Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What composition selection buys in heterogeneous ensembles, and why oracle bounds overstate it

Muhammetalp Erdem

alperdemdev@gmail.com

Trabzon University

Research Article

Keywords: Ensemble learning, Model selection, Heterogeneous ensembles, Selection bias, Oracle bounds, Forecast combination

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10666378/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

What composition selection buys in heterogeneous ensembles, and why oracle bounds overstate it

Muhammetalp Erdem[1*]

[1*]Computer Engineering, Trabzon University, Söğütlü, Trabzon, 61000, Turkey.

Corresponding author(s). E-mail(s): alperdemdev@gmail.com;

Abstract

Heterogeneous ensembles are commonly built by pooling members drawn from several homogeneous families, and a line of work selects how many members to draw from each family by searching the composition simplex. We ask what that search buys. Across 30 public classification datasets and 10 seeds, with the ensemble budget fixed at 300 members split among random forests, extremely randomised trees and bagged nearest neighbours, we compare six selectors, ranging from a fixed uniform mix to bagged and shrinkage variants of the validation argmax. Any selection beats a uniform mix by 0.68 accuracy points (Wilcoxon $p = 2.4 \times 10^{-5}$). However, no composition search beats simply selecting the best single family: all five selectors fall within the Nemenyi critical difference of one another. We then show that the headroom which motivates better selectors does not exist. An oracle chosen on the test set appears to leave 1.08 accuracy points unclaimed, whereas a cross-fitted oracle, which selects on one half of the test set and is evaluated on the other, does not beat a single argmax. The apparent headroom is therefore selection bias rather than unrealised potential. We recommend that oracle and best-subset bounds over large correlated candidate sets be cross-fitted before they are reported, and that practitioners choose a family rather than a composition.

Keywords: Ensemble learning, Model selection, Heterogeneous ensembles, Selection bias, Oracle bounds, Forecast combination

1 Introduction

An ensemble whose members are all produced by a single learning algorithm, such as a random forest or a bagged ensemble of trees, is homogeneous. Combining members from several such families produces a heterogeneous ensemble, and a natural design question follows immediately: under a fixed budget of T members, how many should be drawn from each family? The answer defines a point on a simplex, and a line of work treats finding that point as a selection problem to be solved by search, typically scoring candidate compositions on held-out or out-of-bag data and taking the best.

The premise of that search is that composition matters and that a better selector would capture more of what it offers. Both halves deserve testing, and the second half is usually tested in a way that guarantees a favourable answer: papers in this area often report an *oracle* or *best-subset* number, computed by taking the maximum over all candidates using the evaluation labels, and treat the gap between the deployed selector and that maximum as recoverable headroom. The candidate set is large and its members are highly correlated, since neighbouring compositions differ by only a few members out of hundreds. The maximum of many correlated noisy estimates therefore exceeds the true maximum. The reported headroom is then partly, and possibly wholly, an artefact of the reporting procedure.

We take as our reference point the formulation of [1], which builds a heterogeneous ensemble by pooling homogeneous ones and selects the proportion assigned to each family by searching the simplex, identifying the optimum out-of-bag. The design is sensible and the reported gains are real. What makes it a useful reference point is that its authors are explicit about three difficulties, all of which our results speak to directly. First, they report that the optimal proportion is strongly problem-dependent, with no universal composition. Second, that the search remains intractable in the number of families, so the method is practical only for small pools. Third, and most tellingly, that on smaller datasets the out-of-bag identification of the optimum simplex point is less accurate.

Read together, these three observations describe a selection problem in which the signal is weak, the search space grows quickly, and the estimate of the objective is unreliable exactly where data is scarce. That is the profile of a problem dominated by estimation noise rather than by the difficulty of the search. If that reading is right, three things should follow, and each is testable: a better-conditioned selector should not help much, selecting at a coarser granularity should do as well as selecting at a fine one, and the apparent headroom above any deployed selector should shrink or vanish once the bound defining it is required to generalise.

We hold the ensemble budget fixed at M members, enumerate the composition simplex, and compare six selectors that differ in how hard and how noisy their selection is. Because a composition is a prefix-average of cached per-member predictions, every selector and every point of the simplex is evaluated from three ensemble fits per split. The entire study runs in under half an hour on a CPU. We then apply to the oracle the diagnostic that this literature applies to its models but not to its bounds: alongside the conventional oracle, we report a cross-fitted oracle that selects on one half of the evaluation set and is scored on the other. The difference between the two is the part of the reported headroom that no selector could ever deliver.

Composition is not irrelevant: a fixed uniform mix loses to every selector by 0.68 accuracy points on average ($p = 2.4 \times 10^{-5}$), and by 1.47 points when a fourth family is added. However, nothing beats selecting a single family. All five selectors fall inside the Nemenyi critical difference of one another, and neither bagged selection nor shrinkage improves on a plain argmax ($p = 0.42$ and $p = 0.93$). The reason is that there is nothing left to win: the conventional oracle appears to leave +0.0108 accuracy on the table, whereas the cross-fitted oracle delivers −0.0012.

The mechanism behind these results is not new and we state its origin at the outset. Since combination weights were introduced [2], the forecasting literature has repeatedly found that equal weighting is competitive with weights estimated from data. This pattern, known as the forecast combination puzzle, has been attributed to estimation error in the weights [3, 4]. It has since been developed into an explicit bias–variance case for shrinking weights toward equality [5]. Our shrinkage selector is that idea transplanted onto a different object: proportions across exchangeable families under a fixed budget, rather than weights on individual models. The contribution here is the measurement, not the mechanism.

Contributions.

1. A controlled comparison of six composition selectors on 30 public datasets with 10 seeds, reporting accuracy, macro-F1 and MCC, with Friedman and Nemenyi analysis (Sections 5.1–5.2). Any selection beats a uniform mix; no search beats selecting a single family.
2. A measurement of selection-on-test bias in oracle bounds over the composition simplex (Section 5.3). The apparent headroom of +0.0108 accuracy becomes −0.0012 under cross-fitting: 111% of it is bias.
3. A geometric account of why (Section 5.4). The accuracy surface is a broad plateau whose internal variation is, on 12 of 30 datasets, smaller than seed-to-seed noise at a fixed composition, with the ratio scaling with sample size.
4. A controlled test of the obvious objection (Section 5.5). Adding a fourth, linear family increases composition signal yet leaves both conclusions intact, and enlarging the candidate set raises the apparent headroom by 25% while the honest headroom moves further below zero.

Two practical recommendations follow. For practitioners: choose the family, not the composition, and do not accept a uniform mix as a default. For the literature: cross-fit oracle and best-subset bounds over large correlated candidate sets, or do not report them as headroom.

We are careful about what is *not* claimed. We do not claim priority over the mechanism, which belongs to the forecasting literature (Section 2.3), nor over cross-fitting, which is standard (Section 2.4). We do not refute [1]: our selector scores on a held-out split rather than out-of-bag, and our families are not theirs (Section 7). The claim is about composition search as a class, under a protocol that measures what any selector could achieve rather than what one selector did.

2 Related work

Four threads bear on this study. The first two supply the object, the third supplies the mechanism we import, and the fourth supplies the diagnostic we apply to the second.

2.1 Homogeneous families and their heterogeneous combination

The families we pool are standard. Bagging [6] perturbs the training sample; the random subspace method [7] perturbs the feature set; random forests [8] combine both, and extremely randomised trees [9] push the randomisation into split selection itself. Each yields *exchangeable* members, which is the property Equation (2) requires.

Combining members of *different* families is older still. Stacked generalisation [10] learns a combiner over base-learner outputs, and modern descendants weight heterogeneous members by cross-validated accuracy estimates [11]. Within the target venue, one line takes the specific form we study: pooling several homogeneous ensembles under a fixed budget and searching the proportion assigned to each family [1].

The general case for ensembles rests on three mechanisms (statistical, computational and representational) [12]. Moreover, the field is surveyed at book length [13] and in review form [14, 15]. What that literature establishes is that *combining* diverse learners helps. It says considerably less about how to divide a fixed budget among them, which is the question here and a strictly narrower one.

Why the theory already predicts a weak effect.

Breiman's variance decomposition [8] gives the ensemble error of a bagged predictor in terms of individual member strength and the mean pairwise correlation $\bar{\rho}$ between members. Families differing in inductive bias help principally by lowering $\bar{\rho}$; once a pool contains such families, the *proportions* among them enter only through their effect on the same aggregate correlation, which is a second-order dependence. This is the theoretical reason to expect a plateau rather than a peak, and Section 5.4 measures one.

The empirical diversity literature points the same way. Pairwise and non-pairwise diversity measures have been catalogued and tested for two decades [16, 17], and the recurring finding is that they predict ensemble accuracy only weakly. Recent theory clarifies why: diversity is a term in a bias–variance–diversity decomposition whose components trade off [18], and its benefit to generalisation is conditional rather than universal [19].

A note on gradient boosting.

The most widely used modern ensemble is not a bagged pool at all: gradient boosting fits members sequentially to the residuals of their predecessors [20]. That construction is why it is absent from our pool (Section 4). Its members are not exchangeable, so "the first c_k members" is a shorter boosting run rather than a random subset, and a composition containing it would not be a composition in the sense of Equation (1). This is a boundary on the scope of our claim rather than an oversight. Work on making boosting cheaper proceeds on a separate track [21].

2.2 Selecting members, and the bounds this literature reports

Selecting a subset of members is a large field. Early work pruned boosted ensembles [22]; Zhou et al. showed that a selected subset can beat the full ensemble [23]; ordered aggregation gives a principled ordering of bagged members [24, 25]; greedy forward selection from a large model library scales the idea to heterogeneous pools [26]; and the problem has been cast as semi-definite programming [27], as directed hill-climbing under an uncertainty-aware criterion [28], and as bi-objective optimisation with guarantees [29]. Surveys organise these into ranking-, clustering- and optimisation-based families [30]. In the target venue, clustering-based forest pruning [31] and clustering-ensemble member selection [32] both appear, the latter explicitly deferring a selective strategy for choosing members to future work.

Two further strands deserve separating from ours, because both are sometimes described as "ensemble selection" and neither addresses composition under a fixed budget. The first is *ordering-based* pruning, which ranks members by a scalar criterion and keeps a prefix. The criterion may be individual contribution [33] or a combination of margin and diversity [34]. Search-based variants replace the ranking with an explicit optimiser [35]. The second is *dynamic* selection, which chooses a different subset for each test instance according to local competence rather than a single subset for all of them [36–38]. Dynamic selection is a genuinely different mechanism: our results bound what a *static* composition can achieve and say nothing about per-instance schemes.

A third, closer question asks not *which* members but *how many*. Ensemble accuracy saturates quickly in the number of trees [39], the gain from tuning that number is small [40], and closed-form guidance exists for the required size [41, 42]. Our fixed-budget design holds this axis constant so that composition is the only thing varying.

These threads are not independent. Ordered aggregation [24, 25], the analysis of how large an ensemble needs to be [41], efficient boosting variants [21] and the composition-search formulation we take as our reference point [1] come largely from one research programme. Our results should therefore be read as engaging with a coherent body of work rather than as isolated criticism: the same group that proposed searching the composition simplex also established that ensemble accuracy saturates quickly and that ordering members by a scalar criterion captures most of the available gain. Those findings and ours point the same way.

The reporting practice this paper examines.

Across much of the selection literature, the quality of a selector is assessed against a best-subset or *oracle* number obtained by maximising a performance estimate over the candidate set using the evaluation data. Such a bound is what motivates further work on selectors: the larger the gap between the deployed selector and the oracle, the more there appears to be left to win. We are not aware of a study in this literature that reports how much of such a bound survives when the oracle is required to generalise. Section 5.3 measures exactly that.

2.3 Combination weights and the forecast combination puzzle

Since Bates and Granger introduced estimated combination weights [2], the forecasting literature has repeatedly found that simple equal weighting is competitive with, and often better than, weights estimated from data, the *forecast combination puzzle*. Reviews documented the pattern [43]; Smith and Wallis attributed it to estimation error in the weights themselves [3]; Claeskens et al. gave a theoretical account showing that estimated weights are unbiased but sufficiently variable to lose to equal weights [4]; and the argument has been developed into an explicit bias–variance case for *shrinking* combination weights toward equality [5].

Our shrinkage selector, Equation (8), is that idea transplanted. What differs is the object. The forecasting literature places weights on individual models, whereas the ensemble literature we study allocates a fixed member budget across families, so the decision variable is a composition on a simplex rather than a weight vector, and the members within a family are exchangeable rather than distinct. Whether the puzzle transfers to that object is an empirical question, and Section 5.2 answers it: it does, and more so, since here even the shrinkage remedy buys nothing.

2.4 Selection bias and the optimism of a maximum

That an estimate chosen as the maximum over many candidates is optimistically biased is well established. In gene expression, selecting features on all the data before validating produces large apparent gains that vanish under honest resampling [44]; in model selection, error estimates from the same data used to select are biased and require nesting [45]; and Cawley and Talbot showed that over-fitting the model selection criterion itself inflates reported performance [46]. The same concern motivates careful protocols for comparing many classifiers over many datasets [47, 48], and critiques of those protocols [49]. Related in spirit, the value of a hyper-parameter search is bounded by how much signal the search space actually carries [50].

Two qualifications matter. First, nesting is not always necessary: when the selection is over few candidates and the sample is large, flat and nested estimates differ little [51]. Our setting is the opposite regime: 496 candidates, and test samples of a few hundred on many of the datasets in Table 1. This is precisely where the inflation is largest. Second, comparisons of learning algorithms are only as replicable as their resampling design allows [52], which is part of why we report ten seeds and treat seed-to-seed variation as an estimand in Section 5.4.

This thread and the previous one have not been brought to bear on ensemble composition. Our contribution is to apply the cross-fitting diagnostic of Section 2.4 to the oracle bounds of Section 2.2, on the object of Section 2.1, and to explain the outcome with the mechanism of Section 2.3.

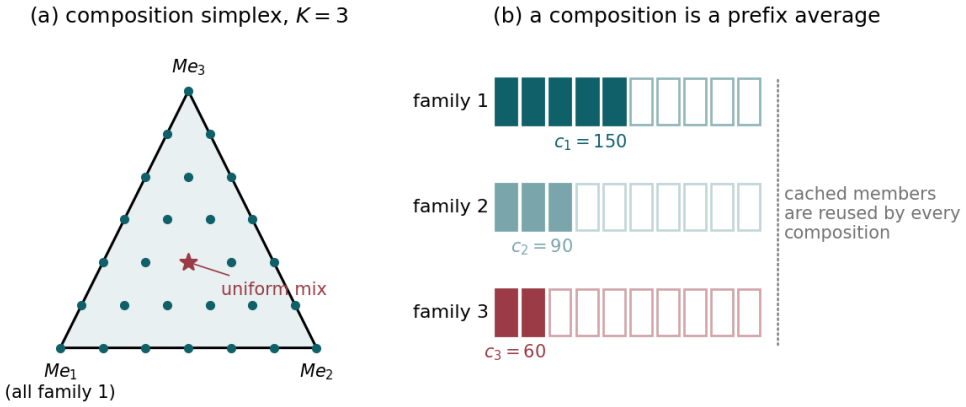


Fig. 1 (a) The composition simplex for $K = 3$ families under a fixed budget. Each dot is a candidate composition and the star marks the uniform mix. (b) A composition takes the first c_k members of each family, so every candidate is a prefix average of one cached set of member predictions.

3 Preliminaries

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ be K homogeneous ensemble families, each able to produce exchangeable members. Fix a total budget M. A *composition* is a vector

$$c = (c_1, \ldots, c_K) \in \mathbb{Z}^K_{\geq 0}, \qquad \sum_{k=1}^{K} c_k = M. \tag{1}$$

Write $p_{k,i}(x) \in \Delta^{C-1}$ for the predicted class distribution of the i-th member of family k. The ensemble under composition c predicts

$$P_c(x) \;=\; \sum_{k=1}^{K} \frac{c_k}{M} \cdot \frac{1}{c_k} \sum_{i=1}^{c_k} p_{k,i}(x) \;=\; \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{c_k} p_{k,i}(x), \tag{2}$$

so a composition is a prefix-average of cached per-member predictions. This is what makes the study cheap: each family is fitted *once* at full size M, per-member predictions are cached, and every point of the simplex is a re-weighting of that cache. With $M = 300$, $K = 3$ and a grid step of $s = 10$ members there are

$$|\mathcal{G}| \;=\; \binom{M/s + K - 1}{K - 1} \;=\; \binom{32}{2} \;=\; 496 \tag{3}$$

candidate compositions, all evaluated at the cost of three ensemble fits.

Let $\hat{R}_S(c)$ denote the accuracy of composition c measured on a sample S. A *selector* is a map from a scoring sample to a composition. We propose four selectors in addition

to the two baselines, and study all six:

$$c^{\text{unif}} = (M/K, \ldots, M/K), \tag{4}$$

$$c^{\text{single}} = M \cdot e_{k^\star}, \qquad k^\star = \arg\max_k \hat{R}_V(Me_k), \tag{5}$$

$$c^{\text{argmax}} = \arg\max_{c \in \mathcal{G}} \hat{R}_V(c), \tag{6}$$

$$c^{\text{bag}} = \frac{1}{B} \sum_{b=1}^{B} \arg\max_{c \in \mathcal{G}} \hat{R}_{V^{(b)}}(c), \tag{7}$$

with $V^{(b)}$ a bootstrap resample of the validation set and $B = 200$. Bagged argmax attacks *estimation* noise while keeping selection hard. Shrinkage instead softens the selection:

$$c^{\text{shrink}}(\tau) = \sum_{c \in \mathcal{G}} w_\tau(c)\, c, \qquad w_\tau(c) \propto \exp\Big(\hat{R}_V(c)/\tau\Big), \tag{8}$$

which recovers c^{argmax} as $\tau \to 0$ and c^{unif} as $\tau \to \infty$. The temperature τ must be chosen on data not used to compute $\hat{R}_V$. Section 4 describes the split. A sixth selector averages (7) and (8).

A worked example.

Take $K = 2$ families and $M = 300$ with a step of 10. The simplex is the segment $c = (m, 300 - m)$ for $m \in \{0, 10, \ldots, 300\}$, so $|\mathcal{G}| = 31$. Two adjacent candidates, $c = (100, 200)$ and $c' = (110, 190)$, share 100 members of family 1 and 190 of family 2, so 290 of their 300 members are common to both. By Equation (2) their predictions differ only through ten members out of three hundred:

$$P_{c'}(x) - P_c(x) = \frac{1}{M} \left[\sum_{i=101}^{110} p_{1,i}(x) - \sum_{i=191}^{200} p_{2,i}(x) \right], \tag{9}$$

a difference of order $\delta/M = 1/30$ in a quantity that is itself an average of bounded terms. Their accuracies are evaluated on the *same* test sample as well, so $\hat{R}_T(c)$ and $\hat{R}_T(c')$ share both most of their members and all of their evaluation noise. The candidate set is therefore not 31 independent estimates but a smooth, heavily dependent field, and at $K = 3$ the same holds across 496 candidates.

Oracles.

The quantity usually reported as an upper bound is

$$c^{\text{oracle}} = \arg\max_{c \in \mathcal{G}} \hat{R}_T(c), \tag{10}$$

selected on the *test* sample T and evaluated on it. Because $|\mathcal{G}| = 496$ and the $\hat{R}_T(c)$ are strongly dependent, $\mathbb{E}[\max_c \hat{R}_T(c)] > \max_c \mathbb{E}[\hat{R}_T(c)]$, so (10) is optimistically biased.

We therefore propose a *cross-fitted oracle* as a diagnostic: split T into halves T_1, T_2, select on one and predict on the other,

$$\hat{R}^{\mathrm{cf}} = \frac{1}{|T|}\Big[|T_2|\hat{R}_{T_2}\big(\arg\max_c \hat{R}_{T_1}(c)\big) + |T_1|\hat{R}_{T_1}\big(\arg\max_c \hat{R}_{T_2}(c)\big)\Big]. \tag{11}$$

Equation (11) is what an oracle selector could actually deliver on unseen data. The gap between (10) and (11) is the selection bias we quantify.

3.1 Why the conventional oracle is biased

Write $\hat{R}_T(c) = R(c) + \varepsilon_c$, where $R(c)$ is the true accuracy of composition c and ε_c is evaluation noise with $\mathbb{E}[\varepsilon_c] = 0$ and $\mathrm{sd}(\varepsilon_c) = \sigma \approx \sqrt{R(1-R)/|T|}$. Then

$$\mathbb{E}\Big[\max_{c\in\mathcal{G}} \hat{R}_T(c)\Big] \;\geq\; \max_{c\in\mathcal{G}} R(c) \;+\; \mathbb{E}\Big[\max_{c\in\mathcal{G}} \varepsilon_c\Big] \;-\; \big(\max_c R(c) - R(c^\dagger)\big), \tag{12}$$

with $c^\dagger$ the argmax of $\hat{R}_T$. The middle term is strictly positive for any non-degenerate noise: the maximum of a collection of mean-zero variables has positive expectation. Its size is governed by σ and by the *effective* number of independent candidates, which the dependence in Equation (9) makes far smaller than $|\mathcal{G}|$, but never one.

Two things follow, and Section 5 tests both. The inflation should grow with the number of candidates, since enlarging $\mathcal{G}$ can only raise a maximum. Accordingly, Section 5.5 measures a 25% increase going from 136 to 816 compositions. In addition, the inflation should be largest exactly where $R(\cdot)$ is flattest, because then the ordering of candidates is determined by ε rather than by R. Section 5.4 shows that the surface is a plateau. The cross-fitted oracle of Equation (11) removes the middle term of (12) by construction: the composition is chosen using ε from one half of the test set and scored against independent noise on the other.

4 Experimental design

Families.

$K = 3$: random forest [8], extremely randomised trees [9], and bagged k-nearest neighbours ($k = 5$, 50% sample fraction, 70% feature fraction) [6, 7]. The pool is chosen so that two families share an inductive bias and one does not, which is the common case in practice. Section 5.5 adds a fourth family with a linear bias to test whether that choice drives the result. All three produce exchangeable members, which Equation (2) requires: a composition takes the first c_k members of family k, so that prefix must be distributionally equivalent to a random subset. Gradient boosting is excluded for exactly this reason.

Budget.

$M = 300$ members in total, held constant across every selector. Because the budget is fixed, no selector can win by spending more inference, and the comparison is at equal deployment cost. The only difference between selectors is how the same 300 members

Table 1 The 30 OpenML classification datasets. IR is the imbalance ratio (largest class / smallest class).

Dataset	n	d	classes	IR
balance-scale	625	4	3	5.9
banknote	1372	4	2	1.2
blood-transfusion	748	4	2	3.2
car	1728	6	4	18.6
churn	5000	20	2	6.1
climate-simulation	540	20	2	10.7
credit-g	1000	20	2	2.3
diabetes	768	8	2	1.9
hill-valley	1212	100	2	1.0
ilpd	583	10	2	2.5
ionosphere	351	34	2	1.8
kr-vs-kp	3196	36	2	1.1
madelon	2600	500	2	1.0
mfeat-karhunen	2000	64	10	1.0
optdigits	5620	64	10	1.0
ozone-level	2534	72	2	14.8
pendigits	10992	16	10	1.1
phoneme	5404	5	2	2.4
qsar-biodeg	1055	41	2	2.0
segment	2310	19	7	1.0
sonar	208	60	2	1.1
spambase	4601	57	2	1.5
steel-plates	1941	33	2	1.9
tic-tac-toe	958	9	2	1.9
vehicle	846	18	4	1.1
vowel	990	12	11	1.0
wall-robot	5456	24	4	6.7
waveform-5000	5000	40	3	1.0
wdbc	569	30	2	1.7
wilt	4839	5	2	17.5

are allocated. The grid step is 10 members, giving 496 compositions by Equation (3). Section 5.5 uses a step of 20 so that $K = 3$ and $K = 4$ are compared at equal granularity.

Data.

30 public OpenML classification datasets (Table 1) spanning $n = 208$ to 10,992, $d = 4$ to 500, 2 to 11 classes, and imbalance ratios up to 18.6. The range is deliberate: the composition surface behaves differently on small noisy problems than on large clean ones (Section 5.4). Features are median-imputed and standardised, targets label-encoded.

Protocol.

Per dataset and seed: a stratified 70/30 train/test split, then a stratified 70/30 fit/-validation split of the training portion, giving roughly 49% fit, 21% validation, 30% test. Families are fitted on the fit portion only, so no member has seen the data used

to select among compositions, and no selector has seen the data used to report. Ten seeds resample all three splits jointly, so seed-to-seed variation includes split variation.

Choosing the shrinkage temperature.

τ cannot be selected on the same data that produced the scores being shrunk. Doing so biases the choice toward $\tau \to 0$, since the shrunk composition that best fits the validation scores is the argmax of those scores, and the selector silently collapses into the one it was meant to improve on. We therefore split the validation portion in half, compute simplex scores on one half, choose $\tau \in \{0.002, 0.005, 0.01, 0.02, 0.05\}$ on the other, and apply the chosen τ to scores from the full validation portion.

Metrics.

Accuracy is the headline. Macro-F1 and MCC are reported alongside because several datasets in Table 1 are imbalanced (the imbalance ratio reaches 18.6). Accuracy alone can conceal a selector that improves on the majority class while degrading the minority. All three orderings agree (Table 3).

Statistics.

We follow the standard protocol for comparing several methods over many datasets [47]: a Friedman test across selectors with datasets as blocks, and where it rejects, a Nemenyi post-hoc test with the critical difference at $\alpha = 0.05$. The same protocol is applied in the target venue to comparisons of this size [53]. With $k = 6$ selectors and $n = 30$ datasets, CD $= 1.377$ rank units. Pairwise comparisons against ArgMax use the Wilcoxon signed-rank test. We note the standing critique of mean-rank post-hoc procedures [49]. However, our central claims concern a group of selectors that is not separated at all rather than a contested ordering within it.

What the design can and cannot detect.

With 30 datasets and CD $= 1.377$, the comparison is powered to detect a selector that is consistently better by roughly one rank unit, not one better by a hundredth of a rank. Accepting a null of "no difference" on that basis would be weak evidence. This is why Section 5.3 does not rest on the failure of the selectors: the cross-fitted oracle bounds what *any* selector could achieve.

Why repeated hold-out rather than k-fold.

Selection and evaluation must not share data, so a nested design is required. We use a repeated stratified hold-out instead of nested k-fold for two reasons. First, ten seeds resample the fit, validation and test partitions jointly, so the seed-to-seed spread we report in Section 5.4 contains split variability as well as estimator randomness. Under k-fold the test partitions are fixed by construction and that component would be suppressed, yet it is precisely the quantity against which the composition surface is measured. Second, the comparison is between selectors evaluated on identical members, and a hold-out lets every selector be scored from one cached set of per-member predictions. Nested k-fold would require refitting the families once per outer fold without changing any conclusion.

Complexity.

Let n be the number of evaluation instances, C the number of classes, M the member budget, K the number of families and $|\mathcal{G}|$ the number of candidate compositions. Evaluating the simplex naively, by forming each composition from scratch, costs

$$O(|\mathcal{G}| \cdot M \cdot n \cdot C) \tag{13}$$

member-prediction accumulations. The prefix-average construction of Equation (2) removes the dependence on $|\mathcal{G}|$: one cumulative pass over the cached members gives every prefix mean in

$$O(K \cdot M \cdot n \cdot C)\,, \tag{14}$$

after which each candidate is a K-term weighted sum, so scoring the whole simplex costs $O(|\mathcal{G}| \cdot K \cdot n \cdot C)$ with no further access to individual members. With $M = 300$, $K = 3$ and $|\mathcal{G}| = 496$ this is a reduction of roughly two orders of magnitude in the dominant term, and it is why every selector and both oracles can be evaluated from a single set of three ensemble fits. Training cost is unchanged and dominates: it is $O(K \cdot M \cdot T_{\text{fit}})$, independent of the selector.

Measured cost.

Three ensemble fits per (dataset, seed) cell, or 900 fits in total, running in 24 minutes on a consumer CPU with no GPU. Every point of the simplex, every selector and both oracles are computed from the same cached per-member predictions, so the comparison is exactly matched and the marginal cost of adding a selector is the $O(|\mathcal{G}| \cdot K \cdot n \cdot C)$ scoring pass alone.

5 Results

5.1 Any selection beats a uniform mix

Mean accuracy across 30 datasets (Table 3): Uniform 0.8781, BestSingle 0.8851, ArgMax 0.8849, BaggedArgMax 0.8843, Shrink 0.8850, ShrinkBagged 0.8849. The uniform mix is worse than the argmax selector by 0.68 accuracy points on average (median 0.18), losing on 25 of 30 datasets, Wilcoxon $p = 2.4 \times 10^{-5}$ (Table 5). In the Friedman analysis (Table 2, Figure 2) Uniform has average rank 5.23 while every selector lies between 2.60 and 3.50, a separation beyond the critical difference of 1.377. Composition is therefore not irrelevant: a fixed equal mix is reliably a poor choice.

5.2 No search beats selecting a single family

The ordering of the remaining five selectors is BestSingle (2.60), ShrinkBagged (3.10), Shrink (3.20), ArgMax (3.37), BaggedArgMax (3.50). The spread is 0.90 rank units, below CD = 1.377: all five are statistically indistinguishable, and the selector that ignores mixing altogether ranks first.

Pairwise against ArgMax (Table 5), nothing is significant: BestSingle +0.0001 ($p = 0.52$), Shrink +0.0000 ($p = 0.93$), ShrinkBagged −0.0001 ($p = 0.54$), BaggedArgMax −0.0006 ($p = 0.42$). Both variance-reduction strategies fail. Moreover, bagged argmax,

Table 2 Friedman average ranks over the six selectors.

Selector	Average rank
BestSingle	2.60
ShrinkBagged	3.10
Shrink	3.22
ArgMax	3.38
BaggedArgMax	3.48
Uniform	5.22

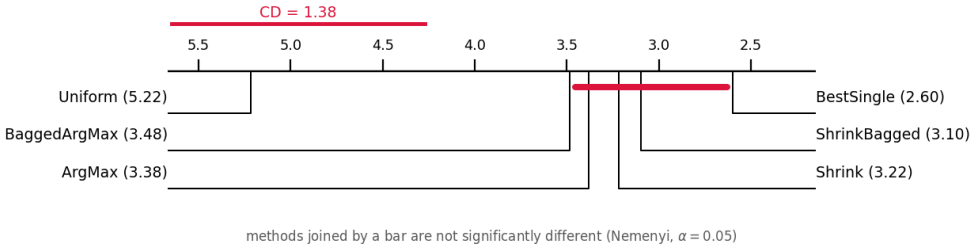


Fig. 2 Critical-difference diagram. Methods joined by a bar are not significantly different (Nemenyi, $\alpha = 0.05$). Every selector except Uniform falls in a single group.

Table 3 All three metrics, averaged over datasets and seeds.

Selector	Accuracy	Macro-F1	MCC
Uniform	0.8781	0.8127	0.6800
BestSingle	0.8851	0.8218	0.6941
ArgMax	0.8849	0.8215	0.6957
BaggedArgMax	0.8843	0.8207	0.6944
Shrink	0.8850	0.8227	0.6966
ShrinkBagged	0.8849	0.8217	0.6955
$\text{Oracle}_{\text{cf}}$	0.8838	0.8205	0.6927
$\text{Oracle}_{\text{test}}$	0.8957	0.8370	0.7235

which targets estimation noise directly, is if anything slightly worse than plain argmax. Shrinkage, with τ chosen out-of-sample, is exactly neutral. The smallest temperature $\tau = 0.002$, effectively argmax, was selected in 196 of 300 cells, so the data itself prefers hard selection. The same ordering holds under macro-F1 and MCC (Table 3).

Table 4 Mean accuracy over 10 seeds. The best selector on each dataset is shown in bold. The two oracles are diagnostics rather than selectors, and are reported separately in Table 6.

Dataset	Uniform	BestSingle	ArgMax	BaggedArgMax	Shrink	ShrinkBagged
balance-scale	0.8463	0.8702	0.8782	0.8787	0.8766	**0.8809**
banknote	0.9934	0.9951	0.9973	0.9961	0.9973	**0.9976**
blood-transfusion	0.7569	0.7609	**0.7716**	0.7711	0.7649	0.7671
car	0.9522	**0.9578**	0.9539	0.9539	0.9570	0.9543
churn	0.9246	**0.9508**	0.9499	0.9490	0.9483	0.9496
climate-simulation	0.9130	0.9130	**0.9136**	0.9136	0.9136	0.9136
credit-g	0.7470	0.7467	0.7467	**0.7477**	0.7467	0.7477
diabetes	0.7554	0.7498	**0.7567**	0.7558	0.7550	0.7537
hill-valley	0.5393	**0.5519**	0.5445	0.5456	0.5445	0.5451
ilpd	0.7040	0.7040	**0.7097**	0.7000	0.7086	0.7017
ionosphere	0.9302	**0.9377**	0.9321	0.9321	0.9358	0.9349
kr-vs-kp	0.9843	**0.9859**	0.9855	0.9845	0.9851	0.9849
madelon	0.6662	**0.6737**	0.6727	0.6708	0.6723	0.6722
mfeat-karhunen	0.9585	0.9618	0.9618	0.9627	**0.9628**	0.9617
optdigits	0.9826	0.9832	0.9826	**0.9833**	0.9833	0.9832
ozone-level	0.9417	0.9415	0.9411	0.9415	**0.9420**	0.9417
pendigits	0.9909	**0.9924**	0.9920	0.9917	0.9912	0.9914
phoneme	0.8954	**0.8980**	0.8966	0.8972	0.8972	0.8972
qsar-biodeg	**0.8707**	0.8678	0.8691	0.8688	0.8707	0.8703
segment	0.9703	**0.9733**	0.9709	0.9706	0.9716	0.9717
sonar	0.7825	0.8016	0.8095	0.8016	**0.8127**	0.8095
spambase	0.9499	**0.9519**	0.9505	0.9515	0.9510	0.9514
steel-plates	0.9969	**0.9986**	0.9962	0.9962	0.9967	0.9969
tic-tac-toe	0.8726	**0.9073**	0.9052	0.9052	0.9062	0.9056
vehicle	0.7441	**0.7528**	0.7406	0.7433	0.7461	0.7465
vowel	0.9118	**0.9411**	0.9367	0.9350	0.9323	0.9340
wall-robot	0.9803	**0.9929**	0.9918	0.9914	0.9920	0.9919
waveform-5000	0.8551	0.8568	**0.8589**	0.8567	0.8563	0.8559
wdbc	0.9520	0.9526	0.9526	**0.9544**	0.9520	0.9544
wilt	0.9755	**0.9803**	0.9799	0.9798	0.9789	0.9793
Mean	0.8781	0.8851	0.8849	0.8843	0.8850	0.8849

Table 5 Difference from ArgMax, over 30 datasets. * denotes $p < 0.05$ (Wilcoxon signed-rank).

Selector	mean Δ	median Δ	wins	p
Uniform	-0.0068	-0.0018	5/30	2.37e-05*
BestSingle	+0.0001	+0.0004	18/30	0.517
BaggedArgMax	-0.0006	-0.0001	11/30	0.424
Shrink	+0.0000	+0.0000	14/30	0.929
ShrinkBagged	-0.0001	+0.0003	17/30	0.539
$\text{Oracle}_{\text{cf}}$	-0.0012	-0.0003	10/30	0.0677
$\text{Oracle}_{\text{test}}$	+0.0108	+0.0081	29/30	2.56e-06*

5.3 The headroom that motivates better selectors is selection bias

The oracle selected on test attains 0.8957, i.e. +0.0108 accuracy over ArgMax, winning on 29 of 30 datasets ($p = 2.6 \times 10^{-6}$). Read naively this says a perfect selector would gain more than a full accuracy point.

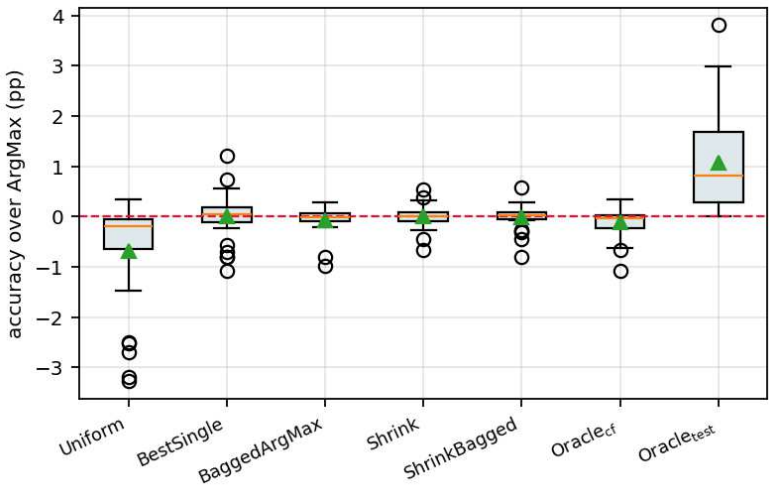


Fig. 3 Distribution across datasets of each selector's accuracy relative to ArgMax.

Table 6 Oracle headroom before and after cross-fitting.

Quantity	mean	median	p
Oracle selected on test, minus ArgMax	+0.0108	+0.0081	2.56e-06
Cross-fitted oracle, minus ArgMax	-0.0012	-0.0003	0.0677
Share of apparent headroom that is selection bias	111%		

The cross-fitted oracle of (11) attains 0.8838, −0.0012 relative to a single argmax ($p = 0.068$), i.e. no better, with a slight negative trend (Table 6, Figure 4):

$$\frac{\text{honest headroom}}{\text{apparent headroom}} = \frac{-0.0012}{+0.0108} \approx -0.11, \tag{15}$$

so 111% of the apparent headroom is selection-on-test bias. There is nothing for a better selector to capture, which is exactly why the variance-reduction strategies of Section 5.2 fail. The two results are one result.

Any oracle, best-subset or upper-bound number computed by maximising over a large correlated candidate set on the evaluation data should be cross-fitted before being reported. Otherwise it measures the size of the candidate set as much as the potential of the method.

5.4 Why: the composition surface is a plateau

Figure 5 plots test accuracy against family shares for four datasets, with a band marking seed-to-seed standard deviation at fixed composition. The surface has structure, since mixtures dominated by the weakest family are worse. However, above that the good region is a broad plateau, and the noise band covers much of it.

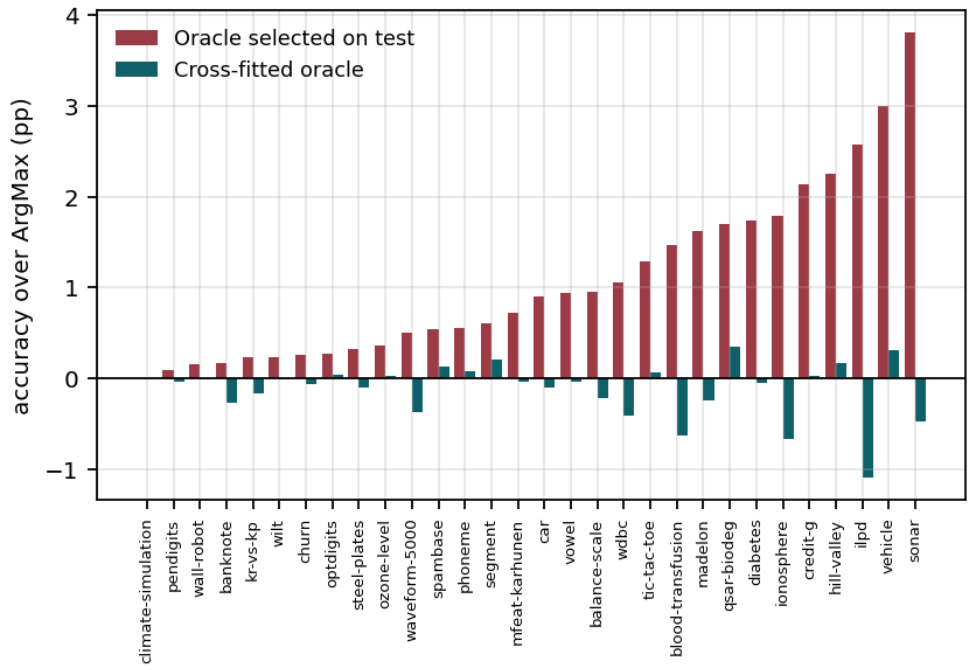


Fig. 4 Oracle advantage over ArgMax per dataset, before and after cross-fitting. The apparent headroom is present on nearly every dataset and disappears under cross-fitting.

Table 7 quantifies this across all 30 datasets. Median composition SD is 0.0080 against a median seed SD of 0.0057, so the typical ratio is 1.25 and the median spread between best and worst composition is 3.7 accuracy points. What the data show is more specific: on 12 of 30 datasets the ratio is below one, so the whole simplex varies less than run-to-run noise. In addition, the ratio scales with sample size (Spearman $\rho = +0.395$, $p = 0.031$; Figure 6). Splitting at the median n, the smaller half has median ratio 0.85 with 9 of 15 below one, the larger half 1.67 with only 3 of 15 below one.

That second result is the reference method's third admitted difficulty, measured. [1] reports that out-of-bag identification of the optimum is less accurate on smaller datasets. Figure 6 shows why, and that it is a property of the problem rather than of their estimator. Below roughly $n = 1500$ the composition signal is routinely submerged in resampling noise, so no selector can locate the optimum reliably, because between resamples there is no stable optimum to locate.

5.5 Robustness to the number and diversity of families

The most likely objection is that $K = 3$ with two tree families and one weak family understates what composition can do. Bagged logistic regression, which has a linear decision boundary unlike either trees or neighbours, was added as a fourth family,

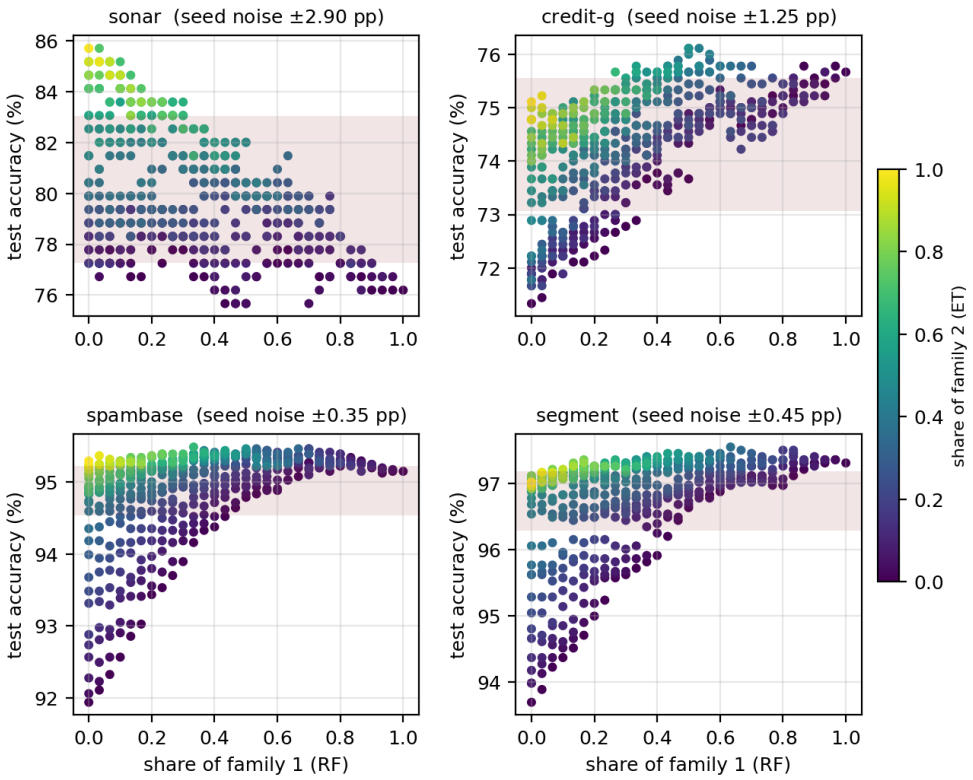


Fig. 5 Accuracy across the composition simplex for four datasets. The shaded band marks seed-to-seed noise at a fixed composition.

and $K = 3$ and $K = 4$ were run under identical conditions (same 30 datasets, same 5 seeds, same budget, grid step 20 for both, giving 136 and 816 compositions).

The objection has partial merit. Adding a diverse family un-flattens the surface: raising the composition-surface standard deviation from 0.0134 to 0.0185 and the best-worst range from 0.063 to 0.101. Both headline conclusions nevertheless survive. The penalty for a uniform mix doubles (-0.0070 to -0.0147), which is expected since a uniform mix over four families commits more budget to weak ones. The cross-fitted oracle still does not beat a single argmax (-0.0019, $p = 0.32$). Selecting a single family remains statistically indistinguishable from searching ($p = 0.69$), holding the best average rank (2.60 vs 2.85) even though ArgMax edges ahead on the mean.

The selection-bias result strengthens. Holding seeds, datasets and protocol fixed, enlarging the candidate set from 136 to 816 compositions raises the apparent oracle headroom from $+0.0103$ to $+0.0129$, a 25% increase, while the honest headroom moves in the opposite direction, from -0.0015 to -0.0019. This is the signature of selection bias rather than of real potential.

Table 7 Composition-surface variation against seed-to-seed noise, all 30 datasets. A ratio below one means the entire simplex varies less than a single composition does between runs.

Dataset	n	seed SD	comp. SD	best−worst	ratio
sonar	208	0.0290	0.0245	0.1005	0.85
ionosphere	351	0.0163	0.0337	0.1415	2.07
climate-simulation	540	0.0006	0.0008	0.0021	1.42
wdbc	569	0.0043	0.0025	0.0136	0.58
ilpd	583	0.0109	0.0070	0.0362	0.64
balance-scale	625	0.0110	0.0181	0.0674	1.65
blood-transfusion	748	0.0227	0.0112	0.0459	0.49
diabetes	768	0.0238	0.0052	0.0274	0.22
vehicle	846	0.0112	0.0147	0.0735	1.31
tic-tac-toe	958	0.0104	0.0343	0.1586	3.30
vowel	990	0.0169	0.0280	0.1953	1.66
credit-g	1000	0.0125	0.0107	0.0478	0.85
qsar-biodeg	1055	0.0096	0.0051	0.0315	0.52
hill-valley	1212	0.0359	0.0082	0.0403	0.23
banknote	1372	0.0056	0.0039	0.0235	0.71
car	1728	0.0163	0.0145	0.1092	0.89
steel-plates	1941	0.0035	0.0028	0.0177	0.79
mfeat-karhunen	2000	0.0038	0.0061	0.0300	1.59
segment	2310	0.0045	0.0077	0.0385	1.72
ozone-level	2534	0.0018	0.0010	0.0053	0.54
madelon	2600	0.0114	0.0238	0.1085	2.09
kr-vs-kp	3196	0.0035	0.0089	0.0574	2.53
spambase	4601	0.0035	0.0066	0.0355	1.92
wilt	4839	0.0030	0.0084	0.0349	2.85
churn	5000	0.0034	0.0194	0.0762	5.70
waveform-5000	5000	0.0032	0.0038	0.0282	1.19
phoneme	5404	0.0058	0.0062	0.0362	1.07
wall-robot	5456	0.0037	0.0247	0.1374	6.66
optdigits	5620	0.0019	0.0031	0.0134	1.67
pendigits	10992	0.0010	0.0012	0.0054	1.20

6 Discussion

6.1 What a practitioner should do

Under a fixed member budget, allocate it to one family chosen on validation data, and do not spend engineering effort searching proportions. A uniform mix is the one choice that is reliably wrong, and it becomes more wrong as the pool grows more diverse. However, this is a narrower recommendation than "heterogeneous ensembles do not help": pooling families is not the same decision as choosing proportions among them, and our design holds the first fixed while varying the second.

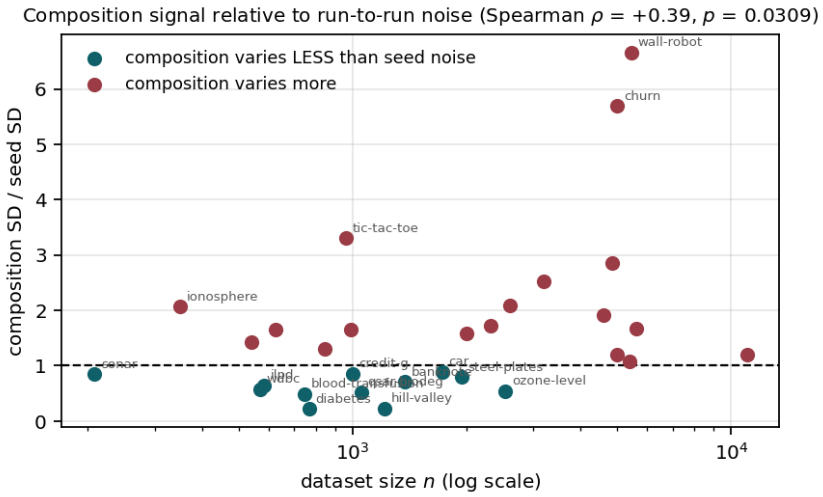


Fig. 6 Composition signal relative to run-to-run noise, all 30 datasets. Points below the dashed line are datasets where the entire simplex varies less than a single composition does between runs.

6.2 What the literature should change

An oracle or best-subset number obtained by maximising a performance estimate over a candidate set, using the labels of the set it is evaluated on, is not an upper bound on what a selector can achieve. It is an upper bound *plus* an inflation term that grows with the number and correlation of candidates. Cross-fitting the oracle costs one extra pass over cached predictions and converts a number that motivates further work into a number that bounds it. The same reporting pattern appears wherever a large correlated candidate set is scored on evaluation data: ensemble pruning over subsets [31], member selection in clustering ensembles [32], sample-selection criteria in self-training [54], and model-size selection more broadly [41, 42].

A recurring move in this literature is to remove a hyper-parameter, leaving one that must still be decided with expertise [55]. Our findings suggest a prior question: whether the surviving parameter carries signal above the noise floor at the sample sizes involved. Where it does not, the honest resolution is a documented default rather than a better selection rule.

6.3 Relation to the reference method's admitted difficulties

The three difficulties [1] reports now have a single explanation. The optimum is strongly problem-dependent because the surface is a plateau whose peak location is largely noise, so the argmax moves between resamples without the underlying optimum moving. Out-of-bag identification is less accurate on smaller datasets because that is exactly where composition variation falls below seed noise. And the intractability in the number of families is real but, on this evidence, not worth solving: enlarging the search space raises the apparent prize while lowering the real one.

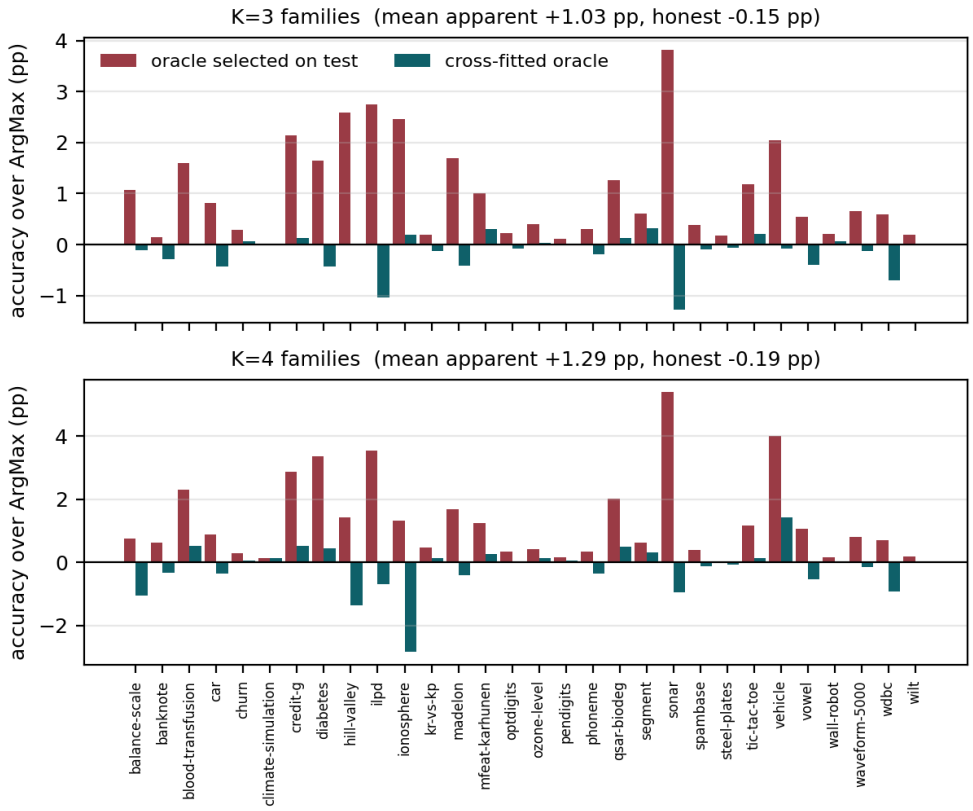


Fig. 7 Oracle advantage over ArgMax per dataset at $K = 3$ and $K = 4$. Enlarging the candidate set raises the apparent headroom, whereas the cross-fitted headroom stays at or below zero.

7 Threats to validity

We do not refute the reference method.

Our selector scores on a held-out validation split rather than out-of-bag, our three families are not identical to any prior study's, and our 30 datasets are our own selection. The claim is about composition search as a class under a controlled protocol.

Family choice bounds the effect, and partially does.

Section 5.5 tests this by adding a linear family. Composition signal does increase with K, so a larger or more complementary pool could increase it further. What does not change is that the extra signal is not recoverable by any selector. We tested $K \leq 4$ only.

The landscape analysis uses 3 seeds, not 10.

Table 7 and Figure 6 estimate a seed-to-seed standard deviation from three resamples. The ratio for any single dataset should be read loosely. The aggregate claims rest on 30 such estimates.

Fixed budget.

Everything is at $M = 300$. Behaviour at much smaller budgets, where individual members carry more weight, is untested.

Cross-fitting halves the selection sample,

so Equation (11) is a mildly conservative estimate of achievable oracle performance, far too small an effect to explain a swing from $+0.0108$ to -0.0012.

Grid granularity.

A step of 10 members out of 300 could miss narrow optima. Figure 5 shows no evidence of structure at that scale.

8 Conclusion

Under a fixed member budget, composition selection in heterogeneous ensembles is worth doing at the coarsest granularity and no finer. Choosing a mix rather than accepting a uniform one is worth about 0.7 accuracy points and is reliably significant. Choosing a *composition* rather than simply the best single family is worth nothing measurable, and neither bagged selection nor shrinkage recovers anything, because there is nothing to recover. The headroom that would justify a better selector, when computed the way this literature computes it, is essentially all selection-on-test bias.

Three boundaries deserve restating. We are not saying heterogeneous ensembles fail to help, since our design pools families in every condition and varies only the proportions. We are not saying selection is futile in general: dynamic selection [38] is a different mechanism our static design cannot bound. And we are not saying the composition surface is featureless. On larger datasets it carries measurable structure, only structure that no selector we tested converts into accuracy.

Composition signal grows with family diversity and with sample size. A pool of genuinely complementary families evaluated on large datasets is the regime where composition search might repay its cost, and our evidence does not extend there. That is the experiment we would run next, and the cross-fitted oracle is the instrument to run it with, because it answers, before any selector is built, whether there is anything for a selector to find.

Declarations

- **Funding.** The author did not receive support from any organization for the submitted work.
- **Competing interests.** The author has no competing interests to declare that are relevant to the content of this article.

- **Ethics approval.** Not applicable; the study uses only public, non-personal benchmark datasets.
- **Data availability.** All 30 datasets are public and retrieved from OpenML by numeric identifier; the identifiers are listed in Table 1 and hard-coded in the released code, so the corpus is reconstructible exactly. No new data were generated.
- **Code availability.** All code, cached per-run results and the scripts that generate every table and figure are released. The complete study reproduces in 24 minutes on a consumer CPU without a GPU.
- **Author contributions.** Muhammetalp Erdem is the sole author and carried out the study design, implementation, analysis and writing in their entirety.

References

[1] Sabzevari, M., Martínez-Muñoz, G., Suárez, A.: Building heterogeneous ensembles by pooling homogeneous ensembles. International Journal of Machine Learning and Cybernetics (2021) https://doi.org/10.1007/s13042-021-01442-1

[2] Bates, J.M., Granger, C.W.J.: The combination of forecasts. Journal of the Operational Research Society (1969) https://doi.org/10.1057/jors.1969.103

[3] Smith, J., Wallis, K.F.: A simple explanation of the forecast combination puzzle*. Oxford Bulletin of Economics and Statistics (2009) https://doi.org/10.1111/j.1468-0084.2008.00541.x

[4] Claeskens, G., Magnus, J.R., Vasnev, A.L., Wang, W.: The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting (2016) https://doi.org/10.1016/j.ijforecast.2015.12.005

[5] Blanc, S.M., Setzer, T.: Bias–variance trade-off and shrinkage of weights in forecast combination. Management Science (2020) https://doi.org/10.1287/mnsc.2019.3476

[6] Breiman, L.: Bagging predictors. Machine Learning (1996) https://doi.org/10.1007/bf00058655

[7] Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence (1998) https://doi.org/10.1109/34.709601

[8] Breiman, L.: Random forests. Machine Learning (2001) https://doi.org/10.1023/a:1010933404324

[9] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning (2006) https://doi.org/10.1007/s10994-006-6226-1

[10] Wolpert, D.H.: Stacked generalization. Neural Networks (1992) https://doi.org/10.1016/s0893-6080(05)80023-1

[11] Large, J., Lines, J., Bagnall, A.: A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. Data Mining and Knowledge Discovery (2019) https://doi.org/10.1007/s10618-019-00638-y

[12] Dietterich, T.G.: Ensemble methods in machine learning. In: Lecture Notes in Computer Science (2000). https://doi.org/10.1007/3-540-45014-9_1

[13] Zhou, Z.: Ensemble Methods: Foundations and Algorithms, (2012)

[14] Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review (2009) https://doi.org/10.1007/s10462-009-9124-7

[15] Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery (2018) https://doi.org/10.1002/widm.1249

[16] Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning (2003) https://doi.org/10.1023/a:1022859003006

[17] BROWN, G.: Diversity creation methods: a survey and categorisation. Information Fusion (2004) https://doi.org/10.1016/s1566-2535(04)00037-5

[18] Wood, D., Mu, T., Webb, A., Reeve, H.W.J., Luján, M., Brown, G.: A unified theory of diversity in ensemble learning. arXiv (Cornell University) (2023) https://doi.org/10.48550/arxiv.2301.03962

[19] Bian, Y., Chen, H.: When does diversity help generalization in classification ensembles? IEEE Transactions on Cybernetics (2022) https://doi.org/10.1109/tcyb.2021.3053165

[20] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics (2001) https://doi.org/10.1214/aos/1013203451

[21] Emami, S., Martínez-Muñoz, G.: Condensed-gradient boosting. International Journal of Machine Learning and Cybernetics (2024) https://doi.org/10.1007/s13042-024-02279-0

[22] Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. In: (venue to Confirm) (1997)

[23] Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artificial Intelligence (2002) https://doi.org/10.1016/s0004-3702(02)00190-x

[24] Martínez-Muñoz, G., Suárez, A.: Pruning in ordered bagging ensembles. In: Proceedings of the 23rd International Conference on Machine Learning - ICML '06 (2006). https://doi.org/10.1145/1143844.1143921

[25] Martinez-Munoz, G., Hernandez-Lobato, D., Suarez, A.: An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009) https://doi.org/10.1109/tpami.2008.78

[26] Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: Twenty-first International Conference on Machine Learning - ICML '04 (2004). https://doi.org/10.1145/1015330.1015432

[27] Zhang, Y., Burer, S., Street, W.N.: Ensemble pruning via semi-definite programming. In: (venue to Confirm) (2006). https://doi.org/10.5555/1248547.1248595

[28] Partalas, I., Tsoumakas, G., Vlahavas, I.: An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Machine Learning (2010) https://doi.org/10.1007/s10994-010-5172-0

[29] Qian, C., Yu, Y., Zhou, Z.H.: Pareto ensemble pruning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2015). https://doi.org/10.1609/aaai.v29i1.9579

[30] Tsoumakas, G., Partalas, I., Vlahavas, I.: An ensemble pruning primer. Studies in Computational Intelligence (2009) https://doi.org/10.1007/978-3-642-03999-7_1

[31] Ragab Hassen, H., Alabdeen, Y.Z., Gaber, M.M., Sharma, M.: D2ts: a dual diversity tree selection approach to pruning of random forests. International Journal of Machine Learning and Cybernetics (2022) https://doi.org/10.1007/s13042-022-01644-1

[32] Alqurashi, T., Wang, W.: Clustering ensemble method. International Journal of Machine Learning and Cybernetics (2018) https://doi.org/10.1007/s13042-017-0756-7

[33] Lu, Z., Wu, X., Zhu, X., Bongard, J.: Ensemble pruning via individual contribution ordering. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2010). https://doi.org/10.1145/1835804.1835914

[34] Guo, H., Liu, H., Li, R., Wu, C., Guo, Y., Xu, M.: Margin & diversity based ordering ensemble pruning. Neurocomputing (2018) https://doi.org/10.1016/j.neucom.2017.06.052

[35] Partalas, I., Tsoumakas, G., Vlahavas, I.: Pruning an ensemble of classifiers via reinforcement learning. Neurocomputing (2009) https://doi.org/10.1016/j.neucom.2008.06.007

[36] Ko, A.H., Sabourin, R., Britto, A.S. Jr.: From dynamic classifier selection

to dynamic ensemble selection. Pattern Recognition (2008) https://doi.org/10.1016/j.patcog.2007.10.015

[37] Britto, A.S., Sabourin, R., Oliveira, L.E.: Dynamic selection of classifiers—a comprehensive review. Pattern Recognition (2014) https://doi.org/10.1016/j.patcog.2014.05.003

[38] Cruz, R.M., Sabourin, R., Cavalcanti, G.D.: Dynamic classifier selection: Recent advances and perspectives. Information Fusion (2018) https://doi.org/10.1016/j.inffus.2017.09.010

[39] Oshiro, T.M., Perez, P.S., Baranauskas, J.A.: How many trees in a random forest? In: Lecture Notes in Computer Science (2012). https://doi.org/10.1007/978-3-642-31537-4_13

[40] Probst, P., Boulesteix, A.: To tune or not to tune the number of trees in random forest? arXiv (Cornell University) (2017) https://doi.org/10.48550/arxiv.1705.05654

[41] Hernández-Lobato, D., Martínez-Muñoz, G., Suárez, A.: How large should ensembles of classifiers be? Pattern Recognition (2013) https://doi.org/10.1016/j.patcog.2012.10.021

[42] Bonab, H., Can, F.: Less is more: A comprehensive framework for the number of components of ensemble classifiers. IEEE Transactions on Neural Networks and Learning Systems (2019) https://doi.org/10.1109/tnnls.2018.2886341

[43] Clemen, R.T.: Combining forecasts: A review and annotated bibliography. International Journal of Forecasting (1989) https://doi.org/10.1016/0169-2070(89)90012-5

[44] Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. In: Proceedings of the National Academy of Sciences (2002). https://doi.org/10.1073/pnas.102102699

[45] Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics (2006) https://doi.org/10.1186/1471-2105-7-91

[46] Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. In: (venue to Confirm) (2010). https://doi.org/10.5555/1756006.1859921

[47] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. In: (venue to Confirm) (2006)

[48] García, S., Herrera, F.: An extension on "statistical comparisons of classifiers over

multiple data sets" for all pairwise comparisons. In: (venue to Confirm) (2008)

[49] Benavoli, A., Corani, G., Mangili, F.: Should we really use post-hoc tests based on mean-ranks? arXiv (Cornell University) (2015)

[50] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. In: (venue to Confirm) (2012)

[51] Wainer, J., Cawley, G.: Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Systems with Applications (2021) https://doi.org/10.1016/j.eswa.2021.115222

[52] Bouckaert, R.R., Frank, E.: Evaluating the replicability of significance tests for comparing learning algorithms. In: Lecture Notes in Computer Science (2004). https://doi.org/10.1007/978-3-540-24775-3_3

[53] Moosaei, H., Hladík, M., Mousavi, A., Gao, Z., Fu, H.: Quadratic surface twin support vector machine for imbalanced data. International Journal of Machine Learning and Cybernetics (2026) https://doi.org/10.1007/s13042-026-03059-8

[54] Tanha, J., Someren, M., Afsarmanesh, H.: Semi-supervised self-training for decision tree classifiers. International Journal of Machine Learning and Cybernetics (2015) https://doi.org/10.1007/s13042-015-0328-7

[55] Yuan, X., Yu, H., Liang, J., Xu, B.: A novel density peaks clustering algorithm based on k nearest neighbors with adaptive merging strategy. International Journal of Machine Learning and Cybernetics (2021) https://doi.org/10.1007/s13042-021-01369-7