OXFORD (GIGA)^n SCIENCE

GigaScience, 2024, **13**, 1–17
DOI: 10.1093/gigascience/giae073
Research

Discovering genotype–phenotype relationships with machine learning and the Visual Physiology Opsin Database (*VPOD*)

Seth A. Frazer [1], Mahdi Baghbanzadeh [2], Ali Rahnavard [2], Keith A. Crandall [2,3], and Todd H. Oakley [1,*]

[1]Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106, USA
[2]Computational Biology Institute, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052, USA
[3]Department of Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20012, USA
*Correspondence address. Todd H. Oakley, Department of Ecology, Evolution, and Marine Biology, University of California Santa Barabara, Santa Barbara, California 93106, USA. E-mail: oakley@ucsb.edu

Abstract

Background: Predicting phenotypes from genetic variation is foundational for fields as diverse as bioengineering and global change biology, highlighting the importance of efficient methods to predict gene functions. Linking genetic changes to phenotypic changes has been a goal of decades of experimental work, especially for some model gene families, including light-sensitive opsin proteins. Opsins can be expressed *in vitro* to measure light absorption parameters, including λ_{max}—the wavelength of maximum absorbance—which strongly affects organismal phenotypes like color vision. Despite extensive research on opsins, the data remain dispersed, uncompiled, and often challenging to access, thereby precluding systematic and comprehensive analyses of the intricate relationships between genotype and phenotype.

Results: Here, we report a newly compiled database of all heterologously expressed opsin genes with λ_{max} phenotypes that we call the Visual Physiology Opsin Database (*VPOD*). *VPOD_1.0* contains 864 unique opsin genotypes and corresponding λ_{max} phenotypes collected across all animals from 73 separate publications. We use *VPOD* data and *deepBreaks* to show regression-based machine learning (ML) models often reliably predict λ_{max}, account for nonadditive effects of mutations on function, and identify functionally critical amino acid sites.

Conclusion: The ability to reliably predict functions from gene sequences alone using ML will allow robust exploration of molecular-evolutionary patterns governing phenotype, will inform functional and evolutionary connections to an organism's ecological niche, and may be used more broadly for *de novo* protein design. Together, our database, phenotype predictions, and model comparisons lay the groundwork for future research applicable to families of genes with quantifiable and comparable phenotypes.

Keywords: machine learning, regression, compiled database, genotype–phenotype relationships, predicting phenotypes, spectral sensitivity, color-vision, opsins, imputation

Key Points:

- We introduce the Visual Physiology Opsin Database (*VPOD_1.0*), which includes 864 unique animal opsin genotypes and corresponding λ_{max} phenotypes from 73 separate publications.
- We demonstrate that regression-based machine learning models can reliably predict λ_{max} from gene sequence alone, predict nonadditive effects of mutations on function, and identify functionally critical amino acid sites.
- We provide an approach that lays the groundwork for future robust exploration of molecular-evolutionary patterns governing phenotype, with potential broader applications to any family of genes with quantifiable and comparable phenotypes.

Introduction

Although critical to progress in drug and vaccine design [1–3], responses to climate change [4–8], and bioengineering [4, 9–11], accurately predicting gene function from sequences remains a significant challenge. While there are many ways to elucidate genotype–phenotype relationships experimentally, including deep mutational scanning, and *in vitro* heterologous expression with phenotyping, these techniques are often tedious and cost-prohibitive, especially when applied to broad comparative studies of gene families. In addition, accurately predicting the phenotype of a protein using computational methods alone is challenging because of data gaps and the sheer complexity of possible relationships between genes and phenotypes, including epistasis and the nonadditive effects of different mutations. Machine learning (ML) is gaining traction for its potential broad biological applications, accessibility, and faster speeds, especially in biological contexts where phenotype data are abundant and quantifiable. Here, classical regression and classification algorithms are sometimes used to train models for phenotype predictions using

Received: February 15, 2024. **Revised:** June 25, 2024. **Accepted:** September 1, 2024

genotype–phenotype data [12, 13], while deep learning models can be used to integrate heterogeneous multilayered omics and environmental data for establishing higher-dimensional genotype–phenotype connections [14, 15] or *de novo* protein design [16]. In broader biological contexts, ML models often inform laboratory experiments to predict directional evolution of diseases and their variants [17–19] or to automate image sorting and animal identification from camera trap data [20–22]. In all cases, ML models are a worthwhile long-term investment for genotype–phenotype studies because models can iteratively improve as empirical data accumulate over time.

Such accumulation of important information is exemplified by decades of laboratory work that has led to significant progress in understanding the genetic basis of phenotypic changes for model gene families such as opsins. Opsins are a family of G-protein coupled receptors (GPCR) that bind to a retinal chromophore. The 2 units together, opsin and chromophore, form visual pigments that absorb photons [23]. Opsins have crucial roles in many organismal functions, including circadian rhythms, phototaxis, and image-forming color vision. A critical opsin phenotype is spectral sensitivity—the range of wavelengths to which a gene or organism is sensitive. The main parameter of opsin spectral sensitivity is λ_{max}, the wavelength of light (in nm) with maximal absorbance [24]. Common methods of characterizing spectral sensitivities and λ_{max} include organ-level electroretinograms (ERGs) [25–27], cell-level microspectrophotometry (MSP) [28–32], purification of heterologously expressed opsins followed by spectrophotometry [33], and heterologous action spectroscopy using light response assays for opsins expressed in immortalized cell lines [34]. Different opsins are tuned by changes in amino acid sequences to respond to different wavelengths of light, and many previous studies have expressed experimentally mutated opsins and measured spectral sensitivities to establish genotype–phenotype connections [34–38]. Although other factors sometimes affect spectral responsiveness, including the type of chromophore to which an opsin is covalently bound (11-*cis* retinal or 11-*cis*-3,4-didehydro retinal) [39, 40], opsins provide a rare case where an intrinsic molecular function extends rather directly to organismal phenotypes, especially those involving color sensitivity. Despite opsins being a well-studied system with an extensive backlog of published literature, some previous authors expressed doubts that sequence data alone could provide reliable computational predictions of λ_{max} phenotypes [41–44]. At the same time, some λ_{max} predictions showed promise, although on the limited scale of vertebrate cone visual pigments via atomistic molecular simulations [45, 46]. Furthermore, only the nonanimal, microbial, or type 1 (T1) opsins have been systematically cataloged and used to examine genotype–phenotype predictive power of ML models [47, 48]. While some researchers have made significant efforts to compile peak sensitivity data for terrestrial animal photopigments [49] and taxon-specific light-sensitivity data for groups like frogs [50, 51] and ray-finned fishes [52, 53], these efforts currently lack direct links to genetic data that are essential for our current study. Consequently, the extensive data on genotype–phenotype associations of animal opsins remain disorganized, decentralized, often in noncomputer readable formats within older literature, and underanalyzed computationally.

Here, we report a genotype–phenotype database for animal opsins called the Visual Physiology Opsin Database (*VPOD*). We used standard literature searches to compile all heterologously expressed animal opsin genes with spectral sensitivity measurements. We used this newly compiled and harmonized database to evaluate ML methods for connecting genotypes and phenotypes. We created 11 subsets of the overall database to examine factors that impact the reliability and performance of ML models and briefly compared ML predictions to phylogenetic imputation [54, 55]. We also examined whether ML can predict intragenic epistasis, and we predicted amino acid sites particularly important for changing λ_{max}. Using our database of 864 unique opsin sequences and corresponding λ_{max} values, we show ML models trained on opsin data accurately predict the λ_{max} of opsins from genetic data alone (highest $R^2 = 0.968$ with a lowest mean absolute error [MAE] of 6.56 nm), especially when ample and diverse training data are available. ML also predicts some known effects of epistatic mutations on λ_{max}. Finally, ML models identify several sites that cause shifts in λ_{max} (e.g., "spectral tuning sites") and sites known to be structurally important, even in the absence of mutant data in training. When training data are sufficient, these results support the use of ML as a reliable and efficient predictor of λ_{max} for previously uncharacterized opsins, as a tool for identifying candidate spectral tuning sites and epistatic interactions, and as a more general method for linking gene sequences and phenotypes.

Methods

Compiling a genotype–phenotype database for animal opsins

We collected λ_{max} data for opsins using typical literature review/search methods, with search engine, keywords, and date of access documented in the "*litsearch*" table of the *VPOD* database (RRID:SCR_025668). We cataloged all usable papers with λ_{max} data in the "*references*" table of *VPOD*, recording DOI and a key to link to the search that found the paper. We documented the details of heterologous expression experiments in the "*heterologous*" table, including species, GenBank accession number for the sequence, mutation(s) (if applicable) using a machine-readable notation, λ_{max}, cell type for expression (e.g., HEK293, COS1, etc.), protein purification method, type of spectrum (e.g., dark or difference spectrum), and a key to link to the corresponding literature source. Note, we did not record the chromophore used to reconstitute the purified opsin protein because 11-*cis* retinal is the standard and all instances thus far recorded in the "*heterologous*" table are from experiments using 11-*cis* retinal (although future iterations of VPOD could record these details if data with alternative chromophores become available). We input opsin genetic data in an "*opsins*" table, recording opsin gene family names (e.g., long-wave sensitive = LWS, short-wave sensitive = SWS1, etc.). We also included specific "*gene names*" (where applicable), phylum, class, species information, accession number, DNA sequence, amino acid sequence, and the database from which sequences were retrieved (e.g., NCBI). We re-created all mutant and chimeric (e.g., 1 or more transmembrane domains of the mutant copied from a different sequence to replace the original) opsin sequences based on literature descriptions using a pair of Python scripts (*mutagenesis.py* and *chimeras.py*) available on our GitHub [56]. We added all heterologously expressed opsins from the literature to *VPOD*; we call this version of the database *VPOD_1.0*. We refer to heterologous data as *VPOD_het_1.0*, which will allow for future additions to the database to link specific opsin sequences to λ_{max} values established with methods other than heterologous expression, including microspectrophotometry or other methods. During the course of manuscript review, we found and entered 259 new heterologously expressed opsins into *VPOD*, an update we call *VPOD_1.1* (Fig. 1). We decided to keep results from *VPOD_1.0* in the main text

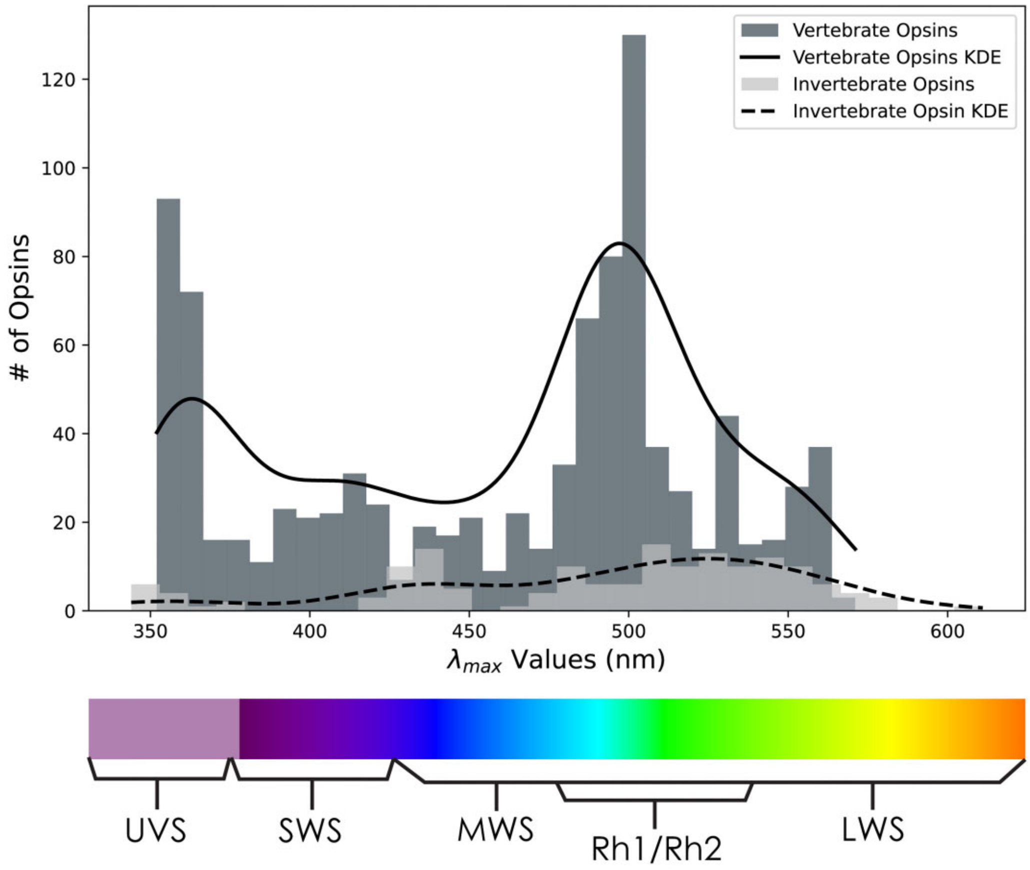

Figure 1: Histogram distributions of vertebrate and invertebrate opsins and absorbance data—λ_{max}—from *VPOD_het_1.1* with a scaled kernel density estimate (KDE) curves overlaid to better visualize the general shape and characteristics of our λ_{max} distributions. Note an obvious data bias for vertebrate opsins, especially those with λ_{max} values between 350–375 nm and 480–510 nm, probably due to focal research on UVS and Rh1 opsins.

because the new data points did not drastically alter any model performances. We also provide this table of performance metrics for *VPOD_1.1* (Supplementary Material 1 (S1)). Therefore, all tests and figures should still be assumed to use *VPOD_1.0* data unless stated otherwise.

Training ML models with *deepBreaks*

We performed all data preprocessing, including data extraction, sequence alignments, and formatting, in the Jupyter notebooks "*opsin_model_wf.ipynb*," available on GitHub. We used 2 multiple sequence alignment methods, MAFFT (RRID:SCR_011811) [57] and MUSCLE (RRID:SCR_011812) [58], and a version of both alignments with a Gblocks (RRID:SCR_015945) [59] refinement (for a total of 4 alignments), all set to their default parameters to begin to test the sensitivity of model performance to different alignments. We then trained various ML models employing a custom version of *deepBreaks* [60], an ML tool designed for exploring genotype–phenotype associations. *deepBreaks* takes aligned genotype data (DNA, RNA, amino acid) and some measure(s) of corresponding continuous or categorical phenotype data as input to train ML models. *deepBreaks* uses one-hot encoding to convert amino acid sequences into numerical values. One consequence of this encoding is any amino acids at a given position in the alignment, which are not present at that position in any training data, will be treated equivalently as unseen. For example, cases of only A and V at a highly conserved site in the training set that are presented with a sequence with T at that site will be considered as no A and no V. The models cannot distinguish the input whether it is T or other unseen amino acids at that site. The results produced by *deepBreaks* encompass a compilation of 12 regression ML models [60], showcasing 10 metrics of cross-validation performance (ranked by R^2) and a feature importance report derived from the top-performing models that ranks amino acid positions by their relative importance to each model (from 0.0–1.0, with 1.0 being a site with the highest relative importance) for the phenotype in question (λ_{max}). The metrics used to determine these relative importance scores of each position vary based on the structure and output of the algorithms used for model training. For example, xgboost [61] and LightGBM [62, 63] use the number of times a feature appears in a tree as a proxy for importance [60], while AdaBoost [64] and random forest [65, 66], use Gini importance, which quantifies a feature's contribution to improving prediction accuracy [60, 67, 68]. For a more detailed explanation on how position importance scores are calculated for different models, refer to the "*Interpretation*" heading under the methods section of the *deepBreaks* publication [60]. In addition to R^2, *deepBreaks* reports the MAE, mean absolute percent error (MAPE), mean square error (MSE), and root mean square error (RMSE) for each of the 12 ML models. We evaluated the performance of algorithms based on their relative ranks to look for patterns in which algorithms performed better for different data subsets and approaches. *deepBreaks* also produces a set of distribution box plots (default is 100) to visualize phenotypes (λ_{max}) associated with a particular amino acid identity at a site of interest, ordered alphabetically.

Table 1: Performance metrics across opsin subsets and top-performing models

Name	Data subset version	# sequences	Top ML algorithm	R^2[a]	MAE (nm)[b]	MAPE (%)[b]	MSE[a]	RMSE[a]
Whole dataset	*VPOD_wds_het_1.0*	864	LGBM	0.947	7.47	1.71	207	13.8
All wild types	*VPOD_wt_het_1.0*	318	Bayesian Ridge	0.902	10	2.18	297	16.5
All mutants	*VPOD_mut_het_1.0*	546	LGBM	0.951	7.89	1.86	194	13.4
Vertebrates	*VPOD_vert_het_1.0*	721	LGBM	0.968	6.56	1.49	111	10.3
WT vertebrates	*VPOD_wt_vert_het_1.0*	274	GBR	0.961	5.46	1.18	82.1	8.36
Invertebrates	*VPOD_inv_het_1.0*	143	LGBM	0.814	14.7	3.22	614	23.1
Rods	*VPOD_rod_het_1.0*	352	Bayesian Ridge	0.834	3.51	0.71	27.7	5.04
WT Rods	*VPOD_wt_rod_het_1.0*	157	GBR	0.783	3.57	0.72	31.9	5.11
MWS/LWS	*VPOD_mls_het_1.0*	91	XGB	0.677	8.77	1.82	317	15
UVS/SWS	*VPOD_uss_het_1.0*	280	GBR	0.821	8.02	2.06	200	13.6
WT UVS/SWS	*VPOD_wt_uss_het_1.0*	66	Adaboost	0.865	7.79	1.87	152	10.6
T1 opsins	*Karyasuyama_T1_ops*	884	Random Forest	0.804	9.41	1.76	186	13.5

[a] R^2, mean square error (MSE), and root mean square error (RMSE) are often interpreted as direct measures of comparing/analyzing model performance and used as training loss terms of the objective function—which measures how well the model fits the training data. One has to often balance between this and the regularization term, which controls the complexity of the model. Thus, a high performance is both simple and predictive, a trade-off referred to as the "*bias-variance*" trade-off.
[b] Mean absolute error (MAE) and mean absolute percent error (MAPE) are in relation to the absolute error λ_{max} predictions and interpreted in the same units of "nm."

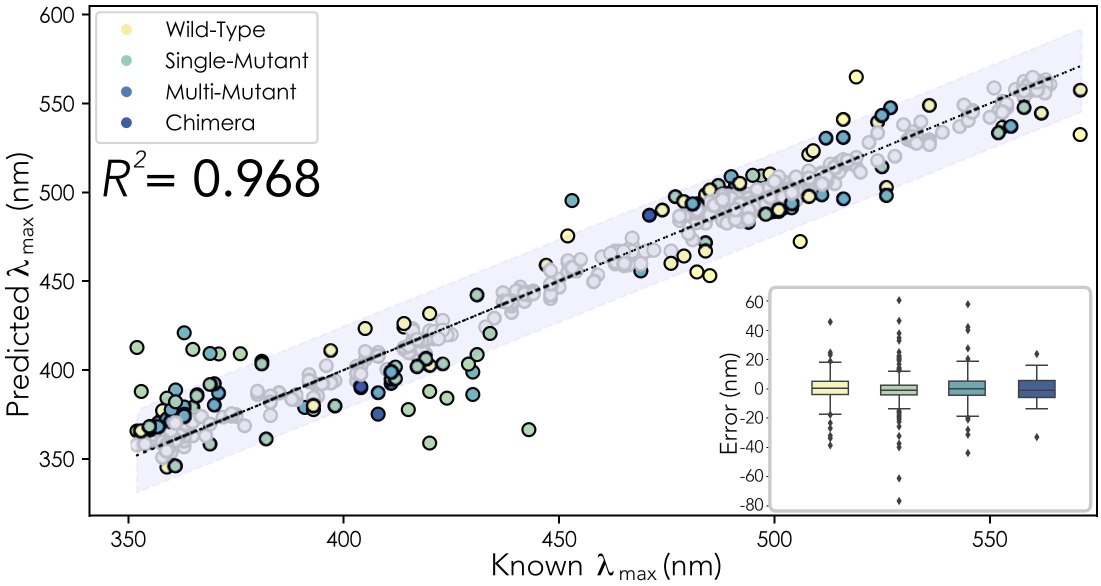

Figure 2: ML model predictions on whole vertebrate opsin dataset, $n = 721$, $R^2 = 0.968$, MAE = 6.68 nm, MAPE = 1.52. Sequences were iteratively and randomly selected to be withheld from the training dataset ($n = 50$) to act as unseen test data. This was repeated until all sequences had been sampled once. Predictions in which the absolute difference between the "known" and "predicted" λ_{max} are <10 nm are represented by gray dots. All predictions in which the absolute difference between the "known" and "predicted" λ_{max} are >10 nm are represented by colored dots. Yellow dots represent WT predictions, mutants with only a single mutation are green, mutants with greater than 1 mutation are light blue, and chimeric opsins are dark blue. The light gray bar surrounding the trend line represents a 95% confidence interval. Inset: Boxplot distribution of prediction error for different opsin data types from the top-performing vertebrate opsin ML model to better visualize our sources of error. Note, the median for each boxplot hovers around 0 nm. Single mutations have the largest spread of error, but this is most likely due to the high abundance of that data type over all others.

Understanding model performance using different subsets of the database

We created 11 data subsets with varying levels of taxonomic and gene family inclusivity (Table 1) to test which factors most impact the reliability/performance of ML methods. We used naming conventions that include versioning to improve reproducibility and reliability of individual datasets and models. For example, 1 subset combines ultraviolet and SWS opsins, which we named *VPOD_uss_het_1.0*. Our convention is to name the subset (in this case USS = "ultraviolet and short-wave sensitive" opsins), name the source of phenotype data (heterologous = het), and record the version number of the dataset (1.0). We also created subsets for medium- and long-wave sensitive opsins (*VPOD_mls_het_1.0*) and all rod (Rh1) and rod-like (Rh2) opsins (*VPOD_rod_het_1.0*). Other subsets use species taxonomy, one for vertebrates (*VPOD_vert_het_1.0*) and another for invertebrates (*VPOD_inv_het_1.0*). For taxonomic subsets, we considered all sequences from phylum Chordata as "vertebrates" and the rest as "invertebrates." Another subset excludes all mutant opsin sequences, called "wild-types" (*VPOD_wt_het_1.0*). A final named subset is the whole dataset (*VPOD_wds_het_1.0*) (Fig. 2).

Using various subsets of data, we performed a number of experiments to better understand the performance of ML models in predicting λ_{max}. First, to better understand how training data relate to model performance, R^2, and training data size, we

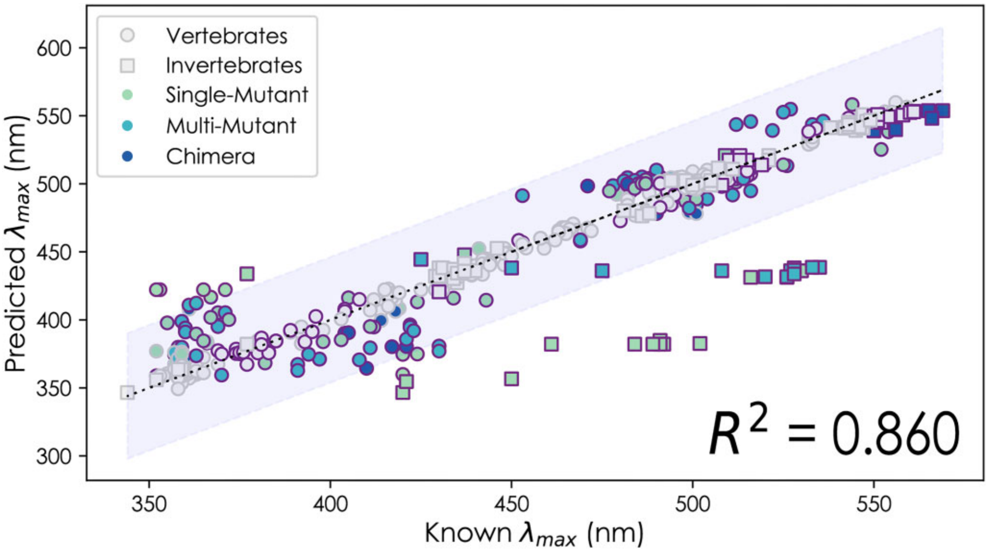

Figure 3: Scatterplot of wild-type model's λ_{max} predictions for 546 mutant opsins, with an R^2 of 0.860, MAE of 12.36 nm, and MAPE of 2.91%. Mutant predictions in which the absolute difference between the "known" and "predicted" λ_{max} are <10 nm are represented by gray dots. All predictions in which the absolute difference between the "known" and "predicted" λ_{max} are >10 nm are represented by colored symbols, further separated by invertebrate (squares) and vertebrate (circles) opsins. Mutants with only a single mutation are green, mutants with greater than 1 mutation are light blue, and chimeric opsins are dark blue. Mutations that caused a shift of >10 nm from the WT are outlined in purple. The light gray bar surrounding the trend line represents a 95% confidence interval.

gradually increased the size of training datasets by starting from zero and incrementally adding between 15 and 50 randomly selected sequences at a time for the whole dataset (WDS), vertebrate, wild-type (WT), and rod subsets separately, repeating the process 3 times per subset (Supplementary Material 2 (S2)). We then analyzed the fit between the size of training datasets (x-axis) and model performance (y-axis), comparing 6 nonlinear models with Akaike information criterion (AIC) to find the model that best explains the observed variation (Supplementary Material 3 (S3). Second, to understand if ML could predict known phenotypic changes due to experimental mutations, we queried the top-performing WT model (which lacks data from artificially mutated sequences) using all experimentally mutated opsins to predict their known phenotypes. We plotted these results using *matplotlib* [69] to visualize characteristics of poorly predicted outliers (e.g., taxonomic bias or sensitivity to mutations, which caused large shifts in λ_{max} from the WT) (Fig. 3). To test further whether including these mutant data significantly improves predictions of λ_{max}, we used the *VPOD_het_1.1* dataset (Supplementary Material 1 (S1)) and a *Wilcoxon signed-rank test* [70, 71] to compare distributions of squared error for predictions by the WDS model (contains mutant data) and WT model (no mutant data) on all mutant data ($n = 761$) and separately comparing only mutants causing the largest phenotypic changes in λ_{max} (>10 nm from the wild-type; $n = 346$). To accomplish this for the WDS models, we iteratively removed 25 mutant opsins at a time from training data, used the same training algorithm (gradient boosted regressor [GBR]), and predicted λ_{max} values of withheld opsins following the completion of model training (withheld opsins are not used as test data during the actual model training), until all mutant opsins were sampled once (this notebook is available on GitHub as "*vpod_wf_iterate_subsample.ipynb*." Third, we examined the ability of our models to predict λ_{max} of 30 invertebrate opsins not in *VPOD_1.0* because they are only known from physiological studies (Supplementary Material 4 (S4), Supplementary Material 5 (S5)). Here, we collected data both characterized by single-cell microspectrophotometry (MSP) or electroretinogram methods and with expression localized to cell type by *in situ* hybridization (ISH), to link λ_{max} to a specific opsin (the sequences and metadata can be found in "*msp_erg_raw.txt*" and "*msp_erg_meta.tsv*," while the resulting predictions can be found under the "*msp_tests*" folder on our GitHub repository). Finally, we directly compared predictive capabilities of models trained on different data subsets by randomly selecting and removing the same 25 wild-type ultraviolet or short-wave sensitive opsins from the training data of the WDS, vertebrate, WT, and ultraviolet sensitive (UVS)/SWS models before training and querying the model with those same sequences following training (Supplementary Material 4 (S4), Supplementary Material 6 (S6)).

Comparing machine learning and phylogenetic imputation

We compared performance of ML models to phylogenetic imputation, which estimates phenotypes using phylogenetic information [54, 55]. Phylogenetic imputation uses maximum likelihood (we will not abbreviate maximum likelihood as ML to avoid confusion with machine learning), usually assuming Brownian motion to predict missing phenotypes using a phylogenetic tree, such that more closely related species or sequences have more similar phenotypes. For the phylogeny, we constructed opsin gene trees in phyML [72], assuming the "WAG" substitution model [73] and a proportion of 0.029 invariable sites, with Gamma as a rate across sites model, and 4 substitution rate classes. We randomly removed 50 opsin sequences and their corresponding λ_{max} values from each of the ML training datasets (with the exception of the smaller medium wavelength-sensitive (MWS)/LWS and invertebrate datasets, where we only removed 15), then estimated the removed λ_{max} values using phylogenetic imputation. We used the phylogenetic imputation submodule of the *phytools* R package [74] for imputation. We compared imputed and actual λ_{max} using regression. Imputation seemed sensitive to input alignment, perhaps caused by very short or zero length branch lengths in the phylogeny, as we could only complete imputation with *phytools* after removing uninformative and heavily gapped regions with Gblocks. To allow direct comparisons of regressions between imputation and ML, we re-created ML training–data alignments using MAFFT, MUSCLE, and Gblocks in the same way as for imputation and predicted λ_{max} for the same sets of sequences as imputation (Supplementary Material 7 (S7)).

Testing ability of ML to account for intragenic epistasis

Functional predictions are often misled by epistasis [41], so we tested the ability of our WDS models to predict the effects of epistatic mutations by haphazardly selecting 3 double mutants with previously demonstrated epistatic effects from training data in which double mutants, each single mutant, and wild-type sequence are all characterized by heterologous expression. The 3 epistatic double mutants are all derived from bovine rhodopsins: D83N_A292S, F261Y_A269T, and A164S_A269T. We removed the double mutants from the training dataset but retained single mutants to test whether the model treats the mutations as additive or epistatic. We hypothesized that the many instances of multimutant sequences with epistatic effects in the training set would allow the model to account for both the magnitude and direction of intragenic epistasis. We then ran a separate test where we removed the same double mutants plus their corresponding single mutants to observe whether the WDS model still predicts epistatic effects from wild-type data alone. We subsequently repeated this same process for the WT and vertebrate models (Supplementary Material 8 (S8)).

We ran an additional experiment to test the general ability to to predict epistatic interactions between mutations for all available data. Here, we identified all multimutants that have phenotype data for each individual component mutation. Next we selected those multimutants with nonadditive (epistatic) interactions between mutations (which we define as >1 nm difference between the actual multimutant phenotype and the sum of changes in phenotype due to the individual mutations). These 111 "epistatic mutants" were then all removed from WDS (*VPOD_wds_het_1.1*) to create a new training dataset called "WDS-minusepi" that lacks evidence of intragenic epistasis. For this test, we hypothesized that if the ML approach can account for epistasis, the RMSE of predictions of the 111 epistatic mutants would be significantly lower for the model trained with WDS-minusepi than the model trained with no mutants at all (WT). We tested for statistically significant differences in the distributions of square error for predictions made by WDS-minusepi versus WT and WDS-minusepi versus the epistasis-free additive mutation values (EAMVs, which represent the expected λ_{max} for mutants if the effects of their singular mutational components were treated as additive). We also predicted a statistically significant difference between predictions made by WT and EAMV only if WT contains enough natural variation (not based on mutants) to observe patterns of intragenic epistasis. These statistical tests assumed a Bonferroni correction for multiple tests.

Identifying known spectral tuning sites

In addition to predicting λ_{max}, we wanted to identify amino acid sites with strong effects on the phenotype, called spectral tuning sites for opsins. To do so, *deepBreaks* produces an "importance report" of the relative importance of amino acid positions within the sequence relative to the phenotype. This report is generated for each of the top 3 performing models, with the addition of a column that calculates the "mean relative importance" value of each individual position. We automated the translation of these feature representations of aligned amino acid positions compared to bovine rhodopsin for the sake of interpretability. We also included the amino acid residue identity at each corresponding position and whether it is in one of the opsin transmembrane domains (TMDs). We used this to provide us with a standardized context for analysis of the most significant positions highlighted by the models, which we could use to compare to published mutants and known spectral tuning sites. We analyzed the importance report for each model to see what positions it highlighted as most important, with an extra emphasis placed on the output for the WT models since it was the least likely to be biased by the presence of already known mutant data (Supplementary Material 9 (S9)), as previous researchers often chose suspected tuning sites for mutagenesis experiments.

Results

Data description: A genotype–phenotype database for animal opsins

VPOD is a new database, available on GitHub and in *GigaDB* [75] that currently includes all heterologously expressed animal opsins. We refer to a subset of the database with only heterologous data as *VPOD_het_1.0*, although for version 1.0, this is synonymous with the entire database. *VPOD_het_1.0* relies on 73 publications, mainly primary sources, with dates ranging from the 1980s to 2023. The database contains opsin sequences and phenotype data from 166 unique species (counting 35 reconstructed ancestors), including fishes, amphibians, reptiles, mammals, crustaceans, and bivalves. Altogether, *VPOD_het_1.0* contains 864 unique opsin sequences and corresponding λ_{max} values. This includes 318 unique WT opsins and 546 unique experimentally mutated opsins (447 from vertebrates and 99 from invertebrates) from 82 species (73 vertebrate and 9 invertebrate species). Of the mutants, 73 are "chimeric," meaning 1 or more transmembrane domains of the mutant are copied from a different opsin to replace the original. Phylogenetically, *VPOD_het_1.0* is mainly vertebrate opsins ($n = 721$), with only 143 unique invertebrate opsins (Supplementary Material 10 (S10)). The vertebrate opsins consist of 113 UVS opsins, 167 SWS opsins, 8 MWS opsins, 83 LWS opsins, 237 rhodopsin (Rh1) opsins, and 113 rhodopsin-like (Rh2) opsins (Supplementary Material 10 (S10)). Phenotypically, *VPOD_het_1.0* spans a range of λ_{max} values from 350 to 611 nm. The highest concentration of phenotype values are between 350–375 nm and 475–525 nm (Fig. 1) due to the literature bias favoring characterization of UVS/SWS opsins and rhodopsins (Rh1).

The data used for model training strongly impact accuracy

Several models trained with different subsets of data predicted λ_{max} with high accuracy (Table 1). The top-performing models from these subsets consistently used the same 5 algorithms, including the gradient boosting regressor (GBR) [68, 76], Bayesian ridge (BR) [77, 78], light gradient boosting machine (LGBM) [79], random forest (RF) [66], and extreme gradient boosted machine (XGB) [61]. For example, *VPOD_vert_het_1.0*—trained with all vertebrate wild-type, mutant, and chimeric opsins—had the highest 10-fold cross-validation (CV) R^2 (0.968) and lowest MAE (6.56 nm) of any models we compared (Fig. 2). Similarly, *VPOD_wds_het_1.0*, trained with the whole dataset, had very high R^2 (0.947) and low MAE (7.47 nm). The 2 data subsets also shared the same 5 top-performing models (GBR, BR, LGBM, RF, and XGB). In addition, *VPOD_wt_het_1.0*—trained without mutants and only wild-type data—had a similarly high R^2 (0.902) and a low MAE (10.3 nm) when predicting unseen wild-type data. Overall, this "wild-type-only" model also fared well, even when predicting mutant data not included in the model (Fig. 3). While these performance metrics are impressive, it is important to remember that phylogenetic relatedness between sequences of a dataset could inflate values, like R^2, when using random sampling for cross-validation because opsins that are more similar to those in the training data will be easier to predict, and phylogenetically clustered sequences will also be more likely to be resampled. Roberts et al. [80] provide a discussion of alternative cross-validation strategies such as "block cross-validation" for nonindependent data types, including phylogenetically related data, which can help mitigate this issue. Despite overall high R^2, we noticed multiple instances where mutations that cause large shifts in λ_{max} (>10 nm) were not well predicted by the wild-type-only model, as indicated by large residual values for the predictions of these mutant sequences (Fig. 3). We found including mutant data significantly improves predictions of λ_{max} when comparing predictions of models trained with (WDS) and without (WT) mutant data and rejecting the null hypotheses of no underlying differences between the distribution of squared error for predictions of all mutants (P = 9.96e-22, WDS RMSE = 12.6 nm, WT RMSE = 17.6 nm) (Supplementary Material 11 (S11)) and when predicting phenotypes of mutants with large shifts in λ_{max} (P = 2.29e-25, WDS RMSE = 17.0 nm, WT RMSE = 24.2 nm) (Supplementary Material 11 (S11)).

In addition to including mutant data, data availability more generally improves predictive power, with performance thresholds and plateaus depending on the genetic diversity of the training data. Overall accuracy in predicting λ_{max} for our models trained on more genotypically and phenotypically complete subsets of data (WDS, vertebrate, WT) improves as a function of the number of sequences in a dataset and shows an initial plateau (R^2 = ~0.80–0.90) of diminishing returns around 120 to 200 sequences that continues to taper off above 200 sequences (Supplementary Material 2 (S2), Supplementary Material 3 (S3). Consistent with a rough performance threshold, we found models from data subsets with fewer than ~200 training sequences to far less accurately predict λ_{max}. For example, *VPOD_mls_het_1.0*—trained only on the 91 MWS/LWS opsins of vertebrates—and *VPOD_inv_het_1.0*—trained only on 144 invertebrate opsins—showed among the lowest R^2 (0.677 and 0.814, respectively; Table 1). For all data subsets, we found the relationship between number of sequences in a dataset and model performance best fits a reciprocal model, which is suitable when the dependent variable plateaus as the independent variable grows larger. We found the coefficients of the reciprocal equations to be different between data subsets and to increase in negative magnitude with a decrease in taxonomic/genetic diversity (the rod model holding the largest negative value of −44). These equations do not account directly for taxonomic, genetic, or phenotypic diversity, as the raw number of genes is the value of the x-axis. Therefore, one should be cautious about applying them to predict model performance based on training data size alone.

The complicated relationship between size of training dataset and predictive power is further illustrated by models from some larger data subsets that resulted in rather poor predictions. One large dataset (884 sequences), the previously published Karyasuyama type 1 opsin dataset (*Karyasuyama_T1_ops* [47]), showed only moderate R^2 (0.804) and MAE (9.41), similar to models from the much smaller invertebrate data (Table 1). One explanation for lower predictive power could be that the very old age of T1 opsins led to a higher complexity and diversity of genotype–phenotype associations, which are not yet completely sampled enough to allow good predictions. In addition, models based on rod, UVS/SWS, and MWS/LWS subsets tend to show lower R^2 than might be at first expected (Supplementary Material 2 (S2), Supplementary Material 3 (S3), especially since these 3 datasets together comprise the training data for the vertebrate model (our highest performing model, R^2 = 0.968). For example, the rod model, with 352 sequences, should have resulted in a model with an R^2 around 0.900 to 0.960 based on the trend lines for the WDS and vertebrate datasets (Supplementary Material 2 (S2), Supplementary Material 3 (S3) but resulted in an R^2 = 0.831. A possible explanation for this lower R^2 value for rod models is the small degree of variability in $\lambda_{max.}$ When variation is low, even very small differences from model predictions could lead to larger differences in R^2. Therefore, when a data subset such as rod opsins contains limited variability in the response variable (λ_{max}), additional metrics that are less sensitive to variance will be important, such as MAE or RMSE, which report the absolute magnitude of errors rather than the proportion of explained variance. To illustrate further, most models tested on their ability to predict the λ_{max} for a set of 25 subsampled WT-SWS opsins from *VPOD* performed relatively poorly based on R^2 alone (Supplementary Material 4 (S4)), with the vertebrate model (R^2 = 0.914, MAE = 7.89) demonstrating a relatively greater predictive power than all other models (Supplementary Material 4 (S4), Supplementary Material 6 (S6)). However, between the vertebrate and lowest performing model (SWS model; R^2 = 0.778, MAE = 11.6 nm), there is only a 3.71-nm increase in MAE, a much less dramatic perceived shift in performance than might be interpreted from R^2 alone.

When predicting λ_{max} of 30 unseen wild-type invertebrate opsins from a separately curated MSP dataset, almost every model performed rather poorly, with exception of the WT model (n = 30, R^2 = 0.887, MAE = 17.5) (Supplementary Material 4 (S4), Supplementary Material 5 (S5)). The best-performing model produced by the sparsely populated "*Invertebrate*" dataset could only predict unseen invertebrate opsins with an R^2 of 0.837 and MAE of 26.3 nm (Supplementary Material 4 (S4), Supplementary Material 6 (S6)). Until the models are trained with more invertebrate (r-opsin) data, we would not put high confidence in the estimates of λ_{max}. Furthermore, these separately curated invertebrate opsins are independent of the phylogenetic relatedness of the data used in model training and therefore provide a less inflated estimate of the ability to predict λ_{max} compared to random resampling of training data. Because of the sparsity of invertebrate data in the training set, this result further highlights that opsins more distantly related to those in the database will be more difficult to predict.

ML predictions of λ_{max} are comparable to phylogenetic imputation

Both ML and phylogenetic imputation were often accurate predictors of λ_{max} (Supplementary Material 7 (S7)). When using the same test data, ML models usually outperformed phylogenetic imputation, however slightly (Supplementary Material 7 (S7)), albeit using far less computational time: ML used on the order of minutes to calculate models, and imputation used on the order of hours to generate opsin phylogenies. The MWS/LWS dataset was the only instance where phylogenetic imputation (R^2 = 0.784) largely outperformed ML (R^2 = 0.512). We found our implementation protocol for phylogenetic imputation required removing aligned sites with extensive gaps (for which we used Gblocks); we speculate this lessened the impacts of very short branch lengths on model fitting during imputation. To allow direct comparisons between approaches, we also used the same trimmed alignments for training ML models. Interestingly, there was a slight but noticeable decrease in ML performance following Gblocks trimming for the invertebrate, MWS/LWS, and UVS/SWS datasets (Supplementary Material 7 (S7)). The R^2 of the MWS/LWS model dropped from 0.677 to 0.645, while the invertebrate model dropped from 0.814 to 0.797 (Supplementary Material 7 (S7)). ML performance remained relatively consistent after tripping for the WT, vertebrate, WDS, SWS/UVS, and rod models, with only a slight reduction in R^2 (<0.01) and slight increase in MAE (±1 nm) for the WT model. We speculate the observed differences in ML performance following Gblocks processing is due to the reduced number of features in the datasets from removing aligned sites.

ML often predicts the effects of epistatic mutations

The WDS successfully predicted 3 out of 3 individual instances of epistasis (Supplementary Material 8 (S8)) using sequences that were removed from the training data before using the model to predict known epistatic phenotypes. For double mutant D83N_A292S, the model predicted 485.2 nm, which was 0.2 nm off the known λ_{max} of 485 nm. If the WDS model believed the sites were additive, the resulting λ_{max} based on adding shifts of single mutants would have been much lower, at 477.5 nm. Second, for mutant F261Y_A269, the model predicted 520.0 nm, for which the known λ_{max} was 520 nm. An additive prediction would have been higher, 524 nm. Third, for mutant A164S_A269T, the model predicted a λ_{max} of 515.5 nm, where the known λ_{max} was 514 nm. This is a special case in which the double mutant experiences a form of epistasis where the effect of mutation A269T (λ_{max} = 514) masks the shift otherwise caused by mutation A164S (λ_{max} = 502 nm). Thus, the model correctly predicted an instance of epistasis in which one mutation masks the effect of another.

We also queried the WT model with these same 3 double mutants to test the importance of mutant sequences in informing the model on epistatic interactions. However, without any mutant data at all, the WT model did not display the same abilities to predict epistasis in any instance. For the double mutant D83N_A292S, the model predicted that neither the individual mutations nor the double mutant would have a significant effect on λ_{max}, and all were predicted to be 499.9 nm. For double mutants F261Y_A269 and A164S_A269T, the WT model successfully predicted all individual mutations would cause a red shift (although F261Y and A269 were >3 nm off their known λ_{max}) but incorrectly treated the mutational effects as additive for the double mutant (Supplementary Material 8 (S8)).

Our broader experiment to test the predictability of epistatic effects using the WDS-minusepi model (which excluded from training all 111 opsins with known nonadditive mutational effects, which we call epistatic opsins) correctly predicted epistasis for 105 of 111 of the epistatic opsins with higher R^2 (0.969) and much lower RMSE (12.4 nm) than predictions by the WT model (R^2 = 0.894, RMSE = 22.3 nm), which contains no experimentally mutated opsins, and the EAMV (R^2 = 0.878, RMSE = 29.8 nm), which ignores epistatic effects, respectively (Supplementary Material 12 (S12)). Our test of the null hypotheses of no underlying differences between the distribution of squared error for predictions of the 111 epistatic mutants were rejected after Bonferroni correction by the WDS-minusepi model versus WT model (P = 1.24e-06) and WDS-minusepi model versus EAMV (P = 2.56e-09) but not rejected for the WT model versus EAMV (P = 0.086) (Supplementary Material 12 (S12)). Together, the large differences in RMSE and the results of the statistical tests strongly support the idea that the inclusion of even single mutants significantly reduces the error of ML models when predicting epistatic interactions between mutations and that this error is also less than the error we would observe if our models simply treated mutations as additive. Nevertheless, the insignificant difference between WT predictions and EAMV indicates there is not enough information about epistatic interactions in wild-type (nonmutant) data alone to accurately predict intragenic epistasis.

ML predicts tuning sites from wild-type sequences alone

The full WT model and its few variants (SWS and rod WT models) predict several previously characterized "spectral tuning sites"—functionally demonstrated to change λ_{max}—even with no information on mutants used in the training data (Fig. 4, Supplementary Material 9 (S9)). For the primary WT model alone, we found 15 of the top 25 amino acid sites, ranked by relative importance to the model (all ≥0.40), were spectral tuning sites previously characterized by mutagenesis and heterologous expression (Supplementary Material 9 (S9)). For example, the especially well-characterized position 308 (p308), known for its role in tuning LWS opsins and considered 1 of the 5 key sites in characterizing LWS opsins under the "five-site rule" [81], had the highest relative importance value of 1.0 when using the full WT model, indicating the amino acid identity at p308 is especially important for predicting λ_{max}. In another example, the full WT model highlighted p181, a phylogenetically conserved counterion in the retinal-opsin Schiff base interaction for all nonvertebrate opsins [82, 83]. Additionally, the transition from E to H at p181 (E181H) is a characteristic of the red-shifted vertebrate LWS opsins [35, 83], easily visualized in Fig. 4C. When predicting λ_{max} of bovine rhodopsin with mutation E181H, the WT model predicted a red shift compared to wild type, as observed with the natural evolution of the LWS opsin lineage. The WT SWS/UVS model similarly highlighted p113, a site functionally characterized as the counterion in the retinal-opsin Schiff base interaction for all vertebrate opsins [35, 83] and as a known spectral tuning site in SWS/UVS opsins [84]. Moreover, even the WT rod model, trained on a mere 157 sequences, identified p292 (Supplementary Material 9 (S9)), another well-characterized and conserved spectral tuning site for vertebrate rhodopsins [85–87], as the site with highest relative importance to its predictions of rhodopsin λ_{max}. These spectral tuning sites are not simply conserved sites, as there is little to no correlation between amino acid sites important to model predictions (importance scores) and their relative *Shannon entropy* [88, 89] scores (R^2 = 0.001). This is somewhat expected as *deepBreaks* drops all conserved ("zero-entropy") sites during preprocessing, because a site with no variation provides no important information about the effects of variation on the resulting phenotype. In addition, we predict any correlation between site conservation and model importance would be for sites that are moderately conserved and in close proximity to opsin–chromophore binding site (position 296) or binding pocket [41, 42, 90].

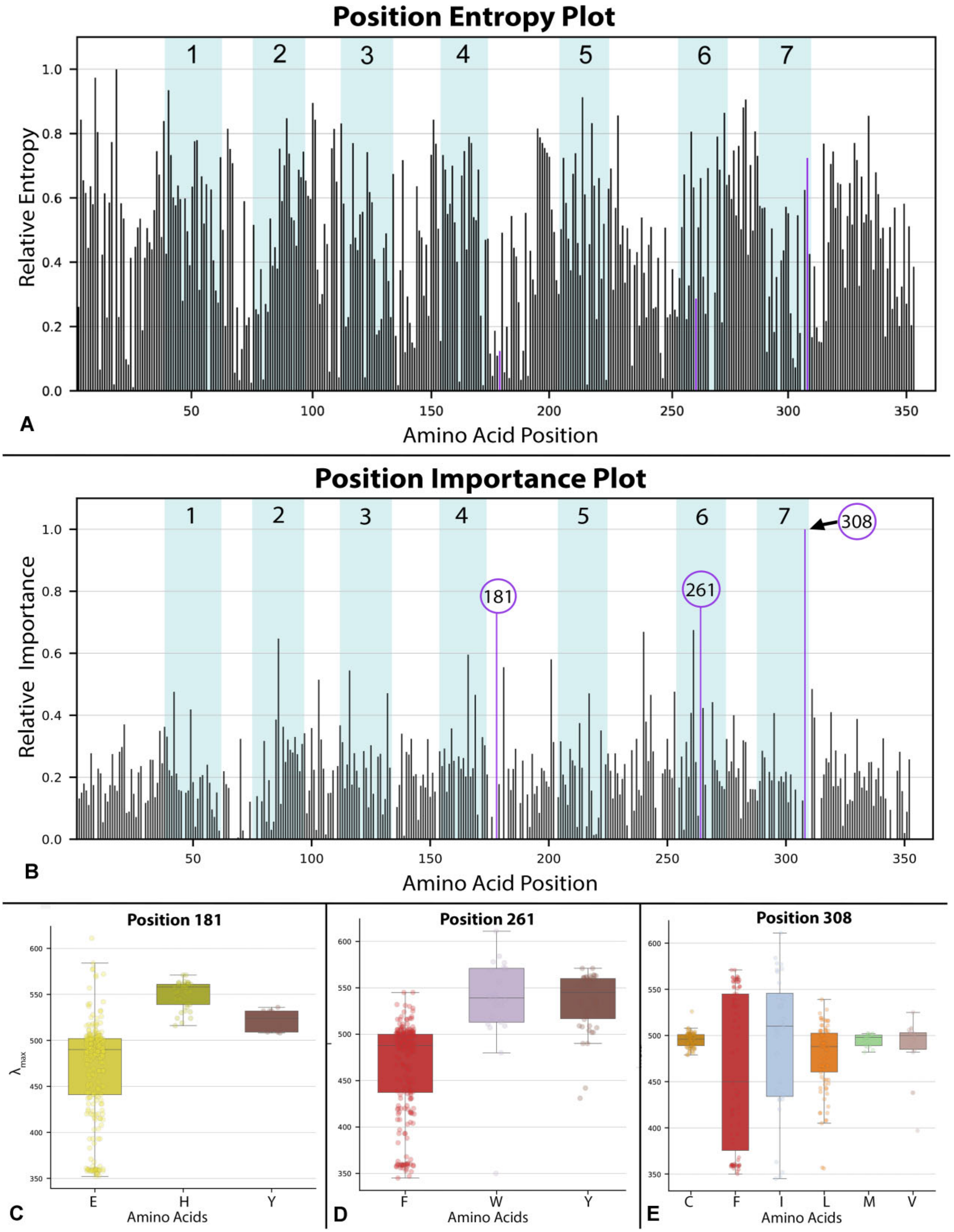

Figure 4: (A, B) Blue bars indicate the 7 transmembrane domain regions of the bovine rhodopsin and are labeled accordingly. Purple bars indicate the top 3 most important positions to predictions of λ_{max} by the "BayesianRidge" ML regression model trained on the WT opsin dataset. (A) Bar graph of relative entropy scores by position calculated via Shannon entropy [71, 88, 89] using the multisequence alignment for the WT data subset. (B) Bar graph of relative importance by position generated via "BayesianRidge" ML regression model trained on the WT opsin dataset. We interpret positions with higher relative importance as having a larger effect or weight on λ_{max} prediction. Positions 181 [35, 83], 261 [87, 91], and 308 [81] are highlighted in purple because they are among the highest scoring sites and have all been previously characterized as functionally important to opsin phenotype and function. Based on an R^2 of 0.001, there is no linear relationship between relative entropy by position and the relative importance of scores by position. (C–E) These distribution box plots provide a visualization for which amino acid (aa) residues at a particular site are associated with different ranges of lambda max at a site of interest, ordered alphabetically, not by frequency (left to right). For a more detailed explanation on how position importance scores are calculated for different models, refer to the "Interpretation" heading under the methods section of the *deepBreaks* publication [60].

Discussion

To better understand methods to connect genes and their functions, we initiate *VPOD*, a database of opsin genes and corresponding spectral sensitivity phenotypes. Here, we used *VPOD_1.0* to examine the ability of ML models to predict functions of opsin genes, predict intragenic epistasis, and identify amino acid sites critical for functional changes. In all cases, ML shows promise, especially when given enough training data.

The important relationship between data availability and predictive power

The predictive power of λ_{max} is often high when using ML for opsins, and it improves with a greater amount and variety of data, albeit with diminishing returns. In particular, the number of opsin genes, their genetic diversity, and the relationship between genetic and phenotypic differences are all critical in determining predictive power. Particularly illustrative of these ideas are our analyses with and without experimentally mutated opsins. Even though we might conceive of all wild-type data as natural mutants chosen by evolution, experimentally induced mutations are particularly important by often changing just 1 amino acid that drastically changes phenotype. As such, we found that including mutant data usually improved predictive power, and conversely, predicting some phenotypes from laboratory mutagenesis was sometimes difficult without including other mutant data in model training (Supplementary Material 11 (S11)). However, relying on published mutant data alone is not optimal because it is derived from a nonrandom subset of species because people continue to work in established systems. Nevertheless, the genotype–phenotype landscape may be sampled well enough using high numbers of only wild-type genes, as evidenced by the small difference in performance when adding mutant data to the wild-type subset of well-sampled vertebrate opsins (Table 1). In contrast, adding mutant data to the sparsely sampled invertebrate opsins made a big difference. For invertebrate opsins, using only wild-type data (ignoring all mutants) led to some very inaccurate predictions, especially of large phenotypic shifts caused by experimental mutagenesis (Fig. 3), indicating the genotype–phenotype space is still undersampled for invertebrates. This is expected since ML learns from patterns in the underlying dataset, making predictions of distantly related opsins from those in *VPOD* more unreliable. We acknowledge this as a significant drawback for the ML approach, especially in systems or taxonomic groups lacking sufficient or reliable data. Thus, given this currently limited dataset, we do not put high confidence in the λ_{max} estimates of either wild-type or mutant invertebrate (rhabdomeric) opsins. Therefore, targeting invertebrate opsins should be a high priority for new additions to *VPOD*.

A large diversity of training data is also critical for reliably predicting intragenic epistasis—the nonadditive effects on a phenotype of interactions between 2 or more mutations within a gene—which is common [10, 41, 43, 44, 92, 93] and an obstacle to connecting genotypes and phenotypes [41, 94–96]. Our most complete datasets (whole dataset and vertebrate dataset) identified known cases of intragenic epistasis, but our models trained without experimental mutagenesis data did not. Moreover, ML demonstrates some capacity to predict the epistatic interactions between mutations, even when only provided with the single mutation components—as is evidenced by our WDS-minusepi dataset test (Supplementary Material 12 (S12)). Similarly to the overall predictive power of λ_{max} above, predicting epistasis probably requires sufficient variation at interacting sites, which seems especially enhanced by experimentally mutated genes.

Variation in the availability of genotype–phenotype data for training impacts not only the predictive power of phenotype but also the converse: the ability to predict amino acid sites that change λ_{max}. Several models, including those trained with the WDS, vertebrate, and WT data, were able to successfully predict previously characterized spectral tuning sites. This is less surprising for models trained with WDS and vertebrate datasets due to the prevalence of data, even including mutants in the training data from experiments that specifically targeted sites thought by researchers to be functionally informative. Yet even without any targeted mutational data, 3 model variants using only wild-type data predicted experimentally well-characterized spectral tuning/functional sites, including sites important to the stability of the opsin–chromophore interaction (P181 and P113). This demonstrates the strong potential for ML models to identify amino acid sites that govern phenotype, leading to predictions of candidate spectral tuning sites, which can be confirmed with mutagenesis experiments [38, 86] if not done so already.

ML algorithm type contributes to the predictive power of ML models

While probably not as important as the training data used, the ML algorithm itself also impacts predictive power. All 5 of the best-performing ML algorithms (GBR, BR, LGBM, RF, and XGB) are variants of the decision tree model architecture (Supplementary Material 13 (S13)), and 3 of 5, including GBR, LGBM, and XGB, are "gradient boosted" decision tree–based ML algorithms. The gradient boosted algorithms all share the same general principles of gradient boosting [76, 97], including the use of ensembles of "weak learners," usually decision trees, which work sequentially and "gradient descent" when minimizing a loss function, to improve ML model performance. While LGBM generally performed best for predicting phenotype, it was not as effective in predicting the epistatic effects of mutations, where GBR and XGB showed the highest performance. This suggests that while LGBM excels in general phenotype prediction, the details of GBR and XGB may be better suited for epistasis prediction. The difference likely arises from the unique aspects of each algorithm's model training and settings of hyperparameters. XGB and LGBM differ from GBR by the addition of a regularization term to the objective function and in the process of ensemble tree construction during model training: GBR and XGB use level-based tree fitting while LGBM uses leaf-based tree fitting. One consequence of leaf-based tree construction is that due to its faster convergence/training time, it can create complex trees that are more prone to overfitting, thereby "learning" patterns that may not exist as it constructs trees on a "best-first basis" with a fixed number of n-terminal nodes [63, 79]. This creates a model that often performs well on training data but may overgeneralize, missing finer grained collinearities and interdependencies, which would be important for predicting epistasis. As such, our models might be improved by fine-tuning hyperparameters (e.g., learning rate, max-depth, and number of estimators), and the choice of which model to use will depend on the end goals of the analysis.

The assumptions of our method and limitations of ML extrapolation

Understanding the limitations and assumptions inherent in predictive modeling is vital for accurately interpreting animal color sensitivity from opsin sequences, especially considering the impact of various factors on sensitivity beyond the opsin itself across multiple levels of biological organization. At the cell level, we assume that λ_{max} measured in cell culture (e.g., HEK293, COS cells) is the same as in living photoreceptor cells. We also assume the photopigment uses 11-*cis*-retinal, as all heterologously expressed opsins in *VPOD* were reconstituted using this chromophore. However, this assumption is violated in some organisms because they use 13-*cis*-retinal as the *in vivo* chromophore [23, 98, 99], which is associated with a red shift in λ_{max} [35, 98]. At the organ level, filters such as oil droplets in bird eyes [100–103], pigments in butterfly eyes [104], or a combination of transmissive filter and narrow band reflector in mantis shrimp larval eyes [105] each may selectively influence light reaching photoreceptor cells and therefore animal color sensitivity. Finally, organismal responses to light involve neural processes, so even if an organism possesses the physiological ability to detect certain wavelengths, it still may not have a use for that ability. Similar considerations for all these assumptions will apply when using ML to infer other functions from other genes. In fact, many genes are more susceptible than opsins (but see [106] showing the pressure of ocean depth may slightly affect λ_{max} phenotypes) to changes in pH, temperature, and other environmental factors [107], such that databases compiling these gene functions should also record these parameters for use in training ML models.

Perhaps the most important caveat of using ML models to accurately predict phenotype or functional sites is that we assume there is a genotype–phenotype association that we can fit to a function and that our models were trained using ample data to capture these associations. Based on the nonlinear fit between size of training dataset and model performance, we estimate that including about 200 sequences (and corresponding λ_{max}) from a taxonomically and phenotypically diverse range still provides improvements to model performance. Above 200 sequences, there is still improvement, but at a diminishing rate consistent with a reciprocal model (Supplementary Material 2 (S2), Supplementary Material 3 (S3). That said, we encourage caution when extrapolating these results to predict model performance on training data size alone as the equations we used do not account directly for taxonomic, genetic, or phenotypic diversity. When using ML for predicting functionally important sites, the addition of experimental mutants to training data that cause large phenotypic changes could heavily bias which sites are selected as “most important” and potentially mask the importance of other sites. Here again, providing a diverse set of genotype–phenotype data should allow for the discovery of new functional sites, even when including known mutants in the training data with large phenotypic effects. Additionally, providing a large number of mutations from a limited breadth of taxa can bias model predictions as not all mutations will have the same effect on different sequences, especially if they are genetically distant. This makes it all the more important to consider the level of genetic diversity used to train a model when extrapolating to find potentially important functional sites (i.e., if identifying tuning sites for rhodopsins, then using a dataset of only rhodopsins would likely be the best approach, but if data are sparse or if looking for sites that may largely impact spectral tuning across opsin subfamilies, a genetically and phenotypically broad dataset may be better).

Conclusion

Using opsin sequence data with *deepBreaks*, we were able to train regression-based ML models to reliably predict λ_{max}, often accounting for nonadditive effects of mutations on function (intragenic-epistasis) and identifying amino acid sites critical for function. We expect future work will improve these already promising results even further through at least 2 general directions. First, adding more data to *VPOD* will improve results, especially adding invertebrate (rhabdomeric opsins) data, as technical knowledge improves for expressing these genes [34]. In addition, phenotypic data—besides the *in vitro* heterologous expression targeted here—is expansive, including λ_{max} measurements from microspectrophotometry and electroretinograms, but will take considerable effort to link these phenotypes to specific opsin genes. Second, our models can be improved to take advantage of more information. One important addition should be inclusion of physicochemical properties of the amino acids [108], as implemented with success on a small scale of only 26 amino acid positions of microbial opsins to predict red-shifted phenotypes for optogenetics [109]. Additionally, information on protein structure could be particularly important, such as the distance of an amino acid from the binding pocket of the chromophore [40]. While there are only a few solved crystal structures for opsins [110, 111] to provide such data, indirect techniques like homology modeling [112] or neural network–based structural prediction [113] might be usable. Other information about opsins could also be predictive, such as which G-protein the opsin signals to, allowing prediction of which amino acids dictate G-protein specificity. Opsin kinetics [e.g., 114], or even the habitat depth at which the animal lives in the ocean, which not only influences light environment but also alters which amino acids are used in opsins [115], could improve predictive power of the ML models. Finally, we once again caution against treating predictions of λ_{max} uncritically, because the quantity and quality of genotype–phenotype data used to train a model—including the taxonomic, genetic, and phenotypic diversity—is integral to the reliability of a model’s predictions. Thus, ML models like those used here can be considered tools to make predictions based on summaries of existing knowledge, thereby complementing traditional experimental methods.

Potential implications

Given the high performance demonstrated in this article, current models are already robust enough to allow several applications. First, predicting λ_{max} will often be useful, especially for vertebrate opsins. For example, ML could provide an estimate of λ_{max} in a hogfish, whose skin expresses an opsin with unknown absorption and where λ_{max} has implications for a conceptual model of chromatophore expansion [116]. Second, estimates of λ_{max} from opsin sequences formed part of an argument that changes in gene expression, not sequence, adapted Amazon fishes to local light environments [117]. On broader taxonomic scales, predictions of λ_{max} from opsin sequences could expand studies of adaptation, molecular evolution, and constraint in comparison to light environments [118]. Another application could be protein design for optogenetics—the use of genetic light sensors to induce and study expression or response pathways [119–121]—including those associated with embryogenesis [122,123], stress and depression [124–126], or neuronal diseases [127, 128]. Finally, our models could be used to simulate molecular evolution under a realistic genotype–phenotype landscape. One shortcoming presently for such simulations is that our models are not trained with nonfunctional opsins, so even nonfunctional genes would be predicted to have functional λ_{max} values. A solution could be to add large-scale mutagenesis data to the training set, such as that from deep mutational scanning [129], although the authors indicated the method is only in a proof-of-concept stage, such that the results are too noisy to be useful for model training. As the *VPOD* database expands, there will be many applications for ML, and similar techniques can also be applied to other gene families such as luciferases [16,130,131].

Availability of Supporting Source Code and Requirements

Project name: The Visual Physiology Opsin Database (VPOD).
Project homepage: https://github.com/VisualPhysiologyDB/visual-physiology-opsin-db [56].
License: GNU General Public License (GPL)—Version 3, 29 June 2007.
RRID: SCR_025668.
Operating system(s): Windows, MacOS, and Linux.
Programming language: Python, R.
Other requirements: Conda 4.9.2, deepBreaks 1.1.2, GBlocks 0.91b, MAFFT 7.520-1, MUSCLE 3.8.31, mySQL workbench 8.0.36, Python 3.9, RStudio 2023.06.2+562.
Docker image of the latest version of the deepBreaks: [132].
The Docker image provided above includes a summary of required package libraries and instructions on how to use it. Along with our existing online materials with tools used, *deepBreaks*, we also have a Jupyter notebook, instructions for Conda installation, and Code Ocean (RRID:SCR_015532) capsule [133], for deepBreaks.
These resources should help practitioners using the main ML program we used, deepBreaks, described elsewhere, use the VPOD database for Opsin applications.

Additional Files

Supplementary Material 1 (S1). Performance metrics across opsin subsets and top performing models for *VPOD_1.1*.
Supplementary Material 2 (S2). Tracking model performance vs. number of sequences in training data.
Supplementary Material 3 (S3). Three functions fitted to visualize the relationship between training data size (number of genotypes and corresponding phenotypes) vs. model performance (R^2) based on results from the vertebrate subset of data. The Akaike information criterion (AIC) is a measure used for model selection when comparing different statistical models, accounting for both the goodness of fit of the model and the simplicity of the model (the number of parameters used). The goal is to find a balance between a model's ability to explain the data and its complexity, preventing overfitting.
Supplementary Material 4 (S4). Comparing ML predictions on invertebrate and vertebrate UVS/SWS opsin MSP data.
Supplementary Material 5 (S5). Graph of WT model predictions for 30 unseen invertebrate opsins, $R^2 = 0.887$, MAE = 17.5 nm, MAPE = 4.05. All the "known" λ_{max} values are from physiological measures, including MSP or ERG measurements (instead of purified heterologously expressed opsins), and are linked to a particular opsin sequence by *in situ* hybridization. The light gray bar surrounding the trend line represents a 95% confidence interval.
Supplementary Material 6 (S6). Graph of vertebrate model predictions for unseen WT-UVS/SWS data, $n = 25$, $R^2 = 0.914$, MAE = 7.89 nm, MAPE = 1.90. All sequences were randomly selected from the UVS/SWS model under the condition that they were WT opsins. The light gray bar surrounding the trend line represents a 95% confidence interval.
Supplementary Material 7 (S7). Comparing performances of ML predictions and phylogenetic imputation on a subsample of opsin data.
Supplementary Material 8 (S8). Results for epistasis test on the WDS, vertebrate, WT, and rod models.
Supplementary Material 9 (S9). Functionally characterized spectral tuning sites predicted by the WT models.
Supplementary Material S10 (S10). Phylogenetic gene tree of all wild-type opsins ($n = 362$), including ancestral constructs (branch lengths = 0), constructed from the VPOD_wt_het_1.1 dataset. In this tree, we have annotated the major opsin groups (c-opsins, r-opsins, and some tetraopsins), then further annotated the c-opsin families (LWS, SWS1, SWS2, Rh1, and Rh2). We have also assigned taxonomic annotations by class, which are color-coded and provided by the key.
Supplementary Material S11 (S11). Including data from experimentally mutated opsin sequences reduces errors in predicting λ_{max}. (A) Distributions of errors from predicting λ_{max} of experimentally mutated opsin sequences. Blue are prediction errors when using the WT model, which lacks experimentally mutated sequences (root mean square error [RMSE] = 17.6 nm). Orange are prediction errors when using the WDS model, which includes experimentally mutated sequences (RMSE = 12.6 nm). (B) Data from experimental mutants significantly improves predictions of λ_{max} when using a model trained with experimental mutants (WDS) compared to a model without (WT) experimental mutant data, rejecting the null hypotheses of no difference between prediction errors based on different models. At top is the distribution of differences between predictions with and without experimental mutants in the training data for large effect mutations (>10 nm). At bottom is the same for all mutations. We plot differences of absolute error instead of squared error in B for easier visualization, although *P* values were calculated using distributions of squared errors. Additionally, plotting raw differences allows seeing most values are below zero, meaning predictions with WDS (which has experimental mutants) have less error than those without experimental data from mutants (WT).
Supplementary Material 12 (S12). Including data from experimentally mutated opsin sequences reduces errors in predicting epistatic effects. (A) We analyzed opsins with multiple mutations whose known effect on λ_{max} phenotype were nonadditive (epistasis). In purple, we plot the difference (absolute error in nm) between known λ_{max} phenotypes with epistasis, compared to λ_{max} phenotypes ignoring epistasis by assuming individual mutations are not additive, which we call epistasis-free additive mutation values (EAMVs). Here root mean square error (RMSE) = 29.8 nm. In blue, we plot errors when predicting epistatic phenotypes using a model trained without opsins containing experimentally generated mutations (WT), which lead to RMSE = 12.4 nm. In orange, we plot errors when predicting epistatic phenotypes using a model trained with opsins containing experimentally generated mutations but excluding those whose mutational effects are nonadditive (WDS-minusepi), which lead to RMSE = 12.4. (B) Our tests of the null hypotheses of no underlying differences between the distribution of squared error for predictions of λ_{max} for the 111 "epistatic opsins" were rejected with Wilcoxon signed-rank tests after Bonferroni correction by the WDS-minusepi model versus WT model ($P = 1.24e\text{-}06$); WDS-minusepi model versus EAV ($P = 2.56e\text{-}09$), but not rejected for the WT model versus EAV ($P = 0.086$). The large differences in RMSE and the results of the statistical comparisons strongly support the idea that the inclusion of even single mutants greatly reduces the error of ML models when predicting epistatic interactions between mutations and

that this error is significantly less than the error we would observe if our models simply treated mutations as additive. Conversely, the insignificant difference between WT predictions and EAMV indicates there is not enough information about epistatic interactions in wild-type (which excludes artificially mutated opsins) data alone to accurately predict intragenic epistasis. As with S11, we plot differences of absolute error instead of squared error in B for easier visualization but use squared error for statistical comparison.

Supplementary Material 13 (S13). Ranked ML algorithm performances.

Abbreviations

Adaboost: adaptive boosting; AIC: Akaike information criterion; COS1: monkey kidney cell line; CV: cross-validation; EAMV: epistasis-free additive mutation value; ERG: electroretinogram; GBR: gradient boosting regressor; GPCR: G-protein coupled receptor; HEK293: human embryonic kidney cell line; ISH: *in situ* hybridization; KDE: kernel density estimate; LGBM: light gradient boosting machine; LWS: long-wave sensitive; MAE: mean absolute error; MAPE: mean absolute percentage error; ML: machine learning; MSE: mean squared error; MSP: microspectrophotometry; MWS: medium wavelength-sensitive; NCBI: National Center for Biotechnology Information; RMSE: root mean square error; SWS: short-wave sensitive; T1: type 1 {microbial opsins}; TMD: transmembrane domain; USS: ultraviolet and short-wave sensitive; UVS: ultraviolet sensitive; VPOD: Visual Physiology Opsin Database; WAG: Whelan and Goldman substitution model; WDS: whole dataset; WT: wild-type; XGB: extreme gradient boosting; λ_{max}: lambda max/wavelength of light with maximal absorbance.

Acknowledgments

We acknowledge use of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara. We acknowledge R. Varney for providing technical support and expertise on the phylogenetic imputation experiments. Thanks to V. Scriven for literature searches and data entry. Thanks to A. Singh and S. Yi for advice. We acknowledge Oakley Lab for comments on an early draft of the manuscript. We thank the reviewers for helpful suggestions to improve the presentation of our work.

Author Contributions

Seth A. Frazer (Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Validation [lead], Visualization [lead], Writing – original draft [equal], Writing – review & editing [equal]), Mahdi Baghbanzadeh (Investigation [supporting], Methodology [supporting], Software [equal], Writing – review & editing [supporting]), Ali Rahnavard (Funding acquisition [equal], Investigation [equal], Methodology [equal], Supervision [equal], Writing – review & editing [equal]), Keith A. Crandall (Conceptualization [equal], Funding acquisition [equal], Investigation [supporting], Methodology [supporting], Resources [supporting], Software [supporting], Writing – review & editing [equal]), and Todd H. Oakley (Conceptualization [lead], Funding acquisition [equal], Formal Analysis [supporting], Investigation [supporting], Visualization [supporting], Writing - original draft [equal], Writing - review & editing [equal]).

Funding

This work was supported by the US National Science Foundation grants DEB-2153773 and IOS-1754770 to THO and DEB-2109688 to A.R. and K.A.C. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The dataset(s) supporting the results and all other code used in this article are available in the "*Visual Physiology Opsin Database*" GitHub repository archived in Zenodo [134] and [135]. A DOME-ML (Data, Optimisation, Model, and Evaluation in Machine Learning) annotation [136] supporting the current study is available for scruitiny [137].

Competing Interests

The authors declare they have no competing interests.

References

1. Ovsyannikova IG, Poland GA. Vaccinomics: current findings, challenges and novel approaches for vaccine development. AAPS J. 2011;13:438–44. https://doi.org/10.1208/s12248-011-9281-x.
2. Steinbrück L, McHardy AC. Inference of genotype–phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. PLoS Comput Biol. 2012;8:e1002492. https://doi.org/10.1371/journal.pcbi.1002492.
3. Roberts JP. Single-cell analysis deepens antibody discovery. Genet Eng Biotechnol News. 2020;40:23–25. https://doi.org/10.1089/gen.40.02.09.
4. Cobb JN, DeClerck G, Greenberg A, et al. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theor Appl Genet. 2013;126:867–87. https://doi.org/10.1007/s00122-013-2066-0.
5. Chevin L-M, Collins S, Lefèvre F. Phenotypic plasticity and evolutionary demographic responses to climate change: taking theory out to the field. Funct Ecol. 2013;27:967–79. https://doi.org/10.1111/j.1365-2435.2012.02043.x.
6. Franks SJ, Weber JJ, Aitken SN. Evolutionary and plastic responses to climate change in terrestrial plant populations. Evol Appl. 2014;7:123–39. https://doi.org/10.1111/eva.12112.
7. Gienapp P, Teplitsky C, Alho JS, et al. Climate change and evolution: disentangling environmental and genetic responses. Mol Ecol. 2008;17:167–78. https://doi.org/10.1111/j.1365-294X.2007.03413.x.
8. Munday PL, Warner RR, Monro K, et al. Predicting evolutionary responses to climate change in the sea. Ecol Lett. 2013;16:1488–500. https://doi.org/10.1111/ele.12185.
9. Singhal A, Simmons M, Lu Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. PLoS Comput Biol. 2016;12:e1005017. https://doi.org/10.1371/journal.pcbi.1005017.
10. Kemble H, Nghe P, Tenaillon O. Recent insights into the genotype-phenotype relationship from massively parallel

genetic assays. Evol Appl. 2019;12:1721–42. https://doi.org/10.1111/eva.12846.

11. Dikicioglu D, Pir P, Oliver SG. Predicting complex phenotype-genotype interactions to enable yeast engineering: Saccharomyces cerevisiae as a model organism and a cell factory. Biotechnol J. 2013;8:1017–34. https://doi.org/10.1002/biot.201300138.
12. Leung MKK, Delong A, Alipanahi B, et al. Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE. 2016;104:176–97. https://doi.org/10.1109/JPROC.2015.2494198.
13. Guzzetta G, Jurman G, Furlanello C. A machine learning pipeline for quantitative phenotype prediction from genotype data. BMC Bioinf. 2010;11:S3. https://doi.org/10.1186/1471-2105-11-S8-S3.
14. Lee B, Zhang S, Poleksic A, et al. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet. 2019;10:1381. https://doi.org/10.3389/fgene.2019.01381.
15. Lee Y-C, Christensen JJ, Parnell LD, et al. Using machine learning to predict obesity based on genome-wide and epigenome-wide gene-gene and gene-diet interactions. Front Genet. 2021;12:783845. https://doi.org/10.3389/fgene.2021.783845.
16. Yeh AH-W, Norn C, Kipnis Y, et al. De novo design of luciferases using deep learning. Nature. 2023;614:774–80. https://doi.org/10.1038/s41586-023-05696-3.
17. Brandes N, Goldman G, Wang CH, et al. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023;55:1512–22. https://doi.org/10.1038/s41588-023-01465-0.
18. Pinto MF, Oliveira H, Batista S, et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. Sci Rep. 2020;10:21038. https://doi.org/10.1038/s41598-020-78212-6.
19. Himmelstein DS, Baranzini SE. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS Comput Biol. 2015;11:e1004259. https://doi.org/10.1371/journal.pcbi.1004259.
20. Sheikh N. Identification and classification of wildlife from camera-trap images using machine learning and computer vision. 2020. https://norma.ncirl.ie/4283/1/nawazsheikh.pdf. Accessed 12 December 2023.
21. Vélez J, McShea W, Shamon H, et al. An evaluation of platforms for processing camera-trap data using artificial intelligence. Methods Ecol Evol. 2023;14:459–77. https://doi.org/10.1111/2041-210x.14044.
22. Kutugata M, Baumgardt J, Goolsby JA, et al. Automatic camera-trap classification using wildlife-specific deep learning in Nilgai management. J Fish Wildlife Manage 2021;12:412–21.https://doi.org/10.3996/JFWM-20-076.
23. Terakita A. The opsins. Genome Biol. 2005;6:213. https://doi.org/10.1186/gb-2005-6-3-213.
24. Govardovskii VI, Fyhrquist N, Reuter T, et al. In search of the visual pigment template. Vis Neurosci. 2000;17:509–28. https://doi.org/10.1017/s0952523800174036.
25. Jacobs GH, Neitz J, Krogh K. Electroretinogram flicker photometry and its applications. J Opt Soc Am A. 1996;13:641. https://doi.org/10.1364/josaa.13.000641.
26. Thomas MM, Lamb TD. Light adaptation and dark adaptation of human rod photoreceptors measured from the a-wave of the electroretinogram. J Physiol. 1999;518:479–96. https://doi.org/10.1111/j.1469-7793.1999.0479p.x.
27. Rocha FA de F, Gomes BD, Silveira LC de L, et al. Spectral sensitivity measured with electroretinogram using a constant response method. PLoS One. 2016;11:e0147318. https://doi.org/10.1371/journal.pone.0147318.
28. Liebman PA. Microspectrophotometry of photoreceptors. In: Abrahamson EW, Baumann C, Bridges CDB, al. et, eds. Photochemistry of vision. Berlin: Springer; 1972.
29. Yewers MS, McLean CA, Moussalli A, et al. Spectral sensitivity of cone photoreceptors and opsin expression in two colour-divergent lineages of the lizard ctenophorus decresii. J Exp Biol. 2015;218:2979. https://doi.org/10.1242/jeb.131854.
30. Kojima D, Fukada Y. Spectroscopic analysis of wavelength sensitivities of opsin-type photoreceptor proteins. In: Hirota T, Hatori M, Panda S, eds. Circadian clocks. New York: Springer US; 2022.
31. Carlson SD. Microspectrophotometry of visual pigments. Quart Rev Biophys. 1972;5:349–93. https://doi.org/10.1017/s0033583500000986.
32. Bowmaker JK. Microspectrophotometry of vertebrate photoreceptors. A brief review. Vis Res. 1984;24:1641–50. https://doi.org/10.1016/0042-6989(84)90322-5.
33. Merbs SL, Nathans J. Absorption spectra of human cone pigments. Nature. 1992;356:433–35. https://doi.org/10.1038/356433a0.
34. Liénard MA, Valencia-Montoya WA, Pierce NE. Molecular advances to study the function, evolution and spectral tuning of arthropod visual opsins. Phil Trans R Soc B. 2022;377:20210279. https://doi.org/10.1098/rstb.2021.0279.
35. Hagen JFD, Roberts NS, Johnston RJ Jr. The evolutionary history and spectral tuning of vertebrate visual opsins. Dev Biol. 2023;493:40–66. https://doi.org/10.1016/j.ydbio.2022.10.014.
36. Yokoyama S, Radlwimmer FB. The molecular genetics and evolution of red and green color vision in vertebrates. Genetics. 2001;158:1697–710. https://doi.org/10.1093/genetics/158.4.1697.
37. Bloch NI. The evolution of opsins and color vision: connecting genotype to a complex phenotype. Acta Biol Colomb. 2016;21:481. https://doi.org/10.15446/abc.v21n3.53907.
38. Rajamani R, Lin Y-L, Gao J. The opsin shift and mechanism of spectral tuning in rhodopsin. J Comput Chem. 2011;32:854–65. https://doi.org/10.1002/jcc.21663.
39. Hárosi FI. An analysis of two spectral properties of vertebrate visual pigments. Vis Res. 1994;34:1359–67. https://doi.org/10.1016/0042-6989(94)90134-1.
40. Wang W, Geiger JH, Borhan B. The photochemical determinants of color vision: revealing how opsins tune their chromophore's absorption wavelength. Bioessays. 2014;36:65–74. https://doi.org/10.1002/bies.201300094.
41. Smedley GD, McElroy KE, Feller KD, et al. Additive and epistatic effects influence spectral tuning in molluscan retinochrome opsin. J Exp Biol. 2022;225:jeb242929. https://doi.org/10.1242/jeb.242929.
42. Nathans J. Determinants of visual pigment absorbance: identification of the retinylidene Schiff's base counterion in bovine rhodopsin. Biochemistry. 1990;29:9746–52. https://doi.org/10.1021/bi00493a034.
43. Yokoyama S, Xing J, Liu Y, et al. Epistatic adaptive evolution of human color vision. PLoS Genet. 2014;10:e1004884. https://doi.org/10.1371/journal.pgen.1004884.
44. Yokoyama S, Altun A, Jia H, et al. Adaptive evolutionary paths from UV reception to sensing violet light by epistatic interactions. Sci Adv. 2015;1:e1500162. https://doi.org/10.1126/sciadv.1500162.
45. Patel D, Barnes JE, Davies WIL, et al. Short-wavelength-sensitive 2 (Sws2) visual photopigment models combined with

atomistic molecular simulations to predict spectral peaks of absorbance. PLoS Comput Biol. 2020;16:e1008212. https://doi.org/10.1371/journal.pcbi.1008212.

46. Patel JS, Brown CJ, Ytreberg FM, et al. Predicting peak spectral sensitivities of vertebrate cone visual pigments using atomistic molecular simulations. PLoS Comput Biol. 2018;14:e1005974. https://doi.org/10.1371/journal.pcbi.1005974.
47. Karasuyama M, Inoue K, Nakamura R, et al. Understanding colour tuning rules and predicting absorption wavelengths of microbial rhodopsins by data-driven machine-learning approach. Sci Rep. 2018;8:15580. https://doi.org/10.1038/s41598-018-33984-w.
48. Adam ZR, Schwieterman EW, Kacar B. Earliest photic zone niches probed by ancestral microbial rhodopsins. Mol Biol. 2022;39:msac100. https://doi.org/10.1093/molbev/msac100.
49. Longcore T. A compendium of photopigment peak sensitivities and visual spectral response curves of terrestrial wildlife to guide design of outdoor nighttime lighting. Basic Appl Ecol. 2023;73:40–50. https://doi.org/10.1016/j.baae.2023.09.002.
50. Schott RK, Fujita MK, Streicher JW, et al. Diversity and evolution of frog visual opsins: spectral tuning and adaptation to distinct light environments. Mol Biol Evol. 2024;41:msae049. https://doi.org/10.1093/molbev/msae049.
51. Schott RK, Perez L, Kwiatkowski MA, et al. Evolutionary analyses of visual opsin genes in frogs and toads: diversity, duplication, and positive selection. Ecol Evol. 2022;12:e8595. https://doi.org/10.1002/ece3.8595.
52. Schweikert LE, Fitak RR, Caves EM, et al. Spectral sensitivity in ray-finned fishes: diversity, ecology and shared descent. J Exp Biol. 2018;221:jeb189761. https://doi.org/10.1242/jeb.189761.
53. Schweikert LE, Caves EM, Solie SE, et al. Variation in rod spectral sensitivity of fishes is best predicted by habitat and depth. J Fish Biol. 2019;95:179–85. https://doi.org/10.1111/jfb.13859.
54. Molina-Venegas R, Moreno-Saiz JC,Castro Parga I, et al. Assessing among-lineage variability in phylogenetic imputation of functional trait datasets. Ecography. 2018;41:1740–49. https://doi.org/10.1111/ecog.03480.
55. Garland T Jr, Ives AR. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat. 2000;155:346–64. https://doi.org/10.1086/303327.
56. Frazer SA, Baghbanzadeh M, Rahnavard A, et al. The Visual Physiology Opsin Database: a database opsin data and machine-learning models to predict phenotype. GitHub. https://github.com/VisualPhysiologyDB/visual-physiology-opsin-db. Accessed 17 September 2024.
57. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.
58. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–97. https://doi.org/10.1093/nar/gkh340.
59. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52. https://doi.org/10.1093/oxfordjournals.molbev.a026334.
60. Baghbanzadeh M, Dawson T, Sayoldin B, et al. DeepBreaks: a machine learning tool for identifying and prioritizing genotype-phenotype associations [PREPRINT]. Research Square. 2023. rs-2534899. https://doi.org/10.21203/rs.3.rs-2534899/v1.
61. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R Package Version 0 4-2. 2015. https://xgboost.readthedocs.io/en/stable/. Accessed 17 September 2024.
62. Sibindi R, Mwangi RW, Waititu AG. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Eng Rep. 2023;5:e12599. https://doi.org/10.1002/eng2.12599.
63. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;31:46–54. https://doi.org/10.5555/3294996.3295074.
64. Schapire RE. Explaining AdaBoost. In: Schölkopf B, Luo Z, Vovk V, eds. Empirical inference: Festschrift in honor of Vladimir N Vapnik. Berlin: Springer; 2013
65. Rigatti SJ. Random forest. J Insur Med. 2017;47:31–39. https://doi.org/10.17849/insm-47-01-31-39.1.
66. Segal MR. Machine learning benchmarks and random forest regression. escholarship. 2004. https://escholarship.org/uc/item/35x3v9t4. Accessed 17 September 2004.
67. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30. https://doi.org/10.5555/1953048.2078195.
68. Prettenhofer P, Louppe G. Gradient boosted regression trees in Scikit-learn. PyData. 2014. https://orbi.uliege.be/handle/2268/16352. Accessed 17 Septmeber 2024.
69. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–95. https://doi.org/10.1109/MCSE.2007.55.
70. Damian Riina M, Stambaugh C, Stambaugh N, et al. Continuous variable analyses: t-test, Mann–Whitney, Wilcoxin rank. In: Eltorai AEM, Bakal JA, Kim DW, *et al.*eds. Translational radiation oncology. London, UK : Academic Press; 2023.
71. Virtanen P, Gommers R,Oliphant E, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.Nature Methods. 2020;17:261–272. https://doi.org/10.1038/s41592-019-0686-2.
72. Silva JO, Orellana ETV, Torres M. Development of a parallel version of PhyML 3.0 using shared memory. IEEE Latin Am Trans. 2017;15:959–67. https://doi.org/10.1109/TLA.2017.7912593.
73. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25:1307–20. https://doi.org/10.1093/molbev/msn067.
74. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3:217–23. https://doi.org/10.1111/j.2041-210x.2011.00169.x.
75. Sneddon TP, Li P, Edmunds SC. GigaDB: announcing the GigaScience database. Gigascience. 2012;1:1–11. https://doi.org/10.1186/2047-217X-1-11.
76. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–1232. https://doi.org/10.1214/aos/1013203451.
77. Bedoui A, Lazar NA. Bayesian empirical likelihood for ridge and lasso regressions. Comput Stat Data Anal. 2020;145:106917. https://doi.org/10.1016/j.csda.2020.106917.
78. Karabatsos G. Fast marginal likelihood estimation of the ridge parameter(s) in ridge regression and generalized ridge regression for big data. arXiv. 2014. https://doi.org/10.48550/arXiv.1409.2437.
79. Fan J, Ma X, Wu L, et al. Light Gradient Boosting Machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agric Water Manage. 2019;225:105758. https://doi.org/10.1016/j.agwat.2019.105758.
80. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic

structure. Ecography 2017;40:913–29.https://doi.org/10.1111/ecog.02881.
81. Yokoyama S, Radlwimmer FB. The "five-sites" rule and the evolution of red and green color vision in mammals. Mol Biol Evol. 1998;15:560–67. https://doi.org/10.1093/oxfordjournals.molbev.a025956.
82. Shichida Y, Matsuyama T. Evolution of opsins and phototransduction. Phil Trans R Soc B. 2009;364:2881–95. https://doi.org/10.1098/rstb.2009.0051.
83. Terakita A, Koyanagi M, Tsukamoto H, et al. Counterion displacement in the molecular evolution of the rhodopsin family. Nat Struct Mol Biol. 2004;11:284–89. https://doi.org/10.1038/nsmb731.
84. Shi Y, Radlwimmer FB, Yokoyama S. Molecular genetics and the evolution of ultraviolet vision in vertebrates. Proc Natl Acad Sci USA. 2001;98:11731–36. https://doi.org/10.1073/pnas.201257398.
85. Sugawara T, Terai Y, Imai H, et al. Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deepwater cichlids from Lakes Tanganyika and Malawi. Proc Natl Acad Sci USA. 2005;102:5448–53. https://doi.org/10.1073/pnas.0405302102.
86. Takenaka N, Yokoyama S. Mechanisms of spectral tuning in the RH2 pigments of Tokay gecko and American chameleon. Gene. 2007;399:26–32. https://doi.org/10.1016/j.gene.2007.04.036.
87. Yokoyama S, Tada T, Zhang H, et al. Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci USA. 2008;105:13480–85. https://doi.org/10.1073/pnas.0802426105.
88. Shannon C. A mathematical theory of communication. Bell System Technical Journal. 1948;3:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
89. Ramazzotti M, Degl'Innocenti D, Manao G, et al. Entropy calculator: getting the best from your multiple protein alignments. Ital J Biochem. 2004;53:16–22. https://pubmed.ncbi.nlm.nih.gov/15356957/.
90. Lin SW, Sakmar TP. Colour tuning mechanisms of visual pigments. Novartis Found Symp. 1999;224: 124–41. https://doi.org/10.1002/9780470515693.ch8.
91. Chan T, Lee M, Sakmar TP. Introduction of hydroxyl-bearing amino acids causes bathochromic spectral shifts in rhodopsin. Amino acid substitutions responsible for red-green color pigment spectral tuning. J Biol Chem. 1992;267:9478–80. https://doi.org/10.1016/S0021-9258(19)50115-6.
92. Orgogozo V, Morizot B, Martin A. The differential view of genotype–phenotype relationships. Front Genet. 2015;6:179. https://doi.org/10.3389/fgene.2015.00179.
93. Baldwin MW, Ko M-C. Functional evolution of vertebrate sensory receptors. Horm Behav. 2020;124:104771. https://doi.org/10.1016/j.yhbeh.2020.104771.
94. Park Y, Metzger BPH, Thornton JW. Epistatic drift causes gradual decay of predictability in protein evolution. Science. 2022;376:823–30. https://doi.org/10.1126/science.abn6895.
95. Lyons DM, Zou Z, Xu H, et al. Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. Nat Ecol Evol. 2020;4:1685–93. https://doi.org/10.1038/s41559-020-01286-y.
96. Gonzalez Somermeyer L, Fleiss A, Mishin AS, et al. Heterogeneity of the GFP fitness landscape and data-driven protein design. eLife. 2022;11:e75842. https://doi.org/10.7554/eLife.75842.
97. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38:367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.
98. Sekharan S, Morokuma K. Why 11-cis-retinal? Why not 7-cis-, 9-cis-, or 13-cis-retinal in the eye? J Am Chem Soc. 2011;133:19052–55. https://doi.org/10.1021/ja208789h.
99. Buczyłko J, Saari JC, Crouch RK, et al. Mechanisms of opsin activation. J Biol Chem. 1996;271:20621–30. https://doi.org/10.1074/jbc.271.34.20621.
100. Das D, Wilkie SE, Hunt DM, et al. Visual pigments and oil droplets in the retina of a passerine bird, the canary Serinus canaria: microspectrophotometry and opsin sequences. Vis Res. 1999;39:2801–15. https://doi.org/10.1016/s0042-6989(99)00023-1.
101. Toomey MB, Collins AM, Frederiksen R, et al. A complex carotenoid palette tunes avian colour vision. J R Soc Interface. 2015;12:20150563. https://doi.org/10.1098/rsif.2015.0563.
102. Hart NS, Vorobyev M. Modelling oil droplet absorption spectra and spectral sensitivities of bird cone photoreceptors. J Comp Physiol A. 2005;191:381–92. https://doi.org/10.1007/s00359-004-0595-3.
103. Toomey MB, Corbo JC. Evolution, development and function of vertebrate cone oil droplets. Front Neural Circuits. 2017;11:97. https://doi.org/10.3389/fncir.2017.00097.
104. Arikawa K, Stavenga D. Random array of colour filters in the eyes of butterflies. J Exp Biol. 1997;200:2501–6. https://doi.org/10.1242/jeb.200.19.2501.
105. Feller KD, Wilby D, Jacucci G, et al. Long-wavelength reflecting filters found in the larval retinas of one mantis shrimp Family (Nannosquillidae). Curr Biol. 2019;29:3101–8.e4. https://doi.org/10.1016/j.cub.2019.07.070.
106. Partridge JC, White EM, Douglas RH. The effect of elevated hydrostatic pressure on the spectral absorption of deep-sea fish visual pigments. J Exp Biol. 2006;209:314–19. https://doi.org/10.1242/jeb.01984.
107. Ogbunugafor CB, Wylie CS, Diakite I, et al. Adaptive landscape by environment interactions dictate evolutionary dynamics in models of drug resistance. PLoS Comput Biol. 2016;12:e1004710. https://doi.org/10.1371/journal.pcbi.1004710.
108. Woolley S, Johnson J, Smith MJ, et al. TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics. 2003;19:671–72. https://doi.org/10.1093/bioinformatics/btg043.
109. Inoue K, Karasuyama M, Nakamura R, et al. Exploration of natural red-shifted rhodopsins using a machine learning-based Bayesian experimental design. Commun Biol. 2021;4:362. https://doi.org/10.1038/s42003-021-01878-9.
110. Palczewski K, Kumasaka T, Hori T, et al. Crystal structure of rhodopsin: a G protein-coupled receptor. Science. 2000;289:739–45. https://doi.org/10.1126/science.289.5480.739.
111. Murakami M, Kouyama T. Crystal structure of squid rhodopsin. Nature. 2008;453:363–67. https://doi.org/10.1038/nature06925.
112. Briscoe AD. Homology modeling suggests a functional role for parallel amino acid substitutions between bee and butterfly red- and green-sensitive opsins. Mol Biol Evol. 2002;19:983–86. https://doi.org/10.1093/oxfordjournals.molbev.a004158.
113. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–89. https://doi.org/10.1038/s41586-021-03819-2.
114. Van Nynatten A, Castiglione GM, de A Gutierrez E, et al. Recreated ancestral opsin associated with marine to freshwater croaker invasion reveals kinetic and spectral adaptation. Mol Biol Evol. 2021;38:2076–87. https://doi.org/10.1093/molbev/msab008.
115. Porter ML, Roberts NW, Partridge JC. Evolution under pressure and the adaptation of visual pigment compressibility in deep-

sea environments. Mol Phylogenet Evol. 2016;105:160–65. https://doi.org/10.1016/j.ympev.2016.08.007.

116. Schweikert LE, Bagge LE, Naughton LF, et al. Dynamic light filtering over dermal opsin as a sensory feedback system in fish color change. Nat Commun. 2023;14:46422023. https://doi.org/10.1038/s41467-023-40166-4.

117. Borghezan E de A, da Silva Pires TH, Zuanon J, et al. Unstable environmental conditions constrain the fine-tune between opsin sensitivity and underwater light in an Amazon forest stream fish. J Evol Biol. 2024;37:212–24. https://doi.org/10.1093/jeb/voae001.

118. Murphy MJ, Westerman EL. Evolutionary history limits species' ability to match colour sensitivity to available habitat light. Proc R Soc B. 2022;289:612. https://doi.org/10.1098/rspb.2022.0612.

119. Kwon E, Heo WD. Optogenetic tools for dissecting complex intracellular signaling pathways. Biochem Biophys Res Commun. 2020;527:331–36. https://doi.org/10.1016/j.bbrc.2019.12.132.

120. Mukherjee A, Repina NA, Schaffer DV, et al. Optogenetic tools for cell biological applications. J Thorac Dis. 2017;9:4867–70. https://doi.org/10.21037/jtd.2017.11.73.

121. Tischer D, Weiner OD. Illuminating cell signalling with optogenetic tools. Nat Rev Mol Cell Biol. 2014;15:551–58. https://doi.org/10.1038/nrm3837.

122. Kaur P, Saunders TE, Tolwinski NS. Coupling optogenetics and light-sheet microscopy, a method to study Wnt signaling during embryogenesis. Sci Rep. 2017;7:16636. https://doi.org/10.1038/s41598-017-16879-0.

123. Fan H, Barnes C, Hwang H, et al. Precise modulation of embryonic development through optogenetics. Genesis. 2022;60:e23505. https://doi.org/10.1002/dvg.23505.

124. Sparta DR, Jennings JH, Ung RL, et al. Optogenetic strategies to investigate neural circuitry engaged by stress. Behav Brain Res 2013;255:19–25. https://doi.org/10.1016/j.bbr.2013.05.007.

125. Belzung C, Turiault M, Griebel G. Optogenetics to study the circuits of fear- and depression-like behaviors: a critical analysis. Pharmacol Biochem Behav. 2014;122:144–57. https://doi.org/10.1016/j.pbb.2014.04.002.

126. Muir J, Lopez J, Bagot RC. Wiring the depressed brain: optogenetic and chemogenetic circuit interrogation in animal models of depression. Neuropsychopharmacol. 2019;44:1013–26. https://doi.org/10.1038/s41386-018-0291-6.

127. LaLumiere RT. A new technique for controlling the brain: optogenetics and its potential for use in research and the clinic. Brain Stimulation. 2011;4:1–6. https://doi.org/10.1016/j.brs.2010.09.009.

128. Montagni E, Resta F, Mascaro ALA, et al. Optogenetics in brain research: from a strategy to investigate physiological function to a therapeutic tool. Photonics. 2019;6:92. https://doi.org/10.3390/photonics6030092.

129. Penn WD, McKee AG, Kuntz CP, et al. Probing biophysical sequence constraints within the transmembrane domains of rhodopsin by deep mutational scanning. Sci Adv. 2020;6:eaay7505. https://doi.org/10.1126/sciadv.aay7505.

130. Hensley NM, Ellis EA, Leung NY, et al. Selection, drift, and constraint in cypridinid luciferases and the diversification of bioluminescent signals in sea fireflies. Mol Ecol. 2021;30:1864–79. https://doi.org/10.1111/mec.15673.

131. Schenkmayerova A, Pinto GP, Toul M, et al. Engineering the protein dynamics of an ancestral luciferase. Nat Commun. 2021;12:3616. https://doi.org/10.1038/s41467-021-23450-z.

132. Baghbanzadeh M, Dawson T, Sayoldin B, et al. Ali Rahnavard: omicseye/deepbreaks-dc—docker image. 2023. https://hub.docker.com/r/omicseye/deepbreaks-dc. Accessed 27 August 2024.

133. Baghbanzadeh M, Dawson T, Sayoldin B, et al. deepBreaks: a machine learning tool for identifying and prioritizing genotype-phenotype associations. CodeOcean. 2024. https://doi.org/10.24433/CO.0636307.v1.

134. Frazer S. VisualPhysiologyDB/visual-physiology-opsin-db: vpod_v1.0_for_publication. Zenodo. https://doi.org/10.5281/zenodo.10667840. Date of Deposit: Feb 15, 2024.

135. Frazer S. VisualPhysiologyDB/visual-physiology-opsin-db: vpod_v1.1_for_publication. Zenodo. https://doi.org/10.5281/zenodo.12213246. Date of Deposit: June 21, 2024 .

136. Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18:1122–27. https://doi.org/10.1038/s41592-021-01205-4.

137. Frazer S.A. et al. Annotations to: "Discovering genotype-phenotype relationships with machine learning and the Visual Physiology Opsin Database (VPOD)" DOME Registry. 2024. https://registry.dome-ml.org/review/88n4lxv68p.

Received: February 15, 2024. **Revised:** June 25, 2024. **Accepted:** September 1, 2024