A framework for simulating genotype by environment interaction using multiplicative models

Jon Bančič (jbancic@ed.ac.uk)
The University of Edinburgh The Roslin Institute https://orcid.org/0000-0001-7077-7163
Gregor Gorjanc
The University of Edinburgh The Roslin Institute
Daniel Tolhurst
The University of Edinburgh The Roslin Institute https://orcid.org/0000-0002-4787-080X

Research Article

Keywords: genotype by environment interaction , multiplicative models , multi-environment trials , plant breeding simulation

Posted Date: January 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3855188/v1

A framework for simulating genotype by environment interaction using multiplicative models

J. Bančič, G. Gorjanc, D.J. Tolhurst

Received: date / Accepted: date

Key message The simulation of genotype by environment interaction using multiplicative models provides a general framework to generate realistic multi-environment datasets and model plant breeding programmes.

Abstract This paper develops a general framework for simulating genotype by environment interaction (GEI) using multiplicative models. Many stages of plant breeding are complicated by GEI, from the selection of potential parents to the development of improved genotypes for release to growers. Despite its importance, however, current plant breeding simulations do not adequately capture the complexity of GEI because they either use unrealistic models to simulate it or they ignore it completely. The framework developed in this paper simulates the two main components of GEI, that is non-crossover and crossover interaction, using the class of multiplicative models. The framework is demonstrated using two working examples supported by R code. The first example embeds the framework into a linear mixed model to generate MET datasets with low, moderate or high GEI, which are then used to compare various statistical models widely used in plant breeding. The results show that the prediction accuracy of all models increases as the level of GEI decreases or the number of sampled environments increases. The second example integrates the framework into a breeding programme simulation to compare genomic and phenotypic selection strategies over time. The results show that genomic selection outperforms phenotypic selection by $1.4-1.8$ times, depending on the level of GEI. These examples demonstrate how the new framework can be used to generate realistic MET datasets and model plant breeding programmes that better reflect the complexity of real-world settings, making it a valuable tool for breeders to optimize their breeding programmes. The framework also has broader applications beyond plant breeding, including animal, aquaculture and tree breeding as well as other analysis comparison settings.

Keywords genotype by environment interaction · multiplicative models · multi-environment trials · plant breeding simulation

J. Bančič D.J. Tolhurst ✉
The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, United Kingdom
E-mail: jbancic@ed.ac.uk E-mail: dtolhurs@ed.ac.uk

1 Introduction

Plant breeding is complicated by the fact that genotypes respond differently to different environments, a phenomenon known as genotype by environment interaction (GEI). Despite its importance, however, current plant breeding simulations do not adequately capture the complexity of GEI because they either use unrealistic models to simulate it or they ignore it completely. The framework developed in this paper simulates GEI using multiplicative models. The framework can be used to simulate realistic multi-environment trial (MET) datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

Plant breeding has been historically shaped by GEI, from the selection of potential parents to the development of improved genotypes for release to growers. GEI can be broadly categorised as either non-crossover or crossover interaction, which reflect changes in the magnitude (scale) of genotype response between environments or changes in genotype rank (Gail and Simon, 1985; Baker, 1988, Fig. 1). Crossover GEI is of particular importance to breeders because their selection decisions are more complicated by changes in rank than changes in scale (Baker, 1990; Eisemann et al., 1990). Plant breeders gauge the magnitude and form of GEI in their programmes by accumulating MET datasets, which contain a sample of environments that generally span multiple years and locations (Smith et al., 2021). An important consideration when constructing a MET dataset is the extent to which it represents the breeder's target population of environments (TPE, Comstock and Moll, 1963; Cooper et al., 1993). This is referred to as the MET-TPE alignment (Cooper et al., 2023).

Multiplicative models have gained popularity in plant breeding because they are effective at capturing non-crossover and crossover GEI. The most general model for GEI is the unstructured model, which fits a separate genetic variance for each environment and a separate genetic covariance for each pair of environments. The unstructured model captures the maximum amount of GEI in the data, however, it becomes computationally prohibitive and unnecessarily complicated as the number of sampled environments increases. These issues can be overcome using *reduced rank* multiplicative models. The appealing feature of multiplicative models is that they capture a large proportion of GEI with a small number of multiplicative terms, where each term is the product of an environmental effect and a genotype effect (Mandel, 1971). Some traditional examples include AMMI (Kempton, 1984; Gauch, 1988), GGE (Cornelius et al., 1996; Yan et al., 2000) and factor analytic models (Piepho, 1997; Smith et al., 2001). These approaches have been shown to provide an informative model for GEI and a good fit to MET datasets in general (Gauch et al., 2008; Kelly et al., 2007). The extensive theory and advantages of multiplicative models provide a good foundation for not only modelling GEI but also simulating it.

Simulations are routinely used in plant breeding as a fast and cost-effective way to compare different statistical approaches. Several studies have generated MET datasets for the purpose of addressing their research objectives. For example, Hartung et al. (2023) take empirical datasets and reproduce these datasets in simulation to assess the efficiency of new statistical approaches (also see Lisle, 2023). However, there is currently no reproducible and scalable framework for simulating MET datasets with different levels of GEI.

Simulations are also routinely used to compare different breeding strategies over time. Numerous simulation packages have been developed to model plant breeding programmes, including AlphaSimR (Gaynor et al., 2021), ADAM-Plant (Liu et al., 2019), ChromaX (Younis et al., 2023), GPOPSIM2 (Li et al., 2021), MOBPS (Pook et al., 2020) and QUGENE (Podlich and Cooper, 1998). These packages provide a fast and cost-effective approach for comparing different breeding strategies, however, they all over-simplify GEI which can result in optimistic projections and spurious comparisons. For example, AlphaSimR, ChromaX and GPOPSIM2 construct a single phenotype for each genotype comprising a main effect, interaction effect and error. The interaction effect is generally modelled through a single multiplicative term, where the environmental effect is randomly sampled and consequently difficult to control. It is important to note that most of these packages do have the functionality to simulate multiple environments as multiple correlated traits, so they do have the potential to implement a more realistic framework for GEI. The examples above highlight the need for a general and scalable framework for simulating GEI in MET datasets and plant breeding programmes.

The aim of this paper is to develop a general framework for simulating GEI using multiplicative models. The framework can be used to simulate various plant breeding settings, including different levels of non-crossover and crossover GEI, different correlated genetic effects and multiple TPE and/or traits. This paper develops the theory of the framework and demonstrates its application using two working examples supported by R code. The first example simulates MET datasets with different levels of GEI, which are then used to compare various statistical models widely used in plant breeding. The second example integrates the framework into a breeding programme simulation to compare different selection strategies over time. Lastly, the framework has broader applications beyond plant breeding as well as other analysis comparison settings.

2 Material and Methods

This section develops the framework for simulating GEI using multiplicative models. The framework is initially

developed for simulating the genetic effects, and then embedded within a linear mixed model to generate phenotypes which also include appropriate non-genetic effects.

The methods consist of two parts:

1. Simulating genotype by environment effects; including how to simulate a between-environment genetic variance matrix with realistic structure and complexity and how to tune this matrix using measures of variance explained.
2. Simulating phenotypes; including how to simulate a breeder's TPE and sample a MET dataset and how to obtain measures of accuracy that describe the phenotypes.

Each part is detailed in the following.

2.1 Simulating genotype by environment effects

The framework simulates genetic effects which capture GEI. Assume the genetic effects are simulated for v genotypes in p environments, hereafter referred to as the genotype by environment (GE) effects. Let the vp-vector of GE effects be given by $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_p^\top)^\top$, where $\mathbf{u}_j$ is the v-vector for the j^{th} environment. The GE effects are simulated as:

$$\mathbf{u} \sim \mathrm{N}\big(\mathbf{0}, \mathbf{G_e} \otimes \mathbf{G}\big), \tag{1}$$

where $\mathbf{G_e}$ is a $p \times p$ between-environment genetic variance matrix and $\mathbf{G}$ is a $v \times v$ genotype relationship matrix. The matrix $\mathbf{G_e}$ is initially formulated according to an unstructured model and then reformulated according to a *reduced rank* multiplicative model. The matrix $\mathbf{G}$ is completely general and may represent a known/simulated pedigree or genomic relationship matrix constructed through a breeding simulation package (see Section 3.2). Both matrices are assumed to be positive (semi)-definite.

The unstructured model provides the most general form for simulating $\mathbf{G_e}$ using all $p(p-1)/2$ parameters. This generates a heterogeneous GEI pattern based on a different genetic variance for each environment, $\sigma^2_{g_j}$, and a different genetic covariance for each pair of environments, $\sigma_{g_{ij}}$. The unstructured model can be written as a multiplicative model with all p terms:

$$\begin{aligned}\mathbf{u} &= \big(\mathbf{s}_1 \otimes \mathbf{f}_1\big) + \ldots + \big(\mathbf{s}_p \otimes \mathbf{f}_p\big) \\ &= \big(\mathbf{S} \otimes \mathbf{I}_v\big)\mathbf{f},\end{aligned} \tag{2}$$

where $\mathbf{S} = [\mathbf{s}_1 \ \ldots \ \mathbf{s}_p]$ is a $p \times p$ matrix of environmental effects (covariates) and $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_p^\top)^\top$ is a vp-vector of genotype effects (slopes). The covariates and slopes are obtained from the eigendecomposition given by:

$$\mathbf{G_e} = \mathbf{ULU}^\top, \tag{3}$$

where $\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_p]$ is an orthogonal matrix of eigenvectors and $\mathbf{L} = \oplus_{r=1}^{p} l_r$ is a diagonal matrix of eigenvalues sorted in decreasing order. The covariates are obtained by setting $\mathbf{S} = \mathbf{U}$ and the slopes are simulated as $\mathbf{f} \sim \mathrm{N}(\mathbf{0}, \mathbf{L} \otimes \mathbf{G})$. The proportion of variance explained by the r^{th} term can be calculated as $l_r / \sum_{r=1}^{p} l_r$, where the denominator is equivalent to the sum of the diagonal elements of $\mathbf{G_e}$ given by $\sum_{j=1}^{p} \sigma^2_{g_j}$. A large proportion of variance in plant breeding data is typically explained by the first few terms, which makes the full rank form in Eq. 2 unnecessary as p increases.

The reduced rank form of Eq. 2 arises from taking the first k eigenvectors in Eq. 3, which gives:

$$\mathbf{u} = \big(\mathbf{S}_k \otimes \mathbf{I}_v\big)\mathbf{f}_k, \tag{4}$$

where $\mathbf{S}_k$ is a $p \times k$ matrix and $\mathbf{f}_k$ is a vk-vector, with $\mathbf{f}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}_k \otimes \mathbf{G})$. The genotype slopes can be simulated independently or with a breeding simulation package by defining each term as a separate trait, with mean vector set to $\mathbf{0}$ and variance matrix set to $\mathbf{L}_k$. This formulation requires just k traits (terms) to be simulated, which makes the reduced rank model in Eq. 4 even more appealing than the full rank model (see Section 3.2).

The GE effects are therefore simulated as:

$$\mathbf{u} \sim \mathrm{N}\big(\mathbf{0}, \mathbf{S}_k\mathbf{L}_k\mathbf{S}_k^\top \otimes \mathbf{G}\big), \tag{5}$$

where $\mathbf{G_e} \simeq \mathbf{S}_k\mathbf{L}_k\mathbf{S}_k^\top$ is a reduced rank between-environment genetic variance matrix. The framework requires $\mathbf{G_e}$ to be previously known/simulated prior to obtaining the environmental covariates and genotype slopes from the eigendecomposition. It is possible to simulate the covariates and slopes directly (see Lisle, 2023), however, this typically leads to an uncontrollable structure for $\mathbf{G_e}$ and spurious correlations between environments.

Simulating between-environment genetic variance matrix

An important feature of the framework is a reproducible approach for simulating $\mathbf{G_e}$ with realistic structure and complexity. This is achieved by explicitly simulating heterogeneity of scale variance and lack of genetic correlation through:

$$\mathbf{G_e} = \mathbf{D_e}^{1/2}\mathbf{C_e}\mathbf{D_e}^{1/2}, \tag{6}$$

where $\mathbf{D_e}$ is a $p \times p$ diagonal genetic variance matrix with diagonal elements given by $\sigma^2_{g_j}$ and $\mathbf{C_e}$ is a $p \times p$ reduced rank between-environment genetic correlation matrix with off-diagonal elements given by $\rho_{ij} = \sigma_{g_{ij}} / \sigma_{g_i}\sigma_{g_j}$.

The genetic variances in $\mathbf{D_e}$ are simulated as $\sigma^2_{g_j} \sim$ Inv-Gamma(α, β), where α is the shape parameter and β is the scale parameter. These parameters are set to $\alpha = 5$ and $\beta = 5$ for all examples in this paper (Fig. 2), but note that other values can be used where required. The inverse gamma distribution was chosen to ensure that the genetic variances are positive and have a skewed distribution, but other distributions can also be used.

Following Hardin et al. (2013), the between-environment genetic correlation matrix is simulated as:

$$\mathbf{C_e} = \rho\mathbf{J}_p + \epsilon\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}, \tag{7}$$

where ρ is the baseline genetic correlation, $\mathbf{J}_p$ is a $p \times p$ matrix of ones, ϵ is the variability of the correlations (magnitude of structured noise) around the baseline and $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_p]$ is a $(k-1) \times p$ matrix of simulated noise in which $\boldsymbol{\lambda}_j$ is the $(k-1)$-vector for the j^{th} environment. The reduced rank form of $\mathbf{C_e}$ arises from the fact that $\mathbf{J}_p$ has rank 1 and $\boldsymbol{\Lambda}$ has rank $k-1$, or more specifically that $\rho\mathbf{J}_p + \epsilon\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$ has rank k when $\rho, \epsilon \neq 0$. Note that other base correlation functions can be used instead of $\mathbf{J}_p$ where required (see Hardin et al., 2013).

The baseline correlation is subject to the constraint $0 \leq \rho < 1$, which ensures $\mathbf{C_e}$ is positive (semi)-definite. If the constraint is not imposed and $-1 < \rho < 0$, spurious and negative definite matrices may be generated that require bending. The noise is also subject to the constraint $\epsilon = 1 - \rho$, which ensures the rank of $\mathbf{C_e}$ equals k when $\rho > 0$. If the constraint is not imposed and $\epsilon < 1 - \rho$, the rank of $\mathbf{C_e}$ will equal p. The first k terms will still capture the majority of variation in $\mathbf{G_e}$, but now the remaining $p - k$ terms will each capture a small proportion of variance given by $1 - \rho - \epsilon$.

An extension of Hardin et al. (2013) is used to simulate the genetic correlations based on a skewed distribution. The columns of $\boldsymbol{\Lambda}$ are simulated as:

$$\boldsymbol{\lambda}_j \sim \begin{cases} U(-1, 1+\gamma) & -1 \leq \gamma \leq 0 \\ U(-1, 1-\gamma) & 0 < \gamma \leq 1, \end{cases} \tag{8}$$

where γ governs the amount of skewness. The $\boldsymbol{\lambda}_j$ are then scaled to unit length, i.e. $\boldsymbol{\lambda}_j^\top\boldsymbol{\lambda}_j = 1$. When $0 < \gamma \leq 1$, the between-environment genetic correlation matrix in Eq. 7 is constructed as $\mathbf{C_e} = -(\rho\mathbf{J}_p + \epsilon\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda})$, which ensures the correct matrix is obtained. Note that when $\gamma = 0$, the baseline correlation ρ equals the mean correlation between environments given by $\bar{p} = \sum_{i<j}^{p} 2\rho_{ij}/p(p-1)$, but not when $\gamma \neq 0$.

Different structure for $\mathbf{C_e}$ can be generated by altering ρ, ϵ, γ and k. The examples in Supplementary Fig. 1 demonstrate that decreasing ρ decreases the mean correlation, increasing ϵ increases the variability of the correlations around the mean, altering γ changes the skew of the distribution and increasing k increases the rank of the noise. The practical implication is that by changing k, the amount of structure in the noise can be changed from high to low or even no structure.

The framework above was used to simulate the three examples of $\mathbf{C_e}$ presented in Fig. 3 and summarised in Table 1. The matrices were constructed with $k = 7$, but with varying ρ, ϵ and γ. All matrices were then multiplied with $\mathbf{D_e}$ in Fig. 2 to create three examples of $\mathbf{G_e}$ using Eq. 6. These matrices form the basis of the low, moderate and high GEI scenarios used throughout the remainder of the paper.

Measures of variance explained

An important supplement to simulating $\mathbf{G_e}$ are measures of variance explained for the GE effects. The measures are used to quantify and tune the proportion of (i) main effect and interaction variances, and (ii) non-crossover and crossover variances simulated in $\mathbf{G_e}$.

1. The proportion of genotype main effect variance is:

$$v_g = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2}, \tag{9}$$

where σ_g^2 is the main effect variance and σ_{ge}^2 is the pooled interaction variance. The main effect variance is obtained as $\sigma_g^2 = \bar{\mathbf{s}}_k\mathbf{L}_k\bar{\mathbf{s}}_k^\top$, where $\bar{\mathbf{s}}_k = \mathbf{1}_p^\top\mathbf{S}_k/p$ is a k row-vector of means for each environmental covariate. The interaction variance is then obtained as the mean diagonal element of $\mathbf{S}_k^*\mathbf{L}_k\mathbf{S}_k^{*\top}$, where $\mathbf{S}_k^* = \mathbf{S}_k - \bar{\mathbf{s}}_k \otimes \mathbf{1}_p$ is a $p \times k$ matrix of column centred environmental covariates. The proportion of interaction variance is therefore given by $v_{ge} = 1 - v_g$.

2. The proportion of non-crossover variance is:

$$v_n = \frac{\sigma_n^2}{\sigma_g^2 + \sigma_{ge}^2}, \tag{10}$$

where σ_n^2 is the non-crossover variance, which is obtained as $\sigma_n^2 = \sigma_g^2 + \bar{\mathbf{s}}_k\mathbf{L}_k\mathbf{S}_k^{*\top}\mathbf{S}_k^*\mathbf{L}_k\bar{\mathbf{s}}_k^\top/p\sigma_g^2$ (see Appendix A). The non-crossover variance captures all variation attributed to perfect positive correlation with the genotype main effects (Tolhurst, 2023). The crossover variance then captures all remaining variation independent of the main effects. The proportion of crossover variance is therefore given by $v_c = 1 - v_n$.

The measures of variance explained are demonstrated for the three examples of $\mathbf{G_e}$ summarised in Table 1. All examples have the same genetic variances given by $\sigma_{g_j}^2 = 0.01 - 9.95$, because they were all constructed with $\mathbf{D_e}$ in Fig. 2. The examples were classified as low, moderate or high GEI by tuning the proportions of non-crossover and crossover variance.

2.2 Simulating phenotypes

The framework for simulating the GE effects can be embedded within a linear mixed model to generate phenotypes which also capture appropriate non-genetic effects. The framework is summarised in Fig. 4, with example R code provided in Appendix B. Let the n-vector of phenotypes be given by $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_p^\top)^\top$, where $\mathbf{y}_j$ is the n_j-vector for the j^{th} environment. The linear mixed model used to simulate $\mathbf{y}$ is given by:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{11}$$

where μ is the overall trait mean, $\boldsymbol{\tau}$ is a p-vector of environmental main effects with $n \times p$ design matrix $\mathbf{X}$, $\mathbf{u}$ is the vp-vector of GE effects with $n \times vp$ design matrix $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$ is the n-vector of plot errors. A randomised complete block design is used for each environment in the example R code, with r replicate blocks of all v genotypes. Simulation of other experimental designs and additional non-genetic effects is straightforward.

The environmental main effects are simulated as:

$$\boldsymbol{\tau} \sim \mathrm{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_p), \tag{12}$$

where σ_e^2 is the main effect variance. This specification will be demonstrated for the MET dataset simulation in Section 3.1, and extended to a regression on environmental covariates for the breeding programme simulation in Section 3.2.

The GE effects are simulated using the framework in Section 2.1, which can be summarised by four key steps:

1. Construct a between-environment genetic variance matrix as $\mathbf{G_e} = \mathbf{D_e}^{1/2}\mathbf{C_e}\mathbf{D_e}^{1/2}$, tuned using the measures of variance explained.
2. Decompose the between-environment genetic variance matrix as $\mathbf{G_e} = \mathbf{ULU}^\top$, taking the first k terms to obtain $\mathbf{U}_k$ and $\mathbf{L}_k$.
3. Set the environmental covariates as $\mathbf{S}_k = \mathbf{U}_k$ and simulate the genotype slopes as $\mathbf{f}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}_k \otimes \mathbf{G})$, either independently or with a simulation package.
4. Construct the GE effects as $\mathbf{u} = (\mathbf{S}_k \otimes \mathbf{I}_v)\mathbf{f}_k$.

The framework can be used to generate additive, dominance and epistatic GE effects, by appropriately defining $\mathbf{G_e}$ and $\mathbf{G}$ for each.

Lastly, the plot errors are simulated as:

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{R}), \tag{13}$$

where σ_ε^2 is the average error variance and $\mathbf{R}$ is a $n \times n$ block diagonal matrix comprising a separate two-dimensional spatial model for each environment. Correlated plot errors can be generated in R using FieldSimR (Werner et al., 2023), but note that the example R code considers independent errors for brevity.

Simulating a TPE and sampling a MET datset

An additional feature of the framework is the ability to simulate a set of environments that represent a breeder's TPE. This process is summarised in Fig. 4 for the between-environment genetic correlation matrix, $\mathbf{C_e}$. Assume the vector of GE effects, $\mathbf{u}$, includes all p environments in the TPE. A subset of p_m environments is then sampled from the TPE, which may represent a subset observed in a particular year or multiple years in a MET dataset. The MET dataset may be used to compare various statistical approaches or constructed within a breeding programme simulation to compare different selection strategies. The number of environments in the TPE is set to $p = 1000$ for all examples in this paper as it produces a sufficiently large distribution for demonstration purposes, but p can be altered as required. Further structure can be applied to $\mathbf{C_e}$ which captures genotype by year and/or genotype by location interaction, which may be particularly appealing for simulating various TPE relevant to a breeding programme. Multiple TPE and multiple phenotypic traits are also considered in Appendix A.

Measures of accuracy

An important supplement to simulating the phenotypes are measures of accuracy. The measures are used to quantify the correlation between the (i) predicted main effects in the MET dataset and the true main effects in the TPE, (ii) true main effects in the MET dataset and TPE, referred to as the MET-TPE alignment (Cooper et al., 2023), and (iii) true and predicted GE effects in the MET dataset. The measures below are the expected values based on the true simulation parameters.

1. The expected main effect accuracy in the TPE is:

$$r_g = \sqrt{\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2/p_m + \sigma_\varepsilon^2/p_m r}}, \tag{14}$$

which is equal to the square root of the line-mean heritability across environments (Cooper and DeLacy, 1994).

2. The expected MET-TPE alignment is:

$$r_{mt} = \sqrt{\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2/p_m}}, \tag{15}$$

which is obtained by setting $\sigma_\varepsilon^2 = 0$ in Eq. 14. This measure will be used in Section 3.1 as the maximum main effect accuracy in the TPE.

The fundamental relationship between Eqs. 14 and 15 is then given by:

$$r_g = r_m \times r_{mt}, \tag{16}$$

where r_m is the expected main effect accuracy in the MET dataset, which is given by:

$$r_m = \sqrt{\frac{\sigma_g^2 + \sigma_{ge}^2/p_m}{\sigma_g^2 + \sigma_{ge}^2/p_m + \sigma_\varepsilon^2/p_m r}}, \tag{17}$$

where $\sigma_g^2 + \sigma_{ge}^2/p_m$ is the expected main effect variance sampled in the MET dataset. The numerator arises from an inflation of the true main effect variance in Eq. 14 by σ_{ge}^2/p_m, which represents the sampling error in the MET dataset. The practical implication is that the expected accuracy observed in the MET dataset will always be higher than in the TPE, i.e. $r_m \geq r_g$.

3. The expected accuracy of the GE effects in the MET dataset is:

$$r_{ge} = \sqrt{\frac{\sigma_g^2 + \sigma_{ge}^2}{\sigma_g^2 + \sigma_{ge}^2 + \sigma_\varepsilon^2/r}}, \tag{18}$$

which is equal to the square root of the line-mean heritability within environments.

The measures of accuracy are presented in Supplementary Fig. 2a-c for different values of σ_g^2, σ_{ge}^2 and p_m. These figures show that increasing the number of environments sampled in the MET dataset will increase the MET-TPE alignment and decrease the sampling error, to a point where the contribution of the interaction variance becomes negligible.

3 Results

The following sections showcase the application of the new framework by simulating realistic MET datasets and by integrating the framework into a breeding programme simulation.

3.1 MET dataset simulation

This section simulates a small example MET dataset using the framework developed in Section 2.1, which can be reproduced with the R code in Appendix B. The simulated MET dataset comprises 400 genotypes evaluated in field trials across 10 environments (Table 2). All trials are generated using a randomised complete block design with 2 replicate blocks comprising 5 columns and 40 rows each. The environmental main effects are sampled from a standard normal distribution. The GE effects are simulated based on a hypothetical trait with overall mean of 4 and genetic variances sampled from an inverse gamma distribution with shape parameter of 5 and scale parameter of 5. The plot errors are simulated assuming independence between plots, based on an overall plot-level heritability of 0.3. This produces heterogeneous environment means, variances and heritabilities.

The framework can also be used to simulate various plant breeding settings, including:

- Low, moderate and high GEI
- Multiple TPE and multiple phenotypic traits (Supplementary Fig. 3a-d)
- Correlated additive, dominance and epistatic genetic effects using AlphaSimR
- Correlated plot errors using FieldSimR
- Unbalanced and incomplete experimental designs.

Further R code is provided in the GitHub repository for some of these examples. The examples demonstrate how the framework provides a flexible approach for simulating realistic MET datasets.

Model comparison

This section compares eight statistical models using MET datasets built on the small example above. Three hypothetical TPEs were simulated with low, moderate and high GEI, each with 1,000 environments in total. These scenarios correspond to the between-environment genetic variance matrices presented in Fig. 3 and summarised in Table 1. Four MET datasets were then constructed for each level of GEI by randomly sampling 5, 10, 20 or 50 environments from each TPE (Fig. 4). This process was replicated 1,000 times, with eight statistical models fitted to each replicate. The models included main effects only, compound symmetry, main effects plus diagonal, diagonal and factor analytic of order one, two, three and four (see Tolhurst et al., 2022). The aim of the analyses was to obtain accurate predictions of the genotype main effects and GE effects. All models were fitted using ASReml-R (Butler et al., 2017).

The true parameters for all 1,000 replicates are summarised in Supplementary Fig. 4a-c. This figure compares the true parameters in each TPE with those sampled in the MET datasets. The parameters become more aligned to the TPE as more environments are sampled. This is the case for all levels of GEI, but note that fewer environments are required to be well-aligned for the low GEI scenario compared to the moderate and high GEI scenarios.

Fig. 5 presents the prediction accuracy of the eight statistical models fitted to simulated MET datasets with different levels of GEI and different numbers of environments. This figure also includes the expected main effect accuracy in the MET dataset and TPE (*dashed black lines*), and the expected MET-TPE alignment (*solid black line*) from Section 2.2. There are five important results:

1. All prediction accuracies decrease as the level of GEI increases. The largest differences occur between models at high GEI.
2. All prediction accuracies increase as the number of sampled environments increases. The largest differences occur between models at 5 environments for the main effects and at 50 environments for the GE effects, particularly for high GEI.
3. The main effect accuracies are higher in the MET dataset than in the TPE. The smallest differences occur between models at 50 environments, where the sampled MET datasets become more aligned to the TPE.
4. The main effect accuracies in the TPE are highest for the factor analytic models of order three and four. The differences between the remaining models are negligible except the factor analytic model of order one.
5. The prediction accuracies of the GE effects are also the highest for the factor analytic models of order three and four. The largest differences occur between models at high GEI and 50 environments.

The simulated MET-TPE alignments for all 1,000 replicates are presented in Supplementary Fig. 5. This figure provides further insight into the extent and form of the GEI simulated in the TPE and sampled in the MET datasets. Not only does the alignment decrease as the level of GEI increases or the number of sampled environments decreases, but the variability around the expected alignment also increases. As a result, well-aligned MET datasets can be constructed with as few as five environments, depending on the level of GEI. This is an interesting result that is the topic of current research.

Key summaries for the eight statistical models are presented in Supplementary Fig. 6a-b. This figure includes measures of reliability and model fit averaged

across all 1,000 replicates. The factor analytic models produce the most reliable estimates of the genetic variances and covariances in terms of root mean square error (RMSE). They also provide a superior fit to the data in terms of AIC and the proportion of variance explained, but have substantially longer running times.

3.2 Breeding programme simulation

This section integrates the framework developed in Section 2.1 into a breeding programme simulation. The three hypothetical TPE from Section 3.1 are again used for demonstration. The simulation involves 20 years of breeding for a hypothetical continuous trait, with 20 environments randomly sampled from the 1,000 environments in each TPE every year (Fig. 4). There are four stages of field evaluation, with an increasing number of environments and a decreasing number of genotypes observed in each stage (Fig. 6). This produces a subset of 400 environments from each TPE and a maximum of 20 environments observed in each stage, every year.

The following workflow was developed to integrate the framework into a simulation package:

1. *Simulate and decompose* a between-environment genetic variance matrix representing a hypothetical TPE. This produces the full set of reduced rank environmental covariates in each TPE, denoted $\mathbf{S}_k$.
2. *Simulate genotype slopes* in a founder population, denoted $\mathbf{f}_k$. This is achieved by defining each multiplicative term as a separate trait in the simulation package. The traits (terms) are simulated with mean vector set to $\mathbf{0}$ and variance matrix set to $\mathbf{L}_k$.
3. *Sample a subset of environments* from the full set of environments in the TPE. This produces a subset of environmental covariates for the current breeding year, denoted $\mathbf{S}_{k_i}$.
4. *Construct GE effects and phenotypes.* The true GE effects are constructed by multiplying $\mathbf{S}_{k_i}$ with the current genotype slopes, denoted $\mathbf{f}_{k_i}$. Phenotypes are then constructed for the genotypes and environments observed in each stage by adding error.
5. *Select and advance superior genotypes* based on the predicted genotype main effects. The predictions are obtained from analysing a MET dataset with a sample of environments that span multiple years and locations.
6. *Track genetic progress* via genetic gain, genetic variance and prediction accuracy. These measures are obtained based on the true GE effects in the MET dataset or TPE, with the latter obtained by multiplying $\mathbf{f}_{k_i}$ from Step 4 with $\mathbf{S}_k$ from Step 1.

Steps 1 and 2 are performed at the beginning of every simulation while Step 3 is performed once every year and Steps 4-6 are performed in each stage, every year for 20 years of breeding. Example R code for integrating the framework into AlphaSimR is available from the GitHub repository, but note that the workflow above can be integrated into many current simulation packages to efficiently simulate GEI at a very large scale.

An important component to integrating the framework into a breeding programme simulation is the ability to track the genetic gain and genetic variance in the MET dataset and TPE. This is achieved by extending the vector of environmental main effects, $\boldsymbol{\tau}$, to a regression on environmental covariates. The regression for the TPE is given by $\boldsymbol{\tau} = \mathbf{S}_k\boldsymbol{\tau}_{\mathbf{s}_{k_i}}$, where $\mathbf{S}_k$ is the $p \times k$ matrix of reduced rank environmental covariates and $\boldsymbol{\tau}_{\mathbf{s}_{k_i}}$ is a k-vector with elements given by the mean response of genotypes to each covariate, i.e. $\boldsymbol{\tau}_{\mathbf{s}_{k_i}} = \mathrm{E}(\mathbf{f}_{k_i})$. The genetic gain and genetic variance for the current breeding year are then calculated as $\mu_{g_i} = \mu + \bar{u}_{g_i}$ and $\sigma^2_{g_i} = \bar{\mathbf{s}}_k\mathbf{L}_{k_i}\bar{\mathbf{s}}_k$, where $\bar{u}_{g_i} = \bar{\mathbf{s}}_k\boldsymbol{\tau}_{\mathbf{s}_{k_i}}$ is the average genotype main effect and $\mathbf{L}_{k_i} = \mathrm{var}(\mathbf{f}_{k_i})$. The regression above can also be applied to the MET dataset simulation to induce a mean-variance ratio.

Breeding programme comparison

This section compares phenotypic and genomic selection strategies using a breeding programme simulation in AlphaSimR built on the workflow above. The key features of the breeding programme are presented in Fig. 6 and detailed in the Supplementary Material. The breeding programme was simulated with no, low, moderate and high GEI, and then phenotypic or genomic selection was applied for 20 years of breeding. This produced eight scenarios that were replicated 20 times. The MET dataset for phenotypic selection comprises the subset of sampled environments for each stage in the current year only (ranging from one for headrow to 20 for elite yield trial), and for genomic selection comprises all stages and sampled environments from the last three years (60 in total). A compound symmetry model was fitted for the genomic selection strategy as it produces sufficient results for demonstration purposes, but ideally factor analytic models should be considered. The aim of the simulation was to track genetic gain, genetic variance and the measures of accuracy in the headrow stage during 20 years of breeding (Figs. 7-9). The scenario without GEI will be used as a baseline for comparison (*solid black line*).

There are three important results for the genetic gain in Fig. 7:

1. The genetic gain decreases as the level of GEI increases. The largest differences occur after 20 years, where the genetic gain is $1.7 - 7.5$ times lower than the baseline for phenotypic selection and $1.6 - 6.2$ times lower for genomic selection.
2. The genetic gain from genomic selection is $1.4 - 1.7$ times higher than phenotypic selection after 20 years. The largest difference occurs for high GEI.
3. There are negligible differences between the genetic gain in the MET dataset and TPE for genomic se-

lection, but there are noticeable differences for phenotypic selection. The largest difference occurs for high GEI with ~ 0.4 units after 20 years.

There are three important results for the genetic variance in Fig. 8:

1. The loss in genetic variance decreases as the level of GEI increases. The largest differences occur after 20 years, where the genetic variance is $1.3 - 1.8$ times higher than the baseline for phenotypic selection and $1.7 - 2.8$ times higher for genomic selection.
2. The loss in genetic variance from genomic selection is $1.8 - 2.8$ times higher than phenotypic selection after 20 years. The largest difference occurs in the absence of GEI.
3. There are negligible differences between the genetic variance in the MET dataset and TPE for genomic selection, but there are substantial differences for phenotypic selection. The largest difference occurs for high GEI with ~ 0.1 units after 20 years.

There are three important results for the measures of accuracy in Fig. 9:

1. The main effect accuracy and MET-TPE alignment decrease as the level of GEI increases.
2. The main effect accuracy and MET-TPE alignment for genomic selection is $1.4 - 3.6$ times higher than for phenotypic selection. The largest difference occurs for high GEI.
3. The MET-TPE alignment for high GEI is much lower and more variable for phenotypic selection than genomic selection.

The genetic gain, genetic variance and measures of accuracy for all stages of the breeding programme are presented in Supplementary Figure 7a-c.

4 Discussion

Simulations are routinely used in plant breeding for optimising various statistical approaches and breeding strategies. Many of the current simulations, however, do not adequately capture the complexity of GEI inherent to real-world settings. The framework developed in this paper simulates GE effects that capture the two main components of GEI, that is non-crossover and crossover interaction, using multiplicative models. The utility of the framework was demonstrated for two working examples that compared different statistical models and breeding strategies in the presence of low, moderate and high GEI.

The framework for simulating GEI can be summarised by four key steps:

1. *Simulate a between-environment genetic variance matrix*, $\mathbf{G_e}$, with heterogeneous genetic variances and genetic correlations. Measures of variance explained were developed to tune the amount of non-crossover and crossover variance. This produces $\mathbf{G_e}$ with the required structure and complexity.
2. *Decompose* $\mathbf{G_e}$ to obtain the eigenvectors and eigenvalues for the first k terms. This produces a reduced rank set of vectors that capture the structure in $\mathbf{G_e}$.
3. *Obtain environmental covariates* as the eigenvectors and *simulate genotype slopes* based on the eigenvalues. This produces a reduced rank set of covariates and slopes, which can be generated using a simulation package.
4. *Construct GE effects* by multiplying the environmental covariates with the genotype slopes. This produces GE effects based on a reduced rank multiplicative model.

The framework can be embedded within a linear mixed model for simulating MET datasets or it can be integrated into simulation packages for modelling plant breeding programmes.

The framework features an approach for simulating heterogeneous genetic variances and genetic correlations in $\mathbf{G_e}$ from Step 1. The genetic variances are simulated from an inverse gamma distribution, which generates positive variances with a skewed distribution. The genetic correlations are then simulated following Hardin et al. (2013), which generates a correlation matrix by adding structured noise to a base correlation function. The rank of the noise dictates its structure while the base function dictates the underlying correlation structure, which can have many different forms, e.g. uniform, compound symmetry and autoregressive. In this paper, the approach of Hardin et al. (2013) was extended for simulating a between-environment genetic correlation matrix, $\mathbf{C_e}$, with genetic correlations based on a controllable skewed distribution. This approach can be used to reproduce known correlation distributions, which is particularly useful to compare different statistical approaches and breeding strategies at a much larger scale. The approach can also be used to add structured noise to $\mathbf{C_e}$ obtained from an empirical analysis. For example, noise can be added to environmental covariates that are either known or latent, e.g. estimated from factor analytic models (Tolhurst et al., 2022). This provides a general approach for simulating $\mathbf{G_e}$ that can be tailored to many plant breeding settings.

Measures of variance explained were developed to tune the parameters responsible for simulating different structure in $\mathbf{G_e}$. The measures quantify the magnitude of main effect and interaction variance, which are mostly controlled by the baseline correlation and skew. The measures also quantify the magnitude of non-crossover and crossover GEI, which reflect changes in the scale and rank of genotype response between environments. The non-crossover variance captures all variation attributed to perfect positive correlation with the genotype main effects (Tolhurst, 2023). The non-crossover variance is also controlled by the baseline correlation and skew as well as the shape and scale parameters of the inverse gamma distribution, which alter the

heterogeneity of scale variance. The crossover variance then includes all remaining variation independent of the main effects. The crossover variance arises from a lack of correlation between environments, which can also be controlled by the parameters above, and to a lesser extent by the variability of correlations and the rank of noise. In this paper, different levels of GEI were generated by adjusting the proportion of non-crossover and crossover variance in $\mathbf{G_e}$. The measures provide control over the construction of $\mathbf{G_e}$ and align the new framework with the historical partitioning of GEI into non-crossover and crossover interaction.

MET dataset simulation

The framework for simulating GEI can be embedded within a linear mixed model to generate realistic MET datasets. This is achieved by combining the simulated GE effects with appropriate non-genetic effects such as environmental main effects and plot errors. A linear mixed model was chosen because it provides a flexible basis to build a wide range of MET datasets. Some typical features include correlated additive, dominance and epistatic GE effects obtained from AlphaSimR (Gaynor et al., 2021), correlated plot errors obtained from FieldSimR (Werner et al., 2023) and unbalanced experimental designs, including incomplete block, *p*-rep and sparse testing. This provides a general approach for simulating MET datasets that can be tailored to many research objectives.

The framework was demonstrated for comparing eight statistical models fitted to simulated MET datasets with low, moderate or high GEI. There are three important results:

1. Gains in accuracy can be achieved by sampling more environments from the TPE in the MET dataset, despite losses in accuracy due to increasing GEI. This provides a framework to devise strategies for optimising the construction of MET datasets.
2. Negligible differences between models are observed for the genotype main effects in the TPE. This indicates that simpler models such as compound symmetry can be used to obtain accurate predictions of the main effects even for MET datasets with few environments.
3. Substantial differences between models are observed for the GE effects in the MET dataset. This indicates that more complex models are required to obtain accurate predictions of the GE effects, particularly for high GEI.

Overall, the results indicate that the factor analytic models of order 3 and 4 are superior, regardless of the level of GEI and the number of environments in the MET dataset. At least three factors were required in all cases to capture the extent of crossover GEI in the different MET datasets.

Breeding programme simulation

The framework for simulating GEI can be integrated within a simulation package to compare different breeding strategies. This is achieved by defining each multiplicative term as a separate trait, rather than defining each environment as a separate trait (see, for example, Liu et al., 2019). The practical implication is that the genetic effects generated during breeding are the genotype slopes, so they capture any changes in population structure over time. There are two appealing features of this approach. Firstly, the reduced rank model only requires a small number of traits (terms) to be simulated rather than a much larger number of environments across all years of breeding, e.g., 400 traits for all examples in this paper. Secondly, the environmental covariates are constructed prior to simulation so they capture the expected GEI patterns for all years of breeding. The observed GEI patterns will differ from the expected patterns, but these differences are a natural consequence of the changing population structure over time.

The framework's integration within a simulation package was demonstrated by comparing phenotypic and genomic selection strategies in the presence of no, low, moderate and high GEI. There are three important results:

1. More realistic projections of genetic gain can be obtained for the GEI patterns observed in plant breeding programmes. This provides a framework for breeders to broadly gauge how much genetic gain can be achieved in their TPE relative to what is observed in the MET dataset.
2. Substantial differences in genetic variance between the MET dataset and TPE are observed for phenotypic selection. This is because the MET dataset for phenotypic selection includes just one environment while the training population for genomic selection includes 60 environments.
3. Substantial fluctuations in MET-TPE alignment are observed when a single environment is sampled in the MET dataset. This highlights the importance of constructing MET datasets that include multiple environments spanning multiple years, particularly in the early stages of a breeding programme.

Overall, the framework for simulating GEI provides opportunities to explore many new research objectives through simulation. One important application is the selection for stability, which requires more complex models to be run in real time. In this paper, selection was based on the genotype main effects only since the MET datasets generated each year were too large to run factor analytic models, reflecting an ongoing challenge in many real-world plant breeding programmes.

Concluding remarks

Simulation continues to serve as a valuable tool for breeders to optimize their breeding programmes. The integration of GEI within simulation represents an important advancement for comparing different statistical approaches and breeding strategies in real time. The new framework can be readily implemented in any software that provides the multi-trait functionality.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contributions

DT conceived the methodology. DT and JB developed the methodology and wrote the manuscript. JB implemented the framework into AlphaSimR and conducted all simulations and analyses. GG provided quantitative genetics perspectives and reviewed the manuscript. All authors have read and approved the final manuscript.

Funding

The authors acknowledge funding that contributed to the development of AlphaSimR and associated research: BBSRC (grants BBS/E/D/30002275, BBS/E/RL/230001A and BBS/E/RL/230001C, BB/R019940/1), Bayer CropSciences, BASF, Data-Driven Innovation - Edinburgh and South East Scotland City Region Deal, Marie Skłodowska-Curie Action and The University of Edinburgh.

Authors note

Acknowledgements

The authors thank Colin Cavanagh, Antje Rhodes and Nicole Cocks for their stimulating discussions.

Code availability

The R scripts for this study are available in the Supplementary Material and at the GitHub repository (https://github.com/HighlanderLab/jbancic_GEIsim).

References

Baker RJ (1988) Tests for crossover genotype-environmental interactions. Canadian Journal of Plant Science 68:405–410, URL https://doi.org/10.4141/cjps88-051

Baker RJ (1990) Crossover genotype-environmental interaction in spring wheat. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding, Louisiana State University, Baton Rouge, Louisiana, pp 42–51

Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2017) ASReml-R Reference Manual Version 4. URL https://vsni.co.uk/software/asreml-r, R package version 4.1.0

Comstock RE, Moll RH (1963) Genotype-environment interactions. In: Hanson WD, Robinson HF (eds) Statistical genetics and plant breeding, National Academy of Sciences - National Research Council, Publication 982, Washington, D.C., pp 164–196

Cooper M, DeLacy IH (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theoretical and Applied Genetics 88:561–572, URL https://doi.org/10.1007/BF01240919

Cooper M, DeLacy IH, Eisemann RL (1993) Recent advances in the study of genotype × environment interactions and their application to plant breeding. In: Imrie BC, Hacker JB (eds) Focused plant improvement: towards responsible and sustainable agriculture. Australian plant breeding conference, Gold Coast, Queensland, pp 116–131

Cooper M, Powell O, Gho C, Tang T, Messina C (2023) Extending the breeder's equation to take aim at the target population of environments. Frontiers in Plant Science 14:1129591, URL https://doi.org/10.3389/fpls.2023.1129591

Cornelius PL, Crossa J, Seyedsadr MS (1996) Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: Kang MS, Gauch HG (eds) Genotype-by-environment interaction, CRC Press, Boca Raton, Florida, pp 199–234

Eisemann RL, Cooper M, Woodruff DR (1990) Beyond the analytical methodology – better interpretation of genotype-by-environment interaction. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding, Louisiana State University, Baton Rouge, Louisiana, pp 108–117

Gail M, Simon R (1985) Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41:361–372, URL https://doi.org/10.2307/2530862

Gauch HG (1988) Model selection and validation for yield trials with interaction. Biometrics 44:705–715, URL https://doi.org/10.2307/2531585

Gauch HG, Piepho HP, Annicchiarico P (2008) Statistical analysis of yield trials by AMMI and GGE:

Further considerations. Crop Science 48:866–889, URL https://doi.org/10.2135/cropsci2007.09.0513

Gaynor RC, Gorjanc G, Hickey JM (2021) AlphaSimR: An R package for breeding program simulations. G3: Genes|Genomes|Genetics 11:jkaa017, URL https://doi.org/10.1093/g3journal/jkaa017

Hardin J, Garcia SR, Golan D (2013) A method for generating realistic correlation matrices. Annals of Applied Statistics 7:1733–1762, URL https://doi.org/10.1214/13-AOAS638

Hartung J, Laidig F, Piepho HP (2023) Effects of systematic data reduction on trend estimation from german registration trials. Theoretical and Applied Genetics 136:21, URL https://doi.org/10.1007/s00122-023-04266-5

Kelly AM, Smith AB, Eccleston JA, Cullis BR (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Science 47:1063–1070, URL http://doi.org/10.2135/cropsci2006.08.0540

Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. The Journal of Agricultural Science 103:123–135, URL http://doi.org/10.1017/S0021859600043392

Li X, Song H, Zhang Z, Huang Y, Zhang Q, Ding X (2021) The theory on and software simulating large-scale genomic data for genotype-by-environment interactions. BMC Genomics 22:877, URL https://doi.org/10.1186/s12864-021-08191-z

Lisle C (2023) Information based diagnostics for the optimal construction of multi-environment trial datasets. PhD thesis, University of Wollongong, URL https://ro.uow.edu.au/theses1/1619

Liu H, Tessema BB, Jensen J, Cericola F, Andersen JR, Sørensen AC (2019) ADAM-Plant: A software for stochastic simulations of plant breeding from molecular to phenotypic level and from simple selection to complex speed breeding programs. Frontiers in Plant Science 9:1926, URL https://doi.org/10.3389/fpls.2018.01926

Mandel J (1971) A new analysis of variance model for non-additive data. Technometrics 13:1–18, URL https://doi.org/10.2307/1267072

Piepho HP (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics 53:761–766, URL https://doi.org/10.2307/2533976

Podlich DW, Cooper M (1998) QU-GENE: A simulation platform for quantitative analysis of genetic models. Bioinformatics 14:632–653, URL https://doi.org/10.1093/bioinformatics/14.7.632

Pook T, Schlather M, Simianer H (2020) MoBPS - Modular breeding program simulator. G3: Genes|Genomes|Genetics 10:1915–1918, URL https://doi.org/10.1534/g3.120.401193

Smith AB, Cullis BR, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138–1147, URL https://doi.org/10.1111/j.0006-341X.2001.01138.x

Smith AB, Ganesalingam A, Lisle C, Kadkol G, Hobson K, Cullis BR (2021) Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs. Frontiers in Plant Science 11:1–13, URL https://doi.org/10.3389/fpls.2020.623586

Tolhurst DJ (2023) Genomic prediction models, selection tools and association studies for genotype by environment data. PhD thesis, University of Edinburgh

Tolhurst DJ, Gaynor RC, Gardunia B, Hickey JM, Gorjanc G (2022) Genomic selection using random regressions on known and latent environmental covariates. Theoretical and Applied Genetics 135:3393–3415, URL https://doi.org/10.1007/s00122-022-04186-w

Waters DL, van der Werf JHJ, Robinson H, Hickey LT, Clark SA (2023) Partitioning the forms of genotype-by-environment interaction in the reaction norm analysis of stability. Theoretical and Applied Genetics 136:99, URL https://doi.org/10.1007/s00122-023-04319-9

Werner C, Gemenet D, Tolhurst DJ (2023) FieldSimR: An R package for simulating plot data in multi-environment field trials. Frontiers in Plant Science Under review

Yan W, Hunt LA, Sheng Q, Szlavnics Z (2000) Cultivar evaluation and mega-environment investigation based on the gge biplot. Crop Science 40:597–605, URL https://doi.org/10.2135/cropsci2000.403597x

Younis OG, Turchetta M, Suarez DA, Yates S, Studer B, Athanasiadis IN, Krause A, Buhmann JM, Corinzia L (2023) ChromaX: a fast and scalable breeding program simulator. Bioinformatics 39:btad691, URL https://doi.org/10.1093/bioinformatics/btad691

Table 1: Summary of the simulated between-environment genetic variance matrices with low, moderate and high GEI.

	Input parameters						Variance explained			
GEI	α	β	ρ	ϵ	γ	k	v_g	v_{ge}	v_n	v_c
Low	5	5	0.50	0.50	0.50	7	0.51	0.49	0.61	0.39
Moderate	5	5	0.20	0.80	0.50	7	0.30	0.70	0.36	0.64
High	5	5	0.00	1.00	0.35	7	0.08	0.92	0.09	0.91

Presented are the shape (α) and scale (β) parameters for simulating the genetic variance matrix, $\mathbf{D_e}$, and the baseline correlation (ρ), magnitude of noise (ϵ), skewness (γ) and rank of noise (k) for simulating the between-environment genetic correlation matrix, $\mathbf{C_e}$. Also presented are the proportion of variance explained by the genotype main effect (v_g) and interaction (v_{ge}) variances as well as the non-crossover (v_n) and crossover (v_c) variances.

Table 2: Summary of the simulated MET dataset.

	Design			Trait			
Env	Genos	Reps	Plots	Mean	h_j^2	$\sigma_{g_j}^2$	$\sigma_{\varepsilon_j}^2$
E1	200	2	400	3.11	0.32	2.24	4.71
E2	200	2	400	3.73	0.33	4.60	9.46
E3	200	2	400	5.54	0.32	4.18	8.88
E4	200	2	400	4.04	0.22	3.22	11.40
E5	200	2	400	4.00	0.22	2.15	7.42
E6	200	2	400	5.66	0.30	2.80	6.44
E7	200	2	400	4.39	0.34	0.39	0.76
E8	200	2	400	2.55	0.39	1.59	2.50
E9	200	2	400	3.24	0.49	2.00	2.09
E10	200	2	400	3.70	0.32	5.57	11.97
Overall	-	**-**	-	**4.00**	**0.32**	**2.87**	**6.56**

Presented are the number of genotypes, replicates and plots in each environment. Also presented for a hypothetical continuous trait are the mean, plot-level heritability (h_j^2), genetic variance ($\sigma_{g_j}^2$) and error variance ($\sigma_{\varepsilon_j}^2$).

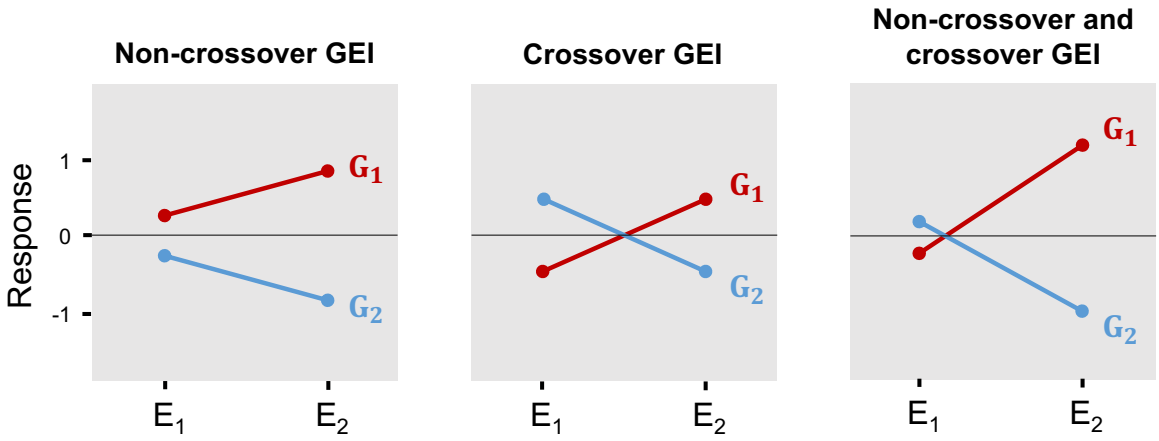

Fig. 1: The response of hypothetical genotypes G1 and G2 in environments E1 and E2 for a hypothetical continuous trait. The figure demonstrates genotype response in terms of non-crossover and crossover GEI, which reflect changes in scale and rank between environments.

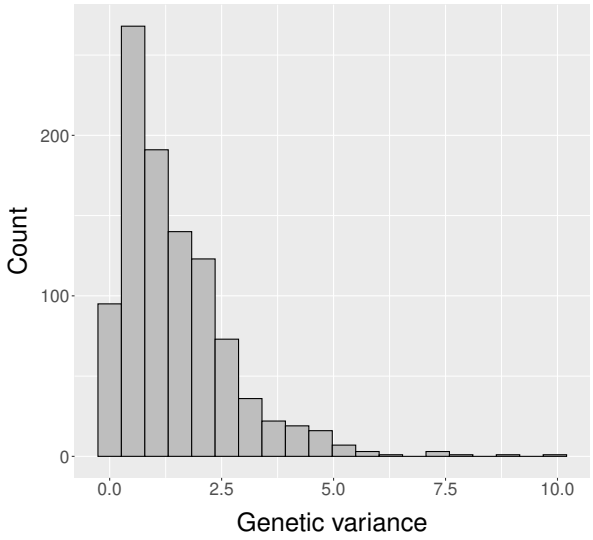

Fig. 2: Simulated genetic variances for 1,000 environments, obtained by sampling from an inverse gamma distribution with shape parameter of 5 and scale parameter of 5.

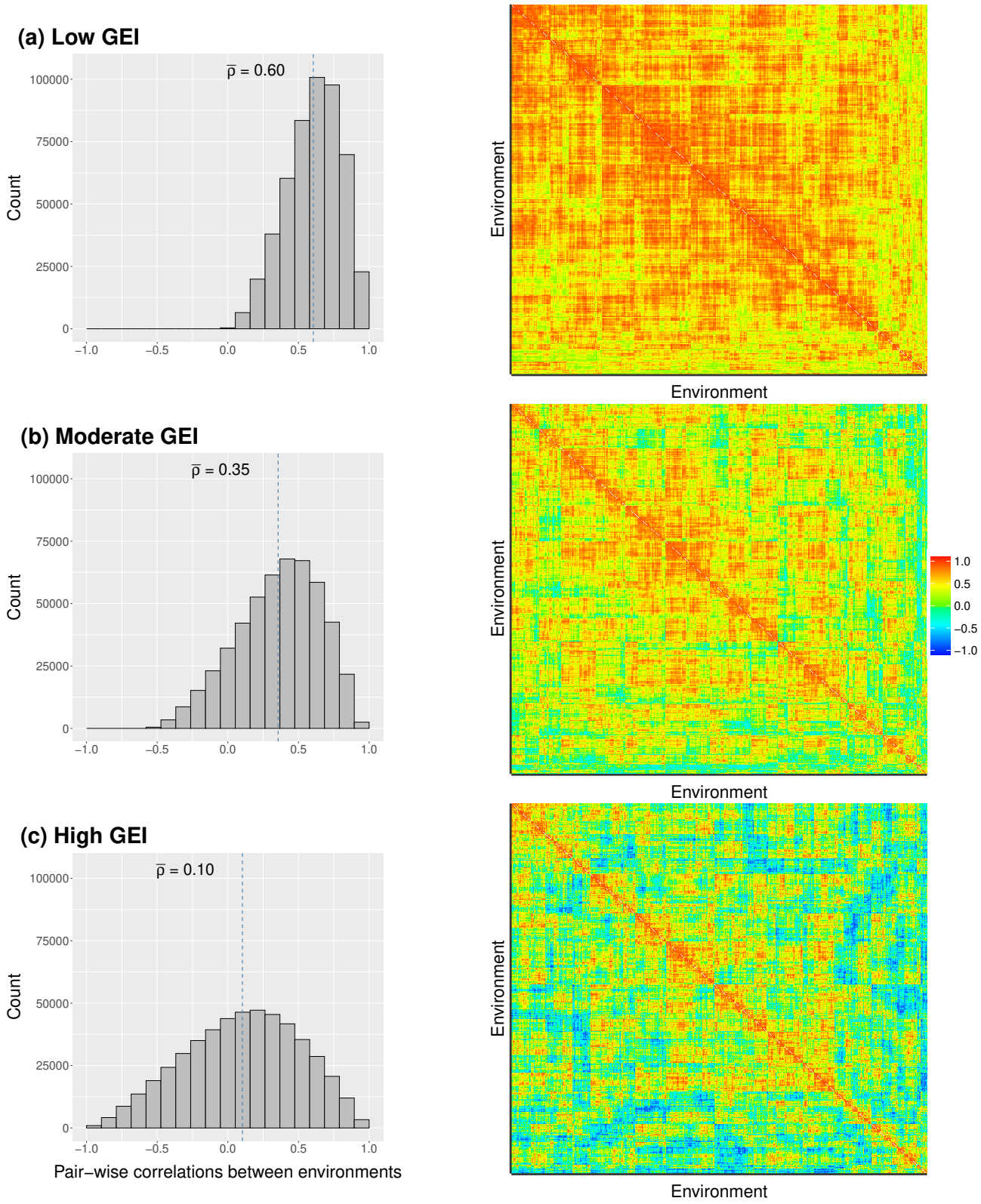

Fig. 3: Simulated between-environment genetic correlation matrices with low, moderate or high GEI. The dashed line in the left panel represents the mean correlation (also labelled). The colour key in the right panel ranges from 1 (agreement in rankings) through 0 (disagreement in rankings) to −1 (reversal of rankings). All matrices are hierarchically ordered based on separate dendrograms.

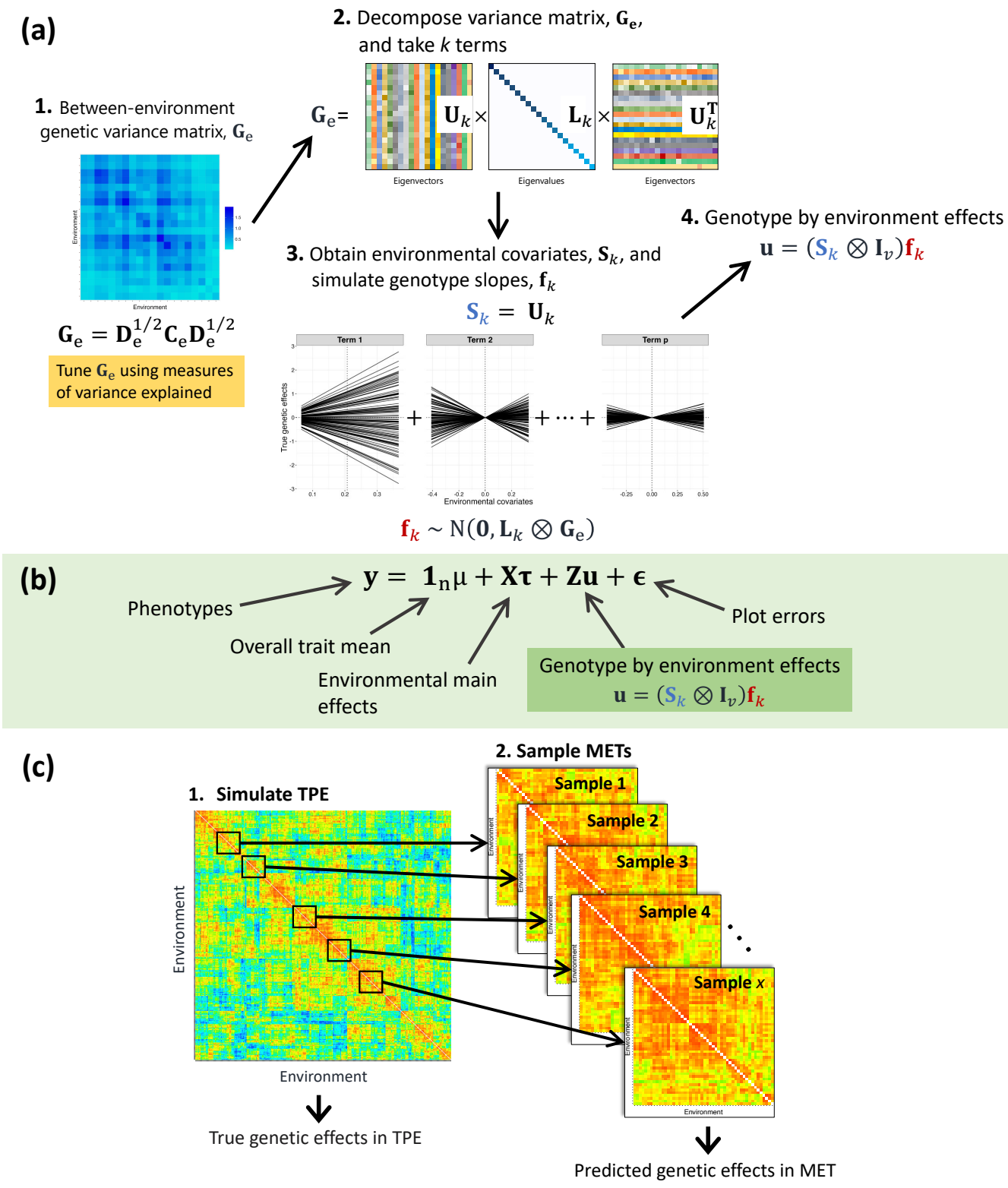

Fig. 4: Overview of the framework for simulating GEI. Presented are **a** the four key steps for simulating the GE effects, **b** how these are embedded within a linear mixed model to generate phenotypes and **c** the process of sampling MET datasets from the simulated TPE.

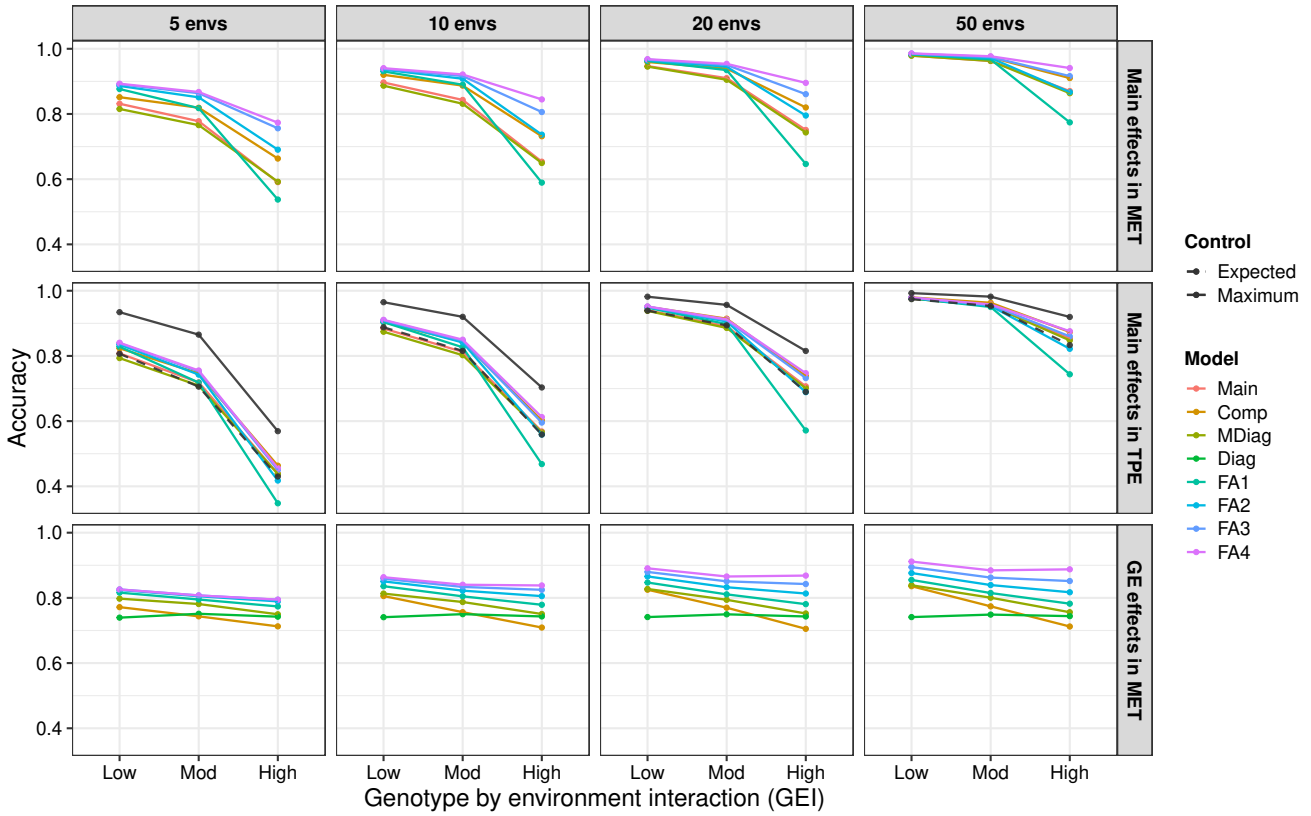

Fig. 5: Prediction accuracy of eight statistical models fitted to simulated MET datasets with low, moderate or high GEI and 5, 10, 20 or 50 environments. The top two panels show the genotype main effect accuracy in the MET dataset and TPE while the bottom panel shows the accuracy of the predicted GE effects in the MET dataset. Note: The main effects in all factor analytic models were obtained as averages across latent covariates. The factor analytic models of order 3 and 4 were fitted without the diagonal term for the 5 environment scenario. The maximum main effect accuracy represents the MET-TPE alignment from Eq. 15. Main - main effects, Comp - compound symmetry, MDiag - main effects plus diagonal, Diag - diagonal, FA - factor analytic.

Year	Stage		Genotypes	Envs	Reps	σ^2_ϵ	Action
1	Crossing	$P_1 \times P_2$	100 crosses				Make crosses
2	F_1/DH		100 FS families				Produce DHs
3	HDRW		100 x 100 DHs	1	1	8	Advance 500 DHs
4	PYT		500 DHs	2	1	4	Yield trial
5	AYT		50 DHs	5	2	2	Yield trial
6	EYT		10 DHs	20	2	2	Yield trial
7	Variety		1 DHs				Release variety

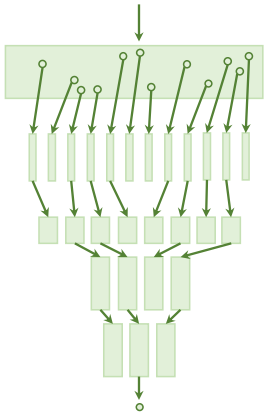

Fig. 6: Key features of the simulated plant breeding programme. Presented are the number of genotypes, environments and replicates as well as the average error variance (σ^2_ε) and the action taken. DH - double haploid, FS - full-sib, HDRW - headrow, PYT - preliminary yield trial, AYT - advanced yield trial, EYT - elite yield trial

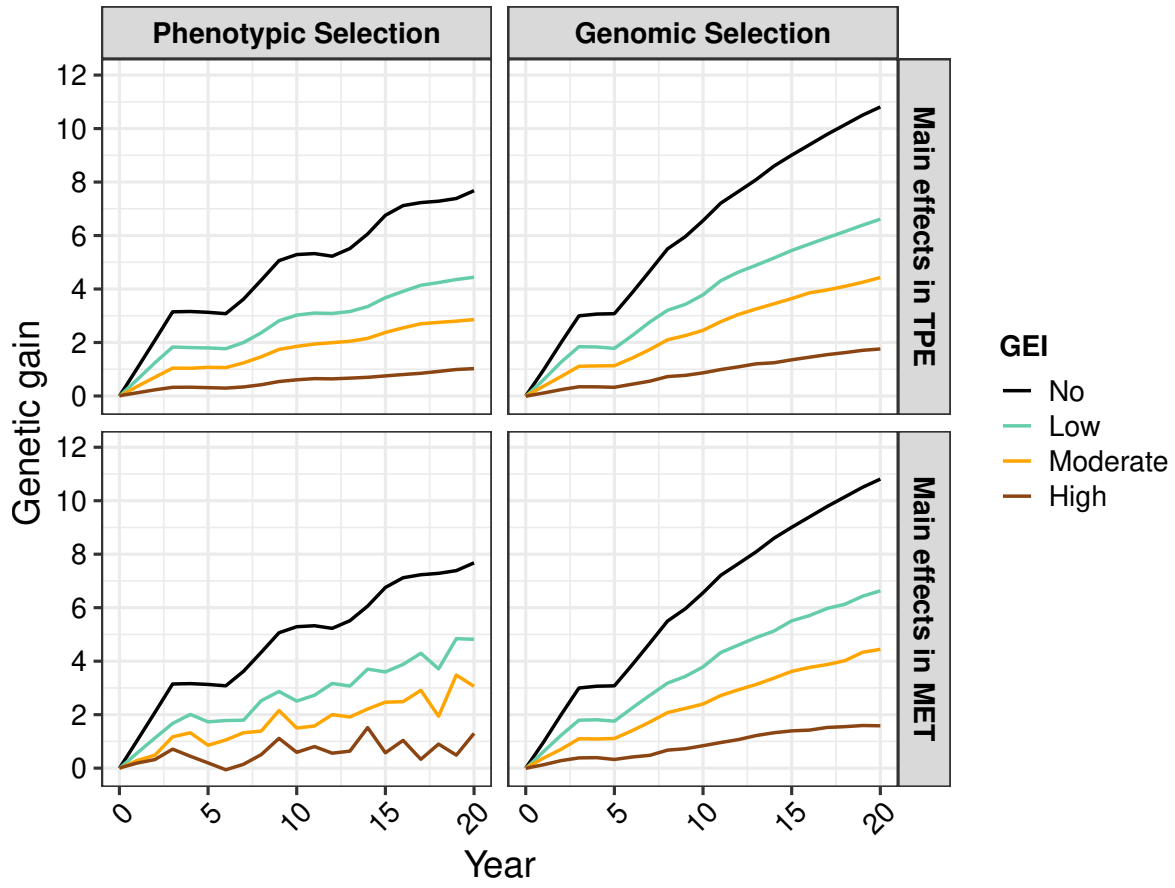

Fig. 7: Genetic gain in the simulated plant breeding programme with no, low, moderate or high GEI. Presented is the genetic gain in the headrow stage for phenotypic and genomic selection strategies. The genetic gain reflects the average genotype main effect in the MET dataset or TPE. Note: The MET dataset is constructed with one environment for phenotypic selection or 60 environments (three years) for genomic selection.

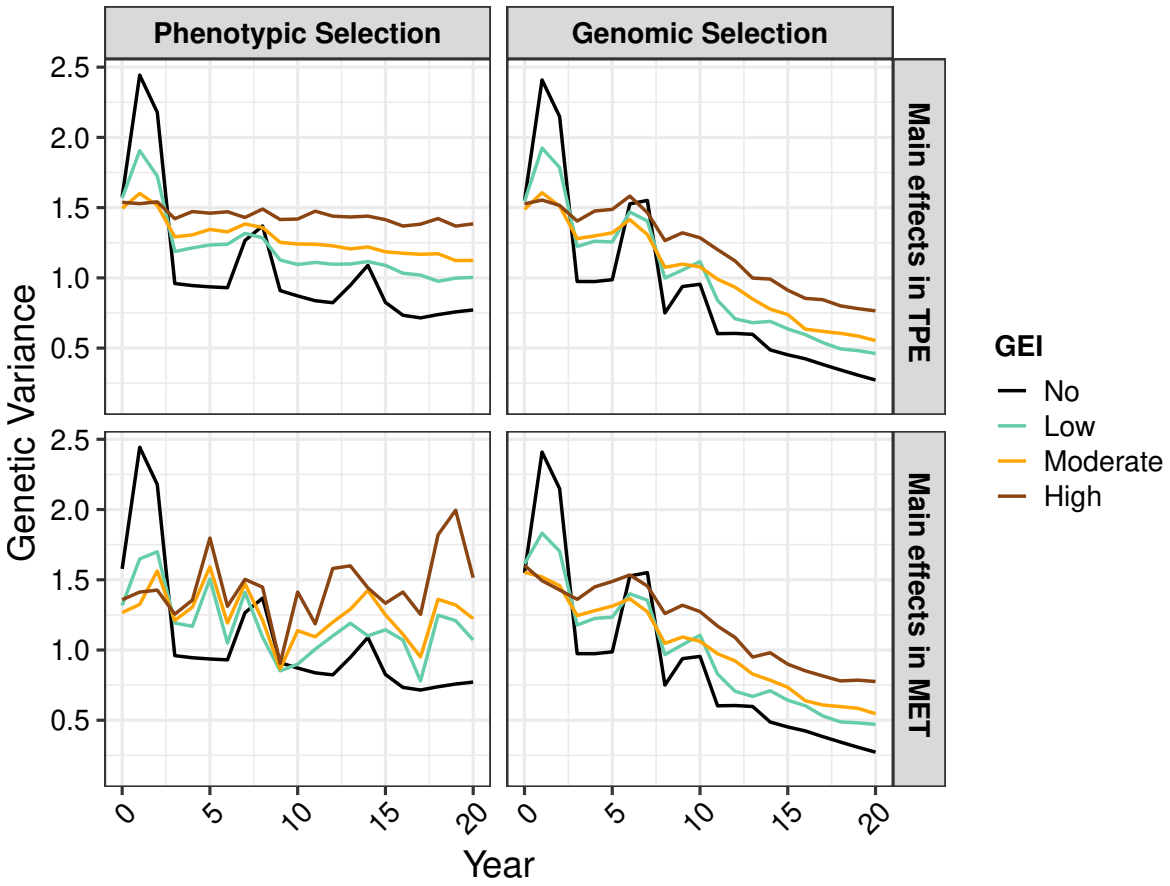

Fig. 8: Genetic variance in the simulated plant breeding programmes with no, low, moderate or high GEI. Presented is the genetic variance in the headrow stage for phenotypic and genomic selection strategies. The genetic variance reflects the variance of the genotype main effects in the MET dataset or TPE. Note: The MET dataset is constructed with one environment for phenotypic selection or 60 environments (three years) for genomic selection.

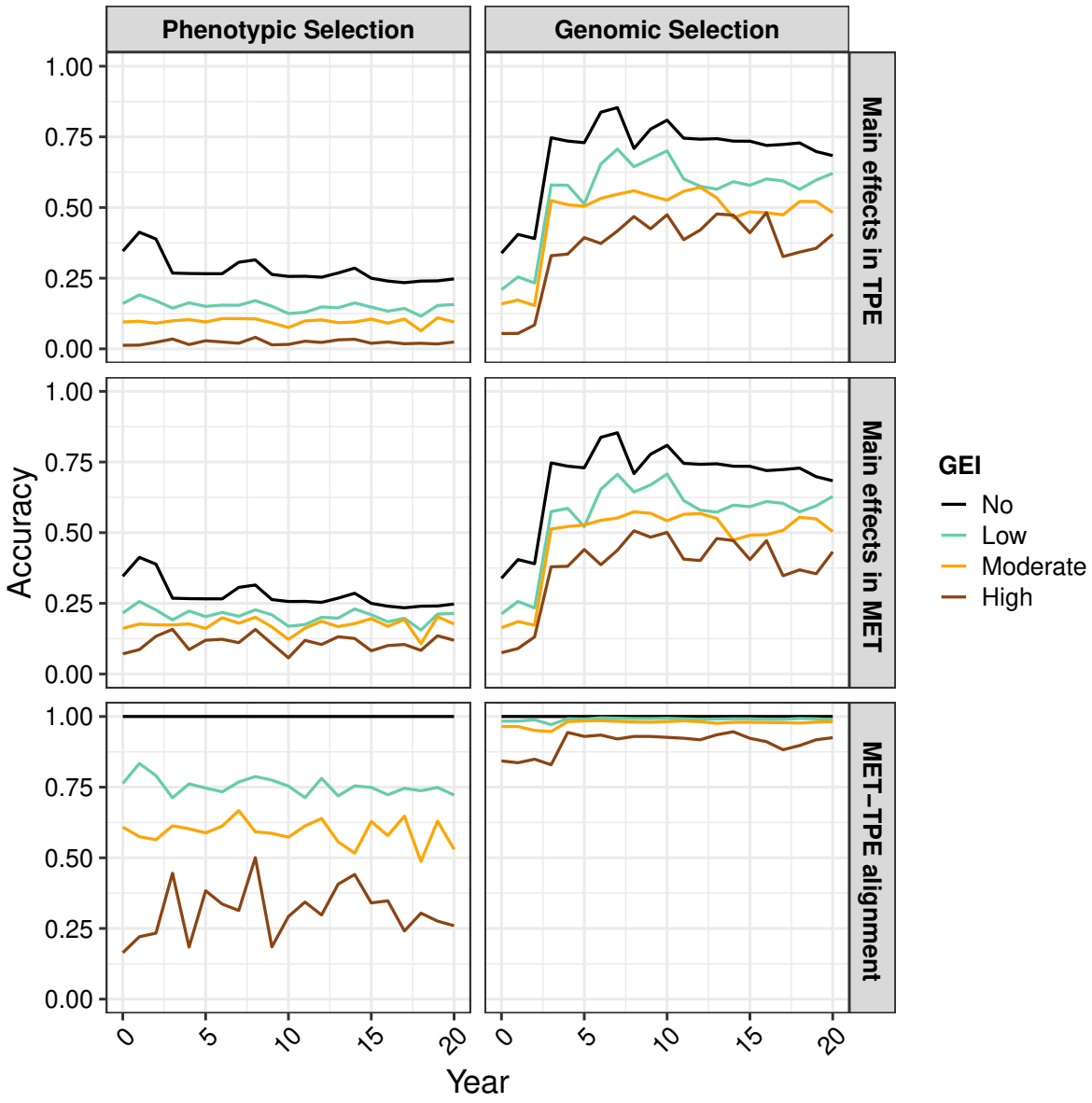

Fig. 9: Measures of accuracy in the simulated plant breeding programmes with no, low, moderate or high GEI. Presented is the prediction accuracy in the headrow stage for phenotypic and genomic selection strategies. The top two panels show the genotype main effect accuracy in the MET dataset and TPE while the bottom panel shows the correlation between the true main effects in the MET dataset and TPE, referred to as the MET-TPE alignment. Note: The MET dataset is constructed with one environment for phenotypic selection or 60 environments (three years) for genomic selection.

Appendix A Simulation features

This appendix demonstrates how the framework can be used to simulate non-crossover and crossover GEI in addition to multiple TPE and/or traits.

A.1 Simulating non-crossover and crossover GEI

An important feature of the new framework is the ability to simulate and tune the magnitude of non-crossover and crossover GEI. In this paper, non-crossover and crossover GEI are defined as the change in scale and rank of genotype response between environments, respectively. The non-crossover variance therefore captures all variation attributed to perfect positive correlation with the genotype main effects while the crossover variance captures all remaining variation independent of the main effects (Tolhurst, 2023).

The reduced rank model in Eq. 4 does not have explicit main effects, which instead arise naturally from the form of $\mathbf{G_e}$ simulated in Eq. 6. Main effects and interaction effects can be obtained after simulation as:

$$\mathbf{u_g} = \left(\bar{\mathbf{s}}_k \otimes \mathbf{I}_v\right)\mathbf{f}_k \quad \text{and} \quad \mathbf{u_{ge}} = \left(\mathbf{S}^*_k \otimes \mathbf{I}_v\right)\mathbf{f}_k, \tag{19}$$

where $\bar{\mathbf{s}}_k = \mathbf{1}_p^\top \mathbf{S}_k/p$ is a k row-vector of means for each environmental covariate and $\mathbf{S}^*_k = \mathbf{S}_k - \bar{\mathbf{s}}_k \otimes \mathbf{1}_p$ is a $p \times k$ matrix of column centred environmental covariates.

It then follows that:

$$\begin{bmatrix} \mathbf{u_g} \\ \mathbf{u_{ge}} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \bar{\mathbf{s}}_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top & \bar{\mathbf{s}}_k \mathbf{L}_k \mathbf{S}_k^{*\top} \\ \mathbf{S}^*_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top & \mathbf{S}^*_k \mathbf{L}_k \mathbf{S}_k^{*\top} \end{bmatrix} \otimes \mathbf{G} \right), \tag{20}$$

where $\sigma^2_g = \bar{\mathbf{s}}_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top$ is the main effect variance, which is equivalent to the mean element of $\mathbf{G_e}$ given by:

$$\sigma^2_g = \sum_{j=1}^{p} \sigma^2_{g_j}/p^2 + 2\sum_{i<j}^{p} \sigma_{g_{ij}}/p^2. \tag{21}$$

The pooled interaction variance is then obtained as the mean diagonal element of $\mathbf{S}^*_k \mathbf{L}_k \mathbf{S}_k^{*\top}$, which is equivalent to the mean diagonal element of $\mathbf{G_e} - \sigma^2_g \mathbf{J}_p$ given by:

$$\sigma^2_{ge} = \sum_{j=1}^{p} \sigma^2_{g_j}/p - \sigma^2_g. \tag{22}$$

Following Tolhurst (2023), generalised main effects and adjusted interaction effects can also be obtained after simulation as:

$$\mathbf{u}^*_\mathbf{g} = \mathbf{1}_p \otimes \mathbf{u_g} + \boldsymbol{\delta}_\mathbf{ge} \quad \text{and} \quad \mathbf{u}^*_\mathbf{ge} = \mathbf{u_{ge}} - \boldsymbol{\delta}_\mathbf{ge}, \tag{23}$$

where $\boldsymbol{\delta}_\mathbf{ge} = \mathbf{S}^*_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top \otimes \mathbf{u_g}/\sigma^2_g$ is a vp-vector which captures variation in the interaction effects attributed to perfect positive correlation with the main effects (also see Waters et al., 2023). The generalised main effects exclusively capture non-crossover GEI while the interaction effects adjusted for the generalised main effects exclusively capture crossover GEI. The non-crossover variance is therefore obtained as $\sigma^2_n = \sigma^2_g + \bar{\mathbf{s}}_k \mathbf{L}_k \mathbf{S}_k^{*\top} \mathbf{S}^*_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top/p\sigma^2_g$ while the crossover variance is obtained as $\sigma^2_c = \sigma^2_{ge} - \bar{\mathbf{s}}_k \mathbf{L}_k \mathbf{S}_k^{*\top} \mathbf{S}^*_k \mathbf{L}_k \bar{\mathbf{s}}_k^\top/p\sigma^2_g$. Note that constraints may be required to ensure that only variation attributed to non-crossover GEI is transferred from the interaction effects to the generalised main effects (see Tolhurst, 2023).

A.2 Simulating multiple TPE

Another important feature of the new framework is the ability to simulate multiple TPE simultaneously. Assume p_{m_i} environments are sampled in each TPE, such that $p = \sum_{i=1}^{s} p_{m_i}$ is the total number of environments sampled across all s TPE. The between-environment genetic correlation matrix in Eq. 7 is now simulated as:

$$\mathbf{C_e} = \rho \mathbf{J}_p + \oplus_{i=1}^{s} \delta_i \mathbf{J}_{p_{m_i}} + \epsilon \boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}, \tag{24}$$

where ρ is the baseline genetic correlation across all TPE, δ_i is the deviation from the baseline for the i^{th} TPE and all other parameters are previously defined. This specification constructs a separate baseline genetic correlation for each TPE given by $\rho_i = \rho + \delta_i$, but the same genetic correlation between TPE given by ρ. The parameters are subject to similar constraints as before, with $\epsilon \leq 1 - p_{max}$, $0 \leq \rho < p_{min}$ and $0 \leq \rho_i < 1$ where $p_{max} = \max\{p_{m_i}\}$ and $p_{min} = \min\{p_{m_i}\}$. The between-environment genetic variance matrix is then constructed using Eq. 6, but note that separate distributions can be used for the genetic variances in each TPE where required. The approach above is analogous to simulating multiple phenotypic traits.

Appendix B R code to simulate an example MET dataset

```
#> Initial parameters
set.seed(123) # do not change
p  = 10   # Environments
b  = 2    # Blocks
v  = 200  # Genotypes
mu = 4    # Trait mean
H2 = 0.3  # Plot-level heritability
k  = 7    # No. of multiplicative terms

#> Simulate environmental effects
X   = kronecker(diag(p), rep(1, v * b))
tau = rnorm(p, 0, 1)

#> Simulate GE effects
# 1. Simulate Ge
Ce = matrix(0, p, p)
Ce[upper.tri(Ce,F)] = runif(p * (p - 1)/2, 0.4, 1)
Ce = Ce + t(Ce); diag(Ce) = 1
De = diag(1/rgamma(p, shape = 5, rate = 5))
Ge = sqrt(De) %*% Ce %*% sqrt(De)

# 2. Decompose Ge and take first k terms
U = svd(Ge)$u[,1:k]
L = diag(svd(Ge)$d[1:k])

# 3. Obtain covariates and slopes
S = U
slopes = scale(matrix(rnorm(k * v), ncol = k))
slopes = c(matrix(slopes, ncol = k) %*% sqrt(L))

# 4. Construct GE effects
Z = kronecker(diag(v * p), rep(1, b))
u = kronecker(S, diag(v)) %*% slopes

# Obtain genotype main effects
ug = rowMeans(matrix(u, ncol = p))

#> Simulate plot errors
H2 = abs(rnorm(p, H2, 0.1))
H2[H2 < 0] = 0; H2[H2 > 1] = 1
R  = diag(diag(De) / H2 - diag(De))
e  = c(matrix(rnorm(p*b*v), ncol = p) %*% sqrt(R))

#> Create phenotypes
y = mu + X %*% tau + Z %*% u + e

# Construct MET dataset
df.MET = data.frame(
  env = factor(rep(1:p, each = v * b)),
  rep = factor(1:b),
  id  = factor(rep(1:v, each = b)),
  y = y,
  u = rep(u, each = b),
  e = e
  )
```

```
# Order MET and randomise by trial
df.MET = df.MET[order(df.MET$env, df.MET$rep),]
for(i in 0:(p * b - 1)){
  df.MET[i*v+1:v,] = df.MET[i*v+1:v,][sample(1:v,v,F),]
}

#> Run model
asr = asreml(y ~ 1 + env,
             random     = ~ rr(env, 4):id +
               diag(env):id,
             residual   = ~ dsum(~units|env),
             na.action = na.method("include"),
             data       = df.MET)
```

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupFiguresGEIframework.pdf