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

Peter D. Tonner^{1,*}, Abe Pressman¹, and David Ross¹

¹ National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899

*Corresponding author

Abstract

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 biochemical system. Recent modeling efforts, however, prioritize predictive accuracy at the expense of model interpretability. Here, we present LANTERN (<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 into the landscape including its inherent dimensionality, a latent space of additive mutational effects, and novel metrics of landscape structure. LANTERN facilitates straightforward discovery of fundamental mechanisms in GPLs, while also reliably extrapolating to unexplored regions of genotypic-space.

1 Introduction

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 – 10^7 distinct genotypes (Kinney et al. 2019). These measurements play an expanding role in understanding biological variation, with applications from engineering to epidemiology (Tack et al. 2021; Starr, Greaney, et al. 2020). Despite their importance, though, 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 – 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 un-measured 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 (Domingo et al. 2019). 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 non-additive interactions due to their close physical proximity in the fold-protein, and GPL models can represent this effect through pair-wise interaction coefficients (Poelwijk et al. 2016). Alternatively, even when individual mutations cause additive changes to an underlying biophysical parameter (e.g. folding free energy), the measured phenotype can be a non-linear function of that parameter (Jakub Otwinowski et al. 2018; Jakub Otwinowski 2018; Starr and Thornton 2016). GPL models that directly represent these epistatic mechanisms have the advantage of being *interpretable* (Lipton 2018). 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 (J. Otwinowski et al. 2014). So, recent efforts to model large-scale GPL measurements have instead

relied on deep neural network (DNN) architectures due to their superior predictive performance (Pokusaeva et al. 2019; Sarkisyan et al. 2016; Valeri et al. 2020; Angenent-Mari et al. 2020; Ching et al. 2018). DNNs make predictions through complicated cascades of non-linear 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* techniques estimate the relevant factors of DNN predictions (Jiménez-Luna et al. 2020). 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 (Rudin 2019). Additionally, these techniques only apply to individual predictions rather than generalizing the whole model (Guidotti et al. 2019). 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 novel GPL modeling approach called LANTERN (a genotype-phenotype 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 non-linear effects of epistasis through a multi-dimensional, non-parametric Gaussian-process model. 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.

2 Results

2.1 Constructing interpretable models of genotype-phenotype landscapes

LANTERN takes as input data a combination of genotypes and their measured phenotypes (Fig 1a). A LANTERN model of this 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. $\mathbf{z}^{(i)}$ for mutation i , Fig 1c). 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 this latent mutational effect space. This underlying additive structure can accurately represent many biophysical phenomena. For example, individual mutations often additively impact the thermodynamic stability of protein folding (Tokuriki et al. 2009). GPLs measurements, however, frequently target a phenotype that is a non-linear function of these additive effects. In the example of thermodynamic folding stability, many protein phenotypes remain robust to small decreases in this stability but rapidly diminish beyond a certain threshold (Starr and Thornton 2016). So, the second component of LANTERN relates measured phenotypes to the combination of latent mutational effects through a smooth, non-linear surface: $f(\mathbf{z})$ (Fig 1d).

These two components, the latent mutational effect space $\mathbf{z}$ and non-linear 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 biochemical mechanisms driving the complexity of GPLs in either the latent mutational space or the non-linear 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. The variance of each dimension determines its relative impact on describing the effect of mutations: dimensions with higher variance result in larger mutational effects while dimensions with variance close to zero are effectively removed from the model (Bishop 1999). 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 placed

a hierarchical prior on the variance of each dimension of the latent mutational effect space (Fig 1d). This prior represents the expectation that only a relatively small number of latent mutational effect dimensions will be necessary to explain the data and ensures that dimensions that are 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 non-linear relationship between latent mutational effects and measured phenotypes, we placed a Gaussian process (GP) prior on the surface $f(\mathbf{z})$ (Fig 1f) (Rasmussen et al. 2006). GPs recover the distribution of 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 (Supplemental 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 (Jakub Otwinowski et al. 2018; Sailer, Shafik, et al. 2020). To learn both components of LANTERN for different GPLs, we apply stochastic variational methods that make inference tractable and scalable to millions of observations (Hoffman et al. 2013).

2.2 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 (Razo-Mejia et al. 2018). 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 2a, Section 4.6). We then randomly combined these mutations to simulate individual protein variants with a mean of 4.38 mutations per variant. For each variant, the resulting perturbed biophysical constants determined the allosteric dose-response parameters via the analytic model. We trained LANTERN only using the variant

genotypes and resulting dose-response parameters, *without* access to the underlying biophysical parameters or the analytical model.

To test LANTERN’s ability to correctly identify GPL dimensionality, we simulated three different GPL datasets. In these datasets, we varied the number of perturbed biophysical constants from one to three (datasets $\mathcal{D}_1$ to $\mathcal{D}_3$, respectively), creating datasets with true dimensionality equal to the number of varying biophysical constants. For each dataset, we considered the learned variance of the latent-space dimensions (Fig 2b) and determined the dimensionality learned by LANTERN by counting the number of dimensions with learned variance above a fixed threshold. Below this threshold, no latent mutational effects remain significantly non-zero (Supplemental Fig S2a). Additionally, when we removed dimensions with variance below this threshold from the model, predictive accuracy remained constant (Supplemental Fig S2b–f). Across all simulated datasets, the learned dimensionality exactly matched the number of biophysical constants perturbed when generating each landscape, demonstrating that LANTERN can recover the true dimensionality of GPL measurements automatically from data.

To determine if 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 parameters, despite the fact that LANTERN had no access to this information (Fig 2c, Supplemental Fig S3). Furthermore, LANTERN learned smooth, interpretable surfaces that accurately predict the allosteric phenotype of each simulated landscape (Fig 2d, Supplemental Figs S4ab, S5). Using the correlation between latent mutational effects learned by LANTERN and the true biophysical constants in the simulation (Fig 2c), we rotated and re-scaled the space of latent mutational effects to directly compare $f(\mathbf{z})$ to the true biophysical surface (Fig 2e, Supplemental Fig S4c). In regions of latent mutational effect space with large sampling coverage, the rotated and re-scaled surface, $f(\mathbf{z})$, matches the surface from the analytical model exactly (Fig 2e). As sampling density decreases, the deviation between predicted and true surfaces increases (Fig 2f). 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 can recover the underlying dimensionality and biophysical mechanisms of our simulations *de-novo* from data, without additional

Table 1: Large-scale genotype phenotype landscape datasets.

Dataset	observations	mutations	phenotype	reference
avGFP	54,025	1,879	fluorescence	(Sarkisyan et al. 2016)
LacI	47,462	2,510	hill coefficient EC_{50}	(Tack et al. 2021)
LacI	47,462	2,510	hill coefficient G_{∞}	(Tack et al. 2021)
SARS-CoV-2	146,437	3,802	ACE2-RBD binding ($\log K_D$)	(Starr, Greaney, et al. 2020)
SARS-CoV-2	177,759	4,002	RBD expression ($\Delta \log \text{MFI}$)	(Starr, Greaney, et al. 2020)

domain specific knowledge.

2.3 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 (Table 1): the fluorescence brightness of green fluorescent protein (**avGFP**; Sarkisyan et al. 2016), the dose-response curves of an allosteric transcription factor (**LacI**; Tack et al. 2021), and the joint phenotypes of ACE2 binding affinity and structural stability for the receptor binding domain of the SARS-CoV-2 spike protein (**SARS-CoV-2**; Starr, Greaney, et al. 2020). 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 ten 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 non-linear models as well as black-box DNNs. Overall, LANTERN achieves a new state-of-the-art in predictive accuracy. In all but one instance, LANTERN equalled or outperformed all alternative approaches with respect to ten-fold cross-validated prediction accuracy (see Methods, Fig 3a). 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 feedforward neural network, which performed equally well. Additionally, LANTERN outperforms all other approaches in pre-

dicting LacI EC₅₀. 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 (Tack et al. 2021). 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 (Jankowiak et al. 2019), we found a substantial increase in LANTERN’s prediction accuracy for G_{∞} (Supplemental Fig S6). Overall, LANTERN provides the most accurate predictions of any single method across a broad benchmark of GPL datasets.

To determine the impact of sample size on LANTERN performance, we evaluated cross-validated accuracy as a function of dataset size (Fig 3b). Due to the formulation 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. The relative gains of LANTERN over alternative approaches become most pronounced as the number of measurements increases. For smaller datasets, e.g. less than five thousand genotypes, LANTERN performs equally well as a simple linear model. Importantly, LANTERN does not underperform the linear model as dataset size decreases, reflecting the ability of LANTERN to straightforwardly scale *down* to relatively small datasets. LANTERN even scales down to small-scale measurements with only five to seven mutations (Supplemental Fig S7).

2.4 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 (Snoek et al. 2019). We therefore compared the prediction uncertainties reported by LANTERN to its actual prediction error. The actual prediction error typically increase with increased mutational distance from the wild-type genotype (Fig 3c). 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 3d). LANTERN does appear to underestimate the overall true error rate however, a known issue with variational methods (Blei et al. 2016). Overall, the degree of confidence placed in each prediction by LANTERN is automatically specified by its uncertainty.

2.5 LANTERN provides interpretable models of genotype-phenotype landscapes

While LANTERN generates reliable predictions, it also straightforwardly explains these predictions through its interpretable components. Here, we demonstrate how this interpretability can be leveraged to understand the models learned by LANTERN for the three large-scale GPL datasets: avGFP, LacI, and SARS-CoV2. Across these datasets, the latent dimensionality learned by LANTERN ranged three to five (Fig 4a). 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 (Supplemental Figs S8—S11).

avGFP To determine how LANTERN makes predictions about avGFP fluorescent brightness, we analyzed the latent mutational effects and surface learned by the model. A sharp boundary along the most relevant dimension, $\mathbf{z}_1$, 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 4b). Maximum brightness occurs in a region near (but not centered on) the wild-type (Fig 4c). Over 78% of mutations (1476/1879) decrease fluorescence because their mutational effect vectors point away from this maximum towards the region of decreased fluorescence (Fig 4d). Previous analysis of this landscape demonstrated that structural stability constituted a major factor in decreased fluorescence, with small increases in folding free energy leading to complete loss of functionality (Sarkisyan et al. 2016). However, discovery of this association depended on fixing the dimensionality of the neural network model used for the analysis. In contrast, LANTERN learns this association directly from data, finding a primary axis of decreased fluorescence along $\mathbf{z}_1$ that recapitulates the relationship between structural stability and brightness.

In order to further understand the LANTERN model of avGFP fluorescence, we analyzed mutations used in engineered variants of blue fluorescent protein (BFP, Yang et al. 1998). The foundational mutation of all BFP variants is Y66H. GFP variants containing Y66H have on-average

fluorescence equal to 0.081 ± 0.010 the wild-type level and LANTERN coordinately predicts Y66H to decrease brightness, as expected in an assay to detect green fluorescence (Fig 4e). However, newer variants of BFP include additional mutations that increase the blue fluorescence, possibly through stabilizing protein structure (Ai et al. 2007). Interestingly, LANTERN predicts all of these mutations (with one exception: S65T) to similarly increase wild-type avGFP fluorescence. Additionally, these mutations point towards the maximum predicted brightness avGFP (Fig 4c). According to LANTERN then, these mutations may generally improve brightness independent of fluorescent color, possibly through improved structural stability.

LacI Tack et al. (2021) measured the allosteric landscape of over 47,000 LacI variants with Hill-equation-like dose-response curves. With LANTERN, we modeled the landscape of these response curves with a multivariate phenotype of EC_{50} and G_{∞} for each variant. From the LANTERN model, the LacI GPL has a dimensionality of three, and the majority of mutational effects closely aligned with the positive or negative $\mathbf{z}_1$ axis (Fig 4a, f-h).

To interpret the latent mutational effect space learned by LANTERN, we compared the effects of changes in the two most relevant latent effect dimensions ($\mathbf{z}_1$ and $\mathbf{z}_2$) with the expectations from an analytic biophysical model for LacI dose-response (Razo-Mejia et al. 2018). Increases in $\mathbf{z}_1$ corresponds to a decreased EC_{50} and a slightly increased G_{∞} . Conversely, decreases in $\mathbf{z}_1$ correspond to an increased EC_{50} and a sharply decreased G_{∞} (Fig 4f,g,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 4i): the allosteric constant ($\Delta\epsilon_{AI}$), the DNA operator affinity ($\Delta\epsilon_{RA}$), or the ligand binding affinity of the inactive (non-operator binding) state (K_I) (Eqs. 21, 22). So, we can interpret the changes in the $\mathbf{z}_1$ direction as free energy changes related to those three biophysical constants. Changes along the $\mathbf{z}_2$ dimension have a different effect: the positive $\mathbf{z}_2$ direction corresponds to a decreased G_{∞} at roughly constant EC_{50} (Fig 4f,g,i), and the negative $\mathbf{z}_2$ direction corresponds 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 4i). So, we can interpret the changes in the $\mathbf{z}_2$ direction 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 distribution 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 4j). Since mutations in this region are unlikely to affect the DNA operator affinity or ligand affinity, these shifts in $\mathbf{z}_1$ could be due to changes in the allosteric constant ($\Delta\epsilon_{AI}$). Conversely, mutations in the ligand-binding region of LacI more often decrease $\mathbf{z}_1$ (Fig 4k). Mutations in this region might be expected to affect the ligand affinity (K_I), though previous work indicates that mutations near the ligand-binding region can also affect the allosteric constant (Chure et al. 2019). Finally, mutations in the DNA-binding domain of LacI generally increase $\mathbf{z}_1$ (Fig 4l). 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 4m). This region of the protein undergoes a conformational shift when the protein switches between the active and inactive states (Lewis et al. 1996). 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 Receptor Binding Domain Starr, Greaney, et al. 2020 measured the effect of over 3,800 distinct mutations in the SARS-CoV-2 spike protein receptor binding domain (RBD) for their binding affinity to ACE2 ($\log K_d$) and structural stability (measured as change in yeast display expression, $\Delta \log \text{MFI}$), sampling in total more than 170,000 unique variants. LANTERN discovered five dimensions in this landscape measurement (Fig 4a). The most common direction of mutational effects roughly follows a gradient of steepest descent for structural stability measured by $\Delta \log \text{MFI}$ (Fig 5a–c). We derived an axis from this direction, which we refer to as the *stability* axis (Fig 5c). The direction of latent effect for nearly 50% of mutations (1842/3798) lie within ten degrees of this axis. We next identified a nearly orthogonal axis that lies along a constant ridge of RBD structural stability, which we call the *binding* axis (Fig 5c). ACE2 binding decreases along both axes, but structural stability only changes along the *stability* axis (Fig 5d–f). This suggests that mutations along the *stability* axis, i.e. 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 5f,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 (Hodcroft 2021; Harvey et al. 2021, Fig 5h). 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 B.1.1.7 variant that has steadily increased in proportion worldwide (Centers for Disease Control and Prevention 2021; Frampton et al. 2021). Given the importance of this mutation to the ongoing pandemic, we analyzed the ten mutations with the most similar latent effects, determined by the similarity of their latent mutational effect directions to that of N501Y. Among these mutations, RBD-ACE2 binding strength increases, and measurements generally confirm stronger ACE2 binding for single-mutants with these mutations (Fig 5i). The LANTERN analysis indicates that these mutations involve similar mechanisms as N501Y, based on the similarity their latent mutational effects. So, these mutations may be particularly important for genomic surveillance of SARS-CoV-2.

2.6 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 GPL metrics provide insight into adaptive evolution and engineering potential (Kondrashov et al. 2015). With LANTERN, we define two novel 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 6a). 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 6b). 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 6c). In the case of SARS-CoV-2, LANTERN predicts high robustness near the predicted maximum of binding strength between the RBD and ACE2 (Fig 6d). 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 (Hie et al. 2021; Atlani-Duault et al. 2021; Harvey et al. 2021).

To similarly quantify additivity in multiple dimensions, we add the curvature across each dimension to compute the Laplacian ($\Delta f(\mathbf{z})$, Fig 6e). 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 EC₅₀ (Fig 6f). Previous analysis of the LacI EC₅₀ phenotype showed that mutational effects in single and double mutants combine with no epistasis (Tack et al. 2021). The additivity surface quantitatively represents this phenomenon, with the EC₅₀ surface having high additivity around the wild-type (Fig 6f). 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.

3 Discussion

LANTERN addresses the need for GPL models that make accurate predictions while remaining highly 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, our comparison to DNN models may remain incomplete because changes in network architecture or training hyperparameters can marginally improve predictive accuracy (Bergstra et al. 2012; Bal-

aparakash et al. 2019; Kandasamy et al. 2018). This further highlights the advantages of LANTERN though, 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 much data is sufficient. In contrast, LANTERN scales to any dataset size (Supplemental Fig S7). 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.

Typically, improvements to model predictive accuracy come at the expense of model interpretability, with an assumption that this trade-off is unavoidable (Lipton 2018). In the case of GPLs, LANTERN shows that this trade-off is unnecessary: optimal predictions do not preclude interpretable modeling. Specifically, due to its construction from easily understandable components, LANTERN automatically provides an explanation for every prediction it makes. Furthermore, we establish that these components also independently provide insight into the biological mechanisms of GPLs. For example, LANTERN learns latent mutational effect spaces that reflect the underlying biophysical process, possibly up to a rotation (Fig 2e). 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 (Fig 4e, 5h) and how these mutations are tied to structural biochemistry (Fig 4j—m, 5g). Finally, LANTERN quantifies the new metrics of local *robustness* and *additivity* that provide novel perspectives for understanding large-scale GPL measurements (Fig 6). Overall, we expect LANTERN’s interpretability will enhance our understanding of diverse landscapes across biology.

LANTERN models GPLs as a non-linear surface over a low-dimensional latent mutational effect space, a form of epistasis where non-additivity arises from global structure (Starr and Thornton 2016; Jakub Otwinowski et al. 2018). 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 (Orr 2005; Tenailon 2014). These models suggest that GPLs will commonly involve low-dimensional structure due to the benefits in adaptive evolution (Sato et al. 2020; Husain et al. 2020). LANTERN will therefore likely have broad applicability across GPL measurements of diverse biological systems (Kinney et al. 2019). Additionally, LANTERN extends beyond existing global epistasis modeling approaches

that suffer from restrictive assumptions, e.g., a predetermined dimensionality or a fixed family of non-linear functions (Jakub Otwinowski et al. 2018; Sailer, Shafik, et al. 2020) (Supplemental Figs S12, S13). LANTERN instead learns these details directly from the data. This enables direct comparisons between GPLs that are not possible with other modeling approaches, like the relative complexity of two landscapes quantified by their dimensionality (Fig 2b, 4a).

Large-scale GPL measurements will increasingly influence bioscience initiatives of the future. Despite the growing magnitude of experimental throughput, the full genotypic space will always remain under-sampled. 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.

4 Methods

4.1 Genotype-phenotype landscape datasets

We aggregated GPL measurements from published sources (Sarkisyan et al. 2016; Tack et al. 2021; Starr, Greaney, et al. 2020). In the case of LacI, we filtered observations to variants with Hill-equation-like dose-response curves to avoid inference on inaccurate hill parameter estimates. We combined the two libraries of SARS-Cov-2 for each of binding and expression measurements to make a single, aggregate dataset for both measurements. In the case of joint modeling, we merged the binding and expression datasets for cases where a variant was present in both.

We prepared all training datasets in a similar fashion. Within each dataset, we one-hot encoded all mutations (Angermueller et al. 2016). Specifically, for p total mutations in a dataset (Table 1), each variant i 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 standard deviation of one. The final dataset for training was then $\mathcal{D} = [\{x_i, y_i\} | 1 \leq i \leq N]$ for N total variants.

4.2 Hierarchical Bayesian modeling of GPLs

We constructed LANTERN with two key components: a latent surface $f(\mathbf{z})$ and a set of latent mutational effects, represented by a matrix $W = [z^{(1)}, z^{(2)}, \dots, z^{(p)}] \in \mathbb{R}^{K \times p}$ where $z^{(i)}$ represents the mutational effect vector of mutation i (Eq 3, Fig 1a). First, we placed a Gaussian process (GP) prior on the latent surface $f(\mathbf{z})$:

$$f(\mathbf{z}) \sim \text{GP}(\mu_f(\mathbf{z}), \kappa_f(\mathbf{z}, \mathbf{z}')), \quad (1)$$

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 : $\text{cov}(f(\mathbf{z}), f(\mathbf{z}')) = \kappa_f(\mathbf{z}, \mathbf{z}')$. We used the rational quadratic covariance function:

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

where σ_κ is an unknown scale parameter and α controls the overall distribution between smooth and rugged regions of f . We did not include a lengthscale parameter for the norm $\|\mathbf{z} - \mathbf{z}'\|$ used in the kernel because we already learn the relative scale between dimensions through the hierarchical prior on dimension variance (Snelson et al. 2006).

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. For a variant i with mutation vector $x_i \in \{0, 1\}^p$ (see Sec 4.1), the latent mutational effect vector z_i (conditional on W) is computed as

$$z_i = Wx_i. \quad (3)$$

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 r of W , w_r , we defined a Gaussian prior for each element of w_r :

$$w_{r,j} \sim N(0, \alpha_r^{-1}), \quad (4)$$

where α_r is the precision (or inverse variance) of dimension i . We assumed each α_r to be drawn from a Gamma distribution (Bishop 1999):

$$\alpha_r \sim \text{Gamma}(\alpha_0, \beta_0), \quad (5)$$

where α_0 and β_0 are model hyper-parameters. We set $\alpha_0 = \beta_0 = 10^{-3}$ in all experiments here, which generally leads to models that minimize the number of effective dimensions (Bishop 1999). Different values of α_0 and β_0 would change the scale learned for the latent dimensional effect space (Supplemental Fig S14). However, the overall scale of the latent dimension is not relevant to the model. Instead, a certain small number of dimensions should have substantially larger magnitude than the rest. So, we use the prior for α_r to ensure that most dimensions are effectively removed from the model while allowing a small number of dimension with non-negligible variance ($1/\alpha_r$). We rank the importance of each dimension by its variance, which we refer to as its *relevance*.

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

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

The likelihood of each phenotype y_i was calculated as

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

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 α , as well as the global phenotype noise σ_y^2 , as unknown variational parameters (Section 4.3) constrained to be positive. These parameters are then still learned for each dataset.

4.3 Approximate inference

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) (Blei et al. 2016). VI recasts the inference procedure as an optimization problem to minimize the KL divergence between an approximate posterior $q(f, W, \boldsymbol{\alpha})$ and the true posterior $p(f, W, \boldsymbol{\alpha} | \mathcal{D})$. We take the typical mean field approach which assumes that q decomposes multiplicatively across unknown variables: $q(f, W, \boldsymbol{\alpha}) = q(f) \prod_i [q(\alpha_i) \prod_j q(w_{ij})]$. See below for the choice of q for each variable in the model. Returning to the objective, which is still intractable, we instead minimized it indirectly through maximization of the evidence lower bound (ELBO, Supplemental Fig S16):

$$\text{ELBO}(q) = \mathbb{E}_q [\log p(\mathcal{D} | f, W, \boldsymbol{\alpha})] - \text{KL}(q(f, W, \boldsymbol{\alpha}) \| p(f, W, \boldsymbol{\alpha})) \quad (9)$$

$$= \mathbb{E}_q [\log p(\mathcal{D} | f, W, \boldsymbol{\alpha})] - \text{KL}(q(f) \| p(f)) - \mathbb{E}_{q(\boldsymbol{\alpha})} [\text{KL}(q(W) \| p(W | \boldsymbol{\alpha}))] - \text{KL}(q(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha})). \quad (10)$$

This objective bounds the model evidence $p(\mathcal{D})$ up to an unknown constant and therefore maximizing the ELBO corresponds to minimizing the KL divergence between q and the true posterior.

In order to make inference scalable to large-scale GPL datasets, we employed stochastic variational inference (SVI) (Hoffman et al. 2013). SVI approximates exact variational inference by sub-sampling the training data in smaller mini-batches and computing a noisy approximation of the gradient of the ELBO:

$$\text{ELBO}(q) = \sum_{k=1}^B \left[\mathbb{E}_q [\log p(\mathcal{D}_k | f, W, \boldsymbol{\alpha})] \right] - \text{KL}(q(f) \| p(f)) - \mathbb{E}_{q(\boldsymbol{\alpha})} [\text{KL}(q(W) \| p(W | \boldsymbol{\alpha}))] - \text{KL}(q(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha})) \quad (11)$$

$$\approx \mathbb{E}_q [\log p(\mathcal{D}_k | f, W, \boldsymbol{\alpha})] - \frac{1}{B} \left[\text{KL}(q(f) \| p(f)) - \mathbb{E}_{q(\boldsymbol{\alpha})} [\text{KL}(q(W) \| p(W | \boldsymbol{\alpha}))] - \text{KL}(q(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha})) \right], \quad (12)$$

where $\mathcal{D}_k$ is the k th mini-batch and there are B total mini-batches.

We rely on variational distributions q for which the KL divergence is analytically tractable (see below). But, computation of the expected log-likelihood of the data remains an analytical challenge. So, we approximate the true expectation through unbiased Monte-Carlo samples. This is made possible by re-parameterizing each variational distribution $q(\theta)$ in terms of a new distribution $q(\epsilon)$ such that $\theta = g(\epsilon)$ and $\mathbb{E}_q[h(\theta)] = \mathbb{E}_\epsilon[h(g(\epsilon))]$. We can then approximate the true expected log-likelihood as

$$\frac{1}{T} \sum_{t=1}^T \log p(\mathcal{D}_k | f_t, W_t, \alpha_t), \quad (13)$$

for T Monte-carlo samples (we set $T = 1$ in all cases) and Monte-carlo samples of each parameter using its reparameterized distribution: $\theta_t = g(\epsilon_t)$ and $\epsilon_t \sim q(\epsilon)$. This greatly reduces the variance in gradient calculations and improves inference (Diederik P Kingma et al. 2013; Blundell et al. 2015).

Variational distributions. We used a variational distribution of $f(\mathbf{z})$ using the methods of (Hensman et al. 2015). Briefly, the complete distribution on f is approximated by learning the distribution of f at specific inducing points Z (which are parameters of the variational distribution and not random variables) and unknown function values $u = f(Z)$. A variational distribution on u is defined as $q(u) \sim N(\mu_u, \Sigma_u)$, the variational distribution on f is

$$q(f) = \int p(f|u)q(u)du, \quad (14)$$

and can be analytically calculated from the GP prior. For M inducing points, the computational cost of this approximation during inference is $O(M^3)$, realizing substantial savings when $M \ll N$. We use the implementation of the approximation approach for $f(\mathbf{z})$ provided by the python library `gpytorch` (Gardner et al. 2018).

For the remaining parameters, weights $w_{i,j}$ and precisions α_i , we set their variational family to the same family as their corresponding prior:

$$q(w_{i,j}) \sim N(\mu_{i,j}, \sigma_{i,j}^2) \quad \text{and} \quad q(\alpha_i) \sim \text{Gamma}(\tilde{\alpha}_i, \tilde{\beta}_i), \quad (15)$$

where $\mu_{i,j}$, $\sigma_{i,j}^2$, $\tilde{\alpha}_i$, and $\tilde{\beta}_i$ are optimized with respect to the ELBO (Eq. 12). Note that $p(\mathbf{w}_i|\alpha_i) = N(\mathbf{w}_i|0, \alpha_i)$ depends on α_i , which we obtain from a Monte Carlo sample from $q(\alpha_i)$.

We implemented LANTERN with the automatic differentiation library `pytorch` (Paszke et al. 2019), with GP components of the model relying on `gpytorch` (Gardner et al. 2018). We trained models on compute systems with NVIDIA TITAN V GPUs, with ≥ 12 GB of memory and 5120 CUDA cores. We trained LANTERN models with mini-batch size 4096 or 8192, depending on GPU capacity, learning rate of 10^{-2} using the Adam optimizer (Diederik P. Kingma et al. 2014), and for 5,000 epochs. We set the maximum number of latent mutational effects dimensions to eight. A total of 800 inducing points were learned for each model, with initial positions uniform randomly sampled over the range $[-10, 10]$.

4.4 Computing *additivity* and *robustness*

Additivity and *robustness* were derived from the Laplacian ($\Delta f(\mathbf{z})$) and gradient ($\nabla f(\mathbf{z})$) posterior predictive distributions, respectively (Section S3). For a location z_k in the latent mutational effect space, we calculate the analytic posterior distribution of the gradient, $q(\nabla f(z_k)) \sim N(\mu_{\nabla}, \Sigma_{\nabla})$, and Laplacian, $q(\Delta f(z_k)) \sim N(\mu_{\Delta}, \sigma_{\Delta}^2)$ (See Equations S2 and S13, respectively). Both *robustness* and *additivity* were quantified from the unnormalized density, or kernel, of their respective differential operator’s posterior predictive distribution at zero:

$$\text{robustness} = \exp\left(-\frac{1}{2}\mu_{\nabla}^T \Sigma_{\nabla}^{-1} \mu_{\nabla}\right). \quad (16)$$

and

$$\text{additivity} = \exp\left(-\frac{\|\mu_{\Delta}\|^2}{2\sigma_{\Delta}^2}\right), \quad (17)$$

Values of *additivity* and *robustness* close to 1 then imply high probability of zero values for the underlying differential operators $\Delta f(z)$ and $\nabla f(z)$.

4.5 Comparison of predictive accuracy

Models. We employed the monotonic I-spline model of (Jakub Otwinowski et al. 2018) via the `dms_variants` library (github.com/jbloomlab/dms_variants), which provides an implementation of the model in python. We trained monotonic spline models on each dataset using default parameters. We used a Gaussian likelihood for all datasets.

We adapted deep neural network architectures from recent publications performing regression on GPL measurements (Sarkisyan et al. 2016; Pokusaeva et al. 2019). These architectures follow a feed-forward structure:

$$y_i = f_{\theta}(x_i) = W_{\theta}^{(2)} \sigma(W_{\theta}^{(1)} \sigma(W_{\theta}^{(0)} \sigma(z) + b^{(0)}) + b^{(1)}) + b^{(2)}, \quad (18)$$

with weights $W^{(k)}$ and biases $b^{(k)}$ and non-linearity σ . These models generally have a low-dimensional initial hidden layer, e.g. $W^{(0)} \in \mathbf{R}^{L \times K}$ and $b^{(0)} \in \mathbf{R}^L$ with $L \ll K$. The primary hidden layer has width w : $W^{(1)} \in \mathbf{R}^{w \times L}$ and $b^{(1)} \in \mathbf{R}^w$. A final linear layer transforms the hidden neurons to the output dimension: $W^{(2)} \in \mathbf{R}^{D \times w}$ and $b^{(2)} \in \mathbf{R}^D$. For this study, models were constructed with an initial latent dimensionality L of either one or eight, sigmoid nonlinearity σ , and hidden width of $w = 32$. We trained neural networks with the Adam optimizer, with mini-batch size of 128, learning rate of 10^{-3} , for 100 epochs. We chose the epoch length as it minimized the held-out validation prediction error. We minimized the mean squared error, with losses weighted inversely proportional to measurement uncertainty.

Cross validation. We evaluated all models with ten-fold cross validation. The same folds for each dataset were re-used across modeling approaches in order to ensure accurate comparisons. We quantified predictive accuracy across folds with the coefficient of determination, $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ for model predictions $\hat{y}_i$ and observations mean $\bar{y}$. In the case of measurement uncertainty, the individual prediction scores were weighted proportionally to the certainty of their measurement:

$$R_{\mathbf{w}}^2 = 1 - \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^n w_i (y_i - \bar{y}_{\mathbf{w}})^2}, \quad (19)$$

for weights $\mathbf{w} = \{\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots\}$ and $\bar{y}_{\mathbf{w}} = \frac{\sum w_i y_i}{\sum w_i}$. In the case of models that are approximated with variational methods, we made predictions with the approximate posterior mean of all unknown random variables.

In order to determine the relationship between predictive accuracy and dataset size, we performed cross-validation training with randomized subsets of the training data. We trained models from data subsampled at 5,000 increments of the full training set size, 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 were taken from the variational posterior: $\tilde{W} \sim q(W)$. This sample was then used to calculate the latent position of every variant: $\tilde{z}_i = \tilde{W}x_i$. Then, the approximate posterior predictive distribution of 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 fifty times to estimate the overall uncertainty of f_i for the variant.

4.6 Simulated biophysical allosteric model

Simulated data was generated using a biophysical model for allosteric transcriptional regulation (Razo-Mejia et al. 2018). That model relates the dose-response curve, $f(c)$, of an allosteric transcription factor to the biophysical parameters such as binding constants and free-energy differences between protein states:

$$f(c) = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\epsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}} \right)^{-1},$$

where c is the ligand concentration, K_A and K_I are the dissociation constants between the ligand and the transcription factor in the active and inactive states, respectively, $\Delta\epsilon_{AI}$ is the free energy difference between the active and inactive states, $\Delta\epsilon_{RA}$ is the free energy difference between specific and non-specific binding of the transcription factor to DNA, R is the number of transcription factor molecules in the cell, N_{NS} is the number of non-specific DNA binding sites, $\beta = (k_B T)^{-1}$, k_B is the Boltzmann constant, and T is the temperature. Simulated results were generated for the baseline

dose-response, $G_0 = f(0)$, the saturated dose-response, $G_\infty = \lim_{c \rightarrow \infty} f(c)$, and the sensitivity, EC_{50} :

$$G_0 = \left(1 + \frac{1}{1 + e^{-\beta\Delta\epsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}}\right)^{-1} \quad (20)$$

$$G_\infty = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{1 + e^{-\beta\Delta\epsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}}\right)^{-1} \quad (21)$$

$$EC_{50} = K_A \left(\frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left(\frac{\left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}}\right) + \left(\frac{K_A}{K_I}\right)^n \left(2e^{-\beta\Delta\epsilon_{AI}} + \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}}\right)\right)}{2\left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}}\right) + e^{-\beta\Delta\epsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\epsilon_{AI}}} \right)^{\frac{1}{n}} - 1 \right) \quad (22)$$

The values assumed for the wild-type parameters were: $K_A = 139 \mu\text{mol/L}$, $K_I = 0.53 \mu\text{mol/L}$, $\beta\Delta\epsilon_{AI} = 4.5$, $R = 150$, $n = 2$, $\beta\Delta\epsilon_{RA} = -15.3$, $N_{NS} = 4.6 \times 10^6$.

Simulated datasets were generated assuming a protein with 300 amino acid positions and six possible amino acid substitutions at each position.

Three different simulated datasets were generated:

- 1) $\mathcal{D}_1$: data for a 1D manifold, varying only the $\beta\Delta\epsilon_{AI}$ parameter
- 2) $\mathcal{D}_2$: data for a 2D manifold, varying $\beta\Delta\epsilon_{AI}$ and $\beta\Delta\epsilon_{RA}$
- 3) $\mathcal{D}_3$: data for a 3D manifold, varying $\beta\Delta\epsilon_{AI}$, $\beta\Delta\epsilon_{RA}$, and $\log_{10}(K_A)$

For each dataset a 300 amino-acid length protein was simulated with 6 possible substitutions for each amino acid position. Random mutational effects were assigned to each possible substitution using the following procedure: First, a mean effect was chosen for each position as a random draw from a normal distribution with zero mean and standard deviation of 1.5. Then, the effect of each specific substitution was determined as the mean effect at the corresponding position multiplied by

a random integer between -1 and 1 plus a random draw from a normal distribution with zero mean and standard deviation of 1.

For each simulated dataset, 100,000 simulated variants were chosen by first assigning the number of amino acid substitutions for each variant based on random draws from the empirical distribution from an experimental dataset (Tack et al. 2021). The positions and identities of simulated substitutions were then randomly chosen as unbiased random draws from the set of positions (without replacement), and possible substitutions (with replacement). The shifts of the biophysical parameters ($\beta\Delta\epsilon_{AI}$, $\beta\Delta\epsilon_{RA}$, and $\log_{10}(K_A)$) were then determined for each simulated variant by summing the effects of each substitution. The effects of changes to $\log_{10}(K_A)$ were much stronger than the effects of the other parameters, so the shift in $\log_{10}(K_A)$ for each variant in the 3rd dataset was scaled by a factor of 0.1. The biophysical parameters for each variant were determined by adding the shift for that parameter-variant combination to the wild-type parameter values listed above. Finally, equations 20–22 were used to calculate values for G_0 , G_∞ , and EC_{50} for each simulated variant.

Code availability

Source code is available at github.com/usnistgov/lantern.

Acknowledgements

We would like to thank Blaza Toman, Swarnavo Sarkar, and Dennis Leber for thoughtful feedback on this manuscript. PDT and ADP were supported through NRC Fellowships. Analysis performed in part on the NIST Nisaba HPC clusters.

Author Contributions

PDT and DJR conceived of the approach
PDT implemented LANTERN and ran analyses
DJR generated simulated biophysical landscapes
PDT, ADP, and DJR wrote the manuscript

7 Disclaimer

The authors declare no competing interests

Certain commercial equipment, instruments, or materials are identified to adequately specify experimental procedures. Such identification neither implies recommendation nor endorsement by the National Institute of Standards and Technology nor that the equipment, instruments, or materials identified are necessarily the best for the purpose.

Correspondence and requests for materials should be addressed to Peter Tonner (peter.tonner@nist.gov)

References

- Ai, Hui-wang et al. (2007). “Exploration of New Chromophore Structures Leads To the Identification of Improved Blue Fluorescent Proteins”. In: *Biochemistry* 46.20, pp. 5904–5910. DOI: [10.1021/bi700199g](https://doi.org/10.1021/bi700199g). URL: <https://doi.org/10.1021/bi700199g>.
- Anderson, Dave W et al. (2015). “Intermolecular Epistasis Shaped the Function and Evolution of an Ancient Transcription Factor and Its DNA Binding Sites”. In: *eLife* 4. DOI: [10.7554/elife.07864](https://doi.org/10.7554/elife.07864). URL: <https://doi.org/10.7554/elife.07864>.
- Angenent-Mari, Nicolaas M. et al. (2020). “A Deep Learning Approach To Programmable RNA Switches”. In: *Nature Communications* 11.1, p. 5057. DOI: [10.1038/s41467-020-18677-1](https://doi.org/10.1038/s41467-020-18677-1). URL: <https://doi.org/10.1038/s41467-020-18677-1>.
- Angermueller, Christof et al. (2016). “Deep Learning for Computational Biology”. In: *Molecular Systems Biology* 12.7, p. 878. DOI: [10.15252/msb.20156651](https://doi.org/10.15252/msb.20156651). URL: <https://doi.org/10.15252/msb.20156651>.
- Atlani-Duault, Laetitia et al. (2021). “Immune Evasion Means We Need a New Covid-19 Social Contract”. In: *The Lancet Public Health* 6.4, e199–e200. DOI: [10.1016/s2468-2667\(21\)00036-0](https://doi.org/10.1016/s2468-2667(21)00036-0). URL: [https://doi.org/10.1016/s2468-2667\(21\)00036-0](https://doi.org/10.1016/s2468-2667(21)00036-0).
- Balaprakash, Prasanna et al. (2019). “Scalable Reinforcement-Learning-Based Neural Architecture Search for Cancer Deep Learning Research”. In: *CoRR*. arXiv: [1909.00311](https://arxiv.org/abs/1909.00311) [cs.LG]. URL: [http://arxiv.org/abs/1909.00311v1](https://arxiv.org/abs/1909.00311).
- Bergstra, James et al. (2012). “Random Search for Hyper-Parameter Optimization”. In: *Journal of Machine Learning Research* 13.10, pp. 281–305. URL: <http://jmlr.org/papers/v13/bergstra12a.html>.
- Bishop, C.M. (1999). “Variational principal components”. In: *9th International Conference on Artificial Neural Networks: ICANN '99*. DOI: [10.1049/cp:19991160](https://doi.org/10.1049/cp:19991160). URL: <https://doi.org/10.1049/cp:19991160>.
- Blei, David M. et al. (2016). “Variational Inference: a Review for Statisticians”. In: *CoRR*. arXiv: [1601.00670](https://arxiv.org/abs/1601.00670) [stat.CO]. URL: [http://arxiv.org/abs/1601.00670v9](https://arxiv.org/abs/1601.00670v9).
- Blundell, Charles et al. (2015). “Weight Uncertainty in Neural Networks”. In: *arXiv preprint arXiv:1505.05424*. URL: <https://arxiv.org/pdf/1505.05424.pdf>.

Centers for Disease Control and Prevention (2021). *CDC COVID Data Tracker*. URL: <https://covid.cdc.gov/covid-data-tracker/#datatracker-home> (visited on 05/01/2021).

Ching, Travers et al. (2018). “Opportunities and Obstacles for Deep Learning in Biology and Medicine”. In: *Journal of The Royal Society Interface* 15.141, p. 20170387. DOI: [10.1098/rsif.2017.0387](https://doi.org/10.1098/rsif.2017.0387). URL: <https://doi.org/10.1098/rsif.2017.0387>.

Chure, Griffin et al. (2019). “Predictive Shifts in Free Energy Couple Mutations To Their Phenotypic Consequences”. In: *Proceedings of the National Academy of Sciences*, p. 201907869. DOI: [10.1073/pnas.1907869116](https://doi.org/10.1073/pnas.1907869116). URL: <https://doi.org/10.1073/pnas.1907869116>.

Domingo, Júlia et al. (2019). “The Causes and Consequences of Genetic Interactions (Epistasis)”. In: *Annual Review of Genomics and Human Genetics* 20.1, pp. 433–460. DOI: [10.1146/annurev-genom-083118-014857](https://doi.org/10.1146/annurev-genom-083118-014857). URL: <https://doi.org/10.1146/annurev-genom-083118-014857>.

Frampton, Dan et al. (2021). “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”. In: *The Lancet Infectious Diseases*. DOI: [10.1016/s1473-3099\(21\)00170-5](https://doi.org/10.1016/s1473-3099(21)00170-5). URL: [https://doi.org/10.1016/s1473-3099\(21\)00170-5](https://doi.org/10.1016/s1473-3099(21)00170-5).

Gardner, Jacob et al. (2018). “GPYtorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 7576–7586. URL: <http://papers.nips.cc/paper/7985-gpytorch-blackbox-matrix-matrix-gaussian-process-inference-with-gpu-acceleration.pdf>.

Guidotti, Riccardo et al. (2019). “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5, pp. 1–42. DOI: [10.1145/3236009](https://doi.org/10.1145/3236009). URL: <https://doi.org/10.1145/3236009>.

Harvey, William T. et al. (2021). “SARS-CoV-2 Variants, Spike Mutations and Immune Escape”. In: *Nature Reviews Microbiology*. DOI: [10.1038/s41579-021-00573-0](https://doi.org/10.1038/s41579-021-00573-0). URL: <https://doi.org/10.1038/s41579-021-00573-0>.

Hensman, James et al. (May 2015). “Scalable Variational Gaussian Process Classification”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon et al. Vol. 38. Proceedings of Machine Learning Research. San Diego,

California, USA: PMLR, pp. 351–360. URL: <http://proceedings.mlr.press/v38/hensman15.html>.

Hie, Brian et al. (2021). “Learning the Language of Viral Evolution and Escape”. In: *Science* 371.6526, pp. 284–288. DOI: [10.1126/science.abd7331](https://doi.org/10.1126/science.abd7331). URL: <https://doi.org/10.1126/science.abd7331>.

Hodcroft, Emma B. (2021). *CoVariants: SARS-CoV-2 Mutations and Variants of Interest*. URL: <https://covariants.org/> (visited on 01/04/2021).

Hoffman, Matthew D. et al. (2013). “Stochastic Variational Inference”. In: *Journal of Machine Learning Research* 14.4, pp. 1303–1347. URL: <http://jmlr.org/papers/v14/hoffman13a.html>.

Husain, Kabir et al. (2020). “Physical Constraints on Epistasis”. In: *Molecular Biology and Evolution* 37.10, pp. 2865–2874. DOI: [10.1093/molbev/msaa124](https://doi.org/10.1093/molbev/msaa124). URL: <https://doi.org/10.1093/molbev/msaa124>.

Jankowiak, Martin et al. (2019). “Parametric Gaussian Process Regressors”. In: *CoRR*. arXiv: [1910.07123](https://arxiv.org/abs/1910.07123) [stat.ML]. URL: <http://arxiv.org/abs/1910.07123v3>.

Jiménez-Luna, José et al. (2020). “Drug Discovery With Explainable Artificial Intelligence”. In: *Nature Machine Intelligence* 2.10, pp. 573–584. DOI: [10.1038/s42256-020-00236-4](https://doi.org/10.1038/s42256-020-00236-4). URL: <https://doi.org/10.1038/s42256-020-00236-4>.

Kandasamy, Kirthevasan et al. (2018). “Neural Architecture Search with Bayesian Optimisation and Optimal Transport”. In: *NeurIPS*.

Kingma, Diederik P et al. (2013). “Auto-Encoding Variational Bayes”. In: *CoRR*. arXiv: [1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML]. URL: <http://arxiv.org/abs/1312.6114v10>.

Kingma, Diederik P. et al. (2014). “Adam: a Method for Stochastic Optimization”. In: *CoRR*. arXiv: [1412.6980](https://arxiv.org/abs/1412.6980) [cs.LG]. URL: <http://arxiv.org/abs/1412.6980v9>.

Kinney, Justin B. et al. (2019). “Massively Parallel Assays and Quantitative Sequence-Function Relationships”. In: *Annual Review of Genomics and Human Genetics* 20.1, annurev-genom-083118–014845. DOI: [10.1146/annurev-genom-083118-014845](https://doi.org/10.1146/annurev-genom-083118-014845). URL: <https://doi.org/10.1146/annurev-genom-083118-014845>.

681 Kondrashov, Dmitry A. et al. (2015). “Topological Features of Rugged Fitness Landscapes in Se-
682 quence Space”. In: *Trends in Genetics* 31.1, pp. 24–33. DOI: [10.1016/j.tig.2014.09.009](https://doi.org/10.1016/j.tig.2014.09.009).
683 URL: <https://doi.org/10.1016/j.tig.2014.09.009>.

684 Lewis, Mitchell et al. (1996). “Crystal structure of the lactose operon repressor and its complexes
685 with DNA and inducer”. In: *Science* 271.5253, pp. 1247–1254.

686 Lipton, Zachary C (2018). “The Mythos of Model Interpretability: In machine learning, the concept
687 of interpretability is both important and slippery.” In: *Queue* 16.3, pp. 31–57. DOI: [10.1145/](https://doi.org/10.1145/3236386.3241340)
688 [3236386.3241340](https://doi.org/10.1145/3236386.3241340). URL: <https://doi.org/10.1145/3236386.3241340>.

689 Orr, H. Allen (2005). “The Genetic Theory of Adaptation: a Brief History”. In: *Nature Reviews*
690 *Genetics* 6.2, pp. 119–127. DOI: [10.1038/nrg1523](https://doi.org/10.1038/nrg1523). URL: <https://doi.org/10.1038/nrg1523>.

691 Otwinowski, J. et al. (2014). “Inferring Fitness Landscapes By Regression Produces Biased Esti-
692 mates of Epistasis”. In: *Proceedings of the National Academy of Sciences* 111.22, E2301–E2309.
693 DOI: [10.1073/pnas.1400849111](https://doi.org/10.1073/pnas.1400849111). URL: <https://doi.org/10.1073/pnas.1400849111>.

694 Otwinowski, Jakub (2018). “Biophysical Inference of Epistasis and the Effects of Mutations on
695 Protein Stability and Function”. In: *Molecular Biology and Evolution* 35.10, pp. 2345–2354.
696 DOI: [10.1093/molbev/msy141](https://doi.org/10.1093/molbev/msy141). URL: <https://doi.org/10.1093/molbev/msy141>.

697 Otwinowski, Jakub et al. (2018). “Inferring the Shape of Global Epistasis”. In: *Proceedings of the*
698 *National Academy of Sciences* 115.32, E7550–E7558. DOI: [10.1073/pnas.1804015115](https://doi.org/10.1073/pnas.1804015115). URL:
699 <https://doi.org/10.1073/pnas.1804015115>.

700 Paszke, Adam et al. (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning
701 Library”. In: *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach et al.
702 Curran Associates, Inc., pp. 8024–8035. URL: [http://papers.neurips.cc/paper/9015-](http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf)
703 [pytorch-an-imperative-style-high-performance-deep-learning-library.pdf](http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf).

704 Poelwijk, Frank J. et al. (2016). “The Context-Dependence of Mutations: a Linkage of Formalisms”.
705 In: *PLOS Computational Biology* 12.6, e1004771. DOI: [10.1371/journal.pcbi.1004771](https://doi.org/10.1371/journal.pcbi.1004771). URL:
706 <https://doi.org/10.1371/journal.pcbi.1004771>.

707 Pokusaeva, Victoria O. et al. (2019). “An Experimental Assay of the Interactions of Amino Acids
708 From Orthologous Sequences Shaping a Complex Fitness Landscape”. In: *PLOS Genetics* 15.4,
709 e1008079. DOI: [10.1371/journal.pgen.1008079](https://doi.org/10.1371/journal.pgen.1008079). URL: <https://doi.org/10.1371/journal.pgen.1008079>.

- Rasmussen, C.E. et al. (2006). *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning series. University Press Group Limited. ISBN: 9780262182539. URL: <http://gaussianprocess.org/gpml/>.
- Razo-Mejia, Manuel et al. (2018). “Tuning Transcriptional Regulation Through Signaling: a Predictive Theory of Allosteric Induction”. In: *Cell Systems* 6.4, 456–469.e10. DOI: [10.1016/j.cels.2018.02.004](https://doi.org/10.1016/j.cels.2018.02.004). URL: <https://doi.org/10.1016/j.cels.2018.02.004>.
- Rudin, Cynthia (2019). “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5, pp. 206–215. DOI: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x). URL: <https://doi.org/10.1038/s42256-019-0048-x>.
- Sailer, Zachary R. and Michael J. Harms (2017). “Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps”. In: *Genetics* 205.3, pp. 1079–1088. DOI: [10.1534/genetics.116.195214](https://doi.org/10.1534/genetics.116.195214). URL: <https://doi.org/10.1534/genetics.116.195214>.
- Sailer, Zachary R., Sarah H. Shafik, et al. (2020). “Inferring a Complete Genotype-Phenotype Map From a Small Number of Measured Phenotypes”. In: *PLOS Computational Biology* 16.9, e1008243. DOI: [10.1371/journal.pcbi.1008243](https://doi.org/10.1371/journal.pcbi.1008243). URL: <https://doi.org/10.1371/journal.pcbi.1008243>.
- Sarkisyan, Karen S. et al. (2016). “Local Fitness Landscape of the Green Fluorescent Protein”. In: *Nature* 533.7603, pp. 397–401. DOI: [10.1038/nature17995](https://doi.org/10.1038/nature17995). URL: <https://doi.org/10.1038/nature17995>.
- Sato, Takuya U. et al. (Feb. 2020). “Evolutionary Dimension Reduction in Phenotypic Space”. In: *Phys. Rev. Research* 2 (1), p. 013197. DOI: [10.1103/PhysRevResearch.2.013197](https://doi.org/10.1103/PhysRevResearch.2.013197). URL: <https://doi.org/10.1103/PhysRevResearch.2.013197>.
- Snelson, Edward et al. (2006). “Variable Noise and Dimensionality Reduction for Sparse Gaussian processes”. In: *ArXiv* abs/1206.6873.
- Snoek, Jasper et al. (2019). “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift”. In: *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach et al. Curran Associates, Inc., pp. 13969–13980. URL: <http://papers.nips.cc/paper/9547-can-you-trust-your-models-uncertainty-evaluating-predictive-uncertainty-under-dataset-shift.pdf>.

- 741 Solak, Ercan et al. (2003). “Derivative observations in Gaussian process models of dynamic sys-
742 tems”. In: *Advances in neural information processing systems*, pp. 1057–1064.
- 743 Starr, Tyler N., Allison J. Greaney, et al. (2020). “Deep Mutational Scanning of SARS-CoV-2
744 Receptor Binding Domain Reveals Constraints on Folding and Ace2 Binding”. In: *Cell* 182.5,
745 1295–1310.e20. DOI: [10.1016/j.cell.2020.08.012](https://doi.org/10.1016/j.cell.2020.08.012). URL: [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.cell.2020.08.012)
746 [cell.2020.08.012](https://doi.org/10.1016/j.cell.2020.08.012).
- 747 Starr, Tyler N. and Joseph W. Thornton (2016). “Epistasis in Protein Evolution”. In: *Protein*
748 *Science* 25.7, pp. 1204–1218. DOI: [10.1002/pro.2897](https://doi.org/10.1002/pro.2897). URL: [2897](https://doi.org/10.1002/pro.
749 <a href=).
- 750 Tack, Drew S et al. (2021). “The Genotype-phenotype Landscape of an Allosteric Protein”. In:
751 *Molecular Systems Biology* 17.3. DOI: [10.15252/msb.202010179](https://doi.org/10.15252/msb.202010179). URL: [15252/msb.202010179](https://doi.org/10.
752 <a href=).
- 753 Tenaillon, O. (2014). “The Utility of Fisher’s Geometric Model in Evolutionary Genetics”. In: *An-*
754 *nual Review of Ecology, Evolution, and Systematics* 45.1, pp. 179–201. DOI: [10.1146/annurev-](https://doi.org/10.1146/annurev-ecolsys-120213-091846)
755 [ecolsys-120213-091846](https://doi.org/10.1146/annurev-ecolsys-120213-091846). URL: [091846](https://doi.org/10.1146/annurev-ecolsys-120213-
756 <a href=).
- 757 Tokuriki, Nobuhiko et al. (2009). “Stability Effects of Mutations and Protein Evolvability”. In:
758 *Current Opinion in Structural Biology* 19.5, pp. 596–604. DOI: [10.1016/j.sbi.2009.08.003](https://doi.org/10.1016/j.sbi.2009.08.003).
759 URL: <https://doi.org/10.1016/j.sbi.2009.08.003>.
- 760 Valeri, Jacqueline A. et al. (2020). “Sequence-To-Function Deep Learning Frameworks for Engi-
761 neered Riboregulators”. In: *Nature Communications* 11.1, p. 5058. DOI: [10.1038/s41467-020-](https://doi.org/10.1038/s41467-020-18676-2)
762 [18676-2](https://doi.org/10.1038/s41467-020-18676-2). URL: <https://doi.org/10.1038/s41467-020-18676-2>.
- 763 Yang, Te-Tuan et al. (1998). “Improved Fluorescence and Dual Color Detection With Enhanced
764 Blue and Green Variants of the Green Fluorescent Protein”. In: *Journal of Biological Chemistry*
765 273.14, pp. 8212–8216. DOI: [10.1074/jbc.273.14.8212](https://doi.org/10.1074/jbc.273.14.8212). URL: [jbc.273.14.8212](https://doi.org/10.1074/
766 <a href=).

Figure 1: Interpretable modeling of genotype-phenotype landscapes. (a) LANtern takes as input genotype-phenotype measurements, where each genetic background (genotype) has 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 to be 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 ($\vec{z}^{(i)}, \vec{z}^{(j)}$). 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) Observed phenotypes are a non-linear function $f(z)$ of the latent mutational effect space, which ultimately results in a non-linear relationship between genotype and phenotype. (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 towards zero, and the smallest dimensionality necessary to explain the data is recovered. (f) Non-linear functions $f(z)$ are modeled with Gaussian process priors. An example GP posterior is shown, fit to observations (black dots, bars show observation uncertainty), where the solid blue line is the posterior predictive mean and the shaded region is the 95% credible region. Dotted lines show example function draws from this posterior.

Figure 2: Recovering biophysical landscapes and dimensions. (a) Simulating allosteric GPL landscapes of varying dimensionality. Randomized genotypes have latent mutational effects on one to three underlying biophysical constants (D_1 to D_3). The analytical model determines the allosteric parameters EC_{50} , G_{∞} , and G_0 for each resulting combinations of biophysical constants. LANtern learns a model for each dataset, without knowledge of the underlying biophysics. (b) Approximate posterior variance of independent dimensions for each allosteric simulations of varying latent dimensionality. Black line shows the prior distribution. To determine the dimensionality learned by LANtern on each simulated dataset, we counted the number of dimensions with posterior mean variance with less than $\sim 0.2\%$ prior probability (red line). (c) Correlation between learned latent dimensions and simulated biophysical parameters. (d) Learned two-dimensional surface $f(\mathbf{z})$ for EC_{50} in $\mathcal{D}_2$. 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 parameter value. (e) Rotation of the learned surface $f(\mathbf{z})$ in (d) to the true biophysical surface. As a function of biophysical parameters $\Delta\epsilon_{AI}$ and $\Delta\epsilon_{RA}$, the image shows the true biophysical landscape and the contours show the posterior mean of $f(\mathbf{z})$ from LANtern. The black vectors mark the rotation of the z_1 and z_2 dimensions to the native biophysical space. (f) Squared residual between posterior mean of $f(\mathbf{z})$ and true EC_{50} surface shown as contours rotated as in (e) overlaid on the sampling density of variants in the latent space. See section 4.6 for method details.

Figure 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 show the distribution of R^2 values while scatter points mark the value for individual folds. (b) Ten-fold CV predictive accuracy as a function of number of observations for LacI EC_{50} . (c) LANTERN root mean squared error (RMSE) as a function of mutational distance from the wild-type sequence. (d) RMSE versus the average posterior predictive uncertainty of LANTERN (see section 4.5) for varying mutational distance from the wild-type.

Figure 5: SARS-Cov2 joint binding and expression landscape. (a, b) Learned surface $f(z)$ for SARS-Cov2 binding (a) and expression (b). Contours show posterior mean of $f(z)$, scatter points are mean posterior z for individual variants colored by their observed mean value, and red crosses mark the wild-type origin. (c) Distribution of latent, single mutant effect directions along the first two dimensions of z , with *stability* axis (green) and *binding* axis (purple). (d, e) Joint binding and expression surfaces with identified axes overlaid. Contours are the mean of the variational posterior of $f(z)$ and the two-dimensional histograms mark the density of observed variants in z -space. (f) The predicted binding and expression as a function of the projection along both identified axes. (g) Structural association of the identified axes. SARS-Cov2 binding the ACE2 RBD and the SARS-Cov2 protein in isolation are shown. Colors indicate the average value of mutational effect projections onto the corresponding axis. Spheres indicate a strong average alignment ($\angle \leq 36^\circ$) between mutational effects at a residue and the corresponding axis. (h) Mutations associated with COVID-19 outbreaks (Hodcroft 2021). Individual mutational effects are shown as vectors. Contours and histogram are as in (d,e). (i) Predicted binding along the mutational effect direction of N501Y. Projection corresponds to the scale factor of the N501Y unit direction vector, solid black line is the posterior mean of $f(z)$, shaded regions is the 95% confidence interval, and 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% confidence interval of the single mutant measurement.

Figure 6: Local robustness and additivity of GPLs. (a) LacI EC₅₀ surface and slope along z_1 . Posterior of $f(\mathbf{z})$ is shown in blue and slope ($\frac{\partial}{\partial z_1} f(\mathbf{z})$) is shown in green. When the slope is zero, the surface is locally *robust* (purple box). (b) LacI EC₅₀ surface and curvature along z_1 . Posterior of $f(\mathbf{z})$ is shown in blue and curvature ($\frac{\partial^2}{\partial z_1^2} f(\mathbf{z})$) 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, though may be non-zero (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 EC₅₀ in multiple dimensions, calculated as the Laplacian ($\Delta f(\mathbf{z})$, see methods). The multidimensional equivalent of *additivity* is when $\Delta f(\mathbf{z}) = 0$. (f) The *additivity* of LacI EC₅₀. Values near one indicate $\Delta f(\mathbf{z})$ is close to zero.