Genetics and Genomics
Evolutionary Biology

Epistasis facilitates functional evolution in an ancient transcription factor

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
March 22, 2024 (this version)

Reviewed preprint version 1
July 12, 2023

Sent for peer review
April 29, 2023

Posted to preprint server
April 20, 2023

Brian P.H. Metzger, Yeonwoo Park, Tyler N. Starr, Joseph W. Thornton

Department of Ecology and Evolution, University of Chicago, 60637 USA • Department of Biological Sciences, Purdue University • Program in Genetics, Genomics, and Systems Biology, University of Chicago, 60637 USA • Center for RNA Research, Seoul National University • Department of Biochemistry and Molecular Biophysics, University of Chicago, 60637 USA • Department of Biochemistry, University of Utah • Department of Human Genetics, University of Chicago, 60637 USA

Abstract

A protein's genetic architecture – the set of causal rules by which its sequence produces its functions – also determines its possible evolutionary trajectories. Prior research has proposed that genetic architecture of proteins is very complex, with pervasive epistatic interactions that constrain evolution and make function difficult to predict from sequence. Most of this work has analyzed only the direct paths between two proteins of interest – excluding the vast majority of possible genotypes and evolutionary trajectories – and has considered only a single protein function, leaving unaddressed the genetic architecture of functional specificity and its impact on the evolution of new functions. Here we develop a new method based on ordinal logistic regression to directly characterize the global genetic determinants of multiple protein functions from 20-state combinatorial deep mutational scanning (DMS) experiments. We use it to dissect the genetic architecture and evolution of a transcription factor's specificity for DNA, using data from a combinatorial DMS of an ancient steroid hormone receptor's capacity to activate transcription from two biologically relevant DNA elements. We show that the genetic architecture of DNA recognition consists of a dense set of main and pairwise effects that involve virtually every possible amino acid state in the protein-DNA interface, but higher-order epistasis plays only a tiny role. Pairwise interactions enlarge the set of functional sequences and are the primary determinants of specificity for different DNA elements. They also massively expand the number of opportunities for single-residue mutations to switch specificity from one DNA target to another. By bringing variants with different functions close together in sequence space, pairwise epistasis therefore facilitates rather than constrains the evolution of new functions.

eLife assessment

This study includes **fundamental** findings on protein evolution, namely that changes in function are largely attributable to pairwise rather than higher-order interactions, and that epistasis potentiates rather than constrains evolutionary paths. **Compelling** evidence supporting the conclusions is provided by applying a new model to a previously generated experimental dataset on deep mutational scanning of the DNA-binding domain (DBD) of steroid hormone receptor. The implications of this work are of considerable interest to protein biochemistry, evolutionary biology, and numerous other fields.

Introduction

A protein's genetic architecture – the set of causal rules by which its sequence determines its functions – is of fundamental importance in genetics, biochemistry, and evolution. These rules include the effect on function of every possible amino acid state at each site in the protein, and the epistatic effects of all possible pairs and higher-order combinations. This genetic architecture determines the distribution of functions across the space of all possible protein sequences and, in turn, the paths accessible to an evolving protein under the influence of mutation, drift, and selection.

The extent and impact of epistatic interactions is of particular interest. If the effects of an amino acid state depend strongly on the states present at other sites in the protein – and especially on higher-order combinations – then the functions of a protein could become difficult to predict from its sequence or those with similar sequences. Epistasis could also constrain evolution by creating a rugged functional landscape in sequence space, in which many optimal or near-optimal genotypes are inaccessible from others under most evolutionary scenarios (1–22). A related topic is potential bias in epistatic interactions: if they predominantly impair function, "diminishing the returns" of otherwise favorable mutations (23–31), then functional variants might be restricted to only those regions of sequence space where each site has the state or states with the most beneficial main effects, further constraining evolutionary paths.

Epistasis is clearly present in the genetic architecture of proteins (32–49), but its overall character and effects on evolutionary trajectories remain murky. Studies of pairs of mutations show that pairwise interactions can block some twostep paths around a designated wild-type protein (50–53) and experiments on higher-order combinations have found that epistasis can reduce the number of functional intermediates that connect a designated pair of starting and ending proteins (54–62). Most combinatorial studies, however, have assessed just two states per site (34, 35, 41, 48, 63–88), thus excluding most of the huge ensemble of potential genotypes and the evolutionary paths among them. Further, most of these studies have measured only a single function, even though many protein families contain members with distinct functional specificities, such as the capacity to bind different substrates or other ligands (but see (57, 89, 90) for recent exceptions); as a result, the genetic architecture of functional specificity and its influence on the evolution of new functions are poorly understood.

Characterizing the genetic architecture of functional specificity and its evolutionary impact requires three things: 1) an experiment that assays all combinations of 20 amino acid states at a designated set of sites for multiple functions; 2) a method to extract from these data the causal rules by which sequence determines function, and 3) a way to characterize the impact of those rules on trajectories through sequence space. Improved technologies have enabled complete scans

of all combinations of 20 amino acids, typically at three or four sites of a priori structural or functional interest (91–98). The key limiting factor has been the lack of effective methods to dissect genetic architecture from such datasets. Most methods for dissecting genetic architecture from combinatorial studies have been limited to just two states per site (but see (99–102)). Another concern is that most studies have modeled the effects of mutations relative to a particular wildtype reference sequence, which yields globally inaccurate estimates of the effects of mutations and combinations when they are introduced into other backgrounds (83).

Here we develop and implement a reference-free method for global dissection of 20-state sequence-function relationships and apply it to a combinatorial DMS dataset generated in our laboratory. The model's terms are encoded relative to the global functional average across all genotypes rather than a particular reference sequence, and they directly express the portion of all functional variation in a protein that is attributable to any particular set of genetic determinants. This way of encoding the model also allows us to directly assess the effect of epistasis on the distribution of multiple functions across sequence space and to assess their accessibility under various evolutionary scenarios.

The data we analyze were generated in a complete combinatorial scan at four sites that are critical for DNA recognition in the DNA-binding domain (DBD) of the reconstructed ancestral steroid hormone receptor (65, 96, 103). This experiment assayed the transcription factor's ability to activate transcription from two different biologically relevant DNA response elements, enabling us to characterize the rules of genetic architecture that define functional specificity. By applying our approach to a reconstructed ancestral protein, we were also able to assess how the protein's historical genetic architecture shaped the maintenance of the ancestral function and the accessibility of the derived function that actually evolved during steroid hormone receptor history.

Results

Experimental data

Steroid hormone receptors are a family of ligand activated transcription factors that are involved in vertebrate development, behavior, and reproduction (Bentley, 1998). There are two major subfamilies of steroid hormone receptors, which recognize distinct palindromic DNA response elements (REs). Estrogen receptors (ERs) bind to a palindrome of the ER response element (ERE, AGGTCA), whereas the ketosteroid receptors (kSRs, including the receptors for androgens, progestogens, glucocorticoids, and mineralocorticoids) strongly prefer the steroid receptor response element (SRE, AGAACA) (104–106). Earlier experiments showed that the reconstructed DBD of the last common ancestor of the two subfamilies (called AncSR1) was ERE-specific, whereas the last common ancestor of the kSRs (An-cSR2) was SRE-specific (103). During the interval between AncSR1 and AncSR2, three substitutions occurred in the protein's recognition helix (RH), which inserts into the DNA major groove and makes contact with the bases that vary between ERE and SRE. These substitutions were shown experimentally to cause the shift in specificity (103).

The data that we analyze here come from a previously published deep mutational scan of AncSR1 (96). The library contained all 160,000 combinations of 20 amino acids at four sites in the RH – the three historically substituted sites plus another site that is variable in closely related nuclear receptors. The library was transformed into yeast strains carrying either an ERE-or SRE-driven GFP reporter and functionally characterized using a FACS-based Sort-seq assay.

To reduce the effect of measurement noise on the characterization of genetic architecture in this project, we transformed the continuous functional data into categorical form: each variant was categorized on each RE as a null, weak, or strong activator relative to negative control (stop-codon-

containing variants) and historically-relevant positive control reference sequences (the RH sequence of AncSR1 and extant ERs on ERE, or that of AncSR2 and extant kSRs on SRE). On each RE, variants were classified as null if their mean fluorescence was indistinguishable from the negative controls, weak if they produced fluorescence between null and the historical reference state, and strong if their fluorescence was as great or greater than the reference. Across both REs, 1342 variants were classified as strong and 3166 as weak activators. Categorization substantially reduced the effect of technical noise in the assay: concordance in the activation class assigned to each variant between replicates was >97%, much better than the between-replicate correlation of mean fluorescence as a continuous variable (R^2 = 0.62 for functional variants, and R^2 = 0.11 when null variants are included).

Ordinal linear regression of genetic architecture

To dissect the genetic architecture of RE binding and specificity by AncSR-DBD, we developed an ordinal regression model and fit it to the experimental data. The model contains terms for the main effect of every possible amino acid at each variable site in the protein and for the epistatic effect of every pair and triplet of sites (**Fig. 1A**). The genetic score of each variant is defined as the sum of the main and epistatic effects of the sequence states and combinations it contains (**Fig. 1B**). The variant's genetic score determines the probability that a variant is a null, weak, or strong activator through an ordinal logistic function: an increase in the genetic score causes a corresponding linear increase in the log of the odds that a variant is in a higher vs. lower activation class. The only other free parameters in the model are the average genetic score across all variants and the threshold scores that represent the boundaries between activation classes.

To incorporate functional specificity, the model contains two terms for every amino acid state or combination: one for its effect on nonspecific DNA binding (the contribution to genetic score averaged over ERE and SRE) and another that reflects its impact on specificity (the difference between its score on each RE and the average over REs). The model also contains a single term for the global effect of SRE vs.

ERE, averaged across all protein variants, which represents the main effect of the RE itself. The total genetic score for any protein:RE complex is the sum of all the main and epistatic effects of its amino acids on nonspecific DNA binding, plus all the main and epistatic effects on RE specificity, plus the global effect of the RE. (The sign of the specificity and RE terms in the summation is determined by whether the complex contains SRE or ERE.)

The model is encoded to give accurate and efficient estimates of the rules of genetic architecture across the entire set of possible sequences. Rather than defining the effects of a state or combination relative to a designated wild-type sequence, we use a reference-free framework in which effects are defined relative to the global average across all variants (**Fig. 1C-D**; for a related approach using DNA sequences, see (107)). The main effect of any amino acid state is the mean genetic score of all variants containing that state minus the global average score. The epistatic effect of a pair of states is the mean genetic score of all variants containing that pair minus the sum of the global average and the main effects of the two states. And the third-order effect of a triplet of states is the average score of all variants containing that triplet minus the global average and the sum of the pairwise and main effects of the component states and pairs. We did not model fourth-order epistatic effects, because these cannot be distinguished from technical error in the assay measurements.

This approach has several advantages. First, it reduces the effect of estimation bias and measurement error, because each effect is averaged across a large number of observations from unique genotypes at other sites (e.g. 8000 for each main effect term, 400 for each pairwise effect) (83). Second, the model's form allows us to use an ANOVA-like framework to directly partition the genetic variance – the variance in genetic score across all protein variant-RE complexes – into that attributable to every causal factor in the model. The genetic variance attributable to any effect

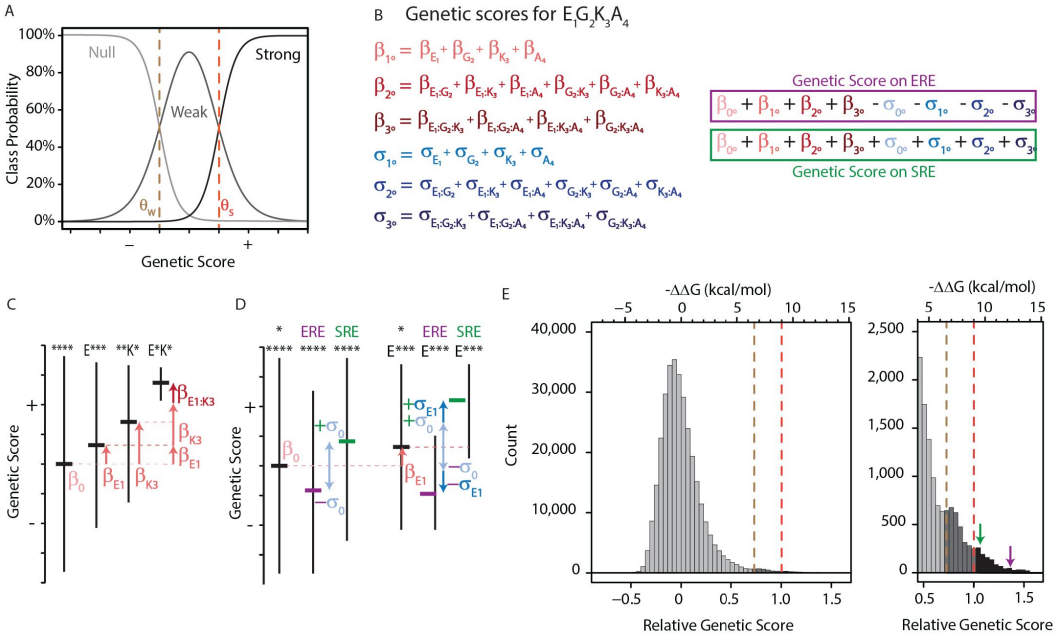

Fig. 1.

An ordinal regression model of the genetic architecture of transcription factor specificity.

A. The ordinal logistic model. The probability that a protein variant is in each binding class is a logistic function of its genetic score. Thresholds between classes (dotted lines) correspond to scores at which the probability of being in two classes is equal. **B**. A variant's genetic score is the sum of the effects of the sequence states and combinations it contains, including main (1º), pairwise (2º), and third-order (3º) effects. Each state and combination has a term for its effect on generic binding averaged over REs (βs, red) and another for its effect on specificity (σs, blue) for ERE vs. SRE; specificity terms are added or subtracted to give the genetic score on SRE or ERE, respectively. All terms are defined relative to the global average over all variants (0º). Components of the score for variant EGKA is shown as an example. **C**. Schematic of reference-free model terms. Each term represents the average deviation of sequences containing a state (or combination) from the sum of the variant's terms at lower orders, including the global average. Vertical and horizontal lines show range and mean of genetic scores for variants containing the specified states; *, any of the 20 amino acid states. **D**. Reference-free specificity effects. Binding effects (red) are defined the same as in C. The global specificity effect (σ_0) is the deviation from the global average for all genotypes on ERE (purple) vs. SRE (green). First order specificity effects (σ_1 s) are additional deviations from this global effect between ERE and SRE for genotypes with a particular amino acid at a particular site. **E**. Distribution of genetic scores after fitting to the experimental DMS data for all 320,000 protein:DNA combinations. Scores were rescaled so that the global average is zero and the threshold score to be classified as a strong binder is 1. Dashed lines, interclass thresholds. Inset, right tail of the distribution. Estimated -ΔΔG values of binding associated with each genetic score are shown along the top, based on calibration to experimental binding data for a subset of variants. Arrows, biological reference sequences GSKV (SRE binder, green) and EGKA (ERE binder, purple).

term in the model is the squared coefficient, weighted by the fraction of all genotypes to which it applies, and the variance explained by a set of terms (e.g., at a site or an epistatic order) is simply the sum of the variance accounted for by all the terms in the set (see Methods and Appendix 1). A third advantage is that the sigmoidal shape of the logistic function may incorporate global nonlinearity in the genotype-phenotype relationship, such as that imposed by limited dynamic range of measurement in the transcriptional assay; this is important, because failing to do so can lead to spurious findings of specific epistatic interactions (5, 20, 108–112).

We fit the model to the DMS data by least-squares ordinal logistic regression (**Fig. 1E**). To reduce overfitting and maximize predictive accuracy, we used L2 regularization with cross-validation to identify the optimal regularization parameters. The model fits the experimental data well. When we used the best-fit model to predict the highest-probability activation class for each protein variant on each RE, the predicted class matched the observed class for 99.5% of all variants. The deviance of the model – an analog of the coefficient of determination used in linear regression – was 0.86 (**Figure 1-Figure Supplement 1**, **Figure 1-Figure Supplement 2**, **Figure 1-Figure Supplement 3**). The remaining variation in function not explained by the model is attributable to technical error in measurement and/or fourth-order epistasis. The genetic score of activators is well correlated with the apparent ΔG of dissociation from DNA, which was previously measured for a subset of variants by fluorescence anisotropy (R^2 = 0.75, **Figure 1-Figure Supplement 4**; this correlation is slightly better than the correlation between the mean fluorescence values used to classify variants and ΔG, R^2 = 0.63).

Low-order genetic architecture

We found that main effects and pairwise epistasis are important in RE recognition, but higher order epistasis is not. Of all the genetic variance explained by the best-fit model, more than half is attributable to the main effects of single amino acids, and virtually all of the remainder is attributable to pairwise interactions or the global effect of the RE. Main effects are the primary determinants of nonspecific DNA binding, but pairwise interactions explain the majority of RE-specificity. Third-order epistasis accounts for only 2% of total variance, with similarly small contributions to both nonspecific binding and specificity (**Fig. 2A-C**). Of all the specific variance explained by epistatic interactions, about half involves sign epistasis – interactions that on average change the direction of a state's effects in the presence of another amino acid – while the other half changes the magnitude but not the direction of effect (**Figure 2D**). Although high-order interactions are unimportant, the model is very dense at low orders. Of the nearly 69,000 possible terms in the complete model, it takes >7,000 to account for 99% of genetic variance ("the 99% set", **Fig. 2E**), including the main effect of every possible amino acid state at all four sites and >80% of all possible second-order epistatic terms. Every amino acid state participates in a minimum of 60 pairwise interactions with states at other sites (**Figure 2-Figure Supplement 1**). By contrast, the 99% set includes only 5% of all possible third-order terms.

The non-sparsity of the DBD's genetic architecture reflects the fact that thousands of variants activate from ERE and/or SRE, and their sequences are very diverse (**Fig. 2F-G**). The effects of most amino acids and pairs are therefore of small to moderate magnitude, changing the genetic score by a median of 3% and 2% of the threshold required for strong activation, respectively. Third-order effects were generally tiny, with a median absolute magnitude of just 0.02% (**Fig. 2H-I**, **Figure 2-Figure Supplement 2**, **Figure 2-Figure Supplement 3**). As a consequence, no single amino acid state or pair is sufficient to make a sequence into a strong activator; rather, numerous favorable states and combinations of small to moderate size at three or all four of the sites are required to generate a functional protein.

The sequence-function relationship of DNA recognition is therefore complex but not idiosyncratic. The functions of any RH sequence can be predicted with good accuracy from its main and pairwise effects alone (**Figure 2-Figure Supplement 4**). Higher-order epistatic terms are largely dispensable, so prediction does not require precise knowledge of the effects of the vast number of

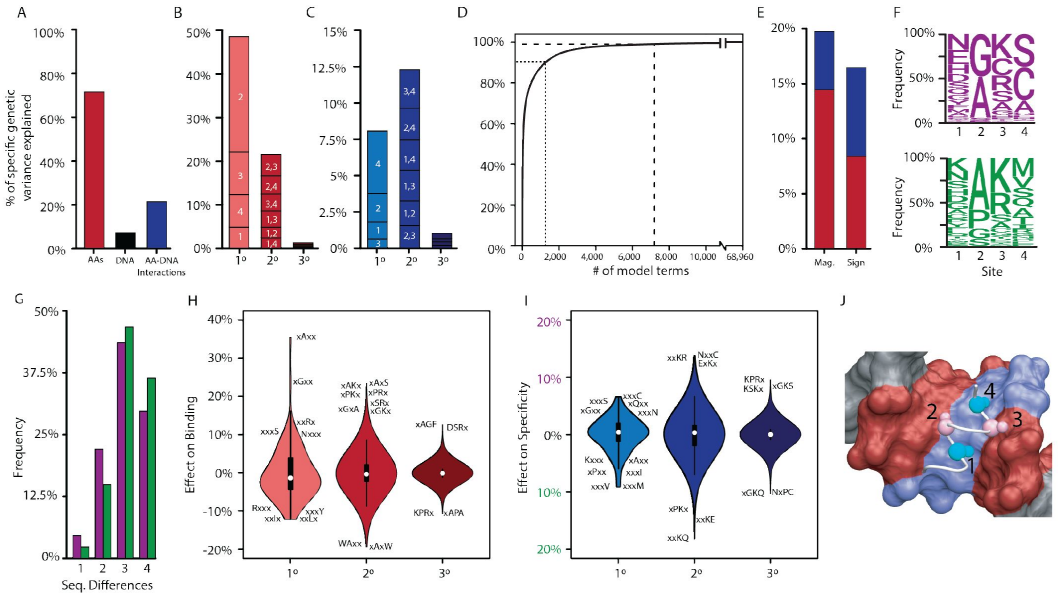

Fig. 2.

Genetic variance explained by components of the model.

A. Fraction of genetic variance explained by all terms in the protein's amino acid sequence (non-REspecific binding, red), the RE's DNA sequence (black), or the interaction between them (specificity, blue). **B**. Fraction of genetic variance explained by terms in each model order for non-RE-specific binding. Indexes within each column show the variance explained by terms at each site or combination of sites. **C**. Fraction of genetic variance explained by specificity terms at each order. **D**. Fraction of genetic variance explained by pairwise epistatic effects that affect only the magnitude or affect the sign of interaction between two states; red and blue, effects on pairwise binding and specificity. **E**. Cumulative fraction of genetic variance explained by model terms, ordered by the fraction of variance explained by each. Dashed lines indicate 90% and 99% of the total variance in genetic score explained by the complete model. **F**. Frequency of amino acid states among sequences predicted to be strong activators on ERE (purple) and SRE (green). **G**. Distribution of amino acid differences among predicted strong activator sequences on ERE (purple) and SRE (green). **H**. Distributions of effect sizes on non-RE-specific binding. Largest-magnitude effects are labeled by amino acid state (x, any state). **I**. Distribution of effect sizes on RE specificity. **J**. Structural view of DBD recognition helix (white tube) and DNA response elements (surface). Blue, nucleotides that differ between ERE and SRE; red, nucleotides identical between ERE and SRE; gray, nucleotides outside of RE. Cα and Cβ atoms of variable residues in the recognition helix are shown as large and small spheres, respectively (cyan, largest effects on RE specificity; pink, largest effects on non-RE-specific binding). Coordinates are from the X-ray crystal structure of AncSR1-DBD, PDB 4OLN.

triplets and quartets (**Figure 2-Figure Supplement 5**). Although low-order prediction is largely sufficient, it cannot be reduced to a few simple rules, such as "must have an arginine at site 3" to bind either RE, or "must have a valine or methionine at site 4" to prefer SRE. Rather, there are hundreds of potential "and" and "or" rules by which RH variants across the massive multidimensional space of the protein's sequence can become a specific or non-specific activator.

The functional effects of variation at each site can to some extent be structurally rationalized (**Fig. 2J**). Sites 2 and 3 have their largest effects through nonspecific binding and their side chains are positioned near the parts of the DNA that are shared between ERE and SRE; the most favorable effects come from having small-volume amino acids at site 2, particularly in combination with a basic residue at site 3, which can potentially hydrogen-bond with a guanine common to both REs. By contrast, the states at sites 1 and 4 have their largest effects on specificity, and these residues are closest to the bases in the major groove that differ between the REs; small hydrophobic residues at site 4 favor SRE binding, where they can pack against the hydrophobic methyl groups that are unique to SRE but leave ERE's hydrogen bond (donors/acceptors) unpaired.

Epistasis generates functional activators

To determine if epistatic effects have a directional bias on functional sequences, we calculated the net contribution of all epistatic terms to the genetic score of each active variant on each RE (**Fig. 3A-B**). Contrary to the expectation of diminishing returns (in which epistasis predominantly impairs function) (23, 29–31), we found that epistatic terms overwhelmingly enhance the function of active variants. Epistatic terms make a net positive contribution to the genetic score of 100% of strong activators, contributing an average of 64% of the score needed to be a strong activator, and the epistatic contribution was necessary to surpass this threshold in every single case. Weak activators also benefit from epistatic interactions, which contribute positively to 98% of variants in this class, but their effect is smaller. Both pairwise and thirdorder effects contribute, but second-order interactions make a far larger positive contribution than third-order effects do (**Figure 3-Figure Supplement 1**).

To directly characterize the effect of epistasis on the number, identity, and function of genotypes, we fit to the data truncated models of genetic architecture that allow only main effects and thus exclude all epistasis – or that allow main and pairwise effects but exclude higher-order interactions. We then used each fitted model to predict which variants would be functional activators. Confirming the key role of epistasis in generating functional proteins, the number of strong activators in the main-effects-only model is reduced by about 25% compared to the third-order model, and the number of weak activators is also substantially reduced (**Fig. 3C**). Most of this increase is caused by pairwise epistasis, because a second-order model contains 98% as many activators as the third-order model (98%, **Figure 3-Figure Supplement 2**). Epistasis also changes which sequences are functional; of variants that are strong activators in the first-order model, only 64% remain strong activators when epistasis is included.

Why does epistasis increase the number of functional variants and change their identity? Even in the first-orderonly model, main effects are mostly of small magnitude (**Figure 3-Figure Supplement 3**): the only way to achieve a genetic score sufficient for activation is to combine the few amino acids that have the largest main effects, and the number of ways to do this is limited. With epistasis, however, positive epistatic terms can supplement main effects, improving the function of sequences that contain states that have weak main effects. Consistent with this explanation, states with positive main effects on binding overwhelmingly have positive pairwise and third-order epistatic interactions (**Fig. 3D**, **Figure 3-Figure Supplement 4**).

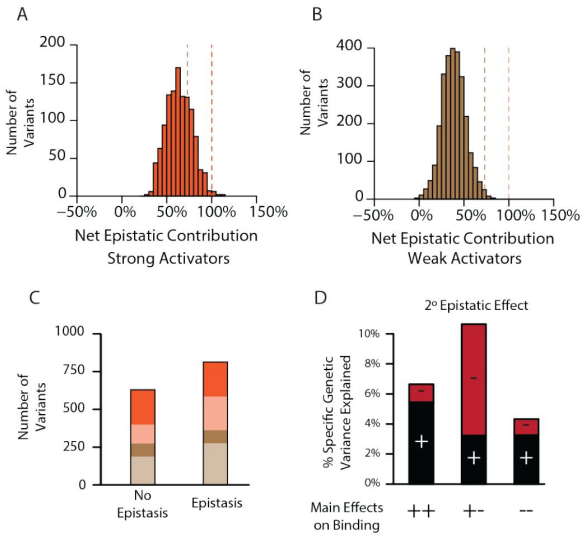

Fig. 3.

Pervasive pairwise epistasis.

A. Net contribution of second- and third-order epistatic effects to the total genetic score of all variants predicted to be strong binders, scaled relative to the threshold for classification as a strong binder. Dashed lines, thresholds between activation classes. **B**. Same as A, but for variants predicted as weak binders. **C**. Number of variants classified as a weak (brown) or strong (orange) binder using terms from a model fit with no epistasis (1º effects only, left) or the full model with epistasis (1º+2º+3º effects, right). Dark colors represent the number of variants that retain the same class membership in both models; pale colors change class. **D**. Percent of genetic variance explained by positive (black) and negative (red) pairwise binding interactions, classified by the sign of the main effects of the states.

The genetic architecture of mutation effects

The genetic architecture of a protein directly describes how the sequence of each protein determines its function, but evolution involves mutations – changes that turn one sequence into another. Exchanging one amino acid for another in a particular protein alters the main effect at that site and all the epistatic interactions involving that site. We used our model to characterize the impacts on the genetic score for ERE and SRE binding of all 1.2 million possible context-specific mutations, defined as one of the 1,520 mutation types (a single-residue exchange from one of the 20 possible starting residues to 19 ending states at each of four sites) introduced into any of the 8,000 possible combinations of states at the other three sites.

We found that there are >200,000 context-mutations that can transform a null protein into a functional activator, and epistasis is critical to these transitions (**Fig. 4A**). The underlying mutations are diverse, with 1160 of the 1520 unique types of amino acid exchanges moving one or more variants into a higher activation class (**Fig. 4B**). In nearly every case (99%), the net epistatic impact of the mutation's effects are positive; in 94% of cases, the epistatic contribution is necessary to move the variant to the higher class, and in almost half of cases it is sufficient (**Fig. 4C-D**). Sign epistasis was rampant among mutations: every single one of the 1,520 mutation types caused positive effects in some backgrounds and negative effects in others, and more than 90% of mutation types (1380 of 1520) had both positive and negative effects on more than 10% of genetic backgrounds (**Fig. 4E**).

Epistasis also expands the mutational opportunities to change the protein's DNA specificity. There are hundreds of context-specific mutations and mutation types that change a strong ERE-specific activator directly into a strong SRE-specific activators; these involve changes at all four sites and use dozens of different amino acid states to generate both ERE and SRE specificity (**Fig. 4F**). In every single case, the net change in epistatic effects caused by the mutation is necessary for the switch in specificity, and in almost half of cases the epistatic effects are also sufficient (**Fig. 4G-H**). Both pairwise and third-order epistasis contributed: in about half of cases, third-order epistasis is necessary, but it is rarely sufficient (**Figure 4-Figure Supplement 1**).

To directly test the role of epistasis in creating mutational opportunities to switch DNA specificity, we compared the mutations available under the best-fit third-order model to those under the best-fit first-order model. Without epistasis, the total number of specificity-switching mutations is reduced by >80%, and the number of mutation types declines from 85 to just three (**Fig. 4F**). Under a model that incorporates pairwise but not third-order epistasis, the number of RE-switching mutations is intermediate, but closer to that in the third-order model (**Figure 4-Figure Supplement 2**). Thus, epistasis – mainly second order – is the primary factor in the opportunity for mutations to change RE-specificity with a single amino acid change.

Evolutionary paths in sequence space

We next assessed trajectories of functional evolution in sequence space, and the effect of epistasis on those trajectories. The RH sequence space consists of all 160,000 possible sequences, each representing one protein variant, with edges connecting nodes that can be directly interconverted by a single nucleotide change given the standard genetic code. We assigned to each node the function(s) predicted by its genetic score under each model. Nonfunctional nodes – those that are null or weak on both REs – were excluded from the network, because purifying selection will remove these genotypes from an evolving population under most circumstances (113). We focused our analysis on evolutionary paths from ERE-specific starting points to SRE-specific endpoints, because this is the functional transition that occurred during history.

Sign epistasis is generally thought to constrain epistasis by creating a fragmented sequence space in which local optima are separated by impassable valleys. While sign epistasis is rampant in our datasets (**Fig. 2D**, **Fig. 4E**), we found that given the best-fit third-order model, 99.8% of

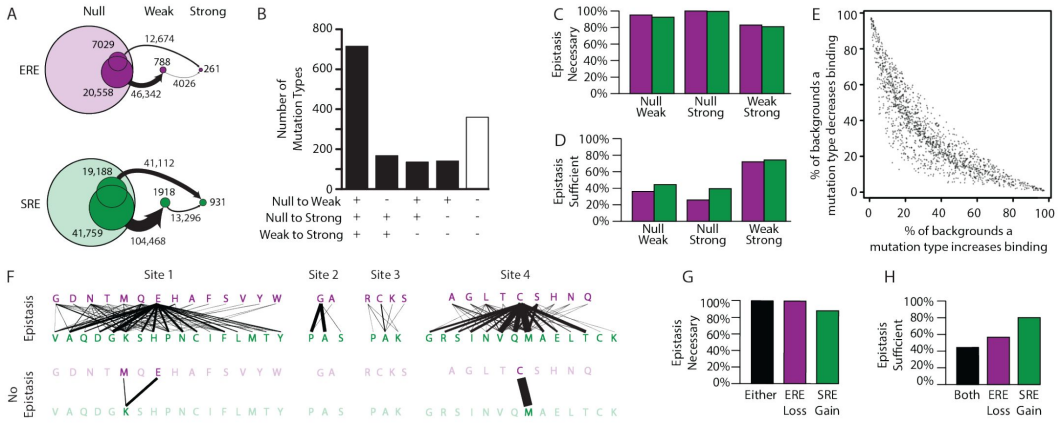

Fig. 4.

Effect of epistasis on the effects of mutations.

A. Number of mutations that can promote protein variants from one activation class to another (circles). Each mutation is a single-amino-acid change from one unique sequence into another. **B**. Effects on activation class of variants by unique mutation types (a change from one amino acid state to another at a particular site, irrespective of the background at other sites). Each column shows the number of mutation types that do (+) or do not (-) promote one or more variants from one class to another. **C, D**. Of mutations that change a variant's binding class, the fraction is shown for which the contribution of epistasis is necessary and/or sufficient (D) to cross the threshold for reclassification. **E**. Context-dependence of mutation effects. Fraction of genetic backgrounds in which a mutation type increases (x-axis) or decreases (y-axis) genetic score. Only mutations that change the genetic score by at least 5% of the threshold to be a strong activator are included. **F**. Single-amino-acid mutations that change DNA specificity from ERE to SRE in the complete model (top) or the first-order-only model (bottom). Lines show mutations from the state in an ERE-specific variant (purple) to that in the SRE-specific variant (green); thickness is proportional to the number of backgrounds at other sites against which the mutation changes specificity. **G, H**. Of mutations that change RE specificity, the fraction for which epistasis is necessary (G) and/or or sufficient (H) to lose ERE binding (purple), gain SRE binding (green) or both (black).

functional sequences are connected to each other in a single mutually accessible network (**Fig. 5A**).

Epistasis also enlarges the network of connected functional sequences, consistent with the function-enhancing effect of epistatic terms on functional variants. The thirdorder model contains 1603 functional nodes (1600 of which are connected in a single network), whereas the first-order epistasis-free model contains only 1080 functional nodes. The second-order model is almost as large as the complete third-order model, indicating that pairwise interactions account for most of the network-enlarging effect of epistasis (**Figure 5-Figure Supplement 1**). In multidimensional sequence space, epistasis therefore does not isolate functional variants: rather, it expands the network of functional sequences, virtually any of which can reach any other through a series of single-nucleotide substitutions.

The probability of reaching any variant from a given starting point depends on the number of substitutions required to reach it. We characterized the evolutionary accessibility of SRE specificity from ERE-specific starting points by simulating evolution from each ERE-specific genotype and measuring the length of the paths required for SRE specificity to evolve. We used two evolutionary scenarios for these simulations. The first represents purifying selection and neutral drift: from any ERE-specific node, every step to any functional neighbor has equal probability, irrespective of which REs it activates, and the path ends as soon as an SRE-specific node is reached (Fig. 5B). In this scenario, including epistasis shortens the average path length to SRE-specificity for 80% of ERE-specific nodes, with an average reduction of 25% (**Fig. 5C**). Accessibility under the second-order model was very similar to the complete model, indicating that the increased accessibility of SRE-specific nodes was primarily caused by pairwise epistasis (**Figure 5-Figure Supplement 2**).

The second evolutionary scenario incorporates positive selection for SRE-specificity by making the probability of visiting each neighbor determined by a selection coefficient that is proportional to the node's preference for SRE over ERE. Compared to the complete third-order model, eliminating epistasis in the first-order model again lengthened the average path taken to SRE-specificity, although this difference shrank as selection became stronger (**Fig. 5D**). This difference, too, was attributable primarily to pairwise epistasis (**Figure 5-Figure Supplement 3**).

Epistasis therefore shortens rather than lengthens the paths taken during the evolution of a new function. To understand why this is so, we examined the distribution of functions across sequence space and its effect on the paths that connect ERE- and SRE-specific sequences. We found several contributing factors. First, the minimum distance from ERE-specific starting points to the nearest SRE-specific variant is on average 10% shorter in the third-order model than when epistasis is eliminated in the first-order model (**Fig. 5E**). A second factor is that there are far more direct single-step mutations from ERE-to SRE-specific nodes: the average number of SRE-specific neighbors per ERE-specific node is 2.5 times higher in the third-order model, and each SRE-specific variant is also more likely to have at least one ERE-specific single-step neighbor (**Fig. 5F**). Third, epistasis reduces the average number of ERE-specific neighbors around ERE-specific nodes, further increasing the probability that a step to SRE specificity will be taken if one is available (**Fig. 5F**). Finally, epistasis increases the number of paths by which SRE specificity can evolve: in the third-order model, each ERE-specific node can reach its closest SRE-specific nodes by an average of 65 different paths; excluding epistasis shrinks this number by about 20%, and these paths are less distinct (**Fig. 5G-H**, **Figure 5-Figure Supplement 4**). Thus, ERE-specific nodes can more easily access nearby SRE-specific nodes when epistasis is present, even though the the average distance from ERE specific genotypes to all SRE specific genotypes increases (**Figure 5-Figure supplement 4**), because epistasis adds many new functional variants to the network, and these contain more diverse combinations of states than in the absence of epistasis.

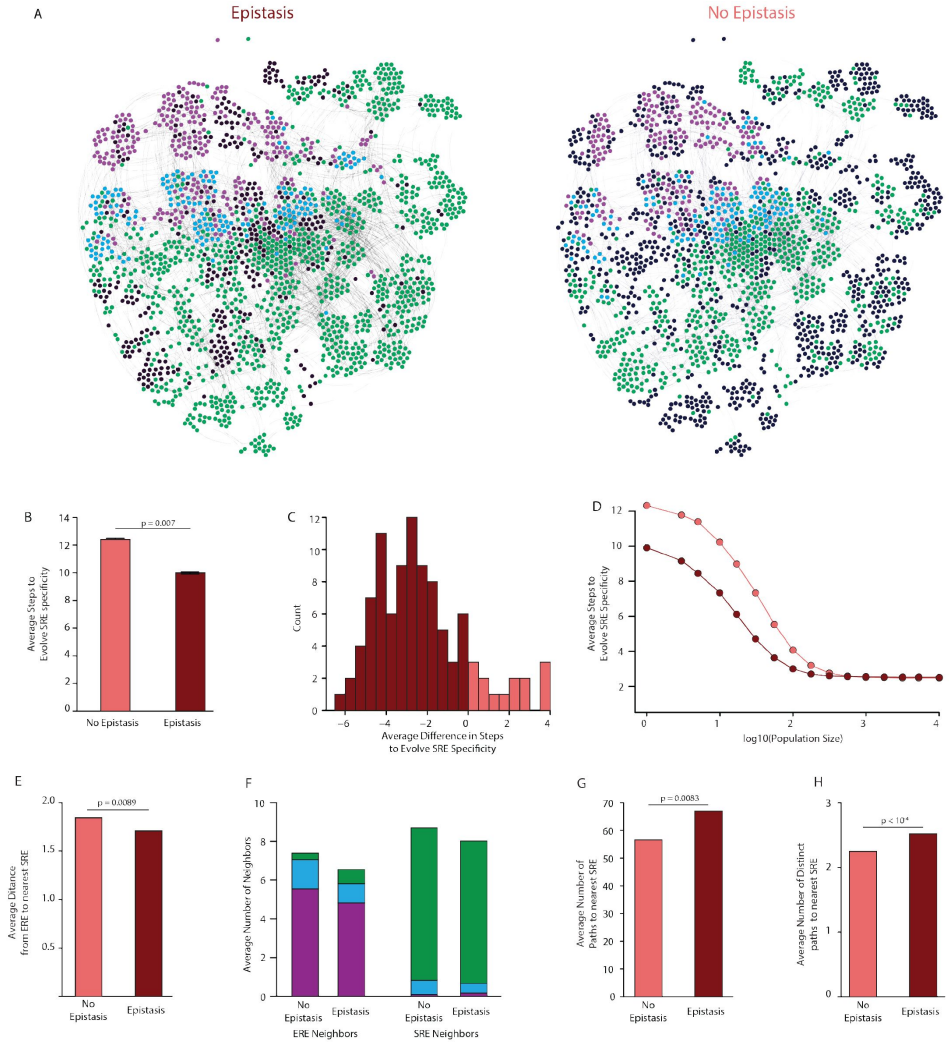

Fig. 5.

Effect of epistasis on the distribution of functions in sequence space.

A. Force-directed graph of all variants that are predicted to be strong activators on one or both REs in the complete model (left) or the first-order-only model (right) after fitting to the DMS data. Lines connect variants that are separated by a single nucleotide change given the standard genetic code. Nodes are colored by their predicted functions given each model: strong activator on ERE (purple), SRE (green), or both (cyan); black nodes are null but predicted to be a strong activator under the other model. Nodes that are null in both models are excluded from the network. **B**. Average number of steps in sequence space for each ERE-specific genotype to reach an SRE-specific genotype when evolution is simulated using a model with only purifying selection and drift. Error bars, 95% confidence interval based on 100 replicates per starting point. **C**. Distribution of the average difference in steps needed to reach an SRE-specific variant from each ERE-specific variant when epistasis is included or excluded, using the same evolutionary process as in B. Light and dark bars show cases in which epistasis makes the path shorter or longer, respectively, from a given ERE-specific variant. **D**. Average number of steps needed to reach an SRE-specific genotype from each ERE-specific genotype in the sequence space of the complete (dark line) and first-order model (light line), when evolution is simulated with positive selection for increased SRE specificity. The “population size” parameter controls the strength of positive selection relative to drift. **E**. Average distance from each ERE-specific variant to the closest SRE-specific genotype in the sequence space of the first-order and complete models. **F**. Average number of single-step neighbors in sequence space per node that is ERE-specific (left) or SRE-specific (right), in the presence or absence of epistasis. Each column shows the average number of neighbors adjacent to the designated node that are strong activators on ERE (purple), SRE (green), or both (cyan). **G**. Average number of unique minimum length paths connecting ERE-specific variants to their nearest SRE-specific neighbor in the first-order and complete models. **H**. Same as G for distinct minimum length paths. All p-values were calculated from a permutation test.

Epistasis therefore facilitates the evolution of new protein functions, rather than constraining it. Epistasis does constrain evolution on a fine scale: by making some of the paths between designated pairs of sequences inaccessible, epistasis can and does make the evolution of particular outcomes contingent on intermediate steps. But epistasis also shapes which pairs exist in the network of functional sequences in the first place and how many steps apart those sequences are. When the genetic architecture includes epistasis, there are more functional sequences; further, similar sequences are more likely to have different functions, because a single amino acid difference can combine with the particular residues at the other sites, sometimes dramatically improving one function and impairing the other. By contrast, if the genetic architecture consists entirely of main effects, variants that share functions will tend to have sequences that are more similar to each other and more different from those with different functions: in this case, each function will be distributed more smoothly across sequence space, and neighboring sequences are less likely to have distinct functions than when epistasis is present.

Genetic architecture of a historical change in function

Finally, we sought to understand how the DBD's genetic architecture shaped the historical transition from ERE-to SRE-specificity that occurred during the phylogenetic interval between AncSR1 and AncSR2. Each of the three amino acid changes from the ancestral RH sequence egKa to the derived GSKV (ancestral in lower case, derived in upper case) can each be conferred by a single nucleotide change. The functions of the intermediates along the direct path from EGKA to GSKV suggest the order in which the substitutions are likely to have occurred: of the three single-mutant variants, only one (GGKA) is functional on either RE – indicating that this was probably the first step – and the same is true of the possible second steps (GGKV being the only accessible functional node, **Fig. 6A**).

The ancestral and derived RH variants that occurred during history – and persist to the present – were accessible only because of pairwise epistasis. Under the main-effects only model, neither egKa nor GSKV is a strong activator on either RE. Including the terms from the pairwise or third-order models restores the function of these variants, because interactions make large and necessary contributions to their functions. For example, the glutamate at site 1 in the ancestral variant has virtually no main effect on ERE activation, but it makes a major contribution via pairwise interactions with sites 2 and 3 (**Fig. 6B**). Similarly, the serine at site 2 in the derived variant has no main effect on SRE specificity, but its interactions with sites 3 and 4 are necessary to make this protein an SRE activator. These RH sequences are conserved today in all known ERs and kSRs: therefore the functionality of both historical and present-day steroid receptors depends critically on pairwise epistasis.

The historical transition from ERE-to SRE-specificity also altered the architecture by which REs are recognized, changing the sites and pairs that confer DNA binding rather than just exchanging amino acid states within an otherwise fixed architecture **Fig. 6B**). Moreover, this architecture continued to change once the RH became SRE-specific: the final substitution in the trajectory (g2S, which had no net effect on specificity) replaced the initial historical determinants of SRE specificity with a different set of favorable interactions. As a result, the final RH sequence – which is now found in all extant kSRs – has determinants of function that are entirely different from those in the ancestral RH and even those that first mediated the acquisition of SRE specificity during history.

Discussion

Many of our findings differ from other studies of the genetic architecture of protein function and evolution. We find only a small role for high-order epistasis, with almost all functional variation among genotypes attributable to main and pairwise effects. Furthermore, the genetic architecture

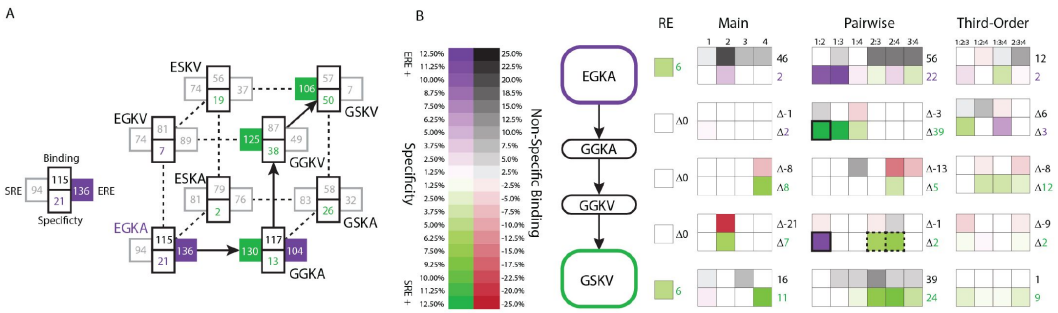

Fig. 6.

Genetic architecture of the historical change in RE specificity.

A. Each intermediate on the set of direct paths from the RH sequence of AncSR1 (EGKA) to AncSR2 (GSKV) is shown at a vertex of the cube. At each vertex, the total genetic score for binding ERE (right) and SRE (left) are shown, scaled to 100 as the threshold for strong activation, and colored dark if the variant is a strong activator on ERE (purple) and/or SRE (green). Contribution of terms to non-specific binding (top) and specificity (bottom, colored by the RE they favor) are shown. Lines connect genotypes that differ by a single amino acid mutation; solid lines connect vertices that are strong activators on one or both REs; dotted lines connect to a vertex that activates on neither RE. **B**. Contribution of historical changes in state to the change in RE specificity. Each genotype along the plausible path from AncSR1 to AncSR2 RH is shown at left. For EGKA and GSKV, the contribution of effects at each order to the total genetic score for non-RE-specific binding (top) and RE-specificity (bottom) is shown, scaled as in A. Each cell is shaded by the magnitude of the effect at that site or combinations, and the sum of effects for the category is shown at right. The middle rows show the effect of each amino acid mutation along the path, defined as the effect of the derived state (or combination) minus that of the ancestral state or combination. Solid outline, mutation that caused the initial acquisition of SRE binding; dashed outlines, mutations that changed the determinants of SRE binding. The RE column shows the global specificity effect. This value is constant during evolution, but is needed to reconstruct the specificity values shown in A.

is dense at these low epistatic orders, instead of sparse as found in prior work. Finally, we find that epistasis is not simply a constraint on evolution, but instead facilitates the evolution of new functions by altering which sequences are functional in the first place and how similar sequences with different functions are. These differences arise primarily because we take a global instead of a local view of epistasis and we expand the scope of analysis to incorporate all possible amino acid states for multiple functions.

A global view of genetic architecture

Most prior studies have attempted to decompose protein genetic architecture by estimating the effects of mutations (and combinations) on a designated reference sequence (27, 54, 59, 61, 62, 110, 114). This strategy provides an accurate local picture of mutation effects when they are introduced into the reference or nearby proteins, but accuracy is poor across other backgrounds and is very sensitive to experimental error (83). A method called background averaging reduces this problem by averaging estimates of mutation effects across sets of variants containing a state or combination of interest, but it still defines mutation effects relative to a particular reference state, which may make it sensitive to error in the measurement of variants containing the reference state (83). In contrast, our approach provides a globally optimal dissection of genetic architecture by minimizing the total deviation between predicted and observed functions across the entire ensemble of variants, regardless of the number of states allowed at each site.

The shift from a reference-based account to a global model of genetic architecture reframes questions about the relationship of sequence to function and its impacts on evolution. In reference-based analyses, the wild-type sequence is taken as a given; the goal is to understand the effect of mutations as the protein moves along paths away from that starting point (usually to a defined endpoint). The starting point itself has no genetic determinants at all. A point mutation can change the function of the original protein via only a single main effect. Never considered are the effect of wildtype states on the protein's functions nor a mutation's interactions with those states. That perspective may be appropriate if one is concerned only about the immediate sequence neighborhood of a particular 'wild-type' protein, but it is not well-suited to characterizing a protein's genetic architecture of function or its evolution across an entire sequence landscape.

A reference-free model of genetic architecture, in contrast, asks how sequence states per se – not just a change from a particular beginning state to a particular end state – determine function. The function of each variant in the entire ensemble of sequences is caused by all the states it contains, not by the differences that separate it from a wild-type sequence. Epistasis not only affects the available paths through sequence space from the reference to other nodes; it also determines which nodes are in the space in the first place – the entire set of possible starting, ending, and intermediate genotypes. The effect of mutations is also conceived differently: in the global model, a pair of mutations replaces the main effect of the ancestral states with two new derived main effects, and replaces the original epistatic interactions with every site in the protein – including with those sites at which the state does not change. A mutation does not simply replace one brick in a static genetic architecture; it also changes many of the second-order interactions that determined the protein's functions but that were invisible if the starting point is viewed as an already functional black-box. As a protein evolves, each substitution can epistatically change the effect on function of the state at many other sites or modify the effect of other future mutations.

Low-order, dense genetic architecture

Like other researchers, we found pervasive epistasis within a protein (54, 59, 61, 110, 114). Unlike most previous studies, we found that higher-order epistasis plays only a tiny role (54, 59, 61, 110, 114) (but see (61)). We did not measure fourth-order epistasis, but it is unlikely that it will account for much global variation: we obtained good predictions of

function from main and pairwise effects alone. Moreover, for fourth-order terms to have much of a global effect, each one would have to be considerably larger on average than thirdorder and second-order terms, but we, like others (61), have observed the opposite pattern – effects of shrinking magnitude as epistatic order increases. Why do our findings concerning higher-order epistasis differ from previous studies? A likely cause is that prior work used reference-based methods and/or they did not adequately account for global nonlinearities. Both of these factors are known to cause spurious inference of higher-order epistasis (83, 110, 112)

Another difference from previous work is that we found a dense genetic architecture at first and second orders, such that main and pairwise effects involving virtually all possible states at all sites make substantial contributions to genetic architecture. Some other studies have found a very sparse architecture, with just a few states or combinations explaining the distribution of functions across variants (83). One cause of this difference may be that we analyzed all 20 amino acid states rather than just two, so the number of possible terms in our model is far larger than studies that analyze only pairs of states present in two high-functioning proteins. Consistent with this idea, combinatorial DMS studies have found that there are often numerous functional sequences that share few to no sequence states at the mutated sites, which necessitates a denser genetic architecture to account for the entire set of functional sequences compared to a set of sequences where only two amino acid states are allowed at each site (94, 96, 97). Another factor may be that all the sites we examined are in the protein's "active site" – in this case, making direct contact with the DNA to which the protein binds.

Epistasis and evolution

Prior work has viewed epistasis as an evolutionary constraint, but we found that the primary effect of epistasis is to facilitate functional evolution. Virtually all earlier studies on this topic considered only the effect of epistasis on direct paths through a binary sequence space between two functional sequences designated a priori as starting and ending points, and they have focused on optimization of a single function. In that context, the major effect of epistasis is to create idiosyncrasies in the functions of the intermediates between two high-functioning proteins; in this framework, epistasis can only constrain evolution along those paths (115, 116). But this narrow perspective never considers the effect of epistasis in making the designated sequences functional in the first place, and it excludes from view the vast number of additional potential starting, ending, and intermediate nodes in a 20-state combinatorial sequence space.

By incorporating a global perspective, our analysis reveals the facilitating effect of epistasis on evolution. The genetic architecture of function generates a huge network of densely connected functional RH sequences that evolution can navigate (117, 118). Epistatic interactions are critical in making protein variants functional, so it dramatically increases the number of functional nodes that can be visited during evolution. Moreover, by determining the protein's DNA specificity, epistasis brings variants with different specificities closer together in sequence space; this makes regions of sequence space with new functions far easier to access from those with ancestral functions. By increasing the number of functional nodes and shortening the paths between those with different functions, epistasis dramatically increases the probability that evolution will arrive at a new function from anywhere in the network of potential starting points.

Overall, epistasis makes the topography of sequence space idiosyncratic enough that new functions are easily accessible from nodes with different functions, but not so idiosyncratic that functional nodes become inaccessible from each other. Although epistatic interactions do constrain trajectories when we limit our view to arrival at a particular destination from a particular origin, interactions open opportunities to evolve new functions that are often just a step away from the trajectories that were realized during evolution.

Limitations and future work

We are aware of several limitations in our study. First, we used an ordinal model to analyze the genetic architecture of functional data transformed into categorical form. This approach could reduce precision in estimating the terms of genetic architecture compared to direct analysis of continuous phenotypic data. The advantage of this approach is that it can reduce the impact of noise. In our dataset, many genotypes are at the lower bound of measurement and much of the variation in measured florescence for these variants is measurement noise; ordinal regression is therefore expected to reduce the impact of noise on estimates of genetic architecture, and it appears to have done so in this case. Another potential issue is nonspecific epistasis, which if not incorporated can result in inflated estimates of specific epistasis. The logistic model assumes that an amino acid state or combination has a consistent effect on the log-odds of changing a variant's functional category, irrespective of the level of function of the background into which it is introduced. It therefore does not explicitly incorporate nonspecific epistasis, such as that imposed by a limited dynamic range of measurement, although the sigmoid shape may fortuitously accommodate this kind of nonspecific context-dependence. Because our categorical analysis of variants' functional class depends only on the functional rank-order of variants and not on their quantitative function per se, the logistic model is expected to be robust to nonspecific epistasis, so long as the underlying relationship between genotype and phenotypic measurement is monotonic.

We studied four sites in one protein, so the generality of our observations is unknown at this point. For example, genetic architecture away from the protein's DNA-binding surface could be sparser than that in the RH. However, complete single-mutant scans of all sites in other proteins (and the few studies of all double mutants) shows that most mutations at most sites affect function, consistent with a generally dense genetic architecture – although perhaps less extremely so than in the RH sites that we studied here (33, 51 – 53, 119–133). Higher-order epistasis might be more important in other parts of the protein than in the active site, but we know of no structural rationale that might support this expectation: the RH sites are all packed closely in space and bind adjacent nucleotides in the response element, a physical architecture that might be expected to enrich for rather than deplete higher-order effects. We know of no structural or functional reason to expect that high-order epistasis will be more important in other proteins. It is possible that third- (and higher-) order interactions could become more important when larger numbers of sites are considered, because the number of third-order terms that affect each variant will increase as the number of variable sites grows (61);

A major challenge in broadening our understanding of protein genetic architecture is the immense size of sequence space. Covering all combinations of states using DMS is currently practical for at most five variable sites at a time, so a complete characterization of genetic architecture at all orders is possible for only tiny protein fragments. Our finding that high-order epistasis is relatively unimportant implies that modeling first- and second-order terms should be sufficient to account for most genetic variation in a protein's functions. If generalizable, the focus could shift from complete sampling of all high-order combinations to covering the much smaller set of pairs. It would still be necessary, however, to analyze the effects of pairs across sufficiently diverse backgrounds at other sites to allow effective global averaging of secondorder terms. Evaluating strategies to efficiently accomplish this end without collapsing back into the local tunnel-vision of reference-based, binary state analyses is a key goal for the future.

We found that within the DBD's recognition helix, the relationship between genetic score and its DNA-binding function is tractable and reasonably simple, allowing us to predict the functions of a genotype with good confidence from its main and pairwise effects. To what extent is this finding generalizable from proteins to phenotypes at higher biological levels? Our comprehensive analysis of genetic architecture was possible because we reduced the problem of genetic causality to the effects of a few sites in a single protein domain on a simple biochemical function that can be measured with a high-throughput laboratory assay. Caution is therefore required before extrapolating our results to phenotypes involving multiple molecules or loci (134). Even in our

dataset, most functional specificity is attributable to pairwise interactions between amino acids: this represents third- (or potentially fourth-) order intermolecular epistatic interactions between amino acid pairs and the variable nucleotides in the RE. Our assay was also carried out in a controlled and constant environment, and our library was combinatorically complete. This design allowed us to dissect genetic causality free of environmental variance, population structure, gene-environment interactions, or gene-environment correlations – all factors that radically complicate the genotype-phenotype relationship in natural populations. While our study implies that the genetic project may be largely tractable at the level of biochemistry, it does not imply that complex organismal phenotypes or biosocial traits – particularly those affected by many loci, many environmental variables, and complex dependencies within and between those categories – are predictable by reference to genotype alone.

Methods

Experimental Data

The experimental data analyzed here were first reported in Starr et al. 2017 (96). Libraries of all combinations of all 20 amino acid states at four variable sites in the SR DBD recognition helix (RH, 160,000 total protein variants) were prepared in two different reconstructed ancestral backgrounds: AncSR1 and AncSR1+11P, which contains 11 historical permissive substitutions, which nonspecifically increase DNA affinity without strongly affecting specificity (103). The libraries were transformed into yeast strains in which a fluorescent reporter is driven by the ancestral estrogen receptor response element (ERE) or the derived ketosteroid receptor response element (SRE). Activation by each variant on each RE was measured using a FACS-assisted sort-seq assay and then discretized into three ordered activation categories on each RE: Null if their mean fluorescence was not significantly greater than stop-codon-containing variants, Strong if their mean fluorescence was significantly greater than stop-codon variants and not less than 80% of the historical reference for that RE (AncSR1 on ERE, AncSR1+11P+GSKV on SRE, where GSKV is the sequence of extant ketosteroid receptors in the RH), or Weak if they were neither Null nor Strong. Further details of the categorization can be found in Starr et al. 2017 (96). For model fitting, we only used data from the combinatorial library created on the AncSR1+11P background as it contained the most balanced number of variants across the three classes.

Model Description

We formulated an ordinal logistic regression model to dissect the effects of amino acid states and combinations on the activation class of RH variants on the two REs. The model is a forward-cumulative model with a proportional log-odds assumption – a standard approach to modeling data discretized into ordered classes of an otherwise continuous variable (in this case, mean fluorescence). The model imposes a linear relationship between the predictors (the genetic states/combinations) and the log-odds that a variant is in a set of higher classes vs. a set of lower classes; log-odds is defined as the logarithm of the ratio of the probabilities of the higher vs. the lower classifications. For three activation classes, the relationship between the predicted class Y (g) of a genetic variant with sequence g is:

$$log\left(\frac{P(Y(g)=Null)}{P(Y(g)=Weak/Strong)}\right)=t_{NW}+\beta X \quad \textbf{(1)}$$

$$log\left(\frac{P(Y(g)=Null/Weak)}{P(Y(g)=Strong)}\right)=t_{WS}+\beta X \quad \textbf{(2)}$$

where t_{NW} and t_{WS} are thresholds that delimit the boundaries between the Null/Weak and Weak/Strong binding classes, X is a matrix of indicator variables that represent all the possible sequence states and combinations, and β (the variables to be estimated) is a set of coefficients that describe the effect of each state or combination on the log-odds of classification. We evaluated the proportional odds assumption by fitting two simple logistic regression models to the observed activation class data, with variants reclassified as null vs. weak-or-strong, and again as null-or-weak vs. strong. Estimated effects across the two thresholds were similar, as required by the proportional odds assumption. (**Figure 1-Figure Supplement 3**). However, this assumption can be relaxed if there is evidence that the effects of amino acid states or combinations is dependent on the binding class.

We used a reference-free encoding of indicator variables and effect coefficients. First, we describe the form of the model that would be used if only a single phenotype were measured; we then expand the description to incorporate two different phenotypes. For the single-phenotype model, the indicator matrix X describes the amino acid states and combinations contained by each of the 160,000 protein variants. Each row in X specifies a protein variant with sequence $q_1q_2q_3q_4$ at the four variable sites, and each column represents a particular state at a site (e.g., q_1 = A) or a particular combination of states at a combination of sites (e.g., q_1q_2 = AA, or $q_1q_2q_3$ = AAA). Each cell is assigned value 1 or 0 to specify whether that state or combination is present or absent in the variant. Each coefficient represents the effect of having one of these states or combinations on the log-odds of being in an activation class, relative to the average across all variants. The complete model therefore consists of 80 first-order effects (20 amino acid states at each of the four sites), 2400 second-order effects (400 pairs of states at each of the six pairs of sites), and 32,000 third-order effects (8000 triplets of states at each of the four triplets of sites). A single global intercept, β_0 describes the average across all variants (and is associated with indicator variable with value 1 in all variants). In principle, the model could contain all fourth-order combinations and effects, but we did not include them, because each fourth-order coefficient would be estimated from the observed activation class of a single variant and so would be indistinguishable from measurement error.

To incorporate the effects of genetic states on two different functions, the indicator matrix is expanded to uniquely identify all 320,000 protein-RE complexes based on the RE it contains and the amino acid states/combinations in the protein variant. Each amino acid state/combination is now associated with two coefficients – one for its average effect across the two REs (β which represents the RE-nonspecific effect on activation) and the other for the difference in effect on ERE vs. SRE (σ representing the effect on specificity). Indicator variables for the RE-nonspecific effects are assigned value 1 if an amino acid state/combination is present in a complex and 0 if absent; indicators for specificity effects are assigned value 1 if the complex contains ERE (and the amino acid state/combination is present), or -1 if it contains SRE. The net effect of any amino acid state or combination on ERE activation is therefore its binding coefficient plus its specificity coefficient; its effect on SRE is the binding coefficient minus the specificity coefficient. A global RE coefficient σ_0 represents the main effect of having ERE vs SRE in the complex, averaged over all protein variants (with indicator 1 or -1 for complexes containing ERE or SRE, respectively). As in the single-phenotype model, a global intercept coefficient represents the global average across all complexes (β_0, with indicator 1 in all complexes).

Although the number of total coefficients in the model is large (68,962 possible), each of the 320,000 complexes is described by just 30 non-zero indicator variables: one for the global intercept, one for the main RE effect, and a binding and specificity coefficient for each of the four

amino acid states, six pairs, and four triplets that it contains. For a particular variant with sequence $q_1q_2q_3q_4$ the complete model is specified as follows, after coefficients are multiplied by the relevant indicator variables:

$$log\left(\frac{P(Y(q_1q_2q_3q_4|ERE)=Null)}{P(Y(q_1q_2q_3q_4|ERE)=Weak/Strong)}\right)=t_{NW}+$$
$$\beta_0+\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}+$$
$$\sigma_0+\sum_i\sigma_{q_i}+\sum_{i<j}\sigma_{q_iq_j}+\sum_{i<j<k}\sigma_{q_iq_jq_k} \quad \textbf{(3)}$$

$$log\left(\frac{P(Y(q_1q_2q_3q_4|SRE)=Null)}{P(Y(q_1q_2q_3q_4|SRE)=Weak/Strong)}\right)=t_{NW}+$$
$$\beta_0+\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}-$$
$$\sigma_0-\sum_i\sigma_{q_i}-\sum_{i<j}\sigma_{q_iq_j}-\sum_{i<j<k}\sigma_{q_iq_jq_k} \quad \textbf{(4)}$$

$$log\left(\frac{P(Y(q_1q_2q_3q_4|ERE)=Null/Weak)}{P(Y(q_1q_2q_3q_4|ERE)=Strong)}\right)=t_{WS}+$$
$$\beta_0+\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}+$$
$$\sigma_0+\sum_i\sigma_{q_i}+\sum_{i<j}\sigma_{q_iq_j}+\sum_{i<j<k}\sigma_{q_iq_jq_k} \quad \textbf{(5)}$$

$$log\left(\frac{P(Y(q_1q_2q_3q_4|SRE)=Null/Weak)}{P(Y(q_1q_2q_3q_4|SRE)=Strong)}\right)=t_{WS}+$$
$$\beta_0+\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}-$$
$$\sigma_0-\sum_i\sigma_{q_i}-\sum_{i<j}\sigma_{q_iq_j}0\sum_{i<j<k}\sigma_{q_iq_jq_k} \quad \textbf{(6)}$$

where | indicates the RE in the complex, and i, j, and k index sites in the RH sequence.

Specific vs non-specific epistasis and the genetic score

In general, epistatic interactions among states can be classified as specific or non-specific (20, 109). Non-specific epistasis occurs if a global non-linear relationship between an underlying physical property and the measured phenotype affects all combinations of states in the same way. Specific epitasis occurs when combinations of states have distinct non-additive effects on the underlying physical properties. Failure to account for non-specific epistasis when modeling epistatic interactions can artificially inflate inferences of specific epistasis (59, 110). Nonspecific epistasis can arise from intrinsically nonlinear relationships between physical properties and measured phenotypes (such as the energy of binding and occupancy of the bound state), limited dynamic range in an assay or biological phenotype, non-linear scaling within the dynamic range, or the imposition of thresholds within the dynamic range for classifying the functionality of variants.

Our model estimates specific and non-specific epistatic effects in a single fitting procedure by including both forms of epistasis in the model. We do so by defining the genetic score y(g) of any variant g with sequence q1q2q3q4 in complex with ERE or SRE as the sum of all the specific genetic effects of the states in g and the complex:

$$y(g|ERE) = \beta_0 + \sum_i \beta_{q_i} + \sum_{i<j} \beta_{q_i q_j} + \sum_{i<j<k} \beta_{q_i q_j q_k} + \sigma_0 + \sum_i \sigma_{q_i} + \sum_{i<j} \sigma_{q_i q_j} + \sum_{i<j<k} \sigma_{q_i q_j q_k} \quad (7)$$

$$y(g|SRE) = \beta_0 + \sum_i \beta_{q_i} + \sum_{i<j} \beta_{q_i q_j} + \sum_{i<j<k} \beta_{q_i q_j q_k} - \sigma_0 - \sum_i \sigma_{q_i} - \sum_{i<j} \sigma_{q_i q_j} - \sum_{i<j<k} \sigma_{q_i q_j q_k} \quad (8)$$

The specific epistatic interactions are estimated as the second- and third-order epistatic terms of the model, which have an additive effect on the genetic score (together with all the non-epistatic first- and zero-order effects). Nonspecific epistasis is then incorporated through the logit function, which makes classification a nonlinear outcome of the genetic score.

Epistatic coding and interpreting coefficient estimates

Most efforts to model sequence-function relationships have encoded genetic effects as the effects of mutations (singly or in combination) relative to a designated reference or “wild-type” sequence. Others have used a Fourier transformation to recode genotypes and estimate the effects of the transformed states relative to the average across all genotypes (84, 99, 107). Here we develop a reference-free approach that directly estimates the effect of any amino acid state or combination on the genetic score relative to the global average across all variants. In our model, the zero-order term is the global average:

$$\beta_0 = \frac{1}{2 \cdot 20^4} \sum_{g \in G} y(g) = \overline{y(G)} \quad (9)$$

where y(g) is the genetic score of a particular genotype g and G is the set of all possible protein sequence/RE element complexes; the factor before the sum averages over all protein genotypes times 2 to reflect the number of REs. The main effect of a particular amino acid state q at a particular site i on the log-odds of activation (not specific to an RE) is the average difference between the genetic score of all variants with that state and the global average:

$$\beta_{q_i} = \overline{y(G_{q_i})} - \beta_0 = \frac{1}{2 \cdot 20^3} \sum_{g \in G_{q_i}} y(g) - \beta_0 \quad (10)$$

where G_{q_i} is the set of genotypes with amino acid state q at site i. Similarly, a pairwise epistatic effect is the difference between the average genetic score of $G_{q_i q_j}$ – the set of variants containing states q_i and q_j at a particular pair of sites i and j – and the expected score if there was no pairwise interaction:

$$\beta_{q_i q_j} = \overline{y(G_{q_i q_j})} - (\beta_0 + \beta_{q_i} + \beta_{q_j}) = \frac{1}{2 \cdot 20^2} \sum_{g \in G_{q_i q_j}} y(g) - (\beta_0 + \beta_{q_i} + \beta_{q_j}) \quad (11)$$

Third-order interactions are the difference between the expected genetic score based on the relevant lower-order effects:

$$\begin{aligned}\beta_{q_iq_jq_l} &= \overline{y(G_{q_iq_jq_k})} - (\beta_0 + \beta_{q_i} + \beta_{q_j} + \beta_{q_k} + \\ &\qquad \beta_{q_iq_j} + \beta_{q_iq_k} + \beta_{q_jq_k}) \\ &= \frac{1}{2 \cdot 20} \sum_{g \in G_{q_iq_jq_k}} y(g) - (\beta_0 + \beta_{q_i} + \beta_{q_j}\beta_{q_k} + \\ &\qquad \beta_{q_iq_j} + \beta_{q_iq_k} + \beta_{q_jq_k})\end{aligned} \tag{12}$$

The specificity coefficients are encoded similarly. The global RE coefficient reflects the average difference between the genetic score of all complexes containing ERE and the global average (which is of equal magnitude but opposite sign as the average difference of all complexes containing SRE from the global average). This equals half the difference between the average genetic scores for all complexes on ERE and all complexes on SRE:

$$\begin{aligned}\sigma_0 &= \overline{y(G^E)} - \overline{y(G)} = -\overline{y(G^S)} - \overline{y(G)} \\ &= \frac{1}{2}(\overline{y(G^E)} - \overline{y(G^S)}) \\ &= \frac{1}{20^4}(\sum_{g \in G^E} y(g^E) - \sum g \in G^S y(g^S)\end{aligned} \tag{13}$$

where $y(g^E)$ is the genetic score when a protein variant with sequence g is bound to ERE and $y(g^S)$ is the genetic score when a genotype g is bound to SRE. The remaining specificity coefficients follow the same pattern, reflecting differences in specificity beyond what is expected from lowerorder effects. For the main specificity coefficients:

$$\sigma_{q_i} = \frac{1}{2}(\overline{y(G^E_{q_i})} - \overline{y(G^S_{q_i})} - \sigma_0 \tag{14}$$

For pairwise specificity coefficients:

$$\sigma_{q_iq_j} = \frac{1}{2}(\overline{y(G^E_{q_iq_j})} - \overline{y(G^S_{q_iq_j})} - (\sigma_0 + \sigma_{q_i} + \sigma_{q_j}) \tag{15}$$

For third order specificity coefficients:

$$\begin{aligned}\sigma_{q_iq_jq_k} &= \frac{1}{2}(\overline{y(G^E_{q_iq_jq_k})} - \overline{y(G^S_{q_iq_jq_k})} - \\ &\qquad (\sigma_0 + \sigma_{q_i} + \sigma_{q_j} + \sigma_{q_k} + \sigma_{q_iq_j} + \sigma_{q_iq_k} + \sigma_{q_jq_k})\end{aligned} \tag{16}$$

Model selection, regularization, and cross-validation

We fit the model to the DMS data using regularized logistic regression in R (135). To avoid overfitting model parameters to measurement noise, we used regularization via ridge regression ($L2$ norm) (**Figure 1-Figure Supplement 1**). To identify the optimal regularization penalty lambda, we used iterated 10-fold cross-validation. Specifically, we fit a series of models with 900 lambda values of decreasing magnitude from $10^{\Delta 1}$ to $10^{\Delta 8}$ on a log_{10} scale. The data were randomly divided into 10 equally sized blocks, and the model at each lambda was fit to 90% of the data, leaving out 10% of the data as a test set, and the activation class of each variant in the test set was predicted on each RE. We compared these predictions to the observed class for each variant, counting misclassification into an adjacent category (weak-null or weak-strong) as a single error and misclassification between null and strong as two errors. We repeated this procedure, using

each of the ten blocks as the test set in turn and calculated the mean misclassification rate. We then repeated this entire procedure ten times, dividing the data into different blocks each time, and using the variation in the estimated misclassification ratio to determine the standard error in our estimate. We applied this procedure to all values of lambda and selected the value with the lowest average misclassification rate.

Model Implementation

To implement the cumulative odds model with three ordered classes and the proportional odds assumption in a framework that allowed for regularization and cross-validation, we made several modifications to the glmnetcr function in the glmnetcr package (136).

Proportional Odds assumption via logistic regression

In general, the cumulative odds model can be viewed as generating a set of binary variables around each threshold, each of which distinguishes between the set of classes lower and the set of classes higher than the threshold. For our three activation classes, we recoded the categories using two binary variables – one that distinguishes the Null class from Weak-or-Strong and another that distinguishes Null-or-Weak from Strong. Null corresponds to the 00 class, weak to the 10 class, and strong to the 11 class. Logistic regression can then be used to estimate the best-fit model parameters that predict these recoded classes. This produces equivalent estimates of model coefficients as ordinal regression.

Sparse Matrix Representation

Manipulating the large model matrix imposed a large computational burden. Since a large portion of the matrix consists of zeros (each row contains only 30 non-zero values), we used the MatrixModels R package (137) and modification of functions in the glmnetcr package to implement the handling of sparse matrices.

Long vector accommodation

The version of R we used fit logistic regression models with regularization using Fortran. To accommodate the large number of observations and covariates, we modified the lognet function from the glmnet R package to handle long vectors and passed this to the underlying Fortan code using the dotCall64 R package (138).

Zero-sum constraint on coefficients

The reference-free framework requires the sum of model coefficients for a given site (or combination of sites) to sum to zero. All marginal subsets of pairwise or third-order coefficients – defined as the subset of coefficients for a given pair (or triplet) of sites that share a particular state (or pair of states) – must also sum to zero. For example, of the 400 pairwise interactions between sites 1 and 2, the entire set must sum to zero; of these, 20 have an A at site 1, and this marginal set must also sum to zero. This constraint guarantees that all coefficients represent the average effect of having some sequence state or combination relative to prediction from applicable lower-order terms and the global average.

Imposing this constraint during regularized regression is computationally expensive. We therefore performed unconstrained regularized regression and then adjusted the unconstrained coefficients post-hoc to impose this constraint. Specifically, we calculated the average of each set of unconstrained coefficients and each marginal set of unconstrained coefficients. We then

subtracted the relevant averages from each unconstrained coefficient, thus recentering each set and marginal subset around an average of zero. For third-order interactions, the constrained estimates are:

$$\hat{\beta}_{q_i q_j q_k} = \beta_{q_i q_j q_k} - \left(\overline{\beta_{*_i q_j q_k}} + \overline{\beta_{q_i *_j q_k}} + \overline{\beta_{q_i q_j *_k}}\right) + \left(\overline{\beta_{*_i *_j q_k}} + \overline{\beta_{*_i q_j *_k}} + \overline{\beta_{q_i *_j *_k}}\right) - \overline{\beta_{*_i *_j *_k}} \quad (17)$$

$$\hat{\sigma}_{q_i q_j q_k} = \sigma_{q_i q_j q_k} - \left(\overline{\sigma_{*_i q_j q_k}} + \overline{\sigma_{q_i *_j q_k}} + \overline{\sigma_{q_i q_j *_k}}\right) + \left(\overline{\sigma_{*_i *_j q_k}} + \overline{\sigma_{*_i q_j *_k}} + \overline{\sigma_{q_i *_j *_k}}\right) - \overline{\sigma_{*_i *_j *_k}} \quad (18)$$

Where $\hat{\beta}$ and $\hat{\sigma}$ are the constrained coefficients, β and σ the unconstrained coefficients, and the averages are over the possible states at a site (or combination) indicated by the *. Pairwise coefficients follow the same pattern, with the total marginal effect of each combination of amino acids to be zero:

$$\begin{aligned}\hat{\beta}_{q_i q_j} = \beta_{q_i q_j} &- \left(\overline{\beta_{*_i q_j}} + \overline{\beta_{q_i *_j}} - \overline{\beta_{*_i *_j}}\right) + \\ &\left(\overline{\beta_{q_i q_j *_k}} - \overline{\beta_{q_i *_j *_k}} - \overline{\beta_{*_i q_j *_k}} + \overline{\beta_{*_i *_j *_k}}\right) + \\ &\left(\overline{\beta_{q_i q_j *_l}} - \overline{\beta_{q_i *_j *_l}} - \overline{\beta_{*_i q_j *_l}} + \overline{\beta_{*_i *_j *_l}}\right)\end{aligned} \quad (19)$$

$$\begin{aligned}\hat{\sigma}_{q_i q_j} = \sigma_{q_i q_j} &- \left(\overline{\sigma_{*_i q_j}} + \overline{\sigma_{q_i *_j}} - \overline{\sigma_{*_i *_j}}\right) + \\ &\left(\overline{\sigma_{q_i q_j *_k}} - \overline{\sigma_{q_i *_j *_k}} - \overline{\sigma_{*_i q *_k}} + \overline{\sigma_{*_i *_j *_k}}\right) + \\ &\left(\overline{\sigma_{q_i q_j *_l}} - \overline{\sigma_{q_i *_j *_l}} - \overline{\sigma_{*_i q_j *_l}} + \overline{\sigma_{*_i *_j *_l}}\right)\end{aligned} \quad (20)$$

Similarly for the main effect terms:

$$\begin{aligned}\hat{\beta}_{q_i} = &\beta_{q_i} - \overline{\beta_{*_i}} + \left(\overline{\beta_{q_i *_j}} - \overline{\beta_{*_i *_j}}\right) + \left(\overline{\beta_{q_i *_k}} - \overline{\beta_{*_i *_k}}\right) \\ &+ \left(\overline{\beta_{q_i *_l}} - \overline{\beta_{*_i *_l}}\right) + \left(\overline{\beta_{q_i *_j *_k}} - \overline{\beta_{*_i *_j *_k}}\right) \\ &+ \left(\overline{\beta_{q_i *_j *_l}} - \overline{\beta_{*_i *_j *_l}}\right) + \left(\overline{\beta_{q_i *_k *_l}} - \overline{\beta_{*_i *_k *_l}}\right)\end{aligned} \quad (21)$$

$$\begin{aligned}\hat{\sigma}_{q_i} = &\sigma_{q_i} - \overline{\sigma_{*_i}} + \left(\overline{\sigma_{q_i *_j}} - \overline{\sigma_{*_i *_j}}\right) + \left(\overline{\sigma_{q_i *_k}} - \overline{\sigma_{*_i *_k}}\right) \\ &+ \left(\overline{\sigma_{q_i *_l}} - \overline{\sigma_{*_i *_l}}\right) + \left(\overline{\sigma_{q_i *_j *_k}} - \overline{\sigma_{*_i *_j *_k}}\right) \\ &+ \left(\overline{\sigma_{q_i *_j *_l}} - \overline{\sigma_{*_i *_j *_l}}\right) + \left(\overline{\sigma_{q_i *_k *_l}} - \overline{\sigma_{*_i *_k *_l}}\right)\end{aligned} \quad (22)$$

The intercepts then capture the average deviations from zero:

$$\begin{aligned}\hat{\beta}_0 = \beta_0 &+ \overline{\beta_{*_1}} + \overline{\beta_{*_2}} + \overline{\beta_{*_3}} + \overline{\beta_{*_4}} + \overline{\beta_{*_1 *_2}} + \overline{\beta_{*_1 *_3}} \\ &+ \overline{\beta_{*_1 *_4}} + \overline{\beta_{*_2 *_3}} + \overline{\beta_{*_2 *_4}} + \overline{\beta_{*_3 *_4}} + \overline{\beta_{*_1 *_2 *_3}} \\ &+ \overline{\beta_{*_1 *_2 *_4}} + \overline{\beta_{*_1 *_3 *_4}} + \overline{\beta_{*_2 *_3 *_4}}\end{aligned} \quad (23)$$

$$\begin{aligned}\hat{\sigma}_0 = \sigma_0 &+ \overline{\sigma_{*_1}} + \overline{\sigma_{*_2}} + \overline{\sigma_{*_3}} + \overline{\sigma_{*_4}} + \overline{\sigma_{*_1 *_2}} + \overline{\sigma_{*_1 *_3}} \\ &+ \overline{\sigma_{*_1 *_4}} + \overline{\sigma_{*_2 *_3}} + \overline{\sigma_{*_2 *_4}} + \overline{\sigma_{*_3 *_4}} + \overline{\sigma_{*_1 *_2 *_3}} \\ &+ \overline{\sigma_{*_1 *_2 *_4}} + \overline{\sigma_{*_1 *_3 *_4}} + \overline{\sigma_{*_2 *_3 *_4}}\end{aligned} \quad (24)$$

This procedure does not change the genetic score of any variant, and it does not change the likelihood of any model. Constraining the model this way merely centers and scales coefficients so that each represents the average effect of an amino acid or interaction relative to the average of the entire data set (and to the average predicted by lower order terms). In practice, we found that

*L*2 regularization brings the average of most sets of coefficients close to zero anyway, so the post-hoc constraint changes most terms by less than 1% of the unconstrained estimates (**Figure 1-Figure Supplement 5**).

Alternative models without epistasis

To understand the effects of specific epistasis on protein function and evolution, we compared the third-order model described above to models that do not incorporate any epistasis, and thus contain only first-order terms (first-order model), or no higher-order epistasis, and thus contains only first- and second-order terms (second-order model). To implement these lower-order models, we modified the matrix of indicator variables by removing all indicators for combinations at the orders to be excluded. These models were then fit to the data using analogous selection procedures for the full model containing first-, second-, and third-order terms.

Comparison to observed DMS data

To estimate the quality of the model fit, we compared the observed class of each protein variant to the observed classification from the empirical data. In each case, we calculated the number of true positives, true negatives, false positives, and false positives for each of the three activation categories, as well as a "non-strong" category (union of null and weak) and an "activators" category (union of weak and strong) (**Figure 1-Figure Supplement 2**).

Comparison to measured ΔG

We compared the estimated genetic score for a subset of protein variants to their energy of binding to ERE and SRE, previously measured using fluorescence anisotropy (139) (**Figure 1-Figure Supplement 4**). We used simple least-squares linear regression to estimate the relationship between the genetic score and the Δ*G* of binding. Because both genetic score and Δ*G* accrue additively, the relationship between these quantities should be linear. The slope and intercept of the relationship between genetic score and Δ*G* was then used it to estimate the Δ*G* of binding for all variants in the data set from their genetic score.

Model variance explained by individual terms

The reference-free coding of model coefficients leads to a simple relationship between the effect size of a term and the fraction of genetic variance in phenotype that it explains. A detailed proof is provided in Appendix 1 and summarized here.

The genetic variance is defined as the variance in pheno-type attributable to all the effects of individual genetic states and combinations; it excludes genetic variance caused by nonspecific epistasis (which is captured by the logit function in our model). The total genetic variance is the average squared deviation of each variant's genetic score from the global mean:

$$Var(y(G)) = \frac{1}{2 \cdot 20^4} \sum_{g \in G} \left(y(g) - \overline{y(G)} \right)^2 \qquad (25)$$

where G is the set of all protein variants on either RE, y(g) is the genetic score of a particular protein-RE complex, and $2 \cdot 20^4$ is the total number of complexes.

The total variance explained by a model can be partitioned into the partial variances explained by each causal factor. In the case of our model, the total genetic variance can be partitioned into the partial variances attributable to the possible states at each site or (combination of states at multiple sites). Organizing these partial variances by the epistatic order and effect type (nonspecific or RE-specific):

$$Var(y(G)) = \sum_i Var(\beta_i) + \sum_{i<j} Var(\beta_{i,j}) + \sum_{i<j,j<k} Var(\beta_{i,j,k}) + Var(\sigma_0) + \sum_i Var(\sigma_i) + \sum_{i<j} + Var(\sigma_{i,j}) + \sum_{i<j,j<k} Var(\sigma_{i,j,k}) \quad (26)$$

where i, j, and k index the four variable sites, the variances are over all the possible states (or combinations), and the summations are over all sites (or combination).

The model coefficients have a straightforward relationship to this partitioned variance. Main-effect coefficients are defined as the average deviation of a subset of variants containing a state or combination relative to the global mean, and higher-order coefficients are defined as deviations from the sum of lower-order terms. The variance attributable to any state (or combination) is therefore simply the square of its coefficients, and the variance attributable to any set of states is the average of the squared coefficients (see proof in Appendix 1). Thus, for any given site i (or site combination i,j or i,j,k), the variance attributable to the set of model terms applicable to that site can be written as follows:

$$Var(\beta_i) = \frac{1}{20} \sum_q \beta_{q_i}^2 \quad (27)$$

$$Var(\beta_{i,j}) = \frac{1}{20^2} \sum_{q^2} \beta_{q_i q_j}^2 \quad (28)$$

$$Var(\beta_{i,j,k}) = \frac{1}{20^3} \sum_{q^3} \beta_{q_i q_j q_k}^2 \quad (29)$$

$$Var(\sigma_0) = \sigma_0^2 \quad (30)$$

$$Var(\sigma_i) = \frac{1}{20} \sum_q \sigma_{q_i}^2 \quad (31)$$

$$Var(\sigma_{i,j}) = \frac{1}{20^2} \sum_{q^2} \sigma_{q_i q_j}^2 \quad (32)$$

(33)

The leading factors in **equations 25-31** can can all be expressed as a function of *O*, the epistatic order of the effect they represent, where *O* = 0 for the global average effect of the RE, and 1, 2, or 3 for the main, pairwise, and third-order epistatic effects of each amino acid combination.

Substituting into **equation 24**, the total genetic variance is therefore the sum of the squared effects of every model term at every order (except for the global average), each divided by the number of model terms at that order:

$$\sum_{q^2}\sigma_{q_1q_2}=\sum_{q^2}\sigma_{q_1q_3}=\sum_{q^2}\sigma_{q_1q_4}=\sum_{q^2}\sigma_{q_2q_3}=\sum_{q^2}\sigma_{q_2q_4}=\sum_{q^2}\sigma_{q_3q_4}=0 \quad \textbf{(16)}$$

We can use this same approach to calculate $F\,(Var(\theta))$, the fraction of the total genetic variance explained by any coefficient θ (a particular β or σ of any order, except the global average):

$$F(Var(\theta))=\frac{\theta^2}{20^{O(\theta)}}Var(y(G)) \quad \textbf{(35)}$$

The fraction of genetic variance explained by any set of coefficients is simply the sum of $F\,(Var(\theta))$ over all the coefficients in the set.

This formalizes the intuition that within an epistatic order, model terms of large magnitude explain more variation than smaller terms; if two terms at different order have the same magnitude, the lower-order term explains more variation than the higher-order term, because the former affects more genotypes than the latter. The leading terms effectively weight each set of coefficients in a set by the number of genotypes to which they apply.

We used this simple relationship between the magnitude of a model coefficient, its order of epistatic effect, and the fraction of genetic variance explained by that coefficient to directly calculate the percent of variance explained by every model term. We then summed the percent variance explained by individual terms to determine the percent of variance explained by sets of terms (**Fig. 2**). Finally, we used this relationship to define the 'important' terms in the model as those terms needed to explain 99% of the model variance. Thus setting all of the 'unimportant' terms to zero reduces the amount of variance explained by less than 1% of the full model (**Figure 2-Figure supplement 1**).

Fraction of genetic score attributable to epistasis or specificity

To find the relative impact of epistasis terms on each variant's genetic score, we calculated the sum of the absolute value of all pairwise and third-order model terms for that genotype and divided it by the sum of the absolute value of all model terms for that genotype (**Figure 3**; **Figure 3-Figure supplement 1**; **Figure 3-Figure supplement 2**). For specificity, we summed the absolute value of only the specificity terms divided by the total absolute sum of all terms.

Types of epistasis

Epistasis can be divided into several subtypes depending on the signs and magnitude of mutational effects on different genetic backgrounds. These distinctions are most commonly made for pairwise epistatic interactions. For example, sign epistasis arises when the direction of effect of a mutation depends on the genetic background. Reciprocal sign epistasis occurs when the direction of effect of both mutations depends on the genetic background. All other instances of epistasis are classified as magnitude epistasis. Within magnitude epistasis, there are two broad forms, diminishing returns and diminishing costs epistasis. The former, diminishing returns, also includes common forms of epistasis such as amplifying costs, differing only on which genotype is considered 'wild-type' or the starting point of a mutant cycle. Likewise, diminishing costs epistasis also covers amplifying returns epistasis, only differing in which genotype is considered the beginning of a mutant cycle.

To classify the types and frequencies of different forms of epistasis, we compared the estimated effect of each pairwise interaction to the estimated effect of each single amino acid involved in that pairwise interaction. The fraction of genetic variance explained by the particular pairwise effect was then added to the appropriate category, either sign, reciprocal sign, diminishing returns, or diminishing costs epistasis. To extend this approach to third order interactions, we sequentially fixed each amino acid in the third order interaction, taking that genotype as the starting genotype and then varied the other two amino acids in a manner analogous to the pairwise epistasis calculation. The fraction of genetic variance explained by a particular third order interaction was then divided by three (as there are three states that can be fixed) and added to the appropriate category (**Figure 3**; **Figure 3-Figure supplement 4**).

Direction of the effect of mutations' effects

The genetic score of a specific genotype is the summation of the effects in that sequence, i.e. the single global term, four single site effects, six pairwise interactions, and four three-way interactions for both binding and specificity. The model thus gives the effect of each amino acid or interaction relative to the average of all genotypes, not the effect of individual mutations from one state to another. To calculate the effect of a mutation in a given genetic background, we calculated the difference in the genetic score between two genotypes with a single amino acid difference. This entails a change in one main effect, three pairwise interactions, and three third-order interactions. In this way, the model directly captures the fact that even a single amino acid change has the potential to alter numerous epistatic interactions. As a result, to determine the effect of a mutation requires measuring its impact across numerous genetic backgrounds. To determine the range of possible effects of a single mutation, we compared the genetic score between two genotypes differing by a single amino acid on each of the 8000 genetic backgrounds on which the particular change from one amino acid to another could occur and determined the frequency in which that particular mutation increased or decreased binding on a particular DNA element. To minimize the effect of small changes in genetic score upon mutation being counted, we only considered changes in genetic score greater than 5% of the difference between the average genotype and the threshold needed to be classified as a strong activator. We also used the fact that the model identifies which epistatic factors (main, pairwise, third-order) are responsible for the change in genetic score between variants, as well as how this effect is partitioned between binding and specificity, to identify when changes in epistasis were necessary, sufficient, or both for the change in genetic score (**Figure 4**; **Figure 4-Figure supplement 1**).

Sequence space analysis

Network Types

To determine how epistasis affected the evolution of DBD binding specificity, we generated connected networks for each model of epistasis using the R package igraph (140). In each case, genotypes were connected if they differed by a single amino acid (hamming distance) or if the single amino acid difference could occur via a single nucleotide mutation (genetic code distance). In the latter case, we accounted for the disjoint nature of serine codons in the genetic code, treating the two groups as non-interchangeable by coding all serine amino acids in both an S and a Z form. Thus, for a sequence containing n serine amino acid, there are 2n versions of that amino acid sequence with the same phenotype that inhabit different locations in sequence space with different connectivity to neighboring sequences. We created networks for the full model, the model containing up to pairwise epistasis, and the model containing no specific epistasis (**Figure 5**; **Figure 5-Figure supplement 1**). We used only those genotypes that were functional in a particular model, i.e. were either weak or strong on either ERE, SRE, or both. In addition, we created a network of all 160,000 genotypes. This latter network contained no information about the function of each genotype and was used to model evolution over sequence space unconstrained by function.

Network distance

Network distances were calculated as the shortest path distance between two genotypes, constrained to only take steps to neighboring genotypes in a network. Because only functional genotypes are present in a network, the distance between genotypes in a network can exceed four, even when connections are made by hamming distance, if there are no direct paths between the genotypes. We also calculated the distance from each ERE-specific genotype to the nearest SRE-specific genotype, of which there may be several, by finding the minimum distance for each ERE-specific genotype (**Figure 5-Figure supplement 2**).

Shared vs unique network genotypes

One of the main effects of epistasis is to influence whether a genotype is functional or not. Thus, many genotypes are only functional in one or a subset of networks. However, even if the same genotypes are functional with and without epistasis, epistasis can influence the paths between them that evolution is likely to take. To distinguish between these effects, we compared the effects of epistasis on path lengths for all functional genotypes and for the set of genotypes that are functional in all three networks.

One-step functional transitions and epistatic necessity/sufficiency

To determine how epistasis influenced changing binding classes or specificity, we identified all instances in which two neighboring genotypes in a network differ in binding class or specificity. These represent portions of sequence space where a single amino acid mutation can move a genotype from one binding class to another or alter specificity from ERE-specific to SRE-specific. For changes between binding classes, we identified all such mutations on the full network with third order epistasis. For changes in specificity, we identified the mutation responsible for all such instances in the full network, as well as the networks that had no third-order epistasis or had no epistatic effects (**Figure 4**; **Figure 4-Figure supplement 2**).

We calculated whether epistasis was necessary, sufficient, or both for transitions in binding classes and specificity. For binding classes, epistasis was categorized as necessary if the main effect of the amino acid change was not large enough to move the genotype into the new binding class without contribution from the epistatic effects. Similarly, epistasis was categorized as sufficient if epistatic changes were large enough to move the genotype into the new binding class without contribution from the main effects. For single step transitions in specificity, epistasis was necessary if the main, non-epistatic, effect of an amino acid change was not large enough to either reduce ERE binding or increase SRE binding enough to switch the binding class from ERE-specific to SRE-specific. Epistasis was sufficient if the epistatic effects alone were large enough to both reduce ERE binding and increase SRE binding enough to switch the binding class from ERE-specific to SRE-specific. In the full model, a similar distinction was made for third order epistasis relative to main effects and pairwise epistasis and calculated in an analogous manner.

Number of paths and path distinctiveness

The number of paths between two genotypes was calculated as the number of distinct shortest paths between them. Two paths were considered distinct if at least one genotype along the paths were different. To account for the fact that many paths traverse similar sets of genotypes, we also developed a metric for distinctiveness of a set of paths between two genotypes. To calculate this metric, we extracted from the larger network the set of genotypes along any shortest path between the starting and ending genotype. Intuitively, for a path with S genotypes, the set of paths between the starting and ending genotypes should be considered more distinct if they contain more unique genotypes or fewer shared edges between intermediate genotypes. To capture these aspects, we calculated two metrics. The first metric relates the effective number of paths (P_g) to the number of

unique genotypes (G) and the path length (S - 1). The second metric relates a different measure of the effective number of paths (P_e) to the number of unique edges (E) and the number of genotypes along a single path (S).

$$P_g = \frac{G-2}{S-1} \quad \textbf{(36)}$$

$$P_e = \frac{E}{S} \quad \textbf{(37)}$$

The first calculation provides the maximum number of completely distinct paths with no shared genotypes that could be created given the path length and the number of genotypes present. It thus provides an upper bound on the number of distinct paths possible. If a set of paths are completely distinct, i.e. sharing no genotypes, then P_g will equal P_e. However, if the paths are not completely distinct, then there will be more edges among the set of paths than if all paths were distinct. Thus, for a given number of genotypes and path length, the greater the number of edges connecting those genotypes, the less distinct the set of paths are. By comparing the difference in these metrics to the maximum possible value, we can estimate the effective number of distinct paths in a set.

$$D = P_g - (P_e - P_g)\frac{P_g}{P_e} = \frac{P_g^2}{P_e} = \frac{S(\frac{G-2}{S-1})^2}{E} \quad \textbf{(38)}$$

We used this equation to calculate the distinctiveness of paths from each ERE-specific genotype to each SRE-specific genotype over the network.

Evolutionary simulations

Neutral Model

To determine whether epistasis increased or decreased the ability of evolution to find a protein variant with altered specificity, we performed forward evolutionary simulations over the previously described genotype networks. These simulations captured neutral wandering among a set of functional variants and are similar to the verbal model of protein evolution described by Maynard Smith (19). In each simulation, evolution was constrained to move among only functional variants, i.e. those classified by the particular model as either ERE-specific, SRE-specific or promiscuous. Genotypes were classified as either ERE-specific if they bound ERE and not SRE, SRE-specific if they bound SRE and not ERE, and promiscuous if they bound both ERE and SRE. Genotypes were considered neighbors if a single nucleotide change could mutate one variant into the other (genetic code) or if they differed by a single amino acid (hamming distance). To estimate the global effect of epistasis, evolution was initiated from each ERE-specific genotype and allowed to take a random step to a neighboring genotype, regardless of DNA binding specificity, until an SRE-specific genotype was reached. This was repeated 100 times for each starting genotype. For the subset of ERE-specific genotypes found in each of the three models with varying amounts of epistasis (Full model, pairwise only, no epistasis), an additional 1000 simulations were performed to estimate how epistasis altered particular evolutionary paths. For each simulation, we recorded the number of steps, ending genotype identity, and the identity of each intermediate genotype during the evolutionary walk.

Model with selection for increased specificity

To estimate how epistasis might interact with selection for increased specificity, we performed additional forward evolutionary simulations. As before, evolution was initiated from ERE-specific genotypes and terminated when an SRE-specific genotype was first encountered. Unlike the

neutral simulations, however, we incorporated selection for increased specificity. At each step, the specificity of each neighboring genotype was calculated from the underlying genetic scores as the difference in binding to SRE and ERE. We then used Kimura's formula for fixation probability to calculate the relative probability of fixation among all possible neighbors (141). The effects of specificity were scaled such that the difference between a non-binder and a strong binder corresponded to a scaled selection coefficient (i.e. Ns) of 1 in a population of size 100. We initiated evolution with a range of population sizes between 1 and 10,000 to capture the transition between neutrality and selection driven changes, performing 100 simulations for each population size on each network using the genetic code to connect genotypes. We noticed during these simulations, that for a small subset of starting genotypes, evolution would often become 'stuck' on local optimum in the pairwise model, continuously cycling between two genotypes. This was more common for larger population sizes and had the effect of drastically increasing the average evolutionary path length away from the median, or typical, path length.These genotypes were substantially more SRE-specific than the neighboring functional sequences. Simulations with strong selection for specificity (i.e. large population sizes) were therefore more likely to continuously cycle between these genotypes. To account for this effect, we also estimated evolutionary path lengths after removing all evolutionary simulations in which more than 20 steps were taken (**Figure 5-Figure Supplement 3**).

Data Availability

The coding scripts for running all analyses are located on Github (*https://github.com/JoeThorntonLab/DBD.GeneticARchitecture*). Initial and intermediate data files can be found at dryad (*https://doi.org/10.5061/dryad.jsxksn0hk*).

Acknowledgements

We thank members of the Thornton lab for helpful comments on the manuscript and the University of Chicago's Research Computing Center for computational resources.

Supplementary Note 1: Supplementary Figures

Supplementary Note 2: Appendix 1

Partitioning of Variance. In our modeling formalism, the importance of each model term – the fraction of total specific genetic variance that it accounts for – can be calculated directly from its magnitude and order. In addition, these microscopic contributions can be summed over any set of terms to yield the fraction of variance accounted for by the set. Specifically, the fraction of variance accounted for by any term is equal to the square of the term, weighted by the number of genotypes to which it applies, which is a function of the term's epistatic order (main, pairwise, third-order, etc.). This property flows ultimately from the constraint we placed on the model coefficients to sum to zero at each site and combination of sites. Here we show why this relationship holds.

We begin with the definition of specific genetic variance, which is the variance in the genetic score among all genotype:

$$Var(y(G)) = \frac{1}{2 \cdot 20^4} \sum_{g \in G} (y(g) - \overline{y(G)})^2 \quad (1)$$

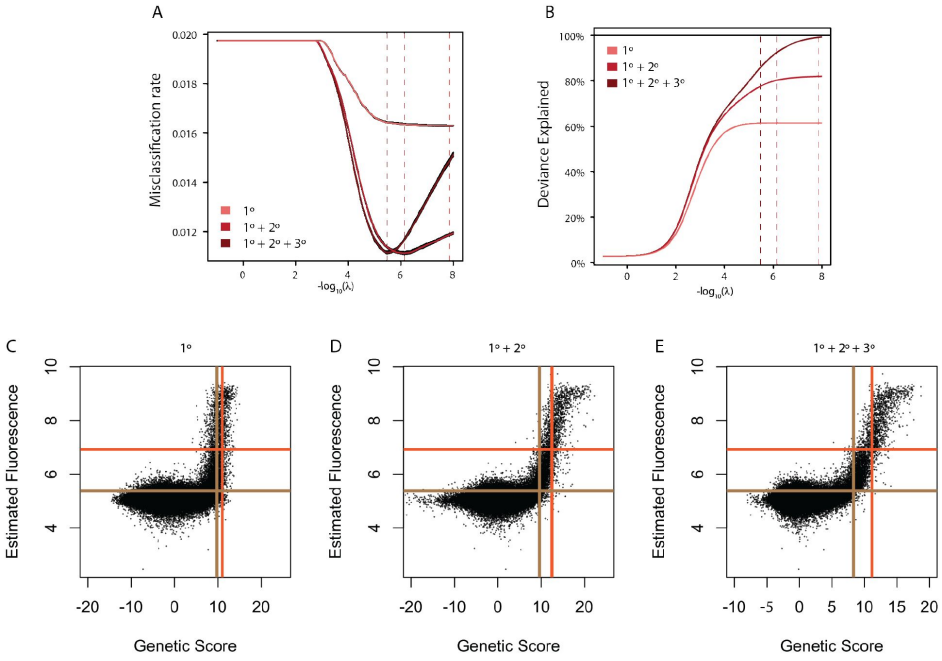

Figure 1 - Figure Supplement 1

A. Average misclassification rate. Misclassification rate was measured by 10-fold cross-validation. The x-axis gives the lambda value used in ridge regression for the models containing main effects (1°, light red), main and pairwise epistatic effects (1°+2°, medium red), and main, pairwise, and third-order epistatic effects (1°+2°+3°, Full model, dark red). In each case, variance around the estimate from 10 independent cross-validation runs is shown in black. Dashed lines show the lambda value at which misclassification error is minimized. **B**. Deviance explained by each model as the lambda ridge regression penalty parameter is reduced. Deviance is measured as 2 times the difference in log(Likelihood) of the model compared to a saturated model that contains a single parameter for every data point and is thus a perfect predictor of genotype binding class. For generalized linear models, deviance is the measure analogous to R2. Dashed lines are the same as in A and thus maximize deviance explained without over-fitting. **C**. Comparison of estimated genetic score and mean fluorescence for the main effects only model (1°). Threshold genetic score between Null and Weak genotypes (vertical brown line) and between Weak and Strong genotypes (vertical orange line) are shown. The dashed horizontal lines show the 10th percentile of variants estimated as weak (orange) or strong (brown) based on their mean fluorescence. **D**. Same as C, but for the model containing pairwise epistasis (1°+2°). **E**. Same as C, but for the full model containing pairwise and third order epistasis (1°+2°+3°)

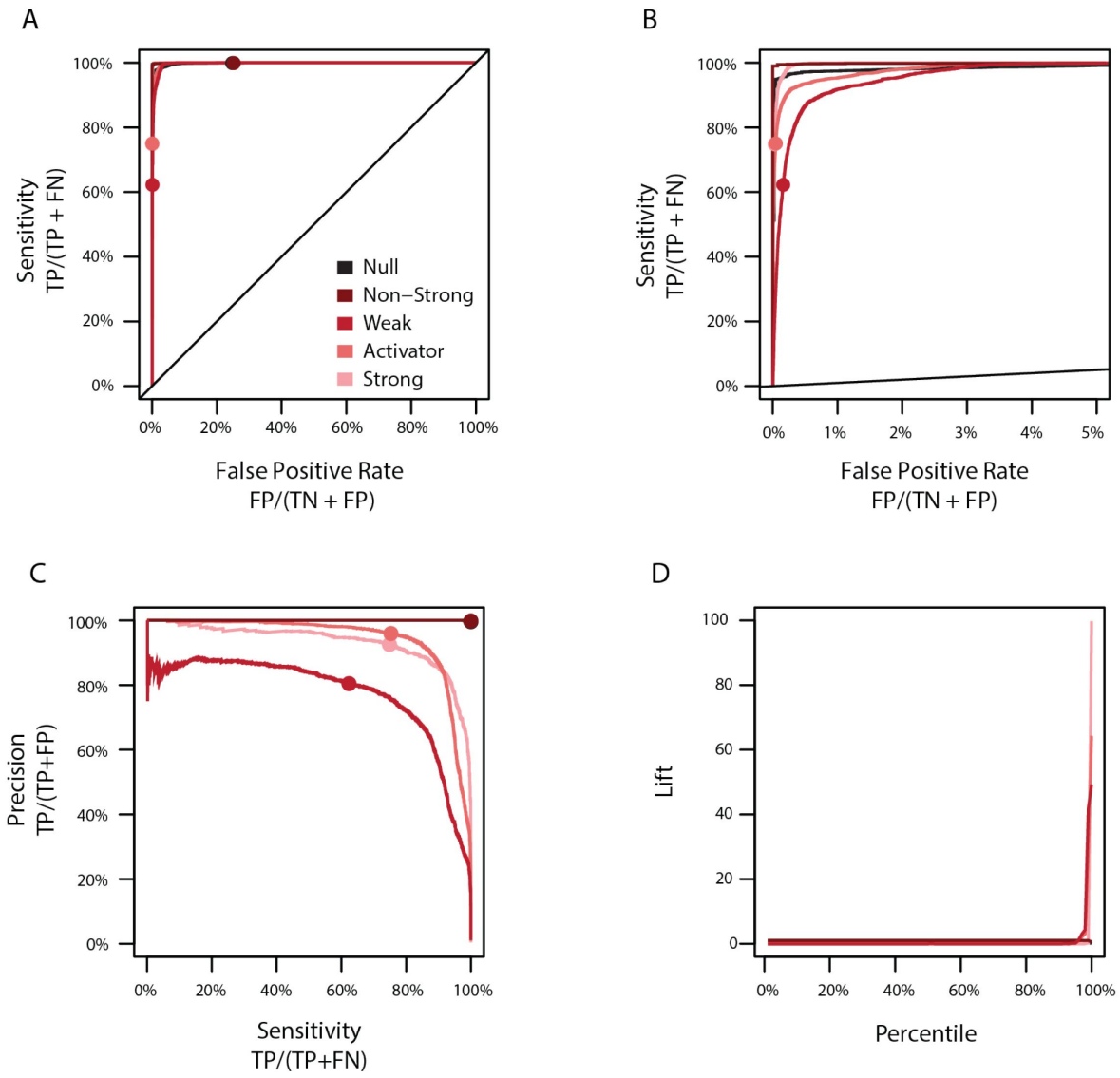

Figure 1 - Figure Supplement 2.

A. Sensitivity versus the false positive rate of the full model for each classification. The non-strong class (dark red) consists of genotypes labeled as either null (black) or weak (red), while the activator class (light red) consists of genotypes labeled as weak or strong (pink). Sensitivity is the percent of genotypes correctly estimated to be in a class (TP) among the set of all genotypes originally assigned to that class (TP + FN). The false positive rate is the percent of genotypes incorrectly estimated to be in a class (FP) among the set of all genotypes not originally assigned to that class (TN+FP). Circles show the sensitivity and false positive rate in the full model. **B**. Same as A, showing a reduced portion of the x-axis. **C**. Precision versus sensitivity of the full model for each classification. Colors are the same as C. Precision is the percent of genotypes correctly estimated to be in a class (TP) among the set of all genotypes estimated to be in that class (TP+FP). **D**. Lift of each percentile for each classification. Genotypes were ordered by genetic score and lift calculated as the frequency of each class in a percentile relative to the average frequency of each class in the entire data set.

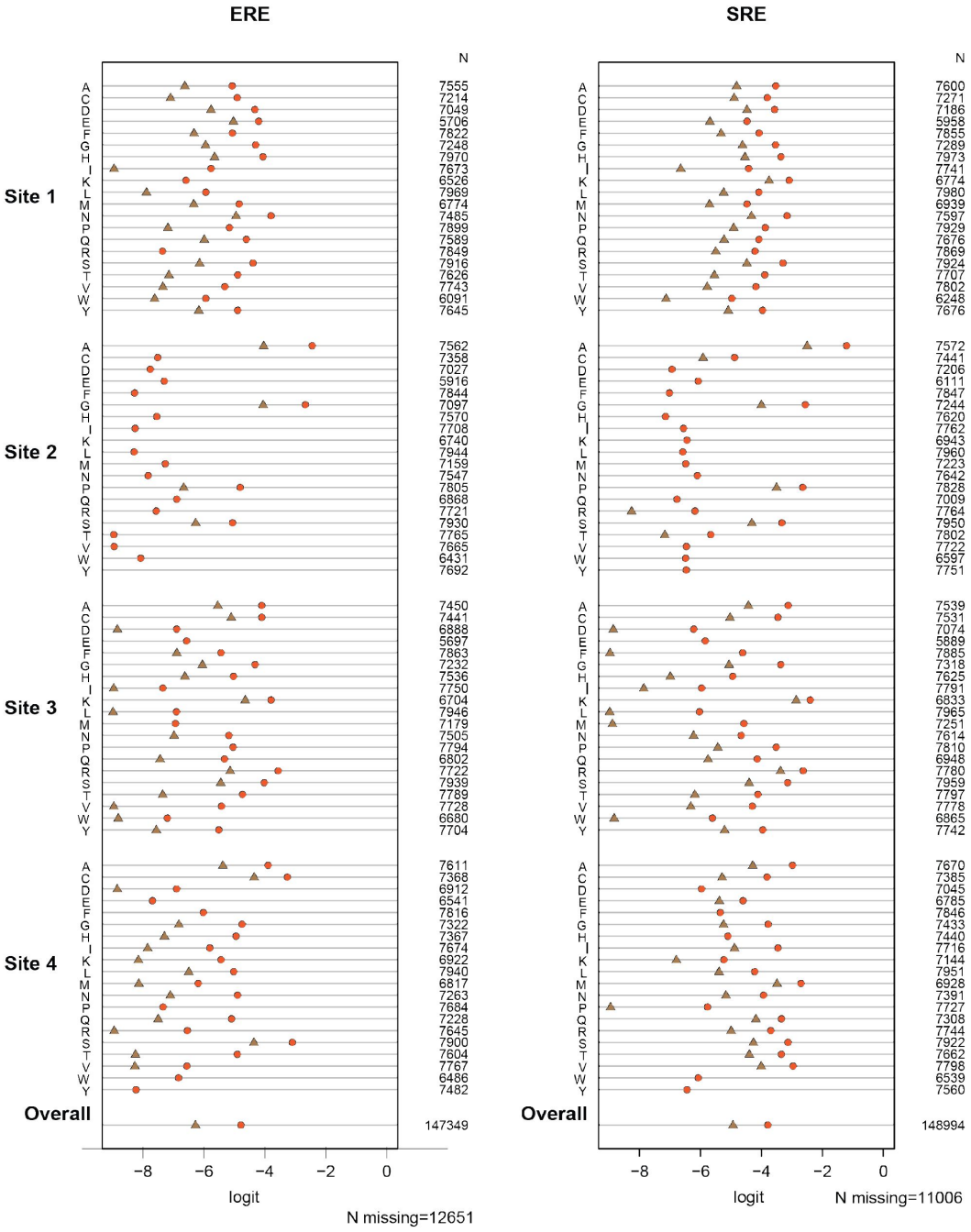

Figure 1 - Figure Supplement 3.

Test of proportional odds assumption. Each panel shows estimated main-effect coefficients for each amino acid at the variable sites using simple logistic regression, when the activation class of variants was coded in two different ways: null vs. weak+strong (brown), or null+weak vs. strong (orange). Under the proportional odds assumption, the difference between the two estimates should be the same across coefficients. Coefficients for the effect on ERE and SRE were estimated separately, and no epistasis was included in this preliminary model. Missing symbols indicate amino acid states never observed among strong activators (no brown) or only observed in null activators (no orange or brown). N gives the number of genotypes containing the specified state for which mean fluorescence values. The weighted average of all estimated coefficients is shown at the bottom.

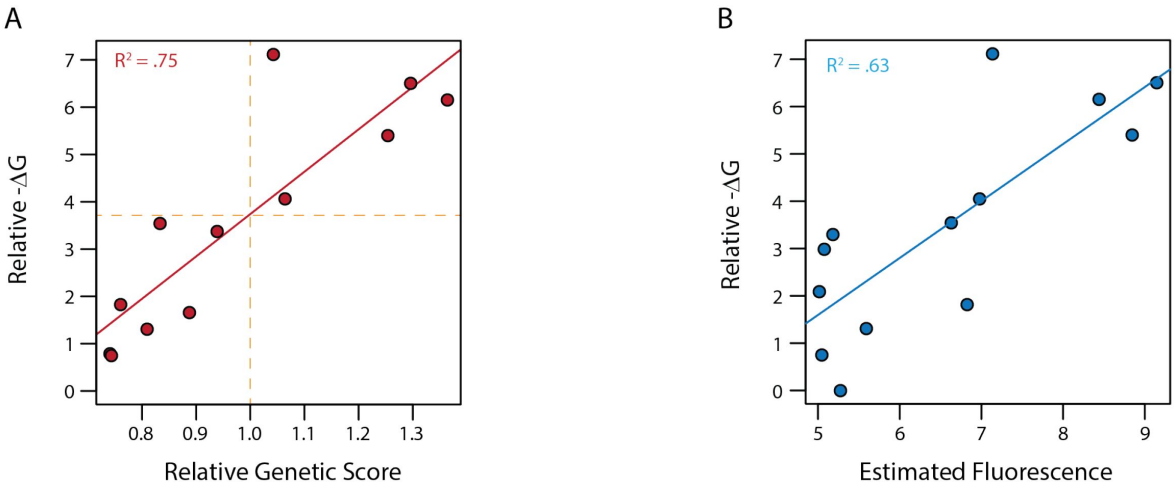

Figure 1 - Figure Supplement 4.

A. *δG* relative to the weakest binder as estimated from fluorescence polarization versus the estimated genetic score. Genetic scores are scaled relative to the threshold to be a strong binder (vertical orange dashed line), with zero being the average genetic score of all genotypes. Dashed horizontal line shows the *δG* values estimated to correspond to this threshold by fitting a linear model (solid red line). **B**. *δG* from fluorescence polarization versus the observed mean fluorescence of each genotype prior to categorization. Linear regression between *δG* and mean fluorescence (solid blue line).

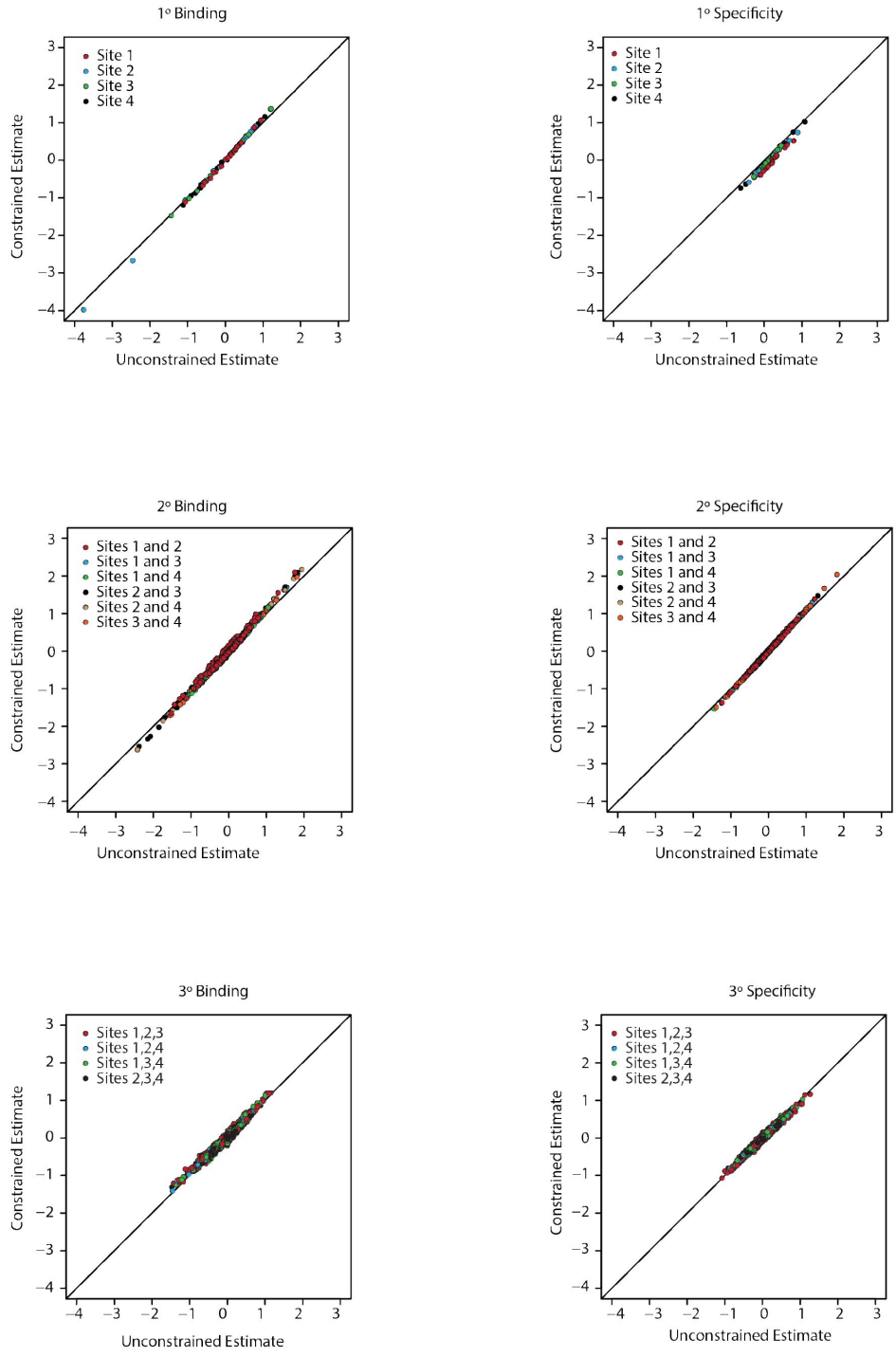

Figure 1 - Figure Supplement 5.

Unconstrained (x-axis) and constrained (y-axis) estimates for all coefficients in the third order model. Each panel is for a particular order of epistasis and either binding or specificity. Colors indicate the site, or combination of sites, for each set of coefficients.

A

Pairwise Epistasis
Number of Important Effects

	Positive Binding				Negative Binding				ERE Specificity				SRE Specificity			
	X1	X2	X3	X4	X1	X2	X3	X4	X1	X2	X3	X4	X1	X2	X3	X4
A	25	24	27	29	22	33	24	26	17	27	23	26	29	27	26	25
C	20	25	18	25	31	27	34	29	21	23	25	28	29	25	20	23
D	16	23	22	16	33	25	30	33	21	21	28	18	26	22	22	30
E	25	30	22	21	25	26	32	22	24	25	28	19	23	29	24	22
F	25	20	20	26	30	18	34	20	26	18	20	26	26	14	31	15
G	23	19	27	29	25	34	24	25	20	30	23	24	27	21	24	28
H	29	22	24	34	20	19	31	16	25	24	26	24	21	11	23	25
I	20	22	23	25	32	32	19	19	23	24	21	14	26	24	17	25
K	26	18	26	20	30	24	32	28	31	19	33	17	25	17	25	28
L	19	23	18	15	31	27	20	31	23	26	24	17	20	19	13	26
M	27	23	28	20	28	31	26	35	29	25	17	28	23	27	35	26
N	23	19	22	28	28	33	26	28	25	27	25	24	23	21	22	25
P	26	22	29	24	24	29	23	26	14	30	26	30	28	20	25	19
Q	22	26	23	29	35	30	30	19	30	25	27	27	23	29	20	18
R	25	29	23	22	25	27	31	28	22	23	23	25	21	30	29	24
S	23	26	19	25	28	27	31	32	26	21	26	24	24	30	21	31
T	17	28	31	20	30	22	23	30	19	29	18	19	22	14	30	28
V	23	15	14	21	29	28	38	30	25	16	27	25	23	24	19	21
W	25	21	19	17	28	25	27	34	25	23	24	22	27	19	20	27
Y	25	21	26	25	28	26	28	17	27	21	30	23	25	22	18	14

B

Third Order Epistasis
Number of Important Effects

	Positive Binding				Negative Binding				ERE Specificity				SRE Specificity			
	X1	X2	X3	X4	X1	X2	X3	X4	X1	X2	X3	X4	X1	X2	X3	X4
A	31	157	56	54	21	130	62	49	26	127	54	47	18	135	46	46
C	30	19	40	61	29	20	33	43	32	19	30	45	21	13	30	50
D	47	12	9	17	53	12	28	21	53	11	22	26	34	9	13	8
E	41	8	17	19	31	20	26	16	35	12	16	20	25	8	21	9
F	28	8	14	23	24	15	19	30	28	12	14	25	16	6	10	21
G	33	106	50	19	33	98	43	35	29	104	42	24	25	87	40	21
H	32	6	23	31	32	12	33	31	31	7	21	25	25	7	23	26
I	32	7	11	31	27	18	6	40	24	10	11	31	29	6	6	29
K	43	12	63	27	42	9	57	23	42	6	57	25	34	9	51	22
L	23	9	9	26	20	15	10	17	23	14	11	24	16	3	7	13
M	23	10	19	45	28	10	19	51	26	12	19	50	19	5	10	34
N	52	22	17	28	55	17	30	33	47	19	19	32	47	17	24	21
P	17	71	38	26	24	64	43	21	20	63	44	23	16	51	27	14
Q	22	14	23	38	34	16	36	44	18	12	29	40	27	13	20	34
R	36	15	104	38	35	18	90	30	31	16	85	27	30	11	89	34
S	30	49	40	47	39	46	44	59	32	41	38	54	30	40	35	41
T	30	8	25	36	23	12	20	37	26	7	19	38	20	7	15	27
V	33	11	18	36	31	17	24	38	37	16	17	31	19	10	15	32
W	23	15	15	20	31	15	23	24	28	17	21	23	17	10	10	11
Y	37	10	35	18	27	8	33	9	22	13	31	12	32	4	33	11

Figure 2 - Figure Supplement 1.

A. Number of pairwise interaction terms involving each amino acid (rows) at each site (columns) that are needed to explain 99% of the full model genetic variance for positive binding, negative binding, ERE specificity, and SRE specificity. In each case, an amino acid can be involved in a maximum of 60 interactions (20 at each of the other three sites). **B**. Same as left, but for third-order interactions. Each amino acid can be involved in a maximum of 1200 interactions (400 at each pair of the other three sites).

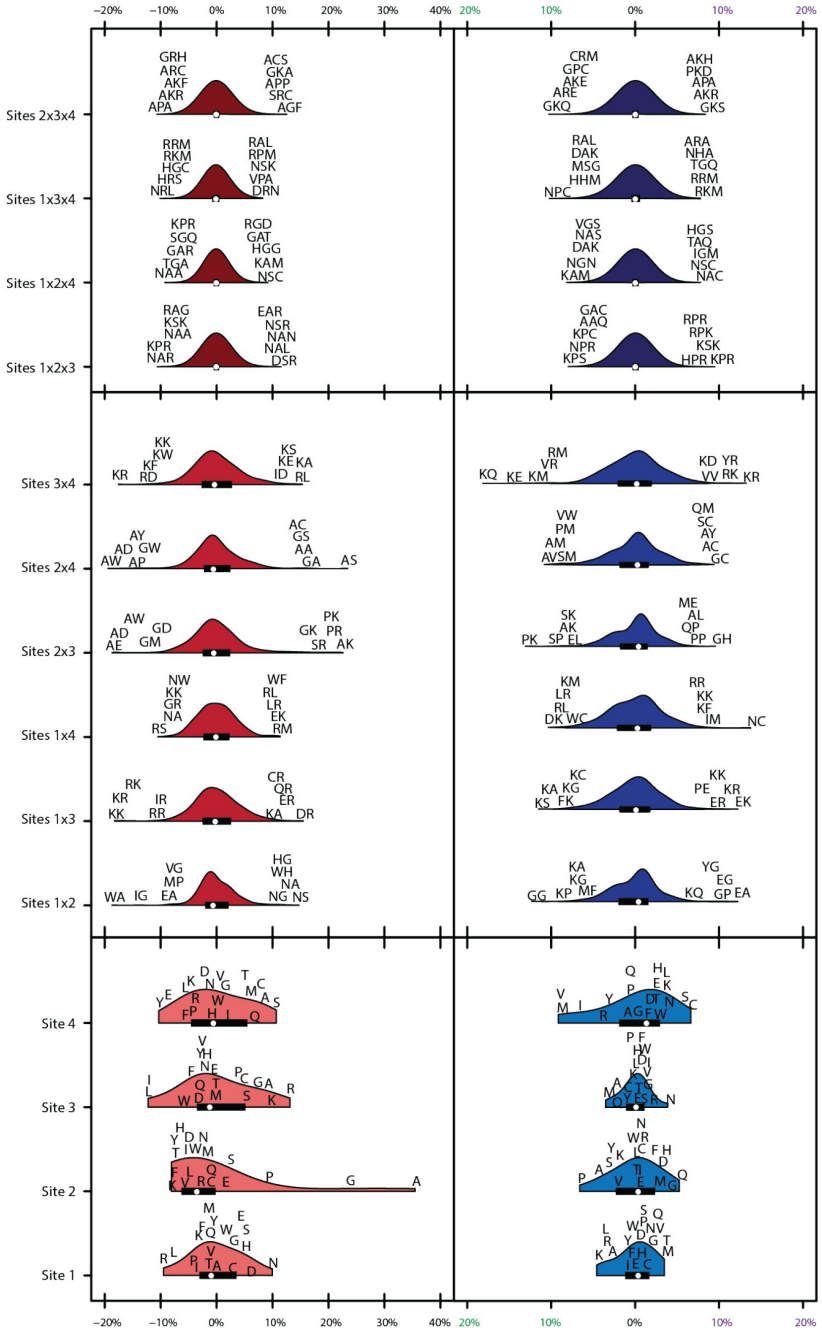

Figure 2 - Figure Supplement 2.

Distribution of effects for each site/site combination for the full model (1°+2°+3°). Effects are scaled to the value needed to be a strong binder. Binding effects (red) and specificity effects (blue) are separated. For main effects (light, bottom), all 20 amino acids are shown with single letter codes. For pairwise epistatic effects (medium, middle) and third-order effects (dark, top), only the most extreme five values for each site combination are shown.

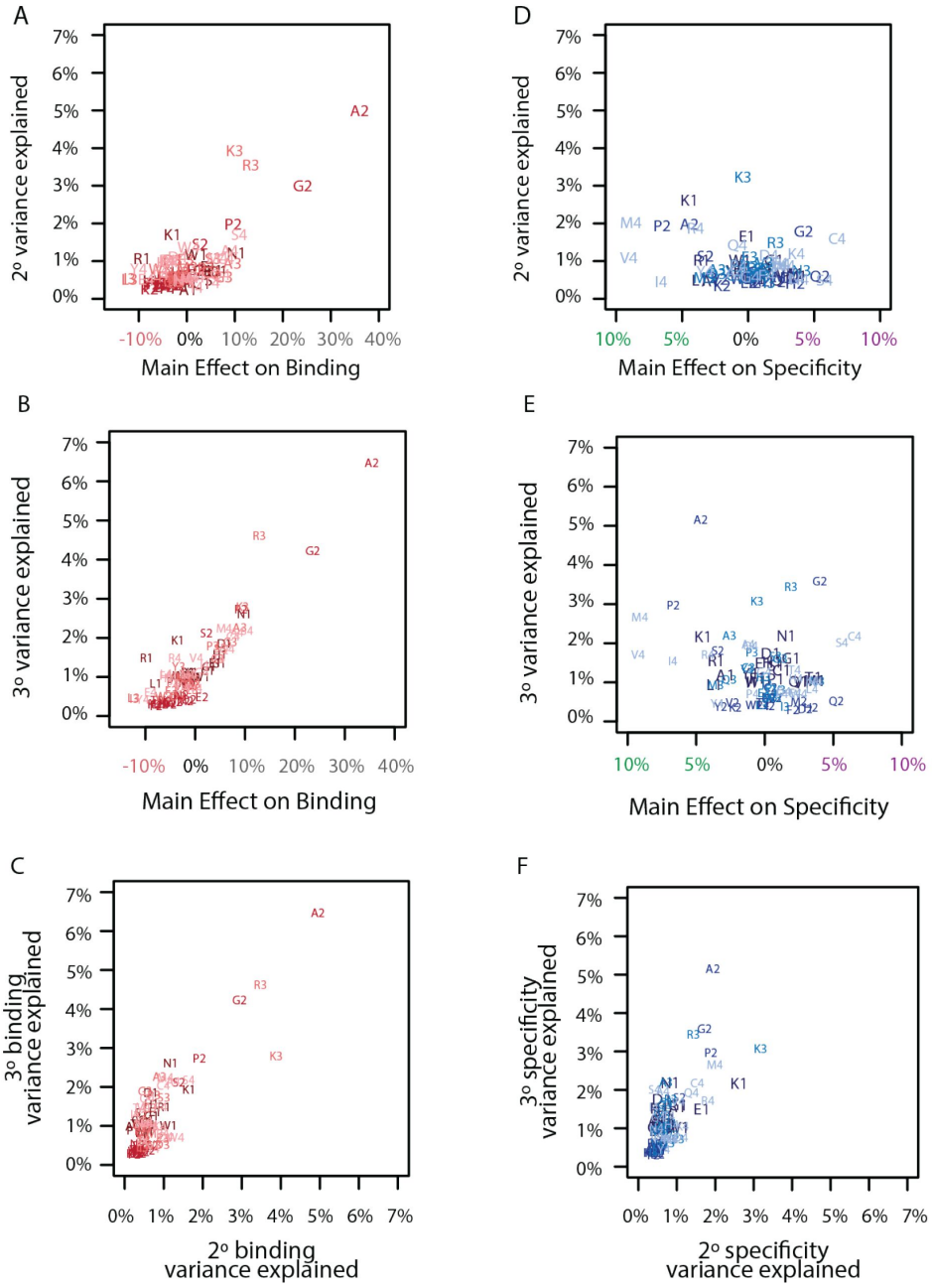

Figure 2 - Figure Supplement 3.

A. Comparison of the main effect of each amino acid on binding with its epistatic effects on binding. Each point is a different amino acid (letter) at a different site (number). Each site is a different color from dark (site 1) to light (site 4). **B**. Same as A, but for third-order interactions. **C**. Same as A, but for the comparison of pairwise and third-order binding effects. **D**. Comparison of the main effect of each amino acid on specificity with its epistatic effects on specificity. Each point is a different amino acid (letter) at a different site (number). Each site is a different color from dark (site 1) to light (site 4). **E**. Same as D, but for third-order interactions. **F**. Same as D, but for the comparison of pairwise and third-order specificity effects.

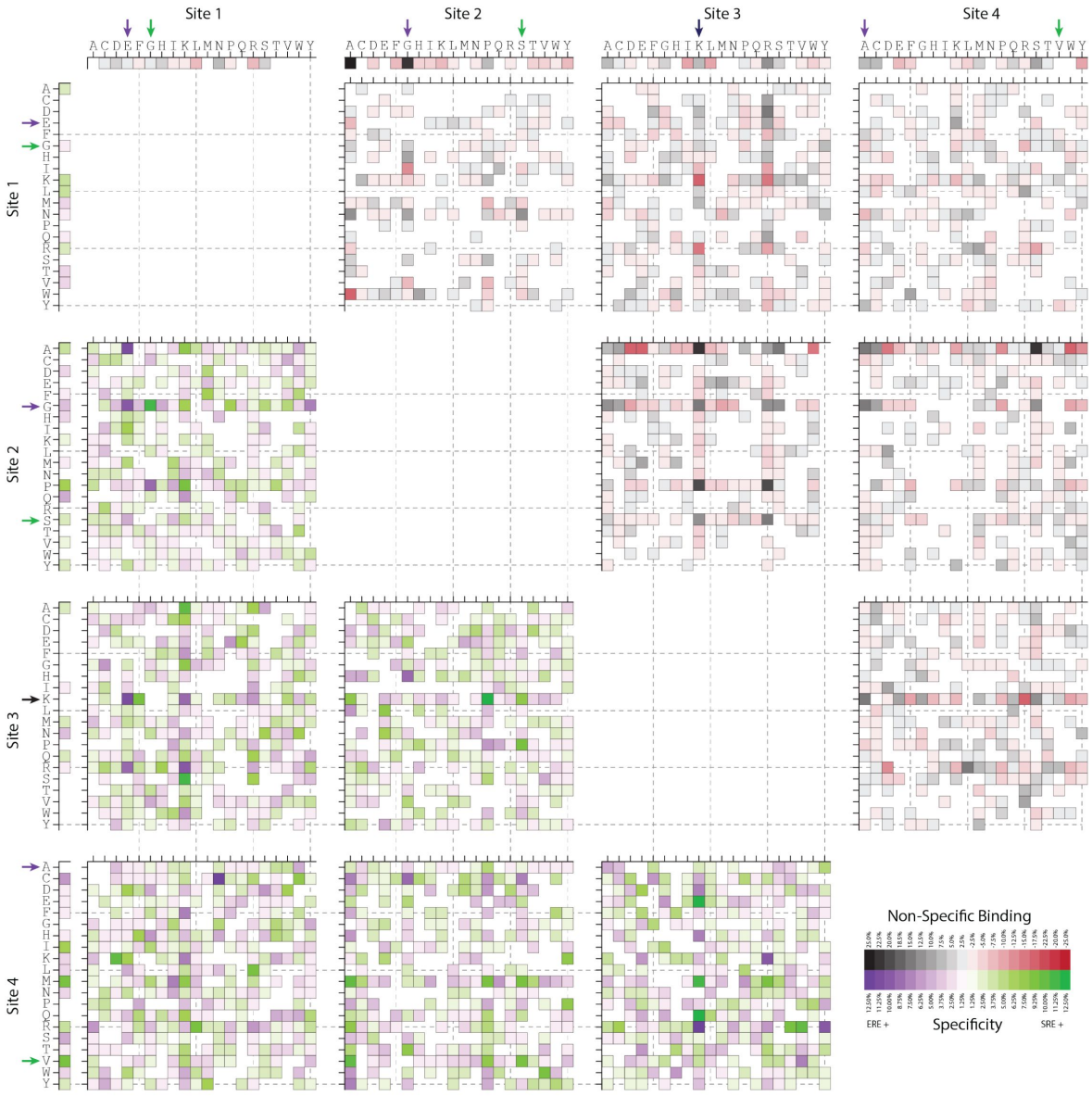

Figure 2 - Figure Supplement 4.

First- and second-order effects on non-RE-specific binding (above the diagonal) and on specificity (below) for all amino acid sites and pairs. Effects are scaled relative to the value necessary to be classified as a strong binder. Arrows indicate the states found among extant steroid receptors (purple, estrogen receptors; green, ketosteroid receptors; black, both categories).

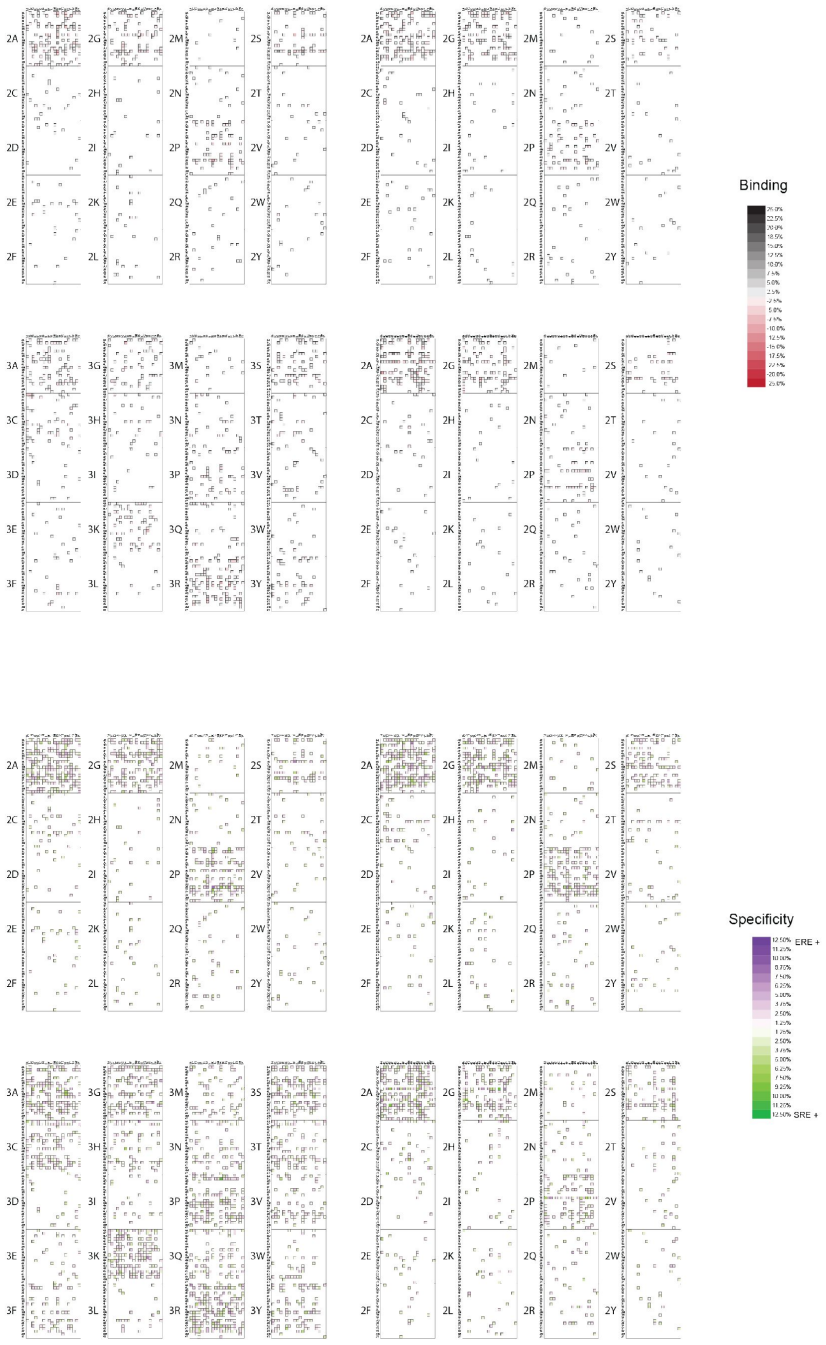

Figure 2 - Figure Supplement 5.

Third-order epistatic effects for binding (top) and specificity (bottom). Effects that increase binding (black), decrease binding (red), prefer ERE (purple) and prefer SRE (green) are shown, with intensity corresponding to strength of effect. All effects are scaled relative to the value necessary to be classified as a strong binder. Scale is the same as **Figure 2-Figure Supplement 4** for comparison. The majority of third order terms are estimated as near zero and not shown.

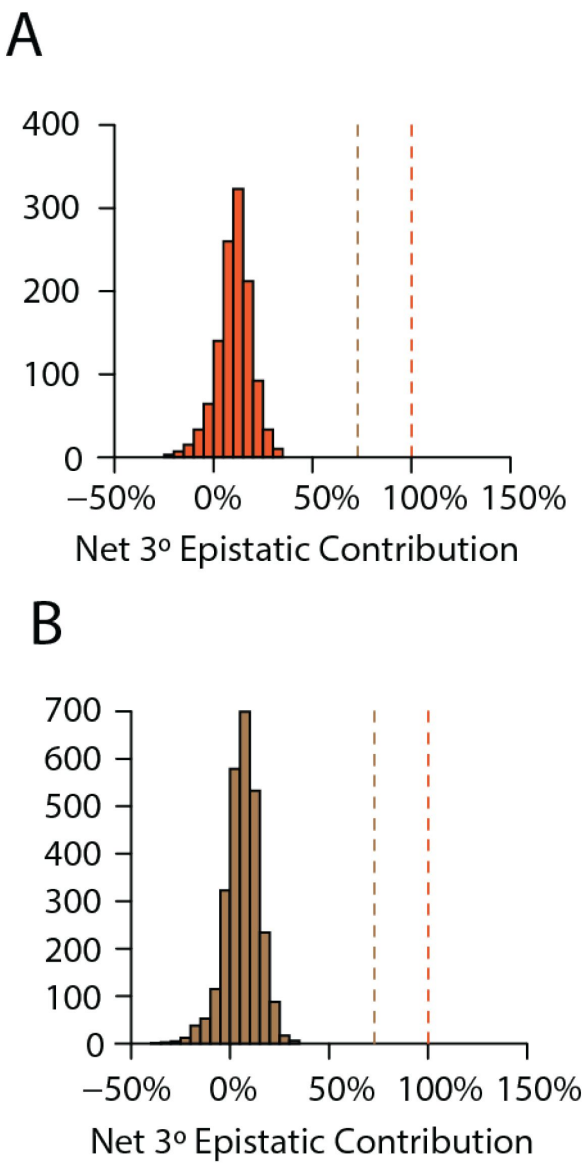

Figure 3 - Figure Supplement 1.

A. Net contribution of third order epistatic effects to the total genetic score of all predicted strong binders. Dashed lines are the thresholds where net third order epistatic effects are sufficient for a genotype to be classified as a weak (brown) or strong (orange) binder. **B**. Same as A, but for predicted weak binders.

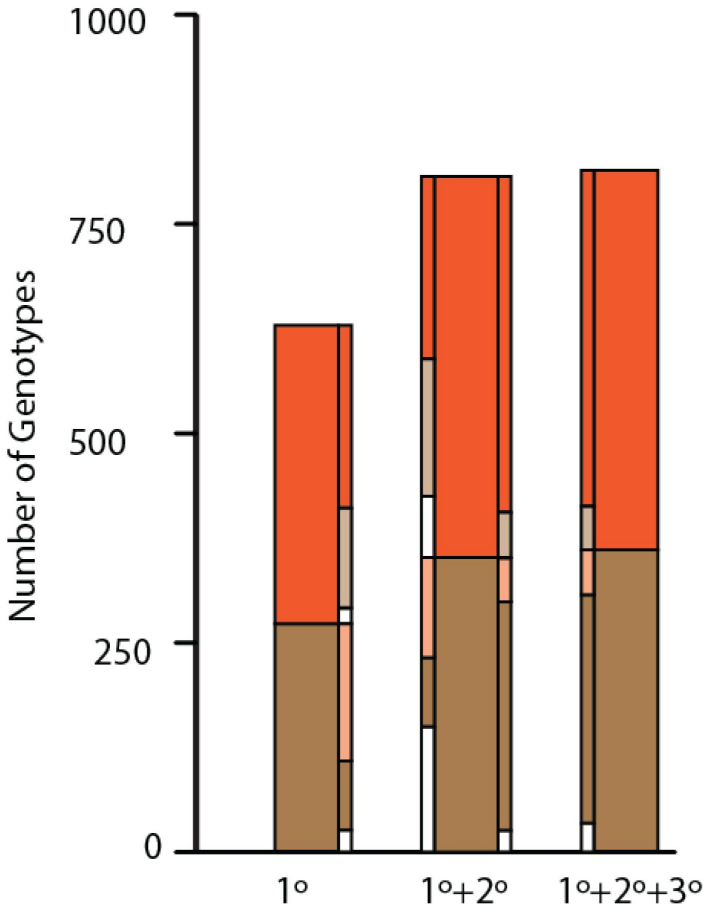

Figure 3 - Figure Supplement 2.

Variants classified as non (white), weak (brown), or strong (orange) binder in a model without epistasis (1°, left), the model with pairwise epistasis (1°+2°, middle), or the full model with epistasis (1°+2°+3° effects, right). Solid colors in skinny bars are genotypes that retain the same class membership as epistatic effects are added (to the right of a bar) or removed (to the left of a bar), while pale colors indicate genotypes that change class membership and what class they are changed to.

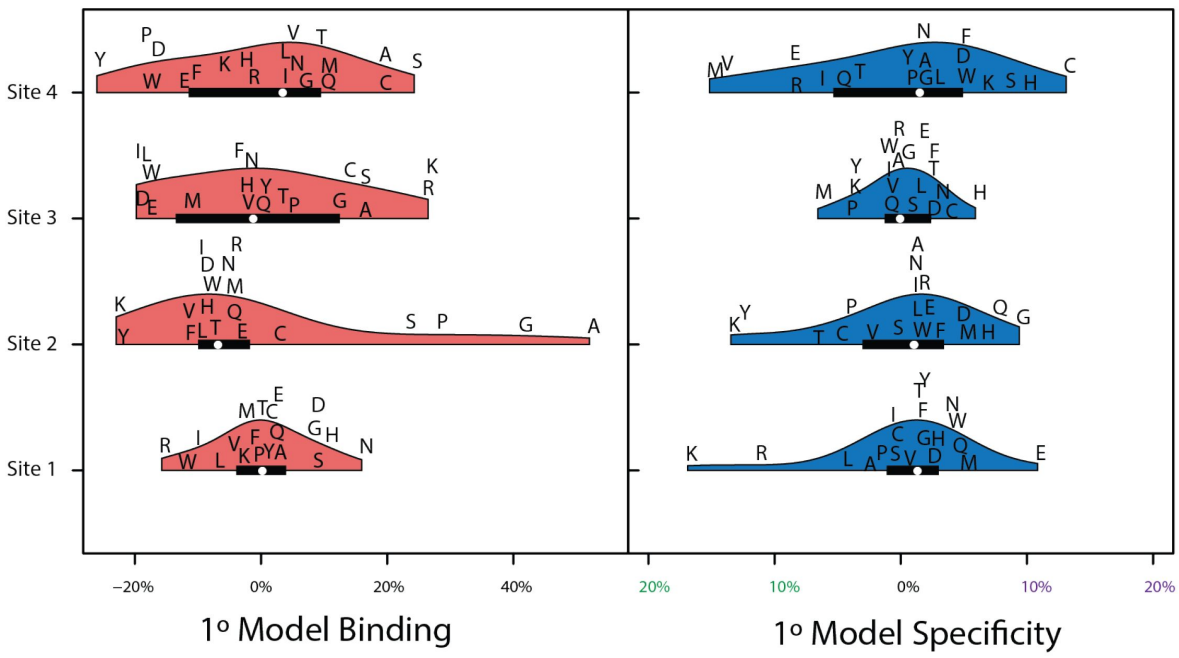

Figure 3 - Figure Supplement 3.

Distribution of effects for each site for the main effect only model (1º). Effects are scaled to the value needed to be a strong binder. Binding effects (red) and specificity effects (blue) are separated.

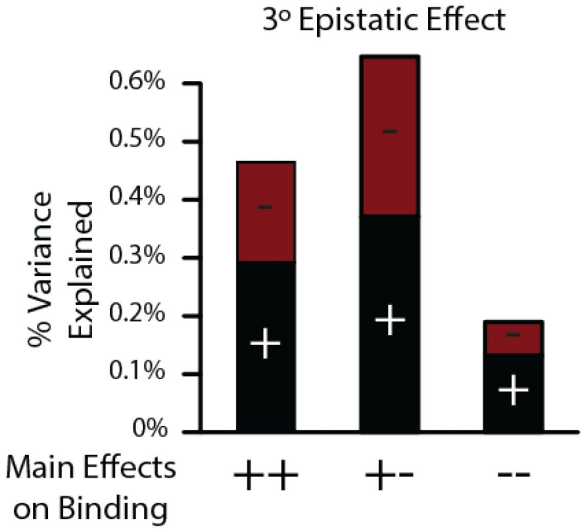

Figure 3 - Figure Supplement 4.

Percent of specific genetic variance explained by positive (black) and negative (red) third-order binding interactions when main effects are positive (++), both are negative (–), or the main effects are of opposite signs (+-).

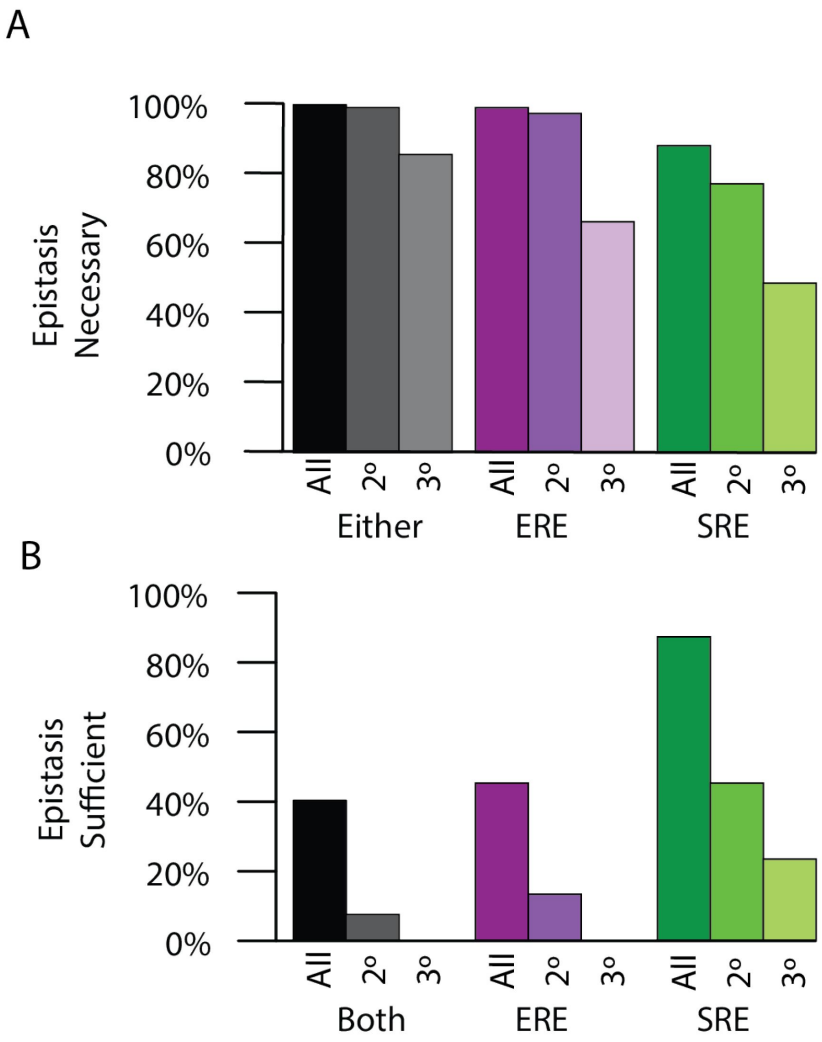

Figure 4 - Figure Supplement 1.

A. Percent of time epistasis is necessary for transition in specificity for the loss of ERE binding activity (purple), the gain of SRE binding activity (green), or either change for either all epistasis (dark), only pairwise epistasis (middle), or only third-order epistasis (light). **B**. Percent of time epistasis is sufficient for the loss of ERE binding activity (purple) and the gain of SRE binding activity (green), or both (black) for either all epistasis (dark), only pairwise epistasis (medium), or only third-order epistasis (light).

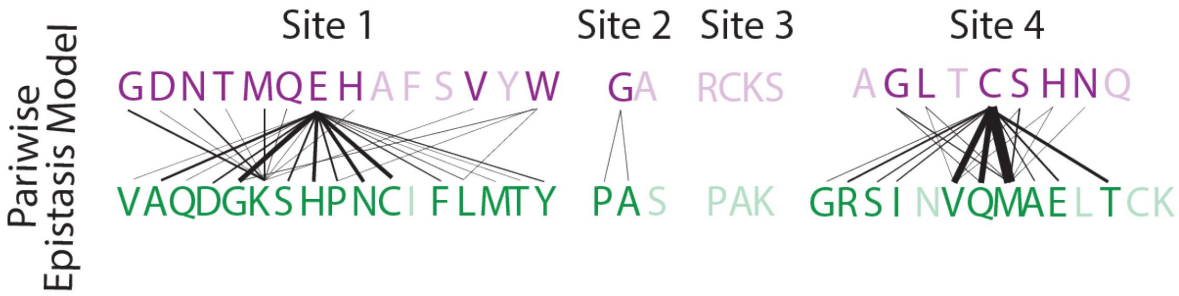

Figure 4 - Figure Supplement 2.

Single step mutations which change DNA specificity in the presence of pairwise epistasis. The ERE-specific amino acid state (purple) is on top, with the SRE-specific amino acid state (green) on the bottom. Line thickness is proportional to the frequency that a mutation causes the change in specificity.

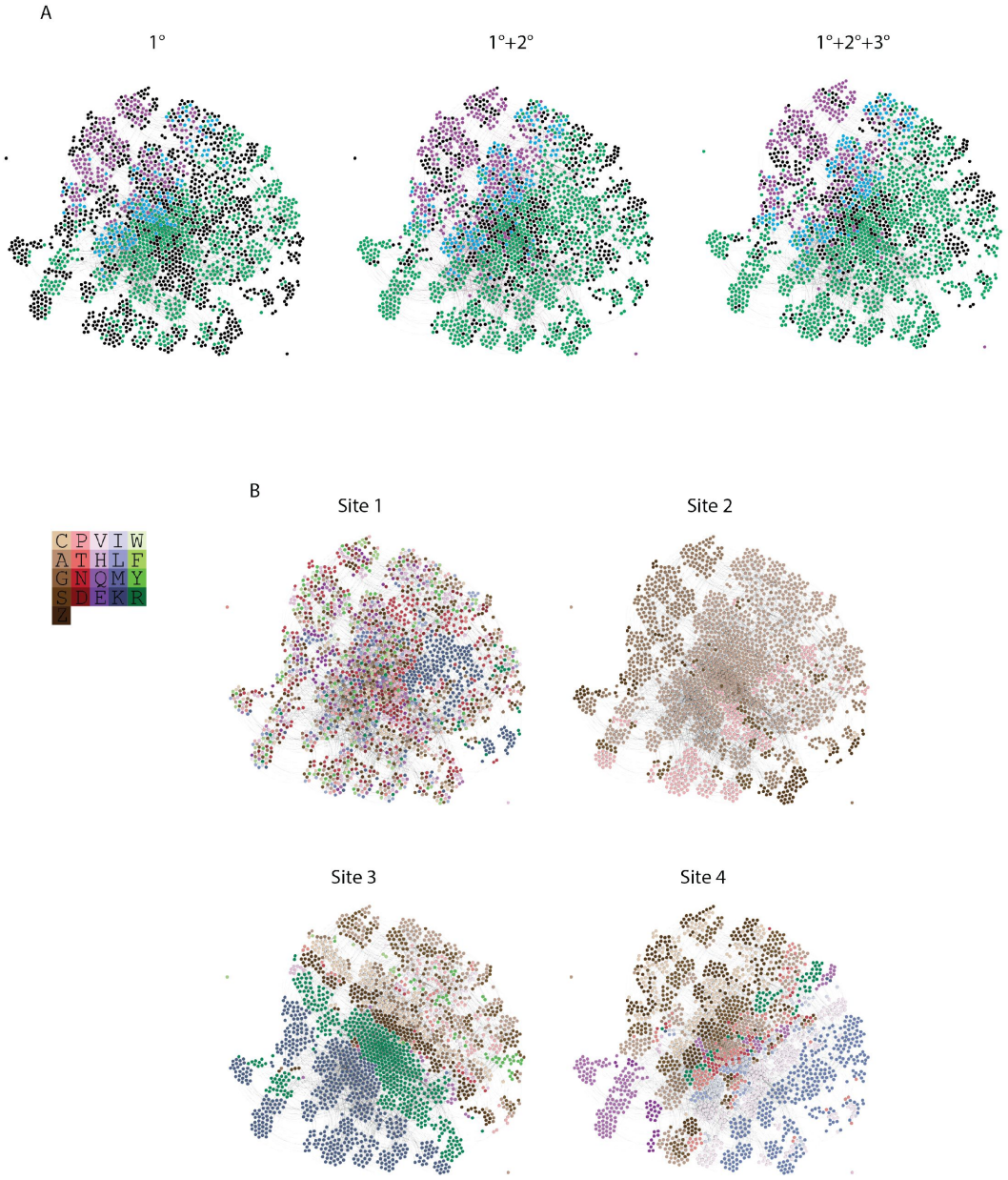

Figure 5 - Figure Supplement 1.

A. Force directed graph of genotypes that are activators in either the absence of epistasis (left), the presence of pairwise epistasis, or both pairwise and third-order epistasis (right). In each case, each node is a genotype, colored by if it is ERE-specific (purple), SRE-specific (green), promiscuous (blue), or a non-binder in the network but an activator in one of the other networks (black). All activator nodes are connected in a given network based on single amino acid changes via the genetic code. **B**. Amino acid state at each node in the force directed graph for each site. Amino acids are colored to reveal clustering in the force directed graph and are labeled by single letter codes. S and Z are used to represent the two non-connected regions of the genetic code containing serine codons.

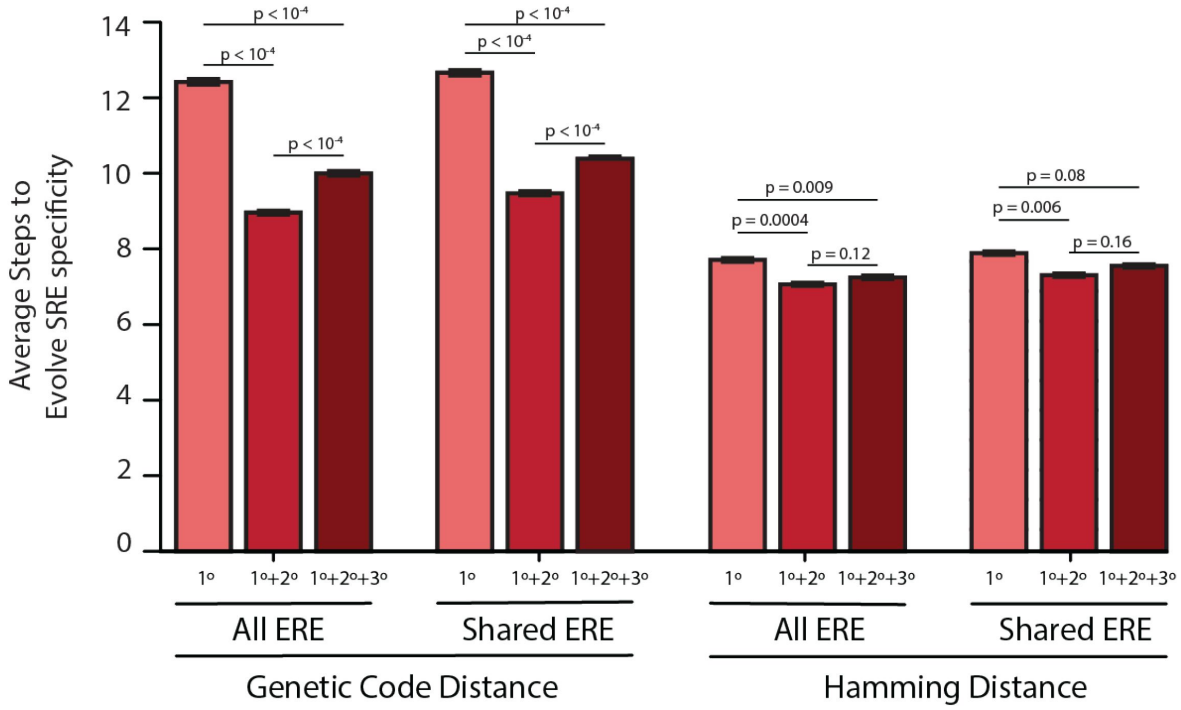

Figure 5 - Figure Supplement 2.

Average number of steps needed to reach an SRE-specific genotype from an ERE-specific genotype in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark) in a model with no positive selection for function. Error bars give a 95% confidence interval on the estimate based on 100 replicates from each ERE-specific genotype. Values are shown for either all ERE-specific starting genotypes or only ERE-specific genotypes found in all three networks when genotypes are connected via either the genetic code or by hamming distance. All p-values were calculated from a permutation test.

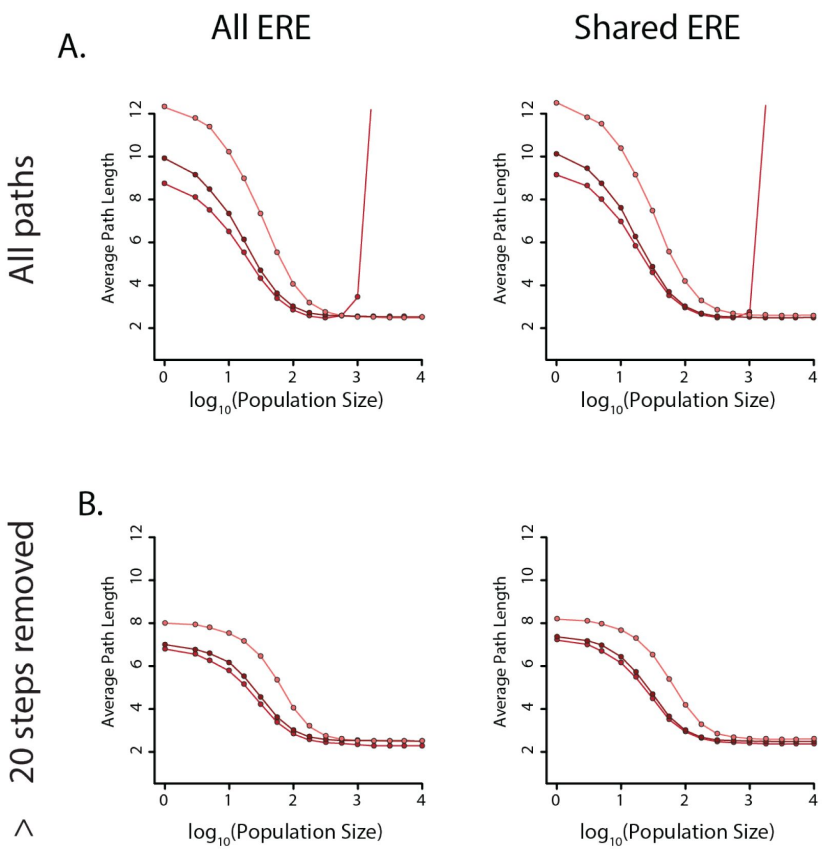

Figure 5 - Figure Supplement 3.

A. Average number of steps needed to reach an SRE-specific genotype from an ERE-specific genotype in the absence of epistasis (light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (dark) in a model with positive selection for increased SRE specificity for all ERE-specific starting genotypes (left) or shared ERE-specific genotypes found in all three networks (right). The population size was varied to control the relative strength of positive selection. **B**. Same as A, but after removing paths that are greater than 20 steps long.

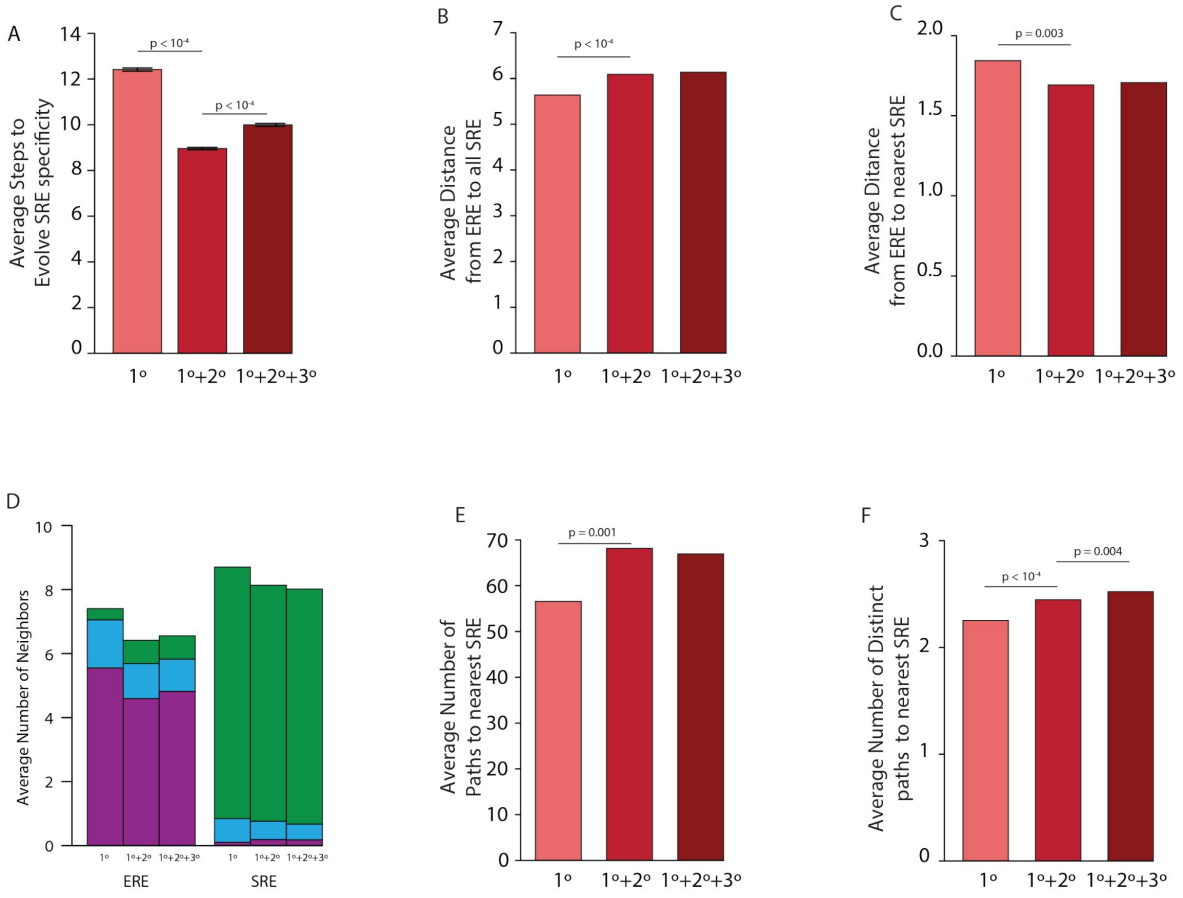

Figure 5 - Figure Supplement 4.

A. Average number of steps needed to reach an SRE-specific genotype from an ERE-specific genotype in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark) in a model with no positive selection for function. Error bars give a 95% confidence interval on the estimate based on 100 replicates from each ERE-specific genotype. **B**. Average distance between all ERE-specific genotypes and all SRE-specific genotypes in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark). **C**. Average distance between all ERE-specific genotypes and the closest SRE-specific genotype in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark). **D**. Average number of single step neighbors for ERE-specific (left) and SRE-specific (right) genotypes in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark). In each case, the average number of neighbors that are ERE-specific (purple), promiscuous (blue), or SRE-specific (green) are shown. **E**. Average number of shortest paths between all ERE-specific and nearest SRE-specific genotype in the absence of epistasis (left, left), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark). **F**. Average number of distinct shortest paths between all ERE-specific and nearest SRE-specific genotype in the absence of epistasis (left, light), the presence of pairwise epistasis (middle), or both pairwise and third-order epistasis (right, dark). All p-values were calculated from a permutation test.

Where g is a particular genotype on a particular DNA element, and y(g) the genetic score (phenotype) of that genotype, and G the set of all possible protein sequences/RE element complexes. The variance is calculated over all genotypes for both ERE and SRE. The phenotype of a particular genotype y(g) can be expressed as the sum of the model coefficients for which that genotype has a particular amino acid state or combination of states. We retain a form of y(g) with both binding and specificity coefficients, although this distinction is not necessary for the proof. The binding coefficients (β s) are the average effect on the genetic score of an amino acid or combination of amino acids, regardless of which DNA element is being considered, while specificity coefficients (σ s) are one half of the difference in the genetic score between binding to ERE and SRE. In the convention we use, binding to ERE adds the specificity effects, while binding to SRE subtracts the specificity terms. Thus, for a particular genotype g with sequence $q_1q_2q_3q_4$ we have:

$$y(g|ERE) = \beta_0 + \sum_i \beta_{q_i} + \sum_{i<j} \beta_{q_iq_j} + \sum_{i<j<k} \beta_{q_iq_jq_k} + \sigma_0 + \sum_i \sigma_{q_i} + \sum_{i<j} \sigma_{q_iq_j} + \sum_{i<j<k} \sigma_{q_iq_jq_k} \quad \textbf{(2)}$$

$$y(g|SRE) = \beta_0 + \sum_i \beta_{q_i} + \sum_{i<j} \beta_{q_iq_j} + \sum_{i<j<k} \beta_{q_iq_jq_k} - \sigma_0 - \sum_i \sigma_{q_i} - \sum_{i<j} \sigma_{q_iq_j} - \sum_{i<j<k} \sigma_{q_iq_jq_k} \quad \textbf{(3)}$$

Here g^E and g^S are genotypes bound to ERE or SRE respectively, i,j, and k index the available sites, and β_{qi}/σ_{qi} are the corresponding model terms based on the state of genotype g at site i.

The global binding intercept β_0 is defined as the average phenotype of all genotypes on the two DNA elements:

$$\beta_0 = \frac{1}{2 \cdot 20^4} \sum_{g \in G} y(g) = \overline{y(G)} \quad \textbf{(4)}$$

Similarly, the global specificity term σ_0 reflects the difference in the average phenotype of genotypes on one element relative to the global average, or one half of the difference in average phenotypic binding between the two DNA elements:

$$\sigma_0 = \frac{1}{20^4} \sum_{g \in G} y(g^E) - \beta_0 = \frac{1}{2 \cdot 20^4} \sum_{g \in G} y(g^E) - y(g^S) = \frac{1}{2}(\overline{y(G^E)} - \overline{y(G^S)}) \quad \textbf{(5)}$$

The remaining model terms capture deviations from the expected phenotype in the absence of progressively higher order interactions:

$$\beta_{q_i} = \overline{y(G_{q_i})} - \beta_0 \quad \textbf{(6)}$$

$$\beta_{q_iq_j} = \overline{y(G_{q_iq_j})} - (\beta_0 + \beta_{q_i} + \beta_{q_j}) \quad \textbf{(7)}$$

$$\beta_{q_iq_jq_k} = \overline{y(G_{q_iq_jq_k})} - (\beta_0 + \beta_{q_i} + \beta_{q_j} + \beta_{q_k} + \beta_{q_iq_j} + \beta_{q_iq_k} + \beta_{q_jq_k}) \quad \textbf{(8)}$$

$$\sigma_{q_i} = \frac{1}{2}\left(\overline{y(G^E_{q_i})} - \overline{y(G^S_{q_i}t)}\right) - \sigma_0 \quad \textbf{(9)}$$

$$\sigma_{q_iq_j} = \frac{1}{2}\left(\overline{y(G^E_{q_iq_j})} - \overline{y(G^S_{q_iq_j})}\right) - (\sigma_0 + \sigma_{q_i} + \sigma_{q_j}) \quad \textbf{(10)}$$

$$\sigma_{q_iq_jq_k} = \frac{1}{2}\left(\overline{y(G^E_{q_iq_jq_k})} - \overline{y(G^S_{q_iq_jq_k})}\right) - (\sigma_0 + \sigma_{q_i} + \sigma_{q_j} + \sigma_{q_k} + \sigma_{q_iq_j} + \sigma_{q_iq_k} + \sigma_{q_jq_k}) \quad \textbf{(11)}$$

Where G_{q_i} is the set of genotypes with state q at site i. Because these terms capture deviations from an average, the set of terms at a site or combination of sites sum to zero by definition:

$$\sum_{q}\beta_{q_1}=\sum_{q}\beta_{q_2}=\sum_{q}\beta_{q_3}=\sum_{q}\beta_{q_4}=0 \tag{12}$$

$$\sum_{q^2}\beta_{q_1q_2}=\sum_{q^2}\beta_{q_1q_3}=\sum_{q^2}\beta_{q_1q_4}=\sum_{q^2}\beta_{q_2q_3}=\sum_{q^2}\beta_{q_2q_4}=\sum_{q^2}\beta_{q_3q_4}=0 \tag{13}$$

$$\sum_{q^3}\beta_{q_1q_2q_3}=\sum_{q^3}\beta_{q_1q_2q_4}=\sum_{q^3}\beta_{q_1q_3q_4}=\sum_{q^3}\beta_{q_2q_3q_4}=0 \tag{14}$$

$$\sum_{q}\sigma_{q_1}=\sum_{q}\sigma_{q_2}=\sum_{q}\sigma_{q_3}=\sum_{q}\sigma_{q_4}=0 \tag{15}$$

$$\sum_{q^3}\sigma_{q_1q_2q_3}=\sum_{q^3}\sigma_{q_1q_2q_4}=\sum_{q^3}\sigma_{q_1q_3q_4}=\sum_{q^3}\sigma_{q_2q_3q_4}=0 \tag{17}$$

(18)

Where q^2 and q^3 are the set of all possible pairs or triplets of amino acid states respectively. Using the definition of β_0 and inserting equations (2) and (3) into equation (1), we can rewrite the specific genetic variance as:

$$\begin{aligned}Var(y(G))=&\frac{1}{2\cdot20^4}\sum_{g\in G^E}\left(\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}+\sigma_0+\sum_i\sigma_{q_i}+\sum_{i<j}\sigma_{q_iq_j}+\sum_{i<j<k}\sigma_{q_iq_jq_k}\right)^2\\&+\frac{1}{2\cdot20^4}\sum_{g\in G^S}\left(\sum_i\beta_{q_i}+\sum_{i<j}\beta_{q_iq_j}+\sum_{i<j<k}\beta_{q_iq_jq_k}-\sigma_0-\sum_i\sigma_{q_i}-\sum_{i<j}\sigma_{q_iq_j}-\sum_{i<j<k}\sigma_{q_iq_jq_k}\right)^2\end{aligned} \tag{19}$$

As expected, the specific genetic variance is independent of the global binding coefficient β_0. However, it is not independent of the global specificity coefficient σ_0. We can rewrite (20) using a more compact notation as:

$$Var\left(y(G)\right)=\frac{1}{2\cdot20^4}\sum_{g\,\in\,G^E}\left(\sum_a\beta_a+\sigma_0+\sum_a\sigma_a\right)^2+\frac{1}{2\cdot20^4}\sum_{g\,\in\,G^S}\left(\sum_a\beta_a-\sigma_0-\sum_a\sigma_a\right)^2 \tag{20}$$

Where a is the set of sites or combinations of sites for all orders other than the intercepts. Thus, a includes all of the q_i first order terms, the q_iq_j second order terms, and the $q_iq_jq_k$ third order terms. We next expand the squared term, rewriting the sum as the sum of each squared coefficient plus the sum of the product of all pairwise combinations of coefficients:

$$\begin{aligned}Var\left(y(G)\right)=&\frac{1}{2\cdot20^4}\sum_{g\,\in\,G^E}\left(\sum_a(\beta_a^2+\sigma_a^2)+\sigma_0^2+2\sum_{a<b}(\beta_a\beta_b+\sigma_a\sigma_b+\beta_a\sigma_b)+2\sum_a(\sigma_0\beta_a+\sigma_0\sigma_a)\right)\\&+\frac{1}{2\cdot20^4}\sum_{g\,\in\,G^S}\left(\sum_a(\beta_a^2+\sigma_a^2)+\sigma_0^2+2\sum_{a<b}(\beta_a\beta_b+\sigma_a\sigma_b-\beta_a\sigma_b)-2\sum_a(\sigma_0\beta_a+\sigma_0\sigma_a)\right)\end{aligned} \tag{21}$$

Where b also indexes the sites/combinations of sites among the different epistatic orders. Distributing the outer sum over these terms gives:

$$Var(y(G)) = \frac{1}{2\cdot 20^4}\left(\sum_a \sum_{g\in G^E}(\beta_a^2+\sigma_a^2) + \sum_{g\in G^E}\sigma_0^2 + 2\sum_{a<b}\sum_{g\in G^E}(\beta_a\beta_b+\sigma_a\sigma_b+\beta_a\sigma_b) + 2\sum_a\sum_{g\in G^E}(\sigma_0\beta_a+\sigma_0\sigma_a)\right)$$
$$+\frac{1}{2\cdot 20^4}\left(\sum_a \sum_{g\in G^S}(\beta_a^2+\sigma_a^2) + \sum_{g\in G^S}\sigma_0^2 + 2\sum_{a<b}\sum_{g\in G^S}(\beta_a\beta_b+\sigma_a\sigma_b-\beta_a\sigma_b) - 2\sum_a\sum_{g\in G^S}(\sigma_0\beta_a+\sigma_0\sigma_a)\right) \quad (22)$$

To simplify this expression, we note that all of the cross-product terms in (23) turn out to be zero. To demonstrate this, we note that the set of all genotypes can be represented by conditioning on a particular amino acid state at a particular site and then taking the union over all possible amino acid states for that site:

$$G = \bigcup_{q_i} G_{q_i} \quad (23)$$

We can thus reorganize a sum of cross product terms over all genotypes in (10) as a sum of cross product terms conditioned on a particular set of amino acid states and then sum over all possible amino acid states. For example, the sum of the crossproducts of the binding coefficients with $a = q_1$ and $b = q_2$ can be rewritten by conditioning on the amino acid states at sites 1 and 2, and then summing over all possible combinations of amino acids for those sites:

$$\sum_{g\in G^E}\beta_{q_1}\beta_{q_2} = \sum_{q_1}\sum_{q_2}\sum_{g\in G^E_{q_1q_2}}\beta_{q_1}\beta_{q_2} \quad (24)$$

Where $G^E_{q_1q_2}$ is the set of genotypes with state q_1 at site 1 and state q_2 at site 2 bound to ERE. From here, we note that once conditioned on a particular set of amino acid states at a set of sites, that combination of amino acid states will appear on all possible combinations of amino acids at the remaining sites. We can thus replace a sum over all genotypes with number of such combinations and then factor the result:

$$\sum_{q_1}\sum_{q_2}\sum_{g\in G^E_{q_1q_2}}\beta_{q_1}\beta_{q_2} = \sum_{q_1}\sum_{q_2}20^2\cdot\beta_{q_1}\beta_{q_2} = 20^2\cdot\sum_{q_1}\beta_{q_1}\sum_{q_2}\beta_{q_2} \quad (25)$$

Since each of these latter sums is over the entire set of coefficients at a site, then by construction they sum to zero, and their product is thus also zero:

$$20^2\cdot\sum_{q_1}\beta_{q_1}\sum_{q_2}\beta_{q_2} = 20^2\cdot\sum_{q_1}\beta_{q_1}\cdot 0 = 0 \quad (26)$$

A similar logic holds for higher order terms, for specificity terms, and for the combination of binding and specificity terms; as long as the conditional sums are over the entire set of amino acid combinations for those sites then we can replace summing over all genotypes with the number of genetic backgrounds a particular combination of states will appear on. For example, for the sum of cross products between binding and specificity terms with $a = q_1q_2$ and $b = q_1q_2q_3$

$$\sum_{g\in G^E}\beta_{q_1q_2}\sigma_{q_1q_2q_3} = \sum_{q_1}\sum_{q_2}\sum_{q_3}\sum_{g\in G^E_{q_1q_2q_3}}\beta_{q_1q_2}\sigma_{q_1q_2q_3} = \sum_{q_1}\sum_{q_2}\sum_{q_3}20\cdot\beta_{q_1q_2}\sigma_{q_1q_2q_3}$$
$$= 20\cdot\sum_{q_1}\sum_{q_2}\beta_{q_1q_2}\sum_{q_3}\sigma_{q_1q_2q_3} = 20\cdot\sum_{q_1}\sum_{q_2}\beta_{q_1q_2}\cdot 0 = 0 \quad (27)$$

Similar logic can be applied to all terms in (23) that are the sum of the product of coefficients at other orders: the terms for any pair or triplet of sites always sum to zero, so the sum of the product of those terms (times any other set of terms) also sums to zero.

Combining equation (23) with this result, we can rewrite the specific genetic variance as:

$$Var\left(y\left(G\right)\right)=\frac{1}{2\cdot 20^4}\left(\sum_{g\in G^E}\sigma_0^2+\sum_a\sum_{g\in G^E}\left(\beta_a^2+\sigma_a^2\right)\right)+\frac{1}{2\cdot 20^4}\left(\sum_{g\in G^S}\sigma_0^2+\sum_a\sum_{g\in G^S}\left(\beta_a^2+\sigma_a^2\right)\right)=\frac{1}{20^4}\left(\sum_{g\in G}\sigma_0^2+\sum_a\sum_{g\in G}\beta_a^2+\sigma_a^2\right) \tag{28}$$

Where the second equality follows because the sets of squared coefficients are the same for binding ERE and SRE. To simplify further, we again use the fact that the set of all genotypes can be partitioned into smaller sets based on the amino acid state at a site. For example, for the squared binding coefficients at site 1:

$$\frac{1}{20^4}\sum_{g\in G}\beta_{q_1}^2=\frac{1}{20^4}\sum_{q_1}\sum_{g\in G_{q_1}}\beta_{q_1}^2=\frac{1}{20^4}\sum_{q_1}20^3\cdot\beta_{q_1}^2=\frac{1}{20}\sum_{q_1}\beta_{q_1}^2 \tag{29}$$

Similarly, when coefficients involve multiple sites:

$$\frac{1}{20^4}\sum_{g\in G}\beta_{q_1q_2}^2=\frac{1}{20^4}\sum_{q_1}\sum_{q_2}\sum_{g\in G_{q_1q_2}}\beta_{q_1q_2}^2=\frac{1}{20^4}\sum_{q_1}\sum_{q_2}20^2\cdot\beta_{q_1q_2}^2=\frac{1}{20^2}\sum_{q\epsilon Q}\beta_{q_1q_2}^2 \tag{30}$$

And the global specificity term:

$$\frac{1}{20^4}\sum_{g\in G}\sigma_0^2=\frac{1}{20^4}\cdot 20^4\cdot\sigma_0^2=\sigma_0^2 \tag{31}$$

Extending this logic to all the squared terms and combining the results of (30), (31), and (32) with equation (29), we get:

$$Var\left(y\left(G\right)\right)=\frac{1}{20}\sum_i\beta_{q_i}^2+\frac{1}{20^2}\sum_{i<j}\beta_{q_iq_j}^2+\frac{1}{20^3}\sum_{i<j<k}\beta_{q_iq_jq_k}^2+\sigma_0^2+\frac{1}{20}\sum_i\sigma_{q_i}^2+\frac{1}{20^2}\sum_{i<j}\sigma_{q_iq_j}^2+\frac{1}{20^3}\sum_{i<j<k}\sigma_{q_iq_jq_k}^2 \tag{32}$$

Which is the sum of the squared coefficients, each normalized by the number of terms at a particular epistatic order. We can rewrite this more compactly as:

$$Var\left(y\left(G\right)\right)=\sum_{\beta\neq\beta_0}\frac{\beta^2}{20^{O(\beta)}}+\sum_\sigma\frac{\sigma^2}{20^{O(\sigma)}} \tag{33}$$

where β and σ represent the individual binding and specificity terms, and O is the epistatic order of each term, with O = 0 for global (intercept) terms, 1 for main effects, 2 for pairwise interactions, and 3 for third order interactions. Said another way, the total variance is the sum of each term's squared effect, weighted by the number of genotypes that the term affects.

One consequence of the relationship between the effect size of a term, its order of epistasis, and the total specific genetic variance explained, is that the fraction of specific genetic variance explained by any single term is readily calculable. For example, the fraction of specific genetic variance explained by each first order binding term at a particular site is:

$$F(Var\left(\beta_{q_i}\right)) = \frac{\frac{1}{20}\beta_{q_i}^2}{\sum_{\beta \neq \beta_0} \frac{\beta^2}{20^{O(\beta)}} + \sum_{\sigma} \frac{\sigma^2}{20^{O(\sigma)}}} = \frac{\frac{1}{20}\beta_{q_i}^2}{Var\left(y\left(G\right)\right)} \tag{34}$$

In addition, since the total specific genetic variance is the simple weighted sum of squared terms, the fraction of specific genetic variance explained by any set of terms is also easy to calculate. For example, the fraction of specific genetic variance explained by all first order binding terms at a particular site is:

$$F(Var\left(\sum_i \beta_{q_i}\right)) = \frac{\frac{1}{20}\sum_i \beta_{q_i}^2}{\sum_{\beta \neq \beta_0} \frac{\beta^2}{20^{O(\beta)}} + \sum_{\sigma} \frac{\sigma^2}{20^{O(\sigma)}}} = \frac{\frac{1}{20}\sum_i \beta_{q_i}^2}{Var\left(y\left(G\right)\right)} \tag{35}$$

Thus, knowing only the value of a model coefficient and the epistatic order it pertains to, we can readily calculate the fraction of the total specific genetic variance contributed by an individual term and thus any set of terms in the model, including across sites, epistatic orders, mechanisms of action, and the particular form of epistasis.

References

1. Breen Michael S., Kemena Carsten, Vlasov Peter K., Notredame Cedric, Kondrashov Fyodor A. (2012) **Epistasis as the primary factor in molecular evolution** *Nature* **490**:535–538 https://doi.org/10.1038/nature11510

2. Carneiro Maurício, Hartl Daniel L. (2010) **Adaptive landscapes and protein evolution** *Proceedings of the National Academy of Sciences* **107**:1747–1751 https://doi.org/10.1073/pnas.0906192106

3. Chen Jong Chen, Conrad Michael (1994) **A multilevel neuromolecular architecture that uses the extradimensional bypass principle to facilitate evolutionary learning** *Physica D: Nonlinear Phenomena* **75**:417–437 https://doi.org/10.1016/0167-2789(94)90295-X

4. Conrad Michael (1998) **Towards High Evolvability Dynamics Introduction** *Evolutionary Systems* :33–43

5. Domingo Júlia, Baeza-Centurion Pablo, Lehner Ben (2019) **The Causes and Consequences of Genetic Interactions (Epistasis)** *Annual Review of Genomics and Human Genetics* **20**:433–460 https://doi.org/10.1146/annurev-genom-083118-014857

6. Franke Jasper, Klözer Alexander, de Visser J. Arjan G. M., Krug Joachim (2011) **Evolutionary Accessibility of Mutational Pathways** *PLoS Computational Biology* **7** https://doi.org/10.1371/journal.pcbi.1002134

7. Gavrilets Sergey (1997) **Evolution and speciation on holey adaptive landscapes** *Trends in ecology & evolution* **12**:307–312 https://doi.org/10.1016/S0169-5347(97)01098-7

8. Gavrilets Sergey, Gravner Janko (1997) **Percolation on the fitness hypercube and the evolution of reproductive isolation** *Journal of Theoretical Biology* **184**:51–64 https://doi.org/10.1006/jtbi.1996.0242

9. Goldstein Richard (2016) **Sequence Entropy and the Absolute Rate of Amino Acid Substitutions** *biorxiv*

10. Kauffman Stuart, Levin Simon (1987) **Towards a General Theory of Adaptive Walks on Rugged Landscapes** *Journal of Theo* **128**:11–45

11. Kauffman S a, Weinberger E D (1989) **The NK model of rugged fitness landscapes and its application to maturation of the immune response** *Journal of Theoretical Biology* **141**:211–245

12. Kondrashov Dmitry A., Kondrashov Fyodor A. (2015) **Topological features of rugged fitness landscapes in sequence space** *Trends in Genetics* **31**:24–33 https://doi.org/10.1016/j.tig.2014.09.009

13. McCandlish David M., Rajon Etienne, Shah Premal, Ding Yang, Plotkin Joshua B. (2013) **The role of epistasis in protein evolution** *Nature* **497**:7–9 https://doi.org/10.1038/nature12219

14. Miton Charlotte M., Buda Karol, Tokuriki Nobuhiko (2021) **Epistasis and intramolecular networks in protein evolution** *Current Opinion in Structural Biology* **69**:160–168 https://doi.org/10.1016/j.sbi.2021.04.007

15. Payne Joshua L., Wagner Andreas (2018) **The causes of evolvability and their evolution** *Nature Reviews Genetics* **20**:24–38 https://doi.org/10.1038/s41576-018-0069-z

16. Pollock David D, Thiltgen Grant, Goldstein Richard A (2012) **Amino acid coevolution induces an evolutionary Stokes shift** *Proceedings of the National Academy of Sciences* **109**:E1352–E1359 https://doi.org/10.1073/pnas.1120084109

17. Romero Philip A, Arnold Frances H (2009) **Exploring protein fitness landscapes by directed evolution** *Nature reviews. Molecular cell biology* **10**:866–876 https://doi.org/10.1038/nrm2805

18. Shah Premal, Plotkin Joshua B, McCandlish David M. (2015) **Contingency and entrenchment in protein evolution under purifying selection** *Proceedings of the National Academy of Sciences* **112**:E3226–E3235 https://doi.org/10.1073/pnas.1412933112

19. Smith J M (1970) **Natural selection and the concept of a protein space** *Nature* **225**:563–564 https://doi.org/10.1038/225563a0

20. Starr Tyler N, Thornton Joseph W. (2016) **Epistasis in protein evolution** *Protein Science* **25**:1204–1218 https://doi.org/10.1002/pro.2897

21. Usmanova Dinara R, Ferretti Luca, Povolotskaya Inna S, Vlasov Peter K, Kondrashov Fyodor A. (2014) **A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution** *Molecular biology and evolution* **32**:542–554 https://doi.org/10.1093/molbev/msu318

22. Whitlock M C, Phillips P C, Moore F B, Tonsor S J (1995) **Multiple Fitness Peaks and Epistasis** *Annual Review of Ecology and Systematics* **26**:601–629 https://doi.org/10.1146/annurev.es.26.110195.003125

23. Chou H.-H. Hsin-Hung, Chiu H.-C. Hsuan-Chao, Delaney Nigel F., Segre D., Marx Christopher J., Segrè Daniel (2011) **Diminishing returns epistasis among beneficial mutations decelerates adaptation** *Science* **332**:1190–2 https://doi.org/10.1126/science.1203799

24. Cvijović Ivana, Ba Alex N. Nguyen, Desai Michael M. (2018) **Experimental Studies of Evolutionary Dynamics in Microbes** *Trends in Genetics* **34**:693–703 https://doi.org/10.1016/j.tig.2018.06.004

25. Johnson Milo S., Martsul Alena, Kryazhimskiy Sergey, Desai Michael M. (2019) **Higherfitness yeast genotypes are less robust to deleterious mutations** *Science* **366**:490–493 https://doi.org/10.1126/science.aay4199

26. Kryazhimskiy Sergey, Rice Daniel P., Jerison Elizabeth R., Desai Michael M. (2014) **Global epistasis makes adaptation predictable despite sequence-level stochasticity** *Science* **344**:1519–1522 https://doi.org/10.1126/science.1250939

27. Lyons Daniel M, Zou Zhengting, Xu Haiqing, Zhang Jianzhi (2020) **Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories** *Nature Ecology and Evolution* https://doi.org/10.1038/s41559-020-01286-y

28. Reddy Gautam, Desai Michael M. (2021) **Global epistasis emerges from a generic model of a complex trait** *eLife* **10** https://doi.org/10.7554/ELIFE.64740

29. Tokuriki Nobuhiko, Jackson Colin J., Afriat-Jurnou Livnat, Wyganowski Kirsten T., Tang Renmei, Tawfik Dan S. (2012) **Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme** *Nature Communications* **3** https://doi.org/10.1038/ncomms2246

30. Wei Xinzhu, Zhang Jianzhi (2019) **Patterns and Mechanisms of Diminishing Returns from Beneficial Mutations** *Molecular Biology and Evolution* **36**:1008–1021 https://doi.org/10.1093/molbev/msz035

31. Wünsche Andrea, Dinh Duy M., Satterwhite Rebecca S., Arenas Carolina Diaz, Stoebel Daniel M., Cooper Tim F. (2017) **Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory** *Nature Ecology and Evolution* **1** https://doi.org/10.1038/s41559-016-0061

32. Ashenberg Orr, Gong Lizhi Ian, Bloom Jesse D (2013) **Mutational effects on stability are largely conserved during protein evolution** *Proceedings of the National Academy of Sciences* **110**:21071–21076 https://doi.org/10.1073/pnas.1314781111

33. Bank Claudia, Hietpas Ryan T., Jensen Jeffrey D., Bolon Daniel N. A. (2014) **A Systematic Survey of an Intragenic Epistatic Landscape** *Molecular Biology and Evolution* **32**:229–238 https://doi.org/10.1093/molbev/msu301

34. Bridgham Jamie T., Carroll Sean M., Thornton Joseph W. (2006) **Evolution of hormonereceptor complexity by molecular exploitation** *Science* **312**:97–101 https://doi.org/10.1126/science.1123348

35. Bridgham Jamie T, Ortlund Eric a, Thornton Joseph W. (2009) **An epistatic ratchet constrains the direction of glucocorticoid receptor evolution** *Nature* **461**:515–519 https://doi.org/10.1038/nature08249

36. Ding David, Green Anna G., Wang Boyuan, Lite Thuy Lan Vo, Weinstein Eli N., Marks Debora S., Laub Michael T. (2022) **Co-evolution of interacting proteins through non-contacting and non-specific mutations** *Nature Ecology and Evolution* **6**:590–603 https://doi.org/10.1038/s41559-022-01688-0

37. Emlaw Johnathon R., Burkett Kelly M., Dacosta Corrie J.B. (2020) **Contingency between Historical Substitutions in the Acetylcholine Receptor Pore** *ACS chemical neuroscience* **11**:2861–2868 https://doi.org/10.1021/ACSCHEMNEURO.0C00410

38. Faber Matthew S., Wrenbeck Emily E., Azouz Laura R., Steiner Paul J., Whitehead Timothy A. (2019) **Impact of in Vivo Protein Folding Probability on Local Fitness Landscapes** *Molecular Biology and Evolution* **36**:2764–2777 https://doi.org/10.1093/molbev/msz184

39. Gong Lizhi Ian, Suchard Marc A., Bloom Jesse D (2013) **Stability-mediated epistasis constrains the evolution of an influenza protein** *eLife* **2** https://doi.org/10.7554/eLife.00631

40. Gong Lizhi Ian, Bloom Jesse D (2014) **Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution** *PLoS genetics* **10** https://doi.org/10.1371/journal.pgen.1004328

41. Kumar Amit, Natarajan Chandrasekhar, Moriyama Hideaki, Witt Christopher C., Weber Roy E., Fago Angela, Storz Jay F. (2017) **Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin** *Molecular Biology and Evolution* **34**:1240–1251 https://doi.org/10.1093/molbev/msx085

42. Lunzer Mark, Golding G. Brian, Dean Antony M. (2010) **Pervasive cryptic epistasis in molecular evolution** *PLoS Genetics* **6** https://doi.org/10.1371/journal.pgen.1001162

43. Ortlund E A, Bridgham J T, Redinbo M R, Thornton Joseph W. (2007) **Crystal structure of an ancient protein: evolution by conformational epistasis** *Science* **317**:1544–1548 https://doi.org/10.1126/science.1142819

44. Park Yeonwoo, Metzger Brian P H, Thornton Joseph W (2022) **Epistatic drift causes gradual decay of predictability in protein evolution** *Science* **376**:823–830

45. Pokusaeva Victoria O. *et al.* (2019) **An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape** *PLoS Genetics* **15** https://doi.org/10.1371/journal.pgen.1008079

46. Starr Tyler N, Flynn Julia M, Mishra Parul, Bolon Daniel N. A., Thornton Joseph W. (2018) **Pervasive contingency and entrenchment in a billion years of Hsp90 evolution** *Proceedings of the National Academy of Sciences* **115**:4453–4458 https://doi.org/10.1073/pnas.1718133115

47. Starr Tyler N. *et al.* (2022) **Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution** *Science* **377**:420–424

48. Tufts Danielle M, Natarajan Chandrasekhar, Revsbech Inge G, Projecto-Garcia Joana, Hoffmann Federico G, Weber Roy E, Fago Angela, Moriyama Hideaki, Storz Jay F (2015) **Epistasis Constrains Mutational Pathways of Hemoglobin Adaptation in High-Altitude Pikas** *Molecular biology and evolution* **32**:287–298 https://doi.org/10.1093/molbev/msu311

49. Wang Lin, Zheng Wei, Zhao Hongyu, Deng Minghua (2013) **Statistical analysis reveals coexpression patterns of many pairs of genes in yeast are jointly regulated by interacting loci** *PLoS genetics* **9** https://doi.org/10.1371/journal.pgen.1003414

50. Adams Rhys M., Kinney Justin B., Walczak Aleksandra M., Mora Thierry (2019) **Epistasis in a fitness landscape defined by antibody-antigen binding free energy** *Cell systems* **8** https://doi.org/10.1016/J.CELS.2018.12.004

51. Diss Guillaume, Lehner Ben (2018) **The genetic landscape of a physical interaction** *eLife* **7**:1–31 https://doi.org/10.7554/eLife.32472

52. Olson C. Anders, Wu Nicholas C., Sun Ren (2014) **A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain** *Current Biology* **24**:2643–2651 https://doi.org/10.1016/j.cub.2014.09.072

53. Salinas Victor H., Ranganathan Rama (2018) **Coevolution-based inference of amino acid interactions underlying protein function** *eLife* **7** https://doi.org/10.7554/eLife.34300

54. Buda Karol, Miton Charlotte M., Tokuriki Nobuhiko (2022) **Higher-order epistasis creates idiosyncrasy, confounding predictions in protein evolution** *bioRxiv* https://doi.org/10.1101/2022.09.07.505194

55. Arjan J, Visser G M De, Krug Joachim (2014) **Empirical fitness landscapes and the predictability of evolution** *Nature Reviews Genetics* **15**:480–490 https://doi.org/10.1038/nrg3744

56. Jochumsen Nicholas, Marvig Rasmus L., Damkiær Søren, Jensen Rune Lyngklip, Paulander Wilhelm, Molin Søren, Jelsbak Lars, Folkesson Anders (2016) **The evolution of antimicrobial peptide resistance in Pseudomonas aeruginosa is shaped by strong epistatic interactions** *Nature Communications* **7**:1–10 https://doi.org/10.1038/ncomms13002

57. Phillips Angela M., Lawrence Katherine R., Moulana Alief, Dupic Thomas, Chang Jeffrey, Johnson Milo S., Cvijovic Ivana, Mora Thierry, Walczak Aleksandra M., Desai Michael M. (2021) **Binding affinity landscapes constrain the evolution of broadly neutralizing antiinfluenza antibodies** *eLife* **10** https://doi.org/10.7554/ELIFE.71393

58. Sailer Zachary R., Shafik Sarah H., Summers Robert L., Joule Alex, Patterson-Robert Alice, Martin Rowena E., Harms Michael J (2020) **Inferring a complete genotypephenotype map from a small number of measured phenotypes** *PLoS Computational Biology* **16** https://doi.org/10.1371/journal.pcbi.1008243

59. Sailer Zachary R., Harms Michael J (2017) **Detecting high-order epistasis in nonlinear genotype-phenotype maps** *Genetics* **205**:1079–1088 https://doi.org/10.1534/genetics.116.195214

60. Szendro Ivan G, Schenk Martijn F, Franke Jasper, Krug Joachim, de Visser J Arjan G M (2013) **Quantitative analyses of empirical fitness landscapes** *Journal of Statistical Mechanics: Theory and Experiment* **2013** https://doi.org/10.1088/1742-5468/2013/01/P01005

61. Weinreich DM, Lan Yinghong, Jaffe Jacob, Heckendorn Robert B. (2018) **The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography** *Journal of Statistical Physics* **172**:208–225 https://doi.org/10.1007/s10955-018-1975-3

62. Weinreich Daniel M., Knies Jennifer L. (2013) **FISHER'S GEOMETRIC MODEL OF ADAP-TATION MEETS THE FUNCTIONAL SYNTHESIS: DATA ON PAIRWISE EPISTASIS FOR FITNESS YIELDS INSIGHTS INTO THE SHAPE AND SIZE OF PHENOTYPE SPACE** *Evolution* **67**:2957–2972 https://doi.org/10.1111/evo.12156

63. Aakre Christopher D. *et al.* (2015) **Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates** *Cell* **163**:1–13 https://doi.org/10.1016/j.cell.2015.09.055

64. Anderson Dave W., Baier Florian, Yang Gloria, Tokuriki Nobuhiko (2021) **The adaptive landscape of a metallo-enzyme is shaped by environment-dependent epistasis** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-23943-x

65. Anderson Dave W, Mckeown Alesia N, Thornton Joseph W. (2015) **Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites** *eLife* **4** https://doi.org/10.7554/eLife.07864

66. De Visser J. Arjan G.M., Park Su Chan, Krug Joachim (2009) **Exploring the effect of sex on empirical fitness landscapes** *American Naturalist* **174**:S15–S30 https://doi.org/10.1086/599081

67. Dutta Sanjib, Gullá Stefano, Chen T. Scott, Fire Emiko, Grant Robert A., Keating Amy E. (2010) **Determinants of BH3 Binding Specificity for Mcl-1 versus Bcl-xL** *Journal of Molecular Biology* **398**:747–762 https://doi.org/10.1016/j.jmb.2010.03.058

68. Field Yair, Kaplan Noam, Fondufe-Mittendorf Yvonne, Moore Irene K, Sharon Eilon, Widom Jonathan, Segal Eran, Lubling Yaniv (2008) **Distinct modes of regulation by chromatin encoded through nucleosome positioning signals** *PLoS computational biology* **4** https://doi.org/10.1371/journal.pcbi.1000216

69. Jenson Justin M., Ryan Jeremy A., Grant Robert A., Letai Anthony, Keating Amy E. (2017) **Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1** *eLife* **6**:1–23 https://doi.org/10.7554/eLife.25541

70. Jiang W, Bikard D, Cox D, Zhang F, Marraffini L A (2013) **RNA-guided editing of bacterial genomes using CRISPR-Cas systems** *Nat Biotechnol* **31**:233–239 https://doi.org/10.1038/nbt.2508

71. Khan Aisha I., Dinh Duy M., Schneider Dominique, Lenski Richard E., Cooper Tim F. (2011) **Negative epistasis between beneficial mutations in an evolving bacterial population** *Science* **332**:1193–6 https://doi.org/10.1126/science.1203801

72. Lee Youn Hyung, DSouza Lisa M., Fox George E. (1997) **Equally parsimonious pathways through an RNA sequence space are not equally likely** *Journal of Molecular Evolution* **45**:278–284 https://doi.org/10.1007/PL00006231

73. Lozovsky Elena R., Chookajorn Thanat, Brown Kyle M., Imwong Mallika, Shaw Philip J., Kamchonwongpaisan Sumalee, Neafsey Daniel E., Weinreich Daniel M., Hartl Daniel L. (2009) **Stepwise acquisition of pyrimethamine resistance in the malaria parasite** *Proceedings of the National Academy of Sciences of the United States of America* **106**:12025–12030 https://doi.org/10.1073/pnas.0905922106

74. Lunzer Mark, Miller Stephen P., Felsheim Roderick, Dean Antony M. (2005) **The biochemical architecture of an ancient adaptive landscape** *Science* **310**:499–501 https://doi.org/10.1126/science.1115649

75. Malcolm Bruce A., Wilson Keith P., Matthews Brian W., Kirsch Jack F., Wilson Allan C. (1990) **Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing** *Nature* **345**:86–89 https://doi.org/10.1038/345086a0

76. Meini M.-R., Tomatis P. E., Weinreich DM, Vila a. J. (2015) **Quantitative Description of a Protein Fitness Landscape based on Molecular Features** *Molecular Biology and Evolution* **32**:1774–1787 https://doi.org/10.1093/molbev/msv059

77. Miton Charlotte M., Tokuriki Nobuhiko (2016) **How mutational epistasis impairs predictability in protein evolution and design** *Protein Science* **0**:1–13 https://doi.org/10.1002/pro.2876

78. Moriuchi Ryota, Tanaka Hiroki, Nikawadori Yuki, Ishitsuka Mayuko, Ito Michihiro, Ohtsubo Yoshiyuki, Tsuda Masataka, Damborsky Jiri, Prokop Zbynek, Nagata Yuji (2014) **Stepwise enhancement of catalytic performance of haloalkane dehalogenase LinB towards β-hexachlorocyclohexane** *AMB Express* **4** https://doi.org/10.1186/S13568-014-0072-5

79. Natarajan C., Inoguchi N., Weber R. E., Fago A., Moriyama Hideaki, Storz Jay F (2013) **Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin** *Science* **340**:1324–1327 https://doi.org/10.1126/science.1236862

80. Noor Sajid, Taylor Matthew C., Russell Robyn J., Jermiin Lars S., Jackson Colin J., Oakeshott John G., Scott Colin (2012) **Intramolecular epistasis and the evolution of a new enzymatic function** *PLoS ONE* **7** https://doi.org/10.1371/journal.pone.0039822

81. O'Maille Paul E, Malone Arthur, Nikki Dellas B, Hess Andes, Smentek Lidia, Sheehan Iseult, Greenhagen Bryan T, Chappell Joe, Manning Gerard, Noel Joseph P (2008) **Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases** *Nature chemical biology* **4**:617–23 https://doi.org/10.1038/nchembio.113

82. Palmer Adam C., Toprak Erdal, Baym Michael, Kim Seungsoo, Veres Adrian, Bershtein Shimon, Kishony Roy (2015) **Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes** *Nature Communications* **6** https://doi.org/10.1038/ncomms8385

83. Poelwijk Frank J., Socolich Michael, Ranganathan Rama (2019) **Learning the pattern of epistasis linking genotype and phenotype in a protein** *Nature Communications* **10** https://doi.org/10.1038/s41467-019-12130-8

84. Poelwijk Frank J., Krishna Vinod, Ranganathan Rama (2016) **The Context-Dependence of Mutations: A Linkage of Formalisms** *PLOS Computational Biology* **12** https://doi.org/10.1371/journal.pcbi.1004771

85. Reetz Manfred T., Sanchis Joaquin (2008) **Constructing and analyzing the fitness landscape of an experimental evolutionary process** *ChemBioChem* **9**:2260–2267 https://doi.org/10.1002/cbic.200800371

86. Weinreich DM, Delaney Nigel F, Depristo Mark a, Hartl Daniel L (2006) **Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins** *Science* **312** https://doi.org/10.1126/science.1123539

87. Yang Gloria, Anderson Dave W, Baier Florian, Dohmen Elias, Hong Nansook, Carr Paul D, Kamerlin Shina Caroline Lynn, Jackson Colin J, Bornberg-Bauer Erich, Tokuriki Nobuhiko (2019) **Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme** *Nature Chemical Biology* **15**:1120–1128 https://doi.org/10.1038/s41589-019-0386-3

88. Zhang Xu-Sheng (2012) **Fisher's geometrical model of fitness landscape and variance in fitness within a changing environment** *Evolution* **66** https://doi.org/10.1111/j.1558-5646.2012.01610.x

89. Moulana Alief, Dupic Thomas, Phillips Angela M., Chang Jeffrey, Roffler Anne A., Greaney Allison J., Starr Tyler N., Bloom Jesse D., Desai Michael M. (2023) **The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution** *eLife* **12** https://doi.org/10.7554/ELIFE.83442

90. Phillips Angela M, Maurer Daniel P, Brooks Caelan, Dupic Thomas, Schmidt Aaron G, Desai Michael M (2023) **Hierarchical sequence-affinity landscapes shape the evolution of breadth in an anti-influenza receptor binding site antibody** *eLife* **12** https://doi.org/10.7554/ELIFE.83628

91. Jalal Adam S.B., Tran Ngat T., Stevenson Clare E., Chan Elliot W., Lo Rebecca, Tan Xiao, Noy Agnes, Lawson David M., Le Tung B.K. (2020) **Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family** *Cell Reports* **32** https://doi.org/10.1016/j.celrep.2020.107928

92. Lite Thúy Lan V., Grant Robert A., Nocedal Isabel, Littlehale Megan L., Guo Monica S., Laub Michael T (2020) **Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library** *eLife* **9** https://doi.org/10.7554/eLife.60924

93. McClune Conor J., Alvarez-Buylla Aurora, Voigt Christopher A., Laub Michael T. (2019) **Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space** *Nature* **574**:702–706 https://doi.org/10.1038/s41586-019-1639-8

94. Podgornaia Anna I., Laub Michael T (2015) **Pervasive degeneracy and epistasis in a proteinprotein interface** *Science* **347**:673–677 https://doi.org/10.1126/science.1257360

95. Ramachandran Srinivas, Henikoff Steven (2016) **Transcriptional Regulators Compete with Nucleosomes Post-replication** *Cell* **165**:580–592 https://doi.org/10.1016/j.cell.2016.02.062

96. Starr Tyler N, Thornton Joseph W. (2017) **Exploring protein sequence–function landscapes** *Nature Biotechnology* **35**:125–126 https://doi.org/10.1038/nbt.3786

97. Bin Wu Carolina Eliscovich, Yoon Young J, Singer Robert H (2016) **Translation dynamics of single mRNAs in live cells and neurons** *Science, page aaf1084* https://doi.org/10.1126/science.aaf1084

98. Yoo Justin I., Daugherty Patrick S., O'Malley Michelle A. (2020) **Bridging non-overlapping reads illuminates high-order epistasis between distal protein sites in a GPCR** *Nature Communications* **11** https://doi.org/10.1038/s41467-020-14495-7

99. Brookes David H., Aghazadeh Amirali, Listgarten Jennifer (2022) **On the sparsity of fitness functions and implications for learning** *Proceedings of the National Academy of Sciences of the United States of America* **119**

100. Faure Andre J., Lehner Ben, Pina Verónica Miró, Colome Claudia Serrano, Weghorn Do-nate (2023) **An extension of the Walsh-Hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity** *bioRxiv* https://doi.org/10.1101/2023.03.06.531391

101. Faure Andre J., Martí-Aranda Aina, Hidalgo-Carcedo Cristina, Schmiedel Jörn M., Lehner Ben (2023) **The genetic architecture of protein stability** *bioRxiv* https://doi.org/10.1101/2023.10.27.564339

102. Park Yeonwoo, Metzger Brian P.H., Thornton Joseph W. (2023) **The simplicity of protein sequence-function relationships** *bioRxiv* https://doi.org/10.1101/2023.09.02.556057

103. McKeown Alesia N., Bridgham Jamie T., Anderson Dave W., Murphy Michael N., Ortlund Eric A., Thornton Joseph W. (2014) **Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module** *Cell* **159**:58–68 https://doi.org/10.1016/j.cell.2014.09.003

104. Chusacultanachai Sudsanguan, Glenn Kevin A., Rodriguez Adrian O., Read Erik K., Gardner Jeffrey F., Katzenellenbogen Benita S., Shapiro David J. (1999) **Analysis of estrogen response element binding by genetically selected steroid receptor DNA binding domain mutants exhibiting altered specificity and enhanced affinity** *Journal of Biological Chemistry* **274**:23591–23598 https://doi.org/10.1074/jbc.274.33.23591

105. So Alex Yick Lun, Chaivorapol Christina, Bolton Eric C., Li Hao, Ya-mamoto Keith R. (2007) **Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor** *PLoS Genetics* **3**:927–938 https://doi.org/10.1371/journal.pgen.0030094

106. Welboren Willem Jan, Sweep Fred C G J, Span Paul N., Stunnenberg Hendrik G. (2009) **Genomic actions of estrogen receptor alpha: What are the targets and how are they regulated?** *Endocrine-Related Cancer* **16**:1073–1089 https://doi.org/10.1677/ERC-09-0086

107. Stormo Gary D. (2011) **Maximally efficient modeling of DNA sequence motifs at all levels of complexity** *Genetics* **187**:1219–1224 https://doi.org/10.1534/genetics.110.126052

108. DePristo Mark A., Weinreich DM, Hartl Daniel L. (2005) **Missense meanderings in sequence space: A biophysical view of protein evolution** *Nature Reviews Genetics* **6**:678–687 https://doi.org/10.1038/nrg1672

109. Harms Michael J, Eick Geeta N, Goswami Devrishi, Colucci Jennifer K, Griffin Patrick R, Ortlund Eric A, Thornton Joseph W. (2013) **Biophysical mechanisms for largeeffect mutations in the evolution of steroid hormone receptors** *Proceedings of the National Academy of Sciences* **110**:11475–11480 https://doi.org/10.1073/pnas.1303930110

110. Otwinowski Jakub, McCandlish David M., Plotkin Joshua B (2018) **Inferring the shape of global epistasis** *Proceedings of the National Academy of Sciences* **115**:E7550–E7558 https://doi.org/10.1073/pnas.1804015115

111. Phillips Patrick C (2008) **Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems** *Nature reviews Genetics* **9**:855–67 https://doi.org/10.1038/nrg2452

112. Sailer Zachary R., Harms Michael J (2017) **Molecular ensembles make evolution unpredictable** *Proceedings of the National Academy of Sciences* :1–34 https://doi.org/10.1073/pnas.1711927114

113. Smith John Maynard (1969) **Natural selection and The Concept of a Protein Space** *Theory and Practice in Language Studies* **2**:1885–1889 https://doi.org/10.4304/tpls.2.9.1885-1889

114. Sailer Zachary R., Harms Michael J (2017) **High-order epistasis shapes evolutionary trajectories** *PLoS Computational Biology* **13** https://doi.org/10.1371/journal.pcbi.1005541

115. Poelwijk Frank J., Kiviet Daniel J., Weinreich Daniel M., Tans Sander J. (2007) **Empirical fitness landscapes reveal accessible evolutionary paths** *Nature* **445**:383–386 https://doi.org/10.1038/nature05451

116. Weinreich DM, Watson Richard A R.A., Chao Lin (2005) **Perspective: Sign epistasis and genetic constraint on evolutionary trajectories** *Evolution* **59**:1165–1174

117. Bendixsen Devin P., Collet James, Østman Bjørn, Hayden Eric J. (2019) **Genotype network intersections promote evolutionary innovation** *PLOS Biology* **17** https://doi.org/10.1371/JOURNAL.PBIO.3000300

118. Payne Joshua L, Wagner Andreas (2014) **Latent phenotypes pervade gene regulatory circuits** *BMC systems biology* **8** https://doi.org/10.1186/1752-0509-8-64

119. Araya Carlos L, Fowler Douglas M, Chen Wentao, Muniez Ike, Kelly Jeffery W, Fields Stanley (2012) **A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function** *Proceedings of the National Academy of Sciences* **109**:16858–16863 https://doi.org/10.1073/pnas.1209751109

120. Bloom Jesse D (2014) **An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit** *Molecular biology and evolution* **31**:1956–1978 https://doi.org/10.1093/molbev/msu173

121. Chen Zhoutao *et al.* (2020) **Ultralow input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely** *Genome Research* https://doi.org/10.1101/gr.260380.119

122. Debartolo Joe, Dutta Sanjib, Reich Lothar, Keating Amy E. (2012) **Predictive Bcl-2 family binding models rooted in experiment or structure** *Journal of Molecular Biology* **422**:124–144 https://doi.org/10.1016/j.jmb.2012.05.022

123. Firnberg Elad, Labonte Jason W., Gray Jeffrey J., Ostermeier Marc (2014) **A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape** *Molecular biology and evolution* **31**:1581–1592 https://doi.org/10.1093/molbev/msu081

124. Fowler Douglas M, Araya Carlos L, Fleishman Sarel J, Kellogg Elizabeth H, Stephany Jason J, Baker David, Fields Stanley (2010) **High-resolution mapping of protein sequencefunction relationships** *Nature methods* **7**:741–746 https://doi.org/10.1038/nmeth.1492

125. Fragata Inês, Matuszweski Sebastian, Schmitz Mark A., Bataillon Thomas, Jensen Jeffrey D., Bank Claudia (2018) **The fitness landscape of the codon space across environments** *bioRxiv* https://doi.org/10.1101/252395

126. Hietpas Ryan T., Jensen Jeffrey D, Bolon Daniel N. A. (2011) **Experimental illumination of a fitness landscape** *Proceedings of the National Academy of Sciences* **108**:7896–901 https://doi.org/10.1073/pnas.1016024108

127. Jr Richard N. McLaughlin, Poelwijk Frank J., Raman Arjun, Gosal Walraj S., Ranganathan Rama (2012) **The spatial architecture of protein function and adaptation** *Nature* **490**:138–142 https://doi.org/10.1038/nature11500

128. Roscoe Benjamin P., Thayer Kelly M., Zeldovich Konstantin B., Fushman David, Bolon Daniel N. A. (2013) **Analyses of the effects of all ubiquitin point mutants on yeast growth rate** *Journal of Molecular Biology* **425**:1363–1377 https://doi.org/10.1016/j.jmb.2013.01.032

129. Sarkisyan Karen S. *et al.* (2016) **Local fitness landscape of the green fluorescent protein** *Nature* **533**:397–401 https://doi.org/10.1038/nature17995

130. Starita Lea M., Pruneda Jonathan N., Lo Russell S., Fowler Douglas M., Kim Helen J., Hiatt Joseph B., Shendure Jay, Brzovic Peter S., Fields Stanley, Klevit Rachel E. (2013) **Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis** *Proceedings of the National Academy of Sciences of the United States of America* **110**:1263–1272 https://doi.org/10.1073/pnas.1303309110

131. Starr Tyler N *et al.* (2020) **Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding** *Cell* **182**:1295–1310 https://doi.org/10.1016/j.cell.2020.08.012

132. Thyagarajan B., Bloom Jesse D (2014) **The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin** *eLife* **3** https://doi.org/10.7554/eLife.03300

133. Whitehead Timothy A *et al.* (2012) **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing** *Nature biotechnology* **30**:543–8 https://doi.org/10.1038/nbt.2214

134. Bakerlee Christopher W., Ba Alex N. Nguyen, Shulgina Yekaterina, Echenique Jose I. Rojas, Desai Michael M. (2022) **Idiosyncratic epistasis leads to global fitness–correlated trends** *Science* **376**:630–635 https://doi.org/10.1126/SCIENCE.ABM4774

135. R Core Team (2023) **R: A language and environment for statistical computing**

136. Archer K. (2010) **glmnetcr: An R Package for Ordinal Response Prediction in High-dimensional Data Settings**

137. Bates D, Maechler M (2022) **MatrixModels: Modeling with Sparse and Dense Matrices**

138. Gerber F., Mösinger K., Furrer R. (2018) **dotCall64: An R package providing an efficient interface to compiled C, C++, and Fortran code supporting long vectors** *SoftwareX* **7**:217–221 https://doi.org/10.1016/J.SOFTX.2018.06.002

139. Anderson Jill T, Lee Cheng-Ruei, Mitchell-Olds Thomas (2014) **Strong selection genomewide enhances fitness trade-offs across environments and episodes of selection** *Evolution* **68**:16–31 https://doi.org/10.1111/evo.12259

140. Csárdi Gábor, Nepusz T. (2006) **The igraph software package for complex network research** *InterJournal*

141. Kimura Motoo (1962) **ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN A POPULATION** *Genetics*

Article and author information

Brian P.H. Metzger
Department of Ecology and Evolution, University of Chicago, 60637 USA, Department of Biological Sciences, Purdue University
ORCID iD: 0000-0002-4670-7509

Yeonwoo Park
Program in Genetics, Genomics, and Systems Biology, University of Chicago, 60637 USA, Center for RNA Research, Seoul National University
ORCID iD: 0000-0002-3665-8002

Tyler N. Starr
Department of Biochemistry and Molecular Biophysics, University of Chicago, 60637 USA, Department of Biochemistry, University of Utah
ORCID iD: 0000-0001-6713-6904

Joseph W. Thornton
Department of Ecology and Evolution, University of Chicago, 60637 USA, Program in Genetics, Genomics, and Systems Biology, University of Chicago, 60637 USA, Department of Human Genetics, University of Chicago, 60637 USA
For correspondence: joet1@uchicago.edu

Copyright

Editors

Reviewing Editor
Armita Nourmohammad
University of Washington, Seattle, United States of America

Senior Editor
Christian Landry
Université Laval, Québec, Canada

Reviewer #1 (Public Review):

Metzger et al develop a rigorous method filling an important unmet need in protein evolution - analysis of protein genetic architecture and evolution using data from combinatorially complete 20^N variant libraries. Addressing this need has become increasingly valuable, as experimental methods for generating these datasets expand in scope and scale. Their method integrates two key features - (1) it reports the effects of mutations relative to the average across all variants, rather than a particular genotype, making it useful for examining global genetic architecture, and (2) it does this for all possible 20 states at each site, in contrast to the binary analyses in prior work. These features are not individually novel but integrating them into a single analysis framework is novel and will be valuable to the protein evolution community. Using a previously published dataset generated by two of the authors, they conclude that (1) changes in function are largely attributable to pairwise but not higher-order interactions, and (2) epistasis potentiates, rather than constrains, evolutionary paths. These findings are well-supported by the data. Overall, this work has important implications for predicting the relationship between genotype and phenotype, which is of considerable interest to protein biochemistry, evolutionary biology, and numerous other fields.

https://doi.org/10.7554/eLife.88737.2.sa1

Reviewer #2 (Public Review):

The authors aimed to understand how epistasis influences the genetic architecture of the DNA-binding domain (DBD) of steroid hormone receptor. An ordinal regression model was developed in this study to analyze a published deep mutational scanning dataset that consists of all combinatorial amino acid variants across four positions (i.e. 160,000 variants). This published dataset measured the binding of each variant to the estrogen receptor response element (ERE, sequence: AGGTCA) as well as the steroid receptor response element (SRE, sequence: AGAACA). This model has major strengths of being reference free and able to account for global nonlinearity in the genotype-phenotype relationship. Thorough analyses of the modelling results have performed, which provided convincing results to support the importance of epistasis in promoting evolution of protein functions. This conclusion is impactful because many previous studies have shown that epistasis constrains evolution. The novelty this study will likely stimulate new ideas in the field. The model will also likely be utilized by other groups in the community.

https://doi.org/10.7554/eLife.88737.2.sa0

Author Response

The following is the authors' response to the original reviews.

We are grateful for the reviewers' comments. We have modified the manuscript accordingly and detail our responses to their major comments below.

> *(1) Reviewer 2 was concerned that transformation of continuous functional data into categorical form could reduce precision in estimating the genetic architecture.*

We agree that transforming continuous data into categories may reduce resolution, but it also improves accuracy when the continuous data are affected by measurement noise. In our dataset, many genotypes are at the lower bound of measurement, and the variation in measured fluorescence among these genotypes is largely or entirely caused by measurement noise. By transforming to categorical data, we dramatically reduced the effect of this noise on the estimation of genetic effects. We modified the results and discussion sections to address this point.

> *(2) Reviewer 2 asked about generalizability of our findings.*

Because our paper is the first use of reference-free analysis of a 20-state combinatorial dataset, generalizability is at this point unknown. However, a recent manuscript from our group confirms the generality of the simplicity of genetic architecture: using reference-free methods to analyze 20 published combinatorial deep mutational scans, several of which involve 20-state libraries, we found that main and pairwise effects account for virtually all of the genetic variance across a wide variety of protein families and types of biochemical functions (Park Y, Metzger BPH, Thornton JW. 2023. The simplicity of protein sequence-function relationships. BioRxiv, 2023.09.02.556057). Concerning the facilitating effect of epistasis on the evolution of new functions, we speculate that this result is likely to be general: we have no reason to think that the underlying cause of this observation – epistasis brings genotypes with different functions closer in sequence space to each other and expands the total number of functional sequences – arises from some peculiarity of the mechanisms of steroid receptor DBD folding or DNA binding. However, we acknowledge that our data involve sequence variation at those sites in the protein that directly mediate specific protein-DNA contact; it is plausible that sites far from the "active site" may have weaker epistatic interactions and therefore have weaker effects on navigability of the landscape. We have addressed these issues in the discussion.

> *(3) Reviewer 3 asked "in which situation would the authors expect that pairwise epistasis does* not *play a crucial role for mutational steps, trajectories, or space connectedness, if it is dominant in the genotype-phenotype landscape?"*

The question addressed in our paper is not whether epistasis shapes steps, trajectories or connectedness in sequence space but how it does so and what its particular effects are on the evolution of new functions. The dominant view in the field has been that the primary role of epistasis is to block evolutionary paths. We show, however, that in multi-state sequence space, epistasis facilitates rather than impedes the evolution of new functions. It does this by increasing the number of functional genotypes and bringing genotypes with different functions closer together in sequence space. This finding was possible because of the

> ***Response to Reviewers' Public Comments***

We are grateful for the reviewers' comments. We have modified the manuscript accordingly and detail our responses to their major comments below.

> *(1) Reviewer 2 was concerned that transformation of continuous functional data into categorical form could reduce precision in estimating the genetic architecture.*

We agree that transforming continuous data into categories may reduce resolution, but it also improves accuracy when the continuous data are affected by measurement noise. In our dataset, many genotypes are at the lower bound of measurement, and the variation in measured fluorescence among these genotypes is largely or entirely caused by measurement noise. By transforming to categorical data, we dramatically reduced the effect of this noise on the estimation of genetic effects. We modified the results and discussion sections to address this point.

> *(2) Reviewer 2 asked about generalizability of our findings.*

Because our paper is the first use of reference-free analysis of a 20-state combinatorial dataset, generalizability is at this point unknown. However, a recent manuscript from our group confirms the generality of the simplicity of genetic architecture: using reference-free methods to analyze 20 published combinatorial deep mutational scans, several of which involve 20-state libraries, we found that main and pairwise effects account for virtually all of the genetic variance across a wide variety of protein families and types of biochemical functions (Park Y, Metzger BPH, Thornton JW. 2023. The simplicity of protein sequence-function relationships. BioRxiv, 2023.09.02.556057). Concerning the facilitating effect of epistasis on the evolution of new functions, we speculate that this result is likely to be general: we have no reason to think that the underlying cause of this observation – epistasis brings genotypes with different functions closer in sequence space to each other and expands the total number of functional sequences – arises from some peculiarity of the mechanisms of steroid receptor DBD folding or DNA binding. However, we acknowledge that our data involve sequence variation at those sites in the protein that directly mediate specific protein-DNA contact; it is plausible that sites far from the "active site" may have weaker epistatic interactions and therefore have weaker effects on navigability of the landscape. We have addressed these issues in the discussion.

> *(3) Reviewer 3 asked "in which situation would the authors expect that pairwise epistasis does* not *play a crucial role for mutational steps, trajectories, or space connectedness, if it is dominant in the genotype-phenotype landscape?"*

The question addressed in our paper is not whether epistasis shapes steps, trajectories or connectedness in sequence space but how it does so and what its particular effects are on the evolution of new functions. The dominant view in the field has been that the primary role of epistasis is to block evolutionary paths. We show, however, that in multi-state sequence space, epistasis facilitates rather than impedes the evolution of new functions. It does this by increasing the number of functional genotypes and bringing genotypes with different functions closer together in sequence space. This finding was possible because of the difference in approach between our paper and prior work: most prior work considered only direct paths in a binary sequence space between two particular starting points – and typically only considering optimization of a single function – whereas we studied the evolution of new functions in a multi-state amino acid space, under empirically relevant epistasis informed by complete combinatorial experiments. The result is a clear demonstration that the net effect of

real-world levels of epistasis on navigability of the multidimensional sequence landscape is to make the evolution of new functions easier, not harder.

> *(4) Reviewer 3 asked for "an explanation of how much new biological results this paper delivers as compared with the paper in which the data were originally published."*

Starr 2017 did not use their data to characterize the underlying genetic architecture of function by estimating main and epistatic effects of amino acid states and combinations; it also did not evaluate the importance of epistasis in generating functional variants, determining the transcription factor's specificity, or shaping evolutionary navigability on the landscape.

> *(5) Reviewer 3 requested an explanation of how the results would have been (potentially) different if a reference-based approach were used, and how reference-based analysis compares with other reference-free approaches to estimating epistasis.*

This topic has been covered in detail in a recent manuscript from our group (Park et al. Biorxiv 2023.09.02.556057). Briefly, reference-free approaches provide the most efficient explanation of an entire genotype-phenotype map, explaining the maximum amount of genetic variance and reducing sensitivity to experimental noise and missing genotypes compared to reference-based approaches. Reference-based approaches tend to infer much more epistasis, especially higher-order epistasis, because measurement error and local idiosyncrasy near the wild-type sequence propagate into spurious high-order terms. Reference-based analyses are appropriate for characterizing only the immediate sequence neighborhood of a particular "wild-type" protein of interest. Reference-free approaches are therefore best suited to understanding genotype-phenotype landscapes as a whole. We have clarified these issues in the revised discussion.

> *(6) Reviewer 3 suggested that the comparison between the full and main-effects-only model should involve a re-estimation of main effects in the latter case.*

This is indeed what we did in our analysis. We have clarified the description in the results and methods sections to make this clear.

> *(7) Reviewer 3 asked about the applicability of the approach to data beyond those analyzed in the present study and requirements to use it.*

Our approach could be used for any combinatorial DMS dataset in which the phenotypic data are categorical (or can be converted to categorical form). Complete sampling is not required: a virtue of reference-free analysis is that by averaging the estimated effects of states and combinations over all variants that contain them, reference-free analysis is highly robust to missing data (except at the highest possible order of epistasis, where only a single variant represents a high-order effect) as long as variant sampling is unbiased with respect to phenotype. All the required code are publicly available at the github link provided in this manuscript. We have also described a general form of reference-free analysis for continuous data and applied it to 20 protein datasets in a recent publication (Park et al. Biorxiv 2023.09.02.556057).

> *(8)Reviewer 3 suggested that the text could be shortened and made less dense.*

We agree and have done a careful edit to streamline the narrative.

> ***Response to Reviewers' Non-Public Recommendations***
>
> *(1) Reviewer 1 noted that specific epistatic effects might in some cases produce global nonlinearities in the genotype-phenotype relationship. They then asked how our results might change if we did not impose a nonlinear transformation as part of the genotype-phenotype model. The reviewer's underlying concern was that the non-specific transformation might capture high-order specific epistatic effects and thus reducing their importance.*

Because our data are categorical, we required a model that characterizes the effect of particular amino acid states and combinations on the probability that a variant is in a null, weak, or strong activation class. A logistic model is the classic approach to this kind of analysis. The model structure assumes that amino acid states and combinations have additive effects on the log-odds of being in one functional class versus the lower functional class(es); the only nonlinear transformation is that which arises mathematically when log-odds are transformed into probability through the logistic link function. Thinking through the reviewer's comment, we have concluded that our model does not make any explicit transformation to account for nonlinearity in the relationship between the effects of specific sequence states/combinations and the measured phenotype (activation class). If additional global nonlinearities are present in the genotype-phenotype relationship – such as could be imposed by limited dynamic range in the production of the fluorescence phenotype or the assay used to measure it – it is possible that the sigmoid shape of the logistic link function may also accommodate these nonlinearities. We have noted this part in the revised manuscript.

> *(2) Reviewer 1 observed that our model seems to prefer sets of several pairwise interactions among states across sites rather than fewer high-order interactions among those same states.*

This finding arises because the pattern of phenotypic variation across genotypes in our dataset is consistent with that which would be produced by pairwise interactions rather than by high-order interactions. In a reference-free framework, these patterns are distinct from each other: a group of second-order terms cannot fit the patterns produced by high-order epistasis, and high-order terms cannot fit the pattern produced by pairwise interactions. Similarly, main-effect terms cannot fit the pattern of phenotypes produced by a pairwise interaction, and a pairwise epistatic term cannot fit the pattern produced by main effects of states at two sites. For example, third-order terms are required when the genotypes possessing a particular triplet of states deviate from that expected given all the main and second-order effects of those states; this deviation cannot be explained by any combination of first- and second-order effects.

We explain this point in detail in our recent manuscript (Park Y, Metzger BPH, Thornton JW. 2023. The simplicity of protein sequence-function relationships. BioRxiv, 2023.09.02.556057) and we summarize it here. Consider the simple example of two sites with two possible states (genotypes 00, 01, 10, and 11). If there are no main effects and no pairwise effects, this architecture will generate the same phenotype for all four variants – the global average (or zero-order effect). If there are pairwise effects but no main effects, this architecture will generate a set of phenotypes on which the average phenotype of genotypes with a 0 at the first site (00 and 01) equals the global average – as does the average of those with 0 at the second site (00 and 10). The epistatic effect causes the individual genotypes to deviate from the global average. This pattern can be fit only by a pairwise epistatic term, not by first-order terms. Conversely, if there are main effects but no pairwise effects, then the average phenotype of genotypes 00 and 01 will deviate from the global average (by an amount equal to the first-order effect), as will the average of (00 and 10): the phenotype of each genotype

will be equal to the sum of the relevant first-order effects for the state it contains. This pattern cannot be fit by second-order model terms. The same logic extends to higher orders: a cluster of second-order terms cannot explain variation generated by third-order epistasis, because third-order variation is by definition is the deviation from the best second-order model.

> *(3) Reviewer 1 suggested several places in the text where citations to prior work would be appropriate.*

We appreciate these suggestions and have modified the manuscript to refer to most of these works.

> *(4) Reviewer 1 pointed to the paper of Gong et al eLife 2013 and asked whether it is known how robust the proteins in our study are to changes in conformation/stability compared to other proteins, and whether this might impact the likelihood of observing higher-order epistasis in this system.*

The DBDs that we study here are very stable, and previous work shows that mutations affect DNA specificity primarily by modifying the DBD's affinity rather than its stability (McKeown et al., Cell 2014). Additionally, Gong et al.'s findings pertain to a globally nonlinear relationship between stability and function, which arises from the Boltzmann relationship between the energy of folding and occupancy of the folded state. Because our data are categorical – based on rank-order of measured phenotype rather than fluorescence as a continuous phenotype – the kind of global nonlinearity observed in Gong's study are not expected to produce spurious estimates of epistasis in our work. We have modified the discussion to discuss the point.

> *(5) Reviewer 1 asked a) why the epistatic models produce landscapes on which variants have fewer neighbors on average than main-effects only models and b) why the average distance from all ERE-specific nodes to all SRE-specific nodes is greater with epistasis (but the average distance from ERE to nearest SRE is lower with epistasis).*

In the main effects-only landscape, the functional genotypes are relatively similar to each other, because each must contain several of the states that contribute the most to a positive genetic score. Moreover, ERE-specific nodes are similar to each other, and SRE-specific nodes are similar to each other, because each must contain one or more of a relatively small number of specificity-determining states. When epistasis is added to the genetic architecture, two things happen: 1) more genotypes become functional because there are more combinations that can exceed the threshold score to produce a functional activator and 2) these additional functional variants are more different from each other – in general, and within the classes of ERE- or SRE-specific variants – because there are now more diverse combinations of states that can yield either phenotype. As a result, a broader span of sequence space is occupied, but ERE- and SRE-specific variants are more interspersed with each other. This means that the average distance between all pairs of nodes is greater, and this applies to all ERE-SRE pairs, as well. However, the interspersing means that the closest single SRE to any particular ERE is closer than it was without epistasis. We have added this explanation to the main text.

> *(6) Reviewer 2 asked us to explain why average path length increases with pairwise epistasis as the strength of selection for specificity increases.*

This behavior occurs because of the existence of a local peak in the pairwise model. Genotypes on this peak contained few connections to other genotypes, all of which were less SRE specific. Thus, with strong selection, i.e. high population size, the simulations became

stuck on the local peak, cycling among the genotypes many times before leaving, resulting in a large increase in the mean step number. As shown in the rest of the figure, when the longest set of paths are removed, there are still differences in the average number of steps with and without epistasis. This issue is described in the methods section.

> *(7) Reviewers made several suggestions for clarity in the text and figures.*

We have modified the paper to address all of these comments.

> *(8) Reviewer 3 stated that the code should be available.*

The code is available at https://github.com/JoeThorntonLab/DBD.GeneticArchitecture.