

# TSOpt: Fast and robust training-set design for genomic selection and sparse testing

Waseem Hussain

waseemhussain@plantnura.com

PlantNura Consulting & Solutions

---

## Research Article

**Keywords:** Genomic selection · Training population optimisation · CDmean · Sparse testing · Multi-environment trials · Rice breeding

**Posted Date:** October 22nd, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11450611/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# **TSOpt: Fast and robust training-set design for genomic selection and sparse testing**

Waseem Hussain

PlantNura Consulting & Solutions, Division of Education, Hyderabad 500032, Telangana, India

**Corresponding author:** Waseem Hussain

Email: waseemhussain@plantnura.com

**ORCID:** Waseem Hussain, <https://orcid.org/0000-0002-6861-0193>

## **Key message**

TSOpt designs genomic-selection training sets exactly in under a second, about 900 times faster than STPGA, robust to unknown heritability, and extends design to sparse multi-environment testing

## **Abstract**

Which candidates are phenotyped largely decides the accuracy of genomic prediction, yet few programmes design training sets formally. Existing tools evaluate the design criterion, the mean generalised coefficient of determination (CDmean), in a reduced marker space or repeatedly inside stochastic searches, optimise for one assumed heritability, and offer little built-in support for multi-environment trials or historical records. TSOpt removes these limits. It updates the exact CDmean by rank-one steps, so a design for several hundred lines takes about 0.1 s, averages the criterion over the unknown heritability, and uses one engine for line  $\times$  environment and line  $\times$  trait units, lines already phenotyped, hybrids, plot costs and major genes. We tested it on 328 elite lines from eight seasons of the International Rice Research Institute's irrigated breeding programme. Twenty-five TSOpt lines predicted untested lines as accurately as 76–167 random lines. TSOpt raised predictive ability over random sampling by 0.097 (95% confidence interval 0.045–0.150), was ahead of STPGA on average, and ran at least 900 times faster. With a quarter of line  $\times$  season combinations tested over four seasons,

its sparse-testing allocations predicted untested combinations better than balanced random allocation, even when scored with correlations unseen by the design; with half tested, it did not. Forward in time, all design methods gained little and the full history remained the best training set. On six public panels TSOpt matched STPGA, TSDFGS and published TrainSel designs, much faster. Lines with the highest mean relationship to the targets never beat random choice.

**Keywords** Genomic selection · Training population optimisation · CDmean · Sparse testing · Multi-environment trials · Rice breeding

## Introduction

Genomic selection has changed how crops are bred. A model trained on phenotyped relatives turns the marker genotypes of every candidate into predictions for lines that have never been grown (Meuwissen et al. 2001; Heffner et al. 2009; Crossa et al. 2017). Because genotyping now costs far less than a field plot, most programmes are limited by how many lines they can phenotype rather than by how many they can genotype. The choice of which candidates get a plot, the *training set*, therefore affects every prediction the programme makes, and choosing well costs nothing beyond the genotypes already in hand.

The theory of this choice is well developed. Rincent et al. (2012) showed that the generalised coefficient of determination (CD) of the mixed model (Laloë 1993), averaged over the lines to be predicted (CDmean), anticipates the accuracy a training set will reach. The criterion was later adapted to structured populations (Isidro et al. 2015; Rincent et al. 2017), to targeted designs aimed at a known set of candidates (Akdemir and Isidro-Sánchez 2019), to the choice of training-set size (Ou and Liao 2019) and to multi-environment trials (Liu and Liao 2026). Three R packages put these ideas into practice. STPGA searches for good sets with a genetic algorithm (Akdemir et al. 2015), and TSDFGS uses exchange algorithms with CD or its r-score (Ou and Liao 2019). TrainSel combines a genetic algorithm with simulated annealing, evaluates the full criterion, and accepts user-defined and multi-objective criteria (Akdemir et al. 2021). Independent comparisons found that these methods add a few points of predictive ability over random sampling, with large differences between datasets, and that targeted CDmean designs were among the most reliable (Isidro y Sánchez and Akdemir 2021; Fernández-González et al. 2023).

Even so, training-set design is still uncommon in day-to-day breeding, and the tools themselves are part of the reason. CDmean requires the inverse of a matrix the size of the training set, and every candidate set visited by a search needs a new one. STPGA and

61 TSDFGS keep this affordable by working on a few principal components of the relationship  
62 matrix, which approximates the criterion, and TrainSel evaluates the full criterion many times  
63 within a stochastic search. A design then takes from seconds to minutes. That is acceptable  
64 once, but not when a breeder wants to compare traits, target sets, budgets and seasons before  
65 planting, or to design for thousands of early-stage candidates. The criterion also depends on  
66 the heritability expected in the coming season, which is rarely known, yet the tools optimise  
67 for one assumed value by default. Real decisions are also more varied than choosing lines  
68 from a single panel. Plots are spread over sites and seasons, hundreds of lines have already  
69 been phenotyped, hybrids are predicted from their parents, and plots differ in cost. In TrainSel  
70 several of these can be written as user-defined criteria, but none of the packages offers them  
71 as built-in designs within one exact framework. Finally, most evaluations have measured the  
72 criterion itself or random cross-validation inside a diversity panel. Breeders predict forward in  
73 time, from the data they hold now to new lines in seasons still to come, and whether designed  
74 training sets help in that setting has rarely been tested on a real pipeline. The same is true of a  
75 common rule of thumb, phenotyping the candidates most related to those to be predicted.  
76 TSOpt was written to close these gaps. Its engine never inverts a matrix during the search. It  
77 updates the prediction error variance one unit at a time with Sherman–Morrison steps, which  
78 gives the exact CDmean, with the intercept and other fixed effects projected out, at the cost of  
79 a matrix–vector product per step. That speed makes two further features practical. The default  
80 criterion is averaged over a range of heritabilities, or over REML estimates from the  
81 programme’s own trials, so the design does not hinge on one guess. And the same engine  
82 works on any unit a breeder might phenotype: a line, a hybrid, a line in one environment, or a  
83 line scored for one trait, together with lines already phenotyped, plot costs, budgets and  
84 protected major-gene alleles. Around the design, TSOpt estimates the variance model from  
85 historical trials, replays a design on the programme’s own history before it is trusted, and

writes plans to field-book apps and breeding databases. In this paper we evaluate the core of this engine: single-trait designs of lines, line  $\times$  season designs for sparse testing, and designs that build on lines already phenotyped. The hybrid, multi-trait and major-gene designs are described but not evaluated here.

We test TSOpt on the multi-environment trials of the irrigated rice breeding programme of the International Rice Research Institute (IRRI) (Spindel et al. 2015a) and ask three practical questions. How much phenotyping does a designed training set save when lines are chosen within a season? Does design still help when the programme's own decisions are replayed forward in time, as a new cohort enters the trials? And can the design decide where to test lines, not only which lines, in sparse multi-environment testing? We then check the results on six public panels of five crops, compare TSOpt with the training sets that TrainSel's authors published for those panels, and ask whether richer genetic models change which lines are worth phenotyping.

## Materials and methods

### *Statistical model and design criterion*

Genomic prediction uses the mixed model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e}$ , with genetic values  $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{K})$ , residuals  $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$ , fixed effects  $\boldsymbol{\beta}$  (at least an intercept) and  $\lambda = \sigma_e^2 / \sigma_g^2 = (1 - h^2) / h^2$ . For a training set  $S$ , the prediction error variance of the genetic values is  $\mathbf{A}(S) = \mathbf{K} - \mathbf{K}_S \mathbf{V}_S^{-1} \mathbf{K}_S$ , with  $\mathbf{V}_S = \mathbf{K}_{SS} + \lambda \mathbf{I}$ . The generalised coefficient of determination of a contrast  $\mathbf{c}$  (Laloë 1993; Rincent et al. 2012) is

$$\text{CD}(\mathbf{c}) = 1 - \frac{\mathbf{c}^\top \mathbf{A}(S) \mathbf{c}}{\mathbf{c}^\top \mathbf{K} \mathbf{c}}.$$

CDmean is the average of CD over a set of contrasts, each comparing one unit with the mean of a reference set. In *targeted* designs the reference set is the targets, the units to be predicted, and there is one contrast per target ( $\mathbf{c}_i = \mathbf{e}_i - \bar{\mathbf{e}}_T$  for target  $i$ , where  $\bar{\mathbf{e}}_T$  has  $1/n_T$  at each

target). In *untargeted* designs the reference set is all candidates, with one contrast per candidate. Rincent et al. (2012) used the whole population as reference; our untargeted criterion uses the candidates.

*Fixed effects.* The intercept, environment means and any other fixed effects  $\mathbf{X}$  are absorbed with a flat prior. The engine runs on  $\mathbf{K} + c\mathbf{X}\mathbf{X}^T$  with  $c$  large, while the contrasts are formed from  $\mathbf{K}$ . As  $c \rightarrow \infty$  this gives the prediction error variance with  $\boldsymbol{\beta}$  estimated by generalised least squares (GLS), so the criterion belongs to the model that will be fitted rather than to one with  $\boldsymbol{\beta}$  known. TSOpt uses  $c = [\text{value of } c \text{ used in TSOpt 0.5.0}]$ , with  $\mathbf{K}$  scaled to a mean diagonal of 1. To check the approximation, we compared it with CDmean computed directly with the intercept estimated by GLS, on the IRRI relationship matrix, for 60 designs (25 or 100 training lines,  $h^2$  of 0.2, 0.5 and 0.8). The largest absolute difference was  $9 \times 10^{-6}$  for  $c = 10^2$ ,  $9 \times 10^{-10}$  for  $c = 10^6$ , and rose again to  $10^{-6}$  for  $c = 10^{10}$  through loss of numerical precision. The same check on designs chosen by TSOpt gave differences below  $3 \times 10^{-10}$  at  $c = 10^6$ , and TSOpt's own CDmean for these designs agreed with the direct GLS value to within  $10^{-15}$  at each of the three heritabilities (Online Resource 4).

*Robust CDmean.* The heritability of a trait in the coming season is unknown, and a design tuned to one value can over-fit it. By default, TSOpt averages CDmean over  $h^2 \in \{0.2, 0.5, 0.8\}$ , a grid that spans the heritabilities of line means typical of field trials. When historical trials are available, this grid is replaced by the 10%, 50% and 90% quantiles of the sampling distribution of the restricted maximum likelihood (REML) estimate (see *Breeding structures and variance components*).

### ***The rank-one engine***

Adding unit  $j$ , with noise variance  $\lambda_j$ , to the training set changes the posterior covariance by one Sherman–Morrison step (Sherman and Morrison 1950):

$$\mathbf{A}_{\text{new}} = \mathbf{A} - \frac{\mathbf{u}\mathbf{u}^\top}{\lambda_j + u_j}, \quad \mathbf{u} = \mathbf{A}_{.j},$$

and removing a unit reverses the step. TSOpt never forms  $\mathbf{A}$ . It keeps only the product  $\mathbf{B} = \mathbf{C}\mathbf{A}$  of the contrasts  $\mathbf{C}$  with  $\mathbf{A}$ , the diagonal of  $\mathbf{A}$ , and  $\mathbf{V}_S^{-1}$ , which is updated by bordering. From these, the gain in CDmean of every candidate is a column-wise weighted sum of squares of  $\mathbf{B}$ , computed for all candidates at once. A step costs  $O(mN)$  operations for  $m$  contrasts and  $N$  units, against  $O(N^3)$  for recomputing the criterion. The three inner operations (the column of  $\mathbf{A}$ , the rank-one update of  $\mathbf{B}$  and the vectorised gains) are compiled C code calling BLAS. The package's test suite checks each against its plain matrix formula, and the results of these checks are given in Online Resource 4.

*Breeding-value and general combining ability targets.* When the targets are breeding values or general combining abilities (GCA) rather than total genetic values, the contrasts are formed from the covariance of the predicted component (additive or GCA), while the engine updates on the total covariance. No further change is needed, because the posterior covariance between targets and training units depends only on their cross-covariance.

*Search.* A greedy forward selection fills the set, and at each step the exact gain of adding any candidate is read from the stored rank-one quantities. CDmean is not submodular in general, so the greedy set carries no optimality guarantee. Under the robust criterion, which is the default, the greedy set is the design. With a single assumed heritability, a rank-one tabu search can then exchange units, again evaluating the exact gain of every possible swap from the same stored quantities. Stratification quotas, costs and budgets, units that must be included or excluded, historical (already phenotyped) units and a maximum number of environments per line enter the search as constraints.

## 156 ***Breeding structures and variance components***

157 *Multi-environment and multi-trait designs.* Here the units are line  $\times$  environment (and line  $\times$   
158 trait) cells, with covariance  $\mathbf{\Omega}_E \otimes \mathbf{K}$  (or  $\mathbf{\Omega}_E \otimes \mathbf{G}_T \otimes \mathbf{K}$ ), a residual variance for each  
159 environment, and environment means as fixed effects. Above 8,000 cells the unit covariance  
160 is held in factorised form and never built. Results match those of the dense form to  $10^{-10}$  in  
161 the tests.

162 *Other structures.* Hybrid and testcross designs use the covariances of GCA and specific  
163 combining ability (SCA) effects between the parents of each cross. Major genes can be  
164 protected by a locus-weighted kernel or by requiring a minimum number of training carriers  
165 for each allele class found among the targets. Other kernels (dominance, epistasis, Gaussian,  
166 pedigree, single-step, haplotype and omics) can be combined, with shares that are either fixed  
167 or averaged over a grid. These designs are exact under their models but are not evaluated on  
168 real data in this paper.

169 *Variance components.* These are estimated by average-information REML for models of the  
170 form  $\mathbf{V} = \sum_k \theta_k \mathbf{M}_k$ , which covers traits, environments and kernel shares, with an active set  
171 for variances that reach zero. The implementation matched a brute-force optimiser of the  
172 same likelihood in all 20 test cases (Online Resource 4).

## 173 ***Software implementation and availability***

174 TSOpt is an R package with a compiled C engine, a command-line tool and a REST interface.  
175 Binary packages are distributed for Windows, macOS and Linux (R 4.5 or later), and at run  
176 time it depends only on data.table, ggplot2 and scales. On one core of an Apple M-series  
177 computer, a dense design for 1,000 lines took 1.3 s, for 2,000 lines 6.8 s and for 5,000 lines 52  
178 s. A factorised design for 1,000 lines in 15 environments (15,000 units) took 2.5 min, and a  
179 low-rank engine designed for 20,000 lines in 28 s.

Besides the design itself, the package estimates the variance model, replays designs on a programme's history, draws sample-size curves, plans stages and seasons, predicts crosses, and budgets genotyping, imputation and omics profiling. Each plan comes with its expected reliability compared with random sampling, the value of every plot (the loss if it were dropped) and the coverage of the targets' alleles and haplotypes. Plans can be laid out in blocks with checks, exported to Field Book or FieldHub (Murillo et al. 2021), and written back to a Breeding API (BrAPI) database such as BMS or Breedbase (Selby et al. 2019). Table 1 compares TSOpt with the published tools.

**Table 1** Training-set design tools compared

|                                                 | STPGA                                                                 | TSDFGS                               | TrainSel                                                                                                     | TSOpt                                                                                              |
|-------------------------------------------------|-----------------------------------------------------------------------|--------------------------------------|--------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| Criterion evaluation                            | Matrix inversion on a feature matrix (typically principal components) | Matrix inversion on a feature matrix | Full or feature-based, user-defined                                                                          | Exact, by rank-one updates of the prediction error variance                                        |
| Fixed effects                                   | Intercept in the reduced space                                        | Not projected                        | Through user criteria (CDmean with fixed effects)                                                            | Any fixed effects, by a flat prior                                                                 |
| Heritability                                    | One assumed value                                                     | One assumed value                    | One value by default (user criteria can average)                                                             | Averaged over a grid, or over REML estimates from the programme's trials                           |
| Search                                          | Genetic algorithm                                                     | Exchange                             | Genetic algorithm + simulated annealing; single or multiple objectives                                       | Greedy rank-one (default); rank-one exchange and tabu; genetic algorithm for non-standard criteria |
| Targeted designs                                | Yes                                                                   | Yes                                  | Yes                                                                                                          | Yes                                                                                                |
| Environments and traits (sparse testing)        | No                                                                    | No                                   | Through user-supplied criteria                                                                               | Built in, with genetic covariances between environments and traits                                 |
| Hybrids, testcrosses, historical data, costs    | No                                                                    | No                                   | Partly, through user code                                                                                    | Built in                                                                                           |
| Major genes and haplotypes                      | No                                                                    | No                                   | No                                                                                                           | Allele-carrier constraints; locus-weighted and haplotype kernels                                   |
| Training-set size                               | No                                                                    | Yes (logistic curve)                 | No                                                                                                           | Exact operating curve and economic optimum                                                         |
| Forward validation, season and genotyping plans | No                                                                    | No                                   | No                                                                                                           | Built in                                                                                           |
| Interface for breeders                          | R functions                                                           | R functions                          | R functions                                                                                                  | R functions, command line, BrAPI connection, field-book exports and plan reports                   |
| Licence                                         | GPL-3                                                                 | GPL ( $\geq 3$ )                     | Proprietary; universities and non-profit organisations; for-profit use needs the authors' written permission | Proprietary; free for all users, including companies (registration)                                |

Entries are taken from each package's documentation at the versions compared (STPGA 5.2.1, TSDFGS 2.0, and TrainSel as described by Akdemir et al. (2021)). "Built in" means available as an option without user-written criteria; only the single-trait, line  $\times$  season and history designs of TSOpt are evaluated in this paper. BrAPI, Breeding API; REML, restricted maximum likelihood

TSOpt is proprietary software of PlantNura Consulting & Solutions and is free for all users, including commercial breeding programmes. Users register by email for a personal licence key, which is checked offline, so no data leave the user's computer. Binaries, user guides, a tutorial with example datasets and the reference manual are available at <https://github.com/whussain2/TSOpt>, and the version used here is archived with a DOI (see Code availability). All analyses used TSOpt 0.5.0 with default settings unless stated otherwise.

### ***Plant material, phenotypes and genotypes***

The data are public under a CC0 licence (Spindel et al. 2015a, b). They come from replicated yield trials of elite lines in IRRI's irrigated rice breeding programme at Los Baños, Philippines, in the dry and wet seasons of 2009–2012, giving eight seasons that we treat as separate test environments. Lines entered the trials in cohorts and, once entered, were tested in the following seasons in two or three replicates. We analysed grain yield ( $\text{kg ha}^{-1}$ ), days to flowering and plant height; plant height was not recorded in the 2009 wet season. The lines were genotyped by sequencing at 73,147 single-nucleotide polymorphisms (SNPs) (Spindel et al. 2015a). After removing markers with minor-allele frequency below 0.05, 39,692 SNPs remained, and we used the 328 lines that had both genotypes and phenotypes. The additive relationship matrix **K** followed VanRaden (2008), scaled to a mean diagonal of 1. It was computed once from all 328 lines, since the genotypes of targets are known when a design is made.

*Phenotypic analysis.* Within each season, line best linear unbiased estimates (BLUEs) came from a linear model with line and replicate as fixed effects, and plot-level repeatability from the same model with line as random (lme4). Across several seasons, line BLUEs were

216 obtained from a two-way model of the season BLUEs with line and season fixed. Genetic  
217 correlations between seasons and genomic heritabilities of season means were estimated by  
218 average-information REML with an unstructured genetic covariance across seasons and a  
219 separate residual variance for each season.

220 *Population structure.* We described population structure by the principal components of **K** for  
221 the 328 lines, with missing genotypes set to the marker mean, and by the mean genomic  
222 relationship within and between the historical lines and the 2011 cohort (Online Resource 1).

### 223 *Design comparison within seasons*

224 In each of 20 random splits, 82 lines (25%) were set aside as targets and never phenotyped,  
225 and the remaining 246 were candidates. Training sets of 25, 50, 75 and 100 lines were chosen  
226 by the following methods:

- 227 • TSOpt with its defaults (robust CDmean, intercept projected out, greedy rank-one  
228 search);
- 229 • STPGA 5.2.1 (Akdemir et al. 2015), a genetic algorithm with the CDMEAN criterion  
230 on the first 30 principal components of **K** (population 100, 5 elite solutions, mutation  
231 probability 0.8, 300 iterations, ridge parameter  $\lambda = 1/30$ ). Because this ridge  
232 parameter corresponds to a high heritability, we repeated the comparison with  $\lambda = 1$   
233 (matched to  $h^2 = 0.5$ ) and 1,000 iterations;
- 234 • TSDFGS 2.0 (Ou and Liao 2019), its exchange algorithm with the CD and r-score  
235 criteria on the same principal components, with  $\min(20,000, 65 \cdot \sqrt{(\text{candidates} \times n)})$   
236 iterations. TSDFGS 2.0 does not declare its compiled library in its namespace, so we  
237 loaded the library explicitly, leaving the package code unchanged;
- 238 • GA+SA, our own search that combines a genetic algorithm with simulated annealing  
239 on the exact CDmean (population 60, up to 150 generations, stopping after 40  
240 generations without improvement). It follows the strategy published for TrainSel but is

not TrainSel, and its settings differ from TrainSel’s defaults. We did not run TrainSel itself (Akdemir et al. 2021), because its licence does not allow use by private or for-profit organisations without the authors’ written permission;

- TSOpt with a single assumed heritability (0.2, 0.5 or 0.8), to test the robust criterion on these data;
- three heuristics: partitioning around medoids (PAM) (Kaufman and Rousseeuw 1990); the lines most related to the targets, defined as the candidates with the highest mean relationship to all targets; and nearest relatives, in which targets take turns, in random order, to add their most closely related candidate not yet chosen;
- random sampling stratified over four PAM clusters, and plain random sampling (10 sets per split).

All sets were scored in the same way. We fitted genomic best linear unbiased prediction (GBLUP), with heritability estimated by REML, to the across-season BLUEs of the training lines. *Predictive ability* was the correlation between predictions and the across-season BLUEs of the targets. The normalised discounted cumulative gain of the top 10% (NDCG@10%) measured how well the best targets were ranked (Liu and Liao 2026). Computing time was wall-clock time on one core of an Apple M-series computer. For STPGA and TSDFGS the principal components were computed once per split outside the timed step; computing them took a median of 0.03 s, so this has no practical effect on the comparison.

*Plots saved.* We extended random sampling up to all 246 candidates. For a TSOpt set of a given size, the number of plots saved is the number of random lines needed to reach its mean predictive ability, found by linear interpolation along the random-sampling curve. Intervals were obtained by resampling the 20 splits with replacement (2,000 bootstrap samples).

## 264 ***Forward validation on the breeding pipeline***

265 By the end of 2010 the programme held phenotypes from 2009–2010 on 209 genotyped lines,  
266 which we call the history. In the 2011 dry season a cohort of 118 new genotyped lines entered  
267 the trials. We replayed two decisions using only information available at that time, and scored  
268 each on the new lines' performance in the *later* seasons (2011 wet, 2012 dry and 2012 wet),  
269 which no design or model had seen.

270 *(a) Which new lines to phenotype first.* From the cohort,  $k = 10, 20, 30, 45$  or 60 lines were  
271 phenotyped in their first season (2011 dry). The model was trained on the history (2009–2010  
272 BLUEs) together with these lines' first-season BLUEs, with a fixed effect for the source of  
273 each record, and used to predict the rest of the cohort. The  $k$  lines were chosen by the  
274 methods of the design comparison, targeted to the cohort. In TSOpt the history was declared  
275 as already phenotyped.

276 *(b) Which historical lines to train on.* Training sets of 40, 80 or 120 historical lines, chosen to  
277 predict the cohort, were compared with training on all 209.

278 Each of 30 replicates used a random 85% of the cohort as candidates and targets. Because all  
279 replicates come from one cohort, we treat their spread as a measure of the stability of each  
280 method, not as independent evidence.

## 281 ***Sparse multi-environment testing***

282 We used the 321 lines tested in all four seasons of 2011–2012 and assumed that plots were  
283 available for 25% or 50% of the line  $\times$  season cells. The allocations compared were:

- 284 • TSOpt, targeted at every line  $\times$  season cell, without a limit on seasons per line;
- 285 • TSOpt with every line tested in the same number of seasons ( $\lceil 4f \rceil$  for a fraction  $f$ ), so  
286 that every line is tested at least once;
- 287 • random cells;
- 288 • random allocation balanced per line, the usual sparse design (Jarquin et al. 2020);

- the overlap/non-overlap designs of Jarquin et al. (2020), in which 10% or 20% of the lines are tested in every season and the remaining plots are spread over the other lines. The design was given only what the programme knew at the end of 2010, namely the dry–dry, wet–wet and dry–wet genetic correlations of grain yield estimated from 2009–2010, applied to the four future seasons by season type. Every allocation was scored by predicting the untested cells with the same multi-environment GBLUP, with covariance  $\mathbf{R}_E \otimes \mathbf{K}$ , season means fixed, and genetic and residual variances estimated by REML (Burgueño et al. 2012; Jarquin et al. 2020). Because a design optimised under one covariance model may look best when scored under the same model, we scored every allocation twice: with the season-type correlations from 2009–2010 used by the design, and with the unstructured correlations estimated from all eight seasons, which the design never saw. Accuracy was the correlation between predicted and observed BLUEs of the untested cells within each season, averaged over seasons; we also report it pooled over seasons after centring each season. Each of 20 replicates used a random 90% of the lines, with five random allocations of each random design per replicate.

#### ***Public panels, published TrainSel designs and richer genetic models***

To see whether the results generalise, we repeated the design comparison on the six public panels with real phenotypes compiled by Fernández-González et al. (2023) (maize, two rice panels, sorghum, switchgrass and white spruce). This gave 33 scenarios: the whole panel as target, and five targeted splits (two for spruce) in which 25% of the lines were never phenotyped. Training sets were 15% of the candidates, and all traits were scored by GBLUP predictive ability and NDCG@10% as above. Times on these panels were measured with eight benchmark processes running in parallel, one core each, and are indicative only. We also recorded, on five panels, the gain of TSOpt designs made under single assumed heritabilities from 0.1 to 0.8, with and without exchange steps after the greedy search. The benchmark

314 scripts and all raw results are public (<https://github.com/whussain2/TSOpt>, folder  
315 benchmark/).

316 *TrainSel's published training sets.* Since we could not run TrainSel, we used the sets its  
317 authors optimised with it on the same panels and published with their comparison (Fernández-  
318 González et al. 2023). For each panel these comprise 40 random splits into candidates (85%  
319 of lines) and test lines (15%), with TrainSel's targeted and untargeted CDmean sets at five  
320 sizes (10, 20, 40, 60 and 80% of the candidates) and the matching random sets. On every split  
321 and at every size, TSOpt chose a set of the same size from the same candidates with its  
322 defaults, once targeted to the test set and once untargeted. All sets, TrainSel's included, were  
323 scored on the test lines with the same GBLUP, so absolute values may differ from those  
324 published, but the paired differences are on a common footing. As a check that our data and  
325 the published sets refer to the same lines, we first recomputed the authors' criterion for their  
326 sets with their code. We recovered the values TrainSel had reported to within 0.01% for five  
327 panels and 0.4% for spruce, whose relationship matrix differs slightly in scale, whereas a  
328 shuffled line order gave errors of about 16%.

329 *Richer genetic models.* Three further tests asked whether designing under a model richer than  
330 the additive default improves realised accuracy: additive  $\times$  additive epistasis (simulated  
331 epistatic values), haplotype kernels (four of the panels) and omics kernels (simulated  
332 transcript-like profiles and near-infrared spectra). These use the benchmark scripts distributed  
333 with the package.

334 ***Statistical analysis***

335 Methods were compared through paired differences within split (or replicate), size and trait,  
336 first averaged within each split or replicate. We regard TSOpt against random sampling as the  
337 primary comparison. All other comparisons are secondary and are reported without  
338 adjustment for multiplicity.

In the within-season comparison the 20 splits are drawn from the same 328 lines, so their training and test sets overlap and the split-level differences are correlated. We therefore used the corrected resampled  $t$ -test of Nadeau and Bengio (2003), in which the variance of the mean difference is  $s^2(1/J + n_2/n_1)$  for  $J$  splits with  $n_2$  test and  $n_1$  training candidates (here  $n_2/n_1 = 82/246$ ), and report 95% confidence intervals (CI) from it. For the published TrainSel designs we applied the same correction within each panel (40 splits,  $n_2/n_1 = 15/85$ ) and summarised over panels both by the mean of the six panel estimates, with its standard error, and by an inverse-variance weighted mean. On the public-panel benchmark the 33 scenarios come from six panels, so standard errors were computed over panel means. For the forward and sparse-testing analyses, which rest on a single cohort and a single set of seasons, we report the mean, the share of replicates in which a method beat its comparator, and the share of individual random designs that it beat, without formal tests. All analysis scripts are provided as Online Resource 3, and the full statistical tables as Online Resource 4.

## Results

### *The IRRI trials and population structure*

The 328 elite lines entered the trials in cohorts. Of these, 209 were tested in 2009–2010 and a new cohort of 118 entered in the 2011 dry season (Fig. 1a), which allowed the programme's own sequence of decisions to be replayed. Plot-level repeatability of season means ranged from 0.34 to 0.90 for grain yield, from 0.94 to 0.98 for days to flowering and from 0.75 to 0.89 for plant height. Genomic heritabilities of season means ranged from 0.19 to 0.76 for grain yield and from 0.84 to 0.96 for days to flowering (Fig. 1c). Genetic correlations for grain yield averaged 0.76 between dry seasons, 0.70 between wet seasons and 0.72 between a dry and a wet season (Fig. 1b). This genotype  $\times$  season structure is what sparse testing exploits.

Population structure was weak. The first two principal components of the genomic relationship matrix explained 6.1% and 4.7% of the variance, the first five together 21.6%, and 24 components were needed to reach half (Online Resource 1). Historical lines and the 2011 cohort overlapped on the first two components. The mean genomic relationship between a historical line and a cohort line ( $-0.007$ ) was close to that within the cohort ( $0.004$ ) and within the history ( $-0.001$ ), as expected for a single, well-connected elite population.

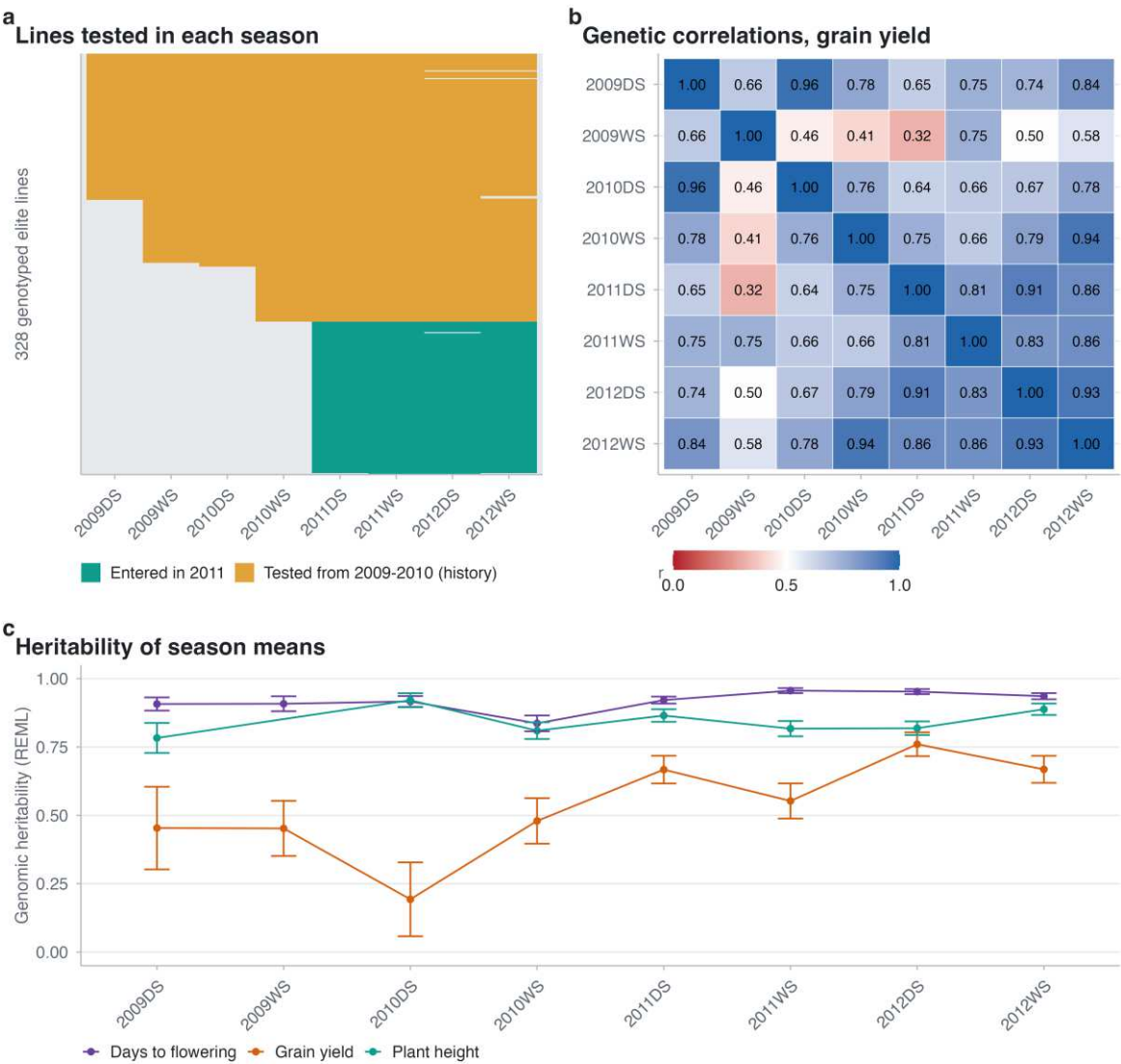


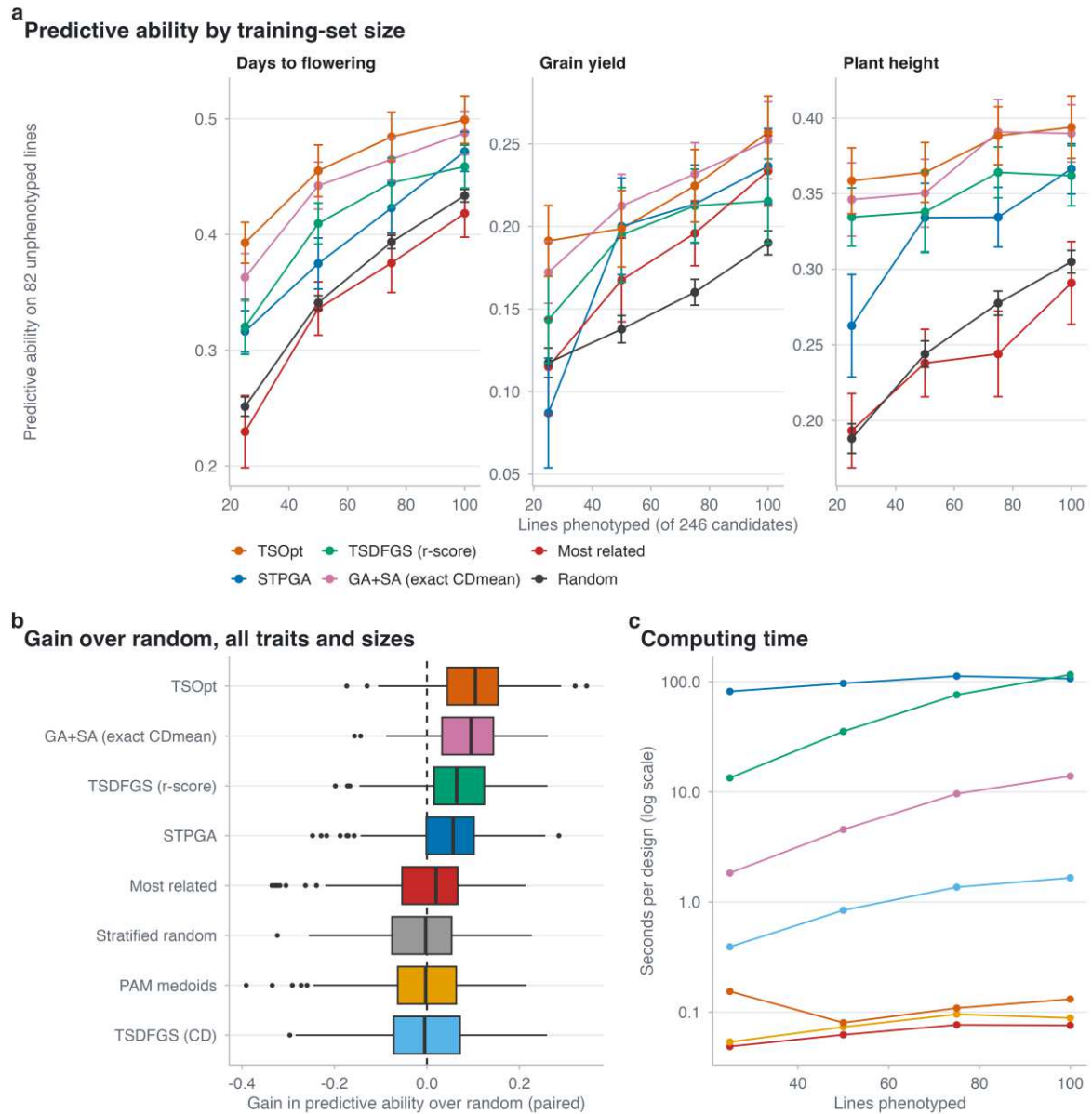
**Fig. 1** The IRRI multi-environment data. **a** The 328 genotyped elite lines (rows) tested in each of the eight seasons (columns). Lines entered the trials in cohorts: 209 were tested in 2009–2010 (the history), and 118 entered in 2011. **b** Genetic correlations of grain yield between seasons (average-information REML on season BLUEs; DS, dry season; WS, wet season). **c** Genomic heritability of season means for the three traits ( $\pm$  SE)

375 ***Designed training sets predicted unphenotyped lines better, and much faster***

376 Across three traits and four training-set sizes (Fig. 2a, b), TSOpt raised predictive ability on  
377 the 82 targets by 0.097 over random sampling (95% CI 0.045–0.150,  $P = 0.001$ ). It was ahead  
378 in 19 of the 20 splits and beat 79% of the 2,400 individual random training sets. GA+SA,  
379 which optimises the same exact criterion, gained 0.089 (CI 0.043–0.134) and TSDFGS with  
380 its r-score 0.063 (CI 0.009–0.117). STPGA gained 0.048, with an interval that included zero  
381 (CI –0.012 to 0.109). TSDFGS with its CD criterion was no better than random (–0.007, CI  
382 –0.052 to 0.039), and neither were the heuristics: PAM medoids changed predictive ability by  
383 –0.008, stratified sampling by –0.010 and the most related lines by 0.000.

384 In direct comparisons, TSOpt was ahead of STPGA by 0.049 on average and in 19 of 20  
385 splits, but once the overlap between splits was taken into account the interval included zero  
386 (CI –0.006 to 0.103,  $P = 0.08$ ). Its margins over TSDFGS were 0.104 with the CD criterion  
387 (CI 0.050–0.158) and 0.034 with the r-score (CI –0.005 to 0.074), and its margin over  
388 GA+SA was small (0.009, CI –0.012 to 0.030). The advantage over random sampling was  
389 largest for small training sets, 0.129 with 25 lines (CI 0.065–0.192) against 0.074 with 100  
390 lines. With 25 lines, TSOpt reached a predictive ability of 0.191 for grain yield against 0.117  
391 for random sets, 0.393 against 0.251 for days to flowering, and 0.359 against 0.188 for plant  
392 height. On these data the assumed heritability hardly mattered: designs made for a single  
393 heritability of 0.2, 0.5 or 0.8 differed from the robust default by –0.004 to 0.001. Part of  
394 STPGA's shortfall came from its settings. With the ridge parameter matched to  $h^2 = 0.5$  and  
395 1,000 iterations it gained 0.072 over random sampling (CI 0.021–0.122), and TSOpt's margin  
396 over it fell to 0.026 (CI –0.015 to 0.066), but each STPGA design then took a median of 32  
397 min. Unlike the lines with the highest mean relationship, nearest relatives, in which each  
398 target in turn adds its closest candidate, beat random sampling by 0.056 (CI 0.003–0.109);  
399 TSOpt was ahead of it by 0.042 (CI –0.013 to 0.097).

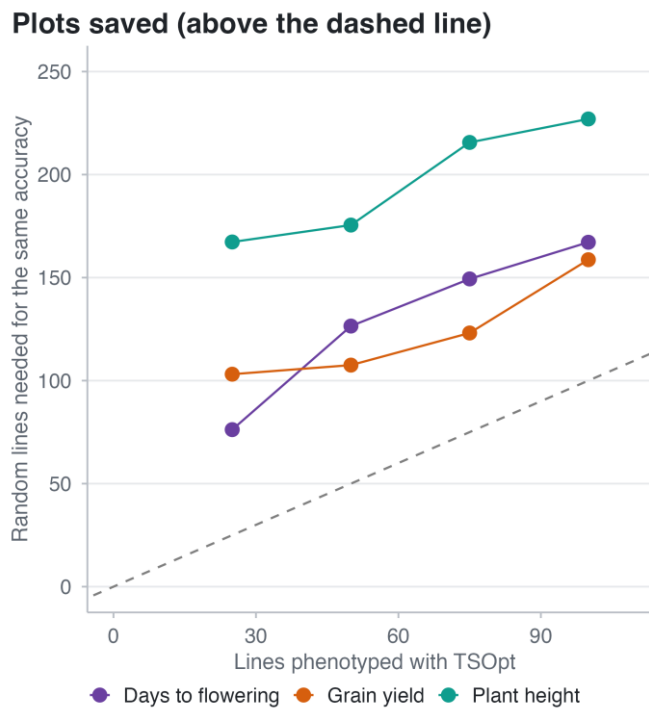
400 The median time per design (Fig. 2c) was 0.11 s for TSOpt, 7.0 s for GA+SA and 101 s for  
 401 STPGA, which is 884 times longer than TSOpt. TSDFGS took 1.1 s with CD and 55.5 s with  
 402 the r-score, and STPGA with 1,000 iterations 1,891 s.



403  
 404 **Fig. 2** Choosing which lines to phenotype. **a** Predictive ability on 82 unphenotyped lines of training  
 405 sets of 25–100 lines chosen by TSOpt, GA+SA on the exact CDmean, STPGA, TSDFGS (r-score), the  
 406 most related lines and random sampling (mean  $\pm$  SE over 20 splits, descriptive). **b** Paired gain over  
 407 random sampling for all methods, over all traits, sizes and splits. **c** Computing time per design for  
 408 grain yield (log scale; principal components for STPGA and TSDFGS not included)

409 ***The same accuracy from far fewer plots***

410 A TSOpt set of 25 lines predicted as well as random sets of about 103 lines for grain yield  
 411 (95% bootstrap interval 73–127), 76 for days to flowering (64–99) and 167 for plant height  
 412 (115–220) (Fig. 3). A TSOpt set of 50 lines matched random sets of 108 (70–150), 126 (97–  
 413 150) and 175 (122–231) lines. Within a season, TSOpt therefore needed 15–46% of the plots  
 414 that random sampling needed for the same predictive ability. Forward in time, where the  
 415 history dominates the training set, the saving was much smaller and inconsistent (Online  
 416 Resource 4).



417  
 418 **Fig. 3** Plots saved. The number of randomly chosen lines needed to reach the predictive ability of a  
 419 TSOpt training set of each size (from Fig. 2a). Points above the dashed line are plots saved. 95%  
 420 bootstrap intervals are given in Online Resource 4

421 ***Forward in time: replaying the programme's decisions***

422 The first decision was which new lines to phenotype first (Fig. 4a). Trained on the history  
 423 alone, predictions for the new cohort in the later seasons reached 0.183 for grain yield, 0.374  
 424 for days to flowering and 0.314 for plant height. When 10–60 new lines were added after

425 phenotyping in their first season, every design method beat random choice on average, but  
426 only by small margins. TSOpt gained 0.020 and beat random choice in 23 of 30 replicates,  
427 STPGA 0.032 (29 of 30), and TSDFGS 0.023 with CD (27 of 30) and 0.018 with the r-score  
428 (26 of 30). STPGA was slightly ahead of TSOpt overall (by 0.012; TSOpt ahead in 7 of 30  
429 replicates). TSOpt was more accurate for all three traits when only 10 new lines were  
430 phenotyped (grain yield 0.250 against 0.232), whereas STPGA was more accurate for grain  
431 yield and plant height with 45–60 lines. Phenotyping the new lines most related to the rest of  
432 the cohort was worse than random choice (–0.019; better in only 9 of 30 replicates).  
433 The second decision was which historical lines to train on (Fig. 4b). Training on all 209  
434 historical lines gave the highest accuracy for every trait (0.183 for grain yield), and no subset  
435 did as well. When only a subset could be kept, for instance because older records had to be  
436 dropped, TSOpt’s subsets beat random subsets of the same size by 0.032 (28 of 30 replicates).  
437 Subsets of the most related lines gained 0.027 (26 of 30), STPGA subsets 0.000 (15 of 30),  
438 and TSDFGS subsets chosen with CD lost 0.047 (0 of 30). With 40 lines, a fifth of the  
439 history, TSOpt’s subsets reached 0.149 for grain yield against 0.096 for random subsets.

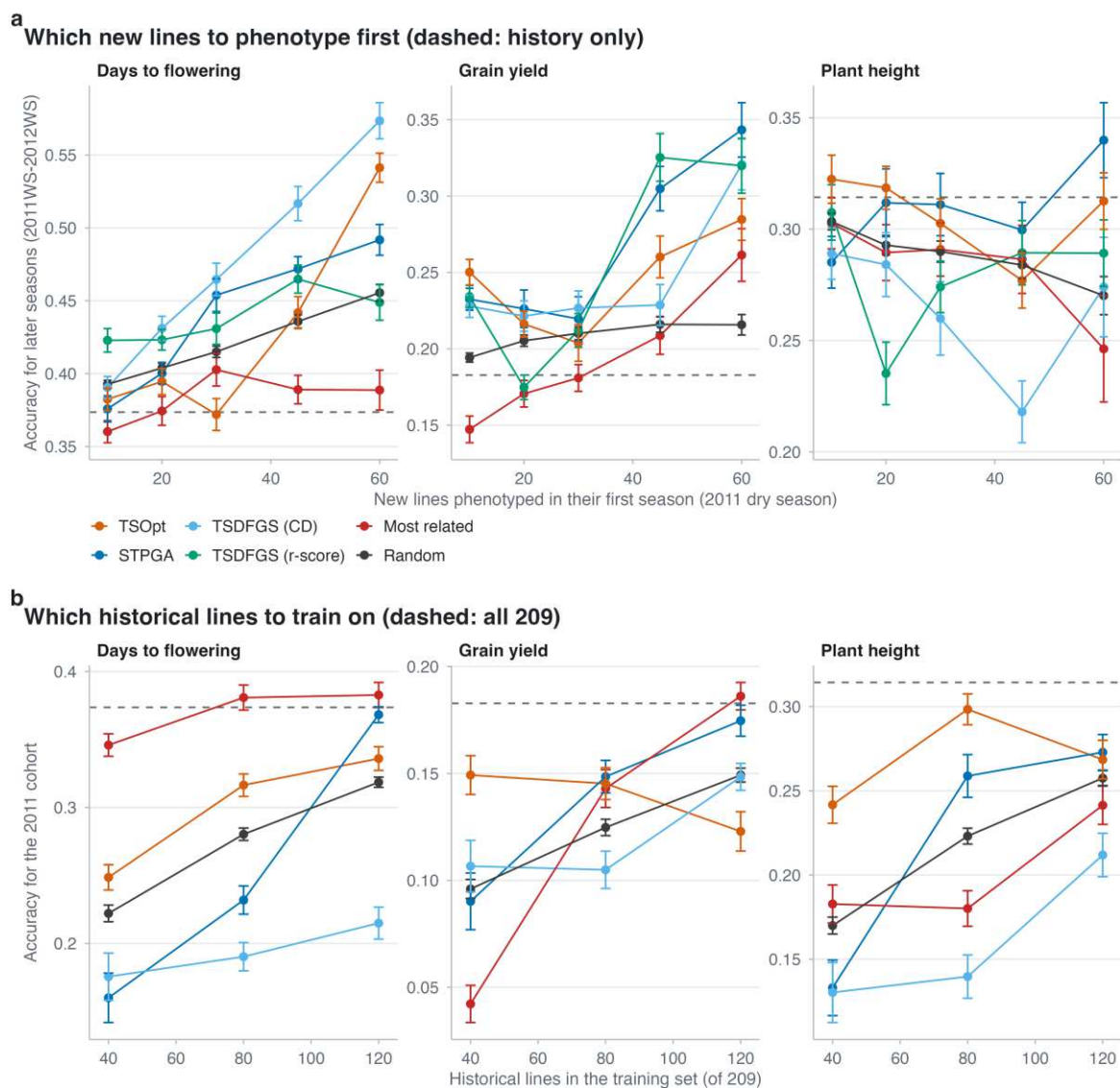


**Fig. 4** Forward in time on IRRI’s breeding pipeline. **a** Accuracy for the new cohort’s later seasons when  $k$  of its lines are phenotyped in their first season and trained together with the 2009–2010 history (mean  $\pm$  SE over 30 replicates of one cohort); the dashed line trains on the history alone. **b** Accuracy when the training set is a subset of the history chosen by each method; the dashed line uses all 209 historical lines

### *Sparse testing across seasons*

With a quarter of the line  $\times$  season cells tested, TSOpt’s allocations predicted the untested cells better than random cells by 0.078 and better than random allocation balanced per line by 0.024 (Fig. 5). TSOpt was ahead of balanced random allocation in all 20 replicates and beat 84% of the 300 individual balanced random allocations. Without a limit on seasons per line,

TSOpt left 8% of the lines untested. When every line had to be tested once, the gain over balanced random allocation was 0.021, again in all 20 replicates and against 80% of individual allocations. With half the cells tested, the gains over balanced random allocation were smaller: 0.010 without a limit (16 of 20 replicates) and 0.008 with every line tested twice (17 of 20). At a quarter of the cells, accuracy for untested cells was 0.424 with TSOpt and 0.392 with balanced random allocation for grain yield, 0.698 and 0.674 for days to flowering, and 0.583 and 0.569 for plant height. Scoring with the genetic correlations estimated from all eight seasons, which the design never saw, reduced these gains by about a fifth but did not remove them at a quarter of the cells. TSOpt was then ahead of balanced random allocation by 0.019 (19 of 20 replicates; 75% of individual allocations beaten), and by 0.015 when every line was tested once (17 of 20; 73%). With half the cells tested, the gain disappeared under this scoring (−0.001 without a limit; 0.006 with every line tested twice, 16 of 20 replicates). Accuracy pooled over seasons gave the same picture (0.018 and 0.013 at a quarter of the cells). The overlap/non-overlap designs, which test some lines in every season and leave others untested, were less accurate than balanced random allocation under both scorings (−0.075 and −0.228 with 10% and 20% overlap at a quarter of the cells; −0.020 and −0.039 at half).

### Sparse testing across four seasons (2011-2012)

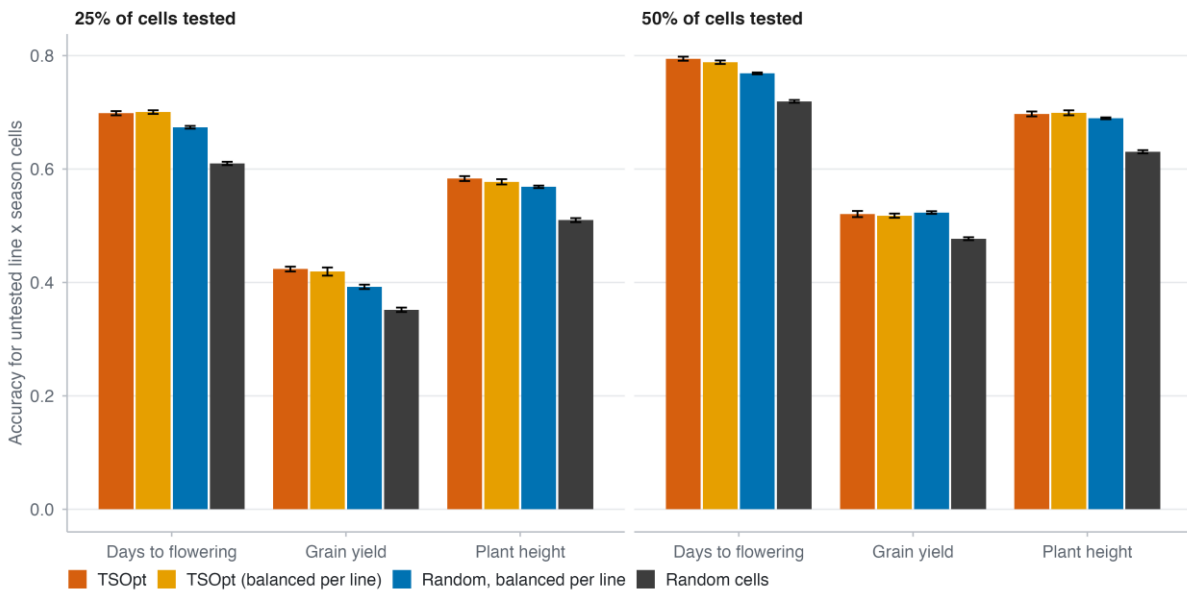


**Fig. 5** Sparse testing across four seasons. Accuracy for untested line  $\times$  season cells in 2011–2012, when 25% or 50% of the cells are tested, by allocation design (mean  $\pm$  SE over 20 replicates). All designs are scored with the same multi-environment GBLUP, here with the 2009–2010 season-type correlations; scoring with correlations the design never saw is reported in Online Resource 4

### *Other crops and richer genetic models*

Across the six public panels (33 scenarios), TSOpt's default raised predictive ability by 0.066 over random sampling and beat random sampling in 88% of scenarios and on all six panels (SE over panels 0.020,  $P = 0.03$ ) (Fig. 6a; Online Resource 2). STPGA gained 0.047, TSDFGS 0.043 (r-score) and GA+SA 0.061, while choosing the most related lines lowered predictive ability by 0.114. TSOpt was ahead of STPGA by 0.020 (SE over panels 0.008; ahead on four of six panels and in 58% of scenarios) and of TSDFGS by 0.024 (SE 0.005; all six panels, 64% of scenarios). On average the three tools were therefore close in accuracy, and TSOpt did not win consistently scenario by scenario. Speed was the clear difference, with 0.03 s per design for TSOpt against 40–120 s for the other two. Ranked by NDCG@10%, TSOpt gained 0.032 over random sampling, compared with 0.033 for STPGA and 0.022 for TSDFGS.

485 Averaging the criterion over heritabilities was ahead of a single assumed heritability of 0.5 by  
 486 0.009 (SE over panels 0.005,  $P = 0.19$ ; four of six panels), a small difference. Its value  
 487 showed when the assumed heritability was wrong. Over five panels, designs made for an  
 488 assumed heritability of 0.1, 0.2 or 0.35 gained 0.028, 0.045 and 0.065 over random sampling,  
 489 against 0.082 for 0.5, 0.073 for 0.8 and 0.084 for the robust average (Online Resource 4).  
 490 Adding exchange steps after the greedy search did not improve realised accuracy (greedy  
 491 minus greedy with exchange 0.006, SE over panels 0.005).  
 492 On identical splits and sizes, TSOpt's targeted designs predicted the test lines better than the  
 493 targeted CDmean sets published by TrainSel's authors by 0.005 on average. The difference  
 494 was positive on four of the six panels but individually clear only for switchgrass, and its  
 495 significance depends on how the panels are combined (mean of panel estimates 0.005, SE  
 496 0.003,  $P = 0.14$ ; inverse-variance weighted 0.005, SE 0.002,  $P = 0.004$ ). The gap came mostly  
 497 from the smallest training sets (0.013 at 10% of the candidates, against 0.001 at 60–80%). The  
 498 untargeted designs differed by 0.019 in TSOpt's favour, on all six panels (SE 0.003,  $P =$   
 499 0.003), although the two packages may define the reference set of an untargeted design  
 500 differently. Both tools improved on the published random sets, by 0.048 for TSOpt and 0.043  
 501 for TrainSel. TSOpt took a median of 0.09 s per design.  
 502 Designing under richer genetic models (additive  $\times$  additive epistasis, haplotype kernels and  
 503 omics kernels) changed predictive ability by at most 0.006 compared with the additive default  
 504 (Fig. 6b). The only difference larger than two standard errors, 0.005 (SE 0.002), was seen  
 505 when breeding values were the targets of prediction, which is consistent with forming the  
 506 contrasts from the additive covariance when that is what will be predicted.

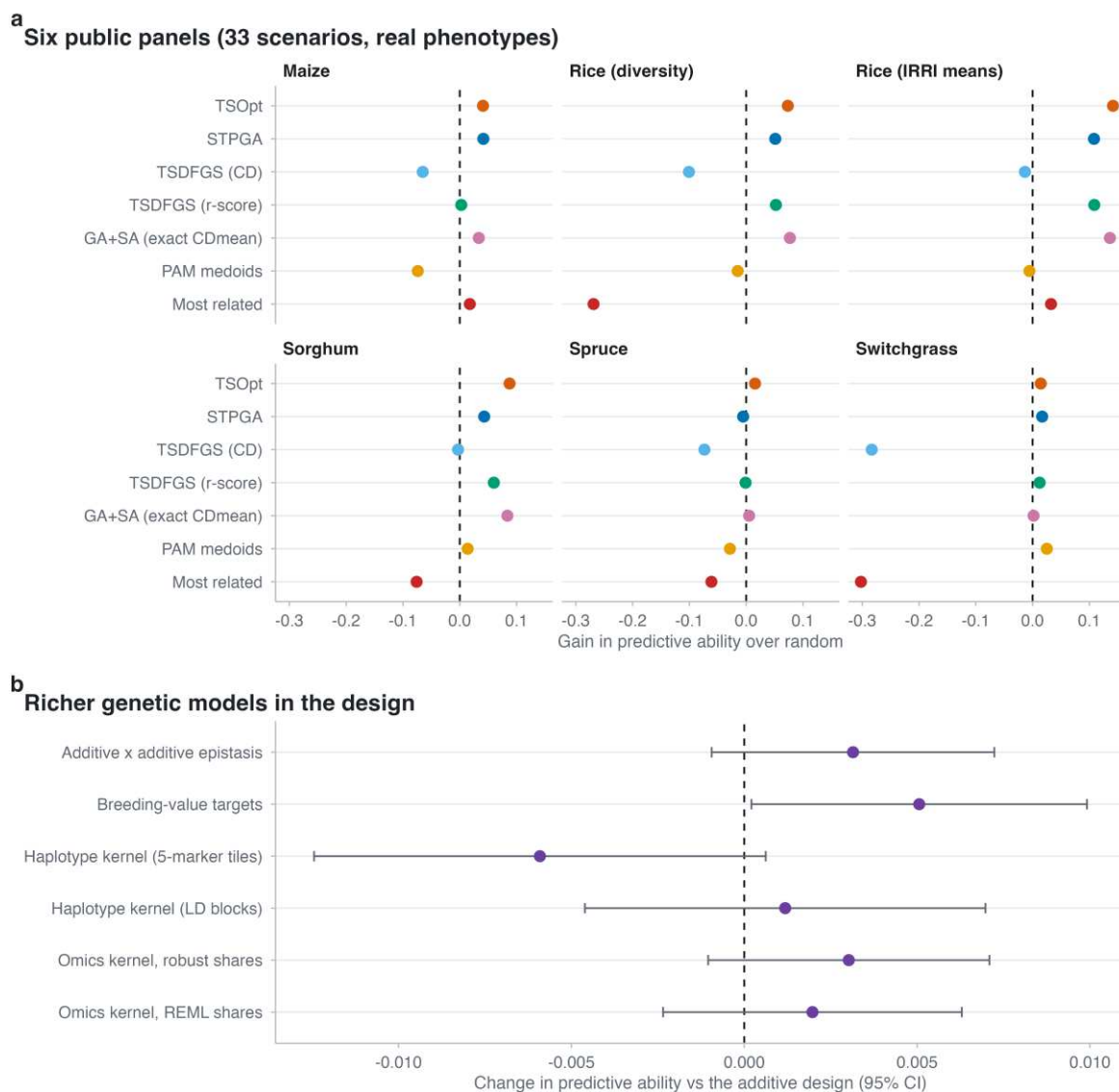


**Fig. 6** Generality. **a** Gain in predictive ability over random sampling on six public panels with real phenotypes (33 scenarios). **b** Change in predictive ability when the design uses a richer genetic model than the additive default (paired differences with 95% confidence intervals)

## Discussion

### *What design buys, and where the gains come from*

On IRRI's elite lines, designed training sets predicted unphenotyped lines better than random ones, and the gain is easiest to appreciate as plots saved. Twenty-five lines chosen by TSOpt predicted grain yield as well as about 100 random lines, and plant height as well as about 170, although the bootstrap intervals show that these figures are uncertain. A programme that

517 chooses which new lines to phenotype within a cycle could save a large part of that  
518 phenotyping this way, with no cost beyond the genotypes it already has.

519 The comparison with other tools is more modest. On IRRI, GA+SA with the same exact  
520 criterion came within 0.009 of TSOpt, and TSOpt was ahead of STPGA and TSDFGS on  
521 average and in almost every split. Once the overlap between splits was accounted for,  
522 however, only the margin over TSDFGS with CD was clearly different from zero, and part of  
523 STPGA's shortfall came from its default settings: with a ridge parameter matched to the  
524 heritability and a longer search, it came within 0.026 of TSOpt, at about half an hour per  
525 design. On the public panels, TSOpt was ahead of STPGA and TSDFGS by about 0.02 on  
526 average but won only 58–64% of scenarios, and against TrainSel's own published designs it  
527 was ahead by 0.005, mostly with small training sets. This fits the comparison by Fernández-  
528 González et al. (2023), in which targeted CDmean designs were among the best methods but  
529 their advantage varied from dataset to dataset. We therefore do not claim that TSOpt is more  
530 accurate than existing tools when they are given the same problem. What it adds is speed and  
531 scope. A design takes about a tenth of a second rather than seconds or minutes, and the same  
532 engine covers sparse testing, historical data and costs. Speed is what turns design into a  
533 routine step, because a breeder can redesign for each trait, target set, budget and season, check  
534 how sensitive a plan is before planting, and design for the thousands of candidates of an early-  
535 stage trial, where a search that takes minutes for 250 lines is not practical.

536 The robust criterion behaved as insurance rather than as a source of extra accuracy. On IRRI,  
537 any assumed heritability between 0.2 and 0.8 gave practically the same designs, and on the  
538 public panels the robust design was ahead of a fixed heritability of 0.5 by a small margin that  
539 was not significant. But on the public panels, when the assumed heritability was too low, the  
540 loss was large: designs made for a heritability of 0.1 or 0.2 gained only a third to a half of  
541 what the robust design gained. Since breeders rarely know the heritability of the coming

season, averaging protects against the costly mistake at little expense. Projecting out the intercept made no measurable difference to realised accuracy. We keep it because it makes the criterion match the model that will be fitted, which matters once environment means or other fixed effects enter, and our check shows that, with  $c$  near  $10^6$ , the flat-prior form reproduces the GLS criterion to within  $10^{-9}$ .

Two further results suggest that optimising the criterion harder does not pay. Exchange steps after the greedy search raise the criterion but did not raise realised accuracy, and with a single assumed heritability they sometimes lowered it. GA+SA, a much longer search on the same criterion, was not better than the greedy design. Beyond a point, a higher CDmean for one assumed model seems to fit that model rather than the data.

The weak population structure of the IRRI lines helps explain why the stratified heuristics failed. Stratified sampling and cluster medoids help mainly when a population falls into distinct groups (Isidro et al. 2015; Rincent et al. 2017). In an elite population whose first principal component explains only 6% of the genomic variance, clusters carry little information. What matters instead is the balance between how closely a candidate is related to the targets and how much it duplicates other candidates, and that balance is what CDmean measures.

### *Forward in time the gains are smaller, and the history matters most*

Replaying IRRI's own decisions gave a more sober picture than cross-validation within seasons. Choosing which new lines to phenotype first still helped on average, but all design methods gained by similarly small margins, TSOpt beat random choice in only 23 of 30 replicates, and STPGA was slightly ahead of TSOpt once 45–60 new lines were phenotyped. Accuracy in the later seasons depended mainly on the history, and all 209 historical lines together beat every subset. Several things probably contribute. A few first-season phenotypes add little to hundreds of historical records. These phenotypes come from a single season and

567 carry its genotype  $\times$  season interaction (Fig. 1b). In addition, the criterion rewards training  
568 lines that predict the targets under the model, not under the conditions of the seasons to come.  
569 For a running programme, the practical message is that design matters most when the training  
570 set is small or must be trimmed, and least when a large history is already in hand. A criterion  
571 built on the genetic covariance between the testing season and the target seasons, which  
572 TSOpt's multi-environment form already allows, is the obvious next step.  
573 Phenotyping the lines most related to the targets, a common rule of thumb, did not help when  
574 the decision was which lines to phenotype. It matched random sampling within seasons  
575 (0.000), was worse for choosing new lines to phenotype first ( $-0.019$ ), and lowered predictive  
576 ability by 0.114 on the public panels. Only when trimming an existing history did related lines  
577 beat random subsets (0.027, close to TSOpt's 0.032). How relatives are chosen matters.  
578 Ranking candidates by their mean relationship to all targets concentrates the training set on a  
579 few central lines and covers little of the genetic space in which the targets lie. Giving each  
580 target in turn its nearest relative spreads the set over all targets and recovered more than half  
581 of TSOpt's gain within seasons (0.056 against 0.097). CDmean goes one step further, because  
582 it also discounts candidates that duplicate lines already chosen.

### 583 *Sparse testing: the most practical gain*

584 Multi-environment programmes decide not only which lines to test but also where to test  
585 them. In IRRI's 2011–2012 seasons, TSOpt's allocations predicted untested line  $\times$  season  
586 cells better than the balanced random allocations recommended for sparse testing (Jarquin et  
587 al. 2020), when plots were scarce, and this held when every line had to be tested at least once,  
588 as most breeders would require. Scoring with correlations the design never saw shrank the  
589 gain by about a fifth, which shows that part of the original advantage came from scoring  
590 under the design's own model, but the rest held at a quarter of the cells. With half the cells  
591 tested, a designed allocation was no better than a balanced random one. The overlap/non-

overlap designs did worse than both, mainly because they leave many lines untested when plots are scarce. The design needed nothing a programme does not already have, only its genotypes and the season-type genetic correlations from earlier years. Of all the results here, this one transfers most directly to practice. It complements the work of Liu and Liao (2026), who optimised CDmean training sets for multi-environment trials by choosing lines rather than line  $\times$  environment cells. The genetic correlations between seasons were high (0.70–0.76), and with stronger genotype  $\times$  environment interaction, or with many locations, the value of a designed allocation may well differ.

### ***Simple models are enough for design***

Designing under models with epistasis, haplotypes or omics profiles did not change which lines were worth phenotyping. Such models can change predictions (Rincent et al. 2018; Schrag et al. 2018), but the additive, heritability-averaged default already gave designs that held up under them. For design, modelling gene action in more detail seems to add little, whereas using the history, the genotype  $\times$  environment structure and an exact criterion adds a good deal.

### ***Limitations***

The data come from late-stage replicated yield trials of one programme, at one site, over four years, more than a decade ago. Breeding pipelines have since moved genomic selection to earlier stages with many more candidates, where the speed of the design matters more than here, and the gains we measured may differ there. The forward and sparse-testing results rest on a single new cohort and a single set of seasons, so their replicates describe the stability of each method, not how results would vary between cohorts or programmes. The 321 lines used for sparse testing are those that remained in the trials for all four seasons and may not represent all entries. Most of the public panels are diversity panels rather than breeding populations. Phenotypes were analysed in two stages through season BLUEs, and a one-stage

analysis might change absolute accuracies, although it should not change the comparisons. Better settings narrowed STPGA's gap, at a large cost in time, and we did not tune TSDFGS. TrainSel's licence did not allow us to run it, so our comparison uses the sets its authors published, which tests TrainSel as its developers used it but only on the public panels. A pre-registered comparison by an independent group licensed to use TrainSel, following the protocol in the TSOpt repository, would settle this. The hybrid, testcross, multi-trait and major-gene designs are exact under their models but were not tested on real data here. Finally, TSOpt is distributed free of charge under a proprietary licence, so reproducing these results requires registration, and the source code is available to reviewers and researchers on request.

#### ***Recommendations for breeding programmes***

Design the training set instead of picking the lines most related to the targets on average; if no design tool is at hand, giving each target its nearest relative is a reasonable second best. Within a breeding cycle, a fraction of a second of computing can stand in for a large share of the phenotyping budget, especially when the training set is small. Train on all historical lines, and use design to choose which new lines to phenotype and, if the history must be trimmed, which lines to keep, while expecting smaller gains forward in time than cross-validation suggests. When plots are scarce, let the design decide how plots are spread across sites and seasons, with every line tested at least once, using genetic correlations estimated from the programme's own trials. Before relying on a plan, replay it on the programme's history, because the size of the gain depends on the population, the trait and the stage of the pipeline. The next steps are forward validation across several programmes, crops and cohorts, larger multi-location networks, and design criteria aimed at the covariance between the seasons in which lines are tested and those in which they will be grown. TSOpt reads from and writes to breeding databases through BrAPI, so both can be done on a programme's live data from one season to the next.

## 642 **Supplementary Information**

643 The online version contains supplementary material.

- 644 • **Online Resource 1** Population structure of the 328 IRRI elite lines: principal  
645 components of the genomic relationship matrix, by cohort (PDF)
- 646 • **Online Resource 2** Training-set design methods on six public panels (33 scenarios):  
647 gain in predictive ability, NDCG@10% and CDmean over random sampling, and  
648 computing time (XLSX)
- 649 • **Online Resource 3** Analysis scripts that reproduce every number and figure from the  
650 public data (ZIP)
- 651 • **Online Resource 4** Statistical details and checks: corrected tests for every  
652 comparison, bootstrap intervals for plots saved, forward and sparse-testing stability,  
653 published TrainSel designs by panel, heritability probes, and exactness checks of the  
654 engine (PDF)

655

656 **Acknowledgements** I thank J. Spindel, S. McCouch and colleagues, and the irrigated rice  
657 breeding programme of the International Rice Research Institute, for making the genotypes  
658 and multi-environment trial data public

659 **Author contribution statement** WH conceived the study, designed and wrote the software,  
660 performed the analyses and wrote the manuscript.

661 **Statements and Declarations**

662 **Funding** This work was funded by PlantNura Consulting & Solutions. No external funding  
663 was received.

664 **Competing interests** The author is the founder of PlantNura Consulting & Solutions, which  
665 develops TSOpt and distributes it free of charge.

666 **Ethics approval** Not applicable.

667 **Use of generative AI**

668 The author used Claude (Anthropic), a large language model, to help and revise and  
669 troubleshoot the codes. The author reviewed, checked, and edited all code and text, and takes  
670 full responsibility for the content of the publication. Grammarly was used to edit and improve  
671 the language of the manuscript

672 **Data availability** The IRRI trial and genotype data are public under a CC0 licence (Spindel et  
673 al. 2015b) and mirrored at Zenodo (record 4934442). The six public panels and the published  
674 TrainSel training sets are available from Fernández-González et al. (2023). The analysis  
675 scripts (Online Resource 3), the benchmark scripts and every raw benchmark result are  
676 available at <https://github.com/whussain2/TSOpt>.

677 **Code availability** TSOpt 0.5.0 is available free of charge to all users, including for  
678 commercial use, as binary packages for Windows, macOS and Linux from  
679 <https://github.com/whussain2/TSOpt>, with user guides, a tutorial with example datasets and a  
680 reference manual. A personal licence key is issued on request by email  
681 ([waseemhussain@plantnura.com](mailto:waseemhussain@plantnura.com)). The source code is available to qualified researchers for

research and verification, on request to the author, under an agreement that does not permit redistribution. Reviewers receive a licence key and, on request, the source code. All results use TSOpt's default criterion (robust CDmean) unless stated otherwise.

## References

- Akdemir D, Isidro-Sánchez J (2019) Design of training populations for selective phenotyping in genomic prediction. *Sci Rep* 9:1446. <https://doi.org/10.1038/s41598-018-38081-6>
- Akdemir D, Rio S, Isidro y Sánchez J (2021) TrainSel: An R package for selection of training populations. *Front Genet* 12:655287. <https://doi.org/10.3389/fgene.2021.655287>
- Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. *Genet Sel Evol* 47:38. <https://doi.org/10.1186/s12711-015-0116-6>
- Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype  $\times$  environment interaction using pedigree and dense molecular markers. *Crop Sci* 52:707–719. <https://doi.org/10.2135/cropsci2011.06.0299>
- Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci* 22:961–975. <https://doi.org/10.1016/j.tplants.2017.08.011>
- Fernández-González J, Akdemir D, Isidro y Sánchez J (2023) A comparison of methods for training population optimization in genomic selection. *Theor Appl Genet* 136:30. <https://doi.org/10.1007/s00122-023-04265-6>
- Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12. <https://doi.org/10.2135/cropsci2008.08.0512>
- Isidro J, Jannink J-L, Akdemir D, et al (2015) Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128:145–158. <https://doi.org/10.1007/s00122-014-2418-4>

707 Isidro y Sánchez J, Akdemir D (2021) Training set optimization for sparse phenotyping in  
 708 genomic selection: A conceptual overview. *Front Plant Sci* 12:715910.  
 709 <https://doi.org/10.3389/fpls.2021.715910>  
 710 Jarquin D, Howard R, Crossa J, et al (2020) Genomic prediction enhanced sparse testing for  
 711 multi-environment trials. *G3 (Bethesda)* 10:2725–2739.  
 712 <https://doi.org/10.1534/g3.120.401349>  
 713 Kaufman L, Rousseeuw PJ (1990) Finding groups in data: An introduction to cluster analysis.  
 714 Wiley, New York  
 715 Laloë D (1993) Precision and information in linear models of genetic evaluation. *Genet Sel*  
 716 *Evol* 25:557–576. <https://doi.org/10.1186/1297-9686-25-6-557>  
 717 Liu Z-J, Liao C-T (2026) Optimizing training sets for genomic selection to identify superior  
 718 genotypes across multiple environments. *G3 (Bethesda)* 16:jkag031.  
 719 <https://doi.org/10.1093/g3journal/jkag031>  
 720 Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using  
 721 genome-wide dense marker maps. *Genetics* 157:1819–1829.  
 722 <https://doi.org/10.1093/genetics/157.4.1819>  
 723 Murillo DA, Gezan SA, Heilman AM, et al (2021) FieldHub: A Shiny app for design of  
 724 experiments in life sciences. *J Open Source Softw* 6:3122.  
 725 <https://doi.org/10.21105/joss.03122>  
 726 Nadeau C, Bengio Y (2003) Inference for the generalization error. *Mach Learn* 52:239–281.  
 727 <https://doi.org/10.1023/A:1024068626366>  
 728 Ou J-H, Liao C-T (2019) Training set determination for genomic selection. *Theor Appl Genet*  
 729 132:2781–2792. <https://doi.org/10.1007/s00122-019-03387-0>  
 730 Rincent R, Charcosset A, Moreau L (2017) Predicting genomic selection efficiency to  
 731 optimize calibration set and to assess prediction accuracy in highly structured

populations. *Theor Appl Genet* 130:2231–2247. <https://doi.org/10.1007/s00122-017-2956-7>

Rincent R, Charpentier J-P, Faivre-Rampant P, et al (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar. *G3 (Bethesda)* 8:3961–3972. <https://doi.org/10.1534/g3.118.200760>

Rincent R, Laloë D, Nicolas S, et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192:715–728. <https://doi.org/10.1534/genetics.112.141473>

Schrag TA, Westhues M, Schipprack W, et al (2018) Beyond genomic prediction: Combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208:1373–1385. <https://doi.org/10.1534/genetics.117.300374>

Selby P, Abbeloos R, Backlund JE, et al (2019) BrAPI—an application programming interface for plant breeding applications. *Bioinformatics* 35:4147–4155. <https://doi.org/10.1093/bioinformatics/btz190>

Sherman J, Morrison WJ (1950) Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann Math Stat* 21:124–127. <https://doi.org/10.1214/aoms/1177729893>

Spindel J, Begum H, Akdemir D, et al (2015a) Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet* 11:e1004982. <https://doi.org/10.1371/journal.pgen.1004982>

755 Spindel J, Begum H, Akdemir D, et al (2015b) Data from: Genomic selection and association  
756 mapping in rice (*Oryza sativa*). Dryad Digital Repository.  
757 <https://doi.org/10.5061/dryad.7369p>  
758 VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci  
759 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>  
760