_gr1a	0.81	1.66	PriceDelayRsq	0.30	1.22	TBIq_6	0.54	1.38
SHR5MO	1.38	2.29	col_gr1a	0.64	1.38	fnl_gr1a	0.74	1.38	VolMkt	0.30	1.01	EG_1	0.53	1.37
CFPJUN	1.33	2.20	qmj_prof	0.62	1.36	noa_gr1a	0.71	1.68	DolVol	0.25	1.02	POA	0.53	1.78
NSIFY	1.32	2.13	turnover_126d	0.60	1.33	mispricing_perf	0.70	1.66	NumEarnIncrease	0.24	0.91	dBE	0.44	1.26
CURRAT	1.14	1.91	zero_trades_21d	0.54	1.24	rd5_at	0.61	1.47	OptionVolume1	0.22	0.94	dWC	0.43	1.20
ROM	1.09	2.04	o_score	0.49	1.35	nncoa_gr1a	0.54	1.46	CBOperProf	0.21	0.98	Rev_6	0.39	1.05
SHR5ANN	1.07	2.04	mispricing_perf	0.48	1.24	qmj_prof	0.52	1.34	DelCOA	0.21	1.04	EG_6	0.37	1.24

(continued)

Table X. Continued

DLRZ			JKP10			JKP3			CZ			HXZ		
Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW	Anomaly	R^2_{OS}	CW
Panel B: Summary statistics														
Average all	0.27	0.56	Average all	−0.30	−0.04	Average all	−0.20	0.07	Average all	−0.43	−0.20	Average all	−0.25	0.02
Average top 15	1.94	2.35	Average top 15	0.95	1.65	Average top 15	0.92	1.79	Average top 15	0.39	1.15	Average top 15	0.79	1.62
# factors	100		# factors	153		# factors	153		# factors	207		# factors	188	
# signif. (raw)	18		# signif. (raw)	5		# signif. (raw)	11		# signif. (raw)	0		# signif. (raw)	8	
% signif. (raw)	18%		% signif. (raw)	3%		% signif. (raw)	7%		% signif. (raw)	0%		% signif. (raw)	4%	
# signif. (Bonf)	0		# signif. (Bonf)	0		# signif. (Bonf)	0		# signif. (Bonf)	0		# signif. (Bonf)	0	
% signif. (Bonf)	0%		% signif. (Bonf)	0%		% signif. (Bonf)	0%		% signif. (Bonf)	0%		% signif. (Bonf)	0%	
# signif. (RW)	0		# signif. (RW)	0		# signif. (RW)	0		# signif. (RW)	0		# signif. (RW)	0	
% signif. (RW)	0%		% signif. (RW)	0%		% signif. (RW)	0%		% signif. (RW)	0%		% signif. (RW)	0%	

significant issue-related predictors in the DLRZ anomaly sample than there are significant predictors in any of the other samples. The DLRZ sample contains the broadest representation of the successful categories of issuance and earnings surprise variables. As many as nine anomalies (out of 100) are directly related to issuance (CEIANN, CEIFY, CEIMO, NSIANN, NSIFY, NSIMO, SHR1, SHR5ANN, SHR5MO) and six more are related to earnings surprises (FERR, CHFEPS, CSUE, SUE, EAR, RSUP).

Notably, not only does the DRLZ sample contain more significant anomalies, but their predictive ability exceeds that of variables from other samples. For example, the top five anomalies in DRLZ have R^2_{OS} ranging from 2.40% (CEIMO) to 4.29% (NSIMO). For comparison, literally, no other anomaly in any of the other datasets reaches a similar level of predictive power—even though some of them reflect similar economic phenomena.

In summary, the power of the original DLRZ sample lies in its composition and design. The DLRZ dataset contains more anomaly categories with superior predictive abilities: issuance and earnings surprises. Moreover, their individual predictive power is significantly higher than that of their counterparts from other datasets.

5.2 Which Anomalies Drive the Models' Forecasts?

So far, we have examined the predictive power of individual anomalies treated on a standalone basis. But which of them contribute to the aggregate predictions of different models? To explore this, we perform several additional analyses.

5.2.a. The importance of individual anomalies

First, we compute the variable importance (VI) of individual anomalies. Specifically, similar to Gu, Kelly, and Xiu (2020) and Leippold, Wang, and Zhou (2022), we define VI as the reduction in the overall predictive R^2_{OS} due to discarding information from a selected anomaly. We perform this exercise for the DLRZ set, that is, the 100 anomalies from the original DLRZ study.

As reported in Table XI, the VI analysis shows that the methods are very well able to concentrate on the relevant variables (signal) and to disregard irrelevant variables (noise). The models extract predictability almost exclusively from the two variable categories that were found to be most important in Table X: issuance and earnings surprises. These two categories dominate the top of the VI ranking. All other anomalies play a minor role. In other words, all return predictability seems to be driven by a handful of closely related signals.

The tests in Sections 5.1 and above highlight the unique role of issuance anomalies. To further illustrate this phenomenon, we zoom in on the monthly rebalanced net stock issuance anomaly (NSIMO), which showed the strongest predictive power in Section 5.1. To illustrate its importance, we contrast it with the aggregate models of DLRZ. Specifically, we run the forecast encompassing tests of Harvey *et al.* (1988). We are interested in knowing whether the forecasting models that aggregate information from 100 anomalies encompass the NSIMO-based forecasts. Table XII reports the results.

As shown in Panel A, no model encompasses the NSIMO predictions. The only exception is OLS, where the results are not significant. On the other hand, the results in Panel B answer the opposite question: Does the NSIMO anomaly encompass the predictions of the aggregate models? In this case, the answer is yes. No single coefficient is statistically significant. In other words, the aggregate models do not provide valuable information about

Table XI. VI for market return predictions

The table presents the VI for the US market excess returns forecast based on various sets of long–short anomaly portfolio return. The VI (ΔR^2_{OS}) is calculated as the reduction in the out-of-sample R^2 coefficients by Campbell and Thompson (2008) due to exclusion of the given variable from the sample. The anomaly sample is the original set of 100 anomalies from DLRZ and the underlying stock universe comprises US stocks. ΔR^2_{OS} is reported in percentage. The study period runs from January 1970 to December 2017; the testing period starts in January 1985.

ENet			Comb		C-Enet		Avg		PC		PLS	
Rank	Anomaly	ΔR^2_{OS}	Anomaly	ΔR^2_{OS}	Anomaly	ΔR^2_{OS}	Anomaly	ΔR^2_{OS}	Anomaly	ΔR^2_{OS}	Anomaly	ΔR^2_{OS}
1	NSIANN	0.67	NSIMO	0.08	NSIMO	0.99	FERR	0.12	FERR	0.08	FERR	0.39
2	NSIMO	0.56	FERR	0.07	FERR	0.95	DEPR	0.08	CHFEPS	0.07	NSIMO	0.19
3	FERR	0.32	NSIANN	0.06	ABSACC	0	CFPIA	0.06	NSIMO	0.06	CHATOIA	0.17
4	RSUP	0.21	CEIANN	0.04	AGE	0	CHATOIA	0.06	CEIANN	0.04	CHFEPS	0.16
5	NOA	0.01	CEIMO	0.04	AGR	0	TANG	0.06	CEIMO	0.04	NSIANN	0.08
6	ABSACC	0	CHFEPS	0.03	BETA1	0	CINVEST	0.05	NSIANN	0.04	PCHGMSALE	0.08
7	ACC	0	CHATOIA	0.02	BETA1LAG	0	CSUE	0.05	ROM	0.03	CSUE	0.07
8	AGE	0	CSUE	0.02	BETA3	0	EAR	0.05	ACC	0.02	SUE	0.07
9	AGR	0	NSIFY	0.02	BETA3LAG	0	INDMOM12M	0.05	AGR	0.02	EAR	0.05
10	BETA1	0	SHR1	0.02	BM	0	MOM36M	0.05	CFPJUN	0.02	HERF	0.05

Table XII. Net share issuance versus 100 anomalies: a forecast encompassing tests

The table reports Harvey, Leybourne, and Newbold (1998) forecast encompassing tests for comparing the predictions from the models of DLRZ and the predictions from a univariate prediction model based on the returns on the monthly rebalanced net share issuance (NSIMO) anomaly. The considered prediction models include: the ENet, simple combination (Comb), C-ENet, predictor average (Avg), PC, and PLS. The models aggregate the returns on the 100 anomalies from the original dataset from DLRZ and the NSIMO returns are obtained from the same source. Panel A verifies the null hypothesis that a given model of DLRZ contains all of the information in the NSIMO anomaly forecasts (i.e., encompasses it) against the alternative hypothesis that the NSIMO model contains additional valuable information for forecasting. Panel B verifies the null hypothesis that the NSIMO anomaly forecasts encompass the forecasts from a given model of DLRZ against the alternative hypothesis that this model contains additional information. λ indicates the forecast weight in the joint model and the numbers in parentheses are the test statistics of Harvey *et al.* (1988). We follow the implementation by Neely *et al.* (2014). The study period is from January 1970 to December 2017 and the testing period starts in January 1985. The asterisks * and ** indicate values significant at the 5% and 1% levels, respectively.

	OLS	ENet	Comb	C-ENet	Avg	PC	PLS
Panel A: Do the models of DLRZ encompass the NSIMO anomaly?							
λ	1.00	1.13**	0.87**	0.92*	0.89**	0.96**	0.91**
	(*1.03*)	(*2.50*)	(*2.93*)	(*2.24*)	(*2.80*)	(*2.95*)	(*2.81*)
Panel B: Does the NSIMO anomaly encompasses the models of DLRZ?							
λ	0.00	−0.13	0.13	0.08	0.11	0.04	0.09
	(*0.40*)	(*−0.29*)	(*0.48*)	(*0.19*)	(*0.34*)	(*0.12*)	(*0.31*)

future returns beyond what is already represented by NSIMO. Simply put, NSIMO alone beats all other forecasting models.

The results in Table XII illustrate the critical dependence of return predictability on factor selection. A single well-designed issuance anomaly from the dataset can beat the aggregate models, rendering them redundant. A natural question follows: What happens if such a specific anomaly is missing from the dataset? Section 5.2.b explores this question in more detail.

5.2.b. Market return predictions with limited anomaly samples

To further understand the importance of different categories of anomalies, we now examine the predictability of market returns based on limited samples of anomalies. Our results so far suggest a unique role for issuance anomalies, followed by the signals associated with equity issuance, as well as earnings surprises and beating analysts' forecasts. Interestingly, these groups of relevant anomalies align with the findings of Daniel, Hirshleifer, and Sun (2020), who demonstrate that share issuance and earnings surprises explain the cross-section of stock returns particularly well. This may suggest that the time-series predictability originates specifically from those anomalies that capture the best cross-sectional variation in stock returns. To examine their impact on the overall results, we now examine the dependence of market return predictability on their inclusion in the anomaly sample.

We begin by creating two groups of predictors that represent similar economic phenomena; we call them issuance (CEIANN, CEIFY, CEIMO, NSIANN, NSIFY, NSIMO, SHR1, SHR5ANN, SHR5MO) and surprise (FERR, CHFEPS, CSUE, SUE, EAR, RSUP) anomalies. The average pairwise correlation coefficient within these two categories is 0.63 and 0.40, respectively, indicating their close internal relationship. Notably, the correlation between certain specific pairs of anomalies is close to one, signaling that they are almost identical. For example, the Pearson correlation coefficient between RSUP and SUE (the sixth and seventh highest R^2_{OS} in Table X) is 0.9996 and the correlation between CEIANN and CEIMO (the third and fifth highest R^2_{OS} in Table X) is 0.9.

Table XIII reports the prediction performance of models aggregating different anomaly sets. In general, we use the DLRZ set, that is, the original 100 anomalies from DLRZ. Specification (1) shows the results for the full sample, with the large and mostly significant R^2_{OS} values. What happens if we restrict the anomaly selection to only the nine issuance anomalies? As we can see in specification (2), the predictive performance improves considerably. The R^2_{OS} values increase noticeably, ranging from 0.73% (OLS) to 3.11% (C-ENet). Furthermore, they are highly significant for all models, even for OLS.

Specification (3) extends the set of factors considered to 15 and also adds the surprise anomalies. In this case, performance is still strong, and for some models, even stronger, exceeding 4% for Avg and PLS. Moreover, all associated timing strategies generate high Sharpe ratios, typically beating those based on naive forecasts.

What happens when we exclude these critical anomalies from the sample? As we can see in specification (4), the picture changes dramatically. As soon as only the issuance anomalies are eliminated, the predictability of returns disappears almost completely. Four out of seven models no longer have a significantly positive R^2_{OS}. Only two models—COMB and Avg—still have positive and significant R^2_{OS}, but even in these two cases, the *t*-statistics are borderline insignificant (1.69, 1.82). Specification (5) illustrates what happens when we also drop the group of six surprise anomalies. The results continue to deteriorate. No prediction model can generate a positive and significant predictive R^2_{OS}. Similarly, the Sharpe ratios do not exceed their counterparts based on naive forecasts.

To summarize our above considerations, the superior predictive abilities of the models in DLRZ derive their strength from the composition of the anomaly sample. Specifically, they over-represent specific categories of anomalies with unique predictive power, such as issuance and earnings surprises. These few correlated predictors alone suffice to generate reliable forecasts that outperform aggregate models. On the other hand, removing them from the sample eliminates any trace of return predictability.[16]

To better understand the nature of the return predictability by the critical anomalies—and the net issuance in particular—we run four additional exercises. First, we explore various research design choices to explore their impact on return predictability by NSIMO, the most prominent issuance anomaly. Second, we extend the tests of NSIMO anomaly to international markets. Third, we examine an extended study period from 1927 to 2021. Fourth, we build on the findings of McLean, Pontiff, and Watanabe (2009) to explore the link between return predictability by issuance anomalies and the country-specific constraints on issuance activity. These tests, described in detail in Supplementary Appendix Section C, reveal the fragility of the observed patterns. The return predictability by issuance

[16] Importantly, Dong *et al.* (2022) perform a series of subsample analyses in their paper based on different groups of anomalies (Supplementary Appendix Table IA.XIII therein). Nonetheless, none of the subgroups effectively eliminates all issuance or surprise anomalies, which prove critical in our case.

Table XIII. Prediction performance of the subsets of DLRZ sample

The table presents the measures of prediction performance for the US market excess returns forecast based on various sets of long–short anomaly portfolio returns: the out-of-sample R^2 coefficients (R^2_{OS}) by Campbell and Thompson (2008) (Panel A), the annualized average utility gains (Panel B), and the annualized Sharpe ratios for an investor who allocates between the US market portfolio and risk-free Treasury bills (Panel C). We consider five different sets of long–short anomaly portfolios: (1) the full original set of 100 DLRZ anomalies; (2) nine issuance anomalies from the DLRZ set only (CEIANN, CEIFY, CEIMO, NSIANN, NSIFY, NSIMO, SHR1, SHR5ANN, SHR5MO); (3) nine issuance and six surprise anomalies only (FERR, CHFEPS, CSUE, SUE, EAR, RSUP); (4) the full sample excluding the nine issuance anomalies; and (5) the full sample excluding the nine issuance and six surprise anomalies. The underlying stock universe comprises US stocks. The R^2_{OS} and utility gains are expressed in percentage terms. The numbers in parentheses are: in Panel A, Clark and West's (2007); and in Panel B, *t*-statistics from the test of Ledoit and Wolf (2008), which compares a given Sharpe ratio with the Sharpe ratio of the portfolio formed using the prevailing mean benchmark forecast (=0.42). The values in bold are significant at the 5% level in standalone tests (*t*-stat $>$ 1.645 in Panel and $|t\text{-stat}| > 1.96$ in Panel C); furthermore, the underline font indicates the 5% significance after the Bonferroni adjustment for multiple testing framework (*t*-stat $>$ 2.45 in Panel A and $|t\text{-stat}| > 2.69$ in Panel C). The study period is from November 1971 to December 2017 and the testing period starts in November 1986.

	Full sample	Issuance anomalies only	Issuance and surprise anomalies only	Issuance anomalies excl.	Issuance and surprise anomalies excl.
Panel A: Predictive R^2 coefficients					
OLS	−2513.86 (0.53)	$\underline{0.73\ (2.54)}$	−$\underline{1.55\ (3.02)}$	−116.54 (1.45)	−51.41 (1.07)
Enet	**2.03 (2.26)**	$\underline{2.74\ (2.67)}$	$\underline{3.26\ (2.79)}$	−1.27 (−0.55)	−0.42 (−1.48)
Comb	$\underline{0.89\ (2.50)}$	$\underline{2.52\ (2.60)}$	$\underline{3.25\ (3.23)}$	**0.48 (1.69)**	0.31 (1.20)
C-ENet	$\underline{2.81\ (2.49)}$	$\underline{3.11\ (2.76)}$	**2.48 (2.13)**	−1.49 (−0.12)	−0.96 (−0.97)
Avg	**1.89 (2.13)**	$\underline{2.02\ (2.62)}$	$\underline{4.52\ (3.02)}$	**1.51 (1.82)**	1.15 (1.62)
PC	**1.25 (1.84)**	$\underline{2.05\ (2.63)}$	$\underline{3.08\ (2.93)}$	0.63 (1.24)	0.52 (1.14)
PLS	$\underline{2.06\ (2.65)}$	$\underline{2.14\ (2.69)}$	$\underline{4.24\ (3.17)}$	**−0.32 (1.82)**	−1.56 (1.26)
Panel B: Utility gains					
OLS	−4.97	5.60	9.12	−0.58	−1.83
Enet	6.26	5.87	9.59	−0.49	−1.28
Comb	2.59	6.23	7.31	1.61	1.22
C-ENet	6.06	7.39	5.78	−1.78	−1.41
Avg	3.74	7.12	9.54	3.71	2.77
PC	3.28	7.23	7.84	2.60	2.45
PLS	6.38	7.18	9.17	3.74	1.77
Panel C: Sharpe ratios					
OLS	0.23 (−0.82)	0.71 (1.24)	**0.89 (1.97)**	0.43 (0.06)	0.36 (−0.18)
Enet	**0.81 (1.99)**	0.74 (1.35)	**0.95 (2.35)**	0.39 (−0.54)	0.34 (−1.71)
Comb	**0.59 (1.96)**	0.79 (1.85)	$\underline{0.88\ (2.73)}$	0.51 (1.43)	0.48 (1.10)
C-ENet	0.79 (1.79)	0.83 (1.73)	0.76 (1.87)	0.33 (−1.23)	0.33 (−1.22)
Avg	0.65 (1.18)	0.82 (1.74)	**0.96 (2.31)**	0.63 (1.18)	0.57 (0.88)
PC	0.62 (1.13)	0.83 (1.76)	0.86 (1.85)	0.57 (0.95)	0.56 (0.93)
PLS	0.80 (1.55)	0.82 (1.69)	**0.93 (2.06)**	0.63 (0.84)	0.52 (0.42)

is highly sensitive to research design and cannot be confirmed under many alternative implementations. Furthermore, we fail to find evidence for a similar phenomenon in international markets and an extended study period for the US market. Third, it does not follow cross-sectional patterns associated with return predictability. Lastly, let us reiterate that—as seen in Table X—even the high R^2_{OS} for issuance anomalies fail to pass the multiple hypotheses testing framework.

Identifying fully the economic mechanism behind the effect of issuance anomalies on future market returns is beyond the scope of our article and should be explored in future studies. While we refrain from making a final judgment, we acknowledge that this may even be a statistical artifact characteristic only for the USA, resulting from studying many anomalies simultaneously or specific to a particular research implementation.

6. Concluding Remarks

In an attempt to link the time-series and cross-sectional asset pricing literature, our study revisits the evidence on market risk premium predictability by equity anomalies. We build on the methodological framework of DLRZ and extend their findings to international markets and alternative implementation settings. We take a holistic approach and examine the prediction performance of seven models applied to up to 153 anomalies in forty-two countries. Our conclusions are disappointing: the anomalies fail to predict market excess returns. Any alleged evidence lacks external validity in two aspects: stock market sample and anomaly selection.

First, it does not extend internationally. In the prevailing majority of countries and tests, the models fail to produce positive and significant R^2_{OS} in international markets. This conclusion is robust to different tests, as well as various methodological modifications.

Second, the predictability does not extend to other anomaly sets. To demonstrate this point, we scrutinize anomalies that come from the well-known samples of HXZ, CZ, and JKP3. To ensure comparability, we now turn to the original study period of DLRZ, that is, from 1970 to 2018. We find that the return predictability critically depends on the choice of anomalies as inputs: it is strong for the factors of DLRZ but does not stretch to other samples.

To further strengthen this evidence, we explore various subsets of anomalies from the abovementioned samples. Specifically, we randomly choose combinations of 100 anomalies and examine their performance. Our simulations show that different anomaly selections can cause a sizeable dispersion in prediction results. Most importantly, however, virtually no combination can generate reliable forecasts.

Finally, we are also interested in the role of anomaly portfolio construction in the determination of return predictability. Ultimately, the long–short strategies could be formed in many ways—leading to variation in their eventual properties. Hence, we consider a battery of choices in anomaly portfolio design, including different weighting schemes, share price filters, winsorization rules, or breakpoint settings. Eventually, we obtain 864 distinct variants of anomaly strategies.

The results indicate that these methodological choices substantially impact the eventual prediction performance. For example, R^2_{OS} for the same model fed with the same selection of anomalies may range from profoundly negative to slightly positive—depending on the portfolio formation choices. Nevertheless, our primary conclusion remains intact: as a rule, anomalies fail to predict market returns. Hardly any design applied to hundreds of anomalies from JKP3 or CZ can generate trustworthy forecasts or match the predictive abilities of the 100 portfolios of DLRZ. Also, there are no systematic differences between significant and insignificant anomalies, casting doubt on the "mispricing correction mechanism," which serves as a theoretical foundation for the market portfolio return predictability by anomalies.

Notably, following the work of DLRZ and Engelberg *et al.* (2023), our article focuses on the implications of cross-sectional predictability for time-series predictability. A potential limitation of our article is that we do not investigate the implications of time-series predictability for cross-sectional predictability. However, theories explaining cross-sectional predictability—whether risk-based, mispricing-based, or data mining—do not necessarily imply a strong or direct relationship between anomaly returns and market returns. Our finding of no predictability thus aligns with all of the above explanations.

To sum up, our conclusions challenge the earlier findings from the US market. The existing evidence strongly depends on the examination settings: study period, market choice, anomaly selection, and portfolio construction. This sensitivity to methodological choices yields a considerable risk of false discoveries—statistical artifacts may be mistaken for real economic phenomena. Once certain methodological choices are modified, the forecasting abilities of stock market anomalies can no longer be confirmed.

Acknowledgements

We thank Pedro M. Barroso, Andrew C. Chang, Andrew Chen, Amit Goyal, Campbell R. Harvey, Erik Hjalmarsson, Alejandro Lopez-Lira, Jeffrey Pontiff, and Mihail Velikov for helpful comments and suggestions. We are solely responsible for any remaining errors.

Supplementary Material

Supplementary data are available at *Review of Finance* online.

Funding

Adam Zaremba acknowledges the support of the National Science Center of Poland [grant no. 2022/45/B/HS4/00451].

Data Availability

The data (and the code to retrieve the data) underlying the empirical sample in this article were obtained from the web pages of the following papers:

- Chen and Zimmermann (2022)
  - webpage: https://www.openassetpricing.com/
- Dong *et al.* (2022)
  - webpage: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13099
- Hou, Xue, and Zhang (2020)
  - webpage: https://global-q.org/testingportfolios.html
- Jensen, Kelly, and Pedersen (2023)
  - webpage: https://github.com/bkelly-lab/ReplicationCrisis

For Chen and Zimmermann (2022) and Jensen, Kelly, and Pedersen (2023), the publicly available code was used to download the data from WRDS databases (e.g., CRSP, Compustat) and calculate factor portfolio returns. In the cases of Dong *et al.* (2022) and Hou, Xue, and Zhang (2020), the factor return data were retrieved directly from the authors' websites.

References

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle, in: Petrov, Boris N. and Csaki, Frigyes (eds.), *Proceedings of the 2nd International Symposium on Information Theory*, Akadémiai Kiadó, Budapest.

Bali, T. G. and Cakici, N. (2008): Idiosyncratic volatility and the cross section of expected returns, *Journal of Financial and Quantitative Analysis* 43, 29–58.

Bali, T. G., Engle, R. F., and Murray, S. (2016): *Empirical Asset Pricing: The Cross Section of Stock Returns*, John Wiley & Sons.

Baltussen, G., Swinkels, L., and Van Vliet, P. (2021): Global factor premiums, *Journal of Financial Economics* 142, 1128–1154.

Cakici, N. and Zaremba, A. (2022): Salience theory and the cross-section of stock returns: international and further evidence, *Journal of Financial Economics* 146, 689–725.

Campbell, J. Y. and Thompson, S. B. (2008): Predicting excess stock returns out of sample: can anything beat the historical average?, *Review of Financial Studies* 21, 1509–1531.

Chen, A. Y. and Velikov, M. (2023): Zeroing in on the expected returns of anomalies, *Journal of Financial and Quantitative Analysis* 58, 968–1004.

Chen, A. Y. and Zimmermann, T. (2022): Open source cross-sectional asset pricing, *Critical Finance Review* 11, 207–264.

Clark, T. E. and West, K. D. (2007): Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* 138, 291–311.

Daniel, K., Hirshleifer, D., and Sun, L. (2020): Short- and long-horizon behavioral factors, *Review of Financial Studies* 33, 1673–1736.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022): Anomalies and the expected market return, *Journal of Finance* 77, 639–681.

Engelberg, J., McLean, R. D., Pontiff, J., and Ringgenberg, M. C. (2023): Do cross-sectional predictors contain systematic information?, *Journal of Financial and Quantitative Analysis* 58, 1172–1201.

Fama, E. F. (1998): Market efficiency, long-term returns, and behavioral finance, *Journal of Financial Economics* 49, 283–306.

Fama, E. F. and French, K. R. (1993): Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, E. F. and French, K. R. (2012): Size, value, and momentum in international stock returns, *Journal of Financial Economics* 105, 457–472.

Fama, E. F. and French, K. R. (2015): A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, E. F. and French, K. R. (2017): International tests of a five-factor asset pricing model, *Journal of Financial Economics* 123, 441–463.

Fama, E. F. and French, K. R. (2018): Choosing factors, *Journal of Financial Economics* 128, 234–252.

Goyal, A. and Wahal, S. (2015): Is momentum an echo?. *Journal of Financial and Quantitative Analysis* 50, 1237–1267.

Goyal, A., Welch, I., and Zafirov, A. (2021): A comprehensive look at the empirical performance of equity premium prediction II. Available at SSRN 3929119.

Gray, P. and Huynh, T. (2021): Treasury rates no longer predict returns: a reappraisal of Breen, Glosten and Jagannathan (1989), *Critical Finance Review* 10, 429–444.

Green, J., Hand, J. R., and Zhang, X. F. (2017): The characteristics that provide independent information about average US monthly stock returns, *Review of Financial Studies* 30, 4389–4436.

Gu, S., Kelly, B., and Xiu, D. (2020): Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Han, Y., He, A., Rapach, D., and Zhou, G. (2023): Cross-sectional expected returns: New Fama–MacBeth regressions in the era of machine learning. Available at SSRN 3185335.

Hanauer, M. X. (2020): A comparison of global factor models. Available at SSRN 3546295.

Harvey, C. R., Liu, Y., and Zhu, H. (2016): … and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1998): Tests for forecast encompassing, *Journal of Business and Economic Statistics* 16, 254–259.

Hjalmarsson, E. and Kiss, T. (2021): Dividend growth does not help predict returns compared to likelihood-based tests: an anatomy of the dog, *Critical Finance Review* 10, 445–464.

Hoerl, A. E. and Kennard, R. W. (1970): Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* 12, 55–67.

Hollstein, F. (2022): Local, regional, or global asset pricing?, *Journal of Financial and Quantitative Analysis* 57, 291–320.

Hollstein, F., Prokopczuk, M., Tharann, B., and Wese Simen, C. (2020): Predicting the equity premium around the globe: comprehensive evidence from a large sample. Available at SSRN 3567622.

Hou, K., Xue, C., and Zhang, L. (2020): Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Hurvich, C. M. and Tsai, C. L. (1989): Regression and time series model selection in small samples, *Biometrika* 76, 297–307.

Jacobs, H. and Müller, S. (2020): Anomalies across the globe: once public, no longer existent?, *Journal of Financial Economics* 135, 213–230.

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2023): Is there a replication crisis in finance?, *Journal of Finance*. https://doi.org/10.1111/jofi.13249.

Kelly, B. and Pruitt, S. (2013): Market expectations in the cross-section of present values, *Journal of Finance* 68, 1721–1756.

Kelly, B. and Pruitt, S. (2015): The three-pass regression filter: a new approach to forecasting using many predictors, *Journal of Econometrics* 186, 294–316.

Ledoit, O. and Wolf, M. (2008): Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Leippold, M., Wang, Q., and Zhou, W. (2022): Machine learning in the Chinese stock market, *Journal of Financial Economics* 145, 64–82.

Linnainmaa, J. T. and Roberts, M. R. (2018): The history of the cross-section of stock returns, *Review of Financial Studies* 31, 2606–2649.

Lo, A. W. and MacKinlay, A. C. (1990): Data–snooping biases in tests of financial asset pricing models, *Review of Financial Studies* 3, 431–467.

Löffler, G. (2022): Equity premium forecasts tend to perform worse against a buy-and-hold benchmark, *Critical Finance Review* 11, 65–77.

McLean, R. D., Pontiff, J., and Watanabe, A. (2009): Share issuance and cross-sectional returns: international evidence, *Journal of Financial Economics* 94, 1–17.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., … and Weitzel, U. (2023): Non-standard errors, *Journal of Finance*. http://dx.doi.org/10.2139/ssrn.3961574.

Neely, C. J., Rapach, D. E., Tu, J., and Zhou, G. (2014): Forecasting the equity risk premium: the role of technical indicators, *Management Science* 60, 1772–1791.

Rapach, D. E. and Zhou, G. (2020): Time-series and cross-sectional stock return forecasting: New machine learning methods, in: *Machine Learning for Asset Management: New Developments and Financial Applications*, edited by Emmanuel Jurczenko, pp. 1–33. WIley.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010): Out-of-sample equity premium prediction: combination forecasts and links to the real economy, *Review of Financial Studies* 23, 821–862.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2013): International stock return predictability: what is the role of the United States?, *Journal of Finance* 68, 1633–1662.

Rapach, D. E., Wohar, M. E., and Rangvid, J. (2005): Macro variables and international stock return predictability, *International Journal of Forecasting* 21, 137–166.

Romano, J. P. and Wolf, M. (2016): Efficient computation of adjusted *p*-values for resampling-based stepdown multiple testing, *Statistics and Probability Letters* 113, 38–40.

Schwert, G. W. (2003): Anomalies and market efficiency, in: *Handbook of the Economics of Finance, Vol.* 1, pp. 939–974.

Soebhag, A., van Vliet, B., and Verwijmeren, P. (2023): Non-standard errors in asset pricing: Mind your sorts. Available at SSRN 4136672.

Walter, D., Weber, R., and Weiss, P. (2022): Non-standard errors in portfolio sorts. Available at SSRN 4164117.

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

Windmüller, S. (2022): Firm characteristics and global stock returns: a conditional asset pricing model, *Review of Asset Pricing Studies* 12, 447–499.

Xue, C. (2022): Release notes for July 2022 update. Available at https://global-q.org/uploads/1/2/2/6/122679606/release_notes_for_july_2022_portfolio_data_update.pdf

Zou, H. and Hastie, T. (2005): Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320.