Machine Learning as Arbitrage:

Can Economics Help Explain AI?

Huahao Lu[*], Matthew Spiegel[†], Hong Zhang[‡]

This Draft: September 9, 2024

Abstract

Machine learning algorithms have shown to be remarkably successful tools for predicting asset returns. However, the underlying economic mechanisms behind their performance remain unclear. This paper proposes a model-based dynamic arbitrage trading strategy that combines economic and statistical nonstationarity to demystify this black box. In predicting stock returns based on 153 firm characteristics (anomalies), our strategy ranks anomalies similarly to neural networks in the cross-section. Overall, it accounts for approximately 87.9 bps monthly alphas of the high-minus-low portfolios selected by neural networks in the time series. When unpublished anomalies and microcap stocks are excluded from trading, this strategy can fully explain the performance of neural networks. Our results reveal three economic sources of neural-network performance: a time varying strategy analogous to dynamic arbitrage, a tendency to weight portfolios on unpublished anomalies, and exposure to microcaps.

Keywords: Machine Learning, Dynamic Trading, Anomalies, Interpretable AI
JEL Codes: C38, C45, G10.

[*] PBC School of Finance, Tsinghua University, Email: luhh.21@pbcsf.tsinghua.edu.cn
[†] School of Management, Yale University, P.O. Box 208200 New Haven, CT 06520-8200, Email: matthew.spiegel@yale.edu.
[‡] Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899; E-mail: hongzhang@smu.edu.sg.

Financial applications of machine learning algorithms have been remarkably rich. The resulting insights have ranged from predicting asset returns to deciphering the performance of institutional investors (see, e.g., Karolyi and Van Nieuwerburgh 2020, and Kelly and Xiu 2023 for recent reviews). However, the economic mechanisms through which these algorithms achieve their performance remain relatively obscure.[4] Our paper aims to shed light on this issue by employing an economically motivated algorithm that produces dynamic portfolios. The results substantially mimic the portfolios and performance generated by machine learning algorithms, helping to shed light on these black box algorithms.

To understand our strategy, recall that in the literature on arbitrage (Ingersoll (1984), and Da, Nagel, and Xiu (2023)) and informed trading (Kyle (1985), and Admati (1985)), a rational investor will optimally set portfolio weights in proportion to each asset's alpha forecast.[5] We extend the same intuition to anomalies as trading signals and, accordingly, set the portfolio weights in proportion to their estimated alpha.[6] We refer to these time-varying portfolios as *Dynamic Arbitrage Portfolios* or DAPs.

Algorithms choose among anomaly strategies by estimating their alphas, likely once per period. Under these conditions estimation error becomes a significant problem. One solution is the back-testing methodology of Mamaysky, Spiegel, and Zhang (2007, 2008). That technique sets the weight of a strategy to zero, if its forecast performance (e.g., estimate based on a rolling period) has a sign opposite that of its most recent one. Mathematically, back-testing is equivalent to a *ReLU* activation function widely used in machine learning tools, such as neural networks.[7] While simple, this filtering rule is among the most robust predictors of mutual fund performance in the literature (e.g., Jones and Mo, 2021) and can be applied to the current setup. Even absent estimation error, dynamic changes to the economy can produce anomalies that generate abnormal returns in the

[4] This is not unique to the finance; interpreting the predictions and underlying mechanisms used by machine learning models has confronted people across many fields. See Li et al. (2022) for a recent survey in the computer science literature.

[5] Applying this insight to mutual funds, Mamaysky, Spiegel, and Zhang (2007; 2008) show that informed mutual fund managers deliver abnormal returns that are quadratic to their private signals, where the signals can be estimated using latent-state techniques such as the Kalman filter.

[6] As detailed in subsequent sections, we estimate the conditional alpha of an anomaly from a rolling window as the High-minus-Low return spread that can be generated by sorted decile portfolios based on the corresponding firm characteristics associated with the anomaly.

[7] The ReLU function of a variable equals to the value of the variable—if it is greater than zero—and zero otherwise. The difference between DAP and machine learning is that the DAP activation function simply screens false positive measurements of economically optimal signals, whereas the *ReLU* activation function enters the optimization process for neural networks.

historical data, but not going forward. (Smith and Timmermann (2021a, 2021b)). Backtesting helps to deal with this case too.

Why does backtesting help? Intuitively, when constructing a portfolio, the important alpha estimates are those in the tails. If an anomaly trade is estimated to have a near zero alpha, it does not matter too much if it is or is not a true positive. So long as the portfolio weights reflect the alpha estimate's magnitude, these anomalies will receive little or no weight. For those in the tails, the situation is very different. These will drive a substantial fraction of the portfolio's return. To make the problem concrete, suppose an anomaly based trade is estimated to have a monthly return of +2% based on a 60-month rolling window (which is not unusual in the literature). If it then generates a -1% return in the latest month, one should be wary about trusting the estimated alpha going forward. The back-testing methodology used here addresses this issue by setting the anomaly's weight to zero. In a sense, the anomaly trade is put into a "penalty box" until it once again produces a return matching its predicted sign.

The DAP strategy with back testing incorporates two economically intuitive elements. First, weights relate to expected alphas, per theoretical models of optimal trading. As a result, DAP portfolio returns become a quadratic function in expected alphas. Second, back testing helps to down weight strategies with alphas that may be due to false-positive signals or no longer valid due to regime shifts. Considering the pivotal role nonlinearity and nonstationarity play in machine learning algorithms, our strategy can potentially provide insights into the economic underpinnings underlying the success seen by machine learning models.

We construct four empirical DAP portfolios to analyze the predictive power of machine learning algorithms. The first two, labeled *Base DAP* and *Backtested DAP*, use the conditional alphas of anomalies as trading signals, either with or without back testing. The third and the fourth, labeled *DAP_N* and *Backtested DAP_N*, further normalize anomaly alphas by their standard deviation, essentially using the information ratio of conditional alphas as trading signals. On the machine learning side, our empirical analysis focuses on *Feedforward Neural Networks* (*NNs*) as our main machine learning candidates due to their advantage of reliably estimating a complex functional relationship among a large set of predictors. Although various machine learning techniques—such as LASSO, Ridge, and Random Forest—can also be employed in financial markets, neural networks are

distinguished by their superior performance in predicting asset returns, as evidenced by Gu, Kelly, and Xiu (2020).

We apply both DAP and NN strategies to the same set of 153 anomalies in US-listed stocks. Our main sample excludes the smallest 20% of all stocks to avoid known biases these stocks impose on machine learning algorithms (e.g., Avramov, Cheng, and Metzker, 2023; Cong et al., 2021). (For the sake of parsimony, this group of stocks is simply referred to as microcap stocks from here on in.) Later sections present results when microcaps are included in the investable universe, to help interpret the economic source of machine learning's performance. Starting from January 1987, we use available information—including stock returns and the set of 153 characteristics—of US stocks within a rolling window, ranging from the prior 60 to 240 months, to train the two types of models. Monthly estimates are then sorted into ten deciles, based on each model's current forecasts.

Using a 60-month training period, without microcaps, the *Backtested DAP* and *Backtested DAP_N* portfolios generate large monthly H-L returns risk adjusted by the "Fama-French seven-factor model" of 1.823% and 1.658% per month respectively (or annualized values of 24.2% and 21.8%).[8] Among the neural networks, the version with one hidden layer (denoted NN1) generates the highest monthly 7-factor-adjusted return: 1.072%. If one includes the *DAP_N* or *Backtested DAP_N* returns as an additional factor, the NN1 alpha becomes insignificant. For instance, the inclusion of *Backtested DAP_N*, leads to an insignificant NN1 alpha of 0.136% per month. Finally, the explanatory power of DAP remains, at a similar magnitude, when microcap stocks are included, or when other neural network algorithms are employed (e.g., NN3, which are neural networks with three hidden layers). Overall, we observe that the inclusion of DAP accounts for the majority of the economic magnitude of the neural network alpha and also absorbs most of its statistical significance.

A caveat to the above tests is that the training period spans 60 months. Although this rolling window is standard in traditional asset pricing studies, it may result in a suboptimal sample size for neural network algorithms. To address this limitation, we systematically expand the rolling window from 90 to 240 months, with the last training window being

[8] The 7-factor model expands the Fama-French five-factor model (Fama and French 2015) with momentum (Carhart 1997) and short-term reversal (Jegadeesh 1990). Compared to the average seven-factor alpha of anomalies (0.887% per month or 11.2% when annualized) and that for DAP and DAP_N portfolios without back testing (approximately 1.15% per month or 14.5% when annualized), we can see that both the economic and statistical dynamics of DAPs help improve the performance of an equally-weighted anomaly portfolio. The statistical dynamics introduced by back testing seems particularly powerful in achieving this goal.

consistent with that used in Gu, Kelly, and Xiu (2020). This extension enables us to understand how data size affects machine learning performance and how the explanatory power of DAPs on machine learning varies accordingly. Figure 1 illustrates the results.

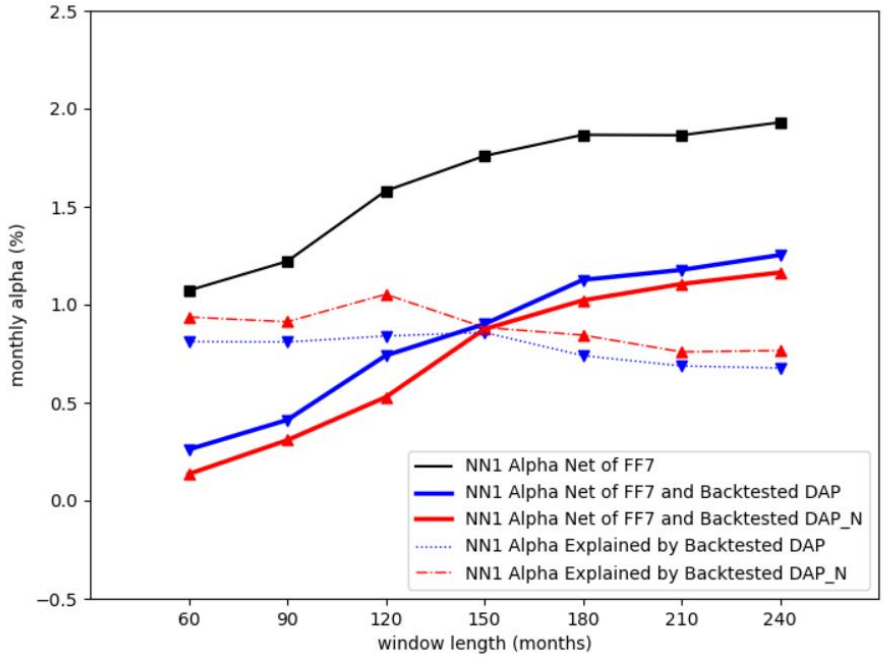


Figure 1: Neural Network Returns: Net of FF7 and Net of FF7 with DAP

The top line plots the NN1 alphas net of the seven factors. The bottom two solid lines plot the NN1 alphas net of both the seven factors and the two DAP strategies. The gap between the top and the bottom two lines, depicted as dotted lines, indicates the explanatory power of the two DAP strategies. While longer training periods benefit the NN1's alpha generation, it does not dramatically change the absolute amount explained by the DAP. As one can see in the relatively level values reflected in the dotted lines.

Focusing on the *Backtested DAP_N* model in Figure 1, the dotted red line averages 87.9, with a peak of 103 bps and a minimum of 75 bps. This stability suggests that the training period length likely alters the universe of factors neural networks focus on, one element it always includes is the economic mechanisms that generate the DAP.

In addition to the above time-series analysis, we also perform cross-sectional tests to compare the relative importance of various factors. If neural networks and DAPs focus on similar factors, that provides additional evidence that their portfolio construction reflects analogous economic principles. We examine this question with a variable gradient analysis (as per Sadhwani et al. 2020; Horel and Giesecke 2020) to estimate each factor's gradient

in the neural network's portfolio's construction. We then rank the factors by the absolute value of their gradients over the testing period to determine their relative importance (labeled *NN Importance Rank*). For DAPs, we rank their relative importance based on the average investment weights of anomalies in DAP portfolios (labeled *DAP Weight Rank*).

We find that neural networks and DAPs load on similar anomalies. For example, when explaining the gradient ranks of NN1 using the weight ranks from *Back tested_DAP_N* over training periods of 60 months (240 months), we observe a highly significant coefficient of 0.939 (0.723) and a univariate R-squared value of 0.882 (0.523). Using DAP anomaly weights to explain the raw NN1 anomaly gradients yields a similar conclusion. This cross-sectional observation further substantiates the explanatory power of DAPs in the context of machine learning algorithms.

Our final set of tests considers McLean and Pontiff's (2016) finding that an anomaly's alpha declines post publication. Do neural networks derive their performance from unpublished anomalies? Can DAPs still account for the performance of neural networks after accounting for anomaly publication? We provide two tests to empirically address these questions. In the first test, we include a Pre-publication Ratio—defined as the fraction of anomalies before publication at any given time—as an additional factor to account for the potential influence of publications. We find that the inclusion of this ratio reduces the seven-factor alpha for NN1 (with a 240-month training window) from 1.930% to 0.786% when microcap stocks are excluded, indicating that neural network algorithms may benefit from unpublished anomalies. A further inclusion of the *Backtested DAP_N* completely absorbs the significance of neural network alphas. In contrast, when microcap stocks are included, neural networks maintain a significant alpha, even after adjusting for the seven factors, the Pre-publication Ratio, and *Backtested DAP_N*.

The second test limits both DAPs and neural networks to trades using published anomalies. Returns from the NN1 algorithm generate a statistically significant seven-factor alpha of 0.514% when microcap stocks are excluded. However, the inclusion of the *Backtested DAP_N* eliminates that statistical significance. As seen in earlier tests, including microcap stocks enables neural networks to continue to generate significant alpha, even when one of the DAPs is included.

Taken together, the two tests suggest three economic mechanisms underpin the performance of neural networks. First, neural networks deliver alpha through a nonlinear

strategy similar to our economic model-based DAP. Second, to the extent that publication impacts neural networks, their performance also stems from an aptitude for data mining that identifies opportunities associated with unpublished anomalies.[9] Thirdly, neural networks are susceptible to the spurious returns generated by microcap stocks when they are included in the analysis. While the last finding aligns with existing studies (e.g., Avramov, Cheng, and Metzker 2023; Cong et al., 2021), the first two mechanisms represent novel contributions to the literature.

The remaining article is organized as follows. Section 1 reviews the literature. Section II describes the data and machine learning models. Section III provides baseline tests for predicting retail investors' returns. Section IV examines the importance of behavioral heuristics and firm characteristics. Section V provides additional tests and robustness checks, followed by a short conclusion with policy implications.

I. Literature Review

Our paper is related to several strands of the literature. Recent studies show that machine-learning algorithms can help predict asset prices in different sectors of the markets.[10] The challenge is deciphering the underlying mechanisms. In the absence of "interpretable AI," these algorithms may appear as "black boxes" that merely exploit nonlinear and intertemporal combinations of signals. A few recent studies have begun tackling this problem by examining the advantages conferred by machine learning's highly parameterized nature (e.g., Kelly, Malamud, Zhou 2024; Didisheim, Ke, Kelly, and Malamud, 2024) and its ability to process weak signals (e.g., Shen and Xiu 2024). Our research extends their work by posing an inverse query: Can established economic principles elucidate the performance of machine-learning models in generating returns? We

[9] Assuming one can include in the neural network's set of possible strategies the unpublished anomalies.
[10] For instance, Freyberger, Neuhierl, and Weber (2020), Gu, Kelly, and Xiu (2020), Bryzgalova, Pelger, and Zhu (2020), Jensen, et al. (2022), and Chen, Pelger, and Zhu (2023) apply machine learning algorithms to predicting stock returns and characteristics. Also see Leippold, Wang, and Zhou (2022) for applications in the Chinese equity market, Li et al (2023) for the global supply chain, Bianchi, Büchner, and Tamoni (2021) for bond risk premium, Easley, López de Prado, O'Hara, and Zhang (2021) for market microstructure, Filippou et al. (2022) for currencies, and Bali, Beckmeyer, Mörke, and Weigert (2023) for option pricing, and Van Binsbergen, Han, and Lopez-Lira, A. (2023) for earnings expectations. Li and Rossi (2020), DeMiguel, Gil-Bazo, Nogales, and Santos (2023), Kaniel, Lin, Pelger, and Van Nieuwerburgh (2023) use machine learning technology to predict mutual fund performance. Wu, Chen, Yang, and Tindall (2021) apply similar technology to hedge funds. Karolyi and Van Nieuwerburgh (2020) and Kelly and Xiu (2023) provide recent reviews.

present evidence that portfolios based on economic concepts can contribute to the quest for interpretable AI.

Our dynamic strategy also underscores the necessity of filtering out false-positive signals that may arise from time-varying risk premiums and regime shifts. In the literature, Smith and Timmermann (2021a, b) identify structural breaks in risk premia, observing a decline over time in the associated risk premia of the four factors. Brandt, Santa-Clara, and Valkanov (2009) examine the evolving optimal weights of a four-factor model for a utility-maximizing investor. Kelly, Pruitt, and Su (2019) introduce time-variant factor loadings as influenced by firm characteristics. Our research diverges by emulating a strategy that informed fund managers might employ (e.g., Mamaysky, Spiegel, and Zhang 2007, 2008; Jones and Mo, 2021). Furthermore, we demonstrate that machine learning algorithms may dynamically incorporate firm characteristics in ways similar to our optimal trading strategy. Hence, the economic and statistical nonlinearity introduced by our strategy provides some initial economic basis for interpreting the performance of machine learning algorithms, thereby advancing the understanding of interpretable AI.

Our results are also related to the anomaly literature (see, e.g., McLean and Pontiff 2016; Harvey, Liu, and Zhu 2016; Hou, Xue, and Zhang 2020; Jensen, Kelly, and Pedersen 2023 for recent evidence), which presents a "multidimensional challenge" to empirical asset pricing studies (Cochrane 2011). Vast studies explore the economic sources of anomalies, such as mispricing, data mining, and systematic risk, and the associated concern of publication bias in reported anomalies (e.g., McLean and Pontiff 2016; Harvey, Liu, and Zhu 2016). Different from these studies, we utilize anomalies as the testing ground to understand machine learning mechanisms. Nonetheless, our results also illuminate in the current setting some of the issues brought up in the anomaly studies. For instance, the publication of anomalies seems to profoundly influence machine learning returns, consistent with mispricing and data mining as the potential economic source of anomalies.

Recently Da, Nagel, and Xiu (2023) examine the limits imposed on arbitragers when faced with a plethora of weak signals. For individual firm data grouped into industries, they show that a machine learning model can come close to generating the optimal feasible Sharpe ratio. Similar to ours, their approach builds on the arbitrage pricing theory (APT) developed by Ross (1976) and the extension of Huberman (1982), Chamberlain and Rothschild (1983), and, particularly, Ingersoll (1984), and sets the portfolio weights proportionally to the estimated alphas. They compare a strategy they develop with those

using ML. However, here the problem we examine is almost the opposite, a setting with a plethora of strong rather than weak signals. We are also interested in determining the degree to which an economically intuitive model mimics the portfolio and returns produced by a machine-learning black box.

II. The analytical framework

This section lays out the framework of our analysis. It first introduces the conceptual framework for an economically motivated dynamic strategy. Then it discusses its estimation and compares it to neural networks.

A. The DAP Framework

To see how dynamics can help filter among trading anomalies, consider a world with three of them labeled A, B, and C as well as two states X and Y. Each anomaly has its independent economic source to generate above-market returns, which we refer to as alphas when there is no confusion. Anomaly A generates an alpha of 3% per year in both states. Anomaly B generates a 6% alpha during state X and 0% during state Y. Anomaly C does the reverse; it generates an alpha of 6% during Y and 0% during X. The state is a Markov process in which, on average, the state remains unchanged from period to period. Unconditionally, however, each state is equally likely.

The traditional way of trading a basket of anomalies would be to hold a diversified anomaly portfolio that equally invests in each of them or is value-weighted by stock market cap. The expected returns would be 3% in our example. However, are these traditional portfolios optimal if we expect the investment-period state to be X? The answer is no. Intuitively, we may not want to invest in anomaly C, as it generates zero return during state X. This intuition implies a trading rule that the investment weight should be related to the conditional alpha generated during the state. For instance, if we adopt a linear rule (i.e., setting the investment weight in proportion to the conditional alphas), we will invest 1/3 into A and 2/3 into B, leading to an expected return of 5% during state X. It is easy to see, this dynamic strategy outperforms traditional methods.

Two important observations emerge from the example proposed above. First, the example can be loosely mapped into Admati's (1985) model of trading with asymmetric information, or the arbitrage models of Ingersoll (1984), and Da, Nagel, and Xiu (2023). As

we will see from the generalized case detailed in the Online Appendix, the optimal trading rule uses weights that are proportional to the value of the trading signal, which is the conditional alpha in our example. Second, a dynamic strategy's success hinges on quickly detecting newly dysfunctional anomalies. In our example, if the state has shifted to Y, continuing to invest in B will yield very poor returns. Below, we discuss the intuition of the generalized trading strategy and a technique to quickly discard anomalies that have an elevated probability of producing negligible returns going forward.

Intuitively, one can view anomalies as assets that generate persistent yet heterogeneous conditional alphas. Assume that the k^{th} anomaly strategy, which refers to a high-minus-low portfolio of the k^{th} characteristic of firms ($k \in [1, K]$), delivers a conditional alpha in period t according to the following conditional CAPM framework,

$$y_{kt} = \alpha_{kt} + \beta_{kt} \times X_t + \epsilon_{kt};$$

where y_{kt} refers to the return of the k^{th} anomaly in period t. When there is no confusion, all returns are in excess of the risk-free rate. In the first line of the equation, α_{kt} is its conditional alpha, X_t refers to a common factor to affect asset returns (e.g., the market return), β_{kt} refers to market exposure, and ϵ_{kt} is an i.i.d. normally distributed noise term (i.e., $\epsilon_{kt} \sim N(0, \sigma_k^2)$). In this framework, the informativeness of an anomaly lies in its conditional risk premium, which persists over time but can be shocked by changes in the economy.[11]

Based on the above assumption, the optimal trading rule following the multiasset framework of Admati (1985), Ingersoll (1984), and Da, Nagel, and Xiu (2023) is to invest capital in proportion to each anomaly's expected conditional alpha. Our Online Appendix further confirms that this optimal rule applies to traders who observe state-contingent anomaly alphas. This rule allows us to write the investment weight as $c_t \times \alpha_{kt-1}$, where c_t is a normalization constant so that the portfolio weights add to 1 and α_{kt-1} is the anomaly portfolio's alpha as of the portfolio's construction date $t-1$. We can write the return of the dynamic arbitrage portfolio—or DAPs—in period t (y_{Pt}) as the sum of the underlying stock returns:

[11] A more general assumption about anomaly returns may take into consideration of the correlation structure across anomalies. However, the literature (e.g., Bai and Ng 2002; Pesaran 2006; Smith and Timmermann 2021a) suggests that one can use a common factor to absorb the cross-sectional correlations. We follow this literature to assume that anomaly returns are affected by the common factor and leave the analysis of a more general cross-anomaly correlation to future research.

$$y_{Pt} = \sum_k \underbrace{c_t \alpha_{kt-1}}_{\substack{DAP\ weight \\ in\ Anomalies}} y_{kt} = \sum_k c_t \alpha_{kt-1} \underbrace{\left(\sum_i w_{i \in k,t} \times r_{it} \right)}_{\substack{Anomaly\ Returns \\ by\ Stocks}} = \sum_i \underbrace{\left(c_t \sum_k \alpha_{kt-1} w_{i \in k,t} \right)}_{\substack{DAP\ weight \\ in\ Stocks}} r_{it}. \quad (1)$$

In the second step, we have used $y_{kt} = \sum_i w_{i \in k,t} \times r_{it}$ as the return generated by anomaly k, where $w_{i \in k,t}$ is the investment weight of the stock in constructing the k^{th} anomaly portfolio—i.e., the exposure of stock i to the k^{th} characteristic.[12] In the last step, we rearrange terms to build the DAP portfolio as a stock portfolio, in which the investment weight of stock i becomes the conditional alpha weighted average of its original investment weight in each anomaly strategy: i.e., $y_{Pt} = \sum_i w_{it,DAP} r_{it}$, where $w_{it,DAP} = c_t \sum_k \alpha_{kt-1} w_{i \in k,t}$. Economically, we can interpret $w_{it,DAP}$ as the conditional alpha-weighted characteristic exposure of a stock.

B. Nonlinearity and Back-testing

We first notice that Equation (1) implies that the DAP portfolios will fail to generate future excess returns, if the conditional alpha does not remain elevated. Economically speaking, the excess returns of a particular anomaly must persist for it to be useful. However, if the alpha is persistent, y_{kt} becomes a linear function of α_{kt-1} and $y_{Pt} = \sum_k c_t \alpha_{kt-1} y_{kt}$ becomes a quadratic function in α_{kt-1}. Hence, DAP portfolio returns are quadratic in alpha, so long as the alpha remains time invariant. This shows how important it is to verify that large, estimated alphas will persist. Squaring means accurate but small alphas are relatively unimportant. But squaring also means large alphas are vastly more influential than their absolute magnitude might initially imply.

What if the conditional alpha is not time invariant? Perhaps it has gone to zero or was simply misestimated and was never positive to begin with. Our initial example reveals a general economic ground for a return-predictive process to generate false-positive signals: regime shifts across different states change a signal's predictive power. Such structural breaks are reported in the literature (e.g., Smith and Timmermann 2021a, b), and can affect a factor's validity and risk premium. Even absent regime changes, false positives will almost certainly arise given the large number of anomaly alphas that need to be estimated each month.

[12] Empirically, stocks in the top (bottom) decile are assigned the weight of one (negative one).

To identify false positives and discard them from trading, we adopt the back-testing methodology of Mamaysky, Spiegel, and Zhang (2007, 2008) to screen out potentially erroneous signals. More explicitly, to construct a DAP in month t, we not only measure α_{kt-1} in the rolling training period as mentioned above. We also estimate the most recent realization of the conditional alpha—i.e., the high-minus-low portfolio return—from the last month of the training period (using the month after the training period yields similar results). If the most recent realization carries the same sign as the rolling-window estimated α_{kt-1}, then the latter estimation is valid to serve as a signal when constructing the DAP in month *t*. Otherwise, the anomaly's weight is set to zero.

By quickly detecting and excluding signals that have likely become false positives due to changes in the economy or arose from estimation error, this methodology is especially suitable for selecting among estimated signals in a wide cross section of returns. Empirically, the literature shows that back testing helps validate the informativeness of institutional investors (e.g., Kacperczyk, Sialm, Zheng 2008; Jones and Mo, 2021) and the usefulness of asset pricing factors (e.g., Boguth and Simutin 2018 and Dong et al. 2021). Indeed, back testing is among the most reliable empirical methods to predict fund returns when market conditions change over time (Jones and Mo, 2021). The application of back testing in our current setup synchronizes these influences.[13]

Conceptually, our back-testing method essentially introduces a ReLU activation function into DAPs. Indeed, we can rewrite the back-testing process for the rolling alpha α_{kt-1} as a *ReLU* activation function. More specifically, back tested alpha can be written as

$$ReLU\left(\alpha_{kt-1} \times H\left(\alpha_{k,-t}\right)\right),$$

where $\alpha_{k,-t}$ is the most recent realization of alpha and $H(x)$ denotes the sign function.[14] Our Online Appendix shows that the more general investment policy of back tested DAPs follows a *ReLU* activation function similar to the activation function typically used by neural networks.

[13] Our Online Appendix further illustrates the superior statistical power of back testing in predicting anomaly returns. For instance, the back testing process can select performance-delivering anomalies with a high success rate of 57.37%. This predictive power allows the back tested DAP to deliver highly significant performance when compared to the simulated distribution of DAP strategies that randomly drop a similar fraction of anomalies as the real back testing process. The simulated p-value for both the seven-factor alpha and t-statistics are 0.00.

[14] $H(x) = 1,\ if\ x \geq 0$, and $-1,\ otherwise$.

Note that the DAP activation function primarily screens economically large signals for false positives. (Small alpha anomalies receive a near-zero weight whether they pass the recent return screen or not.) To the degree that the resulting DAP portfolios mimic those of neural networks, this provides a way to interpret how a neural network selects among anomalies over even short periods of time.

C. The Empirical Process of Constructing DAP Portfolios

Of course, Equation (1) cannot be directly employed since the α_{kt-1} are unknown and must be estimated. To map $w_{it,DAP} = \sum_k c_t \alpha_{kt-1} w_{i\in k,t}$ into the empirical high-minus-low specification, we let $w_{i\in k,t}$ represent the investment rank of stock i when all stocks are sorted based on the k^{th}-characteristic. The rank is normalized within each anomaly to be disturbed between 0 and 1. The corresponding anomaly strategy buys (shorts) stocks within top-decile (bottom) ranks.

We measure the DAP trading signals via two methods. The first follows the literature's convention by computing the signal as an anomaly's conditional alpha, which we empirically estimate as the average return spread generated by the top-minus-bottom portfolio during the rolling training period. The second method then normalizes the conditional alpha by its standard deviation estimated in the rolling window.[15] Economically speaking, it can be interpreted as the information ratio of the top-minus-bottom anomaly portfolio returns. Hence, it also provides a good signal for DAPs to trade anomalies. The two measures of trading signals are then combined with back testing to generate the following set of DAP trading strategies:

1) *Base DAP*: using conditional alphas as signals without back testing;
2) *Backtested DAP*: using conditional alphas as signals with back testing;
3) *DAP_N*: using normalized conditional alphas as signals without back testing;
4) *Backtested DAP_N*: using the conditional alpha as signals with back testing;

We drop anomalies where less than half of the necessary data in the historical window exists with which to estimate its alpha. Moreover, for any anomaly delivering negative

[15] Mathematically, $\alpha_{k,t-1} = \frac{\overline{\alpha}}{\frac{1}{T}\sqrt{a_1^2+\cdots+a_T^2}}$, where $\overline{\alpha}$ is the average returns generated by the top-minus-bottom portfolio, and $a_1^2, \ldots, a_T^2$ denote the squared return in any particular month during the training window, which consists of T months.

returns during the estimation period, we invert the ranks of its associated portfolios. This simply shorts the high-minus-low portfolio if an anomaly's estimated α leads to an expected negative risk premium. Hence, a high-minus-low spread is always associated with a positive risk premium in our empirical analysis. Finally, in calculating the above DAP signals, the anomaly top-decile (bottom) portfolio returns can be either value-weighted or equal-weighted. Equal-weighted signals are closer to the treatment of signals in neural networks and provide a better ground to explain the performance of neural networks. However, this treatment may load on the returns of microcap stocks that are difficult to obtain in practice, giving rise to inflated return estimations. Indeed, machine learning algorithms are known to be vulnerable to this issue (see, e.g., Avramov, Cheng, and Metzker 2023; Cong et al., 2021). Due to this tradeoff, we use equal-weighted signals for our main analysis and focus on the sample of firms that exclude the microcap stocks. As a robustness check, we also present results that include all stocks. These tests use value-weighted signals for the DAPs to reduce potentially spurious trading signals associated with microcap stocks (using equal-weighted signals for the whole sample will not change our conclusions).

Once we obtain $\alpha_{k,t-1}$, we use it to weight the rank of any particular stock across valid anomalies, which gives an ultimate $w_{it,DAP} = \sum_k c_t \alpha_{kt-1} w_{i\in k,t}$ for each stock as its DAP investment rank. We finally sort all stocks according to $w_{it,DAP}$ into ten deciles, with decile ten consisting of stocks that have the largest values of $w_{it,DAP}$. A value-weighted DAP then takes a long (short) position in the market value-weighted portfolio of decile ten (one) stocks, similar to the VW momentum case. (The portfolios are value-weighted to further mitigate the influence of small stocks. Values are based on period $t-1$ market caps.) All $w_{it,DAP}$ sorted decile portfolios are rebalanced every month, with $\alpha_{k,t-1}$ re-estimated based on the new rolling window.

The above process is close to Stambaugh and Yuan (2017) for aggregating characteristics. The difference is that we use the conditional premium to value-weight them. In other words, our approach nests Stambaugh and Yuan (2017) if all conditional premiums are the same. However, since different characteristics are likely to exhibit different conditional premiums, our approach essentially allows more informative anomalies to play a more important role than less informative ones.

We also estimate the stock-level trading signals from neural networks with firm characteristics as the most important input. Since the process is quite standard (see, e.g., Gu, Kelly, and Xiu 2020), the details are relegated to the Online Appendix. Based on these signals stocks are sorted into ten deciles. A value-weighted NN portfolio then takes a long (short) position in the market value-weighted portfolio of decile ten (one) stocks.

III. Machine Learning vs. DAPs (60-month Training Windows)

This section examines the performance of DAPs and neural networks using a 60-month rolling training period. It begins with a description of the data and then examines the full sample performance estimates.

A. Data and Empirical Methods in Constructing DAP and NN Portfolios

Our sample includes the common equity of U.S. firms trading on NYSE, Amex, or Nasdaq. We obtain the market information of these stocks from the Center for Research in Security Prices (CRSP) and accounting information from Compustat. We then follow Jensen, Kelly, and Pedersen (2023) and use a set of 153 anomalies. Our main out-of-sample return testing period spans from 1987 to 2022, a total of 432 months. That is, starting from January 1987, we train the two types of models based on the information during the rolling training window, which spans 60 months, to create ten value-weighted portfolios. We rebalance the portfolios monthly and hold them until December 2022. We then calculate the out-of-sample returns generated by these portfolios, as well as those for the high-minus-low (H-L) portfolios. The returns generated by these deciles and H-L portfolios are out-of-sample.

Risk-adjusted returns are estimated via the standard factor models. The literature typically uses the Fama-French three-factor model (Fama and French 1992, 1993), the Fama-French five-factor model (Fama and French 2015), momentum-enhanced Fama-French-Carhart four-factor model (Carhart 1997), and momentum-enhanced Fama-French six-factor model (Kelly, Pruitt, and Su 2019). To further account for adjustments that others have used when creating alpha estimates, we also consider the short-term reversal factor (e.g., Jegadeesh 1990; Feng, Giglio, and Xiu, 2020). Our main analysis thus adopts a momentum and short-term reversal enhanced Fama-French seven-factor model. These factors are obtained from Kenneth French's Data Library.

B. The Performance of DAPs and NNs

We start with the performance of the economically motivated DAP portfolios. Table 1 reports summary statistics on returns generated by averaging the 153 anomalies, and by constructing portfolios using either DAP, or a NN. In each panel, columns 1-6 report the returns when portfolios are constructed based on the main sample of stocks that exclude the microcap stocks. In particular, the first two columns report the deciles with the lowest and highest returns (Low and High). The third column reports the return spread generated by the H-L portfolio. Columns 4-6 report the risk-adjusted returns for the H-L portfolios based on the Fama-French three-factor, five-factor, and seven-factor models. As a robustness check, columns 7-12 tabulate the results when all stocks are included. The layouts are similar to that of the first six columns.

From columns 1-6, we observe that stocks in the DAP-selected top decile deliver much higher returns than those in the bottom DAP decile, allowing the DAP H-L portfolios to generate significant out-of-sample risk-adjusted returns. Compared to the average H-L returns that can be generated by the baseline anomalies, DAPs perform significantly better. Furthermore, among all the DAP H-L returns, back tested DAPs also outperform the corresponding DAPs without back-testing. The back tested DAP (labeled "*Backtested DAP*" in the panel) and back tested and normalized DAP (labeled "*Backtested DAP_N*"), for instance, can generate a monthly H-L return of as high as 1.888% and 1.689% (or 25.1% and 22.3% annualized returns), respectively.

Risk adjustments based on various factor models have little impact on the magnitude: the two models still deliver a monthly seven-factor alpha of 1.823% and 1.658%, respectively, or 24.2% and 21.8% when annualized. These performances are higher than those of non-back tested DAPs—i.e., a monthly seven-factor alpha of 1.145% and 1.150% for DAP and DAP_N without back test—which are still better than the typical performance of anomalies (e.g., a monthly alpha of 0.887). Given that microcap stocks are already excluded, the economic magnitudes of such back tested DAPs are comparable to the performance of some of the best machine-learning performances reported in the literature.

Columns 7-12 demonstrate that adding back microcap stocks can marginally improve the returns of back tested DAPs. The back tested DAP and Backtested DAP_N now generate a monthly H-L return of 1.893% and 2.048%, respectively. Risk adjustments based on Fama-French's seven factors lead to a monthly alpha of 1.803% and 2.061% (or 23.9% and 27.7% annualized). As such, including microcap stocks mainly improves the performance of the Backtested DAP_N. In other words, Backtested DAP_N might be

exposed to issues associated with small stocks, but back tested DAP itself is not vulnerable to this potential bias. This contrasts with the findings in several papers showing that many machine learning strategies load up on microcaps (see, e.g., Avramov, Cheng, and Metzker 2021; Cong et al., 2020). Above all, our results suggest that the economically motivated dynamic trading strategy in this paper can deliver very robust out-of-sample returns.

We now move on to report the performance of neural networks based on similar conditions. Panel C demonstrates that even with a training window as short as 60 months, neural networks can select portfolios with significant out-of-sample performance. For instance, when microcaps are removed, the H-L portfolio selected by the neural network algorithm with one hidden layer (labeled NN1) and three hidden layers (labeled NN3) generate a significant monthly seven-factor alpha of 1.072% and 0.678%, respectively.

There are two other interesting observations from this table. First, NN3 is vulnerable to microcaps. Removing microcaps reduces its seven-factor alpha from 0.926% to 0.678%. However, NN1 is less subject to this well-documented vulnerability. Indeed, NN1 outperforms NN3 (and NN2, the neural network with 2 hidden layers) in this test. Conceptually, the success of NN1 in our anomaly-pricing framework is consistent with Didisheim, Ke, Kelly, and Malamud's (2024) theoretical argument that factor investments can be mathematically mapped into a one-layer neural network. Empirically, the outperformance of NN1 suggests that over-parameterization may be an issue for neural network estimation. For these reasons, our later analysis will focus on NN1 as the representative candidate of machine learning, though we will also demonstrate the robustness of our results based on NN3.

Figure 2 plots the cumulative returns that can be generated by the H-L portfolio selected by the two back tested DAP strategies and NN1 over our main testing period. The most interesting observation from these plots is that DAP and NN1 returns are intuitively correlated. For instance, both types of strategies witnessed a surge in return around 2009, a drawdown in 2021, and a bounce toward the end of our testing period. This correlation suggests that NN1 may be exposed to the economic mechanisms of DAPs, allowing the latter to help explain its performance. The next section takes on this task to examine the explanatory power of DAPs on neural networks.

C. The Performance of NNs Explained by DAPs

Before we report the results, it is worth noting that a 60-month training period may not exhaust the predictive power of neural networks. But this step fits our goal of moving toward interpretable AI, since it helps show how machine learning performs across a wide range of conditions. If our economic model helps explain machine learning across these settings, then it likely captures important elements of the latter.

Table 2 reports the results when we use DAP returns as an additional factor to explain neural network returns on top of the Fama-French seven-factor model. Panels A and B tabulate the analysis for NN1 and NN3. Since the results are largely consistent, we mainly focus on Panel A to discuss our main findings.

The first five columns focus on the sample when the microcap stocks are excluded. The first column reports the seven-factor alpha as well as the risk exposure of the H-L portfolio selected by NN1 when we regress its portfolio returns on the Fama-French seven factors. The monthly seven-factor alpha itself, 1.072%, has also been reported in the previous table. Importantly, columns 2-5 include various DAP H-L portfolio returns as an additional factor to explain NN1 returns. We observe significant coefficients across all DAP returns, which also significantly reduce the magnitude and significance of NN1 alphas. For instance, if we further include *the Backtested DAP_N* as an additional factor, the NN1 alpha becomes an insignificant 0.136% per month.

The inclusion of *Backtested DAP_N* absorbs approximately 94 basis points (in absolute terms) of monthly alpha generated by NN1.[16] That is approximately 92% of the NN1's monthly H-L excess return. Models 6-10 confirm that DAP can still absorb the monthly alpha of NN1 when all stocks are used. Panel B also illustrates a similar explanatory power of DAP on NN3 performance.

If DAP returns can explain the NN's performance, can NN returns help explain the DAP's performance? If so, the two are empirically equivalent. To address this question, we conduct the reverse exercise of regressing back tested DAP and *Backtested DAP_N* returns (proxied by the H-L return generated by the respective model) on the returns of neural network-selected H-L portfolios as additional factors.

Table 3 reports the results, with Panels A and B tabulating the regression coefficients for back tested DAP and *Backtested DAP_N* returns, respectively. The first model of Panel

[16] The difference is calculated as annualized seven-factor alpha minus the annualized alpha netting out both seven factors and the Backtested DAP_N factor.

A again illustrates how we have calculated the seven-factor adjusted alpha for back tested DAP portfolios when microcap stocks are excluded. The monthly alpha of 1.823% has been reported in Table 1 (Panel B). Columns 2-4 include the portfolio returns of NN1, NN2, NN3 as an additional factor, and Column 5 tests the joint explanatory power of all three. We observe that NN models also help explain DAP returns. Adding the NN1 return as a factor, reduces the DAP H-L excess return from 1.823 to 1.399, which is a 23% reduction. Nevertheless, including NN models in addition to the original seven factors still leaves a statistically significant DAP return.

Collectively, therefore, our economically motivated strategy performs at least as well as the tested machine learning algorithms. Furthermore, the latter seems to load significantly on the former, suggesting that the economic mechanisms underlying DAPs could help elucidate neural network alphas.

IV. Machine Learning vs. DAPs with Extended Training Windows

A caveat of the above test is that the training period spans 60 months. Although this rolling window is standard in traditional asset pricing studies, neural network models might need a longer time series to realize their full potential. As mentioned above, we are interested in the explanatory power of DAPs in this case as it reveals important, albeit preliminary, information regarding the economic mechanisms of NN algorithms.

To explore the mechanisms going beyond this limit, we now systematically expand the rolling window to a spectrum of 90, 120, …, and 240 months, with the last training window consistent with that used in Gu, Kelly, and Xiu (2022). This extension enables us to understand how data size affects machine learning performance and the explanatory power of DAPs.

A. The Performance of DAPs and NNs with 240-month Rolling Training Window

Table 4 tabulates the performance of the DAP (in Panel A) and NN portfolios (in Panel B) when the rolling training window is extended to 240 months. The table layouts are like those used for Table 1. Columns 1 and 6 of Panel A report that, absent microcaps, the two back tested DAP monthly H-L returns (column (3)) increase to 2.133% and 2.010% (from the Table 1 values of 1.888% and 1.689% respectively). Although the return increase is substantial, the seven-factor alphas—i.e., 1.854% and 1.724%—are similar to those in

Table 1, suggesting that much of the return gains are related to risk exposure. This implies that a 60-month training horizon is already long enough for back tested DAPs to obtain stable and reliable trading signals in the absence of microcap stocks. Based on the similarity of the Table 1 and Table 4 DAP returns, the alphas are relatively similar whether the training windows is 60 or 240 months long.

The next two columns report that, when microcaps are added back, the extended rolling training window benefits back tested DAPs in both returns and alphas. *Backtested DAP* and *Backtested DAP_N*, for instance, generate a monthly H-L return of 2.268% and 2.178% (column (9)) on the whole sample (compared to the analogous Table 1 values of 1.893% and 2.048%). Even after removing the Fama-French seven factor model's return, the two back tested DAP portfolios deliver a monthly alpha of 2.056% and 2.031%, respectively (column (12)). Across this and the first table, it seems that the longer training horizon may benefit DAPs by increasing the quality of trading signals related to microcaps. Nevertheless, since microcaps are difficult to trade, it is not clear to what degree this can be exploited.

Panel B reveals a very different picture for neural networks. Focusing on the main sample excluding microcap stocks, we observe that neural network alphas increase in the length of the training period. For instance, the seven-factor alpha of NN1 increases from 1.072% (Table 1) to 1.930% (Table 4) when the training period is extended from 60 to 240 months. This substantial increase indicates that neural networks benefit from a longer training window. As discussed earlier, Figure 1 plots the NN1 seven-factor alphas across different lengths of training windows. It shows that neural network alphas increase monotonically in the length of the training period.

B. The Performance of NNs Explained by DAPs

Table 5 reports the results when we use DAP returns as an additional factor to explain neural network returns based on this extensive training window. The layout is like that of Table 2. We again focus on Panel A (NN1) to discuss our main findings—the results of NN3, as reported in Panel B are very similar.

The first column confirms that the H-L portfolio selected by NN1 generates a monthly seven-factor alpha of 1.930% (excluding microcap stocks)—and it also reports the exposure of NN1 returns to these factors. Columns 2-5 include various DAP H-L portfolio returns as an additional factor to explain NN1 returns. We observe significant coefficients across all

DAP returns, which also significantly reduce the magnitude of NN1 alphas. For instance, if we further include the *Backtested DAP_N* as an additional factor, the NN1 alpha equals 1.164% per month. Hence, the inclusion of *Backtested DAP_N* absorbs approximately 77 basis points (bps) of monthly alpha generated by NN1, a 40% reduction in its alpha.

Models 6-10 suggest that DAP continues to absorb a statistically significant fraction of monthly alphas of NN1 when all stocks are used. However, the fraction of the alpha explained by the DAP returns declines. The NN1 model goes from generating an H-L excess return of 2.499 to 1.853 when the *Backtested DAP_N* is added as a factor, a 26% reduction. Panel B also illustrates similar results of the DAP's relationship to the NN3 performance.

Table 6 reports the reverse test when back tested DAP (in Panel A) and *Backtested DAP_N* returns (in Panel B) are regressed against the NN portfolio returns. The first model in Panel A reports a monthly back tested DAP alpha of 1.854% when microcap stocks are excluded, which has been reported in Table 5 (Panel B). Columns 2-4 suggest that including the portfolio returns of NN1, NN2, or NN3 as an additional factor does not fully absorb the DAP alphas. Column 5 confirms that even including all three NN models in addition to the original seven factors fails to eliminate the *Backtested DAP*'s alpha. It remains at 91 bps, which comes to 11.5% when annualized.

C. The General Explanatory Power of DAPs Across Different Training Horizons

Thus far, our results show that an economically motivated strategy can explain a significant fraction of the neural network alphas for training windows of 60 and 240 months. What about training windows whose lengths lie in between?

As discussed in the introduction, Figure 1 provides the answer. The top solid line in black displays the seven-factor NN1 alphas across training windows of varying lengths. Across the same training windows, the bottom solid line in blue (red) displays the NN1 returns net of both the seven factors and *Backtested_DAP* (*Backtested DAP_N*). The difference between the top black line and those in blue and red indicates the degree to which the seven-factor NN1 strategy pulls information overlapped with the DAPs. The magnitude of this difference, illustrated as blue and red dotted lines in Figure 1, remains fairly stable across different training periods, ranging from a low of 76 bps to a high of

94.[17] This implies that part of the seven-factor NN1 that can be potentially explained by the DAP is similar for all training windows.

V. The Analysis of Publication

Our last set of tests aims to provide a more complete picture regarding the economic grounds behind the performance of neural networks by revisiting the issue of an anomaly's publication date. McLean and Pontiff (2016) document that the trading profitability of anomalies declines—though still remains significant—post publication. Can neural networks generate significant alphas post publications? Can DAPs explain much of the neural networks' alpha in this case? This section provides two tests addressing this question. To allow for the maximum power of neural networks, the tests employ a 240-month training window for this part of the analysis.

A. Pre-Publication Ratio as an Additional Factor

Table 7 conducts the first test by including a Pre-publication Ratio as an additional factor to capture the potential influence of publications. The ratio is defined as the fraction of anomalies before publication at any given time. Our Online Appendix (Figure IN1) plots the time series of this ratio.

Columns 1-4 tabulate the results when microcap stocks are excluded. The Pre-publication Ratio itself is positively associated with neural network returns, suggesting that neural network algorithms benefit from focusing on unpublished anomalies. We also observe that the inclusion of this ratio reduces the neural network seven-factor alpha of NN1 from 1.930% (column 1) to 0.786% (column 3). When the *Backtested DAP_N* is added to the list of controls, then the Pre-publication ratio reduces the neural network seven-factor alpha from 1.164% (column 1) to 0.296% (column 3), which becomes statistically insignificant.

Columns 5-8 suggest that when microcap stocks are included, neural networks can still produce a significant alpha even after adjusting for the influence of the seven factors, the Pre-publication Ratio, and *Backtested DAP_N*.

B. NN and DAP Returns Using Published Anomalies

[17] Our Online Appendix (Figure 1 Supplementary Table) provides more detailed information about NN1 alphas and the part of NN1 alphas explainable by DAPs.

The second test restricts trades to published anomalies for both the DAPs and neural networks. The results are tabulated in Table 8, Panel A for NN1 and Panel B for NN3. We again focus on NN1 to explain our main findings since NN3 portfolios yield similar conclusions.

Column 1 suggests that restricting neural networks to trading on published anomalies significantly reduces their performance. Although NN1 still delivers a significant seven-factor alpha, its magnitude drops to 0.514% when microcap stocks are excluded. Furthermore, columns 2-5 suggest that the inclusion of DAPs further reduces the NN1 alphas by a substantial amount. The NN1 alpha goes from statistically significant without the *Backtested DAP_N*, to insignificant with it, as seen in column 5. Columns 6-10 suggest that similarly to earlier tests, the inclusion of microcap stocks allows neural networks to generate significant alpha no matter what DAP return is included. Although in this case it is the *Backtested DAP* that yields the largest NN1 alpha reduction.

Taken together, the two tests suggest three economic mechanisms underpin the performance of natural networks. First, a time varying set of portfolio weights akin to those generated by the DAP algorithm enables neural networks to persistently deliver alpha. Second, restricting trades to unpublished anomalies hurts a neural network's performance. One can interpret this as an aptitude for identifying opportunities aligned with unpublished anomalies, assuming they can be included pre-publication into the list of potential strategies. Thirdly, neural networks are vulnerable to loading on the spurious returns generated by microcap stocks when allowed to use them as part of their portfolio's construction. While the last observation is consistent with existing studies (e.g., Avramov, Cheng, and Metzker (2021); Cong et al., (2020).), the first two mechanisms are novel in the literature.

VI. Additional Analyses and Robustness Checks

This section provides additional tests to check the robustness of the results and explores other avenues of potential interest.

A. The Cross-sectional Importance of Factors

While our main tests on the relationship between DAPs and neural networks focus on time-series analysis, we can also perform cross-sectional tests to compare the relative importance of various factors. The goal is to determine the degree to which neural networks and DAPs

consistently focus upon the same set of factors. If so, this provides additional evidence that they generate weights using similar economic principles.

To achieve this goal, this section begins by looking at the relative importance of individual anomalies in both NN1 and DAPs. For NN1, we first employ a variable gradient analysis (as per Sadhwani et al. 2020; Horel and Giesecke 2020) to estimate the gradient of each factor in the neural network analyses. We then rank the factors by the absolute value of their gradients in each cross-section. Finally, we calculate the overall ranks of anomalies—labeled *NN Importance Rank*, which ranges from 1 to 153, with a higher number indicating greater importance—based on their average gradient ranks during the testing period from 1987-2022. For DAPs, we first rank the relative importance of anomalies in each cross-section based on their investment weights in the DAP portfolios. We then calculate the overall ranks of anomalies, labeled *DAP Weight Rank*, based on their average weight ranks during the testing period. These tests all exclude the microcaps.

Panel A in Table 9 reports the results of a univariate cross-sectional regression, which uses the *DAP Weight Rank* to explain the *NN Importance Rank*. In every test the *DAP Weight Rank* is significant at the 1% level, and the *Adj R2* exceeds 0.50. For example, the row labeled *Backtested DAP_N* lists the coefficient estimates for the regression of the *NN Importance Rank* of NN1 using the *DAP Weight Rank* from *Backtested_DAP_N* . When the rolling training is 60 months (240 months), the model produces a significant coefficient of 0.939 (0.723) and a univariate R-squared value of 0.882 (0.523).

Figure 3 visualizes this relationship by plotting the *DAP Weight Rank* (in the x-axis) versus the *NN1 Importance Rank* for the 60-month training window. It is easy to see that the two variables lie around a 45-degree line. Panel B of Table 9 further uses the average DAP investment weights of anomalies to explain the average gradients of the anomalies employed in NN1. This analysis yields a similar conclusion, though the explanatory power is slightly smaller in magnitude.

Overall, the importance ranks of anomalies are remarkably consistent across neural networks and DAPs. This cross-sectional observation further substantiates the ability of DAPs to mimic the portfolios produced by machine learning algorithms.

B. Extended Performance Evaluation Period

Figure 4 extends the initial year of the performance-evaluation period from 1987 to 1977. Hence, the DAP and NN portfolios are constructed and rebalanced each month from 1977 to 2022. The explanatory power of DAP over this extended period is even more prominent and persistent, between 0.79 and 1.06 bps per month. The last two rows in Panel B, indicate that across training windows, the DAPs account for between 0.73 bps and 0.89 bps of the monthly return from the NN1, which itself varies from 0.95 bps to 1.79 bps per month.

C. Alternative DAP Estimations

Our main DAP estimates set the back-testing month as the last month of the training period. Our Online Appendix shows that allowing an independent back-testing month before the portfolio construction month and after the training period yields very similar results. In particular, Table IN1 reports that this alternative back-testing convention yields similar performance of DAP-sorted portfolios using 60 or 240 months of training horizon. Table IN2 suggests that the DAP returns have a similar explanatory power toward neural networks as Table 2. Finally, Figure IN2 plots the explanatory power of NN alphas as Figure 1. We again see that DAP returns again explain a consistent portion of the NN alphas, in absolute terms.

D. Post-Publication Tests: Explaining DAP Returns by NN Returns

Table 7 assesses the potential influence of publication on NN returns by including a Pre-publication Ratio as an additional factor. Table IN3 reverses the dependent and independent variables to produce the equivalent explaining DAP returns as a function of those produced by the NN algorithm. We observe that DAP alphas remain significant after including the Pre-publication Ratio.

Another test (Table 8) explains NN returns when after restricting both DAPs and neural networks to trade exclusively on published anomalies. Table IN4 conducts the opposite test of using NN to explain DAP returns in this setup. DAP alphas are not fully explained by NN and the seven factors in this setup.

E. Subsample Analysis with 50% Pre-Publication Ratio as the Cutoff

Tables IN5 and IN6 provide another test to illuminate the potential influence of publication. There, we conduct subsample analyses (of Tables 4 and 5) based on a 50% Pre-Publication Ratio as the cutoff year. We observe that the NN generated significantly better performance

in the first subperiod when more anomalies were unpublished. For instance, the monthly NN alpha after adjusting for the seven factors and *Backtested DAP_N* is 1.554%. However, in the second period, when there were more published than unpublished anomalies, the monthly alpha drops to 0.667%.

Looking at the *Backtested DAP_N* alphas (adjusted for the seven factors and all three NN returns), they also drop across the 50% publication threshold periods. In this case from 1.283% in the first to 0.704% in the second. However, the magnitude of the alpha's decrease is much smaller than that of the NN alpha. These tests suggest that DAPs are more robust to the potential influence of publication than NNs.

VII. Conclusions

Machine learning algorithms have performed remarkably well in many academic fields, including predicting asset returns. However, the mechanisms behind their success remain an open question, making "interpretable AI" one of the leading fields in machine learning research.

This paper employs an economically motivated dynamic trading strategy to interpret the performance of neural networks in trading anomalies. Our main findings are threefold. First, our strategy carries two important aspects that are widely believed to underly the success of machine learning algorithms: dynamic anomaly weighting and a ReLU activation function. Both aspects have a clear economic motivation in our strategy, which provides a way to interpret what is going on in the black box underlying machine learning models.

Second, we observe that neural networks deliver significant alphas, which also increase over longer training windows. Nevertheless, across training windows our strategy can absorb a significant fraction of the alphas generated by neural network alphas. For a window of 60 months, the *Backtested DAP_N* reduces the NN alpha by 93 bps per month, or 87% of the total alpha value. With a 240-month window, the reduction is 62 bps per month, or 32% of the total alpha. Lastly, we show that even with a 240-month training window, the DAP strategy's portfolio returns explain a large fraction of the neural network alphas when both are restricted to trade only on published anomalies, excluding microcap stocks. For example, the *Backtested DAP_N* portfolio return explains 27 bps per month of the NN1's alpha, which is 53% of its value.

Our analyses reveal three main economic sources of neural network performance: a dynamic factor weighting strategy akin to our economic model, a general weighting towards unpublished anomalies, and a loading on the return of microcaps. The first two are novel in the literature. Our results call for further research, potentially utilizing economic principles to interpret the performance of state-of-the-art machine learning tools.

VIII. References

Admati, A., 1985. A Noisy Rational Expectations Equilibrium for Multi-asset Securities Markets. Econometrica 53(3):629–57.

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. Management Science, 69(5):2587–2619.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. Econometrica 70:191–221.

Boguth, O. and M. Simutin, 2018, Leverage constraints and asset prices: Insights from mutual fund risk taking, Journal of Financial Economics 127, 325–341.

Brandt, M., P. Santa-Clara and R. Valkanov, 2009, Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns, Review of Financial Studies 22(9), 3411-3447.

Carhart, M. 1997. On Persistence in Mutual Fund Performance. Journal of Finance 52(1):57–82.

Chamberlain, G. and M. Rothschild, 1983, Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51, 1281-1304.

Chordia, T., A. Goyal and A. Saretto, 2020. Anomolies and False Rejections. The Review of Financial Studies 33, 2134-2179

Cochrane, J. H., 2011. Presidential address: Discount rates. Journal of Finance 66:1047–108.

Cong, L., Tang, K., Wang, J. & Zhang, Y., 2021, AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI, working paper.

Da, R., S. Nagel, and D. Xiu, 2023, The Statistical Limit of Arbitrage, working paper.

Daniel, K., and Moskowitz, T. J., 2016, Momentum crashes. Journal of Financial Economics 122, 221-247.

DeMiguel, V., Martin-Utrera, A., Nogales, F.J. , Uppal, R., 2020. A transaction-cost perspective on the multitude of firm characteristics. Review of Financial Studies, forthcoming.

Didisheim, Antoine and Ke, Shikun and Kelly, Bryan T. and Malamud, Semyon, Complexity in Factor Pricing Models (March 13, 2023). Swiss Finance Institute Research Paper No. 23-19.

Dong, X., S. Feng and R. Sadka, 2021. Liquidity Risk and Mutual Fund Performance. Management Science 65(3):1020-1041.

Fama, E. F., and K. R. French. 1992. The cross-section of expected stock returns. Journal of Finance 47:427-65.

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

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

Feng, G., S. Giglio, and D. Xiu. 2020. Taming the factor zoo. Journal of Finance 75: 1327-1370.

Freyberger, J., A. Neuhierl, and M. Weber. 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326-77.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average U.S. monthly stock returns. Review of Financial Studies 30:4389–436.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. Review of Financial Studies, forthcoming.

Haddad, V., S. Kozak, and S. Santosh. 2020. Factor Timing. The Review of Financial Studies, 33. 1980–2018.

Harvey, C. R., Y. Liu, and H. Zhu, 2015a. . . . and the cross-section of expected returns, Review of Financial Studies 29, 5–68.

Harvey, C. R., Y. Liu, and H. Zhu, 2015b. Digesting anomalies: An investment approach. Review of Financial Studies 28:660-705.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting Anomalies: An Investment Approaceh. Review of Financial Studies 28: 650-705.

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

Huberman, G., 1982, A simple approach to arbitrage pricing theory. Journal of Economic Theory 28: 183-191.

Ingersoll, J. E., 1984, Some results in the theory of arbitrage pricing. Journal of Finance 39: 1021-1039.

Jegadeesh, N., (1990), Evidence of Predictable Behavior of Security Returns, Journal of Finance, 45, issue 3, p. 881-98.

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market e_ciency. Journal of Finance 48:65-91.

Jensen, T.I., Kelly, B. and Pedersen, L.H. (2023), Is There a Replication Crisis in Finance?. J Finance, 78: 2465-2518.

Jones, Christopher S. and Mo, Haitao, 2021, Out-of-Sample Performance of Mutual Fund Predictors. Review of Financial Studies, 34, 149–193.

Kacperczyk, M., C. Sialm, L. Zheng. 2008. Unobserved actions of mutual funds, Review of Financial Studies, 21 (6), 2379-2416.

Karolyi, G. A., & Van Nieuwerburgh, S. 2020. New methods for the cross-section of returns. The Review of Financial Studies, 33(5), 1879-1890.

Kelly, B., Malamud, S. and Zhou, K. (2024), The Virtue of Complexity in Return Prediction. J Finance, 79: 459-503.

Kelly, B., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501-34.

Kelly, B., & Xiu, D. 2023. Financial machine learning. Foundations and Trends® in Finance, 13(3-4), 205-363.

Kim, S., R. Korajczyk, and A. Neuhierl 2020, Arbitrage Portfolios, Review of Financial Studies, forthcoming.

Kozak, S., S. Nagel, and S. Santosh. 2020. Shrinking the cross section. Journal of Financial Economics 135, 271–292.

Kyle, A. S. 1985. Continuous Auctions and Insider Trading. Econometrica, 53(6), 1315–1335.

Li, X., Xiong, H., Li, X. et al. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64, 3197–3234 (2022).

Lou, D., Polk, C., and Skouras, S., 2019, A Tug of War: Overnight vs. Intraday Expected Returns, Journal of Financial Economics 134, 192-213.

Mamaysky, H., Spiegel, M. and Zhang, H., 2007, Improved Forecasting of Mutual Fund Alphas and Betas, Review of Finance 11: 359-400

Mamaysky, H., Spiegel, M. and Zhang, H., 2008, Estimating the Dynamics of Mutual Fund Alphas and Betas, Review of Financial Studies 21: 233-264.

McLean, R. D., and J. Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability?, Journal of Finance 71:5–31.

Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74:967–1012.

Ross, S. A., 1976, The arbitrage theory of capital asset pricing. Journal of Economic Theory 13: 341-360.

Shen, Z., and D. Xiu, 2024 Can Machines Learn Weak Signals? (February 11, 2024). Chicago Booth Research Paper No. 24-03.

Smith, S. and A. Timmermann, 2021a, Break Risk, Review of Financial Studies 34:2046-2100.

Smith, S. and A. Timmermann, 2021b, Have Risk Premia Vanished? Journal of Financial Economics, forthcoming.

Stambaugh, R. F., and Y., Yu, 2017, Mispricing Factors. Review of Financial Studies 30, 1270-1315.

Tobek, O., and M. Hronec, 2021, Does It Pay to Follow Anomalies Research? Machine Learning Approach with International Evidence, Journal of Financial Markets 56.

Table 1. Out-of-sample Performance Delivered by DAP Portfolios (with 60-month Rolling Windows)

This table reports the returns generated by various strategies. For each strategy, we tabulate the returns for the Low and High deciles, as well as risk-adjusted returns (i.e., alphas) for the High-minus-Low portfolio. Panels A, B, and C report results for equally-weighted anomalies, DAP- and neural network-sorted decile portfolios, where the latter two types of models are trained from a rolling window of 60 months. The OLS standard error is used to construct t-stats. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Exclude Small Stocks (20%)						Whole Sample					
	Low	High	H-L	FF3	FF5	FF7	Low	High	H-L	FF3	FF5	FF7
Panel A. Returns Generated by Anomalies												
Avg Anomaly	-0.571	0.870***	1.441***	1.942***	1.271***	0.887***	0.140	1.140***	0.997***	1.360***	1.036***	0.657**
	(-1.32)	(4.12)	(4.03)	(6.16)	(4.19)	(3.52)	(0.37)	(4.07)	(2.93)	(4.27)	(3.38)	(2.48)
Panel B. Returns Generated by DAP Portfolios												
Base DAP	-0.711	0.927***	1.638***	2.228***	1.573***	1.145***	-0.908*	0.803***	1.711***	2.342***	1.583***	1.084***
	(-1.55)	(4.28)	(4.26)	(6.61)	(4.77)	(4.24)	(-1.86)	(3.55)	(4.16)	(6.40)	(4.42)	(3.93)
Backtested DAP	-0.828*	1.060***	1.888***	2.384***	1.828***	1.823***	-0.844*	1.048***	1.893***	2.399***	1.859***	1.803***
	(-1.87)	(4.30)	(4.87)	(6.73)	(5.27)	(7.12)	(-1.82)	(4.03)	(4.60)	(6.28)	(4.83)	(6.22)
DAP_N	-0.510	0.976***	1.486***	1.994***	1.481***	1.150***	-0.615	0.871***	1.486***	2.036***	1.446***	1.042***
	(-1.22)	(4.61)	(4.48)	(6.83)	(5.24)	(4.84)	(-1.38)	(3.82)	(4.10)	(6.26)	(4.47)	(3.97)
Backtested DAP_N	-0.605	1.084***	1.689***	2.144***	1.656***	1.658***	-0.968**	1.080***	2.048***	2.548***	2.096***	2.061***
	(-1.50)	(4.82)	(5.17)	(7.31)	(5.82)	(7.71)	(-2.23)	(4.31)	(5.55)	(7.50)	(6.12)	(8.13)
Panel C. Returns Generated by Neural Network Portfolios												
NN1	-0.368	1.029***	1.397***	1.846***	1.532***	1.072***	0.115	1.308***	1.194***	1.453***	1.462***	1.088***
	(-0.91)	(4.03)	(4.08)	(5.72)	(4.69)	(3.95)	(0.32)	(4.53)	(3.78)	(4.72)	(4.84)	(3.99)
NN2	-0.376	0.946***	1.322***	1.799***	1.439***	0.964***	0.080	1.160***	1.080***	1.372***	1.189***	0.771***
	(-0.90)	(3.86)	(3.78)	(5.52)	(4.35)	(3.47)	(0.22)	(4.13)	(3.34)	(4.39)	(3.84)	(2.82)
NN3	-0.330	0.900***	1.229***	1.632***	1.150***	0.678**	-0.102	1.226***	1.328***	1.598***	1.365***	0.926***
	(-0.80)	(3.63)	(3.49)	(4.87)	(3.41)	(2.38)	(-0.29)	(4.22)	(4.06)	(5.05)	(4.32)	(3.37)

Table 2. NN Returns Explained by FF7 and DAPs (60 months)

This table reports the results of the test in which we use the FF7 factors as well as the H-L DAP returns to explain the time series of H-L returns generated by the neural networks. In Panel A, the dependent variable is the H-L returns generated by NN1. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha with no DAP added. For columns 2-5, we add baseline DAP, Backtested DAP, DAP with normalized rolling alphas (DAP_N), and backtested DAP_N as explanatory variables sequentially. Panel B applies the same test to the H-L returns generated by NN3. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: NN1 Returns Explained by FF7 and DAPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)					Whole Sample				
Alpha	1.072***	0.223	0.261	0.081	0.136	1.088***	0.641**	0.693**	0.302	0.324
	(3.95)	(1.19)	(1.00)	(0.44)	(0.52)	(3.99)	(2.54)	(2.50)	(1.56)	(1.32)
SMB	-0.107	0.054	0.011	0.169**	0.044	0.004	0.101	0.076	0.060	-0.028
	(-1.09)	(0.80)	(0.12)	(2.58)	(0.50)	(0.04)	(1.12)	(0.78)	(0.87)	(-0.32)
HML	-0.198*	-0.040	-0.085	0.045	-0.012	0.018	0.161	0.083	0.104	0.042
	(-1.68)	(-0.50)	(-0.78)	(0.57)	(-0.11)	(0.15)	(1.48)	(0.71)	(1.25)	(0.41)
RMW	-0.134	-0.424***	-0.202*	-0.214***	-0.206*	-0.809***	-1.132***	-0.886***	-0.565***	-0.733***
	(-1.08)	(-4.97)	(-1.78)	(-2.61)	(-1.84)	(-6.45)	(-9.50)	(-7.20)	(-6.39)	(-6.70)
CMA	0.652***	-0.101	0.168	-0.282**	0.074	0.755***	0.487***	0.614***	-0.297**	0.119
	(3.77)	(-0.83)	(1.02)	(-2.35)	(0.45)	(4.34)	(3.03)	(3.58)	(-2.26)	(0.74)
MKT	-0.266***	0.002	-0.219***	-0.011	-0.189***	-0.251***	-0.137**	-0.241***	-0.023	-0.207***
	(-4.04)	(0.05)	(-3.64)	(-0.25)	(-3.19)	(-3.80)	(-2.23)	(-3.74)	(-0.48)	(-3.59)
MOM	0.859***	0.231***	0.720***	0.282***	0.725***	0.644***	0.232***	0.552***	0.258***	0.536***
	(13.91)	(4.60)	(12.41)	(5.96)	(12.76)	(10.38)	(3.25)	(8.74)	(5.48)	(9.79)
SREV	-0.054	0.134**	0.544***	0.177***	0.552***	0.206**	0.353***	0.511***	0.318***	0.731***
	(-0.65)	(2.39)	(5.56)	(3.22)	(5.86)	(2.49)	(4.60)	(5.03)	(5.48)	(8.61)
Base DAP		0.742***					0.412***			
		(22.52)					(9.43)			
Backtested DAP			0.445***					0.219***		
			(9.50)					(4.91)		
DAP_N				0.862***					0.829***	
				(23.66)					(21.09)	
Backtested DAP_N					0.565***					0.567***
					(10.29)					(11.67)
Adj R2	0.446	0.748	0.543	0.761	0.556	0.345	0.678	0.470	0.680	0.503
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: NN3 Returns Explained by FF7 and DAPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)							Whole Sample		
Alpha	0.678**	-0.152	-0.087	-0.188	-0.070	0.926***	0.448*	0.464*	0.206	0.331
	(2.38)	(-0.72)	(-0.31)	(-0.83)	(-0.24)	(3.37)	(1.79)	(1.68)	(0.97)	(1.26)
SMB	-0.150	-0.018	-0.039	0.092	-0.029	0.223**	0.327***	0.307***	0.011	0.199**
	(-1.45)	(-0.20)	(-0.40)	(1.12)	(-0.29)	(2.25)	(3.65)	(3.17)	(0.12)	(2.16)
HML	-0.163	-0.074	-0.056	0.049	-0.015	-0.067	0.086	0.008	-0.174*	-0.048
	(-1.31)	(-0.77)	(-0.48)	(0.50)	(-0.12)	(-0.56)	(0.79)	(0.07)	(-1.80)	(-0.44)
RMW	0.201	-0.000	0.136	0.131	0.144	-0.398***	-0.744***	-0.488***	-0.064	-0.339***
	(1.53)	(-0.00)	(1.12)	(1.28)	(1.16)	(-3.15)	(-6.29)	(-3.98)	(-0.45)	(-2.90)
CMA	0.757***	-0.034	0.301*	-0.059	0.295	0.900***	0.613***	0.735***	0.025	0.404**
	(4.17)	(-0.66)	(1.70)	(-0.39)	(1.62)	(5.14)	(3.84)	(4.30)	(0.48)	(2.35)
MKT	-0.193***	0.376***	-0.148**	0.030	-0.131**	-0.184***	-0.062	-0.172***	0.411***	-0.150**
	(-2.79)	(7.26)	(-2.30)	(0.55)	(-2.00)	(-2.77)	(-1.02)	(-2.69)	(8.00)	(-2.43)
MOM	0.866***	0.272***	0.735***	0.362***	0.759***	0.765***	0.324***	0.658***	0.298***	0.681***
	(13.37)	(4.32)	(11.84)	(6.15)	(12.11)	(12.26)	(4.57)	(10.45)	(4.70)	(11.62)
SREV	0.008	0.147*	0.571***	0.209***	0.492***	0.195**	0.353***	0.553***	0.275***	0.604***
	(0.09)	(1.96)	(5.46)	(3.06)	(4.73)	(2.34)	(4.63)	(5.46)	(3.64)	(6.66)
Base DAP		0.725***					0.441***			
		(19.47)					(10.17)			
Backtested DAP			0.420***					0.256***		
			(8.37)					(5.78)		
DAP_N				0.753***					0.760***	
				(16.62)					(17.69)	
Backtested DAP_N					0.451***					0.442***
					(7.45)					(8.51)
Adj R2	0.425	0.696	0.505	0.651	0.490	0.379	0.646	0.459	0.642	0.468
Observations	432	432	432	432	432	432	432	432	432	432

Table 3: DAP Returns Explained by NN Models (60 months)

This table reports the results of the revert test in which we use the FF7 factors as well as the H-L returns of various NN models to explain the time series of H-L returns generated by DAP. In Panel A, the dependent variable is the H-L returns generated by backtested DAP, when EW rolling-window alphas are used. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha of DAP when no NN is added. For columns 2-5, we add NN1, NN2, and NN3 as explanatory variables. Panel B applies the same test to the H-L returns generated by backtested DAP with normalized conditional alphas when EW rolling-window alphas are used. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: BackTested DAP Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.823***	1.399***	1.496***	1.593***	1.406***	1.803***	1.535***	1.610***	1.539***	1.493***
	(7.12)	(5.91)	(6.19)	(6.67)	(6.00)	(6.22)	(5.34)	(5.66)	(5.44)	(5.24)
SMB	-0.265***	-0.223***	-0.179**	-0.214**	-0.213**	-0.327***	-0.328***	-0.337***	-0.391***	-0.374***
	(-2.86)	(-2.64)	(-2.06)	(-2.49)	(-2.53)	(-3.12)	(-3.21)	(-3.30)	(-3.85)	(-3.67)
HML	-0.255**	-0.177*	-0.167	-0.200*	-0.173*	-0.295**	-0.300**	-0.278**	-0.276**	-0.279**
	(-2.29)	(-1.74)	(-1.60)	(-1.93)	(-1.72)	(-2.34)	(-2.44)	(-2.27)	(-2.27)	(-2.30)
RMW	0.153	0.206*	0.155	0.085	0.158	0.353***	0.553***	0.481***	0.467***	0.522***
	(1.30)	(1.92)	(1.41)	(0.78)	(1.47)	(2.65)	(4.07)	(3.65)	(3.60)	(3.88)
CMA	1.087***	0.829***	0.868***	0.831***	0.775***	0.643***	0.457**	0.403**	0.386**	0.352*
	(6.66)	(5.50)	(5.63)	(5.38)	(5.16)	(3.48)	(2.48)	(2.17)	(2.10)	(1.91)
MKT	-0.106*	-0.001	-0.006	-0.041	-0.000	-0.046	0.016	0.017	0.006	0.024
	(-1.71)	(-0.02)	(-0.11)	(-0.71)	(-0.00)	(-0.66)	(0.22)	(0.25)	(0.09)	(0.34)
MOM	0.312***	-0.027	0.019	0.019	-0.076	0.418***	0.259***	0.240***	0.200***	0.177**
	(5.36)	(-0.43)	(0.29)	(0.29)	(-1.17)	(6.35)	(3.61)	(3.28)	(2.70)	(2.36)
SREV	-1.342***	-1.321***	-1.356***	-1.344***	-1.326***	-1.396***	-1.446***	-1.465***	-1.451***	-1.466***
	(-17.30)	(-18.73)	(-18.78)	(-18.70)	(-18.94)	(-15.90)	(-16.80)	(-16.94)	(-17.04)	(-17.12)
NN1		0.395***			0.299***		0.247***			0.075
		(9.50)			(4.25)		(4.91)			(0.92)
NN2			0.339***		-0.025			0.251***		0.057
			(8.12)		(-0.33)			(5.03)		(0.67)
NN3				0.339***	0.176***				0.286***	0.200***
				(8.37)	(3.16)				(5.78)	(2.89)
Adj R2	0.616	0.683	0.667	0.670	0.690	0.639	0.670	0.664	0.657	0.670
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: Backtested DAP_N Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.658***	1.278***	1.389***	1.484***	1.285***	2.061***	1.797***	1.873***	1.790***	1.749***
	(7.71)	(6.52)	(6.82)	(7.28)	(6.56)	(8.13)	(7.20)	(7.58)	(7.34)	(7.12)
SMB	-0.268***	-0.230***	-0.197***	-0.230***	-0.232***	-0.247***	-0.248***	-0.257***	-0.313***	-0.299***
	(-3.44)	(-3.30)	(-2.69)	(-3.13)	(-3.30)	(-2.69)	(-2.80)	(-2.89)	(-3.57)	(-3.40)
HML	-0.329***	-0.259***	-0.257***	-0.287***	-0.261***	-0.352***	-0.356***	-0.335***	-0.332***	-0.336***
	(-3.52)	(-3.08)	(-2.92)	(-3.25)	(-3.11)	(-3.19)	(-3.34)	(-3.15)	(-3.17)	(-3.20)
RMW	0.126	0.174*	0.128	0.075	0.153*	0.218*	0.414***	0.343***	0.334***	0.383***
	(1.28)	(1.96)	(1.39)	(0.80)	(1.72)	(1.87)	(3.51)	(2.99)	(2.99)	(3.30)
CMA	1.023***	0.792***	0.843***	0.829***	0.769***	0.615***	0.432***	0.381**	0.351**	0.325**
	(7.47)	(6.35)	(6.48)	(6.30)	(6.14)	(3.81)	(2.70)	(2.36)	(2.22)	(2.04)
MKT	-0.136***	-0.042	-0.053	-0.086*	-0.044	-0.095	-0.034	-0.033	-0.041	-0.027
	(-2.60)	(-0.87)	(-1.07)	(-1.74)	(-0.92)	(-1.54)	(-0.56)	(-0.55)	(-0.70)	(-0.45)
MOM	0.237***	-0.067	-0.004	0.015	-0.086	0.354***	0.197***	0.180***	0.130**	0.111*
	(4.86)	(-1.26)	(-0.07)	(0.27)	(-1.57)	(6.14)	(3.17)	(2.84)	(2.04)	(1.72)
SREV	-1.073***	-1.054***	-1.085***	-1.075***	-1.053***	-1.285***	-1.335***	-1.352***	-1.342***	-1.353***
	(-16.46)	(-18.05)	(-17.82)	(-17.52)	(-18.01)	(-16.72)	(-17.85)	(-18.00)	(-18.28)	(-18.33)
NN1		0.354***			0.338***		0.243***			0.072
		(10.29)			(5.76)		(5.57)			(1.02)
NN2			0.279***		-0.053			0.245***		0.037
			(7.94)		(-0.85)			(5.63)		(0.51)
NN3				0.257***	0.091*				0.293***	0.222***
				(7.45)	(1.94)				(6.87)	(3.72)
Adj R2	0.618	0.694	0.667	0.661	0.695	0.584	0.611	0.612	0.625	0.626
Observations	432	432	432	432	432	432	432	432	432	432

Table 4. Out-of-sample Performance Delivered by DAP Portfolios (240 months)

This table reports the returns generated by DAPs and NNs when both types of models are trained from a rolling window of 240 months. For each DAP strategy, we tabulate the returns for the Low and High deciles, as well as risk-adjusted returns (i.e., alphas) for the High-minus-Low portfolio. Panels A and B report results for DAP and neural network-sorted decile portflios, respectively. The OLS standard error is used to construct t-stats. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Exclude Small Stocks (20%)						Whole Sample					
	Low	High	H-L	FF3	FF5	FF7	Low	High	H-L	FF3	FF5	FF7
Panel A. Returns Generated by DAP Portfolios												
Base DAP	-0.820*	0.931***	1.751***	2.170***	1.256***	0.964***	-0.933*	0.927***	1.860***	2.394***	1.437***	1.075***
	(-1.79)	(5.01)	(4.68)	(8.54)	(6.03)	(5.65)	(-1.86)	(4.82)	(4.51)	(7.96)	(5.53)	(5.05)
Backtested DAP	-0.935**	1.199***	2.133***	2.575***	1.848***	1.854***	-1.034**	1.234***	2.268***	2.737***	2.010***	2.056***
	(-2.12)	(5.37)	(5.78)	(8.10)	(6.11)	(7.88)	(-2.20)	(5.23)	(5.73)	(7.91)	(5.95)	(7.90)
DAP_N	-0.594	0.986***	1.581***	1.933***	1.178***	0.913***	-0.791*	0.999***	1.790***	2.274***	1.429***	1.132***
	(-1.46)	(5.13)	(5.09)	(8.68)	(6.15)	(5.70)	(-1.72)	(5.12)	(5.03)	(8.82)	(6.40)	(5.98)
Backtested DAP_N	-0.808**	1.202***	2.010***	2.369***	1.703***	1.724***	-0.939**	1.239***	2.178***	2.617***	1.955***	2.031***
	(-2.01)	(5.55)	(6.18)	(8.44)	(6.40)	(8.41)	(-2.16)	(5.32)	(6.14)	(8.35)	(6.40)	(8.86)
Panel B. Returns Generated by Neural Network Portfolios												
NN1	-0.844**	1.660***	2.503***	2.672***	2.285***	1.930***	-0.637*	2.290***	2.926***	2.935***	2.654***	2.499***
	(-2.21)	(5.77)	(9.14)	(10.29)	(8.71)	(8.45)	(-1.77)	(6.45)	(10.69)	(10.98)	(9.78)	(9.37)
NN2	-0.780**	1.557***	2.337***	2.647***	2.106***	1.777***	-0.426	1.851***	2.278***	2.313***	1.971***	1.781***
	(-1.97)	(6.45)	(7.85)	(10.10)	(8.21)	(7.85)	(-1.26)	(6.04)	(8.39)	(8.99)	(7.60)	(7.28)
NN3	-0.540	1.291***	1.831***	2.087***	1.647***	1.424***	-0.443	1.722***	2.165***	2.202***	1.833***	1.597***
	(-1.41)	(5.40)	(6.51)	(8.59)	(6.85)	(6.23)	(-1.30)	(5.77)	(7.85)	(8.39)	(6.94)	(6.32)

Table 5: NN Returns Explained by FF7 and DAPs (240 months)

This table reports the results of the test in which we use the FF7 factors as well as the H-L DAP returns to explain the time series of H-L returns generated by the neural networks. In Panel A, the dependent variable is the H-L returns generated by NN1. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha with no DAP added. For columns 2-5, we add baseline DAP, Backtested DAP, DAP with normalized rolling alphas (DAP_N), and backtested DAP_N as explanatory variables sequentially. Panel B applies the same test to the H-L returns generated by NN3. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: NN1 Returns Explained by FF7 and DAP

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)					Whole Sample				
Alpha	1.930***	1.401***	1.253***	1.309***	1.164***	2.499***	2.265***	2.065***	2.064***	1.853***
	(8.45)	(6.48)	(5.53)	(6.28)	(5.14)	(9.37)	(8.36)	(7.38)	(7.71)	(6.62)
SMB	-0.001	0.303***	0.091	0.252***	0.074	0.086	0.231**	0.167*	0.274***	0.169*
	(-0.02)	(3.68)	(1.17)	(3.30)	(0.97)	(0.89)	(2.23)	(1.73)	(2.78)	(1.80)
HML	0.315***	0.138	0.342***	0.152*	0.314***	0.136	0.113	0.172	0.108	0.207*
	(3.17)	(1.48)	(3.71)	(1.71)	(3.44)	(1.17)	(0.98)	(1.51)	(0.97)	(1.84)
RMW	0.270**	-0.446***	0.064	-0.364***	0.017	0.038	-0.275*	-0.120	-0.410***	-0.173
	(2.58)	(-3.62)	(0.64)	(-3.36)	(0.17)	(0.31)	(-1.85)	(-0.95)	(-2.91)	(-1.40)
CMA	0.523***	0.071	0.086	-0.077	0.058	0.862***	0.716***	0.671***	0.555***	0.587***
	(3.59)	(0.50)	(0.60)	(-0.56)	(0.40)	(5.07)	(4.15)	(3.90)	(3.22)	(3.44)
MKT	-0.094*	0.077	-0.044	0.089*	-0.052	-0.053	0.024	-0.032	0.088	-0.021
	(-1.70)	(1.42)	(-0.85)	(1.73)	(-1.01)	(-0.81)	(0.36)	(-0.50)	(1.32)	(-0.34)
MOM	0.630***	0.325***	0.543***	0.293***	0.548***	0.211***	0.062	0.168***	-0.003	0.169***
	(12.15)	(5.63)	(11.03)	(5.34)	(11.29)	(3.49)	(0.85)	(2.79)	(-0.04)	(2.87)
SREV	0.098	0.143**	0.500***	0.125**	0.541***	0.332***	0.355***	0.602***	0.353***	0.721***
	(1.42)	(2.25)	(6.24)	(2.05)	(6.72)	(4.11)	(4.44)	(5.97)	(4.53)	(7.04)
Base DAP		0.549***					0.217***			
		(9.24)					(3.62)			
Backtested DAP			0.365***					0.211***		
			(8.35)					(4.33)		
DAP_N				0.681***					0.384***	
				(11.17)					(5.83)	
Backtested DAP_N					0.444***					0.318***
					(8.95)					(5.84)
Adj R2	0.419	0.522	0.458	0.496	0.466	0.164	0.187	0.197	0.224	0.224
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: NN3 Returns Explained by FF7 and DAPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)					Whole Sample				
Alpha	1.424***	0.878***	0.948***	0.945***	0.868***	1.597***	1.416***	1.300***	1.215***	1.140***
	(6.23)	(4.09)	(4.01)	(4.28)	(3.67)	(6.32)	(5.50)	(4.86)	(4.77)	(4.23)
SMB	-0.350***	-0.035	-0.285***	-0.155*	-0.295***	0.016	0.128	0.071	0.181*	0.075
	(-4.23)	(-0.43)	(-3.53)	(-1.92)	(-3.70)	(0.17)	(1.30)	(0.77)	(1.93)	(0.82)
HML	0.161	-0.022	0.180*	0.035	0.160*	0.251**	0.233**	0.275**	0.226**	0.301***
	(1.62)	(-0.24)	(1.88)	(0.38)	(1.68)	(2.28)	(2.13)	(2.53)	(2.13)	(2.78)
RMW	0.646***	-0.093	0.501***	0.157	0.462***	0.174	-0.068	0.067	-0.219	0.025
	(6.15)	(-0.76)	(4.79)	(1.37)	(4.41)	(1.50)	(-0.48)	(0.55)	(-1.63)	(0.21)
CMA	0.326**	-0.141	0.019	-0.137	-0.011	0.867***	0.753***	0.736***	0.596***	0.671***
	(2.24)	(-1.00)	(0.12)	(-0.93)	(-0.08)	(5.39)	(4.59)	(4.47)	(3.64)	(4.09)
MKT	-0.219***	-0.042	-0.184***	-0.078	-0.188***	-0.059	0.000	-0.045	0.065	-0.037
	(-3.95)	(-0.79)	(-3.41)	(-1.43)	(-3.53)	(-0.97)	(0.01)	(-0.74)	(1.02)	(-0.62)
MOM	0.388***	0.073	0.327***	0.128**	0.328***	0.357***	0.241***	0.327***	0.168**	0.326***
	(7.48)	(1.27)	(6.37)	(2.21)	(6.48)	(6.21)	(3.48)	(5.68)	(2.56)	(5.76)
SREV	0.097	0.143**	0.380***	0.117*	0.418***	0.348***	0.366***	0.533***	0.366***	0.624***
	(1.40)	(2.26)	(4.54)	(1.82)	(4.97)	(4.55)	(4.81)	(5.53)	(4.94)	(6.32)
Base DAP		0.566***					0.168***			
		(9.60)					(2.95)			
Backtested DAP			0.257***					0.144***		
			(5.64)					(3.10)		
DAP_N				0.525***					0.338***	
				(8.13)					(5.38)	
Backtested DAP_N					0.322***					0.225***
					(6.22)					(4.29)
Adj R2	0.419	0.522	0.458	0.496	0.466	0.263	0.276	0.278	0.309	0.292
Observations	432	432	432	432	432	432	432	432	432	432

Table 6: NN explain DAP (240 months)

This table reports the results of the revert test in which we use the FF7 factors as well as the H-L returns of various NN models to explain the time series of H-L returns generated by DAP. In Panel A, the dependent variable is the H-L returns generated by backtested DAP, when EW rolling-window alphas are used. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha of DAP when no NN is added. For columns 2-5, we add NN1, NN2, and NN3 as explanatory variables. Panel B applies the same test to the H-L returns generated by backtested DAP with normalized conditional alphas when EW rolling-window alphas are used. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: BackTested DAP Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.854***	1.106***	1.102***	1.466***	0.910***	2.056***	1.554***	1.734***	1.810***	1.480***
	(7.88)	(4.69)	(4.79)	(6.18)	(3.90)	(7.90)	(5.54)	(6.36)	(6.71)	(5.23)
SMB	-0.253***	-0.252***	-0.211***	-0.158*	-0.186**	-0.382***	-0.400***	-0.391***	-0.385***	-0.400***
	(-2.97)	(-3.19)	(-2.70)	(-1.88)	(-2.36)	(-4.05)	(-4.32)	(-4.20)	(-4.12)	(-4.33)
HML	-0.074	-0.196**	-0.150	-0.118	-0.192**	-0.172	-0.200*	-0.201*	-0.211*	-0.212*
	(-0.72)	(-2.04)	(-1.60)	(-1.19)	(-2.05)	(-1.52)	(-1.80)	(-1.79)	(-1.87)	(-1.90)
RMW	0.565***	0.461***	0.368***	0.389***	0.326***	0.746***	0.739***	0.717***	0.720***	0.722***
	(5.23)	(4.56)	(3.64)	(3.58)	(3.18)	(6.24)	(6.30)	(6.06)	(6.06)	(6.14)
CMA	1.195***	0.993***	0.862***	1.106***	0.869***	0.904***	0.731***	0.732***	0.771***	0.668***
	(7.97)	(7.03)	(6.08)	(7.60)	(6.20)	(5.45)	(4.36)	(4.29)	(4.54)	(3.91)
MKT	-0.138**	-0.101*	-0.034	-0.078	-0.034	-0.097	-0.086	-0.071	-0.088	-0.074
	(-2.42)	(-1.91)	(-0.64)	(-1.40)	(-0.63)	(-1.54)	(-1.40)	(-1.13)	(-1.41)	(-1.19)
MOM	0.239***	-0.005	-0.010	0.133**	-0.062	0.205***	0.162***	0.164***	0.150**	0.145**
	(4.47)	(-0.09)	(-0.18)	(2.42)	(-1.09)	(3.46)	(2.76)	(2.76)	(2.45)	(2.40)
SREV	-1.101***	-1.139***	-1.130***	-1.128***	-1.146***	-1.280***	-1.347***	-1.382***	-1.334***	-1.390***
	(-15.46)	(-17.20)	(-17.32)	(-16.35)	(-17.76)	(-16.23)	(-17.09)	(-16.66)	(-16.68)	(-16.90)
NN1		0.387***			0.168**		0.201***			0.152***
		(8.35)			(2.58)		(4.33)			(2.66)
NN2			0.423***		0.254***			0.181***		0.096
			(9.16)		(3.75)			(3.54)		(1.56)
NN3				0.272***	0.118**				0.154***	0.016
				(5.64)	(2.40)				(3.10)	(0.25)
Adj R2	0.643	0.692	0.701	0.667	0.709	0.619	0.634	0.629	0.627	0.635
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: Backtested DAP_N Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.724***	1.032***	1.070***	1.354***	0.858***	2.031***	1.445***	1.645***	1.735***	1.353***
	(8.41)	(5.07)	(5.33)	(6.59)	(4.26)	(8.86)	(5.96)	(6.95)	(7.38)	(5.55)
SMB	-0.170**	-0.170**	-0.133*	-0.079	-0.107	-0.262***	-0.282***	-0.273***	-0.265***	-0.283***
	(-2.29)	(-2.49)	(-1.96)	(-1.09)	(-1.57)	(-3.16)	(-3.53)	(-3.37)	(-3.26)	(-3.56)
HML	0.003	-0.110	-0.063	-0.039	-0.109	-0.224**	-0.256***	-0.258***	-0.270***	-0.272***
	(0.04)	(-1.33)	(-0.77)	(-0.45)	(-1.35)	(-2.24)	(-2.66)	(-2.65)	(-2.75)	(-2.83)
RMW	0.570***	0.473***	0.399***	0.402***	0.353***	0.664***	0.655***	0.628***	0.631***	0.634***
	(6.05)	(5.43)	(4.52)	(4.27)	(3.98)	(6.30)	(6.45)	(6.10)	(6.10)	(6.25)
CMA	1.046***	0.859***	0.756***	0.961***	0.764***	0.867***	0.665***	0.661***	0.706***	0.586***
	(8.00)	(7.05)	(6.11)	(7.63)	(6.31)	(5.93)	(4.59)	(4.46)	(4.77)	(3.98)
MKT	-0.095*	-0.062	-0.005	-0.039	-0.005	-0.098*	-0.086	-0.066	-0.087	-0.071
	(-1.92)	(-1.35)	(-0.11)	(-0.79)	(-0.12)	(-1.76)	(-1.60)	(-1.22)	(-1.60)	(-1.31)
MOM	0.186***	-0.040	-0.031	0.085*	-0.089*	0.135**	0.085*	0.086*	0.068	0.063
	(3.99)	(-0.81)	(-0.63)	(1.79)	(-1.79)	(2.58)	(1.67)	(1.66)	(1.28)	(1.21)
SREV	-0.996***	-1.031***	-1.021***	-1.021***	-1.039***	-1.224***	-1.302***	-1.346***	-1.289***	-1.355***
	(-16.03)	(-18.04)	(-17.94)	(-17.11)	(-18.63)	(-17.62)	(-19.09)	(-18.67)	(-18.48)	(-19.11)
NN1		0.358***			0.188***		0.235***			0.172***
		(8.95)			(3.34)		(5.84)			(3.49)
NN2			0.368***		0.180***			0.217***		0.117**
			(9.14)		(3.09)			(4.89)		(2.20)
NN3				0.260***	0.128***				0.186***	0.025
				(6.22)	(3.00)				(4.29)	(0.46)
Adj R2	0.650	0.705	0.707	0.679	0.720	0.633	0.659	0.651	0.647	0.663
Observations	432	432	432	432	432	432	432	432	432	432

Table 7: Portfolio Returns and Publication Bias (240-mon Training Window)

This table reports the results when a pre-publication ratio ("pre_pub"), which denotes for each year the fraction of anomalies that are before their formal publication, is further included to explain the H-L portfolio returns of NN1. The inclusion of the pre-publication ratio expands the previous tests (Table 7, Panel A) to explain NN1 returns by FF7 and backtested DAP with normalization. The first four columns report the results when 20% of microcap stocks are excluded, whereas all stocks are included in the test reported in the next four columns. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Excluding 20% Microcap Stocks				All Stocks			
VARIABLES	nn1	nn1	nn1	nn1	nn1	nn1	nn1	nn1
Constant	1.930***	1.164***	0.786**	0.296	2.499***	2.037***	1.332***	1.126***
	(8.45)	(5.14)	(2.07)	(0.83)	(9.37)	(7.75)	(2.99)	(2.64)
pre_pub			2.283***	1.786***			2.329***	1.856***
			(3.73)	(3.15)			(3.25)	(2.69)
Backtested DAP_N		0.444***		0.428***		0.414***		0.397***
		(8.95)		(8.68)		(6.78)		(6.51)
smb	-0.001	0.074	0.000	0.073	0.086	0.019	0.088	0.023
	(-0.02)	(0.97)	(0.00)	(0.96)	(0.89)	(0.20)	(0.92)	(0.25)
hml	0.315***	0.314***	0.294***	0.298***	0.136	-0.020	0.114	-0.030
	(3.17)	(3.44)	(3.00)	(3.29)	(1.17)	(-0.18)	(0.99)	(-0.27)
rmw	0.270**	0.017	0.278***	0.032	0.038	-0.154	0.046	-0.140
	(2.58)	(0.17)	(2.69)	(0.32)	(0.31)	(-1.29)	(0.38)	(-1.18)
cma	0.523***	0.058	0.525***	0.077	0.862***	0.438**	0.865***	0.458***
	(3.59)	(0.40)	(3.67)	(0.54)	(5.07)	(2.52)	(5.14)	(2.66)
mkt_rf	-0.094*	-0.052	-0.094*	-0.053	-0.053	-0.053	-0.053	-0.053
	(-1.70)	(-1.01)	(-1.73)	(-1.05)	(-0.81)	(-0.87)	(-0.82)	(-0.87)
mom	0.630***	0.548***	0.612***	0.536***	0.211***	0.252***	0.192***	0.235***
	(12.15)	(11.29)	(11.91)	(11.13)	(3.49)	(4.35)	(3.19)	(4.06)
st_rev	0.098	0.541***	0.087	0.516***	0.332***	0.629***	0.321***	0.608***
	(1.42)	(6.72)	(1.27)	(6.45)	(4.11)	(7.11)	(4.01)	(6.89)
Observations	432	432	432	432	432	432	432	432
r2_a	0.389	0.485	0.407	0.495	0.164	0.244	0.182	0.255

Table 8: Explaining Neural Network Returns Generated by Published-Anomalies (240 months)

This table reports the results of the test in which we use the FF7 factors as well as the H-L DAP returns to explain the time series of H-L returns generated by the neural networks. Unlike the previous setup, for DAP and NN, we only utilized currently published anomalies during estimation. In Panel A, the dependent variable is the H-L returns generated by NN1. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha with no DAP added. For columns 2-5, we add baseline DAP, Backtested DAP, DAP with normalized rolling alphas (DAP_N), and backtested DAP_N as explanatory variables sequentially. Panel B applies the same test to the H-L returns generated by NN3. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: NN1 Returns Explained by FF7 and DAP

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)						Whole Sample			
Alpha	0.514***	0.330**	0.213	0.368**	0.241	0.884***	0.487***	0.385**	0.781***	0.787***
	(2.61)	(1.98)	(1.08)	(2.18)	(1.23)	(4.49)	(2.79)	(2.30)	(3.95)	(3.98)
SMB	-0.151**	0.022	-0.048	-0.149**	-0.171**	0.277***	-0.068	0.006	0.227***	0.236***
	(-2.12)	(0.36)	(-0.79)	(-2.17)	(-2.48)	(3.89)	(-0.99)	(0.10)	(3.14)	(3.28)
HML	0.453***	0.121	0.164**	0.417***	0.393***	0.547***	0.238***	0.219***	0.481***	0.502***
	(5.30)	(1.59)	(2.13)	(5.04)	(4.72)	(6.38)	(3.01)	(2.92)	(5.50)	(5.81)
RMW	0.686***	0.199**	0.336***	0.613***	0.612***	0.197**	0.142*	0.116	0.184**	0.171*
	(7.59)	(2.35)	(4.08)	(6.97)	(6.94)	(2.18)	(1.80)	(1.54)	(2.05)	(1.90)
CMA	0.495***	0.253**	0.196*	0.301**	0.328***	0.559***	-0.001	-0.162	0.437***	0.425***
	(3.95)	(2.36)	(1.78)	(2.40)	(2.63)	(4.45)	(-0.01)	(-1.39)	(3.36)	(3.21)
MKT	-0.185***	-0.070*	-0.130***	-0.167***	-0.181***	-0.163***	-0.130***	-0.120***	-0.179***	-0.182***
	(-3.88)	(-1.71)	(-3.16)	(-3.62)	(-3.93)	(-3.42)	(-3.13)	(-3.01)	(-3.77)	(-3.81)
MOM	0.639***	0.208***	0.271***	0.583***	0.576***	0.061	0.160***	0.174***	0.090**	0.083*
	(14.29)	(4.16)	(5.60)	(13.20)	(12.92)	(1.35)	(4.01)	(4.57)	(2.00)	(1.85)
SREV	0.391***	0.344***	0.360***	0.608***	0.620***	0.560***	0.404***	0.418***	0.628***	0.627***
	(6.56)	(6.83)	(7.04)	(8.85)	(8.81)	(9.39)	(7.53)	(8.24)	(9.99)	(9.92)
Base DAP		0.535***					0.481***			
		(13.16)					(11.76)			
Backtested DAP			0.192***					0.656***		
			(5.76)					(13.81)		
DAP_N				0.540***					0.108***	
				(12.42)					(3.14)	
Backtested DAP_N					0.210***					0.114***
					(5.64)					(2.99)
Adj R2	0.593	0.711	0.623	0.702	0.622	0.432	0.572	0.608	0.444	0.443
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: NN3 Returns Explained by FF7 and DAPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)						Whole Sample			
Alpha	0.724***	0.427**	0.492***	0.328	0.348	0.601***	0.233	0.087	0.447**	0.440**
	(3.27)	(2.53)	(2.66)	(1.51)	(1.61)	(3.17)	(1.37)	(0.55)	(2.39)	(2.36)
SMB	-0.121	0.096	-0.014	-0.117	-0.151**	0.228***	-0.091	-0.051	0.154**	0.160**
	(-1.51)	(1.55)	(-0.22)	(-1.54)	(-1.99)	(3.33)	(-1.35)	(-0.87)	(2.25)	(2.36)
HML	0.615***	0.162**	0.267***	0.572***	0.530***	0.629***	0.343***	0.291***	0.530***	0.554***
	(6.40)	(2.09)	(3.19)	(6.24)	(5.77)	(7.62)	(4.46)	(4.17)	(6.41)	(6.80)
RMW	0.482***	-0.149*	0.061	0.398***	0.374***	0.302***	0.251***	0.218***	0.283***	0.259***
	(4.75)	(-1.75)	(0.67)	(4.09)	(3.84)	(3.47)	(3.27)	(3.11)	(3.33)	(3.05)
CMA	0.433***	0.083	0.047	0.176	0.203	0.629***	0.111	-0.114	0.448***	0.408***
	(3.07)	(0.76)	(0.39)	(1.27)	(1.47)	(5.20)	(0.96)	(-1.05)	(3.64)	(3.27)
MKT	-0.241***	-0.087**	-0.168***	-0.217***	-0.234***	-0.170***	-0.140***	-0.126***	-0.194***	-0.202***
	(-4.50)	(-2.09)	(-3.74)	(-4.25)	(-4.61)	(-3.70)	(-3.44)	(-3.39)	(-4.31)	(-4.48)
MOM	0.689***	0.115**	0.243***	0.617***	0.602***	0.048	0.139***	0.165***	0.092**	0.085**
	(13.72)	(2.29)	(4.61)	(12.61)	(12.24)	(1.10)	(3.59)	(4.65)	(2.14)	(2.00)
SREV	0.114*	0.104**	0.122**	0.403***	0.434***	0.546***	0.401***	0.400***	0.647***	0.657***
	(1.71)	(2.05)	(2.19)	(5.26)	(5.55)	(9.51)	(7.69)	(8.48)	(10.86)	(11.02)
Base DAP		0.727***					0.445***			
		(17.65)					(11.16)			
Backtested DAP			0.661***					0.675***		
			(13.78)					(15.29)		
DAP_N				0.247***					0.160***	
				(6.76)					(4.93)	
Backtested DAP_N					0.290***					0.189***
					(7.00)					(5.23)
Adj R2	0.560	0.746	0.696	0.603	0.606	0.507	0.619	0.682	0.534	0.537
Observations	432	432	432	432	432	432	432	432	432	432

Table 9: Explaining the Importance of Anomalies in Neural Networks

This table compares the relative importance of various factors between the neural network NN1 and DAPs. Panel A presents the results of the following cross-sectional regression:

$$NN_ImportanceRank_k = \alpha + \beta \times DAP_WeightRank_k + \epsilon_k,$$

where $NN_ImportanceRank_k$ refers to the rank of the k^{th} anomaly in neural network estimations (NN1) based on the relative importance of this anomaly in variable gradient analysis and $DAP_WeightRank_k$ is the rank of the anomaly based on its investment weights in the DAP portfolios. More explicitly, in variable gradient analysis, we first calculate k^{th} anomaly's gradient as follows,

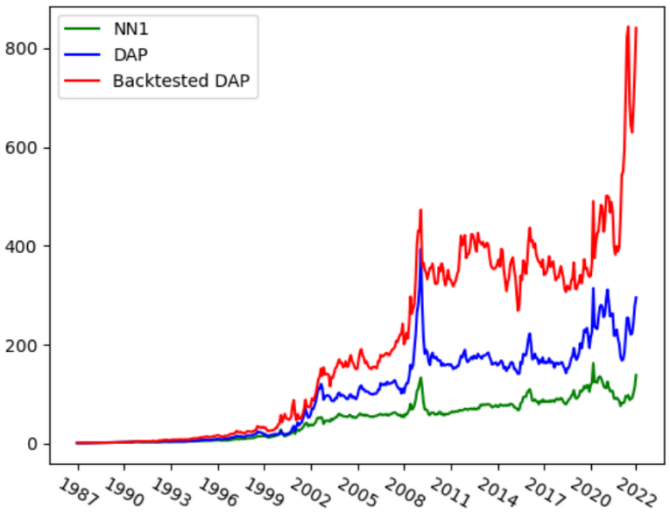
$$Grad_k = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{\partial R_{i,t+1}^{pred,NN}}{\partial x_{i,k,t}}$$

where $R_{i,t+1}^{pred,NN}$ is the expected return for stock i on time $t+1$ by NN1 and $x_{i,k,t}$ is the k^{th} characteristic for stock i on time t. We then rank anomalies by the absolute value of their gradients in each cross-section and obtain their overall rank over the testing period. For DAPs, we first rank anomalies in each cross-section based on their average investment weights in DAP portfolios and then obtain their overall rank over the testing period. Panel B replaces $NN_ImportanceRank_k$ with the average absolute value of the gradient for the k^{th} anomaly and $DAP_WeightRank_k$ with its average value of investment weights in DAP portfolios. This analysis excludes 20% of microcap stocks. The testing period is from 1987 to 2022, with a rolling training window of 60 or 240 months. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

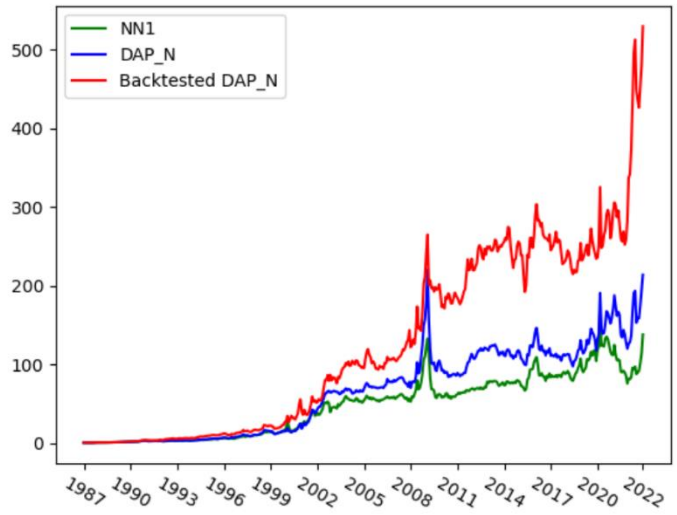
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	60-month Training Period				240-month Training Period			
Panel A: NN1 Importance Ranks Explained by DAP Weight Ranks								
Base DAP	0.953***				0.751***			
	(38.61)				(13.99)			
Backtested DAP		0.952***				0.750***		
		(38.34)				(13.93)		
DAP_N			0.936***				0.712***	
			(32.56)				(12.46)	
Backtested DAP_N				0.939***				0.723***
				(33.53)				(12.86)
Constant	3.627*	3.674*	4.960*	4.703*	19.148***	19.256***	22.173***	21.327***
	(1.66)	(1.67)	(1.94)	(1.89)	(4.02)	(4.03)	(4.37)	(4.27)
Adj R2	0.908	0.907	0.875	0.882	0.564	0.562	0.507	0.523
Observations	432	432	432	432	153	153	153	153
Panel B: NN1 Gradients Explained by DAP Weights								
Base DAP	0.855***				0.660***			
	(38.50)				(13.86)			
Backtested DAP		1.136***				0.908***		
		(38.72)				(13.77)		
DAP_N			0.837***				0.633***	
			(32.66)				(12.48)	
Backtested DAP_N				1.114***				0.882***
				(33.69)				(12.91)
Constant	0.001***	-0.001***	0.001***	-0.001***	0.002***	0.001	0.002***	0.001
	(5.99)	(-4.39)	(5.79)	(-3.28)	(6.24)	(1.29)	(6.32)	(1.60)
Adj R2	0.808	0.808	0.776	0.783	0.46	0.457	0.408	0.425
Observations	432	432	432	432	153	153	153	153

Figure 2: Cumulative Returns of DAP and Neural Network Strategies

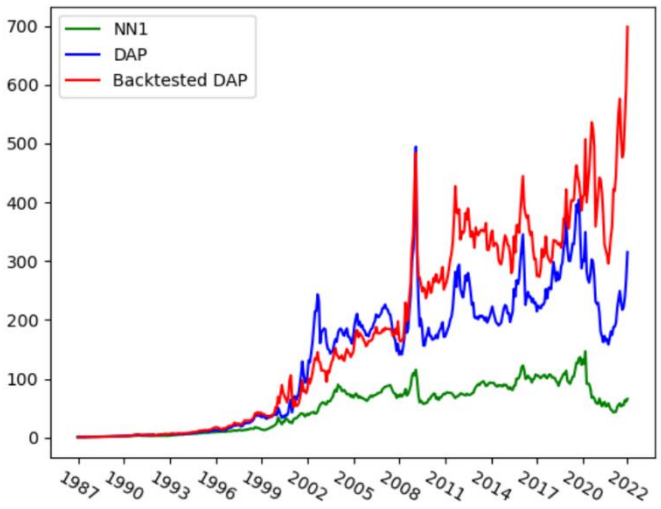
This figure plots the cumulative returns of the High-minus-Low (H-L) portfolios constructed by DAP and NN1 from January 1987 to December 2022. We set January 1987 as the baseline with a value of 1.



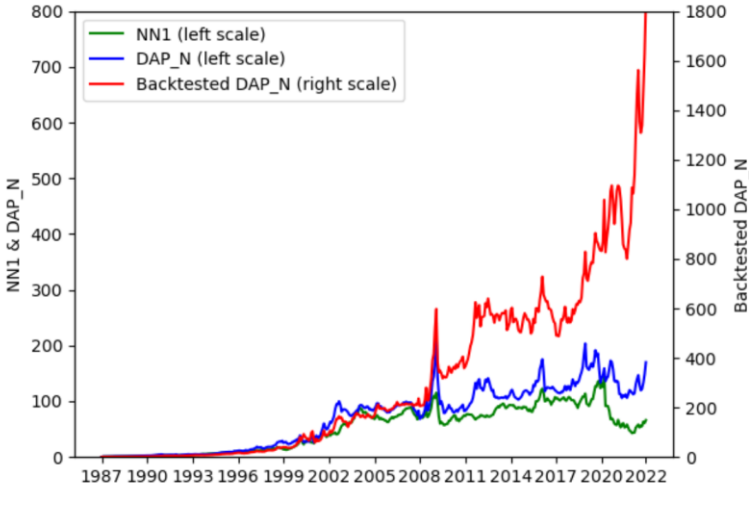
Panel A: Backtested DAP and NN1 (Exclude Small Stocks)



Panel B: Backtested DAP_N and NN1 (Exclude Small Stocks)



Panel C: Backtested DAP and NN1 (Whole Sample)



Panel D: Backtested DAP_N and NN1 (Whole Sample)

Figure 3: Comparing the Importance of Anomalies (NN1 vs. DAPs)

This figure plots the relationship between the importance rank of anomalies in NN1 and the weight rank of anomalies in DAPs. We first use variable gradient analysis to calculate the gradient of anomalies. We then rank anomalies by the absolute value of their gradients in each cross-section and obtain their overall rank over the testing period (labeled *NN Importance Rank*). For DAPs, we first rank anomalies in each cross-section based on their average investment weights in DAP portfolios and then obtain their overall rank over the testing period (labeled *DAP Weight Rank*). This analysis excludes 20% of microcap stocks. The testing period is from 1987-2022, with a 60-month rolling training window.

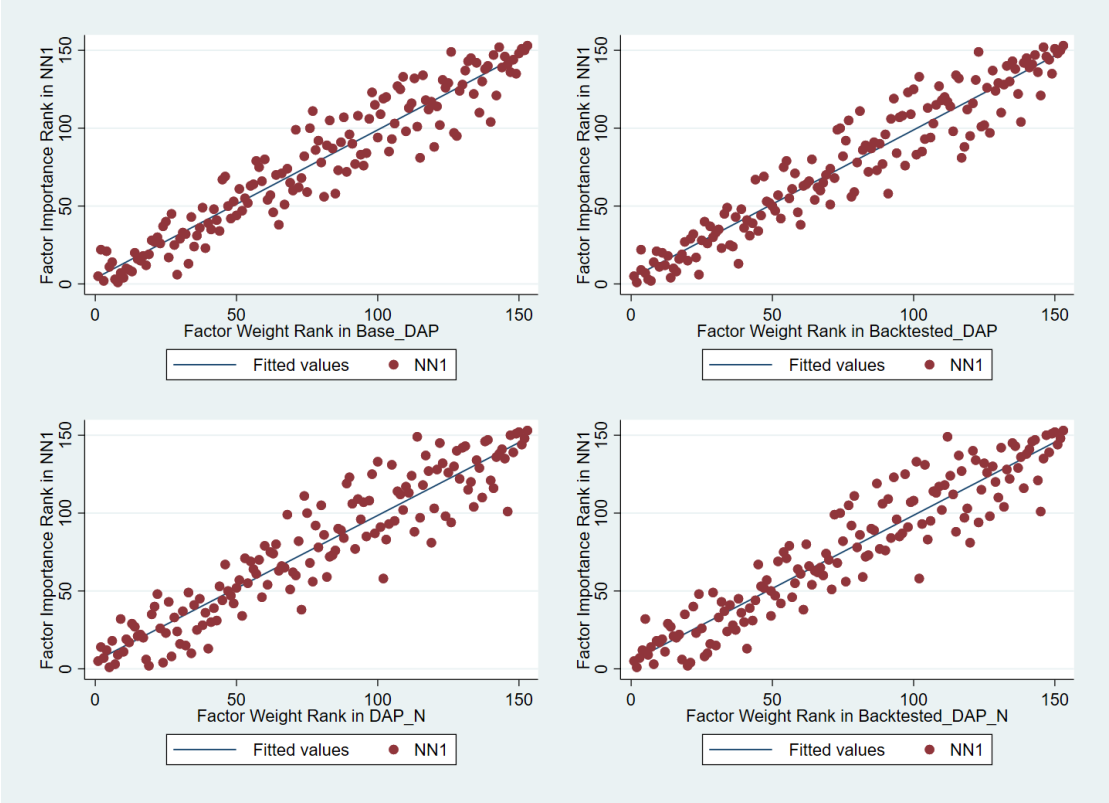
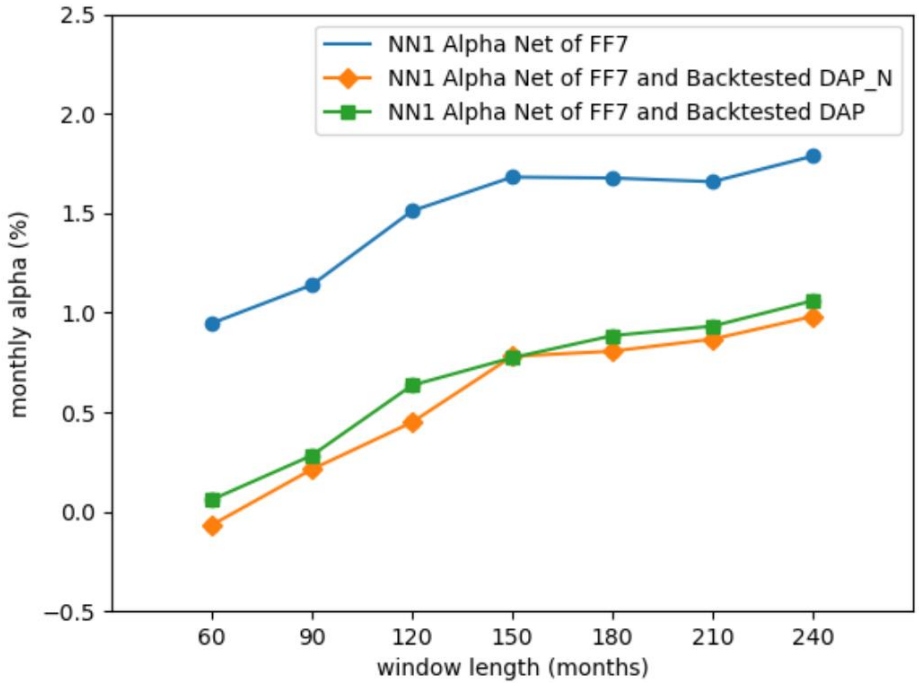


Figure 4: Explaining Neural Network Returns Across Different Training Windows（1977 – 2022）

This figure plots various returns of the H-L portfolios selected by neural networks across various lengths of training windows (i.e., 60, 90, 120,…, 240 months). The line "NN1 Alpha Net of FF7" plots the FF7-adjusted returns for the NN1 selected H-L portfolios over the testing period from January 1977 to December 2022. In Panel A, the top line ("NN1 Alpha Net of FF7") plots the monthly Fama-French seven factor-adjusted returns. The bottom two lines ("NN1 Alpha Net of FF7 and Backtested DAP" and "NN1 Alpha Net of FF7 and Backtested DAP_N") plot the NN1 alpha when backtested DAP and backtested DAP_N H-L portfolio returns are also included as an additional factor on top of the seven factors. Panel B tabulates the magnitude of these alphas as well the magnitude of NN1 alpha explained by the two DAP models, (which are 0.827% and 0.911%, respectively, in the last two rows).

Panel A: The Plot of NN1 Alphas



Panel B: The Value of NN1 Alphas

Monthly NN1 Alphas (in %)	Estimation Window (months)						
	60	90	120	150	180	210	240
NN2 Alphas Net of FF7	0.95	1.14	1.51	1.68	1.68	1.66	1.79
	(4.11)	(5.30)	(7.18)	(8.21)	(8.78)	(8.70)	(9.52)
Net of FF7 and Backtested DAPs	0.06	0.28	0.64	0.77	0.88	0.93	1.06
	(0.28)	(1.33)	(3.05)	(3.92)	(4.75)	(4.89)	(5.65)
Net of FF7 and BN Backtested DAPs	-0.07	0.21	0.45	0.78	0.81	0.87	0.98
	(-0.31)	(1.01)	(2.25)	(3.96)	(4.30)	(4.58)	(5.25)
Explained by Backtested DAPs	0.89	0.86	0.87	0.91	0.8	0.73	0.73
Explained by Backtested DAP_Ns	1.02	0.93	1.06	0.9	0.87	0.79	0.81

Online Appendix

Machine Learning as Arbitrage:
Can Economics Help Explain AI?

In this Online Appendix, we first provide details of our analytical framework. We then tabulate a list of additional tables that are discussed in the main text.

Part I. Online Appendix for the Analytical Framework

This section lays out the framework of our analysis. We first introduce the conceptual framework of our economically motivated dynamic strategy. We then discuss how we estimate it for empirical analysis and compare it with neural networks.

A. Traditional Anomaly Portfolios

Let $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})'$ be stock i's rank vector across K characteristics. For instance, if the k^{th} factor is BM, then $z_{ik} \in [0,1]$ represent the normalized rank of the i^{th} stock for BM in the cross-section. The traditional anomaly strategy directly forms the anomaly portfolio based on z_{ik}. For instance, it typically longs the stocks when $z_{ik} > 0.9$ and short the stock when $z_{ik} < 0.1$. Hence, if we denote $w_i = (w_{i1}, w_{i2}, \dots, w_{iK})'$ be stock i's investment weight in the multitude of factors, then we can write the stock's weight in k^{th} factor as $w_{ik} = c_k \times (I\{z_{ik} > 0.9\} - I\{z_{ik} < 0.1\})$, where $I\{.\}$ Denotes an indication function and c_k is the normalization parameter to construct the long-short portfolio.

B. Neural Networks

Following Gu, Kelly, and Xiu (2020), we adopt the *ReLU* activation function for Neural Networks:

$$Relu(x) = \begin{cases} x, \ if \ x \geq 0 \\ 0, \ otherwise \end{cases}. \qquad \text{(A1)}$$

Let $K^{(l)}$ denote the number of neurons in each layer $l = 1, \dots, L$. Define the output of neuron k in layer l as x_k^l and the outputs for this layer as $x^{(l)} = (1, x_1^l, \dots, x_{K^{(l)}}^l)$. The initial value of the input layer is $x^0 = (1, z_1, \dots, z_N)$, where $z_1, \dots, z_N$ are the N initial predictors. Let $\theta_k^{(l)}$ denote the weight vector that links $x^{(l-1)}$ to the x_k^l. Then the Neural Network follows the following recursive output formula:

$$x_k^l = ReLU\left(x^{(l-1)'}\theta_k^{(l-1)}\right), \qquad \text{(A2)}$$

with the final output

$$g(z;\theta) = x^{(L-1)'}\theta_k^{(L-1)}. \qquad \text{(A3)}$$

We can apply the Neural Network methods to characteristics. Based on the initial input of $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})'$ as defined above, neural network algorithms will update according to equation (A2) and ultimately give the final signal g_k for each stock as the predicted returns.

We train the model using all the characteristics and stock returns available during the training period. In particular, we employed a gradient-based approach to train the neural network model. There are various training strategies for neural networks, and a common solution is to utilize the Adam optimizer. To enhance the optimization speed and performance of the model, the Adam optimizer randomly selects a subset of samples (batch) from the training data for gradient updates in each iteration. A key parameter of the Adam optimizer is the learning rate, which dictates the step size for updates along the gradient direction. A well-chosen learning rate involves a trade-off between convergence speed and avoiding overshooting. Thus, it is essential to dynamically adjust the learning rate based on the training process's state. Therefore, we implemented a learning rate scheduler during training. A learning rate scheduler is a predefined framework that modifies the learning rate between epochs or iterations as the training advances. In this context, we employed a learning rate decay strategy, gradually reducing the learning rate as the training progresses.

Neural networks often exhibit strong expressive power and the ability to fit any arbitrary function, but they are also susceptible to the issue of overfitting. Overfitting occurs when a neural network performs well on the training data but poorly on unseen test data. Previous research generally attributes overfitting to the model memorizing the noise and details of the training data excessively while neglecting the overall distribution of the data, resulting in a decrease in the model's generalization ability. To mitigate overfitting, we employed EarlyStopping. EarlyStopping is a regularization technique in model training. If the model's performance on the validation dataset does not improve consistently, training is halted to prevent the model from excessively fitting the training data.

Gu, Kelly, and Xiu (2020) have also adopted *Batch Normalization* (Ioffe and Szegedy 2015) for neural networks and a long period to validate the estimation. In particular, for each hidden unit in each training step (a "batch"), the *Batch Normalization* process cross-sectionally demeans and variance standardizes the batch inputs to restore the representation power of the unit. Given that our estimation spans a wide range of lengths for training periods, we do not adopt these empirical specifications. Instead, we always train neural network algorithms over the training window as specified above and use one more month following the training period (and prior to the out-of-sample testing period) to validate the estimated parameters of neural networks. The out-of-sample predicting power of our estimation process is on par with that of Gu, Kelly, and Xiu (2020).

The training process gives the optimal signal g_k for each stock as the predicted returns. We then sort stocks into deciles according to their g_k values to form the neural network-predicted portfolios and estimate the *value-weighted* out-of-sample returns generated by the stocks within each decile. We then take the long (short) positive of the deciles with the highest (lowest) predicted returns to form our baseline testing portfolios (i.e., the high-minus-low or "H-L" portfolio). All decile portfolios are rebalanced monthly till the end of the testing period.

C. Dynamic Arbitrage Portfolios (DAPs) and Back testing

We now provide the theoretical framework for DAPs. Below, we first present arbitrageurs' trading strategy in the APT framework (e.g., Ingersoll 1984 and Da, Nagel, and Xiu 2023) and then discuss the strategy in a more dynamic setting in the spirit of Kyle (1985) and Admati (1985).

To understand how arbitrageurs can potentially trade anomalies, we can view high-minus-low portfolios of characteristics, which we refer to as "anomalies" when there is no confusion, as synthetic assets in an APT framework. In particular, we assume that anomalies deliver returns in the following linear pricing framework (see, Ingersoll 1984 and Da, Nagel, and Xiu 2023):

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta} \times (\mu + v_t) + \boldsymbol{\epsilon}_t \qquad (A4)$$

where $\boldsymbol{y}_t$ is the $K \times 1$ vector of anomaly returns in a given period t (with its k-th element representing the return of the k-th anomaly), $\boldsymbol{\alpha}$ is the $K \times 1$ vector of pricing errors (i.e., alphas), $\boldsymbol{\beta}$ refers to the exposure of the anomaly to the systematic risk, μ is the risk premium, v_t is the innovation of systematic risk, and $\boldsymbol{\epsilon}_t$ is the $K \times 1$ vector of idiosyncratic shocks to anomalies, which are independent of factors. In an economy with N factors, μ and v_t denote the vector of N and $\boldsymbol{\beta}$ is a matrix of $K \times N$. Further, assume that $E[v] = 0$, $E[\boldsymbol{\epsilon}_k] = 0$, $E[vv'] = \mathbb{I}$, and the covariance matrix of $\boldsymbol{\epsilon}_k$ is $\boldsymbol{\Omega}$.

Ingersoll (1984) shows that, under these assumptions, the no-arbitrage condition holds when the weighted sum of the squared pricing errors is uniformly bounded. Mathematically, this means the existence of a positive number V such that $\boldsymbol{\alpha}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}' \leq V < \infty$. In contrast, if asymptotic arbitrage opportunities do exist, then a general arbitrage portfolio can be constructed to deliver positive expected returns without loading on risk.

Important, Ingersoll (1984) points out that the weight of the arbitrage portfolio is determined by $\boldsymbol{w} = \boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$. In our setting, the arbitrage portfolio invests in anomalies in proportion to the pricing errors—or alphas that can be delivered by these anomalies. This property provides an APT-based foundation for our alpha-weighted DAP strategy.

One advantage of the above APT approach is that the arbitrage portfolio is "preference-free". That is, regardless of the preferences of arbitrageurs, they will adopt the above alpha-weighted arbitrage portfolio. But as pointed out by Ingersoll (1984), APT is also consistent with CAPM-type asset pricing models. Hence, we can also view the above alpha-weighted arbitrage as an optimal strategy from the perspective of a CAPM-type investor with mean-variance preference.

To see this intuition, we assume that arbitrate opportunities exist in an economy with one factor. An investor observing the information set of (A4) constructs an anomaly-based arbitrage portfolio to maximize her mean-variance utility function: $Max_{\boldsymbol{w_t}}\ \boldsymbol{w}'\overline{\boldsymbol{y}}_t - \frac{1}{2}\gamma\, \boldsymbol{w}'\Sigma_y \boldsymbol{w}$, subject to $\boldsymbol{w}'\mathbb{I} = 0$ and $\boldsymbol{w}'\boldsymbol{\beta} = 0$. Here, $\boldsymbol{w} = [w_1, \ldots, w_K]'$ is the vector of investment weights in the K-anomalies (we omit the subscript t for easy notation), $\overline{\boldsymbol{y}}_t$ is the vector of expected anomaly returns ($\overline{\boldsymbol{y}}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\mu$ based on the information set of the investor, with μ being the premium of the systematic risk), $\mathbb{I}$ is

the vector of ones, $\boldsymbol{\beta}$ refers to the vector of systematic risk exposure of anomalies, γ is the risk aversion parameter, and Σ_y is the covariance matrix of the anomaly returns. The two constraints, $\boldsymbol{w}'\mathbb{1} = 0$ and $\boldsymbol{w}'\boldsymbol{\beta} = 0$, mean to form zero-cost portfolio with zero systematic risk exposure. The optimal portfolio is $\boldsymbol{w}^* = \frac{1}{\gamma}\Sigma_y^{-1}(\overline{\boldsymbol{y}}_t + \lambda\mathbb{1} + \theta\boldsymbol{\beta})$, where $\lambda = \frac{ED-FB}{\Delta}$ and $\theta = \frac{BD-AE}{\Delta}$ are two constants in the Lagrangian representing the two constraints. In addition, $A = \mathbb{1}'\Sigma_y^{-1}\mathbb{1}$, $B = \mathbb{1}'\Sigma_y^{-1}\overline{\boldsymbol{y}}_t$, $D = \mathbb{1}'\Sigma_y^{-1}\boldsymbol{\beta}$, $E = \boldsymbol{\beta}'\Sigma_y^{-1}\overline{\boldsymbol{y}}_t$, $F = \boldsymbol{\beta}\Sigma_y^{-1}\boldsymbol{\beta}$ are all constants; $\Delta = AF - D^2 > 0$ due to the Cauchy-Schwarz inequality when Σ_y is nonsingular and all assets do not have the same mean. Since $\overline{\boldsymbol{y}}_t \propto \boldsymbol{\alpha}$, the optimal weight is in proportion to alphas (i.e., $\boldsymbol{w}^* \propto \boldsymbol{\alpha}$), which gives rise to an alpha-weighted optimal strategy.

We lastly show that the above intuition can also be extended to a dynamic setup based on Kyle (1985) and, particularly, Admati (1985). Assume that the returns of the k^{th} anomaly strategy, which is the high-minus-low portfolio of the k^{th} characteristic of firms ($k \in [1, K]$), satisfy the following conditional pricing framework:

$$y_{kt} = \alpha_{kt} + \beta_{kt} \times X_t + \epsilon_{kt} \qquad (A4')$$

where y_{kt} and α_{kt} refer to the return of the k^{th} anomaly and its conditional alpha in period t, and X_t refers to the systematic factor.

Next, assume that alphas are conditional and depend on a hidden state variable F_t:

$$\alpha_{kt} = \alpha_k \times F_t + \eta_{kt}, \qquad (A5)$$

and that the state variable follows an AR(1) process:

$$F_t = \Gamma \times F_{t-1} + \delta_t, \qquad (A6)$$

where $\Gamma > 0$ is a positive constant, $\eta_{kt} \sim N(0, \sigma_\eta^2)$ and $\delta_t \sim N(0, \sigma_\delta^2)$ are two i.i.d random noises.

Equations (A5) and (A6) lay out the framework for the state-contingent alphas of anomalies that generalize the example provided in the main text. The state variable F_t can be intuitively thought of as the indicator for high-alpha and low-alpha states.

Now assume that some investors can observe the conditional alphas, and thus become informed in the spirit of Kyle (1985) and Admati (1985). The remaining investors are noise traders. More explicitly, β_{kt} and X_t are common knowledge to both types of investors. But informed investors observe the signal of α_{kt-1} at the beginning of period t, from which they can infer the expected conditional premium as $E[\alpha_{kt}|t-1] = \Gamma\alpha_{kt-1} + e_{kt-1}$, where $e_{kt-1} = -\Gamma\eta_{kt-1}$ is the noise term. Noise traders do not observe this signal or beware of the existence of conditional alphas. Based on these assumptions, we can quantify the optimal investment strategy of informed investors as follows.

Lemma 1 (Dynamic Arbitrage Portfolio): *Denote w_{kt} the fraction of the total capacity invested in the k^{th} anomaly in period t with its normalization condition holds:*

$\sum_k w_{kt} = 1$. *We have* $w_{kt} = c_t \alpha_{kt-1}$, *where* $c_t = (\sum_k \alpha_{kt-1})^{-1}$ *is a normalization coefficient.*

Proof: The optimization problem of an informed trader resembles that of a fund manager receiving signals about conditional alpha in Mamaysky, Spiegel, and Zhang (2008). The proposition holds under the general setting in Admati (1985) following Mamaysky, Spiegel, and Zhang (2008). QED.

The above lemma suggests that, similar to APT arbitragers, informed investors in dynamic trading also optimally set up their portfolio weights in proportion to conditional alphas. The intuition is as follows. In informed trading models such as Kyle (1985) and Admati (1985), informed traders set up their optimal trading weights in proportion to the value of the private signals that they observe. In our setup, conditional alpha is the value of the private signals. Therefore, informed traders set up their trading weights in proportion to conditional alphas.

Hence, across APT (Ingersoll 1984) and informed trading framework (Admati 1985), arbitrage portfolios are featured with weights in proportion to alphas. Without the loss of generality, we adopt the alpha-weighted arbitrage strategy to construct dynamic arbitrage portfolios (DAPs) as our benchmark to interpret the performance of machine learning algorithms.

Our empirical DAP strategies also feature two properties. First, the payoffs of DAP portfolios are a quadratic function of alphas, as summarized below.

Property 1 (DAP Returns): *The optimal DAPs can generate positive market-adjusted returns that follow a quadratic function of conditional alphas.*

Proof: The expected return delivered by the optimal DAP is

$$\begin{aligned} y_{Pt} &= \sum_k w_{kt} \times y_{kt} = \sum_k c_t \alpha_{kt-1} \times y_{kt} \\ &= \sum_k c_t \alpha_{kt-1} \times \alpha_{kt} + \sum_k c_t \alpha_{kt-1} \times \beta_{kt} X_t + \sum_k c_t \alpha_{kt-1} \times \epsilon_{kt} \\ &\equiv \alpha_{Pt} + \beta_{Pt} \times X_t + \epsilon_{Pt}. \end{aligned}$$

Hence, the expected *market-adjusted* returns of the above DAP becomes $\alpha_{Pt} = \sum_k c_t \alpha_{kt-1} \times \alpha_{kt} = \sum_k c_t \Gamma \alpha_{kt-1}^2$. Easy to see, $\alpha_{Pt} > 0$ if and only if $\Gamma > 0$. QED.

Next, we adopt backtesting when implementing the DAP trading strategy. Back testing does not affect the analytical solution of the optimal trading strategy. Rather, it aims to screen out false-positive signals from empirically estimated conditional alphas. For easy notation, we ignore the time label and use $\boldsymbol{A} = (\alpha_1, \ldots, \alpha_K)'$ to denote the vector of estimated conditional alpha in a given state, where the k^{th} element of the vector (i.e., α_k) denotes the estimated conditional alpha of the k^{th} characteristic. Without the loss of generality, we require $\alpha_k > 0$. In the empirical estimation, if α_n turns out to be negative, we can redefine $\hat{\alpha}_k$ and $\hat{z}_{ik}$ as $\hat{\alpha}_k = -\alpha_k$ and $\hat{z}_{ik} = 1 - z_{ik}$ when $\alpha_n < 0$

(and $\hat{\alpha}_k = \alpha_k$ and $\hat{z}_{ik} = z_{ik}$ when $\alpha_k \geq 0$). This transformation ensures all conditional alpha values are positive.

Using this notation, the investment weight of the DAP in stock i can be denoted as (when ignoring the scaling constant):

$$w_{i,DAP} = z_i' \boldsymbol{A},$$

where $w_{i,DAP} = \left(w_{i1,DAP}, \dots, w_{iK,DAP}\right)'$ is the vector of the investment weight of the stock in characteristics with $w_{ik,DAP} = z_{ik}\alpha_k$.

Denote $\alpha_{k,-t}$ as the most recent realization of alpha. The backtest requires that the conditional alpha estimated during the rolling training period and its most recent realization must have the same sign. Since $\alpha_k > 0$, our backtest essentially quires $\alpha_{k,-t} > 0$. The following property suggests that the back testing request is equivalent to imposing a *ReLU* activation function on the original DAP investment weight.

Property 2 (Back testing and the ReLU Activation Function): *The back testing process imposes* the following *ReLU* activation function on the original DAP investment weight:

$$w_{i,DAP} = ReLU\left(z_i' H\left(Diag(\boldsymbol{A}_{-t})\right)\boldsymbol{A}\right), \qquad (A7)$$

where $A_{-t} = \left(\alpha_{1,-t}, \dots, \alpha_{K,-t}\right)'$ and $Diag(\boldsymbol{A}_{-t})$ be the $N \times N$ diagonal matrix of $\boldsymbol{A}_{-t}$

Proof: Let $H(\alpha_{k,-t})$ be the sign function of $\alpha_{k,-t}$, where

$$H(x) = \begin{cases} 1, \ if \ x \geq 0 \\ -1, \ otherwise' \end{cases}$$

the back test imposes the following *ReLU* activation function for the investment weight of stock i in the k^{th} characterister:

$$w_{ik,DAP} = z_{ik} ReLU\left(\alpha_k \times H(\alpha_{k,-t})\right). \qquad (A8)$$

With this notion, we can rewrite the back tested DAP signal for the i^{th} stock as:

$$\begin{aligned} w_{i,DAP} &= z_i' ReLU\left(A' \times H\left(Diag(A_{-t})\right)\right) \\ &= ReLU\left(z_i'\left(A' \times H\left(Diag(A_{-t})\right)\right)\right) \\ &= ReLU\left(z_i' H\left(Diag(A_{-t})\right)A\right), \end{aligned}$$

where in the second step we moved z_i' into the ReLU function because $z_{ik}' > 0$. QED.

D. The Statistical Power of Back test in Predicting Anomaly Returns

Our main text demonstrates that Backtested DAP typically delivers a higher alpha than the baseline DAP. In this section, we further validate the statistical power of back testing in predicting anomaly returns.

We first examine how accurate the back testing process is in predicting anomaly returns. To achieve this goal, we first estimate, for each anomaly k in a given month t during the out-of-sample performance test period from January 1987 to December 2022, the 60-month rolling anomaly alpha, α_{kt-1}, and the most recent realization of alpha, $\alpha_{k,-t}$ of the anomaly as specified in our main text. However, instead of applying DAP and back test to construct the portfolio of anomalies, here we want to compare these two alpha values to the realized anomaly alphas in month t, α_{kt}. In particular, we want to understand how accurate the back testing process (which is based on the signs of α_{kt-1} and $\alpha_{k,-t}$) is in predicting the sign of α_{kt}.

Table IN10 tabulates the results. In Columns (1) and (2), we split the sample of all anomaly-month observations into two groups based on the outcomes of back testing (i.e., to keep or to drop the anomaly in DAP portfolios). More specifically, conditional on $\alpha_{kt-1} > 0$, the back testing process keeps an anomaly for the month-t DAP portfolio when $\alpha_{k,-t} > 0$ (i.e., $H(\alpha_{kt-1}) = H(\alpha_{k,-t})$, where $H(x)$ denotes the sign function) and masks it otherwise. Column (2) reports that the back testing process on average drops 39.76% anomalies. The remaining 60.24% anomalies are used to construct Back tested DAP portfolios.

Next, we assess the accuracy of such selection when compared to the realized anomaly alphas in month t, α_{kt}. Columns (3) and (4) further split each of the two back testing groups into two subgroups depending on whether the realized anomaly alpha in month t has a positive sign or not (i.e., whether the sign of the realized alpha aligns with the selection of the back testing process).

Take the group of anomalies selected by back testing (60.24%) first. Column (4) reports that while most selected anomalies represent "good selections" as they deliver positive alphas, back testing may still make "wrong selections" on anomalies that fail to deliver positive alpha. The two subgroups of good selections and wrong selections account for 37.88% and 22.36% of all anomaly-month observations. It is important to note that the success rate (i.e., the 37.88% good selections) is more than 50% higher than the failure rate (i.e., the 22.36% wrong selections), confirming that the back testing helps select "good" anomalies that can deliver positive alpha. Column (4) reports a similar split for the group of anomalies dropped by back testing (39.76%), when back testing may drop anomalies properly or incorrectly. These two subgroups account for 19.48% and 20.27% of all anomaly-month observations.

Columns (5) and (6) synchronize the statistical interpretations of the four subgroups, including:

1) correct predictions made by back testing (i.e., the back testing process has properly selected "good" anomalies that can realize positive alphas in the testing period or properly rejected "wrong" anomalies with negative future alphas);
2) Type I Error of back testing (false positive, i.e., the back testing process selects "wrong" anomalies that fail to deliver positive alpha); and
3) Type II Error of back testing (false negative, i.e., the back testing process rejects "good" anomalies).

Column (6) reports that the back testing process can make proper predictions and selection 57.37% of the time, indicating a superior predictive power of back testing on anomaly returns.

To further illustrate how the predictive power of back testing transforms into performance, we next consider a simulation analysis to nail down the statistical power of the alpha delivered by Backted DAP. To achieve this goal, we construct simulated DAP portfolios with Pseudo-Back testing as follows. In each simulation, we follow the construction of Back tested DAP in Table 1, except that we randomly drop 40% of the anomalies each month (instead of applying the real back testing). This dropping rate is on par with the real dropping rate of the back testing process. We then calculate the seven-factor adjusted performance (i.e., FF7 Alpha) and its t-statistics of the time-series returns delivered by the DAP portfolios with Pseudo-Back testing in the test period from January 1987 to December 2022. After repeating the simulation 1000 times, we obtain a stable distribution of DAP alphas with Pseudo-Back testing.

Figure IN3 reports the distribution of simulated monthly alphas (Panel A) and corresponding t-statistics (Panel B). In each figure, the dashed and dotted lines indicate the positions of the statistics from our original DAP and Backtested DAP portfolios as reported in Table 1. We observe that a random drop of 40% anomalies in general hurt the performance of DAPs. For instance, the simulated Pseudo-Back testing DAPs deliver a seven-factor adjusted monthly alpha of 1.02%, below our original DAP alpha (1.145%). Moreover, our original DAP alpha has a p-value of 11.5% when compared to the distribution of the simulated alphas, suggesting that randomly dropping anomalies marginally reduces the DAP performance.

More importantly, the Backtest DAP portfolios can deliver superior performance that is difficult to achieve by Pseudo-Back testing DAPs. Indeed, the Backtested DAP alpha exhibits a highly significant p-value of 0.00 when compared to the distribution of the simulated alphas. Panel B also reports a similar p-value for the t-statistics that can be generated by the Backtested DAP alphas. These observations suggest that the positive influence of the back test on anomaly portfolio returns cannot be practically achieved by randomly dropping a similar fraction of anomalies. Collectively, Table IN10 and

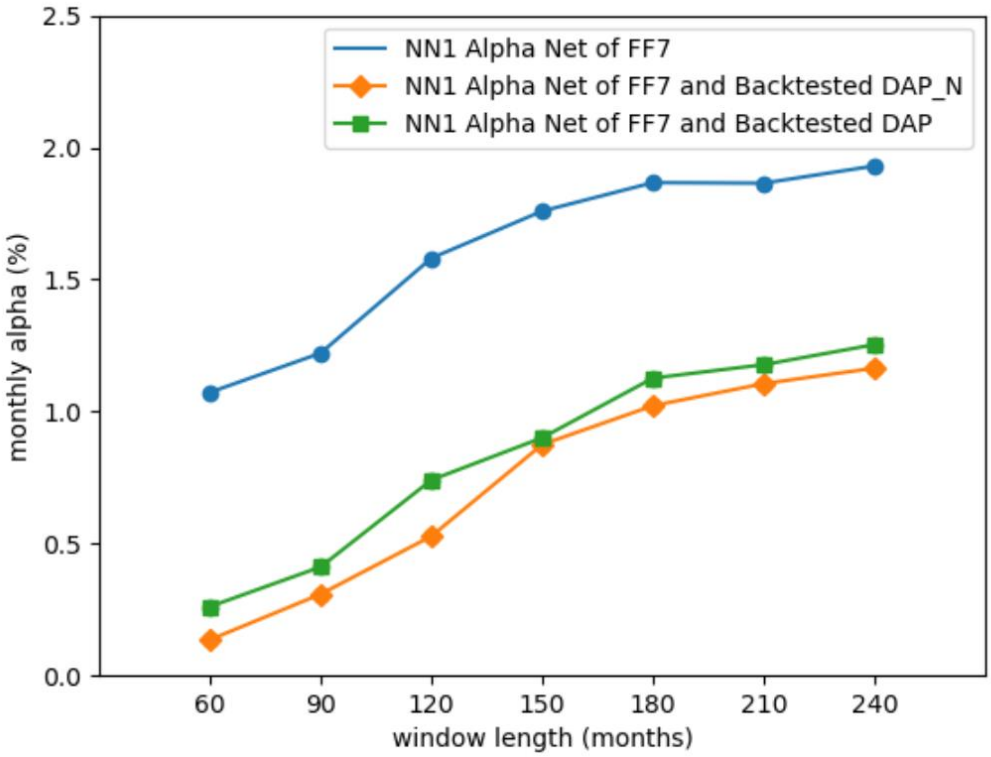
Figure IN3 suggest that our back test can help select the proper set of anomalies to enhance the performance of anomaly portfolios.

Part II. Online Appendix for Additional Empirical Analysis

Figure 1 Supplementary Table: Explaining Neural Network Returns

This figure reports how various factors affect the returns of H-L portfolios selected by neural networks with one hidden layer (NN1). The testing period is from January 1987 to December 2022. The x-axis denotes the length of the rolling training windows (i.e., 60, 90, 120, …, 240 months). In Panel A, the top line ("NN1 Alpha Net of FF7") plots the monthly Fama-French seven factor-adjusted returns. The bottom two lines ("NN1 Alpha Net of FF7 and Backtested DAP" and "NN1 Alpha Net of FF7 and Backtested DAP_N") plot the NN1 alpha when backtested DAP and BN backtested DAP H-L portfolio returns are also included as an additional factor on top of the seven factors. Panel B tabulates the magnitude of these alphas as well the magnitude of NN1 alpha explained by the two DAP models (which are 0.774% and 0.879%, respectively, in the last two rows).

Panel A: The Plot of NN1 Alphas



Panel B: The Value of NN1 Alphas

Monthly NN1 Alphas (in %)	Estimation Window (months)						
	60	90	120	150	180	210	240
NN1 Alphas Net of FF7	1.07	1.22	1.58	1.76	1.87	1.86	1.93
	(3.95)	(4.71)	(6.20)	(7.04)	(8.05)	(8.12)	(8.45)
Net of FF7 and Backtested DAP	0.26	0.41	0.74	0.90	1.13	1.18	1.25
	(1.00)	(1.63)	(2.98)	(3.79)	(5.04)	(5.21)	(5.53)
Net of FF7 and Backtested DAP_N	0.14	0.31	0.53	0.87	1.02	1.11	1.16
	(0.52)	(1.23)	(2.21)	(3.70)	(4.56)	(4.93)	(5.14)
Explained by Backtested DAP	0.81	0.81	0.84	0.86	0.74	0.68	0.68
Explained by Backtested DAP_N	0.93	0.91	1.05	0.89	0.85	0.75	0.77

Table IN1. Out-of-sample Performance Delivered by DAP Portfolios (60 months and 240 months rolling window)

This table reports the returns generated by DAPs with a rolling window of 60 or 240 months. For each DAP strategy, we tabulate the returns for the Low and High deciles, as well as risk-adjusted returns (i.e., alphas) for the High-minus-Low portfolio. The OLS standard error is used to construct t-stats. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Exclude Small Stocks (20%)						Whole Sample					
	Low	High	H-L	FF3	FF5	FF7	Low	High	H-L	FF3	FF5	FF7
Panel A. Returns Generated by DAP Portfolios (60 months training period)												
Base DAP	-0.703	0.871***	1.574***	2.128***	1.467***	0.999***	-0.711	0.837***	1.549***	2.165***	1.486***	0.912***
	(-1.55)	(3.99)	(4.14)	(6.32)	(4.44)	(3.64)	(-1.46)	(3.65)	(3.76)	(5.86)	(4.05)	(3.19)
Backtest DAP	-0.736*	1.098***	1.834***	2.311***	1.783***	1.746***	-0.868*	1.042***	1.910***	2.438***	1.840***	1.768***
	(-1.70)	(4.57)	(4.95)	(6.86)	(5.39)	(6.98)	(-1.88)	(4.06)	(4.73)	(6.57)	(4.97)	(6.43)
BN DAP	-0.539	0.906***	1.444***	1.926***	1.398***	1.018***	-0.457	0.847***	1.304***	1.847***	1.262***	0.787***
	(-1.31)	(4.26)	(4.41)	(6.65)	(4.97)	(4.26)	(-1.02)	(3.74)	(3.56)	(5.60)	(3.84)	(2.92)
BN Backtest	-0.532	1.091***	1.623***	2.066***	1.581***	1.576***	-0.860**	1.053***	1.913***	2.398***	1.960***	1.906***
	(-1.34)	(4.88)	(5.08)	(7.19)	(5.66)	(7.35)	(-2.02)	(4.22)	(5.33)	(7.23)	(5.87)	(7.85)
Panel B. Returns Generated by DAP Portfolios (240 months training period)												
Base DAP	-0.755	0.906***	1.660***	2.201***	1.338***	0.961***	-0.854*	0.913***	1.767***	2.273***	1.361***	0.978***
	(-1.58)	(4.70)	(4.31)	(7.90)	(5.50)	(5.02)	(-1.73)	(4.74)	(4.36)	(7.65)	(5.24)	(4.65)
Backtest DAP	-0.913**	1.066***	1.979***	2.438***	1.771***	1.785***	-1.035**	1.238***	2.273***	2.736***	1.984***	2.025***
	(-2.10)	(4.87)	(5.48)	(7.88)	(5.88)	(8.07)	(-2.22)	(5.30)	(5.81)	(7.99)	(5.98)	(7.91)
BN DAP	-0.676	0.977***	1.653***	2.124***	1.357***	1.060***	-0.783*	0.989***	1.772***	2.269***	1.433***	1.123***
	(-1.54)	(4.97)	(4.92)	(8.75)	(6.35)	(5.86)	(-1.71)	(5.05)	(5.01)	(8.73)	(6.32)	(5.83)
BN Backtest	-0.800*	1.127***	1.927***	2.347***	1.746***	1.795***	-0.917**	1.203***	2.120***	2.579***	1.909***	1.954***
	(-1.91)	(5.13)	(5.69)	(8.04)	(6.11)	(8.73)	(-2.08)	(5.27)	(5.90)	(8.18)	(6.23)	(8.47)

Table IN2. NN Returns Explained by FF7 and DAPs (60 months)

This table reports the results of the test in which we use the FF7 factors as well as the H-L DAP returns to explain the time series of H-L returns generated by the neural networks. In Panel A, the dependent variable is the H-L returns generated by NN1. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha with no DAP added. For columns 2-5, we add baseline DAP, Backtested DAP, DAP with BN normalized rolling alphas (DAP_BN), and backtested DAP_BN as explanatory variables sequentially. Panel B applies the same test to the H-L returns generated by NN3. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: NN1 Returns Explained by FF7 and DAPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)					Whole Sample				
Alpha	1.072***	0.330*	0.233	0.186	0.138	1.088***	0.561***	0.473*	0.417**	0.382
	(3.95)	(1.81)	(0.90)	(1.04)	(0.54)	(3.99)	(3.06)	(1.93)	(2.18)	(1.56)
SMB	-0.107	0.085	0.027	0.204***	0.051	0.004	-0.072	-0.033	0.097	0.023
	(-1.09)	(1.29)	(0.31)	(3.15)	(0.58)	(0.04)	(-1.10)	(-0.38)	(1.43)	(0.27)
HML	-0.198*	-0.034	-0.091	0.073	-0.005	0.018	0.055	0.118	0.012	0.071
	(-1.68)	(-0.43)	(-0.86)	(0.95)	(-0.05)	(0.15)	(0.70)	(1.13)	(0.15)	(0.69)
RMW	-0.134	-0.463***	-0.205*	-0.265***	-0.229**	-0.809***	-0.483***	-0.731***	-0.599***	-0.804***
	(-1.08)	(-5.54)	(-1.82)	(-3.29)	(-2.06)	(-6.45)	(-5.71)	(-6.63)	(-6.88)	(-7.38)
CMA	0.652***	-0.094	0.161	-0.296**	0.074	0.755***	-0.206*	0.024	-0.185	0.097
	(3.77)	(-0.79)	(0.99)	(-2.50)	(0.45)	(4.34)	(-1.68)	(0.14)	(-1.45)	(0.60)
MKT	-0.266***	-0.003	-0.210***	-0.009	-0.189***	-0.251***	-0.042	-0.210***	-0.011	-0.210***
	(-4.04)	(-0.07)	(-3.54)	(-0.20)	(-3.22)	(-3.80)	(-0.92)	(-3.61)	(-0.23)	(-3.64)
MOM	0.859***	0.213***	0.695***	0.242***	0.716***	0.644***	0.193***	0.591***	0.230***	0.553***
	(13.91)	(4.32)	(12.04)	(5.13)	(12.75)	(10.38)	(4.22)	(10.82)	(4.91)	(10.15)
SREV	-0.054	-0.026	0.526***	-0.022	0.555***	0.206**	0.065	0.674***	0.100*	0.673***
	(-0.65)	(-0.48)	(5.64)	(-0.41)	(6.05)	(2.49)	(1.18)	(8.06)	(1.74)	(8.20)
DAP		0.743***					0.740***			
		(23.41)					(23.10)			
BackTest DAP			0.481***					0.460***		
			(10.16)					(11.26)		
BN DAP				0.871***					0.860***	
				(24.43)					(21.57)	
BN BackTest DAP					0.593***					0.566***
					(10.90)					(11.74)
Adj R2	0.446	0.758	0.554	0.770	0.567	0.345	0.710	0.495	0.687	0.505
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: NN3 Returns Explained by FF7 and DAPs

Table IN3: Applying Pre-Publication Ratio to DAPs (240-mon)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks (20%)						Whole Sample			
Alpha	0.678**	-0.047	-0.158	-0.139	-0.072	0.926***	0.449**	0.434*	0.333	0.371
	(2.38)	(-0.23)	(-0.58)	(-0.64)	(-0.25)	(3.37)	(2.19)	(1.67)	(1.55)	(1.42)
SMB	-0.150	0.037	-0.016	0.137*	-0.023	0.223**	0.155**	0.194**	0.306***	0.239***
	(-1.45)	(0.50)	(-0.17)	(1.75)	(-0.23)	(2.25)	(2.09)	(2.11)	(3.98)	(2.60)
HML	-0.163	-0.003	-0.057	0.087	-0.008	-0.067	-0.033	0.013	-0.072	-0.025
	(-1.31)	(-0.03)	(-0.50)	(0.94)	(-0.07)	(-0.56)	(-0.38)	(0.12)	(-0.78)	(-0.23)
RMW	0.201	-0.120	0.131	0.080	0.125	-0.398***	-0.103	-0.336***	-0.213**	-0.394***
	(1.53)	(-1.27)	(1.10)	(0.82)	(1.02)	(-3.15)	(-1.08)	(-2.87)	(-2.17)	(-3.38)
CMA	0.757***	0.029	0.268	-0.117	0.293	0.900***	0.031	0.315*	0.069	0.383**
	(4.17)	(0.22)	(1.55)	(-0.82)	(1.63)	(5.14)	(0.23)	(1.79)	(0.48)	(2.22)
MKT	-0.193***	0.064	-0.137**	0.045	-0.131**	-0.184***	0.005	-0.151**	0.028	-0.152**
	(-2.79)	(1.26)	(-2.17)	(0.85)	(-2.01)	(-2.77)	(0.11)	(-2.45)	(0.53)	(-2.46)
MOM	0.866***	0.235***	0.703***	0.297***	0.752***	0.765***	0.357***	0.723***	0.399***	0.693***
	(13.37)	(4.20)	(11.47)	(5.23)	(12.07)	(12.26)	(6.97)	(12.47)	(7.56)	(11.92)
SREV	0.008	0.035	0.585***	0.037	0.496***	0.195**	0.068	0.570***	0.101	0.562***
	(0.09)	(0.56)	(5.90)	(0.58)	(4.88)	(2.34)	(1.09)	(6.43)	(1.57)	(6.41)
DAP		0.726***					0.669***			
		(20.10)					(18.62)			
BackTest DAP			0.479***					0.369***		
			(9.52)					(8.50)		
BN DAP				0.803***					0.760***	
				(18.68)					(16.93)	
BN BackTest DAP					0.476***					0.445***
					(7.89)					(8.64)
Adj R2	0.425	0.705	0.525	0.684	0.497	0.379	0.658	0.468	0.629	0.471
Observations	432	432	432	432	432	432	432	432	432	432

This table reports the results when a pre-publication ratio ("pre_pub"), which denotes for each year the fraction of anomalies that are before their formal publication, is further included to explain the H-L portfolio returns of backtested DAP with BN normalization. The first four columns expand the previous tests (Table 8, Panel B) to explain the

backtested DAP with BN normalization returns by FF7, NN returns, and the pre-publication ratio. Columns 5-8 examine the H-L portfolio returns of backtested DAP with BN normalization based on VW characteristic alphas estimated from the rolling window. We exclude 20% of microcap stocks in this analysis. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Backtested DAP with BN normalization				Backtested DAP with BN and VW Char-Alphas			
VARIABLES	bn_backtest	bn_backtest	bn_backtest	bn_backtest	bn_backtest	bn_backtest	bn_backtest	bn_backtest
Constant	1.724***	1.144***	0.866***	0.683**	1.841***	1.225***	1.056***	0.844**
	(8.41)	(3.31)	(2.70)	(2.17)	(8.96)	(3.54)	(3.12)	(2.56)
pre_pub		1.158**	0.351	0.371		1.230**	0.740	0.773
		(2.09)	(0.67)	(0.73)		(2.21)	(1.35)	(1.45)
smb	-0.170**	-0.169**	-0.169**	-0.107	-0.344***	-0.343***	-0.344***	-0.274***
	(-2.29)	(-2.29)	(-2.48)	(-1.57)	(-4.63)	(-4.64)	(-4.77)	(-3.84)
hml	0.003	-0.007	-0.112	-0.111	-0.092	-0.104	-0.167*	-0.165*
	(0.04)	(-0.08)	(-1.34)	(-1.37)	(-1.03)	(-1.16)	(-1.90)	(-1.94)
rmw	0.570***	0.574***	0.476***	0.355***	0.594***	0.599***	0.539***	0.403***
	(6.05)	(6.11)	(5.45)	(4.01)	(6.29)	(6.37)	(5.84)	(4.32)
cma	1.046***	1.047***	0.862***	0.766***	0.738***	0.739***	0.626***	0.513***
	(8.00)	(8.05)	(7.06)	(6.33)	(5.63)	(5.67)	(4.86)	(4.03)
mkt_rf	-0.095*	-0.095*	-0.062	-0.006	-0.114**	-0.114**	-0.094*	-0.029
	(-1.92)	(-1.93)	(-1.35)	(-0.12)	(-2.29)	(-2.30)	(-1.94)	(-0.59)
mom	0.186***	0.176***	-0.040	-0.088*	0.179***	0.169***	0.038	-0.018
	(3.99)	(3.78)	(-0.80)	(-1.79)	(3.84)	(3.62)	(0.72)	(-0.34)
st_rev	-0.996***	-1.002***	-1.032***	-1.040***	-1.185***	-1.192***	-1.210***	-1.219***
	(-16.03)	(-16.17)	(-18.04)	(-18.63)	(-19.03)	(-19.20)	(-20.00)	(-20.76)
nn1			0.354***	0.182***			0.214***	0.011
			(8.68)	(3.19)			(4.98)	(0.18)
nn2				0.183***				0.223***
				(3.12)				(3.63)
nn3				0.126***				0.135***
				(2.96)				(3.01)
Observations	432	432	432	432	432	432	432	432
r2_a	0.650	0.653	0.705	0.720	0.681	0.684	0.701	0.719

Table IN4: Explaining DAPs Returns Generated by Published-Anomalies

(240 months training window)

This table reports the results of the revert test in which we use the FF7 factors as well as the H-L returns of various NN models to explain the time series of H-L returns generated by DAP. Unlike the previous setup, for DAP and NN, we only utilized currently published anomalies during estimation. In Panel A, the dependent variable is the H-L returns generated by backtested DAP, when EW rolling-window alphas are used. Columns 1-5 report the tests when the 20% microcaps are excluded in portfolio construction. Columns 6-10 tabulate the results when all stocks are included. Within the first five models, Column 1 is the FF7 alpha of DAP when no NN is added. For columns 2-5, we add NN1, NN2, and NN3 as explanatory variables. Panel B applies the same test to the H-L returns generated by backtested DAP with BN normalized conditional alphas when EW rolling-window alphas are used. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: BackTest DAP Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.597***	1.408***	1.325***	1.312***	1.249***	0.961***	0.774***	0.675**	0.758***	0.717***
	(5.73)	(5.18)	(4.90)	(4.89)	(4.67)	(3.48)	(2.77)	(2.50)	(2.78)	(2.65)
SMB	-0.016	0.040	0.073	0.032	0.067	0.462***	0.404***	0.409***	0.385***	0.415***
	(-0.16)	(0.40)	(0.75)	(0.33)	(0.69)	(4.62)	(4.01)	(4.23)	(3.90)	(4.25)
HML	0.176	0.010	-0.043	-0.066	-0.115	0.615***	0.499***	0.439***	0.402***	0.404***
	(1.46)	(0.08)	(-0.35)	(-0.54)	(-0.95)	(5.12)	(4.01)	(3.68)	(3.22)	(3.29)
RMW	0.341***	0.088	0.048	0.151	0.033	0.120	0.078	0.049	0.017	0.018
	(2.66)	(0.67)	(0.37)	(1.21)	(0.25)	(0.95)	(0.62)	(0.40)	(0.14)	(0.15)
CMA	1.036***	0.854***	0.864***	0.866***	0.811***	1.129***	1.011***	0.875***	0.916***	0.838***
	(5.83)	(4.88)	(5.01)	(5.07)	(4.74)	(6.41)	(5.67)	(5.00)	(5.18)	(4.79)
MKT	-0.098	-0.030	-0.017	-0.003	0.014	0.150**	0.185***	0.207***	0.208***	0.213***
	(-1.45)	(-0.45)	(-0.26)	(-0.05)	(0.22)	(2.24)	(2.75)	(3.17)	(3.14)	(3.28)
MOM	0.291***	0.056	0.015	0.020	-0.059	-0.278***	-0.291***	-0.313***	-0.294***	-0.312***
	(4.59)	(0.75)	(0.20)	(0.27)	(-0.76)	(-4.42)	(-4.67)	(-5.15)	(-4.80)	(-5.18)
SREV	-1.169***	-1.313***	-1.320***	-1.214***	-1.296***	-0.626***	-0.744***	-0.825***	-0.811***	-0.839***
	(-13.84)	(-15.32)	(-15.63)	(-15.06)	(-15.22)	(-7.48)	(-8.18)	(-9.44)	(-9.04)	(-9.34)
NN1		0.369***			0.070		0.212***			-0.226**
		(5.54)			(0.74)		(3.14)			(-2.25)
NN2			0.406***		0.188**			0.359***		0.378***
			(6.24)		(1.97)			(5.88)		(4.26)
NN3				0.394***	0.257***				0.339***	0.238**
				(6.76)	(3.44)				(4.93)	(2.56)
Adj R2	0.547	0.578	0.586	0.591	0.600	0.420	0.433	0.464	0.451	0.474
Observations	432	432	432	432	432	432	432	432	432	432

Panel B: Backtested DAP_BN Explained by FF7 and NNs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Exclude Small Stocks					Whole Sample				
Alpha	1.295***	1.122***	1.068***	1.036***	0.993***	0.850***	0.690***	0.649***	0.657***	0.654***
	(5.28)	(4.71)	(4.47)	(4.40)	(4.23)	(3.43)	(2.75)	(2.63)	(2.70)	(2.67)
SMB	0.105	0.156*	0.179**	0.148*	0.173**	0.360***	0.310***	0.323***	0.287***	0.307***
	(1.18)	(1.81)	(2.07)	(1.75)	(2.04)	(4.01)	(3.43)	(3.67)	(3.25)	(3.46)
HML	0.294***	0.141	0.112	0.074	0.038	0.399***	0.300***	0.275**	0.197*	0.202*
	(2.76)	(1.33)	(1.04)	(0.70)	(0.36)	(3.70)	(2.68)	(2.53)	(1.76)	(1.81)
RMW	0.374***	0.143	0.130	0.201*	0.109	0.228**	0.193*	0.178	0.131	0.131
	(3.31)	(1.23)	(1.12)	(1.83)	(0.95)	(2.00)	(1.70)	(1.59)	(1.17)	(1.17)
CMA	0.793***	0.626***	0.650***	0.638***	0.590***	1.173***	1.072***	0.994***	0.970***	0.936***
	(5.07)	(4.08)	(4.26)	(4.26)	(3.93)	(7.42)	(6.69)	(6.23)	(6.13)	(5.90)
MKT	-0.026	0.037	0.042	0.061	0.074	0.167***	0.196***	0.206***	0.221***	0.222***
	(-0.43)	(0.63)	(0.72)	(1.05)	(1.29)	(2.77)	(3.25)	(3.47)	(3.74)	(3.76)
MOM	0.300***	0.085	0.070	0.053	-0.008	-0.197***	-0.208***	-0.222***	-0.213***	-0.221***
	(5.38)	(1.29)	(1.05)	(0.84)	(-0.12)	(-3.50)	(-3.72)	(-4.01)	(-3.88)	(-4.04)
SREV	-1.100***	-1.232***	-1.226***	-1.141***	-1.207***	-0.584***	-0.686***	-0.725***	-0.760***	-0.768***
	(-14.80)	(-16.38)	(-16.42)	(-16.14)	(-16.13)	(-7.78)	(-8.38)	(-9.09)	(-9.47)	(-9.42)
NN1		0.337***			0.105		0.181***			-0.153*
		(5.77)			(1.27)		(2.99)			(-1.68)
NN2			0.338***		0.099			0.253***		0.192**
			(5.88)		(1.19)			(4.54)		(2.38)
NN3				0.358***	0.251***				0.322***	0.297***
				(7.00)	(3.82)				(5.23)	(3.52)
Adj R2	0.558	0.591	0.592	0.604	0.610	0.410	0.422	0.438	0.446	0.454
Observations	432	432	432	432	432	432	432	432	432	432

Table IN5: Subsample Analysis Based on High or Low Pre-publication Ratio (240 month, exclusing 20% microcaps)

This table reports the results of subsample tests when we use the FF7 factors as well as the H-L DAP returns to explain the time series of H-L returns generated by the neural networks (NN1). Particularly, we use the pre-publication ratio to split the whole testing period into two subsamples, 1987-2005 and 2006-2022, for which the pre-publication ratio is above and below the cutoff threshold of 50%. We then apply the previous test (Table 7, Panel A) to these subperiods, in which we explain NN1 returns by FF7 and backtested DAP with BN normalization. Panel A excludes 20% microcaps in constructing portfolios. Panel B includes all stocks. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Before cutoff year					After cutoff year		
VARIABLES	nn1	nn1	nn1	nn1	nn1	nn1	nn1	nn1	nn1	nn1
Constant	2.674***	2.097***	1.847***	1.678***	1.554***	1.078***	0.740***	0.732***	0.776***	0.667**
	(7.94)	(6.06)	(5.67)	(4.76)	(4.38)	(3.70)	(2.89)	(2.88)	(2.70)	(2.34)
smb	-0.054	0.207*	0.231**	0.089	0.115	-0.051	0.264**	0.124	-0.014	-0.066
	(-0.50)	(1.74)	(2.17)	(0.86)	(1.11)	(-0.39)	(2.20)	(1.08)	(-0.11)	(-0.54)
hml	0.132	0.049	0.002	0.257	0.284*	0.386***	0.193*	0.260**	0.382***	0.336***
	(0.77)	(0.30)	(0.01)	(1.61)	(1.79)	(3.17)	(1.78)	(2.46)	(3.27)	(2.92)
rmw	0.180	-0.365**	-0.389**	-0.010	-0.040	0.641***	-0.218	-0.045	0.397**	0.301*
	(1.26)	(-2.01)	(-2.55)	(-0.07)	(-0.29)	(3.92)	(-1.23)	(-0.28)	(2.38)	(1.81)
cma	0.691***	0.338	0.130	0.195	0.155	0.358*	-0.171	-0.254	0.048	-0.006
	(3.26)	(1.56)	(0.63)	(0.92)	(0.73)	(1.84)	(-0.95)	(-1.40)	(0.24)	(-0.03)
mkt_rf	-0.082	0.057	0.097	-0.019	0.003	-0.192***	0.029	0.012	-0.132*	-0.142**
	(-0.97)	(0.66)	(1.20)	(-0.24)	(0.04)	(-2.62)	(0.42)	(0.17)	(-1.85)	(-2.05)
mom	0.699***	0.443***	0.366***	0.603***	0.609***	0.489***	0.211***	0.196***	0.437***	0.421***
	(9.60)	(4.96)	(4.53)	(8.70)	(8.91)	(6.77)	(2.95)	(2.74)	(6.22)	(6.09)
st_rev	0.049	0.108	0.105	0.512***	0.556***	0.223**	0.201**	0.190**	0.466***	0.529***
	(0.53)	(1.21)	(1.25)	(4.47)	(4.81)	(2.16)	(2.24)	(2.14)	(4.10)	(4.67)
base		0.426***					0.591***			
		(4.60)					(8.10)			
bn			0.644***					0.657***		
			(7.12)					(8.38)		
backtest				0.386***					0.264***	
				(6.09)					(4.34)	
bn_backtest					0.474***					0.355***
					(6.45)					(5.22)
r2_a	0.420	0.469	0.527	0.502	0.510	0.357	0.517	0.525	0.411	0.433
N	228	228	228	228	228	204	204	204	204	204

Table IN6: Subsample Analysis Applying to DAPs (240 month, excluding 20% microcaps)

This table reports the results of subsample tests when we use the FF7 factors as well as the NN H-L portfolio returns to explain the time series of H-L returns generated by backtested DAPs. Particularly, we use the pre-publication ratio to split the whole testing period into two subsamples based on the cutoff year when the pre-publication ratio is 50%. We exclude 20% microcaps in this test. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
			Before Cutoff year						After Cutoff year			
VARIABLES	backtest	backtest	backtest	bn_backtest	bn_backtest	bn_backtest	backtest	backtest	backtest	bn_backtest	bn_backtest	bn_backtest
Constant	2.581***	1.576***	1.399***	2.361***	1.461***	1.283***	1.141***	0.782**	0.713**	1.158***	0.786***	0.704**
	(7.76)	(4.51)	(4.02)	(8.32)	(4.94)	(4.30)	(3.49)	(2.41)	(2.25)	(4.03)	(2.82)	(2.55)
smb	-0.371***	-0.350***	-0.299***	-0.356***	-0.337***	-0.287***	-0.140	-0.123	-0.011	0.044	0.061	0.142
	(-3.44)	(-3.51)	(-3.02)	(-3.87)	(-3.99)	(-3.38)	(-0.95)	(-0.88)	(-0.08)	(0.34)	(0.51)	(1.18)
hml	-0.324*	-0.373**	-0.387**	-0.321**	-0.365***	-0.375***	0.017	-0.112	-0.149	0.142	0.009	-0.009
	(-1.92)	(-2.38)	(-2.58)	(-2.23)	(-2.76)	(-2.90)	(0.12)	(-0.83)	(-1.14)	(1.18)	(0.08)	(-0.08)
rmw	0.492***	0.424***	0.322**	0.464***	0.403***	0.318***	0.924***	0.711***	0.431**	0.957***	0.736***	0.516***
	(3.47)	(3.22)	(2.47)	(3.84)	(3.62)	(2.84)	(5.04)	(3.90)	(2.27)	(5.94)	(4.68)	(3.11)
cma	1.286***	1.026***	0.937***	1.132***	0.899***	0.851***	1.172***	1.053***	0.823***	1.024***	0.901***	0.702***
	(6.15)	(5.17)	(4.88)	(6.34)	(5.36)	(5.15)	(5.39)	(5.01)	(3.81)	(5.36)	(4.98)	(3.73)
mkt_rf	-0.163*	-0.133*	-0.060	-0.181**	-0.153**	-0.094	-0.227***	-0.163**	-0.120	-0.140*	-0.074	-0.028
	(-1.94)	(-1.70)	(-0.77)	(-2.52)	(-2.32)	(-1.41)	(-2.75)	(-2.03)	(-1.48)	(-1.94)	(-1.07)	(-0.40)
mom	0.249***	-0.014	-0.062	0.190***	-0.045	-0.085	0.198**	0.035	-0.033	0.192***	0.024	-0.035
	(3.46)	(-0.17)	(-0.80)	(3.10)	(-0.67)	(-1.28)	(2.44)	(0.41)	(-0.38)	(2.70)	(0.32)	(-0.46)
st_rev	-1.201***	-1.219***	-1.184***	-1.070***	-1.086***	-1.067***	-0.920***	-0.994***	-0.992***	-0.864***	-0.941***	-0.954***
	(-13.18)	(-14.43)	(-14.50)	(-13.76)	(-15.19)	(-15.21)	(-7.94)	(-8.86)	(-8.97)	(-8.48)	(-9.72)	(-9.89)
nn1		0.376***	0.095		0.336***	0.149**		0.333***	0.181*		0.345***	0.194**
		(6.09)	(1.11)		(6.45)	(2.03)		(4.34)	(1.74)		(5.22)	(2.13)
nn2			0.374***			0.238***			-0.072			0.022
			(4.41)			(3.26)			(-0.61)			(0.21)
nn3			0.037			0.065			0.420***			0.284***
			(0.63)			(1.29)			(3.87)			(3.01)
Observations	228	228	228	228	228	228	204	204	204	204	204	204
r2_a	0.683	0.728	0.750	0.705	0.751	0.764	0.606	0.639	0.664	0.609	0.655	0.672

Table IN7: Explaining Neural Network Returns Across Different Training Windows

This table reports how various factors affect the returns of H-L portfolios selected by neural networks with one hidden layer (NN1). The testing period is from January 1987 to December 2022. The x-axis denotes the length of the rolling training windows (i.e., 60, 90, 120, …, 240 months). The first column ("NN1 Alpha Net of FF7") reports the monthly Fama-French seven factor-adjusted returns. The next two columns ("NN1 Alpha Net of FF7 and Backtest" and "NN1 Alpha Net of FF7 and BN Backtest") report the NN1 alpha when backtested DAP and BN backtested DAP H-L portfolio returns are also included as an additional factor on top of the seven factors.

Estimation Window (month)	(1) NN1 Alpha Net of FF7	(2) NN1 Alpha Net of FF7 and Backtest	(3) NN1 Alpha Net of FF7 and BN Backtest
60	0.95	0.15	-0.06
	(4.11)	(0.28)	(-0.27)
90	1.14	1.19	0.20
	(5.30)	(1.33)	(0.93)
120	1.51	2.90	0.42
	(7.18)	(3.05)	(2.13)
150	1.68	3.97	0.77
	(8.21)	(3.92)	(3.93)
180	1.68	4.43	0.83
	(8.78)	(4.75)	(4.44)
210	1.66	4.73	0.88
	(8.70)	(4.89)	(4.63)
240	1.79	5.58	0.98
	(9.52)	(5.65)	(5.22)
Observations	432	432	432

Table IN8: BackTest DAP Group Statistics (60 months)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	STD	SR	TVR	Mean	STD	SR	TVR
	Exclude Small Stocks (20%)				Whole Sample			
Low	-0.83	8.884	-0.320	1.400	-0.84	9.658	-0.303	1.436
2	-0.15	7.560	-0.068	1.634	-0.27	8.163	-0.115	1.653
3	0.11	6.704	0.055	1.706	-0.04	7.439	-0.017	1.700
4	0.38	5.972	0.223	1.744	0.19	6.533	0.103	1.731
5	0.49	5.502	0.307	1.751	0.48	5.922	0.282	1.740
6	0.57	5.117	0.388	1.746	0.63	5.314	0.408	1.727
7	0.72	4.909	0.510	1.716	0.68	5.120	0.460	1.713
8	0.76	4.760	0.553	1.670	0.75	4.918	0.527	1.664
9	0.94	4.885	0.666	1.579	0.89	5.033	0.611	1.572
High	1.06	5.244	0.668	1.143	1.05	5.403	0.672	1.143
H-L	1.89	7.726	0.821	1.272	1.89	8.554	0.766	1.289
Observations	432	432	432	432	432	432	432	432

Table IN9: BN BackTest DAP Group Statistics (60 months)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	STD	SR	TVR	Mean	STD	SR	TVR
	Exclude Small Stocks (20%)				Whole Sample			
Low	-0.60	8.402	-0.249	1.324	-0.97	9.038	-0.371	1.436
2	0.03	6.949	0.017	1.597	-0.08	7.426	-0.039	1.678
3	0.17	6.159	0.094	1.647	0.10	6.604	0.051	1.723
4	0.47	5.459	0.298	1.706	0.31	6.178	0.174	1.752
5	0.55	5.230	0.361	1.724	0.54	5.395	0.346	1.763
6	0.56	5.012	0.389	1.723	0.64	5.131	0.432	1.761
7	0.69	4.666	0.513	1.679	0.63	4.872	0.445	1.723
8	0.79	4.661	0.591	1.625	0.80	4.863	0.572	1.692
9	0.99	4.704	0.726	1.520	0.93	4.861	0.662	1.595
High	1.08	4.674	0.803	1.123	1.08	5.212	0.718	1.178
H-L	1.69	6.784	0.862	1.223	2.05	7.665	0.926	1.319
Observations	432	432	432	432	432	432	432	432

Table IN10: The Predicting Accuracy of Back Testing

This table reports the accuracy of backtesting predictions when applying to the 60-month DAPs. For each month t during the test period from January 1987 to December 2022 and each anomaly k, we first estimate its 60-month rolling anomaly alpha, α_{kt-1} and its most recent realization of alpha, $\alpha_{k,-t}$, as specified in our main text. In Columns (1) and (2), we split the sample of all anomaly-month observations into two groups based on the outcomes of backtesting (i.e., to keep or mask the anomaly in DAP portfolios). More specifically, conditional on $\alpha_{kt-1} > 0$, the backtesting process keeps an anomaly for the month-t DAP portfolio when $\alpha_{k,-t} > 0$ (i.e., $H(\alpha_{kt-1}) = H(\alpha_{k,-t})$, where $H(x)$ denotes the sign function) and masks it otherwise. Column (2) reports the fraction of anomaly-month observations that have been kept and masked in our out-of-sample tests. Next, Columns (3) and (4) further split each of the two backtesting groups into two subgroups depending on whether the realized anomaly alphas in month t has a positive sign or not (i.e., whether the sign of the realized alpha aligns with the selection of the backtesting process). Column (4) reports the fraction of all anomaly-month observations that fall into each of the four subgroups. Lastly, Columns (5) and (6) provide the statistical interpretations of the four subgroups, including 1) correct predictions made by backtesting (i.e., the backtesting process has properly selected "good" anomalies that can realize positive alphas in the testing period or properly rejected "wrong" anomalies with negative future alphas); 2) Type I Error of backtesting (false positive, i.e., the backtesting process selects "wrong" anomalies that fail to deliver positive alpha); and 3) Type II Error of backtesting (false negative, i.e., the backtesting process rejects "good" anomalies).

Backtesting based on the signs of α_{kt-1} and $\alpha_{k,-t}$ (1)	Fraction of Observations (2)	Realized Anomaly Alphas (α_{kt}) (3)	Fraction of Observations (4)	Statistical Interpretation (5)	Fraction of Observations (6)
Keep: $\alpha_{k,-t} \geq 0$	60.24%	$\alpha_{kt} < 0$	22.36%	Type I Error	22.36%
		$\alpha_{kt} \geq 0$	37.88%	Correct	57.37%
Drop: $\alpha_{k,-t} < 0$	39.76%	$\alpha_{kt} < 0$	19.48%		
		$\alpha_{kt} \geq 0$	20.27%	Type II Error	20.27%

Figure IN1: Pre-publication Ratio

This figure plots the time series of the pre-publication ratio. For each year, this ratio denotes the fraction of anomalies that are before their formal publication.

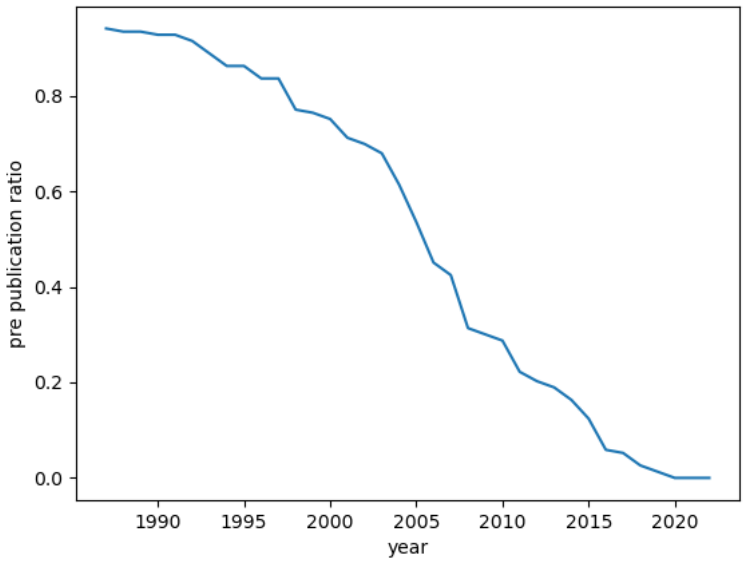


Figure IN2: Explaining Neural Network Returns Across Different Training Windows

This figure plots various returns of the H-L portfolios selected by neural networks across various lengths of training windows (i.e., 60, 90, 120,…, 240 months). The line "NN1 Alpha Net of FF7" plots the FF7-adjusted returns for the NN1 selected H-L portfolios over the testing period from January 1987 to December 2022. The line "NN1 Alpha Net of FF7 and Backtest" plots the returns for the NN1 selected H-L portfolios when both FF7 factors and the H-L portfolio return of the backtested DAP are adjusted. The line "NN1 Alpha Net of FF7 and BN Backtest" plots the returns for the NN1 selected H-L portfolios when both FF7 factors and the H-L portfolio return of the backtested DAP using BN-normalized rolling alpha are adjusted.

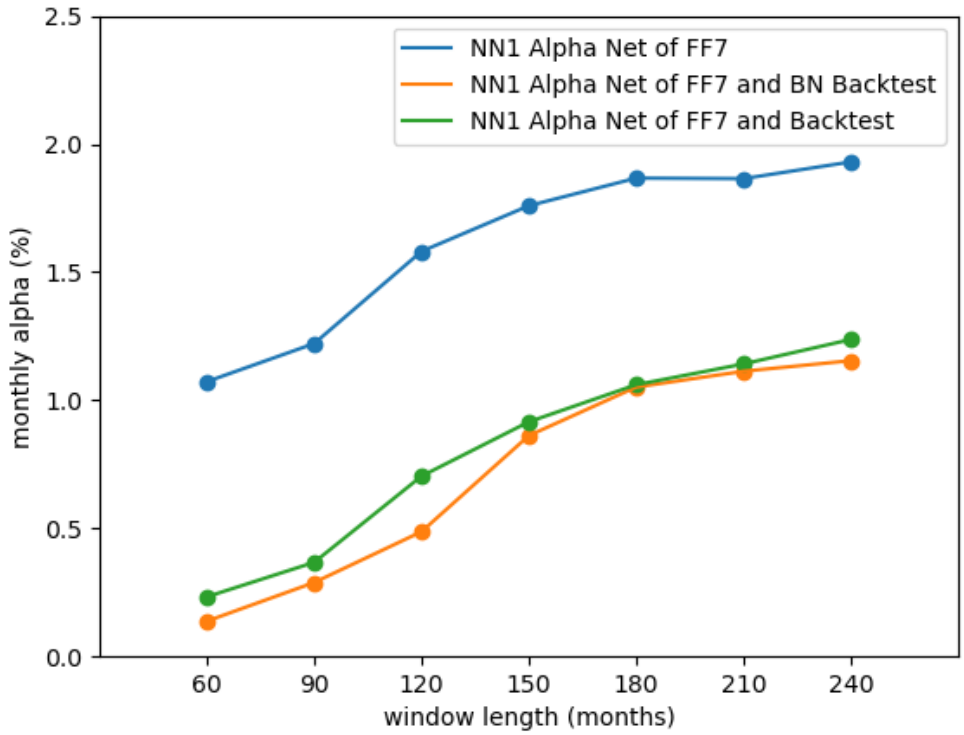
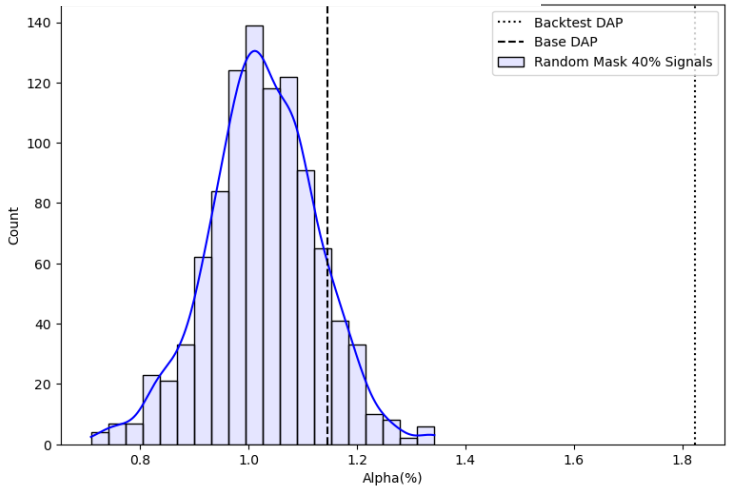


Figure IN3: The Distribution of Simulated Pseudo-Backtesting

This figure reports the distribution of simulated DAP portfolios with Pseudo-Backtesting. For each simulation, we randomly drop 40% of the signals (i.e., rolling alphas) when constructing the DAP portfolio each month. All other conditions are similar to the construction of Backtested DAP in Table 1. We then calculate the seven-factor adjusted performance (i.e., FF7 Alpha) and its t-statistics of the time-series returns delivered by the DAP portfolio with Pseudo-Backtesting in the test period. The test period spans from January 1987 to December 2022. Panel A plots the distribution of FF7 Alphas. The dashed and dotted lines indicate the positions of the alphas or our original DAP and Backtest DAP portfolios. Panel B plots the distribution of the t-statistics.

Panel A: Monthly Alpha Net of FF7



Panel B: T-statistics of Monthly Alpha Net of FF7

