PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY
APPLIED MATHEMATICS OPEN ACCESS

Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power

Peter D. Tonner[a,1], Abe Pressman[b], and David Ross[b]

Edited by Wing Hung Wong, Stanford University, Stanford, CA; received August 3, 2021; accepted March 4, 2022

Large-scale measurements linking genetic background to biological function have driven a need for models that can incorporate these data for reliable predictions and insight into the underlying biophysical system. Recent modeling efforts, however, prioritize predictive accuracy at the expense of model interpretability. Here, we present LANTERN (landscape interpretable nonparametric model, https://github.com/usnistgov/lantern), a hierarchical Bayesian model that distills genotype–phenotype landscape (GPL) measurements into a low-dimensional feature space that represents the fundamental biological mechanisms of the system while also enabling straightforward, explainable predictions. Across a benchmark of large-scale datasets, LANTERN equals or outperforms all alternative approaches, including deep neural networks. LANTERN furthermore extracts useful insights of the landscape, including its inherent dimensionality, a latent space of additive mutational effects, and metrics of landscape structure. LANTERN facilitates straightforward discovery of fundamental mechanisms in GPLs, while also reliably extrapolating to unexplored regions of genotypic space.

interpretability | machine learning | genotype–phenotype landscape | epistasis

Significance

A critical challenge across biological disciplines is understanding how mutations in genetic sequence change downstream biological function. Measurements linking genotype to phenotype are fundamentally limited, however, because the space of possible genetic sequences is massive. So, practitioners increasingly rely on machine learning models to facilitate discoveries and predict new phenotypes. Most machine learning models make a trade-off between predictive accuracy and interpretability, that is, the degree to which we can understand a model's predictions. Here, we develop a machine learning approach called LANTERN (landscape interpretable nonparametric model) that is fully interpretable. Importantly, LANTERN's predictive accuracy equals or exceeds alternative approaches across a broad benchmark of genotype–phenotype measurements, representing state-of-the-art prediction connecting genotype to phenotype.

Genotype–phenotype landscapes (GPLs) characterize the relationship between a gene's mutations and its function. Driven by reduced sequencing costs and growing experimental throughput, the scale of available GPL measurements has increased dramatically, with recent measurements sampling as many as 10^4 to 10^7 distinct genotypes (1). These measurements play an expanding role in understanding biological variation, with applications from engineering to epidemiology (2, 3). Despite their importance, however, even the highest-throughput experimental measurements cannot overcome the massive combinatorial size of GPLs. For example, meaningful changes in engineered function or virulence are often the result of three or more mutations. For a typical protein, there are approximately 10^{11} potential triple mutants, so a large-scale GPL measurement on the order of 10^4 to 10^7 observations can only sample a tiny fraction of the relevant mutational space around the native sequence. Since this makes complete experimental coverage of GPLs unrealistic, a full understanding of landscape spaces can only come from estimates of unmeasured genotype–phenotype combinations, using models and their predictions.

Predicting phenotypes for genotypes with multiple mutations is challenging because the effect of each mutation depends on which other mutations are present. So, the phenotypic change due to multiple mutations is not simply the sum of the changes from each single mutation (4). This context dependence, referred to as epistasis, arises from many biological mechanisms and has motivated diverse modeling approaches. For example, specific pairs of protein residues can have strong nonadditive interactions due to their close physical proximity in the folded protein structure, and GPL models can represent this effect through pair-wise interaction coefficients (5). Alternatively, even when individual mutations cause additive changes to an underlying biophysical parameter (e.g., folding free energy), the measured phenotype can be a nonlinear function of that parameter (6–8). GPL models that directly represent these epistatic mechanisms have the advantage of being interpretable (9). Interpretability comes from the correspondence between model components and biological mechanisms and can provide practitioners with an intuitive understanding of each component of the larger model. Additionally, clear explanations of how the model generates each prediction increases interpretability by aiding the diagnosis of inaccurate predictions, increasing the reliability of decisions made from predictions, and increasing trust in the model.

Existing interpretable GPL models often have limitations on their predictive accuracy (10). So, recent efforts to model large-scale GPL measurements have, instead, relied on deep neural network (DNN) architectures, due to their superior predictive performance (11–15). DNNs make predictions through complicated cascades of nonlinear computation with large numbers of parameters that lack direct biological motivation, operating, essentially, as a black box. To make DNNs more interpretable, post hoc explainability

Author contributions: P.D.T. and D.R. designed research; P.D.T. and D.R. performed research; P.D.T. analyzed data; and P.D.T., A.P., and D.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: peter.tonner@nist.gov.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2114021119/-/DCSupplemental.

Published June 21, 2022.

techniques estimate the relevant factors of DNN predictions (16). By necessity, however, these methods only approximate the actual important features, and incorrect conclusions can be drawn by mischaracterization of the details in the explanation (17). Additionally, these techniques often apply to individual predictions rather than generalizing the whole model (18). So, explaining these black-box models does not straightforwardly scale to the billions of predictions that may be of interest for GPL-dependent research. Overall, DNNs provide a useful approach for maximizing predictive accuracy but force users to make compromises on model interpretability.

Here, we address the conflict between predictive accuracy and interpretability by developing a GPL modeling approach called LANTERN (landscape interpretable nonparametric model). LANTERN learns interpretable models of GPLs by finding a latent, low-dimensional space where mutational effects combine additively. LANTERN then captures the nonlinear effects of epistasis through a multidimensional, nonparametric Gaussian process (GP) model. This approach generalizes the modeling of global epistatic relationships in two ways. First, we allow for multiple different biophysical mechanisms to influence biological function by modeling multiple latent dimensions (19, 20). Second, we avoid any strong assumptions on the shape of the nonlinear surface, instead learning it directly from the data. In a benchmark across multiple protein GPLs, LANTERN achieves predictive accuracy as good as or better than existing models, including DNNs. Importantly, LANTERN automatically provides interpretable explanations of these predictions. LANTERN therefore remains highly interpretable while maximizing predictive power, and can thus increase the coverage of GPLs by orders of magnitude while simultaneously distilling complex landscapes into their fundamental structure.

Results

Constructing Interpretable Models of GPLs. LANTERN takes, as input data, a combination of genotypes and their measured phenotypes (Fig. 1*A*). A LANTERN model of these data has two key components. First, LANTERN decomposes genetic mutations onto a latent mutational effect space, where individual mutations are each represented by a vector (e.g., $\vec{\mathbf{z}}^{(i)}$ for mutation i; Fig. 1*C*). Vectors provide interpretable comparisons between mutations, with two vectors in the same direction implying similar function and vector magnitude representing strength of effect. Importantly, we assume that individual mutations combine additively, represented as vector addition in the latent effect space. This additive structure recapitulates many biophysical phenomena. For example, individual mutations often additively impact the thermodynamic stability of protein folding (21). GPL measurements, however, frequently target a phenotype that is a nonlinear function of these additive effects. For example, many protein phenotypes remain robust to small decreases in folding stability but rapidly diminish beyond a certain threshold (8). So, the second component of LANTERN relates measured phenotypes to the combination of latent mutational effects through a smooth,

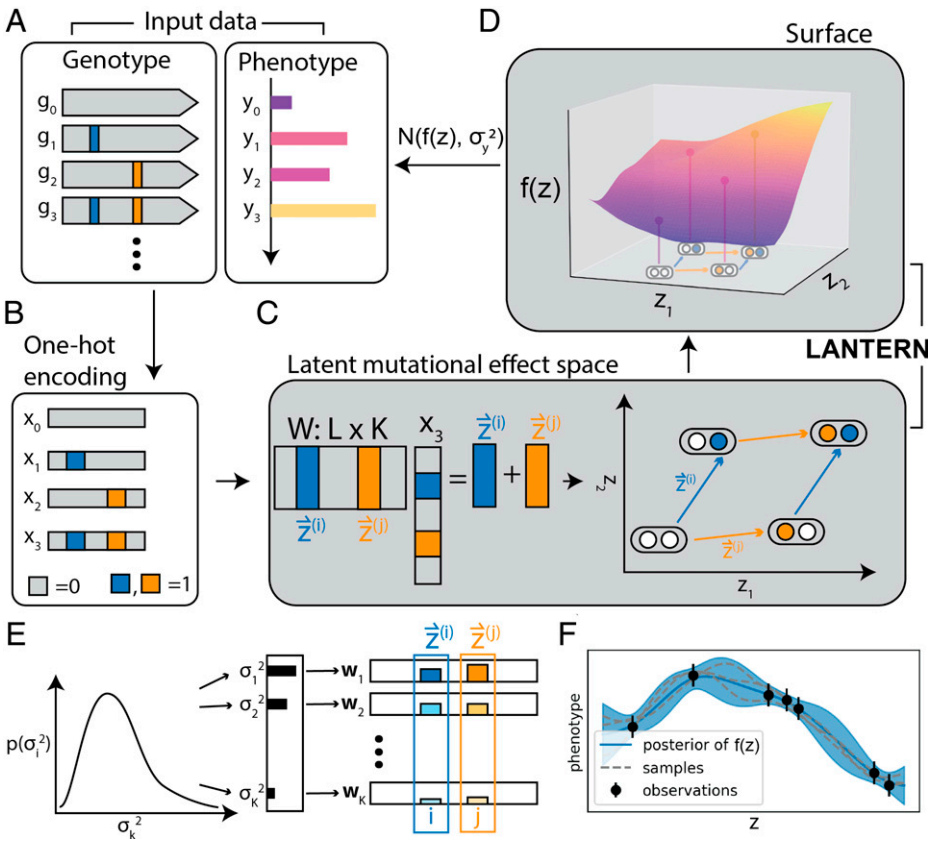

Fig. 1. Interpretable modeling of GPLs. (*A*) LANTERN takes, as input data, genotypes with a corresponding measured phenotype. (*B*) LANTERN converts the genotype of each variant k into a one-hot encoded vector $x_k \in \{0,1\}^p$, where p is the total number of mutations observed across all variants, and $x_{kl} = 1$ implies the presence of mutation l in variant k ($x_{kl} = 0$ otherwise). (*C*) LANTERN predicts the position of variant k in the latent mutational effect space as a linear combination of mutation effect vectors with an unknown matrix W that is learned from the data. This formulation represents the assumption that mutations combine additively in the latent space. Additionally, individual mutations have an interpretable representation in the model in the form of their mutational effect vector ($\vec{\mathbf{z}}^{(i)}$, $\vec{\mathbf{z}}^{(j)}$, etc.). Computing the location of each variant k in the latent mutational effect space then requires simply adding the mutational effect vectors for mutations present in the variant. (*D*) Phenotypes are a nonlinear function, $f(\mathbf{z})$, of the latent mutational effects, $\mathbf{z}$. This results in a nonlinear relationship between genotype and phenotype. Additionally, the observed phenotypes are assumed to be measured with iid Gaussian noise with unknown variance σ_y^2. (*E*) Dimensionality of the model is controlled through a hierarchical prior on latent mutational effect dimensions. The variance of individual dimensions in the latent space (e.g., the rows of W) have a prior skewed to small values. This results in the majority of dimensions effectively shrinking toward zero variance, and the smallest dimensionality necessary to explain the data is recovered. (*F*) Nonlinear surfaces $f(\mathbf{z})$ are modeled with GP priors. An example one-dimensional GP posterior of $f(\mathbf{z})$ is shown, fit to observations (black dots, bars show observation 95% CI), where the solid blue line is the posterior predictive mean, and the shaded region is the 95% credible region. Dashed lines show example function draws from this posterior.

nonlinear surface: $f(\mathbf{z})$ (Fig. 1*D*). Phenotype measurements are inherently noisy, so we assume $f(\mathbf{z})$ is measured indirectly with independent identically distributed (iid) Gaussian noise.

These two components, the latent mutational effect space $\mathbf{z}$ and nonlinear surface $f(\mathbf{z})$, make LANTERN interpretable. Both components clearly correspond to general biophysical mechanisms often seen in GPL measurements, and therefore have intuitive explanations for their role in the model. Additionally, they explain the predictions made by LANTERN straightforwardly: Each prediction results by first combining the latent mutational effects and then transforming through the surface $f(\mathbf{z})$. Notably, LANTERN makes no assumptions about the specific biophysical mechanisms driving the complexity of GPLs in either the latent mutational space or the nonlinear response surface. Instead, LANTERN learns these relationships directly from the GPL data after assuming the general structure outlined above. Therefore, while each component learned by LANTERN may have a direct correspondence to the biophysical mechanisms of the systems, these connections must be determined through additional analysis.

For application to GPL data, we implemented LANTERN as a hierarchical Bayesian model. As part of this hierarchy, we employed an approach for determining the dimensionality of any GPL directly from the data. Specifically, we learn a relative scale between each dimension of the latent mutational effect space ($\mathbf{z}$) in the form of each dimension's variance. Higher variance results in larger mutational effects across the corresponding dimension, while dimensions with variance close to zero are effectively removed from the model. The variance of each dimension therefore reflects its relative impact on the model, which we refer to as the relevance of the dimension.

To ensure that LANTERN learns the minimal number of dimensions necessary to explain the data, we adapted the prior on variance from a Bayesian treatment of principal components analysis (PCA; Fig. 1*E*) (22). With this prior, we enforce assumptions similar to those used for PCA: The latent mutational effect dimensions should be uncorrelated, and there should be decreasing variance (i.e., relevance) with each dimension. This prior also has the added advantage of enforcing the assumption that a relatively small number of latent dimensions will be necessary to explain the data, ensuring that dimensions not supported by the data are removed. LANTERN therefore learns the minimal number of dimensions necessary to explain the data, which we refer to as the dimensionality of the GPL.

To learn the nonlinear relationship between latent mutational effects and measured phenotypes, we placed a GP prior on the surface $f(\mathbf{z})$ (Fig. 1*F*) (23). GPs learn the distribution over possible functions that best explain the data, rather than specifying a parametric form to the underlying relationship between $\mathbf{z}$ and the observed phenotypes governed by f (*SI Appendix*, Fig. S1). This ensures that LANTERN can learn the surface $f(\mathbf{z})$ of any GPL automatically from the data rather than relying on expert knowledge to choose an appropriate parametric form (6, 24). To learn both components of LANTERN for different GPLs, we apply stochastic variational methods that make inference tractable and scalable to millions of observations (25).

LANTERN Learns Biophysical Mechanisms. To determine how well LANTERN can discover true biophysical mechanisms from GPL data, we evaluated its performance on simulated data from an analytic model of protein allostery (26). Allosteric proteins regulate cellular processes in response to changes in ligand concentration, and the analytic model describes this response as a function of the underlying biophysical constants such as ligand–protein binding constants and free-energy differences between protein states. In the analytic model, the biophysical constants determine three phenotypic parameters that characterize the allosteric dose–response curve: the basal and saturating transcription levels (G_0 and G_∞, respectively), and the concentration of ligand where transcription is halfway between minimum and maximum (EC_{50}).

Briefly, we constructed synthetic GPL measurements by simulating a set of mutations that additively shift the wild-type biophysical constants (Fig. 2*A* and *SI Appendix*, section 1). We then randomly combined these mutations to simulate individual protein variants with a mean number of mutations matching that of an existing allosteric landscape dataset (4.38 mutations per variant, on average) (2). For each variant, the resulting perturbed biophysical constants determined the allosteric dose–response parameters via the analytic model. We trained LANTERN using only the variant genotypes and resulting dose–response parame-

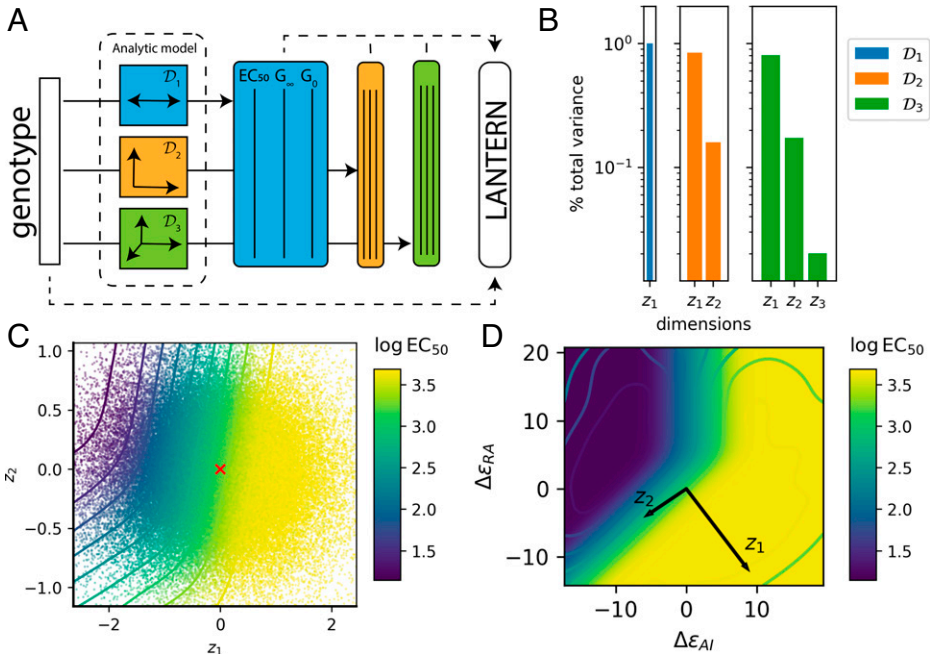

Fig. 2. Recovering biophysical landscapes and dimensionality. (*A*) Simulated GPLs of varying dimensionality. Three simulated datasets were generated (datasets $\mathcal{D}_1$ to $\mathcal{D}_3$), with latent mutation effects on one, two, or three underlying biophysical constants. The biophysical constants determine the allosteric dose–response phenotypes EC_{50}, G_∞, and G_0 based on a biophysical model of allosteric protein function (see *SI Appendix*, section 1 for details). (*B*) Percentage of total mutational effect variance for each dimension learned by LANTERN for $\mathcal{D}_1$ to $\mathcal{D}_3$. For each dataset, only variances up to the dimensionality learned by LANTERN are shown (*SI Appendix*, Fig. S2). (*C*) Learned two-dimensional surface $f(\mathbf{z})$ for EC_{50} from dataset $\mathcal{D}_2$. Contours show the posterior mean of $f(\mathbf{z})$, shading shows the relative variance of $f(\mathbf{z})$, and scatter points are observations positioned by their latent $\mathbf{z}$ value and colored by their observed EC_{50} value Red cross marks the wild-type origin ($\mathbf{z}_{wt} = \mathbf{0}$). (*D*) Rotation of the $f(\mathbf{z})$ surface shown in *C* to align with the underlying biophysical constants $\Delta\epsilon_{AI}$ and $\Delta\epsilon_{RA}$. The color map shows the true EC_{50} surface, and the contours show the posterior mean of $f(\mathbf{z})$ from LANTERN. The black vectors mark the rotation of the $\mathbf{z}_1$ and $\mathbf{z}_2$ dimensions to the native biophysical space.

ters, without access to the underlying biophysical constants or the analytical model.

To test LANTERN's ability to correctly identify GPL dimensionality, we simulated three different GPL datasets. In these datasets, we controlled the true dimensionality by varying the number of perturbed biophysical constants from one to three (datasets $\mathcal{D}_1$ to $\mathcal{D}_3$, respectively). To extract the dimensionality learned by LANTERN, we quantified how much each dimension increases the log-likelihood of the data (see *Materials and Methods*). If the log-likelihood increases when an additional dimension is included in the model, then that dimension is important for explaining the data. So, we count the total number of dimensions where evidence increases, and call this the dimensionality of the landscape learned by LANTERN. With this procedure, we correctly recovered the true dimensionality of the simulated biophysical landscapes as well as higher-dimensional simulated datasets (Fig. 2*B* and *SI Appendix*, section 2 and Figs. S2 and S3). So, LANTERN recovers the true GPL dimensionality automatically from data.

To determine whether LANTERN also learns representations of GPLs that agree with the underlying biophysical mechanisms, we compared the model learned by LANTERN to the true biophysical system. LANTERN learned latent mutational effects that strongly correlated with true biophysical constants, despite the fact that LANTERN had no access to this information (*SI Appendix*, Fig. S4). Furthermore, LANTERN learned smooth, interpretable surfaces that accurately predict the allosteric phenotype of each simulated landscape (Fig. 2*C* and *SI Appendix*, Fig. S5). Using the correlation between latent mutational effects learned by LANTERN and the true biophysical constants in the simulation, we rotated and rescaled the space of latent mutational effects to directly compare $f(\mathbf{z})$ to the true biophysical surface (Fig. 2*D* and *SI Appendix*, Fig. S6). In regions of latent mutational effect space with large sampling coverage, the rotated and rescaled surface, $f(\mathbf{z})$, matches the surface from the analytical model exactly (Fig. 2*D*). As sampling density decreases, the deviation between predicted and true surfaces increases (*SI Appendix*, Fig. S6). LANTERN therefore approximates a rotated and scaled version of the true biophysical model in regions with high experimental coverage, but balances this against uncertainty in more sparsely sampled regions of parameter space. Overall, this demonstrates that LANTERN recovered the underlying dimensionality and biophysical mechanisms of our simulations de novo from data, without additional domain specific knowledge.

LANTERN Outperforms Alternative Predictive Methods. To compare LANTERN's predictive accuracy to alternative models with real GPL data, we analyzed three published large-scale GPL datasets (*SI Appendix*, Table S1): the fluorescence brightness of green fluorescent protein [avGFP (12)], the dose–response curves of an allosteric transcription factor [LacI (2)], and the joint phenotypes of ACE2 binding affinity and structural stability for the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein [SARS-CoV-2 (3)]. Each dataset samples a large number of genotypes, between ∼50,000 and ∼170,000 distinct genotypes and ∼1,800 to ∼4,000 unique mutations, making them ideal candidates for evaluating predictive performance of different GPL models. We evaluated GPL model performance on each dataset through predictive accuracy (R^2) cross-validated over 10 random splits of the data into test and training sets.

To provide the broadest comparison of LANTERN to alternatives, we tested multiple modeling approaches with each dataset, including interpretable linear and nonlinear models as well as black-box DNNs. In all but one instance, LANTERN equaled or outperformed all alternative approaches with respect to 10-fold cross-validated prediction accuracy (Fig. 3*A*). For avGFP, all of the tested models except for a simple linear approach have comparable predictive accuracy. In this case, simpler models than LANTERN are sufficient to make accurate predictions. Importantly, LANTERN achieves the same predictive accuracy as these approaches, meaning that LANTERN does not overfit the data despite its capacity to learn more complex models. For SARS-CoV-2, LANTERN has higher predictive accuracy than all alternatives other than a dense feed-forward neural network, which performed equally well. Additionally, LANTERN outperforms all other approaches in predicting LacI effective concentration, 50% (EC_{50}). In one case, the G_∞ of LacI, a one-dimensional DNN predicts out-of-sample phenotypes more accurately than LANTERN. The G_∞ phenotype poses unique challenges, due to the complex relationship between observed G_∞ and phenotype uncertainty: Values of G_∞ most distinct from the wild-type are also the most uncertain (2). This may explain the slight decrease in predictive accuracy for LANTERN in this case. Notably, when adapting the variational loss to more robustly handle highly uncertain measurements (27), we found a substantial increase in LANTERN's prediction accuracy for G_∞ (*SI Appendix*, Fig. S7). Finally, predictive performance did not increase with deeper DNN models, suggesting we are evaluating the best possible performance for our chosen DNN architecture (*SI Appendix*, Fig. S8). Overall, LANTERN provides the most accurate predictions of any single method across a broad benchmark of GPL datasets and therefore achieves state of the art predictive accuracy.

To determine the impact of sample size on LANTERN performance, we evaluated cross-validated accuracy as a function of dataset size. As a Bayesian model, LANTERN straightforwardly balances the complexity of the model against the available GPL data. Consequently, the same training procedure for LANTERN applies regardless of dataset size. So, LANTERN maintained higher predictive accuracy compared to alternative approaches regardless of the number of observations (*SI Appendix*, Fig. S9). Importantly, LANTERN does not underperform the simpler linear model with decreased data, reflecting LANTERN's ability to scale down to relatively small datasets without overfitting. LANTERN even scales down to small-scale measurements with only five to seven mutations (*SI Appendix*, Fig. S10). The improvement of LANTERN over simpler models also increases when predicting larger combinations of mutations, which is critical for many applications (*SI Appendix*, Fig. S11).

LANTERN Quantifies Predictive Uncertainty. As a Bayesian modeling approach, LANTERN directly quantifies the uncertainty of its predictions. In order for these uncertainties to provide useful information, however, they must accurately reflect the degree of certainty that should be placed in each prediction when tested against real data (28). We therefore compared the prediction uncertainties reported by LANTERN to its actual prediction error. Prediction error typically increases with increased mutational distance from the wild-type genotype (Fig. 3*B*). This is due, at least in part, to the decrease in measurement coverage for variants with more mutations. A well-calibrated prediction uncertainty should similarly increase with mutational distance to reflect the decrease in available data. LANTERN's predictions directly reflect this phenomenon, with uncertainty growing proportional to the prediction error (Fig. 3*C*). LANTERN does appear to underestimate the overall true error rate, however, a known issue with variational methods (29). The difference between observed and actual error appears

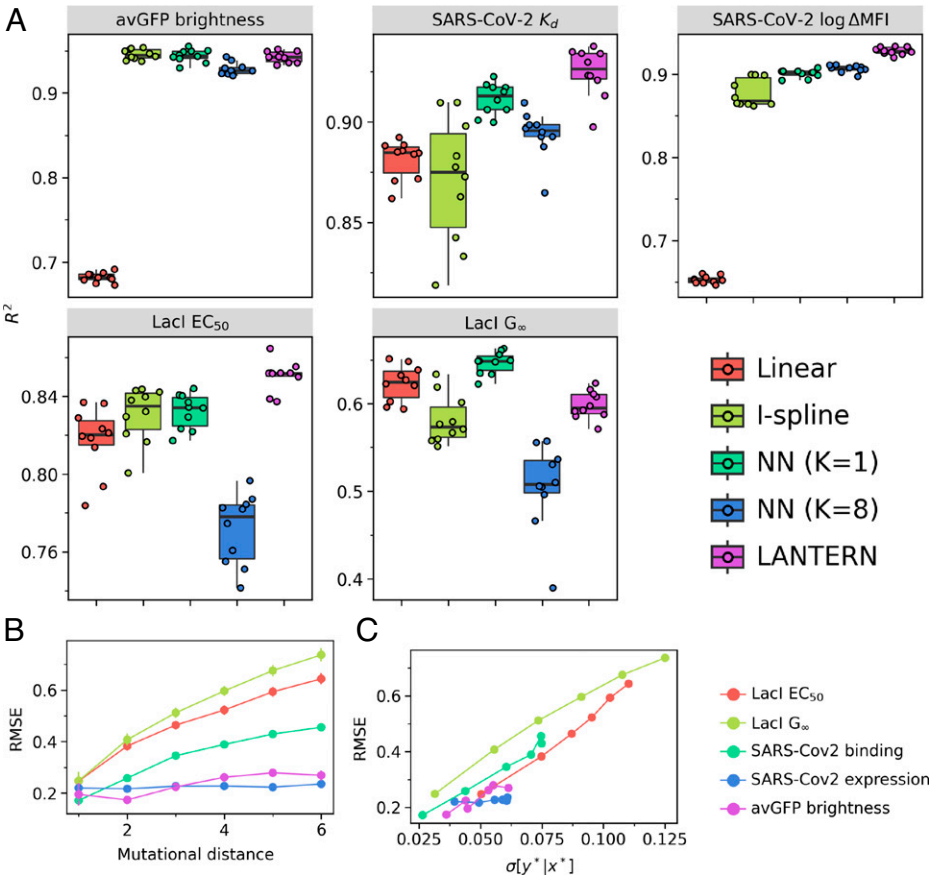

Fig. 3. LANTERN equals or outperforms alternative models in predictive accuracy. (*A*) Ten-fold cross-validated predictive accuracy, as determined by R^2 for different models across high-throughput GPL measurements. Box plots summarize the distribution of R^2 values: median (center line), interquartile range (hinges), and farthest point within 1.5 of the interquartile range of the each hinge (whiskers). Scatter points mark the R^2 value for individual folds. (*B*) LANTERN root-mean-squared error (RMSE) as a function of mutational distance from the wild-type sequence. (*C*) RMSE versus the average posterior predictive uncertainty from LANTERN. Plotted points are for different mutational distances from the wild type. In *B* and *C*, each point is the bootstrapped mean cross-validated RMSE, and error bars represent the bootstrap estimate of the 95% CI.

linear, however, so a possible correction could be learned independently for each dataset. Additionally, the posterior predictive intervals are miscalibrated, further reflecting the overconfidence of the model (*SI Appendix*, Fig. S12). Generally, LANTERN provides a degree of confidence for each prediction, but improvements to uncertainty quantification would benefit downstream applications.

LANTERN Provides Interpretable Models of GPLs. While LANTERN generates reliable predictions, it also straightforwardly explains these predictions through its interpretable components: an additive latent mutational effect space ($\mathbf{z}$) and nonlinear surface ($f(\mathbf{z})$). Here, we demonstrate how to leverage this interpretability in the analysis of the three large-scale GPL datasets: avGFP, LacI, and SARS-CoV-2. Across these datasets, the latent dimensionality learned by LANTERN ranged from three to five (Fig. 4*A* and *SI Appendix*, Fig. S13). We focus our analysis on the two most relevant dimensions of each dataset, because additional dimensions capture an exponentially decreasing fraction of the total variation in the latent space of mutational effects, and correspondingly represent more minor details of each landscape (*SI Appendix*, Figs. S14–S18).

avGFP. For avGFP, the surface $f(\mathbf{z})$ contains a sharp boundary along the most relevant dimension, $\mathbf{z}_1$, that divides the latent mutational effect space into two regions: one with near wild-type fluorescent levels, and a second with complete loss of fluorescence (Fig. 4*B*). Maximum brightness occurs in a region near (but not centered on) the wild type (Fig. 4*C*). Two additional local maxima were identified in regions not centered on the wild type, but these are potentially artifacts arising from irregularities in a small number of variants (*SI Appendix*, Figs. S19 and S20). The majority of mutations decrease fluorescence, because their mutational effect vectors point toward a large region of decreased fluorescence (Fig. 4*D*). Previous analysis found a single neural network neuron accurately predicted avGFP brightness, with the neuron possibly representing the effect of each mutation on structural stability (12). However, discovery of this association depended on fixing the dimensionality of the neural network. In contrast, LANTERN learns this directly from data, finding a primary axis of decreased fluorescence along $\mathbf{z}_1$.

To further understand how LANTERN models avGFP fluorescence, we analyzed mutations from engineered variants of blue fluorescent protein (BFP) (30). LANTERN predicts the foundational mutation of all BFP variants, Y66H, to decrease brightness, as expected from an assay to detect green fluorescence (Fig. 4*E*). However, improved variants of BFP include additional mutations that increase the blue fluorescence, possibly through increased folding stability (31). Interestingly, LANTERN predicts nearly all of these mutations to similarly increase wild-type avGFP fluorescence, as they point toward the maximum predicted brightness (Fig. 4*C*). According to LANTERN, then, these mutations may generally improve brightness independent of fluorescent color, possibly through improved structural stability.

LacI. Tack et al. (2) measured over 47,000 LacI variants with Hill equation–like dose–response curves. We modeled the landscape of these response curves as a multivariate phenotype of EC_{50} and G_∞ for each variant. From LANTERN, the LacI GPL has a dimensionality of three, and the majority of mutational effects closely align with the $\mathbf{z}_1$ axis (Fig. 4 *A* and *F*–*H*).

To interpret the latent space learned by LANTERN, we compared the effects of mutations in the two most relevant latent dimensions ($\mathbf{z}_1$ and $\mathbf{z}_2$) with the expectations from an analytic biophysical model for LacI dose–response (26). Increases in $\mathbf{z}_1$ correspond to a decreased EC_{50} and a slightly increased G_∞. Conversely, decreases in $\mathbf{z}_1$ correspond to an increased EC_{50} and

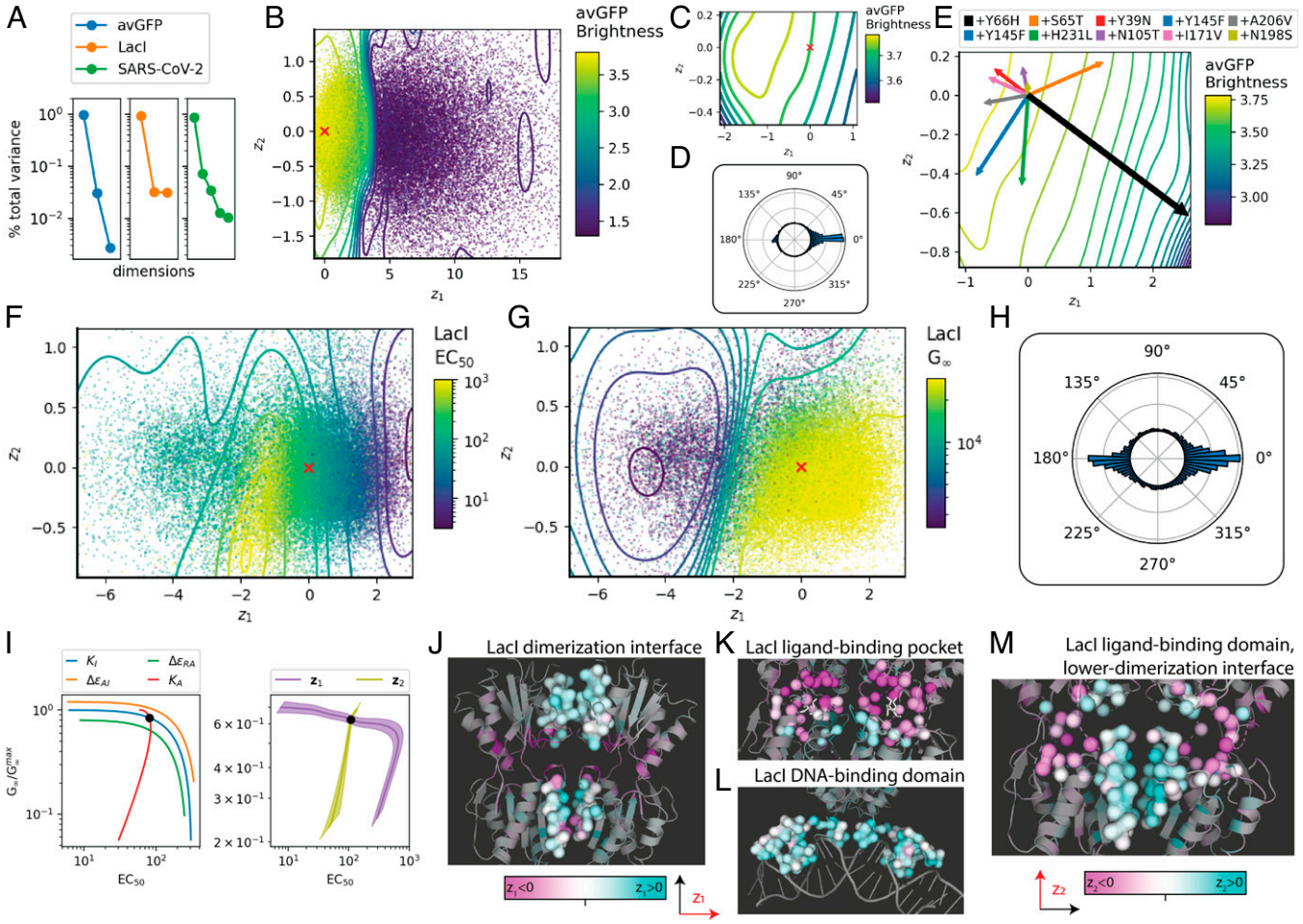

Fig. 4. Interpretable models of avGFP brightness and LacI dose–response. (*A*) Percentage of total mutational effect variance for each dimension in each large-scale GPL measurement. (*B*) Learned avGFP brightness surface $f(\mathbf{z})$ along the two highest relevance dimensions learned by LANTERN ($\mathbf{z}_1$ and $\mathbf{z}_2$). Contours show equal values of the posterior mean of $f(\mathbf{z})$, scatter points show the learned posterior mean of observations colored by their observed brightness, and red cross marks the wild-type origin ($\mathbf{z}_{wt} = \mathbf{0}$). (*C*) Focused region of the fluorescence surface, highlighting the region of maximal brightness off-center from the wild type (red cross). (*D*) The distribution of latent, single-mutant directions learned by LANTERN across $\mathbf{z}_1$ and $\mathbf{z}_2$; 78% (1,476/1,879) of mutations increase along $\mathbf{z}_1$. (*E*) Mutations in blue fluorescent variants of avGFP. BFP is created from avGFP by the Y66H mutation (black vector). Additional mutations that increase BFP blue fluorescence are predicted to similarly increase avGFP green fluorescence (with one exception: S65T). These mutations may then have general structural stabilizing effects independent of fluorescent color. (*F* and *G*) The learned joint surface of LacI EC_{50} (*F*) and G_∞ (*G*). Contours show constant posterior mean of $f(\mathbf{z})$, scatter points are the posterior mean $\mathbf{z}$ value of variants in the dataset colored by their observed phenotype values, and red cross marks the wild-type origin. (*H*) The distribution of latent, single-mutant effect directions across $\mathbf{z}_1$ and $\mathbf{z}_2$ learned by LANTERN for LacI. The majority of mutational effects (1,374/2,510 mutations within $\measuredangle \leq 25^\circ$ of $\mathbf{z}_1$) lie near the positive (804 mutations) or negative (570) $\mathbf{z}_1$ axis. (*I*) Comparison of relationship between EC_{50} and G_∞ when varying biophysical constants in an analytic model of allostery (*Left*) and over $\mathbf{z}_1$ and $\mathbf{z}_2$ (*Right*). G_∞ is normalized by its maximum value, G_∞^{max}, because that is the value reported from the analytic biophysical model. Each line represents the joint predicted value of EC_{50} and G_∞ as the corresponding biophysical constant or latent dimensional effect ($\mathbf{z}_1$ or $\mathbf{z}_2$) is varied. Plots for $\mathbf{z}_1$ and $\mathbf{z}_2$ show predicted mean as solid line and 95% credible interval as a shaded region. For visual clarity, the lines for $\Delta\epsilon_{AI}$ and $\Delta\epsilon_{RA}$ are shifted above or below their true value, respectively, because they lie along the same values as the curve for K_I. (*J*–*M*) Association of latent mutational effects in $\mathbf{z}_1$ (*J*–*L*) and $\mathbf{z}_2$ (*M*) to different regions of the LacI protein. Highlighted regions are shown as connected surfaces in each panel. Residues are colored by their average posterior mean for the corresponding latent effect dimension. Highlighted regions are the dimerization interface (*J*), ligand binding pocket (*K*), DNA-binding domain (*L*), and a mixture of ligand-binding and lower-dimerization interface (*M*). All structures represent the active, DNA-bound state, with the exception of *K*, which shows the inactive, ligand-bound state. In each panel, two LacI proteins are shown as a dimer. Surfaces in *B*, *F*, and *G* show the latent space centered on the central 99% of observations and at a cross-section through the $\mathbf{z}_1$–$\mathbf{z}_2$ plane. For additional views of the surface, see *SI Appendix*, Figs. S14 and S18.

a sharply decreased G_∞ (Fig. 4 *F*, *G*, and *I*). This relationship between EC_{50} and G_∞ along $\mathbf{z}_1$ is consistent with changes to three biophysical constants of the analytic model (which are indistinguishable with respect to their effects on EC_{50} and G_∞; Fig. 4*I*): the allosteric constant ($\Delta\epsilon_{AI}$), the DNA operator affinity ($\Delta\epsilon_{RA}$), or the ligand binding affinity of the inactive (nonoperator binding) state (K_I). So, we can interpret mutations causing changes in $\mathbf{z}_1$ as free-energy changes related to those three biophysical constants. Changes along $\mathbf{z}_2$ have a different effect: Increases in $\mathbf{z}_2$ correspond to a decreased G_∞ at roughly constant EC_{50} (Fig. 4 *F*, *G*, and *I*), and decreases in $\mathbf{z}_2$ correspond to EC_{50} and G_∞ that remain near the wild-type values. This is consistent with changes to a fourth biophysical constant of the analytic model: the ligand binding affinity of the active (operator binding) state (K_A; Fig. 4*I*). So, we can interpret mutations with $\mathbf{z}_2$ effects as free-energy changes related to the active state ligand affinity.

To further understand the connection between the latent mutational effects and the biophysics of LacI, we analyzed the association of mutational effects in the $\mathbf{z}_1$ and $\mathbf{z}_2$ dimensions across the LacI protein structure. Mutations in the dimerization region of LacI generally increase $\mathbf{z}_1$ (Fig. 4*J*). Since mutations in this region are unlikely to affect the DNA operator affinity or ligand affinity, these shifts in $\mathbf{z}_1$ are likely due to changes in the allosteric constant ($\Delta\epsilon_{AI}$). Conversely, mutations in the ligand-binding region of LacI more often decrease $\mathbf{z}_1$, consistent with the potential for mutations in this region to affect ligand affinity (K_I; Fig. 4*K*). However, mutations near the ligand-binding region can

also affect the allosteric constant (32). Finally, mutations in the DNA-binding domain of LacI generally increase $\mathbf{z}_1$ (Fig. 4*L*). This is consistent with a decrease in the DNA operator affinity ($\Delta\epsilon_{RA}$). Surprisingly, mutations that increase $\mathbf{z}_2$, corresponding to an apparent decrease in the active-state ligand binding constant (K_A), are not typically found near the ligand-binding region. Instead, they largely occur at the N-terminal half of the dimerization interface (Fig. 4*M*). This region of the protein undergoes a conformational shift when the protein switches between the active and inactive states (33). So, mutations here could affect K_A via changes in allosteric communication between the DNA-binding and ligand-binding domains of the protein.

SARS-CoV-2 RBD. Starr et al. (3) measured the effect of over 3,800 distinct mutations in the SARS-CoV-2 spike protein RBD for their binding affinity to ACE2 ($\log K_d$) and structural stability (measured as change in yeast display expression, $\Delta\log$MFI), sampling more than 170,000 unique variants (3). LANTERN learned five dimensions in this landscape measurement (Fig. 4*A*). The most common direction of mutational effects roughly follows a gradient of steepest descent for structural stability measured by $\Delta\log$MFI (Fig. 5 *A*–*C*). We derived an axis from this direction, which we refer to as the stability axis (Fig. 5*C*). The direction of latent effect for nearly 50% of mutations (1,842/3,798) lies within 10° of this axis. We next identified a second axis that lies along a constant ridge of RBD structural stability, which we call the binding axis (Fig. 5*C*). ACE2 binding decreases along both axes, but structural stability only changes along the stability axis (Fig. 5 *D*–*F*). This suggests that mutations along the stability axis, that is, most mutations of the RBD, decrease structural stability. This decrease in structural stability then disrupts RBD binding to the ACE2 receptor, since the spike protein must fold correctly before it can bind to ACE2. Conversely, mutations along the binding axis do not impact structural stability and may be particularly important in forming the RBD–ACE2 complex independent of structural stability. This interpretation is supported by the location of mutations within the different protein structure domains: The majority of mutations with latent effects along the binding axis are near the RBD–ACE2 interface, while mutations along the stability axis are distributed throughout the core RBD domain (Fig. 5*G*).

To demonstrate what LANTERN can reveal about clinically relevant mutations, we analyzed mutations that have been found in recently identified SARS-CoV-2 variants of concern (Fig. 5*H*) (34, 35). The latent mutational effects of these mutations all point in distinct directions within the latent space learned by LANTERN, and each direction corresponds to higher binding affinity than the wild type. So, although each of these mutations has a distinct impact on the protein's function in terms of latent mutational effect, each is predicted to increase ACE2 binding affinity. One mutation in particular, N501Y, has a mutational effect vector that points directly toward the predicted maximum of RBD–ACE2 binding affinity. N501Y occurs in the alpha (B.1.1.7), beta (B.1.351), and gamma (P.1) variants of SARS-CoV-2 that increased in proportion worldwide throughout late

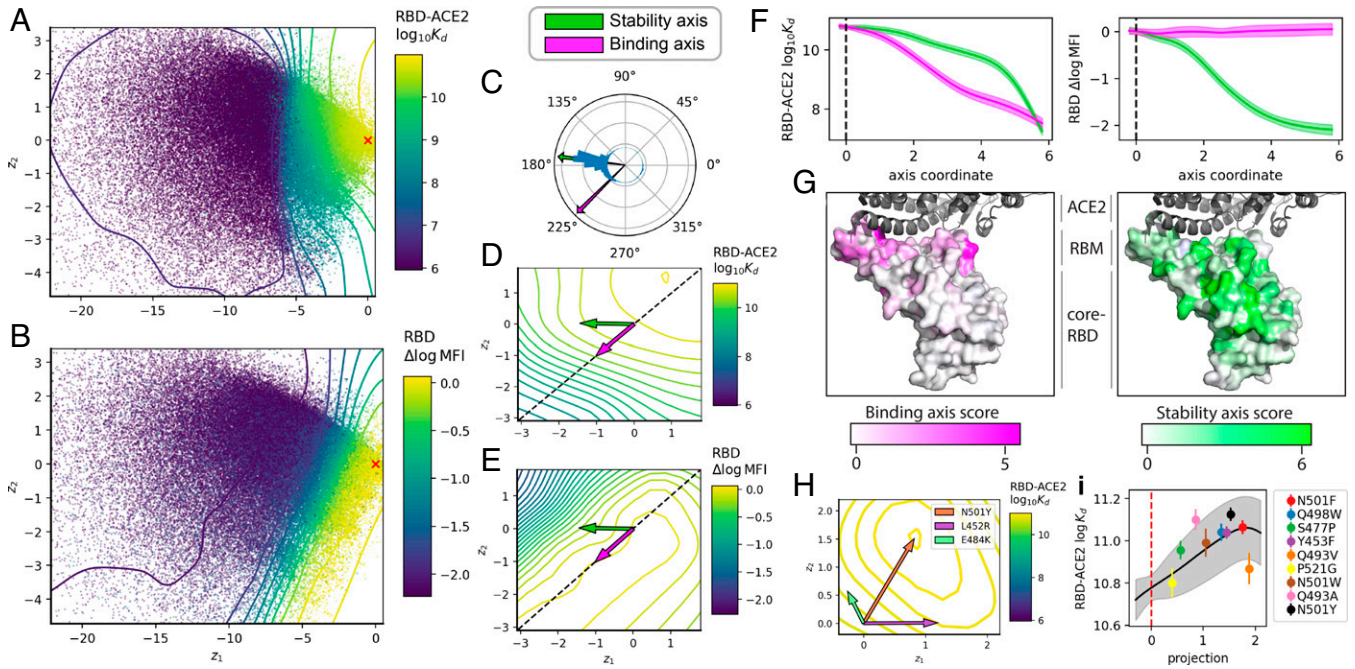

Fig. 5. SARS-Cov2 joint RBD-ACE2 binding and RBD expression landscape. (*A* and *B*) Learned surface $f(\mathbf{z})$ for SARS-Cov2 RBD–ACE2 binding ($\log K_d$ (*A*) and RBD expression (Δ log MFI (*B*). Contours show posterior mean of $f(\mathbf{z})$, scatter points are mean posterior $\mathbf{z}$ for individual variants colored by their observed mean phenotype value, and red crosses mark the wild-type origin. Surfaces in *A* and *B* show the latent space centered on the central 99% of observations and a cross-section through the $\mathbf{z}_1$–$\mathbf{z}_2$ plane. For additional views of the surface, see *SI Appendix*, Figs. S15–S18. (*C*) Distribution of latent, single-mutant effect directions along the first two dimensions of $\mathbf{z}$, with stability axis (green) and binding axis (pink). (*D* and *E*) Joint RBD–ACE2 binding and RBD expression surfaces overlaid with identified axes. Contours are the mean of the variational posterior of $f(\mathbf{z})$. (*F*) The predicted RBD–ACE2 binding and RBD expression as a function of one-dimensional subspaces along both axes. Specifically, for each axis $\vec{v}_i$, we plot the predicted surface along the subspace of $\mathbf{z}$: $f(c\vec{v}_i)$ for $c \in \mathbb{R}$. Solid line shows the posterior mean of $f(\mathbf{z})$, and the shaded area is the 95% credible interval. (*G*) Structural association of the stability and binding axes to the RBD. The average mutation effect vector at each residue is represented in a new basis defined by the stability and binding axes (*SI Appendix*, Fig. S21 and section 3), and the coordinates of the vector in the new basis are shown as a score. Higher score values mean that mutation effect vectors at the residue induce larger changes in $\mathbf{z}$ along the direction of the corresponding axis. We highlight the components of the structure corresponding to the core RBD and receptor-binding motif (RBM) of the SARS-CoV-2 spike protein RBD, as well as the region of ACE2 that contacts the RBD. (*H*) Mutations of interest associated with COVID-19 outbreaks. Individual mutational effects are shown as vectors. Contours are the mean of the variational posterior of $f(\mathbf{z})$. (*I*) Predicted binding along the mutational effect direction of N501Y. The *x* axis corresponds to the scalar value when projecting each mutation effect vector onto N501Y's mutation vector. Solid black line is the posterior mean, and shaded regions are the 95% CI of $f(\mathbf{z})$ along the axis defined by N501Y's latent mutation effect vector. Solid red line is the wild-type origin. Single-mutant variants that align with the virulence axis and have a positive projection value are shown with error bars representing a 95% CI of the single mutant measurement.

2020 and early 2021 (36, 37). Given the importance of this mutation to the ongoing pandemic, we analyzed the 10 mutations with the most similar latent effects, determined by the magnitude of the projection score of each mutation's latent effect vector onto that of N501Y. Among these mutations, LANTERN predicts increased RBD–ACE2 binding strength, and measurements generally confirm stronger ACE2 binding for single mutants with these mutations (Fig. 5*I*). The LANTERN analysis indicates that these mutations involve mechanisms similar to N501Y, based on the similarity their latent mutational effects. So, these mutations may be particularly important for genomic surveillance of SARS-CoV-2. Finally, we analyzed the L452R mutation, present in the now prevalent delta (B.1.617.2) variant. The mutational effect vector for L452R points toward a region of increased binding but may also contribute to the spike protein's stability. Specifically, the vector points in the negative direction of the identified stability axis (Fig. 5 *C* and *H*).

LANTERN Quantifies Local Robustness and Additivity. Despite the complexity of GPLs, quantitative metrics can summarize their most important features and simplify their analysis. For example, global metrics like landscape ruggedness can provide insight into adaptive evolution and engineering potential (38). With LANTERN, we define two metrics based on local properties of the landscape: slope and curvature. The slope of the landscape describes the rate of change of the surface at each position in latent mutational effect space (Fig. 6*A*). With zero slope, the phenotype remains constant in response to small latent mutational effects. So, the (inverse) slope is associated with robustness. The curvature of the landscape reflects the rate of change of the slope, and zero curvature implies that mutations have a constant effect on phenotype (Fig. 6*B*). In regions with zero curvature, mutations have no epistasis. So, the (inverse) curvature is associated with additivity.

To apply the concept of robustness to the multidimensional mutational effect spaces learned by LANTERN, we generalized the slope to multiple dimensions as the surface gradient $\nabla f(\mathbf{z})$. The gradient represents the rate of change in each dimension as a vector, and a gradient with values near zero represents multidimensional robustness (Fig. 6*C*). In the case of SARS-CoV-2, LANTERN predicts high robustness near the predicted maximum of binding strength between the RBD and ACE2 (Fig. 6*D*). This region of latent mutational effect space poses a potential clinical threat, as it represents genetically stable SARS-CoV-2 infectivity through strong affinity between the RBD and ACE2. Variants in this region may then have more mutational flexibility to evade immune response while maintaining infectivity (35, 39, 40).

To similarly quantify additivity in multiple dimensions, we computed the curvature in the form of the Laplacian ($\Delta f(\mathbf{z})$; Fig. 6*E*). A value of zero for the Laplacian indicates a constant rate of change in all dimensions and constitutes multidimensional additivity. As an evaluation of additivity as a useful metric, we analyzed the additivity surface of LacI EC_{50} (Fig. 6*F*). Previous analysis of the LacI EC_{50} phenotype showed that mutational effects in single and double mutants combine with very little epistasis (2). The additivity surface quantitatively represents this phenomenon, with the EC_{50} surface having high additivity around the wild type (Fig. 6*F*). Notably, we draw this conclusion directly from the additivity surface predicted by LANTERN, rather than through combinatorial screening of epistatic effects in single and double mutants.

Both robustness and additivity are local properties of the surface: They describe differential behaviors for infinitesimally small changes in the latent space $\mathbf{z}$. These local properties can reveal useful insights into the general structure of each surface, as shown above. However, the change in $\mathbf{z}$ from individual mutations is nonnegligible, quantified by the magnitude of their mutation effect vector. So, local robustness and additivity cannot predict the impact of mutational effects with large magnitude (*SI Appendix*, Fig. S22). The magnitude of most mutational effect vectors is small, however, so robustness and additivity still provide useful approximations to the effect of individual mutations in local regions of the landscape.

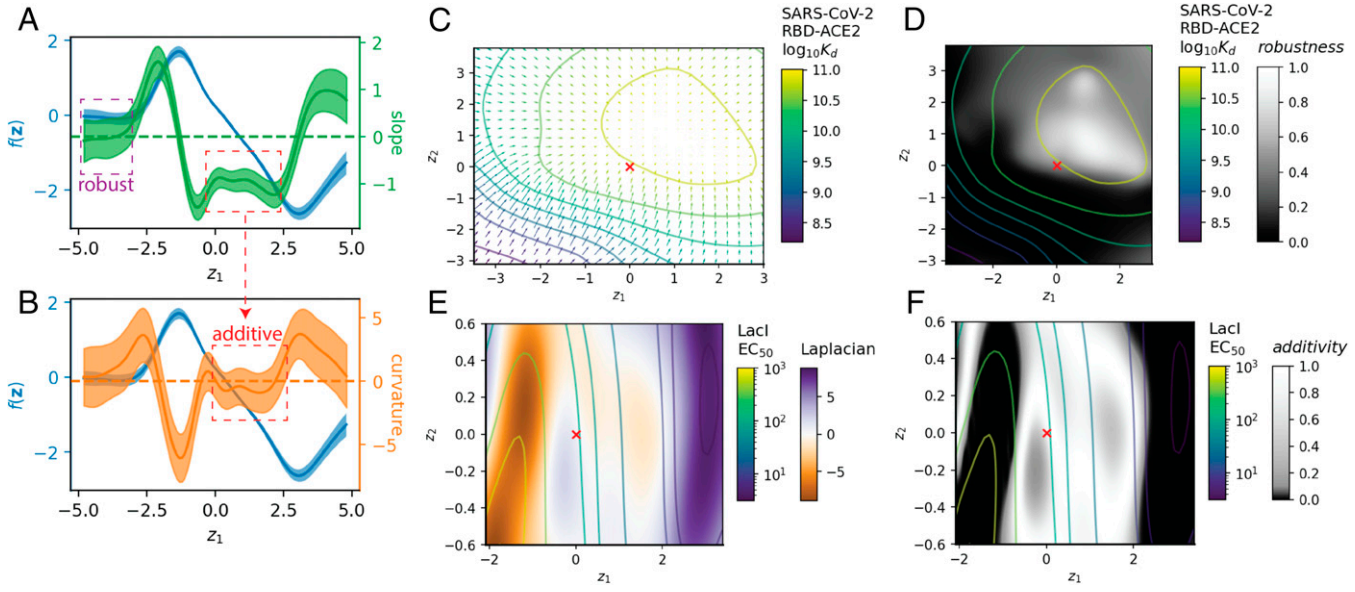

Fig. 6. Local robustness and additivity of GPLs. (*A*) LacI EC_{50} surface and slope along $\mathbf{z}_1$. Posterior of $f(\mathbf{z})$ is shown in blue, and slope ($[\partial/\partial\mathbf{z}_1]f(\mathbf{z})$) is shown in green. When the slope is zero, the surface is locally robust (purple box). (*B*) LacI EC_{50} surface and curvature along $\mathbf{z}_1$. Posterior of $f(\mathbf{z})$ is shown in blue, and curvature ($[\partial^2/\partial\mathbf{z}_1^2]f(\mathbf{z})$) is shown in orange. When the curvature is zero, the surface is locally additive (red box). A curvature of zero also implies the slope is constant, although it may be nonzero (*A*, red box). In both *A* and *B*, solid lines are posterior mean, and shaded regions are 95% credible intervals of $f(\mathbf{z})$. (*C*) The gradient of SARS-CoV-2 binding. Arrows show the posterior mean of the gradient ($\nabla f(\mathbf{z})$), and the contours mark the posterior mean of $f(\mathbf{z})$. The multidimensional equivalent of robustness is when $\nabla f(\mathbf{z}) = 0$. (*D*) The robustness of SARS-CoV-2 binding. Values near one indicate near-zero gradient. (*E*) The curvature of LacI EC_{50} in multiple dimensions, calculated as the Laplacian ($\Delta f(\mathbf{z})$). The multidimensional equivalent of additivity is when $\Delta f(\mathbf{z}) = 0$. (*F*) The additivity of LacI EC_{50}. Values near one indicate $\Delta f(\mathbf{z})$ is close to zero. In *C*–*F*, red cross marks the wild-type origin ($\mathbf{z}_{wt} = \mathbf{0}$).

Discussion

LANTERN addresses the need for GPL models that make accurate predictions while remaining interpretable. We show that, in a benchmark across multiple GPLs, LANTERN achieves equal or better predictive accuracy compared with state-of-the-art DNN models (Fig. 3). However, beyond our analysis of varying DNN depth (*SI Appendix*, Fig. S8), our comparison to DNN models may remain incomplete, because changes in network architecture or training hyperparameters can marginally improve predictive accuracy (41–43). This further highlights the advantages of LANTERN, however, because LANTERN provides a single modeling interface for any GPL measurement, with no additional tuning necessary. Additionally, DNNs typically require large-scale measurements to ensure satisfactory performance but provide no clear cutoff for how many data are sufficient. In contrast, LANTERN scales to any dataset size (*SI Appendix*, Figs. S9 and S10). Overall, we expect LANTERN will provide accurate predictions in a broad class of GPLs and impact engineering and research endeavors that depend on extrapolation to new genotypes.

With black-box models, improvements to predictive accuracy come at the expense of model interpretability. This trade-off is also widely assumed to be unavoidable (9). In the case of GPLs, LANTERN shows that this is not the case: LANTERN makes state-of-the-art predictions (i.e., equal to or better than alternative approaches) while remaining fully interpretable. So, LANTERN automatically explains every prediction it makes through its construction from easily understandable components. So, LANTERN achieves state-of-the-art prediction without any trade-off in interpretability.

The dimensionality of a GPL, determined by the number of biophysical parameters influenced by mutations, contributes to landscape ruggedness and the distribution of epistatic interactions (44). LANTERN automatically determines the dimensionality, which we validated with simulations (*SI Appendix*, Figs. S2 and S3). However, along with simply detecting the number of latent dimensions, LANTERN provides guidance on the degree of importance for each latent dimension in the form of their relevance, similar to the decreasing variance explained in PCA (22). Additionally, the dimensionality learned by LANTERN is empirical: LANTERN learns the number of latent dimensions sufficient to explain the data. As new measurements become available, the dimensionality may increase—reflecting the discovery of new structure in the landscape. LANTERN also benefits from a rich history of research in dimensionality reduction (22, 45, 46), in some cases, recovering latent spaces similar to alternative approaches (*SI Appendix*, section 4 and Fig. S23). But, LANTERN is unique for jointly learning a nonlinear surface and low-dimensional latent space.

Beyond dimensionality, LANTERN learns latent mutational effect spaces that reflect the underlying biophysical process, possibly up to a rotation (Fig. 2*D*). But, in cases where the effect of mutations on multiple biophysical parameters are correlated, LANTERN will struggle to separate these effects into different dimensions. For example, multiple biophysical parameters (or a combination of them) may explain the $\mathbf{z}_1$ axis of the LacI landscape (Fig. 4*I*). Without additional prior information, like the plausibility of correlation between the biophysical parameters in response to mutations, no data-driven approach can resolve this ambiguity. This further emphasizes the value of LANTERN, however, because this issue is only made apparent through the interpretability of LANTERN.

LANTERN also presents straightforward explanations of the effects of individual mutations through their latent effect vector. These vectors facilitate clear understanding of how each mutation contributes to an observed phenotype (Figs. 4*E* and 5*H*) and how these mutations are tied to structural biophysics (Figs. 4 *J*–*M* and 5*G*). LANTERN also quantifies the uncertainty about the effect of each mutation, with uncertainty decreasing the more times a mutation is observed in the dataset (*SI Appendix*, Fig. S24). From this uncertainty, we can determine which mutations have effects that are statistically different from zero. For the large-scale GPL datasets considered here, the majority of mutations (ranging from 55 to 93% of all mutations) have a significant effect (*SI Appendix*, Fig. S25). This agrees with the expectation that most genes and proteins have been optimized via natural selection, so that mutations will disrupt or otherwise impact the protein's function (47). So, modeling approaches that include an assumption of sparsity (i.e., that most mutations have no effect on function) would not be appropriate for analyzing GPL data.

LANTERN quantifies the metrics of local robustness and additivity, which provide perspectives for understanding large-scale GPL measurements (Fig. 6). Rather than describing behavior with regards to any one genotype, they represent the expected effects of underlying parameters on a global phenotype in one small evolutionary region.

LANTERN models GPLs as a nonlinear surface over a low-dimensional latent mutational effect space, a form of epistasis where nonadditivity arises from global structure (6, 8). Certain theoretical models of GPLs, which, until recently, could not be verified, due to the lack of sufficient experimental data, similarly concentrated on the existence of low-dimensional manifolds that explain the complexity of GPLs (48, 49). These models suggest that GPLs will commonly involve low-dimensional structure, due to the benefits in adaptive evolution (20, 50). LANTERN will therefore likely have broad applicability across GPL measurements of diverse biological systems (1).

LANTERN also extends beyond existing global epistasis modeling approaches that employ stronger assumptions, for example, a predetermined dimensionality or a fixed family of nonlinear functions (6, 24). LANTERN, instead, learns these details directly from the data. In some cases, the model learned by LANTERN agrees strongly with these existing approaches (*SI Appendix*, Fig. S26).

One example is avGFP, where the 95% posterior credible interval of the surface $f(\mathbf{z})$ along $\mathbf{z}_1$ largely overlaps with the result from a one-dimensional monotonic I-spline model (6), and there is strong correlation between the mutational effects learned by both approaches (*SI Appendix*, Fig. S26). Given that $\mathbf{z}_1$ constitutes the vast majority of mutational effect variance for avGFP (96.7% for $\mathbf{z}_1$ versus 3.0% and 0.3% for $\mathbf{z}_2$ and $\mathbf{z}_3$; Fig. 4*A*), LANTERN has largely verified that, for avGFP, the most important dimension (and nonlinearity across it) is captured by the monotonic I-spline approach. This also explains the similar predictive accuracy between these two models for avGFP (Fig. 3): It can be difficult to resolve the predictive advantage of including these second and third dimensions, since so much predictive power is realized just by identifying $\mathbf{z}_1$. The two additional dimensions identified by LANTERN should then be regarded as additional potential mechanisms involved in the brightness phenotype that should receive additional exploration in future studies.

LANTERN also assumes global epistasis is the primary factor in GPL data. Clearly, if this assumption is incorrect, then alternative models may be preferable. In particular, if specific

epistatic interactions are dominant, then models designed to capture those interactions might be necessary (5), and future extensions of LANTERN could potentially include those effects. However, when assessing the predictive accuracy of specific-epistasis models on the large-scale datasets considered here, these models performed poorly (*SI Appendix*, Fig. S27). So, at least for these datasets, LANTERN appears to be the better modeling approach.

We placed a symmetric Gaussian prior on the latent mutational effect space, reflecting the assumption that the effect of mutations will be equally distributed in both directions (Eq. **4**). In the case of avGFP and SARS-CoV-2, the distribution of effects appears nonsymmetric in some dimensions (Figs. 4*D* and 5*C*). Asymmetry may be a common feature of mutational effects, in particular when evaluating the effect of mutations on a sequence that has undergone heavy selection (48). In this case, most mutations will be deleterious, and mutational effects will be asymmetrically biased in one direction. Therefore, future extensions to LANTERN may benefit from updating the prior on mutational effects to account for the potential presence of asymmetry. But, as seen for avGFP and SARS-CoV-2, the detection of asymmetrical mutational effects is not prevented by our choice of symmetric priors.

We also model the nonlinear surface with a GP, ensuring a balance between model complexity and data fit (23). So, while an unconstrained nonlinear surface could exactly interpolate between data points, LANTERN's surfaces remain smooth, finding parsimonious explanations of the data (Figs. 4 and 5). These learned surfaces are also reproducible, with surfaces remaining constant in response to multiple perturbations (*SI Appendix*, section 5 and Figs. S28 and S29). Finally, these surfaces reveal GPL structure not supported by previous parametric models, including nonmonotonic effects of mutations on phenotype (*SI Appendix*, Fig. S30).

But, care should be taken when drawing conclusions on the overall structure of the nonlinear surface $f(\mathbf{z})$, in particular when evaluating regions of the latent space with sparse experimental coverage. In these regions, the posterior on $f(\mathbf{z})$ will be uncertain, and the posterior mean will revert to the overall mean of the observed phenotype. In cases where stronger assumptions are justified for $f(\mathbf{z})$, for example, monotonicity, those assumptions could be incorporated into the model. The posterior mean of $f(\mathbf{z})$ is also not a complete representation of the distribution over possible nonlinear surfaces, so, even when the mean of $f(\mathbf{z})$ departs from stronger assumptions (like monotonicity), a certain proportion of samples from the posterior of $f(\mathbf{z})$ may still exhibit these properties.

Large-scale GPL measurements will increasingly influence bioscience initiatives of the future. Despite increases in experimental throughput, the full genotypic space will always remain undersampled. To overcome this fundamental limitation, LANTERN facilitates progress through reduction of landscapes to their minimal complexity. In this way, LANTERN transforms the intractable challenge of exhaustive genotypic sampling to a manageable exploration of a low-dimensional space. GPL-enabled investigations can then explore this space in an efficient manner, relying on LANTERN's guidance toward uncharted regions of phenotypic diversity.

Materials and Methods

GPL Datasets. We aggregated GPL measurements from published sources (2, 3, 12). In the case of LacI, we filtered observations to variants with Hill equation–like dose–response curves to avoid inference on inaccurate Hill equation parameter estimates. We combined the two libraries of SARS-Cov-2 for each of the binding and expression measurements to make a single, aggregate dataset for both measurements, including variants that were present in both binding and expression datasets.

We prepared all training datasets in a similar fashion. Within each dataset, we one-hot encoded all mutations (51). Specifically, for p total mutations in a dataset, each variant l was represented as the one-hot encoded vector $x_i \in [0,1]^p$. For each corresponding phenotype $y_i \in \mathbb{R}^D$ with phenotype dimensionality D, we standardized each phenotype dimension separately to a mean of zero and SD of one. Each final dataset for training was then $\mathcal{D} = \left[\{x_i, y_i\} | 1 \leq i \leq N\right]$ for N total variants.

LANTERN. We constructed LANTERN with two key components: a latent surface $f(\mathbf{z})$ and a set of latent mutational effects, represented by a matrix $W = [\vec{\mathbf{z}}^{(1)}, \vec{\mathbf{z}}^{(2)}, \ldots, \vec{\mathbf{z}}^{(p)}] \in \mathbb{R}^{K \times p}$, where $\vec{\mathbf{z}}^{(k)}$ represents the mutational effect vector of mutation k (Fig. 1*A*). First, we placed a GP prior on the surface $f(\mathbf{z})$,

$$f(\mathbf{z}) \sim \mathrm{GP}\Big(\mu_f(\mathbf{z}), \kappa_f(\mathbf{z}, \mathbf{z}')\Big), \quad [1]$$

with mean and kernel functions μ_f and κ_f, respectively. We set the mean function μ_f as an unknown constant value, $\mu_f(\mathbf{z}) = \hat{f}$. The kernel function κ_f describes the covariance between different observations of f: $\mathrm{cov}\big(f(\mathbf{z}), f(\mathbf{z}')\big) = \kappa_f(\mathbf{z}, \mathbf{z}')$. We used the rational quadratic covariance function,

$$\kappa_f(\mathbf{z}, \mathbf{z}') = \sigma_\kappa^2 \left[1 + \frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2\eta}\right]^{-\eta}, \quad [2]$$

where σ_κ is an unknown scale parameter, and η controls the overall distribution between smooth and rugged regions of f. We did not include a length-scale parameter for the norm $\|\mathbf{z} - \mathbf{z}'\|$ because we already learn the relative scale between dimensions through the hierarchical prior on dimension variance (52). So, even though the kernel function is isotropic, the nonlinear surfaces learned by LANTERN are not constrained to isotropic functions (*SI Appendix*, section 6).

Next, we specified the hierarchical prior for the unknown mutation effects $W \in \mathbb{R}^{K \times p}$ for K latent mutational effect dimensions and p mutations. In all cases, we set $K = 8$ (*SI Appendix*, section 7). For a variant i with mutation vector x_i, the latent mutational effect vector z_i (conditional on W) is

$$z_i = W x_i. \quad [3]$$

Note that z_i, representing the combination of latent mutational effects for each mutation in variant i, is distinct from a dimension of the latent mutational effect space (e.g., $\mathbf{z}_1$).

For each row k of W, w_k, we defined a Gaussian prior for each element of w_k,

$$w_{kj} \sim N(0, \alpha_k^{-1}), \quad [4]$$

where α_k is the precision (or inverse variance) of dimension k. We place a Gamma distribution prior on each α_k (22),

$$\alpha_k \sim \mathrm{Gamma}(\gamma_0, \beta_0), \quad [5]$$

where γ_0 and β_0 are model hyperparameters. We set $\gamma_0 = \beta_0 = 10^{-3}$ in all experiments here, which generally leads to models that minimize the number of dimensions (22). We rank the importance of each dimension by its variance, which we refer to as its relevance.

To combine these components when modeling a GPL dataset $\mathcal{D}$, we assume that each phenotype y_i is conditionally independent given the unknown variables,

$$p(\mathcal{D} | f, W, \boldsymbol{\alpha}) = \prod_{i=1}^{N} p(y_i | f, W, \boldsymbol{\alpha}, x_i), \quad [6]$$

with likelihood of each phenotype y_i as a normal distribution

$$p(y_i | f, W, \boldsymbol{\alpha}, x_i) = N(y_i | f_i, \sigma_y^2 + \sigma_i^2), \quad [7]$$

where $f_i = f(z_i) = f(Wx_i)$, σ_y^2 is an unknown variance parameter estimated from the data, and σ_i^2 is any additional measurement uncertainty provided for each variant in the dataset.

We treat the kernel parameters σ_κ and η, mean surface value $\hat{f}$, and the global phenotype noise σ_y^2 as unknown variational parameters constrained to be positive and learned for each dataset.

Variational Inference. Given the specified model, inference involves the recovery of the posterior distribution,

$$p(f, W, \boldsymbol{\alpha}|\mathcal{D}) = \frac{p(\mathcal{D}|f, W, \boldsymbol{\alpha})p(f)p(W|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\mathcal{D})}. \quad [8]$$

The exact posterior is analytically intractable, and we, instead, relied on an approximation through variational inference (VI) (29). VI recasts the inference procedure as an optimization problem to minimize the Kullback-Leibler divergence between an approximate posterior $q(f, W, \boldsymbol{\alpha})$ and the true posterior $p(f, W, \boldsymbol{\alpha}|\mathcal{D})$ (*SI Appendix*, section 8).

Implemenation. We implemented LANTERN with the automatic differentiation library pytorch (53), with GP components of the model relying on gpytorch (54). We trained on computers with GPUs with $\geq$12 GB of memory and 5,120 CUDA cores. We trained LANTERN models with minibatch size 4,096 or 8,192, depending on GPU capacity, learning rate of 10^{-2} using the Adam optimizer (55), and for 5,000 epochs. We set the maximum number of latent mutational effect dimensions to eight. A total of 800 inducing points were learned for each model, with initial positions uniform randomly sampled over the range $[-10, 10]$.

Determining Dimensionality. For each dimension with learned precision α_k, we sort the dimensions such that $\sigma_k^2 = 1/\alpha_k$ decreases with k (i.e., $\sigma_1^2 > \sigma_2^2 > \dots$). Then, we calculate the expected log-likelihood of each observation with an increasing number of dimensions included in the model,

$$\ell_{ik} = E_q[\log p(y_i|f, W^{(k)})], \quad [9]$$

where $W^{(k)}$ is the subspace of the full latent space defined by W up to dimension k (i.e., the first k rows of W). We assess the impact of including dimension k in the model, by testing whether the evidence for the data (ℓ_{ik}) has increased compared to a model with $k-1$ dimensions. Specifically, we apply a one-sided, two-sample Kolmogorov–Smirnov test to compare the empirical distributions of ℓ_{ik} and ℓ_{ik-1}, and consider the dimension k necessary for the model if those distributions are significantly different, that is, $p \leq 0.05$ (*SI Appendix*, Figs. S2, S3, and S13). We then define the dimensionality as the maximum dimension k where this is true.

Computing Additivity and Robustness. Additivity and robustness were derived from the Laplacian ($\Delta f(\boldsymbol{z})$) and gradient ($\nabla f(\boldsymbol{z})$) posterior predictive distributions, respectively (*SI Appendix*, section 9). For a location z_k in the latent mutational effect space, we calculate the analytic posterior distribution of the gradient, $q(\nabla f(z_k)) \approx N(\mu_\nabla, \Sigma_\nabla)$, and Laplacian, $q(\Delta f(z_k)) \approx N(\mu_\Delta, \sigma_\Delta^2)$. Both robustness and additivity were quantified from the unnormalized density, or kernel, of their respective differential operator's posterior predictive distribution at a value of zero for the corresponding differential operator,

$$\text{robustness} = \exp\left(-\mu_\nabla^T \Sigma_\nabla^{-1} \mu_\nabla / 2\right) \quad [10]$$

and

$$\text{additivity} = \exp\left(-\|\mu_\Delta\|^2 / 2\sigma_\Delta^2\right), \quad [11]$$

Values of additivity and robustness close to one then imply near-zero values for the underlying differential operators $\Delta f(\mathbf{z})$ and $\nabla f(\mathbf{z})$.

Models Used in Comparison with LANTERN.

Linear. A linear GPL model [implemented in pytorch (53)]

$$y_i = \beta^T x_i, \quad [12]$$

assuming β_k is the average effect of mutation k.

I-spline. We used the monotonic I-spline model of ref. 6 via the python library dms_variants (https://github.com/jbloomlab/dms_variants), using default parameters and a Gaussian likelihood for all datasets.

DNN. We adapted DNN architectures from recent publications performing regression on GPL measurements (11, 12). These architectures follow a feed-forward structure,

$$y_i = f_\theta(x_i) = W_\theta^{(2)} \sigma(W_\theta^{(1)} \sigma(W_\theta^{(0)} \sigma(z) + b^{(0)}) + b^{(1)}) + b^{(2)}, \quad [13]$$

with weights $W^{(k)}$, biases $b^{(k)}$, and nonlinearity σ. These models generally have a low-dimensional initial hidden layer, for example, $W^{(0)} \in \mathbb{R}^{L \times K}$ and $b^{(0)} \in \mathbb{R}^L$ with $L \lll K$. The primary hidden layer has width w: $W^{(1)} \in \mathbb{R}^{w \times L}$ and $b^{(0)} \in \mathbb{R}^w$. A final linear layer transforms the hidden neurons to the output dimension: $W^{(2)} \in \mathbb{R}^{D \times w}$ and $b^{(2)} \in \mathbb{R}^D$. For this study, models were constructed with an initial hidden layer width L of either one or eight, ReLU nonlinearity σ, and a hidden width of $w = 32$. We trained neural networks with the Adam optimizer, with a minibatch size of 128, learning rate of 10^{-3}, for 100 epochs. We chose the epoch length to minimize the held-out validation prediction error. We minimized the mean-squared error, with losses weighted inversely proportional to measurement uncertainty. To evaluate increased depth of DNN architectures (*SI Appendix*, Fig. S8), new layers were added matching the architecture of the first hidden layer.

Cross-Validation. We evaluated all models with 10-fold cross-validation. We quantified predictive accuracy across folds with the weighted coefficient of determination,

$$R^2_{\boldsymbol{\eta}} = 1 - \frac{\sum_{i=1}^n \eta_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^n \eta_i (y_i - \bar{y}_{\boldsymbol{\eta}})^2}, \quad [14]$$

for model predictions $\hat{y}_i$, $\bar{y}_{\boldsymbol{\eta}} = (\sum \eta_i y_i)/(\sum \eta_i)$, and individual predictions weighted inversely proportionally to their measurement certainty (σ_i^2 for each observation i): $\boldsymbol{\eta} = \{1/\sigma_1^2, 1/\sigma_2^2, \dots\}$. Predictions from models approximated via variational methods were made with the posterior mean of all unknown random variables. To determine the relationship between predictive accuracy and dataset size, we trained models from data subsampled at 5,000 increments of the full training set size for each cross-validation fold, and evaluated performance on the held-out test data for all resulting models.

Uncertainty Quantification. We quantified predictive uncertainty of LANTERN with Monte Carlo draws from the approximate variational posterior. Specifically, a Monte Carlo sample of the unknown mutation effects was taken from the variational posterior: $\tilde{W} \sim q(W)$. This sample was then used to calculated the latent position of every variant: $\tilde{z}_i = \tilde{W} x_i$. Then, the approximate posterior predictive distribution of the surface f was used to sample a predictive phenotype for the variant: $\tilde{f}_i \sim q(f|z_i)$. This two-stage sampling process was repeated 50 times to estimate the overall uncertainty of f_i for each variant.

Predictive intervals for evaluating uncertainty calibration were calculated from the approximate posterior predictive distribution of f_i. From the Monte Carlo samples, we approximate the posterior predictive distribution for y_i as $N(\hat{\mu}_i, \hat{\sigma}_i^2 + \sigma_y^2)$, where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the mean and variance of the posterior predictive distribution, respectively, for f_i. From each distribution, we define the cumulative distribution function $F_i(y_i)$ and the corresponding quantile function $F_i^{-1}(p) = \inf\{y : y \leq F_i(y)\}$ for $p \in [0, 1]$. Then, for all p, we say a model is calibrated if

$$C_p = \frac{\sum_{i=1}^N \mathbb{I}(y_i \leq F_i^{-1}(p))}{N} = p, \quad [15]$$

where $\mathbb{I}$ is an indicator function and is equal to one if the inequality is satisfied, and is zero otherwise (56). Model calibration was determined by comparing the observed coverage of the interval (C_p) versus the expected coverage for a calibrated model (p) (*SI Appendix*, Fig. S12). We also corrected for biases in the density of observations to more accurately assess the calibration when predicting phenotypes different from the wild type (*SI Appendix*, section 10).

Simulated Biophysical Allosteric Model. Simulated data were generated using a biophysical model for allosteric transcriptional regulation (26). The model relates the dose–response curve of an allosteric transcription factor to biophysical parameters such as binding constants and free-energy differences between protein states. These parameters determine the baseline dose–response (G_0), the saturated dose–response (G_∞), and the sensitivity (EC_{50}). Simulated datasets were generated assuming a protein with 300 amino acid positions and six possible amino acid substitutions at each position. In each dataset, the number

of biophysical parameters influenced by these substitutions varied from one to three ($\mathcal{D}_1$ to $\mathcal{D}_3$, respectively).

For each simulated dataset, 100,000 simulated variants were generated by first assigning the number of amino acid substitutions for each variant based on an empirical distribution from an experimental dataset (2). The positions and identities of simulated substitutions were then randomly chosen as uniform random draws from possible positions (without replacement), and possible substitutions (with replacement). Shifts in biophysical parameters were then determined for each simulated variant by summing the effects of each substitution, and the resulting biophysical parameter values were used to calculate G_0, G_∞, and EC_{50} for each simulated variant. For more details, see *SI Appendix*, section 1.

Code Availability. Source code of the LANTERN library is available at https://github.com/usnistgov/lantern. Analysis code of the manuscript is available at https://github.com/usnistgov/lantern/tree/master/manuscript.

Data Availability. Code and processed data are available in the GitHub repository, https://github.com/usnistgov/lantern. The GPL data that support the findings of this study are available at Figshare for avGFP (DOI: 10.6084/m9.figshare.3102154), the National Institute of Standards and Technology (NIST) Public Data Repository for LacI (DOI: 10.18434/M32259), and GitHub for SARS-CoV-2 (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS).

ACKNOWLEDGMENTS. We thank Drew Tack, Blaza Toman, Swarnavo Sarkar, and Dennis Leber for thoughtful feedback. P.D.T. and A.P. were supported through National Research Council Fellowships. Analysis was performed, in part, on the NIST Nisaba clusters.

Author affiliations: [a]Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899; and [b]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD 20899

1. J. B. Kinney, D. M. McCandlish, Massively parallel assays and quantitative sequence–function relationships. *Annu. Rev. Genomics Hum. Genet.* **20**, 99–127 (2019).
2. D. S. Tack *et al.*, The genotype-phenotype landscape of an allosteric protein. *Mol. Syst. Biol.* **17** (2021).
3. T. N. Starr *et al.*, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and Ace2 binding. *Cell* **182**, 1295–1310.e20 (2020).
4. J. Domingo, P. Baeza-Centurion, B. Lehner, The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).
5. F. J. Poelwijk, V. Krishna, R. Ranganathan, The context-dependence of mutations: A linkage of formalisms. *PLOS Comput. Biol.* **12**, e1004771 (2016).
6. J. Otwinowski, D. M. McCandlish, J. B. Plotkin, Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7550–E7558 (2018).
7. J. Otwinowski, Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
8. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
9. Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **16**, 31–57 (2018).
10. J. Otwinowski, J. B. Plotkin, Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2301–E2309 (2014).
11. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLOS Genet.* **15**, e1008079 (2019).
12. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
13. J. A. Valeri *et al.*, Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat. Commun.* **11**, 5058 (2020).
14. N. M. Angenent-Mari, A. S. Garruss, L. R. Soenksen, G. Church, J. J. Collins, A deep learning approach to programmable RNA switches. *Nat. Commun.* **11**, 5057 (2020).
15. T. Ching *et al.*, Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
16. J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
17. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
18. R. Guidotti *et al.*, A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 1–42 (2019).
19. R. N. Gutenkunst *et al.*, Universally sloppy parameter sensitivities in systems biology models. *PLOS Comput. Biol.* **3**, 1871–1878 (2007).
20. T. U. Sato, K. Kaneko, Evolutionary dimension reduction in phenotypic space. *Phys. Rev. Res.* **2**, 013197 (2020).
21. N. Tokuriki, D. S. Tawfik, Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
22. C. Bishop, "Variational principal components" in *9th International Conference on Artificial Neural Networks: ICANN '99* (Institution of Engineering and Technology, 1999) , vol. 1, pp. 509–514.
23. C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning* (Adaptative Computation and Machine Learning Series, University Press Group Limited, 2006).
24. Z. R. Sailer *et al.*, Inferring a complete genotype-phenotype map from a small number of measured phenotypes. *PLOS Comput. Biol.* **16**, e1008243 (2020).
25. M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
26. M. Razo-Mejia *et al.*, Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. *Cell Syst.* **6**, 456–469.e10 (2018).
27. M. Jankowiak, G. Pleiss, J. R. Gardner, "Parametric Gaussian process regressors." in *Proceedings of the 37th International Conference on Machine Learning*, H. Daumé III, A. Singh, Eds. (PMLR, 2020), vol. 119, pp. 4702–4712.
28. J. Snoek *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates, 2019), vol. 32, pp. 13969–13980.
29. D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
30. T. T. Yang *et al.*, Improved fluorescence and dual color detection with enhanced blue and green variants of the green fluorescent protein. *J. Biol. Chem.* **273**, 8212–8216 (1998).
31. H. W. Ai, N. C. Shaner, Z. Cheng, R. Y. Tsien, R. E. Campbell, Exploration of new chromophore structures leads to the identification of improved blue fluorescent proteins. *Biochemistry* **46**, 5904–5910 (2007).
32. G. Chure *et al.*, Predictive shifts in free energy couple mutations to their phenotypic consequences. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18275–18284 (2019).
33. M. Lewis *et al.*, Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**, 1247–1254 (1996).
34. E. B. Hodcroft, *Covariants : SARS-CoV-2 mutations and variants of interest.* https://covariants.org. Accessed 4 January 2021.
35. W. T. Harvey *et al.*; COVID-19 Genomics UK (COG-UK) Consortium, SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
36. Centers for Disease Control and Prevention, *CDC COVID data tracker*. https://covid.cdc.gov/covid-data-tracker/#datatracker-home. Accessed 5 January 2021.
37. D. Frampton *et al.*, Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: A whole-genome sequencing and hospital-based cohort study. *Lancet Infect. Dis.* **21**, 1246–1256 (2021).
38. D. A. Kondrashov, F. A. Kondrashov, Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
39. B. Hie, E. D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
40. L. Atlani-Duault, B. Lina, F. Chauvin, J. F. Delfraissy, D. Malvy, Immune evasion means we need a new COVID-19 social contract. *Lancet Public Health* **6**, e199–e200 (2021).
41. J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
42. P. Balaprakash *et al.*, *Scalable Reinforcement-Learning-Based Neural Architecture Search for Cancer Deep Learning Research* (Association for Computing Machinery, New York, NY, 2019).
43. K. Kandasamy, W. Neiswanger, J. Schneider, B. Póczos, E. P. Xing, "Neural architecture search with Bayesian optimization and optimal transport" in NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, S. Bengio, H. M. Wallach, Eds. (Curran Associates, 2018), pp. 2020–2029.
44. S. Hwang, S. C. Park, J. Krug, Genotypic complexity of Fisher's geometric model. *Genetics* **206**, 1049–1079 (2017).
45. K. C. Li, Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316–327 (1991).
46. J. P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learn. Res.* **16**, 2859–2900 (2015).
47. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
48. H. A. Orr, The genetic theory of adaptation: A brief history. *Nat. Rev. Genet.* **6**, 119–127 (2005).
49. O. Tenaillon, The utility of Fisher's geometric model in evolutionary genetics. *Annu. Rev. Ecol. Evol. Syst.* **45**, 179–201 (2014).
50. K. Husain, A. Murugan, Physical constraints on epistasis. *Mol. Biol. Evol.* **37**, 2865–2874 (2020).
51. C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
52. E.L. Snelson, Z. Ghahramani, "Variable noise and dimensionality reduction for sparse Gaussian processes." in *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence* (AUAI Press, 2006).
53. A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates, 2019), vol. 32, pp. 8024–8035.
54. J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, A. G. Wilson, "Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration" in *Advances in Neural Information Processing Systems*, S. Bengio *et al.*, Eds. (Curran Associates, 2018), vol. 31, pp. 7576–7586.
55. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv [Preprint] (2014). https://doi.org/10.48550/arXiv.1412.6980 (Accessed 16 May 2022).
56. V. Kuleshov, N. Fenner, S. Ermon, "Accurate uncertainties for deep learning using calibrated regression." in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (PMLR, 2018), vol. 80, pp. 2796–2804.