Understanding the Role of Functional Diversity in Weight-Ensembling with Ingredient Selection and Multidimensional Scaling

Alex Rojas [1] **David Alvarez-Melis** [1 2]

Abstract

Weight-ensembled models, formed when the parameters of multiple neural networks are directly averaged into a single model, have demonstrated a generalization capability in-distribution (ID) and out-of-distribution (OOD) which is not completely understood, though weight-ensembles are thought to successfully exploit functional diversity allotted by each distinct model. Given a collection of models, it is also unclear which combination leads to the optimal weight-ensemble; the SOTA is a linear-time "greedy" method. We introduce two novel methods with targeted model-selection mechanisms to study the link between method-performance dynamics and the nature of how each method decides to use apply the functionally diverse components. We develop a visualization tool to explain how each algorithm explores various domains defined via pairwise-distances to further investigate selection and algorithms' convergence. Empirical analyses shed perspectives which reinforce how high-diversity enhances weight-ensembling while qualifying the extent to which diversity alone improves accuracy and demonstrate that sampling positionally distinct models can contribute just as meaningfully to improvements in a weight-ensemble.

1. Introduction

Model ensembling plays a crucial role in enhancing the performance and robustness of machine learning models. Combining the information learned by models pre-trained (or fine-tuned) with different configurations or on different tasks can reduce overfitting to any particular hyperparameter or dataset choice, leading to better generalization. The traditional approach to model ensembling in machine learning relies on averaging the *predictions* of the various models, which suffers from high memory and requires the inference-time computational cost of many modern neural networks.

Recent work by Wortsman et al. (2022) demonstrates the promise of an alternative approach to model ensembling by instead averaging the *parameters* of the models: the weight-ensembling or weight-averaging (WA) approach. Their "model-soups" approach results in a single model (the WA), addressing many of the computational limitations of prediction-based ensembling. The highly non-convex neural network loss landscape suggests that this approach should fail; yet, linear mode connectivity (LMC) (Frankle et al., 2020) finds that the interpolation between models sharing a stable region of a weight-space training trajectory remains in a low-error region. An example includes when models are fine-tuned from a shared foundation; models were empirically demonstrated in this setting to reside in convex low-loss basins where weight-ensembling does not incur significant loss barriers (Neyshabur et al., 2021).

The literature in such weight-ensembling approaches typically assume access to multiple models of identical architecture fine-tuned from a shared initialization varying hyperparameters such as learning rate and data randomness. Prior work has relied on a "greedy" approach to construct the ensemble, whereby models (the "ingredients") are considered only once to be sequentially added to the "soup" based on validation set performance of the candidate WA, and thrown out of the soup if validation performance declines (Wortsman et al., 2022; Rame et al., 2022). These greedy WA models have shown remarkable performance across complex tasks such as ImageNet (Deng et al., 2009) and DomainBed (Gulrajani & Lopez-Paz, 2020), as shown by Wortsman et al. (2022) and Rame et al. (2022) respectively.

Although the empirical effectiveness of weight-ensembling has been attributed to their finding weights in flatter regions of the loss landscape (Wortsman et al., 2022; Cha et al., 2021) and the inclusion of functionally diverse ingredients (Rame et al., 2022), the mechanics of how the greedy algorithm elicits these phenomena is not clear. Additionally, it is unknown if this algorithm is particularly well-suited to find diverse ingredient-sets, or whether there exists other methods better suited to this goal.

[1]Harvard University, Cambridge, MA, USA [2]Microsoft Research, Cambridge, MA, USA. Correspondence to: Alex Rojas <alexnrojas5@gmail.com>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

In this work, we study the link between functional diversity and weight-ensembling to further understand the effectiveness of weight-ensembling. We use quantitative distance measures (ratio-error, Euclidean distance) as proxies for functional diversity. By characterizing how various weight-ensembling algorithms leverage these relationships, we can reason about the utility of these measures. Namely, we propose the novel "greedier" algorithm, specifically designed to allow more flexibility in selecting new candidates, and we show that when extra flexibility is offered the algorithm tends to select more diverse ingredients. Yet, the novel "ranked" algorithm, designed to select diverse candidates, demonstrates that there is a limit to the benefit of diversity even in the setting where ingredients remain averageable under LMC. In summary, our contributions are:

- We develop the "greedier" algorithm for weight-ensembling which has the flexibility at each iteration to add any model to the set of ingredients.
- We develop the "ranked" algorithm for weight-ensembling which considers ingredients in order of decreasing diversity from the current WA and greedily selects the first candidate which improves performance
- We empirically demonstrate that higher Euclidean and diversity-distances robustly explain the selection mechanism of our "greedier" algorithm compared to "greedy", implying that having these distances between ingredients help improve accuracy efficiently. However, the even larger diversities selected by the "ranked" algorithm, which perform similarly to "greedy," qualify the extent to which selecting for maximal diversity is useful.
- We introduce a form of qualitative visualization that provide additional insights on the connection between these weight-ensembling algorithms and loss landscapes.

2. Methodology

2.1. Distance Measures

Functional diversity, as measured by the ratio-error (Aksela, 2003) was claimed in Rame et al. (2022) to be the driving force behind the improvements of weight-ensembling over standard (SGD-found) models because of analysis claiming that functional diversity decorrelates model predictions. They also demonstrate a positive correlation of the average pairwise diversity of a set of models with the accuracy gain of the corresponding WA over the mean accuracy over the individual ingredients, a statistic which does not directly imply that there is a lack of functional redundancy in the collection, only in the average case of a pair. For this reason, when analyzing weight-ensembling algorithms iteration-by-iteration in this work, we also pay attention to the diversity between a selected candidate ingredient and the current WA (in the context of unselected ingredients' distances). Less lossy than average pairwise diversity, we analyze these pairwise relationships jointly in the visualization method.

Definition 2.1. We use the convention of ratio-error (Aksela, 2003) to measure diversity following Rame et al. (2022). For models θ_A and θ_B, where N_{uns} and N_{sha} refer to the number of unshared and shared errors on a labelled dataset, we refer to the diversity distance as $d_D(\theta_A, \theta_B) = \frac{N_{uns}}{N_{sha}}$

Motivated by the finding of loss basins in fine-tuning (Neyshabur et al., 2021), convex regions in which models sharing initialization remain essentially linearly mode connected, we also use a Euclidean geometry-inspired measure to determine the extent to which specific weight-space geometry can explain weight-ensembling approaches. The Euclidean metric allows us to explore how different weight-ensembling traverse the basin and evaluate whether sampling candidates different parts of a loss basin sufficiently improves the WA, recasting the pursuit diversity as a weight-space traversal question.

Definition 2.2. As such, we define the Euclidean distance between two neural network parameters $\theta_A, \theta_B \in \Theta$ to be $d_E(\theta_A, \theta_B) = ||\text{vec}(\theta_A) - \text{vec}(\theta_B)||_2^2$

2.2. The Greedy Weight-Ensembling Technique

The greedy souping method (Wortsman et al., 2022; Rame et al., 2022) sorts the individual models by decreasing validation accuracy. Starting from the single highest-accuracy model, we sequentially consider adding the remaining models, only adding to the ingredients list when a candidate WA improves training-domain validation accuracy and otherwise throwing out the failed candidate in a linear pass.

2.3. A Greedier Weight-Ensembling Technique

A novel weight-ensembling algorithm to this work, the *greedier* algorithm's ingredient-set also initializes to the top validation-accuracy ingredient. At each step, we consider the inclusion of every remaining ingredient to the set, aggregating the candidate whose WA maximally performed to the set if we have outperformed the current set's WA accuracy. If no candidate set's WA has outperformed the current set WA's accuracy, the algorithm terminates. See Algorithm 1 for granular details. The core difference between the algorithms is that instead of the greedy algorithm's one single linear pass through the models sorted by individual performance, the greedier algorithm can add models in any order if they still contribute positively to the soup. The similarity is that both algorithms initialize the ingredients list to the maximal performing model.

While this algorithm has a costly runtime, it serves to illuminate the measures which maximally explain the selection mechanism of the algorithm. This will help diagnose what drives the selection of new ingredients to understand

what relationships between ingredients and the current WA contribute to maximal improvements in the WA. Treating greedier as a "gold-standard" benchmark also allows us to correlate other algorithms' selection mechanisms with their performance characteristics.

2.4. Ranked Weight-Ensembling

The next novel weight-ensembling algorithm to this work, the *ranked* algorithm, is initialized identically to previous algorithms. At each step, we sort remaining ingredients by decreasing distance from the current set's WA and proceed through these rankings considering the addition of each ingredient to the set individually. The first ingredient whose candidate set improves training domain accuracy is accepted, and the rejected models are stashed for the next iteration. See Algorithm 2. The intent of this method is to make salient the effect of biasing diversity into the selection mechanism, which allows us to tease out benefits which stem from the inclusion of diverse candidates and less of the confounding factors which may affect the greedier and greedy algorithms.

3. Experiments

3.1. Experimental Setting

We adopt the DomainBed setting (Gulrajani & Lopez-Paz, 2020) used by Rame et al. (2022), honing the focus of our experiments to the OfficeHome dataset (Venkateswara et al., 2017), a domain generalization dataset containing four test environments. We refer the reader to Section 3.2 of Rame et al. (2022) which established the setting adopted here, including the random initialization fine-tuning setup for ResNet50 (He et al., 2016) trained on ImageNet (Deng et al., 2009) as a foundation model varying hyperparameters and randomness to obtain distinct fine-tuned models. Holding out one test environment at a time as OOD set, we run ten trials of fine-tuning on the three ID environments 40 models per trial. After the fine-tuning, we run the greedier, greedy, and ranked weight-ensembling algorithms. For the ranked method, we run two versions, one which ranks ingredients by ratio-error which we call "diversity-ranked" and the other using Euclidean distance called "Euclidean-ranked."

3.2. Weight-Ensemble Algorithm Performance Results

To compare performance the performance of the methods, for each trial in all test environments we calculate the difference in accuracy at each iteration from the other weight-ensembling algorithms to the greedier method, visualizing the average difference in Figure 1. This amounts to benchmarking the performance of other methods relative to the greedier method. As some algorithms terminate before the ultimate time step $t = 39$, we propagate the terminal accuracy value forward through time in order to still be able to run our aforementioned calculation for subsequent time steps. Initially, the increasingly negative values show that the greedier algorithm gains accuracy faster than the ranked and greedy methods at the outset. As more ingredients are included past $t = 4$, ranked and greedy methods start to recover, with ranked methods suffering fewer losses and rebounding faster. This rebound occurs after many greedier runs have flatlined due to termination. Both ranked performances are similar, initially falling behind greedier but slowly recovering later on; greedy follows a similar trajectory, thus demonstrating that the greedier algorithm utilizes fewer ingredients effectively. ID validation ("training") accuracy for the non-greedier methods finish below the accuracy of greedier at a statistically significant level, while for OOD accuracy ("testing"), greedy closes below greedier at a statistically significant level with the ranked algorithms' intervals just barely enveloping the performance of greedier in the upper bound.

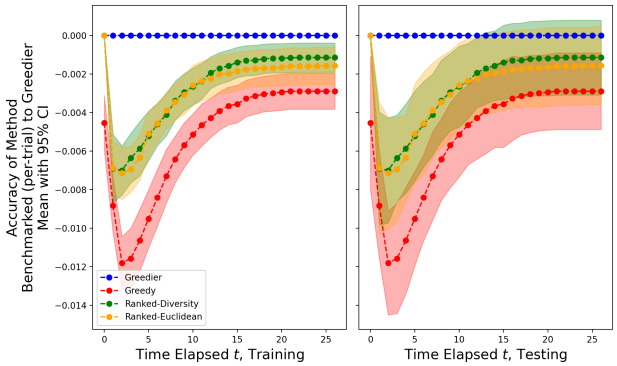

Figure 1. Difference between greedier accuracy and other methods' accuracy averaged across all trials with 95% confidence interval. Training at left, testing at right. Terminal value carried forward.

4. Explaining the Role of Diversity

4.1. Distributions of Quantiles over Selected Models

We next probe the extent to which distance measures (between the current WA to the candidate calculated ID when prediction is necessary) were associated with each algorithm's selection mechanism by binning the quantiles of distances of selected candidates to the current WA at each iteration time t. Since both greedier and ranked have a discrete host of models from which they can add a candidate at each step (but only select one), we first bin the quantile of the distance instead of the distance itself to make a well-posed statement about whether the more or less distant ingredients at some step were selected. For example, if θ_j had been the second least diverse from the current WA and had gotten selected for inclusion at time t, we would have binned $\frac{2}{\text{Num Remaining}}$ at time t. We can thus view each addition of a candidate as the discrete choice which was most useful to the WA. A similar procedure may be repeated for the greedy algorithm, where we only advance t in the

figure when candidates are actually accepted by greedy; if a candidate is skipped (getting thrown out for the remainder of the procedure), we do not advance t because in this visualization we are interested in the distances of *selected* ingredients with the current WA.

The distributions of the diversity-quantiles of selected models over time for each method is visualized in Figure 2, with mean trends compared side-by-side in Figure 6 of the Appendix. Over the greedier algorithm's first few iterations $t = 1$ to $t = 4$, we see that the distribution of quantiles is skewed to select ingredients which are *more-diverse* than random chance, reinforced by the confidence interval in Figure 6. This comes in direct opposition to the greedy algorithm, which specifically selects *less-diverse* candidates from its first iteration $t = 1$ up to iteration $t = 7$ as seen in Figure 6. As expected, the ranked-diversity algorithm tends to select ingredients which are most diverse from the current WA.

Contextualizing early-stage diversity-quantile results with the ID and OOD accuracy gains that greedier and ranked make relative to greedy between $t = 1$ to $t = 4$ in Figure 1, this shows a clear association between selecting the most diverse points and rapid accuracy improvement. Yet, the algorithm designed to select the most diverse candidates, ranked-diversity, still underperforms the greedier method. This comes in spite of ranked-diversity having ingredient models which have the highest average pairwise diversity, the proxy of diversity used in Rame et al. (2022) as we see in Figure 13. Although ranked-diversity's selection of more diverse ingredients seen in Figure 2 led to ranked-diversity WAs having the highest average pairwise diversity as seen in Figure 13, the fact that the greedier method outperforms the ranked-diversity method while having a less diversity-inducing selection mechanism indicate that the greedier method benefits from some force beyond what is provided for by diversity. Diversity correlates with the benefits realized by the greedier algorithm, but it does not quite encapsulate the full power of the greedier algorithm because building it into the selection mechanism with ranked-diversity does not achieve competitive accuracy. That in the latter stages past $t = 5$ of the greedier routine, both the diversity-quantiles of ingredients that we select for and the average pairwise diversity of the WA that is accepted seems to have saturated further evidence that the benefits of diversity are capped.

The quantiles of selected Euclidean distances are given by Figure 7 in the Appendix. In the figure, we observe an even stronger association between high-quantile Euclidean distance and selection by our greedier algorithm, with each boxplot living well-above random chance up to $t = 4$ when greedier's is making gains on algorithms' accuracies. The result is correlated with diversity selection, although clear differences in ranked-diversity and ranked-Euclidean distributions in Figures 6, 7 indicate some decoupling between the selecting for divesity and Euclidean distance. The strong association of Euclidean distance and greedier decision-making demonstrate that sampling far-apart ingredients in a loss basin can be just as powerful as selecting for diversity when it is quantified by ratio-error.

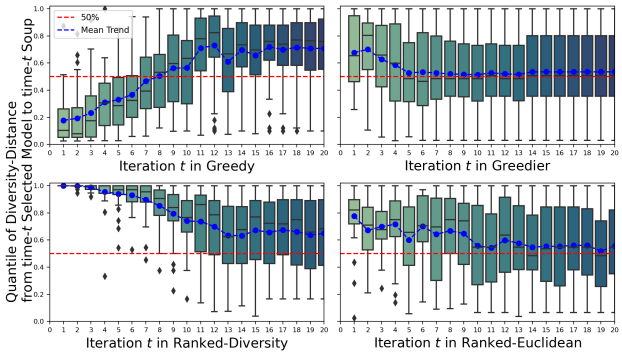

Figure 2. Box-plot of quantiles of diversity distance between the current WA and the selected model at each iteration t of each algorithm across the 40 trials. Dashed red-line at 50% indicates random selection.

4.2. Dynamics of Errors

In Section 4.1, we concluded that diverse ingredients are a beneficial component for a WA, although WAs may improve through other means and the benefit stemming from diversity is likely to saturate relatively quickly. In this section, we analyze the dynamics of how the errors that WAs make evolve as ingredients are added to the WA. The comparison of such dynamics across algorithms will make conspicuous the direct benefits and limitations of diversity for ID and OOD prediction, and how the greedier approach may exploit new ingredients better than ranked-diversity.

To this end, at each time step t (starting with $t = 1$ with 2 ingredients) we split up the ID and OOD sets into data points disjointly into four sets: 1) points which the WA at time t had classified incorrectly but the time-t selected ingredient classified correctly (not yet included in the latter WA), denoted "t-incorrect ingredient-correct", and similarly the disjoint sets 2) "t-correct ingredient-incorrect", 3) "t-correct ingredient-correct", and 4) "t-incorrect and ingredient-correct". For each of these sets are interested in the probability for some data point that the ingredient's outcome takes hold in the $t + 1$ WA: such as for 1) the probability that the time $t + 1$ WA is correct given that the time t WA was incorrect and the ingredient was correct.

As in previous analysis, given a weight-ensembling method and experimental trial, we benchmark each series with respect to the greedier method's series to contrast the methods. We carry forward the terminal values of each series before averaging. In Figure 3, we plot the first 10 iterations of the differenced series corresponding to t-incorrect ingredient-

correct. We plot the first ten only as the greedier method typically terminates within 10 time steps. For these early time periods, it is clear that the greedier algorithm makes relatively better use of its new ingredients in both the ID and OOD setting. Noticeably, we observe the importance of diversity for generalization in the OOD setting (right), where for the first several iterations algorithms ($t = 2, t = 3$ when the impact of adding a new ingredient is greatest due to weighting), selecting for diversity induces approximately a 10% greater probability of being the current-step WA's mistake corrected by the ingredient. The decreasing trend of other non-greedier beyond $t = 5$ correspond to many cases in which greedier has terminated, and other methods (which may still be running) already include more ingredients and thus it is more difficult for any newly-added individual to strongly impact the result. Figures 15, 16 in the appendix demonstrate that the selection of diverse candidates makes scenarios 2) and 4), in which new errors appear or existing ones are retained, less likely.

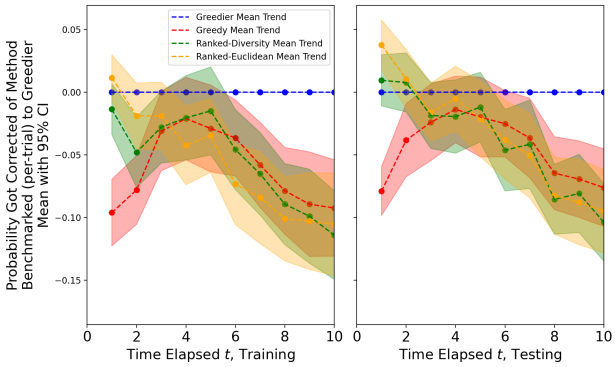

Figure 3. t-incorrect ingredient-correct: Probabilities that the next-step WA predicts correctly given that the current-step WA was incorrect and the ingredient was correct, difference from greedier and other methods' averaged across all trials with 95% confidence interval. Training at left, testing at right. Terminal value carried forward.

5. Visualization of Weight-Ensembling

5.1. MDS Visualization of the Greedier Algorithm

We develop a visualization method to capture all pairs of distance relationships jointly without reducing the diversity to a single number, as does average pairwise diversity. We do so with Multidimensional Scaling (MDS), a dimensionality reduction algorithm designed to preserve pairwise distances.

After running a weight-ensembling algorithm on the set of k neural networks, we calculate all pairwise distances between the models evaluated in the experiment. We store these distances in a symmetric distance matrix, then use this matrix as a plug-in to MDS. We use metric MDS for Euclidean distance, and for diversity we use non-metric MDS. At each iteration t, we reveal in the decomposed space the candidate WAs that we considered at this time using our current set of ingredients and the remaining ingredients. Due to the structure of MDS, we thus visualize the pairwise distances between the candidate WAs, the current and past WAs, and the individual candidates which shed qualitative insight on the selection mechanism of the greedier algorithm.

We demonstrate the progression of the greedier algorithm through Euclidean distance in one example in Figure 11 (in Appendix due to figure size). Consistent with the results of quantile-ranks in Euclidean space, we see for iterations $t = 2, 3, 4$ that we have selected points in space which were on the larger end of Euclidean distance from the current WA; at $t = 5$ we have saturated accuracy and the algorithm has terminated. We can see convergence in decomposed weight space, reflecting the diminishing returns from adding more candidates after location in Euclidean space saturates. Turning to the diversity distance Figure 12 we observe that the WA points do not converge as nicely as time passes through the algorithm, although we still manage to select candidates which tend to lie on the farther side from our current WA through 4 iterations. Finally, we observe the intuition allotted by the visualization technique: we have selected truly distinct points in weight space, as the selected candidates exhibit separation in both decomposed spaces, as opposed to the comparitively reductive average pairwise diversity or Euclidean distance in the literature.

6. Discussion

We have introduced the greedier algorithm which at each iteration selects the ingredient which maximally improved the WA's ID validation performance. When treated as a "gold-standard," greedier's decision-making helps to uncover how relationships between ingredients and a WA are leveraged to best improve the WA's performance. We also propose the ranked algorithm, which at each iteration sorts ingredients by their distance from the WA and selects the first to improve ID validation accuracy. We can contrast greedier results with the ranked and greedy algorithms, using different behaviors to reason about the role that diversity plays in performance dynamics. Leveraging this structure, we identify that both high diversity distance and high Euclidean distance explain the selection method when performance improves the fastest, implying that selecting diverse or spaced out candidates contributes rapidly towards improving the WA. Yet, that the ranked-diversity algorithm does not match the greedier method ID or OOD limit the extent to which diversity plays a role in this performance. We finally introduce a method by which we can examine how our algorithms selection traverses a loss basin and whether our candidates are truly diverse by not reducing our distance relationships but rather leveraging pairwise structure in a decomposition for qualitative analysis.

Impact Statement

This work provides a new weight-ensembling algorithm and explanations for why a WA procedure improves accuracy when able to choose which model to add to the list of candidates using the greedier algorithm. By shedding light on what may cause WA to improve, we provide new support to an existing avenue by which they can improve inference using the WA efficiently. Such work has the potential to improve deep learning algorithms in all facets of society, both for clearly positive (such as medicine) and more nuanced to negative (such as surveillance) purposes.

References

Aksela, M. Comparison of classifier selection methods for improving committee performance. In *International Workshop on Multiple Classifier Systems*, pp. 84–93. Springer, 2003.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear Mode Connectivity and the Lottery Ticket Hypothesis, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Neyshabur, B., Sedghi, H., and Zhang, C. What is Being Transferred in Transfer Learning?, 2021.

Rame, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., and Cord, M. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

A. Quantiles and Distance Distributions from Iterations of Weight-Ensemble Algorithms to Selected Ingredients

In this section, we provide the boxplots of distances and quantiles of the selected models at each iteration using our weight-space geometric distance measures. We also plot their mean trends together for a closer comparison.

A.1. Distance Distributions

In Section 4.1, the quantiles of the distances of selected ingredients from a current soup has the advantage of elucidating within a set of ingredients whether the more or less diverse ingredients were utile. In contrast, we examine here the raw trends (without taking quantiles) to ensure the robustness of the analysis.

For example, we can explain the first iteration of this process through the lens of diversity distance. Following notation in Algorithm 1, at $t = 1$ we know the current WA is equal to the model average(ingredients) $= \theta_1$. Then if we selected θ_j to add to the ingredients using the greedier method, we store the result $d_D(\theta_1, \theta_j)$ in our bin for $t = 1$. We proceed this binning from $t = 1$ to $t = T_{max}$ where T_{max} is equal to the largest amount of time any greedier algorithm instance ran for. We then boxplot over each t. As in the quantile example, we have an analogous implementation for the greedy algorithm. These results are similarly in favor of the higher diversity and Euclidean distances being selected by the greedier algorithm in earlier iterations. In both but especially in the Euclidean res ult of Figure 8 we see a steep drop-off in selected distances after $t = 4$. Such a dropoff is intuitive because as we roughly move towards center of the the points in weight space, our distance to unincorporated but likely related points will also decrease.

A.2. Diversity Distance

The distributions over selected-ingredient diversity distances is visualized in Figure 4. We also plot the mean trends in selected diversity distance with confidence intervals in one plot in Figure 5. Elaborating on Figure 2, we plot the mean trend in the quantiles of the distances of selected ingredients in one plot in Figure 6.

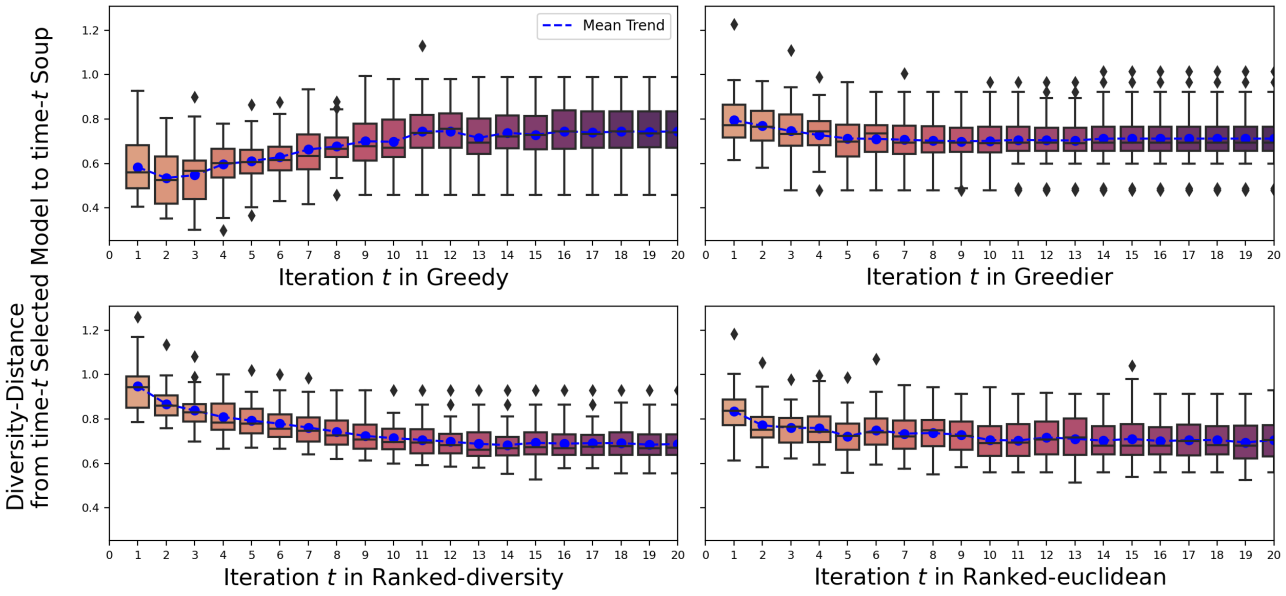

Figure 4. Box-plot of diversity distance between the current WA and the selected model at each iteration t of each algorithm across the 40 trials.

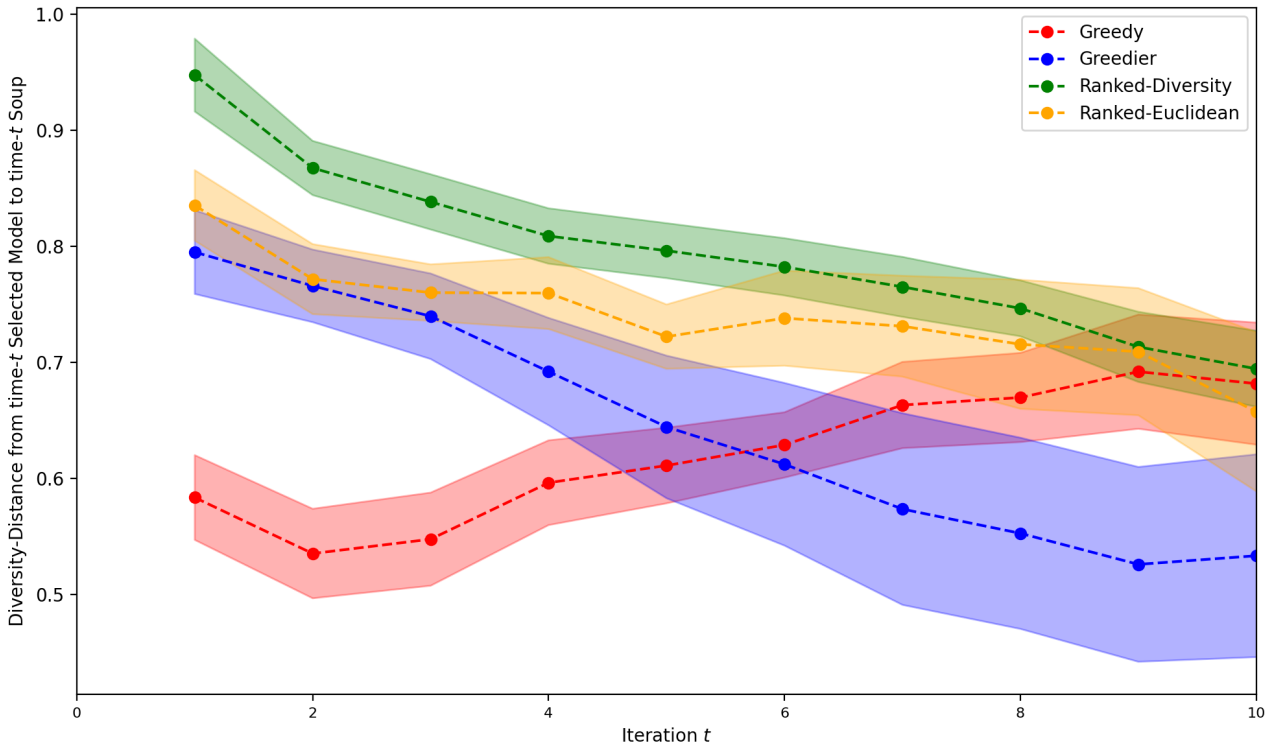

Figure 5. Diversity distance between the current WA and the selected model at each iteration t of the greedy, greedier, ranked-diveristy, and ranked-Euclidean algorithms averaged across all trials with 95% confidence interval. First ten iterations plotted.

A.3. Euclidean Distance

The quantiles over selected-ingredient Euclidean distances is visualized in Figure 7. The distributions over selected-ingredient Euclidean distances is visualized in Figure 8. We also plot the mean trends in selected diversity distance with confidence intervals in one plot in Figure 9. Elaborating on Figure 7, we plot the mean trend in the quantiles of the distances of selected ingredients in one plot in Figure 10.

B. MDS Extended

We demonstrate an example of the MDS visualization in Euclidean space (Figure 11) and diversity space (Figure 12).

Remaining experiments will be attached as supplementary material.

C. Analyzing Diversity and Errors

C.1. Average Pairwise Diversity

Given a collection of models $\{\theta_1, \ldots, \theta_k\}$ it is unclear how to measure the total diversity of the collection because the ratio-error diversity is defined via pairwise relationships. As such, (Rame et al., 2022) choose to represent the diversity of the collection as the average pairwise diversity between any two distinct models in the collection. In these algorithms at any time step, we can use the ingredients selected so far to calculate the average pairwise diversity. In Figure 13 we plot the average pairwise diversity up to time step 20.

In only the greedier algorithm at each time-step do we have access to the result of including each remaining ingredients with the current set. As such, we may calculate the average pairwise diversity of every WA from the time step and bin the quantile of the average pairwise diversity of the selected model. This allows us to see whether WAs with a higher or lower

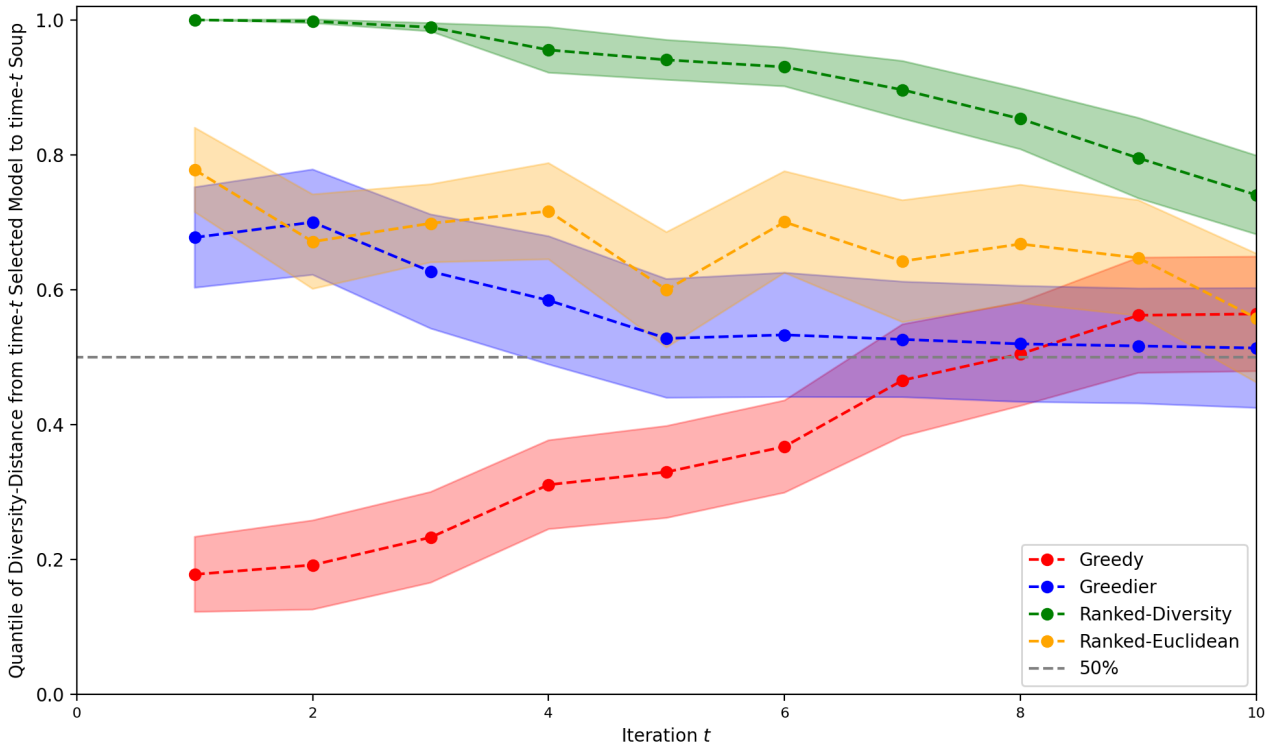

Figure 6. Quantiles of Diversity distance between the current WA and the selected model at each iteration t of the greedy, greedier, ranked-diveristy, and ranked-Euclidean algorithms averaged across the 40 runs with 95% confidence interval. First ten time steps plotted.

average pairwise diversity are selected. Figure 14 visualizes the mean trend in the quantile of the new WA at each time step.

C.2. Dynamics of Errors

Similar to the benchmarking of the probability trend in Figure 3, we plot the results for 2) t-correct ingredient-incorrect in Figure 15, and 4) t-incorrect and ingredient-correct in Figure 16. We do not plot 3) t-correct ingredient-correct due to a lack of observable trend.

From Figure 15, we observe in the ID and OOD case for early iterations up to $t = 4$ the greedy method is signifcantly more likely than greedier to make a "new" error when the ingredient made the error as well. Most of the distribution of values for the diversity-ranked also lie below that of the greedy algorithm, although there is some intersection of the confidence intervals. This evidences that diversity between a WA and an ingredient which makes mistakes may contribute to being more robust to the next-step WA making the same error.

In Figure 16, we observe that the the diversity-selecting methods and the greedier algorithm lower probabilities of retaining a current WA's errors when the ingredient was also correct. This demonstrates that if an ingredient (though well-trained) is incorrect, the WA's performance on previously-misclassified points will benefit more from its inclusion if the model is diverse from the current WA.

D. Algorithms

We formalize our greedier algorithm below as Algorithm 1. Notation for Algorithm 1 is drawn from (Wortsman et al., 2022).

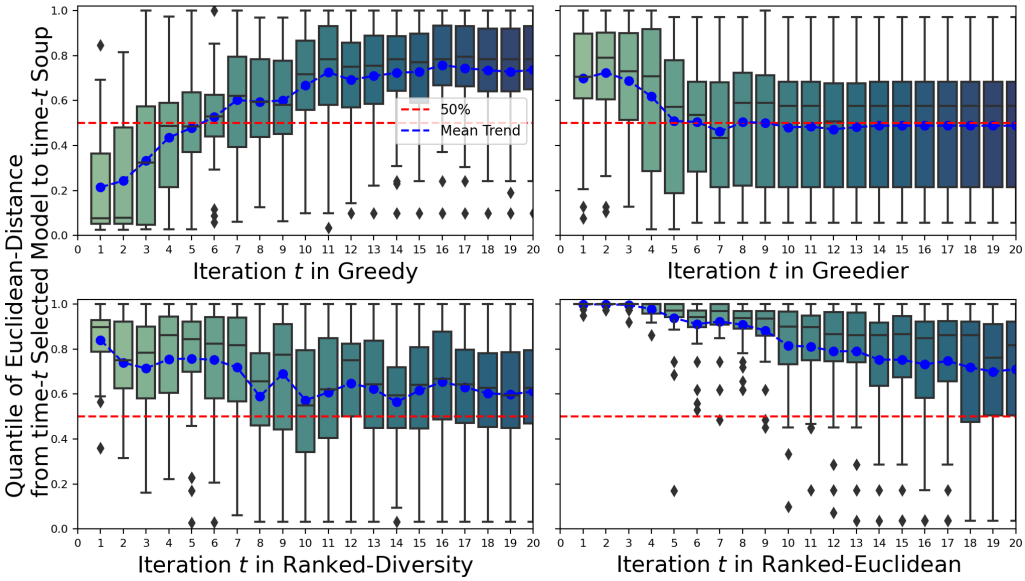

Figure 7. Box-plot of quantiles of Euclidean distance between the current WA and the selected model at each iteration t of each algorithm across the 40 trials. Dashed red-line at 50% indicates random selection.

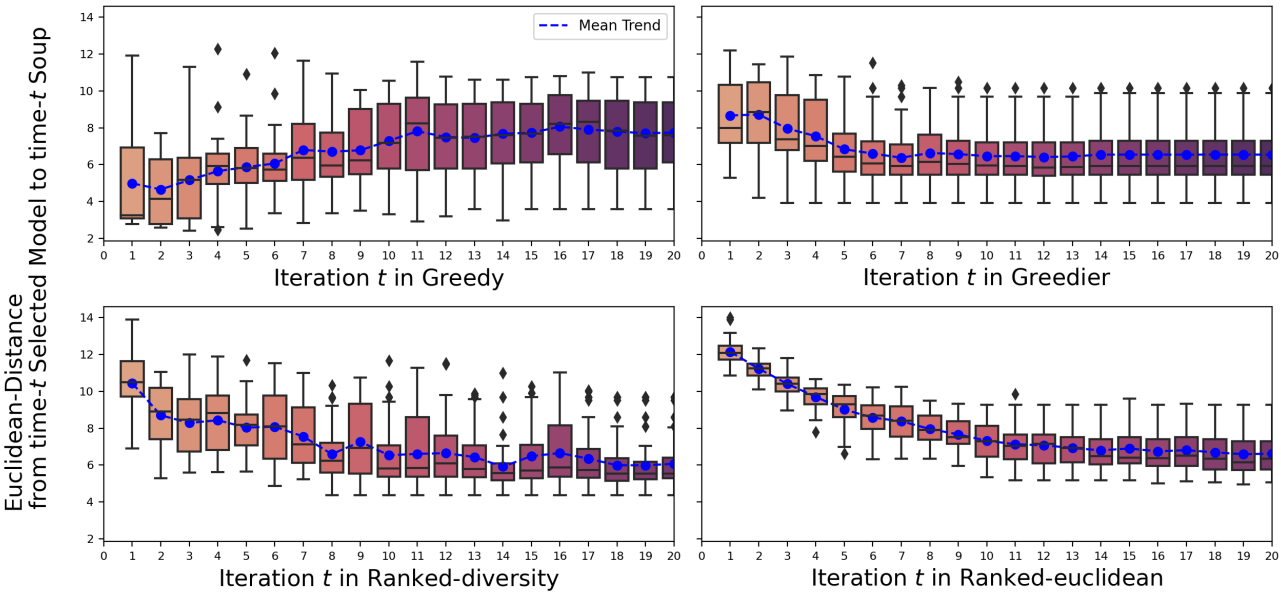

Figure 8. Box-plot of Euclidean distance between the current WA and the selected model at each iteration t of each algorithm across the 40 trials.

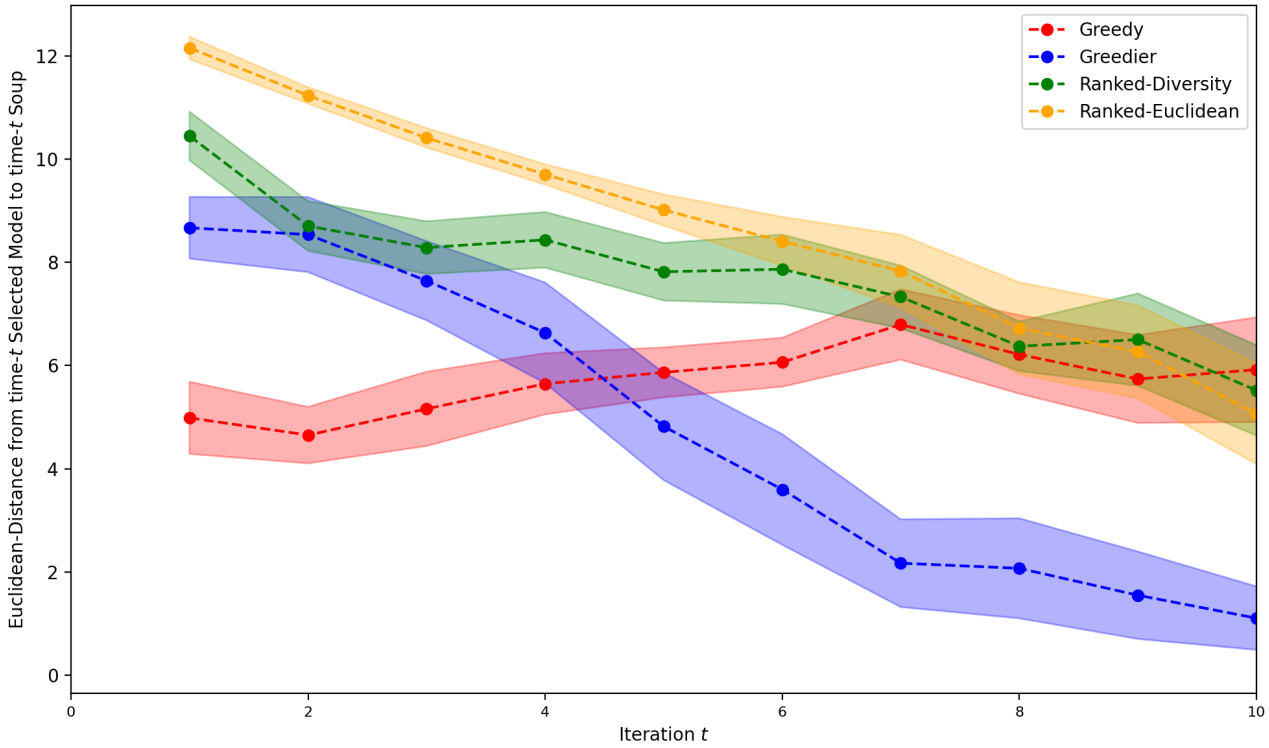

Figure 9. Euclidean distance between the current WA and the selected model at each iteration t of the greedy, greedier, ranked-diveristy, and ranked-Euclidean algorithms averaged across all trials with 95% confidence interval. First ten iterations plotted.

Algorithm 1 Greedier Algorithm for Weight-Ensembling

Input: Fine-tuned, potential ingredients $\{\theta_1, \ldots, \theta_k\}$, sorted by ID validation accuracy in the training domain, and a choice of diversity metric.
Initialize ingredients $\leftarrow \{\theta_1\}$
Initialize remaining ingredients $\leftarrow \{\theta_2, \ldots, \theta_j\}$
Initialize MaxAcc $\leftarrow$ ValAcc(θ_1).
for $i = 2$ **to** k **do**
  Best$_i \leftarrow 0$
  for θ_j in remaining ingredients **do**
    if ValAcc(average(ingredients $\cup \{\theta_j\}$)) $\geq$ MaxAcc **then**
      MaxAcc $\leftarrow$ ValAcc(average(ingredients $\cup \{\theta_j\}$))
      Best$_i \leftarrow j$
    end if
  end for
  if Best$_i > 0$ **then**
    ingredients $\leftarrow$ ingredients $\cup \{\theta_j\}$
    remaining ingredients $\leftarrow$ remaining ingredients $\setminus \{\theta_j\}$
  else
    return average(ingredients)
  end if
end for
return average(ingredients)

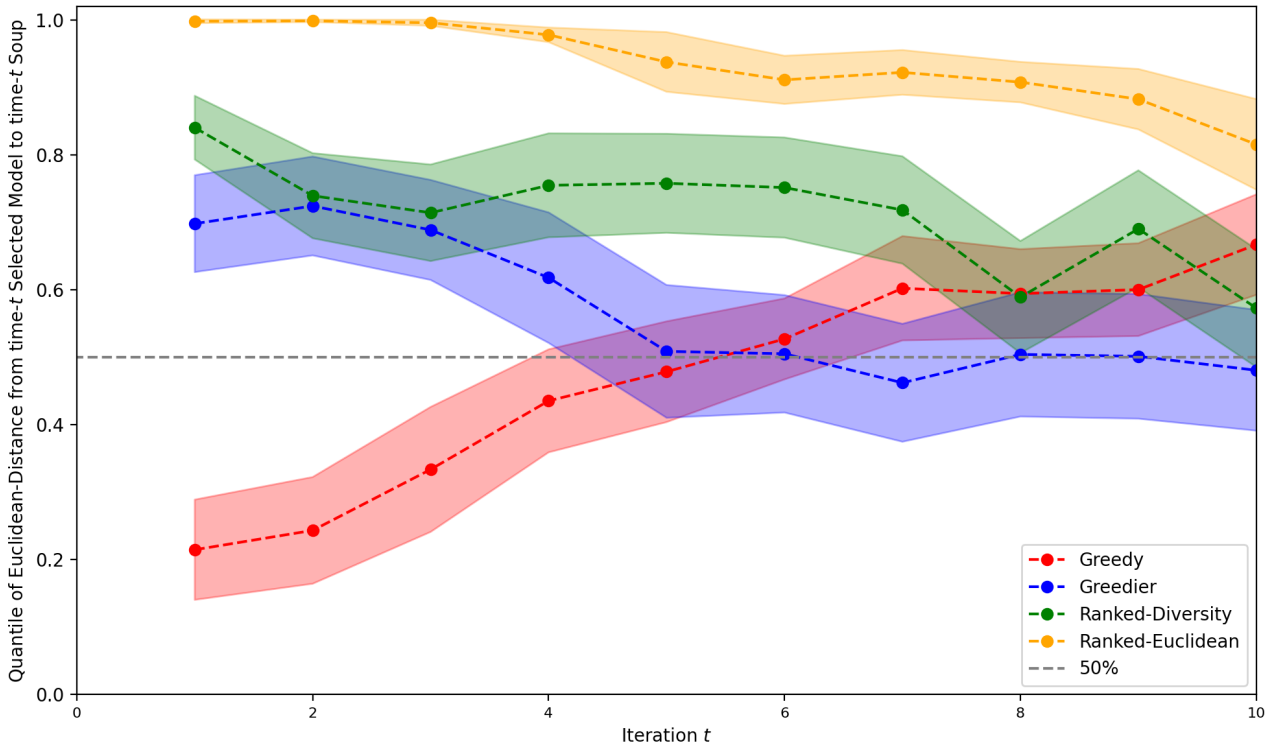

Figure 10. Quantiles of Euclidean distance between the current WA and the selected model at each iteration t of the greedy, greedier, ranked-diveristy, and ranked-Euclidean algorithms averaged across the 40 runs with 95% confidence interval. First ten time steps plotted.

Algorithm 2 Ranked Algorithm for Weight-Ensembling

Input: Fine-tuned, potential ingredients $\{\theta_1, \ldots, \theta_k\}$, sorted by ID validation accuracy in the training domain
Initialize ingredients $\leftarrow \{\theta_1\}$
Initialize remaining ingredients $\leftarrow \{\theta_2, \ldots, \theta_j\}$
Initialize MaxAcc $\leftarrow$ ValAcc(θ_1).
for $i = 2$ **to** k **do**
  Best$_i \leftarrow 0$
  Calculate diversity-metric between average(ingredients $\cup \{\theta_j\}$) and each remaining ingredient
  for θ_j Diversity-Ranked-Descending(remaining ingredients) **do**
    if ValAcc(average(ingredients $\cup \{\theta_j\}$)) $\geq$ MaxAcc **then**
      MaxAcc $\leftarrow$ ValAcc(average(ingredients $\cup \{\theta_j\}$))
      Best$_i \leftarrow j$
      Break loop
    end if
  end for
  if Best$_i > 0$ **then**
    ingredients $\leftarrow$ ingredients $\cup \{\theta_j\}$
    remaining ingredients $\leftarrow$ remaining ingredients $\setminus \{\theta_j\}$
  else
    return average(ingredients)
  end if
end for
return average(ingredients)

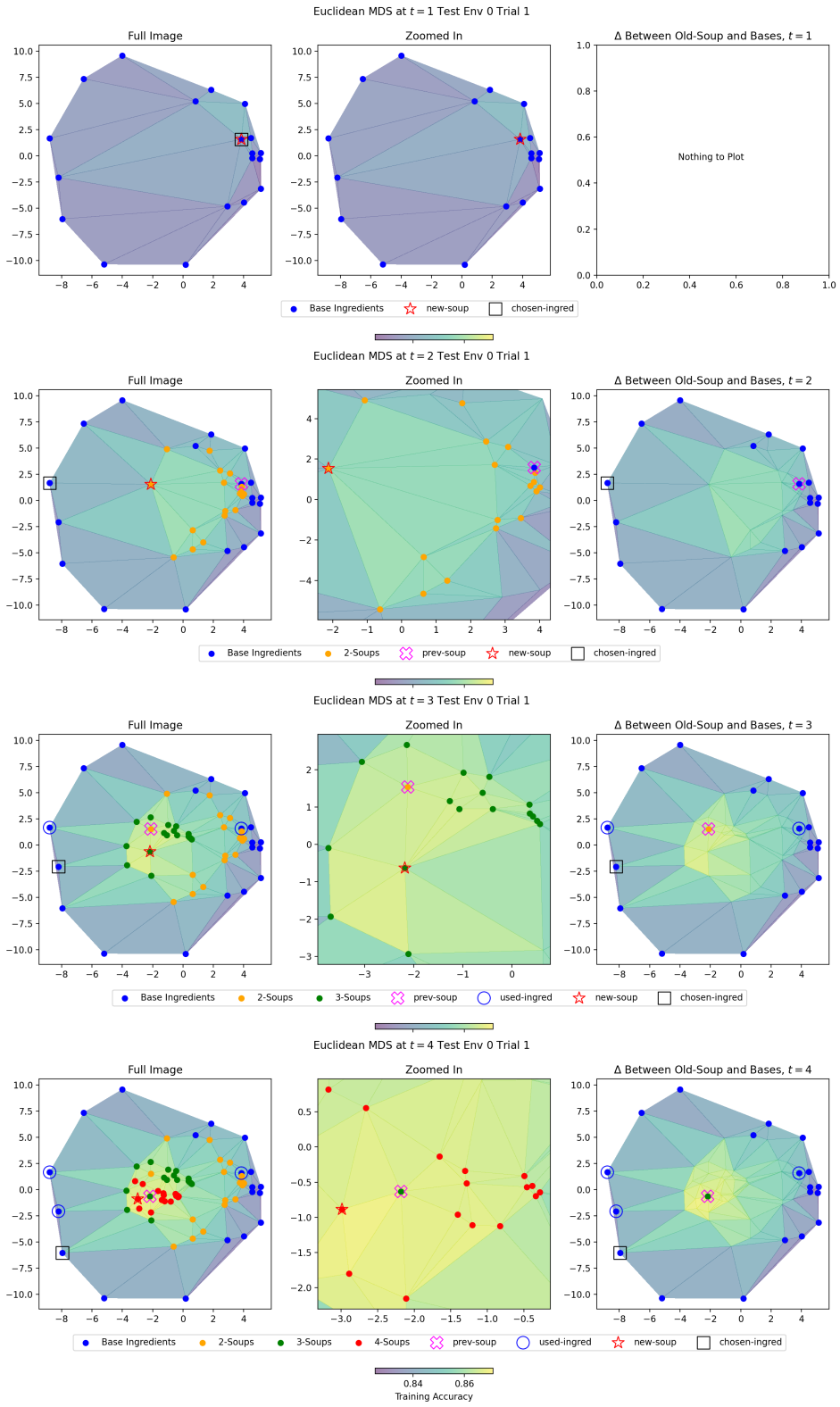

Figure 11. Euclidean distances between models in the greedier procedure plugged into MDS, with points color-coded by number of ingredients. Previously used ingredients are circled, currently selected ones are in a square, and the current WA is in an x. Backdropped by triangulated accuracy. Left: all points up to and including time t. Center: zoomed in on previous WA and time t candidate WAs. Right: current WA and individual candidates. Smaller experiment (20 candidates) is visualized.

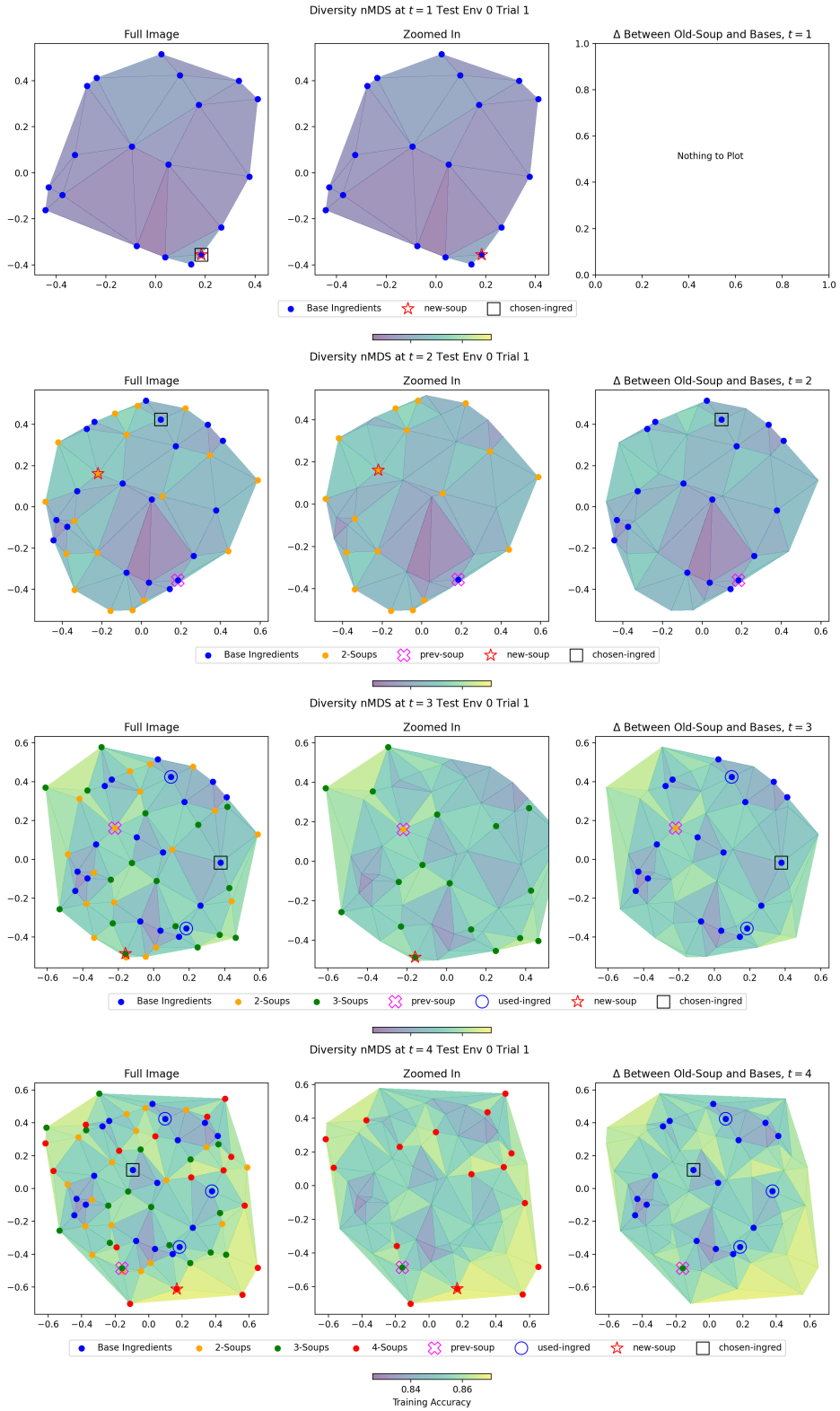

Figure 12. Diversity distances between models in the greedier procedure plugged into non-metric MDS, with points color-coded by number of ingredients. Previously used ingredients are circled, currently selected ones are in a square, and the current WA is in an x. Backdropped by accuracy. Left: all points up to and including time t. Center: zoomed in on previous WA and time t candidate WAs. Right: current WA and individual candidates. Smaller experiment (20 candidates) is visualized.

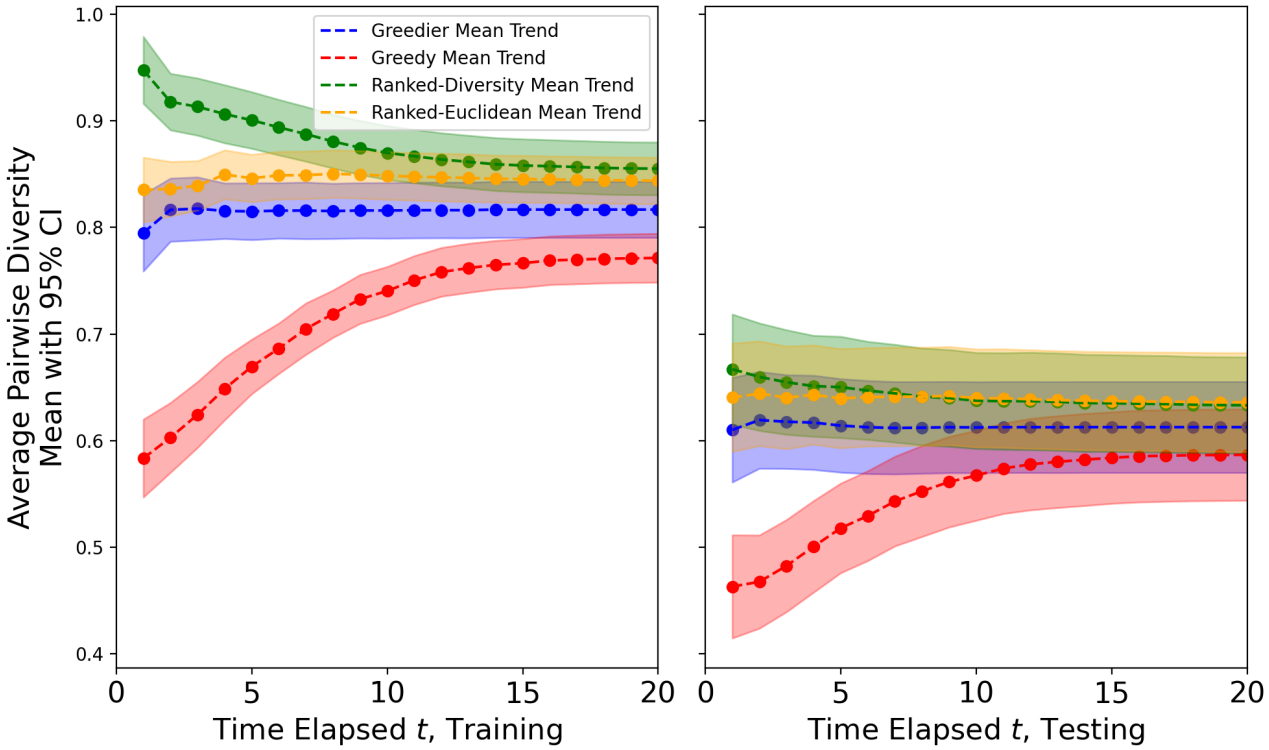

Figure 13. Average pairwise diversity through time for all algorithms across the 40 trials.

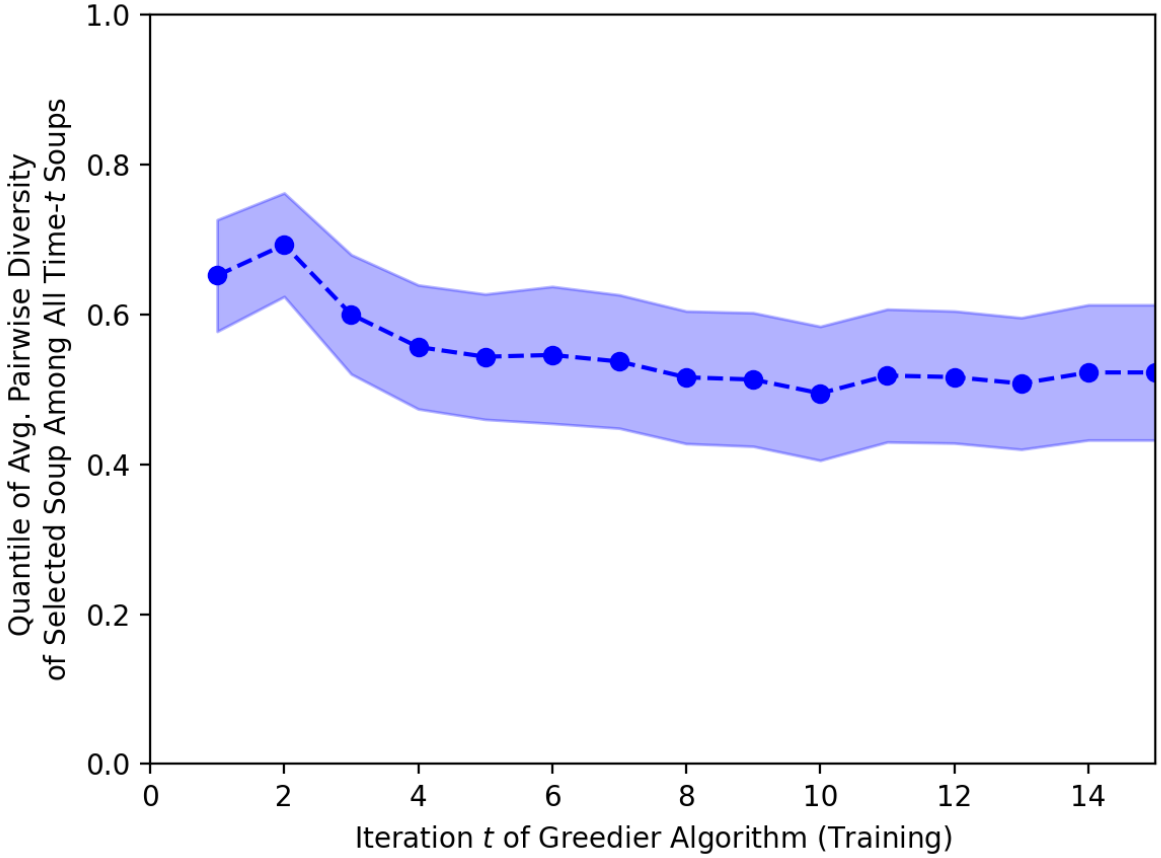

Figure 14. Quantile of average pairwise diversity of selected model through time for all algorithms across the 40 runs.

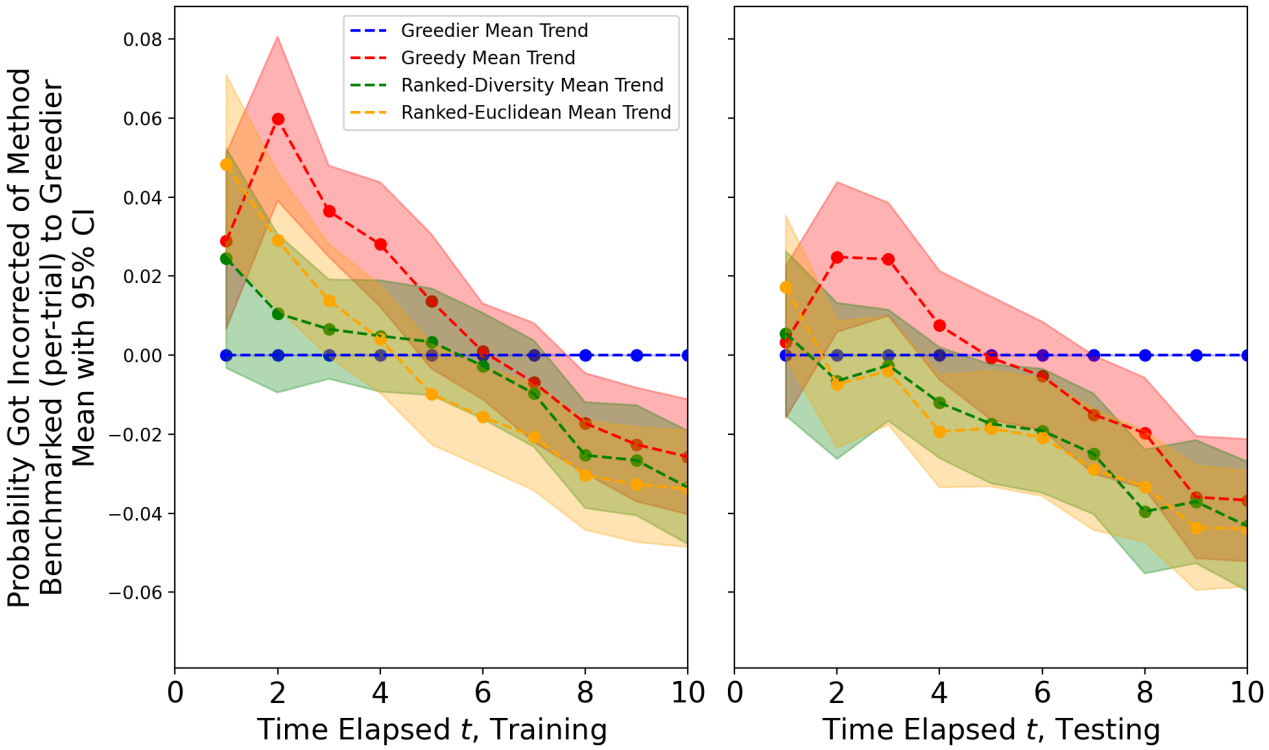

Figure 15. t-correct ingredient-incorrect: Series of probabilities that the next-step WA predicts incorrectly given that the current-step WA was correct and the ingredient was incorrect, difference from greedier and other methods' averaged across all trials with 95% confidence interval. Training at left, testing at right. Terminal value carried forward.

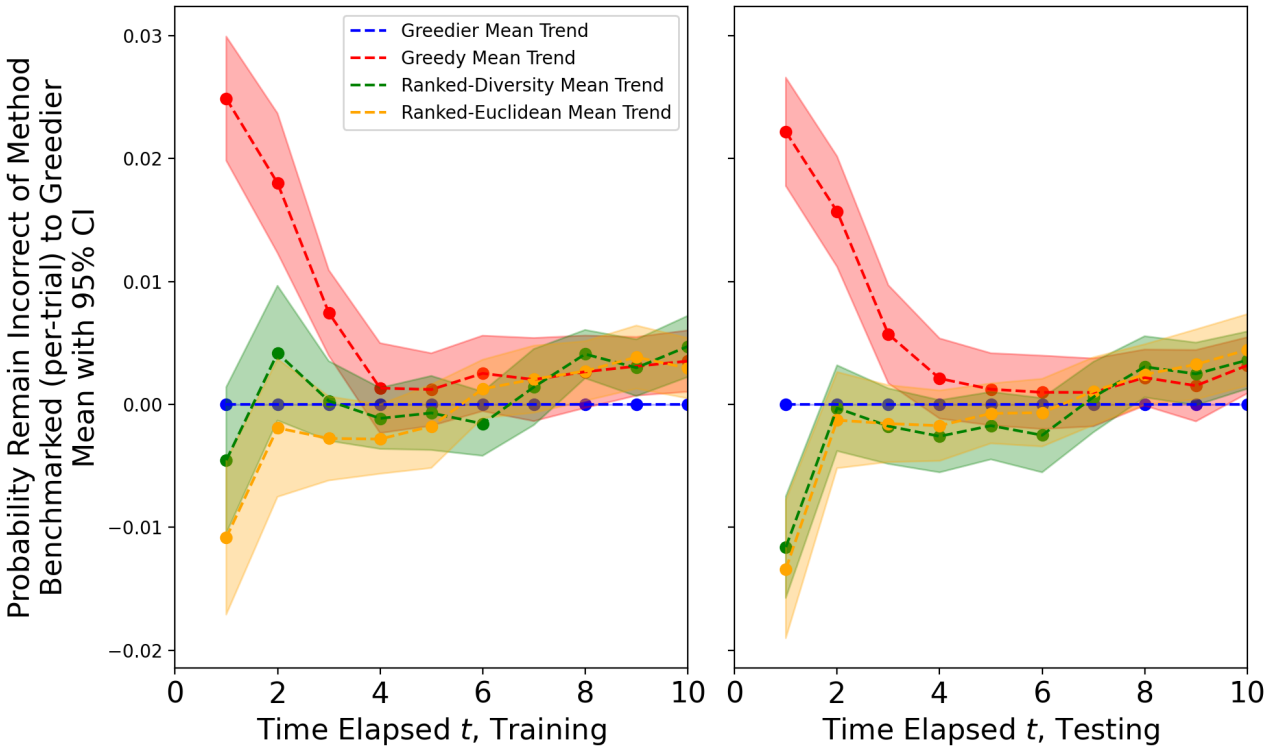

Figure 16. t-incorrect ingredient-incorrect: Probabilities that the next-step WA predicts incorrectly given that the current-step WA was incorrect and the ingredient was incorrect, difference from greedier and other methods' averaged across all trials with 95% confidence interval. Training at left, testing at right. Terminal value carried forward.