Title

Generative AI-based design of hybrid transcriptional activator proteins with new DNA-binding specificity

Authors

Sota L Okuda[1,†], Atsushi Minami[2,†], Michio Aiko[3,†], Kohei Uetsuka[1], Kazuteru Miyazaki[4], Kazumasa Ohtake[1,2], and Daisuke Kiga[1,2,3*]

Affiliations

1 Department of Electrical Engineering and Bioscience, Waseda University, Shinjuku-ku, Tokyo 169-8050, Japan.

2 Research Institute for Science and Engineering, Waseda University, Shinjuku-ku, Tokyo 169-8050, Japan.

3 The Smart Life Science Institute, Advanced Collaborative Research Organization for Smart Society, Waseda University, Shinjuku-ku, Tokyo 169-8050, Japan.

4 Research Department, National Institution for Academic Degrees and Quality Enhancement of Higher Education, Kodaira-shi, Tokyo 187-8587, Japan

† These authors contributed equally to this work.

* Correspondence to:

Daisuke Kiga (kiga@waseda.jp).

Abstract

Transcriptional control arises from the specific recognition of promoter DNA by transcription factors (TFs), forming the basis of cellular information processing and gene regulation. In synthetic biology, TF-promoter interactions are assembled into genetic circuits to program cellular behaviors. To ensure expected circuit performance, most synthetic genetic circuits rely on combinations of well-characterized and orthogonal transcriptional regulations. This reliance minimizes crosstalk but constrains circuit complexity and information integration. Creating hybrid TFs that combine or interpolate promoter specificities could therefore expand the design space of synthetic regulatory systems. However, it remains unclear whether hybrid functions can be created by mixing amino acid sequences, and how such functional integration could be achieved in a principled manner. Here we show that a variational autoencoder (VAE) trained on LuxR-family DNA-binding domains can generate transcription factors with hybrid and partially novel promoter recognition properties. By sampling intermediate regions of the VAE-learned latent space, we designed hybrid TFs that activate both the *lux* and *las* promoters. High-throughput sort-seq assays together with individual *in vivo* assays revealed that a subset of functional variants exhibited dual-responsive behavior while maintaining sequence-selective DNA recognition. Together, these results provide a data-driven strategy for exploring functional intermediate sequences between closely related proteins.

Introduction

Transcriptional control, mediated by the precise recognition of promoter DNA by transcription factors (TFs), is fundamental to information processing in natural cells and to the operation of synthetic genetic circuits. In synthetic biology, which aims to engineer biological systems by assembling standardized genetic parts into programmable networks[1], TF-promoter interactions serve as the fundamental building blocks of synthetic genetic circuits[2,3]. Within such circuits, transcriptional regulation plays a central role, with TFs acting as molecular switches that convert environmental or intracellular cues into gene expression outputs. To simplify design and minimize crosstalk, well-characterized and orthogonal genetic parts have been predominantly used for constructing synthetic genetic circuits[4–10]. While this strategy supports reliable circuit behavior, it also constrains circuit complexity and information density. To construct more sophisticated biological computation, increasingly complex information-processing architectures have been developed across molecular and cellular scales.

Recently, neural network-like pattern recognition and decision-making have been demonstrated using multi-input and multi-output genetic circuits *in vitro*[11–13] and in living cells[14–21]. Yet most of these systems still depend on assembling multiple components each of which has its major target, rather than using regulatory elements that combine multiple input/output specificities within a single element. Therefore, enabling compact genetic circuits with multi-input and multi-output transcriptional logic requires transcription factors and promoters that interact across multiple regulatory axes. One promising strategy to achieve this is the design of dual-active hybrid TFs that combine recognition features of multiple parental regulators (**Fig. 1a**).

Given that protein function is ultimately dictated by its amino acid sequence,

68 including DNA recognition by TFs, could hybrid sequences give rise to hybrid functions?
69 Several approaches have been explored to address this question, including recombination-
70 based methods such as domain swapping[22] and DNA shuffling[23]. Ancestral sequence
71 reconstruction (ASR)[24] offers a conceptually distinct route, inferring intermediate
72 sequences by tracing evolutionary trajectories within a protein family. These methods
73 have enabled the creation of chimeric proteins or functional intermediates between
74 homologous proteins. However, these approaches do not account for the residue-residue
75 interactions and sequence context that collectively determine folding, stability, and DNA-
76 recognition specificity in TFs.

77 Recent advances in deep generative modeling offer a fundamentally different
78 way to mix sequence information. Variational autoencoders (VAEs) are well known from
79 image generation for their ability to interpolate smoothly between examples, such as
80 producing intermediate handwritten digits by sampling along continuous trajectories in
81 latent space[25]. VAEs have recently been applied to protein sequence design, where VAEs
82 trained on natural sequence families can learn evolutionary constraints and generate
83 functional variants of enzymes, signal peptides, and nanobodies without requiring
84 structural information[26–33]. Notably, however, these applications have focused on
85 producing alternative variants within a single functional class; whether latent-space
86 interpolation can generate meaningful "intermediate" sequences that combine functional
87 features of two homologous transcription factors remains to be elucidated.

88 In this study, we employed a VAE to generate hybrid transcription factor
89 sequences derived from the parental TFs; LuxR and LasR. By constructing a diverse
90 library of VAE-generated proteins and assessing their transcriptional activation abilities,
91 we identified several proteins exhibiting dual promoter-recognition capabilities.

92 Furthermore, by combining a randomized promoter library with deep sequencing (sort-
93 seq), we systematically profiled the promoter recognition specificities of each variant.
94 Together with structural analyses based on AlphaFold3 predictions and molecular
95 dynamics simulations, we reveal the molecular basis underlying the DNA-binding
96 properties of these hybrid proteins. Our results provide valuable insights into the rational
97 design of proteins with hybrid functions.
98

Results

Design of hybrid transcription factors using a variational autoencoder (VAE)

We first sought to design hybrid transcription factors by applying a variational autoencoder (VAE) to the DNA-binding domains (DBDs) of LuxR-family regulators. LuxR-type transcription factors consist of an N-terminal ligand-binding receiver domain and a C-terminal helix-turn-helix DBD that mediates promoter recognition[34]. Since the DBD alone is sufficient to bind target DNA and can activate transcription independently of ligand binding[35,36], we focused our design efforts exclusively on this domain. This allowed us to directly probe how sequence-level changes within the DBD alter promoter specificity.

To generate hybrid sequences, we constructed a curated multiple sequence alignment (MSA) of LuxR-family DBDs[37] and trained an MSA-VAE on this dataset[27,29,33]. The VAE comprises an encoder that compresses each aligned sequence into a low-dimensional latent representation and a decoder that reconstructs sequences from these representations. During training, the model simultaneously optimizes a reconstruction objective and a regularization term that encourages the latent variables to follow a multivariate Gaussian distribution, resulting in a structured latent space that captures biophysical constraints in the input sequences (**Fig. 1b**). When encoded by the trained model, the parental LuxR and LasR sequences occupied distinct regions of the latent space (**Fig. 1c**). Other representative members of the LuxR family, including RhlR and CinR, were also mapped to distinct positions, with clusters of related natural sequences forming around each representative point (**Fig. 1c**). This segmentation indicates that the diversity present in the training dataset is faithfully preserved in the latent space and that the VAE captures family-level sequence relationships rather than collapsing them into a

single region. The clustering among these subfamilies also enabled the definition of an interpolation point between the LuxR and LasR latent vectors, providing a principled way to explore sequence variants that combine features of both regulators.

Using this latent-space representation, we sampled sequences from a hyperspherical region centered at the midpoint between the LuxR and LasR latent vectors (0.5-0.5 interpolation). This strategy enabled targeted sampling of hybrid sequences, in contrast to prior VAE studies that sampled amino acid sequences closely related or neighboring locations in the latent space for collecting similar variants[27,29,32,33]. Sampling across the interpolation region produced a diverse set of candidate sequences that contain novel combinations of parental residues while retaining global amino-acid usage patterns of the LuxR family. From each of the top-performing VAE models, 20,000 sequences were decoded from the local interpolation region. After removing redundant and gap-containing sequences using CD-HIT, we obtained a pool of unique candidate hybrids (**Fig. 1d**).

Fig. 1 | Schematic of the VAE-based sequence generation strategy for hybrid TFs

a, Design and dual-sequence recognition of the hybrid transcription factor. Parental natural transcription factors (natural TFs: blue and orange) specifically bind to their respective target DNA sequences. This study aims to design the hybrid TF (colored in rainbow), formed by mixing of natural TFs, that exhibits dual recognition capability, enabling it to bind to both sequence and activate downstream gene expression for both targets. **b,** Schematic overview of the MSA-VAE architecture used for hybrid sequence design. Aligned LuxR-family DNA-binding domains (DBDs) are encoded into a low-dimensional latent representation and decoded back into sequence space. **c,** t-SNE projection of representative LuxR-family transcription factors. The positions of LuxR and LasR are indicated by a pink square and a green square, respectively. Additional representative family members are plotted using X-shaped markers. Natural sequences are shown as individual points. **d,** Sequence sampling from the LuxR-LasR interpolation region. Hybrid sequences were generated from a hyperspherical region centered at the midpoint between the LuxR and LasR latent vectors (0.5-0.5 interpolation), decoded, and filtered for redundancy using CD-HIT.

158 ***Functional characterization of VAE-designed TFs toward lux and las promoters***

159 To determine whether the VAE-designed hybrid sequences exhibit functional
160 activity *in vivo*, we selected 9 candidates at random from the generated sequences and
161 evaluated their ability to activate transcription from the native *lux* and *las* promoters in
162 *Escherichia coli*. For each designed variant, we constructed two coding sequences that
163 differed only in their immediate N-terminal context: one beginning with Met ("M+CDS"),
164 and the other beginning Met followed by Ala ("MA+CDS") to account for potential N-
165 terminal context effects[38,39], resulting in a total of 18 constructs. The mixed
166 oligonucleotide pool was cloned and expressed in *E. coli*, and its ability to activate
167 transcription from the native *lux* and *las* promoters was quantified using GFP reporter
168 assays (**Fig. 2a**). Some of individual fluorescent colonies were isolated, and the
169 corresponding coding sequences were identified by sequencing prior to quantitative
170 analysis. Variants that were not recovered for quantitative analysis likely reflect a
171 combination of factors, including severe growth defects, unstable or insufficient
172 expression, lack of detectable promoter activation, or sampling bias inherent to colony-
173 based isolation. The impact of the N-terminal Ala insertion was heterogeneous across
174 variants, with no consistent trend observed between the two forms.

175 The quantitative assay using a flow-cytometry exhibited a wide range of
176 transcriptional activities for each TF across the two promoters (**Fig. 2b**). Under identical
177 assay conditions, wild-type LuxR showed robust activation of the *lux* promoter but
178 negligible activity at the *las* promoter, whereas wild-type LasR activated the *las* promoter
179 strongly and displayed weak but detectable activity toward the *lux* promoter. Notably, a
180 subset of constructs activated both promoters to varying degrees, revealing balanced dual-
181 responsive behavior that is not observed for the parental LuxR or LasR proteins. Among

182 the variants that activated both promoters, four constructs (8M, 5MA, 1M, and 9M)
183 displayed distinct activity profiles. Variants 8M and 5MA showed higher activation at the
184 *las* promoter than at the *lux* promoter, whereas 9M exhibited the opposite trend and 1M
185 showed an intermediate balance between the two parental proteins (**Fig. 2b**). Consistent
186 with these functional differences, an unrooted phylogenetic analysis placed 8M and 5MA
187 closer to LasR, while 9M was positioned nearer to LuxR, with 1M occupying an
188 intermediate branch (**Fig. 2c**). Sequence alignment of these variants further revealed
189 residue-level features that parallel their activity patterns: 9M uniquely retained Arg30
190 conserved in LuxR subfamily, whereas the other three variants carried the Ala at this
191 position conserved in LasR subfamily; and both 8M and 5MA preserved LasR-associated
192 Arg40 and Arg41 (**Supplementary Fig. S1**). Together, these observations indicate that
193 promoter preference in the VAE-designed hybrids reflects combinations of parental
194 sequence features at key DNA-contacting positions.
195

Fig. 2 | Functional validation of VAE-generated hybrid TFs

a, Schematic overview of the GFP reporter assay used to quantify transcriptional activation by VAE-designed transcription factors. Each construct was expressed in *E. coli*, and promoter activity was measured from GFP driven by the native *lux* or *las* promoter. **b,** Transcriptional activation activities of VAE-designed variants and wild-type LuxR and LasR toward the *lux* and *las* promoters. Bars represent the geometric mean of GFP fluorescence measured by a flow-cytometry. A strain harboring pET16b (an empty vector) is shown as a control (Ctrl). Data are shown as the mean of three biological replicates, and error bars indicate the standard error of the mean. **c,** (left) Unrooted phylogenetic tree of the parental LuxR-family proteins and the four dual-responsive variants (8M, 5MA, 1M, and 9M). The tree was inferred from the LuxR-family DBD alignment and visualized without an assigned root. (right) Multiple sequence alignment of the variants shown in the phylogenetic tree together with the wild-type LuxR and LasR DBDs. Positions corresponding to residues 30, 40, and 41 discussed in the main text are highlighted with orange triangles.

213 ***Promoter recognition profiles of VAE-designed proteins revealed by sort-seq***

214 Building on the initial characterization of clones from 18 variants pool, we next
215 assessed promoter recognition across a larger set of VAE-designed sequences using a
216 massively parallel reporter assay (MPRA)[40,41]. Sixty additional VAE-designed DBD
217 sequences were randomly selected from the VAE-generated candidate sequences, and
218 two versions of each were created: one with M+CDS and the other with MA+CDS,
219 yielding a total of 120 unique variants. These sequences were synthesized as an
220 oligonucleotide pool and collectively cloned into the expression vector to construct a
221 pooled plasmid library, which was then transformed into *E. coli* harboring a GFP reporter
222 plasmid under the control of either *lux* or *las* promoter. Fluorescence from each cell was
223 measured by a fluorescence-activated cell sorter (FACS), and cells were sorted into
224 predefined fluorescence bins for each promoter. Plasmid DNAs from each bin were
225 extracted and subjected to deep-sequencing (sort-seq)[42,43], with activity of each variant
226 quantified by counting exact matches to the designed sequences across bins, normalizing
227 these counts by the fraction of cells in each bin of histogram from FACS analysis, and
228 calculating a geometric mean activity score for each promoter (**Fig. 3a; Supplementary**
229 **Fig. S2**).

230 Sixteen variants of the 120 VAE-designed proteins in the sort-seq exhibited
231 measurable activation toward either promoter. Notably, many of these active variants
232 showed detectable responses at both promoters, indicating that dual-responsiveness was
233 a frequent outcome within the functional subset (**Fig. 3b**). These variants displayed a
234 broad dynamic range of activities, spanning strongly *lux*-biased, strongly *las*-biased, and
235 dual-responsive profiles. Despite this diversity, dual-responsive variants were more
236 prevalent than strictly promoter-specific ones. Overall, the population-level trends

237 captured by the sort-seq reflected the behaviors identified in the single-construct assays,
238 supporting the robustness of the functional patterns observed.

239 To validate the activities inferred from the sort-seq, we individually constructed
240 plasmids encoding selected variants (No. 20, 22, and 24, designated as 20L, 22L, and 24L,
241 respectively) and measured their promoter activation using the GFP reporter assay
242 described above (**Fig. 2a**). The individually tested clones reproduced the activity patterns
243 observed in the pooled screen, confirming the dual-responsive behaviors inferred from
244 the sort-seq data (**Fig. 3c**). These results demonstrate that the sort-seq reliably captures
245 the promoter-recognition properties of VAE-designed transcription factors.

246

247 ***Structural insights into promoter recognition by the transcription factors***

248 To understand how LuxR-family transcription factors recognize their cognate
249 promoters and to identify residues potentially responsible for the altered specificity
250 observed in the VAE-designed variants, we modeled the complexes between LuxR or
251 LasR and the *lux* or *las* promoter sequences using AlphaFold3[44]. Four protein-DNA
252 complexes (LuxR•P*lux*, LuxR•P*las*, LasR•P*lux*, and LasR•P*las*) were initially predicted,
253 and the resulting structures were subjected to molecular dynamics simulations to examine
254 their stability and residue-level interactions.

255 Although global structural metrics such as RMSD and overall binding energetics
256 did not fully distinguish the four complexes (**Supplementary Fig. S3a**), the simulations
257 revealed reproducible patterns in specific residues that contribute to nucleotide
258 recognition in each protein (**Fig. 3d**). In LuxR, Arg30, described above, formed the only
259 consistently stable hydrogen bonds with specific bases (dT4, dG5, and dT6) within the
260 major groove, identifying it as a primary determinant of LuxR-specific base recognition.

261 This position is Ala30 in LasR, explaining the weaker and less specific base contacts
262 predicted for LasR at the same sites. In contrast, LuxR Thr31 engaged the DNA phosphate
263 backbone, suggesting a role in stabilizing DNA positioning rather than dictating base
264 identity; LasR carries an Asn31 at this site, which may alter backbone interactions without
265 imposing strong base preferences. A conserved Phe34 in both LuxR and LasR was
266 positioned for potential T-shaped π-π stacking with DNA bases (**Supplementary Fig.**
267 **S3b**), consistent with analogous interactions reported in another transcription factor
268 MarA[45]. LasR also contained two positively charged residues, Arg40 and Arg49, located
269 near the major groove; although direct hydrogen bonds were not consistently observed in
270 MD trajectories, both residues contributed strongly to total DNA-binding energy, whereas
271 the corresponding LuxR residues (Gln40 and Cys49) provided less energetic stabilization.
272 These differences highlight distinct recognition strategies employed by the two parental
273 regulators.

274 Notably, the VAE-designed variants 20L and 22L combined features of both
275 parental modes of DNA engagement (**Supplementary Fig. S3c**). At the Arg/Ala30
276 position, both variants carried the LasR-type Ala30, consistent with reduced LuxR-
277 specific base contacts. However, they also preserved the LuxR-type Thr31, potentially
278 maintaining similar DNA-backbone interactions. The LasR-associated Arg40 was
279 retained in both variants, while position 49 was occupied by Gln, intermediate in polarity
280 between the LuxR and LasR residues (Cys and Arg, respectively), suggesting partial
281 recapitulation of LasR-like electrostatic stabilization. Together with the conservation of
282 Phe34 and the universally retained Arg at position 48, these patterns suggest that 20L and
283 22L adopt hybrid combinations of the parental DNA-contacting residues, providing a
284 structural basis for their mixed and broadened promoter-recognition profiles.

Fig. 3 | Massively parallel reporter assays and structural features of VAE-designed TFs

a, Schematic overview of the sort-seq workflow. A total of 120 VAE-designed variants were constructed as a pooled plasmid library, transformed into *E. coli*, and sorted into fluorescence bins based on reporter output from the *lux* or *las* promoter. Plasmid DNA recovered from each bin was deep-sequenced, and activity scores were calculated for each variant from exact sequence counts across bins. **b,** Promoter-activation profiles of the 120 VAE-designed variants obtained from one of two biological replicate sort-seq experiments. The result of the replicate is shown in **Supplementary Fig. S4**. Each point represents the geometric mean activity score of a variant toward the *lux* or *las* promoter. Variants showing measurable activation include lux-biased, las-biased, and dual-responsive sequences. **c,** Validation of pooled-screen results by individual clone assays. Variants 20, 22, and 24 (20L, 22L, and 24L) were cloned individually and assessed using the GFP reporter assay. The individually measured activities recapitulate the promoter-recognition patterns inferred from the pooled deep-sequencing data. **d,** Representative structural models of parental transcription factors bound to their cognate promoters. AlphaFold-based models of LuxR bound to the *lux* promoter (left) and LasR bound to the *las* promoter (right) after molecular dynamics refinement. Proteins are shown in cartoon representation with selected DNA-contacting residues highlighted as sticks, and DNA is shown as a double helix.

309 ***Promoter recognition profiles of VAE-designed proteins revealed by deep sequencing***

310 Since the VAE-designed proteins exhibited broadened activities at the native *lux*
311 and *las* promoters, it was essential to determine whether their apparent dual
312 responsiveness reflected genuine sequence preferences or merely nonspecific DNA
313 binding. To address this, we generated randomized promoter libraries by mutating
314 nucleotide positions presumed to influence recognition by LuxR-family regulators and
315 then assessed promoter-recognition specificity for wild-type LuxR, wild-type LasR, and
316 VAE-designed selected variants (20L and 22L) using an MPRA. The randomized
317 promoter libraries were synthesized from random oligonucleotides and cloned upstream
318 of a GFP reporter to create pooled plasmid libraries. Either wild-type LuxR, wild-type
319 LasR, 20L, or 22L was expressed in *E. coli* containing the promoter library, and
320 fluorescence output from individual cells was quantified by FACS (**Fig. 4a**). Cells were
321 sorted into multiple fluorescence bins for each transcription factor, and promoter
322 sequences from each bin were recovered and deep-sequenced (sort-seq). For each
323 promoter variant, we calculated an enrichment-based activity score from its distribution
324 across fluorescence bins.

325 Analysis of the randomized promoter libraries revealed that the VAE-designed
326 variants 20L and 22L did not simply exhibit indiscriminate DNA binding. Although the
327 VAE-designed variants showed higher mean activities across the promoter library, their
328 activity distributions exhibited greater variance than those of the wild-type proteins. This
329 broader spread reflects an expanded dynamic range of promoter activation rather than
330 loss of sequence-selective recognition (**Fig. 4b**). To further examine how their sequence
331 preferences were altered, we quantified nucleotide-level enrichment across randomized
332 positions for each transcription factor (**Fig. 4c**). Both 20L and 22L displayed preference

333 profiles broadly similar to those of LasR, especially in the *las* promoter library. These
334 trends align with structural modeling, which suggested that LasR makes broader and less
335 restrictive base contacts than LuxR, providing a plausible route by which LasR-like
336 variants could acquire activity at the *lux* promoter (**Fig. 3c, d**). In contrast, analysis of the
337 *lux* promoter library revealed that LuxR-like nucleotide preferences were partially
338 retained. For example, at position 2, LasR favors G over T, whereas LuxR most strongly
339 prefers T; notably, 20L and 22L similarly exhibited a T > G preference at this position.
340 At position 1, both 20L and 22L retained a primary preference for A, similar to LasR,
341 while additionally exhibiting increased tolerance for G, which is most preferred by LuxR.
342 The nucleotide preference patterns observed for wild-type LuxR and LasR were
343 consistent with those reported in previous studies, supporting the validity of the assay[46,47].
344 Notably, furthermore, the VAE variants preferentially activated a subset of sequences not
345 strongly activated by both wild-type TF, as revealed by mapping promoter activities into
346 a four-dimensional activity landscape where LuxR and LasR activities are plotted on two
347 axes and the 20L or 22L activity is represented by point coloration (**Fig. 4d**). This pattern
348 indicates that the VAE-designed factors do not simply interpolate between LuxR and
349 LasR specificity but instead acquire hybrid or partially novel recognition profiles.
350

Fig. 4 | Specificity of VAE-designed transcription factors toward randomized promoter libraries

a, Schematic of the promoter library construction used for sort-seq. Randomized promoter libraries were generated by mutating selected nucleotide positions within the *lux* and *las* promoter regions and cloned upstream of a GFP reporter. Cells expressing wild-type LuxR, wild-type LasR, or VAE-designed variants (20L or 22L) were sorted into multiple fluorescence bins by FACS, and promoter sequences from each bin were recovered and analyzed by deep sequencing to calculate enrichment-based activity scores. **b,** Violin plots showing the distribution of activity scores across all randomized promoter variants for each transcription factor. Each violin illustrates the full activity distribution for wild-type LuxR, wild-type LasR, 20L, or 22L. Horizontal bars indicate group means. Pairwise differences in mean activity among the four DBDs were assessed by one-way ANOVA followed by post hoc multiple-comparison testing. **c,** Nucleotide-enrichment profiles derived from the randomized promoter libraries. Position-wise nucleotide preferences are shown for wild-type LuxR, wild-type LasR, 20L and 22L. The nucleotide identities of the native *lux* and *las* promoters are indicated by black outlines at positions previously reported to be conserved based on sequence logo analyses[46,47], shown in the corresponding LuxR and LasR heatmaps, respectively. **d,** Four-dimensional activity landscape of promoter responses to wild-type and VAE-designed transcription factors. Promoter activities for wild-type LasR and LuxR are plotted on the horizontal and vertical axes, respectively. Activities for 20L or 22L are encoded by color, and negative-control fluorescence is represented by point size.

376 **Discussion**

377 Our study demonstrates that a variational autoencoder can effectively design
378 hybrid transcription factors that bridge functional features of distinct LuxR-family
379 regulators. By training an MSA-VAE on a curated alignment of LuxR-family DNA-
380 binding domains and sampling sequences from the interpolated region between LuxR and
381 LasR in latent space, we generated a series of hybrid DBDs that exhibited functional
382 promoter activation *in vivo*. Notably, several of these VAE-designed proteins showed
383 dual-responsive activity toward both the *lux* and *las* promoters, which is not naturally
384 observed in either parental protein. This finding provides direct experimental support for
385 the idea that intermediate sequences between homologous transcription factors can
386 manifest intermediate or even expanded promoter-recognition capacities.

387 Conventional strategies for combining functionalities from homologous
388 transcription factors such as domain swapping or DNA shuffling are inherently limited in
389 how finely they can mix parental features. Because these approaches recombine blocks
390 of residues rather than individual amino acids, they often disrupt local residue-residue
391 and residue-nucleotide interactions that are essential for DNA recognition. Ancestral
392 sequence reconstruction (ASR) offers an alternative, evolutionarily grounded route
393 toward "intermediate" sequences, but in our experiments the ASR-derived candidates
394 showed LuxR-like behavior: one reconstructed sequence expressed poorly, and the
395 remaining two activated only the *lux* promoter (**Supplementary Fig. S5**). This limited
396 divergence may reflect the intrinsic constraints of ASR, which infers residues
397 independently at each position and therefore cannot capture co-variation of residues or
398 inter-residue interactions that support stable DNA-binding interfaces. Moreover, ASR
399 reconstructs the most probable ancestral states given the multiple sequence alignment;

400 consequently, biases in the underlying sequence dataset may be propagated into the
401 reconstructed sequences.

402 Beyond classical approaches, emerging protein design technologies also offer
403 alternative routes for engineering of transcription factors. For example, large protein
404 language models, trained on millions of diverse sequences, generate latent spaces that
405 capture broad biochemical trends[48,49]. However, it may not resolve local functional
406 distinctions among closely related regulators such as LuxR and LasR. Interpolation in
407 such global resolution is therefore unlikely to yield meaningful mixtures of parental
408 recognition features without additional constraints or task-specific fine-tuning.
409 Additionally, *de novo* DNA-binding domain design[50] enables the creation of entirely new
410 HTH scaffolds with programmable nucleotide preferences. As *de novo* frameworks
411 continue to improve, the design of hybrid or multi-specific transcription factors may
412 become achievable, providing an exciting future direction for synthetic genetic circuit
413 engineering.

414 Structural modeling and MD simulations provided a mechanistic basis for the
415 promoter-recognition behaviors observed in the VAE-designed transcription factors.
416 LuxR and LasR differ at several DNA-contacting residues that shape their respective
417 base-recognition modes. LuxR forms more restrictive base-specific interactions, and
418 LasR rather relies on a broader and more permissive contact network (**Fig. 3d**). These
419 structural trends align with the sort-seq promoter library data: the VAE-designed variants
420 (20L and 22L) exhibited LasR-like nucleotide preferences while also activating a subset
421 of promoter sequences not efficiently recognized by either wild-type factor. Visualization
422 of promoter activities in a multidimensional activity landscape further revealed that these
423 engineered proteins occupy promoter-recognition regimes intermediate to, yet distinct

424 from, both parents. Together, the structural and sort-seq results suggest that the VAE-
425 generated hybrids achieve dual responsiveness not through indiscriminate DNA binding,
426 but by combining permissive LasR-like contacts with specific contributions from
427 individual residues, enabling access to a partially novel specificity space.

428 While this study demonstrates that a VAE trained on natural LuxR-family
429 sequences can generate transcription factors with hybrid or partially novel promoter
430 specificities, several limitations remain. Our experimental validation focused exclusively
431 on the LuxR-LasR axis, and the generality of this framework across other transcription
432 factor families remains to be tested. In addition, VAE-based design relies on the
433 evolutionary diversity present in the underlying sequence alignment; families with sparse
434 sequence representation may prove more challenging. Despite these constraints,
435 extending this framework to larger families, integrating richer functional readouts, and
436 iteratively updating the model with experimental data will further enhance its capacity to
437 design proteins such as transcriptional regulators for increasingly complex synthetic
438 genetic circuits.
439

440 **Methods**

441 ***Creation of a dataset of LuxR-family proteins***

442 LuxR-family protein sequences were collected based on the LuxI/R dataset
443 reported by Miranda *et al*[37]. The representative LuxI/R query sequences used in that study,
444 together with the LuxR and LasR sequences (NCBI RefSeq accession ID:
445 WP_213090604.1 and WP_003082999.1, respectively), were added to the dataset and
446 aligned using MAFFT[51]. Because all sequences in the dataset are arranged as LuxI (N-
447 terminal) followed by LuxR (C-terminal), the boundary between the LuxI and LuxR
448 regions was manually determined by referencing conserved N-terminal residues among
449 the representative receiver proteins. The LuxI region was then removed to retain only the
450 LuxR-family portion. Sequences that became excessively short after gap removal (shorter
451 than 200 aa), contained large internal gaps, or were duplicates were iteratively filtered out
452 using CD-HIT[52]. MAFFT alignment and filtering were repeated for several rounds to
453 obtain a consistent LuxR-family multiple sequence alignment.

454 The DNA-binding domain (DBD) dataset was derived from the LuxR-family
455 alignment using experimentally validated N-terminal deletion boundaries of the LuxR
456 and LasR variants used in our laboratory. These representative sequences were added to
457 the dataset, realigned with MAFFT, and the conserved leucine residue marking the N-
458 terminal DBD boundary was identified. In this study, this conserved leucine was defined
459 as residue 1 for all subsequent numbering. Regions upstream of the aligned leucine
460 position were removed, and C-terminal trimming was performed to match the boundaries
461 used by Miranda *et al*. Sequences shorter than 56 aa after trimming or containing
462 extensive gaps were removed, and identical sequences were clustered using CD-HIT.
463 MAFFT alignment and gap-based filtering were repeated twice to increase alignment

464 consistency and information density. The resulting dataset contained only the DBD
465 regions of LuxR-family proteins.
466
467 ***Model building and validation using VAE***
468 We implemented a multiple-sequence-alignment variational autoencoder (MSA-
469 VAE) in PyTorch following architectures described by previous studies[29,33]. The model
470 was trained to maximize the evidence lower bound (ELBO), consisting of a binary cross-
471 entropy (BCE) reconstruction term and a Kullback-Leibler divergence (KLD)
472 regularization term:
473 $$\text{Loss} = \beta\ \text{KLD} + \text{BCE},$$
474 where β is a positive weighting coefficient. The KLD term encourages the encoder
475 posterior $q(z|x)$ to remain close to a standard normal prior, thereby promoting a smooth
476 and generalizable latent space. BCE was computed per aligned position using one-hot
477 encoded amino-acid categories[25].
478 To identify suitable model configurations, we performed a two-stage grid search
479 over (i) the dimension of the latent space, (ii) the size of hidden layers, (iii) the β
480 coefficient for the KLD term, and (iv) random seeds (For detail, see **Supplementary**
481 **Methods**). Models were trained with a batch size of 32, a learning rate of 1×10^{-3}, and the
482 Adam optimizer. Early stopping was applied when the validation ELBO failed to improve
483 for 5 consecutive epochs.
484 Model quality was assessed using the pairwise amino-acid frequency correlation
485 between generated sequences and the natural sequences in the training dataset, as
486 previously described[29]. Higher correlation values indicate that the model captures
487 evolutionary constraints, including covariation patterns. Unlike previous studies that

488 sample uniformly across the latent space, we generated sequences from interpolation
489 regions between two latent vectors corresponding to selected parental sequences[33,37].
490 Because models with similar hyperparameters can nonetheless yield distinct latent-space
491 geometries, we selected the top 10 models ranked by the evaluation metric and used all
492 of them for sequence generation (**Supplementary Table S1**). For each model, the LuxR
493 and LasR sequences were encoded to obtain their latent means z_{LuxR} and z_{LasR}. A latent
494 interpolation vector was then defined as:

495 $$z_{\mathrm{mix}} = 0.5\, z_{\mathrm{LuxR}} + 0.5\, z_{\mathrm{LasR}}$$

496 Sequences were sampled from a hyperspherical region centered at z_{mix} with a radius R
497 was defined as:

498 $$R = \frac{\|z_{\mathrm{LuxR}} - z_{\mathrm{LasR}}\|}{8}$$

499 Sampling was performed uniformly within this region by drawing random directions and
500 radii, followed by decoding the resulting latent points into amino acid sequences using
501 the VAE decoder. From this distribution, 20,000 sequences were generated per model.
502 After removing redundant sequences using CD-HIT[52], nine sequences were randomly
503 selected and used for experimental characterization.

504

505 ***Structural prediction and MD simulations***

506 The structural models of the parental transcription factors (LuxR and LasR) in
507 complex with *lux* or *las* promoter sequences were predicted using AlphaFold3[44]. Since
508 LuxR-family regulators function as homodimers, all predictions were performed in the
509 dimeric state with the corresponding DNA duplex included as input.

510 MD simulations were performed using GROMACS 2023.2 with the

511 AMBER99SB-ILDN force field[53,54] as described previously[55]. Each complex was
512 solvated using the SPC216 water model in a cubic box and neutralized with 150 mM
513 NaCl. Electrostatic and van der Waals interactions were truncated at 1.0 nm, and long-
514 range electrostatics were calculated with the particle mesh Ewald method[56]. Energy
515 minimization was performed using the steepest descent algorithm, and the system was
516 then equilibrated at 300 K and 1 bar for 100 ps each in two steps: first under constant
517 volume (NVT), then under constant pressure (NPT). Temperature and pressure were
518 controlled using the V-rescale thermostat and Parrinello-Rahman barostat, respectively[57].
519 The LINCS algorithm was used to constrain bonds involving hydrogen atoms[58]. The
520 integration time step was set as 2 fs. Production runs of 100 ns were conducted under the
521 NPT ensemble without restraints, with three independent replicates for each complex.
522 The RMSD and hydrogen bonds profiles were generated by GROMACS modules. Images
523 were generated by PyMOL (The PyMOL Molecular Graphics System, Version 3.1.0
524 Schrödinger, LLC).
525
526 ***Construction of plasmids***
527 Plasmids and primers used in this study were listed in **Supplementary Table S2**
528 **and S3**, respectively. An expression vector for LuxR DBD was constructed from pPtet
529 LuxR (Ptet-LuxR on pSB3K3)[59]. First, the P*tet-luxR* region fragment was excised using
530 EcoRI and PstI and subcloned into the corresponding restriction sites of pSB6A1. Next,
531 inverse PCR was performed to delete the N-terminal region of LuxR (amino acid residues
532 2-182), followed by self-ligation of the PCR product. An expression vector for the LasR
533 DBD was constructed using oligonucleotides synthesized as an oligo pool (Twist
534 Bioscience, United States). Briefly, double-stranded DNA was generated by primer

535 extension with a complementary oligonucleotide primer, purified, and digested with NheI
536 and HindIII. The resulting fragment was cloned into the corresponding restriction sites of
537 pTrc99A-DsRedExpress2 (for detailed procedures, see **Supplementary Methods**).
538 For VAE-designed proteins, coding sequences were selected at random from the
539 model-generated sequence pool, codon usage of each sequence is optimized, and a
540 ribosome-binding site was individually designed using the RBS Calculator (v2.1)[60] to
541 achieve comparable predicted translation initiation rates across variants. Resultant
542 sequences were synthesized as an oligonucleotide library (Twist Bioscience, United
543 States). Double-stranded DNA was generated by primer extension with a complementary
544 oligonucleotide primer, purified, and inserted between the NheI and HindIII sites of
545 pTrc99A-DsRedExpress2, as described above.
546 Promoter-reporter plasmids were constructed as follows. The gene fragment
547 containing the *las* promoter followed by *gfp* (iGEM Parts Registry: BBa_K649001) was
548 digested with XbaI and SpeI and cloned into the corresponding restriction sites of the
549 pSB3K3 vector. Subsequently, the *las* promoter region was replaced by digestion with
550 XbaI and SalI, and the *lux* promoter (iGEM Parts Registry: BBa_R0062) was inserted
551 into the same restriction sites. Randomized promoter libraries were generated
552 analogously using synthesized oligonucleotides in which selected nucleotide positions of
553 the *lux* or *las* promoter were randomized (FASMAC Co., Ltd.), followed by primer
554 extension, digestion with XbaI and SalI, and insertion into the same pSB3K3-P*lux*-GFP
555 or pSB3K3-P*las*-GFP backbone using the corresponding restriction sites. For functional
556 assays, each expression plasmid (or the empty vector pET16b(+)) and a reporter plasmid
557 were co-transformed into *Escherichia coli* DH5α.
558

559 ***In vivo assays of transcription activation abilities***

560 For individual reporter assays, *E. coli* DH5α cells carrying an expression plasmid
561 and either the reporter plasmid were grown overnight in LB medium supplemented with
562 50 μg/mL of carbenicillin and 25 μg/mL of kanamycin. Overnight cultures were diluted
563 into fresh LB medium and further grown 1 h at 37°C with shaking. After dilution of fresh
564 culture, protein expression was induced by adding IPTG at a final concentration of 1 mM
565 or 0.33 mM, followed by incubation for 6 h at 37°C. GFP fluorescence and cell density
566 were quantified using either a plate reader (Synergy, Agilent BioTek) or a flow cytometer
567 (CytoFLEX S, Beckman Coulter).

568 For MPRAs, *E. coli* DH5α cells were similarly co-transformed with a TF
569 expression plasmid and a reporter plasmid. For pooled TF library MPRA, the pooled
570 VAE-designed TF plasmid library was transformed into cells harboring either P*lux* or
571 P*las* reporter plasmid. For the randomized promoter-library MPRA, the promoter-GFP
572 library was transformed into cells harboring an individual TF expression plasmid (LuxR,
573 LasR, 20L, or 22L). In both cases, transformant colonies were collected from agar plates
574 and resuspended into fresh LB medium. In pooled TF library MPRA, resulting suspension
575 was grown 1 h at 37°C with shaking, induced by adding IPTG at a final concentration of
576 1 mM, followed by incubation for 6 h at 37°C. In the randomized promoter-library MPRA,
577 resulting suspension was induced by adding IPTG at a final concentration of 1 mM,
578 followed by incubation for 10 h at 15°C and further 7.5 h at 37°C. After induction, cells
579 were diluted and subjected to fluorescence-activated cell sorting (FACS). Cells were
580 sorted based on GFP-A fluorescence intensity using a BD FACSAria Fusion Cell Sorter
581 (BD Biosciences, United States) equipped with a 100-μm nozzle. Sorting was performed
582 using the 'Purity' precision mode. The GFP-A histogram was partitioned into 8

583 consecutive gates (bins), and a target of 10,000 cells from each bin were collected into
584 individual wells of a 96-well plate when sufficient events were available, with fewer cells
585 collected for sparsely populated high-fluorescence bins (**Supplementary Fig. S2 and S6**).
586 Each well contained 1/10-strength LB medium supplemented with 50 μg/mL
587 carbenicillin and 25 μg/mL kanamycin. Cells collected from each bin were pelleted,
588 resuspended in 18 μL of Milli-Q water and lysed by heating at 94°C for 5 min. The
589 variable region derived from plasmid DNA (TF coding sequence for the TF library;
590 promoter insert for the promoter library) was amplified by PCR to attach adapter
591 sequences using primers compatible with adapters. To maintain the original population's
592 profile, the resulting PCR products from each bin were pooled according to the cell
593 distribution ratio observed in the FACS histogram. The pooled products were sequenced
594 using an Illumina NextSeq 1000 Sequencing System, with sequencing services provided
595 by Bioengineering Lab. Co., Ltd. Sequence counts were obtained by exact-match parsing,
596 and enrichment-based activity scores were calculated for each variant by comparing its
597 representation across fluorescence bins.
598
599 ***Statistical analysis of promoter library activities***
600 To assess whether activity distributions differed significantly among
601 transcription factors, both parametric and nonparametric statistical tests were performed.
602 One-way analysis of variance (ANOVA) was applied to raw activity values, followed by
603 the Kruskal-Wallis test as a distribution-free alternative. Post hoc pairwise comparisons
604 were conducted using Dunn's test with Bonferroni correction. Homogeneity of variances
605 was assessed using Levene's test.
606 For constructing feature matrix X, each sequence was encoded using one-hot

607 encoding at the variable positions. For each variable position, four binary features
608 corresponding to nucleotides A, C, G, and T were generated. The response matrix Y
609 consisted of activity values for LuxR, LasR, 20L, and 22L. Raw activity values were
610 transformed to reduce skewness. The transformed values were then standardized (zero
611 mean and unit variance) independently for each transcription factor prior to multivariate
612 analysis. Partial least squares (PLS) regression was performed to model the relationship
613 between sequence features and transcriptional activity. PLS was chosen to extract latent
614 components of the sequence feature space that maximize covariance with the activity
615 measurements. The one-hot encoded sequence matrix X was used as the predictor matrix,
616 and the standardized activity matrix Y was used as the response matrix. Two PLS
617 components were retained for analysis. Model fitting was performed using the
618 PLSRegression implementation from scikit-learn. PLS regression coefficients were
619 extracted to quantify the contribution of each nucleotide at each variable position to
620 transcriptional activity. Coefficients were reshaped into matrices representing nucleotide
621 preferences (A, C, G, T) at each variable position.
622
623 **Data availability**
624 All custom code used for VAE model training and latent-space sampling and
625 datasets curated for training the VAE models are available at GitHub (kigalab/VAE-
626 hybridTF). Raw sequencing data from the pooled transcription factor library sort-seq
627 experiments and randomized promoter-library sort-seq experiments have been deposited
628 in Zenodo (10.5281/zenodo.18798256). Additional data supporting the findings of this
629 study are available at Supplementary Information.
630

Acknowledgments

We thank R. Yoshida, K. Takahashi, Y. Okada and T. Ueda for critical discussion of the manuscript and valuable suggestions; and the members of the Laboratory of Synthetic Biology (the D. Kiga group) for their support. In particular, we are grateful to H. Shimokawa for assistance with model construction and coding, and to R. Sekine, Y. Takizawa, S. Nishimura and A. Sasaki for providing materials. This work was financed by the JSPS KAKENHI (Grant numbers 19H00985, 21K19831, 21H05228, 22H04892, 23H04431, 24K03036, and 25K22848) to D.K., by JST CREST (Grant number JPMJCR21N4) to D.K. including AIP challenge program to S.L.O. and K.U., and by NINS Astrobiology Center program research (Grant numbers AB041011, AB0602, and AB0712) to D.K.

CRediT Author contributions

S.L.O.: conceptualization, data curation, investigation, formal analysis, methodology, funding acquisition, software, visualization, writing – original draft, review and editing. A.M.: formal analysis, investigation, methodology, software, visualization, writing – original draft, review and editing. M.A.: formal analysis, investigation, data curation, writing – original draft, review and editing. K.U.: investigation, funding acquisition. K.M.: resources. K.O.: supervision, writing – review and editing. D.K.: conceptualization, formal analysis, methodology, funding acquisition, project administration, resources, supervision, writing - review and editing.

Competing interests

The authors declare no competing interests.

656 **References**

657 1. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology.
658 *Nat. Rev. Microbiol.* **12**, 381–390 (2014).
659 2. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional
660 regulators. *Nature* **403**, 335–338 (2000).
661 3. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch
662 in Escherichia coli. *Nature* **403**, 339–342 (2000).
663 4. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in
664 Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.
665 *Nucleic Acids Res.* **25**, 1203–1210 (1997).
666 5. Zhan, J. *et al.* Develop reusable and combinable designs for transcriptional logic
667 gates. *Mol. Syst. Biol.* **6**, 388 (2010).
668 6. Slusarczyk, A. L., Lin, A. & Weiss, R. Foundations for the design and
669 implementation of synthetic genetic circuits. *Nat. Rev. Genet.* **13**, 406–420 (2012).
670 7. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science* **352**, aac7341
671 (2016).
672 8. Chen, Y. *et al.* Tuning the dynamic range of bacterial promoters regulated by ligand-
673 inducible transcription factors. *Nat. Commun.* **9**, 64 (2018).
674 9. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A.
675 Escherichia coli "Marionette" strains with 12 highly optimized small-molecule
676 sensors. *Nat. Chem. Biol.* **15**, 196–204 (2019).
677 10. Du, P. *et al.* De novo design of an intercellular signaling toolbox for multi-channel
678 cell-cell communication and biological computation. *Nat. Commun.* **11**, 4226 (2020).
679 11. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand

displacement cascades. *Nature* **475**, 368–372 (2011).

12. Cherry, K. M. & Qian, L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature* **559**, 370–376 (2018).

13. Okumura, S. *et al.* Nonlinear decision-making with enzymatic neural networks. *Nature* **610**, 496–501 (2022).

14. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science* **333**, 1307–1311 (2011).

15. Didovyk, A. *et al.* Distributed classifier based on genetically engineered bacterial cell cultures. *ACS Synth. Biol.* **4**, 72–82 (2015).

16. Samaniego, C. C., Moorman, A., Giordano, G. & Franco, E. Signaling-based neural networks for cellular computation. in *2021 American Control Conference (ACC)* 1883–1890 (IEEE, 2021).

17. Kanakov, O. *et al.* Multi-input distributed classifiers for synthetic genetic circuits. *PLoS One* **10**, e0125144 (2015).

18. Li, X. *et al.* Synthetic neural-like computing in microbial consortia for pattern recognition. *Nat. Commun.* **12**, 3139 (2021).

19. Rizik, L., Danial, L., Habib, M., Weiss, R. & Daniel, R. Synthetic neuromorphic computing in living cells. *Nat. Commun.* **13**, 5602 (2022).

20. Chen, Z. *et al.* A synthetic protein-level neural network in mammalian cells. *Science* **386**, 1243–1250 (2024).

21. Tang, L., Wang, J., Xu, K., Li, Z. & Song, J. Upgraded circular single-stranded DNA regulators for multiple-input multiple-output gene circuits in mammalian cells. *Sci. Adv.* **11**, eadv3396 (2025).

704 22. Bashton, M. & Chothia, C. The generation of new protein functions by the
705 combination of domains. *Structure* **15**, 85–99 (2007).
706 23. Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: in vitro
707 recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10747–
708 10751 (1994).
709 24. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral
710 sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141
711 (2021).
712 25. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]*
713 (2013).
714 26. Strokach, A. & Kim, P. M. Deep generative modeling for protein design. *Curr. Opin.*
715 *Struct. Biol.* **72**, 226–236 (2022).
716 27. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic
717 variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
718 28. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein
719 folds using variational autoencoders. *Sci. Rep.* **8**, 16189 (2018).
720 29. Hawkins-Hooker, A. *et al.* Generating functional protein variants with variational
721 autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
722 30. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: Generative modeling of protein
723 structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**, e1010271
724 (2022).
725 31. Sevgen, E. *et al.* ProT-VAE: Protein Transformer Variational AutoEncoder for
726 functional protein design. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2408737122 (2025).
727 32. Boob, A. G. *et al.* Design of diverse, functional mitochondrial targeting sequences

728 across eukaryotic organisms using variational autoencoder. *Nat. Commun.* **16**, 4151
729 (2025).
730 33. Kohyama, S., Frohn, B. P., Babl, L. & Schwille, P. Machine learning-aided design
731 and screening of an emergent protein function in synthetic cells. *Nat. Commun.* **15**,
732 2010 (2024).
733 34. Tsai, C.-S. & Winans, S. C. LuxR-type quorum-sensing regulators that are detached
734 from common scents: LuxR-type apoproteins. *Mol. Microbiol.* **77**, 1072–1082
735 (2010).
736 35. Choi, S. H. & Greenberg, E. P. The C-terminal region of the Vibrio fischeri LuxR
737 protein contains an inducer-independent lux gene activating domain. *Proc. Natl.*
738 *Acad. Sci. U. S. A.* **88**, 11115–11119 (1991).
739 36. Kiratisin, P., Tucker, K. D. & Passador, L. LasR, a transcriptional activator of
740 Pseudomonas aeruginosa virulence genes, functions as a multimer. *J. Bacteriol.* **184**,
741 4912–4919 (2002).
742 37. Wellington Miranda, S. *et al.* A covariation analysis reveals elements of selectivity
743 in quorum sensing systems. *Elife* **10**, (2021).
744 38. Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G. & Blanquet, S. Extent of N-
745 terminal methionine excision from Escherichia coli proteins is governed by the side-
746 chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 8247–
747 8251 (1989).
748 39. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria.
749 *Science* **254**, 1374–1377 (1991).
750 40. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays.
751 *Genomics* **106**, 159–164 (2015).

752 41. White, M. A. Understanding how cis-regulatory function is encoded in DNA
753 sequence using massively parallel reporter assays and designed sequences.
754 *Genomics* **106**, 165–170 (2015).
755 42. Kinney, J. B., Murugan, A., Callan, C. G., Jr & Cox, E. C. Using deep sequencing
756 to characterize the biophysical mechanism of a transcriptional regulatory sequence.
757 *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9158–9163 (2010).
758 43. Barnes, S. L., Belliveau, N. M., Ireland, W. T., Kinney, J. B. & Phillips, R. Mapping
759 DNA sequence to transcription factor binding energy in vivo. *PLoS Comput. Biol.*
760 **15**, e1006226 (2019).
761 44. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
762 AlphaFold 3. *Nature* **630**, 493–500 (2024).
763 45. Corbella, M. *et al.* The N-terminal helix-turn-helix motif of transcription factors
764 MarA and Rob drives DNA recognition. *J. Phys. Chem. B* **125**, 6791–6806 (2021).
765 46. Antunes, L. C. M., Ferreira, R. B. R., Lostroh, C. P. & Greenberg, E. P. A mutational
766 analysis defines Vibrio fischeri LuxR binding sites. *J. Bacteriol.* **190**, 4392–4397
767 (2008).
768 47. Gilbert, K. B., Kim, T. H., Gupta, R., Greenberg, E. P. & Schuster, M. Global
769 position analysis of the Pseudomonas aeruginosa quorum-sensing transcription
770 factor LasR. *Mol. Microbiol.* **73**, 1072–1085 (2009).
771 48. Kyro, G. W., Qiu, T. & Batista, V. S. A model-centric review of deep learning for
772 protein design. *arXiv [cs.LG]* (2025).
773 49. Xiao, Y. *et al.* Protein large language models: A comprehensive Survey. *arXiv [q-*
774 *bio.BM]* (2025) doi:10.48550/arXiv.2502.17504.
775 50. Glasscock, C. J. *et al.* Computational design of sequence-specific DNA-binding

776 proteins. *Nat. Struct. Mol. Biol.* **32**, 2252–2261 (2025).
777 51. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid
778 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**,
779 3059–3066 (2002).
780 52. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets
781 of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
782 53. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through
783 multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25
784 (2015).
785 54. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically
786 disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
787 55. Shigematsu, M. *et al.* RNase 2/EDN cleaves tRNA anticodon loops to generate
788 immunoactive RNAs. *bioRxiv* 815803 (2025) doi:10.1101/815803.
789 56. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for
790 Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
791 57. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new
792 molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
793 58. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear
794 constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
795 59. Sekine, R. *et al.* Tunable synthetic phenotypic diversification on Waddington's
796 landscape through autonomous signaling. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17969–
797 17973 (2011).
798 60. Salis, H. M. The ribosome binding site calculator. *Methods Enzymol.* **498**, 19–42
799 (2011).

Fig. 1

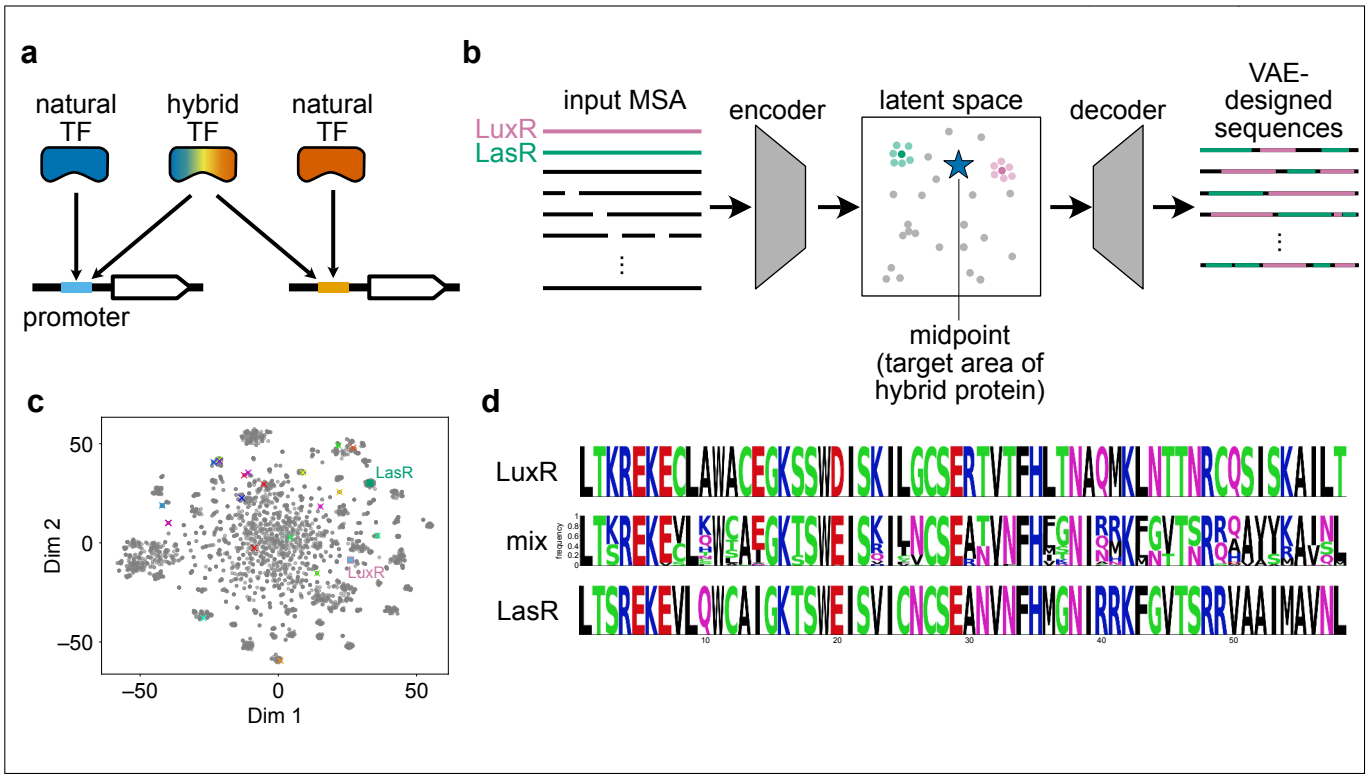

a
natural TF
hybrid TF
natural TF
promoter
b
input MSA
LuxR
LasR
encoder
latent space
decoder
VAE-designed sequences
midpoint (target area of hybrid protein)
c
Dim 2
Dim 1
50
0
–50
–50
0
50
LasR
LuxR
d
LuxR
mix
LasR

Fig. 2

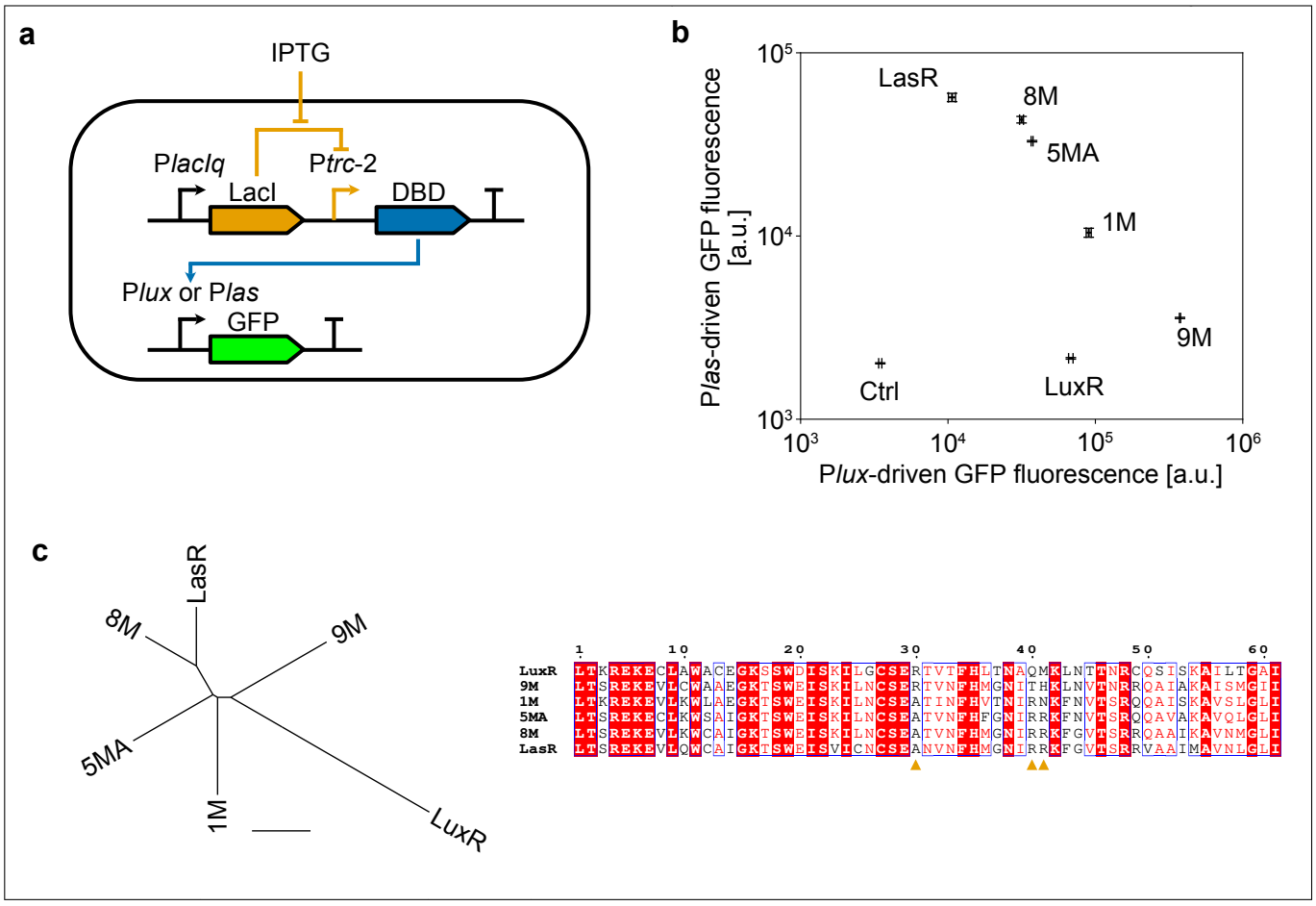

a
IPTG
PlacIq
PtrC-2
LacI
DBD
Plux or Plas
GFP
b
LasR
8M
5MA
1M
9M
Ctrl
LuxR
Plas-driven GFP fluorescence [a.u.]
Plux-driven GFP fluorescence [a.u.]
c
LasR
8M
9M
5MA
1M
LuxR
LuxR
9M
1M
5MA
8M
LasR

Fig. 3

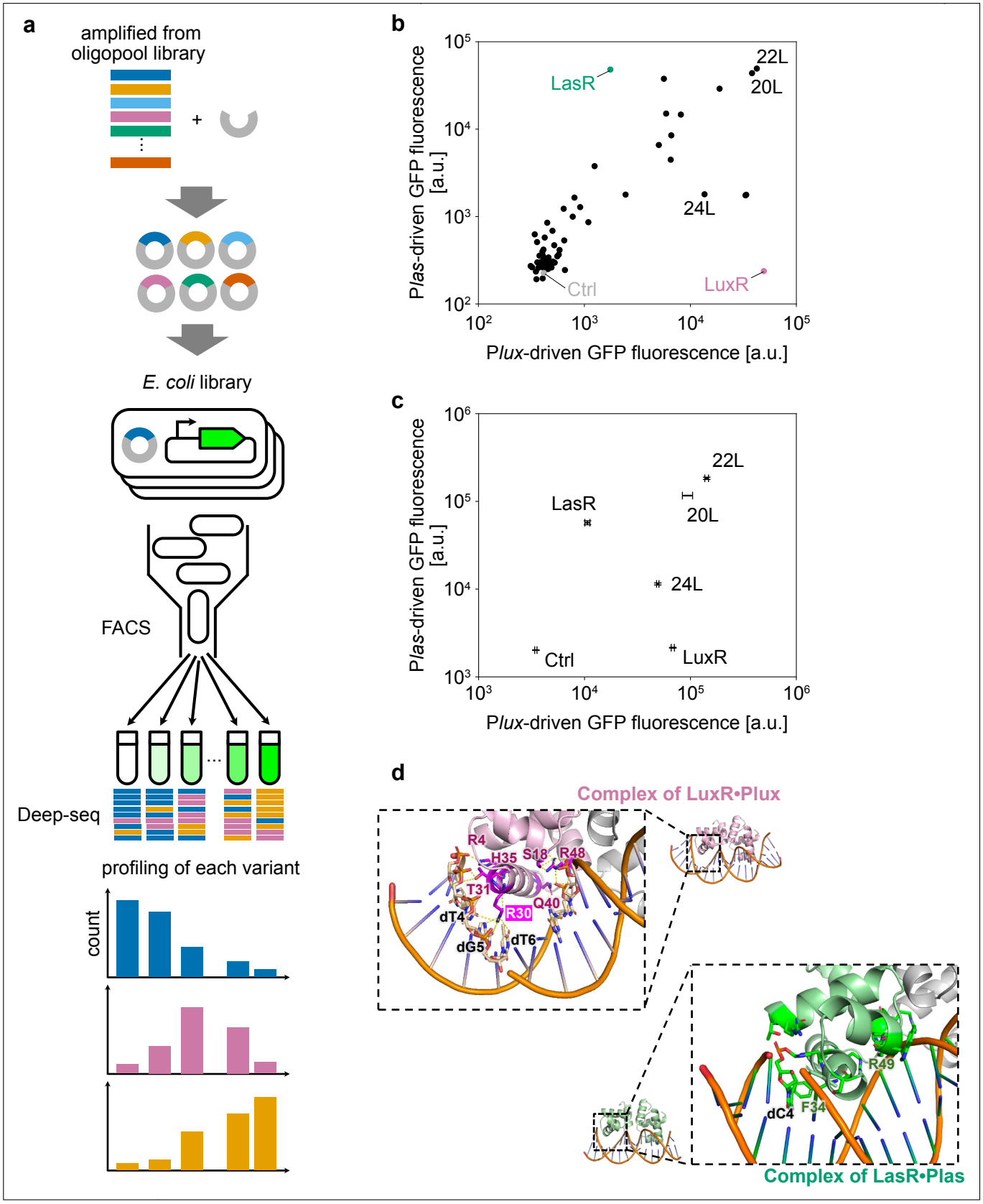
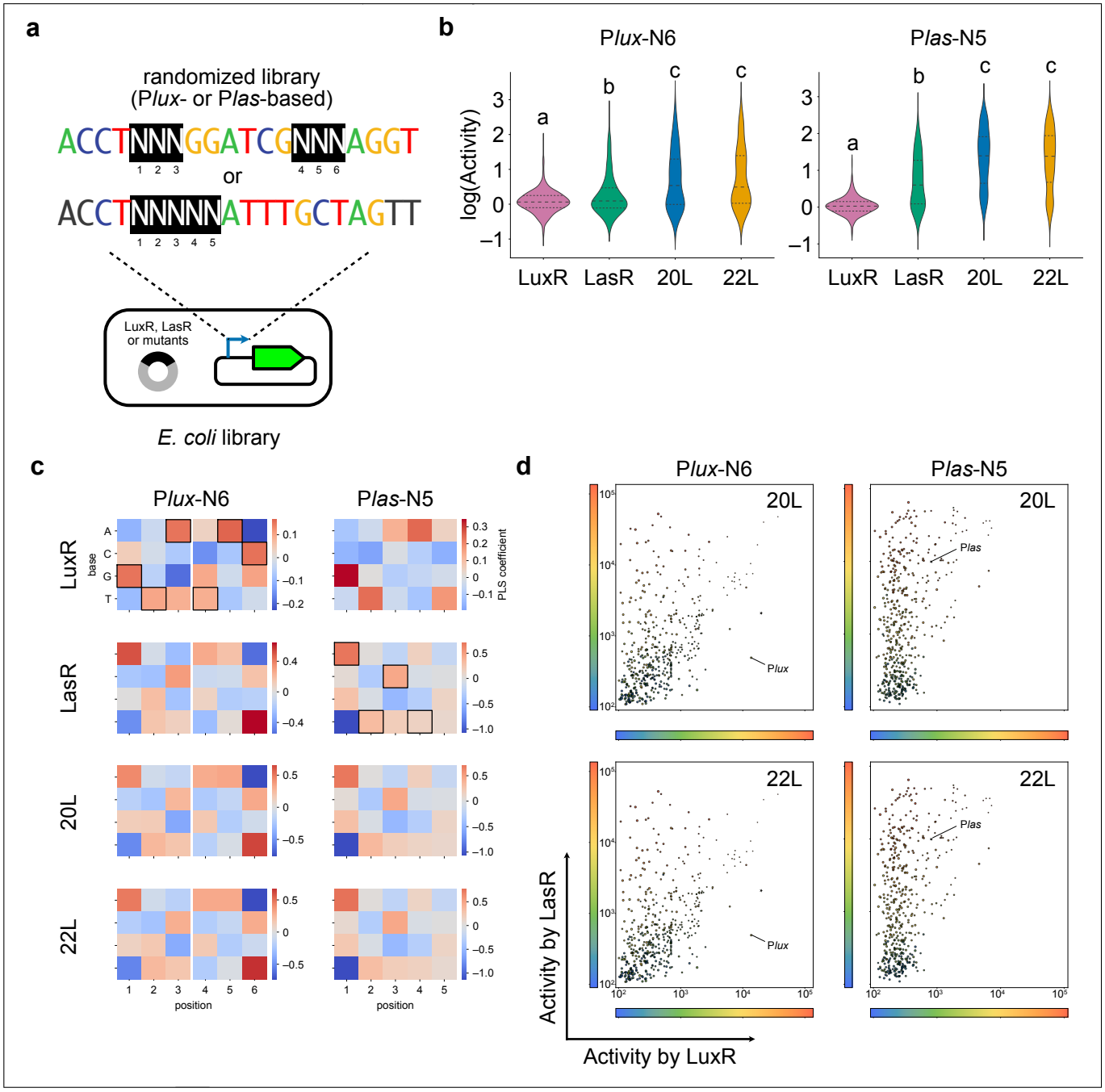

a
amplified from oligopool library
+
E. coli library
FACS
Deep-seq
profiling of each variant
count
b
P*las*-driven GFP fluorescence [a.u.]
P*lux*-driven GFP fluorescence [a.u.]
LasR
22L
20L
24L
Ctrl
LuxR
c
P*las*-driven GFP fluorescence [a.u.]
P*lux*-driven GFP fluorescence [a.u.]
LasR
22L
20L
24L
Ctrl
LuxR
d
Complex of LuxR•Plux
R4
H35
S18
R48
T31
Q40
R30
dT4
dG5
dT6
R49
F34
dC4
Complex of LasR•Plas

Fig. 4


a
randomized library
(Plux- or Plas-based)
ACCTNNNGGATCGNNNAGGT
1 2 3
4 5 6
or
ACCTNNNNNATTTGCTAGTT
1 2 3 4 5
LuxR, LasR
or mutants
E. coli library
b
Plux-N6
Plas-N5
log(Activity)
LuxR
LasR
20L
22L
c
Plux-N6
Plas-N5
LuxR
LasR
20L
22L
base
A
C
G
T
PLS coefficient
position
d
Plux-N6
Plas-N5
20L
22L
Plux
Plas
Activity by LasR
Activity by LuxR