# Modularity, Limited Sparsity, and Extensive Pleiotropy in a Genotype–Phenotype–Fitness Map of Drug Resistance

Mohammad Hossein Donyavi[1,2], Reza Ghelich[1,2], Kara Schmidlin[1,2], Grant Kinsler[3], and Kerry Geiler-Samerotte[1,2,*]

[1]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe AZ
[2]School of Life Sciences, Arizona State University, Tempe, AZ
[3]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[*]Corresponding author: kerry.samerotte@asu.edu

September 16, 2025

**Abstract**

Understanding which phenotypic effects of mutation impact fitness is a central question in evolutionary biology. Yet genotype–phenotype relationships are not bipartite; a single mutation can influence hundreds of molecular-level phenotypes. This complexity poses a major challenge for predicting evolutionary outcomes, such as the relative fitness advantage of one mutant over another. Despite this, genotype–fitness relationships can sometimes be approximated by low-dimensional models that predict the fitness of diverse mutants in novel environments. But are these successes idiosyncratic to a few datasets, or do they reveal general features of genotype–phenotype maps? Here, we highlight the generality of this result and its limits in the context of drug resistance. We apply Singular Value Decomposition (SVD) to a dataset of 800 yeast mutants evolved in various concentrations and combinations of antifungals. This analysis identifies a small number of phenotypic dimensions that predict mutant fitness with up to 70% accuracy in unseen drug combinations. Predictive power plateaus rapidly with the addition of new training environments or model components, indicating that low dimensionality is not an artifact of sampling, but instead represents the funneling of mutational effects through a small number of major phenotypic axes. Our results support longstanding ideas that modularity and canalization constrain phenotypic variation, making evolutionary outcomes more predictable. Beyond modularity, our dataset also sheds light on pleiotropy. Unlike other systems where mutations affect only a subset of dimensions, we find that most mutations perturb most inferred phenotypic axes. This lack of sparsity suggests that drug environments may provoke unusually pleiotropic responses, perhaps because cells rarely encounter these stresses in nature. Natural environments, such as nutrient limitation, may instead reveal a sparser wiring where mutations act through more specific phenotypic routes. Finally, our results highlight that low-dimensional linear models do not capture everything: some drug combinations remain difficult to predict. In these environments, subsets of mutants often show nonmonotonic fitness responses to increasing drug concentrations, and their prediction accuracy does not improve simply by adding more training data. These limitations highlight the need for nonlinear approaches and models that integrate high-dimensional phenotypic measurements to fully uncover how mutations shape cellular processes and fitness.

# 1 Introduction

Predicting biological outcomes from genotype remains a central challenge across biology, from assessing disease risk to engineering resilient microbes and anticipating evolutionary responses [1–5]. This difficulty stems from the complex, context-dependent relationship between genotype, phenotype, and fitness. A single mutation often alters multiple molecular and physiological traits, a phenomenon known as pleiotropy [6–9]. Moreover, the fitness impact of these phenotypic changes often varies across environments, in different ways for different mutants (i.e. GxE interactions) [9–11]. These many-to-many mappings convolute efforts to predict how mutations shape phenotypes and, ultimately, fitness [5, 12].

Despite the complex many-to-many mapping from genotype-to-phenotype-to-fitness, recent studies show that much of the variation in mutant fitness across environments can be captured by low-dimensional linear models [9, 13–16]. Rather than attempting to measure every underlying trait, these models infer abstract 'phenotypic' axes, which reflect dominant patterns of variation, from fitness data alone.

One common explanation for the success of low-dimensional models is modularity (Figure 1A). If the myriad molecular effects of mutation ultimately converge on a limited number of physiological modules, such as drug efflux or stress-response activation, then the fitness consequences of mutation may be canalized and predictable [17, 18]. Yet, the extent of modularity in biological systems has been debated for decades [7, 8, 18–20]. Clear evidence has been elusive, in part because it requires massive datasets measuring the phenotypic and/or fitness effects of many mutants; only then can one ask whether a system as a whole is modular.

A related question is sparsity (Figure 1B): even if modularity exists, do all mutants influence every phenotypic module, and does every module matter for fitness [9, 15]? Sparsity can be visualized as the "wiring diagram". In a sparse system, mutations perturb only a limited subset of phenotypic modules. This selectivity can make the modules easier to interpret because the mutations that affect them point back to specific underlying biological processes. For example, if sequenced mutants that affect a phenotypic module are enriched for mutations in efflux pumps, sparsity would suggest that the module represents efflux activity. Sparsity in the mapping from phenotypic modules to fitness provides similar interpretive power: if only a few modules consistently influence fitness in a given environment, those modules can be directly linked to the processes most relevant for adaptation in that environment. In this way, sparsity does more than simplify prediction, it also helps reveal the biological meaning of the inferred modules. Both modularity and sparsity have emerged as common features in the few high-dimensional datasets where they have been tested [9, 15, 16], but their generality remains unknown.

These concepts may seem abstract, but they have direct consequences, for example, in predicting antimicrobial resistance. Mutations that confer resistance in one drug environment often act pleiotropically, imposing collateral costs or unexpected benefits in others [14, 21, 22]. The resulting tradeoffs differ across mutants depending on which phenotypes they alter, how those phenotypes interact, and how they shape fitness in different environments. In some cases, subsets of mutants even show nonmonotonic fitness responses across drug concentrations, complicating predictions because the same mutation can be beneficial at one dose yet deleterious at another. This complexity raises a critical question: Is it futile to leverage these tradeoffs in treatment design, because it is impossible to know which phenotypic modules a given mutation affects and how those modules contribute to fitness in unseen environments? Or conversely, can modularity and sparsity make such predictions tractable by funneling mutational effects into just a few possible outcomes?

To address this, we previously evolved 774 diverse drug-resistant yeast strains each with distinct adaptive mutations, pleiotropic effects, and fitness tradeoffs across environments [11]. Here, we use Singular Value Decomposition (SVD) to ask to what extent the major axes of phenotypic variation in this dataset reveal modularity and sparsity, and whether, as a result,

low-dimensional models can predict fitness in held-out drug combinations. Each mutant carries, on average, a single adaptive mutation, and the targets of adaptation span diverse pathways, from efflux pumps to stress response regulators, offering a broad window into the genotype-phenotype landscape of resistance [11].

Despite this genetic diversity, we find that linear models with four abstract phenotypic axes can capture the majority of the variation in fitness across mutants and environments. Further, the predictive power of our low-dimensional model saturates quickly with added environments, a hallmark of modularity. These results suggest that, beneath biological complexity, mutations often act through a small set of core modules—an architecture that constrains variation, makes phenotypes more predictable, and has been theorized to support both robustness and evolvability [17, 18, 23, 24].

However, unlike other datasets, the mapping from mutations to phenotypic modules in our system is less sparse: most mutations perturb multiple inferred phenotypic axes. Thus, even though some sequenced mutants carry mutations in genes with clear functions, such as drug efflux pumps, no phenotypic axis can be uniquely associated with a specific class of mutants or the underlying processes they influence. In contrast, the mapping from phenotypic axes to fitness is sparse, with only a subset of phenotypic axes shaping fitness in a given environment. This asymmetry may explain why the inferred phenotypic axes map more readily onto genotype–environment interactions, such as tradeoffs between fluconazole and radicicol resistance, than onto recognizable biological processes. More broadly, our findings raise the question of whether sparsity is a general feature of the genotype–phenotype map, or whether drug environments, which cells may rarely encounter in nature, provoke unusually pleiotropic responses compared to natural conditions such as nutrient limitation.

Still, our results support the emerging view that modularity, and in some cases sparsity, constrain variation and funnels mutational effects into a tractable set of outcomes, making phenotypes more predictable and supporting both robustness and evolvability [17, 18, 23, 24]. Yet some environments remain unpredictable, signaling biological complexity beyond the reach of linear models. In particular, some mutants show nonmonotonic fitness responses across drug concentrations, where the same mutation can help at one dose but harm at another. Different approaches, including nonlinear models [25–27] and models that incorporate high-dimensional phenotype data [28], may be required to capture the full spectrum of cell-level effects of mutation and their consequences for fitness.

# 2 Results

## 2.1 Low-dimensional models predict fitness in novel drug combinations

We asked whether a low-dimensional, linear model could predict mutant fitness in drug combinations using only single-drug data (Figure **??**A–B). Our dataset included 774 diverse drug-resistant yeast mutants for which fitness was measured across 12 environments, three concentrations each of fluconazole and radicicol, plus two controls, as well as four double-drug combinations (Figure **??**B) [11]. The yeast mutants we study here were generated in our previous study by evolving a barcoded ancestor in each of the 12 environments [11]. Each one contains, on average, a single adaptive mutation. The targets of adaptation are diverse, including mutations to drug efflux pumps, protein folding chaperones, glucose sensing pathways and cell membrane components [11].

Using Singular Value Decomposition (SVD) [9, 16], we built low dimensional models summarizing fitness variation across random subsets of 100 mutants in the 8 single-drug and control conditions. Then we predicted the fitness of the remaining 664 mutants in the 4 double-drug combinations. The fitness predictions we report in Figure **??** represent the average across 100 iterations. In each iteration, we inferred latent phenotypic dimensions that minimized the mean

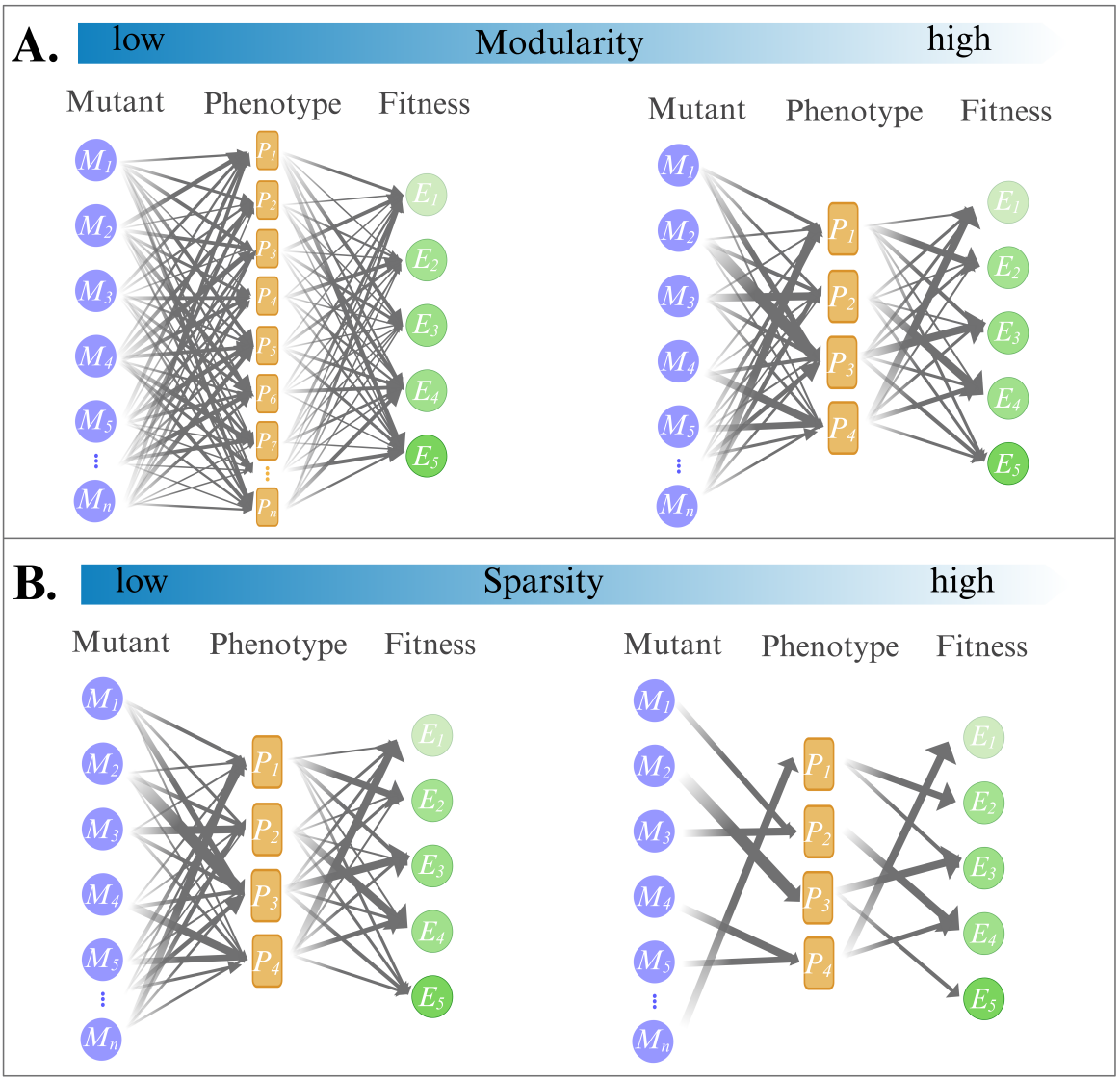



Figure 1: **A toy model of the genotype-phenotype-fitness (GPF) map that highlights two key features: modularity and sparsity.** These GPF map models are modified from previous work [9, 15] to highlight focal concepts in our study. Each map consists of different mutants (blue circles) that influence one or more orthogonal phenotypic modules (yellow squares) to differing degrees (thickness of grey arrows). The phenotypes each affect fitness in one or more environments (green circles) to differing degrees. **A)** We use the term "modularity" to refer to the number of orthogonal phenotypic axes that mediate fitness (depicted as the number of yellow squares) [9]. In a map with low modularity (left), the effects of mutations are diffuse, mapping onto a large and complex set of phenotypes that each contribute to fitness in many environments. This represents an architecture where predicting fitness would be more difficult. With high modularity (right), the effects of different mutations converge on a few phenotypic axes that dictate fitness. This simplifies the map by reducing its effective dimensionality, which could make fitness in unseen environments more predictable. **B)** We use the term "sparsity" to refer to the density of connections between mutants, phenotypes, and environments (depicted as the number of grey arrows) [15]. Connections are dense with a low sparsity structure (left). In a high sparsity map (right) a given mutant may only affect a few phenotypes, and a phenotype may only be relevant for fitness in a specific subset of environments.

squared error of the fitness predictions. These latent dimensions each represent an axis of variation in mutant fitness. They can be thought of as abstract phenotypes inferred from the data (Figure **??**B). Each one may correspond to a unique combination of cellular traits that affect fitness.

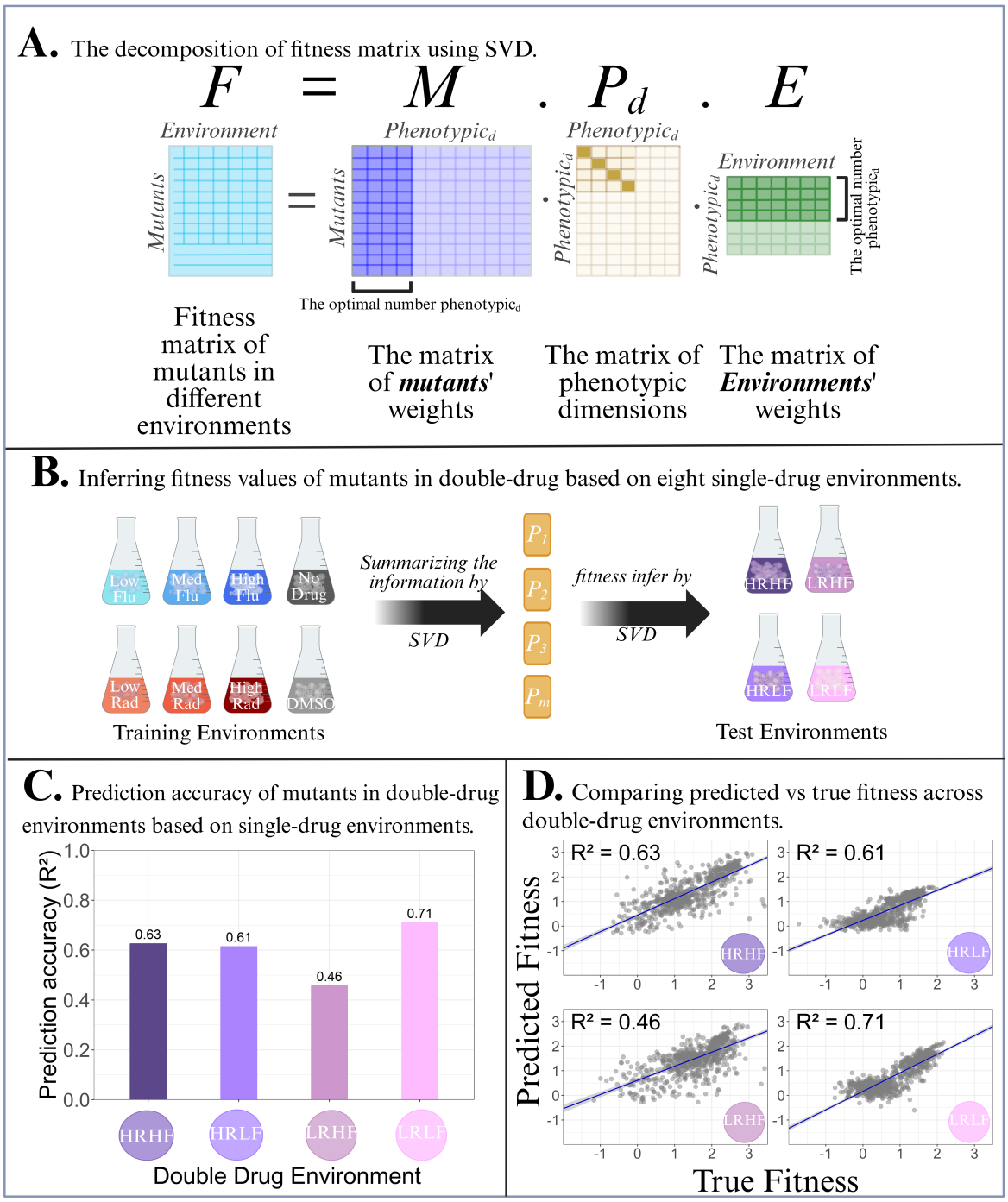
This linear modeling approach predicted fitness in some double-drug environments more accurately than others (Figure **??**C–D). Predictive performance was especially strong in the LRLF condition, combining low concentrations of both drugs. Here, prediction corresponds to measured values with an $R^2$ of 70% (Figure **??**D; bottom right). This result demonstrates that genotype-fitness relationships under drug stress can be partially captured by a linear model, echoing earlier work showing that such models can capture genotype-phenotype relationships under glucose limitation [9, 16]. Such low-dimensional behavior often reflects modular structure, where a limited set of phenotypic processes (or "modules") consistently shape fitness across mutants and environments.

However, in other drug combinations, prediction accuracy declined. For example, in the LRHF condition, combining low radicicol with high fluconazole, the model's performance dropped below 50% (Figure **??**D; bottom left). There are several reasons why a linear model might fail to explain most of the fitness variation in this environment:

First, previous work shows that biological responses differ across mutants [11, 29]. For example, some mutants (including those with sequenced mutations to HDA1) perform worse in multidrug conditions than in either single-drug condition. Other mutants (including those with sequenced mutations to GPB2) do not show this type of GxE interaction. Since these mutants have similar GxE interactions in the 8 training environments, the model cannot learn and predict these differences between mutant classes. By this logic, adding more diverse training environments might uncover additional latent phenotypic axes and improve fitness predictions.

A second, though not mutually exclusive, reason why SVD struggles to predict fitness in some conditions may be that linear models are fundamentally limited. Even with more training environments, some genotype–fitness relationships may be difficult to model because they are inherently nonlinear. For example, the combined effect of stressors may involve threshold responses, feedback loops, or nonmonotonic phenotype-fitness mappings that are not well described by linear models [13, 30–32]. Another possibility is that the phenotypic effects of a mutation change across environments. In modular systems, we expect a mutation to affect the same module (e.g., membrane function or stress response) across environments. But if the modules themselves reconfigure depending on the environment [8], or if mutations switch modules, then even modularity won't help predict fitness from past data using linear models.

In short, if adding more environments doesn't help, the genotype-phenotype map may not just be high-dimensional, but also nonlinear, context-dependent, and reconfigurable. To further these ideas, next we ask whether adding more training environments helps to boost prediction accuracy.


A. The decomposition of fitness matrix using SVD.
F = M . Pd . E
Environment
Mutants
Phenotypic_d
The optimal number phenotypic_d
Fitness matrix of mutants in different environments
The matrix of mutants' weights
The matrix of phenotypic dimensions
The matrix of Environments' weights
B. Inferring fitness values of mutants in double-drug based on eight single-drug environments.
Low Flu
Med Flu
High Flu
No Drug
Low Rad
Med Rad
High Rad
DMSO
Training Environments
Summarizing the information by
SVD
P1
P2
P3
Pm
fitness infer by
SVD
HRHF
LRHF
HRLF
LRLF
Test Environments
C. Prediction accuracy of mutants in double-drug environments based on single-drug environments.
Prediction accuracy (R²)
0.63
0.61
0.46
0.71
HRHF
HRLF
LRHF
LRLF
Double Drug Environment
D. Comparing predicted vs true fitness across double-drug environments.
Predicted Fitness
R² = 0.63
R² = 0.61
R² = 0.46
R² = 0.71
True Fitness

176

Figure 2: **A low-dimensional linear model predicts mutant fitness in double-drug environments from single-drug fitness. (A)** Schematic of Singular Value Decomposition (SVD) applied to a fitness matrix (F). The matrix is decomposed into mutant weights (M), phenotypic dimension weights (Pd), and environmental weights (E), reducing the data into a lower-dimensional space that captures key genotype–phenotype–fitness relationships. **(B)** Training and testing environments. The SVD model was trained on fitness measured in eight single-drug environments [11]: three Radicicol concentrations (red; 5 $\mu$M, 15 $\mu$M, 20 $\mu$M), three Fluconazole concentrations (blue; 4 $\mu$g/ml, 6 $\mu$g/ml, 8 $\mu$g/ml), and two controls (gray; DMSO, no drug). Model accuracy was tested in four double-drug environments (purple): HRHF (15 $\mu$M Rad + 8 $\mu$g/ml Flu), HRLF (15 $\mu$M Rad + 4 $\mu$g/ml Flu), LRHF (5 $\mu$M Rad + 8 $\mu$g/ml Flu), and LRLF (5 $\mu$M Rad + 4 $\mu$g/ml Flu). **(C)** Prediction accuracy of the SVD model in the four double-drug environments. **(D)** Scatter plots underlying the $R^2$ values in panel C, comparing predicted and measured fitness for each of 774 mutants. Predictions were generated by repeatedly training on random subsets of 100 mutants and testing on the remainder; each mutant's reported prediction is the average across all iterations where it was held out of training.

## 2.2 Rapid Saturation of Prediction Accuracy Reveals Modular Structure and Its Limits

To test whether yeast genotype-fitness relationships reflect an underlying phenotypic modularity, we examined how prediction accuracy changes as we systematically increase the number of training environments. We did this in our dataset (12 environments; Figure 3A) and in a second, larger dataset (45 environments; Figure 3B) [9]. In both datasets, prediction accuracy rises steeply with the first few added environments, then quickly saturates (Figure 3; right side). This rapid saturation suggests that a small, representative subset of environments captures most of the predictable signal. This is a strong indication of modularity in the genotype-phenotype-fitness map. In other words, the same small set of causal phenotypes appears to consistently contribute to fitness across these environments, such that once the phenotypic effects of mutation are learned in some conditions, fitness can be predicted in others.

Supporting this interpretation, the number of phenotypic dimensions needed to achieve high predictive power also plateaus rapidly (Figure 3; left side). Even across model iterations where the 'optimal' number of dimensions (the number that minimized mean squared error) was greater, predictive power does not improve (Figure 3; 3D plots). The observation that adding more model components (either dimensions or environments) does not improve predictability is consistent with a modular view. Rather than affecting countless unrelated molecular traits, the 800 mutants we study seem to influence a small set of shared phenotypic modules that explain the majority of the variation in fitness. We will explore the identity of those modules in the next figures.

First, it may be interesting to compare the extent of modularity in the two datasets presented in Figure 3. The larger 45-environment dataset saturated faster than the 12-environment dataset (Figure S1), despite our initial expectation that it would require more environments to maximize prediction accuracy. We had assumed that a more diverse set of environments would reveal a more complex phenotype space, requiring more data to model. But the opposite trend (Figure S2) suggests that prediction is shaped not only by environmental diversity, but also by the diversity of the mutants themselves. Our focal dataset (Figure 3A) includes 774 mutants pooled from 12 different evolution experiments, some evolved in response to radicicol, others to fluconazole. In contrast, the 292 mutants in the second dataset (Figure 3B) all evolved under a single selective pressure: glucose limitation [33]. Because the ability to uncover latent phenotypes depends on perturbing them, the diversity of mutants affects which phenotypic modules the model can detect. Even if a module contributes to fitness, it will remain invisible to the model unless at least some mutations perturb it. This finding underscores a key feature of the genotype–phenotype map: to reveal the underlying structure, one must probe it with sufficiently

diverse mutations and informative environments.

Though both datasets show rapid saturation of predictive accuracy, certain environments remain consistently unpredictable regardless of how many training environments are included or how many phenotypic dimensions the model allows. In particular, LRHF (low radicicol, high fluconazole) again stands out as a persistent outlier (Figure 3A; rose). Even with maximal training environments and expanded model complexity, predictions plateau below an $R^2$ of 0.5, suggesting that linear models miss key features of this condition. These persistent gaps in predictability prompted us to investigate which training environments drive prediction success or failure.

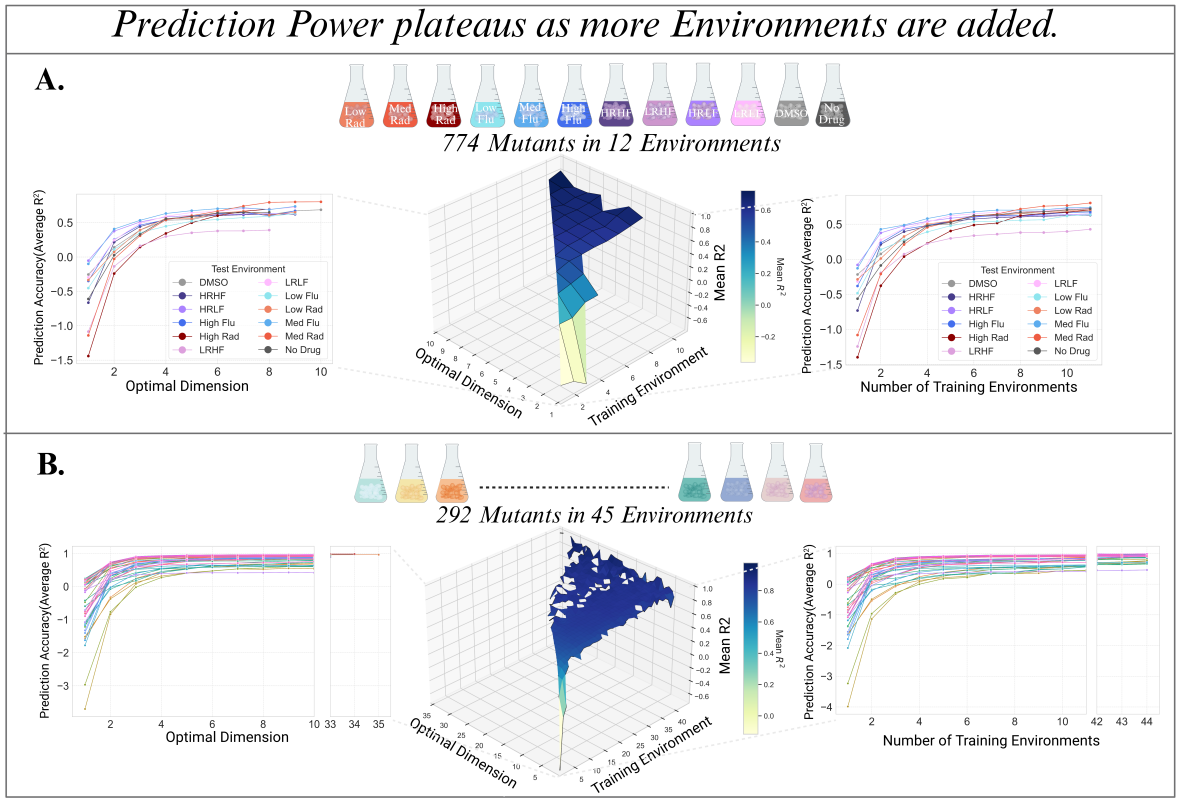



Figure 3: **Prediction accuracy saturates rapidly, revealing a low-dimensional and modular genotype-phenotype map.** To test how prediction accuracy depends on training data and model complexity, we varied the number of training environments ($k$) from 1 to $n-1$, where $n$ is the total number of environments. For each $k$, 100 iterations were performed in which $k$ environments were randomly selected for training, and predictive accuracy ($R^2$) was calculated on the held-out environments and averaged across iterations. Training mutants were chosen as in previous work [9], by selecting a small but diverse subset representing different G×E interactions. This differs from Figure 2 where training mutants were chosen at random, with 100 iterations. We applied this analysis to assess the impact of adding more training environments was applied to two datasets: **(A)** fitness of 774 mutants across 12 environments [11], and **(B)** fitness of 292 mutants across 45 environments [9]. The right-most plots show that prediction accuracy rises rapidly and then saturates after just a few training environments; fewer in the bottom graph. This rapid saturation indicates that a small subset of environments captures the majority of the information needed to predict fitness. Similarly, the left-most plots show that prediction accuracy plateaus after just a few phenotypic dimensions, fewer in the bottom graph. This suggests the fitness effects of these mutations are explained by a small number of phenotypic dimensions. The central 3D surface integrates the left and right plots, showing the prediction accuracy (average $R^2$) as a function of both the number of training environments and optimal dimensions (i.e. the number of dimensions chosen after each iteration of bi-cross validation). The surface rises quickly and then forms a broad plateau, demonstrating that high predictive accuracy is achieved with a limited number of both environments and dimensions, with further additions making negligible improvement. This reinforces the conclusion that the system's dimensionality is low.

## 2.3 Interpreting phenotypic modules by examining informative environments and mutants

To better understand the phenotypes that drive fitness in multidrug environments, we asked which single-drug environments contribute most to predictive accuracy. If a particular condition disproportionately shapes the inferred phenotypic dimensions, it may reveal a key axis of variation in the genotype–phenotype–fitness map. We systematically tested the contribution of single-drug environments to predictive accuracy by asking how removing or adding back environments affected the model's performance.

This analysis revealed strong environmental dependencies. Fluconazole-containing environments and especially low-dose fluconazole were the most informative. Removing all fluconazole environments caused a steep drop in predictive accuracy, while removing radicicol had more modest effects (Figure 4A). This suggests that the phenotypic modules activated by radicicol are either less informative or already partially captured by other environments. The most dramatic impact came from low-dose fluconazole: removing this single environment reduced prediction accuracy substantially, while adding it back increased $R^2$ by 0.2 for the LRLF condition (Figure 4B).

To understand why low fluconazole had such a strong effect, we examined the mutants whose predicted fitness changed the most when low fluconazole was excluded from the training data. Two distinct patterns emerged (Figures 4C–E). The 50 mutants most underpredicted without low fluconazole (Figures 4C–D; blue) exhibit a characteristic nonmonotonic response to changes in fluconazole concentration: they have high fitness only in low-dose fluconazole, but ancestral-like fitness in both higher doses and in fluconazole's absence (Figure 4C; blue). Without low fluconazole in the training data, the model inferred ancestral-like fitness for these mutants in LRLF (Figure 4C; green). Including low fluconazole resulted in a higher-fitness prediction that better matched the observed data (Figure 4C; light blue).

Conversely, the 50 most overpredicted mutants (Figures 4D–E; red) showed the opposite nonmonotonic trend. These mutants had elevated fitness across all conditions except low-dose fluconazole, where they were closer in fitness to the ancestor. Without this key environment in the training data, the model overestimated their fitness in LRLF (Figure 4E; green).

These nonlinear dose-response patterns, where fitness effects change nonmonotonically across drug concentrations (Figures 4C and 4E), suggest that low fluconazole is not just a milder version of higher concentrations, but taps into a distinct physiological regime. Supporting this idea, the two mutant groups we identified closely match previously reported mutant groupings identified via Louvain clustering performed on G×E interaction profiles [11] (Figure S3). That these same groups are recovered through our approach, focused on predictive power rather than clustering, highlights their robustness and reinforces the critical role of low fluconazole in teasing apart groups of mutants with different effects on phenotype. Supporting the idea that low fluconazole engages a distinct physiological regime, it consistently proves critical for accurate prediction of multidrug fitness, even when the training set is substantially altered (Figure S4). In addition, fitness in low fluconazole is the least correlated with any other single-drug environment (Figure S5), underscoring its unique effects.

We repeated a similar analysis to understand why including high fluconazole improves prediction accuracy in the LRHF environment (Figure 4B). LRHF is the environment where prediction accuracy is consistently the worst (Figures 2–4; rose). Examining the 100 mutants whose LRHF predictions improved the most when high fluconazole was included in training again revealed distinct groups with characteristic G×E patterns, suggesting that the difficulty in predicting fitness in LRHF is not exclusively due to measurement noise (Figure S6). Of particular interest are the mutants that become overpredicted when high fluconazole is excluded from the training set. Their G×E interactions mirror those observed in Figure 4C, where mutants show a nonmonotonic fitness advantage in low fluconazole. Here, this generates the opposite type of prediction error: excluding high fluconazole causes LRHF overprediction, while excluding low

fluconazole causes LRLF underprediction (Figure S6).

Together, these data suggest that the non-monotonic G×E interactions present across fluconazole concentrations are important to modeling the relationship from genotype to fitness. They further suggest that environments containing low (4 $\mu$g/ml) versus high (8 $\mu$g/ml) dose fluconazole are qualitatively different, and may either engage distinct phenotypic modules, shift the balance among existing ones, or provoke interactions between modules that are not the same. More broadly, this analysis illustrates how examining prediction performance can yield biological insight, highlighting environments that probe key phenotypic axes and mutants that differentiate among them, and helping us interpret the phenotypes that connect genotype to fitness.

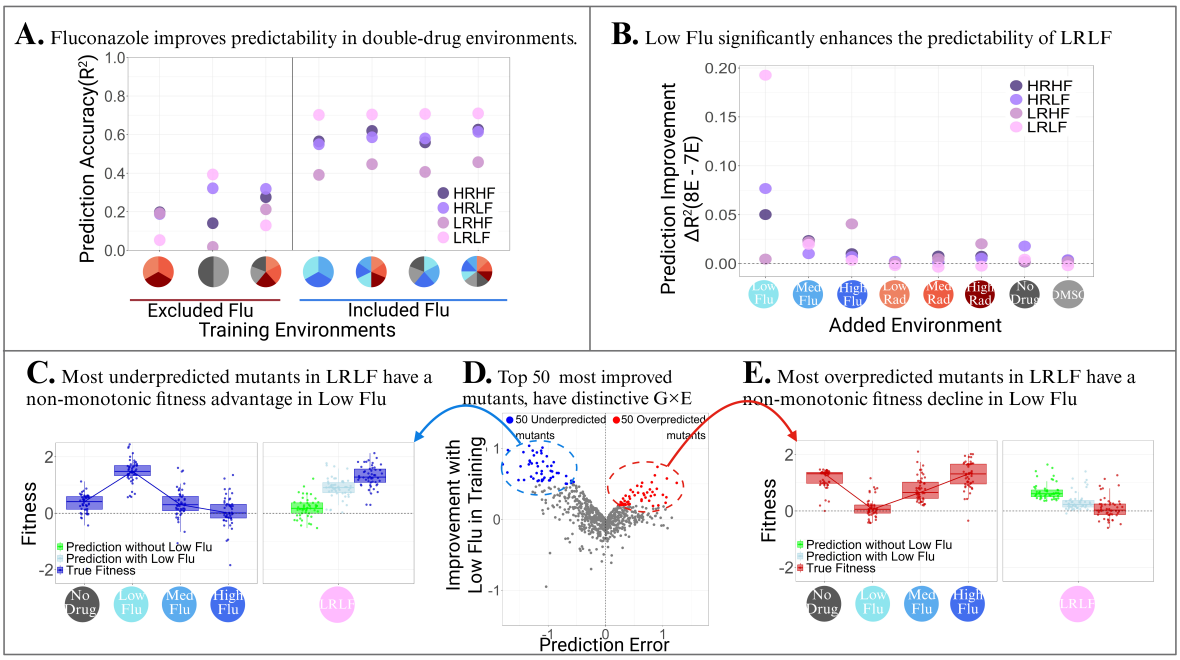



Figure 4: **Low-dose fluconazole is a uniquely informative environment.** To identify which environments provide the most information for predicting fitness in double-drug combinations, we systematically eliminated environments from the training datasets. **(A)** Prediction accuracy improves when fluconazole (Flu) environments are included in the training set. Accuracy was quantified as $R^2$ between predicted and measured fitness, averaged over 100 training runs, where in each run a random set of 100 mutants was used to predict the fitness of the remaining 664 (as in Figure 2). **(B)** The specific contribution of each single-drug environment to prediction accuracy in each double-drug environment. The y-axis shows the improvement in $R^2$ when a given environment on the x-axis is added to a base training set of 7 single-drug environments. Adding Low Flu provides the highest improvement in prediction accuracy, specifically for the LRLF test environment (light pink). **(C)** Fitness profiles of the 50 most underpredicted mutants (points) when Low Flu is removed from the training set display a non-monotonic response to fluconazole concentration, with a distinct fitness advantage in Low Flu. The model trained without Low Flu fails to capture this advantage and underpredicts fitness in LRLF (green). Including Low Flu in the training set allows the model to better approximate fitness in LRLF (light blue). **(D)** Relationship between prediction error when Low Flu is excluded from training, and prediction improvement when Low Flu is included (for the LRLF condition). Colored points display the 100 mutants for which predictions improve the most upon adding Low Flu to the training set (blue when excluding Low Flu leads to underprediction, red when it leads to overprediction). **(E)** Fitness profiles of the 50 most overpredicted mutants from panel D. These mutants exhibit an opposite non-monotonic pattern to those in panel C, with a fitness decline in the Low Flu environment compared to other Flu concentrations. The model trained without Low Flu misses this decline and overpredicts their fitness in LRLF (green).

## 2.4 Lower sparsity than observed in other datasets

Having shown that a low-dimensional model can predict mutant fitness in novel environments, we next asked: what can such models reveal about the phenotypes that shape fitness across environments? In this section, we shift from prediction to interpretation. Rather than holding out double-drug environments to evaluate predictive accuracy, we now train the model on the full set of 12 single-drug environments. Our goal is to examine the major phenotypic axes that organize mutant responses across drug conditions.

To this end, we applied Singular Value Decomposition (SVD) to the complete matrix of 774 mutant fitness values across 12 environments (Figure S7). This analysis revealed that four latent phenotypic dimensions capture 95% of the variance in the data. The overall reconstruction quality was high ($R^2 = 0.87$), indicating that the main patterns of drug response can be compressed into a compact, low-dimensional representation. Note that while the $R^2$ values in previous figures refer to prediction accuracy, here we assess reconstruction quality, how well the inferred model recapitulates the fitness values it was trained on, rather than predictive accuracy on held-out environments.

Not all environments were equally well captured by this model. In particular, fitness responses in the LRHF condition (Figure S7; lowest bar, rose color), a multidrug environment already shown to defy accurate prediction, were also poorly reconstructed. This discrepancy could reflect measurement noise that our model cannot capture. Yet LRHF was not the environment with the highest replicate error [11]. That distinction belonged to medium fluconazole (MF), which was nonetheless reconstructed with the third-highest accuracy (Figure S7; $R^2 = 0.895$). Instead, the model's difficulty reconstructing fitness in certain environments, even when trained on all available data, may evoke nonlinear or context-specific responses that resist compression into linear, low-dimensional space.

Still, the fact that just four latent phenotypic dimensions reconstruct most of the variation in fitness underscores that these axes carry a strong signal worth interpreting. To enhance interpretability, we applied a sparsity-constrained variant of SVD, *Sparse Structure Discovery (SSD)* [15] (Figures 5 and S7). SSD enforces sparsity by pruning links in the genotype–phenotype–fitness map (Figure 5B; grey arrows). This improves interpretability by highlighting which mutants perturb particular latent dimensions and in which environmental contexts they matter most.

In datasets with truly sparse structure, random rotations of the fitness matrix $F$ (Figure **??**A) will destroy the alignment between mutants (or environments) and the inferred phenotypic dimensions. As a result, more inferred dimensions (less sparsity) is required to achieve similar reconstruction accuracy to the original, unrotated data. This is indeed what we observe for the mapping from phenotypic dimensions to environments, where there is a clear gap between the original and rotated data (Figure 5C). This suggests a sparse structure, where not every phenotype affects fitness in every environment. This corresponds to fewer grey arrows connecting phenotypes to environments in our conceptual model (Figure 5B).

By contrast, rotation tests had less effect on the alignment from mutants to inferred phenotypic dimensions (Figure 5A). This lack of a significant gap between the original and rotated data suggests that the mapping from mutants to phenotypes is less sparse. In other words, pleiotropy is higher, such that each mutation affects many phenotypes, as depicted by the many grey arrows connecting mutants to phenotypes in our conceptual model (Figure 5B). This finding differs from what was observed in three previous data sets [15], perhaps suggesting that the mutations that contribute to drug resistance are more pleiotropic than others. However, in one previous dataset that also focused on adaptive yeast mutants [9], sparsity from mutants to inferred axes was driven by one common type of mutation, and largely disappeared when that type was downsampled [15]. This could suggest that higher pleiotropy is a common feature among laboratory-adapted yeast mutants.

In terms of interpretability, the SSD axes largely overlapped with those from standard SVD (Figure S7), regardless of which sparse solution we chose. This convergence of the two methods is a non-trivial result that speaks to the underlying structure of the genotype-phenotype map. In a system where the dominant axes of variation are not well-aligned with axes that reflect a sparse solution, SVD and SSD would yield different solutions, as SSD would find a sparser but rotated space at a cost to reconstruction error. The fact that they yield nearly identical axes in our system (Figure S7) implies that the directions of maximal variance, which SVD is designed to find, are already aligned with the sparsest possible description of the system's modularity.

This means the original vectors returned by SVD are not arbitrary but represent a stable, biologically relevant solution that the sparsity-constrained search of SSD independently confirms. This robustness is significant because it suggests the low-dimensional structure we identified is not a fragile statistical abstraction but a primary feature of our studied system. It reinforces the conclusion that while mutations are broadly pleiotropic at the molecular level, reflected in a less sparse mutant-phenotype map, their net effects on fitness are channeled through a limited set of nearly orthogonal environmental modules. This architecture, where a rather dense core of pleiotropy projects onto a sparse modular output, may be a general feature of adaptation to complex stressors like antifungal drugs.

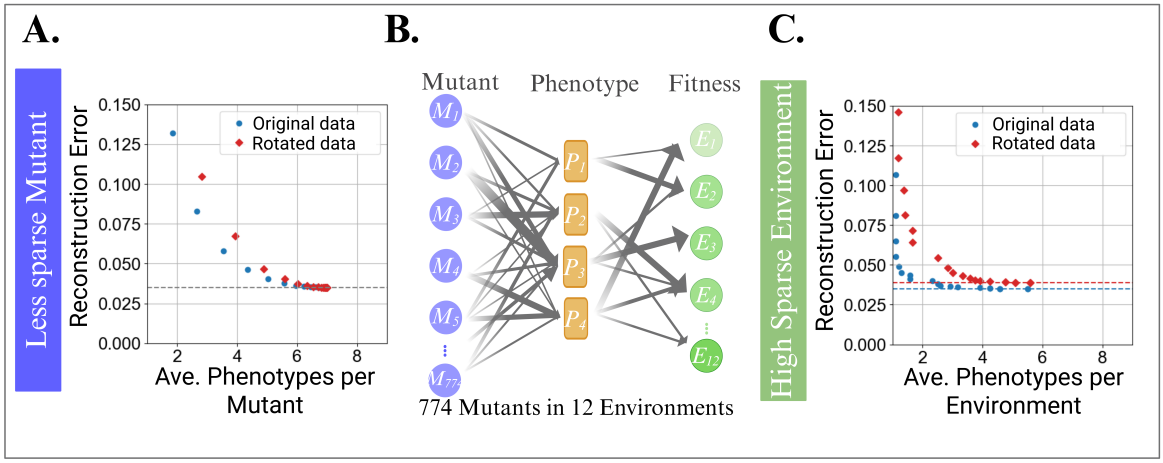



Figure 5: **A rotation test reveals a less sparse mutant-to-phenotype map (blue) but an inherently sparse environment-to-phenotype map (green).** To formally test for the presence of inherent sparsity in our experimental dataset, we applied a rotation test using a sparse discovery algorithm (SSD). The rotation test tries to find a sparse, low-error solution for the original fitness matrix, $F$, versus a rotated matrix ($F' = FO$), which preserves the correlational structure but removes any special sparse basis. A significant gap between the solution curves for the original and rotated data would indicate the presence of a sparse structure. **(A)** The rotation test versus original data for the mutant-to-phenotype mapping. The solution curves for the original and rotated data largely overlap. **(B)** A conceptual genotype-phenotype-fitness map that represents the amount of sparsity found within our dataset. Many grey arrows connect mutants to phenotypes, while fewer connect phenotypes to environments, depicting how the latter mapping is sparser. **(C)** The rotation test versus the original data for the phenotype-to-environment mapping. Each point represents a solution for a different regularization parameter. This plot shows the trade-off between reconstruction error and the average number of inferred phenotypes per environment. A distinct gap exists between the solutions for the original data (blue) and the rotated data (red). The original data can be explained by a much sparser model (fewer phenotypes per environment) without reducing accuracy, a clear hallmark of an inherently sparse structure in the environment-phenotype map.

## 2.5 Phenotypic axes correspond to gene-by-environment interaction patterns

The phenotypic dimensions recovered by both SSD and SVD are similar (Figure S7) and difficult to interpret in terms of discrete cellular functions, perhaps due to the lower sparsity in the mapping from mutants to inferred phenotypes. For example, although 35 of our 774 mutants perturb regulators of multidrug efflux, there is no single phenotypic dimension defined by these mutants. They predominantly fall into group 2 (Figure 6A; green) and contribute strongly to

phenotypic dimension 1 (Figures 6A–B). Consistent with lower sparsity, many other mutants have similar effects on this phenotypic dimension. These effects scale with average fitness advantage across drug environments. Thus, phenotypic dimension 1 seems to capture a common type of gene-by-environment interaction (G×E) whereby fitness is high in environments that contain drugs, and low in environments that do not (Figure 6B).

Similarly, the other phenotypic dimensions seem to capture G×E interactions more so than underlying physiological mechanisms by which mutants affect fitness. For example, phenotypic dimension 2 separates mutants based on relative advantage in fluconazole versus radicicol (Figure 6C). Dimensions 3 and 4 capture patterns of G×E related to changes in fluconazole concentration (Figures 6D–E).

These results suggest that the extracted dimensions likely reflect shared fitness outcomes of mutations, rather than the unique cellular process that underlie those outcomes. Interpreting the underlying biological mechanisms that link genotype to fitness may therefore require integrating additional information, such as mutant genotypes or transcriptional profiles. Still, the fact that just four dimensions explain the vast majority of fitness variation, despite the genetic diversity present across this collection of 774 mutants, implies that genotype-by-environment interactions are canalized such that every perturbation does not have a distinct pattern of G×E. This canalization might arise from constraints imposed by regulatory network architecture or trade-offs in resource allocation, which may in turn restrict the evolutionary routes available for drug resistance. Whatever the cause, this structure is encouraging because it shows that even though mutations are inherently pleiotropic, influencing countless cellular traits, their fitness effects often funnel through a small number of shared axes. This implies that mapping additional phenotypes onto these axes, identifying the core processes that shape fitness, and ultimately predicting evolution, may be more tractable than one might expect given the apparent complexity of all that a mutation could, in principle, alter inside a cell.

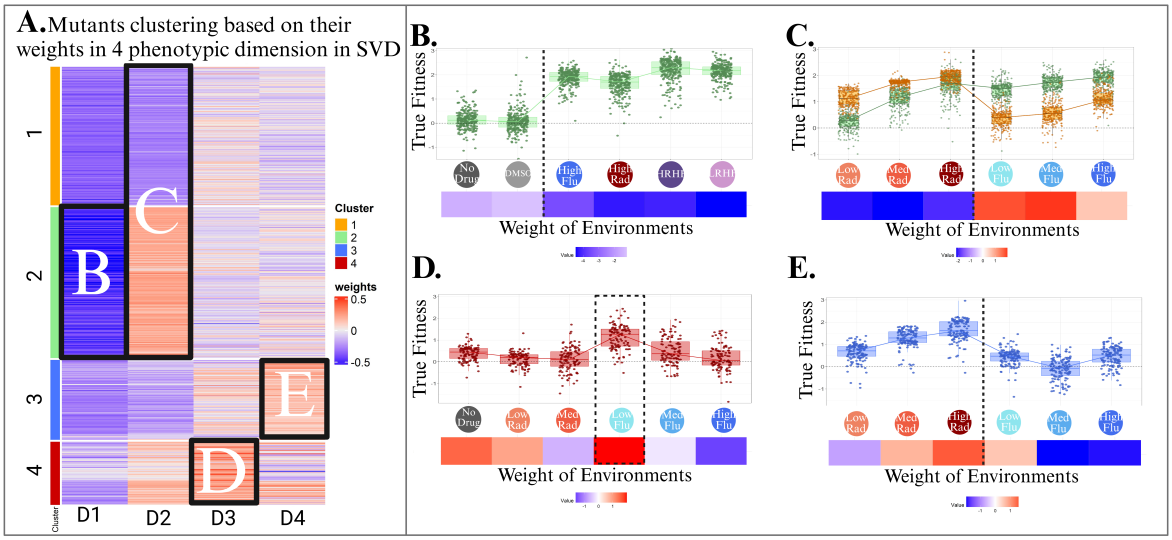



Figure 6: **A four-dimensional model decomposes the fitness matrix into phenotypic axes that capture gene by environment interactions.** Singular Value Decomposition (SVD) was applied to the complete fitness matrix of 774 mutants across all 12 environments to interpret the underlying structure of the genotype-phenotype map. **(A)** A heatmap of the contribution of all 774 mutants to each inferred phenotypic dimension (red vs. blue indicate strong contribution in opposite directions). Mutants are clustered based on their weight profiles (calculated based on $M \times \sqrt{P}$). Boxes highlight the mutant clusters detailed in panels B–E. **(B)** This phenotypic dimension is most strongly affected by mutants that have fitness advantages across environments containing drugs that are lost in environments without drugs. Mutants with amino acid changes in the PDR1 and PDR3 genes, which are regulators of drug efflux pumps, largely fall into this category. **(C)** This dimension is strongly influenced in opposite directions by two groups of mutants: one that resists flu more than rad (green), and another that resists rad more than flu (orange). **(D)** This dimension captures a specific response to low-dose fluconazole, which has the strongest positive environmental weight. Mutants with high weights in this dimension show a distinct fitness advantage in this specific environment. **(E)** This dimension reflects a distinct nonmonotonic response to fluconazole: fitness is lowest at intermediate concentrations. This pattern differs from the nonmonotonic responses in Figure 4, where mutants are instead most sensitive in low fluconazole.

# 3 Discussion

In this study, we demonstrated that despite the apparent complexity of genotype-to-phenotype-to-fitness mappings in drug-resistant yeast, low-dimensional linear models using only singular value decomposition (SVD) capture a surprisingly large portion of fitness variation across environments. This supports the view that genotype–phenotype relationships may be modular such that the effects of many diverse perturbations are canalized into a few common outcomes. This framework resonates with prior quantitative genetics models suggesting that while pleiotropy is pervasive, it may be highly structured or restricted, allowing evolution to proceed without overwhelming numbers of trade-offs [19, 34].

Our results also echo ideas from the omnigenic perspective on complex traits, which holds that mutations in many genes, even those only indirectly linked to a phenotype, can shape trait variation through dense regulatory networks [20, 35]. In our case, these widespread effects did not scatter unpredictably but instead collapsed onto a small set of shared axes. A recent quantitative omnigenic model formalized this principle, showing that network-level complexity

can be summarized in low-dimensional, interpretable structures [36]. The key difference is that while the quantitative omnigenic framework derives axes from regulatory network topology, our SVD approach infers them from observed fitness patterns alone. Yet both converge on the idea that molecular complexity funnels through a limited set of modules that enable prediction.

Looking forward, several questions remain. How general is the low-dimensional modular structure across timescales, selective pressures, or organisms [8, 9, 18, 34]? What molecular or physiological variables underlie the axes we infer, and how can we map them to interpretable traits [12, 15, 28]? To what extent do nonlinear or environment-specific modules emerge in more complex conditions [16, 26]? Addressing these questions will require integrating multi-omics data, dynamic phenotyping, and nonlinear modeling frameworks. Such approaches may bridge the gap between predictive low-dimensional structure and the underlying mechanisms, illuminating how mutations ripple through networks to produce fitness effects.

# 4 Methods

## 4.1 Fitness data, yeast strains and experimental design

The yeast strains and experimental methods employed in this study were previously described in detail [11]. Briefly, all yeast strains originated from a single ancestral "landing pad strain" (SHA185), a MAT$\alpha$ haploid strain characterized by the genetic background: `ura3D0,ybr209w::Gal-Cre-KanMX-1/2URA3-loxP-Barcode-1/2URA3-HygMX-lox66/71`.

To generate the mutants, around 300,000 unique neutral DNA barcodes were introduced into this strain using a Cre-lox recombination system [33]. These barcoded yeast were evolved in 12 environments. These 12 evolution experiments were conducted in a glucose-limited M3 medium lacking uracil, supplemented with fluconazole, radicicol, or DMSO at the following concentrations: Low Flu (4 $\mu$g/ml), Med Flu (6 $\mu$g/ml), High Flu (8 $\mu$g/ml), Low Rad (5 $\mu$M), Med Rad (15 $\mu$M), High Rad (20 $\mu$M), LRLF (5 $\mu$M + 4 $\mu$g/ml), LRHF (5 $\mu$M + 8 $\mu$g/ml), HRLF (15 $\mu$M + 4 $\mu$g/ml), HRHF (15 $\mu$M + 8 $\mu$g/ml), 0.5% DMSO (0.5%) and No Drug (no supplement). Drug concentrations were chosen to apply mild selective pressures without dramatically reducing barcode diversity [11]. Herein, we refer to the 4 environments containing both fluconazole and radicicol as *"double-drug environments"* and all others as *"single-drug environments."* In previous work, adaptive yeast mutants, isolated at intermediate evolutionary time points to ensure maximal barcode diversity, were combined from all 12 evolution experiments into a single pooled population. Then they were competed against an ancestral strain in each of the 12 aforementioned experimental conditions. Fitness was inferred from log-linear changes in each mutant's barcode frequency, over time, relative to the ancestor. Barcode frequencies were measured by Illumina sequencing after sequential batch culture transfers. Only barcodes meeting strict sequencing coverage criteria ($\geq$500 counts across time points) were included, resulting in a dataset comprising fitness values for 774 adaptive yeast mutants across all 12 conditions.

Sixty of these 774 adaptive mutant lineages have been previously sequenced [11]. Most differ from the ancestor by only one or two amino acid changes. Thirty five have (different) amino acid changing mutations to either PDR1 and PDR3, known regulators of a drug efflux pump (PDR5). These 774 adaptive mutant lineages have also been previously clustered based on their G$\times$E interaction profiles [11]. These clusters were used to pick a balanced set of mutants with diverse and representative G$\times$E interactions for the analysis described in Figure 3.

For a complete description of yeast strain construction, barcode library generation, and detailed fitness assay protocols, please refer to Schmidlin et al. 2024 [11].

## 4.2 Dimension reduction of the fitness matrix via singular value decomposition (SVD)

We applied singular value decomposition (SVD) to reduce the dimensionality of a fitness matrix consisting of 774 mutants measured across 12 environments. In matrix notation,

$$F = U\Sigma V^T \tag{1}$$

where $F$ is the $N \times E$ fitness matrix ($N = 774$, $E = 12$), $U$ is an $N \times r$ matrix of left singular vectors (mutant loadings), $V$ is an $E \times r$ matrix of right singular vectors (environment loadings), and $\Sigma$ is an $r \times r$ diagonal matrix of singular values.

In this framework, the columns of $U$ describe how individual mutants load onto each latent dimension, while the columns of $V$ describe how environments load onto those same dimensions. The diagonal entries of $\Sigma$ quantify the amount of variation explained by each dimension. Together, these matrices define abstract phenotypic dimensions that summarize systematic patterns of genotype–environment interactions.

Depending on the analysis, different combinations of mutants and environments were used to train the SVD model. These differences are described in the main text, the figure legends, and below.

All analyses were conducted using custom Python scripts. The full code, including a Jupyter notebook demonstrating the SVD workflow, has been made publicly available at `https://github.com/Russellrezagh/sparse-structure-discovery.git`

## 4.3 Bi-cross-validation of SVD models

In most figures, a bi-cross-validation approach was used to assess the predictive performance of the SVD-based model [9, 37]. This approach involves choosing a subset of mutants and environments as the training data, building a low dimensional model, and then using that model to predict the fitness of the held out mutants and environments, which are referred to as the test data.

In bi-cross-validation, the fitness data matrix (matrix $F$) is divided into four submatrices: (submatrix A) test mutants in test environments, (submatrix B) test mutants in training environments, (submatrix C) training mutants in test environments, and (submatrix D) training mutants in training environments. SVD is applied to the training submatrix (submatrix D), producing singular values and singular vectors that represent latent dimensions in the training data. Predictions for the held-out submatrix (submatrix A) are then generated using the relationship

$$\hat{A} = B(D^{(k)})^{+}C \tag{2}$$

where $D^{(k)}$ is the rank-$k$ approximation of $D$, obtained by retaining the top $k$ singular values and corresponding vectors, and $(D^{(k)})^{+}$ is its Moore–Penrose pseudo-inverse. This pseudo-inverse step is equivalent to solving a least-squares regression problem, in which the test mutants (matrix B) and test environments (matrix C) are projected into the training-dimensional space identified by SVD.

## 4.4 Predicting fitness in double-drug conditions from single-drug data (Figure 2)

**Training data:** In Figure 2, we implemented bi-cross validation (described above) while using eight single-drug environments in the training data. These environments included three concentrations of fluconazole (Low, Medium, and High) and three concentrations of radicicol (Low, Medium, and High), together with two control environments (No Drug and DMSO). We also

included 100 mutants in the training data, selected randomly from the full set of 774. The selection of training mutants was resampled 100 times, and all subsequent procedures were repeated each time, to ensure that results did not depend on the particular mutants chosen.

**Test data:** In each iteration, the remaining 674 mutants were held out as the test set, as were the four multidrug environmental conditions (LRLF, LRHF, HRLF, HRHF).

**Choosing the optimal number of phenotypic dimensions:** In each iteration, we chose the number of latent phenotypic dimensions ($r$) that minimized the mean squared error on the fitness predictions. We chose this approach because prediction accuracy initially improves as dimensions are added because biologically meaningful patterns are captured. However, the prediction accuracy eventually begins to decline with $k$ when additional phenotypic dimensions incorporate noise rather than signal. Therefore, the optimal number of dimensions was identified by minimizing mean squared error (MSE), as in previous work [9]. The optimal number of dimensions varied across our 100 iterations, as did the fitness predictions for each held out mutant in each held out environment. We averaged the fitness predictions pertaining to each mutant across all iterations where it was held out. Comparing these averages to the measured fitness values provides the robust estimates of predictive performance that we report in Figures 2D–E

### 4.5 Evaluating the effect of environmental complexity on predictive power (Figure 3)

To test whether genotype–fitness relationships are organized by a small set of phenotypic modules, we asked how quickly prediction accuracy improves, and then plateaus, as (i) we add more training environments and (ii) more latent dimensions are used in the model. Rapid saturation along either axis is expected if a limited set of phenotypic processes consistently shapes fitness across conditions.

We performed the analysis on two independent datasets including the one that is the focus of the rest of the study (774 drug-resistant mutants $\times$ 12 drug environments) [11] and our previous dataset which surveys fitness across a larger number of environments (292 low-glucose-evolved mutants $\times$ 45 diverse environments) [9].

Since we wanted to focus on the effects of altering the number of training environments, we held the number of training mutants constant. We also balanced the training mutant to contain an even number from different mutants classes. Using balanced mutants reduces the chance that missing a mutant class masquerades as a need for more environments. From the 774 drug-resistant mutants, we chose 100 mutants that evenly represent previously defined mutant clusters with characteristic patterns of G$\times$E [11]. From the 292 low-glucose-evolved mutants, we chose the same balanced set as was used previously, which includes 60 mutants balanced across genotypes [9]. These training mutant sets were fixed (unlike in Figure 2, where training mutants were repeatedly re-sampled at random).

Here, the training environments were repeatedly re-sampled at random. For each dataset with $n$ total environments, we varied the number of training environments ($k$) from 1 to $(n-1)$. For each $k$, we drew 100 random subsets of $k$ environments for training; the remaining $n-k$ environments were held out for testing. For each iteration, we performed bi-cross validation (see above) choosing the number of latent phenotypic dimensions ($r$) that minimized the mean squared error on the fitness predictions. We then recorded the corresponding $R^2$ between predicted and measured fitness for all test mutants in the test environments. Averaging $R^2$ across 100 iterations for each level of training environments (for $k = 1$ through $k = n-1$) gives $R^2$ for each level of training environments, which is displayed in the rightmost plots in Figure 3. Averaging across all iterations for each number of optimal dimensions (for $r = 1$ through $r =$ maximum $r$) gives $R^2$ for each number of optimal dimensions, which is displayed in the leftmost plots in Figure 3. To capture both effects together, we also plotted the average

$R^2$ as a function of both $k$ and $r$; these central surface plots show how predictive accuracy rises quickly with additional environments and dimensions before reaching a broad plateau.

### 4.6 Evaluating the effect of environmental complexity on number of dimensions (Figure S2)

To continue examining the effect of adding training environments on model complexity, we performed a similar analysis to that in Figure 3 in Figure S2. Here, instead of evaluating how predictive power ($R^2$) changes as we add training environments, we evaluate how the number of dimensions required for a reconstruction of the training data changes as we add training environments (Figure S2). For each number of environments ($k = 2$ through $k = n - 1$), we determined the minimum number of phenotypic dimensions necessary to capture 95% of the data variability. For robustness, each incremental increase in environment count was iterated 100 times, with a different random subset of environments selected for each iteration. Since this was a reconstruction, no mutant sampling was performed, all 774 or 292 mutants were included, depending on the dataset. The number of dimensions required to reconstruct the training data increased with the number of environments more quickly in the dataset pertaining to the 774 drug-resistant mutants $\times$ 12 drug environments [11]. The same result was true when we hold the number of environments constant (at $n$) and test the effect of increasing the number of mutants on model complexity. The dataset pertaining to drug resistant mutants requires more dimensions to capture 95% of the variation in fitness data than does the dataset pertaining to mutants evolved under glucose limitation (Figure S2).

### 4.7 Evaluating each training environment's influence on prediction accuracy (Figure 4)

To assess how certain types of training environments impact predictive performance of SVD, we trained our model with smaller training sets including some that omitted radicicol, some that omitted fluconazole, some that omitted the control conditions, and some other combinations (Figure 4A). To assess how specific training environments impact the predictive performance of our SVD model, we systematically included and excluded each single-drug environment from the training set and calculated the prediction accuracy using the coefficient of determination ($R^2$). We assessed the improvement ($\Delta R^2$) in prediction accuracy for each double-drug environment that resulted upon adding individual single-drug environments to the 7-environment training set. We do so 100 times, each time using 100 random mutants in our training set. For each mutant, we averaged the resulting fitness predictions. then calculated the $\Delta R^2$ values between each mutant's average prediction and measured fitness. The values on the vertical axis of Figure 4B represent these averages.

After identifying that including the low fluconazole environment in the training set notably improved predictions in the LRLF condition, we further analyzed which mutants contributed most significantly to this improvement. Specifically, we a identified two groups of mutants: the 50 mutants whose fitness was underpredicted when low fluconazole was excluded and the 50 mutants whose fitness was overpredicted when low fluconazole was excluded. We examined their gene-by-environment interaction profiles and observed characteristic behaviors that echo those detected in an earlier study using different methods (Figure S3) [11].

To robustly evaluate the effect of low fluconazole, we extended our drop-out analysis beyond single environments. In addition to testing the effect of adding one environment at a time to the baseline training set, we also systematically added all unique pairs and triplets of environments. For each case, the improvement in prediction accuracy ($\Delta R^2$) was recalculated relative to the 5- or 6- environment baseline (Figure S4).

### 4.8 Using SVD as an interpretation tool rather than for making predictions (Figure 5–6)

We conducted a comprehensive dimensionality analysis using the entire dataset to evaluate the interpretative power of singular value decomposition (SVD) in capturing phenotypic drivers of fitness. Rather than using bi-cross-validation, here we applied SVD directly to the entire fitness dataset, without separating training and test subsets, to identify abstract phenotypic dimensions. We performed SVD on the complete data matrix (774 mutants by 12 environments) and identified the number of phenotypic dimensions needed to explain at least 95Analysing sparsity with SSD models (Figure 5) To enhance our ability to interpret the biological meaning of the inferred phenotypic dimensions, and confirm the robustness of the identified dimensions, we sought to resolve the inherent ambiguity in standard SVD. While SVD excels at identifying the optimal low-dimensional subspace that captures the maximal variance in fitness, its solution is rotationally invariant. This means that any orthogonal rotation of its basis vectors (the phenotypic axes) within that subspace is equally valid from a reconstruction standpoint, leaving the specific axes returned by the algorithm biologically arbitrary. To arrive at a more biologically meaningful set of axes, one could introduce a guiding prior that selects for a particular kind of structure. Here, we adopt the principle of sparsity as a biologically motivated prior, operationalized through the Sparse Structure Discovery (SSD) framework (Petti et al. 2023). This approach assumes that the genotype-phenotype-fitness map is not a dense, all-to-all network. Instead, it posits that fitness in any given environment is shaped by only a small subset of core phenotypic axes (environment-sparsity), and conversely, that any given mutation perturbs only a few of these axes (mutant-sparsity). This framework translates the search for a sparse decomposition into a concrete optimization problem. We seek to find matrices for environmental weights (W) and mutant weights (M) that minimize the objective function:

$$C(W, M) = ||F - WM||_F^2 + \lambda_W ||W||_1 + \lambda_M ||M||_1 \tag{3}$$

where the first term is the squared Frobenius norm representing the reconstruction error, and the subsequent terms are $L_1$-norm penalties that enforce sparsity on $W$ and $M$, respectively. The regularization parameters, $\lambda_W$ and $\lambda_M$, allow us to tune the trade-off between reconstruction accuracy and the desired level of sparsity. Crucially, the appropriateness of imposing such a sparsity prior is not an axiom but an empirically testable hypothesis. We validated this assumption for our dataset using the rotation test [15]. This procedure tests whether sparsity is an intrinsic, basis-dependent property of the data. By applying a random orthogonal rotation to the fitness matrix (e.g., F' = FO), we preserve its underlying correlational geometry and SVD reconstruction error while shattering any privileged sparse basis. If the original data matrix F can be approximated by a substantially sparser model than its rotated counterpart F' for a given error budget, it provides strong evidence for an inherent sparse structure. Our analysis revealed a striking asymmetry in the structure of our system. The rotation test for the environment-phenotype map demonstrated a significant sparsity-error gap, confirming that fitness in each environment is indeed governed by a small number of phenotypic axes (Figure 5C). In contrast, the test for the mutant-phenotype map showed no clear gap; the original and rotated data were statistically less distinguishable, indicating a less sparse mapping (Figure 5A). This finding suggests that the drug-resistant mutations in our dataset have pervasive pleiotropy, each influencing nearly all of the underlying phenotypic axes. Informed by these results, we leveraged SSD to derive a model that respects this observed asymmetrical structure, yielding a more constrained but not necessarily more interpretable representation of the genotype-phenotype-fitness architecture (Figure S7).

### 4.9 Visualizing loading matrices to interpret the meaning of each phenotypic axis (Figure 6)

To facilitate biological interpretation, we computed the loading matrices, which represent the contribution (weights) of each mutant and environment to these four abstract phenotypic dimensions. Specifically, mutant loadings were calculated by multiplying the mutant singular vectors ($M$) by the square root of the diagonal singular-value matrix ($\Sigma$), and environmental loadings were calculated by multiplying the environmental singular vectors ($E$) by the square root of $\Sigma$. These loading matrixes allowed us to interpret each phenotypic dimension regarding underlying biological mechanisms related to drug resistance (Figure 6).

## References


[1] Michael Lässig, V. Mustonen, and A. M. Walczak. Predicting evolution. *Nature Ecology & Evolution*, 1:77, 2017.

[2] Y. Eguchi, G. Bilolikar, and K. Geiler-Samerotte. Why and how to study genetic changes with context-dependent effects. *Current Opinion in Genetics & Development*, 58-59:95–102, 2019.

[3] M. T. Wortel, D. Agashe, S. F. Bailey, et al. Towards evolutionary predictions: Current promises and challenges. *Evolutionary Applications*, 16:3–21, 2023.

[4] H. Zhang, M. Kabir, S. Ahmed, and M. Vihinen. There will always be variants of uncertain significance. analysis of vuss. *NAR Genomics and Bioinformatics*, 6:lqae154, 2024.

[5] F. Khodaee, R. Zandie, and E. R. Edelman. Multimodal learning for mapping genotype-phenotype dynamics. *Nature Computational Science*, 5:333–344, 2025.

[6] D. Houle, D. R. Govindaraju, and S. Omholt. Phenomics: the next challenge. *Nature Reviews Genetics*, 11:855–866, 2010.

[7] A. B. Paaby and M. V. Rockman. The many faces of pleiotropy. *Trends in Genetics*, 29: 66–73, 2013.

[8] K. A. Geiler-Samerotte, S. Li, C. Lazaris, et al. Extent and context dependence of pleiotropy revealed by high-throughput single-cell phenotyping. *PLoS Biology*, 18:e3000836, 2020.

[9] G. Kinsler, K. Geiler-Samerotte, and D. A. Petrov. Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. *eLife*, 9, 2020. doi: 10.7554/eLife.61271.

[10] S.-A. A. Chen, A. F. Kern, R. M. L. Ang, et al. Gene-by-environment interactions are pervasive among natural genetic variants. *Cell Genomics*, 3(100273), 2023.

[11] K. Schmidlin, S. Apodaca, D. Newell, et al. Distinguishing mutants that resist drugs via different mechanisms by examining fitness tradeoffs across hundreds of fluconazole-resistant yeast strains, 2024.

[12] S. Manrubia, J. A. Cuesta, J. Aguirre, et al. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Physics of Life Reviews*, 38: 55–106, 2021.

[13] J. Otwinowski, D. M. McCandlish, and J. B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115:E7550–E7558, 2018.

[14] S. M. Ardell and S. Kryazhimskiy. The population genetics of collateral resistance and sensitivity. *eLife*, 10, 2021. doi: 10.7554/eLife.73250.

[15] S. Petti, G. Reddy, and M. M. Desai. Inferring sparse structure in genotype-phenotype maps. *Genetics*, 2023. doi: 10.1093/genetics/iyad127.

[16] O. M. Ghosh, G. Kinsler, B. H. Good, and D. A. Petrov. Low-dimensional genotype-fitness mapping across divergent environments suggests a limiting functions model of fitness. bioRxiv, 2025.

[17] G. P. Wagner and L. Altenberg. Perspective: Complex adaptations and the evolution of evolvability. *Evolution*, 50:967–976, 1996.

[18] G. P. Wagner, M. Pavlicev, and J. M. Cheverud. The road to modularity. *Nature Reviews Genetics*, 8:921–931, 2007.

[19] W. G. Hill and X.-S. Zhang. On the pleiotropic structure of the genotype–phenotype map and the evolvability of complex organisms. *Genetics*, 190:1131–1137, 2012.

[20] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169:1177–1186, 2017.

[21] C. Pál, B. Papp, and V. Lázár. Collateral sensitivity of antibiotic-resistant microbes. *Trends in Microbiology*, 23:401–407, 2015.

[22] D. Nichol, J. Rutter, C. Bryant, et al. Antibiotic collateral sensitivity is contingent on the repeatability of evolution. *Nature Communications*, 10:334, 2019.

[23] D. Melo, A. Porto, J. M. Cheverud, and G. Marroig. Modularity: genes, development and evolution. *Annual Review of Ecology, Evolution, and Systematics*, 47:463–486, 2016.

[24] T. Praczyk. Emerging modularity during the evolution of neural networks. *Journal of Artificial Intelligence and Soft Computing Research*, 13:107–126, 2023.

[25] Z. R. Sailer and M. J. Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205:1079–1088, 2017.

[26] P. Avasthi, D. G. Mets, and R. York. Harnessing genotype-phenotype nonlinearity to accelerate biological prediction, 2023.

[27] K. Rijal, C. M. Holmes, S. Petti, et al. Inferring genotype-phenotype maps using attention models. arXiv, 2025.

[28] C. Martí-Gómez, J. Zhou, W.-C. Chen, et al. Inference and visualization of complex genotype-phenotype maps with gpmap-tools. bioRxiv, 2025.

[29] K. Schmidlin, C. B. Ogbunugafor, and K. Geiler-Samerotte. Environment by environment interactions (exe) differ across genetic backgrounds (exexg). bioRxiv, 2024.

[30] Q. Zhang and M. E. Andersen. Dose response relationship in anti-stress gene regulatory networks. *PLoS Computational Biology*, 3:e24, 2007.

[31] L. Perfeito, S. Ghozzi, J. Berg, et al. Nonlinear fitness landscape of a molecular pathway. *PLoS Genetics*, 7:e1002160, 2011.

[32] M. Srivastava and J. L. Payne. On the incongruence of genotype-phenotype and fitness landscapes. *PLoS Computational Biology*, 18:e1010524, 2022.

[33] S. F. Levy, J. R. Blundell, S. Venkataram, et al. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519:181–186, 2015.

[34] H. A. Orr. Adaptation and the cost of complexity. *Evolution*, 54:13–20, 2000.

[35] X. Liu, Y. I. Li, and J. K. Pritchard. Trans effects on gene expression can drive omnigenic inheritance. *Cell*, 177:1022–1034.e6, 2019.

[36] K. Ružičková, M. Hledík, and G. Tkačik. Quantitative omnigenic model discovers interpretable genome-wide associations. bioRxiv, 2024.

[37] Art B. Owen and Patrick O. Perry. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3:564–594, 2009.