Published on Sep 22, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-5953-995f

Harnessing genotype-phenotype nonlinearity to accelerate biological prediction

It is commonly assumed that phenotypes arise from the cumulative effects of many independent genes. However, we show that by accounting for dependent and nonlinear biological relationships, we can generate models that predict phenotypes with great accuracy.

Contributors (A-Z)

Prachee Avasthi, Feridun Mert Celebi, Megan L. Hochstrasser, David G. Mets, Ryan York

Version 4 · Mar 31, 2025

Purpose

A core focus of genetics is understanding the relationship between genetic variation (genotypes) and biological traits (phenotypes). Efforts as diverse as tracing the evolution of complex phenotypes, identifying disease-causing genes, and understanding how organisms are built are all contingent on deciphering the mapping between genotype and phenotype.

Our results show that assumptions underlying many current genotype-phenotype models (namely that genotypes are additive and linear) do not reflect the nonlinearities present in biology. Non-additive relationships between genes are well known — one gene can influence the effects of another (epistasis), and some genes have multiple phenotypic effects (pleiotropy). By accounting for such nonlinear interactions between genes and phenotypes, we show that we can accurately predict suites of simulated phenotypes.

These findings should be of interest to anyone whose work relies on accurately modeling genotype-phenotype relationships, especially those in the fields of quantitative, population, and human genetics. Additionally, we are excited to get feedback on how this work might help contribute to these fields and possible refinements or extensions of its utility.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.
- All associated **code** is available in this GitHub repository.
- **Data** from this pub, including **empirical** and **simulated phenotypes**, are available on Zenodo.

Background and goals

For several centuries now, scientists have attempted to decode how biology emerges from genetic information. Some of the models that have come out of this assume that traits are associated with infinitesimally complex genetic bases **[1]**; others hold that distinct clusters of few, but highly impactful genes drive each aspect of an organism's growth and function **[2]**. Many (if not most) of these models share a common feature: no matter their complexity, phenotypes can be understood by simply adding up the effects of their genetic contributors.

A single human trait best demonstrates this tendency: height. Dozens, if not hundreds, of genetic studies have been conducted with the goal of mining ever greater amounts of genetic "gold dust" **[2]** predicting variation in height **[3][4]**. Through these efforts,

two things have become clear: 1) height is highly heritable (> 80% by a recent study) **[4]** but 2) so much of the genome is involved that identifying a discrete molecular basis seems extremely unlikely.

In response to findings such as these (along with those gleaned from other human traits), some researchers have begun favoring the use of "polygenic (risk) scores" **[5]**. By aggregating the effects of many genomic loci, these scores can explain increasing amounts of trait variation (albeit at the cost of biological interpretability). Similarly, the "omnigenic model" **[6]** proposes that small sets of core, trait-determining genes work in parallel with many other "peripheral" genes. Through their sheer number, these peripheral loci thus also substantially contribute to trait variation. Importantly, in this model, contributions of core and peripheral genes are entangled and can't be distinguished, ultimately implicating vast swaths of the genome. Both polygenic risk scores and the omnigenic model assume the same inevitable conclusion: identifying the molecular drivers of complex traits is exceedingly difficult, if not impossible **[7]**. But what if the problem isn't how we are conceiving of the genetic bases of complex traits; what if the problem actually has to do with how we are thinking of traits *themselves*?

Height really *is* a complex trait, one that involves a myriad of interacting biological processes (development, metabolism, physiology, and so on). Each process is, in turn, regulated by its own sets of genes and it is likely that at least some of these gene sets relate to each other in complicated ways. Indeed, many complex phenotypes result from a variety of interconnected, nonlinear biological networks and genotypes that can relate to each other in a variety of directions (e.g., linear, nonlinear) and manners (e.g., additive, subtractive, dominant) **[8][9]**.

We believe that these points, if considered seriously, may have substantial implications for genetics. How often are phenotypes and genotypes nonlinearly correlated relative to being purely additive? Can information about any one phenotype help to predict another? Does accounting for phenotypic relationships increase predictive power? With these questions in mind, we decided to see if an approach explicitly capturing the complex relationships among phenotypes might provide some useful insights for genetic analysis writ large. Specifically, we focus on simultaneous analysis of groups of multiple phenotypes (here referred to as "**polyphenotypes"**). We examine how levels of **pleiotropy** (the impact of single genes on multiple phenotypes) and **gene-gene interaction** (the non-linear impact of combinations of genes on phenotypes) structure polyphenotypes.

Polyphenotype

Any grouping of multiple organismal phenotypes. Here, we argue that polyphenotypes have multiple uses. On one hand, they can be used to understand the relationships between multiple phenotypes. At the same time, they are also useful for developing accurate predictions of any single phenotype (by allowing one to control for potential false positives/negatives arising from phenotypic relationships).

The approach

For reasons both causal and correlative, phenotypes co-vary. For example, as referenced above, height is likely correlated with other phenotypes such as mass or metabolic rate. In this pub, we quantify the nature and prevalence of phenotype-phenotype relationships within large groups of phenotypes (what we refer to as polyphenotypes) to gain insight into the processes that cause these phenotypes. We find non-linear relationships among phenotypes in natural populations to be widespread. Furthermore, we find that, in simulations, the degree of non-linear phenotypes is modulated by the degree of gene-gene interaction and pleiotropy. We then demonstrate that, where present, phenotype-phenotype relationships can be leveraged to increase prediction accuracy for individual phenotypes.

Data collection/generation

All the data we used to study empirical variation across sets of phenotypes are publicly available. Sources and details for these data are available in Table 1. We chose data sets on the basis of phenotype number, sample size, and population type. We sought data sets in which a minimum of 15 phenotypes were measured for at least 100 individuals of the same species or interspecies cross. We also generated a set of random, unrelated phenotypes to compare with the observed phenotypic relationships contained within these datasets. To do so, for a single “phenotype”, we randomly generated integer values (values could be any integer between 1 and 1,000) 600 times. This process thus resulted in 600 simulated “individuals”, each with a randomly chosen phenotypic value. This was repeated to ultimately generate 30 simulated

phenotypes, each composed of 600 individual observations. After filtering on data completeness (see below for details on data set-specific filtering) we imputed missing values using the mean value for each phenotype and then performed rank normalization using the R function `RankNorm` from the package RNOmni.

Below are descriptions of data set-specific filtering. We tailored filtering parameters to each study given variation in sample size and the rate of missing data.

Arabidopsis: We excluded individuals if they had NAs for more than 20 phenotypic measurements. Similarly, we excluded phenotypes with more than 20 NAs. In addition, we removed non-continuous phenotypes (at least five unique values required per phenotype).

Yeast: We removed non-continuous phenotypes (at least five unique values required per phenotype).

C. elegans: We removed non-continuous phenotypes (at least five unique values required per phenotype).

Mouse (AIL): We removed non-continuous phenotypes (at least five unique values required per phenotype).

Mouse (JAX): We excluded samples that were missing more than 100 measurements. Similarly, we excluded phenotypes missing more than 100 measurements. In addition, we removed non-continuous phenotypes (at least five unique values required per phenotype).

Fruit fly: We excluded samples that were missing more than 50 measurements. Similarly, we excluded phenotypes missing more than 50 measurements. We reduced the dimensionality of gene expression values from Huang et al. 2015 **[10]** using PCA (we extracted the first 30 PCs). In addition, we removed non-continuous phenotypes (at least five unique values required per phenotype).

Name	Main reference	Type	N samples	N phenos
Arabidopsis	**[11]**	Natural strains (accessions)	514	110
Yeast	**[12]**	F1 segregant	13,950	40

Name	Main reference	Type	N samples	N phenos
C. elegans	**[13]**	Recombinant inbred lines (RIL)	2,017	19
Mouse (AIL)	**[14]**	Advanced intercross line (AIL)	1,063	133
Mouse (JAX)	**[15][16][17]**	Laboratory strains	106	271
Fruit fly	**[18]**	Inbred lines	147	270
Random	This pub	–	600	30

Table 1. Data sets we used for empirical analyses of nonlinearity.

All of these data are available on Zenodo.

We simulated 100 phenotypes for 121 populations (N individuals per population = 500). These populations were created by first simulating genetic data and deriving the phenotypes from these genotypes. For each individual, we randomly assigned one of three allelic states at each of 300 loci (e.g., homozygous reference, heterozygous, homozygous alternate). Then, we generated a genetic architecture for each phenotype by randomly assigning 100 loci to that phenotype and giving each possible allele at each locus a weight of influence between zero and 10.

We modeled effects of pleiotropy and gene-gene interaction on phenotypes, varying the impact of each systematically across populations such that each population had a unique pairing of the probability pleiotropy and the probability of gene-gene interactions. These probabilities were per gene-phenotype pair or gene-gene pair and ranged from 0–1 in increments of 0.1, thus forming an 11 by 11 grid with one simulated population for each pairing. For example: population 1 has probabilities *P(pleiotropy)* = 0, *P(epistasis)* = 0; population 2 has probabilities *P(pleiotropy)* = 0.1, *P(epistasis)* = 0; and so on.

To model pleiotropy, for each individual population for each phenotype, we assigned each locus already determined to influence a phenotype (100 loci per phenotype, 9900 locus-phenotype pairs) to be involved in pleiotropy with a population specific-probability as defined above. If we determined the locus-phenotype pair to be involved in pleiotropy, the weights assigned to that locus were included in the calculation of that

phenotype. Similarly, to create gene-gene interactions (e.g., epistasis) that varied across populations, we assigned each gene-gene pair (N = 4,950) to be involved in an interaction with a population-specific probability as defined above. If we determined a locus was involved in an interaction, we randomly assigned that interaction to one of the six possible pairs of alleles (i.e., interaction among loci here occurs only between single pairs of alleles). We then multiplied the weights of those alleles. Finally, we calculated the phenotypes for each individual by summing the weights at loci influencing that phenotype.

Creating the autoencoder

We implemented a neural network called a denoising autoencoder to test the utility of examining multiple phenotypes for phenotypic prediction **[19]**. Autoencoders consist of two networks — first, an encoder that forces the data through an information bottleneck, and then a decoder that takes information compressed through that bottleneck and tries to reconstruct the data. Accurate reconstruction of the data following the compression through the information bottleneck suggests that the network has learned a representation of that data. During training, noise is added to the input data, preventing a common failure mode in which the learned representation does not extrapolate to data that was not in the training set; the learned model fails to generalize.

> All **code** associated with this pub — including **analysis notebooks**, the **synthetic phenotype generator**, and the **autoencoder model** — is available in this GitHub repository (DOI: 10.5281/zenodo.8371249).

Briefly, our autoencoder consisted of an encoder with two fully connected rectified linear layers that we subjected to batch normalization and a similarly structured decoder. The latent space separating the two networks contained 32 nodes. We conducted the training with phenotypic data from 80% of the simulated individuals over 100 epochs with a batch size of 16. To all training data, we added 0.1 standard deviation of noise. Following training, we predicted phenotypes on the remaining 20% of the data. To evaluate the utility of increasing the number of phenotypes under different values of pleiotropy and interaction, we trained individual models using 5, 10, 20, and 30 phenotypes, and evaluated the model accuracy on five phenotypes. We

calculated prediction error as the mean absolute percentage error. We implemented the autoencoder in PyTorch **[20]**.

Analysis of empirical phenotypes

After filtering, imputation, and rank normalization (see "Data collection/generation" section) we computed the frequency of nonlinear phenotypic relationships for each data set. To do so, we fit a linear and a nonlinear model for all possible phenotypic pairs within the data set. We generated the linear model with a linear regression (`lm` function in R). W generated the nonlinear model using a generative additive model (`gam` function in the R package mgcv) with a single smoothing spline term (via the mgcv function `s`). We compared model fits using the Akaike information criterion (AIC) and considered three possible outcomes: a tie (equal AIC), the nonlinear model is a better fit (nonlinear = lower AIC), or the linear model is a better fit (linear = lower AIC). We then calculated the frequency of nonlinearity from the ratio of the number of cases in which the nonlinear model had lower AIC compared to the full number of phenotypic comparisons.

Given the possible diversity of phenotypic relationships within any given data set, and to facilitate the measurement of variance in nonlinearity rates, we used a permutation-based approach to calculate nonlinearity across subsets of each data set. To do so, we calculated the nonlinearity rate for 1,000 random sets of phenotypes for each data set (data proportion per random set = 0.25). We visualized this distribution using violin plots (as in Figure 1, A). We then measured the variation of these permutation distributions using the R function `var` (as in Figure 1, B) and calculated the correlation between all phenotype pairs using Pearson's correlation (as in Figure 1, C).

Analysis of synthetic phenotypes

To further dissect patterns of phenotypic nonlinearity, we generated 10,201 phenotypic matrices spanning possible combinations between gene-gene interaction and pleiotropy probabilities (each ranging from zero to one in increments of 0.01). We measured the nonlinearity rate of each phenotypic matrix using the same approach outlined above. We visualized the distribution of nonlinearity as a function of gene-gene interaction and pleiotropy by creating a generalized additive model (GAM). Here,

nonlinearity was treated as a response variable predicted by gene-gene interaction and pleiotropy and was implemented using the `gam` function in the R package mgcv (as in Figure 2, B). The predicted nonlinearity values are visualized in two dimensions, representing all possible combinations gene-gene interaction and pleiotropy probabilities.

We next wanted to characterize the entropy of full phenotypic data sets. Taking influence from the phenotypic integration literature **[21]**, we first calculated the "generalized variance" (the determinant of the variance-covariance matrix) for each phenotypic matrix. Generalized variance is a useful measure in that it allows us to directly compare phenotypic data sets with different dimensionalities **[21]**. To extract a single-vector descriptor, we then calculated the eigenvector of the generalized variance matrix using spectral decomposition (`eigen` R function). We then calculated the entropy of the leading eigenvector using the R function `entropy.empirical` from the R package entropy **[22]**. We could thus use the resulting entropy estimate to infer the overall information contained among an arbitrarily large set of phenotypic measurements.

We next developed a method to infer the correlational structure of a phenotypic set by calculating entropy across increasingly large, random subsets of phenotypes. Broadly, this method sweeps through pre-set portions of a data set, randomly selects a set of phenotypes for each portion, and calculates entropy using the method described above. We applied this test to phenotypic matrices with varying probabilities of pleiotropy (probability zero to one, 0.01 increments) by calculating entropy for increasingly large proportions of samples (10% to 90%, 10% increments). For each portion, we analyzed 10 permuted sets of phenotypes and calculated their mean entropy. The results of this analysis appear in Figure 3, A. We extracted slopes of the resulting entropy distributions from a linear regression (`lm` function in R) comparing pleiotropy probability and entropy (Figure 3, B).

> **SHOW ME THE DATA**: You can find all the data used in this pub, including **empirical** and **simulated phenotypes**, on Zenodo (DOI: 10.5281/zenodo.8298808).

Autoencoder analyses

We calculated autoencoder prediction error as the mean absolute percent error between prediction and ground truth. We conducted autoencoder training using 80% of the individuals in the data set and evaluated accuracy on the remaining 20%.

We calculated entropy as above using the same parameters (portions: 10% to 90% of samples in 10% increments; 10 permutations per portion).

> All **code** associated with this pub — including **analysis notebooks**, the **synthetic phenotype generator**, and the **autoencoder model** — is available in this GitHub repository (DOI: 10.5281/zenodo.8371249).

The results

Nonlinearity is prevalent among biological traits

To our knowledge, it remains unclear just how common additivity and linearity are in genetic systems. To address this, we compiled a data set of "**polyphenotypes**" (see definition above) from a diverse set of interbreeding species populations (see Approach for details). We reasoned that inferring the rate of nonlinear phenotypic relationships would allow us to glean how well linear/additive models would fit these populations.

Using a simple test (see Approach) we determined the (non)linearity of pairwise phenotypic relationships within each species population. We found that all species display rates of nonlinearity that are significantly greater than expected by chance ($p < 0.001$, Kruskal-Wallis test) (Figure 1, A), ranging from 43.5% (fruit flies) to 84.4% (*Arabidopsis*) (Figure 1, A). These observations support the idea that nonlinearity is a prevalent feature of biological phenotypes and contributes to a substantial portion of species' phenotypic relationships.

The range of nonlinearity also differs greatly across populations. For example, nematodes display almost 10× more variation in phenotypic relationships than fruit

flies (*C. elegans* = 0.19, fruit flies = 0.029; mean normalized standard deviation) (Figure 1, B), while randomly generated data display the greatest degree of relative variation (0.39; mean normalized standard deviation). Interestingly, these randomly generated data should be largely independent of each other and, thus, may be considered representative of a set of non-pleiotropic, additive traits. Supporting this idea, we found that the mean pairwise correlation of the random phenotypes is significantly less than that of the species data (Figure 1, C). Overall, these observations suggest that complex aspects of phenotypic relationships may be inferred using a set of relatively simple descriptive statistics.

However, given their heterogeneity, determining how epistasis and pleiotropy might affect the frequency of phenotypic nonlinearity is hard using these datasets. Some data come from advanced genetic crosses (e.g. the DGRP and JAX data) while other data sets sampled variants from a diverse natural population (*Arabidopsis*). In addition, the polyphenotypes reflect the interests of the original studies and, therefore, occupy somewhat random and undetermined regions of phenotypic space. Therefore, while it is apparent that nonlinearity exists in a variety of quantitative genetic data sets, it is difficult to use these data to develop strong intuitions about its sources.

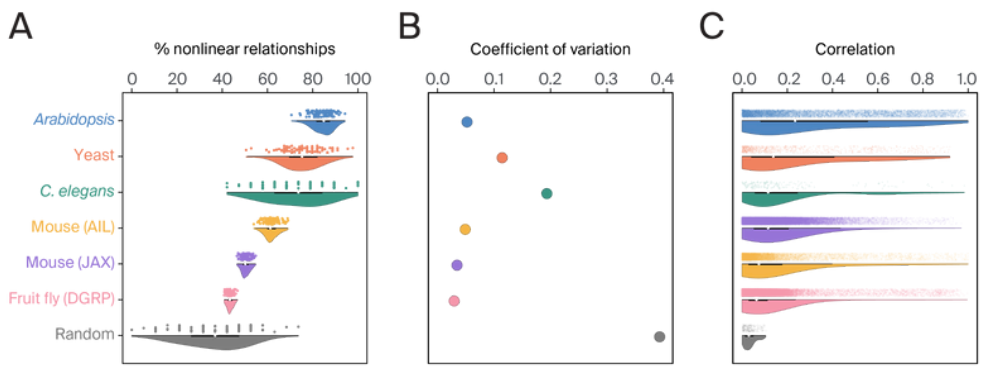

Figure 1

Characterizing phenotypic nonlinearity across species.

(A) Half-violin plots comparing the percentage of nonlinear relationships among phenotypes for six species and a random data set. For each species, we present full sets of data points to the left and show the distribution in the half-violin plot on the right.

(B) The degree of variation in nonlinear relationships for each species.

(C) Half-violin plots comparing the distribution of phenotypic correlations for each species.

Synthetic phenotypes let us interrogate nonlinear effects

With this in mind, we sought to control many of these covariates to allow more direct interrogation of pleiotropy, epistasis, and phenotypic structure. Using a novel approach, we generated a series of polyphenotypes from simulated genotypic data. Briefly, we generated *n* random genotypes from a probability distribution for a given number of individuals (see Approach for a more in-depth description). Each genotype could influence an output phenotype given a set probability distribution and could interact with others via a predefined probability. We also allowed genes to influence more than one phenotype with a set probability, letting us vary the amount of epistasis (probability of gene-gene interactions) and pleiotropy (probability of phenotype-phenotype interactions) in the data. Using this approach, we generated a data set in

which all combinations of epistatic and pleiotropic probabilities were considered (from P = 0 to P = 1, 0.01 increments). This produced a final set of 10,201 polyphenotypes, each containing 20 synthetic phenotypes measured across 600 simulated individuals.

A main goal in generating this synthetic data set was for it to capture a broad range of nonlinear relationships. To assess how well the data set accomplishes this, we used the same test as above (see Figure 1; Approach) to calculate the rate of nonlinearity for each of the 10,201 polyphenotypes. Notably, these percentages span the values observed among the empirical phenotypes, with a mean nonlinearity rate of 47.53% (min = 16%, max = 100%; Figure 2, A; Figure 1, A). In addition, nonlinearity varies smoothly across the distribution of pleiotropy/gene interaction probabilities (Figure 2, B). Together, these observations suggest that our data generation approach successfully produced a naturalistic range of nonlinearity from which to sample.

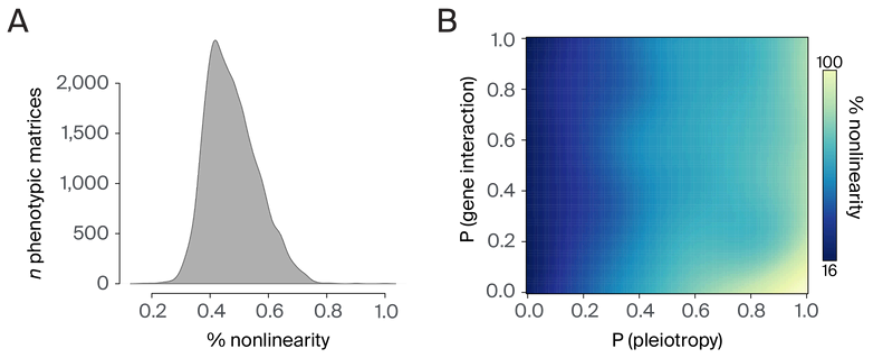

Figure 2

The distribution of nonlinearity within a simulated phenotypic data set.

(A) Histogram of nonlinearity for all 10,201 simulated phenotypic data matrices.

(B) The 2D distribution of nonlinearity (represented by color) as a function of pleiotropy and gene-gene interaction probabilities. We calculated the distribution using a generative additive model.

Entropy and nonlinearity capture diverse phenotypic interactions

How can we best identify drivers of biological nonlinearity? Can we statistically decouple the effects of different genetic and phenotypic interactions? Motivated by our efforts to apply information theory to genetics (more on this in a companion pub coming soon), we hypothesized that we may start to untangle some of the drivers of nonlinearity by measuring the information content (or "entropy" as defined in information theory) of polyphenotypes. We reasoned that entropy may be informative in multiple ways. First, overall entropy reflects the interrelatedness of polyphenotypes. Lower values may reflect a set of phenotypes that are driven by the same underlying biology (e.g., multiple, correlated measurements of a trait such as finger length). On the other hand, higher values may indicate that polyphenotype data contain measurements from multiple, orthogonal features of biology (e.g., finger length and education level). The second way entropy may be informative is through its distribution across different portions of a polyphenotype. Consider the case of completely orthogonal phenotypes. If we select random combinations of orthogonal phenotypes and measure their entropy, it should be the case that entropy proportionally increases as we analyze larger and larger sets of phenotypes (i.e., more new information is being added with each increase in the number of randomly chosen phenotypes). In contrast, for a set of strongly correlated phenotypes (e.g., in the case of pleiotropy), one should expect entropy to stay constant as we analyze larger sets of the phenotypes (i.e., no new information is added).

Applying this framework to all 10,201 polyphenotypes, we calculated entropy across increasing proportions of randomly selected phenotypes (see Approach). We found that entropy distributions strongly vary with the probability of pleiotropy. Increasing pleiotropy equates with a flattening of the distribution (Figure 3, A). This point is further demonstrated by an extremely strong relationship between entropy distribution slopes and pleiotropy (Pearson's $r = -0.95$; Figure 3, B). These observations support the notion that entropy is a reliable measure of the interrelatedness of a set of phenotypes. What's more, this suggests that, by analyzing the within-polyphenotype distribution of entropy, we may infer the amount of phenotypic pleiotropy with minimal knowledge of the underlying genetics.

Are similar measures available for determining the frequency of gene-gene interactions? Taking a hint from the previously identified relationship between

nonlinearity and gene-gene interactions/pleiotropy in Figure 2, we found that nonlinearity varies strongly in the absence of gene-gene interactions but decreases in dynamic range as interaction probability increases (Figure 3, C). Furthermore, comparing the relationship between the entropy slope and percentage of nonlinearity reveals an interesting trade-off between gene-gene interactions and pleiotropy (Figure 3, D). There is an overall negative relationship (Pearson's $r = -0.82$) suggesting that, as pleiotropy increases, so too do nonlinear phenotypic interactions. Furthermore, as gene-gene interactions increase (as indicated by point color in Figure 3, D), the variance of entropy/nonlinearity relationships decreases.

Taken together, these results suggest that pleiotropy leads to increasingly nonlinear phenotypic relationships, especially in the absence of genetic interactions. Furthermore, we can study this trade-off via entropy and nonlinearity, which are both non-genetic measures. Finally, these patterns indicate that phenotypic nonlinearity — like that observed both here and among real phenotypes (Figure 1) — also reflects genetic nonlinearities, hinting at potential insufficiencies of additive/linear models for capturing the genetic components of biological traits.

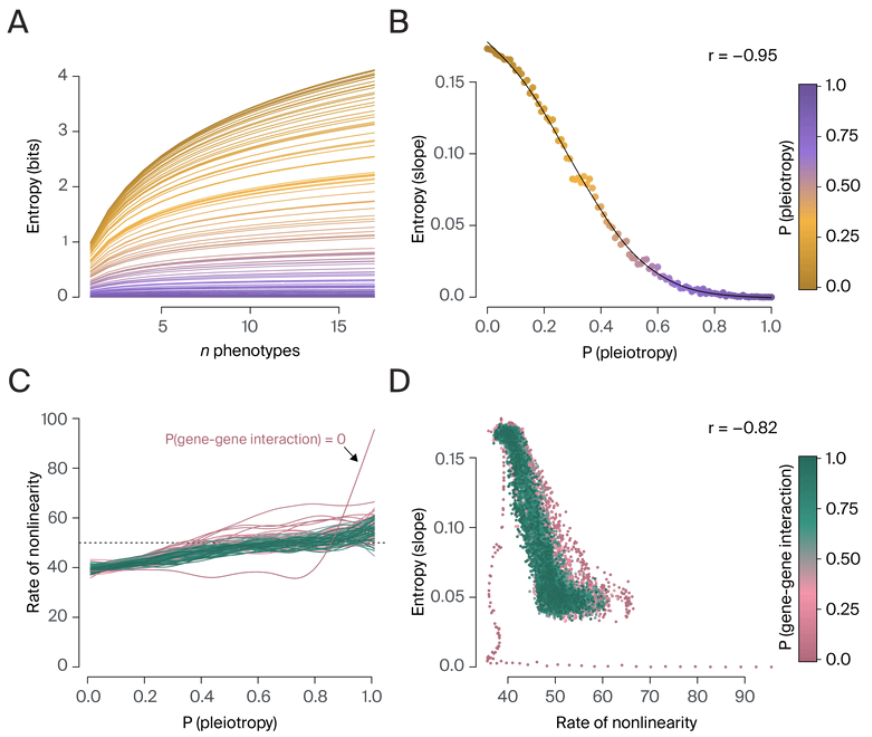

Figure 3

Characteristics of gene interaction and pleiotropy variation.

(A) Shannon entropy calculated across data set proportions for all pleiotropy probabilities. Entropy is measured in bits.

(B) Scatter plot of the slopes for each distribution plotted in Figure 3, A. r: Pearson's correlation.

(C) The percentage of nonlinear interactions between phenotypes as a function of pleiotropy probability (x-axis).

(D) The joint distribution of entropy slope and % nonlinear interactions for all 10,201 synthetic phenotype matrices. r: Pearson's correlation.

We indicate pleiotropy probability (A and B) and the probability of genetic interactions (C and D) using a color scale.

Accounting for phenotypic nonlinearity significantly increases predictive power

If genetic nonlinearities are truly prevalent across many types of biological traits, which types of models might be better suited for capturing their effects? Neural network-based strategies are enticing options for several reasons. Neural networks inherently learn nonlinearities across their layers, letting them model complex interactions between inputs and outputs (e.g., between phenotypes and genotypes). In addition, they can model multiple inputs and outputs, facilitating nonlinear mapping of multiple phenotypes at once. We therefore hypothesized that neural network strategies might help us determine the benefit of accounting for complex, nonlinear interactions between phenotypes.

To do this, we constructed an autoencoder for modeling and predicting phenotypic relationships (see Approach; Figure 4, A). Taking simulated polyphenotypes as input, the model encoded phenotypic relationships into a lower-dimensional latent space and generated predictions via a decoder (Figure 4, A). We used this strategy to predict aspects of all 10,201 polyphenotypes. Specifically, we generated four sets of predictions for each, varying the number of input phenotypes (n = 5, 10, 20, 30) used to predict an output set (n = 5; Figure 4, A). We then assessed the accuracy of each model by calculating the percent error between observed and predicted phenotypes.

We found that the autoencoder approach predicted phenotypes with extremely high accuracy (mean error = 1.76%; Figure 4, B) and that models become more accurate when the number of input phenotypes increases (Figure 4, B). In fact, all models using more than five input phenotypes display significantly decreased error distributions (Kruskal-Wallis test followed by Dunn's test; Figure 4, B). It's also apparent that the error distributions themselves display a degree of heterogeneity, with some clear outliers displaying error percentages above 4% (Figure 4, B). Plotting percent error as a function of pleiotropy shows that these outliers are associated with cases in which the probability of pleiotropic interactions was very low (Figure 4, C), indicating that pleiotropic interactions can help increase the accuracy of multi-phenotype models. More broadly, it's apparent that by accounting for nonlinearities such as pleiotropy, autoencoder strategies are able to predict individual phenotypes with great accuracy.

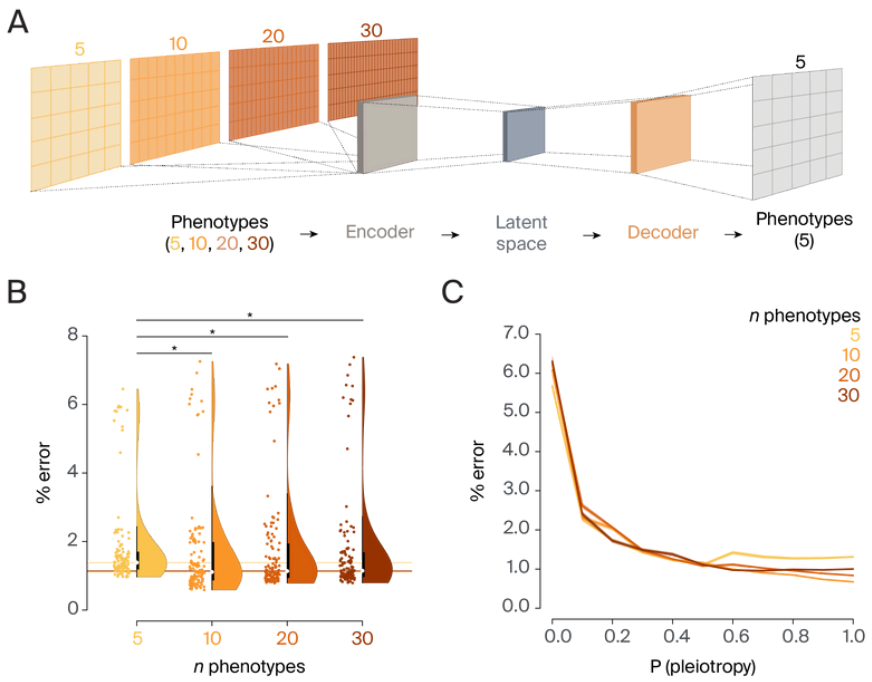

Figure 4

An autoencoder for phenotypic prediction.

(A) Cartoon overview of the autoencoder architecture and analytical design. We used phenotypic matrices of different sizes (5, 10, 20, 30 phenotypes) as input to predict an output set of five phenotypes.

(B) Violin plots comparing percent error for all predictions from the four different model classes (5, 10, 20, 30 phenotypes). Asterisks indicate significant differences according to a Kruskal-Wallis test followed by Dunn's test.

(C) Model error as a function of pleiotropy probability for the four model classes.

Finally, we explored whether these models could generate realistic polyphenotypes. To do so, we measured the entropy of each set of predicted phenotypes. Overall entropy decreases as the probability of pleiotropy increases (Figure 5, A), reflecting the patterns observed among the input phenotypes (Figure 4, A). There is also a similarly tight relationship between percent error and entropy across all models (Pearson's r = 0.78; Figure 5, B). When comparing mean percent error and entropy slope, we found a

strong separation between the five-phenotype model and all others (Figure 5, C). Models with more input phenotypes display lower entropy slopes and greater degrees of model accuracy. Furthermore, the 10-phenotype model displayed the lowest error rate and entropy slope (Figure 5, C), a pattern that is also apparent in the comparisons of percent error across the models (Figure 4, B). It is interesting to consider that this may reflect something important about the structure of the synthetic phenotypes. Specifically, the 10-phenotype model may represent a better trade-off between input phenotype information content and the overall model complexity (i.e., the 20- and 30-phenotype models may just be adding redundant information).

In total, these findings suggest that the autoencoder did indeed create realistic polyphenotypes with expected entropy distributions. Given this, we conclude that models 1) accommodating polyphenotypic designs and 2) accounting for biological nonlinearity provide opportunities to greatly increase the predictive capacity of genetic analysis.

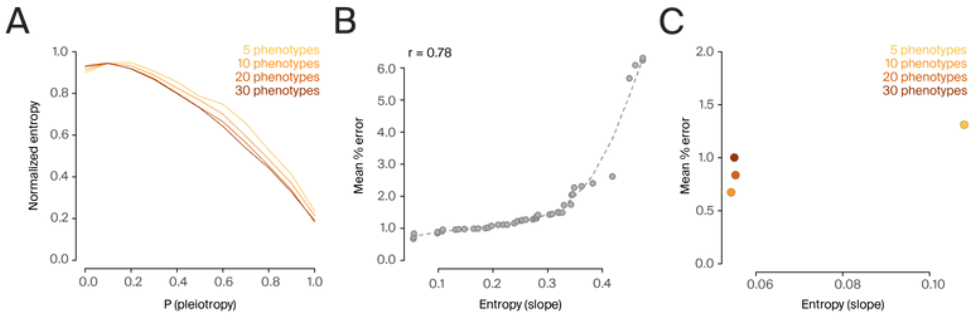

Figure 5

Characterizing autoencoder output.

(A) Entropy compared to pleiotropy probability for the four model classes.

(B) The relationship between entropy slope and model error across all runs. r: Pearson's correlation.

(C) Entropy slope compared to mean error for each model class.

Key takeaways

- Nonlinearity is a prevalent feature of biological phenotypes (Figure 1)
- Phenotypic nonlinearity varies as a function of genetic and phenotypic interactions (Figure 2)
- Measures from information theory, such as entropy, can quantify the structure of phenotypic interactions (Figure 3)
- Models that account for nonlinearity and phenotypic interactions have increased predictive potential and improve as the number of phenotypes increases (Figure 4)
- Model accuracy varies as a function of the information content of phenotype sets (Figure 5)

Implications

Biology is in an age of increasingly large, high-dimensional, and complex data sets. Endeavors such as AlphaFold are attempting to map the full universe of protein structures **[23][24]**. Similarly, a number of multi-team efforts are characterizing human cell type diversity via a host of omics and cell biological data types **[25]**. These data sets — and others like them — contain (or will contain) a diversity of phenotypic measurements possessing unknown and complex inter-relationships. A goal for many of these efforts will be to identify these relationships and, ultimately, use them to decode the function of complex biological systems (e.g., identifying how RNA expression, chromatin accessibility, and cell morphology interact to generate a cell type). This undertaking butts up against a statistical sampling problem: is there enough data available to power such analysis for the system you're interested in? Put another way, have you sampled enough of "phenotypic space" to account for the biology in question?

These are hard questions to answer *a priori*. However, asking them is useful. If it's possible to efficiently sample phenotypic space, minimizing measurement redundancy, then scalability and cost-effectiveness would correspondingly increase. It's interesting to consider how the aggregated results of this pub may help. Our framework predicts that samplings of different parts of phenotypic space should be associated with correspondingly variable parameter combinations (Figure 6). We'd predict that sampling a single phenotype would be associated with uniformly low

values of entropy, nonlinearity, and predictiveness (Figure 6, A). On the other hand, if multiple correlated phenotypes (perhaps due to pleiotropy) were sampled, the rate of nonlinearity and predictiveness would increase, but entropy would not (Figure 6, B; Figure 4, C). If multiple orthogonal phenotypes were measured, we'd find increased entropy and predictiveness and a modest amount of nonlinearity (Figure 6, C; Figure 4, C). Due to their numerical generality, it should be possible to measure the entropy and nonlinearity of most (if not all) polyphenotypes.

Given this, we propose that these measures may be implemented as a general-purpose toolkit for inferring the phenotypic structure, predictiveness, and even genetic patterns (i.e., gene-gene interactions and pleiotropy) associated with a given polyphenotypic data set. There are likely many useful extensions of this. For one, we can use phenotypic entropy to measure the complexity of a phenotypic data set. We could therefore determine if ongoing collection is adding new or informative dimensions to a data set. Similarly, we may use entropy and nonlinearity to estimate the number of generative biological processes associated with a polyphenotype (if one, then entropy should be low and nonlinearity high; if multiple, entropy will increase). Indeed, the entropy of a polyphenotype is the number of bits of information necessary to capture the phenotypic structure and, as a result, the generative processes that drive that structure. With these measures in hand, it's possible to hypothesize, *a priori,* the structure of genetic mapping results and, by factoring in these patterns, design studies around minimally necessary and maximally informative polyphenotypes.

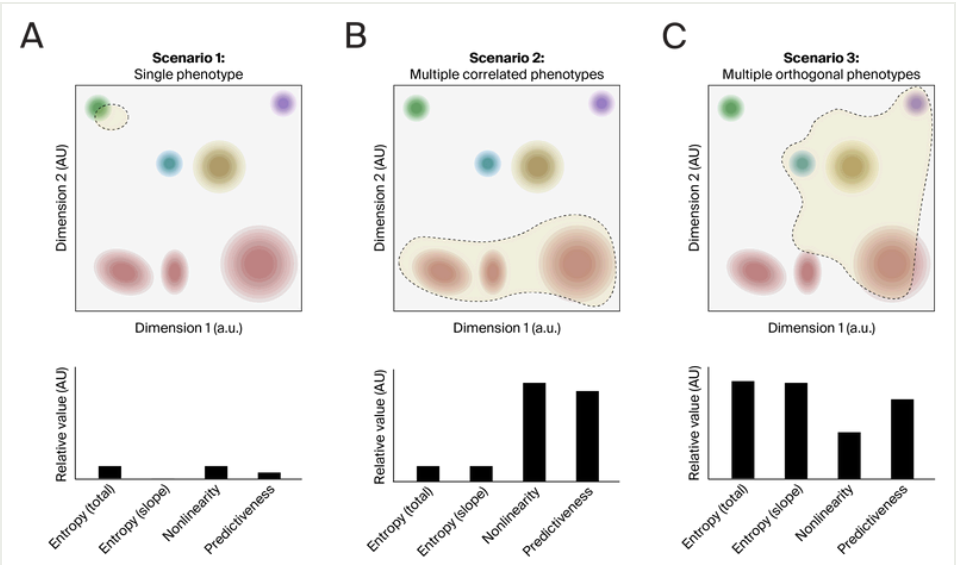

Figure 6

Mapping phenotypic space.

(A) (Top) A cartoon representation of phenotypic space. Here, we assume that we can use some form of high-dimensional measurement to separate and cluster phenotypes. Individual phenotypes are represented by colored shapes with increasing densities (darker color) toward the center of the shape. The amount of phenotypic space sampled is represented by the filled dotted line. In Scenario 1, measurements for only a single phenotype are available. (Bottom) Predicted measurements for total entropy, the slope of entropy, nonlinearity, and overall predictiveness (i.e., ability to predict a sampled phenotype given the amount of space that has been sampled). Values are represented as arbitrary units (AU).

(B) In Scenario 2, we've measured multiple correlated phenotypes (as reflected by their same color). In this case, nonlinearity increases due to the inter-phenotype correlation, as does predictiveness.

(C) In Scenario 3, we've sampled multiple orthogonal (differently colored) phenotypes, Here, entropy increases and nonlinearity is somewhat lower than in Scenario 2 (since orthogonal phenotypes tend to display higher rates of linear relationships). We can still predict phenotypes with accuracy, but

proportionally less than in the situation of multiple correlated phenotypes.

More generally, a “phenotype-forward” framework that allows for complex nonlinear relationships between traits (as we suggest here) has the prospect of reflecting organismal structure that is likely missed when we examine phenotypes individually. For example, modeling height and weight simultaneously likely provides more biological insight and predictive ability than modeling them independently, as some of their causal mechanisms are shared. The neural network method we use here explicitly captures these relationships and has the possibility of “encoding” the generative processes for sets of phenotypes with at least partially overlapping causes.

Treating the organism as a system in this way has the potential to answer more complex questions than modeling individual phenotypes alone. Such approaches may prove critical to leveraging the increasing amount of phenotypic data to achieve better biological understanding and outcomes across a host of problems.

References

1. Fisher RA. (1930). The genetical theory of natural selection. https://doi.org/10.5962/bhl.title.27468
2. Rockman MV. (2011). THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT’S GOLD DOES NOT GLITTER. https://doi.org/10.1111/j.1558-5646.2011.01486.x
3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. (2010). Common SNPs explain a large proportion of the heritability for human height. https://doi.org/10.1038/ng.608
4. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, Graff M, Eliasen AU, Jiang Y, Raghavan S, Miao J, Arias JD, Graham SE, Mukamel RE, Spracklen CN, Yin X, Chen S-H, Ferreira T, Highland HH, Ji Y, Karaderi T, Lin K,

Lüll K, Malden DE, Medina-Gomez C, Machado M, Moore A, Rüeger S, Sim X, Vrieze S, Ahluwalia TS, Akiyama M, Allison MA, Alvarez M, Andersen MK, Ani A, Appadurai V, Arbeeva L, Bhaskar S, Bielak LF, Bollepalli S, Bonnycastle LL, Bork-Jensen J, Bradfield JP, Bradford Y, Braund PS, Brody JA, Burgdorf KS, Cade BE, Cai H, Cai Q, Campbell A, Cañadas-Garre M, Catamo E, Chai J-F, Chai X, Chang L-C, Chang Y-C, Chen C-H, Chesi A, Choi SH, Chung R-H, Cocca M, Concas MP, Couture C, Cuellar-Partida G, Danning R, Daw EW, Degenhard F, Delgado GE, Delitala A, Demirkan A, Deng X, Devineni P, Dietl A, Dimitriou M, Dimitrov L, Dorajoo R, Ekici AB, Engmann JE, Fairhurst-Hunter Z, Farmaki A-E, Faul JD, Fernandez-Lopez J-C, Forer L, Francescatto M, Freitag-Wolf S, Fuchsberger C, Galesloot TE, Gao Y, Gao Z, Geller F, Giannakopoulou O, Giulianini F, Gjesing AP, Goel A, Gordon SD, Gorski M, Grove J, Guo X, Gustafsson S, Haessler J, Hansen TF, Havulinna AS, Haworth SJ, He J, Heard-Costa N, Hebbar P, Hindy G, Ho Y-LA, Hofer E, Holliday E, Horn K, Hornsby WE, Hottenga J-J, Huang H, Huang J, Huerta-Chagoya A, Huffman JE, Hung Y-J, Huo S, Hwang MY, Iha H, Ikeda DD, Isono M, Jackson AU, Jäger S, Jansen IE, Johansson I, Jonas JB, Jonsson A, Jørgensen T, Kalafati I-P, Kanai M, Kanoni S, Kårhus LL, Kasturiratne A, Katsuya T, Kawaguchi T, Kember RL, Kentistou KA, Kim H-N, Kim YJ, Kleber ME, Knol MJ, Kurbasic A, Lauzon M, Le P, Lea R, Lee J-Y, Leonard HL, Li SA, Li Xiaohui, Li Xiaoyin, Liang J, Lin H, Lin S-Y, Liu Jun, Liu X, Lo KS, Long J, Lores-Motta L, Luan J, Lyssenko V, Lyytikäinen L-P, Mahajan A, Mamakou V, Mangino M, Manichaikul A, Marten J, Mattheisen M, Mavarani L, McDaid AF, Meidtner K, Melendez TL, Mercader JM, Milaneschi Y, Miller JE, Millwood IY, Mishra PP, Mitchell RE, Møllehave LT, Morgan A, Mucha S, Munz M, Nakatochi M, Nelson CP, Nethander M, Nho CW, Nielsen AA, Nolte IM, Nongmaithem SS, Noordam R, Ntalla I, Nutile T, Pandit A, Christofidou P, Pärna K, Pauper M, Petersen ERB, Petersen LV, Pitkänen N, Polašek O, Poveda A, Preuss MH, Pyarajan S, Raffield LM, Rakugi H, Ramirez J, Rasheed A, Raven D, Rayner NW, Riveros C, Rohde R, Ruggiero D, Ruotsalainen SE, Ryan KA, Sabater-Lleal M, Saxena R, Scholz M, Sendamarai A, Shen B, Shi J, Shin JH, Sidore C, Sitlani CM, Slieker RC, Smit RAJ, Smith AV, Smith JA, Smyth LJ, Southam L, Steinthorsdottir V, Sun L, Takeuchi F, Tallapragada DSP, Taylor KD, Tayo BO, Tcheandjieu C, Terzikhan N, Tesolin P, Teumer A, Theusch E, Thompson DJ, Thorleifsson G, Timmers PRHJ, Trompet S, Turman C, Vaccargiu S, van der Laan SW, van der Most PJ, van Klinken JB, van Setten J, Verma SS, Verweij N, Veturi Y, Wang CA, Wang C, Wang L, Wang Z, Warren HR, Bin Wei W, Wickremasinghe AR, Wielscher M, Wiggins KL, Winsvold BS, Wong A, Wu Y, Wuttke M, Xia R, Xie T, Yamamoto K, Yang Jingyun, Yao J, Young H, Yousri NA, Yu L, Zeng L, Zhang W, Zhang X, Zhao J-H, Zhao W, Zhou W, Zimmermann ME, Zoledziewska M, Adair LS, Adams HHH, Aguilar-Salinas CA, Al-Mulla F, Arnett DK, Asselbergs FW, Åsvold BO, Attia J, Banas B, Bandinelli S, Bennett DA, Bergler T, Bharadwaj D, Biino G, Bisgaard H, Boerwinkle E, Böger CA, Bønnelykke K, Boomsma DI, Børglum AD, Borja JB, Bouchard C, Bowden DW,

Brandslund I, Brumpton B, Buring JE, Caulfield MJ, Chambers JC, Chandak GR, Chanock SJ, Chaturvedi N, Chen Y-DI, Chen Z, Cheng C-Y, Christophersen IE, Ciullo M, Cole JW, Collins FS, Cooper RS, Cruz M, Cucca F, Cupples LA, Cutler MJ, Damrauer SM, Dantoft TM, de Borst GJ, de Groot LCPGM, De Jager PL, de Kleijn DPV, Janaka de Silva H, Dedoussis GV, den Hollander AI, Du S, Easton DF, Elders PJM, Eliassen AH, Ellinor PT, Elmståhl S, Erdmann J, Evans MK, Fatkin D, Feenstra B, Feitosa MF, Ferrucci L, Ford I, Fornage M, Franke A, Franks PW, Freedman BI, Gasparini P, Gieger C, Girotto G, Goddard ME, Golightly YM, Gonzalez-Villalpando C, Gordon-Larsen P, Grallert H, Grant SFA, Grarup N, Griffiths L, Gudnason V, Haiman C, Hakonarson H, Hansen T, Hartman CA, Hattersley AT, Hayward C, Heckbert SR, Heng C-K, Hengstenberg C, Hewitt AW, Hishigaki H, Hoyng CB, Huang PL, Huang W, Hunt SC, Hveem K, Hyppönen E, Iacono WG, Ichihara S, Ikram MA, Isasi CR, Jackson RD, Jarvelin M-R, Jin Z-B, Jöckel K-H, Joshi PK, Jousilahti P, Jukema JW, Kähönen M, Kamatani Y, Kang KD, Kaprio J, Kardia SLR, Karpe F, Kato N, Kee F, Kessler T, Khera AV, Khor CC, Kiemeney LALM, Kim B-J, Kim EK, Kim H-L, Kirchhof P, Kivimaki M, Koh W-P, Koistinen HA, Kolovou GD, Kooner JS, Kooperberg C, Köttgen A, Kovacs P, Kraaijeveld A, Kraft P, Krauss RM, Kumari M, Kutalik Z, Laakso M, Lange LA, Langenberg C, Launer LJ, Le Marchand L, Lee H, Lee NR, Lehtimäki T, Li H, Li L, Lieb W, Lin X, Lind L, Linneberg A, Liu C-T, Liu Jianjun, Loeffler M, London B, Lubitz SA, Lye SJ, Mackey DA, Mägi R, Magnusson PKE, Marcus GM, Vidal PM, Martin NG, März W, Matsuda F, McGarrah RW, McGue M, McKnight AJ, Medland SE, Mellström D, Metspalu A, Mitchell BD, Mitchell P, Mook-Kanamori DO, Morris AD, Mucci LA, Munroe PB, Nalls MA, Nazarian S, Nelson AE, Neville MJ, Newton-Cheh C, Nielsen CS, Nöthen MM, Ohlsson C, Oldehinkel AJ, Orozco L, Pahkala K, Pajukanta P, Palmer CNA, Parra EJ, Pattaro C, Pedersen O, Pennell CE, Penninx BWJH, Perusse L, Peters A, Peyser PA, Porteous DJ, Posthuma D, Power C, Pramstaller PP, Province MA, Qi Q, Qu J, Rader DJ, Raitakari OT, Ralhan S, Rallidis LS, Rao DC, Redline S, Reilly DF, Reiner AP, Rhee SY, Ridker PM, Rienstra M, Ripatti S, Ritchie MD, Roden DM, Rosendaal FR, Rotter JI, Rudan I, Rutters F, Sabanayagam C, Saleheen D, Salomaa V, Samani NJ, Sanghera DK, Sattar N, Schmidt B, Schmidt H, Schmidt R, Schulze MB, Schunkert H, Scott LJ, Scott RJ, Sever P, Shiroma EJ, Shoemaker MB, Shu X-O, Simonsick EM, Sims M, Singh JR, Singleton AB, Sinner MF, Smith JG, Snieder H, Spector TD, Stampfer MJ, Stark KJ, Strachan DP, 't Hart LM, Tabara Y, Tang H, Tardif J-C, Thanaraj TA, Timpson NJ, Tönjes A, Tremblay A, Tuomi T, Tuomilehto J, Tusié-Luna M-T, Uitterlinden AG, van Dam RM, van der Harst P, Van der Velde N, van Duijn CM, van Schoor NM, Vitart V, Völker U, Vollenweider P, Völzke H, Wacher-Rodarte NH, Walker M, Wang YX, Wareham NJ, Watanabe RM, Watkins H, Weir DR, Werge TM, Widen E, Wilkens LR, Willemsen G, Willett WC, Wilson JF, Wong T-Y, Woo J-T, Wright AF, Wu J-Y, Xu H, Yajnik CS, Yokota M, Yuan J-M, Zeggini E, Zemel BS, Zheng W, Zhu X, Zmuda JM, Zonderman AB, Zwart J-A, Partida GC, Sun Y, Croteau-Chonka D,

Vonk JM, Chanock S, Le Marchand L, Chasman DI, Cho YS, Heid IM, McCarthy MI, Ng MCY, O'Donnell CJ, Rivadeneira F, Thorsteinsdottir U, Sun YV, Tai ES, Boehnke M, Deloukas P, Justice AE, Lindgren CM, Loos RJF, Mohlke KL, North KE, Stefansson K, Walters RG, Winkler TW, Young KL, Loh P-R, Yang Jian, Esko T, Assimes TL, Auton A, Abecasis GR, Willer CJ, Locke AE, Berndt SI, Lettre G, Frayling TM, Okada Y, Wood AR, Visscher PM, Hirschhorn JN. (2022). A saturated map of common genetic variants associated with human height. https://doi.org/10.1038/s41586-022-05275-y

5 Wray NR, Lee SH, Mehta D, Vinkhuyzen AAE, Dudbridge F, Middeldorp CM. (2014). Research Review: Polygenic methods and their application to psychiatric traits. https://doi.org/10.1111/jcpp.12295

6 Boyle EA, Li YI, Pritchard JK. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. https://doi.org/10.1016/j.cell.2017.05.038

7 Sella G, Barton NH. (2019). Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. https://doi.org/10.1146/annurev-genom-083115-022316

8 Mackay TFC. (2013). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. https://doi.org/10.1038/nrg3627

9 Kauffman SA. (1993). The Origins of Order: Self-organization and Selection in Evolution. https://books.google.com/books/about/The_Origins_of_Order.html?id=lZcSpRJz0dgC

10 Huang Y, Ng FS, Jackson FR. (2015). Comparison of Larval and Adult *Drosophila* Astrocytes Reveals Stage-Specific Gene Expression Profiles. https://doi.org/10.1534/g3.114.016162

11 Team 500 Genomes Field Experiment. (n.d.). Natural selection on the Arabidopsis thaliana genome in present and future climates. https://doi.org/10.1038/s41586-019-1520-9

12 Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, Kruglyak L. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. https://doi.org/10.7554/elife.49212

13 Snoek BL, Volkers RJM, Nijveen H, Petersen C, Dirksen P, Sterken MG, Nakad R, Riksen JAG, Rosenstiel P, Stastna JJ, Braeckman BP, Harvey SC, Schulenburg H, Kammenga JE. (2019). A multi-parent recombinant inbred line population of C. elegans allows identification of novel QTLs for complex life history traits. https://doi.org/10.1186/s12915-019-0642-8

14 Gonzales NM, Seo J, Hernandez Cordero AI, St. Pierre CL, Gregory JS, Distler MG, Abney M, Canzar S, Lionikas A, Palmer AA. (2018). Genome wide association analysis in a mouse advanced intercross line. https://doi.org/10.1038/s41467-018-07642-8

15 Bogue MA, Churchill GA, Chesler EJ. (2015). Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. https://doi.org/10.1007/s00335-015-9595-6

16 Bogue MA, Philip VM, Walton DO, Grubb SC, Dunn MH, Kolishovski G, Emerson J, Mukherjee G, Stearns T, He H, Sinha V, Kadakkuzha B, Kunde-Ramamoorthy G, Chesler EJ. (2019). Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. https://doi.org/10.1093/nar/gkz1032

17 Bogue MA, Ball RL, Philip VM, Walton DO, Dunn MH, Kolishovski G, Lamoureux A, Gerring M, Liang H, Emerson J, Stearns T, He H, Mukherjee G, Bluis J, Desai S, Sundberg B, Kadakkuzha B, Kunde-Ramamoorthy G, Chesler EJ. (2022). Mouse Phenome Database: towards a more FAIR-compliant and TRUST-worthy data repository and tool suite for phenotypes and genotypes. https://doi.org/10.1093/nar/gkac1007

18 Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RRH, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu L-L, Qu C, Ràmia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu Y-Q, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA. (2012). The Drosophila melanogaster Genetic Reference Panel. https://doi.org/10.1038/nature10811

19 Vincent P, Larochelle H, Bengio Y, Manzagol P-A. (2008). Extracting and composing robust features with denoising autoencoders. https://doi.org/10.1145/1390156.1390294

20 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/ARXIV.1912.01703

21 O'Keefe FR, Meachen JA, Polly PD. (2021). On Information Rank Deficiency in Phenotypic Covariance Matrices. https://doi.org/10.1093/sysbio/syab088

22 Hausser J, Strimmer K. (2008). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. https://doi.org/10.48550/ARXIV.0811.3579

23 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

24 Barrio-Hernandez I, Yeo J, Jänes J, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. (2023). Clustering predicted structures at the scale of the known protein universe. https://doi.org/10.1101/2023.03.09.531927

25 Fleck JS, Camp JG, Treutlein B. (2023). What is a cell type? https://doi.org/10.1126/science.adf6162