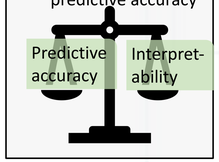
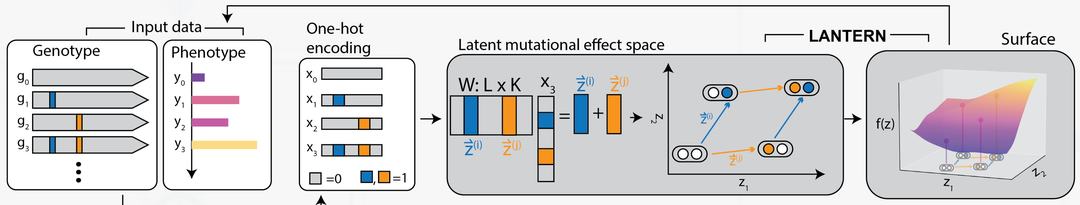
Interpretable Modeling of Genotype-Phenotype Landscapes with State-of-the-Art Predictive Power

Problem

- Large-scale genotype-phenotype measurements (GPLs) require predictive modeling to enable design new sequences
- Current approaches make a trade-off between model interpretability and predictive accuracy



Approach: Interpretable genotype-phenotype modeling with LANTERN

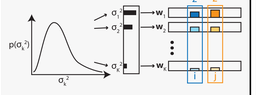


LANTERN models genotype-phenotype landscapes. Genotype-phenotype landscapes (GPLs) measure the joint relationship between genetic background (genotype) and downstream biological function (phenotype). LANTERN models this data through a predictive relationship from genotype to phenotype through two key components.

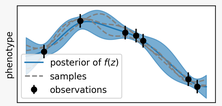
Latent mutational effect space (z). LANTERN models the effect of individual mutations in a latent space where mutations combine *additively*. Individual mutations (i and j) are represented by their corresponding mutational effect vectors ($z^{(i)}$ and $z^{(j)}$). LANTERN computes the **z** position of a variant combining multiple mutations (g_3 = {i, j}) through simple addition of both vectors ($z^{(i)} + z^{(j)}$).

Non-linear surface: f(z). LANTERN models a smooth, non-linear surface over the latent **z** space to predict the phenotype for each variant - given it's learned position in the latent space.

Inference



Dimensionality selection. A hierarchical prior on the variance of each latent dimension ensures LANTERN learns the dimensionality directly from the data



Gaussian process (GP) prior LANTERN places a GP prior on f(**z**) to learn the non-linear surface directly from that data without requiring a predefined parametric form

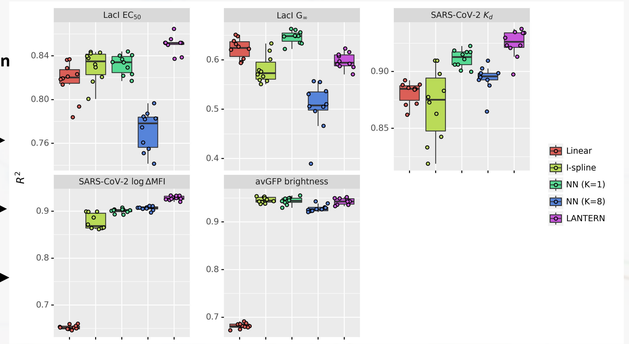
Prediction: LANTERN achieves unprecedented predictive power.

LacI
Mutations: 2,501
Phenotypes: EC50, Ginf
Observations: 47,462
Reference: Tack et al., 2021

SARS-CoV-2
Mutations: 4,002
Phenotypes: K_d, log MFI
Observations: 177,759
Reference: Starr et al., 2020

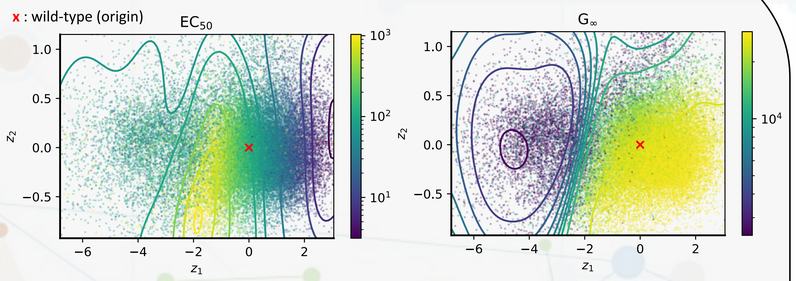
avGFP
Mutations: 1,879
Phenotypes: brightness
Observations: 54,025
Reference: Sarkisyan et al., 2015

10-fold Cross validation

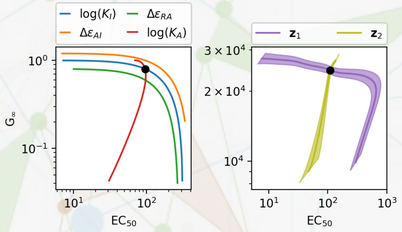


In a benchmark across large-scale GPL measurements, LANTERN equals or outperforms alternative predictive models - including deep neural networks - in ten-fold cross-validation.

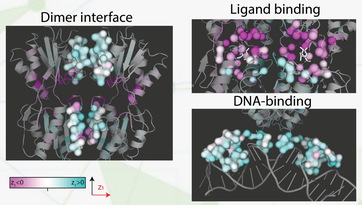
LANTERN discovers mechanisms of LacI allostery



Joint surface of LacI allostery. LANTERN learns a multidimensional surface predicting EC_{50} and G_∞ as a function of **z**. Individual variants are shown by their predicted location in **z** and colored by their observed phenotype. Contours show the predicted posterior mean of f(**z**).

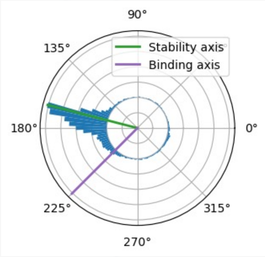


Landscapes align with biophysics. Along z_1 and z_2, LANTERN discovers distinct modes of variation between EC_{50} and G_∞ that match the impact of biophysical constants of an analytic model

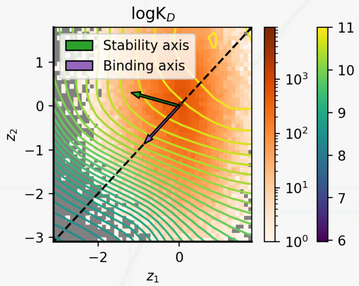


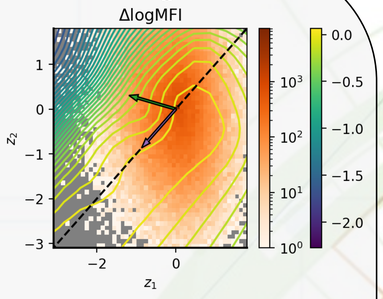
Structural association of latent dimensions. The strongest mutational effects along z_1 are clustered in structural regions of known function: the dimer interface, ligand-binding pocket, and DNA-binding domain

LANTERN decomposes structural details of SARS-CoV-2 ACE2 binding

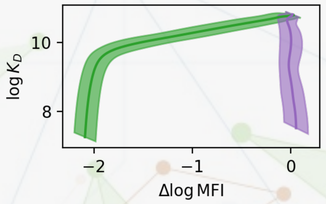


Distribution of mutational effects. The distribution of single mutant effects is enriched along a **stability** axis, while a near orthogonal **binding** axis is substantially less common





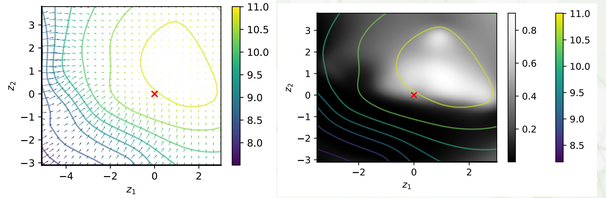
The joint landscape of SARS-CoV-2 binding (log K_d) and expression (Δlog MFI). Along the **stability** axis, changes in log K_d are combined with changes to Δlog MFI. Along the **binding** axis, however, the expression (Δlog MFI) stays near constant.



Distinct mechanisms of disrupted ACE2 binding. The predicted surface along the **stability** and **binding** axes show that decreased ACE2 binding can arise with (**stability**) or without (**binding**) decreased expression

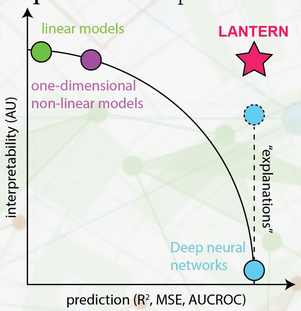
Structural associations of each axis. Structural regions of the SARS-CoV-2 receptor binding domain (RBD) are enriched for mutations aligned with the **stability** axis throughout the core RBD while mutations most strongly aligned with the **binding** axis are closer to the primary ACE contact region.

LANTERN quantifies robustness and additivity



Local properties of the surface f(**z**) quantify important behavior of phenotypes. For example, the gradient (left) describes that rate of change of the phenotype at each **z**. When the gradient is near-zero, we call this *robustness* (right) because the phenotype is robust to changes in **z**. We also quantify the *additivity* of the landscape similarly (not shown).

Impact: Accurate predictions with automatic interpretability



- Rather than making a trade-off between prediction accuracy and interpretability, LANTERN can provide state-of-the-art predictions that are inherently interpretable
- LANTERN enables discovery of underlying mechanisms driving complex GPL measurements
- LANTERN scales to millions of observations

National Institute of Standards and Technology
U.S. Department of Commerce

preprint: biorxiv.org/content/10.1101/2021.06.11.448129v1
source code: https://github.com/usnistgov/lantern

Authors: **Peter D Tonner**, Abe Pressman (poster #92), David Ross (poster #96)