Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics

Claire N. Bedbrook [1], Kevin K. Yang[2,3], J. Elliott Robinson [1,3], Elisha D. Mackey[1], Viviana Gradinaru [1*] and Frances H. Arnold [1,2*]

We engineered light-gated channelrhodopsins (ChRs) whose current strength and light sensitivity enable minimally invasive neuronal circuit interrogation. Current ChR tools applied to the mammalian brain require intracranial surgery for transgene delivery and implantation of fiber-optic cables to produce light-dependent activation of a small volume of tissue. To facilitate expansive optogenetics without the need for invasive implants, our engineering approach leverages the substantial literature of ChR variants to train statistical models for the design of high-performance ChRs. With Gaussian process models trained on a limited experimental set of 102 functionally characterized ChRs, we designed high-photocurrent ChRs with high light sensitivity. Three of these, ChRger1–3, enable optogenetic activation of the nervous system via systemic transgene delivery. ChRger2 enables light-induced neuronal excitation without fiber-optic implantation; that is, this opsin enables transcranial optogenetics.

ChRs are light-gated ion channels found in photosynthetic algae. Transgenic expression of ChRs in the brain enables light-dependent neuronal activation[1]. These channels are widely applied as tools in neuroscience research[2]; however, limitations of available ChRs prohibit a number of optogenetic applications. These channels have broad activation spectra in the visible range and require high-intensity light for activation (~1 mW mm^{-2}). As ChRs are naturally low-conductance channels, 10^5–10^6 functional ChRs need to be expressed in the plasma membrane of a neuron to produce sufficient light-dependent depolarization and to induce neuronal activation[3]. In the mouse brain, ChRs require ~1–15 mW light delivered ~100 μm from the target cell population to reliably activate action potentials[4–6]. This confines light-dependent activation to a small volume of brain tissue (~1 mm^3)[7]. Enabling optogenetics for large brain volumes without the need to implant invasive optical fibers for light delivery would be highly desirable.

Engineering ChRs to overcome limits in conductance and light sensitivity and extend the reach of optogenetic experiments requires overcoming three major challenges. First, rhodopsins are trans-membrane proteins that are inherently difficult to engineer because the sequence and structural determinants of membrane protein expression and plasma membrane localization are highly constrained and poorly understood[8,9]. Second, because properties of interest for neuroscience applications are assayed using low-throughput techniques, such as patch-clamp electrophysiology, engineering by directed evolution is not feasible[10]. And third, in vivo applications require retention and optimization of multiple properties; for example, localization in mammalian cells while simultaneously tuning kinetics, photocurrents and spectral properties[6].

Diverse ChRs have been published, including variants discovered from nature[11,12], variants engineered through recombination[9,13] and mutagenesis[14,15], as well as variants resulting from rational design[16]. Studies of these coupled with structural information[17] and molecular dynamic simulations[18] have established some understanding of the mechanics and sequence features important for specific ChR properties[1,16]. Despite this, it is still not possible to predict functional properties of uncharacterized ChR sequences.

Our approach has been to leverage the substantial literature of ChRs to train statistical models that enable design of highly functional ChRs. These models take as their input the sequence and structural information for a given ChR variant and then predict its functional properties. The models use training data to learn how sequence and structural elements map to ChR function. Once known, that mapping can be used to predict the behavior of untested ChR variants.

Our trained models accurately predict the properties of untested ChR sequences. We used these models to engineer 30 'designer' ChR variants with specific combinations of desired properties, a number of which have unprecedented photocurrent strength and light sensitivity. We characterized these low-light-sensitive, high-photocurrent ChRs for applications in the mammalian brain and demonstrate their potential for minimally invasive activation of populations of neurons in the brain enabled by systemic transgene delivery with the engineered adeno-associated virus (AAV), rAAV-PHP.eB (ref. [19]). This work demonstrates how a machine learning-guided approach can enable engineering of proteins that have been challenging to engineer.

Results

Functional characterization of ChR variants for machine learning. In previous work, we explored structure-guided recombination[20,21] of three highly functional ChR parents (CsChrimsonR (CsChrimR)[11], C1C2 (ref. [17]) and CheRiff (ref. [22])) by designing two ten-block recombination libraries with a theoretical size of ~120,000 (that is, 2×3^{10}) chimeric variants[9]. These recombination libraries are a rich source of functionally diverse sequences[9]. Previously,

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA. [3]These authors contributed equally: Kevin K. Yang, J. Elliott Robinson. *e-mail: viviana@caltech.edu; fha@cheme.caltech.edu

we generated 102 ChR variants selected from these recombination libraries[8,9]. Of these, 76 were previously characterized for photocurrent properties (with patch-clamp electrophysiology) and 26 we characterized here. Together, these 102 ChR recombination variants provide the primary dataset used for model training. We supplemented this dataset with data from other published sources, including 19 ChR variants from nature, 14 single-mutant ChR variants and 28 recombination variants from other libraries (Supplementary Data 1). As the data produced by other laboratories were not collected under the same experimental conditions as data collected in our hands, they cannot be used for comparison for absolute ChR properties (that is, photocurrent strength); however, these data do provide useful binary information on whether a sequence variant is functional or not. Thus, we used published data from other sources when training binary classification models for ChR function.

Because our goal was to optimize photocurrent strength, wavelength sensitivity and off-kinetics, we used these measured properties to train machine-learning models (Fig. 1a). Enhancing ChR photocurrent strength would enable reliable neuronal activation even under low-light conditions. Different off-rates can be useful for specific applications: fast off-kinetics enable high-frequency optical stimulation[23], slow off-kinetics are correlated with increased light sensitivity[3,14,15] and very slow off-kinetics can be used for constant depolarization (step-function opsins[14]). In addition to opsin functional properties, it is also necessary to optimize or maintain plasma membrane localization, a prerequisite for ChR function[8].

Training Gaussian process (GP) classification and regression models. Using the ChR sequence and structure as well as functional data as inputs, we trained GP classification and regression models (Fig. 1). GP models successfully predicted thermostability, substrate binding affinity and kinetics for several soluble enzymes[24], and ChR membrane localization[8]. For a detailed description of the GP model architecture used for protein engineering see refs. [8,24]. Briefly, these models infer predictive values for uncharacterized sequences from training examples by assuming that similar inputs (ChR sequence variants) will have similar outputs (photocurrent properties). To quantify the relatedness of inputs (ChR sequence variants), we compared both sequence and structure. ChR sequence information is encoded in the amino acid sequence. For structural comparisons, we convert the three-dimensional crystal-structural information into a 'contact map' that is convenient for modeling. Two residues are considered to be in contact and potentially important for structural and functional integrity if they have any nonhydrogen atoms within 4.5 Å in the C1C2 crystal structure (3UG9.pdb)[17]. We defined the sequence and structural similarity between two variants by aligning them and counting the number of positions and contacts at which they are identical[24].

We trained a binary classification model to predict whether a ChR sequence will be functional using all 102 training sequences from the recombination library (Supplementary Data 2) as well as data from 61 variants published by others (Supplementary Data 1). We then used this trained classification model to predict whether uncharacterized ChR sequence variants were functional (Fig. 1b). To test prediction accuracy, we performed 20-fold cross-validation on the training dataset and achieved an area under the receiver operator curve of 0.78, indicating good predictive power (Supplementary Table 1).

Next, we trained three regression models, one for each of the ChR photocurrent properties of interest: photocurrent strength, wavelength sensitivity of photocurrents and off-kinetics (Fig. 1c). Once trained, we used these models to predict photocurrent properties of untested ChRs sequence variants. To test prediction accuracy, we performed 20-fold cross-validation on the training dataset and observed high correlation between predicted and measured properties for all models (Pearson correlation coefficient (R)

between 0.77 and 0.9; Supplementary Tables 1–2). Models built using contact maps from either the ChR2 crystal structure[25] or C1Chrimson crystal structure[26] perform as well as models built with a contact map from the C1C2 structure[17] (Supplementary Table 3 and Supplementary Fig. 1c,d), even though these maps share only 82% and 89% of their contacts with the C1C2 map, respectively (Supplementary Fig. 1a,b).

Selection of designer ChRs using trained models. To select 'designer' ChRs (that is, ChRs predicted to have a useful combination of properties), we used a tiered approach (Fig. 1d). First, we eliminated all ChR sequences predicted to not localize to the plasma membrane or predicted to be nonfunctional. To do this, we used classification models of ChR localization[8] and function to predict the probability of localization and function for each ChR sequence in the 120,000-variant recombination library. Not surprisingly, most ChR variants were predicted to not localize and not function. To focus on ChR variants predicted to localize and function, we set a threshold for the product of the predicted probabilities of localization and function (Fig. 1b); any ChR sequence above that threshold would be considered for the next tier of the process. We selected a conservative threshold of 0.4.

The training data made clear that the higher the mutation distance from one of the three parents, the less likely it was that a sequence would be functional; however, we expect that more diverse sequences would also have the more diverse functional properties. To explore diverse sequences predicted to function, we selected 22 ChR variants that passed the 0.4 threshold and were multi-block-swap sequences containing on average 70 mutations from the closest parent. We synthesized these 22 sequences, expressed them in HEK cells and measured their photocurrent properties with patch-clamp electrophysiology. Of the tested sequences 59% were functional (Fig. 1e), compared with 38% of the multi-block-swap sequences randomly selected (that is, not selected by the model) and having comparable average mutation level. This validates the classification model's ability to make useful predictions about uncharacterized functional sequences, even for sequences that are very distant from those previously tested. We updated the models by including data from these 22 sequences for future rounds of predictions.

From the 120,000-variant recombination library, 1,161 chimeric sequence variants passed the 0.4 predicted localization and function threshold (Fig. 1). For the second tier of the selection process, we used the three regression models trained on all functional variants collected up to this point to predict the photocurrent strength, wavelength sensitivity of photocurrents and off-kinetics for each of these 1,161 ChR sequence variants (Supplementary Data 3). We selected 28 designer ChRs predicted to be highly functional with different combinations of properties, including those predicted to have the highest photocurrent strength, most red-shifted or blue-shifted activation wavelengths and off-kinetics from very fast to very slow (Supplementary Figs. 2 and 3).

We synthesized genes encoding the 28 selected designer ChR variants, expressed them in HEK cells and characterized them for their photocurrent properties with patch-clamp electrophysiology (Figs. 1 and 2). All 28 designer ChRs selected using the updated classification model above the 0.4 threshold both localize and function. For each of the designer ChR variants, the measured photocurrent properties correlated well with the model predictions ($R > 0.9$ for all models) (Fig. 1f and Supplementary Table 1), demonstrating the power of this data-driven predictive method for engineering designer ChRs. As a negative control, we selected two ChR variant sequences from the recombination library that the model predicted would be nonfunctional (ChR_29_10 and ChR_30_10). These sequences resulted from a single-block swap from two of the most highly functional ChR recombination variants tested. As predicted, these sequences were nonfunctional (Fig. 2b), which shows that
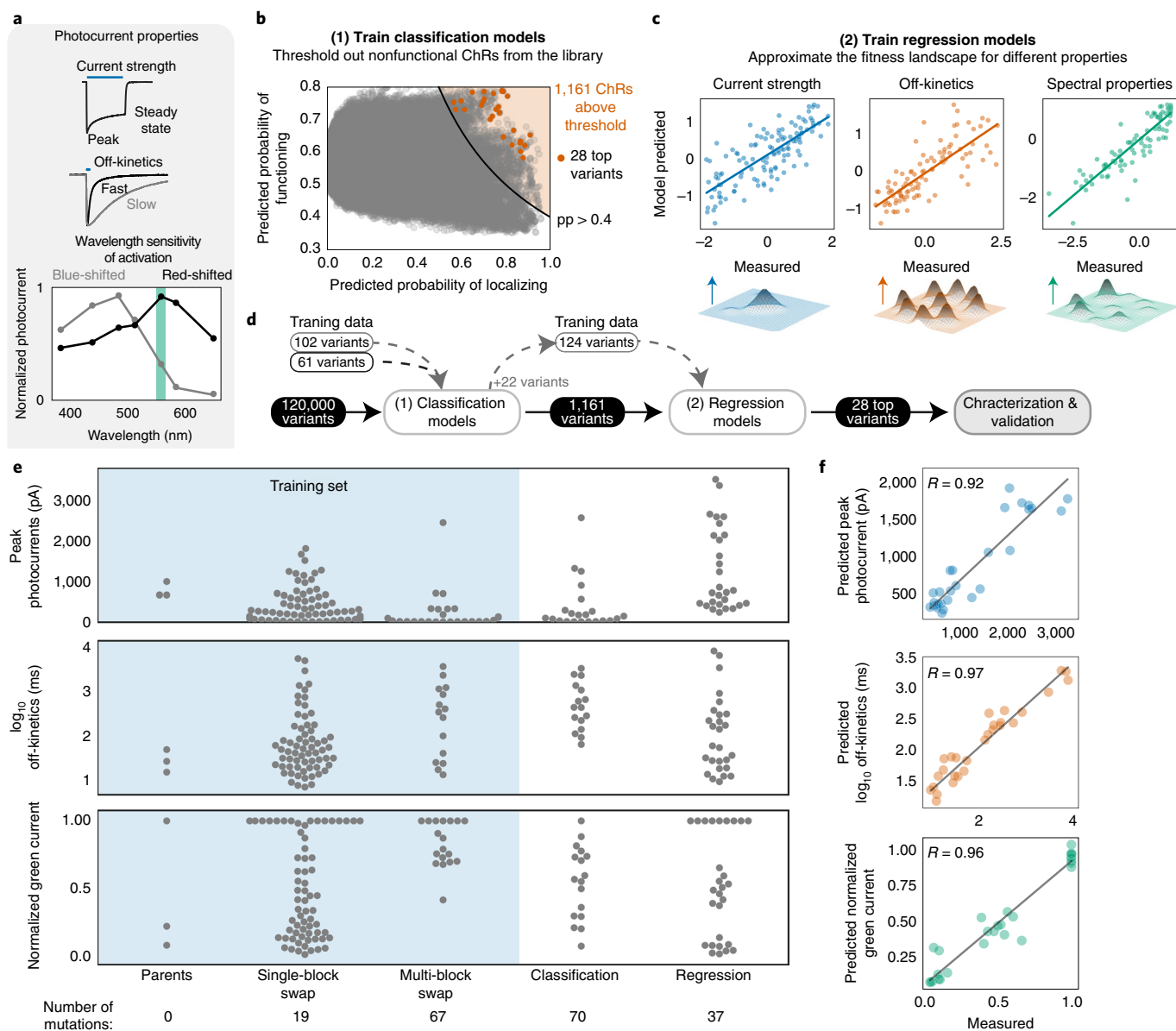
Fig. 1 | Machine learning-guided optimization of ChRs. a, ChR properties used for training classification and regression models. **b**, The classification models' predicted probabilities of function and localization. A threshold for the product of the predicted probabilities (pp) of 0.4 to eliminate nonfunctional ChRs is indicated. **c**, Regression model predictions versus measured photocurrent property on the training set data (20-fold cross-validation). Regression models approximate the fitness landscape for each property of interest (insets show hypothetical fitness landscapes). **d**, Machine-learning pipeline to identify 28 top ChR variants. **e**, Measurements of training set and model-predicted ChR photocurrent properties ($n = 3$–8 cells per variant; Supplementary Data 2). Each gray-colored point is a ChR variant. Training set data are shaded in blue. Mean number of mutations for each set is below the plots. **f**, Regression model predictions versus measured photocurrent property for each of the 28 designer ChRs (that is, the test set).

ChR functionality can be attenuated by incorporating even minimal diversity at certain positions.

Sequence and structural determinants of ChR functional properties. We used L1-regularized linear regression models to identify a limited set of residues and structural contacts that strongly influence ChR photocurrent strength, spectral properties and off-kinetics (Supplementary Fig. 4a). We can assess the relative importance of each of these sequence and structural features by weighting their contributions using L2-regularized linear regression (Supplementary Data 4 and Supplementary Fig. 4). For each functional property, we identified a set of important residues and contacts and their respective weights. A specific residue or contact at a given position is weighted as likely to lead to, for example, low

(negative weight) or high (positive weight) photocurrents. A number of residues and contacts most important for tuning spectral properties are proximal to the retinal-binding pocket, including the blue-shifting contact between A206 and F269 and the blue-shifting contact between F265 and I267, which are conserved in the blue-shifted parents C1C2 and CheRiff, while the red-shifting contact between F201 and Y217 originates from the red-shifted CsChrimR parent (Supplementary Fig. 4). The most heavily weighted contact contributing to off-kinetics includes the reside D195 (that is, D156 according to ChR2 numbering) (Supplementary Fig. 4), a residue that is part of the DC-gate[1]. Mutation of either the aspartic acid or cysteine within the DC-gate has been shown to decrease off-kinetic speed[14,27]. While the cysteine in the DC-gate is conserved in all three ChR parents, the aspartic acid at position 195 is only conserved in
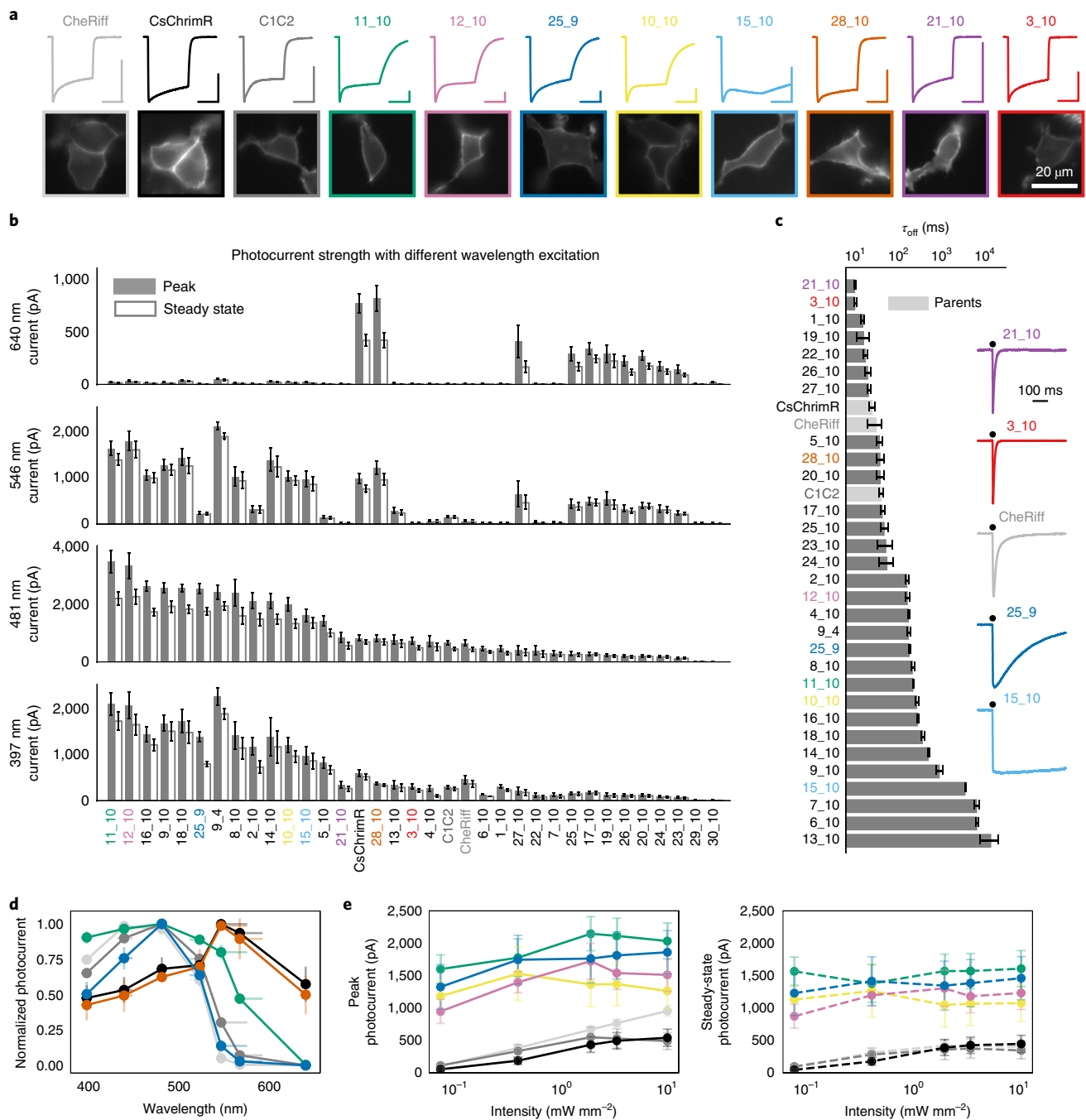
Fig. 2 | The model-predicted ChRs exhibit a large range of functional properties often far exceeding the parents. a, Representative current traces after 0.5 s light exposure for select designer ChR variants with corresponding expression and localization in HEK cells. Vertical colored scale bar for each ChR current trace represents 500 pA, and horizontal scale bar represents 250 ms. The variant color presented in **a** is constant throughout panels. **b**, Measured peak and steady-state photocurrents with different wavelengths of light in HEK cells ($n = 4$–8 cells, see Supplementary Data 2); 397 nm light at 1.5 mW mm^{-2}, 481 nm light at 2.3 mW mm^{-2}, 546 nm light at 2.8 mW mm^{-2} and 640 nm light at 2.2 mW mm^{-2}. **c**, Off-kinetics decay rate (τ_{off}) following a 1 ms exposure to 481 nm light at 2.3 mW mm^{-2} ($n = 4$–8 cells, see Supplementary Data 2). Parent ChRs are highlighted in light gray. Insets show representative normalized current traces with 1 ms light exposure for select ChRs revealing distinct profiles: ChR_21_10 and ChR_3_10 turn off rapidly, ChR_25_9 turns off more slowly and ChR_15_10 exhibits little decrease in photocurrent 0.5 s after the light exposure. **d**, Wavelength sensitivity of activation for select ChRs compared with parental ChRs (CheRiff, $n = 5$ cells; CsChrimR, $n = 5$ cells; C1C2, $n = 4$ cells; 28_10, $n = 5$ cells; 11_10, $n = 5$ cells; 25_9, $n = 5$ cells). **e**, Peak and steady-state photocurrent strengths with varying light irradiances compared with parental ChRs (CheRiff, $n = 6$ cells; CsChrimR, $n = 5$ cells; C1C2, $n = 4$ cells; 11_10, $n = 6$ cells; 12_10, $n = 7$ cells; 25_9, $n = 5$ cells; 10_10, $n = 4$ cells). Top variants ChR_9_4, ChR_25_9 and ChR_11_10 are named ChRger1, ChRger2 and ChRger3 in subsequent figures. Plotted data are mean ± s.e.m.

CheRiff and C1C2 but not in CsChrimR, which has a cysteine at that position. D195 is also part of a contact with L192 that contributes strongly to photocurrent strength (Supplementary Fig. 4).

A number of contacts proximal to retinal contribute strongly to photocurrent strength. For example, the most heavily weighted contact includes A295 (from CsChrimR), which is adjacent to the

conserved lysine residue that covalently links retinal (Supplementary Fig. 4). This position is a serine in both C1C2 and CheRiff.

Machine-guided search identifies ChRs with a range of useful functional properties. We assessed photocurrent amplitude, wavelength sensitivity and off-kinetics of the designer ChRs and the three parental ChRs (Fig. 2). In addition to the 28 regression model-predicted ChRs, we also assessed the two top-performing ChRs from the classification models' predictions (ChR_9_4 and ChR_25_9), for a total of 30 model-predicted highly functional ChRs as well as the two negative control ChRs (ChR_29_10, ChR_30_10). Of the 30 model-predicted ChRs, we found 12 variants with at least two times higher blue-light-activated photocurrents than the top-performing parent (CsChrimR) (Fig. 2b). Three variants exhibit at least 1.7 times higher green-light-activated photocurrents than CsChrimR. Eight variants have larger red-light-activated photocurrents when compared with the blue-light-activated parents (CheRiff and C1C2), though none outperform CsChrimR. Both ChR variants predicted to be nonfunctional by the models produce <30 pA currents.

Designer ChRs' off-kinetics span three orders of magnitude ($\tau_{off} = 10$ ms to >10 s) (Fig. 2c). This range is quite remarkable given that all designer ChRs are built from sequence blocks of three parents that have similar off-kinetics ($\tau_{off} = 30$–50 ms). We found that five designer ChRs have faster off-kinetics than the fastest parent, while 16 have more than five times slower off-kinetics. The two fastest variants, ChR_3_10 and ChR_21_10, exhibit $\tau_{off} = 13 \pm 0.9$ ms and 12 ± 0.4 ms, respectively (mean ± s.e.m.). Four ChRs have particularly slow off-kinetics with $\tau_{off} > 1$ s, including ChR_15_10, ChR_6_10 and ChR_13_10 ($\tau_{off} = 4.3 \pm 0.1$ s, 8.0 ± 0.5 s and 17 ± 7 s, respectively). Two ChRs with very large photocurrents, ChR_25_9 and ChR_11_10, exhibit $\tau_{off} = 220 \pm 10$ ms and 330 ± 30 ms, respectively.

Three designer ChRs exhibit interesting spectral properties (Fig. 2d and Supplementary Fig. 5). ChR_28_10's red-shifted spectrum matches that of CsChrimR, demonstrating that incorporating sequence elements from blue-shifted ChRs into CsChrimR can still generate a red-shifted activation spectrum. ChR_11_10 has a broad activation spectrum relative to the parental spectra, with similar steady-state current strength from 400 to 546 nm light and strong currents (700 ± 100 pA) when activated with 567 nm light. ChR_25_9, on the other hand, exhibits a narrow activation spectrum relative to the parental spectra, with a peak at 481 nm light.

We assessed the light sensitivity of select designer ChRs. Compared with CsChrimR, CheRiff and C1C2, the designer ChRs have at least nine times larger currents at the lowest intensity of light tested (10^{-1} mW mm^{-2}), larger currents at all intensities of light tested and minimal decrease in photocurrent magnitude over the range of intensities tested (10^{-1}–10^1 mW mm^{-2}), suggesting that photocurrents were saturated at these intensities and would only attenuate at much lower light intensities (Fig. 2e). These select designer ChRs are expressed at levels similar to the CsChrimR parent (the highest-expressing parent), indicating that the improved photocurrent strength of these ChRs is not solely due to improved expression (Supplementary Figs. 6 and 7).

We compared three of the top designer ChRs (ChR_9_4, ChR_25_9 and ChR_11_10) with ChR2(H134R) (ref. [6,28]), an enhanced photocurrent single mutant of ChR2 commonly used for in vivo optogenetics, and CoChR (from *Chloromonas oogama*)[11], reported to be one of the highest-conducting ChRs activated by blue light[11] (Supplementary Fig. 8). The selected designer ChRs produce three to six times larger photocurrents than ChR2(H134R) when exposed to high-intensity (2.2 mW mm^{-2}) 481 nm light and 10–18 times larger photocurrents than ChR2(H134R) when exposed to low-intensity (6.5×10^{-2} mW mm^{-2}) 481 nm light (Supplementary Fig. 8d–g). Although CoChR produced peak currents of similar magnitude to the designer ChRs, CoChR decays to a much lower steady-state level (Supplementary Fig. 8d,e), with the designer ChRs producing two to three times larger steady-state photocurrents than CoChR when exposed to high-intensity light and three to four times larger steady-state photocurrents than CoChR when exposed to low-intensity light (Supplementary Fig. 8f,g and Supplementary Table 4). The increased low-light sensitivity of these select designer ChRs is likely due in part to their relatively slow off-kinetics, leading to the increased accumulation of the open state under low-light conditions[14].

Validation of designer ChRs for neuroscience applications. For further validation we selected three of the top high-conductance ChRs, ChR_9_4, ChR_25_9 and ChR_11_10, and renamed them ChRger1, ChRger2 and ChRger3, respectively, for channelrhodopsin Gaussian process-engineered recombinant opsin (Fig. 3 and Supplementary Fig. 9). When expressed in cultured neurons, the ChRgers display robust membrane localization and expression throughout the neuron soma and neurites (Fig. 3b). The ChRgers outperform both CoChR and ChR2(H134R) in photocurrent strength with low-intensity light in neuronal cultures (Fig. 3c). The designer ChRgers require one to two orders of magnitude lower light intensity than CoChR and ChR2(H134R) for neuronal activation (Fig. 3d and Supplementary Fig. 8h).

Next, we performed direct intracranial injections into the mouse prefrontal cortex (PFC) of rAAV-PHP.eB packaging either ChRger1–3 or ChR2(H134R) under the hSyn promoter (Supplementary Table 5). After 3–5 weeks of expression, we measured light sensitivity in ChR-expressing neurons in acute brain slices. We observed greater light sensitivity for the ChRgers compared with ChR2(H134R) (Fig. 3g,h). The ChRgers exhibit >200 pA photocurrent at 10^{-3} mW mm^{-2}, while at the equivalent irradiance ChR2(H134R) exhibits undetectable photocurrents. The ChRgers reach >1,000 pA photocurrents with ~10^{-2} mW mm^{-2} light, a four-fold improvement over ChR2(H134R)'s irradiance-matched photocurrents (Fig. 3g). Our characterization of ChR2(H134R)'s light sensitivity and photocurrent strength is consistent with previous results from other laboratories[6,22].

Designer ChRs and systemic AAVs enable minimally invasive optogenetic excitation. We investigated whether these light-sensitive, high-photocurrent ChRs could provide optogenetic activation coupled with minimally invasive gene delivery. Previous reports of noninvasive optogenetics relied on invasive intracranial virus delivery, which results in many copies of virus per cell and thus very high expression levels of the injected construct[29]. The AAV capsid rAAV-PHP.eB (ref. [19]) produces broad transduction throughout the central nervous system with a single minimally invasive intravenous injection in the adult mouse[30,31]. Use of rAAV-PHP.eB for optogenetic applications has been limited, however, by the low multiplicity of infection with systemically delivered viral vectors, resulting in insufficient opsin expression and light-evoked currents to control neuronal firing with commonly used channels (for example, ChR2).

We hypothesized that the ChRgers could overcome this limitation and allow large-volume optogenetic excitation following systemic transgene delivery. We systemically delivered rAAV-PHP.eB packaging ChRger1–3, CoChR or ChR2(H134R) under the hSyn promoter and observed broad expression throughout the brain (Fig. 3i). Using cell-attached recordings in acute brain slices, we measured the fraction of opsin-expressing cells with sufficient opsin-mediated currents for light-induced firing (Fig. 3j). Only 4% of ChR2(H134R)-expressing neurons produced light-induced firing, while 77% of CoChR-expressing neurons, 89% of ChRger1-expressing neurons and 100% of ChRger2- or ChRger3-expressing neurons produced light-induced activity. With systemic delivery, we observed superior light sensitivity of ChRgers compared with CoChR in both photocurrent strength (Fig. 3k) and spike
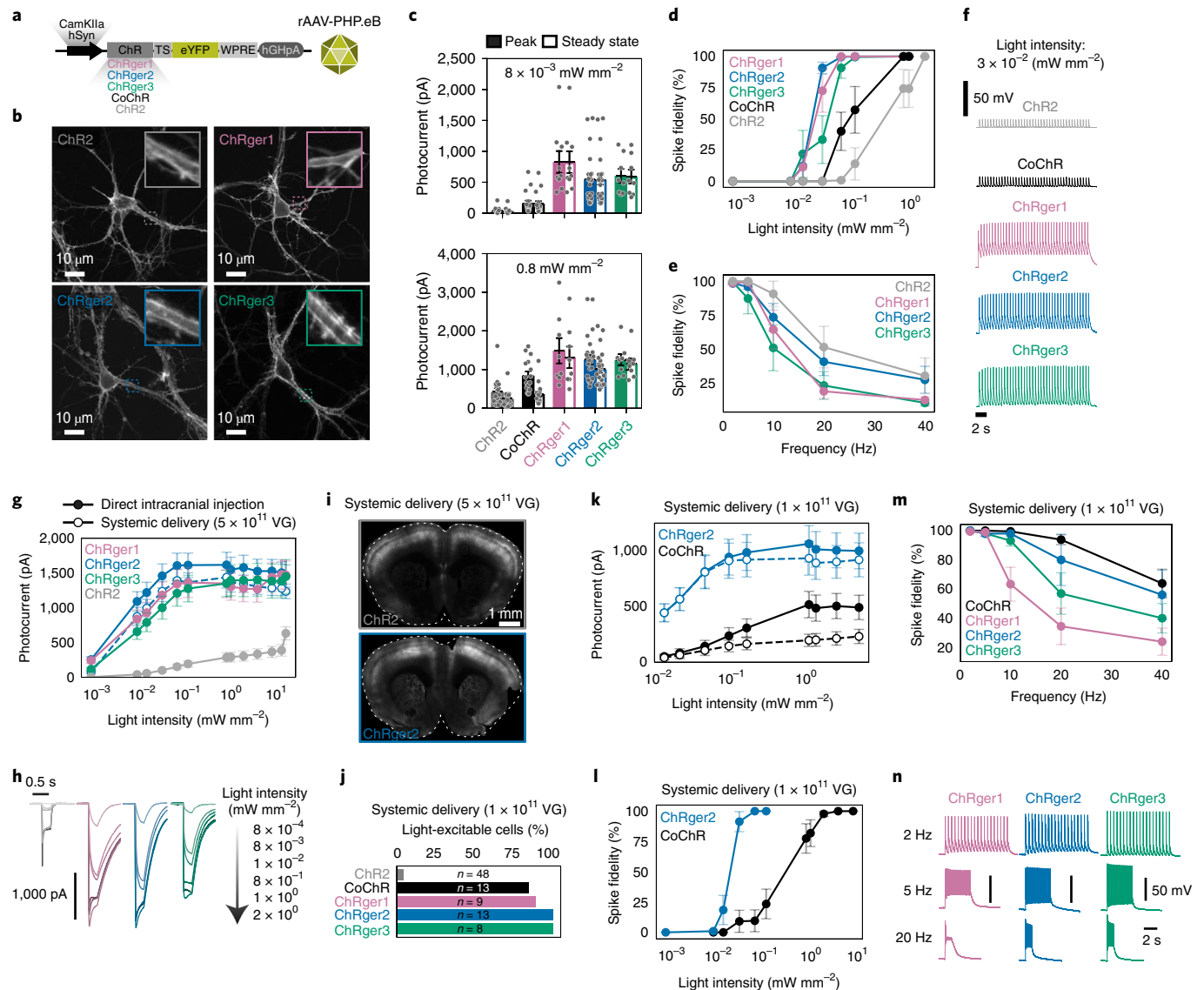
Fig. 3 | ChRger variants in cultured neurons and in acute brain slices outperform ChR2(H134R) and CoChR. a, ChRs were cloned into an AAV vector with either the hSyn or CamKIIa promoter and packaged into rAAV-PHP.eB for expression in culture and in vivo. **b**, Cultured neurons expressing ChRgers and ChR2(H134R) under the hSyn promoter (repeated independently six times per construct with similar results). **c**, Peak and steady-state photocurrent with low-intensity (8×10^{-3} mW mm^{-2}) and moderate-intensity (0.8 mW mm^{-2}) light in cultured neurons (ChR2, $n = 16$ cells; CoChR, $n = 17$ cells; ChRger1, $n = 9$ cells; ChRger2, $n = 24$ cells; ChRger3, $n = 9$ cells). **d**, Spike fidelity with varying-intensity light for 5 ms light-pulse width at 2 Hz stimulation (ChRger1, $n = 6$ cells; ChRger2, $n = 6$ cells; ChRger3, $n = 6$ cells; CoChR, $n = 7$ cells; ChR2, $n = 7$ cells). **e**, Spike fidelity with varying stimulation frequency with 2 ms light-pulse width in cultured neurons (ChRger1, $n = 9$ cells; ChRger2, $n = 12$ cells; ChRger3, $n = 7$ cells; ChR2, $n = 8$ cells). **f**, Representative voltage traces of ChRgers, ChR2(H134R) and CoChR at 2 Hz with 5 ms pulsed low-intensity blue-light stimulation (3×10^{-2} mW mm^{-2}) show robust neuronal firing for ChRgers while ChR2(H134R) and CoChR exhibit only subthreshold light-induced depolarization. **g**, Photocurrent strength with varying light irradiances in acute brain slice after direct injection of rAAV-PHP.eB-packaged hSyn-ChR constructs into the PFC (ChRger1, $n = 11$ cells; ChRger2, $n = 11$ cells; ChRger3, $n = 11$ cells; ChR2, $n = 9$ cells) or after systemic delivery of CamKIIa-ChRger2 (ChRger2, $n = 6$ cells; 5×10^{11} VG per animal). **h**, Representative current traces of ChRgers and ChR2(H134R) with a 300 ms light pulse at varying light irradiances in acute brain slice after direct injection. **i**, Systemic delivery of rAAV-PHP.eB-packaged hSyn-ChRger2 or hSyn-ChR2(H134R) resulted in broad expression throughout the cortex (5×10^{11} VG per animal; repeated independently five times per construct with similar results). **j**, The fraction of light-excitable neurons in the PFC after systemic delivery of hSyn-ChRs measured by cell-attached recording in acute slice targeting only neurons expressing the eYFP marker (1×10^{11} VG per animal). **k,l**, Peak (solid line) and steady-state (dashed line) photocurrent strength (**k**) and spike fidelity (**l**) with varying light irradiances in acute brain slice after systemic delivery (1×10^{11} VG per animal) of hSyn-ChRger2 ($n = 13$ cells) and hSyn-CoChR ($n = 14$ cells) (recorded in PFC neurons). **m**, Spike fidelity with varying stimulation frequency in acute brain slice after systemic delivery (1×10^{11} VG per animal) (CoChR, $n = 15$ cells; ChRger1, $n = 9$ cells; ChRger2, $n = 5$ cells; ChRger3, $n = 8$ cells) with 1 mW mm^{-2} light intensity. **n**, Representative voltage traces with blue-light-driven (1 mW mm^{-2}) spiking at the indicated frequencies. VG, viral genomes. Plotted data are mean ± s.e.m.

fidelity (Fig. 3l). ChRger2-expressing neurons exhibit healthy membrane properties similar to CoChR- or ChR2(H134R)-expressing neurons both in culture and in slice (Supplementary Fig. 10 and

Supplementary Table 6). These results demonstrate the need for light-sensitive and high-photocurrent opsins for applications where systemic delivery is desired.
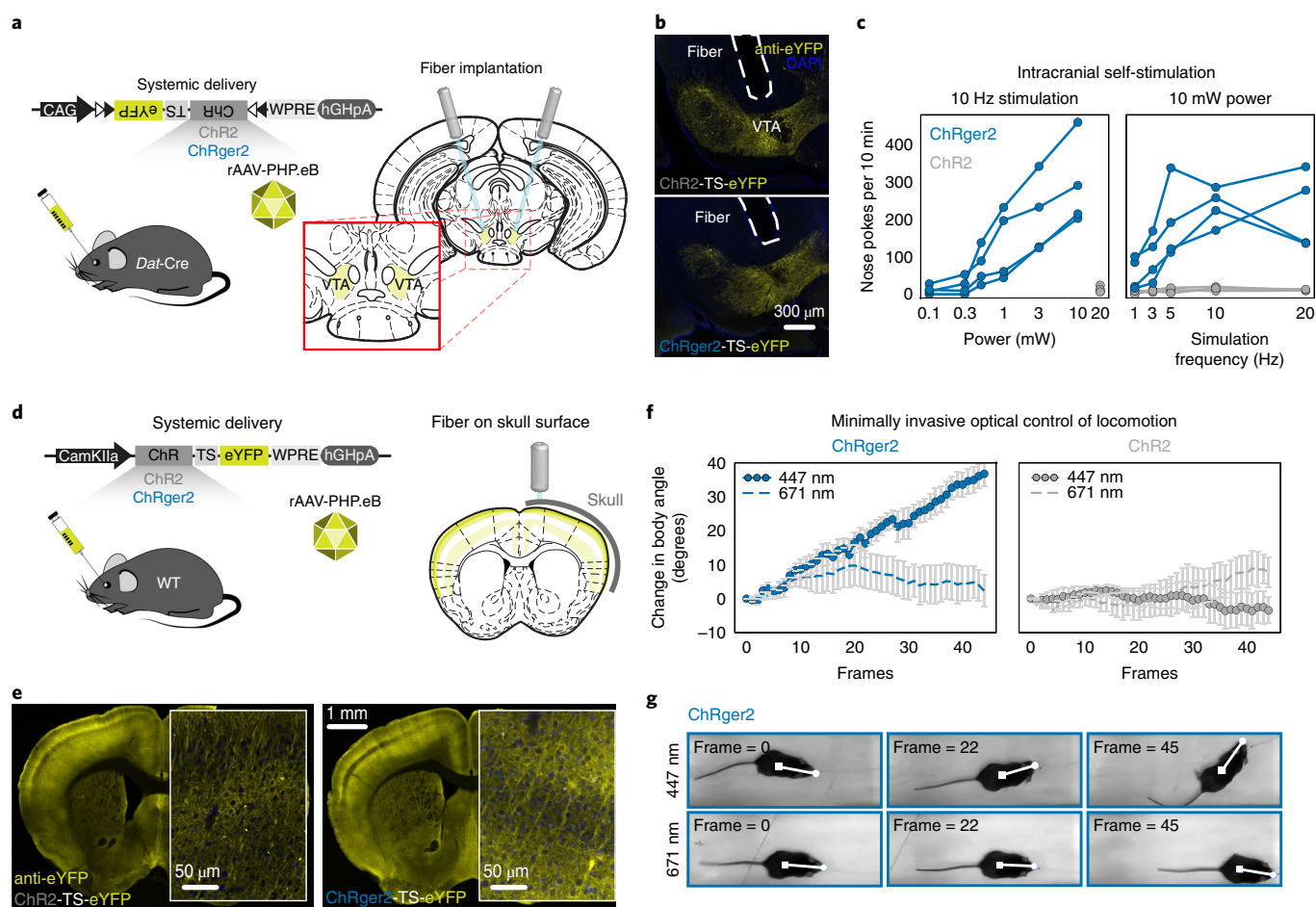
Fig. 4 | Validation of ChRger2 for minimally invasive optogenetic behavioral modulation. a, Systemic delivery of rAAV-PHP.eB-packaged CAG-DIO ChRger2-TS-eYFP or ChR2(H134R)-TS-eYFP (3×10[11] VG per mouse) into *Dat*-Cre animals coupled with fiber-optic implantation above the VTA. **b**, Images show fiber placement and opsin expression for ChR2(H134R) (top) and ChRger2 (bottom). **c**, Blue-light-induced intracranial self-stimulation (ten 5 ms laser pulses) with ChRger2 and ChR2(H134R) at varying light power and varying stimulation frequencies. ChRger2, *n* = 4 animals; ChR2(H134R), *n* = 4 animals. **d**, Minimally invasive, systemic delivery of rAAV-PHP.eB-packaged CaMKIIa ChRger2-TS-eYFP or ChR2(H134R)-TS-eYFP (5×10[11] VG per mouse) into wild-type (WT) animals (left) coupled with surgically secured fiber-optic cannula guide to the surface of the skull above the right M2 that had been thinned to create a level surface for the fiber–skull interface (right). **e**, Expression of ChR2(H134R) (left) and ChRger2 (right) in coronal slices throughout the cortex, including higher-magnification insets of M2. **f,g**, Turning behavior during unilateral light stimulation of M2 at the indicated wavelengths with ChRger2 and ChR2(H134R). ChRger2, *n* = 5 animals; ChR2(H134R), *n* = 5 animals. Plotted data are mean ± s.e.m.

We systemically delivered rAAV-PHP.eB packaging ChRger1–3 under the CaMKIIa promoter. With systemic delivery of ChRger2, we observed photocurrent strength similar to results observed after direct injection into the PFC in acute brain slices (Fig. 3g). When expressed in pyramidal neurons in the cortex, ChRger2 and ChRger3 enabled robust optically induced firing at rates between 2 and 10 Hz, although spike fidelity was reduced at higher frequency stimulation (Fig. 3m,n). ChRger2 performed best with higher frequency stimulation while ChRger1 performed worst. CoChR has better spike fidelity than the ChRgers at higher frequency stimulation (20–40 Hz) (Fig. 3m).

We next evaluated the optogenetic efficiency of ChRger2 using optogenetic intracranial self-stimulation of dopaminergic neurons of the ventral tegmental area (VTA)[32]. We systemically delivered rAAV-PHP.eB packaging a double-floxed inverted open reading frame (DIO) containing either ChRger2 or ChR2(H134R) into *Dat*-Cre mice (Fig. 4a,b and Supplementary Table 5). At 3 weeks after systemic delivery and stereotaxic implantation of fiber-optic cannulas above the VTA, we placed mice in an operant box and conditioned them to trigger a burst of 447 nm laser stimulation via nose poke. Animals expressing ChRger2 displayed robust

optogenetic self-stimulation in a frequency-dependent and laser power-dependent manner. Higher frequencies (up to 20 Hz) and higher light power (up to 10 mW) promoted greater maximum operant response rates (Fig. 4c). Conversely, laser stimulation failed to reinforce operant responding in ChR2(H134R)-expressing animals (Fig. 4c); these results were consistent with results in acute slice where the light-induced currents of ChR2(H134R) are too weak at the low copy number produced by systemic delivery for robust neuronal activation.

To determine whether ChRger2 would enable both minimally invasive transgene delivery and minimally invasive optical excitation, we assayed directional control of locomotion in freely moving animals by optogenetic stimulation of the right secondary motor cortex (M2)[33]. In this assay, unilateral stimulation of M2 disrupts motor function in the contralateral lower extremities, causing mice to turn away from the stimulation side. We systemically administered rAAV-PHP.eB packaging either ChRger2 or ChR2(H134R) under a CaMKIIa promoter for transgene expression in excitatory pyramidal neurons in the cortex (Fig. 4d and Supplementary Table 5). We observed broad expression throughout the cortex for both ChRger2- and ChR2(H134R)-injected animals

(Fig. 4e and Supplementary Fig. 11). We secured a fiber-optic cannula guide to the surface of the thinned skull above M2 without puncturing the dura and therefore leaving the brain intact (Fig. 4e), which we consider to be minimally invasive. Despite the presence of the optically scattering calavarial bone, stimulation with 20 mW 447 nm light-induced left-turning behavior in animals expressing ChRger2 but not in animals expressing ChR2(H134R) (Fig. 4f,g and Supplementary Videos 1 and 2). Left-turning behavior terminated on conclusion of optical stimulation (Supplementary Video 1). Behavioral effects were seen at powers as low as 10 mW. To ensure that the turning behavior was not due to visual stimuli or heating caused by the stimulation laser, we repeated treadmill experiments using 671 nm light, which is outside the excitation spectrum of both opsins. Furthermore, 20 mW 671 nm light failed to induce turning in both ChRger2 and ChR2(H124R). Overall, these experiments demonstrate that ChRger2 is compatible with systemic gene delivery and can enable minimally invasive optogenetic excitation.

Discussion

We demonstrated a data-driven approach to engineering ChR properties that enables efficient discovery of highly functional ChR variants based on data from relatively few variants. In this approach we approximate the ChR fitness landscape for a set of ~120,000 chimeric ChRs and use it to efficiently search sequence space and select top-performing variants for a given property[10,24,34]. By first eliminating the vast majority of nonfunctional sequences, we can focus on local peaks scattered throughout the landscape. Then, using regression models, we predict which sequences lie on the fitness peaks.

Machine learning provides a platform for simultaneous optimization of multiple ChR properties that follow engineering specifications. Application of this machine-learning pipeline (limited data collection from diverse sequences, model training and validation, and prediction and testing of new sequences) has great potential to optimize other neuroscience tools; for example, anion-conducting ChRs[12], calcium sensors, voltage sensors[35] and AAVs[30].

We designed high-performance ChRs (ChRger1–3) with unprecedented light sensitivity and validated ChRger2's application for in vivo optogenetics. The high-photocurrent properties of these ChRs overcome the limitation of low per-cell copy number after systemic delivery. Systemically delivered ChRgers have great potential for optogenetic activation of neuronal populations that are difficult to access surgically (for example, DRG, nodose ganglia, sympathetic chain ganglia and cardiac ganglia) or are widely distributed (for example, the enteric nervous system). Coupling ChRgers with recently reported upconversion nanoparticles may allow for noninvasive optogenetics in deep brain areas with systemic transgene delivery and tissue-penetrating near-infrared light for neuronal excitation[29]. ChRger2 enabled neuronal excitation with high temporal precision without invasive intracranial surgery for virus delivery or fiber-optic implantation for superficial brain areas, extending what is currently possible for optogenetics experiments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0583-8.

References

1. Deisseroth, K. & Hegemann, P. The form and function of channelrhodopsin. *Science* **357**, eaan5544 (2017).
2. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in neural systems. *Neuron* **71**, 9–34 (2011).
3. Lin, J. Y. A user's guide to channelrhodopsin variants: features, limitations and future developments. *Exp. Physiol.* **96**, 19–25 (2011).
4. Zhang, F. et al. Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. *Nat. Protoc.* **5**, 439–456 (2010).
5. Gradinaru, V. et al. Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141**, 154–165 (2010).
6. Mattis, J. et al. Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nat. Methods* **9**, 159–172 (2011).
7. Chuong, A. S. et al. Noninvasive optical inhibition with a red-shifted microbial rhodopsin. *Nat. Neurosci.* **17**, 1123–1129 (2014).
8. Bedbrook, C. N., Yang, K. K., Rice, A. J., Gradinaru, V. & Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* **13**, e1005786 (2017).
9. Bedbrook, C. N. et al. Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proc. Natl Acad. Sci. USA* **114**, E2624–E2633 (2017).
10. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
11. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
12. Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. Natural light-gated anion channels: a family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647–650 (2015).
13. Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).
14. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12**, 229–234 (2009).
15. Lin, J. Y., Lin, M. Z., Steinbach, P. & Tsien, R. Y. Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophysical J.* **96**, 1803–1814 (2009).
16. Berndt, A. et al. Structural foundations of optogenetics: determinants of channelrhodopsin ion selectivity. *Proc. Natl Acad. Sci. USA* **113**, 822–829 (2016).
17. Kato, H. E. et al. Crystal structure of the channelrhodopsin light-gated cation channel. *Nature* **482**, 369–374 (2012).
18. Wietek, J. et al. Conversion of channelrhodopsin into a light-gated chloride channel. *Science* **344**, 409–412 (2014).
19. Chan, K. Y. et al. Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).
20. Smith, M. A., Romero, P. A., Wu, T., Brustad, E. M. & Arnold, F. H. Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Sci.* **22**, 231–238 (2013).
21. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002).
22. Hochbaum, D. R. et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nat. Methods* **11**, 825–833 (2014).
23. Gunaydin, L. A. et al. Ultrafast optogenetic control. *Nat. Neurosci.* **13**, 387–392 (2010).
24. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl Acad. Sci. USA* **110**, E193–E201 (2013).
25. Volkov, O. et al. Structural insights into ion conduction by channelrhodopsin 2. *Science* **358**, eaan8862 (2017).
26. Oda, K. et al. Crystal structure of the red light-activated channelrhodopsin Chrimson. *Nat. Commun.* **9**, 3949 (2018).
27. Bamann, C., Gueta, R., Kleinlogel, S., Nagel, G. & Bamberg, E. Structural guidance of the photocycle of channelrhodopsin-2 by an interhelical hydrogen bond. *Biochemistry* **49**, 267–278 (2010).
28. Nagel, G. et al. Light activation of channelrhodopsin-2 in excitable cells of *Caenorhabditis elegans* triggers rapid behavioral responses. *Curr. Biol.* **15**, 2279–2284 (2005).
29. Chen, S. et al. Near-infrared deep brain stimulation via upconversion nanoparticle-mediated optogenetics. *Science* **359**, 679–684 (2018).
30. Bedbrook, C. N., Deverman, B. E. & Gradinaru, V. Viral strategies for targeting the central and peripheral nervous systems. *Annu. Rev. Neurosci.* **41**, 323–348 (2018).
31. Challis, R. C. et al. Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nat. Protoc.* **14**, 379–414 (2019).
32. Pascoli, V., Terrier, J., Hiver, A. & Luscher, C. Sufficiency of mesolimbic dopamine neuron stimulation for the progression to addiction. *Neuron* **88**, 1054–1066 (2015).
33. Gradinaru, V. et al. Targeting and readout strategies for fast optical neural control in vitro and in vivo. *J. Neurosci.* **27**, 14231–14238 (2007).

34. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
35. Flytzanis, N. C. et al. Archaerhodopsin variants with enhanced voltage-sensitive fluorescence in mammalian and *Caenorhabditis elegans* neurons. *Nat. Commun.* **5**, 4894 (2014).

Author contributions

C.N.B., K.K.Y., V.G. and F.H.A. conceptualized the project. C.N.B. coordinated all experiments and data analysis. C.N.B. and K.K.Y. built machine-learning models. C.N.B. performed construct design and cloning. C.N.B. and E.D.M. performed AAV production. E.D.M. prepared cultured neurons. C.N.B. and J.E.R. conducted electrophysiology. C.N.B. and J.E.R. performed injections. J.E.R. performed fiber cannula implants and behavioral experiments. C.N.B. performed all data analysis. C.N.B. wrote the manuscript with input and editing from all authors. V.G. supervised optogenetics/electrophysiology, and F.H.A. supervised the protein engineering.

Competing interests

A provisional patent application (CIT File No.: CIT-8092-P) has been filed by Caltech based on these results. C.N.B., K.K.Y., V.G. and F.H.A. are inventors on this provisional patent.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0583-8.

Correspondence and requests for materials should be addressed to V.G. or F.H.A.

Peer review information Nina Vogt was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Construct design and cloning. The design, construction and characterization of the recombination library of chimeras are described in detail by Bedbrook et al.[9]. The ten-block contiguous and ten-block noncontiguous recombination libraries were designed and built using SCHEMA recombination[9]. Software packages for calculating SCHEMA energies are openly available at cheme.che.caltech.edu/groups/fha/Software.htm. Each chimeric ChR variant in these libraries is composed of blocks of sequence from the parental ChR (CsChrimR (ref. [11]), C1C2 (ref. [17]) and CheRiff (ref. [22])), including chimeras with single-block swaps (chimeras consisting of nine blocks of one parent and a single block from one of the other two parents) and multi-block-swap chimera sequences.

Selected ChR variant genes were inserted into a constant vector backbone (pFCK from Addgene plasmid no. 51693 (ref. [22])) with a Golgi export trafficking signal sequence (KSRITSEGEYIPLDQIDINV)[5] and fluorescent protein (mKate). All ChR variants contain the SpyTag sequence following the N-terminal signal peptide for the SpyTag/SpyCatcher labeling assays used to characterize ChR membrane localization[9,36]. The C1C2 parent for the recombination libraries is mammalian codon-optimized. ChR variant sequences used in this study are documented in Supplementary Data 2. All selected ChR genes were synthesized and cloned in the pFCK mammalian expression vector by Twist Bioscience. For visualization, sequence alignment between C1C2 and designer ChRs were created using ClustalΩ and visualized using ENDscript[37] (Supplementary Fig. 3).

For characterization in neurons, selected ChR variants (ChRger1, ChRger2, ChRger3, CoChR[11] and hChR2(H134R)) were inserted into a pAAV-hSyn vector backbone (Addgene plasmid no. 26973), a pAAV-CamKIIa vector backbone (Addgene plasmid no. 51087) and a pAAV-CAG-DIO vector backbone (Addgene plasmid no. 104052). In all backbones, each ChR was inserted with a trafficking signal sequence[5] and fluorescent protein (eYFP).

HEK293T cell and primary neuronal cultures. The culturing and characterization of ChRs in HEK cells is described by Bedbrook et al.[9,36]. Briefly, HEK cells were cultured at 37 °C and 5% CO_2 in D10 (DMEM supplemented with 10% (vol/vol) fetal bovine serum, 1% sodium bicarbonate and 1% sodium pyruvate). HEK cells were transfected with purified ChR variant DNA using FuGENE 6 reagent according to the manufacturer's (Promega) recommendations. Cells were given 48 h to express the ChRs before photocurrent measurements. Primary hippocampal neuronal cultures were prepped from C57BL/6N mouse embryos 16–18 d post-fertilization (Charles River Laboratories) and cultured at 37 °C in the presence of 5% CO_2 in Neurobasal media supplemented with glutamine and B27. Cells were transduced 3–4 d after plating with rAAV-PHP.eB packaging ChR2(H134R), CoChR, ChRger1, ChRger2 or ChRger3. Whole-cell recordings were performed 5–10 d after transduction.

Patch-clamp electrophysiology. Whole-cell patch-clamp and cell-attached recordings were performed in transfected HEK cells, transduced cultured neurons and acute brain slices to measure light-activated inward currents or neuronal firing. For electrophysiological recordings, cultured cells were continuously perfused with extracellular solution at room temperature (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10 glucose; pH 7.35) while mounted on the microscope stage. For slice recordings, 32 °C artificial cerebrospinal fluid was continuously perfused over slices. Artificial cerebrospinal fluid contained 127 mM NaCl, 2.5 mM KCl, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 12 mM D-glucose, 0.4 mM sodium ascorbate, 2 mM $CaCl_2$ and 1 mM $MgCl_2$ and was bubbled continuously with 95% oxygen/5% CO_2. Firing and photocurrent measurements were performed in the presence of 3 mM kynurenic acid and 100 μM picrotoxin to block optically evoked ionotropic glutamatergic and GABAergic currents, respectively.

Patch pipettes were fabricated from borosilicate capillary glass tubing (1B150-4; World Precision Instruments) using a model P-2000 laser puller (Sutter Instruments) to resistances of 3–6 MΩ. Pipettes were filled with K-gluconate intracellular solution containing the following (in mM): 134 K-gluconate, 5 EGTA, 10 HEPES, 2 $MgCl_2$, 0.5 $CaCl_2$, 3 ATP and 0.2 GTP. Whole-cell patch-clamp and cell-attached recordings were made using a Multiclamp 700B amplifier (Molecular Devices), a Digidata 1440 digitizer (Molecular Devices) and a PC running pClamp (version 10.4) software (Molecular Devices) to generate current injection waveforms and to record voltage and current traces.

Photocurrents were recorded from cells in voltage clamp held at −60 mV. Neuronal firing was measured in current-clamp mode with current injection for a −60 mV holding potential. Access resistance (R_a) and membrane resistance (R_m) were monitored throughout recording, and cells were discarded if R_a or R_m changed more than 15%. During ChR variant functional screening in HEK cells, photocurrents were only recorded from cells that passed our recording criteria: $R_m > 200$ MΩ and holding current > −100 pA. Our measured membrane properties of ChR-expressing neurons were consistent with previous literature of opsin-expressing cells[11] and are also consistent with previous reports of properties of cultured hippocampal neurons[38,39] and PFC neurons in slice[40,41] (Supplementary Fig. 10). For cell culture experiments, the experimenter was blinded to the identity of the ChR being patched but not to the fluorescence level of the cells. For acute slice recordings, the experimenter was not blinded to the identity of the ChR.

Light delivery and imaging. Patch-clamp recordings were done with short light pulses to measure photocurrents. Light-pulse duration, wavelength and power were varied depending on the experiment (as described in the text). Light pulses were generated using a Lumencor SPECTRAX light engine. The illumination/output spectra for each color were measured (Supplementary Fig. 5). To evaluate normalized green photocurrent, we measured photocurrent strength at three wavelengths (peak ± half width at half maximum): (red) 640 ± 3 nm, (green) 546 ± 16 nm and (cyan) 481 ± 3 nm with a 0.5 s light pulse. Light intensity was matched for these measurements, with 481 nm light at 2.3 mW mm^{-2}, 546 nm light at 2.8 mW mm^{-2} and 640 nm light at 2.2 mW mm^{-2}. For full spectra measurements depicted in Fig. 2d, we measured photocurrents at seven different wavelengths (peak ± half width half maximum): (red) 640 ± 3 nm, (yellow) 567 ± 13 nm, (green) 546 ± 16 nm, (teal) 523 ± 6 nm, (cyan) 481 ± 3 nm light-emitting diode (LED), (blue) 439 ± 8 nm LED and (violet) 397 ± 3 nm with a 0.5 s light pulse for each color. Light intensity is matched across wavelengths at 1.3 mW mm^{-2}.

Imaging of ChR variant expression in HEK cells was performed using an Andor Neo 5.5 sCMOS camera and Micro-Manager Open Source Microscopy Software (v1.4). Imaging of ChR expression in neuronal cultures and in brain slices was performed using a Zeiss LSM 880 confocal microscope and Zen software (v2.3).

Electrophysiology data analysis. Electrophysiology data were analyzed using Clampfit 10.7 (Molecular Devices) and custom data-processing scripts written using open-source packages in the Python programming language to perform baseline adjustments, find the peak and steady-state inward currents, perform monoexponential fits of photocurrent decay for off-kinetic properties and quantify spike fidelity. Only neurons with an uncompensated series resistance between 5 and 25 MΩ, $R_m > 90$ MΩ and holding current > −150 pA (holding at −60 mV) were included in data analysis (Supplementary Fig. 10). The photocurrent amplitude was not adjusted for expression level since both expression and conductance contribute to the in vivo utility of the tool. However, comparisons of expression with photocurrent strength for all ChR variants tested are included in Supplementary Fig. 6 and 7.

On light exposure, ChRs open and reach a peak inward current and then desensitize reaching a lower steady-state current. As metrics of photocurrent strength, peak and steady-state photocurrent were used (Fig. 1a). As a metric for the ChR wavelength sensitivity of activation spectrum (that is, spectral properties), the normalized current strength induced by exposure to green light (546 nm) was used (Fig. 1a), which easily differentiates blue-shifted ChRs (peak activation: ~450–480 nm) and red-shifted ChRs (peak activation: ~520–650 nm). Two parameters were used to characterize ChR off-kinetics: the time to reach 50% of the light-activated current after a 0.5 s light pulse and the photocurrent decay rate (τ_{off}) after a 1 ms light exposure (Fig. 1a).

AAV production and purification. Production of recombinant AAV-PHP.eB packaging pAAV-hSyn-X-TS-eYFP-WPRE, pAAV-CAG-DIO[X-TS-eYFP]-WPRE and pAAV-CaMKIIa-X-TS-eYFP-WPRE (X = ChR2(H134R), CoChR, ChRger1, ChRger2 and ChRger3) was done following the methods described by Deverman et al.[42] and Challis et al.[31]. Briefly, triple transfection of HEK293T cells (ATCC) was performed using polyethylenimine. Viral particles were harvested from the media and cells. Virus was then purified over iodixanol (Optiprep, Sigma; D1556) step gradients (15%, 25%, 40% and 60%). Viruses were concentrated and formulated in PBS. Virus titers were determined by measuring the number of DNase I-resistant viral genomes using quantitative PCR with linearized genome plasmid as a standard.

Animals. All procedures were approved by the California Institute of Technology Institutional Animal Care and Use Committee. *Dat*-Cre mice (006660) and C57Bl/6J mice (000664) were purchased from the Jackson Laboratory.

Intravenous injections, stereotactic injections and cannula implantation. Intravenous administration of rAAV vectors was performed by injecting the virus into the retro-orbital sinus at viral titers indicated in the text. There were no observed health issues with animals after systemic injection of virus at the titers presented in the paper. Mice remained healthy >6 months after systemic delivery of ChR2 and ChRgers. With slice electrophysiology, there was no observed indication of poor cell health due to viral-mediated expression, which was quantified by measuring the membrane resistance (R_m), leak current (holding at −60 mV) and resting membrane potential (Supplementary Fig. 10). Local expression in the PFC was achieved by direct stereotactic injection of 1 μl of purified AAV vectors at 5×10^{12} viral genomes per ml targeting the following coordinates: anterior-posterior (AP), −1.7; media-lateral (ML), ±0.5; and dorsal-ventral (DV), −2.2. For stimulation of the VTA, 300 μm outer diameter mono fiber-optic cannulae (Doric Lenses, MFC_300/330-0.37_6mm_ZF1.25_FLT) were stereotaxically implanted 200 μm above the VTA bilaterally targeted to the following coordinates: AP, −3.44 mm; ML, ±0.48 mm; DV, 4.4 mm. For stimulation of the right M2, 3-mm-long, 400 μm mono fiber-optic cannulae (Doric Lenses, MFC_400/430-0.48_3mm_ZF1.25_FLT) were surgically secured to the surface of the skull above M2 (unilaterally) targeted to the following coordinates: AP, 1 mm; ML, 0.5 mm.

The skull was thinned ~40–50% with a standard drill to create a level surface for the fiber–skull interface. Light was delivered from either a 447 nm or 671 nm laser (Changchun New Industries Model with PSU-H-LED) via mono fiber-optic patch cable(s) (Doric Lenses, MFP_400/430/1100-0.48_2m_FC-ZF1.25) coupled to the fiber-optic cannula(e). Fiber-optic cannulae were secured to the skull with Metabond (Parkel, SKU S396) and dental cement.

Analysis of behavioral experiments was performed using the open-source MATLAB program OptiMouse[43] to track mouse nose, body and tail position while the mouse was running on the treadmill. Optogenetic intracranial self-stimulation was performed using a mouse modular test chamber (Lafayette Instruments, Model 80015NS) outfitted with an infrared nose port (Model 80116TM).

GP modeling. Both the GP regression and classification modeling methods applied in this paper are based on work detailed in refs.[8,24]. For modeling, all sequences were aligned using multiple sequence comparison by log-expectation (MUSCLE) (https://www.ebi.ac.uk/Tools/msa/muscle/). For modeling, aligned sequences were truncated to match the length of the C1C2 sequence, eliminating N- and C-terminal fragments with poor alignment quality due to high sequence diversity (Supplementary Data 1 and 2). Structural encodings (that is, the contact map) use the C1C2 crystal structure (3UG9.pdb) and assume that ChR chimeras share the contact architecture observed in the C1C2 crystal structure. Models built using structural encodings built from the ChR2 structure (6EID.pdb) and the C1Chrimson structure (5ZIH.pdb) performed as well as models using the C1C2 structure (Supplementary Fig. 1c,d). The models are robust to differences in contact maps because they use both sequence and structural information, which is somewhat redundant.

For a given ChR, the contact map is simply a list of contacting amino acids with their positions. For example, a contact between alanine at position 134 and methionine at position 1 of the amino acid sequence would be encoded by (('A134'), ('M1')). Both sequence and structural information were one-hot encoded. Regression models for ChR properties were trained to predict the logarithm of the measured properties. All training data were normalized to have mean of zero and standard deviation of 1.

GP regression and classification models require kernel functions that measure the similarity between protein sequences. Learning involves optimizing the form of the kernel and its hyperparameters (Supplementary Table 2). The Matérn kernel was found to be optimal for all ChR properties (Supplementary Table 1).

For classification model training, all 102 functionally characterized ChR variants from our recombination libraries (Supplementary Data 2) were used as well as data from 61 sequence variants published by others (Supplementary Data 1). The model was then updated with data collected from the 22 additional ChR recombination variants with high sequence diversity (~70 mutations from the closest parent) and predicted to be functional (Fig. 1d). For training the regression models, all 102 functionally characterized training sequences (Supplementary Data 2) were initially used and then the models were updated with data collected from the 22 additional ChR variants (Fig. 1d).

GP regression. In regression, the goal is to infer the value of an unknown function $f(x)$ at a novel point $\mathbf{x}_\ddagger$ given observations $\mathbf{y}$ at inputs X. Assuming that the observations are subject to independent and identically distributed Gaussian noise with variance σ_n^2, the posterior distribution of $f_\ddagger = f(\mathbf{x}_\ddagger)$ for GP regression is Gaussian with mean

$$\bar{f}_\ddagger = \mathbf{k}_\ddagger^{\mathrm{T}}\left(K + \sigma_n^2 I\right)^{-1}\mathbf{y} \tag{1}$$

and variance

$$v_\ddagger = k\left(\mathbf{x}_\ddagger, \mathbf{x}_\ddagger\right) - \mathbf{k}_\ddagger^{\mathrm{T}}\left(K + \sigma_n^2 I\right)^{-1}\mathbf{k}_\ddagger \tag{2}$$

where I is the identity matrix. K is the symmetric, square covariance matrix for the training set: $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for $\mathbf{x}_i$ and $\mathbf{x}_j$ in the training set. $\mathbf{k}_\ddagger$ is the vector of covariances between the novel input and each input in the training set, and $\mathbf{k}_{\ddagger i} = k(\mathbf{x}_\ddagger, \mathbf{x}_i)$. The hyperparameters in the kernel functions and the noise hyperparameter σ_n were determined by maximizing the log marginal likelihood:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\left(K + \sigma_n^2 I\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|K + \sigma_n^2 I\right| - \frac{n}{2}\log 2\pi \tag{3}$$

where n is the dimensionality of the inputs. Regression was implemented using open-source packages in the SciPy ecosystem[44–46].

GP classification. In binary classification, instead of continuous outputs $\mathbf{y}$, the outputs are class labels $y_i \in \{+1, -1\}$, and the goal is to use the training data to make probabilistic predictions $\pi(\mathbf{x}_\ddagger) = p(y_\ddagger = +1|\mathbf{x}_\ddagger)$. We use Laplace's method to approximate the posterior distribution. Hyperparameters in the kernels are found by maximizing the marginal likelihood. Classification was implemented using open-source packages in the SciPy ecosystem[44–46]. The binary classification model was trained to predict whether a ChR sequence is or is not functional. A ChR sequence was considered to be functional if its photocurrents were >100 pA

on light exposure, a threshold set as an approximate lower bound for current necessary for neuronal activation.

GP kernels for modeling proteins. GP regression and classification models require kernel functions that measure the similarity between protein sequences. A protein sequence s of length L is defined by the amino acid present at each location. This can be encoded as a binary feature vector $\mathbf{x}_{se}$ that indicates the presence or absence of each amino acid at each position resulting in a vector of length 20L (for 20 possible amino acids). Likewise, the protein's structure can be represented as a residue–residue contact map. The contact map can be encoded as a binary feature vector $\mathbf{x}_{st}$ that indicates the presence or absence of each possible contacting pair. Both the sequence and structure feature vectors were used by concatenating them to form a sequence–structure feature vector.

Three types of kernel functions $k(s_i, s_j)$ were considered: polynomial kernels, squared exponential kernels and Matérn kernels. These different forms represent possible functions for the protein's fitness landscape. The polynomial kernel with degree d is defined as:

$$k(s, s') = \left(\sigma_0^2 + \sigma_p^2 \mathbf{x}^{\mathrm{T}}\mathbf{x}'\right)^d \tag{4}$$

where s is a protein sequence and $\mathbf{x}$ its corresponding encoding and σ_0 and σ_p are hyperparameters. We considered polynomial kernels with $d = 3$. The squared exponential kernel is defined as:

$$k(s, s') = \sigma_p^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{l}\right) \tag{5}$$

where l and σ_p are also hyperparameters and $|\cdot|_2$ is the L2 norm. Finally, the Matérn kernel with $v = \frac{5}{2}$ is defined as:

$$k(s, s') = \left(1 + \frac{\sqrt{5\|\mathbf{x} - \mathbf{x}'\|_2^2}}{l} + \frac{5\|\mathbf{x} - \mathbf{x}'\|_2^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5\|\mathbf{x} - \mathbf{x}'\|_2^2}}{l}\right) \tag{6}$$

where l is once again a hyperparameter.

L1 regression feature identification and weighting. L1 regression was used to identify residues and contacts in the ChR structure most important for each ChR functional property of interest. First, residues and contacts that covary were identified using the concatenated sequence and structure binary feature vector for each of the training set ChR variants. Each set of covarying residues and contacts was combined into a single feature. L1 linear regression was used to select the features that contribute most to each ChR functional property of interest. The level of regularization was chosen by maximizing the log marginal likelihood of the GP regression model trained on the features selected at that level of regularization. We then performed Bayesian ridge regression on the selected features using the default settings in scikit-learn[47]. Residues and contacts with the largest absolute Bayesian ridge linear regression weights were plotted onto the C1C2 structure (Supplementary Fig. 4). For feature identification and weighting, models were trained on both the training set and also the test set of 28 ChR sequences predicted to have useful combinations of diverse properties.

Statistical analysis. Plotting and statistical analysis were done in Python 2.7 and 3.6 and GraphPad Prism 7.01. For statistical comparisons, we first performed a D'Agostino and Pearson normality test. If the P value of a D'Agostino and Pearson normality test was <0.05, the nonparametric Kruskal–Wallis test with Dunn's multiple comparisons post hoc test was used. If the data passed the normality test, a one-way analysis of variance was used.

Accession codes. GenBank: ChRger1, MN340983; ChRger2, MN340984; ChRger3, MN340985.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The authors declare that data supporting the findings of this study are available within the paper and its Supplementary information files. Source data for classification model training are provided in Supplementary Data 1 and 2. Source data for regression model training are provided in Supplementary Data 2. DNA constructs for the ChRger variants are deposited for distribution at Addgene (http://www.addgene.org, plasmid numbers 127237-44).

Code availability
Code used to train classification and regression models can be found at https://github.com/fhalab/channels.

References

36. Bedbrook, C. N. et al. Genetically encoded spy peptide fusion system to detect plasma membrane-localized proteins in vivo. *Chem. Biol.* **22**, 1108–1121 (2015).

37. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

38. Fan, J. et al. Reduced hyperpolarization-activated current contributes to enhanced intrinsic excitability in cultured hippocampal neurons from PrP$^{-/-}$ mice. *Front. Cell. Neurosci.* **10**, 74 (2016).

39. Slomowitz, E. et al. Interplay between population firing stability and single neuron dynamics in hippocampal networks. *eLife* **4**, e04378 (2015).

40. Kroon, T., van Hugte, E., van Linge, L., Mansvelder, H. D. & Meredith, R. M. Early postnatal development of pyramidal neurons across layers of the mouse medial prefrontal cortex. *Sci. Rep.* **9**, 5037 (2019).

41. van Aerde, K. I. & Feldmeyer, D. Morphological and physiological characterization of pyramidal neuron subtypes in rat medial prefrontal cortex. *Cereb. Cortex* **25**, 788–805 (2015).

42. Deverman, B. E. et al. Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34**, 204–209 (2016).

43. Ben-Shaul, Y. OptiMouse: a comprehensive open source program for reliable detection and analysis of mouse body and nose positions. *BMC Biol.* **15**, 41 (2017).

44. Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

45. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

46. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).

47. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

nature research

Corresponding author(s):   Frances H. Arnold
Viviana Gradinaru

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Software to acquire confocal images with the Zeiss LSM 880 was Zen (version 2.3 from Zeiss). Software to acquire epifluorescence HEK cell images with the Andor Neo 5.5 sCOMS camera was Micro-Manager Open Source Microscopy Software version 1.4. Software for electrophysiology was pClamp (version 10.4 from Molecular Devices).
Data analysis	Gaussian Process regression and classification was implemented by custom code with open-source packages in the SciPy ecosystem (0.19.1) in the Python programming language. Electrophysiology data were analyzed using custom data-processing scripts written using open-source packages in the Python programming language to do baseline adjustments, find the peak and steady state inward currents, perform monoexponential fits of photocurrent decay for off-kinetic properties, and quantify spike fidelity. Animal behavior was analyzed using open-source OptiMouse software (version 3.0, available at: https://github.com/yorambenshaul/optimouse) using Matlab. Plotting and statistical analysis were done in Python 2.7 and 3.6 and GraphPad Prism 7.01.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

> The authors declare that data supporting the findings of this study are available within the paper and its supplementary information files. Source data for classification model training are provided in Dataset 1 and Dataset 2. Source data for regression model training are provided in Dataset 2.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](#)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. We followed the standard sample sizes used for ChR comparisons in culture and in vivo.
Data exclusions	No data were excluded.
Replication	Each ChR variant was characterized by measurements made from many HEK cells. The ChRgers were characterized in multiple vector backbones. The ChRgers were characterized in HEK cells, neurons in culture, neurons slice, and in animals. In culture, multiple cells were recorded per transfection/transduction and multiple replicates of transfection or transduction were performed. In slice experiments, multiple animals were used and multiple cells measured per slice. In in vivo experiments, multiple animals were used. All attempts at replication were successful.
Randomization	Mice were randomly assigned to groups before injection and fiber placement.
Blinding	For all experiments in HEK cells authors were blind to the identity of the ChR being tested as all ChRs were given code names and the sequence identity of each ChR was only checked after data analysis. No blinding was performed for behavioral experiments because no human subjective qualitative metrics were assessed or reported.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	Unique biological materials
☐	Antibodies
☐	Eukaryotic cell lines
☒	Palaeontology
☐	Animals and other organisms
☒	Human research participants

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging

Antibodies

Antibodies used	Green florescent protein (GFP) anitbody from aves labs inc. catalog # GFP-1020. lot # GFP879484. Antibody staining was done at a 1:1000 dilution. http://www.aveslab.com/products/epitope-tag-and-gfp-antibodies/anti-gfp-green-fluorescent-protein-antibodies-2/
Validation	Quality control of the GFP antibody was done by aves labs: antibodies were analyzed by western blot analysis (1:5000 dilution) and immunohistochemistry (1:500 dilution) using transgenic mice expressing the GFP gene product. Western blots were

performed using BlokHen® (Aves Labs) as the blocking reagent, and HRP-labeled goat anti-chicken antibodies (Aves Labs, Cat. #H-1004) as the detection reagent. Immunohistochemistry used tetramethyl rhodamine-labeled anti-chicken IgY.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	human embryonic kidney cells containing the SV40 T-antigen: HEK293T cells ATCC® CRL-3216
Authentication	Cell line used was not authenticated by us
Mycoplasma contamination	Cell line was not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	This study did not use misidentified cell lines

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Dat-Cre C57Bl/6J mice (006660) and wild-type C57Bl/6J mice (000664) were purchased from Jackson Laboratory. Mice were injected at ages between 2-4 months and behavior was performed in animals between 3-5 months. Both males and females were used for experiments.
Wild animals	The study did not involve wild animals
Field-collected samples	The study did not involve field-collected samples