Title

MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect

Authors

Ammar Tareen[1], Mahdi Kooshkbaghi[1], Anna Posfai[1], William T. Ireland[2,3], David M. McCandlish[1], Justin B. Kinney[1,*]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

[2] Department of Physics, California Institute of Technology, Pasadena, CA, 91125

[3] Present address: Department of Applied Physics, Harvard University, Cambridge, MA, 02134

[*] To whom correspondence should be addressed: jkinney@cshl.edu

Abstract

Multiplex assays of variant effect (MAVEs) are diverse techniques that include deep mutational scanning (DMS) experiments on proteins and massively parallel reporter assays (MPRAs) on cis-regulatory sequences. MAVEs are being rapidly adopted in many areas of biology, but a general strategy for inferring quantitative models of genotype-phenotype (G-P) maps from MAVE data is lacking. Here we introduce a conceptually unified approach for learning G-P maps from MAVE datasets. Our strategy is grounded in concepts from information theory, and is based on the view of G-P maps as a form of information compression. We also introduce MAVE-NN, an easy-to-use Python package that implements this approach using a neural network backend. The ability of MAVE-NN to infer diverse G-P maps—including biophysically interpretable models—is demonstrated on DMS and MPRA data in a variety of biological contexts. MAVE-NN thus provides a unified solution to a major outstanding need in the MAVE community.

Main Text

Introduction

Over the last decade, the ability to quantitatively study genotype-phenotype (G-P) maps has been revolutionized by the development of multiplex assays of variant effect (MAVEs), which can measure molecular phenotypes for thousands to millions of genotypic variants in parallel.[1,2] MAVE is an umbrella term that describes a diverse set of experimental methods, three examples of which are illustrated in **Fig. 1**. Deep mutational scanning (DMS) experiments[3] are a type of MAVE commonly used to study protein sequence-function relationships. These assays work by linking variant proteins to their coding sequences, either directly or indirectly, then using deep sequencing to assay which variants survive a process of activity-dependent selection (e.g., **Fig. 1a**). Massively parallel reporter assays (MPRAs) are another major class of MAVE, and are commonly used to study DNA or RNA sequences that regulate gene expression at a variety of steps, including transcription, mRNA splicing, cleavage and polyadenylation, translation, and mRNA decay.[4–7] MPRAs typically rely on either an RNA-seq readout of barcode abundances (**Fig. 1c**) or the sorting of cells expressing a fluorescent reporter gene (**Fig. 1e**).

Most computational methods for analyzing MAVE data have focused on accurately quantifying the activity of individual assayed sequences.[8–14] However, MAVE measurements like enrichment ratios or cellular fluorescence levels usually cannot be interpreted as providing direct quantification of biologically meaningful activities, due to the presence of experiment-specific nonlinearities and noise. Moreover, MAVE data is usually incomplete, as one often wishes to understand G-P maps over vastly larger regions of sequence space than can be exhaustively assayed. The explicit quantitative modeling of G-P maps can address both the indirectness and incompleteness of MAVE measurements.[1,15] The goal here is to determine a mathematical function that, given a sequence as input, will return a quantitative value for that sequence's molecular phenotype. Such quantitative modeling has been of great interest since the earliest MAVE methods were developed,[16–18] but no general-use software has yet been described for inferring G-P maps of arbitrary functional form from MAVE data.

Here we introduce a unified conceptual framework for the quantitative modeling of MAVE data. This framework is based on the use of latent phenotype models, which assume that each assayed sequence has a well-defined latent phenotype (specified by the G-P map), of which the MAVE experiment provides an indirect readout (described by the measurement process). The quantitative forms of both the G-P map and the measurement process are then inferred from MAVE data simultaneously. We further introduce an information-theoretic approach for separately assessing the performance of the G-P map and the measurement process components of latent phenotype models. This strategy is implemented in an easy-to-use open-source Python package called MAVE-NN, which is built on a TensorFlow 2 backend.[19] In what follows, we expand on this unified MAVE modeling strategy and apply it to a diverse array of DMS and MPRA datasets. Along the way we note the substantial advantages that MAVE-NN provides over other MAVE modeling methods, illustrate how the capabilities of MAVE-NN can inform experimental design going forward, and highlight new biological insights that our quantitative modeling of MAVE data reveals.

Results

Latent phenotype modeling strategy

MAVE-NN supports the analysis of MAVE data on DNA, RNA, and protein sequences, and can accommodate either continuous or discrete measurement values. Given a set of sequence-measurement pairs, MAVE-NN aims to infer a probabilistic mapping from sequence to measurement. Our primary enabling assumption, which is encoded in the structure of the latent phenotype model (**Fig. 2a**), is that this mapping occurs in two stages. Each sequence is first mapped to a latent phenotype by a deterministic G-P map, then this latent phenotype is mapped to possible measurement values via a stochastic measurement process. During training, the G-P map and measurement process are simultaneously learned by maximizing a regularized form of likelihood. Our initial implementation of MAVE-NN assumes that latent phenotypes are one-dimensional quantities, but multidimensional latent phenotypes are fully compatible within this conceptual framework.[20,21]

MAVE-NN includes four types of built-in G-P maps: additive, neighbor, pairwise, and black box. Additive G-P maps assume that each character at each position within a sequence contributes independently to the latent phenotype. Neighbor G-P maps incorporate interactions between nearest-neighbor characters, while pairwise G-P maps include interactions between all pairs of characters regardless of their position. Black box G-P maps have the form of a densely connected multilayer perceptron, the specific architecture of which can be controlled by the user. MAVE-NN also supports custom G-P maps that can be used, e.g., to represent specific biophysical hypotheses about the mechanisms of sequence function.

To handle both discrete and continuous measurement values, two different strategies for modeling measurement processes are provided. Measurement process agnostic (MPA) regression uses techniques from the biophysics literature[15,16,20,22] to analyze MAVE datasets that report discrete measurements. Here the measurement process is represented by an overparameterized neural network that takes the latent phenotype value as input and outputs the probability of each possible measurement value (**Fig. 2b**). Global epistasis (GE) regression, by contrast, leverages ideas previously developed in the evolution literature[23–26] for analyzing datasets that contain continuous measurements (**Fig. 2c**). Here, the latent phenotype is nonlinearly mapped to a prediction that represents the most probable measurement value. A noise model is then used to describe the distribution of likely deviations from this prediction. MAVE-NN supports both homoscedastic and heteroscedastic noise models based on three different classes of probability distribution: Gaussian, Cauchy, and skewed-t. We note that the skewed-t distribution, introduced by Jones and Faddy,[27] reduces to Gaussian and Cauchy distributions in certain limits while also accommodating asymmetric experimental noise. **Fig. 2d** shows an example of a GE measurement process with a heteroscedastic skewed-t noise model.

Information-theoretic measures of model performance

We further propose three distinct quantities for assessing the performance of latent phenotype models (**Fig. 2e**). These quantities are motivated by thinking of G-P maps in terms of information compression. In information theory, a quantity called mutual information quantifies the amount of information that one variable encodes about another.[28,29] Unlike standard metrics of model performance, like accuracy or R^2, mutual information can be computed between any two types of variables (discrete, continuous, multi-dimensional, etc.). This property makes the information-based quantities we propose below applicable to all MAVE datasets, regardless of the specific type of experimental readout used. We note, however, that accurately estimating mutual information and related quantities from finite data is nontrivial and that MAVE-NN uses a variety of approaches to do this.

Intrinsic information, I_{int}, is the mutual information between the sequences and measurements contained within a MAVE dataset. This quantity provides a benchmark against which to compare the performance of inferred G-P maps. Predictive information, I_{pre}, is the mutual information between MAVE measurements and the latent phenotype values predicted by a G-P map of interest. This quantifies how well the G-P map preserves sequence-encoded information that is determinative of experimental measurements. When evaluated on test data, I_{pre} is bounded above by I_{int}, and equality obtains only when the latent phenotype losslessly encodes relevant sequence-encoded information. Variational information, I_{var}, is a linear transformation of log likelihood that provides a variational lower bound on I_{pre}.[30–32] The difference between I_{pre} and I_{var} quantifies how accurately the inferred measurement process matches the observed distribution of measurements and latent phenotypes (see **Supplemental Information**).

MAVE-NN infers model parameters by maximizing a (lightly) regularized form of likelihood. These computations are performed using the standard backpropagation-based training algorithms provided within the TensorFlow 2 backend. With certain caveats noted (see **Methods**), this optimization procedure maximizes I_{pre} while avoiding the costly estimates of mutual information at each iteration that have hindered the adoption of previous mutual-information-based modeling strategies.[16]

Application: deep mutational scanning assays

We now demonstrate the capabilities of MAVE-NN on three DMS datasets, starting with the study of Olson et al.[33] on pairwise epistasis in protein G. Here the authors measured the effects of all single and nearly all double mutations to residues 2-56 of the IgG binding domain. This domain, called GB1, has long served as a model system for studying protein sequence-function relationships. To assay the binding of GB1 variants to IgG, the authors combined mRNA display with ultra-high-throughput DNA sequencing (**Fig. 1a**). The resulting dataset reports log enrichment values for all 1,045 single- and 530,737 double-mutant GB1 variants (**Fig. 1b**).

Inspired in by the work of Otwinowski et al.,[26] we used MAVE-NN to infer a latent phenotype model comprising an additive G-P map and a GE measurement process. This inference procedure required only about 3 minutes on a standard laptop computer (**Supplemental Fig. S1**). **Fig. 3a** illustrates the inferred additive G-P map via the effects that every possible single-residue mutation has on the latent phenotype. From this heatmap of additive effects, we can immediately identify all of the critical GB1 residues, including residues

27, 31, 41, 43, and 52. We also observe that missense mutations to proline throughout the GB1 domain tend to negatively impact IgG binding, as expected due to this amino acid's exceptional conformational rigidity. **Fig. 3b** illustrates the corresponding GE measurement process, revealing a sigmoidal relationship between log enrichment measurements and the latent phenotype values predicted by the G-P map. Nonlinearities like this are ubiquitous in DMS data due to the presence of background and saturation effects. Unless they are explicitly accounted for in one's quantitative modeling efforts, as they are here, these nonlinearities can greatly distort the parameters of inferred G-P maps. **Fig. 3c** shows that accounting for this nonlinearity yields predictions that correlate quite well with measurement values. Moreover, every latent phenotype model inferred by MAVE-NN can be used as a MAVE dataset simulator (see Methods). By analyzing simulated data generated by our inferred model for this GB1 experiment, we further observed that MAVE-NN can accurately and robustly recover the GE nonlinearity and ground-truth G-P map parameters (**Supplementary Fig. S1**).

Fig. 3d summarizes the values of our information-theoretic metrics for model performance. On held-out test data, we find that $I_{\mathrm{var}} = 2.194 \pm 0.020\ \mathrm{bits}$ and $I_{\mathrm{pre}} = 2.220 \pm 0.008\ \mathrm{bits}$ and. The similarity of these two values suggests that the inferred GE measurement process, which includes a heteroscedastic skewed-t noise model, has nearly sufficient accuracy to fully describe the distribution of residuals. We further find that $2.680 \pm 0.008\ \mathrm{bits} \leq I_{\mathrm{int}} \leq 3.213 \pm 0.033\ \mathrm{bits}$ (see **Methods**), meaning that the inferred G-P map accounts for 70%-84% of the total sequence-dependent information in the dataset. While this performance is impressive, the additive G-P map evidently misses some relevant sequence features. This observation motivates the more complex biophysical model for GB1 discussed later in **Results**.

The ability of MAVE-NN to deconvolve experimental nonlinearities from additive G-P maps requires that some of the assayed sequences contain multiple mutations. This is because such nonlinearities are inferred by reconciling the effects of single mutations with the effects observed for combinations of two or more mutations. To investigate how many multiple-mutation variants are required, we performed GE inference on subsets of the GB1 dataset containing all 1,045 single-mutation sequences and either 50,000, 5,000, or 500 double-mutation sequences (see **Methods**). The shapes of the resulting GE nonlinearities are illustrated in **Figs. 3e-g**. Remarkably, MAVE-NN is able to recover the underlying nonlinearity using only about 500 randomly selected double mutants, which represent only ~0.1% of all possible double mutants. The analysis of simulated data also supports the ability to accurately recover ground-truth model predictions using highly reduced datasets (**Supplemental Fig. S1**). These findings have important implications for the design of DMS experiments: even if one only wants to determine an additive G-P map, including a modest number of multiple-mutation sequences in the assayed library is often advisable because it may allow the removal of artifactual nonlinearities.

To test the capabilities of MAVE-NN on less complete DMS datasets, we analyzed recent experiments on amyloid beta (Aβ)[34] and TDP-43,[35] both of which exhibit aggregation behavior in the context of neurodegenerative diseases. Like with GB1, the variant libraries used in both experiments included a substantial number of multiple-mutation sequences: 499 single- and 15,567 double-mutation sequences for Aβ; 1,266 single- and 56,730 double-mutation sequences for TDP-43. But unlike with GB1, these datasets are highly incomplete due to the use of mutagenic PCR for variant library creation.

We used MAVE-NN to infer additive G-P maps from these two datasets, adopting the same type of latent phenotype model used for GB1. **Fig. 4a** illustrates the additive G-P map inferred from aggregation measurements of Aβ variants. In agreement with the original study, we see that most amino acid mutations between positions 30-40 have a negative effect on nucleation, suggesting that this region plays a major role in nucleation behavior. **Fig. 4b** shows the corresponding measurement process. Even though these data are much sparser than the GB1 data, the inferred model performs well on held-out test data ($I_{\mathrm{var}} = 1.147 \pm 0.043$ bits,

$I_{\text{pre}} = 1.254 \pm 0.024$ bits, $R^2 = 0.793 \pm 0.071$). Similarly, **Figs. 4c-d** show the G-P map parameters and GE measurement process inferred from toxicity measurements of TDP-43 variants, revealing among other things the toxicity-determining hot-spot observed by Bolognesi et al.[35] at positions 310-340. The resulting latent phenotype model performs well on held-out test data ($I_{\text{var}} = 1.806 \pm 0.018$ bits, $I_{\text{pre}} = 2.011 \pm 0.019$ bits, $R^2 = 0.912 \pm 0.052$).

Application: a massively parallel splicing assay

Exon/intron boundaries are defined by 5´ splice sites (5´ss), which bind the U1 snRNP during the initial stages of spliceosome assembly. To investigate how 5´ss sequence quantitatively controls alternative mRNA splicing, Wong et al.[36] used a massively parallel splicing assay (MPSA) to measure percent-spliced-in (PSI) values for nearly all 32,768 possible 5´ss of the form NNN/GYNNNN in three different genetic contexts (**Fig. 1c,d**). Applying MAVE-NN to data from the BRCA2 exon 17 context, we inferred four different types of G-P maps: additive, neighbor, pairwise, and black box. As with GB1, these G-P maps were each inferred using GE regression with a heteroscedastic skewed-t noise model. For comparison, we also inferred an additive G-P map using the epistasis package of Sailer and Harms.[25]

Fig. 5a compares the performance of these G-P map models on held-out test data, while **Figs. 5b-d** illustrate the corresponding inferred measurement processes. We observe that the additive G-P map inferred using the epistasis package[25] exhibits less predictive information ($I_{\text{pre}} = 0.220 \pm 0.012$ bits) than the additive G-P map found using MAVE-NN ($P = 0.007$, two-sided z-test). This is likely because the epistasis package estimates the parameters of the additive G-P map prior to estimating the GE nonlinearity. We also note that, while the epistasis package provides a variety of options for modeling the GE nonlinearity, none of these options appear to work as well as our mixture-of-sigmoids approach (compare **Figs. 5b,c**). This finding again demonstrates that the accurate inference of G-P maps requires the explicit and simultaneous modeling of experimental nonlinearities.

We also observe that increasingly complex G-P maps exhibit increased accuracy. For example, the additive G-P map gives $I_{\mathrm{pre}} = 0.262 \pm\ 0.011\ \mathrm{bits}$, whereas the pairwise G-P map (**Figs. 5e,f**) attains $I_{\mathrm{pre}} = 0.367 \pm 0.015$ bits. We note that the parameters of the pairwise G-P map appear to be very precisely determined, as MAVE-NN was able to accurately recover ground-truth parameters from simulated datasets of the same size (**Supplemental Fig. S2**). The black box G-P map, which is comprised of 5 densely connected hidden layers of 10 nodes each, performed the best of all four G-P maps, achieving $I_{\mathrm{pre}} = 0.489 \pm 0.012\ \mathrm{bits}$. Remarkably, this last predictive information value exceeds the lower bound of $I_{\mathrm{int}} \geq 0.461 \pm 0.007\ \mathrm{bits}$, which was estimated from replicate experiments (see **Methods**). We thus conclude that pairwise interaction models are not flexible enough to fully account for how 5´ss sequences control splicing. More generally, these results underscore the need for software that is capable of inferring and assessing a variety of different G-P maps through a uniform interface.

Application: biophysically interpretable G-P maps

Biophysical models, unlike the phenomenological models considered thus far, have mathematical structures that reflect specific hypotheses about how sequence-dependent interactions between macromolecules mechanistically define G-P maps. Thermodynamic models, which rely on a quasi-equilibrium assumption, are the most commonly used type of biophysical model.[37–39] Previous studies have shown that precise thermodynamic models can be inferred from MAVE datasets,[16] but no software intended use by the broader MAVE community has yet been developed for doing this. MAVE-NN meets this need by enabling the inference of custom G-P maps. We now demonstrate this biophysical modeling capability in the contexts of protein-ligand binding (using DMS data; **Fig. 1a**) and bacterial transcriptional regulation (using sort-seq MPRA data; **Fig. 1e**).

Otwinowski[40] showed that a three-state thermodynamic G-P map (**Fig. 6a**), one that accounts for GB1 folding energy in addition to GB1-IgG binding energy,[41] can explain the DMS data of Olson et al.[33] better than a simple additive G-P map does. This biophysical model subsequently received impressive confirmation in the work of Nisthal et al.,[42] who measured the thermostability of 812 single-mutation GB1 variants. We tested the ability of MAVE-NN to recover the same type of thermodynamic model that Otwinowski had inferred using custom analysis scripts. Our analysis yielded a G-P map with significantly improved performance on the data of Olson et al. ($I_{\text{var}} = 2.353 \pm 0.012$ bits, $I_{\text{pre}} = 2.373 \pm 0.009$ bits, $R^2 = 0.948 \pm 0.002$) relative to the additive G-P map of **Fig. 3**. **Fig. 6b** shows the two inferred energy matrices that respectively describe the effects of every possible single-residue mutation on the Gibbs free energies of protein folding and protein-ligand binding. The folding energy predictions our model also correlate as well with the data of Nisthal et al. ($R^2 = 0.548 \pm 0.050$) as the predictions of Otwinowski's model does ($R^2 = 0.517 \pm 0.058$). This demonstrates that MAVE-NN can infer accurate and interpretable quantitative models of protein biophysics.

To test MAVE-NN's ability to infer thermodynamic models of transcriptional regulation, we first re-analyzed the MPRA data of Kinney et al.,[16] in which random mutations to a 75 bp region of the *Escherichia coli lac* promoter were assayed. This promoter region binds two regulatory proteins, σ^{70} RNA polymerase (RNAP) and the transcription factor CRP. As in Kinney et al.,[16] we proposed a four-state thermodynamic model that quantitatively explains how promoter sequences control transcription rate (**Fig. 6c**). The parameters of this G-P map include the Gibbs free energy of interaction between CRP and RNAP, as well as energy matrices that describe the CRP-DNA and RNAP-DNA interaction energies. Because the sort-seq MPRA of

Kinney et al. yielded discrete measurement values (**Figs. 1e,f**), we used an MPA measurement process in our latent phenotype model (**Fig. 6d**). The biophysical parameter values we thus inferred (**Fig. 6e**) largely match those of Kinney et al., but were obtained far more rapidly (in ~10 min versus multiple days) thanks to the use of stochastic gradient descent rather than Metropolis Monte Carlo.

Next we analyzed sort-seq MPRA data obtained by Belliveau et al.[43] for the *xylE* promoter, which had no regulatory annotation prior to that study and for which no biophysical model had yet been developed. Based on their MPRA data, as well as follow-up mass spectrometry experiments, Belliveau et al. proposed that *xylE* is regulated by RNAP, CRP, and the locus-specific regulator XylR. These findings motivated us to propose and train an eight-state thermodynamic model describing how interactions between these three regulatory proteins might control *xylE* expression (**Fig. 6f**). The resulting quantitative model includes energy matrix descriptions for RNAP, CRP, and XylR binding to DNA, as well as Gibbs free energy values for the CRP-XylR and XylR-RNAP interactions (**Fig. 6g)**. From this model we see that XylR activates RNAP through what appears to be a class II activation mechanism,[44] as energetic contributions from the -35 region of the RNAP binding site are markedly reduced in the *xylE* context relative to the *lac* context (**Fig. 6e**). We also see that CRP—a homodimer with dyadic symmetry—binds its site with remarkable asymmetry (again, compare to **Fig. 6e**). The biophysical factors that determine whether symmetric transcription factors like CRP interact with DNA in symmetric or asymmetric poses are poorly understood, and represent just one avenue of investigation opened up by the capabilities of MAVE-NN. More generally, these results provide a proof-of-principle demonstration of how MAVE-NN can be used, together with MPRA experiments, to establish biophysical models for previously uncharacterized gene regulatory sequences.

Discussion

In this work we have presented a unified strategy for inferring quantitative models of G-P maps from diverse MAVE datasets. At the core of our approach is the conceptualization of G-P maps as a form of information compression, i.e., that the G-P map first compresses an input sequence into a latent phenotype value, which the MAVE then reads out indirectly via a noisy nonlinear measurement process. By explicitly modeling this measurement process, one can remove potentially confounding effects from the G-P map, as well as accommodate diverse experimental designs. We have also introduced three information-theoretic metrics for assessing the performance of the resulting models. These capabilities have been implemented within an easy-to-use Python package called MAVE-NN.

We have demonstrated the capabilities of MAVE-NN in diverse biological contexts, including in the analysis of both DMS and MPRA data. We have also demonstrated the superior performance of MAVE-NN relative to the epistasis package of Sailer and Harms.[25] Along the way, we observed that MAVE-NN can deconvolve experimental nonlinearities from additive G-P maps when a relatively small number of sequences containing multiple mutations are included in the assayed libraries. This capability provides a compelling reason for experimentalists to include such sequences in their MAVE libraries, even if they are primarily interested in the effects of single mutations. Finally, we showed how MAVE-NN can learn biophysically interpretable G-P maps from both DMS and MPRA data.

Applying MAVE-NN to the MPSA data of Wong et al.,[36] we discovered that pairwise interaction models are not sufficient to describe how 5´ss sequences govern alternative mRNA splicing, and that higher-order epistatic interactions are needed to describe this critical aspect of eukaryotic biology. We also inferred the first biophysical model for transcriptional regulation by the *xylE* promoter. This biophysical model reveals that the well-studied transcription factor CRP binds its target site with surprising asymmetry *in vivo*, an intriguing phenomenon about which much remains to be learned.

MAVE-NN thus fills a critical need in the MAVE community, providing user-friendly software capable of learning quantitative models of G-P maps from diverse MAVE datasets. MAVE-NN has a streamlined user interface, is thoroughly tested, and is readily installed from PyPI by executing “pip install mavenn” at the command line. Comprehensive documentation, worked examples, and step-by-step tutorials are available at http://mavenn.readthedocs.io.

Acknowledgements. This work was supported by NIH grant 1R35GM133777 (awarded to JBK), NIH Grant 1R35GM133613 (awarded to DMM), an Alfred P. Sloan Research Fellowship (awarded to DMM), a grant from the CSHL/Northwell Health partnership, and funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

Author contributions. AT, WTI, DMM, and JBK conceived the project. AT and JBK wrote the software with assistance from AP and MK. WTI and JBK wrote a preliminary version of the software. AT, MK, and JBK performed the data analysis. AT, DMM, and JBK wrote the manuscript with contributions from MK and AP.

Conflicts of interest. The authors declare that they have no known conflicts of interest.

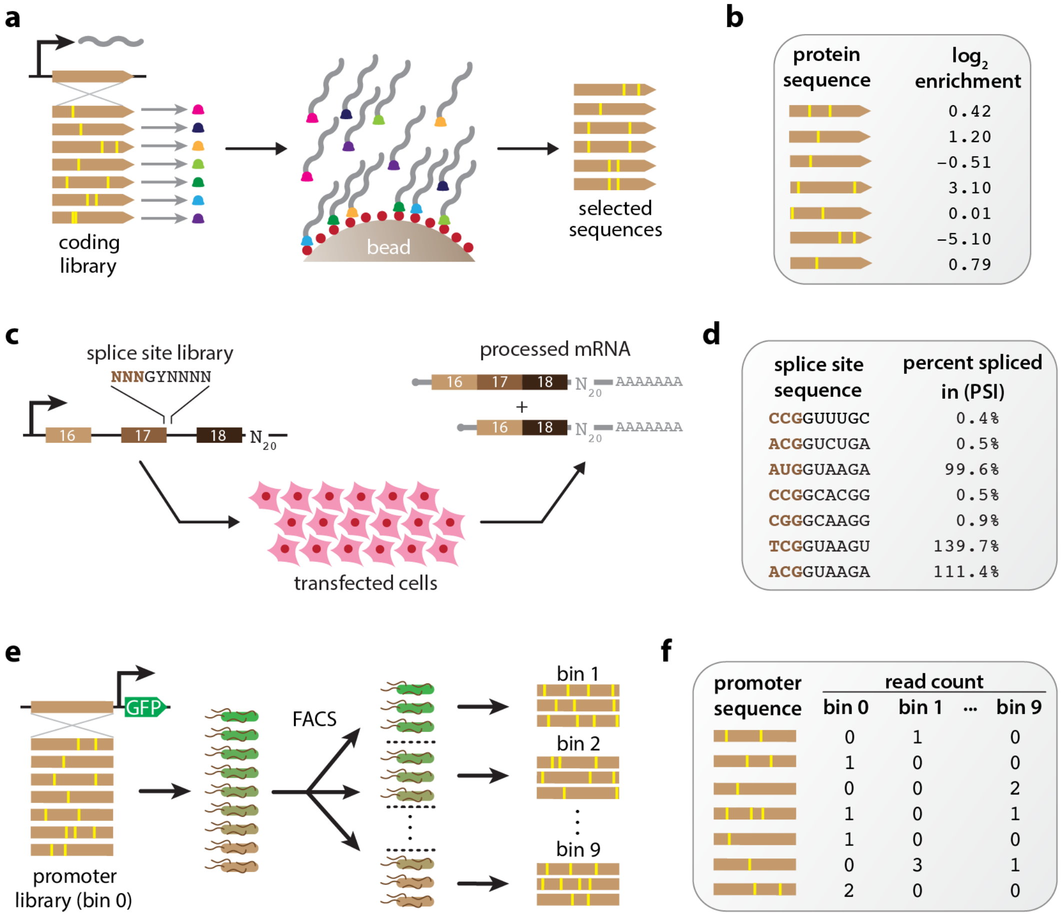

Figure 1. Three example MAVEs. (**a**) The DMS assay of Olson et al..[33] A library of variant GB1 proteins were covalently linked to their coding mRNAs using mRNA display. Functional GB1 proteins were then enriched using IgG beads, and deep sequencing was used to determine an enrichment ratio for each GB1 variant. (**b**) The resulting DMS dataset consists of variant protein sequences and their corresponding log enrichment values. (**c**) The MPSA of Wong et al..[36] A library of 3-exon minigenes was constructed from exons 16, 17, and 18 of *BRCA2*, with each minigene having a variant 5´ss at exon 17 and a random 20 nt barcode in the 3' UTR. This library was transfected into HeLa cells, and deep sequencing was used to quantify mRNA isoform abundance. (**d**) The resulting MPSA dataset comprises variant 5´ss with (noisy) PSI values. (**e**) The sort-seq MPRA of Kinney et al..[16] A plasmid library was generated in which randomly mutagenized versions of the *Escherichia coli lac* promoter drove the expression of GFP. Cells carrying these plasmids were sorted using FACS, and the variant promoters in each bin of sorted cells as well as the initial library were sequenced. (**f**) The resulting dataset comprises a list of variant promoter sequences, as well as a matrix of counts for each variant in each FACS bin. MAVE: multiplex assay of variant effect; DMS: deep

mutational scanning; MPSA: massively parallel splicing assay; 5´ss: 5´ splice site(s); PSI: percent spliced in; GFP: green fluorescent protein; FACS: fluorescence-activated cell sorting.

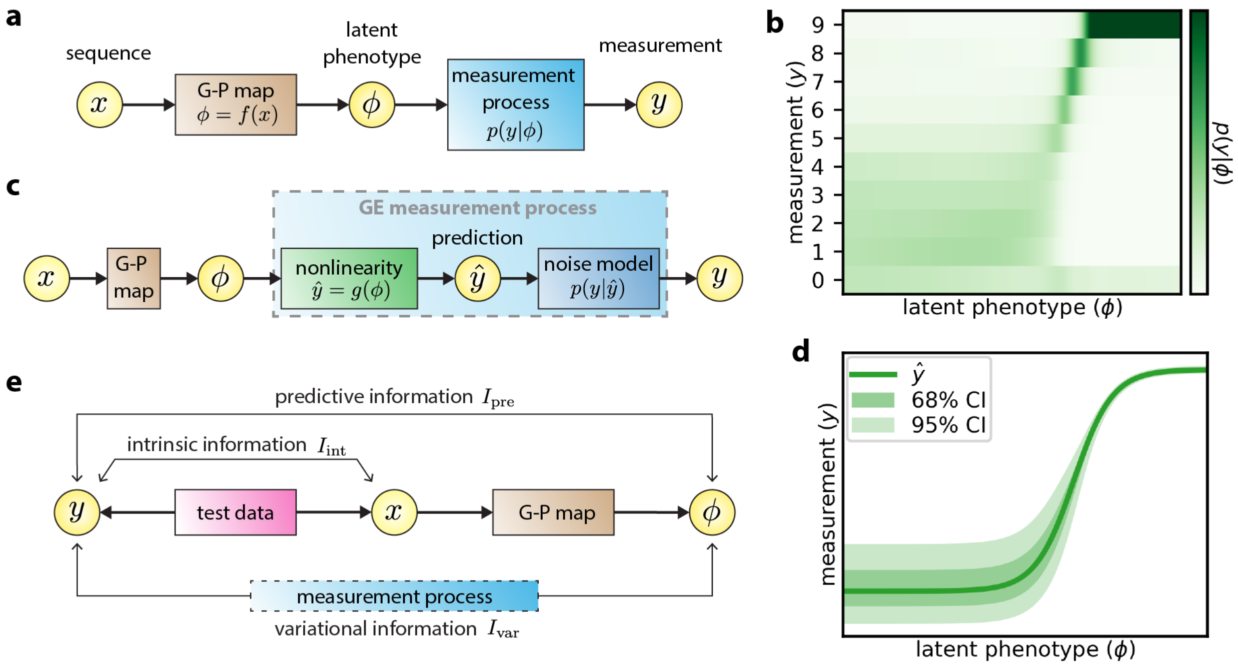

Figure 2. MAVE-NN quantitative modeling strategy. (**a**) Structure of latent phenotype models. A G-P map $f(x)$ maps each sequence x to a latent phenotype ϕ, after which a measurement process $p(y|\phi)$ determines the measurement y. (**b**) Example of an MPA measurement process inferred from the sort-seq MPRA data of Kinney et al..[16] MPA measurement processes are used when y values are discrete. (**c**) Structure of a GE regression model, which is used when y is continuous. A GE measurement process assumes that the mode of $p(y|\phi)$, called the prediction $\hat{y}$, is given by a nonlinear function $g(\phi)$, and the scatter about this mode is described by a noise model $p(y|\hat{y})$. (**d**) Example of a GE measurement process inferred from the DMS data of Olson et al..[33] Shown is the nonlinearity, the 68% CI, and the 95% CI. (**e**) Information-theoretic quantities used to assess model performance. Intrinsic information, I_{int}, is the mutual information between sequences x and measurements y. Predictive information, I_{pre}, is the mutual information between measurements y and the latent phenotype values ϕ assigned by a model. Variational information, I_{var} , is a linear transformation of log likelihood. The inequality $I_{\text{int}} \geq I_{\text{pre}} \geq I_{\text{var}}$ always holds on test data (modulo finite data uncertainties), with $I_{\text{int}} = I_{\text{pre}}$ when the G-P map is correct, and $I_{\text{pre}} = I_{\text{var}}$ when the measurement process correctly describes the distribution of y conditioned on ϕ. G-P: genotype-phenotype; MPA: measurement process agnostic; GE: global epistasis; CI: confidence interval.

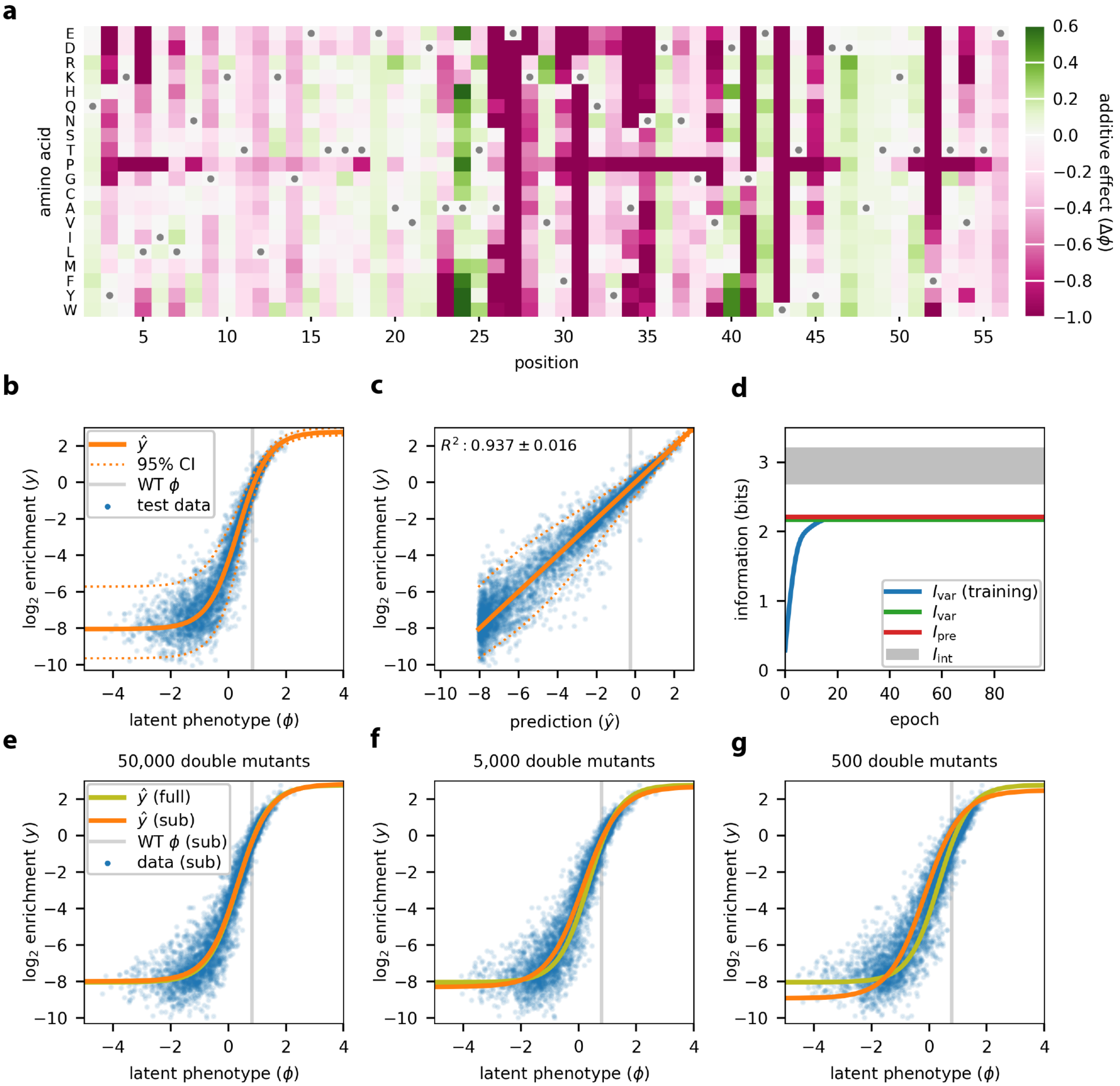

Figure 3. Analysis of DMS data for protein GB1. MAVE-NN was used to infer a latent phenotype model, consisting of an additive G-P map and a GE measurement process having a heteroskedastic skewed-t noise model, from the DMS data of Olson et al..[33] All 1,045 single variants and 530,737 pairwise variants reported for positions 2 to 56 of the GB1 domain were analyzed. Data were split 80:10:10 into training, validation, and test sets. (**a**) The G-P map parameters inferred from all pairwise variants. Gray dots indicate wildtype residues. Amino acids are ordered as in Olson et al..[33] (**b**) GE plot showing measurements versus predicted latent phenotype values for 5,000 randomly selected test-set

sequences (blue dots), alongside the inferred nonlinearity (solid orange line) and the 95% CI (dashed lines) of the noise model. Gray line indicates the latent phenotype value of the wildtype sequence. (**c**) Measurements plotted against $\hat{y}$ predictions for these same sequences. Dashed lines indicate the 95% CI of the noise model. Gray line indicates the wildtype sequence $\hat{y}$. (**d**) Corresponding information metrics computed during model training (using training data) or for the final model (using test data); uncertainties in these estimates are roughly the width of the plotted lines. Gray shaded area indicates allowed values for intrinsic information based on upper and lower bounds estimated as described in **Methods**. (**e-g**) Test set predictions (blue dots) and GE nonlinearities (orange lines) for models trained using subsets of the GB1 data containing all single mutants and 50,000 (**e**), 5,000 (**f**), or 500 (**g**) double mutants. The GE nonlinearity from panel **b** is shown for reference (yellow-green lines). Uncertainties reflect standard errors. GE: global epistasis; G-P: genotype-phenotype; CI: confidence interval.

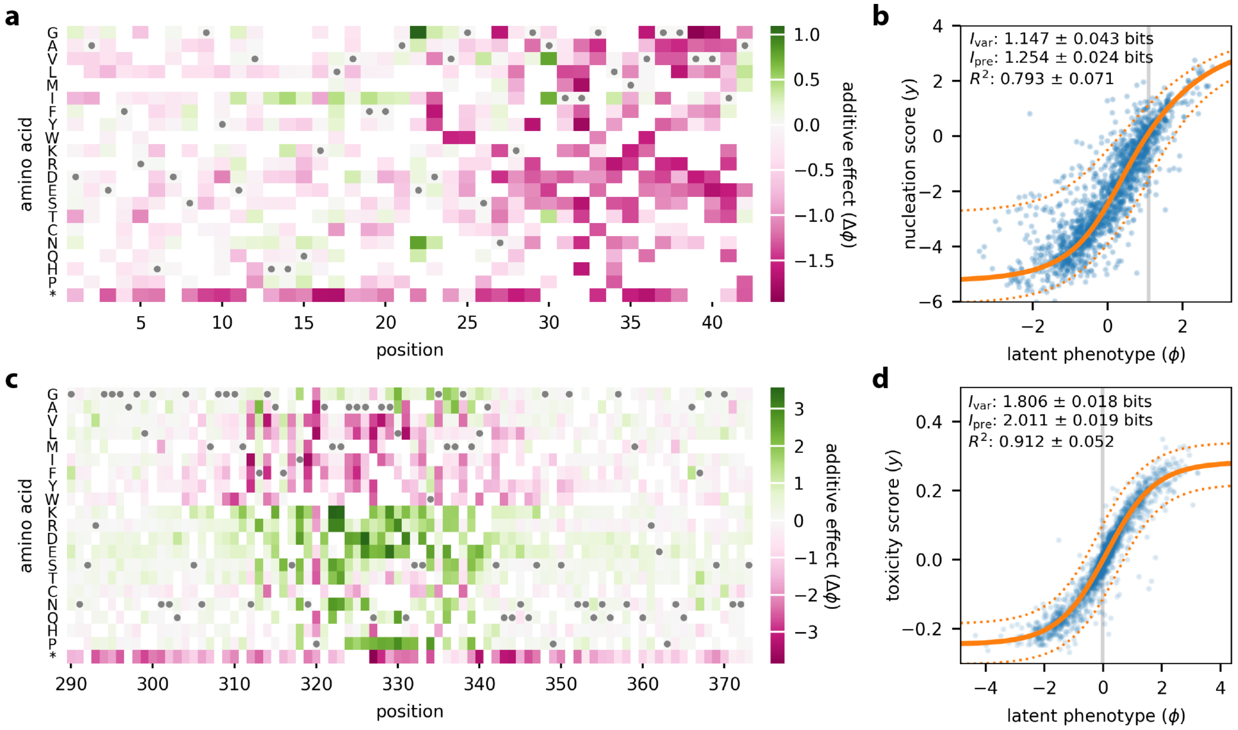

Figure 4. Analysis of DMS data for Aβ and TDP-43. (**a,b**) Seuma et al.[34] measured nucleation scores for 499 single mutants and 15,567 double mutants of Aβ. These data were used to train a latent phenotype model comprising (**a**) an additive G-P map and (**b**) a GE measurement process with a heteroskedastic skewed-t noise model. (**c,d**) Bolognesi et al.[35] measured toxicity scores for 1,266 single mutants and 56,730 double mutants of TDP-43. The resulting data were used to train (**c**) an additive G-P map and (**d**) a GE measurement process of the same form as in panel **b**. In both cases, data were split 90:5:5 into training, validation, and test sets. In (**a,c**), gray dots indicate the wildtype sequence, amino acids are ordered as in the original publications, and * indicates a stop codon. In (**b,d**), blue dots indicate latent phenotype values versus measurements for held-out test data, gray line indicates the latent phenotype value of the wildtype sequence, solid orange line indicates the GE nonlinearity, and dashed orange lines indicate a corresponding 95% CI for the inferred noise model. Values for I_{var}, I_{pre}, and R^2 (between y and $\hat{y}$) are also shown. Uncertainties reflect standard errors. **Supplemental Fig. S3** shows measurements plotted against the $\hat{y}$ predictions of these models. Aβ: amyloid beta; TDP-43: TAR DNA-binding protein 43; G-P: genotype-phenotype; GE: global epistasis; CI: confidence interval.

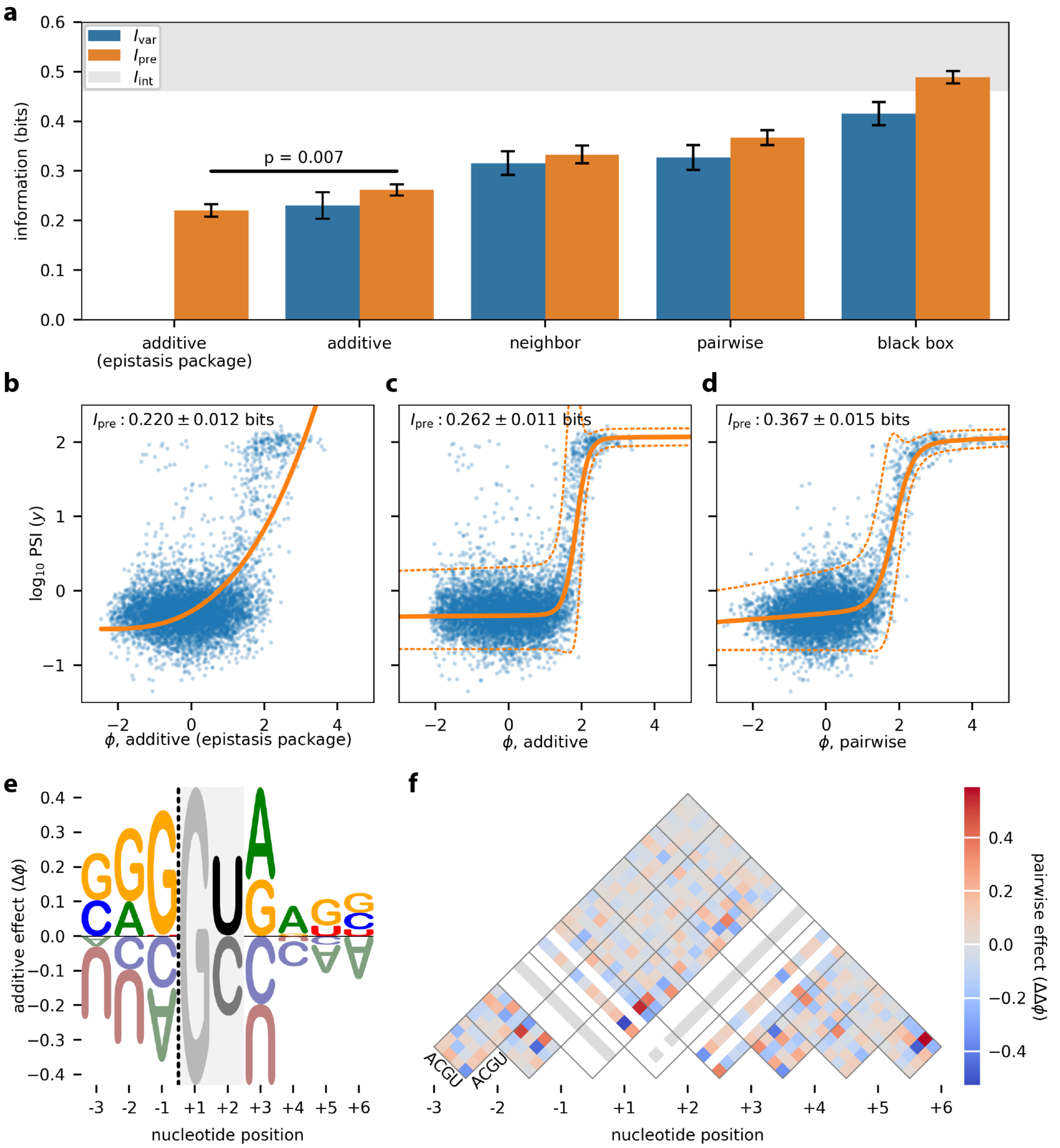

Figure 5. Analysis of MPSA data from Wong et al..[36] This dataset reports PSI values, measured in the *BRCA2* exon 17 context, for nearly all 32,768 variants 5´ss of the form NNN/GYNNNN. Data were split 60:20:20 into training, validation, and test sets. Latent phenotype models with one of four types of G-P map (additive, neighbor, pairwise, or black box), as well as a GE measurement process with a heteroscedastic skewed-t noise model, were inferred. The epistasis package of Sailer and Harms[25] was also used to infer an additive G-P map and GE nonlinearity. (**a**)

Performance of trained models as quantified by I_{var} and I_{pre}, computed on test data. The lower bound on I_{int} was estimated from experimental replicates (see **Methods**). p-value reflects a two-sided z-test. I_{var} was not computed for the additive (epistasis package) model because that package does not infer an explicit noise model. (**b-d**) Measurement values versus latent phenotype values, computed on test data, using the additive (epistasis package) model (**b**), the additive model (**c**), and the pairwise model (**d**). The corresponding GE measurement processes are also shown. (**e**) Sequence logo[45] illustrating the additive effects component of the pairwise G-P map. Dashed line indicates the exon/intron boundary. G at +1 serves as a placeholder because no other bases were assayed at this position. Only values for U and C at +2 were inferred. (**f**) Heatmap showing the pairwise effects component of the pairwise G-P map. White diagonals correspond to unobserved bases. Error bars indicate standard errors. MPSA: massively parallel splicing assay; PSI: percent spliced in; G-P: genotype-phenotype; GE: global epistasis.

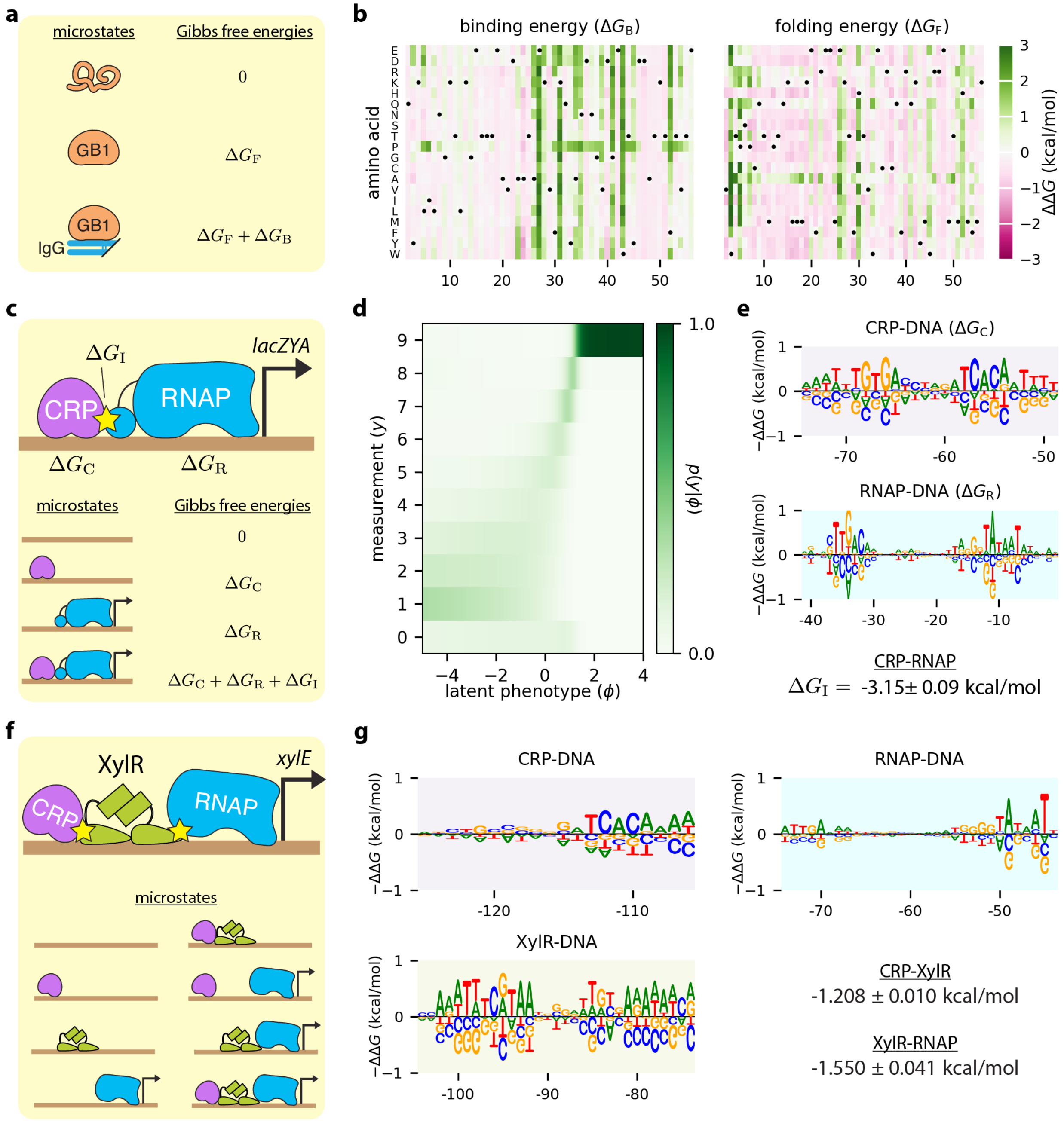

Figure 6. Biophysical models inferred from DMS and MPRA data. (**a**) Thermodynamic model for IgG binding by GB1. This model comprises three GB1 microstates (unfolded, folded-unbound, and folded-bound). The Gibbs free energies of folding (ΔG_F) and binding (ΔG_B) are computed from sequence using additive models called energy matrices. The latent phenotype is given by the fraction of time GB1 is in the folded-bound state. (**b**) The $\Delta\Delta G$ parameters of the energy matrices for folding and binding, inferred from the data of Olson et al.[33] using GE regression. **Supplemental Fig. S5** plots folding energy predictions against the measurements of Nisthal et al..[42] (**c**) A four-state thermodynamic

model for transcriptional activation at the *E. coli lac* promoter. The Gibbs free energies of RNAP-DNA binding (ΔG_{R}) and CRP-DNA binding (ΔG_{C}) are computed using energy matrices, whereas the CRP-RNAP interaction energy ΔG_{I} is a scalar. The latent phenotype is the fraction of time a promoter is bound by RNAP. (**d,e**) The latent phenotype model inferred from the sort-seq MPRA of Kinney et al.,[16] including both the MPA measurement process (**d**) and the parameters of the thermodynamic G-P map (**e**). (**f**) An eight-state thermodynamic model for transcriptional activity at the *xylE* promoter. (**g**) Corresponding G-P map parameters inferred from the sort-seq MPRA data of Belliveau et al..[43] These parameters include energy matrices describing the CRP-DNA, RNAP-DNA, and XylR-DNA interactions, as well as scalar values for the CRP-XylR and XylR-RNAP interaction free energies. Supplemental **Fig. S4** provides detailed definitions of the thermodynamic models in panels **a,c,f**. In panels **e,g**, sequence logos were generated using Logomaker,[45] and standard errors for protein-protein interactions energies were determined by analyzing simulated data. GE: global epistasis. RNAP: RNA polymerase. MPA: measurement-process agnostic. G-P: genotype-phenotype.

Online Methods

Notation

We represent each MAVE dataset as a set of N observations, $\{(x_n, y_n)\}_{n=0}^{N-1}$, where each observation consists of a sequence x_n and a measurement y_n. Here, y_n can be either a continuous real-valued number, or a nonnegative integer representing the "bin" in which the nth sequence was found. Note that, in this representation the same sequence x can be observed multiple times, potentially with different values for y due to experimental noise.

G-P maps

We assume that all sequences have the same length L, and that at each of the L positions in each sequence there is one of C possible characters. MAVE-NN represents sequences using a vector of one-hot encoded features of the form

$$x_{l:c} \quad = \quad \begin{cases} 1 & \text{if character } c \text{ occurs at position } l \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $l = 0,1,\ldots,L-1$ indexes positions within the sequence, and c indexes the C distinct characters. MAVE-NN supports built-in alphabets for DNA, RNA and protein (with or without stop codons), as well as user-defined sequence alphabets.

We assume that the latent phenotype is given by a linear function $\phi(x;\theta)$ that depends on a set of G-P map parameters θ. As mentioned in the main text, MAVE-NN supports four types of G-P map models, all of which can be inferred using either GE regression or MPA regression. The additive model is given by,

$$\phi_{\text{additive}}(x;\theta) \quad = \quad \theta_0 + \sum_{l=0}^{L-1} \sum_{c} \theta_{l:c}\, x_{l:c}\,, \tag{2}$$

and thus each position in x contributes independently to the latent phenotype. The neighbor model is given by,

$$\phi_{\text{neighbor}}(x;\theta) = \theta_0 + \sum_{l=0}^{L-1}\sum_{c}\theta_{l:c}\,x_{l:c} + \sum_{l=0}^{L-2}\sum_{c,c'}\theta_{l:c,l+1:c'}\,x_{l:c}x_{l+1:c'}\,, \tag{3}$$

and further accounts for potential epistatic interactions between neighboring positions. The pairwise model is given by,

$$\phi_{\text{pairwise}}(x;\theta) = \theta_0 + \sum_{l=0}^{L-1}\sum_{c}\theta_{l:c}\,x_{l:c} + \sum_{l=0}^{L-2}\sum_{l'=l+1}^{L-1}\sum_{c,c'}\theta_{l:c,l':c'}\,x_{l:c}x_{l':c'}\,, \tag{4}$$

and includes interactions between all pairs of positions. Note our convention of requiring $l' > l$ in the pairwise parameters $\theta_{l:c,l':c'}$.

Unlike these three parametric models, the black box G-P map does not have a fixed functional form. Rather, it is given by a multilayer perceptron that takes a vector of sequence features (additive, neighbor, or pairwise) as input, contains multiple fully-connected hidden layers with nonlinear activations, and has a single node output with a linear activation. Users are able to specify the number of hidden layers, the number of nodes in each hidden layer, and the activation function used by these nodes.

MAVE-NN further supports custom G-P maps that users can define by subclassing the G-P map base class. These G-P maps can have arbitrary functional form, e.g., representing specific biophysical hypotheses of sequence function. This feature of MAVE-NN is showcased in the analyses of **Fig. 6**.

Gauge modes and diffeomorphic modes

G-P maps typically have non-identifiable degrees of freedom that must be fixed, i.e., pinned down, before the values of individual parameters can be meaningfully interpreted or compared between models. These degrees of freedom come in two flavors: gauge modes and diffeomorphic modes. Gauge modes are changes to θ that do not alter the values of the latent phenotype ϕ. Diffeomorphic modes[15,20] are changes to θ that do alter ϕ, but do so in ways that can be undone by transformations of the measurement process $p(y|\phi)$. As shown by Kinney and Atwal,[15,20] the diffeomorphic modes of linear G-P maps like those considered here will in general correspond to affine transformations of ϕ, although additional unconstrained modes can occur in special situations.

MAVE-NN fixes both gauge modes and diffeomorphic modes of inferred models (except when using custom G-P maps). The diffeomorphic modes of G-P maps are fixed by transforming θ via

$$\theta_0 \rightarrow \theta_0 - a\,, \tag{5}$$

and then

$$\theta \rightarrow \frac{\theta}{b}, \tag{6}$$

where $a = \mathrm{mean}(\{\phi_n\})$ and $b = \mathrm{std}(\{\phi_n\})$ are the mean and standard deviation of ϕ values computed on the training data. This produces a corresponding change in latent phenotype values $\phi \rightarrow (\phi - a)/b$. To avoid altering likelihood values, MAVE-NN makes a corresponding transformation to the measurement process $p(y|\phi)$. In GE regression this is done by adjusting the GE nonlinearity via

$$g(\phi) \rightarrow g(a + b\phi)\,, \tag{7}$$

while keeping the noise model $p(y|\hat{y})$ fixed. In MPA regression MAVE-NN transforms the full measurement process via

$$p(y|\phi) \rightarrow p(y|a + b\phi)\,. \tag{8}$$

For the three parametric G-P maps, gauge modes are fixed using what we call the "hierarchical gauge." Here, the parameters θ are adjusted so that the lower-order terms in $\phi(x;\theta)$ account for the highest possible fraction of variance in ϕ. This procedure requires a probability distribution on sequence space with respect to which these variances are computed. MAVE-NN assumes that such distributions factorize by position, and can thus be represented by a probability matrix with elements $p_{l:c}$, denoting the probability of character c at position l. MAVE-NN provides three built-in choices for this distribution: uniform, empirical, or wildtype. The corresponding values of $p_{l:c}$ are given by

$$p_{l:c} \;=\; \begin{cases} 1/C & \text{for uniform} \\ n_{l:c}/N & \text{for empirical} \\ x^{\text{wt}}_{l:c} & \text{for wildtype} \end{cases}, \tag{9}$$

where $n_{l:c}$ denotes the number of sequences (out of N total) that have character c at position l, and $x^{\text{wt}}_{l:c}$ is the one-hot encoding of a user-specified wildtype sequence. In particular, the wildtype gauge was used for illustrating the additive G-P maps in **Fig. 3** and **Fig. 4**, while the uniform gauge was used for illustrating the pairwise G-P map in **Fig. 5** and the energy matrices in **Fig. 6**. After a sequence distribution is chosen, MAVE-NN fixes the gauge of the pairwise G-P map by transforming

$$\begin{aligned} \theta_0 \;\rightarrow\; & \theta_0 \\ & + \sum_{l} \sum_{c'} \theta_{l:c'}\, p_{l:c'} \\ & + \sum_{l} \sum_{l'>l} \sum_{c,c'} \theta_{l:c,l':c'}\, p_{l:c}\, p_{l':c'}\,, \end{aligned} \tag{10}$$

$$\begin{aligned}\theta_{l:c} \quad \rightarrow \quad & \theta_{l:c} \\ & -\sum_{c'} \theta_{l:c'}\, p_{l:c'} \\ & +\sum_{l'>l} \sum_{c'} \theta_{l:c,l':c'}\, p_{l':c'} \\ & +\sum_{l'<l} \sum_{c'} \theta_{l':c',l:c}\, p_{l':c'} \\ & -\sum_{l'>l} \sum_{c',c''} \theta_{l:c',l':c''}\, p_{l:c'}\, p_{l':c''} \\ & -\sum_{l'<l} \sum_{c',c''} \theta_{l':c'',l':c'}\, p_{l:c'}\, p_{l':c''}\,,\end{aligned} \tag{11}$$

and

$$\begin{aligned}\theta_{l:c,l':c'} \quad \rightarrow \quad & \theta_{l:c,l':c'} \\ & -\sum_{c''} \theta_{l:c'',l':c'}\, p_{l:c''} \\ & -\sum_{c''} \theta_{l:c,l':c''}\, p_{l':c''} \\ & +\sum_{c'',c'''} \theta_{l:c'',l':c'''}\, p_{l:c''}\, p_{l':c'''}\,.\end{aligned} \tag{12}$$

This transformation is also used for the additive and neighbor G-P maps, but with $\theta_{l:c,l':c'} = 0$ for all l, l' (additive) or whenever $l' \neq l + 1$ (neighbor).

GE nonlinearities

GE models assume that each measurement y is a nonlinear function of the latent phenotype $g(\phi)$ plus some noise. In MAVE-NN, this nonlinearity is represented as a sum of tanh sigmoids:

$$g(\phi; \alpha) = a + \sum_{k=0}^{K-1} b_k \tanh(c_k \phi + d_k)\,. \tag{13}$$

Here, K specifies the number of hidden nodes contributing to the sum, and $\alpha = \{a, b_k, c_k, d_k\}$ are trainable parameters. We note that this mathematical form is an example of the bottleneck architecture previously used by[21,24] for modeling GE nonlinearities. By default, MAVE-NN constrains $g(\phi; \alpha)$ to be monotonic in ϕ by requiring all $b_k \geq 0$ and $c_k \geq 0$, but this constraint can be relaxed.

GE noise models

MAVE-NN supports three types of GE noise model: Gaussian, Cauchy, and skew-t. These all support the analytic computation of quantiles and confidence intervals, as well as the rapid sampling of simulated measurement values. The Gaussian noise model is given by

$$p_{\mathrm{gauss}}(y|\hat{y}; s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left[-\frac{(y-\hat{y})^2}{2s^2}\right], \tag{14}$$

where s denotes the standard deviation. Importantly, MAVE-NN allows this noise model to be heteroskedastic by representing s as an exponentiated polynomial in $\hat{y}$, i.e.,

$$s(\hat{y}) = \exp\left[\sum_{k=0}^{K} a_k\, \hat{y}^k\right], \tag{15}$$

where K is the order of the polynomial and $\{a_k\}$ are trainable parameters. The user has the option to set K, and setting $K = 0$ renders this noise model homoscedastic. Quantiles are computed using $y_q = \hat{y} + s\sqrt{2}\,\mathrm{erf}^{-1}(2q-1)$ for user-specified values of $q \in [0,1]$. Similarly, the Cauchy noise model is given by

$$p_{\mathrm{cauchy}}(y|\hat{y}; s) = \left[\pi s\left(1 + \frac{(y-\hat{y})^2}{s^2}\right)\right]^{-1}, \tag{16}$$

where the scale parameter s is an exponentiated K'th order polynomial in $\hat{y}$, and quantiles are computed using $y_q = \hat{y} + s\tan\left[\pi(q - \frac{1}{2})\right]$.

The skew-t noise model is of the form described by Jones and Faddy,[27] and is given by

$$p_{\text{skewt}}(y|\hat{y}; s, a, b) = s^{-1} f(t; a, b)\,, \tag{17}$$

where

$$t = t^* + \frac{y - \hat{y}}{s}, \quad t^* = \frac{(a-b)\sqrt{a+b}}{\sqrt{2a+1}\sqrt{2b+1}}\,, \tag{18}$$

and

$$\begin{aligned} f(t; a, b) \quad = \quad & \frac{2^{1-a-b}}{\sqrt{a+b}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left[1 + \frac{t}{\sqrt{a+b+t^2}}\right]^{a+\frac{1}{2}} \\ & \times \left[1 - \frac{t}{\sqrt{a+b+t^2}}\right]^{b+\frac{1}{2}}\,. \end{aligned} \tag{19}$$

Note that the t statistic here is an affine function of y chosen so that the distribution's mode (corresponding to t^*) is positioned at $\hat{y}$. The three parameters of this noise model, $\{s, a, b\}$, are each represented using K-th order exponentiated polynomials with trainable coefficients. Quantiles are computed using

$$y_q = \hat{y} + \left(t_q - t^*\right)s\,, \tag{20}$$

where

$$t_q = \frac{\left(2x_q - 1\right)\sqrt{a+b}}{\sqrt{1 - \left(2x_q - 1\right)^2}}, \quad x_q = I_q^{-1}(a, b)\,, \tag{21}$$

and I^{-1} denotes the inverse of the regularized incomplete Beta function $I_x(a, b)$.

MPA measurement process

In MPA regression, MAVE-NN directly models the measurement process $p(y|\phi)$. At present, MAVE-NN only supports MPA regression for discrete values of y indexed using nonnegative integers. MAVE-NN supports two alternative forms of input for MPA regression. One is a set of sequence-measurement pairs $\{(x_n, y_n)\}_{n=0}^{N-1}$, where N is the total number of reads, $\{x_n\}$ is a set of (typically) non-unique sequences, each $y_n \in \{0,1,\ldots,Y-1\}$ is a bin number, and Y is the total number of bins. The other is a set of sequence-count-vector pairs $\{(x_m, c_m)\}_{m=0}^{M-1}$, where M is the total number of unique sequences and $c_m = (c_{m0}, c_{m1}, \ldots, c_{m(Y-1)})$ is a vector that lists the number of times c_{my} that the sequence x_m was observed in each bin y. MPA measurement processes are represented as multilayer perceptron with one hidden layer (having tanh activations) and a softmax output layer. Specifically,

$$p(y|\phi) = \frac{w_y(\phi)}{\sum_{y'} w_{y'}(\phi)}\,, \qquad (22)$$

where

$$w_y(\phi) = \exp\left[a_y + \textstyle\sum_{k=0}^{K-1} b_{yk} \tanh\left(c_{yk}\phi + d_{yk}\right)\right] \qquad (23)$$

and K is the number of hidden nodes per value of y. The trainable parameters of this measurement process are $\eta = \{a_y, b_{yk}, c_{yk}, d_{yk}\}$.

Loss function

Let θ denote the G-P map parameters, and η denote the parameters of the measurement process. MAVE-NN optimizes these parameters using stochastic gradient descent on a loss function given by

$$\mathcal{L} = \mathcal{L}_{\text{like}} + \mathcal{L}_{\text{reg}}\,, \qquad (24)$$

where $\mathcal{L}_{\text{like}}$ is the negative log likelihood of the model, given by

$$\mathcal{L}_{\text{like}}[\theta,\eta] = -\sum_{n=0}^{N-1} \log\left[p(y_n|\phi_n;\eta)\right] \qquad (25)$$

where $\phi_n = \phi(x_n;\theta)$, and $\mathcal{L}_{\text{reg}}$ provides for regularization of the model parameters.

In the context of GE regression, we can write $\eta = (\alpha, \beta)$ where α represents the parameters of the GE nonlinearity $g(\phi;\alpha)$, and β denotes the parameters of the noise model $p(y|\hat{y};\beta)$. The likelihood contribution from each observation n then becomes $p(y_n|\phi_n;\eta) = p(y_n|\hat{y}_n;\beta)$ where $\hat{y}_n = g(\phi_n;\alpha)$. In the context of MPA regression with a dataset of the form $\{(x_m, c_m)\}_{m=0}^{M-1}$, the loss function simplifies to

$$\mathcal{L}_{\text{like}}[\theta,\eta] \quad = \quad -\sum_{m=0}^{M-1}\sum_{y=0}^{Y-1} c_{my} \log[p(y|\phi_m;\eta)] \tag{26}$$

where $\phi_m = \phi(x_m;\theta)$. For the regularization term, MAVE-NN uses an L_2 penalty of the form

$$\mathcal{L}_{\text{reg}}[\theta,\eta] = \lambda_\theta|\theta|^2 + \lambda_\eta|\eta|^2\,, \tag{27}$$

where the user-adjusted parameters λ_θ and λ_η respectively control the strength of regularization for the G-P map and measurement process parameters.

Predictive information

In what follows, we use $p_{\text{model}}(y|\phi)$ to denote a measurement process inferred by MAVE-NN, whereas $p_{\text{true}}(y|\phi)$ denotes the empirical conditional distribution of y and ϕ values that would be observed in the limit of infinite test data.

Predictive information $I_{\text{pre}} = I[y;\phi]$, where $I[\cdot;\cdot]$ represents mutual information computed on data not used for training (i.e., a held-out test set or data from a different experiment), I_{pre} provides a measure of how strongly a G-P map predicts experimental measurements. Importantly, this quantity does not depend on the corresponding measurement process $p_{\text{model}}(y|\phi)$. To estimate I_{pre}, we use k'th nearest neighbor (kNN) estimators of entropy and mutual information adapted from the NPEET Python package.[46] Here, the user has the option of adjusting k, which controls a variance/bias tradeoff. When y is discrete (MPA regression), I_{pre} is computed using the classic kNN entropy estimator[47,48] via the decomposition $I[y;\phi] = H[\phi] - \sum_y p\,(y) H_y[\phi]$, where $H_y[\phi]$ denotes the entropy of $p_{\text{true}}(\phi|y)$. When y is continuous (GE regression), $I[y;\phi]$ is estimated using the kNN-based Kraskov Stögbauer Grassberger (KSG) algorithm.[48] This approach optionally supports the local nonuniformity correction of Gao et al.,[49] which is important when y and ϕ exhibit strong dependencies, but which also requires substantially more time to compute.

Variational information

We define variational information as an affine transformation of $\mathcal{L}_{\text{like}}$,

$$I_{\text{var}} = H[y] - \frac{\log_2(e)}{N}\mathcal{L}_{\text{like}}\,. \tag{28}$$

Here, $H[y]$ is the entropy of the data $\{y_n\}$, which is estimated using the k'th nearest neighbor (kNN) estimator from the NPEET package.[46] Noting that this quantity can also be written as $I_{\text{var}} = H[y] - \text{mean}(\{Q_n\})$, where $Q_n = -\log_2 p(y_n|\phi_n)$, we estimate the associated uncertainty using

$$\delta I_{\text{var}}[y;\phi] = \sqrt{\delta H[y]^2 + \frac{\text{var}(\{Q_n\})}{N}}\,. \tag{29}$$

The inference strategy used by MAVE-NN is based on the fact that I_{var} provides a tight variational lower bound on I_{pre}.[30] Indeed, in the large data limit,

$$I_{\text{pre}} = I_{\text{var}} + D_{\text{KL}}(p_{\text{true}}||p_{\text{model}})\,, \tag{30}$$

where $D_{\text{KL}}(\cdot) \geq 0$ is the Kullback-Leibler divergence, and thus quantifies the accuracy of the inferred measurement process. From **Eq. 30** one can see that, with appropriate caveats, maximizing I_{var} (or equivalently, $\mathcal{L}_{\text{like}}$) will also maximize I_{pre}.[20] But unlike I_{pre}, I_{var} is readily compatible with backpropagation and stochastic gradient descent. See Supplemental Information for a derivation of **Eq. 30** and an expanded discussion of this key point. Note: Sharpee et al.[50] cleverly showed that I_{pre} can, in fact, be optimized using stochastic gradient descent. Computing gradients of I_{pre}, however, requires a time-consuming density estimation step. Optimizing I_{var}, on the other hand, can be done using standard per-datum backpropagation.

Intrinsic information

Intrinsic information, $I_{\text{int}} = I[x;y]$, is the mutual information between the sequences x and measurements y in a dataset. This quantity is somewhat tricky to estimate due to the high-dimensional nature of sequence space. We instead used three different methods to obtain the upper and lower bounds on I_{int} shown in **Fig. 3d** and **Fig. 5a**. More generally, we believe the development of both computational and experimental methods for estimating I_{int} is be an important avenue for future research.

To compute the upper bound on I_{int} for GB1 data (in **Fig. 3d**), we used the fact that

$$I[x;y] = H[y] - \langle H_x[y] \rangle_x \,, \tag{31}$$

where $H[y]$ is the entropy of all measurements y, $H_x[y]$ is the entropy of $p(y|x)$ for a specific choice of sequence x, and $\langle \cdot \rangle_x$ indicates averaging over all sequences x. In this dataset, the measurement values were computed using

$$y = \log_2 \left[\frac{c_s + 1}{c_i + 1} \right] , \tag{32}$$

where c_i is the input read count and c_s is the selected read count. $H[y]$ was estimated using the KNN estimator.[47] We estimated the uncertainty in y by propagating errors expected due to Poisson fluctuations in read counts, which gives

$$\delta y = \log_2(e)\sqrt{\frac{1}{c_s+1}+\frac{1}{c_i+1}}\,. \tag{33}$$

Then, assuming $p(y|x)$ to be approximately Gaussian, we find the corresponding conditional entropy to be

$$H_x[y] = \frac{1}{2}\log_2(2\pi e\,\delta y^2)\,. \tag{34}$$

These $H[y]$ and $H_x[y]$ values were then used in **Eq. 31** to estimate I_{int}. This should provide an upper bound on the true value of I_{int} because uncertainty in y must be at least that expected under Poisson sampling of reads. We note, however, that the use of linear error propagation and the assumption that $p(y|x)$ is approximately Gaussian complicate this conclusion. Also, when applied to MPSA data, this method yielded an upper bound of 0.96 bits. We believe this value is likely to be far higher than the true value of I_{int}, and that this mismatch probably resulted from read counts in the MPSA data being over-dispersed.

To compute the lower bound on I_{int} for GB1 data (**Fig. 3d**) we used the predictive information I_{pre} (on test data) of a GE regression model having a blackbox G-P map. This provides a lower bound because $I_{\text{int}} \geq I_{\text{pre}}$ for any model (when evaluated on test data) due to the Data Processing Inequality and the Markov Chain nature of the dependencies $y \leftarrow x \rightarrow \phi$ in **Fig. 2e**.[20,29]

To compute a lower bound on I_{int} for MPSA data (**Fig. 5c**), we leveraged the availability of replicate data in Wong et al..[36] Let y and y' represent the original and replicate measurements obtained for a sequence x. Because $y \leftarrow x \rightarrow y'$ forms a Markov chain, $I[x;y] \geq I[y;y']$.[29] We therefore used an estimate of $I[y;y']$, computed using the KSG method,[46,48] as the lower bound for I_{int}.

Uncertainties in kNN estimates

MAVE-NN quantifies uncertainties in $H[y]$ and $I[y;\phi]$ using multiple random samples of half the data. Let $\mathcal{D}_{100\%}$ denote a full dataset, and let $\mathcal{D}_{50\%,r}$ denote a 50% subsample (indexed by r) of this dataset. Given an estimator $E(\cdot)$ of either entropy or mutual information, as well as the number of subsamples R to use, the uncertainty in $E(\mathcal{D}_{100\%})$ is estimated as

$$\delta E(\mathcal{D}_{100\%}) = \frac{1}{\sqrt{2}} \mathrm{std} \left[\left\{ E\left(\mathcal{D}_{50\%,r}\right) \right\}_{r=0}^{R-1} \right] . \qquad (35)$$

MAVE-NN uses $R = 25$ by default. We note that computing such uncertainty estimates substantially increases computation time, as $E(\cdot)$ needs to be evaluated $R + 1$ times instead of just once. We also note that bootstrap resampling[51,52] is often inadvisable in this context, as it systematically underestimates $H[y]$ and overestimates $I[y;z]$.

Datasets

For the GB1 DMS dataset of Olson et al.,[33] measurements were computed using

$$y_n = \log_2 \frac{(c_n^{\text{out}}+1)/(c_{\text{WT}}^{\text{out}}+1)}{(c_n^{\text{in}}+1)/(c_{\text{WT}}^{\text{in}}+1)} ,$$

where c_n^{in} and c_n^{out} respectively represent the number of reads from the input and output samples (i.e., pre-selection and post-selection libraries), and $n = \text{WT}$ represents the 55 aa wildtype sequence, corresponding to positions 2-56 of the GB1 domain. To infer the model in **Fig. 3b** and to compute the information metrics in **Fig. 3c**, only double-mutant sequences with $c_n^{\text{in}} \geq 10$ were used; these represent 530,737 out of the 536,085 possible double mutants. For the models in **Figs. 3d-f**, y_n values for the 1045 single-mutant were also used in the inference procedure.

For the Aβ DMS data of Seuma et al.[34] and TDP-43 DMS data of Bolognesi et al.,[35] y_n values respectively represent nucleation scores and toxicity scores reported by the authors.

For the MPSA data of Wong et al.,[36] we used the data of library 1 replicate 1 obtained for the *BRCA2* minigene data. Measurements were computed as

$$y_n = \log_{10}\left[100 \times \frac{(c_n^{\mathrm{inc}}+1)/(c_{\mathrm{CONS}}^{\mathrm{inc}}+1)}{(c_n^{\mathrm{tot}}+1)/(c_{\mathrm{CONS}}^{\mathrm{tot}}+1)}\right],$$

where c_n^{inc} and c_n^{tot}respectively represent the number of barcode reads obtained from exon inclusion isoforms and from total mRNA, and $n = \mathrm{CONS}$ corresponds to the consensus 5´ss sequence CAG/GUAAGU. Corresponding PSI values were computed as $\mathrm{PSI}_n = 10^{y_n}$. Only sequences with $c_n^{\mathrm{tot}} \geq 10$ were used, representing 30,483 of the 32,768 possible sequences of the form NNN/GYNNNN.

For the lac promoter sort-seq MPRA data of Kinney et al.,[16] we used data from the "full-wt" experiment (available at https://github.com/jbkinney/09_sortseq). For the *xylE* promoter sort-seq MPRA data of Bellilveau et al.,[43] we used data kindly provided by the authors.

References

1. Kinney, J. B. & McCandlish, D. M. Massively parallel assays and quantitative sequence-function relationships. *Annu Rev Genom Hum G* **20**, 99-127 (2019).

2. Starita, L. M. *et al.* Variant interpretation: functional assays to the rescue. *Am J Hum Genetics* **101**, 315-325 (2017).

3. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801-807 (2014).

4. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* **15**, 453-468 (2014).

5. White, M. A. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* **106**, 165-170 (2015).

6. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159-164 (2015).

7. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).

8. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430-3431 (2011).

9. Alam, K. K., Chang, J. L. & Burke, D. H. FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol Ther-Nucleic Acids* **4**, e230 (2015).

10. Bloom, J. D. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* **16**, 168 (2015).

11. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol* **18**, 1-15 (2017).

12. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol* **20**, 183 (2019).

13. Niroula, A., Ajore, R. & Nilsson, B. MPRAscore: robust and non-parametric analysis of massively parallel reporter assays. *Bioinformatics* **35**, 5351-5353 (2019).

14. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol* **21**, 207 (2020).

15. Atwal, G. S. & Kinney, J. B. Learning quantitative sequence-function relationships from massively parallel experiments. *J Stat Phys* **162**, 1203-1243 (2016).

16. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* **107**, 9158-9163 (2010).

17. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-277 (2012).

18. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23**, 1908-1915 (2013).

19. Abadi, M. *et al.* TensorFlow: A Systems for Large-Scale Machine Learning. in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* (2016).

20. Kinney, J. B. & Atwal, G. S. Parametric inference in the large data limit using maximally informative models. *Neural Comput* **26**, 637-653 (2014).

21. Pokusaeva, V. O. *et al.* An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet* **15**, e1008079 (2019).

22. Kinney, J. B., Tkačik, G. & Callan, C. G. Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* **104**, 501-506 (2007).

23. Otwinowski, J. & Nemenman, I. Genotype to phenotype mapping and the fitness landscape of the *E. coli lac* promoter. *PLoS ONE* **8**, e61570 (2013).

24. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397-401 (2016).

25. Sailer, Z. R. & Harms, M. J. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205**, 1079-1088 (2017).

26. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc Natl Acad Sci USA* **115**, E7550-E7558 (2018).

27. Jones, M. C. & Faddy, M. J. A skew extension of the t-distribution, with applications. *J Roy Stat Soc B* **65**, 159-174 (2003).

28. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci USA* **111**, 3354-3359 (2014).

29. Cover, T. M. & Thomas, J. A. *Elements of information theory*. (Wiley, 2006).

30. Barber, D. & Agakov, F. The IM algorithm: a variational approach to information maximization. *Advances in neural information processing systems 16.* (2004).

31. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep variational information bottleneck. *arXiv:1612.00410* (2016).

32. Chalk, M., Marre, O. & Tkacik, G. Relevant sparse codes with variational information bottleneck. *arXiv:1605.07332* (2016).

33. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* **24**, 2643-2651 (2014).

34. Seuma, M., Faure, A., Badia, M., Lehner, B. & Bolognesi, B. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *eLife* **10**, e63364 (2021).

35. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. *Nat Commun* **10**, 4162 (2019).

36. Wong, M. S., Kinney, J. B. & Krainer, A. R. Quantitative activity profile and context dependence of all human 5´ splice sites. *Mol Cell* **71**, 1012-1026.e3 (2018).

37. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**, 116-124 (2005).

38. Sherman, M. S. & Cohen, B. A. Thermodynamic state ensemble models of cis-regulation. *Plos Comput Biol* **8**, e1002407 (2012).

39. Wong, F. & Gunawardena, J. Gene Regulation in and out of equilibrium. *Annu Rev Biophys* **49**, 199-226 (2020).

40. Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol Biol Evol* **35**, 2345-2354 (2018).

41. Manhart, M. & Morozov, A. V. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc Natl Acad Sci USA* **112**, 1797-1802 (2015).

42. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci USA* **116**, 16367-16377 (2019).

43. Belliveau, N. M. *et al.* Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc Natl Acad Sci USA* **115**, 201722055 (2018).

44. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol* **14**, 638-650 (2016).

45. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272-2274 (2020).

46. Steeg, G. V. Non-Parametric Entropy Estimation Toolbox (NPEET). https://www.isi.edu/~gregv/npeet.html (2014).

47. Vasicek, O. A test for normality based on sample entropy. *J Roy Stat Soc B* **38**, 54-59 (1976).

48. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys Rev E* **69**, 066138 (2004).

49. Gao, S., Steeg, G. V. & Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. *In Artificial intelligence and statistics (pp. 277-286). PMLR.*

50. Sharpee, T., Rust, N. C. & Bialek, W. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput* **16**, 223-250 (2004).

51. Efron, B. Bootstrap methods: another look at the jackknife. *Ann Stat* **7**, 1-26 (1979).

52. Efron, B. & Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* **1**, 54-75 (1986).