# Methodology, Design, and Data Integrity Validation Study of Turing Technology's 2024 Ensemble Active Management White Paper

Prof. David Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205
(404)822-8949
sman@gatech.edu

**Executive Summary**: The goal of the project was to validate the methodology, design, and data integrity that Turing has used to arrive at the published results of their January 2024 White Paper entitled "*Ensemble Active Management: AI's Transformation of Active Management*". In particular, we examined: (i) background methodology underlying Turing's work; (ii) statistical/randomness aspects of Turing's fund selection strategies involving the Ensemble Active Management (EAM) and underlying Portfolio of Funds (POF) construction methodologies; and (iii) a representative sampling of performance characteristics of the various portfolios, including estimated performance of strategies and comparison among strategies.

We found that the underlying methodology is sound. Standard sampling/randomness protocols were followed, appropriate randomness protocol for the underlying POF construction was carried out properly, EAM analytics and construction methodology was performed properly, and EAM and POF performance has been properly interpreted by Turing, including bias analysis and mitigation.

## 1. Introduction and Outline

Returns from professionally managed financial portfolios have proven to be less than those resulting from traditional index-based benchmarks, a phenomenon due to a combination of less-than-optimal selection strategies coupled with trade and management charges. This state of affairs is unfortunate given the importance such portfolios play in the broader society's financial health.

Turing has developed modern, state-of-the-art Machine Learning (ML)-based procedures that aim to improve performance when measured against competing strategies. Machine Learning / Ensemble Methods have been used in recent years to enhance the performance of classical statistical analysis methods such as regression, analysis of variance, principal component analysis, and factor analysis. Such methods combine information and predictions from multiple underlying tools, for example, linear and nonlinear regression, "regularized" methods (e.g., subset selection, ridge regression, and Lasso methods), decision trees, Bayesian methods, bootstrapping, and neural networks, among others. In addition to enhancing the toolkit of analysis techniques to work with, these modern methods are relied on to reduce model bias and variance and to enhance model robustness.

In fact, it is well known and accepted by the research community that ensemble methods can be used to extract additional information that classical statistical methods may miss or may not have the data or computational power to otherwise achieve. ML/ensemble methods are now incorporated in a wide variety of practical settings ranging from medical applications (e.g., obtaining the best organ transplantation policy) to sports prediction (e.g., determining the win probability of a team in real time

as a game progresses) to financial engineering (e.g., finding improved portfolios). Turing is concerned with the latter setting in their White Paper.

## 2. Review of Methodology

We were asked to examine the following areas of interest:

Item 1: Review of methodology.

Item 2: Fund selection methodology.

- Potential biases involved in the inclusion of available funds within the study.
- Confirm (through sampling or database reports) that the stated fund selection methodology was followed.

Item 3: Portfolio of Funds construction methodology.

- Confirm through sampling and potentially review of code that the construction of the 60,000 bundles of 12 funds each were random.
- Statistically sample the output results to ensure that the final results were within a reasonable error range.

Item 4: Ensemble Active Management (EAM) construction methodology.

- Sample EAM portfolios to ensure that security selection per portfolio was consistent to standard EAM methodology.
- Ensure that all Portfolio of Funds were translated into EAM portfolios.

Item 5: Performance calculation.

- Sample Portfolio of Fund (POF) performance data based on the underlying set of 12 fund returns.
- Sample EAM portfolio performance data to ensure that the resulting set of stocks and weights are accurately translated into EAM daily returns.

Item 6: Confirmation of Performance Outcomes.

- Confirm EAM and POF summary data as presented by Turing is accurate based on the data sets generated.
- Conduct any additional statistical testing required to verify data accuracy.

In addition to standard statistical methodology references on ML / Ensemble Methods (e.g., the classic texts Hastie, Tibshirani, and Friedman (2009) and James et al. (2021) as well as the fundamental paper Breiman (2001)), we read material directly pertinent to the Turing methodology; the specific materials perused included:

- Pinsky (9/4/18), "Mathematical Foundation for Ensemble Machine Learning and Ensemble Portfolio Analysis". (This paper provided theoretical background for some of the methodology in play. Besides being an interesting piece to read, we found the paper to be rigorous and mathematically correct, with apt data-driven examples that properly motivated the findings.)
- "2023 White Paper Validation Study – Draft Project Plan". (This document provided background information as well as a data limitations analysis.)
- "Ensemble Active Management: Reinventing Active Investment Management" (December 2023 white paper).
- "Core Patent without Claims". (This document provided background material.)
- Many Excel files (discussed below).
- Some Python code supplied by Turing.

- The Turing website, https://turingtechnologyassociates.com/.

Turing provided complete transparency to us regarding all data and code associated with the White Paper; the only information not accessed was code related to their Hercules.ai fund replication technology. ***We determined that the material that was provided during the course of this study was more than sufficient to completely understand the methodology, conduct the appropriate statistical validation, and draw sound conclusions about the methodology's performance.***

## 3. Fund Selection Methodology

Turing maintains an extensive database of replicated mutual funds, including detailed information on daily holdings and weights. The selection of mutual funds that are resident in Turing's database reflect the decision-making of Turing's clients as they build EAM investment portfolios. Collectively, the replicated funds in Turing's database represent more than $4 trillion in actively managed fund assets.

Turing built the underlying POFs that were used in the White Paper analysis by randomly selecting 12 mutual funds that were (i) resident in the database at the time of the analysis, and (ii) adhered to the screening methodology as detailed in Section 4 below. One area of bias that we investigated related to the construction of the pool of available mutual funds that the random POFs were built from.

The fact that a particular fund appears in the set means that at least one client regarded the fund as sufficiently desirable to include it in an EAM portfolio. Thus, the available pool of funds available to the analysis reflected some minor selection bias; however, the set of funds available in the selection pool comprises greater than $3 trillion in fund assets and represents approximately 65% of all actively managed mutual funds in the industry. ***Based on the scale of available funds, we are satisfied that this mitigates against any potential selection bias issue.***

The pool of funds that were available for use in the White Paper analysis contained 405 different funds as defined by unique "fund tickers". Turing clarified to us that approximately 15% of these funds are considered to be redundant in the sense that they are merely different "share classes" of the same underlying fund; but Turing compensated for this redundancy by establishing a POF construction protocol that allowed for at most one fund per fund family – which effectively prevented redundancy of funds within any POF. As such, redundant funds are not overweighted in the subsequent analyses, and the POF construction algorithms work on the 333 independent funds.

The database of 333 funds is divided into six general but commonly denominated style boxes ("style boxes" are defined by a market capitalization categorization of 'large,' 'mid,' or 'small,' and an investment style categorization of 'value,' 'blend,' or 'growth'): Large Cap Value (LCV), Large Cap Blend (LCB), Large Cap Growth (LCG), Small Cap Value (SCV), Small Cap Blend (SCB), and Small Cap Growth (SCG). For example, within the available pool of funds used in the analysis, there were 65 LCV funds (out of the 333) in the database.

For purposes of the subsequent analysis, we note that the superset of 333 funds consists of subsets arising from 142 different fund families, e.g., American Beacon, Putnam, etc. For instance, the LCV subset of funds is itself comprised of 65 funds arising from 46 fund families; and continuing the example, of those 46 families, the American Beacon family is comprised of a single LCV fund (AADEX), while Putnam has three funds (PEIYX, PEQSX, and PEYAX).

Based on Turing's methodology, any Turing EAM portfolio will contain at most one fund from any specific fund family. By definition, a POF used in this analysis would comprise 12 funds from 12 distinct fund

families. This methodology approach is consistent with best practices in the use of Ensemble Methods construction techniques in that it avoids the introduction of unanticipated positive correlation that might arise from the selection of multiple funds from a single family.

***Our analysis found that Turing adhered to the above ground rules for selecting their Turing portfolios.***

We comment on statistical aspects of the selection process below.

### 4. Portfolio of Funds (POF) Construction Methodology

Turing's methodology relied on constructing multiple realizations of underlying Portfolio of Fund portfolios, each consisting of 12 funds within a particular style box (LCV, LCB, LCG, SCV, SCB, and SCG), and then weighting the various assets within the multiple realizations.

The first task (construction of each POF) is straightforward. Namely, for a specific asset class,

- Randomly select 12 funds from distinct fund families (e.g., select one fund from each of 12 randomly selected families out of the 46 families comprising the LCV asset class). We noted that the random selection is weighted so as to slightly overweight fund families having more than one fund, but were comfortable that this slight bias did not inappropriately distort the results. Thus, for example, we would more likely see a representative fund from the Putnam LCV family (one of PEIYX, PEQSX, and PEYAX) than the single American Beacon member AADEX. This makes sense since such multi-fund families tend to be larger and play a bigger role in the market.
- Repeat this construction procedure for N = 10,000 independent replications of the asset class (though this could be a larger number, if desired).

We conducted our own Monte Carlo (MC) simulation experiments to replicate the results of Turing's POF construction process. (Exact calculations are not so straightforward due to the combinatorial nature of the POF construction process, so MC simulation is indeed the best way to proceed.) At that point, we were able undertake statistical goodness-of-fit tests to compare Turing's results with ours, and ***this enabled us to make the key finding that Turing's random sample generation was carried out properly and produced results that were reasonable and consistent.***

## 5. Ensemble Active Management Construction Methodology

EAM construction translates information from the replicated holdings and weights of the ensemble of POFs (12 funds per POF) as detailed in Section 4 into specific stock holdings. Specifically, for a particular POF, one calculates a weighted set of the top 50 equities that appear in that portfolio based on the highest consensus agreement of a manager's level of positive or negative conviction related to each security. (We refer readers to Turing's White Paper for a more-detailed explanation of how they extract measures of manager conviction from the holdings and weights of actively managed mutual funds.)

The EAM construction methodology is straightforward:

- For each of the 12 funds, remove cash holdings, and normalize the weights over all equities within each fund.
- Average the weights over all funds (where the weight for an equity not appearing in a particular fund is simply taken to be zero for that fund).
- Remove all security holdings not in the benchmark, and re-normalize.
- Take the top 50 averages and re-normalize once again.
- Re-do this weighting and selection exercise every two weeks.

***Based on spot checks we conducted for code correctness, we have concluded that the code was implemented properly. Namely, proper weightings were calculated for the 50 securities that comprise the EAM portfolios.***

Although not explicitly included in the purview of this report, we nevertheless make the following comment on the mathematical integrity underlying Turing's methodology. ***The reasoning behind Turing's weighting strategy is apt***, as it reflects 1) core principals of Ensemble Methods – proven mathematical techniques for combining multiple predictive engines to create a statistically stronger aggregated predictive engine, and 2) overall expert analyst choices regarding the stocks that are included in their portfolios. By taking a weighted average of the 50 top securities from each of a sample of 12 respected portfolios, Turing (i) covers a reasonably sized consensus within a given asset class and (ii) is designed to capture significantly more positively informed security choices from top experts.

## 6. Performance Calculation

In Sections 6 and 7 (below) we discuss POF and EAM performance, both individually and comparatively. We will comment here on some specific findings that will prove useful in summarizing results.

We were presented with all performance-related output data, representing more than 60 files and 13 gigabytes of data. All files were perused and all data was sampled. In particular, Turing provided data on daily performance of EAM, POF, and corresponding benchmark portfolios over the 7-year period 2016 to 2022 for each of the fund style boxes LCV, LCB, LCG, SCV, SCB, and SCG.

***We determined that the sample size of data generated for performance evaluation was adequate and sufficient.*** Turing generated 60,000 total EAM and POF portfolios, with performance covering the 7-year period referenced above. This translates to 420,000 unique calendar years' worth of performance data points for both EAM and POF portfolios. As context, from the turn of the current century there have been approximately 19,000 unique calendar years' worth of performance for all actively managed US equity mutual funds in the entire industry. Thus Turing generated 20-times more performance data points than the entire fund industry has in the past 24 years.

Our goal was to independently validate the output performance data, but given the sheer size of the performance data set, a comprehensive sampling effort was deemed inefficient. Therefore, we condensed the data set evaluation by independently constructing 'batch data' sets based on 7 calendar years of data (Section 6, see below), and then compared those batch output results to the output file generated by Turing (Section 7).

**Notes on performance calculation.** The POF data was calculated from the published daily performance of each fund using a daily rebalance methodology. The benchmark returns are actuals from the corresponding indexes represented by each of the fund style boxes. The EAM returns were based on the performance of the daily set of 50 stocks per generated EAM portfolio. As mentioned, the EAM and POF portfolios reflected 10,000 independent combinations of EAM and POF performance for each of the style boxes (60,000 in total), while there was only a single real-life benchmark realization from each style box.

As to be expected, for any particular day there was some variability among the 60,000 realizations of the EAM and POF portfolios; but that variability was small compared to the day-to-day variability of the portfolios (due to the usual day-to-day market conditions), so we will not comment further on that finding.

In our subsequent analysis, for any given day, we averaged the 60,000 replications for both the EAM and POF portfolios. This resulted in a 7-year time series of EAM averages and a 7-year time series of POF averages. *It is these time series that we compared to the benchmarks.*

Before establishing the results from our statistical analysis, we note that the study of day-to-day portfolio data presents challenges due to the facts that day-to-day returns are:

(i) not normally distributed;

(ii) not identically distributed (e.g., the underlying returns may change due to seasonality or long-term trends); and

(iii) serially correlated (e.g., day *n* values are not independent of those of day *n*+1).

On the one hand, these challenges preclude the use of elementary statistical analysis tools. On the other hand, issue (i) is not a material problem in the current application since each EAM and POF data point in our analysis (as explained above) is itself the average of 10,000 replications per style box, and is well-approximated by a normal distribution (via what is known as the Central Limit Theorem); and issue (ii) mostly goes away when we *compare* competing portfolios that are sampled under identical conditions (e.g., when comparing EAM and POF over the same time period). This leaves issue (iii), related to the serial correlation of the returns. A standard, *provably rigorous* methodology for dealing with serial correlation is to divide the data into contiguous *batches* of some large size and to assume that the sample means of these batches are themselves approximately normal, identically distributed, and (importantly) independent of each other.

In the current application, we have various 7-year time series of EAM, POF, and benchmark values. To proceed with the analysis, as a first pass, for a specific time series, we used a batch size of length one year (resulting in 7 batches of observations); and then we calculated each of the mean returns of the 7 batches by taking the averages of all observations from the time series over each corresponding year. (Other batch sizes could be used as well, e.g., ½-year, ¼-year, etc.) We then conducted and passed statistical tests to verify that the batch means were approximately independent.

Statistical Analysis of Performance*: Individual Portfolios*

Using the method of batch means, we obtained the following 95% confidence intervals for the mean performance of the various EAM, POF, and benchmark portfolios. (The interpretation is that we are 95% confident that the true mean lies within the given interval.)

| | | | | | |
|---|---|---|---|---|---|
| **Large Cap Value EAM**: | [-0.009, 0.321] | **POF**: | [-0.021, 0.260] | **Benchmark**: | [-0.035, 0.248] |
| **Large Cap Blend EAM:** | **[-0.006, 0.290]** | **POF:** | [-0.022, 0.269] | **Benchmark:** | [-0.040, 0.308] |
| **Large Cap Growth EAM**: | [-0.060, 0.421] | **POF**: | [-0.085, 0.379] | **Benchmark**: | [-0.067, 0.395] |
| **Small Cap Value EAM**: | [-0.039, 0.313] | **POF**: | [-0.058, 0.288] | **Benchmark**: | [-0.077, 0.284] |
| **Small Cap Blend EAM**: | [-0.019, 0.391] | **POF**: | [-0.049, 0.285] | **Benchmark**: | [-0.066, 0.269] |
| **Small Cap Growth EAM**: | [-0.068, 0.480] | **POF**: | [-0.090, 0.376] | **Benchmark**: | [-0.100, 0.299] |

Thus, for the **red highlighted** example above, we are 95% confident that the true mean return for the LCB EAM portfolio is somewhere between -0.006 and 0.290, with an estimated average of about **0.142** (14.2% in annual return), the midpoint of the interval.

In general, the bounds on the mean rates of return tend to be quite positive, with more upside (and a bit more potential downside) on Growth portfolios. In addition, the lower and upper bounds for EAM are almost always greater than those of the competing portfolios, sometimes significantly so.

Statistical Analysis of Relative Performance*: Comparison of Portfolios*

The findings on the *individual portfolios* (taken in isolation) do not tell the whole story. What is more important is *how the portfolios compare on a relative basis against each other*. To this end, we also provide 95% confidence intervals for the *mean differences in portfolio returns*. Below, we present comparisons for EAM vs POF, and EAM vs the benchmark.

| | | |
|---|---|---|
| **Large Cap Value:** | **EAM-POF:** [-0.006, 0.078] | **EAM-Benchmark:** [ 0.015, 0.084] |
| **Large Cap Blend:** | **EAM-POF:** [-0.024, 0.061] | **EAM-Benchmark:** [-0.070, 0.086] |
| **Large Cap Growth:** | **EAM-POF:** [ 0.005, 0.064] | **EAM-Benchmark:** [-0.034, 0.068] |
| **Small Cap Value:** | **EAM-POF:** [-0.028, 0.073] | **EAM-Benchmark:** [-0.007, 0.074] |
| **Small Cap Blend:** | **EAM-POF:** [ 0.023, 0.114] | **EAM-Benchmark:** [ 0.035, 0.133] |
| **Small Cap Growth:** | **EAM-POF:** [ 0.007, 0.118] | **EAM-Benchmark:** [ 0.002, 0.211] |

Thus, for the **red highlighted** example above, we are 95% confident that the difference in true mean returns between the EAM and benchmark SCB portfolios is somewhere between 0.035 and 0.133, with an estimated average of about **0.084 (840 basis points, or 8.4% in annual excess return),** the midpoint of the interval.

***The midpoints of the above confidence intervals represent tremendous improvements borne out by EAM.*** That being said, we note that the confidence intervals are subject to sampling error, limited data, and modeling issues related to the changing environment of the stock market. But there is no denying that the trends apparent in the above table clearly indicate the efficacy of EAM.

## 7. Confirmation of Performance Outcomes

This section describes the activities we undertook to verify Turing's performance outcomes.

- Comparative Returns: The analysis of the batch data sets described in Section 6 above shows that the expected excess rates of returns from EAM vs POF and EAM vs benchmarks are very likely to be significantly positive. ***In fact, our numbers (based on simple batch means confidence intervals) are actually slightly more optimistic than Turing's.***
    - In particular, we found (see the midpoints of the confidence intervals from the above table) that the average excess return over all portfolio fund classes was, remarkably, 404 basis points (4.04%) for EAM vs POF, and 498 basis points (4.98%) for EAM vs the benchmarks.
    - We also carried out similar analyses and obtained qualitatively similar results (not reported here) related to 3-year and 7-year returns for EAM rates.
    - **In all cases, *these results indicated that Turing's claims on comparative returns are reasonable.***
- Transaction Cost Consequences: Since EAM encourages periodic, dynamic portfolio adjustments (for their analysis, every two weeks), we undertook a rough analysis regarding the consequences

of such adjustments. ***We determined that the costs of such trade adjustment costs are not significant***, especially compared to the estimated performance gains achieved by EAM: (i) the portfolio itself is not likely to vary significantly from time period to time period (being based on a weighted selection of 50 popular securities derived from 12 funds in the same family): and (ii) transaction costs are themselves not overbearing these days.

- Success Rates: Turing defined 'Success Rates' as the percent of rolling 1-, 3-, or 5-year periods where EAM outperformed a 'target' investment return (i.e., either POFs or benchmarks). ***We verified Turing's numbers regarding the success rates (SR) in head-to-head EAM vs POF and EAM vs benchmark comparisons***. Using a method similar to batch means, we spot-checked that EAM achieved an SR of 81% vs POF and an SR of 83% vs the benchmarks for rolling 1-year time periods. These values fall in the mid-range of Turing's claims, and so ***we deem those SR claims as reasonable***.

- Fund Variability: We were interested in the variability of EAM vs that of the selection pool of 405 mutual funds – after all, additional risk might hurt the desirability of slightly augmented expected gains. We found from our batch means analysis that the variance of EAM's yearly performance was, on average, more-or-less the same as (though sometimes a bit higher than) those of POF and the benchmarks. However:

    - POFs, as a group of 12 underlying mutual funds, will see a reduced risk profile (i.e., reduced tails of a distribution of relative returns) based on the natural diversification of a portfolio of 12 entities. The reduced risk profile is measured as the square root of the number of entities; so with 12 entities the reduced risk profile is 1/3.46.

    - Since the EAM portfolios risk profile was similar to POFs, and POFs by definition will have lower risk than the individual underlying funds, ***we confirm that EAM's risk levels, as measured by tail risk, are less than the individual funds as a full cohort***.

    - Further, given that the expected returns of EAM were often significantly higher than POFs and the benchmarks, we remark that ***Turing's claims about reduced relative risk are valid*.**

    - This opinion is further vindicated by 3- and 7-year return results, where we have found that EAM variability is still about the same as POF and benchmark variability; but now, average returns are somewhat higher than one-year returns – ***indicating that "long-term" risk is mitigated by EAM.***

**In summary, in each of the performance items under study (comparison of returns, consequences of transaction costs, success rates, and fund variability), we feel that *Turing's claims are valid and hold up well* to scrutiny.**

## 8. Conclusions

We were tasked with validating the methodology, design, and data integrity that Turing has used to arrive at the published results of their January 2024 White Paper entitled "*Ensemble Active Management: AI's Transformation of Active Management*". In particular, we focused on published results involving EAM and POF returns, including especially positive outcomes involving extensive apples-to-apples Monte Carlo comparisons of EAM vs POF performance and EAM vs benchmark performance. We found through a Monte Carlo (MC) evaluation that Turing's analysis was conducted correctly: the random sampling of funds to be included in the EAM portfolio was carried out properly; and the

conclusions reached by the MC analysis were valid and (almost always) statistically significant. In particular, across all portfolio fund style boxes the EAM portfolio has an overall expected performance benefit of 400–500 basis points when compared against the corresponding POF and benchmark classes (some style boxes will be below that range, others a bit above). Similar success stories manifest for Success Rates and risk considerations.

Of course, mean performance alone does not guarantee that one investment portfolio will perform better than another over a short time horizon. But we also validated evidence indicating that EAM's performance becomes more robust and stable as it is applied over longer time periods, reflecting the compounding of excess relative returns.

***Our summary conclusions are that EAM and POF performance has been properly interpreted by Turing, including bias analysis and mitigation. Turing's claims that EAM performance is comparatively better than traditional active management and standard industry benchmarks were also substantiated.***

## References


L. Breiman (2001), "Random Forests," *Machine Learning*, 45, 5–32.

T. Hastie, R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition, Springer, New York.

G. James, D. Witten, T. Hastie, and R. Tibshirani (2021), *An Introduction to Statistical Learning with Applications in R*, second edition, Springer, New York.