TITLE

Ultra-high throughput mapping of genetic design space

AUTHORS

Kshitij Rai[1,4,*], Ronan W. O'Connell[1,3,*], Trenton C. Piepergerdes[1,3], Yiduo Wang[1,3], Lucas B. C. Brown[1,4], Kian D. Samra[1], Jack A. Wilson[1], Shujian Lin[1], Thomas H. Zhang[1], Eduardo M. Ramos[1], Andrew Sun[1], Bryce Kille[5], Kristen D. Curry[5], Jason W. Rocks[6], Todd J. Treangen[5], Pankaj Mehta[6,7,8], and Caleb J. Bashor[1,2,§]

AFFILIATIONS

[1]Department of Bioengineering, Rice University, Houston, TX 77030, USA
[2]Department of Biosciences, Rice University, Houston, TX 77030, USA
[3]Graduate Program in Bioengineering, Rice University, Houston, TX 77030, USA
[4]Graduate Program in Systems, Synthetic and Physical Biology, Rice University, Houston, TX 77030, USA
[5]Department of Computer Science, Rice University, Houston, TX 77005, USA
[6]Department of Physics, Boston University, Boston, MA 02215, USA
[7]Biological Design Center, Boston University, Boston, MA 02215, USA
[8]Faculty of Computing and Data Science, Boston University, Boston, MA 02215, USA
[§]Correspondence. Email: caleb.bashor@rice.edu
*These authors contributed equally to this work

ABSTRACT

Massively parallel genetic screens have been used to map sequence-to-function relationships for a variety of genetic elements. However, because these approaches only interrogate short sequences, it remains challenging to perform high throughput (HT) assays on constructs containing combinations of multiple sequence elements arranged across multi-kb length scales. Overcoming this barrier could accelerate synthetic biology; by screening diverse gene circuit designs and learning "composition-to-function" mappings that reveal genetic part composability rules and enable rapid identification of behavior-optimized variants. Here, we introduce CLASSIC, a genetic screening platform that combines long- and short-read next-generation sequencing (NGS) modalities to quantitatively assess pools of constructs of arbitrary length containing diverse part compositions. We show that CLASSIC can measure expression profiles of >10^5 gene circuit designs (from 5-20 kb) in a single experiment in human cells. The resulting datasets can be used to train ML models that accurately predict circuit behavior across expansive circuit design landscapes, revealing part composability rules that govern circuit performance. Our work shows that by expanding the throughput of each design-build-test-learn (DBTL) cycle, CLASSIC enhances the pace and scale of synthetic biology and establishes an experimental basis for data-driven design of complex genetic systems.

65 **MAIN TEXT**

66 Synthetic gene circuits are constructed by assembling DNA-encoded genetic parts into
67 multi-gene programs that perform computational tasks in living cells[1-3]. Over the past two dec-
68 ades, gene circuits have emerged as important models for understanding native gene regulation[4]
69 and have been used to create powerful biotechnologies by enabling user-defined control over
70 cellular behavior[5-7]. Despite this progress, the design of quantitatively precise circuit behavior re-
71 mains challenging. Regulatory interactions within a circuit must be carefully tuned, often through
72 multiple iterative DBTL cycles, before part compositions that support a desired circuit behavior
73 are identified[8]. Additionally, since genetic parts must work in close physical proximity to one an-
74 other, as well as within a crowded intracellular environment[9], incidental molecular coupling can
75 occur between parts and with host cell regulatory machinery[10-12]. These context-dependent inter-
76 actions can confound the ability to design circuits using biophysical and mechanistic modeling
77 frameworks, further extending the number of DBTL cycles required to achieve a target behavior[13].

78 One potential strategy for increasing the pace of gene circuit engineering is to expand the
79 number of circuits tested in each cycle by performing HT functional screens on pooled circuit
80 libraries. By measuring large collections of circuits in a single experiment, such an approach could
81 be used for rapidly profiling complex circuit design spaces to identify part compositions with de-
82 sired quantitative behaviors. In contrast to traditional circuit engineering approaches, collecting
83 large amounts of circuit activity data could also facilitate the training of ML/AI models that are
84 capable of inferring context-specific part function and more accurately predicting the behavior of
85 untested circuit designs. HT screening approaches that utilize short-read NGS as a readout[14-18]
86 have been used to generate detailed sequence-to-function mappings for multiple genetic part
87 classes, including promoters[19-23], terminators and 3' UTRs[24,25], transcription factors (TFs)[26,27], nu-
88 cleic acid switches[28], and receptors[29]. However, high-depth functional profiling of libraries of DNA
89 constructs long enough (>1 kb) to encode entire circuits can be costly and technically challenging.
90 Methods have emerged that permit multiplexed analysis of long constructs by appending short

91 barcode index sequences that facilitate amplicon-based readout by short-read NGS[30]. This ap-
92 proach has been used in array formats in which each construct is assembled and barcoded sep-
93 arately, or through nested assembly schemes[30,31], limiting library design flexibility and assay
94 throughput. More recently, random barcoding and indexing via long-read single-molecule real-
95 time (SMRT) sequencing (PacBio) has been used to characterize libraries of variants ranging
96 from 10^2 to 10^4 constructs, mostly for individual ORFs of up to 6 kb in length[32-34].

97 To perform indexed multiplexing for construct libraries containing combinations of genetic
98 parts at larger length scales, we devised an approach that combines long-read nanopore and
99 short-read Illumina NGS to analyze construct libraries generated via pooled part assembly (**Fig.**
100 **1A**). Our assembly scheme incorporates structured barcodes that can be used to create a con-
101 struct composition-to-barcode index by nanopore sequencing. The construct libraries can then be
102 introduced into cells, binned based on functional phenotype, and analyzed by short-read (Illumina)
103 amplicon sequencing to produce a barcode-to-phenotype index. A map that matches construct
104 composition to phenotype can be made by comparing the two indices. Using this technique, which
105 we refer to as CLASSIC (combining long- and short-range sequencing to investigate genetic com-
106 plexity), it is possible to obtain high-depth phenotypic expression data for large libraries (>10^5) of
107 genetic part compositions of arbitrary length using standard phenotypic selection or flow sorting
108 experiments.

109 To configure CLASSIC to quantitatively profile libraries of constructs containing diverse
110 genetic part combinations in human cells (**Fig. 1B**), we designed a custom hierarchical golden-
111 gate[35] cloning scheme in which library diversity is programmed through a series of pooled DNA
112 assembly steps. Diversified pools of 5', coding, and 3' gene elements are first generated through
113 the assembly of input part fragments (level 0 to 1), and subsequently combined to create single-
114 gene expression unit (EU) pools (level 1 to 2). The EU pools are then combined with a BFP EU
115 plasmid pool that expresses a transcript containing semi-degenerate barcode sequences in the
116 3' UTR, yielding barcoded multi-EU circuit libraries (level 2 to 3). Following nanopore sequencing

117 and data analysis (see **Methods**), the libraries are genomically integrated at single copy into a
118 HEK293T cell line harboring a custom "landing pad" cassette (HEK-LP)[36,37] (see **Methods**). Li-
119 brary-integrated cells are then flow-sorted based on circuit expression output, followed by Illumina
120 NGS to quantitate bin distributions of circuit-associated barcodes[14,15] (see **Methods**). We vali-
121 dated this workflow with a small (n=384) expression unit (EU) library containing shuffled regula-
122 tory elements (promoter[38,39], Kozak sequence[40], and terminator[24,41,42]), demonstrating that our as-
123 sembly scheme yields complete, part-balanced library coverage and proportional assortment of
124 unique barcodes. HEK-LP library integration and flow-seq measurements were compared with
125 flow cytometry measurements made with a set of randomly isolated, clonally expanded variants.
126 These "ground truth" measurements showed excellent agreement with corresponding CLASSIC-
127 derived values (mean absolute error [*MAE*]=0.07), and demonstrated that measurement precision
128 is enhanced by mapping multiple unique barcodes to each composition. We also collected repli-
129 cate libraries to demonstrate that CLASSIC is technically (r^2=0.99) and biologically (r^2=0.97) re-
130 producible. Finally, we showed that the data can be used to quantify relative part contributions to
131 expression level, as well as identify unexpected interference between parts.

132 To evaluate whether CLASSIC can be used to experimentally profile a complex design
133 space for a multi-gene circuit, we constructed a library of small-molecule inducible, single-input
134 transcriptional switches comprising >10^5 potential unique compositions. Circuits in this library con-
135 sist of two EUs (**Fig. 2A**): one encoding a constitutively expressed synthetic Cys2-His2 zinc-finger
136 (ZF)-based transcription factor (synTF)[43-45] with appended transcriptional activation domains
137 (TAs), and the other encoding a reporter gene harboring cognate synTF binding motifs (BMs)
138 located upstream of a minimal promoter driving eGFP expression[43]. Transcriptional induction oc-
139 curs upon addition of the small molecule 4-hydroxytamoxifen (4-OHT), which binds to a mutant
140 version of the human estrogen receptor (ERT2)[46] appended to the synTF, facilitating its translo-
141 cation from the cytoplasm into the nucleus to activate reporter transcription. As we and others
142 have demonstrated[43,47-49], identifying designs that are optimized for high fold-change (HFC) (**Fig.**

143 **2A, right**) expression can be challenging in mammalian systems; to ensure robust expression of
144 a transgene exclusively in the presence of inducer, basal and induced expression levels must be
145 respectively minimized and maximized through tuning of both the transcriptional regulatory fea-
146 tures of the locus and the molecular properties of the synTF.

147 To create a library that diversifies the design of this circuit architecture, we identified 10
148 genetic part categories that we hypothesized could play a role in circuit tuning to achieve HFC
149 behavior. This included 4 different transcriptional activation domains (TAs)[50-52], 3 ZF affinities[43],
150 and a set of 4 intrinsically disordered protein (IDP) domains[53-55] that have been shown to facilitate
151 liquid-liquid phase condensation of nuclear-localized TFs[53,54] (**Fig. 2B, left**). We also varied tran-
152 scriptional regulatory features of the circuit, including 4 promoters and 4 terminators in the synTF
153 coding EU, as well as 3 core promoters[56-58] and different numbers of BMs (2, 4, 8, or 12) in the
154 reporter EU (**Fig. 2B, left**). Additionally, we varied the spacing between the EUs (0, 250, or 500
155 bp) along with their 5'-to-3' orientation (**Fig. 2B, left**) to yield an overall circuit design space of
156 165,888 compositions (**Fig. 2B, right**). This library was constructed in 3 steps by first assembling
157 protein domain parts to create a level 1 synTF ORF pool, then conducting parallel assemblies to
158 generate level 2 pools of synTF coding and reporter EUs, and finally combining EU and barcode
159 pools into the level 3 destination vector (**Fig. 2B right**).

160 Nanopore sequencing of the pooled library yielded barcode assignments for 95.3% of total
161 compositions (**Fig. 2C**), with a mean of 8.4 barcodes for each circuit (**Fig. 2C bottom left**). We
162 integrated the library into HEK-LP cells and sorted un-induced and 4-OHT-induced populations
163 (8.6 and 15.7 million cells, respectively) separately into 8 bins based on eGFP fluorescence (**Fig.
164 2D left**). Following analysis by Illumina NGS, we matched a total of 121,292 (73% of design
165 space) compositions to barcodes that were detected in the sorted populations (mean barcodes
166 circuit^{-1}=2.25) (**Fig. 2C**). We then used barcode bin distributions to compute basal, induced, and
167 fold-change expression values for each variant. CLASSIC-derived fold-change values demon-
168 strated excellent overall agreement (*MAE*=0.15) (**Fig. 2D top middle**) with those of random

169 isolates (n=40) (**Fig. 2D bottom**), and values for both basal and induced expression showed
170 comparably high percent similarities (**Fig. 2D top right**). These results confirm that CLASSIC can
171 generate quantitatively accurate measurements of libraries large enough to profile complex ge-
172 netic design spaces.

173 Since an incomplete mapping of circuit design space (27% was unmeasured) could limit
174 our ability to identify HFC-optimized compositions, we tested whether a complete mapping could
175 be obtained by training an ML model with our data and then using it to predict the behavior of
176 unmeasured compositions. We trained several model classes and found that multi-layer percep-
177 tron (MLP) (90/10 train/validation split, with high-quality test set r^2=0.86 for basal 0.90 for induced
178 data) and random forest (0.84 and 0.86) models significantly out-performed biophysically-con-
179 sistent (0.46 and 0.27) and unconstrained linear regression models (0.53 and 0.64), indicating
180 that the complex, non-linear interactions between genetic parts can be more accurately captured
181 by advanced model architectures. We selected the MLP to move forward with (**Fig. 3A**) due to its
182 predictive accuracy and potential for extensibility (**Fig. 3A)** (see **Methods**). We used the MLP to
183 impute basal and induced eGFP expression values to yield a complete design space mapping of
184 all 165,888 compositions (**Fig. 3B**). We then analyzed the distribution of values across a 2-dimen-
185 sional projection of behavior space (basal vs. induced expression), examining the relative densi-
186 ties of compositions in 3 regions of interest: low basal (<500 AU), high induced (>70,000 AU), and
187 HFC expression (>25x fold-change) (**Fig. 3B, right**). We observed a greater proportion of library
188 members in the low basal than in the high induced region (~4.3x), with ~50% of compositions
189 showing fold-change values of <3x. Only a small fraction of the library (~8%) fell within the HFC
190 region. While a comparison between MLP-modeled and CLASSIC-measured distributions re-
191 vealed similar global features, we observed high absolute error (>2) between the model predic-
192 tions and CLASSIC-measured values at the periphery of behavior space (~3% of variants, includ-
193 ing many in the HFC region), suggesting either CLASSIC measurement errors or poor model
194 prediction in these regions.

195 To experimentally assess if our model can accurately predict behavior across circuit de-
196 sign space, we constructed cell lines harboring individual CLASSIC-measured (n=81) and un-
197 measured (n=15) compositions, including many predicted to be among the highest fold-change
198 compositions in the design space (60-80x) (**Fig. 3C**). Flow cytometry analysis of the cell lines
199 closely corresponded with model predictions for both measured (*MAE*=0.18) and unmeasured
200 (*MAE*=0.19) compositions, confirming the behavior of numerous HFC compositions predicted to
201 be in the 60-80x range, near the upper boundary of behavior space. Cell lines constructed for
202 circuit configurations that correspond to apparent high-error CLASSIC measurements also
203 showed strong overall agreement with the model, confirming that it performs accurate error cor-
204 rection for measurement outliers. These results validate the ability of our CLASSIC-trained ML
205 model to enable gene circuit design space mapping through accurate imputation of inaccurately
206 sampled or unsampled compositions.

207 With a quantitatively accurate mapping of single-input design space in hand, we sought to
208 uncover part composition rules governing circuit behavior. First, we examined relative part usage
209 frequencies in design space regions of interest (**Fig. 3D top**). We observed distinct part usage in
210 categories associated with the synTF protein (TA, IDP, and ZF affinity), as well as the promoter
211 driving synTF expression, the number of BMs, the core promoter, and EU orientation. For exam-
212 ple, VPR, a strong TA, is used extensively in the high induced region, but is nearly absent amongst
213 low basal circuits in favor of a weaker TA, p65, while intermediate-strength VP16 and VP64 are
214 enriched amongst HFC circuits. By the same token, lower- and medium-activity core promoters
215 (miniTK and ybTATA) are excluded from high activity circuits, while the low basal region has few
216 compositions with mCMV (high activity core promoter), and HFC circuits utilize a mix of ybTATA
217 and mCMV. HFC circuits favored compositions with the reporter EU in the 5' position, an orienta-
218 tion that appears to minimize basal circuit expression across design space, potentially by elimi-
219 nating transcriptional crosstalk from the upstream promoter.

220 To gain further insight into the design rules underlying HFC behavior, we used our trained
221 MLP to compute Shapley additive explanation (SHAP) values, a widely-used approach that pro-
222 vides an objective metric of each part category's relative importance in determining target behav-
223 ior (**Fig. 3E**)[59]. SHAP analysis revealed that HFC behavior is primarily influenced by part selection
224 within six categories: TA, IDP, ZF affinity, and synTF expression and core promoters. To investi-
225 gate part usage co-variance, we computed cross-category mutual information (MI) for composi-
226 tions within the HFC region (**Fig. 3F**). This analysis revealed strong coupling among the six SHAP-
227 important categories, indicating that their co-optimization is critical for achieving HFC behavior.
228 To further explore this interdependence, we applied UMAP-assisted K-means clustering[60] on HFC
229 compositions, identifying three clusters (**Fig. 3G**). Interestingly, all three had distinct synTF de-
230 signs, but common locus features. This included features that maximize activation like high BM
231 valency (n=8 or 1), a strong core promoter (mCMV), and terminator preference, as well as those
232 that minimize basal expression such reporter-to-coding orientation, medium synTF expression,
233 and extended spacing (500 bp) between EUs (**Fig. 3G**). The largest cluster (A) featured medium-
234 activity promoters driving synTF expression (hEF1a1 and hPGK) and medium-strength TAs
235 (VP64 and VP16) paired with HNR IDP, and medium-affinity ZFs. Performing a dose-response
236 analysis on a representative composition from this cluster revealed a Hill coefficient (n_H) of 2.13,
237 indicating moderate cooperativity. In contrast to cluster A, compositions in clusters B and C fa-
238 vored strong expression (CMV promoter), with B containing medium-affinity ZFs and a weak TA
239 (p65), and C showing enrichment for weak-affinity ZFs and medium-strength TA (VP16). Taken
240 together, these observations suggest that HFC behavior is achieved via a common locus design
241 that minimizes basal reporter expression, and synTF designs that carefully balance expression
242 level, ZF affinity, and transcriptional activation. These results demonstrate that analysis of CLAS-
243 SIC-enabled design space mappings can provide novel insight into how part composition enables
244 target circuit behavior.

245 While our results establish that accurate design space mappings can be obtained for re-
246 gimes with balanced training-to-prediction ratios, we realized that models trained with CLASSIC
247 data would offer greater utility if they could generalize to a broader range of out-of-sample test
248 cases. To evaluate the model's generalization capacity, we progressively downsampled the train-
249 ing data while performing validation with the held-out test set (**Fig. 3H, left**). Surprisingly, we
250 found that only ~9% of the data was sufficient to achieve an r^2 that was 95% of the maximum
251 (>0.81). We then examined whether the design space mapping could be expanded by incorpo-
252 rating additional genetic part measurement data (**Fig. 3H, right**). To do this, we assembled two
253 libraries, each containing approximately 3.5k members: one incorporating a previously unmeas-
254 ured TA (NFZ)[61] and another with a 10-GS linker replacing the IDP. Following measurement of
255 these libraries, we fine-tuned the existing "base" model with the new data to predict the expanded
256 design space, which now included the new parts (approximately 259k compositions). While the
257 base model alone showed good predictions of new part behavior (r^2 of ~0.82 for basal and 0.87
258 for induced conditions for NFZ) the addition of the new data improved model generalization,
259 achieving 95% of maximum accuracy (r^2 of ~0.88 for basal and 0.92 for induced conditions) with
260 just ~1,000 instances. These results show that our modeling framework is highly data-efficient
261 and can easily and rapidly incorporate new parts and design features through fine-tuning. This
262 suggests its potential, through multiple cycles of library construction and testing, to generalize to
263 design spaces that are much larger than the datasets used for training.

264 We attempted to evaluate these capabilities for design space expansion by mapping a
265 design space for multi-input circuits with two inducible synTFs, one regulated by 4-OHT and the
266 other by Grazoprevir (GZV), a small molecule protease inhibitor that blocks autolytic cleavage of
267 a NS3 protease-synTF fusion[62] (**Fig. 4A**). In this 3-EU architecture, co-regulation of the reporter
268 locus by the 2 synTFs could potentially support Boolean logic, including AND-gate function if the
269 synTFs cooperatively activate the reporter, and OR-gate function if they independently maximize
270 activation. We hypothesized that this architecture could be tuned to achieve these behaviors by

271 diversifying a set of 68 design features, including many that were chosen based on our results
272 from the single-input library, as well as synTF BMs arranged in four distinct patterns, yielding in
273 an overall design space of approximately 3.4 billion compositions (**Fig. 4B**). To create a mapping
274 for this space, we employed a two-step data collection strategy, first measuring a densely sam-
275 pled ”base library” space comprising a 44-feature subset (1.1M compositions), and then obtaining
276 data for an “expansion library” that sparsely sampled the remaining features (**Fig. 4B**). During
277 base library construction, we deliberately constrained diversity to ensure high read depth, obtain-
278 ing barcode indexes for ~7% (~112k compositions) of the 1.1M potential compositions (**Fig. 4C**),
279 with a mean of 3.2 barcodes for each circuit (**Fig. 4C bottom left**). The expansion library was
280 constructed by assembling 24 independent 1,500-member pools, each containing one of the ex-
281 pansion features held constant against a random sampling of all other base library features. Both
282 libraries were integrated separately into HEK-LP cells and treated with 4-OHT, GZV, or both in-
283 ducers. Each treatment was sorted into 8 bins based on eGFP fluorescence and isolates were
284 obtained from the base (n=14) and expansion libraries (n=13). As before, we performed Illumina
285 NGS to obtain expression values from both libraries (~83k and ~45k respectively for base and
286 expansion libraries) (**Fig. 4C, bottom**), once again observing overall agreement between CLAS-
287 SIC and isolate measurements (*MAE*=0.28, 0.23) (**Fig. 4D**).

288 To create a multi-input circuit design space mapping, we first trained a 5-layer MLP (**Fig.**
289 **5A**) with the base library data, yielding a model with strong predictive accuracy (90:10 train/vali-
290 dation split with test set r^2 values of 0.80/0.90/0.86/0.88 for basal/4-OHT/GZV/both inducer pre-
291 dictions). We then extended the model to the full 3.4B-member design space by fine-tuning with
292 expansion library data, resulting in good predictive performance (r^2 values of 0.79/0.84/0.84/0.82)
293 (**Fig. 5B, left**). We calculated similarity of circuit composition behavior across the design space
294 to that of idealized AND- and OR-gates using Kullback-Leibler divergence (see **Methods**). We
295 found that, while only a small region of base library space demonstrated Boolean-like behavior
296 (D_{KL} values of <0.12 for AND, <0.05 for OR), these regions expanded in the complete design

297 space (**Fig. 5B, right**). To validate the model predictions, we constructed 36 individual composi-
298 tions spanning the behavior space, 35 of which were unmeasured, including many predicted to
299 exhibit AND- and OR-like function (**Fig. 5C**). As before, we observed good agreement between
300 measurements of individual compositions and the model (*MAE* = 0.32) (**Fig. 5C**), confirming the
301 effectiveness of our data sampling and training approaches.

302 Similar to the single-input design space, we observed strong polarization in part usage
303 within both AND- and OR-like regions (**Fig. 5D**). While compositions from both regions were en-
304 riched for a high number of BMs (n=8 or 12) and strong core promoters (mCMV or SCP2), AND-
305 like compositions predominantly utilized clustered BM patterns, while OR gates used exclusively
306 interspersed patterns. In both regions, the 4-OHT-inducible synTF (synTF A) was expressed at
307 intermediate levels (hEF1a1) while the GZV-inducible synTF (synTF B) showed stronger expres-
308 sion and greater overall part diversity. SHAP analysis revealed that core promoter and inter-EU
309 spacing were common drivers of both behaviors (**Fig. 5E**), and OR-like behavior was impacted
310 by more features than AND-like behavior, indicating a greater degree of feature interactivity. This
311 was consistent with part usage coupling analysis, which showed a lower degree of interdepend-
312 ence amongst AND-like compositions, with strong coupling between BM pattern, both synTF ex-
313 pression promoters, and synTF A affinity (**Fig. 5F, left**). The higher inter-category coupling in the
314 OR-like region was driven by strong interactions between features on different synTFs (promoters
315 and TAs) (**Fig. 5F, right**), suggesting co-optimization of expression level and activation strength
316 are important for tuning OR-like circuit behavior.

317 Clustering analysis of AND-like compositions yielded 3 major clusters. As with HFC cir-
318 cuits, locus design remained constant across clusters (**Fig. 5G**). The largest subset of AND-like
319 compositions (clusters A and B) featured a low-affinity synTF A binding to BMs near the promoter,
320 while synTF B with a medium-affinity ZF and a weak TA bound distally. This suggests a mecha-
321 nism in which synTF B, separated from the core promoter by ~110 bp, is unable to independently
322 activate transcription but can facilitate activation by reinforcing weakly-binding synTF A,

323 potentially through IDR-mediated interactions (HNR on synTF A, and FusN [cluster A] or HNR
324 [cluster B] on synTF B) (**Fig. 5G, right**). While this "asymmetric" activation mechanism comprises
325 the majority of AND-like composition, we observed a third, minority solution family (14%) that
326 utilizes balanced synTF expression levels and affinities to achieve AND-like behavior. Clustering
327 of OR-like compositions revealed 4 clusters containing designs allowing synTFs A and B to inde-
328 pendently activate expression. This is likely facilitated by interspersed BM binding arrays, which
329 provide both synTFs with similar access to the core promoter (**Fig. 5H**). Medium-affinity synTF
330 binding is a common feature of all 4 clusters, and a lack of IDRs for synTF A suggesting that
331 synTF interaction plays a limited role in reporter activation. In summary, our mapping of multi-
332 input circuit design space allowed for identification of non-intuitive compositional rules, and of-
333 fered insights into the role that different genetic parts play in determining circuit function.

334 Here, we have established the utility of combining long- and short-read NGS modalities to
335 perform massively parallel quantitative profiling of multi-kb length-scale genetic part compositions
336 in human cells. CLASSIC holds potential for exploring the emergence of phenotypic function from
337 genetic module composition of across a range of organizational scales and phylogenetic contexts,
338 including for viruses[63], bacterial operons[64], and chromatin domains[64-66]. Additionally, the flexibility
339 of our assembly workflow makes CLASSIC adaptable to a wide array of optimization tasks and
340 phenotypic screening strategies. As we show in this piece, by enabling HT profiling of diverse
341 genetic part compositions, our approach significantly expands the scope of inquiry for synthetic
342 biology projects, potentially minimizing the number of DBTL cycles and cost required to identify
343 behavior-optimized part compositions for projects that involve exploration of medium- to large-
344 scale design spaces. We envision that CLASSIC could be useful for particularly challenging de-
345 sign tasks in underdetermined systems, wherein iterative rounds of design space expansion can
346 enable convergence to difficult-to-access target behaviors. Though, for smaller-scale projects
347 (<100 constructs) that can leverage well-defined design rules, traditional approaches may be
348 more cost-effective. Furthermore, since CLASSIC leverages extant, broadly-accessible molecular

349 cloning and experimental analysis pipelines, we anticipate it will increase the pace and scale of
350 genetic design for diverse synthetic biology applications across a spectrum of organismal hosts,
351 including the development of bioproduction strains[67] and multigenic cell therapy programs[5], as
352 well as other applications that require probing complex genetic design spaces to achieve multi-
353 dimensional phenotypic optima.

354 Finally, we showed that data acquired using CLASSIC can not only train ML models to
355 accurately make predictions for out-of-sample and edge-case circuit behaviors but, with properly
356 crafted training strategies, to generalize to orders-of-magnitude larger genetic design spaces. Our
357 results demonstrate the utility of using a data-driven approach to solve complex circuit design
358 challenges, showcasing the ability to identify designs that support target circuit behavior, while
359 also revealing underlying design principles from across design space that are non-intuitive and
360 challenging to capture using biophysical modeling alone. While extensive recent work has lever-
361 aged ML approaches to develop sequence-to-function models for various classes of genetic
362 parts[28,68-70], our work serves as a starting point for developing AI-based models of gene circuit
363 function that use part compositions as learned features. In the future, multi-modal integration of
364 CLASSIC data with sequence-to-function data could be used to train high-capacity deep-learning
365 architectures (e.g., transformers) to enable the generative design of circuits to user-specific func-
366 tion. Such approaches could work in a black-box fashion, without the incorporation of regulatory
367 or biophysical priors, or synergistically with existing mechanistic frameworks to create interpreta-
368 ble models that provide deeper insights into genetic design[71].

369

370 **METHODS**

371 ***Plasmid library construction***

372 Circuit libraries were cloned using a custom hierarchical golden gate[35] assembly scheme,
373 which enables rapid, modular cloning of complex gene circuits starting from individual genetic part
374 sequences. Briefly, input DNA fragments are amplified from genomic or commercial DNA sources

375 and cloned into kanamycin resistant (KanR) sub-part entry vectors using BpiI (Thermofisher)
376 (*level 0*). Sequence-verified sub-part plasmids are then used as inputs for assembly into carben-
377 icillin-resistant (AmpR) entry vectors using BsaI-v2 HF (NEB) to yield genetic part plasmids (i.e.,
378 promoters, ORFs, terminators) (*level 1*). EUs are then constructed by assembling promoter, ORF,
379 and terminator part plasmids into a kanamycin-resistant entry vector using Esp3I (Thermofisher)
380 (*level 2*). EUs are subsequently combined into multi-unit arrays by assembling into a carbenicillin
381 – or spectinomycin-resistant (SpectR) destination vectors using BpiI (Thermofisher) (*level 3*).
382 In this system, barcode pools are incorporated into library assemblies at level 2 and 3.
383 Level 2 barcoded EU pools were generated by first constructing a destination vector carrying a
384 ccdB placeholder expression cassette downstream of the BFP stop codon. A semi-degenerate
385 18 bp barcode oligo pool (IDT)[72] was polymerase-extended and cloned in place of the ccdB cas-
386 sette using BsaI-v2 (NEB) in a 20 µL golden gate reaction: 2 min at 37 °C and 5 min at 16 °C for
387 a number of cycles equal to 10x the number of input fragments, followed by a 30 min digestion
388 step at 37 °C and subsequent sequential 15 min denaturation steps at 65 °C and 80 °C. Resulting
389 assemblies were purified using a miniprep column (Epoch Life Sciences), electroporated (BioRad)
390 into 100 µL NEB 10-Beta electrocompetent *E. coli* (NEB), and plated onto a custom 30 in x 24 in
391 LB_{Kan} agar plate, yielding ~120M colonies. Plates were grown at 37 °C for 16 h, at which time
392 colonies were scraped into LB_{Kan} (20 mL), incubated for 30 min at 37 °C, and the plasmid library
393 extracted via miniprep (Qiagen). For large-scale libraries, level 3 barcode pools were generated
394 in order to ensure a unique barcode-to-circuit mapping. To create level 3 barcode pools, a semi-
395 degenerate 18 bp oligo pool (IDT) was polymerase-extended and cloned via golden gate assem-
396 bly into a SpectR destination vector carrying a placeholder ccdB cassette using PaqCI (NEB)
397 following the assembly protocol above. Resulting assemblies were transformed into 100 µL in-
398 house generated chemically competent[73,74] Stbl3 *E. coli* cells and plated onto a 10 cm LB_{Spect} agar
399 plate. Plates were grown at 37 °C for 20 h to yield approximately 15,000 colonies, followed by
400 colony scraping and plasmid DNA extraction, as described above.

401 Part pools (levels 1), EU pools (level 2) and circuit libraries (level 3) were constructed by
402 combining input plasmids at 50 fmol for each part category (15 µL total volume) using the above-
403 described cycling protocol. Transformations varied by library size: level 2 and 3 assemblies for
404 the EU library were respectively transformed into 100 µL or 200µL of chemically competent *E.*
405 *coli* (DH5a and Stbl3) and plated on LB_{Kan} or LB_{Carb} agar plates for 12-16 h (37 °C) to yield ~15,000
406 colonies, respectively. Colonies were scraped into 13 mL LB_{Kan} or LB_{Carb} and miniprepped using
407 a Qiagen kit. Level 1 and 2 assemblies for the single-input library (**Fig. 2B**) were transformed
408 using 100 µL and 400 µL of chemically competent *E. coli* (Stbl3) and grown on LB_{Carb} and LB_{Kan}
409 agar plates at 37 °C for 12-16 h to yield ~5,000 and ~20,000 colonies, respectively. Level 3 single-
410 input base library assemblies were purified using a miniprep column (Epoch Life Sciences), elec-
411 troporated (BioRad) into 100 µL 10-beta electrocompetent cells (NEB), and then plated on 10 15
412 cm LB_{Spect} agar plates for 16-20 h (37 °C) to yield ~3M colonies. Each level 3 single-input fine-
413 tuning library assembly was transformed into 500 µL of chemically competent *E. coli* (Stbl3) and
414 grown on LB_{Spect} agar plates at 37 °C for 12-16 h to yield ~40,000 colonies.
415 Level 1 assemblies for the multi-input base library (**Fig. 4B**) were treated identically to
416 those from the single-input library. All level 2 assemblies associated with the base design space
417 were transformed into 100µL of chemically competent *E. coli* (DH5a) and grown on LB_{Kan} agar
418 plates at 37 °C for 12-16 h to yield ~200 colonies. 20 (reporter EUs) or 32 (synTF A or B EUs)
419 colonies from each level 2 library were picked separately into 1 mL of TB_{Kan} and shaken overnight
420 at 37 °C and 900 rpm. 200µL from the 20 or 32 cultures associated with each level 2 library was
421 combined and miniprepped using a Qiagen kit. Separate level 3 base library assemblies were
422 carried out for each of the 6 possible circuit orientations, and purified using a miniprep column
423 (Epoch Life Sciences), electroporated (BioRad) into 15 µL of electrocompetent cells (10-beta,
424 NEB), and then plated on a 15 cm LB_{Spect} agar plates for 16-20 h (37 °C) to yield ~200k colonies.
425 Cells were scraped and miniprepped separately using a Qiagen kit.

426 Level 1 assemblies for the multi-input expansion library were generated separately for
427 each new part (24 in total) by combining balanced pools of level 0 part plasmids while holding the
428 new parts constant (**Fig. 4B**). After transforming, scraping, and miniprepping these new part li-
429 braries as described above, Level 2 assemblies were generated for each new part, transformed
430 into 20 µL of chemically competent *E. coli* (DH5a), and grown on LB_{Kan} agar plates at 37 °C for
431 12-16 h to yield ~50-100 colonies. To sample compositional diversity within each new part library,
432 10 colonies from each plate were picked into 1 mL of TB_{Kan}, shaken overnight at 37 °C and 900
433 rpm, pooled, and miniprepped using the Star Liquid Handler (Hamilton) to generate 72 level 2
434 new part libraries. Each of these libraries were then assembled with 5-member level 2 pools from
435 the base library that corresponded to positions not containing the EUs harboring the new parts,
436 along with the BFP barcode EU pool. These level 3 assemblies were then pooled based on their
437 orientation, column purified, electroporated (BioRad) into 7.5 µL of 10-beta electrocompetent cells
438 (NEB), and then plated on a 10 cm LB_{Spect} agar plates for 16-20 h (37 °C) to yield ~100k colonies.
439 The libraries were scraped and miniprepped separately by Qiagen kit.
440
441 ***Long-read plasmid sequencing***
442 To generate indices linking assembled constructs to their associated DNA barcodes, we
443 used Oxford Nanopore Technology (ONT) long-read sequencing[75,76]. Libraries were prepared for
444 long-read sequencing by digesting 2 µg of a level 3 plasmid library using Esp3I (Thermofisher)
445 and purifying linearized fragments using a 0.5x volume of magnetic beads (Omega Bio-Tek) by
446 volume. Nanopore sequencing adapters were added to the linearized DNA pool using the LSK-
447 112 or LSK-114 ligation sequencing kit (ONT) and sequenced using on a MinION Mk1b device
448 (ONT) equipped with an R10 flow cell (FLO-MIN112) or R10.4.1 flow cell (FLO-MIN114). Base-
449 calling was performed using Guppy or Dorado (ONT, super high accuracy mode) running on a
450 GPU (Nvidia RTX 3090). Composition and barcode assignment was performed using WIMPY
451 (what's in my pot, y'all), a custom Matlab analysis pipeline composed of six discrete modules that

452 converts basecalled Fastq files into a part composition array and barcode sequence for each
453 Nanopore read. Briefly, WIMPY first uses Fastqall (1) to import and consolidate basecalled fastq
454 files, followed by Bowtiles (2) to index the reads to a constant region in the level 3 plasmid back-
455 bone. In this study, the puromycin resistance cassette (PuroR) was designated as the constant
456 region, and after indexing to PuroR, reads were filtered to remove incomplete assemblies and
457 non-library fragments. WIMPY then calls Tilepin (3), which uses localized containment searching
458 (aka tiling) to identify important genetic landmarks, such as ORFs, to define the different sections
459 of the sequence (e.g., reporter array, synthetic transcription factor coding region, or barcoded
460 BFP expression unit). Chophat (4) then truncates the reads into smaller arrays based on the
461 previously identified landmarks to aid later identification of individual parts. WIMPY then calls
462 Viscount (5a) or FASTar (5b). Viscount determines compositions for filtered reads by identifying
463 and assigning genetic parts through a combination of general Smith-Waterman alignments[77] and
464 localized containment searches. This is done by splitting the reference sequence into 6-10bp
465 "tiles" with a 1bp stride, after which nanopore reads are queried for the number of tiles contained
466 for each part reference sequence. FASTar uses the same localized containment searches but
467 indexes the location and number of observed reference sequences, and is used to identify re-
468 peated sequences, such as number of BMs in the single-input library, or BM patterns in the dual-
469 input library. Reads containing >3% of tiles for a reference sequence are assigned while reads
470 with more than one reference assignment are discarded. Barcode sequences for each read are
471 concurrently determined by Barcoat (5c), in which the sequence section downstream of the BFP
472 EU is aligned to a degenerate reference sequence using a custom alignment matrix.
473
474 ***Cell culture***
475 HEK293T cells (ATCC® CRL-11268™) used in this study were cultured under humidity
476 control at 37 °C with 5% CO_2 in media containing Dulbecco's modified Eagle medium (DMEM)
477 with high glucose (Gibco, 12100061) supplemented with 10% Fetal Bovine Serum (FBS;

478 GeminiBio, 900-108), 50 units/ml penicillin, 50 µg/ml streptomycin (Pen Strep; Gibco, 15070063),
479 2 mM L-Alanyl-L-Glutamine (Caisson labs, GLL02), referred to hereafter as complete DMEM.
480 HEK293T-LP cells with and without integrated libraries were maintained in complete DMEM sup-
481 plemented with 50 µg/mL Hygromycin B (Sigma, H3274) and 1 µg/mL Puromycin (Sigma, P8833),
482 respectively.
483
484 ***Single-copy library integration***
485 To establish a landing pad (LP) cell line, low passage HEK293T cells were co-transfected
486 with a dual Cas9 and gRNA expression vector (pX330-T2) targeting the human *AAVS1* locus[78]
487 and a PmeI-linearized (NEB) repair template comprising a constitutively expressed YFP-HygR-
488 expression cassette with a BxB1 *attP* recognition site flanked by 800bp homology arms for the
489 human *AAVS1* locus (pROC079). Genomic integration events were selected using 50 µg/mL Hy-
490 gromycin B, and YFP+ cells were flow sorted to isolate clones (WOLF, NanoCellect). Clones were
491 expanded to 24-well plates and tested for integration competency by integration of an *attB*-con-
492 taining destination vector and subsequent flow cytometry and the presence of the intact LP cas-
493 sette at the *AAVS1* locus by PCR. gDNA was extracted from each clonal population (Omega Bio-
494 Tek) and 50 ng of gDNA was used as the template in a 25 µL reaction along with 0.25 µL of Q5
495 polymerase (NEB), 0.5 µL of 10mM dNTPs, 5 µL of 5X Q5 polymerase buffer, and 1 µL of each
496 1 mM genotyping. Primers were specific to the native AAVS1 locus, Landing Pad, or destination
497 vector. The thermocycling protocol was 2 min at 98 °C, then 30 cycles of 98 °C for 20 s, 65 °C for
498 20 s, 72 °C for 1 min and then a final step of 72 °C for 5 min. We identified a clone (LP79 #36)
499 that demonstrated BxB1-dependent, single copy integration of vectors harboring an *attB* recogni-
500 tion site and used this cell line for all following experiments. We verified single-copy genomic
501 integration of the landing pad cassette by ddPCR using primer sets for the engineered locus (Hy-
502 gro) and a pair of previously validated native genes (Beta Actin and POTP). Briefly, each 25 µL
503 reaction was composed of 12.5µL of EvaGreen Supermix (BioRad), 1.25µL of each primer

504 (diluted to 2μM), 20ng of gDNA, and 0.1 μL of HindIII (NEB). Droplets were generated in the
505 QX600 Droplet Generator (BioRad) by combining 20 μL of each reaction with 70 μL of Droplet
506 Generation Oil (BioRad). PCR was performed on 40 μL of the emulsified reaction using the fol-
507 lowing protocol: 5 min at 95 °C, then 40 cycles of 95 °C for 30 s and 60 °C for 1 min, followed by
508 5 min at 4 °C and 5 min at 90 °C. The PCR results were then analyzed using the QX600 Droplet
509 Reader (BioRad). All resulting concentrations were normalized to the HEK293T Beta Actin con-
510 trol.

511 To integrate individual constructs, 1.25 X 10^5 cells were plated into a 24-well plate one
512 day prior to transfection. The well was co-transfected with 250 ng of a level 3 vector containing a
513 construct of interest and 125 ng of BxB1 expression plasmid (Addgene #51271) using JetPrime
514 (VWR). pCAG-NLS-HA-Bxb1 was a gift from Pawel Pelczar (Addgene plasmid #51271)[79]. Two
515 days after transfection, cells were passaged into complete DMEM containing 1 μg/mL Puromycin
516 and selected for 10 days to achieve homogenous expression. Library integration was performed
517 by co-transfecting HEK293T-LP with 250 ng of plasmid library and 125 ng of BxB1 per 200k cells
518 (384-member library = 1M cells transfected; 166k-member base library = 100M cells transfected;
519 166k-member sub-library = 10 M cells transfected; 3.4B-member base library = 200 M cells trans-
520 fected; 3.4B-member expansion library = 200 M cells transfected) using JetPrime. Media ex-
521 change was performed after 8 h and cells were expanded for an additional 40 h, split into five
522 culture dishes, and grown for 6-8 days under Puromycin selection. Cells were passaged 1:3 and
523 cultured for an additional 5 days under puromycin selection, and then combined in complete
524 DMEM for flow sorting (10M cells/mL).

525 To determine the clonal variability within LP-integrated cell populations, we integrated a
526 constitutively expressed mCherry EU following the protocol described above. We then randomly
527 sorted 23 clones from the population (WOLF, NanoCellect) and measured their mCherry expres-
528 sion. We calculated the MAE for this set (*MAE = 0.08*) and used this value to define an expected
529 error range from clonal heterogeneity (ERCH), which represents the expected variability of cells

530 from an integrated population. For the 384-member library, the expression ERCH was set to 0.08
531 on either side of the expected value. For all libraries constructed as part of the single-inducer
532 circuit class, the fold-change ERCH was calculated to be 0.16 on either side of the expected
533 value. For libraries associated with the multi-input design space, the expression level ERCH was
534 calculated to be 0.16 on either side of the expected value.
535
536 ***Flow sorting***
537 To prepare libraries for flow sorting, cells were lifted using TrypLE (Gibco), washed twice
538 with PBS, and resuspended in complete DMEM (10M cells/mL). For the 384-member library, the
539 mRuby expression distribution was divided into 10 evenly log-spaced bins on a Sony MA900 set
540 to purity mode. The number of cells collected for each bin was proportional to the percentage of
541 the mRuby distribution expressed in that bin, with a target of ~125k cells sorted in the most pop-
542 ulous bin [e.g., 2,252 cells were collected for bin 1 (0.53% of library), 126,562 for bin 4 (29.78%
543 of library)]. For the single-input base and fine-tuning libraries, cells were split and either grown in
544 DMEM containing 1μM 4-OHT (Sigma Aldrich) or without 4-OHT for 72 h. The eGFP expression
545 distributions for both conditions were then divided into 8 evenly log-spaced bins. Cells were sorted
546 into bins in proportion to the relative abundance of the library in each bin, with the most populous
547 bin set to collect 3 M cells and ~0.5 M cells in the single-input and fine-tuning libraries (**Fig. 2D**),
548 respectively. For the multi-input base and expansion libraries, cells were split and either grown in
549 DMEM containing 1μM 4-OHT (Sigma Aldrich), 20μM GZV (MedChemExpress), 1μM 4-OHT and
550 20μM GZV or without inducers for 72 h. eGFP expression distributions for each condition were
551 then divided into 8 evenly log-spaced bins and sorted in proportion to abundance, with the most
552 populous bin set to collect ~1.5 M cells and ~1 M cells for the base and expansion libraries,
553 respectively. For all libraries, the sorted cells were plated into 96-, 24- or 6-well plates to achieve
554 a plating density of 10-30%, grown under puromycin selection for 2 d, passaged and washed with
555 PBS, and grown for an additional 3 d. Cells were lifted using TrypLE and total RNA was extracted

556 using the Takara RNA plus kit following manufacturer's instructions (Takara). 5 µL of total mRNA
557 was converted to total cDNA using random hexamers following manufacturer's instructions
558 (Verso, Life Technologies).
559
560 ***Short-read (Illumina) sequencing***
561 To prepare the sorted libraries for short-read sequencing, barcode regions were PCR am-
562 plified from cDNA in a 15 µL reaction with Phantamax polymerase (Vazyme) and custom bin-
563 specific primer sets (IDT) using the following thermocycling protocol: 2 min at 98 °C, then 12-18
564 cycles (depending on the library) of 98 °C for 20 s, 60 °C for 10 s, 66 °C for 10 s, 72 °C for 20 s
565 and then a final step of 72 °C for 2 min. The resulting amplicon pool was extracted from a 2%
566 agarose gel. A second PCR step, which used 6-10 cycles (library dependent) of the thermocycling
567 protocol for PCR step 1, added Illumina sequencing adapters (i5 and i7) and sample-specific
568 sequencing barcodes using Phantamax and custom primer sets (IDT), and the amplified product
569 was extracted from a 2% agarose gel. Purified amplicons from each bin were pooled at equimolar
570 concentrations and sequenced using an Illumina MiSeq (kit v2, 300 cycles), NovaSeq6000 (Sp
571 v1.5 kit, 300 cycles) and the NextSeq 2000 (P2 kit, 200 cycles) for the 384-, 166k-, and 3.4B-
572 member (both base and expansion) libraries, respectively. Illumina data were analyzed using a
573 custom analysis pipeline (Matlab). Briefly, fastq files are imported and split according to their
574 sample-specific sequencing barcode. Barcode sequencing data is converted to average expres-
575 sion for all compositions/variants as follows: (1) Counts for each barcode in each sample/bin are
576 normalized by the percentage of the library distribution in that bin to account for differences in
577 read depth across bins. (2) A weighted average of number of reads in each bin for each barcode
578 is then calculated, to assign an average expression level (using a bin → expression conversion
579 table) to each barcode. (3) Barcode expression levels are then cross-referenced with the Na-
580 nopore indices to assign average expression values for all barcodes associated with a given var-
581 iant in a n-by-y cell array, where n is the number of library variants and y is the (variable) number

582 of barcodes for each variant. (4) For each variant, a gaussian kernel density estimation is per-
583 formed over the log-normalized expression values from all barcodes associated with each variant,
584 using the Matlab function *ksdensity*. The expression value corresponding to the kernel peak is
585 assigned as the mean expression value for each variant.
586
587 ***Experimental validation of CLASSIC measurements***
588 To assess the accuracy of CLASSIC measurements for both the EU and single-input li-
589 braries, single cells were sorted from the integrated cell populations into 96-well plates (MA900)
590 and clonally expanded from across the mRuby and eGFP expression distributions, respectively.
591 Isolated clonal populations, which are referred to as 'isolates', were expanded to 24-well plates
592 and the fluorescence distribution was measured by flow cytometry (MA900). For the EU library,
593 one 96-well plate of cells was sorted based on mRuby expression. For the single-input base li-
594 brary, two 96-well plates of cells were sorted based on eGFP expression from the uninduced cell
595 library. For the single-input fine-tuning library, one 96-well plate of cells was sorted based on
596 eGFP expression from the uninduced population. For both the multi-input base and expansion
597 libraries, two 96-well plates of cells were sorted based on eGFP expression from the uninduced
598 cell libraries. Following fluorescence measurement under library-specific experimental conditions,
599 gDNA was extracted from each clonal population (Omega Bio-Tek). For the EU and single-input
600 libraries, the genetic variant in each isolate was identified by PCR-based amplification and Na-
601 nopore sequencing of the LP locus using a forward primer in puromycin resistance cassette and
602 a reverse primer in the BFP ORF. A set of 30 different forward primers, each with defined bar-
603 codes, were paired with a common reverse primer to amplify the genetic variants from each iso-
604 late and associate the genetic variant with the isolate during Nanopore sequencing. 50 ng of
605 gDNA was used as the template in a 20 μL reaction along with 10 μL of 2X PhantaMax Master
606 Mix (Vazyme), 1 μL nmol of primer, and brought up to 20 μL in nuclease-free H_2O. The thermo-
607 cycling protocol was 2 min at 98 °C, then 30 cycles of 98 °C for 20 s, 65 °C for 20 s, 72 °C for 5

608 min and then a final step of 72 °C for 10 min. Each PCR reaction was extracted from a 1% agarose
609 gel and column purified using QG buffer (Qiagen), following manufacture instructions. Purified
610 amplicons were pooled and prepared for Nanopore sequencing using LSK-112 or LSK-114 liga-
611 tion sequencing kit (ONT) and sequenced with the R10 (FLO-MIN112) or R10.4.1 (FLO-MIN114)
612 MinION flow cell. Basecalling was performed using Guppy or Dorado (ONT, super high accuracy
613 mode) running on a GPU (Nvidia RTX 3090). Composition and barcode assignment was per-
614 formed using WIMPY, as described above. This process yielded 15 isolates from the EU library,
615 40 isolates from the single-input base library (**Fig. 2D, middle**), and 8 isolates from the single-
616 input fine-tuning. Isolates from the base and expansion 3.4B-member libraries, circuit identity was
617 determined by amplification of the barcode sequence from gDNA, following the PCR protocol
618 described above, extraction from a 2% agarose gel, and Sanger. Using the CLASSIC composi-
619 tion-to-barcode IDs to assign barcodes to circuit variants yielded 14 and 13 mapped isolates from
620 the base and expansion libraries, respectively.
621
622 ***MLP deep neural network regression for circuit behavior prediction***
623 *Model architecture and parameters.* Fully connected multi-layer perceptron (MLP) neural network
624 architectures were constructed for single- and multi-input design spaces. The single-input model
625 had 4 hidden layers with 160/80/40/20 neurons, a 39 x 1 input channel, and 2 output channels
626 for log10 transformed basal and induced expression. The multi-input model consisted of 5 layers
627 160/80/40/20/10 neurons, a 68 x 1 input channel, and 4 output channels. Input compositions were
628 one-hot encoded for each part category and flattened into a 39 x 1 shape for the single-input and
629 68 x 1 shape for the multi-input library. Stochastic gradient descent with momentum (sgdm) was
630 used to optimize the root mean squared error loss function during training (momentum parame-
631 ter=0.9), learning rate of 5×10^{-2}, and a piecewise learn rate scheduler that drops the learning
632 rate by a factor of 0.5 every 10 epochs. The batch size was 256.

633 *Data splits for model training.* High-quality test sets were constructed from the single- and multi-
634 input base library datasets (121,292 and 88,551 measured compositions, respectively), by ran-
635 domly sampling 15% of compositions with >12 and >50 barcode-indexed reads (1,868 composi-
636 tions and 953 compositions respectively). Flow cytometry measurements for isolates were re-
637 moved from training datasets to allow for unbiased comparisons against these measurements
638 later. The remaining data was randomly split 90:10 into training:validation sets containing 103,878
639 compositions and validation sets containing 11,532 compositions for the single-input library, and
640 48,031 and 5,336, respectively, for multi-input library. The validation set was used to monitor
641 training progress, and for early stopping with a patience parameter of 300 iterations. The model
642 with the best validation performance was returned and used to evaluate performance on the high-
643 quality test and isolate sets.
644 *Data downsampling.* For demonstrating the single-input model's tolerance to lower amounts of
645 training data (**Fig. 3F**), compositions were downsampled in increasing order of the number of
646 barcoded reads per composition, retaining compositions from ≥ 1 read to ≥ 40 reads to simulate
647 scenarios in which lower amounts of higher-quality data is collected. The model was re-initialized
648 and trained with a new 90:10 training-validation split from the remaining dataset for each of the
649 40 downsampling iterations, which were all evaluated using the high quality test set mentioned
650 above.
651 *Design space expansion via neural network fine-tuning.* To demonstrate single-input model fine-
652 tuning, an MLP model with the same architecture and hyperparameters was trained on the original
653 library, with two extra input encoding slots to represent new parts that would be added during fine-
654 tuning. For the multi-input library, the base model already contained extra input slots for new
655 parts. A high-quality test set was created, as described above, from the expansion library da-
656 tasets, with an increased barcoded read count threshold of 40 (single-input) or 50 (multi-input).
657 The remaining data was split 90:10 into a training and validation temp sets. These sets were
658 combined with 2,000 compositions from the base training set and 500 compositions from the base

validation set to avoid memory loss during training. The combined validation set (with base and new compositions) was used for early stopping and all hyperparameter optimization as described above, except for the initial learning rate, which was changed to 5 x 10^{-3}. Model performance was evaluated on the high-quality test sets and used to construct individual variants across the behavior space.

Mutual information calculation

For a given pair of features in a region of design space, mutual information was computed using the following formula:

$$\sum_{x,y} p\,(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$$

where x and y are 2 different features of the library, and p(x, y) represents the joint frequency for specific x and y combinations divided by the total number of points in the dataset. p(x) and p(y) represent the fractional frequency of the corresponding parts by themselves.

SHAP & UMAP clustering

For both the single- and multi-input design spaces, SHAP values were used to assess the contribution of each feature in the HFC and AND/OR-like regions of behavior space, respectively. For each model, SHAP values were computed by fitting a SHAP explainer model across 500 randomly sampled training points and then predicting on 200 randomly points sampled from the regions of interest. These values were averaged across expression conditions and within part categories to generate region-specific SHAP scores for each part category.

For the 166k-member library, UMAP dimensional reduction was performed on part composition data from variants in the HFC (>25-fold region) region of the behavior space, using the *run_umap* function in Matlab with a Jaccard distance metric. To identify the optimal number of clusters, a gap test was carried out on the 2D projection using the *clusterevaluation* function in

Matlab. HFC compositions were assigned to clusters using the *kmeans* function in Matlab. Cluster stability was determined by performing 100 independent rounds of UMAP projection and k-means clustering, recording cluster affiliation in each iteration, and computing the cluster stability index (CSI) and adjusted rand index (ARI). We then assigned compositions to clusters the most stable UMAP projection from the 100 runs mentioned above. Clustering for AND-like and OR-like circuits from the multi-input library was performed as described above for the single-input library, including the computation of cluster stability. For display purposes, boundaries have been manually drawn around each cluster.

KL-divergence computation

For multi-input circuits, expression levels were normalized to relative abundances by dividing them by the summed expression across the four induction conditions (no inducer/4-OHT/GZV/both) creating probability distributions suitable for computing Kullback-Leibler divergence (D_{KL}) against "ideal" AND and OR gates, using the following equation:

$$D_{KL} = \sum p(x).\, ln\left(\frac{p(x)}{q(x)}\right)$$

Where *p(x)* represents the normalized distribution across the four conditions, and *q(x)* refers to the probability of the ideal AND [0.01, 0.01, 0.01, 0.97] and OR [0.01, 0.33, 0.33, 0.33] gate across the four induction conditions. Low D_{KL} values indicate high similarity to ideal AND and OR distributions.

ACKNOWLEDGEMENTS

We thank Oleg Igoshin, Ankit Patel, Yashwanth Lagisetty, Satpreet Singh, and members of the Bashor lab for helpful discussions. This work was supported by grants from NIH R01 EB029483 (C.J.B.), NIH R01 EB032272 (C.J.B.), ONR N00014-21-1-4006 (C.J.B.), and funding from the Robert J. Kleberg Jr. and Helen C. Kleberg Foundation (C.J.B.). R.W.O. was supported by a

graduate fellowship from the American Heart Association (917746). B.K. was supported by a NLM Training Program in Biomedical Informatics and Data Science fellowship (T15LM007093-31) and by NIH grant P01-AI15299901. K.D.C. was supported by NSF EF-2126387 and the Ken Kennedy Institute Computational Science & Engineering Recruiting Fellowship. P.M. and J.W.R were supported by NIH R35GM119461 (P.M.)

AUTHOR CONTRIBUTIONS

R.W.O., K.R., and C.J.B. conceived of the study, R.W.O. and K.R. carried out the experiments and developed the analysis software, with assistance from T.C.P., Y.W., L.C.B, K.D.S., J.A.W., S.L., T.H.Z., E.M.R., and A.S. R.W.O., and T.C.P. developed the modular cloning scheme and landing pad cell line, with assistance from S.L. B.K., K.C., and T.J.T. helped develop the barcoding scheme and analysis software. R.W.O., K.R., T.C.P., Y.W., J.W.R., P.M, and C.J.B. analyzed the data. C.J.B. supervised the study. R.W.O., K.R., and C.J.B wrote the manuscript, with input from all authors.

DATA AVAILABILITY

All raw data used in this study is available upon request. Accession codes will be made available prior to publication.

CODE AVAILABILITY

All custom scripts used for Nanopore sequencing data analysis are available on GitHub (github.com/cbashorlab/WIMPY). Code associated with Illumina data analysis and model training are available upon request.

CORRESPONDING AUTHOR

Correspondence and requests for materials should be addressed to Caleb J. Bashor.

735

736 **COMPETING INTERESTS**

737 The authors declare no competing interests.

738

739 **Figure 1. Using CLASSIC to systematically map the design space of complex genetic pro-**
740 **grams.** (**A**) Overview of CLASSIC. Pooled assembly of genetic parts with DNA barcode se-
741 quences yields libraries of barcode-indexed constructs of arbitrary length and complexity. Long-
742 read nanopore sequencing is used to create an index matching construct composition to an as-
743 sociated barcode. In parallel, libraries introduced into cells undergo sorting or selection to bin
744 expression phenotypes. Barcode amplicons generated for each bin are subjected to short-read
745 NGS to quantify expression phenotype, which is then mapped to construct composition via bar-
746 code indexes. (**B**) Application of CLASSIC to profile a synthetic gene circuit design space. Hier-
747 archical golden gate assembly is used to compose libraries of multi-EU circuits with combinatori-
748 ally varied part compositions and circuit designs. Sequence fragment pools for different parts
749 categories from level 0 are combined to yield level 1 pools of promoters, open reading frames
750 (ORF), and terminators (Term.), which are then combined to yield level 2 EUs (square brackets:
751 fragment/part pools). Barcode (BC) pools harbored on BFP EUs are combined with library EU
752 pools to create indexed multi-EU circuit libraries (level 3) that are integrated into HEK293T-LP
753 cells placed at the *AAVS1* locus in chromosome 19 via expression of the BxB1 recombinase (top
754 right). Library analysis by a combination of nanopore and flow-seq yields composition-to-function
755 mapping.

756

757 **Figure 2. Using CLASSIC to quantitatively profile a synthetic gene circuit design land-**
758 **scape.** (**A**) Inducible synTF circuit diagram. Left: The circuit contains two EUs: one codes for the
759 synTF and the other is an eGFP reporter. Without inducer, the synTF localizes to the cytoplasm.
760 Upon addition of 4-OHT (input), the synTF translocates into the nucleus to bind to and activate

761 reporter eGFP expression (output). Right: Expression fold change is the ratio of expression levels
762 in the presence or absence of inducer. (**B**) synTF circuit design space. Left: SynTF diversity (ar-
763 ranged N-to-C term): of 4 TAs, 4 IDPs, and 3 ZF affinity mutants (high: WT, med: ΔC, low: ΔC
764 6x). synTF coding EU diversity: 4 constitutive promoters, 4 terminators. Reporter EU diversity: 4
765 BM number variants, 3 minimal core promoters. This unit contains a constant terminator. Both
766 EUs have 3' spacing sequences of 0bp, 250bp or 500bp downstream. Right: EUs assembled from
767 input parts and combined in 2 different 5'-to-3' orientations, to generate a combinatorial diversity
768 of 165,888 possible circuit variants. (**C**) Balance of circuit library assembly and indexing. Left:
769 data for ~121k out of 166k variants (73%) were recovered by CLASSIC; compositions, light grey;
770 barcodes, dark grey. Right: Nanopore sequencing data of the library were analyzed to assess
771 library composition count (mean and standard deviation [SD] computed on $\log_{10}$ scale) (top), com-
772 position/barcode balance (reads per composition [light grey], unique barcodes per composition
773 [dark grey]; data plotted in rank order of reads per composition) (bottom). (**D**) Circuit library sorting
774 and measurement. Top right: eGFP expression of circuit library in presence (solid green) and
775 absence (dotted green line) of 4-OHT (top, center) is shown, along with boundaries of flow sorting
776 bins (vertical grey dotted lines); grey histogram, empty HEK293T-LP cells. Top middle: Fold-
777 change values for 40 clonal isolates plotted against CLASSIC-derived values. Black dots, isolates
778 displayed at the bottom of the panel; *MAE*, mean average error; grey region, ERCH, error range
779 of clonal heterogeneity AU, arbitrary fluorescence units. Top right: % similarity between basal and
780 induced values for CLASSIC and isolates. Bottom: Flow data from 4 isolates (green dotted line,
781 uninduced; green solid, induced) are shown with corresponding CLASSIC-computed distributions
782 [kernel density (black line) calculated from normalized barcode read count (grey vertical bars)].
783 Parts combinations corresponding to each index are shown.
784
785 **Figure 3. ML-aided mapping of single-input circuit design space reveals gene circuit de-**
786 **sign rules.** (**A**) Left: Multi-layer perceptron (MLP) neural network model of inducible synTF circuit

787 behavior space. Circuit part inputs are one-hot encoded, flattened and passed through a 4-layer
788 MLP with 2 output heads that predict basal and induced expression. Right: ML task description.
789 From the complete design space (gray) of ~166k compositions, data from 121k was collected
790 (green) and used to train the ML model to re-fit circuit behavior for measured compositions and
791 infer unmeasured space (pink). (**B**) Left: MLP predictions on a held-out high-quality test set (see
792 **Methods**) against basal and induced CLASSIC measurements (purple, basal; blue, induced) (r^2,
793 Pearson's r^2). Right: Contour plot of MLP modeled basal and induced expression behavior space
794 (contours, light to dark: 97.5%, 90%, 70%, 50% of total measurements). Regions of interest
795 bounded by dotted lines: low basal (<500 AU), purple arrow; high induction (>70k AU) blue; high
796 fold change (HFC) (>25x, green). Values in the plot indicate the number of compositions in each
797 region. (**C**) Experimentally validating ML prediction of circuit function. Top left: Fold-change values
798 for cell lines from classic measured configurations (green filled) and unmeasured configurations
799 (pink outlined) were plotted against CLASSIC-derived values. Black dots, isolates displayed at
800 the periphery of the panel; *MAE*, mean average error; grey region, ERCH, error range of clonal
801 heterogeneity (see **Methods**); green box, HFC behavior region. Right and bottom: Flow cytometry
802 analysis of cell line behavior (green dotted line, basal; solid green, induced) and associated circuit
803 compositions (below each plot). Horizontal pink bars indicate ML predicted fold change. AU, ar-
804 bitrary fluorescence units. (**D**) Genetic part usage in highlighted regions of behavior space; left to
805 right: low basal, high induced, HFC. Part fold enrichment is calculated by dividing observed part
806 occurrence by expected part occurrence from a balanced library (**E**) Feature importance for all
807 part categories based on absolute SHAP values computed in the HFC region (see **Methods**).
808 Larger SHAP values indicate greater stronger contribution to the model's predictions of HFC be-
809 havior. (**F**) Mutual information between part categories in different regions of behavior space. MI
810 between part categories is denoted by red line thickness. (**G**) Clustering analysis of HFC circuit
811 designs. Part features from HFC designs were subjected to UMAP dimensional reduction followed
812 by K-means clustering with optimal cluster number determined using the gap test (see

813 **Supplementary text**). Top left: UMAP projection of HFC designs. All clusters displayed significant
814 part occurrence similarities in the locus design (top middle), and inter cluster differences in synTF
815 expression level, binding affinity, and activation domain preference (bottom). Right: Mechanism
816 for achieving HFC behavior based on clustering results. Locus features are optimized to reduce
817 basal and maximize induced expression levels, while cluster specific choices in synTF design
818 tend to maximize induced expression. (**H**) Model efficiency and expansion to new features via
819 fine-tuning. Left: The model training set was down-sampled by removing compositions and re-
820 training on smaller percentage of the design space. Middle left: Model performance (Pearson's
821 r^2) on compositions in the high-quality test set, evaluated on basal (purple line) and induced (green
822 line) expression levels as a function of percentage of total compositions removed from the training
823 set. Dark gray box: maximum observed r^2 ± 5%. Dotted line: number of compositions needed to
824 obtain 95% of maximum observed r^2 value. Middle right: Expansion of the design space by fine-
825 tuning the model on data collected from a library of compositions containing a new part feature
826 (NFZ activation domain). Right: Model performance (test set Pearson's r^2) for basal and induced
827 expression as a function of increasing number of compositions used in the training set during fine-
828 tuning. Dark gray box: maximum observed r^2 ± 5%. Dotted line: number of compositions needed
829 to obtain 95% of maximum observed r^2 value
830
831 **Figure 4. ML guided exploration of multi-input gene circuit behavior.** (**A**) Multi-input inducible
832 synTF circuit diagram. Left: the circuit contains three EUs: a 4-OHT inducible synTF A (navy), a
833 GZV-inducible synTF B (teal), and a GFP reporter unit (green). As in **Fig. 2A**, 4-OHT (input A)
834 binds to the ERT2 domain (light blue), allowing synTF A to translocate into the nucleus and acti-
835 vate expression of the GFP (output). GZV (input B) inhibits the proteolytic activity of NS3 (orange),
836 preventing the fragmentation of the synTF B and allowing it to GFP expression (output). Right:
837 the multi-input circuit architecture enables cells to perform Boolean logical operations, including
838 OR (top) and AND-logic (bottom). (**B**) The genetic components that comprise the multi-input

839 design space, organized into synTF parts (top), coding parts (middle left), and reporter parts (mid-
840 dle right). Schematic of the assembly workflow, from protein domains to full-length gene circuits
841 (bottom), where the numbers inside each plasmid or plasmid pool represent the number of vari-
842 ants included in the base/full design space. (**C**) Coverage, balance, and indexing for multi-input
843 circuit libraries. Left: circuit (light gray text, bottom) and barcode (dark gray text, top) recovery at
844 each step of the CLASSIC workflow for the base (light blue boxes) and expansion (light gray
845 boxes) libraries. Top right: Nanopore sequencing data of the library to assess composition counts,
846 revealing 7% coverage in the base (light blue histogram) and 0.002% coverage of the expansion
847 (light gray histogram) design spaces. Bottom right: Nanopore sequencing data plotted in rank
848 order by reads per composition (base library: light blue, expansion library: light gray) and bar-
849 codes per composition (base library: dark blue, expansion library: dark gray). (**D**) Circuit library
850 sorting and measurement. Top left: GFP expression histograms for base (dotted line) and expan-
851 sion (shaded) circuit libraries in the presence of no inducer (gray, top left), 4-OHT only (navy, top
852 right), GZV only (orange, bottom left), or both inducers (green, bottom right), overlaid with the
853 boundaries of flow sorting binds (vertical grey dotted lines) and empty HEK293T-LP cells (light
854 gray histogram). Top right: CLASSIC versus ground-truth GFP expression measurements of ran-
855 domly isolated compositions from the base (n = 14, left) and expansion (n = 13, right) design
856 spaces for all four inducer conditions (no inducer: gray, 4-OHT only: navy, GZV-only: orange, both
857 inducers: green). Bottom: Flow cytometry histograms (top) and CLASSIC-computed distributions
858 (bottom, kernel density (solid line) calculated from normalized barcode read count (shaded verti-
859 cal bars) for one composition from the base library (dark gray fill in scatter) are shown alongside
860 the accompanying composition ID.
861
862 **Figure 5. Analysis of digital logic gene circuit design rules in >10^9 member design space.**
863 (**A**) Left: ML model for predicting multi-input circuit behaviors (basal: light gray, navy: 4-OHT only,
864 orange: GZV only, green: both inducers) using OHE part compositions from base (navy) and

expansion (dark gray) library data (left). Right: the architecture of the MLP enables exploration of large design spaces (light gray) by first training on a base library (green, top) to infer a small section of the design space (pink, middle) and then fine-tuning using additional data (green, middle) to infer the entire design space (pink, bottom). (**B**) Left: scatter plots of the test set r^2 from the fine-tuned model for each of the four input conditions (basal: light gray, navy: OHT only, orange: GZV only, green: both inducers). Right: compositions from the base (dark gray) and full (light gray) design spaces are plotted as Kullback-Leibler divergences (D_{KL}), which describe the similarity between the model-computed output behaviors and the archetypal AND and OR behaviors. Compositions to the left of the vertical dotted line or below the horizontal dotted line represent the most AND- or OR-like designs, respectively. (**C**) Scatter plot of predicted versus ground-truth GFP expression measurements for 36 individually constructed compositions from across the behavior space (white dots: basal, navy dots: 4-OHT only, orange dots: GZV only, green dots: both inducers) (*MAE = 0.32*) (top). Only 1 of the 36 compositions was measured by CLASSIC. The most AND-like (thick red outline) and OR-like (thick purple outline) designs are highlighted in the scatter plot (top) and broken out below (middle and bottom, respectively). (**D**) Genetic part usage in the AND- (left) and OR-like (right) regions of design space, grouped according to the synTF (synTF A: left, synTF B: right). Part fold enrichment is calculated by dividing the observed part occurrence by expected part occurrence from a balance library. (**E**) Feature importance for all part categories based on absolute SHAP values computed in either AND-like (top) or OR-like (bottom) regions. Larger SHAP values indicate greater contribution to the model's predictions of AND- or OR-like behavior (navy: synTF A, dark gray: synTF coding, teal: synTF B, light blue: synTF B coding, green: reporter). (**F**) Mutual information between part categories in the AND- (left) and OR-like (right) regions of behavior space, as indicated by red line thickness (navy: synTF A, dark gray: synTF coding, teal: synTF B, light blue: synTF B coding, green: reporter). (**G**) Clustering analysis of the AND-like circuit designs. Top left: UMAP projection of AND-like designs (light red: Cluster A, dark red: cluster B, blue: cluster C). Top right: schematic of the most common locus design

observed in AND-like compositions (navy: synTF A, green: reporter, teal: synTF B). Bottom left: schematics of the part usage in synTF B (left) and synTF A (middle) in clusters A and B (light red: Cluster A, dark red: cluster B) and an overview of the dominant strategy for achieving AND-like behavior (right). Bottom right: schematics of the part usage in synTF B (left) and synTF A (right) in the minority cluster (blue: cluster C). (**H**) Clustering analysis of the OR-like circuit designs. (Top left) UMAP projection of AND-like designs (blue: Cluster A, light red: Cluster B, dark red: cluster C, dark blue: cluster D). Top right: schematic of the most common locus design observed in AND-like compositions (navy: synTF A, green: reporter, teal: synTF B). Bottom: schematics of the part usage in synTF B and synTF A (left, middle) in all clusters and an overview of the strategy for achieving OR-like behavior (right) (blue: Cluster A, light red: Cluster B, dark red: cluster C, dark blue: cluster D).

REFERENCES

1 English, M. A., Gayet, R. V. & Collins, J. J. Designing Biological Circuits: Synthetic Biology Within the Operon Model and Beyond. *Annu Rev Biochem* **90**, 221-244, doi:10.1146/annurev-biochem-013118-111914 (2021).

2 Mahata, B. *et al.* Compact engineered human transactivation modules enable potent and versatile synthetic transcriptional control. *bioRxiv*, 2022.2003.2021.485228, doi:10.1101/2022.03.21.485228 (2022).

3 Slusarczyk, A. L., Lin, A. & Weiss, R. Foundations for the design and implementation of synthetic genetic circuits. *Nat Rev Genet* **13**, 406-420, doi:10.1038/nrg3227 (2012).

4 Bashor, C. J. & Collins, J. J. Understanding Biological Regulation Through Synthetic Biology. *Annu Rev Biophys* **47**, 399-423, doi:10.1146/annurev-biophys-070816-033903 (2018).

5 Bashor, C. J., Hilton, I. B., Bandukwala, H., Smith, D. M. & Veiseh, O. Engineering the next generation of cell-based therapeutics. *Nat Rev Drug Discov*, doi:10.1038/s41573-022-00476-6 (2022).

6 Beitz, A. M., Oakes, C. G. & Galloway, K. E. Synthetic gene circuits as tools for drug discovery. *Trends Biotechnol* **40**, 210-225, doi:10.1016/j.tibtech.2021.06.007 (2022).

922 7 Kitada, T., DiAndreth, B., Teague, B. & Weiss, R. Programming gene and engineered-cell
923 therapies with synthetic biology. *Science* **359**, doi:10.1126/science.aad1067 (2018).
924 8 Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat Rev*
925 *Microbiol* **12**, 381-390, doi:10.1038/nrmicro3239 (2014).
926 9 Yeung, E. *et al.* Biophysical Constraints Arising from Compositional Context in Synthetic
927 Gene Networks. *Cell Syst* **5**, 11-24 e12, doi:10.1016/j.cels.2017.06.001 (2017).
928 10 Lou, C., Stanton, B., Chen, Y. J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts
929 buffer synthetic circuits from genetic context. *Nat Biotechnol* **30**, 1137-1142,
930 doi:10.1038/nbt.2401 (2012).
931 11 Muller, I. E. *et al.* Gene networks that compensate for crosstalk with crosstalk. *Nat*
932 *Commun* **10**, 4028, doi:10.1038/s41467-019-12021-y (2019).
933 12 Prindle, A. *et al.* Rapid and tunable post-translational coupling of genetic circuits. *Nature*
934 **508**, 387-391, doi:10.1038/nature13238 (2014).
935 13 Shaw, W. M. *et al.* Engineering a Model Cell for Rational Tuning of GPCR Signaling. *Cell*
936 **177**, 782-796 e727, doi:10.1016/j.cell.2019.02.023 (2019).
937 14 Kinney, J. B., Murugan, A., Callan, C. G., Jr. & Cox, E. C. Using deep sequencing to
938 characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc*
939 *Natl Acad Sci U S A* **107**, 9158-9163, doi:10.1073/pnas.1004290107 (2010).
940 15 Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of
941 thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521-530,
942 doi:10.1038/nbt.2205 (2012).
943 16 Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by
944 FACS-seq. *Mol Syst Biol* **10**, 748, doi:10.15252/msb.20145136 (2014).
945 17 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and
946 translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 14024-14029,
947 doi:10.1073/pnas.1301301110 (2013).
948 18 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-
949 seq. *Science* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).
950 19 de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random
951 promoters. *Nat Biotechnol* **38**, 56-65, doi:10.1038/s41587-019-0315-8 (2020).
952 20 Taskiran, II *et al.* Cell-type-directed design of synthetic enhancers. *Nature* **626**, 212-220,
953 doi:10.1038/s41586-023-06936-2 (2024).
954 21 Gosai, S. J. *et al.* Machine-guided design of cell-type-targeting cis-regulatory elements.
955 *Nature* **634**, 1211-1220, doi:10.1038/s41586-024-08070-z (2024).
956 22 Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat Genet* **54**,
957 283-294, doi:10.1038/s41588-021-01009-4 (2022).
958 23 Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory
959 elements. *Nature*, doi:10.1038/s41586-024-08430-9 (2025).
960 24 Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting
961 and Engineering Alternative Polyadenylation. *Cell* **178**, 91-106 e123,
962 doi:10.1016/j.cell.2019.04.046 (2019).
963 25 Khoroshkin, M. *et al.* A generative framework for enhanced cell-type specificity in
964 rationally designed mRNAs. *bioRxiv*, doi:10.1101/2024.12.31.630783 (2024).

26 Gera, T., Jonas, F., More, R. & Barkai, N. Evolution of binding preferences among whole-genome duplicated transcription factors. *Elife* **11**, doi:10.7554/eLife.73225 (2022).
27 DelRosso, N. *et al.* Large-scale mapping and systematic mutagenesis of human transcriptional effector domains. *bioRxiv*, 2022.2008.2026.505496, doi:10.1101/2022.08.26.505496 (2022).
28 Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable RNA switches. *Nat Commun* **11**, 5057, doi:10.1038/s41467-020-18677-1 (2020).
29 Jones, E. M. *et al.* Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9**, doi:10.7554/eLife.54895 (2020).
30 Zhou, Y. *et al.* Encoding Genetic Circuits with DNA Barcodes Paves the Way for Machine Learning-Assisted Metabolite Biosensor Response Curve Profiling in Yeast. *ACS Synth Biol* **11**, 977-989, doi:10.1021/acssynbio.1c00595 (2022).
31 Wong, A. S., Choi, G. C., Cheng, A. A., Purcell, O. & Lu, T. K. Massively parallel high-order combinatorial genetics in human cells. *Nat Biotechnol* **33**, 952-961, doi:10.1038/nbt.3326 (2015).
32 Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874-882, doi:10.1038/s41588-018-0122-z (2018).
33 Castellanos-Rueda, R. *et al.* speedingCARs: accelerating the engineering of CAR T cells by signaling domain shuffling and single-cell sequencing. *Nat Commun* **13**, 6555, doi:10.1038/s41467-022-34141-8 (2022).
34 Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *mSystems* **3**, doi:10.1128/mSystems.00143-17 (2018).
35 Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, e16765, doi:10.1371/journal.pone.0016765 (2011).
36 O'Gorman, S., Fox, D. T. & Wahl, G. M. Recombinase-mediated gene activation and site-specific integration in mammalian cells. *Science* **251**, 1351-1355, doi:10.1126/science.1900642 (1991).
37 Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res* **42**, 13440-13451, doi:10.1093/nar/gku1082 (2014).
38 Norrman, K. *et al.* Quantitative comparison of constitutive promoters in human ES cells. *PLoS One* **5**, e12413, doi:10.1371/journal.pone.0012413 (2010).
39 Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* **5**, e10611, doi:10.1371/journal.pone.0010611 (2010).
40 Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci U S A* **110**, 11284-11289, doi:10.1073/pnas.1305590110 (2013).
41 Petitclerc, D. *et al.* The effect of various introns and transcription terminators on the efficiency of expression vectors in various cultured cell lines and in the mammary gland of transgenic mice. *J Biotechnol* **40**, 169-178, doi:10.1016/0168-1656(95)00047-t (1995).
42 Wang, X. Y. *et al.* Enhanced Transgene Expression by Optimization of Poly A in Transfected CHO Cells. *Front Bioeng Biotechnol* **10**, 722722, doi:10.3389/fbioe.2022.722722 (2022).

1008 43 Khalil, A. S. *et al.* A synthetic biology framework for programming eukaryotic transcription
1009 functions. *Cell* **150**, 647-658, doi:10.1016/j.cell.2012.05.045 (2012).
1010 44 Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K.
1011 Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized
1012 zinc-finger arrays. *Nat Protoc* **4**, 1471-1501, doi:10.1038/nprot.2009.98 (2009).
1013 45 Li, H. S. *et al.* Multidimensional control of therapeutic human cell function with synthetic
1014 gene circuits. *Science* **378**, 1227-1234, doi:10.1126/science.ade0156 (2022).
1015 46 Feil, R., Wagner, J., Metzger, D. & Chambon, P. Regulation of Cre recombinase activity by
1016 mutated estrogen receptor ligand-binding domains. *Biochem Biophys Res Commun* **237**,
1017 752-757, doi:10.1006/bbrc.1997.7124 (1997).
1018 47 Bashor, C. J. *et al.* Complex signal processing in synthetic gene circuits using cooperative
1019 regulatory assemblies. *Science* **364**, 593-597, doi:10.1126/science.aau8287 (2019).
1020 48 Donahue, P. S. *et al.* The COMET toolkit for composing customizable genetic programs in
1021 mammalian cells. *Nat Commun* **11**, 779, doi:10.1038/s41467-019-14147-5 (2020).
1022 49 Muldoon, J. J. *et al.* Model-guided design of mammalian genetic programs. *Sci Adv* **7**,
1023 doi:10.1126/sciadv.abe9375 (2021).
1024 50 Kabadi, A. M. & Gersbach, C. A. Engineering synthetic TALE and CRISPR/Cas9 transcription
1025 factors for regulating gene expression. *Methods* **69**, 188-197,
1026 doi:10.1016/j.ymeth.2014.06.014 (2014).
1027 51 La Russa, M. F. & Qi, L. S. The New State of the Art: Cas9 for Gene Activation and
1028 Repression. *Mol Cell Biol* **35**, 3800-3809, doi:10.1128/MCB.00512-15 (2015).
1029 52 Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent
1030 transcriptional activator. *Nature* **335**, 563-564, doi:10.1038/335563a0 (1988).
1031 53 Shin, Y. *et al.* Spatiotemporal Control of Intracellular Phase Transitions Using Light-
1032 Activated optoDroplets. *Cell* **168**, 159-171 e114, doi:10.1016/j.cell.2016.11.054 (2017).
1033 54 Schneider, N. *et al.* Liquid-liquid phase separation of light-inducible transcription factors
1034 increases transcription activation in mammalian cells and mice. *Sci Adv* **7**,
1035 doi:10.1126/sciadv.abd3568 (2021).
1036 55 Boeynaems, S. *et al.* Phase Separation in Biology and Disease; Current Perspectives and
1037 Open Questions. *J Mol Biol* **435**, 167971, doi:10.1016/j.jmb.2023.167971 (2023).
1038 56 Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by
1039 tetracycline-responsive promoters. *Proc Natl Acad Sci U S A* **89**, 5547-5551,
1040 doi:10.1073/pnas.89.12.5547 (1992).
1041 57 Hansen, J. *et al.* Transplantation of prokaryotic two-component signaling pathways into
1042 mammalian cells. *Proc Natl Acad Sci U S A* **111**, 15705-15710,
1043 doi:10.1073/pnas.1406482111 (2014).
1044 58 Ede, C., Chen, X., Lin, M. Y. & Chen, Y. Y. Quantitative Analyses of Core Promoters Enable
1045 Precise Engineering of Regulated Gene Expression in Mammalian Cells. *ACS Synth Biol* **5**,
1046 395-404, doi:10.1021/acssynbio.5b00266 (2016).
1047 59 Chen, H., Covert, I. C., Lundberg, S. M. & Lee, S.-I. Algorithms to estimate Shapley value
1048 feature attributions. *Nature Machine Intelligence* **5**, 590-601, doi:10.1038/s42256-023-
1049 00657-x (2023).

1050 60 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
1051 Projection for Dimension Reduction. arXiv:1802.03426 (2018).
1052 <https://ui.adsabs.harvard.edu/abs/2018arXiv180203426M>.
1053 61 Tycko, J. *et al.* Development of compact transcriptional effectors using high-throughput
1054 measurements in diverse contexts. *Nat Biotechnol*, doi:10.1038/s41587-024-02442-6
1055 (2024).
1056 62 Tague, E. P., Dotson, H. L., Tunney, S. N., Sloas, D. C. & Ngo, J. T. Chemogenetic control of
1057 gene expression and cell signaling with antiviral drugs. *Nat Methods* **15**, 519-522,
1058 doi:10.1038/s41592-018-0042-y (2018).
1059 63 Wimmer, E., Mueller, S., Tumpey, T. M. & Taubenberger, J. K. Synthetic viruses: a new
1060 opportunity to understand and prevent viral disease. *Nat Biotechnol* **27**, 1163-1172,
1061 doi:10.1038/nbt.1593 (2009).
1062 64 Brophy, J. A. & Voigt, C. A. Principles of genetic circuit design. *Nat Methods* **11**, 508-520,
1063 doi:10.1038/nmeth.2926 (2014).
1064 65 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers
1065 at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
1066 66 Pinglay, S. *et al.* Synthetic regulatory reconstitution reveals principles of mammalian Hox
1067 cluster regulation. *Science* **377**, eabk2820, doi:10.1126/science.abk2820 (2022).
1068 67 Voigt, C. A. Synthetic biology 2020-2030: six commercially-available products that are
1069 changing our world. *Nat Commun* **11**, 6379, doi:10.1038/s41467-020-20122-2 (2020).
1070 68 Valeri, J. A. *et al.* Sequence-to-function deep learning frameworks for engineered
1071 riboregulators. *Nat Commun* **11**, 5058, doi:10.1038/s41467-020-18676-2 (2020).
1072 69 Hollerer, S. *et al.* Large-scale DNA-based phenotypic recording and deep learning enable
1073 highly accurate sequence-function mapping. *Nat Commun* **11**, 3551, doi:10.1038/s41467-
1074 020-17222-4 (2020).
1075 70 LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic
1076 promoters to control transcriptional profiles in bacteria. *Nat Commun* **13**, 5159,
1077 doi:10.1038/s41467-022-32829-5 (2022).
1078 71 Rai, K., Wang, Y., O'Connell, R. W., Patel, A. B. & Bashor, C. J. Using Machine Learning to
1079 Enhance and Accelerate Synthetic Biology. *Current Opinion in Biomedical Engineering*
1080 (2024).
1081 72 Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular
1082 identifiers with Nanopore or PacBio sequencing. *Nat Methods* **18**, 165-169,
1083 doi:10.1038/s41592-020-01041-y (2021).
1084 73 Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent Escherichia
1085 coli: transformation and storage of bacterial cells in the same solution. *Proc Natl Acad Sci*
1086 *U S A* **86**, 2172-2175, doi:10.1073/pnas.86.7.2172 (1989).
1087 74 Parrish, J. R. *et al.* High-throughput cloning of Campylobacter jejuni ORfs by in vivo
1088 recombination in Escherichia coli. *J Proteome Res* **3**, 582-586, doi:10.1021/pr0341134
1089 (2004).
1090 75 Currin, A. *et al.* Highly multiplexed, fast and accurate nanopore sequencing for verification
1091 of synthetic DNA constructs and sequence libraries. *Synth Biol (Oxf)* **4**, ysz025,
1092 doi:10.1093/synbio/ysz025 (2019).

1093 76 De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-
1094 read sequencing. *Nat Rev Genet* **22**, 572-587, doi:10.1038/s41576-021-00367-3 (2021).
1095 77 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol*
1096 *Biol* **147**, 195-197, doi:10.1016/0022-2836(81)90087-5 (1981).
1097 78 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-
1098 823, doi:10.1126/science.1231143 (2013).
1099 79 Hermann, M. *et al.* Binary recombinase systems for high-resolution conditional
1100 mutagenesis. *Nucleic Acids Res* **42**, 3894-3907, doi:10.1093/nar/gkt1361 (2014).
1101

Fig. 1

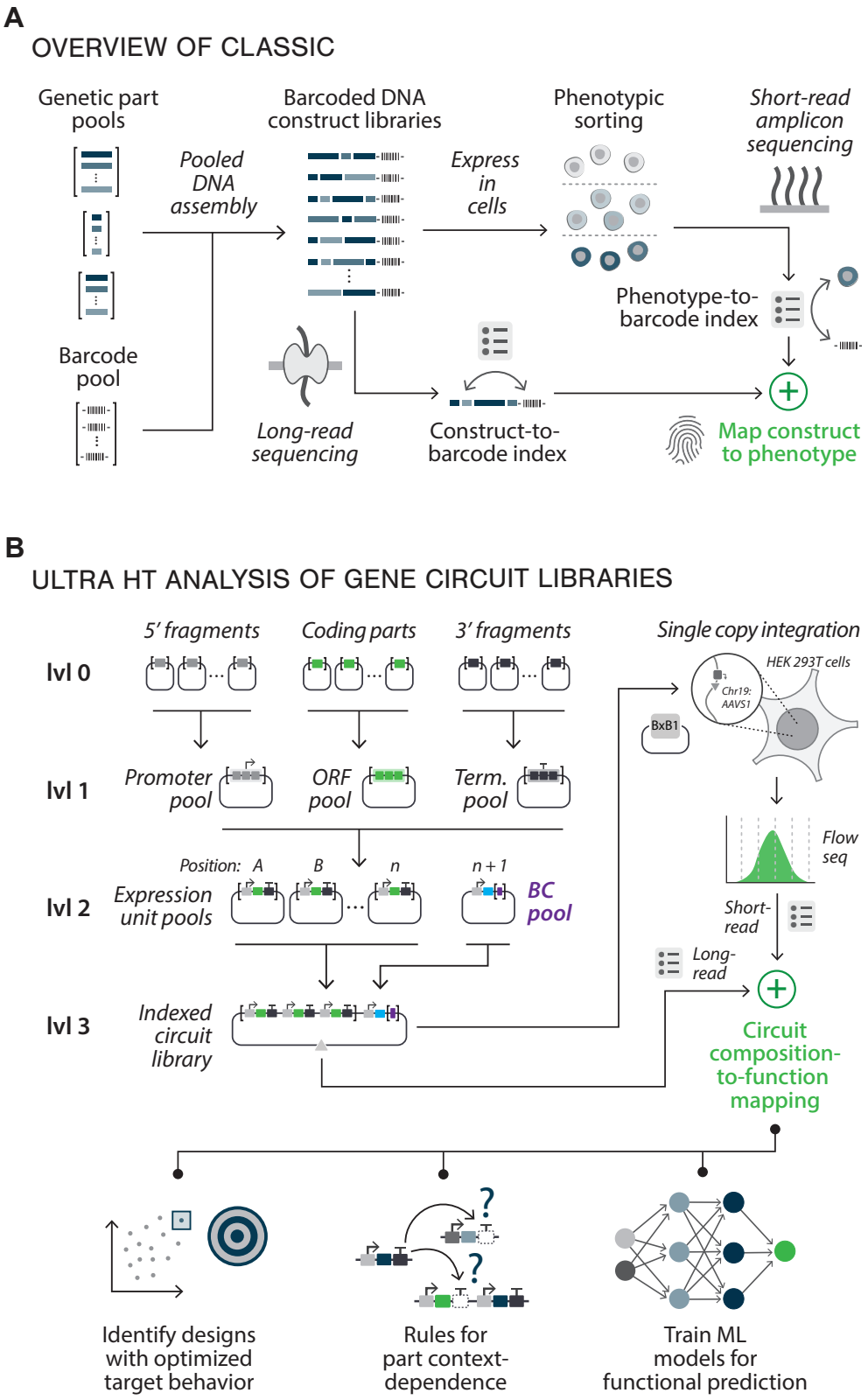

A
OVERVIEW OF CLASSIC
Genetic part pools
Barcoded DNA construct libraries
Phenotypic sorting
Short-read amplicon sequencing
Pooled DNA assembly
Express in cells
Phenotype-to-barcode index
Barcode pool
Long-read sequencing
Construct-to-barcode index
Map construct to phenotype
B
ULTRA HT ANALYSIS OF GENE CIRCUIT LIBRARIES
5′ fragments
Coding parts
3′ fragments
Single copy integration
lvl 0
lvl 1
lvl 2
lvl 3
Promoter pool
ORF pool
Term. pool
Position: A
B
n
n + 1
Expression unit pools
BC pool
Indexed circuit library
HEK 293T cells
Chr19: AAVS1
BxB1
Flow seq
Short-read
Long-read
Circuit composition-to-function mapping
Identify designs with optimized target behavior
Rules for part context-dependence
Train ML models for functional prediction

Fig. 2

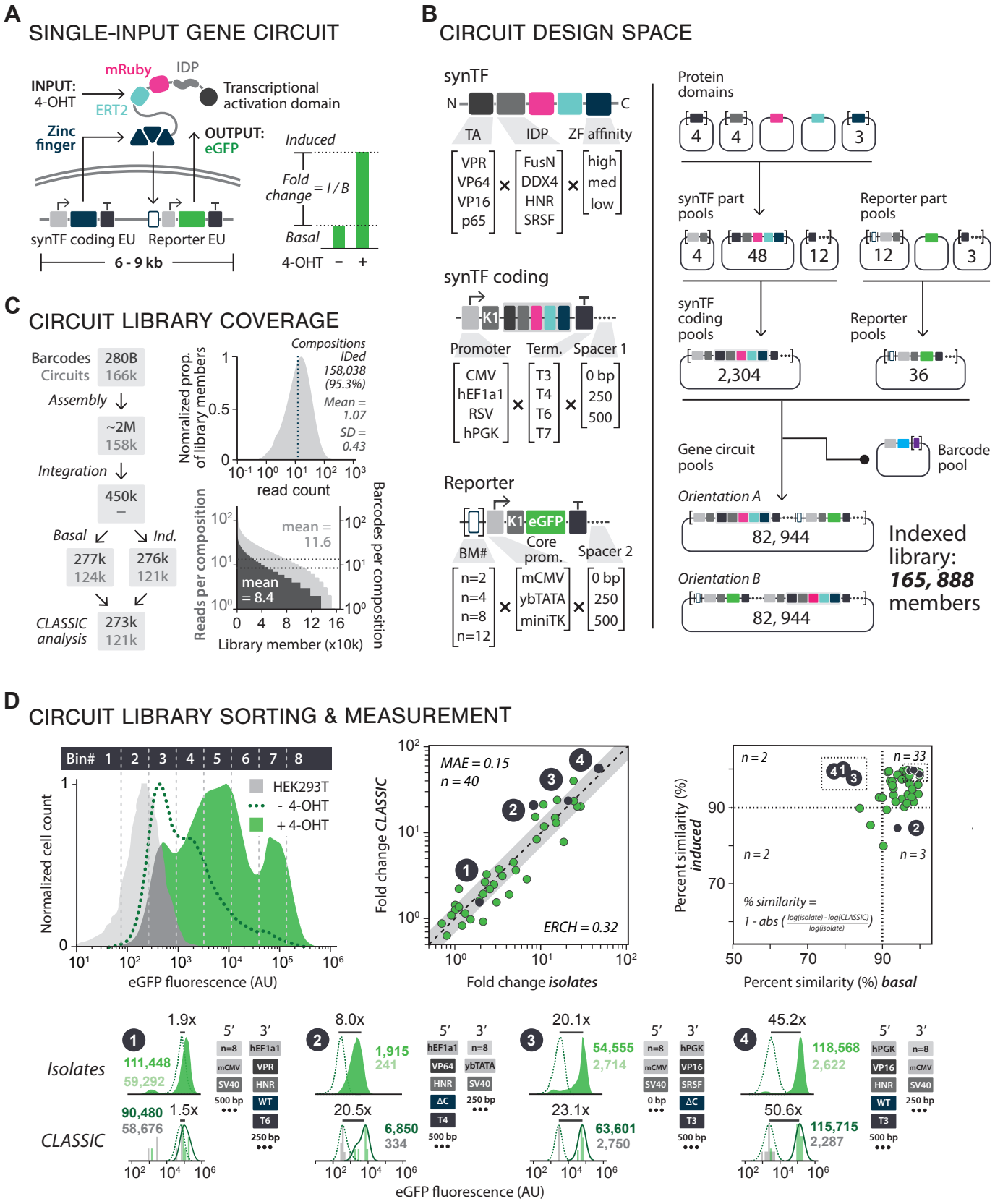
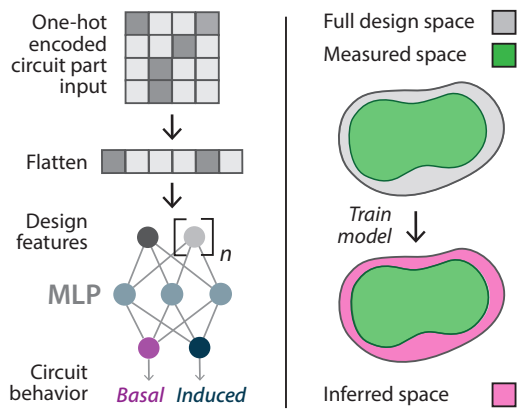
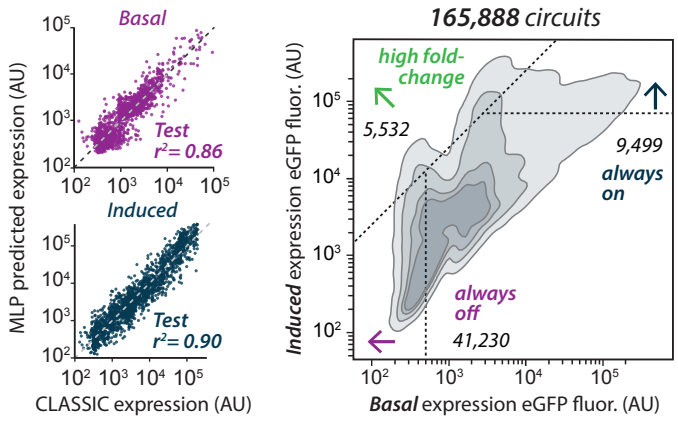

A
SINGLE-INPUT GENE CIRCUIT
INPUT: 4-OHT
mRuby
IDP
ERT2
Transcriptional activation domain
Zinc finger
OUTPUT: eGFP
synTF coding EU
Reporter EU
6 - 9 kb
Induced
Fold change = I / B
Basal
4-OHT − +
B
CIRCUIT DESIGN SPACE
synTF
N
C
TA
IDP
ZF affinity
VPR
VP64
VP16
p65
FusN
DDX4
HNR
SRSF
high
med
low
synTF coding
Promoter
Term.
Spacer 1
CMV
hEF1a1
RSV
hPGK
T3
T4
T6
T7
0 bp
250
500
Reporter
BM#
Core prom.
Spacer 2
n=2
n=4
n=8
n=12
mCMV
ybTATA
miniTK
Protein domains
synTF part pools
Reporter part pools
synTF coding pools
Reporter pools
Gene circuit pools
Barcode pool
Orientation A
Orientation B
82, 944
2,304
36
Indexed library: 165, 888 members
C
CIRCUIT LIBRARY COVERAGE
Barcodes 280B
Circuits 166k
Assembly
~2M
158k
Integration
450k
Basal
Ind.
277k
124k
276k
121k
CLASSIC analysis
273k
121k
Compositions IDed 158,038 (95.3%)
Mean = 1.07
SD = 0.43
Nomralized prop. of library members
read count
mean = 11.6
mean = 8.4
Reads per composition
Barcodes per composition
Library member (x10k)
D
CIRCUIT LIBRARY SORTING & MEASUREMENT
Bin# 1 2 3 4 5 6 7 8
HEK293T
- 4-OHT
+ 4-OHT
Normalized cell count
eGFP fluorescence (AU)
MAE = 0.15
n = 40
ERCH = 0.32
Fold change CLASSIC
Fold change isolates
n = 2
n = 33
n = 3
Percent similarity (%) induced
Percent similarity (%) basal
% similarity = 1 - abs (log(isolate) - log(CLASSIC) / log(isolate))
Isolates
CLASSIC
1.9x
111,448
59,292
90,480
1.5x
58,676
8.0x
1,915
241
20.5x
6,850
334
20.1x
54,555
2,714
23.1x
63,601
2,750
45.2x
118,568
2,622
50.6x
115,715
2,287
5′
3′
n=8
hEF1a1
mCMV
VPR
SV40
HNR
WT
T6
500 bp
250 bp
VP64
ybTATA
ΔC
T4
hPGK
VP16
SRSF
T3
0 bp
eGFP fluorescence (AU)

Fig. 3

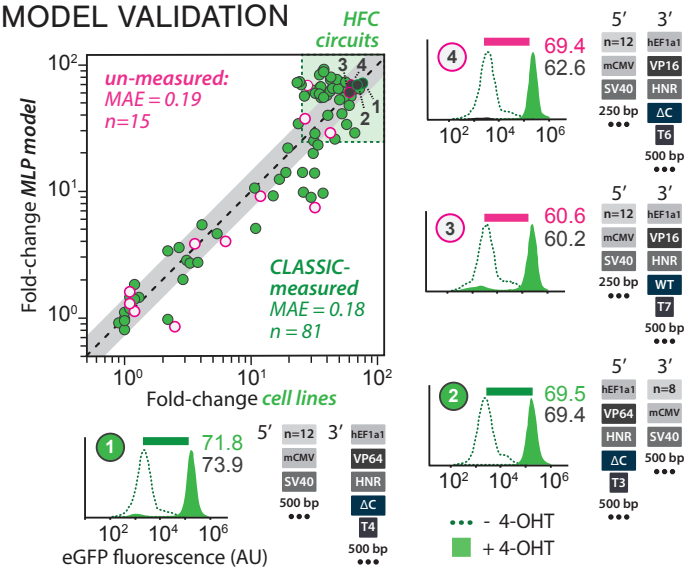
A ML–GUIDED DESIGN SPACE MAPPING



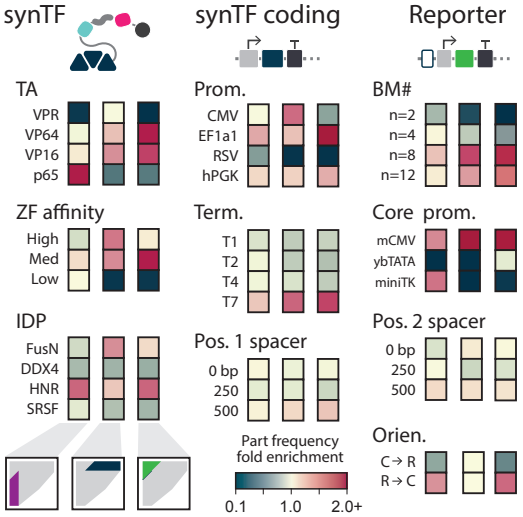
B ML MODEL OF SINGLE-INPUT DESIGN SPACE



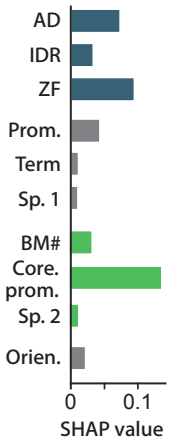
C MODEL VALIDATION



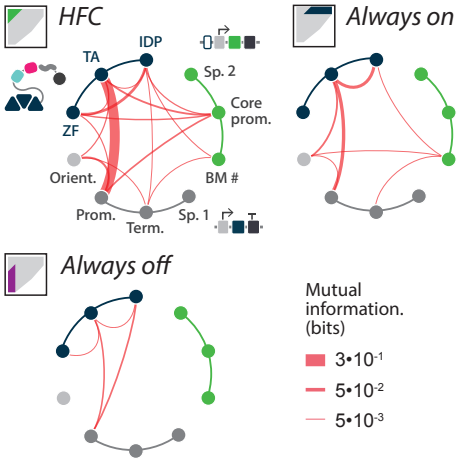
D SINGLE-INPUT PART USAGE



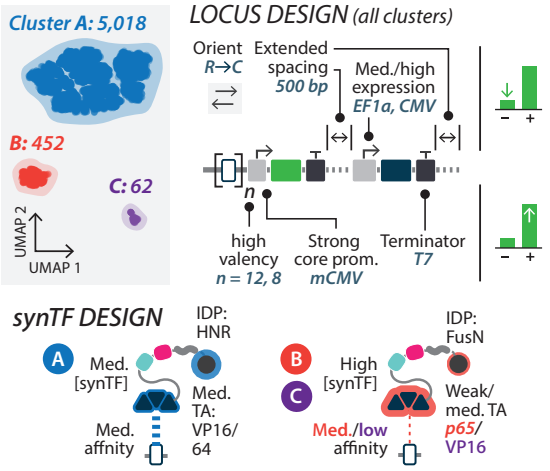
E FEATURE IMPORTANCE



F MUTUAL INFORMATION



G HFC DESIGN RULES



H MODEL EFFICIENCY AND ADAPTABILITY

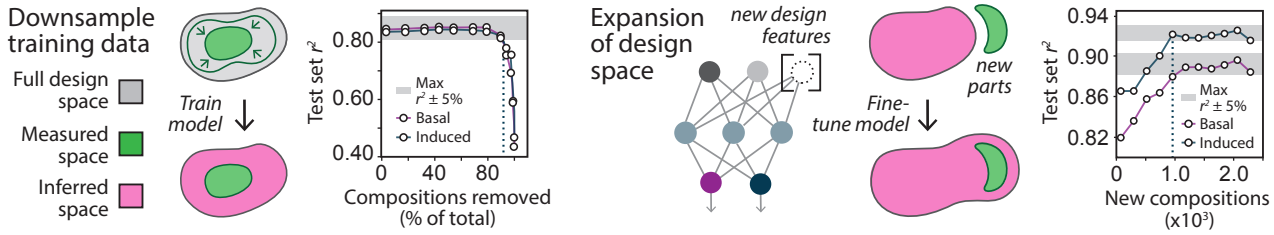


Fig. 4

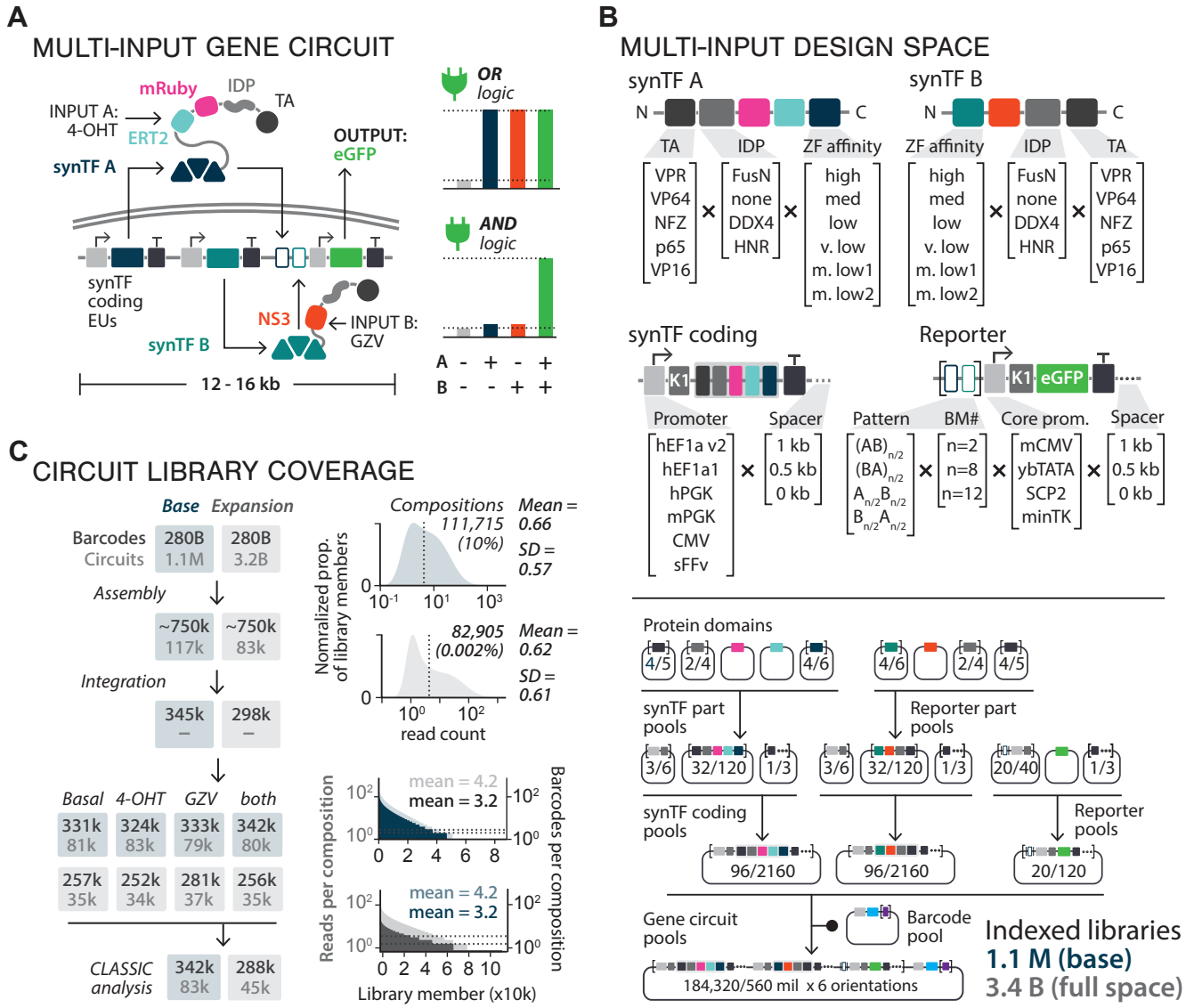
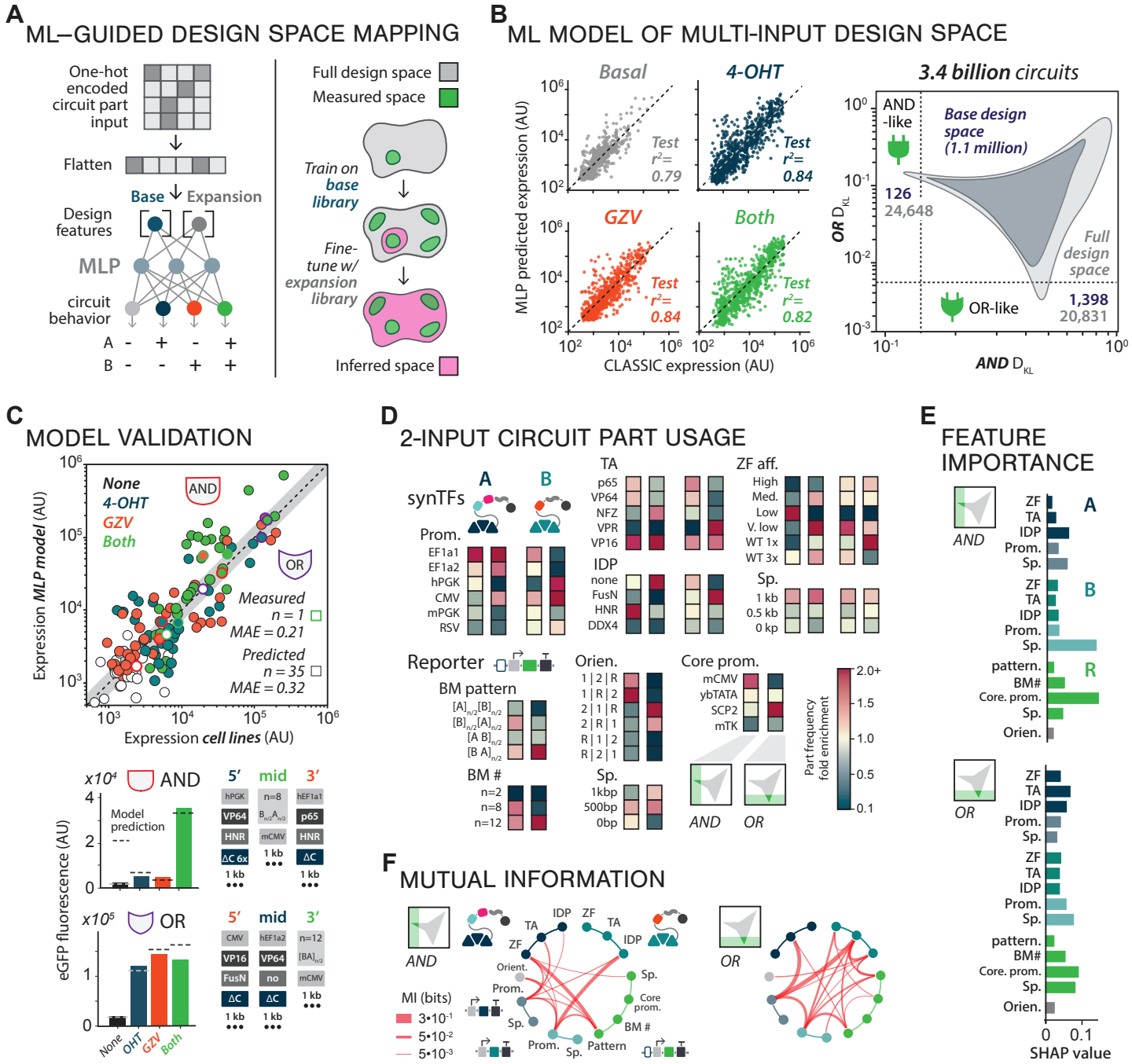

A
MULTI-INPUT GENE CIRCUIT
mRuby
IDP
TA
INPUT A: 4-OHT
ERT2
synTF A
OUTPUT: eGFP
synTF coding EUs
NS3
INPUT B: GZV
synTF B
12 - 16 kb
OR logic
AND logic
A - + - +
B - - + +
B
MULTI-INPUT DESIGN SPACE
synTF A
synTF B
N
C
TA
IDP
ZF affinity
VPR
VP64
NFZ
p65
VP16
FusN
none
DDX4
HNR
high
med
low
v. low
m. low1
m. low2
synTF coding
Reporter
K1
eGFP
Promoter
Spacer
Pattern
BM#
Core prom.
hEF1a v2
hEF1a1
hPGK
mPGK
CMV
sFFv
1 kb
0.5 kb
0 kb
n=2
n=8
n=12
mCMV
ybTATA
SCP2
minTK
Protein domains
synTF part pools
Reporter part pools
synTF coding pools
Reporter pools
Gene circuit pools
Barcode pool
184,320/560 mil x 6 orientations
Indexed libraries
1.1 M (base)
3.4 B (full space)
C
CIRCUIT LIBRARY COVERAGE
Base
Expansion
Barcodes
Circuits
280B
1.1M
280B
3.2B
Assembly
~750k
117k
~750k
83k
Integration
345k
298k
Basal
4-OHT
GZV
both
331k
81k
324k
83k
333k
79k
342k
80k
257k
35k
252k
34k
281k
37k
256k
35k
CLASSIC analysis
342k
83k
288k
45k
Compositions
111,715
(10%)
Mean = 0.66
SD = 0.57
82,905
(0.002%)
Mean = 0.62
SD = 0.61
Nomralized prop. of library members
read count
mean = 4.2
mean = 3.2
Reads per composition
Barcodes per composition
Library member (x10k)
D
CIRCUIT LIBRARY SORTING & MEASUREMENT
Bin#1 2 3 4 5 6 7 8
No inducer
HEK293T
Base
Expansion
4-OHT
GZV
Both
Normalized cell count
eGFP fluorescence (AU)
Base library
MAE = 0.28
n = 14
ERCH = 0.32
Expansion library
MAE = 0.23
n = 13
CLASSIC expression (AU)
Isolate expression (AU)
Isolate
CLASSIC
3,052
19,186
44,288
128,992
4,432
15,520
48,980
118,358
5′
mid
3′
hEF1a
p65
no
ΔC2
WT
n=8
mCMV
500

Fig. 5


A
ML–GUIDED DESIGN SPACE MAPPING
One-hot encoded circuit part input
Flatten
Base
Expansion
Design features
MLP
circuit behavior
A - + - +
B - - + +
Full design space
Measured space
Train on base library
Fine-tune w/ expansion library
Inferred space
B
ML MODEL OF MULTI-INPUT DESIGN SPACE
Basal
4-OHT
GZV
Both
Test r²= 0.79
Test r²= 0.84
Test r²= 0.84
Test r²= 0.82
MLP predicted expression (AU)
CLASSIC expression (AU)
3.4 billion circuits
AND -like
Base design space (1.1 million)
126
24,648
Full design space
OR-like
1,398
20,831
OR D_KL
AND D_KL
C
MODEL VALIDATION
None
4-OHT
GZV
Both
AND
OR
Measured
n = 1
MAE = 0.21
Predicted
n = 35
MAE = 0.32
Expression MLP model (AU)
Expression cell lines (AU)
x10⁴
AND
Model prediction
5′
mid
3′
hPGK
VP64
HNR
ΔC 6x
n=8
mCMV
1 kb
hEF1a1
p65
HNR
ΔC
x10⁵
OR
CMV
VP16
FusN
ΔC
hEF1a2
VP64
no
n=12
eGFP fluorescence (AU)
None
4-OHT
GZV
Both
D
2-INPUT CIRCUIT PART USAGE
synTFs
A
B
Prom.
EF1a1
EF1a2
hPGK
CMV
mPGK
RSV
TA
p65
VP64
NFZ
VPR
VP16
ZF aff.
High
Med.
Low
V. low
WT 1x
WT 3x
IDP
none
FusN
HNR
DDX4
Sp.
1 kb
0.5 kb
0 kp
Reporter
BM pattern
Orien.
1 | 2 | R
1 | R | 2
2 | 1 | R
2 | R | 1
R | 1 | 2
R | 2 | 1
Core prom.
mCMV
ybTATA
SCP2
mTK
BM #
n=2
n=8
n=12
Sp.
1kbp
500bp
0bp
AND
OR
Part frequency fold enrichment
2.0+
1.5
1.0
0.5
0.1
E
FEATURE IMPORTANCE
AND
A
B
R
ZF
TA
IDP
Prom.
Sp.
pattern.
BM#
Core. prom.
Orien.
OR
SHAP value
F
MUTUAL INFORMATION
AND
OR
MI (bits)
3•10⁻¹
5•10⁻²
5•10⁻³
TA
IDP
ZF
Orient.
Prom.
Sp.
Core prom.
BM #
Pattern
G
AND GATE DESIGN RULES
UMAP 2
UMAP 1
Cluster A: 11,167
C: 3,627
B: 9,854
LOCUS DESIGN (all clusters)
Orient A→R→B
Clustered BM pattern B n/2 A n/2
Strong core prom. mCMV
Extended spacing 1 kb
Extended spacing 1 kb
high valency n = 12, 8
Medium spacing 0-0.5 kb
synTF DESIGNS
A B Asymmetric
Exp. level
[Med]
[High]
[High]
FusN, HNR
HNR
p65/ VP16
VP16/ VP64
Affinity
Med.
Weak
+4-OHT
Weak binding
+GZV
Distal binding limits activation
+both
synTFs synergize
C Balanced
[High]
[Med-low]
any
HNR
p65/ VP16
VP16/ VP64
Med.
Med.
H
OR GATE DESIGN RULES
Cluster A: 9,240
B: 7,138
D: 1,545
C: 2,908
UMAP 2
UMAP 1
LOCUS DESIGN (all clusters)
Orient B→A→R
High expression hEF1a1/2, CMV
Strong core prom. SCP2
Medium/ extended spacing 0.5-1 kb
Extended spacing 1 kb
Interpersed BM pattern [B A] n/2
High valency n = 12
synTF DESIGNS
A D
Exp. level
[High]
[High]
[Low]
FusN
VP16
None
NFZ/ VP16
Affinity
Med.
Med.
B C
[High]
[High]
HNR-FusN
VPR
None
VP16, NFZ, p65
Med.
Med.
+4-OHT
+GZV
+both
Either synTF activates (balance part activities)
no synTF synergy