

Karlin Labs



5' ATG CGT ACG ATC 3'

Mutagenesis

Mutagenesis is a canonical method for understanding and engineering proteins because it connects sequence changes to measurable properties. Proteins are coupled physical systems built from discrete sequences, so the effect of a substitution can depend on the surrounding residues and on other changes. Deep mutational scanning (DMS) addresses this experimental scaling bottleneck by combining pooled variant libraries, functional assays and DNA sequencing to compare many variants in parallel.¹

For green fluorescent protein (GFP), cells expressing different variants are sorted by brightness, and sequencing identifies the variants in each group so that a fluorescence level can be estimated for each sequence. Sarkisyan and colleagues measured 51,715 GFP protein sequences in this way, producing a substantial but incomplete view of how sequence variation shapes this phenotype.^{2,3}

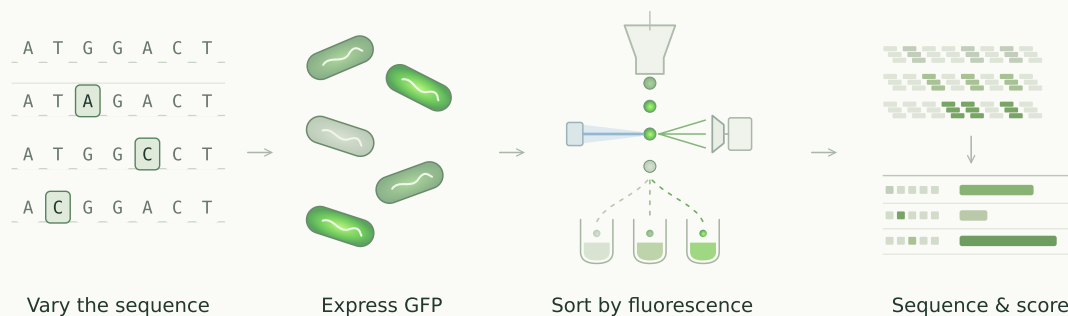


Figure 1. A GFP DMS experiment links sequence changes to fluorescence through expression, sorting and sequencing. The drawing is schematic, with three sorting bins shown for clarity.

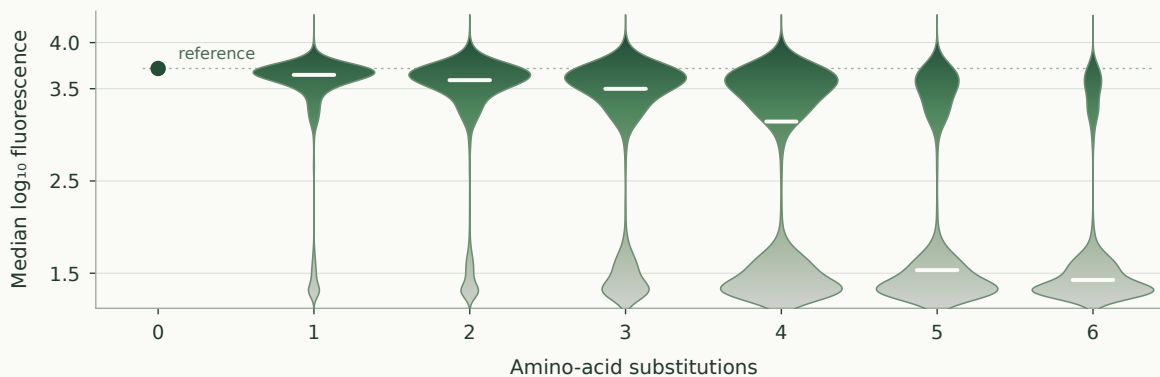


Figure 2. Measured GFP fluorescence from Sarkisyan et al.^{2,3} Width shows the distribution of log₁₀ brightness within each substitution count, with white marks at the medians and the dotted line at the reference. Distributions are normalized separately. Dim fluorescence alone does not identify its molecular cause.

For a 238-residue protein, choosing k positions and one of 19 alternatives at each gives $\binom{238}{k} 19^k$ possible variants: 4,522 single substitutions, 10,181,283 double substitutions and 15,217,624,324 triple substitutions. Even a large DMS library covers only part of this space, while interactions between changes make combination phenotypes harder to infer from single-mutant measurements.

The model

We fit Quadrant-0 to one protein, one reference sequence and one experimental readout, using measured substitutions to predict missing alternatives at positions already sampled. The assay supplies relational data: substitutions share chemical properties, and their measured effects reveal how those properties matter at different sites. Estimating these relationships before fitting the final predictor gives each candidate a description grounded in both chemistry and experimental evidence.

Featurization translates a candidate substitution into numerical descriptors of its position, original and replacement residues, and changes in charge, side-chain volume and hydrophobicity. If $c(a)$ gives the properties of residue a , then $\delta_{ia} = c(a) - c(r_i)$ describes the chemical change from the reference residue r_i . Measurements at the site help estimate how that change affects the chosen phenotype.

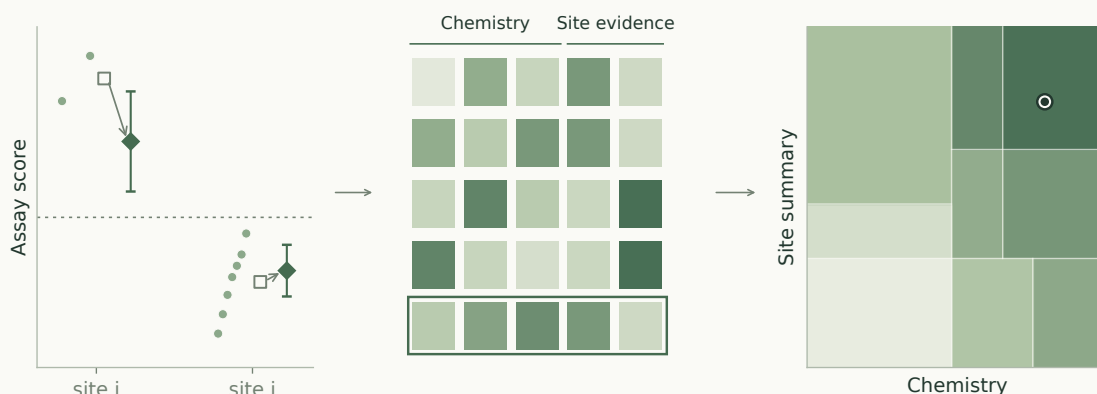


Figure 3. Left, site estimates move toward a shared fit when evidence is sparse. Center, chemical descriptors and site summaries form the candidate's features. Right, the tree ensemble partitions this space to predict the phenotype. Schematic values.

We can represent these relationships through a hierarchical regression in which shared terms describe the overall connection between chemistry and phenotype, while site-specific terms allow each position to have its own baseline and sensitivity:

$$y_{ia} = \mu + \alpha_i + (\beta + u_i)^\top \delta_{ia} + \epsilon_{ia}$$

Here, y_{ia} is the measurement for residue a at site i , while α_i and u_i describe departures from the shared baseline μ and slopes β . The residual ϵ_{ia} captures variation left unexplained. Bayesian partial pooling pulls uncertain site estimates toward the shared fit while allowing well-supported sites to depart from it, so posterior estimates, their uncertainty and observation counts can join the chemical descriptors as features.

A boosted tree ensemble then combines these features into a phenotype prediction.⁷ Each tree makes successive comparisons between feature values, while boosting adds trees that improve the current fit, giving a final score equal to a baseline plus their weighted outputs.

Our proposed contribution is to construct these features from the assay itself, then simplify them before refitting. Comparing chemical descriptors alone with the added experimental summaries and the reduced representation tests which stages improve prediction.



Interpreting the model

Because several descriptors can encode overlapping information, we center and scale the training features before using principal component analysis (PCA) to describe their shared variation.⁶ The resulting loadings identify groups of related descriptors, suggesting smaller representations that can be tested by refitting the predictor.

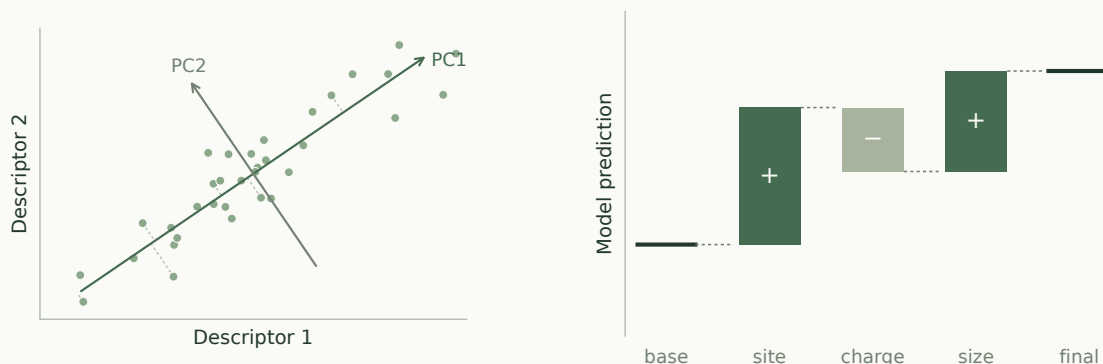


Figure 4. Left, principal axes describe variation in the input features. Right, positive and negative SHAP contributions add to a baseline to recover a prediction. Both views are schematic.

For a standardized matrix $\tilde{Z}$ with n variant rows, the covariance C has eigenvectors V and eigenvalues Λ . Selecting the first r eigenvectors in decreasing order of eigenvalue gives V_r , which projects a standardized candidate row $\tilde{z}$ to coordinates t :

$$C = \frac{\tilde{Z}^\top \tilde{Z}}{n-1} = V \Lambda V^\top, \quad t = \tilde{z} V_r$$

A large eigenvalue marks a direction with substantial variation across variants, but this need not imply predictive relevance because PCA does not use the phenotype. We therefore compare reduced feature sets on validation data, favoring original descriptors when they preserve accuracy and keep explanations tied to recognizable properties.

SHAP complements this view by decomposing a fitted prediction into a baseline and feature contributions, showing how the model combines substitution chemistry and site evidence.⁴ Correlated features can share attribution, and these contributions describe the fitted model rather than molecular causation. Together, PCA and SHAP guide which features to examine, while validation determines whether removing them preserves accuracy.

Phenotype-derived summaries require cross-fitting, with each training example receiving estimates computed without its fold and each held-out example using only the corresponding training partition. Scaling and PCA are fitted within the same boundary, while feature selection and tuning use inner validation folds so that the final test set remains untouched.

The same recipe can be refit across proteins and readouts, with accuracy and calibration assessed alongside fit time, memory, throughput and cost. Once features are available, scoring K candidates with T trees of maximum depth d takes $\mathcal{O}(KTd)$ node evaluations. Recording the datasets, splits, seeds, configurations and hardware makes these comparisons reproducible, while error measured against training-set size reveals how predictive performance scales with experimental coverage.



Results

GFP baseline

The saved GFP comparison fits 37,236 variants, reserves 4,137 for validation and holds out 10,342 for testing, stratified by mutation count. All seven methods below are evaluated on the same 8,507 test variants above the assay’s measurement floor.⁵

Model	Spearman ρ $\uparrow$	RMSE $\downarrow$
Quadrant-0	0.8727	0.2574
Reference ESM-2 interpolator	0.8688	0.2520
Additive curve	0.8684	0.2849
Shared pair factors	0.8286	0.3831
Sequence-embedding Ridge	0.8123	0.4973
Sequence-embedding MLP	0.7991	0.4358
Sequence-embedding linear	0.7840	0.4625

Table 1. Predictions for held-out GFP variants. Higher Spearman correlation indicates better ranking, while lower RMSE indicates smaller errors in log-brightness. Including the 1,835 floor-censored test observations gives a full-set Spearman correlation of 0.8717 and RMSE of 0.2542 for our saved fit.

Our fit has the highest rank correlation and a 9.7% lower RMSE than the additive curve, using assay measurements without a pretrained protein-language-model embedding. The ESM-2 reference uses the same GFP split but adds a frozen representation learned from a much larger sequence corpus.¹² This prior may help explain its slightly lower RMSE, although the comparison does not isolate that effect. Lower RMSE rewards closeness to measured brightness, whereas our higher Spearman correlation reflects better ordering of variants.

These results concern held-out variants that may combine substitutions. Quadrant-1 aims to extend measured-site interpolation to unmeasured positions and reliable interaction prediction, which requires evaluating departures from additivity directly.

For two substitutions, epistasis is the double-mutant effect relative to the reference minus the sum of the single-mutant effects:

$$\epsilon_{ij} = y_{ij} - y_i - y_j + y_0$$

Here, y_0 is the reference phenotype, y_i and y_j are single-mutant values, and y_{ij} is the double-mutant value. The quantity depends on the phenotype scale, which must be fixed when comparing observed and predicted interactions.



Results

Nipah

Published Nipah Malaysia scans add cell entry beside the GFP comparison.^{13,14} The receptor-binding protein (RBP), residues 71 to 602, and the fusion protein (F), residues 29 to 481, are fit separately on entry into CHO cells expressing bat ephrin-B2. The two proteins do not share a head, and neither is combined with GFP. The score is the published log₂-style effect versus wild type, and a negative value means worse entry. An unseen amino acid is a substitution held out at a position where other amino acids were measured, the measured-site task for Quadrant-0. A withheld site is a position absent from training, the unmeasured-position task for Quadrant-1. No double mutant was measured.

Comparison	Spearman ρ $\uparrow$	RMSE $\downarrow$	NLL $\downarrow$
RBP, unseen amino acid, chemistry	0.710	0.915	0.392
RBP, unseen amino acid, edited-site	0.724	0.866	0.357
RBP, withheld site, chemistry	0.457	1.241	0.760
RBP, withheld site, edited-site	0.475	1.234	0.711
F, unseen amino acid, chemistry	0.723	1.126	0.617
F, unseen amino acid, sequence head	0.742	1.061	0.560
F, withheld site, chemistry	0.459	1.461	0.881
F, withheld site, edited-site	0.418	1.558	0.944

Table 2. Held-out single substitutions for Nipah cell entry, condition CHO-bEFNB2. Higher Spearman correlation means better agreement in ranking the published score. Lower RMSE means smaller error in score units. NLL is the mean Gaussian negative log likelihood of that score, with the constant term omitted. Unseen-amino-acid rows contain 2,068 RBP variants and 1,812 F variants. Withheld-site rows contain 1,965 and 1,724. Edited-site values are the median prediction of four seeds. The F sequence head reads a frozen ESM-2 encoding of the whole mutant through a width-32 layer.

At a measured site, an untested amino acid can be ranked. Chemistry uses the substitution, nearby reference residues and, on RBP, a bit recording a match to Hendra virus. The edited-site head adds the frozen ESM-2 state at the changed position and reports the median of four seeds. Other measurements at the site, excluding the query, reach 0.724 on RBP and 0.728 on F. A width-32 head on the complete mutant reaches 0.742 on F and 0.716 on RBP. RMSE from 0.866 to 1.126 means the ranking is usable and the entry value is approximate.

A site absent from training stays near Spearman correlation 0.46. The edited-site head moves RBP to 0.475, and each of its four seeds is at least 0.462; on F the same head falls to 0.418. An evidence head needs other measurements at the query site, so a fully withheld site gives it none. On withheld F sites that head ranks at 0.471 and raises RMSE to 1.633. Encoding the whole mutant leaves the gap.

The receptor is *Pteropus alecto* bat ephrin. Human ephrin, antibody escape, binding, stability and designed combinations lie outside this table. A score counts only against a published value withheld from training. These are the saved chemistry and embedding fits, separate from the GFP models in Table 1.

Next Steps

Extending the recipe from fluorescence to binding, activity or expression requires refitting it to the relevant assay, while changing the perturbation also changes the representation. Nucleotide substitutions require base identity and sequence context, as in saturation genome-editing studies.⁸ Alanine scanning instead samples one replacement across positions, revealing sensitivity to alanine without supplying the full range of amino-acid substitution effects.⁹



Figure 5. Left, an alanine scan fills one row of a substitution map. Center, a nucleotide scan uses a four-base alphabet. Right, a CRISPR screen links interventions to cellular outcomes and measured phenotypes. Schematic coverage and values.

In CRISPR screens, the input may describe an assigned intervention rather than a verified sequence change. Perturb-seq pairs interventions with single-cell expression, but guide assignment alone establishes neither the resulting genotype nor perturbation efficacy, making cell context and evidence of an effective perturbation important to interpretation.^{10,11} Extending the recipe to image-derived or time-dependent phenotypes would also require outputs and error measures suited to those observations.

Descartes

Descartes would use an LLM to autonomously coordinate a portfolio of assay-specific Quadrant models, preparing datasets, scheduling fits and checking held-out performance before managing predictions and updates. Keeping each job's data, configuration and evaluation record would let the same system operate across proteins and readouts.

The data layer

We see two linked opportunities: models for biological customers and experimentally grounded datasets for frontier-model developers. Customer partnerships or assays commissioned through CROs would supply measurements, from which models could interpolate a defined collection of unmeasured variants. Distributed datasets would distinguish measured from predicted labels and include assay context and uncertainty estimates, with reuse governed by the source agreement.

Areas of interest

Cursory data partnerships with A-Alpha Bio, Twist and Ginkgo on antibody affinity and specificity, polyreactivity, developability, cross-reactivity and enzyme activity, for Quadrant-0 and working toward Quadrant-1.

References

1. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* **11**, 801-807 (2014). doi:10.1038/nmeth.3027
2. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397-401 (2016). doi:10.1038/nature17995
3. Author-deposited GFP data, *Figshare* (2016), CC BY 4.0. Figure 2 replots measured median brightness after excluding stop-containing sequences. doi:10.6084/m9.figshare.3102154
4. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67 (2020). doi:10.1038/s42256-019-0138-9
5. Research repository. Saved GFP comparison and evaluation records, accessed 24 October 2026. Numerical provenance and figure-generation sources accompany this brief.
6. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions A* **374**, 20150202 (2016). doi:10.1098/rsta.2015.0202
7. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**, 1189-1232 (2001). doi:10.1214/aos/1013203451
8. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217-222 (2018). doi:10.1038/s41586-018-0461-z
9. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-1085 (1989). doi:10.1126/science.2471267
10. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016). doi:10.1016/j.cell.2016.11.038
11. Papalexi, E. et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics* **53**, 322-331 (2021). doi:10.1038/s41588-021-00778-2
12. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023). doi:10.1126/science.ade2574
13. Larsen, B. B. et al. Functional and antigenic landscape of the Nipah virus receptor-binding protein. *Cell* **188**, 2480-2494.e22 (2025). doi:10.1016/j.cell.2025.02.030
14. Larsen, B. B. et al. Functional and antigenic constraints on the Nipah virus fusion protein. *Proceedings of the National Academy of Sciences* **123**, e2529505123 (2026). doi:10.1073/pnas.2529505123

