Title

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

Authors

Sota 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 gene circuits to program cellular behaviors. To ensure reliable circuit performance, most synthetic gene circuits rely on well-characterized and orthogonal regulatory parts. 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 gene 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 gene circuits[2,3]. Within such circuits, transcriptional regulation plays a central role, with TFs acting as molecular switches that translate environmental or intracellular cues into gene expression outputs. To simplify design and minimize crosstalk, well-characterized and orthogonal parts have been predominantly used for constructing synthetic gene 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, multi-input and multi-output genetic circuits have demonstrated neural network-like pattern recognition and decision-making *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 gene 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, including DNA recognition by TFs, is ultimately

68 dictated by its amino acid sequence, 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[26–32]. 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. Since the DBD alone is sufficient to bind target DNA and can activate transcription independently of ligand binding[33,34], 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[35] and trained an MSA-VAE on this dataset[27,29,36]. 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

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

126 Using this latent-space representation, we sampled sequences from a
127 hyperspherical region centered at the midpoint between the LuxR and LasR latent vectors
128 (0.5-0.5 interpolation). This strategy enabled targeted sampling of hybrid sequences, in
129 contrast to prior VAE studies that sampled amino acid sequences closely related or
130 neighboring locations in the latent space for collecting similar variants[27,29,32,36]. Sampling
131 across the interpolation region produced a diverse set of candidate sequences that contain
132 novel combinations of parental residues while retaining global amino-acid usage patterns
133 of the LuxR family. From each of the top-performing VAE models, 20,000 sequences
134 were decoded from the local interpolation region. After removing redundant and gap-
135 containing sequences using CD-HIT, we obtained a pool of unique candidate hybrids (**Fig.**
136 **1d**).
137

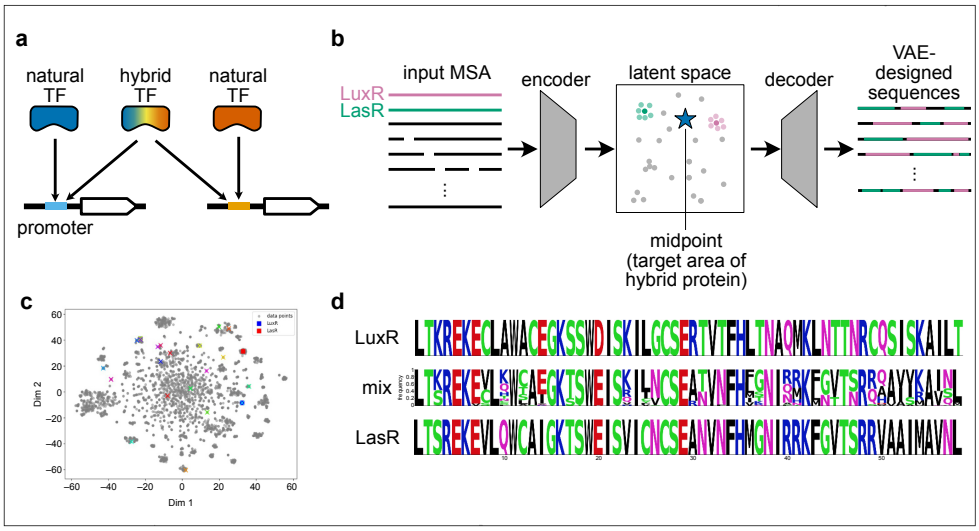


Fig. 1 | Schematic of the VAE-based sequence generation strategy for hybrid TFs

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,** Latent-space projection of representative LuxR-family transcription factors. The positions of LuxR and LasR are indicated by a blue square and a red 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.

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

160 To determine whether the VAE-designed hybrid sequences exhibit functional
161 activity *in vivo*, we selected 9 candidates at random from the generated sequences and
162 evaluated their ability to activate transcription from the native *lux* and *las* promoters in
163 *Escherichia coli*. For each designed variant, we constructed two coding sequences that
164 differed only in their immediate N-terminal context: one beginning with Met ("M+CDS"),
165 and the other beginning Met followed by Ala ("MA+CDS") to account for potential N-
166 terminal context effects[37,38], resulting in a total of 18 constructs. The mixed
167 oligonucleotide pool was cloned and expressed in *E. coli*, and its ability to activate
168 transcription from the native *lux* and *las* promoters was quantified using GFP reporter
169 assays (**Fig. 2a**). Some of individual fluorescent colonies were isolated, and the
170 corresponding coding sequences were identified by sequencing prior to quantitative
171 analysis. Variants that were not recovered for quantitative analysis likely reflect a
172 combination of factors, including severe growth defects, unstable or insufficient
173 expression, lack of detectable promoter activation, or sampling bias inherent to colony-
174 based isolation. The impact of the N-terminal Ala insertion was heterogeneous across
175 variants, with no consistent trend observed between the two forms.

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

183 the variants that activated both promoters, four constructs (8M, 5MA, 1M, and 9M)
184 displayed distinct activity profiles. Variants 8M and 5MA showed higher activation at the
185 *las* promoter than at the *lux* promoter, whereas 9M exhibited the opposite trend and 1M
186 showed an intermediate balance between the two parental proteins (**Fig. 2b**). Consistent
187 with these functional differences, an unrooted phylogenetic analysis placed 8M and 5MA
188 closer to LasR, while 9M was positioned nearer to LuxR, with 1M occupying an
189 intermediate branch (**Fig. 2c**). Sequence alignment of these variants further revealed
190 residue-level features that parallel their activity patterns: 9M uniquely retained Arg30
191 conserved in LuxR subfamily, whereas the other three variants carried the Ala at this
192 position conserved in LasR subfamily; and both 8M and 5MA preserved LasR-associated
193 Arg40 and Arg41 (**Supplementary Fig. S1**). Together, these observations indicate that
194 promoter preference in the VAE-designed hybrids reflects combinations of parental
195 sequence features at key DNA-contacting positions.
196

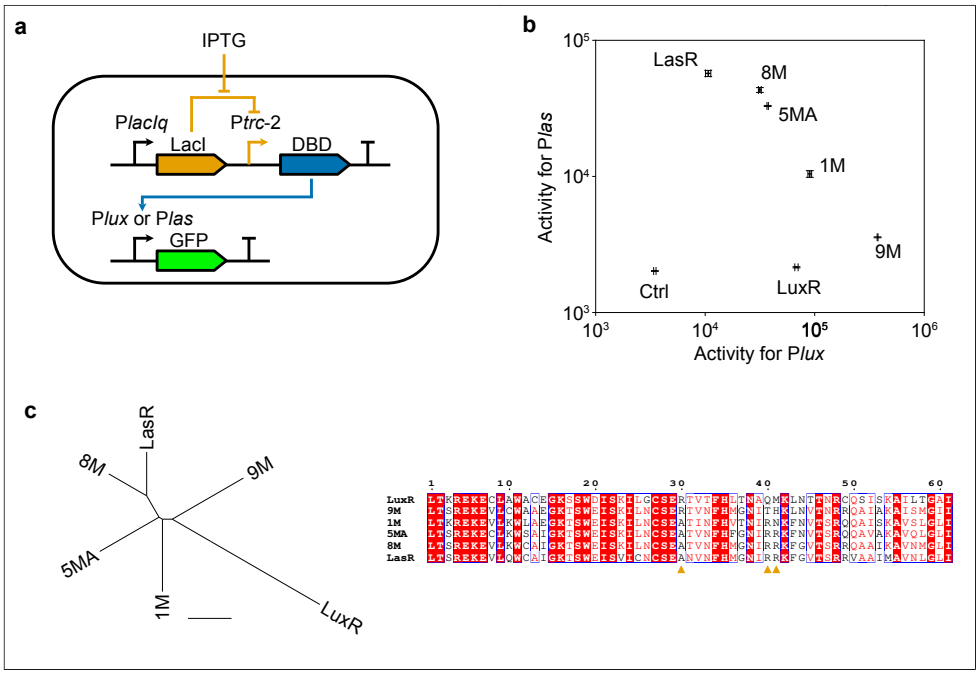


Fig. 2 | Functional validation of VAE-generated hybrid TFs

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.

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

216 Building on the initial characterization of clones from 18 variants pool, we next
217 assessed promoter recognition across a larger set of VAE-designed sequences using a
218 massively parallel reporter assay (MPRA)[39,40]. Sixty additional VAE-designed DBD
219 sequences were randomly chosen, and two versions of each were created: one with
220 M+CDS and the other with MA+CDS, yielding a total of 120 unique variants. These
221 sequences were synthesized as an oligonucleotide pool and collectively cloned into the
222 expression vector to construct a pooled plasmid library, which was then transformed into
223 *E. coli* harboring a GFP reporter plasmid induced by either *lux* or *las* promoter.
224 Fluorescence from each cell was measured by a fluorescence-activated cell sorter (FACS),
225 and cells were sorted into predefined fluorescence bins for each promoter. Plasmid DNAs
226 from each bin were recovered and deep-sequenced (sort-seq)[41,42], with activity of each
227 variant quantified by counting exact matches to the designed sequences across bins,
228 normalizing these counts by the fraction of cells in each bin of histogram from FACS
229 analysis, and calculating a geometric mean activity score for each promoter (**Fig. 3a;**
230 **Supplementary Fig. S2**).

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

239 supporting the robustness of the functional patterns observed.

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

247

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

249 To understand how LuxR-family transcription factors recognize their cognate
250 promoters and to identify residues potentially responsible for the altered specificity
251 observed in the VAE-designed variants, we modeled the complexes between LuxR or
252 LasR and the *lux* or *las* promoter sequences using AlphaFold3[43]. Four protein-DNA
253 complexes (LuxR•P*lux*, LuxR•P*las*, LasR•P*lux*, and LasR•P*las*) were initially predicted,
254 and the resulting structures were subjected to molecular dynamics simulations to examine
255 their stability and residue-level interactions.

256 Although global structural metrics such as RMSD and overall binding energetics
257 did not fully distinguish the four complexes (**Supplementary Fig. S3a**), the simulations
258 revealed reproducible patterns in specific residues that contribute to nucleotide
259 recognition in each protein (**Fig. 3d**). In LuxR, Arg30, described above, formed the only
260 consistently stable hydrogen bonds with specific bases (dT4, dG5, and dT6) within the
261 major groove, identifying it as a primary determinant of LuxR-specific base recognition.
262 This position is Ala30 in LasR, explaining the weaker and less specific base contacts

263 predicted for LasR at the same sites. In contrast, LuxR Thr31 engaged the DNA phosphate
264 backbone, suggesting a role in stabilizing DNA positioning rather than dictating base
265 identity; LasR carries an Asn31 at this site, which may alter backbone interactions without
266 imposing strong base preferences. A conserved Phe34 in both LuxR and LasR was
267 positioned for potential T-shaped π-π stacking with DNA bases (**Supplementary Fig.**
268 **S3b**), consistent with analogous interactions reported in another transcription factor
269 MarA[44]. LasR also contained two positively charged residues, Arg40 and Arg49, located
270 near the major groove; although direct hydrogen bonds were not consistently observed in
271 MD trajectories, both residues contributed strongly to total DNA-binding energy, whereas
272 the corresponding LuxR residues (Gln40 and Cys49) provided less energetic stabilization.
273 These differences highlight distinct recognition strategies employed by the two parental
274 regulators.

275 Notably, the VAE-designed variants 20L and 22L combined features of both
276 parental modes of DNA engagement (**Supplementary Fig. S3c**). At the Arg/Ala30
277 position, both variants carried the LasR-type Ala30, consistent with reduced LuxR-
278 specific base contacts. However, they also preserved the LuxR-type Thr31, potentially
279 maintaining similar DNA-backbone interactions. The LasR-associated Arg40 was
280 retained in both variants, while position 49 was occupied by Gln, intermediate in polarity
281 between the LuxR and LasR residues (Cys and Arg, respectively), suggesting partial
282 recapitulation of LasR-like electrostatic stabilization. Together with the conservation of
283 Phe34 and the universally retained Arg at position 48, these patterns suggest that 20L and
284 22L adopt hybrid combinations of the parental DNA-contacting residues, providing a
285 structural basis for their mixed and broadened promoter-recognition profiles.
286

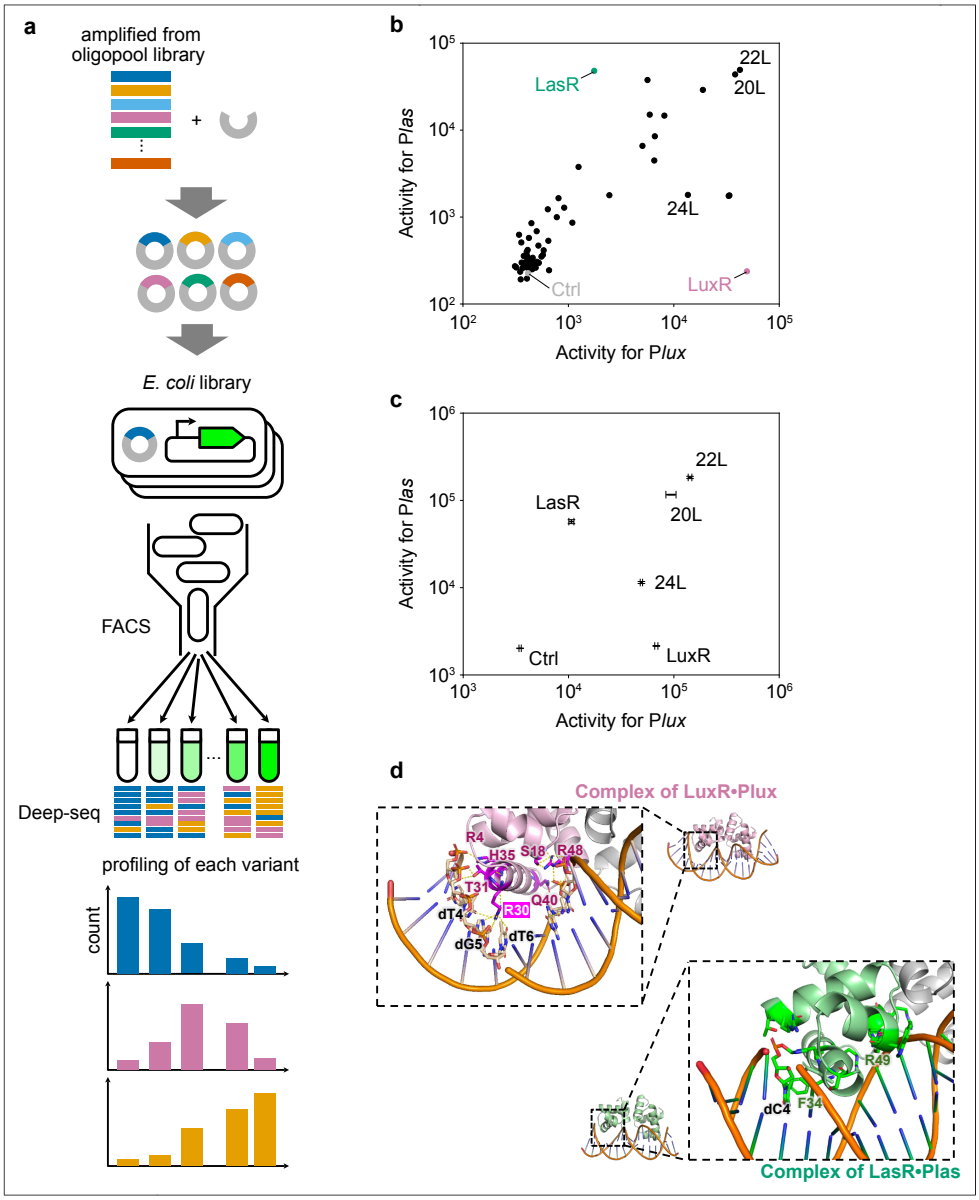


Fig. 3 | Massively parallel reporter assays and structural features of VAE-designed TFs

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.

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

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

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

335 profiles broadly similar to those of LasR, especially in the *las* promoter library. These
336 trends align with structural modeling, which suggested that LasR makes broader and less
337 restrictive base contacts than LuxR, providing a plausible route by which LasR-like
338 variants could acquire activity at the *lux* promoter (**Fig. 3c, d**). In contrast, analysis of the
339 *lux* promoter library revealed that LuxR-like nucleotide preferences were partially
340 retained. For example, at position 2, LasR favors G over T, whereas LuxR most strongly
341 prefers T; notably, 20L and 22L similarly exhibited a T > G preference at this position.
342 At position 1, both 20L and 22L retained a primary preference for A, similar to LasR,
343 while additionally exhibiting increased tolerance for G, which is most preferred by LuxR.
344 The nucleotide preference patterns observed for wild-type LuxR and LasR were
345 consistent with those reported in previous studies, supporting the validity of the assay[45,46].
346 Notably, furthermore, the VAE variants preferentially activated a subset of sequences not
347 strongly activated by both wild-type TF, as revealed by mapping promoter activities into
348 a four-dimensional activity landscape where LuxR and LasR activities are plotted on two
349 axes and the 20L or 22L activity is represented by point coloration (**Fig. 4d**). This pattern
350 indicates that the VAE-designed factors do not simply interpolate between LuxR and
351 LasR specificity but instead acquire hybrid or partially novel recognition profiles.
352

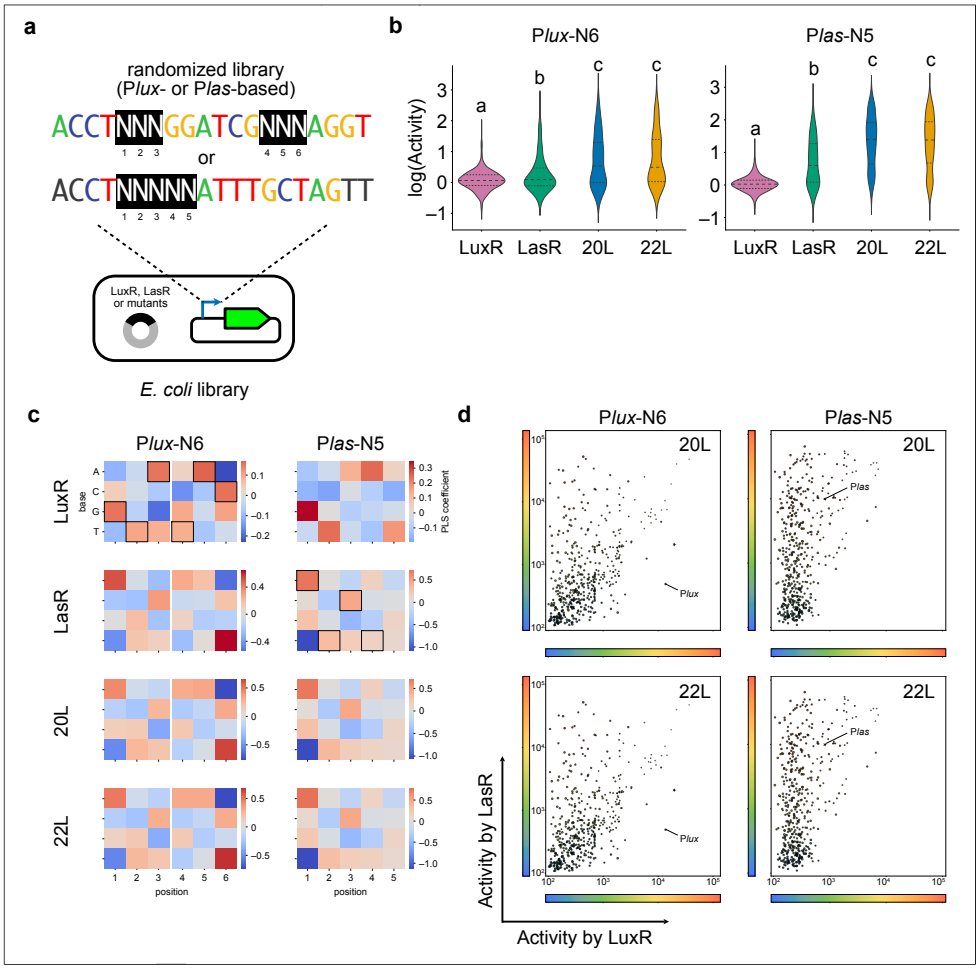


Fig. 4 | Specificity of VAE-designed transcription factors toward randomized promoter libraries

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

372 based on sequence logo analyses[45,46], shown in the corresponding LuxR and
373 LasR heatmaps, respectively. **d,** Four-dimensional activity landscape of promoter
374 responses to wild-type and VAE-designed transcription factors. Promoter
375 activities for wild-type LasR and LuxR are plotted on the horizontal and vertical
376 axes, respectively. Activities for 20L or 22L are encoded by color, and negative-
377 control fluorescence is represented by point size.
378

379 **Discussion**

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

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

403 consequently, biases in the underlying sequence dataset may be propagated into the
404 reconstructed sequences.

405 Beyond classical approaches, emerging protein design technologies also offer
406 alternative routes for engineering of transcription factors. For example, large protein
407 language models, trained on millions of diverse sequences, generate latent spaces that
408 capture broad biochemical trends[47,48]. However, it may not resolve local functional
409 distinctions among closely related regulators such as LuxR and LasR. Interpolation in
410 such global resolution is therefore unlikely to yield meaningful mixtures of parental
411 recognition features without additional constraints or task-specific fine-tuning.
412 Additionally, *de novo* DNA-binding domain design[49] enables the creation of entirely new
413 HTH scaffolds with programmable nucleotide preferences. As *de novo* frameworks
414 continue to improve, the design of hybrid or multi-specific transcription factors may
415 become achievable, providing an exciting future direction for synthetic gene circuit
416 engineering.

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

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

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

442

443 **Methods**

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

445 LuxR-family protein sequences were collected based on the LuxI/R dataset
446 reported by Miranda *et al*[35]. The representative LuxI/R query sequences used in that study,
447 together with the LuxR and LasR sequences (NCBI RefSeq accession ID:
448 WP_213090604.1 and WP_003082999.1, respectively), were added to the dataset and
449 aligned using MAFFT[50]. Because all sequences in the dataset are arranged as LuxI (N-
450 terminal) followed by LuxR (C-terminal), the boundary between the LuxI and LuxR
451 regions was manually determined by referencing conserved N-terminal residues among
452 the representative receiver proteins. The LuxI region was then removed to retain only the
453 LuxR-family portion. Sequences that became excessively short after gap removal (shorter
454 than 200 aa), contained large internal gaps, or were duplicates were iteratively filtered out
455 using CD-HIT[51]. MAFFT alignment and filtering were repeated for several rounds to
456 obtain a consistent LuxR-family multiple sequence alignment.

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

467 consistency and information density. The resulting dataset contained only the DBD
468 regions of LuxR-family proteins.
469
470 ***Model building and validation using VAE***
471 We implemented a multiple-sequence-alignment variational autoencoder (MSA-
472 VAE) in PyTorch following architectures described by previous studies[29,36]. The model
473 was trained to maximize the evidence lower bound (ELBO), consisting of a binary cross-
474 entropy (BCE) reconstruction term and a Kullback-Leibler divergence (KLD)
475 regularization term:
476 $$\text{Loss} = \beta\ \text{KLD} + \text{BCE},$$
477 where β is a positive weighting coefficient. The KLD term encourages the encoder
478 posterior $q(z|x)$ to remain close to a standard normal prior, thereby promoting a smooth
479 and generalizable latent space. BCE was computed per aligned position using one-hot
480 encoded amino-acid categories[25].
481 To identify suitable model configurations, we performed a two-stage grid search
482 over (i) the dimension of the latent space, (ii) the size of hidden layers, (iii) the β
483 coefficient for the KLD term, and (iv) random seeds. Models were trained with a batch
484 size of 32, a learning rate of 1×10^{-3}, and the Adam optimizer. Early stopping was applied
485 when the validation ELBO failed to improve for 5 consecutive epochs.
486 Model quality was assessed using the pairwise amino-acid frequency correlation
487 between generated sequences and the natural sequences in the training dataset, as
488 previously described[29]. Higher correlation values indicate that the model captures
489 evolutionary constraints, including covariation patterns. Unlike previous studies that
490 sample uniformly across the latent space, we generated sequences from interpolation

regions between two latent vectors corresponding to selected parental sequences[35,36]. Because models with similar hyperparameters can nonetheless yield distinct latent-space geometries, we selected the top 10 models ranked by the evaluation metric and used all of them for sequence generation. For each model, the LuxR and LasR sequences were encoded to obtain their latent means z_{LuxR} and z_{LasR}. A latent interpolation vector was then defined as:

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

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

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

Sampling was performed uniformly within this region by drawing random directions and radii, followed by decoding the resulting latent points into amino acid sequences using the VAE decoder. From this distribution, 20,000 sequences were generated per model. After removing redundant sequences using CD-HIT[51], nine sequences were randomly selected and used for experimental characterization.

Structural prediction and MD simulations

The structural models of the parental transcription factors (LuxR and LasR) in complex with *lux* or *las* promoter sequences were predicted using AlphaFold3[43]. Since LuxR-family regulators function as homodimers, all predictions were performed in the dimeric state with the corresponding DNA duplex included as input.

MD simulations were performed using GROMACS 2023.2 with the AMBER99SB-ILDN force field[52,53] as described previously[54]. Each complex was

514 solvated using the SPC216 water model in a cubic box and neutralized with 150 mM
515 NaCl. Electrostatic and van der Waals interactions were truncated at 1.0 nm, and long-
516 range electrostatics were calculated with the particle mesh Ewald method[55]. Energy
517 minimization was performed using the steepest descent algorithm, and the system was
518 then equilibrated at 300 K and 1 bar for 100 ps each in two steps: first under constant
519 volume (NVT), then under constant pressure (NPT). Temperature and pressure were
520 controlled using the V-rescale thermostat and Parrinello-Rahman barostat, respectively[56].
521 The LINCS algorithm was used to constrain bonds involving hydrogen atoms[57]. The
522 integration time step was set as 2 fs. Production runs of 100 ns were conducted under the
523 NPT ensemble without restraints, with three independent replicates for each complex.
524 The RMSD and hydrogen bonds profiles were generated by GROMACS modules. Images
525 were generated by PyMOL (The PyMOL Molecular Graphics System, Version 3.1.0
526 Schrödinger, LLC).
527
528 ***Construction of plasmids***
529 An expression vector for LuxR DNA-binding domain (DBD) was constructed
530 from Ptet-LuxR on pSB6A1[59] by inverse PCR. An expression vector for the LasR DBD
531 was derived from a LasR construct on a pBR-origin, ampicillin-resistant vector under the
532 control of a constitutive promoter (BioBrick parts). The LasR expression cassette was
533 subsequently inserted into pTrc99A vector by inverse PCR. For VAE-designed proteins,
534 coding sequences were selected at random from the model-generated sequence pool,
535 codon usage of each sequence is optimized, and a ribosome-binding site was individually
536 designed using the RBS Calculator (v2.1)[60] to achieve comparable predicted translation
537 initiation rates across variants. Resultant sequences were synthesized as an

538 oligonucleotide library (Integrated DNA Technologies, United States). Each sequence
539 was PCR-amplified, purified from PAGE gel, and inserted between the NheI and HindIII
540 sites of pTrc99A-DsRedExpress2. Promoter-reporter plasmids were constructed by
541 cloning GFP between the BamHI and PstI sites of the pSB3K3 backbone, followed by
542 inserting either the *lux* or *las* promoter (iGEM Parts Registry: BBa_R0062 or
543 BBa_K091117, respectively) between the EcoRI and BamHI sites to generate P*lux*-GFP
544 and P*las*-GFP reporters. Randomized promoter libraries were generated analogously
545 using synthesized oligonucleotides in which selected nucleotide positions of the *lux* or
546 *las* promoter were randomized (FASMAC Co., Ltd.), followed by PCR amplification and
547 insertion into the same pSB3K3-GFP backbone using the corresponding restriction sites.
548 For functional assays, each expression plasmid (or the empty vector pET16b(+)) and a
549 reporter plasmid were co-transformed into *Escherichia coli* DH5α.
550
551 ***In vivo assays of transcription activation abilities***
552 For individual reporter assays, *E. coli* DH5α cells carrying an expression plasmid
553 and either the reporter plasmid were grown overnight in LB medium supplemented with
554 100 μg/mL of carbenicillin and 25 μg/mL of kanamycin. Overnight cultures were diluted
555 into fresh LB medium and further grown 1 h at 37°C with shaking. After dilution of flesh
556 culture, protein expression was induced by adding IPTG at a final concentration of 1 mM
557 or 0.33 mM, followed by incubation for 6 h at 37°C. GFP fluorescence and cell density
558 were quantified using either a plate reader (Synergy, Agilent BioTek) or a flow cytometer
559 (CytoFLEX, Beckman Coulter).
560 For MPRAs, *E. coli* DH5α cells were similarly co-transformed with a TF
561 expression plasmid and a reporter plasmid. For pooled TF library MPRA, the pooled

562 VAE-designed TF plasmid library was transformed into cells harboring either P*lux* or
563 P*las* reporter plasmid. For the randomized promoter-library MPRA, the promoter-GFP
564 library was transformed into cells harboring an individual TF expression plasmid (LuxR,
565 LasR, 20L, or 22L). In both cases, transformant colonies were collected from agar plates
566 and were grown overnight under the same conditions described above. Overnight cultures
567 were diluted into fresh LB medium and further grown 1 h at 37°C with shaking. After
568 dilution of flesh culture, protein expression was induced by adding IPTG at a final
569 concentration of 1 mM, followed by incubation for 10 h at 15°C and further 7.5 h at 37°C.
570 After induction, cells were washed, resuspended in PBS, and subjected to fluorescence-
571 activated cell sorting (FACS). For each assay, cells were sorted into multiple predefined
572 fluorescence bins based on GFP intensity, corresponding to different promoter-activation
573 levels (**Supplementary Fig. S2 and S6**). Cells collected from each bin were pelleted and
574 processed independently. The variable region derived from plasmid DNA (TF coding
575 sequence for the TF library; promoter insert for the promoter library) was amplified by
576 PCR using primers compatible with adapters. Amplicons from each bin were indexed,
577 pooled, and sequenced using an Illumina NextSeq 1000 Sequencing System, with
578 sequencing services provided by Bioengineering Lab. Co., Ltd. Sequence counts were
579 obtained by exact-match parsing, and enrichment-based activity scores were calculated
580 for each variant by comparing its representation across fluorescence bins, with counts
581 normalized by the fraction of cells collected in each bin.
582
583 ***Statistical analysis of promoter library activities***
584 To assess whether activity distributions differed significantly among
585 transcription factors, both parametric and nonparametric statistical tests were performed.

One-way analysis of variance (ANOVA) was applied to raw activity values, followed by the Kruskal-Wallis test as a distribution-free alternative. Post hoc pairwise comparisons were conducted using Dunn's test with Bonferroni correction. Homogeneity of variances was assessed using Levene's test.

For constructing feature matrix X, each sequence was encoded using one-hot encoding at the variable positions. For each variable position, four binary features corresponding to nucleotides A, C, G, and T were generated. The response matrix Y consisted of activity values for LuxR, LasR, 20L, and 22L. Raw activity values were transformed to reduce skewness. The transformed values were then standardized (zero mean and unit variance) independently for each transcription factor prior to multivariate analysis. Partial least squares (PLS) regression was performed to model the relationship between sequence features and transcriptional activity. PLS was chosen to extract latent components of the sequence feature space that maximize covariance with the activity measurements. The one-hot encoded sequence matrix X was used as the predictor matrix, and the standardized activity matrix Y was used as the response matrix. Two PLS components were retained for analysis. Model fitting was performed using the PLSRegression implementation from scikit-learn. PLS regression coefficients were extracted to quantify the contribution of each nucleotide at each variable position to transcriptional activity. Coefficients were reshaped into matrices representing nucleotide preferences (A, C, G, T) at each variable position.

Data availability

Data are available from the corresponding author upon request.

Acknowledgments

We thank R. Yoshida 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), particularly H. Shimokawa, Y. Takizawa, S. Nishimura and A. Sasaki. 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.O. and K.U., and by NINS Astrobiology Center program research (Grant numbers AB041011, AB0602, and AB0712) to D.K.

CRediT Author contributions

S.O.: conceptualization, data curation, investigation, methodology, funding acquisition, software. A.M.: formal analysis, investigation, methodology, software, visualization, writing – original draft, review and editing. M.A.: formal analysis, investigation, writing – 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.

633 **References**

634 1. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology.
635 *Nat. Rev. Microbiol.* **12**, 381–390 (2014).
636 2. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional
637 regulators. *Nature* **403**, 335–338 (2000).
638 3. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch
639 in Escherichia coli. *Nature* **403**, 339–342 (2000).
640 4. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in
641 Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.
642 *Nucleic Acids Res.* **25**, 1203–1210 (1997).
643 5. Zhan, J. *et al.* Develop reusable and combinable designs for transcriptional logic
644 gates. *Mol. Syst. Biol.* **6**, 388 (2010).
645 6. Slusarczyk, A. L., Lin, A. & Weiss, R. Foundations for the design and
646 implementation of synthetic genetic circuits. *Nat. Rev. Genet.* **13**, 406–420 (2012).
647 7. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science* **352**, aac7341
648 (2016).
649 8. Chen, Y. *et al.* Tuning the dynamic range of bacterial promoters regulated by ligand-
650 inducible transcription factors. *Nat. Commun.* **9**, 64 (2018).
651 9. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A.
652 Escherichia coli "Marionette" strains with 12 highly optimized small-molecule
653 sensors. *Nat. Chem. Biol.* **15**, 196–204 (2019).
654 10. Du, P. *et al.* De novo design of an intercellular signaling toolbox for multi-channel
655 cell-cell communication and biological computation. *Nat. Commun.* **11**, 4226 (2020).
656 11. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand

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

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

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

662 14. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-input
663 RNAi-based logic circuit for identification of specific cancer cells. *Science* **333**,
664 1307–1311 (2011).

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

667 16. Samaniego, C. C., Moorman, A., Giordano, G. & Franco, E. Signaling-based neural
668 networks for cellular computation. in *2021 American Control Conference (ACC)*
669 1883–1890 (IEEE, 2021).

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

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

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

676 20. Chen, Z. *et al.* A synthetic protein-level neural network in mammalian cells. *Science*
677 **386**, 1243–1250 (2024).

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

681 22. Bashton, M. & Chothia, C. The generation of new protein functions by the
682 combination of domains. *Structure* **15**, 85–99 (2007).
683 23. Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: in vitro
684 recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10747–
685 10751 (1994).
686 24. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral
687 sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141
688 (2021).
689 25. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]*
690 (2013).
691 26. Strokach, A. & Kim, P. M. Deep generative modeling for protein design. *Curr. Opin.*
692 *Struct. Biol.* **72**, 226–236 (2022).
693 27. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic
694 variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
695 28. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein
696 folds using variational autoencoders. *Sci. Rep.* **8**, 16189 (2018).
697 29. Hawkins-Hooker, A. *et al.* Generating functional protein variants with variational
698 autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
699 30. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: Generative modeling of protein
700 structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**, e1010271
701 (2022).
702 31. Sevgen, E. *et al.* ProT-VAE: Protein Transformer Variational AutoEncoder for
703 functional protein design. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2408737122 (2025).
704 32. Boob, A. G. *et al.* Design of diverse, functional mitochondrial targeting sequences

705 across eukaryotic organisms using variational autoencoder. *Nat. Commun.* **16**, 4151
706 (2025).
707 33. Choi, S. H. & Greenberg, E. P. The C-terminal region of the Vibrio fischeri LuxR
708 protein contains an inducer-independent lux gene activating domain. *Proc. Natl.*
709 *Acad. Sci. U. S. A.* **88**, 11115–11119 (1991).
710 34. Kiratisin, P., Tucker, K. D. & Passador, L. LasR, a transcriptional activator of
711 Pseudomonas aeruginosa virulence genes, functions as a multimer. *J. Bacteriol.* **184**,
712 4912–4919 (2002).
713 35. Wellington Miranda, S. *et al.* A covariation analysis reveals elements of selectivity
714 in quorum sensing systems. *Elife* **10**, (2021).
715 36. Kohyama, S., Frohn, B. P., Babl, L. & Schwille, P. Machine learning-aided design
716 and screening of an emergent protein function in synthetic cells. *Nat. Commun.* **15**,
717 2010 (2024).
718 37. Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G. & Blanquet, S. Extent of N-
719 terminal methionine excision from Escherichia coli proteins is governed by the side-
720 chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 8247–
721 8251 (1989).
722 38. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria.
723 *Science* **254**, 1374–1377 (1991).
724 39. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays.
725 *Genomics* **106**, 159–164 (2015).
726 40. White, M. A. Understanding how cis-regulatory function is encoded in DNA
727 sequence using massively parallel reporter assays and designed sequences.
728 *Genomics* **106**, 165–170 (2015).

729 41. Kinney, J. B., Murugan, A., Callan, C. G., Jr & Cox, E. C. Using deep sequencing
730 to characterize the biophysical mechanism of a transcriptional regulatory sequence.
731 *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9158–9163 (2010).
732 42. Barnes, S. L., Belliveau, N. M., Ireland, W. T., Kinney, J. B. & Phillips, R. Mapping
733 DNA sequence to transcription factor binding energy in vivo. *PLoS Comput. Biol.*
734 **15**, e1006226 (2019).
735 43. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
736 AlphaFold 3. *Nature* **630**, 493–500 (2024).
737 44. Corbella, M. *et al.* The N-terminal helix-turn-helix motif of transcription factors
738 MarA and Rob drives DNA recognition. *J. Phys. Chem. B* **125**, 6791–6806 (2021).
739 45. Antunes, L. C. M., Ferreira, R. B. R., Lostroh, C. P. & Greenberg, E. P. A mutational
740 analysis defines Vibrio fischeri LuxR binding sites. *J. Bacteriol.* **190**, 4392–4397
741 (2008).
742 46. Gilbert, K. B., Kim, T. H., Gupta, R., Greenberg, E. P. & Schuster, M. Global
743 position analysis of the Pseudomonas aeruginosa quorum-sensing transcription
744 factor LasR. *Mol. Microbiol.* **73**, 1072–1085 (2009).
745 47. Kyro, G. W., Qiu, T. & Batista, V. S. A model-centric review of deep learning for
746 protein design. *arXiv [cs.LG]* (2025).
747 48. Xiao, Y. *et al.* Protein large language models: A comprehensive Survey. *arXiv [q-*
748 *bio.BM]* (2025) doi:10.48550/arXiv.2502.17504.
749 49. Glasscock, C. J. *et al.* Computational design of sequence-specific DNA-binding
750 proteins. *Nat. Struct. Mol. Biol.* **32**, 2252–2261 (2025).
751 50. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid
752 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**,

753 3059–3066 (2002).
754 51. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets
755 of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
756 52. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through
757 multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25
758 (2015).
759 53. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically
760 disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
761 54. Shigematsu, M. *et al.* RNase 2/EDN cleaves tRNA anticodon loops to generate
762 immunoactive RNAs. *bioRxiv* 815803 (2025) doi:10.1101/815803.
763 55. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for
764 Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
765 56. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new
766 molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
767 57. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear
768 constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
769 58. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A. & Moreno, E.
770 Gmx_MMPBSA: A new tool to perform end-state free energy calculations with
771 GROMACS. *J. Chem. Theory Comput.* **17**, 6281–6291 (2021).
772 59. Sekine, R. *et al.* Tunable synthetic phenotypic diversification on Waddington's
773 landscape through autonomous signaling. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17969–
774 17973 (2011).
775 60. Salis, H. M. The ribosome binding site calculator. *Methods Enzymol.* **498**, 19–42
776 (2011).

Fig. 1

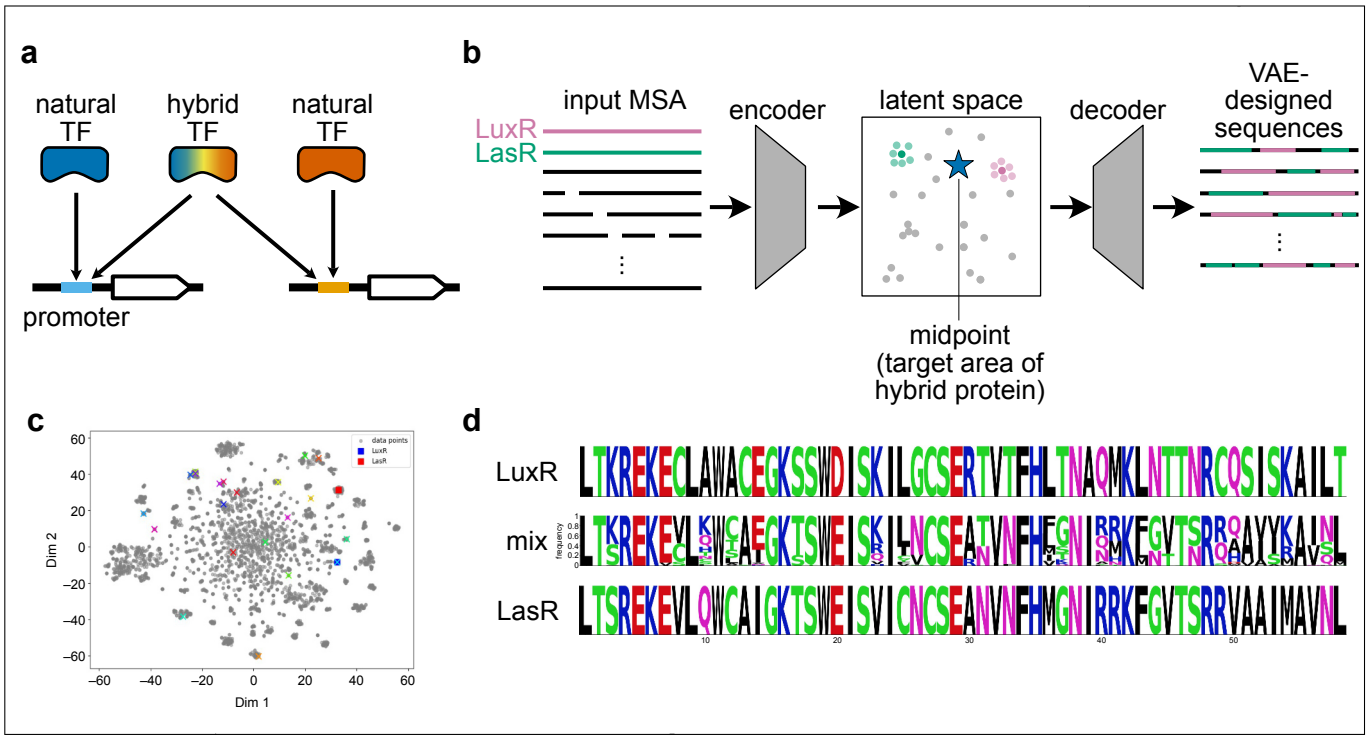

a
natural TF
hybrid TF
natural TF
promoter
b
input MSA
LuxR
LasR
encoder
latent space
decoder
VAE-designed sequences
midpoint (target area of hybrid protein)
c
data points
LuxR
LasR
Dim 1
Dim 2
d
LuxR
mix
LasR

Fig. 2

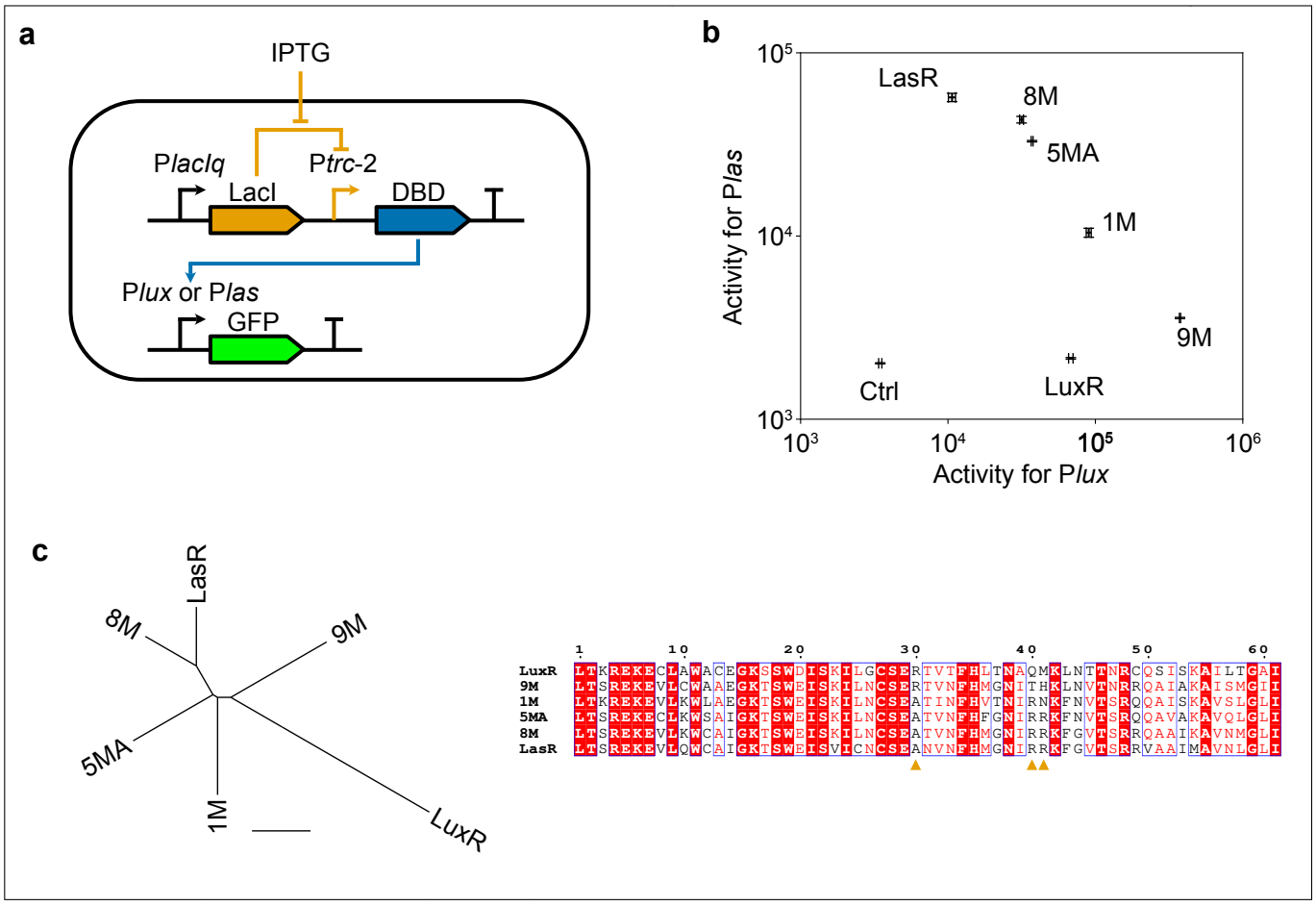

a
IPTG
PlacIq
PtrC-2
LacI
DBD
Plux or Plas
GFP
b
LasR
8M
5MA
1M
9M
Ctrl
LuxR
Activity for Plas
Activity for Plux
c
LasR
8M
9M
5MA
1M
LuxR
LuxR
9M
1M
5MA
8M
LasR

Fig. 3

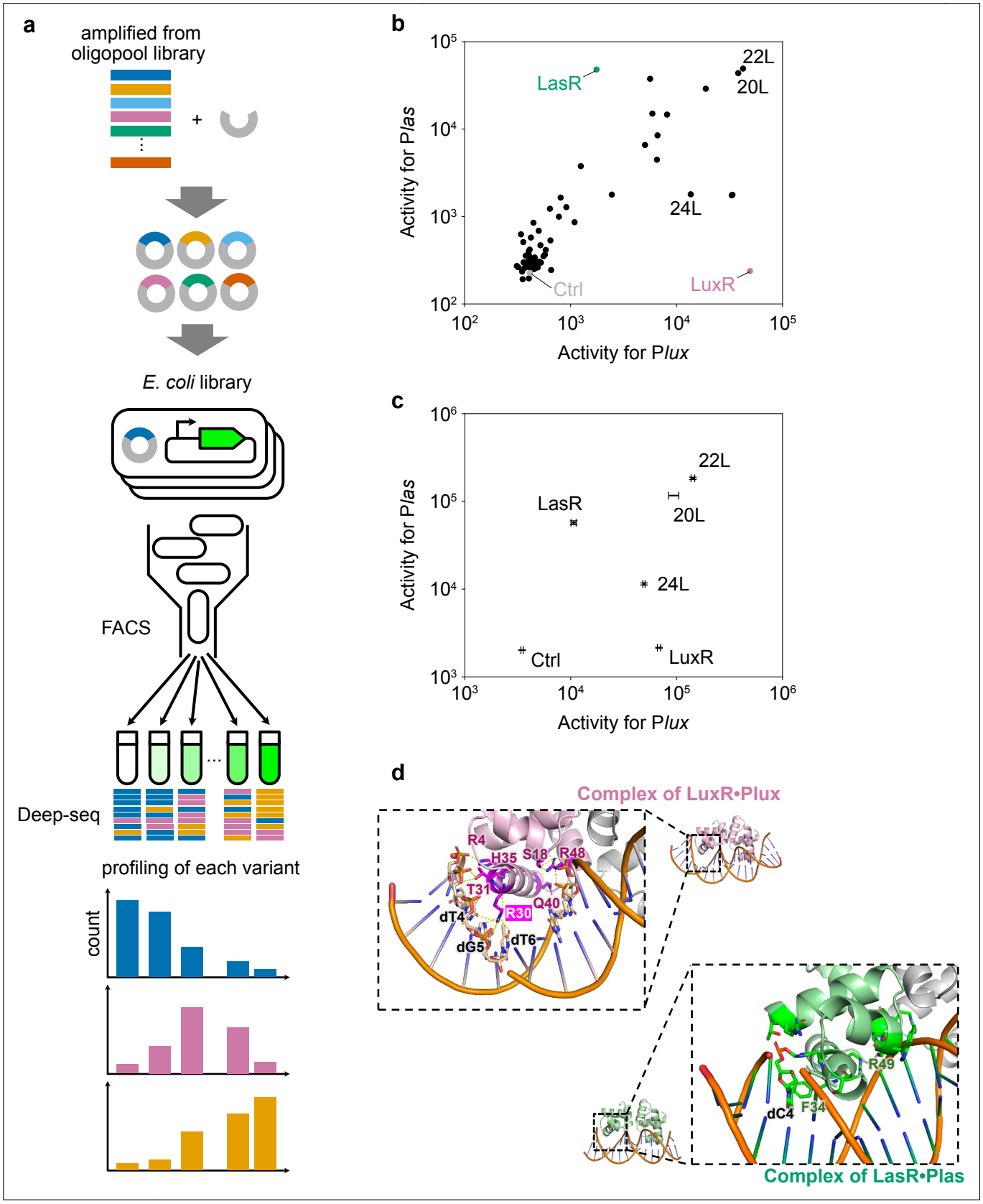
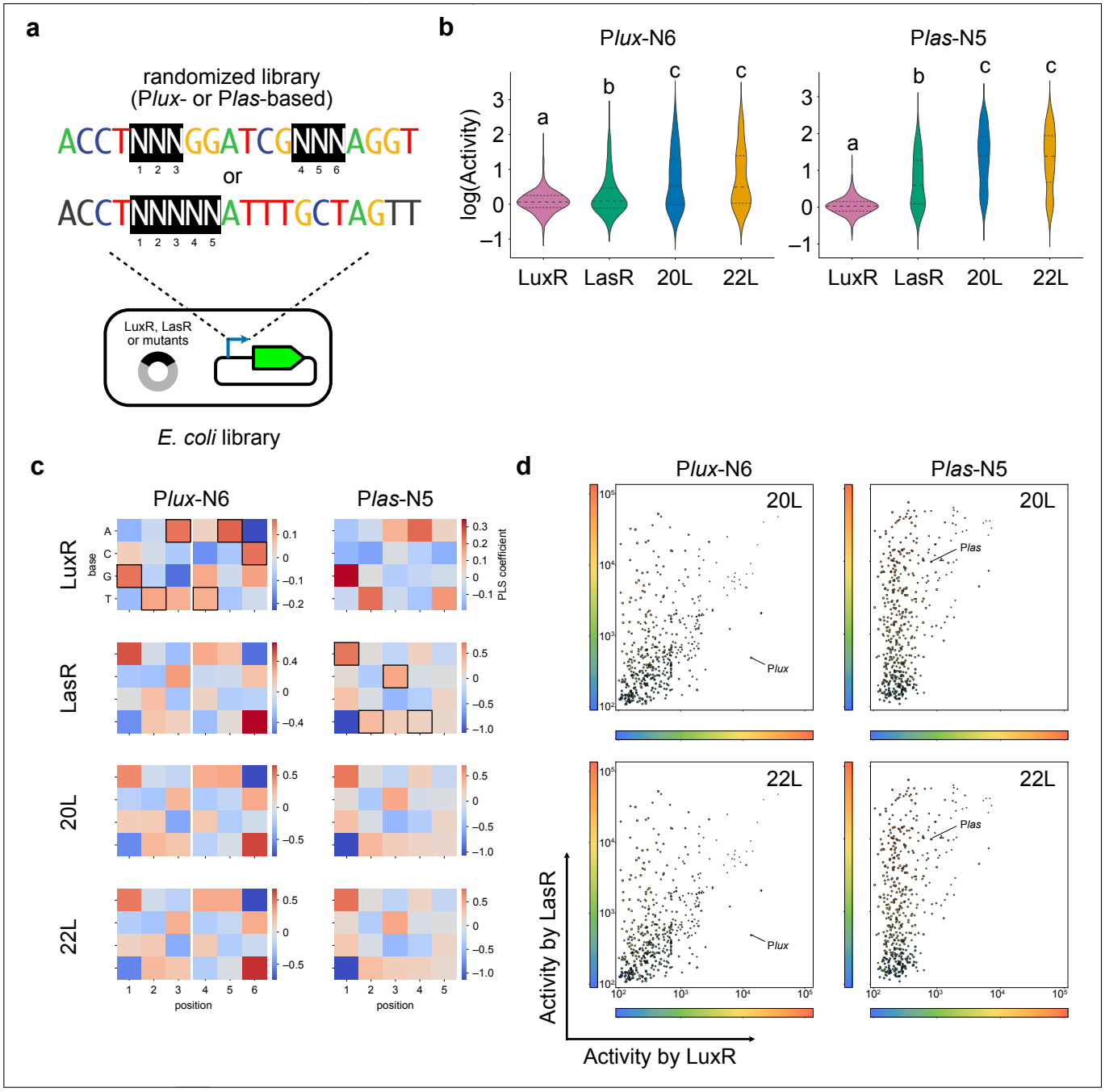

a
amplified from oligopool library
+
E. coli library
FACS
Deep-seq
profiling of each variant
count
b
Activity for Plas
Activity for Plux
LasR
22L
20L
24L
Ctrl
LuxR
c
Activity for Plas
Activity for Plux
LasR
22L
20L
24L
Ctrl
LuxR
d
Complex of LuxR•Plux
R4
H35
S18
R48
T31
Q40
R30
dT4
dG5
dT6
Complex of LasR•Plas
R49
F34
dC4

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
PLS coefficient
position
d
Plux-N6
Plas-N5
20L
22L
Plux
Plas
Activity by LasR
Activity by LuxR