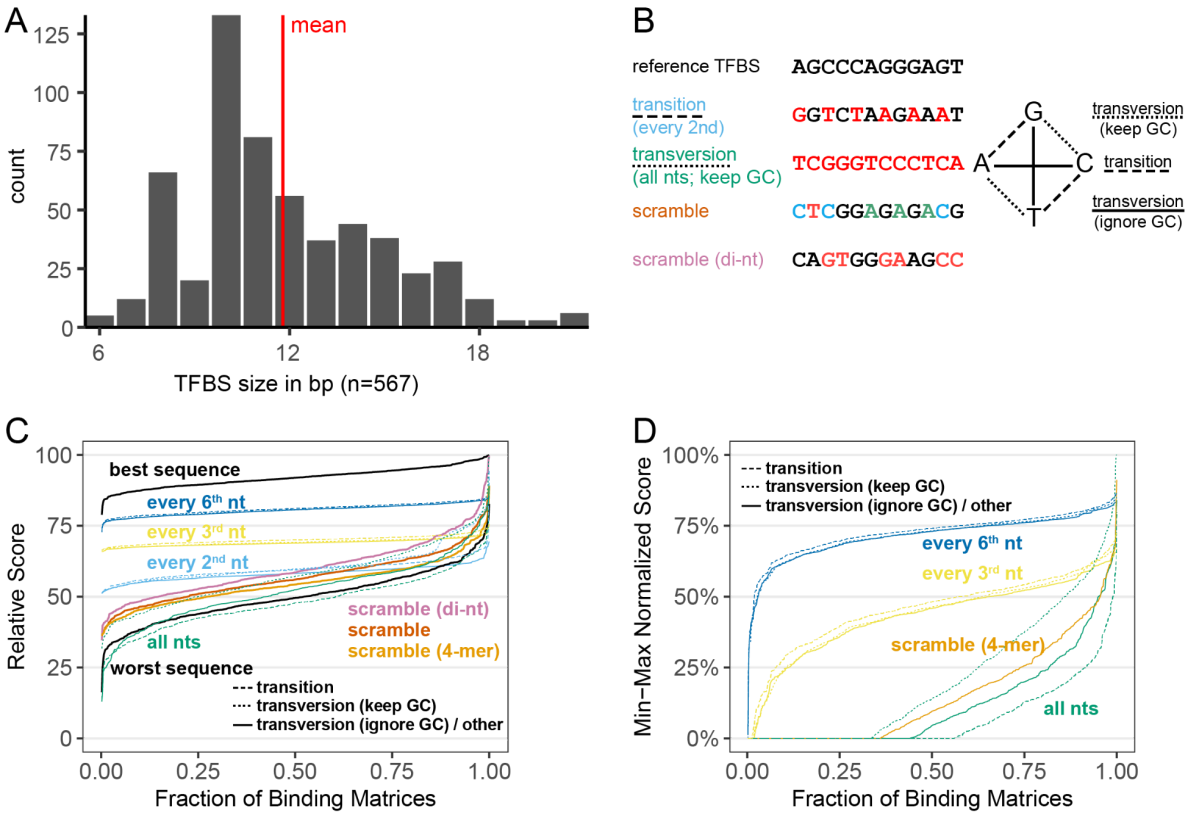
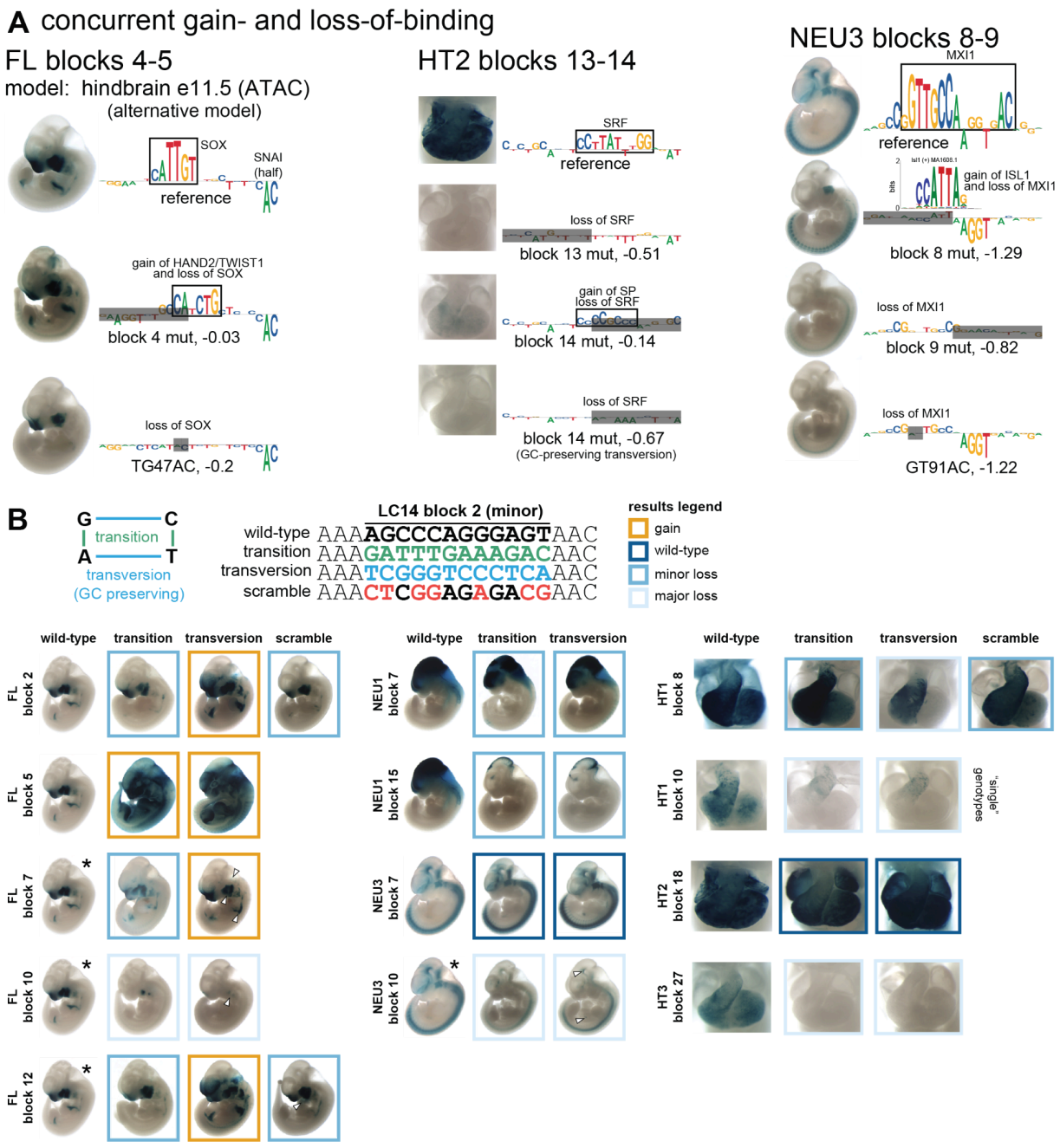
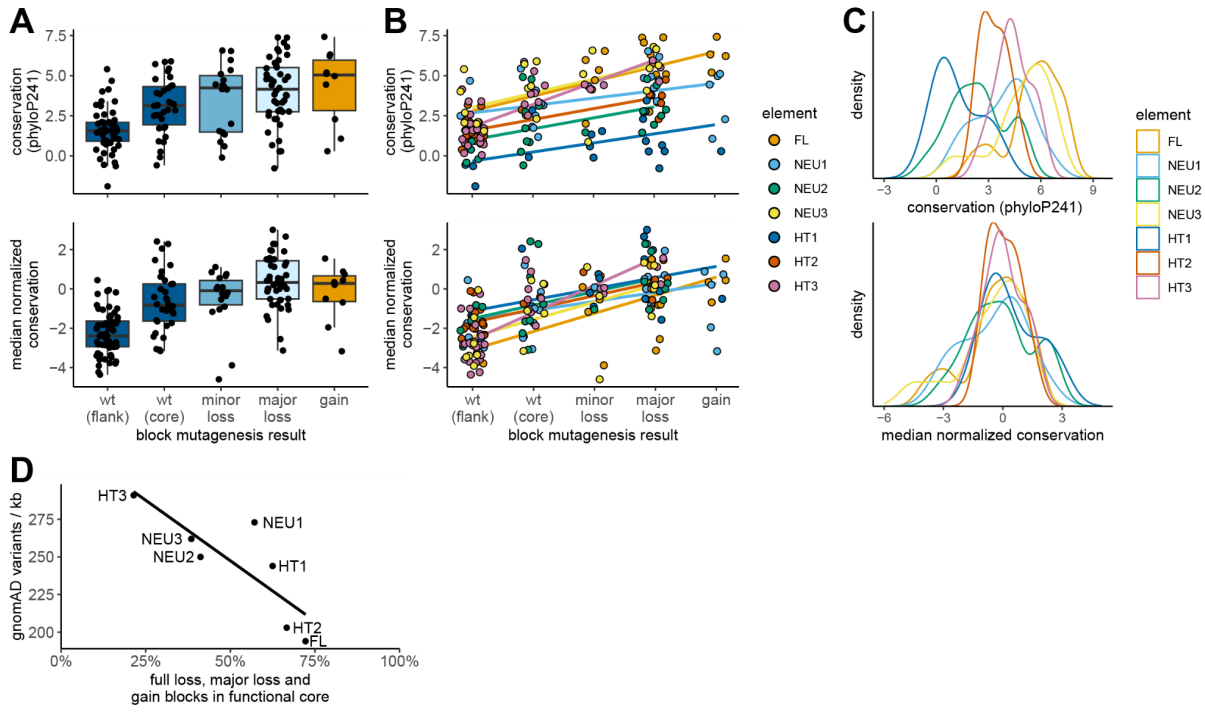
Supplementary Figures



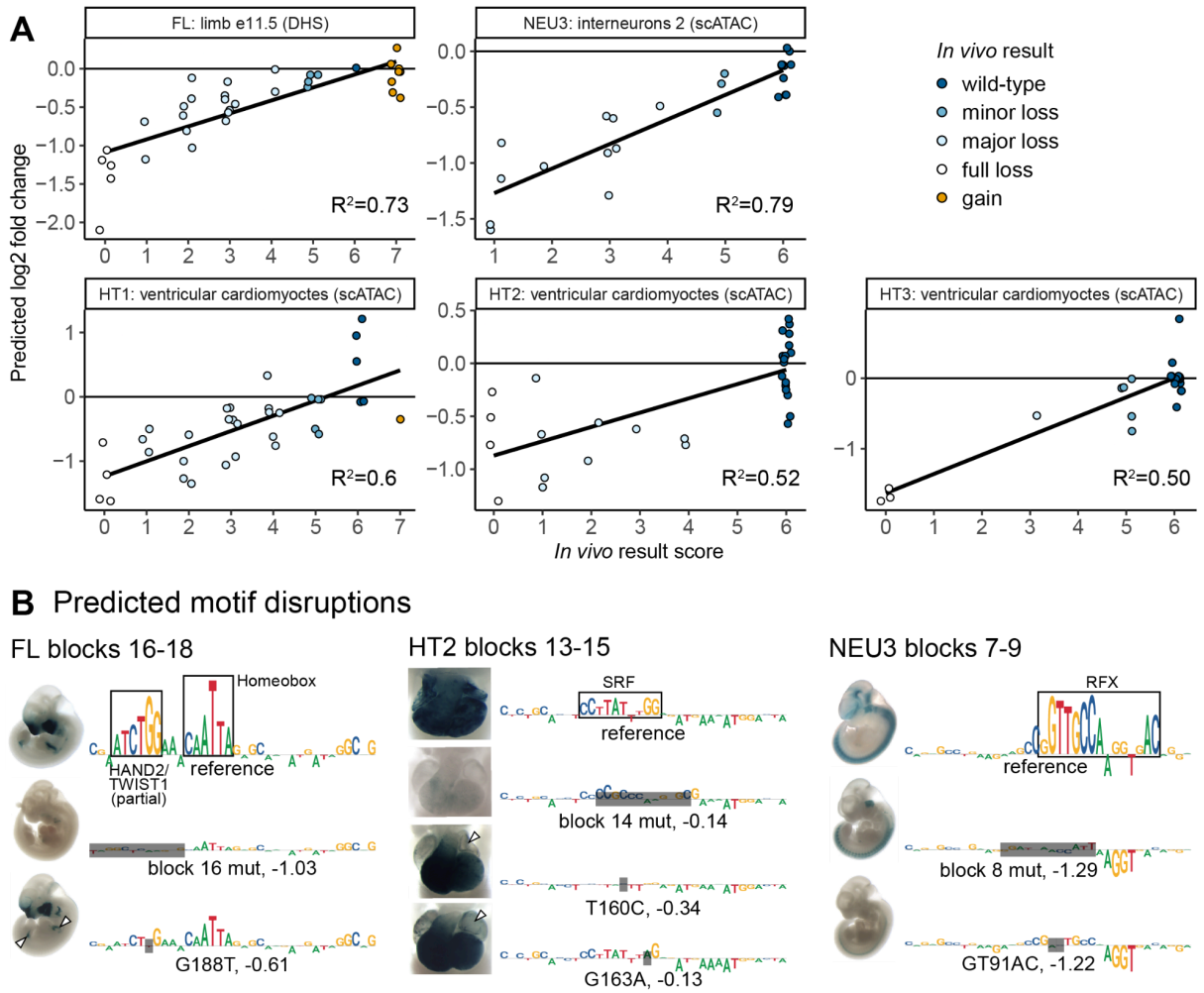
Supplementary Figure 1. Choice of mutagenesis strategy. (A) Size distribution of all JASPAR TF binding motifs. (B) Visualization of *in silico* mutagenesis schemes. (C) Relative score of matches between original TF PWM and mutagenized sequence. (D) Match score min-max normalized to that of best and worst sequence for a given TF PWM. See text for details. Observations are ordered on x-axis by score, so each position does not correspond to the same TF PWM.



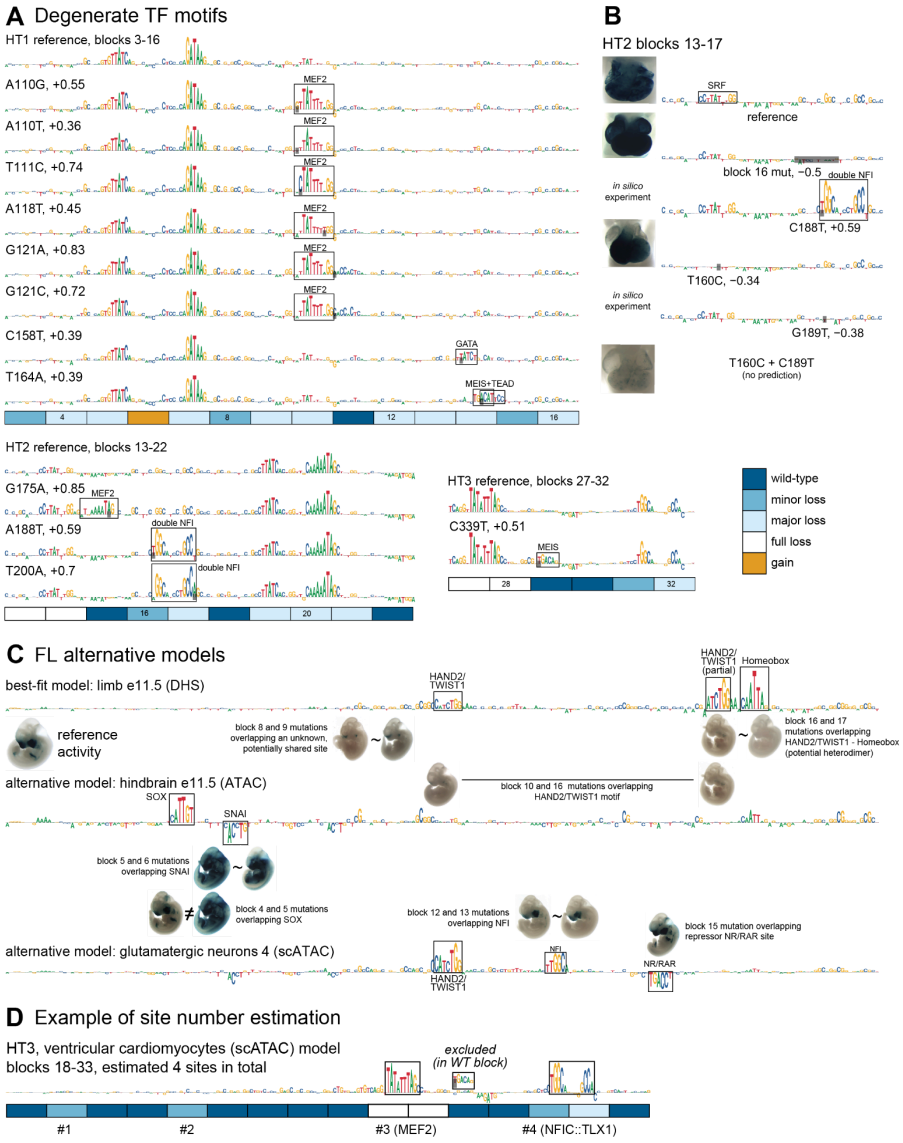
Supplementary Figure 2. Validation of transition mutagenesis scheme. (A) Three blocks with suspected gain-of-binding events or mismatch between adjacent blocks overlapping the same predicted binding motif were tested using alternative mutagenesis scheme or targeted 2bp mutations. In all cases, a result confirming gain-of-binding was obtained. (B) Unbiased testing using alternative mutagenesis schemes. Blocks were mutagenized using both a deterministic transition scheme (default for this study) and a GC-preserving transversion scheme, with selected blocks also mutagenized through random scrambling. Tandem embryos are displayed, except when indicated otherwise (see Methods for genotype definitions). White arrowheads indicate regions in which results of alternative mutagenesis mismatch those of transition mutagenesis (blocks marked with asterisk). See Supplementary Note 2 for details. Related to Figure 1.



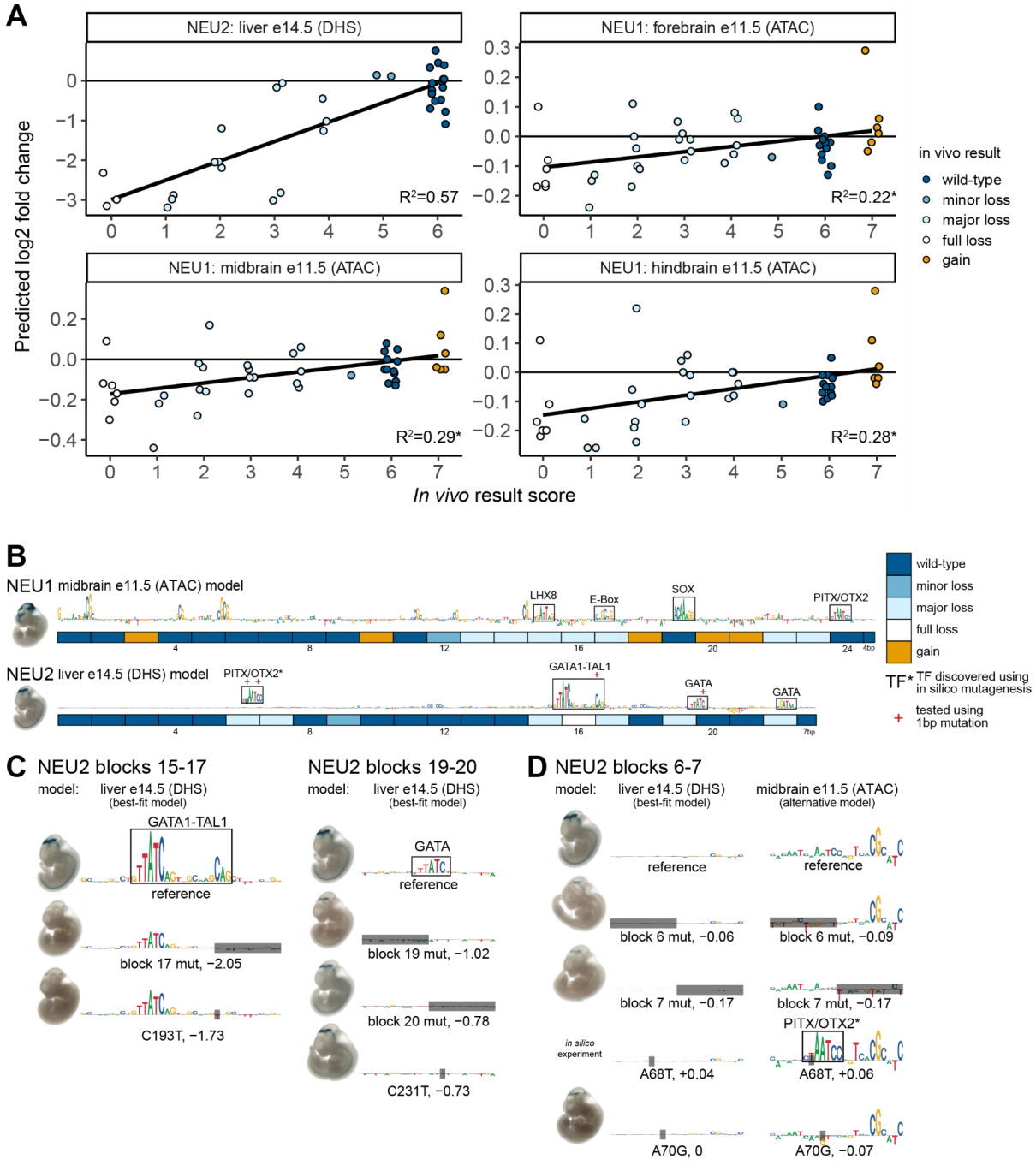
Supplementary Figure 3. Conservation score normalization and analysis including flanking wild-type blocks. (A) Conservation score boxplots by block mutagenesis result. (B) Same as A, but colored by enhancer. Linear regression line is added. (C) Density of conservation scores, colored by enhancer. Each dot in A and B is a 12bp block (N=167). Top panels use raw mammalian conservation score (phyloP241), bottom panels use raw score normalized for median of functional core (per enhancer). Minor loss, major loss and gain blocks were each more conserved, after median normalization, than either wild-type flanking blocks or all wild-type blocks combined. Only major loss blocks were more conserved than wild-type core blocks (FDR<0.05, 9 comparisons, 7 significant). (D) Correlation between density of gnomAD variants and fraction of functional blocks in functional core (Pearson R^2=68%, p-value<0.05). Related to Figure 1.



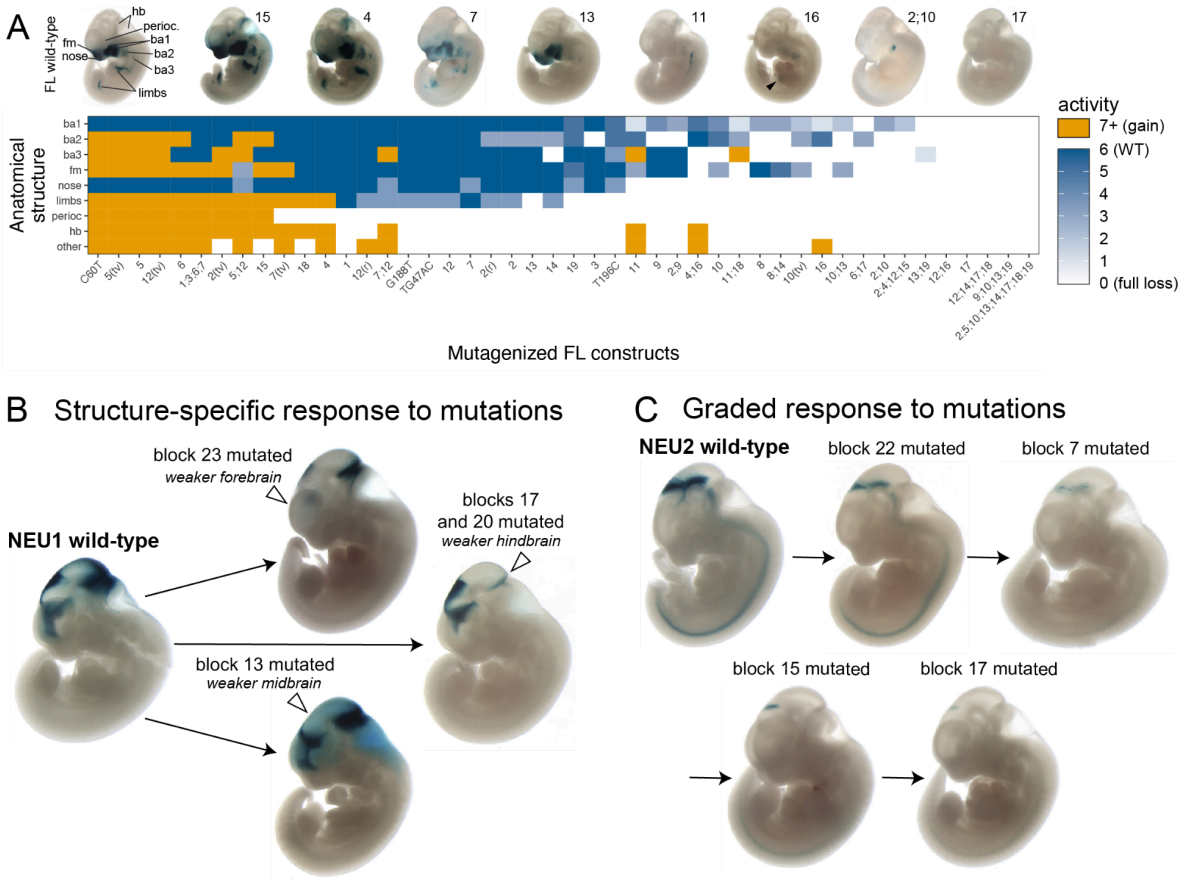
Supplementary Figure 4. Machine learning model selection and validation. (A) Correlations between model predictions and *in vivo* results. Dots = mutagenized constructs. Black fit line is linear regression. R^2 is Spearman correlation. (B) Remaining predicted motif disruptions. Related to Figure 2.



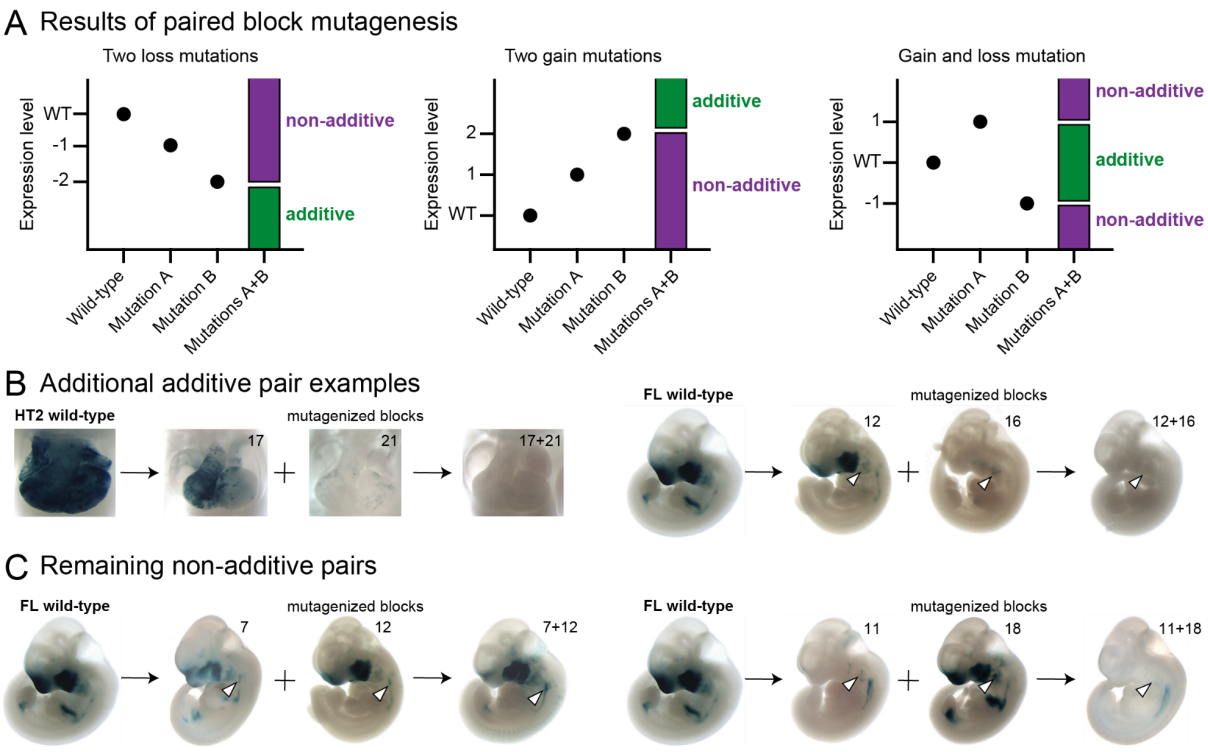
Supplementary Figure 5. Degenerate TFs and alternative models. (A) Contribution score tracks for wild-type sequences and *in silico* mutated constructs which were predicted both to increase the open chromatin signal by at least 25% (log2 fold change > 0.32) and to feature a novel cluster of high, positive scores. Three of six discovered sites were validated experimentally. Two of the unverified sites that overlapped wild-type blocks were classified as false positive predictions (MEF2 in HT2 and MEIS in HT3). (B) Validation of double NFI site predicted in blocks 16-17 of enhancer HT2 by *in silico* saturation mutagenesis. Combined 1bp mutations in SRF site (T160C) and in the predicted double NFI site (C189T) led to a more pronounced loss of function than SRF mutation alone. This validated the double NFI site and led to reassessment of block 16 as (at least) minor loss. Supplement to Figure 3A. (C) Exploration of alternative models for enhancer FL. Block mutations overlapping the same binding motif show very similar activity impacts, with exception of block 4 and 5 (see Supplementary Figure 2 and Supplementary Note 2). (D) Example of total site count for enhancer HT3 (all functional blocks shown). Total site count = all predicted sites – predicted sites in wild-type blocks + blocks without site predictions (4=3-1+2 in this case).



Supplementary Figure 6. Rejected best-fit machine learning models for enhancers NEU1 and NEU2. (A) Correlations between model predictions and *in vivo* results. Dots = mutagenized constructs. Black fit line is linear regression. R^2 is Spearman correlation. Asterisk = non-significant (FDR>0.01). (B) Final TF binding motif and activity map including verified binding motifs discovered through *in silico* saturation mutagenesis (NEU2 PITX/OTX2 site marked with asterisks). (C) Predicted motif disruptions. Note that validation of the GATA motif in blocks 19-20 did not succeed. (D) Discovery and validation of an additional PITX/OTX2 site in enhancer NEU2.



Supplementary Figure 7. Additional examples of multi-tissue responses to mutations. (A) Illustration of paired block mutagenesis outcomes for all possible combinations of loss and gain mutations. Bars represent ranges of possible outcomes that would be classified as additive or non-additive. Redundant is a special case of non-additive in which combined mutagenesis of two blocks resulted in an outcome exactly as severe as the most severe of individual block outcomes. (B) Additional additive pair examples. (C) Remaining three non-additive pairs. White arrowheads indicate loss of function. Black arrowhead indicates gain of function. Related to Figure 4.



Supplementary Figure 8. Additional results of paired block mutagenesis. (A) Illustration of paired block mutagenesis outcomes for all possible combinations of loss and gain mutations. Bars represent ranges of possible outcomes that would be classified as additive or non-additive. (B) Additional additive pair examples. (C) Remaining non-additive pairs. Combined mutagenesis of enhancer FL blocks 7 and 12 resulted in higher branchial arch 3 activity, while no change in activity in these structures was observed in constructs with single block mutations (see also hindbrain activity). Combined mutagenesis of enhancer FL blocks 12 and 18 resulted in lower activity in branchial arch 2 compared to constructs mutated in block 12 only, while mutation of block 18 in isolation did not appreciably change the activity of this structure (compare also hindbrain activity). White arrowheads highlight structures of interest. Related to Figure 5.

Supplementary Notes

Supplementary Note 1: Deterministic transition mutations are the best strategy for eliminating existing TF binding motifs using block mutagenesis

In designing the mutagenesis scheme for this study, we aimed to achieve two goals - reduce the number of experiments necessary to comprehensively map the functional parts of chosen enhancers while retaining a reasonable sequence resolution and to avoid both false positives (calling a functional site in absence of function) and false negatives (calling a site that has function wild-type). We reasoned that mutagenizing sequences in blocks of 12bp, the average size of a TF PWM (Supplementary Figure 1A) strikes a good balance between the resolution and the throughput of the experiment. We speculated this would make it uncommon to deactivate two binding motifs by chance and if such contingency occurred, it would be rare enough to disambiguate using additional targeted mutations.

To choose a block mutagenesis scheme that prevents false negatives that may arise from TF binding motifs being accidentally recreated by mutations, we run an *in silico* experiment. We avoided indel schemes, reasoning that they could lead to changes in activity due to changes in spacing in between TF binding sites, which would make data interpretation difficult. We also avoided "homopolymer schemes", e.g. replacing every basepair in a block with Ts, as that might substantially affect GC-content of the sequence and secondary DNA structure, leading to effects unrelated to changes in TF binding motifs. In the end, we chose to compare various deterministic and random scrambling strategies.

To validate our simulation, we mutagenized all, or every 6th, 3rd and 2nd nucleotide of TF binding site in JASPAR database[41] and found that, as expected, the denser mutation schemes make it less likely for TF binding motifs sequence to retain match with the original PWM (Supplementary Figure 1B-D). We also tested three scramble schemes - simple randomization ('scramble'), randomization in blocks of two nucleotides often used in MPRA experiments ('scramble (di-nt)') and a novel scramble scheme designed to randomize the sequence without recreating any of the 4-mers originally present ('scramble (4-mer)'). As expected, the di-nucleotide scramble was most likely to preserve TF binding motifs match, followed by random and 4-mer scramble. For non-scramble schemes, we tried all three possible deterministic mutations - transitions (A=G, C=T), GC-content preserving transversions (A=T, C=G) and transversions that did not preserve GC-content (A=C, T=G). In line with experimental results[42], we found that the latter transversion scheme had a slight advantage over the other schemes when not mutagenizing every nucleotide. Surprisingly, when mutagenizing all nucleotides, transition scheme was much more potent than the two transversion schemes. Importantly, it was the most effective scheme across a range of TF binding motifs, more effective than 4-mer scramble. We select this scheme for our experiment.

Our simulations did not address the risk of false positives, ie mutagenesis creating a novel site resulting in a false functional call. We reasoned this will require both sacrificing the consistency

of the deterministic scheme as well as an assumption that a good fraction of transcription factor binding motifs involved in the activity of all seven enhancers we have mutagenized are known. We decided that estimating this rate by employing alternative mutagenesis schemes post-factum to functional blocks detected by transition scheme is a better way of addressing this issue (Supplementary Figure 2).

Supplementary Note 2: Transition scheme validation and gain-of-binding events

Transition block mutagenesis was primarily expected to lead to loss of existing binding motifs without creation of new binding motifs. To test this assumption, we used machine learning predictions from best-fit and alternative machine learning models to detect likely gain-of-binding events and conducted an unbiased survey of a selection of blocks using alternative mutagenesis schemes. For simplicity, these results are incorporated into the first section of the manuscript, even though machine learning models are only introduced later.

Machine learning predictions of gain-of-binding and inconsistent staining in blocks overlapping the same predicted binding motif led us to suspect that three transition block mutations caused a simultaneous loss and gain-of-binding. Using alternative GC-preserving transversion mutagenesis scheme (HT2 block 14) and targeted 2bp mutations (FL block 4, NEU3 block 8), we concluded that was likely the case and updated our block assessment accordingly (Supplementary Figure 2A).

An additional unbiased survey of 13 blocks (1 gain, 2 wild-type, 6 minor loss, 4 major loss) using alternative GC-content preserving transversion and scrambling block mutagenesis schemes revealed no major differences, overall validating original transition scheme as primarily causing loss-of-binding (Supplementary Figure 2B). Specifically, in 9/13 cases transition (the default scheme for this study) matched the transversion or scramble result perfectly. In 2 of 4 remaining cases, the difference was very minor and resulted in no change in score. In particular, for FL block 12 scrambling mutagenesis induced a weak gain of heart staining, which is likely explained by an accidental introduction of a GATA motif. The rest of the staining pattern was identical between transition and scrambling mutagenesis. For FL blocks 7 and 10 the differences between transition and transversion mutagenesis were more pronounced, first one changing the direction of effect (from minor loss to minor gain) and the other only changing the magnitude (major loss score 3 to score 2; Supplementary Figure 2B). We conservatively decided to use the transition result as final functional block annotation in all 13 cases.

Finally, classification of HT2 block 16 was updated from wild-type (single block transition mutagenesis result) to minor loss. This was based on the fact that 1bp mutation of a predicted double NFI site overlapping that block had a strong additional loss-of-function effect in combination with 1bp mutation targeting SRF site (see Supplementary Figure 5B).

In conclusion, we updated the classification of four blocks as follows: FL block 4 gain -> minor loss, NEU3 block 8 major loss score 3 to major loss score 1, HT2 block 14 major loss score 2 -> full loss (score 0) and HT2 block 16 wild-type -> minor loss.

Supplementary Note 3: Rejected best-fit models for enhancers NEU1 and NEU2. Alternative NEU3 models.

Two enhancers with brain activity in transgenic assay were not included in the analysis of machine learning models results due to low correlation of model predictions with *in vivo* results (NEU1) or lack of tissue-appropriate models (NEU2). This supplementary note contains additional analysis of these enhancers and their models.

Best model for **NEU1** enhancer was derived from midbrain ATAC-seq dataset at E11.5, with R^2=0.29 (Supplementary Figure 6A, FDR > 0.01, not significant), far below the worst best-fit model included in the main analysis (HT3, R^2=0.5). Log2 fold change predictions from this NEU1 model had a relatively narrow range, with most predictions being above -0.3, compared to best-fit models predictions below -1. This implied limited sensitivity of NEU1 models (Supplementary Figure 6A).

Mutations in different blocks of NEU1 affected different brain structures specifically (fore-, mid- and hindbrain), which could explain poor correlations based on whole-embryo assessment of *in vivo* mutational effect (e.g. mutations that abolished either fore- or hindbrain activity would both be classified as "major loss"). We examined fore-, mid- and hindbrain models, looking for tissue-specific prediction outliers and did not find any that would explain the poor fit (Supplementary Figure 6A). The strongest prediction outliers were shared by all three models and involved gain-of-function mutation of block 21 (singly or together with other blocks), which did not have tissue-specific impact. We conclude tissue-specificity was not the main driver of poor model fit.

For completion, we examined the contribution score predictions of the midbrain NEU1 model. The prediction contained many isolated CG-dinucleotides that did not appear to be TF binding motifs, along with LHX8, E-box, SOX and PITX/OTX2 motif predictions (Supplementary Figure 6B). While LHX8 and E-box sites overlapped a major loss block, the SOX and PITX/OTX2 sites were found within wild-type blocks, calling these limited predictions into question. We speculate NEU1's cell type(s) of activity is poorly represented in whole-tissue samples on which we have built our models, leading to poor correlations and limited predictive power.

All five significantly correlating models of brain-active **NEU2** enhancer were derived from liver and heart tissues (R^2=0.51-0.57), in which this enhancer had no activity. The best neuronal/brain model for this enhancer was based on interneuron 4 cluster of brain scATAC dataset, with (statistically insignificant) R^2 of just 0.1. Models based on bulk ATAC-seq and DHS brain datasets showed even lower correlations.

It could be speculated that NEU2 is active in a non-neuronal cell type that is rare in bulk brain samples, thus neither contributing enough signal to these samples, nor sharing the regulatory logic with the most common cell types in the brain. Further, it could be speculated that the cell type in which NEU2 is active shares some similarities with liver and heart samples. Therefore, it should be theoretically possible to learn aspects of NEU2's function from its best-fit open chromatin liver model (R^2=0.57; Supplementary Figure 6A).

Best-fit model predicted two isolated GATA sites and one GATA1-TAL site, accounting for 5/7 major or full loss blocks (Supplementary Figure 6B). Targeted 1bp mutation of TAL1 part of GATA1-TAL site yielded a similar result to the block mutation 17 that encompassed the TAL1 part (Supplementary Figure 6C, left), supporting the model. However, of the two block mutations overlapping the first isolated GATA site (19-20), only the first one resulted in major loss of function, while the other did not affect the expression of the construct (despite strong prediction of -0.78 log-fold change in signal). Considering the possibility that the results of mutagenizing the second block were confounded by a simultaneous loss of GATA and gain of another (hypothetical) site, we ablated the putative GATA site directly through 1bp mutation. This perturbation resulted in no change of activity (Supplementary Figure 6C, right), strongly arguing against GATA TF binding this site. With virtually all other predicted sites being GATA, this result called the entire model prediction into question. Finally, we also observed that the model made three more strong log2 fold change predictions of block effect (blocks 4, 13 and 18, absolute effects of 0.39 or more), which were not borne out by experimental data, further invalidating the model.

An independent TF motif scan suggested that two adjacent blocks 6-7, which did not have any contribution score predictions in the liver model, could bind a weak PITX/OTX2 site, in line with brain activity of the enhancer. An *in silico* saturation mutagenesis on the e11.5 midbrain ATAC-seq model supported that hypothesis, with PITX/OTX2-like contribution scores appearing upon *in silico* mutation that would strengthen the existing site (Supplementary Figure 6D, A68T). No other sites were discovered in that screen. Experimental introduction of a 1bp mutation designed to destroy the PITX/OTX2 site (A70G) largely recapitulated the effect of block mutations overlapping the PITX/OTX2 site. We conclude that NEU2 is unlikely to share a liver/heart GATA-driven logic, but may use a neuron-like logic, for which we currently lack a suitable model.

Interestingly, the third enhancer active in the brain in our study, **NEU3**, had open chromatin signal in neural tissues, but also in the face and limbs, in which no *in vivo* activity was observed. In other words, NEU3 appeared to be poised in face and limbs and to share some of their functional logic. Models derived from brain tissues made remarkably similar predictions to models derived from face and limb tissues. For example, correlation (R^2) between contribution scores of the best-fit neural NEU3 model to contribution scores from limb and face datasets was 0.57-0.74, compared to correlation with other brain/neural datasets of 0.58-0.88. Conversely, "face E11.5 (ATAC)" (model with highest correlation of 0.8 to *in vivo* results for NEU3 out of face/limb models), had R^2 of 0.54-0.92 to brain/neuronal datasets. This implies the same factors mediate chromatin openness in both tissue types, and that some, yet unidentified factor makes the enhancer active only in the brain (or specifically inactive in the face and limbs).

Altogether, these results indicate that transcriptional activity can sometimes be learned from open chromatin signal at poised loci, but caution and experimental confirmation is needed in such cases. Until tissue-appropriate, activity-based models are trained, this form of "transfer learning" may be practically useful for prioritizing experiments and fine mapping of human variants.

Supplementary Note 4: Systematic assessment of signal change based machine learning model predictions

We used machine learning models primarily to find TF binding motifs, which are revealed by the contributions scores. The models were evaluated on their ability to detect a site in reference sequences consistent with functional annotation of the block. However, the models could theoretically correctly predict the effect of introduced mutations without detecting the presence of a binding motif. Furthermore, in our model assessment we did not take into account the creation of novel binding motifs, which would only be present in the contribution score tracks of the mutated, but not the reference sequence. This Supplementary plementary Note provides an additional analysis of the five enhancers and their best-fit models, using direct model predictions of signal change.

We predicted signal change for each single block transition mutation. Since these predictions are continuous, we binarized them using a threshold. We picked an absolute log2 fold change signal cutoff of 0.32, corresponding to 25% change, so as to correctly classify at least 90% of wild-type blocks. For FL, we used the most extreme (absolute) prediction of the three selected models (limb, hindbrain and glutamatergic neurons). We referred to this prediction set as the "cutoff method" and compared it to the "contribution method", the final set of reference sequence binding motif predictions, which includes alternative models and degenerate binding motifs.

Overall, the cutoff method performed similarly to the contribution method, with higher fraction of correctly classified major loss blocks (78% vs 69%), but lower of minor loss (31% vs 38%) and gain blocks (20% vs 60%; Supplementary Table 5), while maintaining a similar specificity (92% vs 94% of correctly classified wild-type blocks).

functional annotation	blocks	cutoff method	contribution method
gain	5	20%	60%
major loss	32	78%	69%
minor loss	16	31%	38%
wild-type	66	92%	94%

Supplementary Table 5. Machine learning model method comparison. Percentages are fraction of the blocks with a given functional annotation that were correctly classified by each of the methods.

Discrepancies between these two methods involved 16 blocks, primarily in enhancers FL and HT1 (6 blocks each). The majority of discrepancies (9/16) involved cutoff method predicting a change, where contribution method predicted no binding motif. In 7 out of these 9 cases, the cutoff method was correct. This implies models contained some information that could not be extracted using contribution scores. Similar "hidden information" was extracted by us as degenerate TF motifs using saturation mutagenesis (see Figure 3), but the result above implies more remains to be discovered.

The remaining 7 cases involved two blocks with correct contribution prediction (MEIS-TEAD degenerate site in HT1), one with incorrect prediction (degenerate MEF2 site in HT2) and four more complex cases (FL blocks 4-6 and HT2 block 14). One complex example involved mutagenesis of FL block 4. This block mutation likely resulted in a simultaneous creation of a TWIST1/HAND2 activator motif and destruction of a SOX activator motif (see Supplementary Figure 2A). With overall outcome being a gain of function, we speculate that the novel TWIST/HAND2 site contributed more to enhancer activity than was lost by ablation of the SOX site. Targeted destruction of the SOX motif through 2bp mutagenesis confirmed that the "true" functional annotation of this block is minor loss of function. While the cutoff method was technically correct in predicting the outcome of block mutagenesis, the contribution method predicted the functional annotation. Contribution method prediction was correct by accident, since for this method we only considered the presence of an activator SOX motif, but not the gain of TWIST1/HAND2. In another complex case, mutagenesis of HT2 block 14 resulted in destruction of a predicted SRF motif and creation of a novel SP/KFL site. The phenotypic result was major loss of function, implying that SRF contributed more to HT2 activity than the novel SP/KFL motif could compensate for, the opposite of FL block 4 case. The cutoff method incorrectly predicted this mutation will lead to no change of function. Again, the contribution method was correct here by accident, as gain of a novel SP/KFL motif was not taken into account when using this method. If it was, the result would be ambiguous, as one activator site being replaced by another one cannot be easily interpreted in terms of overall direction of change, without making assumptions about relative magnitude of contribution scores.

We conclude that predictions of mutation effects based on signal change ("cutoff method") overall yielded results more closely aligned with outcomes of our experiments than predictions of binding motifs in reference sequences ("contribution method"). This was for the most part due to contribution scores not detecting binding motifs where the model strongly and correctly predicted a change of function. In practice, both methods complement each other, since signal change needs to be interpreted as either gain or loss of binding by the contribution scores and contribution scores may sometimes be unable to extract information available to the model.