Classification and regression trees clarify the role of epistasis and environment in genotype–phenotype maps

Sudam Surasinghe[1,2, †], Swathi Nachiar Manivannan[1,†], Lorin Crawford[3], and C. Brandon Ogbunugafor[1,2,4 ✉]

[1] Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT, 06511, USA
[2] Public Health Modeling Unit, Yale School of Public Health, New Haven, CT, 06511 USA
[3] Microsoft Research, Cambridge, MA, 02142, USA
[4] Santa Fe Institute, Santa Fe, NM, 87501, USA
[†] These authors contributed equally to this work.

✉CBO: brandon.ogbunu@yale.edu

Abstract | Understanding how genetic variation translates into phenotypic outcomes is central to various subfields of genetics. This task is complicated by a range of forces–including epistasis, environmental modulation of mutation effects, and ecological influences–that complicate the process of mapping from genotype to phenotype. In this study, we apply a unified decision tree approach, classification and regression trees (CART), to model genotype-phenotype relationships across protein fitness landscapes across a diversity of organisms: (i) a fluorescent protein isolated from *Entacmaea quadricolor* (bubble-tip anemone), (ii) antifolate resistance in *Plasmodium falciparum* (malaria parasite) dihydrofolate reductase (DHFR) under drug concentration gradients, (iii) allelic variants from the long-term evolution experiment (LTEE) in *Escherichia coli*, (iv) proteostasis-modulated drug resistance phenotypes in three bacterial orthologues of DHFR, and (v) chemotypic diversification of sesquiterpene synthases in *Nicotiana tabacum* (cultivated tobacco). Our results demonstrate that decision trees can effectively capture higher-order interactions between mutations and environments, uncovering nonlinear dependencies and contingencies that are often missed by traditional parametric models. By enabling clear visualization of interaction hierarchies, CART serves as both a predictive tool and an explanatory framework for genotype-phenotype mapping. This approach has use cases across the spectrum, from resolving the genomic architecture of biological traits, to personalized medicine, and varied applications in bioengineering.

Summary

How do genes and environments interact to shape traits like drug resistance, protein function, or plant chemistry? This study introduces a simple but powerful tool—decision trees—to help answer that question. Decision trees, also called classification and regression trees (CART), work by repeatedly splitting data into "if–then" rules that are easy to visualize. We applied this approach to five diverse biological systems: fluorescent proteins in sea anemones, malaria parasites under drug pressure, evolving bacteria, enzyme stability in different species, and chemical diversity in tobacco plants. In each case, decision trees uncovered patterns that traditional models often miss—such as when combinations of mutations only matter under certain environments, or when multiple genetic changes must act together to produce an observable effect. Because the method is interpretable, it doesn't just predict outcomes, but explains them, showing which genetic and environmental factors matter most, and under what conditions. This makes decision trees useful for both basic and applied scenarios in medicine, agriculture, and biotechnology, where understanding complex genetic interactions is key to predicting and controlling biological outcomes.

1 Introduction

The role of genetic variation and environmental context in shaping phenotypic traits is at the very center of many sub-fields of genetics (1,2). These questions manifest at all levels of biological complexity: from the study of gene expression (3) to protein function (4), the study of fitness landscapes (5), and genome-wide analyses of genotype-phenotype maps in complex organisms (6). Specifically, questions about the appropriate methods for disentangling the many forces that shape genotype-phenotype maps permeate all corners of population, evolutionary, and quantitative genetics. Although additive models often serve as a first approximation for trait variation across genotype-phenotype maps, many traits arise from non-linear interactions between genes or mutations (epistasis) (G×G) (7,8) as well as gene-by-environment interactions (G×E), and gene-by-gene-by environment interactions (G×G×E) (9–13) . These effects are central to phenomena such as cryptic variation (14,15), missing or phantom heritability (16), and evolutionary contingency (17,18). Yet, standard models often fail to capture these influences because they often assume linearity or require interaction terms to be explicitly specified in advance (19,20) (see Table 1 for key definitions).

Traditional regression methods such as ordinary least squares (OLS) assume a fixed parametric relationship between predictors and outcomes with coefficients estimated to minimize prediction error (21, Section 8.3.3). While statistically convenient, these models capture non-additive or higher-order effects only when such terms are explicitly specified. Variable-selection heuristics such as stepwise regression or the least absolute shrinkage and selection operator (LASSO) can aid in large predictor spaces, but they typically approach interaction discovery as a heuristic process and may fail to identify complex conditional structures (19). Nonparametric methods (e.g., kernel regression, nearest neighbors) relax functional assumptions but frequently trade interpretability and stability for flexibility, especially in high dimensions (22). Likewise, high-dimensional machine-learning models (neural networks) can achieve strong predictive performance in genotype-phenotype tasks but are often black boxes, limiting biological insight and hypothesis generation (23–29). Theoretical approaches that expand epistatic effects as series over locus combinations provide useful formalisms (e.g., Balvert et al. (2024) (30)) but become computationally onerous as loci increase and typically assume static environments, overlooking the conditional (environment-dependent) nature of many genetic effects (30–33). This tension between flexibility and interpretability is central in modern quantitative genetics.

Classification and regression trees (CART) provide a pragmatic and interpretable approach for genotype-phenotype mapping (34). By recursively partitioning genotypes, environments and organismal backgrounds into regions of similar outcome, CART uncovers thresholds, nonlinearities and higher order interactions. Decision trees (25,35–38) and tree ensemble methods, such as random forests (39–50), have been widely used in genetic studies, particularly for analyses of singlenucleotide polymorphisms (SNPs). Recent studies of protein mutational landscapes have also shown that tree methods can recover meaningful epistatic patterns while remaining experimentally tractable (51,52). Furthermore, on rugged fitness landscapes, decision tree structures often capture local interaction dynamics more effectively than pairwise approximations (53,54). However, these implementations of decision trees have been limited in their ability in breaking down continuous traits, capturing G×E complexity and analyzing sparse datasets (55,56).

Here, we implement a unified CART framework across a heterogeneous collection of genotype-phenotype datasets, constituting 5 case studies (see Table 3): fluorescent protein libraries in *Entamacea quadricolor* (bubble-tip anemone; Case 1) (57), drug-resistance evolution in *Plasmodium falciparum* (malaria parasite; Case 2) (12), differential proteostasis backgrounds in *Escherichia coli* (Case 3) (58), long-term evolution experiment of *E. coli* under chemical stress (Case 4) (11,59), and terpene synthase chemotype variation in *Nicotiana tabacum* (cultivated tobacco; Case 5) (60). These datasets vary in structure and complexity; by applying regression trees consistently, we aim not only to predict but to reveal nonlinear, high-order, and/or context-dependent interactions in an interpretable manner that facilitates mechanistic hypotheses and experimental validation.

2 Results

To examine structural—potentially non-linear and high-order—relationships between traits and their genetic and environmental determinants, we applied the classification and regression trees (CART) algorithm to several empirical datasets. These include mutational scans of fluorescent proteins (57), genotype-by-environment fitness data in microbes (11,12,58,59), and enzyme activity profiles of plant sesquiterpene synthases (STS) (60). Each serves as a case study that combines genotypic variation (typically binary mutation indicators) with environmental covariates (e.g., drug gradients, media composition). Response variables span continuous measures (fluorescence, fitness, metabolite abundance) and categorical traits (color class). A full summary is provided in Section 4.1 and Table 3.

Our goal is not predictive optimization, but rather to illustrate how CART reveals interpretable, rule-based structures that capture interactions and non-linear dependencies among genotype, environment, and phenotype. Details of the method and implementation appear in Sections S4.2.3 and 4.2.

Table 1. Key terminology and definitions.

Term	Definition
Epistasis (G×G)	The "surprise at the phenotype when mutations are combined, given the constituent mutations' individual effects" (*7*), i.e., when the combined effect of two (or more) mutations or gene variants on a phenotypic trait deviates from the expected sum or product of individual mutational effects (*61*).
Genotype-by-environment or gene by environment interactions (G×E)	When the phenotypic effect of one genotype across different environments varies from the phenotypic effect of another genotype across different environments, i.e., non-parallel reaction norms among individuals with different genotypes in response to different environmental conditions. (*62*,*63*)
Environmental epistasis (also referred to as G×G×E)	Non-additive interactions among alleles whose combined phenotypic effect depends on environmental context; the magnitude or sign of epistasis changes across environments (*10*).
Reaction norm or norm of reaction	A depiction used in ecology and quantitative genetics that depicts how a phenotype for a given genotype is shaped by aspects of the environment. Crossing reaction norms between two genotypes are often diagnostic of G×E (*64*–*66*).
Fitness or adaptive landscape	A genotype-to-fitness map, analogized to a physical surface defined by fitness "peaks" and "valleys." They are often used a framework for understanding evolutionary dynamics in biological systems. (*67*–*69*)
Classification tree (CART)	Tree-based model that recursively partitions predictor space to predict a *categorical* outcome; splits are chosen to increase class purity (e.g., minimize Gini impurity or entropy), and each terminal node predicts the majority class (*34*).
Regression tree (CART)	Tree-based model that recursively partitions predictor space to predict a *continuous* outcome; splits are chosen to reduce within-node variance (minimize sum-of-squared errors), and each terminal node predicts the mean response of observations in that leaf (*34*).

2.1 Case 1: Classification tree analysis of the fluorescent protein fitness landscape in *Entacmaea quadricolor*

The dataset from Poelwijk et al. (2019) (*57*) comprises all $2^{13} = 8192$ genotypes generated by introducing amino acid substitutions at 13 variable loci that distinguish two parental *Entacmaea quadricolor*-derived fluorescent proteins: mTagBFP2 (blue) and mKate2 (red) (Figure 1). Each genotype is encoded as a 13-bit binary string, from C-terminal (locus 1: K231R) to N-terminal (locus 13: D20N), representing specific amino acid changes on the mTagBFP2 backbone. These genotypes were expressed in *E. coli* and evaluated using two-color flow cytometry, generating red and blue emission intensities. Following normalization and nonlinear transformation, Poelwijk et al. (2019) derived a continuous scalar fluorescence phenotype. For our analysis, however, we use a discretized classification scheme, assigning genotypes (see Figure 1) to one of three color classes—*red*, *blue*, or *black*—based on emission dominance and intensity thresholds (as described in (*57*, Section S4.2.3)).

Applying the CART algorithm (details of which are provided in Section 4.2) to this dataset reveals a compact decision tree that accurately predicts fluorescence class and isolates key loci responsible for color transitions (see Figure 2). Among the 13 positions, loci 9 (F143), 4 (Y197), 11 (L63), 5 (A174), and 12 (V45) emerge as critical decision nodes (see Figure 2b). The absence of mutations at both F143 and Y197 consistently prevents the appearance of red fluorescence, confirming their role as essential gatekeepers. Crucially, the tree indicates that at least three specific mutations are required to induce red fluorescence—most commonly including residues 4, 9, and 11—demonstrating the importance of higher-order epistasis (see Figure 2c). Genotypes lacking these coordinated changes predominantly exhibit blue or ambiguous (black) phenotypes. The hierarchical structure of the tree reflects non-additive mutational interactions, distinguishing root-level determinants from context-dependent modulators. These findings are consistent with patterns observed in the mutational interaction network (see Figure S2), which highlights that single mutations rarely suffice to produce red fluorescence and that higher-order interactions are critical. For a detailed interpretation of the CART decision tree structure, probability distributions at internal nodes, and a comprehensive epistatic interaction analysis, we refer readers to the Supporting Information (Section S4.2.3).

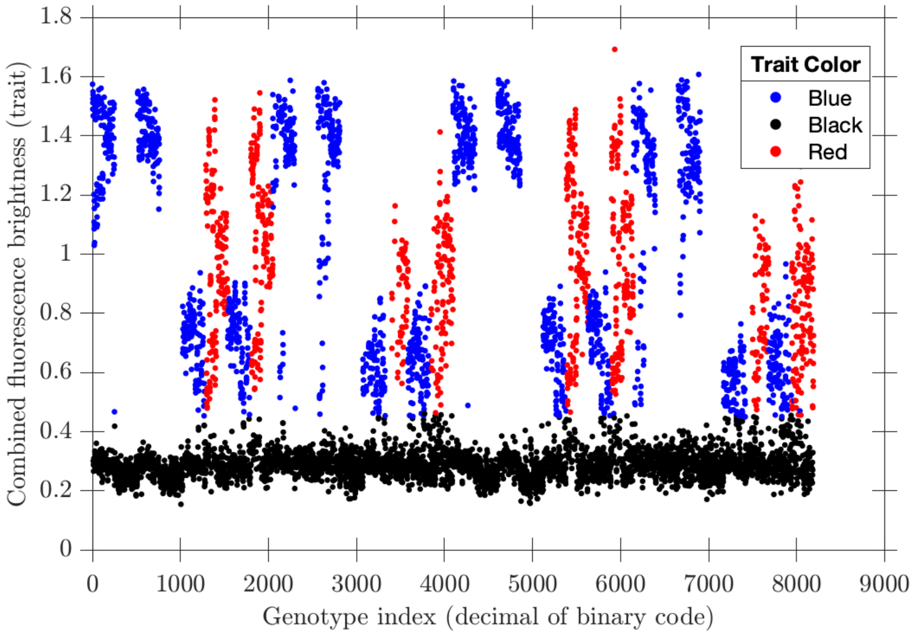

Fig. 1. Data for Case 1: Genotype-fluorescence map of the *Entacmaea quadricolor*-derived fluorescent protein dataset from Poelwijk et al. (2019) (*57*). The dataset includes all $2^{13} = 8192$ genotypes produced by combinatorially introducing amino acid substitutions at 13 variable positions distinguishing two fluorescent proteins derived from *Entacmaea quadricolor*, mTagBFP2 (blue-emitting) and mKate2 (red-emitting). Each genotype is encoded as a 13-bit binary string, with bits indicating the presence (1) or absence (0) of a substitution relative to the mTagBFP2 backbone, ordered from the C-terminal (bit 1: K231R) to the N-terminal (bit 13: D20N). Genotypes were expressed in *E. coli*, and red and blue fluorescence intensities were measured using two-color flow cytometry. These intensity values were normalized and nonlinearly transformed to obtain a scalar brightness phenotype. For visualization, genotypes are indexed on the horizontal axis by the decimal equivalent of their binary code, and their combined fluorescence brightness is shown on the vertical axis. Genotypes are mapped to one of three color classes—*blue*, *red*, or *black*—based on relative red and blue emission levels and total brightness: *blue* denotes high blue and low red emission; *red* denotes high red and low blue; and *black* indicates low overall brightness.

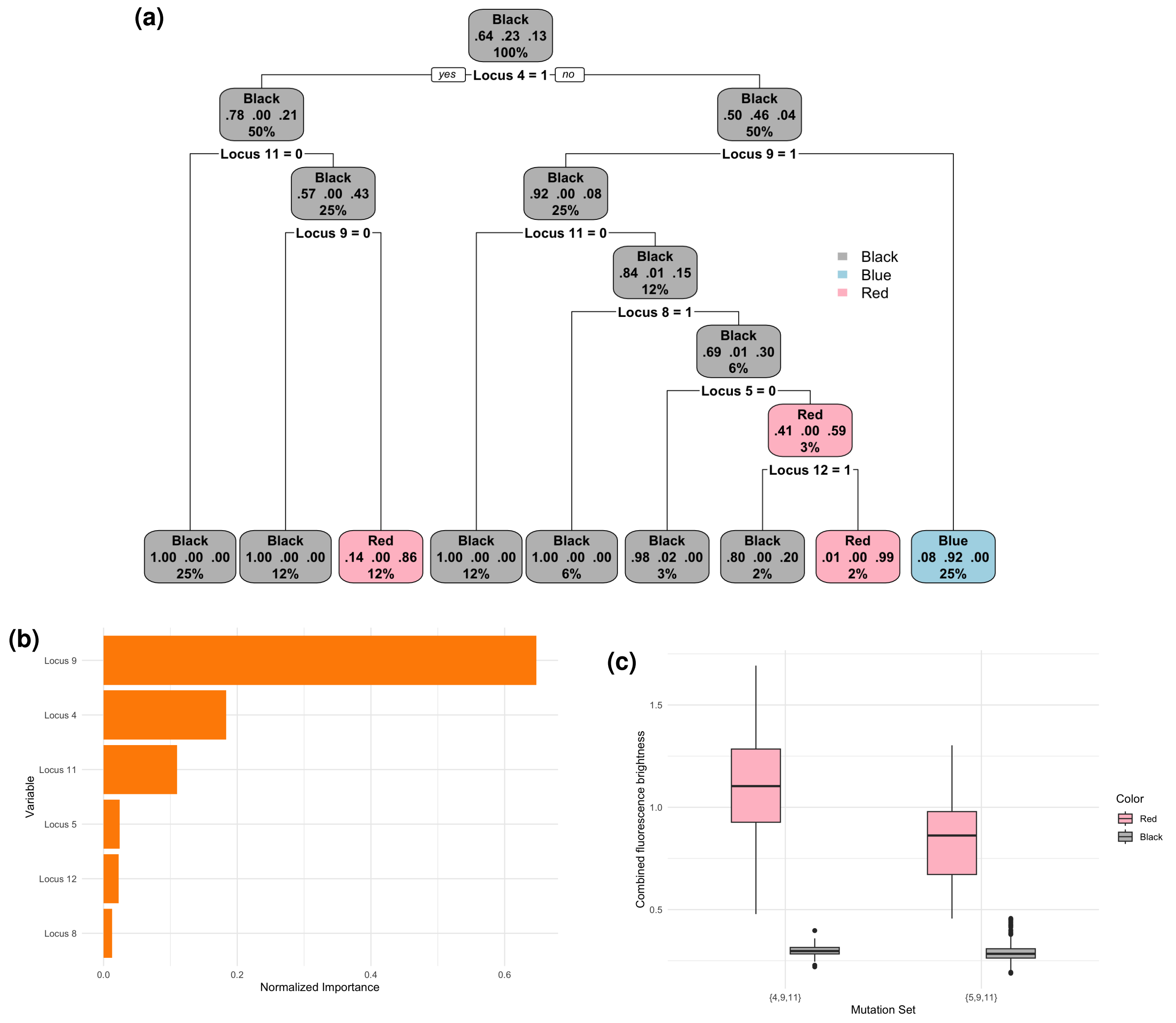

Fig. 2. Case 1 analysis: Resolving G x G interactions (epistasis) in *Entacmaea quadricolor*. Overview of the genotype-to-color classification framework based on the dataset from Poelwijk et al. (2019) (*57*), comprising all $2^{13} = 8192$ genotypes generated via amino acid substitutions at 13 loci distinguishing the parental fluorescent proteins mTagBFP2 and mKate2. Each genotype, encoded as a binary vector, is associated with dual-channel fluorescence intensities measured via two-color flow cytometry and normalized to parental brightness levels. Genotypes are classified into three discrete color categories (*black*, *blue*, or *red*) using polynomial-based thresholding of normalized emission intensities. **(a)** Decision tree (CART) predicting color class from genotype. The proportions represent the phenotypic class probabilities (from left to right: black, blue, red) and the percentages represent the proportion of observations (i.e. genotypes) passing through each node. **(b)** Variable importance scores indicating locus-level contributions. **(c)** Box plots of trait values for red and black genotypes within each mutational combination. Each x-axis group ({4, 9, 11} and {5, 9, 11}) contains two box plots: one for red (colored red) and one for black (colored black) genotypes. Genotypes with mutations at {4, 9, 11} exhibit significantly higher trait values (combined fluorescent brightness) compared to genotypes with mutations at {5, 9, 11}, validating the stronger contribution of locus 4 to the red fluorescence phenotype.

2.2 Regression tree analysis of environmental modulation of mutation effects (Cases 2 and 3)

While the fluorescent protein dataset (*57*) enabled the analysis of higher-order epistatic interactions within a fixed biochemical context (see Section 2.1), many biological systems exhibit phenotypic variability shaped by interactions between genetic background and environmental conditions. The sesquiterpene synthase dataset from O'Maille et al. (2008) (*60*) similarly demonstrates how combinatorial mutations in the enzyme's active site influence product specificity, reflected in the relative production of compounds such as 5-epi-aristolochene, 4-epi-eremophilene, and premnaspirodiene (see Section S4.2.3). Although this variation arises from internal structural constraints rather than external stimuli, it illustrates how context—whether environmental or molecular—modulates mutational effects.

In this section, we examine systems where environmental conditions are explicitly varied, enabling the investigation of genotype-by-environment (G×E) interactions. We apply classification and regression trees (CART) to two genotype-environment-phenotype datasets (see Section 4.1), each integrating mutational profiles with environmental gradients. The first dataset (Case 2) examines antifolate drug resistance in the *Plasmodium falciparum* dihydrofolate reductase (DHFR) gene (*12*), focusing on mutations at four key loci and their interactions with varying concentrations of pyrimethamine and cycloguanil. The second dataset (Case 3) originates from the Long-Term Evolution Experiment (LTEE), which tracks adaptive mutations in *Escherichia coli* under minimal glucose conditions and chemically perturbed environments (*11*). An additional analysis of genotype-by-environment and genotype-by-species interactions under altered proteostasis conditions—derived from mutational variation in the DHFR gene across three bacterial species and proteostatic backgrounds (*58*)—is presented in the Supporting Information (see Section S4.2.3), with corresponding decision tree results also summarized in Table 2.

Notably, the environmental variable in the malaria DHFR (Case 2) dataset is measured continuously (drug concentration), whereas the LTEE and proteostasis (Case 3 and 4, respectively) datasets treat environment/species as categorical factors. CART regression trees naturally accommodate both numeric and categorical predictors, using threshold-based splits for continuous variables and level-based splits for categorical variables.

In Cases 2-4, continuous phenotypic traits serve as response variables, enabling the use of CART regression trees to dissect the hierarchical contribution of genetic and environmental factors. This integrative approach allows us to identify critical loci and environmental conditions that drive quantitative trait variation and to unravel the structure of genotype-by-environment interactions across diverse biological systems. A summary of CART-derived decision tree structures, including environmental thresholds and interaction patterns, is provided in Table 2.

2.2.1 Case 2: Epistasis and drug-environment interactions in the *P. falciparum* dihydrofolate reductase

The dataset (see Figure 3 for a visualization) from Ogbunugafor (2022) (*12*) investigates environmentally modulated epistatic interactions in the *dihydrofolate reductase* (DHFR) gene of *Plasmodium falciparum*, a protozoan parasite and the primary causative agent of malaria. DHFR is an essential enzyme in folate metabolism and DNA synthesis, and is the molecular target of antifolate drugs such as pyrimethamine and cycloguanil. In the original study, both drugs were profiled, but here, we focus on cycloguanil as a representative example of drug-mediated environmental stress. The study considers all $2^4 = 16$ genotypes generated by combinatorial substitution at four key amino acid positions in DHFR:

$$(\text{N51I},\ \text{C59R},\ \text{S108N},\ \text{I164L}) \in \{0,1\}^4,$$

where each mutation is encoded as a binary variable (1 = derived, resistance-associated allele; 0 = ancestral, wild-type residue).

Environmental Variable E Each genotype was assayed under a gradient of 10 discrete cycloguanil concentrations, ranging from 0 to $10^6\,\mu\text{M}$ on a logarithmic scale:

$$E \in \{0,\ 10^{-2},\ 10^{-1},\ 10^{0},\ 10^{1},\ \dots,\ 10^{6}\}\,\mu\text{M}.$$

This drug concentration axis enables modeling how increasing chemical pressure modulates genotype-specific fitness and uncovers epistatic interactions.

Dependent Trait t Fitness was quantified as the parasite's exponential growth rate r, then log-transformed to stabilize variance and accommodate zero growth:

$$t = \ln(r + 1).$$

This continuous phenotype is well suited for downstream analyses such as regression trees and landscape visualization.

All 16 DHFR genotypes were assayed across 10 cycloguanil concentrations, yielding $16\times 10 = 160$ genotype-environment fitness measurements. This dense, fully factorial design for a single drug provides sufficient statistical power to dissect both individual mutational effects and higher-order genotype-by-environment interactions under cycloguanil pressure.

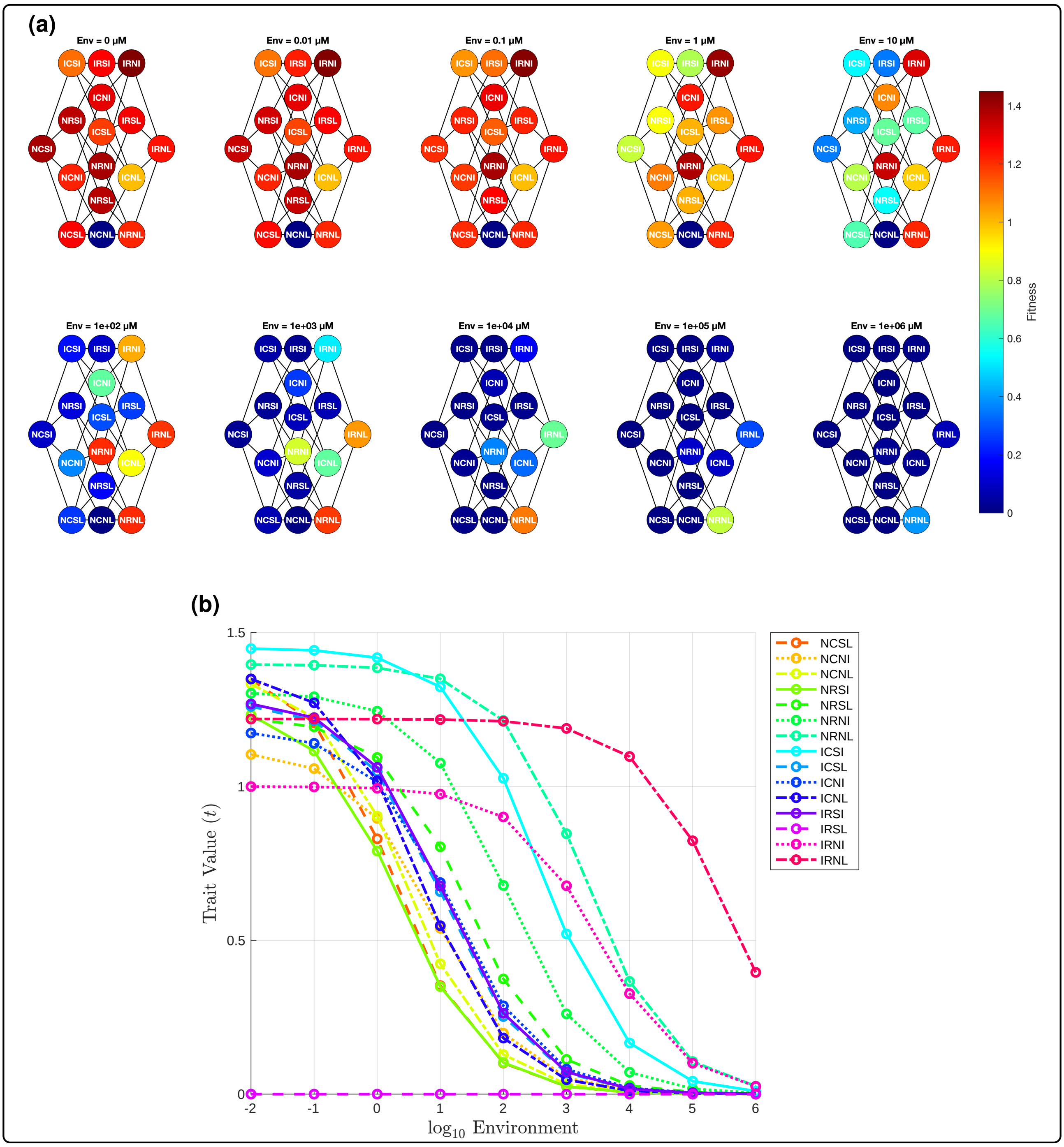

Fig. 3. Data for Case 2: Environmentally-dependent fitness landscapes and reaction norms of *Plasmodium falciparum* dihydrofolate reductase (DHFR) mutants associated with resistance to cycloguanil. The dataset from Ogbunugafor (2022) (*12*) comprises all $2^4 = 16$ drug-resistant dihydrofolate reductase (DHFR) genotypes in *Plasmodium falciparum*—a protozoan parasite and a primary cause of malaria in humans—defined by binary substitutions at four residues (N51I, C59R, S108N, I164L). Fitness $t = \ln(r + 1)$, where r is exponential growth rate, was measured across 10 cycloguanil concentrations ranging from 0 to $10^6\,\mu$M on a logarithmic scale: 0, 10^{-2}, 10^{-1}, 10^0, 10^1, … , $10^6\,\mu$M. **(a)** Genotype networks (4-dimensional hypercubes) for each drug concentration. Each node represents one of the 16 genotypes, arranged horizontally by the number of mutations relative to the wild type (0–4 mutations), and vertically for visual clarity. Edges connect genotypes differing by a single mutation. Node color encodes the log-transformed fitness (t) at that drug level, using a uniform colormap across panels. **(b)** Reaction norms showing how fitness (t) varies across drug concentrations for all genotypes. Each line represents a genotype's fitness trajectory, illustrating genotype–environment interactions and the nonlinear effects of increasing drug concentration.

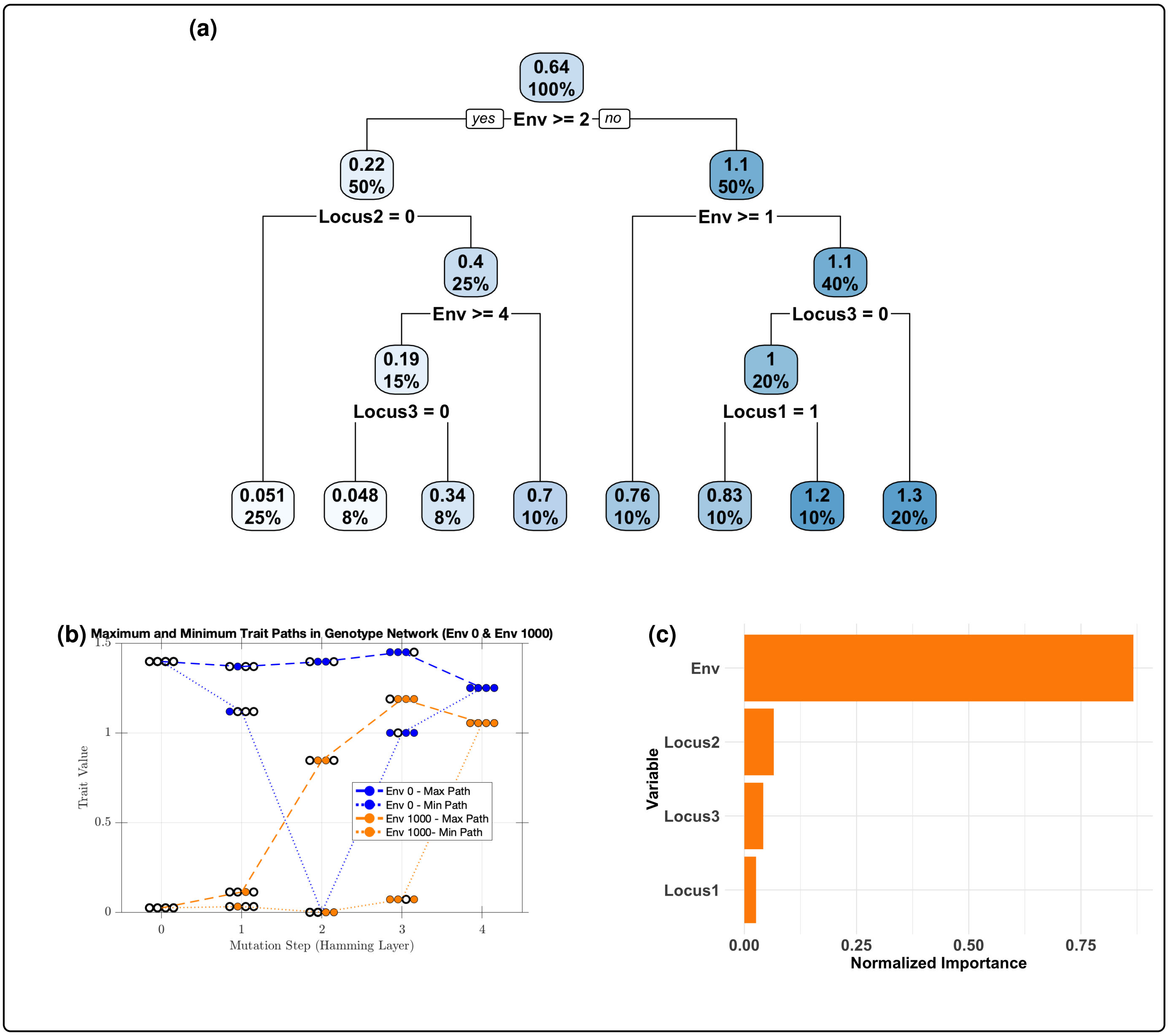

Fig. 4. Case 2 analysis: Genotype–phenotype mapping under cycloguanil reveals dominant environmental control and conditional mutational effects. (a) A decision tree predicting log-transformed growth ($\ln(r+1)$) from four binary DHFR loci—Locus 4 (N51I), Locus 3 (C59R), Locus 2 (S108N), and Locus 1 (I164L)—and $\log_{10}$-scaled cycloguanil concentration. Each internal node shows the split criterion (e.g., *Env* ≥ 2 corresponds to $\geq 100\mu\mathrm{M}$), the mean fitness at that node, and the percentage of observations routed there. The top split on concentration (mean fitness 0.64, 100% of data) partitions high- and low-drug regimes; subsequent splits reveal that C59R effects dominate at low concentration, while S108N effects emerge only at high concentration. Terminal nodes list mean fitness and proportion of genotypes in each final partition. **(b)** Fitness trajectories across mutational distance (Hamming steps from wild type) for the highest- and lowest-performing genotypes at two cycloguanil levels (0 and 1000 μM). At each step, four markers (ordered as Locus 4, 3, 2, 1 from left to right) are filled when the locus is mutated and open otherwise, illustrating the shift from C59R-driven adaptation in drug-free conditions to S108N-driven adaptation under high drug concentrations. **(c)** Variable importance scores from the decision tree model, normalized so that cycloguanil concentration accounts for about 85% of the total importance, followed by S108N (Locus 2) and C59R (Locus 3). This underscores the critical role of environmental drug pressure in modulating the fitness contributions of specific DHFR mutations. Data was obtained from the cycloguanil dataset of Ogbunugafor (2022) ([12]).

Table 2. Summary of decision tree results across datasets: environmental influence and epistatic interactions.See Table 3 for full description for all case data set.

Dataset	**Environment at Root or Early Split?**	**Environmental Threshold(s) Identified**	**Notable Mutation–Environment or Mutation–Species Interactions**	**Epistatic, G×E, or G×G×E Insights from Tree Structure**
Case 2: Malarial DHFR-antifolate (*12*) (see Section 2.2.1)	Yes	$\sim$ 10 μM drug concentration	At low drug levels, mutation at C59R improves fitness; S108N becomes critical above 10 μM for maintaining high growth.	Growth rate is primarily shaped by drug concentration. Mutation at C59R is beneficial without drug, while S108N is essential at higher levels. Growth declines above 10^4 μM.
Case 3: Bacterial LTEE-Chemical Stress (*11*) (see Section 2.2.2)	Yes	DM25+guanazole splits off low-fitness genotypes; DM25+EGTA yields highest fitness when paired with mutation at *glmUS*	Across all three environments, genotypes with a mutation in *glmUS* (locus 4) yields higher fitness than genotypes without a mutation in *glmUS*, but magnitude of effects are environment-dependent.	G×E interaction between environment and glmUS (locus 4).
Case 4: Bacterial DHFR-Proteostasis (*58*) (see Section S4.2.3)	No	Not applicable	In *E. coli*, L28R confers high resistance regardless of proteostasis context. In *L. grayi* and *C. muridarum*, L28R effects are modulated by proteostasis.	Species is the top-level determinant of resistance. Within *E. coli*, mutations dominate; within other species, proteostasis plays a stronger role. Reveals species-specific mutation effects and G×E, G×G×E interactions.

Building on the fully factorial cycloguanil assay described, our decision-tree analysis of *Plasmodium falciparum* DHFR provides a detailed view of how specific resistance mutations interact with cycloguanil concentration to shape parasite fitness (Figure 4). In particular, the first split in the tree occurs at approximately 100 μM ($\log_{10}E \approx 2$), underscoring that cycloguanil concentration alone explains over 75% of the model's predictive power (Figure 4c). Downstream of this primary bifurcation, the presence of the C59R mutation (Locus 3) defines the highest-fitness subpopulation under low-drug conditions, whereas in the high-drug branch the S108N mutation (Locus 2) becomes the critical determinant of elevated growth (Figure 4a). These context-dependent effects are echoed in the mutational-distance trajectories (Figure 4b): at 0 µM cycloguanil, the path of maximal fitness accrues C59R substitutions, while at 1000 µM only genotypes bearing S108N maintain high growth rates. Together, these findings illustrate pronounced genotype-by-environment interactions in the DHFR enzyme of the primary malaria agent, *P. falciparum*, revealing that the adaptive benefit of resistance-associated alleles shifts sharply with cycloguanil dosage. This allows for the clear interpretation of where mutational effects shift in response to environmental conditions, and the detection of conditional epistasis (i.e., epistatic effects with respect to the environment). The decision tree analysis also supports the concept of mutation-effect reaction norms

(Mu-RNs) discussed in the original dataset (*12*), where bifurcations on the decision tree define regimes where mutation effects change sign or magnitude.

2.2.2 Case 3: Environmental modulation of epistasis in *E. coli* from the LTEE

We applied CART to the comprehensive genotype-environment dataset used in Khan et al. (2011) (*59*) and Flynn et al. (2013) (*11*), which includes all 32 genotype combinations across five loci (*rbs*, *topA*, *spoT*, *glmUS*, *pykF*) in *E. coli*, measured across three defined environments: standard minimal medium (DM25), DM25 supplemented with EGTA, and DM25 supplemented with guanazole (see Figure 5 for a visualization). Fitness was quantified as the relative growth rate of each evolved genotype compared to the ancestral strain in direct competition assays, typically log-transformed to approximate normality.

Utilising the CART framework, the fitted regression tree identified environment as the most influential factor, with the root node splitting on whether the condition was DM25+guanazole (Figure 6a). This environment consistently led to lower log-transformed growth rates compared to the others. In DM25+guanazole, the presence or absence of mutation at *glmUS* (locus 4) was critical: strains lacking the mutation exhibited the lowest fitness, whereas strains with the mutation achieved relatively higher growth rates. In contrast, the branch corresponding to DM25 and DM25+EGTA showed overall elevated fitness values, with *glmUS* mutation again associated with higher fitness, particularly when the environment was DM25+EGTA. The beneficial effect of glmUS is dependent on the environmental context, thus highlighting a G×E interaction between the environment and glmUS. The variable importance scores (Figure 6b) confirmed that environment was the dominant predictor of fitness, followed by *glmUS* (locus 4) and *topA* (locus 2).

3 Discussion

Our study demonstrates the versatility and interpretability of decision tree methodologies, which encompass both classification and regression trees, in dissecting complex genotype-phenotype relationships across a spectrum of biological systems. These systems range from protein engineering efforts aimed at tuning fluorescent properties, through enzyme evolution influencing metabolic product specificity, to microbial adaptation under varying environmental and genetic backgrounds.

Our methods complement others used to study nonlinear interactions in genetic systems (e.g., Fourier Transform (*12*,*70*), linear models (*58*)), but do so in a manner that provides ease-of-interpretation, along with a visual representation that allows geneticists to identify factors that shape the genotype-phenotype map.

Uncovering environmental modulation of mutation effects and epistasis Disentangling the role of genetic and environmental effects is as fundamental as it gets in evolutionary and population genetics. This is highly relevant for questions ranging from the genetics of agricultural breeding, to personalized medicine, and modern efforts to engineer genomes with desirable phenotypes.

Our regression tree analyses of genotype-phenotype maps reveals the influence of environmental factors and genetic background on phenotypic expression and epistatic interactions. Unlike traditional parametric models that often treat environment as a linear co-variate or require explicit interaction terms, regression trees naturally incorporate environmental variables as partitioning factors, thereby detecting critical thresholds and nonlinear responses within complex genotype-genotype and genotype-environment spaces.

For example, analysis of the FP611 protein (Case 1) demonstrates the ability of CART to resolve how epistasis shapes fluorescence patterns. The analyses of both the *Plasmodium falciparum* DHFR dataset (Case 2) and the long-term evolution experiment (LTEE) data (Case 3) highlight how environment is a principal factor influencing fitness outcomes. In Case 2, the decision tree highlights pivotal thresholds in drug levels where the fitness effect of key resistance mutations such as S108N and C59R shifts markedly; in Case 3, the decision tree shows how effect of glmUS on the growth rates of *E. coli* is dependent on the environmental context. In both cases, the CART analysis indicates strong genotype-by-environment (G×E) interactions, thus emphasizing the dynamic nature of adaptive landscapes under varying environmental conditions (Figures 3 to 6).

Similarly, in the comprehensive dataset from Guerrero et al. (2019) (Case 4) (*58*), species-specific genetic background emerges as the main determinant of antibiotic resistance phenotypes, with subsequent splits by proteostasis environment and individual mutations refining fitness partitions. This hierarchy uncovers how species and cellular context modulate the penetrance and effect size of resistance-conferring mutations, illustrating intricate multilayered epistasis. The regression tree approach elucidates these conditional dependencies without the

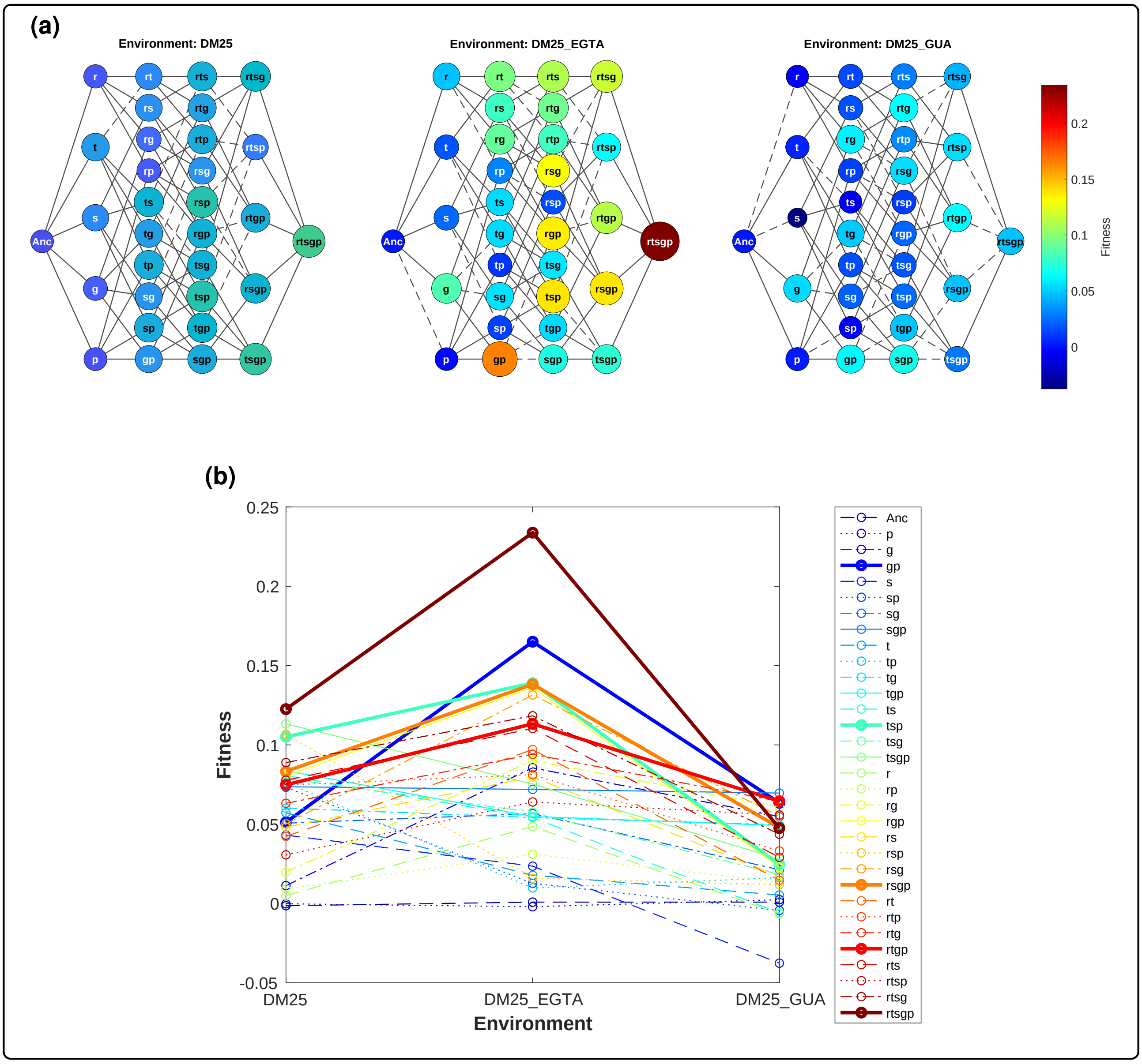

Fig. 5. Data for Case 3: Fitness landscapes and reaction norms across relevant environments in the LTEE (*E. coli*) dataset from (*59*) and (*11*). (a) Data for Case 3: Genotype networks representing 32 constructed genotypes, each corresponding to a unique combination of five mutations in *E. coli*. The ancestral genotype is labeled *Anc*, while other node labels denote mutations at the following loci: *r* (rbs), *t* (topA), *s* (spoT), *g* (glmUS), and *p* (pykF). Each network corresponds to one of three environments: DM25 (glucose-limited medium), DM25_EGTA (DM25 + EGTA), and DM25_gua (DM25 + guanazole). Genotypes are horizontally arranged by their mutational distance from the ancestor (0 to 5 mutations), and vertically offset for clarity. Node color and size represent the same quantity: the $\log_{10}$-transformed relative fitness of each genotype compared to the ancestor. Directed edges denote single-mutation transitions; solid edges indicate fitness gains, dotted edges indicate loss in fitness. **(b)** Reaction norms of the same genotypes across environments (from left to right: DM25, DM25 + EGTA, DM25 + guanazole). The y-axis displays $\log_{10}$-transformed relative fitness, and the x-axis corresponds to the three environments. Each line represents one genotype, with the top five (based on overall mean trait performance) highlighted using distinct solid lines. The remaining genotypes are shown using varied dashed and dotted styles to improve visual distinction.

need for pre-specified interaction terms, offering a transparent mapping of complex genetic and environmental interdependencies (Figure S3).

More generally, these findings are in conversation with studies that examine the role of environment and context in shaping adaptive evolution. For example, the notion of the fitness “seascape” has added even more nuance, offering that the shape of fitness landscapes is crafted by particulars of the environment (*71*,*72*). Tools such as CART allow us to understand the underlying questions about evolvability in fitness landscapes, which capture an essential piece necessary for evolutionary prediction.

In addition to this, the insight offered by CART can be applied for several practical purposes, such as evolutionary control (*73*,*74*) and personalized medicine. The regression tree helps identify the specific actors that

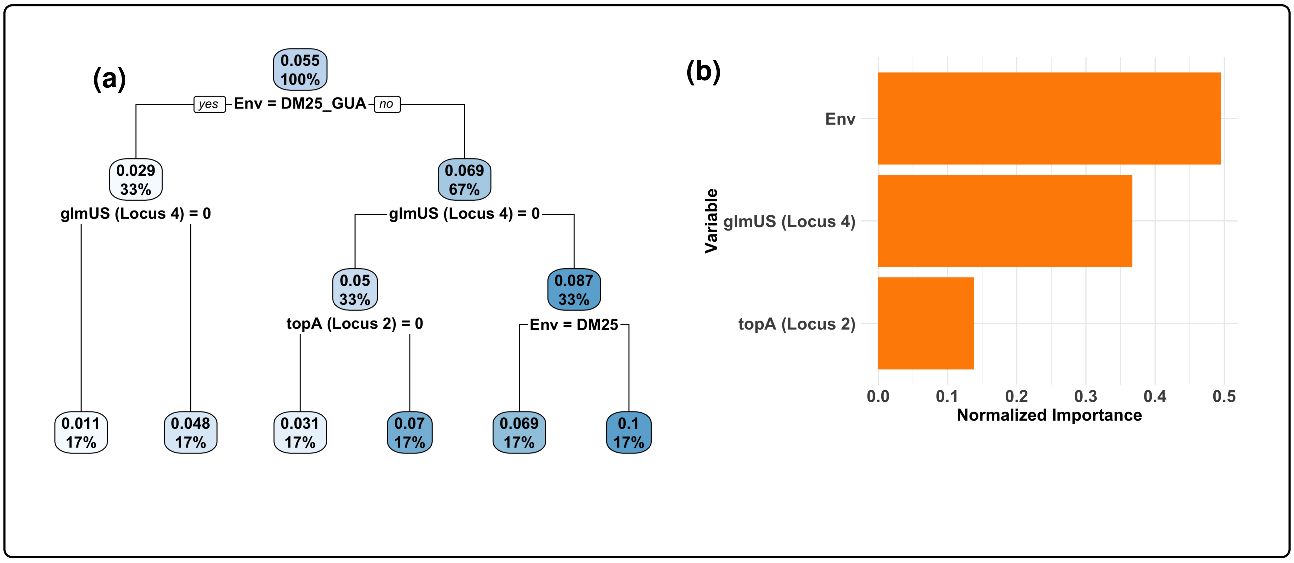

Fig. 6. Case 3 analysis: CART analysis of the LTEE *E. coli* dataset. (a) Regression tree predicting log-transformed growth rate based on genotype and environment. The root node splits on environment, with DM25+guanazole associated with reduced fitness. Within this environment, mutation at *glmUS* (locus 4) distinguishes low-fitness genotypes from moderately fit genotypes. In contrast, the other two environments (DM25 and DM25+EGTA) are associated with higher overall fitness, particularly when *glmUS* is mutated and the medium is DM25+EGTA. (b) Variable importance scores from the fitted CART model, highlighting environment as the dominant predictor, followed by *glmUS* (locus 4) and *topA* (locus 2). The dataset was obtained from by Khan et al. (2011) (*59*) (DM25) and Flynn et al. (2013) (*11*) (DM25 + EGTA, DM25 + guanazole).

shape genotype-phenotype maps and potentially utilize this information to engineer natural systems and to diagnose and treat disease.

Interpretability and mechanistic insight from classification and regression trees The classification tree analysis of the fluorescent protein mutational landscape exemplifies how a relatively simple, rule-based model can distill a high-dimensional combinatorial genotype space into a structured decision framework. This framework not only identifies a minimal subset of key loci but also reveals the hierarchical and conditional nature of their interactions, highlighting critical third-order epistatic effects that are difficult to capture with linear or additive models. Importantly, the classification tree provides an intuitive visualization of how specific mutational combinations lead to distinct fluorescence phenotypes, enabling direct biological interpretation and guiding future experimental design (Figure S2).

Similarly, the regression tree approach applied to the multivariate chemotype landscapes of sesquiterpene synthase enzymes offers nuanced insight into how specific amino acid substitutions and their interactions drive the production of diverse metabolic products. By separately modeling each enzymatic product as a quantitative trait, the regression trees illuminate threshold effects and context-dependent mutation roles, underscoring the complexity of biochemical epistasis shaping enzymatic specificity and promiscuity. This approach complements traditional parametric analyses by uncovering nonlinearities and higher-order interactions without requiring prior specification, thereby enhancing the interpretability of genotype-phenotype maps in enzyme evolution (Figures S6 and S7).

Together, these examples underscore the power of decision trees to bridge the gap between data-driven predictive modeling and mechanistic biological understanding. Their ability to represent complex, nonlinear genetic architectures in an accessible hierarchical format makes them valuable tools for researchers aiming to unravel the intricate patterns of genetic variation and function in diverse biological contexts.

Advantages of regression trees over classical parametric models Regression trees provide key benefits over classical parametric models such as ordinary least squares (OLS) or generalized linear models (GLMs), particularly for complex genotype–phenotype landscapes. Parametric models require explicit specification of main effects and interactions, which can lead to misspecification when unknown or higher-order interactions are present. Linear Mixed Models (LMMs) account for hierarchical structure, repeated measures, and genetic relatedness via random effects or kinship matrices (*75*–*77*), but they retain limitations in exploratory genotype-phenotype mapping (78) because they assume linear functional forms and require that interactions and nonlinear transforms be specified in advance, potentially missing abrupt regime changes or high-order epistasis. By contrast, regression trees are non-parametric and data-driven, automatically capturing nonlinearities, threshold effects, and interaction hierarchies without pre-specified terms.

Our supplementary LASSO analyses (Supplementary Section 4.2.3) serve as a linear reference model for comparison, showing that LASSO recovers many of the same influential loci identified by trees when predictors are ranked by absolute coefficient values. CART captures nonlinearities and threshold effects without feature engineering, as in the DHFR dataset where a sharp split at the environmental variable $(\log_{10} E \approx 2)$ explains the near-zero linear environmental term in LASSO. Trees reveal hierarchical and pathway-like dependencies, clarifying how the same set of substitutions can act differently across mutational contexts, as demonstrated in the FP611 fluorescence landscape. For complex chemotype distributions in terpene synthases, CART implicitly models high-order interactions that LASSO must explicitly encode, yielding compact and interpretable decision rules rather than large multi-way coefficients. Trees accommodate mixed predictor types, tolerate modest missingness and outliers, and reduce preprocessing demands, as illustrated in the LTEE analyses. Single-tree outputs provide explicit, mechanistic rules that are readily interpretable by domain experts, offering intuitive visualizations of how mutations and environmental factors interact to shape phenotype (79). Regression trees can also reveal biologically meaningful thresholds, such as concentration cutoffs where resistance mutations shift in effect, uncovering adaptive constraints and complex emergent patterns that classical models may overlook.

Limitations Although our study examines genotype-phenotype maps of various sorts, most of the data sets are in the form of empirical fitness landscapes (see Table 1 for definitions), relatively small in size. Case 1, however, provides a larger genotype-phenotype map (over 8000 genotypes). But we must acknowledge that real-world data sets can be much larger. The CART method is flexible with regard to the size of the data set, and future efforts can apply the method to larger-scale genotype-phenotype maps.

Although CART provides a flexible and interpretable framework for analyzing genotype-phenotype maps, several limitations warrant attention. A single CART can produce spurious splits when the predictor space is large relative to sample size, when predictors are rare or highly categorical, or when predictors are strongly correlated (34,80,81). To mitigate these risks researchers can use cross validated pruning, constrain tree depth and minimum node size, apply conditional inference trees with permutation based split tests, assess rule stability by subsampling or stability selection, employ ensemble methods with permutation based variable importance, and control multiple testing or seek external replication for reported rules (19,82). Additionally, overfitting remains a central concern in high dimensional, low sample contexts (22,34). To mitigate this, ensemble techniques such as random forests (40) or gradient boosting (83) can improve generalization while hybrid approaches that combine trees with kernels or splines (84–86) can better capture smooth effects.

Finally, current applications often rely on structured datasets, which may not capture the complexity of natural systems. Extending tree-based methods to dynamic, ecological, or population-genomic data and integrating multi-omics modalities offers a promising direction for mapping genotype-phenotype relationships.

Conclusion In this study, we show that classification and regression trees (CART), applied in a unified framework across diverse genotype-phenotype datasets, provide interpretable, context-sensitive models that capture key loci and nonlinearity, higher-order epistasis, and environment-dependent effects with minimal parametric assumptions. It can be readily integrated into hybrid or ensemble pipelines and supports experimental design and hypothesis generation (e.g., for selecting informative mutation sets). More generally, combining tree-based discovery with causal inference and targeted experimentation offers a practical route from descriptive models to explanatory insights, with applications spanning protein engineering, evolutionary biology, and precision interventions.

4 Methods

The primary objective of this study is to explore the structural relationships (potentially including non-linear and higher-order) between phenotypic trait values and a set of predictor variables that include both genotypic variation at specific loci and relevant environmental attributes. In this article, we propose the use of a supervised learning algorithm, namely classification and regression trees (CART) (21,34), as a simple, interpretable, and structurally informative tool for identifying genotype-environment-trait relationships.

Let the dataset be denoted by $\{(\mathbf{X}_i,\ t_i)\}_{i=1}^{M}$, where t_i is the trait value for the i^{th} observation, and $\mathbf{X}_i = (X^{(1)},\ X^{(2)},\ \ldots,\ X^{(N)})_i$ represents a vector of predictor variables that may include both continuous environmental covariates and binary indicators of mutations at specific loci. The goal of supervised learning in this context is to estimate a mapping f such that $\hat{t} = f(\mathbf{X})$, where $\hat{t}$ denotes the predicted trait value based on the predictor variables $\mathbf{X}$ (see Figure 7).

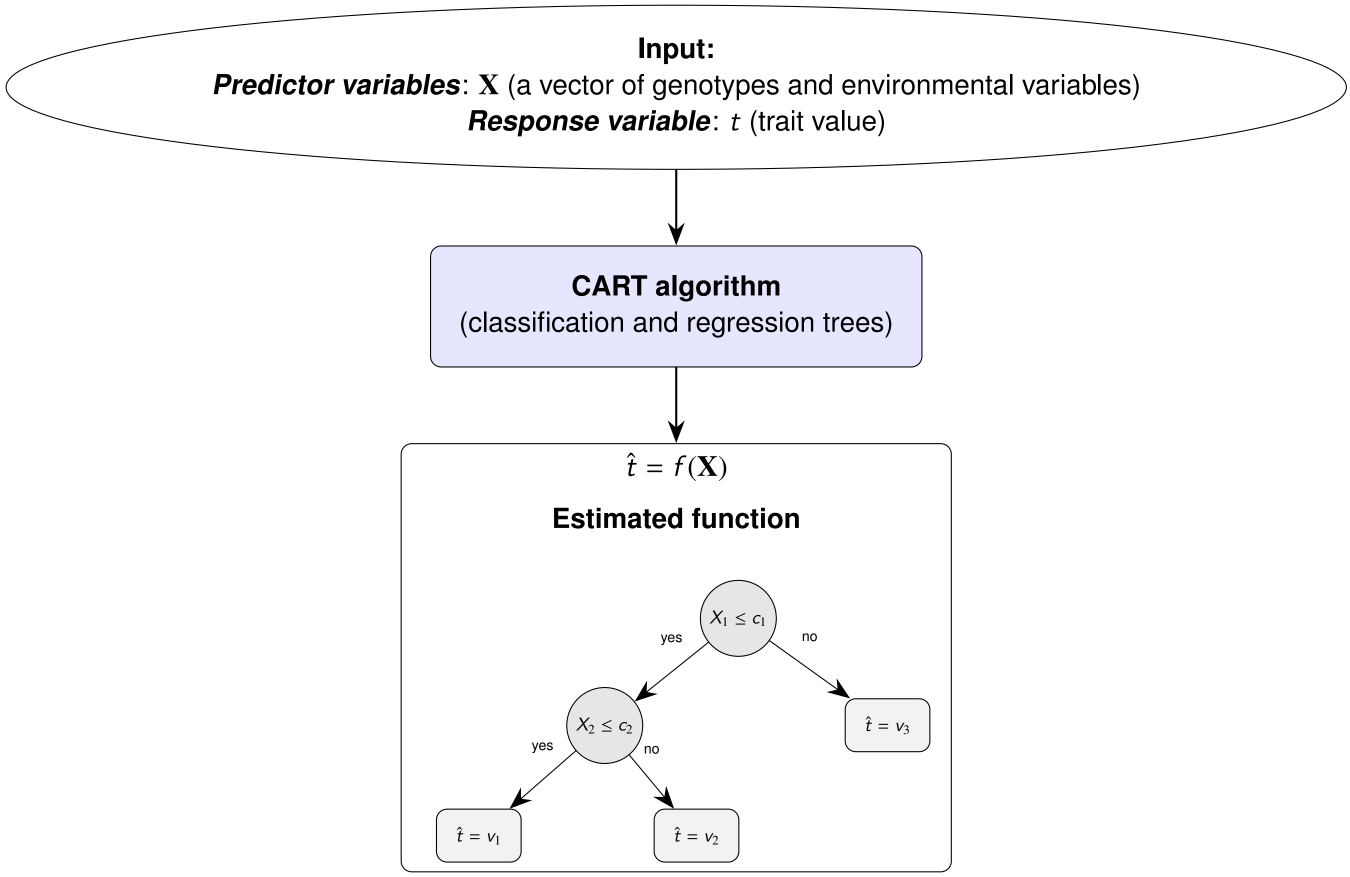
Input:
Predictor variables: $\mathbf{X}$ (a vector of genotypes and environmental variables)
Response variable: t (trait value)
CART algorithm
(classification and regression trees)
$\hat{t} = f(\mathbf{X})$
Estimated function
$X_1 \leq c_1$
yes
no
$X_2 \leq c_2$
yes
no
$\hat{t} = v_3$
$\hat{t} = v_1$
$\hat{t} = v_2$

Fig. 7. Overview of the CART-Based trait estimation framework. Schematic representation illustrating how both the predictor variables $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(N)}$—where each $X^{(j)}$ may represent genotypic information at specific loci or environmental attributes—and the response variable t (trait value) are used as inputs to the CART algorithm. The algorithm yields an estimated function f, which is represented as a decision tree. Here, c_1 and c_2 denote the threshold values for the splitting criteria, while v_1, v_2, and v_3 correspond to the predicted values at the terminal nodes.

While the CART algorithm is commonly applied in predictive modeling tasks, it is equally well-suited for exploratory analysis due to its ability to reveal complex structural patterns within the data (*34*). In this study, we utilize CART primarily as an exploratory tool to uncover interpretable decision tree structures that illuminate the relationships between genotypic and environmental predictors and phenotypic trait values. The resulting tree structures capture interactions and non-linear effects through recursive binary partitioning, providing an intuitive and visual representation of the relationships between predictors and the response (*87*,*88*). We apply the CART algorithm to several empirical datasets obtained from the literature, using the full set of features and observations without partitioning for model validation. Since our emphasis is on interpretability rather than predictive optimization, we focus on analyzing the learned structures for biological relevance and consistency with existing domain knowledge.

In this section, we first introduce the empirical datasets used in this study. We then review the theoretical foundations and methodological implementation of the CART algorithm, followed by a demonstration of its application using publicly available tools in R.

4.1 Empirical Data Sources

To investigate the structural relationships between phenotypic trait values and a set of predictor variables that include both genotypic variation at specific loci and relevant environmental attributes, we compiled a diverse collection of empirical datasets from the literature. Each dataset encompasses measurements across different biological systems and contains trait values alongside genetic (all cases) and environmental (cases 2,3,4) predictors. This subsection provides a concise summary of the datasets in relation to our analytical objectives. For complete methodological details and domain-specific context, readers are referred to the original sources.

A summary of the key features of each dataset is provided in Table 3. The table includes information on the predictor variables (e.g., specific mutations or environmental treatments), the corresponding response variable (typically a phenotypic trait or fitness measurement), the statistical type of the response (categorical or continuous), a brief contextual description, and the original data source.

The predictor variables across datasets vary in nature and scale, encompassing binary indicators of amino acid substitutions, drug concentration gradients, and environmental stress conditions such as protein folding environments or media composition. These features allow for the exploration of diverse forms of gene-environment interactions. The trait values, used as the response variable in our analyses, are derived from experimental measurements such as growth rates, drug resistance (e.g., IC_{50}), or relative product abundance in biosynthetic pathways. Several datasets include log-transformed or normalized versions of these traits to facilitate interpretation. Where appropriate, trait values are treated either as continuous variables (e.g., log-growth or production percentage) or categorical variables (e.g., phenotypic classes like red, blue, or black colony colors), depending on the structure of the original study and the modeling goal. This flexibility allows us to apply both classification and regression trees to highlight different aspects of the genotype-environment-trait relationship. Together, these datasets provide a diverse and representative foundation for evaluating the capacity of CART to uncover complex, potentially nonlinear and higher-order structural relationships in biological systems.

Table 3. Summary of empirical datasets used in the analysis

Dataset	Predictor Variables	Response Variable	Trait Type	Description	Source
Case 1: Bubble-tip anemone. FP611 protein (CART results in Section 2.1)	Binary indicators of amino acid substitutions at 13 loci (from mTagBFP2 to mKate2)	fluorescence intensity or color class (red, blue, black) in *Nicotiana tabacum*	Continuous or Categorical	Exhaustive mutational library covering all $2^{13} = 8192$ genotypes between two fluorescent proteins differing at 13 positions. Trait values include quantitative brightness and visual color class, enabling both regression and classification tasks.	Poelwijk et al. (2019) (*57*)
Case 2: Malaria parasite. DHFR–Antifolate (CART results in Section 2.2.1)	Binary indicators for mutations at four loci in the dihydrofolate reducase in *P. falciparum*; drug concentration gradients of pyrimethamine and cycloguanil	Log-transformed growth rate r, $\ln(r+1)$	Continuous	Fitness measured as growth rate under drug treatments; transformation accounts for zero growth rates. Dataset captures genotype-by-environment interactions involving drug concentration gradients.	Ogbunugafor (2022) (*12*)
Case 3: Bacteria, long-term evolution experiment. LTEE–chemical stress (CART results in Section 2.2.2) in *E.coli*	Binary indicators of amino acid substitutions at five loci arising from the long-term evolution experiment (LTEE); environmental conditions: DM25, DM25 + EGTA, DM25 + guanazole	Fitness: $\log_{10}$(growth rate)	Continuous	Evaluates epistatic interactions and environmental effects on fitness landscapes; fitness data log-transformed as per Weinreich et al. (2018)	Flynn et al. (2013) (*11*)
Case 4: Bacteria, proteostasis environments. DHFR-Proteostasis (CART results in Section S4.2.3)	Three biallelic sites/loci in dihydrofolate reductase across three bacterial species, (*E. coli*, *L. grayi*, *C. muridarum*); three proteostatic environments: WT, GroEL/ES overexpression, and lon protease absence	Fitness measured as $\ln(\text{IC}_{50}+1)$ or $\ln(\text{DHFR abundance})$	Continuous	Fitness data include antibiotic resistance (IC50) and DHFR protein abundance, with log transformation applied to IC50 to handle zeros. Study investigates genotype-by-environment interactions under diverse proteostasis conditions.	Guerrero et al. (2019) (*58*)
Case 5: Cultivated tobacco. STS–Chemotype Mapping (CART results in Section S4.2.3)	Binary indicators of amino acid substitutions at 6 loci (I372V, S402T, L406Y, T438I, L439I, I516V); subset of genotypes with fixed A274T, V291A, S406N retained; data from *N. tabacum*	Relative abundance (percentage) of sesquiterpene products: premnaspirodiene (PSD), 4-epi-eremophilene (4-EE), and 5-epi-aristolochene (5-EA)	Continuous (proportion)	Analyzes structural–functional mapping in terpene synthases using a genotype–phenotype–chemotype framework across a reduced fitness landscape	O'Maille et al. (2008) (*60*)

4.2 Classification and regression trees (CART)

The classification and regression trees (CART) algorithm, developed by Breiman et al. (1984), is a non-parametric method used for both classification and regression tasks. It constructs binary decision trees by recursively partitioning the feature space to maximize the homogeneity of the target variable within each partition ([21,34]). As the name suggests, the CART algorithm (usage demonstrated in Section 4.2.3 with the `rpart` package in `R`) is applied to *classification* tasks when the response variable is categorical and to *regression* tasks when the response variable is continuous.

Furthermore, CART is capable of handling mixed data types—both categorical and continuous—among the predictor variables. It is particularly well-suited for problems involving non-linear relationships and high-order interactions among predictors ([88]). One of CART's key advantages over many traditional statistical methods lies in its non-parametric nature ([87,88]), meaning that it does not require assumptions about the underlying distributions of the predictor or response variables. A major strength of the CART algorithm in the context of understanding the structural relationship between genotypes, environmental variables, and trait values is its interpretability ([88,89]). The resulting decision tree is intuitive and easy to visualize, offering valuable insights into how various predictors influence the response. Additionally, CART is robust to missing data and does not require extensive data preprocessing, which enhances its practical applicability. Most importantly, the algorithm naturally captures interaction effects and non-linear relationships through its hierarchical splitting mechanism ([34,87,88, 90]).

Depending on the context and the trait under investigation, trait values may be either categorical or continuous ([91,92]). Based on the data type of the response variable, i.e., whether it is categorical or continuous, the CART algorithm selects an appropriate model: a classification tree for categorical outcomes, or a regression tree for continuous outcomes. In what follows, we briefly review the general methodology for constructing both classification and regression trees. For a more detailed mathematical formulation and theoretical analysis, the reader is referred to ([34], Chapter 7) for classification trees and ([34], Chapter 8) for regression trees.

4.2.1 Classification trees

In the classification setting where the response variable (in this case, the trait value) is categorical, the CART algorithm constructs a binary decision tree by recursively partitioning the data to increase the homogeneity of class labels within the resulting nodes. Each observation is assumed to be an independent realization of the random vector $(\mathbf{X}, t)$ and is represented by the pair $(\mathbf{X}_i, t_i)$, where $\mathbf{X}_i = (X^{(1)}, X^{(2)}, \ldots, X^{(N)})$ denotes the vector of predictor variables and t_i the corresponding response variable with K classes. The algorithm proceeds via binary recursive partitioning, successively splitting a node into two child nodes based on a single predictor variable, chosen to minimize a measure of impurity such as the Gini index or entropy. Variables may be reused across different levels of the tree or omitted entirely, depending on their relevance. This process continues until a stopping condition (user define threshold) is met, resulting in a hierarchical structure that encodes the classification rules inferred from the data ([88]).

Classification tree construction and splitting criteria. Let us consider a decision tree f, and let d denote one of its nodes. Mathematically, the tree f defines a mapping that assigns each input sample $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(N)})$ to a terminal node. Equivalently, the tree can be viewed as a function that produces a predicted class label $k_i = \hat{t}_i = f(\mathbf{X}_i)$ for each observation i (see Figure 7). To quantify class impurity within a node d, let $p(k \mid d)$ denote the proportion of samples in node d that belong to class k. Two commonly used impurity measures are the entropy and the Gini index. The entropy at node d is defined as

$$E_d = -\sum_{k=1}^{K} p(k \mid d) \log_2 p(k \mid d),$$

with the convention that $x \log_2 x = 0$ when $x = 0$. The Gini index is given by

$$G_d = 1 - \sum_{k=1}^{K} p(k \mid d)^2.$$

Both impurity measures attain a value of zero when node d contains samples from only a single class, and they reach their maximum when all classes are equally represented. At each node d, the CART algorithm evaluates all possible binary splits based on thresholds applied to the predictor variables. The goal is to divide the current (parent) node into two child nodes—denoted d_L (left) and d_R (right)—in a way that maximizes the reduction in impurity.

This reduction is defined as the difference between the impurity of the parent node and the weighted sum of the impurities of the two resulting child nodes. When observations are treated as independent samples, the weights correspond to the proportion of samples in each child node relative to the total number in the parent node. A split is implemented only if this reduction is positive. For instance, the reduction in impurity, as measured by the Gini index, is computed as (similarly for entropy, by substituting G with E):

$$\Delta G = G_d - \frac{N(d_L)}{N(d)} G_{d_L} - \frac{N(d_R)}{N(d)} G_{d_R},$$

where $N(\cdot)$ denotes the number of samples in a node.

The recursive partitioning procedure continues until no further admissible splits can be made. Each terminal node is then assigned the class most frequently represented among its samples (i.e., the conditional mode). However, fully grown trees often overfit the training data, leading to high prediction error, defined as

$$R(f) = P\{f(\mathbf{X}) \neq t\},$$

where f denotes the tree-based classifier.

Pruning and model selection in classification trees. The primary goal in constructing a classification tree is to derive a function f that minimizes the prediction error $R(f)$. To prevent overfitting, pruning is applied to produce a subtree f_s with lower expected risk. Since the true distributions of t and $\mathbf{X}$ are generally unknown, pruning relies on minimizing the empirical risk

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{f(\mathbf{X}_i) \neq Y_i\},$$

where $\mathbb{I}(\cdot)$ is the indicator function and n is the sample size. CART employs cost-complexity pruning to balance empirical risk against model complexity by minimizing the penalized risk

$$R_\alpha(f) = R(f) + \alpha|f|,$$

where $R(f)$ is the resubstitution error, $|f|$ is the number of terminal nodes, and $\alpha \geq 0$ controls the tradeoff between fit and simplicity. Optimal subtrees are typically selected via cross-validation (34,88).

4.2.2 Regression trees

When the response variable t is continuous, the CART algorithm constructs a *regression tree* by recursively partitioning the predictor space to minimize within-node variance (34). Each observation $(\mathbf{X}_i, t_i)$, where $\mathbf{X}_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(N)})$, is assumed to be an independent realization of a random vector. The goal is to derive a function $f : \mathbb{R}^N \to \mathbb{R}$ that approximates the conditional expectation $\mathbb{E}[t \mid \mathbf{X}]$.

In our context, the predictor vector $\mathbf{X}$ may contain a mixture of continuous and categorical variables. The CART algorithm accommodates such heterogeneity by applying variable-type-specific splitting rules, while maintaining a unified recursive partitioning framework (93).

Regression tree construction splitting criterion. At each internal node d, CART evaluates candidate splits for all predictor variables. The algorithm selects the split that maximizes the reduction in node impurity, measured by the residual sum of squares (RSS). Specifically:

- **For continuous predictors:** CART considers binary splits of the form

  $$X^{(j)} < s,$$

  where s is a threshold chosen from the set of observed values of $X^{(j)}$. Each such threshold induces a partition into two child nodes, and the split that results in the greatest reduction in impurity is selected (34).

- **For categorical predictors:** For a variable $X^{(j)}$ with W distinct categories, CART evaluates all non-trivial binary partitions of the category set $C = \{c_1, c_2, \dots, c_W\}$ into subsets S and $C \setminus S$. Each split is of the form

  $$X^{(j)} \in S \quad \text{vs.} \quad X^{(j)} \notin S,$$

  and the best partition is selected based on impurity reduction (93).

Let $N(d)$ denote the number of observations in node d, and let $\bar{t}_d = \frac{1}{N(d)} \sum_{\mathbf{X}_i \in d} t_i$ denote the mean response in that node. The impurity of node d is given by:

$$\mathrm{RSS}(d) = \sum_{\mathbf{X}_i \in d} (t_i - \bar{t}_d)^2.$$

For a candidate split of node d into child nodes d_L and d_R, the reduction in impurity is defined as:

$$\Delta\mathrm{RSS} = \mathrm{RSS}(d) - [\mathrm{RSS}(d_L) + \mathrm{RSS}(d_R)] \, .$$

The split that yields the maximum $\Delta\mathrm{RSS}$ is selected and applied, provided that the reduction is positive.

This recursive partitioning process continues until a stopping criterion is met, such as a minimum node size or a minimum reduction in RSS. Once the tree is fully grown, each terminal node is assigned the average of the response values of the observations it contains. The resulting regression function is thus a piecewise constant estimator:

$$f(\mathbf{X}) = \bar{t}_d, \quad \text{for } \mathbf{X} \in d.$$

Pruning and model selection in regression trees. As in classification trees, regression trees are prone to overfitting when fully grown. To mitigate this, CART employs *cost-complexity pruning*, which seeks a subtree f' that optimally balances goodness of fit and model complexity. The penalized risk is defined as:

$$R_\alpha(f) = \sum_{d \in \mathcal{T}_f} \mathrm{RSS}(d) + \alpha |f|,$$

where $\mathcal{T}_f$ denotes the set of terminal nodes in tree f, $|f|$ is the number of terminal nodes, and $\alpha \geq 0$ is a complexity parameter. The optimal subtree is selected via cross-validation by identifying the value of α that minimizes the estimated prediction error (34).

4.2.3 Variable importance in CART

An important feature of CART models is the ability to quantify the relative importance of predictor variables in determining the response. Variable importance provides insights into which variables contribute most to predicting the response, aiding interpretability and feature selection. The *variable importance score* VI_j for predictor $X^{(j)}$ is computed by aggregating the impurity reduction attributable to splits involving $X^{(j)}$ across all nodes in the tree. Formally, for each node d in the CART model f (whether classification or regression) where a split is performed on the variable $X^{(j)}$, let $\Delta I(d)$ denote the resulting reduction in impurity.

- For **classification trees**, $I(d)$ is typically the Gini index or entropy.
- For **regression trees**, $I(d)$ corresponds to the residual sum of squares (RSS).

The total importance for variable $X^{(j)}$ is then the sum of $\Delta I(d)$ over all nodes d split on $X^{(j)}$:

$$VI_j = \sum_{\substack{d \in f \\ \text{split on } X^{(j)}}} \Delta I(d).$$

This raw importance score is often normalized by dividing by the sum of all variable importances, yielding relative importance scores that sum to 1:

$$VI_j^{\mathrm{rel}} = \frac{VI_j}{\sum_{k=1}^{N} VI_k}.$$

Variables with higher VI_j^{rel} are considered more influential in predicting the response variable. This unified framework applies seamlessly to both classification and regression trees, differing only in the choice of impurity measure. For further details and applications of variable importance in CART, see (34,93).

Data Availability

Data and code can be found at https://github.com/OgPlexus/Cartepistasis1

Author Contributions

Project conception: SS, SNM, CBO; Data curation; SNM; Analysis: SS and CBO; Interpretation: SS, SNM, LC, CBO; Writing, original draft: SS, SNM, CBO; Writing, revision: SS, SNM, LC, CBO; Supervision: LC and CBO.

Acknowledgements

We thank the organizers and participants of the 2024 Modeling and Theory in Population Biology workshop at the Banff International Research Station (BIRS) for Mathematical Innovation and Discovery, where many of the ideas in this manuscript were first discussed. We are also grateful to the organizers and attendees of the 2022 workshop Reimagining the Central Dogma at The Foundations Institute, University of California, Santa Barbara, for conversations that informed this work. Finally, we acknowledge the Santa Fe Institute working group Biological Information and Environmental Uncertainty (May 2025), whose discussions further shaped the development of this article.

Funding support

This work was supported by the A*STAR National Science Scholarship, Singapore (S.N.M.); the National Science Foundation's Division of Environmental Biology Award Number 2142720 (C.B.O.); The National Institutes of Health R01AI168166 (C.B.O.); and by the Stephen Doro MD, PhD Family Private Foundation Fund (SS and CBO).

Conflicts of Interest

None reported.

References

1. Mackay TF, Stone EA and Ayroles JF (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10(8)**:565–577

2. Hill WG, Goddard ME and Visscher PM (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4(2)**:e1000008

3. Smith EN and Kruglyak L (2008). Gene–environment interaction in yeast gene expression. *PLoS biology* **6(4)**:e83

4. Starr TN and Thornton JW (2016). Epistasis in protein evolution. *Protein science* **25(7)**:1204–1218

5. Li C and Zhang J (2018). Multi-environment fitness landscapes of a tRNA gene. *Nature ecology & evolution* **2(6)**:1025–1032

6. Thomas D (2010). Gene–environment-wide association studies: emerging approaches. *Nature reviews genetics* **11(4)**:259–272

7. Weinreich DM, Lan Y, Wylie CS and Heckendorn RB (2013). Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development* **23(6)**:700–707

8. Kerwin RE, Feusier J, Muok A, Lin C, Larson B, Copeland D, Corwin JA, Rubin MJ, Francisco M et al. (2017). Epistasis× environment interactions among *Arabidopsis thaliana* glucosinolate genes impact complex traits and fitness in the field. *New Phytologist* **215(3)**:1249–1263

9. Remold SK and Lenski RE (2004). Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nature genetics* **36(4)**:423–426

10. Lindsey HA, Gallie J, Taylor S and Kerr B (2013). Evolutionary rescue from extinction is contingent on a lower rate of environmental change. *Nature* **494(7438)**:463–467

11. Flynn KM, Cooper TF, Moore FB and Cooper VS (2013). The environment affects epistatic interactions to alter the topology of an empirical fitness landscape. *PLoS genetics* **9(4)**:e1003426

12. Ogbunugafor CB (2022). The mutation effect reaction norm (mu-rn) highlights environmentally dependent mutation effects and epistatic interactions. *Evolution* **76(s1)**:37–48

13. Hall AE, Karkare K, Cooper VS, Bank C, Cooper TF and Moore FBG (2019). Environment changes epistasis to alter trade-offs along alternative evolutionary paths. *Evolution* **73(10)**:2094–2105

14. Paaby AB and Rockman MV (2014). Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics* **15(4)**:247–258

15. Zebell SG, Martí-Gómez C, Fitzgerald B, Cunha CP, Lach M, Seman BM, Hendelman A, Sretenovic S, Qi Y et al. (2025). Cryptic variation fuels plant phenotypic change through hierarchical epistasis. *Nature* pp. 1–9

16. Zuk O, Hechter E, Sunyaev SR and Lander ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109(4)**:1193–1198

17. Mackay TF (2014). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* **15(1)**:22–33

18. Starr TN, Flynn JM, Mishra P, Bolon DN and Thornton JW (2018). Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proceedings of the National Academy of Sciences* **115(17)**:4453–4458

19. Cordell HJ (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10(6)**:392–404

20. Lou XY (2014). Gene-Gene and Gene-Environment Interactions Underlying Complex Traits and their Detection. *Biometrics & biostatistics international journal* **1(2)**:00007

21. Breiman L, Friedman J, Olshen RA and Stone CJ (2017). *Classification and regression trees*. Routledge

22. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J and Laviolette F (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific reports* **9(1)**:4071

23. Perelygin V, Kamelin A, Syzrantsev N, Shaheen L, Kim A, Plotnikov N, Ilinskaya A, Ilinsky V, Rakitko A et al. (2025). Deep Learning captures the effect of epistasis in multifactorial diseases. *Frontiers in Medicine* **11**:1479717

24. da Costa WG, de Oliveira Celeri M, de Paiva Barbosa I, Silva GN, Azevedo CF, Borem A, Nascimento M and Cruz CD (2022). Genomic prediction through machine learning and neural networks for traits with epistasis. *Computational and Structural Biotechnology Journal* **20**:5490–5499

25. Behr M, Kumbier K, Cordova-Palomera A, Aguirre M, Ronen O, Ye C, Ashley E, Butte AJ, Arnaout R et al. (2024). Learning epistatic polygenic phenotypes with Boolean interactions. *Plos one* **19(4)**:e0298906

26. Lee HJ, Lee JH, Gondro C, Koh YJ and Lee SH (2023). deepGBLUP: joint deep learning networks and GBLUP framework for accurate genomic prediction of complex traits in Korean native cattle. *Genetics Selection Evolution* **55(1)**:56

27. Montesinos-Lopez A, Crespo-Herrera L, Dreisigacker S, Gerard G, Vitale P, Saint Pierre C, Govindan V, Tarekegn ZT, Flores MC et al. (2024). Deep learning methods improve genomic prediction of wheat breeding. *Frontiers in Plant Science* **15**:1324090

28. Montesinos-López A, Montesinos-López OA, Ramos-Pulido S, Mosqueda-González BA, Guerrero-Arroyo EA, Crossa J and Ortiz R (2025). Artificial intelligence meets genomic selection: comparing deep learning and GBLUP across diverse plant datasets. *Frontiers in Genetics* **16**:1568705

29. Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1(5)**:206–215

30. Balvert M, Cooper-Knock J, Stamp J, Byrne RP, Mourragui S, van Gils J, Benonisdottir S, Schlüter J, Kenna K et al. (2024). Considerations in the search for epistasis. *Genome Biology* **25(1)**:296

31. Chicco D and Faultless T (2021). Brief survey on machine learning in epistasis. *Epistasis* pp. 169–179

32. Ansarifar J and Wang L (2019). New algorithms for detecting multi-effect and multi-way epistatic interactions. *Bioinformatics* **35(24)**:5078–5085

33. Niel C, Sinoquet C, Dina C and Rocheleau G (2015). A survey about methods dedicated to epistasis detection. *Frontiers in genetics* **6**:285

34. Breiman L, Friedman J, Olshen R and Stone C (1984). Classification and regression trees–crc press. *Boca Raton, Florida* **685**

35. Huang J, Lin A, Narasimhan B, Quertermous T, Hsiung CA, Ho LT, Grove JS, Olivier M, Ranade K et al. (2004). Tree-structured supervised learning and the genetics of hypertension. *Proceedings of the National Academy of Sciences* **101(29)**:10529–10534

36. Assareh A (2012). *Optimizing decision tree ensembles for gene-gene interaction detection*. Kent State University

37. Chen Q, Zhang X and Zhang R (2019). Privacy-preserving decision tree for epistasis detection. *Cybersecurity* **2(1)**:7

38. Novielli P, Romano D, Pavan S, Losciale P, Stellacci AM, Diacono D, Bellotti R and Tangaro S (2024). Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: a case study with a dataset from an almond germplasm collection. *Frontiers in Plant Science* **15**:1434229

39. Lewis Schmalohr C, Grossbach J, Clément-Ziza M and Beyer A (2018). Detection of epistatic interactions with Random Forest. *BioRxiv* p. 353193

40. Breiman L (2001). Random forests. *Machine learning* **45(1)**:5–32

41. Saha S, Perrin L, Röder L, Brun C and Spinelli L (2022). Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests. *Nucleic Acids Research* **50(19)**:e114–e114

42. Jiang R, Tang W, Wu X and Fu W (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics* **10(Suppl 1)**:S65

43. Li J, Horstman B and Chen Y (2011). Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics* **27(13)**:i222–i229

44. Holliday JA, Wang T and Aitken S (2012). Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (Picea sitchensis) using random forest. *G3: Genes| genomes| genetics* **2(9)**:1085–1093

45. Orlenko A and Moore JH (2021). A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. *BioData mining* **14(1)**:9

46. Stephan J, Stegle O and Beyer A (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature communications* **6(1)**:7432

47. Verma SS, Lucas A, Zhang X, Veturi Y, Dudek S, Li B, Li R, Urbanowicz R, Moore JH et al. (2018). Collective feature selection to identify crucial epistatic variants. *BioData mining* **11(1)**:5

48. Alves AAC, da Costa RM, Fonseca LFS, Carvalheiro R, Ventura RV, Rosa GJdM and Albuquerque LG (2022). A Random Forest-based genome-wide scan reveals fertility-related candidate genes and potential inter-chromosomal epistatic regions Associated with Age at First Calving in Nellore cattle. *Frontiers in genetics* **13**:834724

49. Yao C, Spurlock D, Armentano L, Page Jr C, VandeHaar M, Bickhart D and Weigel K (2013). Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *Journal of Dairy Science* **96(10)**:6716–6729

50. Yoshida M and Koike A (2011). SNPInterForest: a new method for detecting epistatic interactions. *BMC bioinformatics* **12(1)**:469

51. Sandhu M (2025). *The Ruggedness of Protein Fitness Landscapes: Implications for Evolution and Machine Learning*. Mphil thesis, Australian National University

52. Catalano GAPI, Brownlee A, Cairns D, Ainslie R and McCall J (2025). Interpretable decision trees to predict solution fitness. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1099–1107

53. Johnston KE, Almhjell PJ, Watkins-Dulaney EJ, Liu G, Porter NJ, Yang J and Arnold FH (2024). A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences* **121(32)**:e2400439121

54. Arthur R and Sibani P (2017). Decision making on fitness landscapes. *Physica A: Statistical Mechanics and its Applications* **471**:696–704

55. Fu G, Dai X, Symanzik J and Bushman S (2017). Quantitative gene–gene and gene–environment mapping for leaf shape variation using tree-based models. *New Phytologist* **213(1)**:455–469

56. Nelson RM, Kierczak M and Carlborg Ö (2013). Higher order interactions: detection of epistasis using machine learning and evolutionary computation. In *Genome-Wide Association Studies and Genomic Prediction*, pp. 499–518. Springer

57. Poelwijk FJ, Socolich M and Ranganathan R (2019). Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications* **10(1)**:4213

58. Guerrero RF, Scarpino SV, Rodrigues JV, Hartl DL and Ogbunugafor CB (2019). Proteostasis environment shapes higher-order epistasis operating on antibiotic resistance. *Genetics* **212(2)**:565–575

59. Khan AI, Dinh DM, Schneider D, Lenski RE and Cooper TF (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332(6034)**:1193–1196

60. O'maille PE, Malone A, Dellas N, Andes Hess Jr B, Smentek L, Sheehan I, Greenhagen BT, Chappell J, Manning G et al. (2008). Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature chemical biology* **4(10)**:617–623

61. Wagner GP (2015). Two rules for the detection and quantification of epistasis and other interaction effects. *Methods in molecular biology* **1253**:145–157

62. Christensen KA, Le Luyer J, Chan MT, Rondeau EB, Koop BF, Bernatchez L and Devlin RH (2021). Assessing the effects of genotype-by-environment interaction on epigenetic, transcriptomic, and phenotypic response in a Pacific salmon. *G3* **11(2)**:jkab021

63. Hansen TF (2023). Variation, inheritance, and evolution: A primer on evolutionary quantitative genetics. In Hansen TF, Houle D, Pavlicev M and Pélabon C (eds.), *Evolvability: A unifying concept in evolutionary biology?*, chap. 5, pp. 73–100. MIT Press

64. Oomen RA and Hutchings JA (2020). *Evolution of Reaction Norms.* Oxford University Press, Oxford, England

65. Schlichting CD and Pigliucci M (1998). *Phenotypic evolution: a reaction norm perspective.* Sinauer Associates Inc.

66. Sae-Lim P, Gjerde B, Nielsen HM, Mulder H and Kause A (2016). A review of genotype-by-environment interaction and micro-environmental sensitivity in aquaculture species. *Reviews in Aquaculture* **8(4)**:369–393

67. Wright S (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* **1**:356–366

68. Gavrilets S (2004). *Fitness landscapes and the origin of species*, vol. 41. Princeton university press

69. McCandlish DM (2011). Visualizing fitness landscapes. *Evolution* **65(6)**:1544–1558

70. Doro S and Herman MA (2022). On the Fourier transform of a quantitative trait: Implications for compressive sensing. *Journal of Theoretical Biology* **540**:110985

71. Mustonen V and Lässig M (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in genetics* **25(3)**:111–119

72. King ES, Stacy AE, Weaver DT, Maltas J, Barker-Clarke R, Dolson E and Scott JG (2025). Fitness seascapes are necessary for realistic modeling of the evolutionary response to drug therapy. *Science Advances* **11(24)**:eadv1268

73. Iram S, Dolson E, Chiel J, Pelesko J, Krishnan N, Güngör Ö, Kuznets-Speck B, Deffner S, Ilker E et al. (2021). Controlling the speed and trajectory of evolution with counterdiabatic driving. *Nature Physics* **17(1)**:135–142

74. Ogbunu CB (2023). The New Quest to Control Evolution. Quanta Magazine, Quantized Columns

75. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C and Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42(4)**:348–354

76. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI and Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nature methods* **8(10)**:833–835

77. Zhou X and Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44(7)**:821–824

78. Jahangiri M, Kazemnejad A, Goldfeld KS, Daneshpour MS, Momen M, Mostafaei S, Khalili D and Akbarzadeh M (2025). Leveraging mixed-effects regression trees for the analysis of high-dimensional longitudinal data to identify the low and high-risk subgroups: simulation study with application to genetic study. *BioData Mining* **18(1)**:22

79. Kuhn L, Page K, Ward J and Worrall-Carter L (2014). The process and utility of classification and regression tree methodology in nursing research. *Journal of Advanced Nursing* **70(6)**:1276–1286

80. Hastie T, Tibshirani R, Friedman J et al. (2009). The elements of statistical learning

81. Strobl C, Boulesteix AL, Zeileis A and Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* **8(1)**:25

82. Hothorn T, Hornik K and Zeileis A (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* **15(3)**:651–674

83. Friedman JH (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics* pp. 1189–1232

84. Wozniak M (2011). A hybrid decision tree training method using data streams. *Knowledge and Information Systems* **29(2)**:335–347

85. Sigrist F (2021). KTBoost: Combined kernel and tree boosting. *Neural Processing Letters* **53(2)**:1147–1160

86. Torgo L (1997). Kernel regression trees. In *Poster papers of the 9th European conference on machine learning (ECML 97)*, pp. 118–127. Prague, Czech Republic

87. Lewis RJ (2000). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, vol. 14. Citeseer

88. Bel L, Allard D, Laurent JM, Cheddadi R and Bar-Hen A (2009). CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Analysis* **53(8)**:3082–3093

89. De'ath G and Fabricius KE (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81(11)**:3178–3192

90. Timofeev R (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin* **54**:48

91. Komatsu KJ, Avolio ML, Padullés Cubino J, Schrodt F, Auge H, Cavender-Bares J, Clark AT, Flores-Moreno H, Grman E et al. (2024). CoRRE Trait Data: A dataset of 17 categorical and continuous traits for 4079 grassland species worldwide. *Scientific Data* **11(1)**:795

92. Li J, Pattaradilokrat S, Zhu F, Jiang H, Liu S, Hong L, Fu Y, Koo L, Xu W et al. (2011). Linkage maps from multiple genetic crosses and loci linked to growth-related virulent phenotype in Plasmodium yoelii. *Proceedings of the National Academy of Sciences* **108(31)**:E374–E382

93. Loh WY (2014). Fifty years of classification and regression trees. *International Statistical Review* **82(3)**:329–348

94. Therneau TM and Atkinson B (2025). *`rpart`: Recursive Partitioning and Regression Trees*. `R` package version 4.1.24

95. Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58(1)**:267–288

96. Friedman JH, Hastie T and Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33(1)**:1–22

97. Huang Y, Tibbe T, Tang A and Montoya A (2023). Lasso and group lasso with categorical predictors: Impact of coding strategy on variable selection and prediction. *Journal of Behavioral Data Science* **3(2)**:15–42

Supplementary Material

Practical Workflow for CART Modeling with `rpart` in `R`

The `rpart` package (*94*) in R offers a flexible and widely used implementation of the CART algorithm for both classification and regression tasks. This section demonstrates how to import real data and fit either a classification tree (for categorical response variables) or a regression tree (for continuous response variables), even when predictor variables include a mix of categorical and continuous types.

Step 1: Installing and Loading Required Packages

R Code Example

```r
# Install packages (run only once if not installed)
install.packages("rpart")         # CART model
install.packages("readr")         # For CSV files
install.packages("readxl")        # For Excel files

# Load libraries
library(rpart)
library(readr)
library(readxl)
```

Step 2: Importing Data

R Code Example

```r
# Load from CSV
df <- read_csv("yourfile.csv")

# OR load from Excel
# df <- read_excel("yourfile.xlsx")
```

Step 3: Inspect and Prepare the Data

R Code Example

```r
# View structure of the dataset
str(df)

# Convert relevant predictors to factors if categorical
df$soil_type <- as.factor(df$soil_type)
df$genotype  <- as.factor(df$genotype)

# Convert trait to factor for classification, keep numeric for regression
# For classification example:
# df$trait <- as.factor(df$trait)
```

Step 4: Fitting a CART Model

(a) Regression Tree (Continuous Trait)

R Code Example

```r
tree_reg <- rpart(trait ~ ., data = df, method = "anova")
rpart.plot(tree_reg, type = 3, extra = 101)
```

(b) Classification Tree (Categorical Trait)

R Code Example

```r
tree_class <- rpart(trait ~ ., data = df, method = "class")
rpart.plot(tree_class, type = 3, extra = 104)
```

Notes:

- Use `method = "anova"` for regression trees (continuous response).
- Use `method = "class"` for classification trees (categorical response).
- The function `rpart()` automatically handles both categorical and numeric predictor variables.
- The resulting decision tree model can be effectively visualized using the plotting functions provided by the `rpart.plot` package. See Figure S1 for an illustration of the internal and terminal node structure in the resulting plot.

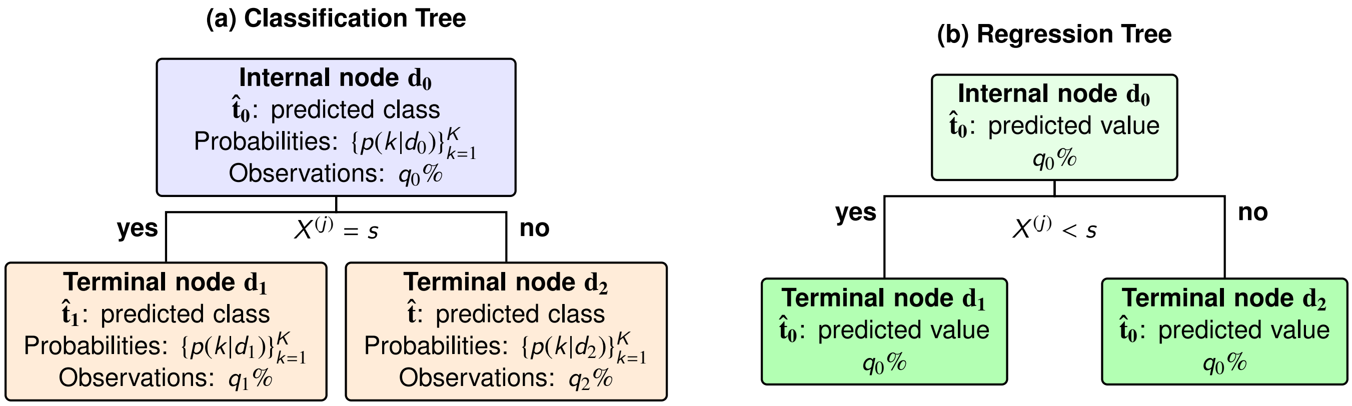
(a) Classification Tree
Internal node $\mathbf{d_0}$
$\hat{\mathbf{t}}_0$: predicted class
Probabilities: $\{p(k|d_0)\}_{k=1}^{K}$
Observations: $q_0\%$
yes
$X^{(j)} = s$
no
Terminal node $\mathbf{d_1}$
$\hat{\mathbf{t}}_1$: predicted class
Probabilities: $\{p(k|d_1)\}_{k=1}^{K}$
Observations: $q_1\%$
Terminal node $\mathbf{d_2}$
$\hat{\mathbf{t}}$: predicted class
Probabilities: $\{p(k|d_2)\}_{k=1}^{K}$
Observations: $q_2\%$
(b) Regression Tree
Internal node $\mathbf{d_0}$
$\hat{\mathbf{t}}_0$: predicted value
$q_0\%$
yes
$X^{(j)} < s$
no
Terminal node $\mathbf{d_1}$
$\hat{\mathbf{t}}_0$: predicted value
$q_0\%$
Terminal node $\mathbf{d_2}$
$\hat{\mathbf{t}}_0$: predicted value
$q_0\%$

Fig. S1. Typical internal and terminal node structures in a classification tree (a) and a regression tree (b), following the style of `rpart.plot`. Each node contains the predicted output: the class label $\hat{t}_i$ and class probabilities $\{p(k|d_i)\}_{k=1}^{K}$ for classification trees, or the predicted numeric value $\hat{t}_i$ for regression trees. The percentage of observations $q_i\%$ passing through each node is also displayed. Internal nodes include a split criterion shown directly below the node box (e.g., $X^{(j)} = s$ or $X^{(j)} < s$), and branches are labeled as **yes** (left) and **no** (right) to indicate the outcome of the split condition.

Detailed tree structure and epistatic analysis of the fluorescent protein landscape in *Entacmaea quadricolor*

For completeness, we provide detailed interpretation of the CART decision tree trained to classify all three fluorescence phenotypes (*red*, *blue*, and *black*). Recall that each genotype is encoded as a 13-bit binary vector representing specific amino acid substitutions on the mTagBFP2 backbone, with the following locus-to-residue mapping:

Locus 1 → K231R, Locus 2 → N207K, Locus 3 → N206D, Locus 4 → Y197R,
Locus 5 → A174L, Locus 6 → A172C, Locus 7 → S168G, Locus 8 → N158A,
Locus 9 → F143S, Locus 10 → T127P, Locus 11 → L63M, Locus 12 → V45A,
Locus 13 → D20N.

In Figure 2(a), the leftmost subtree highlights how a mutation at locus 4 (Y197) increases the likelihood of red fluorescence, but only in combination with mutations at loci 9 (F143) and 11 (L63). The probabilities at internal nodes indicate partial class transitions, suggesting necessary but insufficient roles of single mutations. The rightmost terminal node includes genotypes with no mutations at either locus 4 or 9, and is overwhelmingly classified as blue (92%) with no red signal, indicating their retention of the original mTagBFP2 phenotype. An exceptional case shows red fluorescence emerging without locus 4 mutation, but only when loci 9, 11, and 5 are simultaneously mutated—again underscoring the necessity of coordinated changes. We further analyze these patterns using a mutational interaction network (see Figure S2), which confirms that no genotype with fewer than three mutations exhibits red fluorescence. Among triple mutants, only specific combinations such as {4, 9, 11} or {5, 9, 11} produce red phenotypes. Notably, the former combination consistently yields higher brightness scores, reflecting stronger synergistic epistasis (see Figure 2(c)). The variable importance scores (see Figure 2(b)) support these findings, ranking F143 (locus 9) highest, followed by Y197 (locus 4), L63 (locus 11), and A174 (locus 5). These residues thus represent key modulators of the color transition, acting within a structured hierarchy of interactions.

Together, these results demonstrate that third-order epistasis is required to generate red fluorescence in this system. Our tree-based approach enables precise mapping from genotypic configurations to phenotypic outcomes, offering a robust and interpretable model for understanding the combinatorial logic underlying spectral tuning in fluorescent proteins.

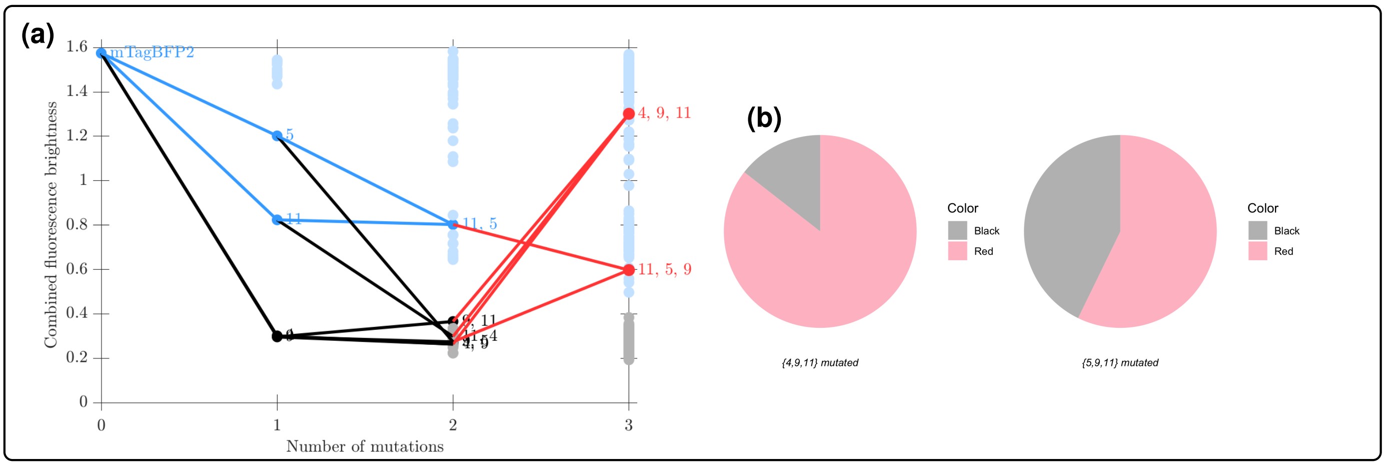

Fig. S2. Case 1, additional analyses: Mutations at specific loci in *Entacmaea quadricolor*(57) drive the transition from blue to red fluorescence through third-order epistatic interactions. **(a)** Visualization of early mutational trajectories from the wild-type mTagBFP2 sequence to red fluorescence. The x-axis represents mutational depth ($x = 0$ to 3), and the y-axis indicates the combined fluorescence trait value. Nodes represent genotypes and are colored by fluorescence class (*blue*, *black*, or *red*); edges indicate single mutational steps and are colored according to the color class of the target node. Only paths involving loci {4, 9, 11} and {5, 9, 11}—the first observed red mutants—are shown. Mutations at loci 4 (Y197) and 9 (F143) alone lead to lower trait values and transitions to the *black* class. The addition of a mutation at locus 11 (L63) results in a transition to the *red* class and a substantial increase in trait value, underscoring the role of third-order epistasis. **(b)** Color class distributions for genotypes with mutations at loci {4, 9, 11} and {5, 9, 11}, shown as pie charts. Only *black* and *red* classes are present, with a higher proportion of red fluorescence observed in the {4, 9, 11} group, supporting its stronger phenotypic effect.

Tree-Based Analyses of Context-Dependent Fitness and Chemotype Variation

This section presents CART-based analyses of two additional datasets not covered in the main text. These include:

- **Case 4: Proteostasis-Dependent Fitness Effects in Bacterial DHFR Mutants** — examining genotype-by-environment and genotype-by-species interactions across three bacterial species and proteostatic conditions.
- **Case 5: Regression Tree Analysis of Multivariate Chemotype Landscapes in Terpene Synthase Evolution** — investigating how amino acid substitutions influence sesquiterpene product profiles in plant terpene synthase enzymes.

These analyses extend our framework to biological systems where environmental, cellular, or biochemical context plays a crucial role in shaping phenotypic outcomes.

Case 4: Proteostasis-Dependent Fitness Effects in Bacterial DHFR Mutants

Comparing simple linear regression and regression trees: In Guerrero et al. (2019) (*58*), generalized linear models—specifically LASSO regression—were employed to estimate the effects of protein quality control, which confers different proteostasis environments, the state of maintaining a balance of properly folded and functional proteins in a cell. In this study, the "environments" are genotypes, corresponding to the presence of different protein quality control actors in the genome: wild type, GroEL$^+$, ΔIon). These were engineered into three species-specific DHFR backgrounds (*E. coli*, *L. grayi*, *C. muridarum*, with 3 point mutations at three loci (P21L, A26T, L28R) that confer different levels of growth in the presence of trimethoprim (an antbiotic) (IC_{50}) (see Figure S3 for a visual representation of the data) and DHFR abundance. This approach was applied separately within each Proteostasis-species combination, enabling rigorous estimation of main effects, pairwise interactions, and higher-order epistatic effects. The authors further incorporated nonlinear correction terms to account for saturation-like effects and ensure robust model fit. Nonetheless, this LASSO-based strategy still relies on pre-specified interaction terms within each context and may therefore overlook emergent nonlinear threshold effects or context-specific conditional interactions not explicitly included in the model.

Classification and regression trees (CART) offer a non-parametric alternative to linear modeling by recursively partitioning the data based on variables such as species background, proteostasis environment, or individual mutation states (e.g., presence or absence of P21L, A26T, or L28R). This method organizes the data into phenotypically homogeneous subsets, guided by the strongest explanatory variables at each node. Unlike parametric approaches that require explicit model specification, CART can uncover threshold-dependent phenotypic bifurcations, such as distinct IC_{50} response regimes observed when the species background is *L. grayi* and the proteostasis environment is wild type, without presuming interaction structures in advance.

Although the primary goal of decision trees is to uncover the structural relationships between variables, regression trees can also be used for estimating or predicting continuous response variables. To illustrate the differences between ordinary least squares (OLS) regression and regression tree predictions, we used a composite mutation code (defined as the base-10 representation of the binary presence/absence of P21L, A26T, and L28R) as the predictor and $\log_{10}$-transformed IC_{50} values as the response variable. Regression curves from both models, along with the data distribution for each mutation code, are shown in Figure S4. Notably, OLS regression is sensitive to outliers, whereas regression trees are more robust to such effects. Moreover, regression trees naturally accommodate categorical predictors and often yield more interpretable partitions of the predictor space than linear regression models. In contrast, linear regression may require variable transformation, careful model specification, and still lacks guaranteed interpretability in complex biological systems.

Regression tree analysis of the complete Guerrero et al. (2019) dataset: We applied CART to the full factorial design of Guerrero et al. (2019) (*58*), encompassing all combinations of proteostasis backgrounds, species backgrounds, and the eight possible combinations of point mutations (P21L, A26T, L28R). Similar to the IC_{50} analysis presented here, the same CART framework can be applied using DHFR protein abundance as the response variable, though we focus in this study on IC_{50} as a primary phenotype relevant to antibiotic resistance.

In contrast to previous LASSO-based analyses that emphasize additive and interaction effects, the regression tree in Figure S5 reveals a distinct hierarchy: the species background is identified as the primary splitting variable, followed by specific point mutations and the proteostasis context. This structure suggests that the species-specific genetic background exerts the strongest influence on resistance phenotypes, with the mutation identity and the proteostasis environment refining phenotypic differences within each species regime. Such nonlinear, context-dependent interactions are challenging to uncover using linear models unless all possible interaction terms are explicitly encoded, whereas CART automatically detects and partitions along these higher-order dependencies.

The structure of the regression tree, as shown in Figure S5, enables straightforward interpretation of phenotype partitions. For example, the left branch of the tree isolates data points from *L. grayi* and *C. muridarum*, which are generally associated with lower IC_{50} values compared to *E. coli*. Within this branch, the next level split is based on proteostasis context: wild-type proteostasis background tends to yield lower resistance levels, while the GroEL+ and ΔIon contexts lead to relatively higher resistance. In both species and across contexts, the presence of the L28R mutation further increases IC_{50}, consistent with its role as a large effect resistance

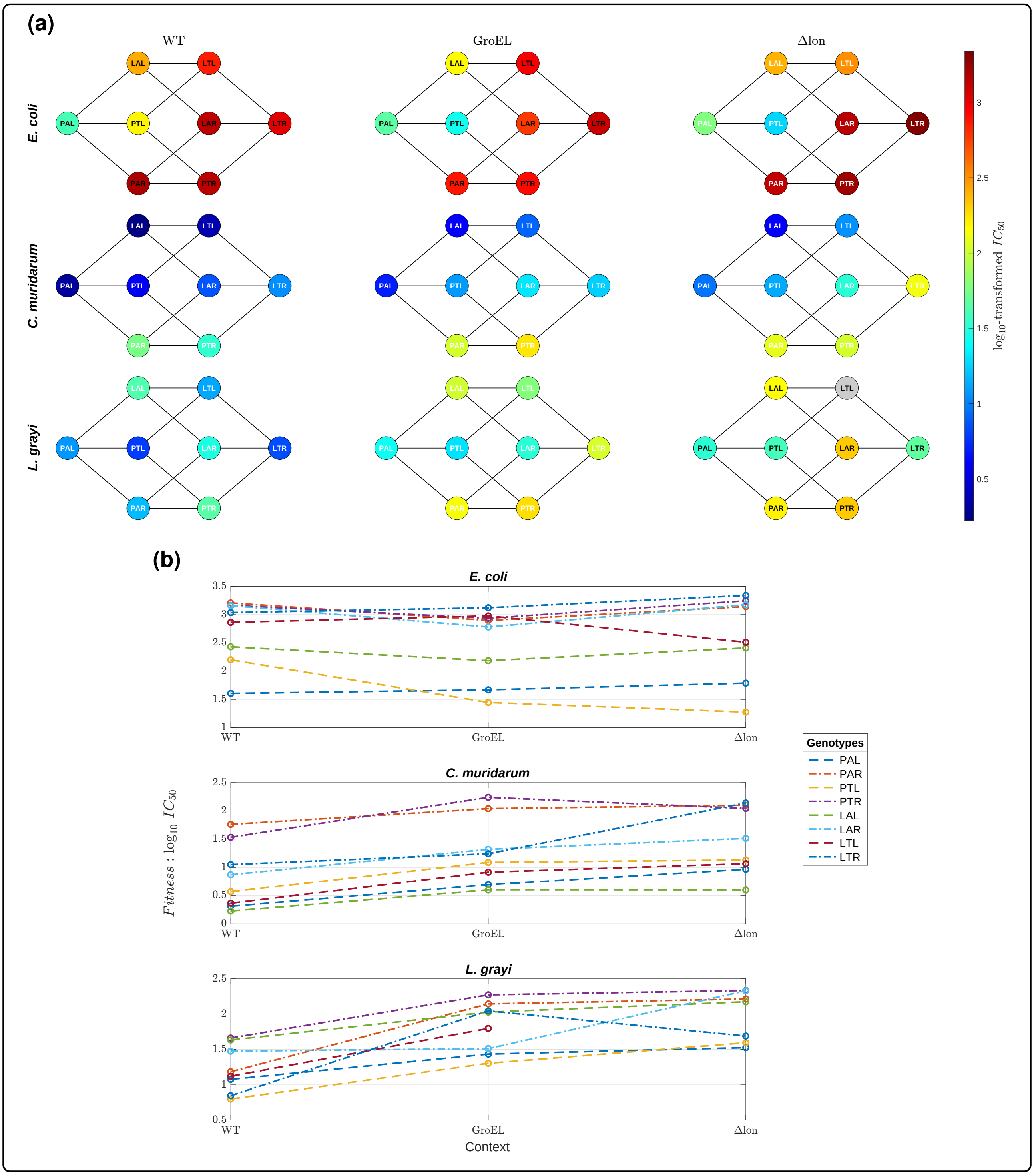

Fig. S3. Data for Case 4: Fitness graphs and reaction norms across species and proteostasis environments. Visualizing the genotype–phenotype relationships from Guerrero et al. (2019) (58) in two complementary ways. **(a)** Genotype networks (hypercubes) for each combination of species (*Escherichia coli*, *Legionella grayi*, and *Chlamydia muridarum*) and proteostasis environment (wild type (WT), GroEL overexpression (GroEL$^+$), and lon deletion (Δlon)). Each 8-node cube represents all possible combinations of three DHFR point mutations (P21L, A26T, L28R); edges connect single–mutation steps, and node color intensity encodes antibiotic resistance measured as $\log_{10}$-transformed IC_{50}. **(b)** Reaction norms for each species, plotting the $\log_{10}$-transformed IC_{50} of every genotype across the three proteostasis contexts. Lines correspond to individual genotypes (mutational backgrounds), illustrating genotype-by-environment interactions in antibiotic resistance.

mutation. Conversely, the right-hand branch of the tree captures the *E. coli* background, which overall exhibits higher IC_{50} values. Within *E. coli*, proteostasis context exerts minimal influence on resistance levels. Instead, the L28R mutation again plays a dominant role in increasing resistance, followed by P21L. This breakdown highlights how CART naturally identifies both species-specific and context-dependent pathways to elevated resistance and can reveal biologically meaningful thresholds without requiring them to be specified in advance.

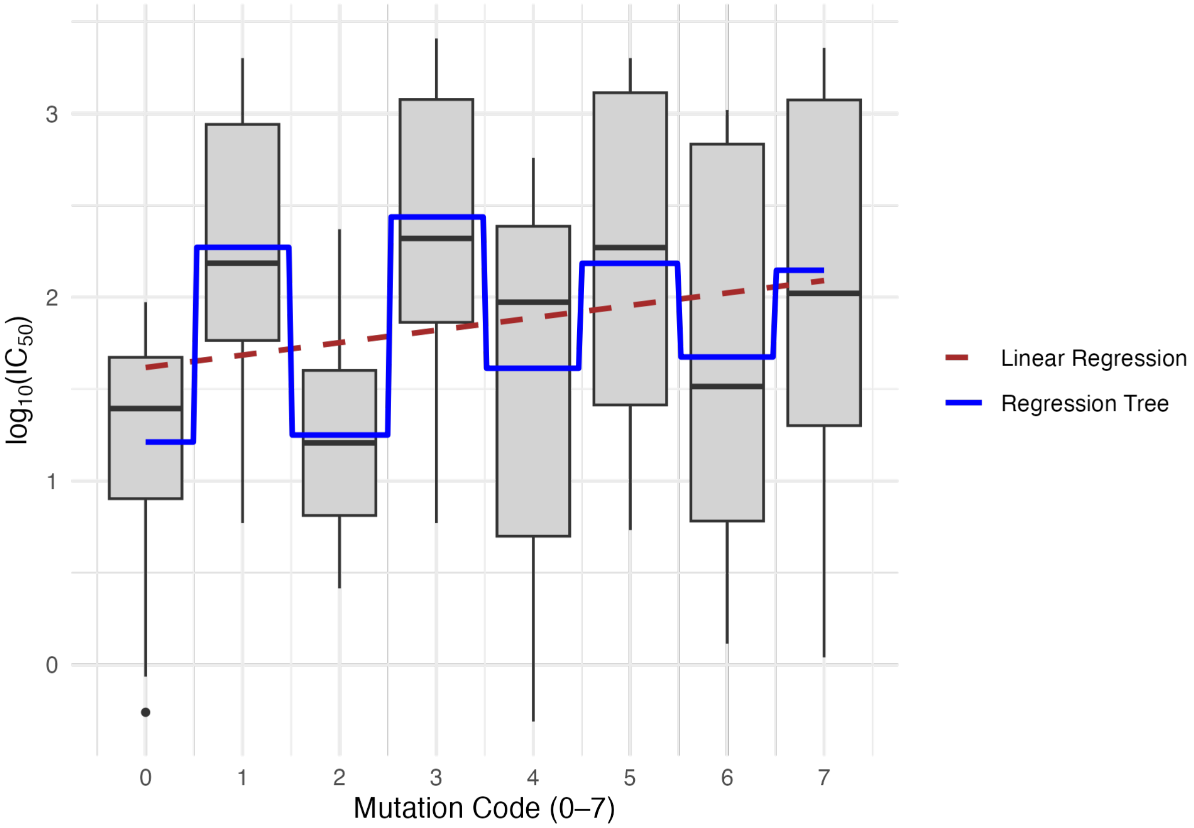

Fig. S4. Case 4, additional analyses. Boxplots of log-transformed IC_{50} values for each mutation code (0–7), defined by the binary presence or absence of mutations P21L, A26T, and L28R. Overlaid are regression curves from a linear model (dashed brown line) and a regression tree model (solid blue line). The regression tree model more flexibly captures threshold effects and nonlinear relationships between mutation combinations and phenotypic resistance. The dataset was originally described by Guerrero et al. (2019) (*58*).

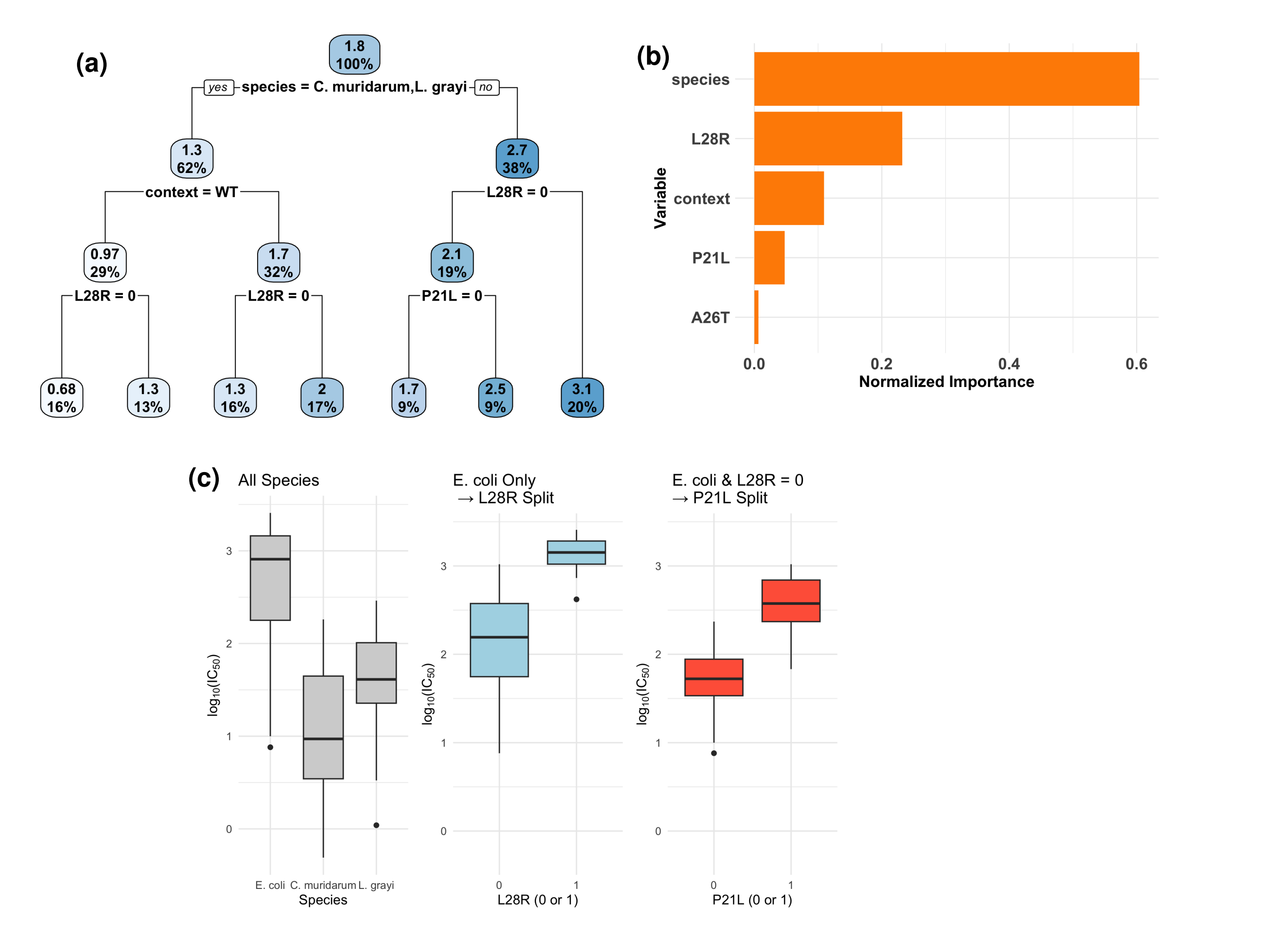

Fig. S5. Case 4 analysis: Decision tree structure and response distribution for DHFR phenotypes. (a) Regression tree fitted to log-transformed IC_{50} as a function of genotype (P21L, A26T, L28R), species background, and proteostasis context, using a maximum tree depth of 3. (b) Normalized variable importance scores from the fitted CART model, identifying *species* as the top predictor, followed by the *L28R* locus and proteostasis context. (c) Boxplot breakdown of response values (IC_{50}) along the right-hand branch of the decision tree. Similar illustrations can be constructed for alternative branches. This visualization supports the CART analysis by showing that among the three species, *E. coli* exhibits the highest IC_{50} values, with a generally elevated distribution compared to *L. grayi* and *C. muridarum*. Within the *E. coli* subset, the proteostasis context has minimal effect, whereas the presence of the L28R mutation is associated with the highest resistance, followed by the P21L mutation. The dataset was originally described by Guerrero et al. (2019) (*58*).

Case 5: Regression tree analysis of multivariate chemotype landscapes in terpene synthase evolution

The dataset (see Figure S6 for a visualization) analyzed originates from the comprehensive combinatorial mutagenesis study of sesquiterpene synthase enzymes described by O'Maille et al. (2008) (*60*). It consists of nine key amino acid substitutions within the enzyme active site region, but we focused on a subset of six amino acid substitutions that conferred functional genotypes: I372V (locus 1), S402T (locus 2), L406Y (locus 3), T438I (locus 4), L439I (locus 5), and I516V (locus 6). Each locus is encoded as a binary variable indicating the presence (mutated) or absence (wild-type) of the respective amino acid change. The response variables measure the relative enzymatic product distribution across four main sesquiterpene compounds: 5-epi-aristolochene (5-EA), 4-epi-eremophilene (4-EE), premnaspirodiene (PSD), and a combined category of minor products. These quantitative phenotypes reflect the catalytic specificity and promiscuity resulting from the combinatorial mutation landscape. Regression tree analysis is applied to elucidate how specific mutations and their interactions partition the genotype space into distinct phenotypic clusters, revealing threshold effects and epistatic relationships. This non-parametric method facilitates interpretable insights into the mutational determinants of terpene biosynthesis, complementing the original study's characterization of enzyme evolutionary trajectories and functional diversification.

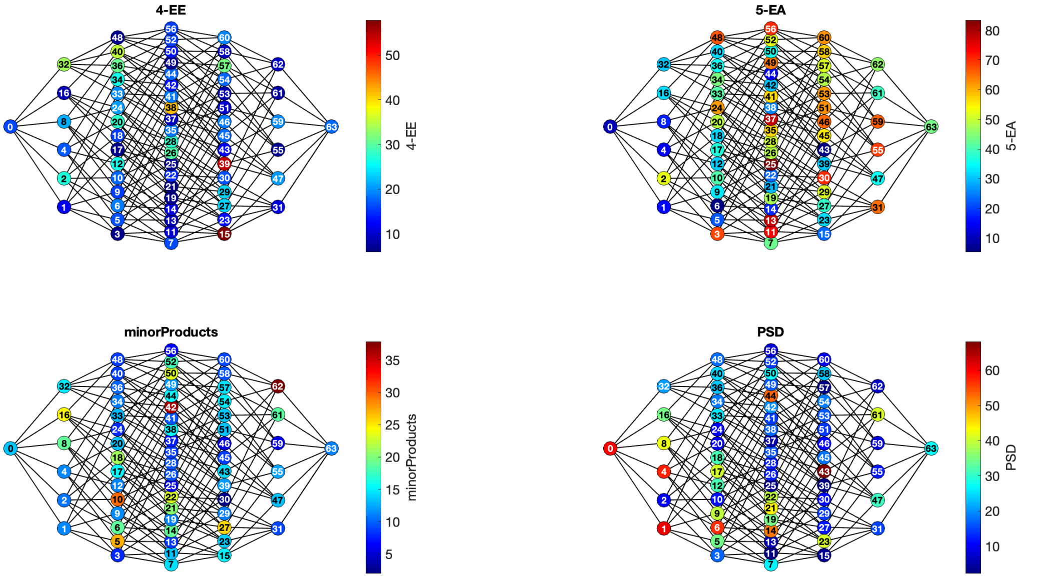

Fig. S6. Data for Case 5: Genotype-phenotype maps for sesquiterpene product specificity in O'Maille et al. (2008) (*60*). Each panel shows the 6-locus genotype hypercube for one quantitative trait—(a) 5-epi-aristolochene (5-EA), (b) 4-epi-eremophilene (4-EE), (c) combined minor products, and (d) premnaspirodiene (PSD) —measured in recombinant sesquiterpene synthases from *Nicotiana tabacum* and *Hyoscyamus muticus*. The 64 nodes correspond to all possible genotypes defined by binary combinations of six active-site mutations (totaling 2^6 genotypes), and are labeled from 0 to 63 for reference. Nodes are arranged horizontally by Hamming distance from the wild-type (node 0) and vertically for visualization clarity. Edges represent single amino acid substitutions between genotypes (Hamming distance = 1), forming a projection of the 6-dimensional hypercube. Node color reflects the normalized production level of the respective compound, according to the panel-specific colorbar. This visualization highlights the nonlinear genotype–phenotype landscape and epistatic effects shaping terpene biosynthesis and enzyme evolution. Note that we focused on a subset of 6 loci from the initial 9-locus dataset from O'Maille et al. (2008) as several genotypes were non-functional in the full dataset.

To further dissect genotype–phenotype mappings, we applied CART individually to each of the four response variables, and the resulting trees are summarized in Figure S7. Each decision tree yields a hierarchical structure of mutation effects, highlighting distinct mutational paths to increased product formation. For **4-EE**, the most informative loci were S402T (locus 2), L406Y (locus 3), and I372V (locus 1), with the highest product levels observed when locus 2 is wild-type and locus 1 is mutated. In the case of **5-EA**, the top predictors were loci 2, 6, and 3, and the highest levels were associated with mutations at both loci 2 and 3. For **minor products**, the most important loci were I516V (locus 6), L406Y (locus 3), and L439I (locus 5); combinations where locus 5 is mutated and locus 6 remains wild-type led to the highest product levels. Lastly, for **PSD**, the decision tree prioritized L406Y (locus 3), S402T (locus 2), and L439I (locus 5), with the combination of wild-type alleles at loci 2 and 5 yielding maximal expression.

The normalized variable importance scores reinforce these observations by quantifying the contribution of each locus to the phenotype-specific trees (see Figure S7(b)). Collectively, these CART models expose the heterogeneous and context-dependent roles of specific amino acid substitutions in shaping enzyme product profiles. The ability of regression trees to handle high-order interactions and provide intuitive branching rules makes them well-suited for uncovering complex biochemical epistasis in enzyme evolution.

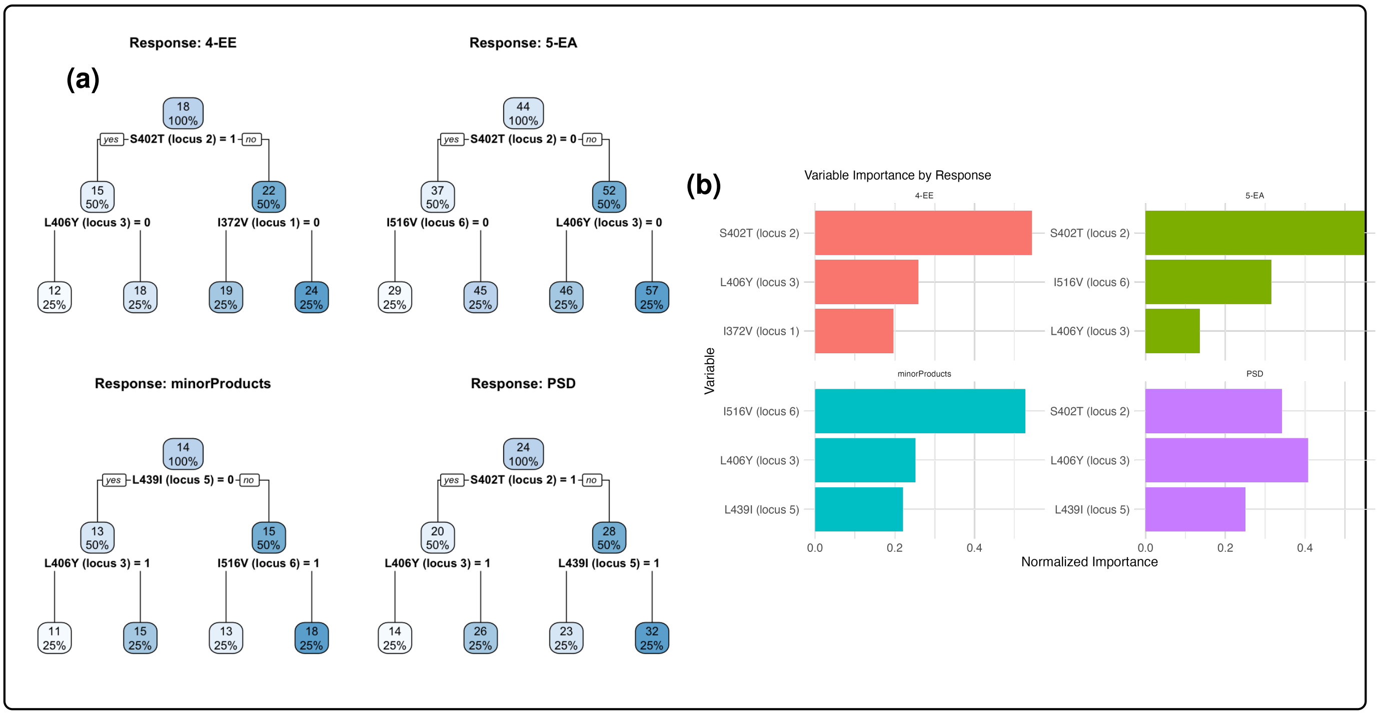

Fig. S7. Case 5 analysis: Regression tree structure and variable importance across four metabolic phenotypes. (a) Regression trees fitted separately to predict each of the four response variables—4-EE, 5-EA, minor products, and PSD—based on the presence or absence of mutations at six loci. Each tree shows hierarchical splits associated with increased or decreased response values. (b) Normalized variable importance scores from the CART models. For each response, the top three contributing loci are: locus 2, 3, and 1 for 4-EE; locus 2, 6, and 3 for 5-EA; locus 6, 3, and 5 for minor products; and locus 3, 2, and 5 for PSD. The dataset was originally described by O'Maille et al. (2008) (*60*).

Supplementary analysis: comparison of CART and LASSO interpretability

To complement the CART-based analyses presented in the main text, we provide a parallel comparison using penalized linear regression with the Least Absolute Shrinkage and Selection Operator (LASSO). Whereas CART identifies predictive variables and their hierarchical interactions via tree structures, LASSO performs feature selection by shrinking regression coefficients toward zero, retaining only a sparse subset of predictors.

This comparison permits an assessment of interpretability for models of the genotype-phenotype map from two complementary perspectives: (i) tree-based paths that represent explicit conditional rules (CART), and (ii) sparse linear models that quantify additive effects of selected predictors (LASSO). For each empirical dataset summarized in Table 3, we fit a LASSO model to the continuous response variable (when available), using standardized predictors. We report the ten largest non-zero regression coefficients (ranked by absolute magnitude) as a concise measure of variable importance for LASSO.

LASSO regression: method summary

To complement the tree-based analyses, we employ penalized linear regression using the least absolute shrinkage and selection operator (LASSO) (*95*). In the general regression framework, the response variable is modeled as

$$y = X\beta + \varepsilon,$$

where $y \in \mathbb{R}^n$ denotes the response vector (phenotype values in our case), $X \in \mathbb{R}^{n \times p}$ is the design matrix whose columns correspond to predictor variables (loci, environmental conditions, and higher-order multiplicative interactions such as Locus$_1$× Locus$_2$, Locus$_1$× Locus$_2$× Locus$_3$, etc.), $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, and ε is the error term.

The LASSO estimator $\hat{\beta}$ is obtained by solving the optimization problem

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

where $\|y - X\beta\|_2^2$ is the residual sum of squares, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1-norm penalty, and $\lambda \geq 0$ is a tuning parameter that controls the degree of regularization. By shrinking some coefficients exactly to zero, LASSO simultaneously performs variable selection and regularization, yielding a sparse and interpretable model of the genotype-phenotype map. Nonzero entries of $\hat{\beta}$ indicate predictors (or explicitly encoded interactions) retained by the model. The magnitude $|\hat{\beta}_j|$ (or the standardized coefficient) serves as a proxy for effect size and a relative measure of variable importance. Unlike CART, which can capture nonlinearities and interactions implicitly through tree splits, LASSO is linear in the supplied features; thus, nonlinear and interaction terms must be explicitly included as columns of X (*95*,*96*). In the Supplement, we compare (i) the top nonzero LASSO coefficients (ranked by absolute magnitude) with (ii) the variables and split-paths identified by CART, to evaluate agreement and differences in interpretability across methods.

Case 1: Sparse linear modeling of a fluorescent protein fitness landscape in *Entacmaea quadricolor* (LASSO results)

In this section, we analyze the FP611 fluorescent-protein fitness landscape of *Entacmaea quadricolor* (*57*), using a combined continuous scalar fluorescence phenotype as the response. The dataset represents an exhaustive 13 site mutational library bridging two fluorescent proteins, with predictors encoded as binary indicators of substitutions at each site. In addition, all possible higher-order interactions among the 13 loci are explicitly included as predictors, encompassing main effects, pairwise interactions, three-way interactions, and up to the 5-way combination. LASSO regression was employed, and the regularization parameter was chosen via cross-validation. We report the ten largest nonzero coefficients (ranked by absolute value), providing a compact representation of the most influential substitutions and interactions. These results are compared with CART based interpretations (Section 2.1) to assess areas of concordance and divergence between sparse linear and tree-based models of the genotype-phenotype map.

Table 4. Case 1: Top ten nonzero LASSO coefficients for FP611 data. Coefficients are reported at the cross-validated minimizer $\lambda_{\min}$. Terms include main effects and interaction terms (up to 5-way).

Variable (term)	**Coefficient** (β)
Locus 4:Locus 9:Locus 11:Locus 12	-0.7762
Locus 4:Locus_9	0.7453
Locus 10	0.6751
Locus 1	0.6407
Locus 5:Locus 12	-0.6405
Locus 8	0.6260
Locus 7	0.6085
Locus 3	0.6009
Locus 13	0.5963
Locus 2	0.5816

The model was fit using cross-validated LASSO (`glmnet`(*96*) package in `R`) with intercept excluded, predictors standardized by default, and coefficients reported at $\lambda_{\min}$. Standardization ensures that the magnitude of each coefficient reflects the relative contribution of the corresponding predictor or interaction term to the phenotype.

LASSO results and comparison with CART Table 4 reports the top ten nonzero LASSO terms (ranked by absolute coefficient). The largest-magnitude term is a negative four-way interaction, Locus 4, 9, 11, and 12 (coef = -0.776), indicating that the joint presence of these four substitutions is associated with a decrease in standardized fluorescence. A strong positive pairwise term, `Locus 4:Locus 9` (coef = 0.745), and several positive main effects (e.g. Locus 10, Locus 1) were also selected, indicating that both localized substitutions and specific combinations contribute to variation in fluorescence.

These LASSO findings are broadly concordant with the CART results presented in the main text: CART identifies loci 9 (F143), 4 (Y197), 11 (L63), 5 (A174) and 12 (V45) as important decision variables (see Figure 2b). In particular, the prominence of loci 4, 9, 11 and 12 in the LASSO top terms accords with the tree-based observation that coordinated changes at these positions are required to produce red fluorescence (see Section 2.1 and Figure 2a).

Limitations and modeling choices Because LASSO is linear in the supplied features, interaction and nonlinear terms must be explicitly encoded in the design matrix. This substantially increases the dimensionality of X and the associated computational burden. For tractability we limited candidate interactions to orders up to five (i.e. main effects, pairwise, three, four and five way interactions). Despite this restriction, LASSO recovered biologically meaningful higher-order terms that overlap with CART identified loci. However, CART provides a complementary, structural representation (split paths and conditional rules) that more directly exposes hierarchical and context-dependent interactions. Unlike regularized linear models, trees implicitly capture nonlinearities and interactions of arbitrary order without any explicit feature engineering, and they readily identify threshold effects and decision boundaries that are often biologically interpretable. The tree topology naturally distinguishes global determinants from context-dependent modulators and yields concise, visual rules that domain experts can inspect and validate. Additional practical advantages include native handling of mixed predictor types and missing values, straightforward computation of variable importance measures, and direct support for visualization (tree plots) that facilitate mechanistic interpretation of epistatic architectures. We note, however, that single trees can be sensitive to sampling variation. To mitigate this we report pruned trees and cross-validated splits, and where appropriate complement single-tree interpretation with ensemble summaries (e.g. random forest or boosted tree variable importance) while preserving the single-tree depiction for interpretability.

Case 2: LASSO analysis of genotype–environment interactions in *P. falciparum* DHFR and cycloguanil

This section reports LASSO results for the DHFR dataset of *Plasmodium falciparum* under cycloguanil exposure (*12*). The decision tree in the main text, Section 2.2.1 and Figure 4, shows that drug concentration explains most model variance, while the contributions of individual mutations change with drug level. To provide a complementary sparse linear view, we fitted LASSO models to the response $t = \ln(r + 1)$ using the same predictor set, that is the four binary loci and the cycloguanil drug concentration. We also included all possible interaction terms among loci and the environment. Models were fitted with `glmnet` (*96*) package with $\alpha = 1$ (for LASSO). Tuning of λ used cross validation. Predictors were standardized by the software prior to fitting and coefficients are reported at $\lambda_{\min}$.

Table 5. Top LASSO coefficients for the DHFR cycloguanil dataset, Case 2. Coefficients are reported at $\lambda_{\min}$. Small values are shown in scientific notation.

Variable (term)	Coefficient(β)
Locus 3	0.404861
Locus 2	0.355494
Locus 4	0.229218
Locus 1	0.124208
Locus 3:Env	-2.17×10^{-7}
Locus 2:Env	-1.51×10^{-7}
Env	-1.47×10^{-7}
Locus4:Env	-1.43×10^{-7}

Interpretation of LASSO and comparison with CART Table 5 shows that LASSO selects the four DHFR loci as positive main effects, with Locus 3 (C59R) and Locus 2 (S108N) the largest contributors in the sparse linear model. This agrees with the CART analysis, where drug concentration is the dominant predictor and

Locus 2 (S108N) and Locus 3 (C59R) have the largest mutational importance in their respective drug regimes, Section 2.2.1 and Figure 4(c). Both methods therefore single out the same residues as key determinants of fitness, while they emphasize different aspects of the genotype by environment relationship.

The environmental term and the locus by environment interaction terms have coefficients close to zero in the fitted LASSO model. There are two main explanations. First, the environmental effect is largely threshold-like, with a sharp split near $\log_{10} E \approx 2$ (here E represent the dru concentration), and a simple linear term will be small if the true relationship is non-linear or piecewise. Second, after standardization and penalization, marginal linear contributions of Env (environmental term) or first-order interactions may be small relative to the main genetic effects, especially when the tree partitions the data into regimes where different mutations dominate. Thus LASSO highlights loci with consistent additive influence across the sampled concentrations, while CART exposes the explicit decision boundaries and the changes in mutational effects that occur across drug regimes.

Notes on modeling choices For this four-locus design the number of candidate interactions is manageable, and LASSO produced a concise set of nonzero coefficients that align with the tree-derived biology. Still, LASSO requires explicit encoding of interactions, which increases the dimension of the design matrix and the computational burden. For systems with strong non-linear or threshold effects, tree-based models often give a more direct and interpretable depiction of conditional effects.

Case 3: LASSO analysis of environmental modulation of epistasis in *E. coli* (LTEE)

This subsection reports LASSO results for the LTEE data set in *E. coli*, which contains all 32 genotype combinations across five loci and measurements in three environments, namely DM25, DM25 plus EGTA, and DM25 plus guanazole (*11*,*59*). The CART analysis in the main text shows that environment is the primary predictor and that *glmUS* and *topA* are the most influential loci, see Fig. 5 and Fig. 6. To provide a sparse linear view we fitted LASSO models to the log transformed fitness response using the same predictor set.

In this data set, the environmental variable is categorical. Standard linear regression requires numerical predictors, so categorical variables must be represented by indicator variables (*97*). We therefore encoded each environment level as dummy indicators and included these columns in the feature design matrix X. In addition, we incorporated all possible interaction terms among the five loci and the environment. Thus, the design matrix X contains both the main effects and all possible interactions of these variables as candidate predictors. This specification allows the linear model to consider every multiplicative combination of loci and environment while leaving variable selection to the penalized estimator. Predictors were standardized by the software prior to fitting. To avoid perfect multicollinearity, the design matrix excluded a redundant reference column. Models were fitted using `cv.glmnet` with LASSO ($\alpha = 1$). The tuning parameter λ was chosen by cross-validation, and coefficients are reported at the cross-validated minimizer $\lambda_{\min}$.

Table 6. Top LASSO coefficients for Case 3, LTEE. Coefficients are reported at $\lambda_{\min}$.

Variable (term)	Coefficient (β)
rbs(Locus 1):topA(Locus 2):spoA(Locus 3):glmUS(Locus 4):pykF(Locus 5):DM25_EGTA	0.042485045
glmUS(Locus 4)	0.030650447
topA(Locus 2)	0.018480689
spoA(Locus 3)	0.018331629
rbs(Locus 1):DM25_EGTA	0.017244808
spoA(Locus 3):DM25_GUA	-0.013966596
pykF(Locus 5)	0.012830050
rbs(Locus 1)	0.010588701
rbs(Locus 1):glmUS (Locus 4):DM25_EGTA	0.008421076
glmUS(Locus 4):pykF (Locus 5):DM25_EGTA	0.007671273

Interpretation of LASSO results and comparison with CART The LASSO model identifies positive main effects for *glmUS* and *topA* (see Table 6). It also selects interaction terms that involve the “DM25 plus EGTA” environment. The top coefficient is a six way interaction that includes all five loci together with the “DM25 plus EGTA” environment, showing a modest positive association with log fitness in that environment. The single locus results for *glmUS* and *topA* agree with the CART findings, where *glmUS* and *topA* are among the most important predictors, see Figure 6. Both methods therefore single out the same residues as primary determinants of fitness, although they present different views of how those residues operate across environments.

Some locus by environment interaction coefficients are small in magnitude. This may reflect the combined effect of standardization and penalization in LASSO, or it may indicate that the environmental influence takes a threshold form that is represented more directly by tree splits. Trees capture abrupt changes and partition the data into regimes where mutation effects change, while LASSO quantifies additive and explicitly encoded interaction effects across the entire sample. For this data set the two approaches are complementary, and each highlights different aspects of genotype by environment relationships.

Notes on modeling choices Including all possible multiplicative interactions rapidly increases the number of candidate predictors. For the five locus design the dimension remains manageable, but computation and memory use grow with interaction order. LASSO produced a concise set of nonzero coefficients that overlap with loci identified by CART. Trees provide a clear depiction of threshold effects and of how splits partition the genotype by environment space, which aids interpretation of regime shifts in mutation effects.

Case 4: LASSO results reported in the original study and comparison with CART

We do not conduct a separate LASSO analysis for this case, as Guerrero et al. (2019) (*58*) provide a comprehensive investigation of the same data set. The data consist of three biallelic DHFR sites, a categorical proteostasis variable with three levels, and a categorical species variable with three levels. In practice, LASSO can be applied to such data by encoding categorical variables as dummy indicators and by incorporating explicit interaction terms in the design matrix, as demonstrated in our Case 3 analysis in the previous section.

Interpretation of LASSO results and comparison with CART Guerrero et al. (*58*) demonstrate through their LASSO analysis that species-specific amino acid background is the dominant driver of the focal trait (IC_{50}). In the full LASSO fits the *C. muridarum* and *L. grayi* backgrounds have the largest negative effects, with estimated effect sizes near -1.44 and -0.90 respectively. The single largest positive main effect is the L28R mutation, with effect size about 1.22. These results indicate that species background and high-effect mutations together shape the trait distribution in a predictable way, and that biochemical properties of the enzyme, such as catalytic efficiency and thermostability, provide a mechanistic basis for these findings. Guerrero et al. (*58*) further show that the total contribution of higher-order interactions exceeds that of main effects. Illustrative examples are a negative interaction between species background and L28R with effect size near -0.69 and a positive interaction between a species background and GroEL overexpression with effect size about 0.41. Some interactions admit straightforward mechanistic explanations while others remain difficult to interpret and therefore suggest directions for future work. Our CART analysis of the same data set identifies species background as the principal partitioning variable and shows how proteostasis environment and specific mutations refine phenotype differences within species regimes. Taken together, the LASSO results reported by Guerrero et al. and the CART results reported here offer complementary perspectives on species background, proteostasis environment, and mutation identity in shaping DHFR phenotypes (*58*).

Notes on modeling choices Guerrero et al. (*58*) fitted regularized linear models both within proteostasis-species strata and in pooled analyses that included explicit interaction terms. Categorical variables were represented by dummy indicators and interaction terms were included in the design matrix so that the penalized estimator could select relevant multiplicative combinations. The authors also incorporated nonlinear correction terms to account for saturation-like effects and to improve model fit. Because this data set has only three loci, the dimensionality of the design matrix remains modest even when higher-order interactions are included, but computation and memory demands grow with interaction order in larger designs. CART complements these linear fits by exposing threshold-like effects and by providing a visual, rule-based representation of how species background, proteostasis environment, and mutations jointly determine phenotypes.

Case 5: LASSO analysis of multivariate chemotype landscapes in terpene synthase evolution

This subsection reports LASSO results for the terpene synthase dataset from O'Maille et al. (2008) (*60*), which contains all 64 genotype combinations across six active-site loci and measurements of four sesquiterpene products: 5-epi-aristolochene (5-EA), 4-epi-eremophilene (4-EE), premnaspirodiene (PSD), and a combined category of minor products. The CART analysis in Section S4.2.3 shows that specific loci such as I372V (locus 1), S402T (locus 2), and L406Y (locus 3) are the most influential for each product, with hierarchical splits revealing threshold effects and epistatic interactions, see Figure S7. To provide a sparse linear perspective on genotype–phenotype relationships, we fitted LASSO models individually for each response variable using the same set of six loci as predictors. The design matrix X contains both the main effects and all possible interactions of these loci as candidate predictors, allowing the linear model to consider every multiplicative combination when selecting relevant terms.

Interpretation of LASSO and comparison with CART Table 7 shows that the LASSO models for the terpene synthase dataset identify strong main effects as well as several higher-order interactions among the six active-site loci. For 4-EE, the largest positive effects correspond to I372V (locus 1), T438I (locus 4), and L406Y (locus 3), while the strongest negative coefficients involve three-way interactions such as I372V:L406Y:L439I. For 5-EA, the dominant main effects are S402T (locus 2), I516V (locus 6), and I372V (locus 1), with several four- and five-way interaction terms carrying negative coefficients that temper these additive contributions. The model for minor products highlights L439I (locus 5), S402T (locus 2), and T438I (locus 4) as the largest positive effects, again accompanied by multi locus interaction terms of opposite sign. Finally, the PSD response stands out in that the largest coefficients correspond not to single loci but to four and five way interaction terms (e.g. I372V:S402T:L406Y:L439I:I516V), alongside strong main effects for T438I (locus 4) and I516V (locus 6). Taken together, these sparse linear models reveal that both main genetic effects and high-order epistasis shape the multivariate chemotype landscape.

The CART analysis in Section S4.2.3 and Figure S7 converges on a similar picture in terms of the most influential loci, but emphasizes context-dependent branching rules. For 4-EE, the tree highlights S402T (locus 2), L406Y (locus 3), and I372V (locus 1) as the most important predictors, in agreement with the strong LASSO coefficients for loci 1 and 3. For 5-EA, loci 2, 6, and 3 dominate both the regression tree and the linear model, while for minor products, both methods highlight contributions from loci 5 and 6. For PSD, CART prioritizes L406Y (locus 3), S402T (locus 2), and L439I (locus 5), partially overlapping with the LASSO findings that assign very large weights to high-order interaction terms involving these same loci. In summary, the two approaches provide complementary insights: LASSO quantifies additive and interaction contributions with sparse coefficients, exposing the importance of high-order epistasis in shaping terpene product specificity, while CART reveals explicit mutational pathways and threshold-like effects by partitioning the genotype space.

Notes on modeling choices With six loci the predictor space expands dramatically, making feature selection essential. While LASSO reduced this space to a sparse set of influential terms, CART offered complementary advantages. Trees provide a direct visualization of threshold effects and clearly delineate how specific substitutions or combinations redirect enzyme outcomes. This partitioning of genotype space highlights alternative evolutionary routes and makes explicit mutational effects. Such interpretability is especially valuable in the discovery of the complex epistatic structure underlying terpene synthase specificity.

Table 7. Non-zero LASSO coefficients for each response variable at the cross-validated minimizer $\lambda_{\min}$.

Response	Variable (term)	Coefficient (β)
4-EE	I372V (locus 1)	9.767
	T438I (locus 4)	9.606
	L406Y (locus 3)	8.929
	I372V (locus 1):L406Y (locus 3):L439I (locus 5)	-7.758
	L439I (locus 5)	7.719
	I372V (locus 1):L406Y (locus 3):T438I (locus 4)	-6.021
	S402T (locus 2):T438I (locus 4):L439I (locus 5)	-5.139
	T438I (locus 4):L439I (locus 5):I516V (locus 6)	3.256
	I372V (locus 1):S402T (locus 2):T438I (locus 4)	-2.088
	S402T (locus 2):T438I (locus 4):I516V (locus 6)	-1.995
5-EA	S402T (locus 2)	21.533
	I516V (locus 6)	19.478
	I372V (locus 1)	17.334
	L406Y (locus 3)	16.148
	I372V (locus 1):L406Y (locus 3):L439I (locus 5):I516V (locus 6)	-13.852
	I372V (locus 1):L406Y (locus 3):I516V (locus 6)	-8.713
	L439I (locus 5)	6.286
	I372V (locus 1):S402T (locus 2):L406Y (locus 3):T438I (locus 4):I516V (locus 6)	-5.059
	S402T (locus 2):T438I (locus 4):I516V (locus 6)	-3.892
	T438I (locus 4)	3.886
minorProducts	L439I (locus 5)	7.177
	S402T (locus 2)	4.895
	T438I (locus 4)	3.913
	I372V (locus 1):S402T (locus 2):L406Y (locus 3):L439I (locus 5):I516V (locus 6)	-3.551
	L406Y (locus 3)	3.330
	I372V (locus 1)	3.128
	S402T (locus 2):T438I (locus 4):L439I (locus 5):I516V (locus 6)	-3.083
	I372V (locus 1):L439I (locus 5):I516V (locus 6)	-2.446
	I516V (locus 6)	2.193
	I372V (locus 1):S402T (locus 2):L406Y (locus 3):T438I (locus 4):L439I (locus 5)	2.007
PSD	I372V (locus 1):L406Y (locus 3):L439I (locus 5):I516V (locus 6)	44.709
	I372V (locus 1):S402T (locus 2):L406Y (locus 3):T438I (locus 4):I516V (locus 6)	32.099
	T438I (locus 4)	29.269
	I372V (locus 1):S402T (locus 2):L406Y (locus 3):L439I (locus 5):I516V (locus 6)	-27.603
	T438I (locus 4):I516V (locus 6)	-25.887
	I516V (locus 6)	25.381
	S402T (locus 2):T438I (locus 4):I516V (locus 6)	19.947
	S402T (locus 2):L406Y (locus 3)	-17.943
	S402T (locus 2):T438I (locus 4)	-17.559
	L406Y (locus 3):I516V (locus 6)	-16.827