FROM BLACK-BOX PREDICTORS TO SCIENTIFIC INSTRUMENTS: BIOLOGICAL RULE FAITHFULNESS IN GENOMIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models are becoming essential tools for AI-driven genomics research. Although their downstream performance has been demonstrated across benchmarks, their black-box nature obscures the biological basis of predictions, limiting assessments of scientific reliability. Here we present a biological rule–based framework for evaluating the scientific reliability of genomic foundation models. Using RNA splicing as a test system, we identify biological component subspaces in frozen genomic foundation models and assess their interactions against experimentally established biological rules. Across six models with distinct architectures and pretraining strategies, we find that similar predictive performance can mask differences in component use and biological rule faithfulness (BRF). To assess whether BRF provides information beyond conventional benchmarks, we examined model pairs with conflicting rankings using measurements reported in independent published experimental studies. Across the evaluated settings, BRF favored the better-performing model in 57.1–88.9% of these comparisons. Grounded in biological rules, our framework reveals differences in underlying representations and task generalization beyond conventional benchmarks. These findings motivate evaluation and training paradigms for developing scientifically reliable foundation models for biological discovery.

1 INTRODUCTION

Genomic foundation models pretrained on large-scale DNA sequence data through self-supervised learning are becoming important tools for AI-driven biological research. Their evaluation, however, still centres on predictive performance in benchmark tasks. High accuracy alone cannot establish whether predictions are based on biological rules or statistical shortcuts. This uncertainty about the basis of model predictions limits their biological interpretation and the use of genomic foundation models as reliable scientific tools.

Interpretability research on genomic foundation models has developed along two broad lines. The first starts with structures discovered within the model. Evo 2 (Brixi et al., 2026), for example, uses sparse autoencoders (SAEs) to extract latent features and subsequently interprets them using biological annotations. Nucleotide dependency analysis (da Silva et al., 2025) characterises relationships between sequence positions through model predictions. The second takes known biological concepts or rules as a reference. TPCAV (Yang & Mahony, 2026) uses concept and control examples to identify representation directions and quantify their contributions to predictions. Sequence tests such as Nullsettes (Jiang et al., 2025) assess model responses to changes expected to disrupt biological function. However, model-driven discovery does not ensure that the recovered features or dependencies cover the key components of a target task or map clearly onto biological concepts. Concept attribution defines these objects more explicitly, but measuring individual contributions does not establish the conditional dependencies or joint effects of different components. Systematic rule testing for a biological process requires both identifying the relevant components and determining whether their functional relationships conform to biological expectations. Although existing studies have advanced analyses at both levels, they have not yet fully brought them together within a single systematic evaluation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

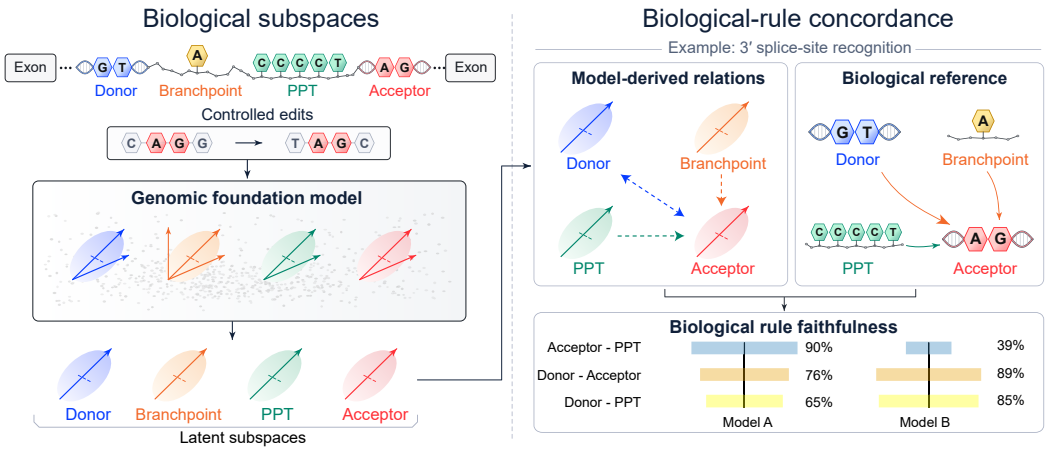


Figure 1: Overview of our evaluation framework. We identify splicing-component subspaces through controlled sequence edits and assess scientific reliability by comparing model-derived component relationships with known biological dependencies.

The Platonic Representation Hypothesis (Huh et al., 2024) proposes that models with different architectures and training objectives may develop similar representations of a shared latent structure. Motivated by this hypothesis, we use known biological concepts as anchors to locate component representations within each model. We examine the roles of these components and their conditional relationships in downstream tasks, testing both against task-specific biological rules. This approach makes biological rule faithfulness (BRF) a measurable dimension of scientific reliability beyond predictive performance. Our main contributions are as follows:

- **We introduce a BRF evaluation framework for assessing the scientific reliability of genomic foundation models.** Using concept-guided subspace discovery and controlled interventions, we examine whether models use and combine component information according to task-specific biological rules. By linking the internal basis of model predictions to these rules, our framework assesses scientific reliability beyond predictive performance.
- **We develop a subspace-based method for evaluating biological component interactions in genomic foundation models.** We identify biological component subspaces and use joint-subspace interventions to assess whether the interactions they encode agree with experimentally measured biological relationships.
- **We show that biological rule faithfulness can inform model evaluation beyond conventional benchmarks.** Using independent published experimental datasets, we find that BRF rankings more often favor models with better performance in the evaluated cases of disagreement. These results support biologically grounded assessment of task-relevant generalization.

2 RELATED WORK

Genomic foundation models and their evaluation. Genomic foundation models aim to acquire biological knowledge through self-supervised pretraining on large-scale genomic sequences and transfer it to downstream applications through task-specific fine-tuning (Ji et al., 2020; Chen et al., 2024; Sanabria et al., 2024; Fishman et al., 2024). DNABERT-2 (Zhou et al., 2024), Nucleotide Transformer (Dalla-torre et al., 2024), HyenaDNA (Nguyen et al., 2023), Evo (Nguyen et al., 2024), Evo 2 (Brixi et al., 2026) and Genos (Lin et al., 2025) have advanced model architectures, training scale and sequence modelling capabilities, supporting increasingly complex biological tasks. Evaluation has likewise expanded from local sequence classification (Grešová et al., 2022) to broader functional tasks. BEND (Marin et al., 2023) covers genomic functional annotation at multiple scales, whereas DART-Eval (Patel et al., 2024) evaluates regulatory feature recognition, activity prediction and responses to counterfactual sequence variants. These benchmarks provide a common basis for comparing downstream performance and help guide model development.

Biological interpretation of genomic foundation models. Mechanistic interpretability seeks to explain the internal computations through which neural networks produce their outputs (Elhage et al., 2021). Two approaches to interpreting genomic foundation models are model-driven feature discovery and concept-driven analysis. Model-driven approaches build on studies of superposition (Elhage et al., 2022) and sparse feature extraction (Cunningham et al., 2023; Templeton et al., 2026; Gao et al., 2024). They use sparse autoencoders (SAEs) to decompose internal activations into features, which are then interpreted using sequence patterns and biological annotations (Nair et al., 2026). Evo 2 (Brixi et al., 2026) and Decode-gLM (Maiwald et al., 2026) identify features associated with regulatory elements and gene structure, and explore their use in functional analysis and prediction steering. Concept-driven approaches instead start from prior knowledge to examine how specified concepts are represented and contribute to predictions. TCAV (Kim et al., 2017) constructs representation directions from concept and control examples, then measures output sensitivity along those directions. TPCAV (Yang & Mahony, 2026) extends this approach to genomic models, enabling attribution to concepts such as motifs, repetitive elements and chromatin accessibility.

Interpreting representations across models. The Platonic Representation Hypothesis (Huh et al., 2024) proposes that large-scale self-supervised training may lead different models towards representations of a shared underlying structure. Cross-model comparisons range from representation similarity measures (Kornblith et al., 2019) to shared neurons (Dravid et al., 2023) and aligned SAE concepts (Thasarathan et al., 2025). In vision, RSVC (Kondapaneni et al., 2025) identifies shared and model-specific concepts and uses concept substitution to examine their relationship to differences in predictions. Chen & Bonner (2024) show that representation dimensions shared more widely across vision models also tend to align more closely with human visual representations. In scientific modelling, Li & Walsh (2025) use reference alignment to reveal shared chemical geometry in interatomic potential models trained with physical supervision. They further link deviations from this geometry to failures in physical prediction. Together, these studies connect shared representations to interpretable domain structure, providing a basis for understanding model differences and evaluating model behaviour.

3 PRELIMINARIES

3.1 RNA SPLICING AND DOWNSTREAM TASKS

During gene expression, DNA is transcribed into a precursor RNA containing both exons and introns. RNA splicing removes the introns and joins the exons to form a mature RNA transcript. Alternative splicing alters splice-site usage or exon inclusion, allowing a single gene to produce distinct mature RNA isoforms.

Splice-site recognition depends on multiple sequence signals, of which we study four core components (Figure 1). Nucleic acid strands have two distinct ends, designated $5'$ and $3'$. Sequences are conventionally written in the $5' \rightarrow 3'$ direction, corresponding to the left-to-right orientation in the figure. Along this direction, the **donor** and **acceptor** mark the beginning and end of an intron, respectively. They are therefore also known as the $5'$ and $3'$ splice sites. In DNA notation, introns typically begin with GT and end with AG, which form the core donor and acceptor signals, respectively. Acceptor recognition also depends on two upstream components, the **branchpoint (BP)** and **polypyrimidine tract (PPT)**. The BP contains a key adenosine (A) that participates in the splicing reaction. The PPT typically lies between the BP and acceptor and is enriched in C and T in DNA notation. Its RNA counterpart binds splicing-associated proteins that facilitate acceptor recognition. The sequence features, relative positions and combinations of these four components jointly influence splice-site selection and exon inclusion.

We use downstream splicing tasks as functional readouts to probe relationships between biological components. By comparing task responses to different combinations of component states, we assess the interactions conveyed through their subspaces.

3.2 COMPONENT SUBSPACES

We aim to identify compact subspaces that capture biological component-associated information in pretrained genomic models. Model-driven approaches such as SAEs discover latent features by

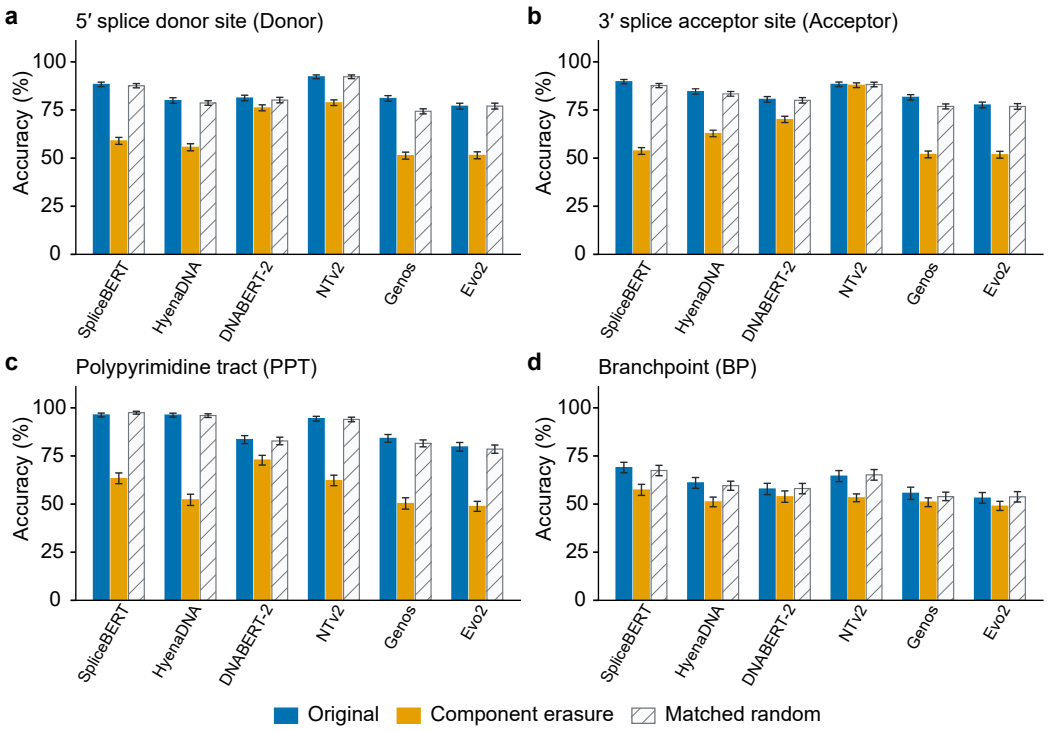


Figure 2: Erasure of the final component subspaces selected according to our selection criteria across six genomic foundation models. Accuracy is compared for original representations, component subspace erasure and matched random subspace erasure for (a) donor sites, (b) acceptor sites, (c) polypyrimidine tracts and (d) branchpoints.

Table 1: Selected component subspaces. In each entry r/ℓ, r is the subspace dimension and ℓ is the selected layer. The original hidden dimension d appears below each model name.

Component	**SpliceBERT** $d = 512$	**HyenaDNA** $d = 256$	**DNABERT-2** $d = 768$	**NTv2** $d = 1024$	**Genos** $d = 1024$	**Evo 2** $d = 4096$
Donor	32 / 1	32 / 1	64 / 3	64 / 3	4 / 1	32 / 1
Acceptor	64 / 1	32 / 1	32 / 1	64 / 1	4 / 1	16 / 1
PPT	64 / 1	4 / 1	64 / 3	32 / 9	64 / 1	32 / 1
BP	32 / 1	32 / 1	32 / 1	64 / 5	32 / 1	32 / 1

decomposing model activations (Cunningham et al., 2023). However, these features may not adequately cover the representational directions associated with a target biological component. We therefore adopt a concept-driven subspace identification strategy inspired by TCAV (Kim et al., 2017). For each component, we construct sequence pairs that share the same background but differ in the target property. We extract the dominant directions of their position-wise representation differences across the full sequence to form a component subspace. Since we focus on component representations that emerge during large-scale self-supervised pretraining, we identify these subspaces in frozen pretrained models. This prevents downstream supervision from altering how component information is encoded and organised. The extraction procedure is detailed below.

For component c, the training set contains n_c sequence pairs $\{(x^+_{c,i}, x^-_{c,i})\}_{i=1}^{n_c}$ of length L. Each pair shares the same sequence background, with the positive and negative labels distinguishing the target component property. Both sequences are passed through the same frozen foundation model. Let $h_{\ell,p}(x) \in \mathbb{R}^d$ denote the hidden representation of sequence x at layer ℓ and DNA position p. Here, d is the hidden dimension and $p = 1, \ldots, L$ indexes positions along the sequence. For models

with multi-nucleotide tokens, each nucleotide position is assigned the hidden representation of its covering token. This mapping is applied separately to each sequence before computing differences at corresponding nucleotide positions. The hidden-state difference between paired sequences at the same layer and position is

$$\Delta h_{c,i,\ell,p} = h_{\ell,p}(x^{+}_{c,i}) - h_{\ell,p}(x^{-}_{c,i}). \tag{1}$$

We assign equal total weight to each sequence background, dividing this weight equally among its perturbation pairs. For pair i, let $m_{c,i}$ denote the number of pairs derived from its background for component c, giving the pair weight $w_{c,i} = 1/m_{c,i}$. We stack the weighted position-wise differences across full sequences to form $D_{c,\ell} \in \mathbb{R}^{n_c L \times d}$, with rows defined by

$$[D_{c,\ell}]_{(i-1)L+p,:} = \sqrt{\frac{w_{c,i}}{L}}\, \Delta h^{\top}_{c,i,\ell,p}. \tag{2}$$

We perform singular value decomposition on this matrix without centring:

$$D_{c,\ell} = U_{c,\ell} \Sigma_{c,\ell} V^{\top}_{c,\ell}, \tag{3}$$

The columns of $U_{c,\ell}$ and $V_{c,\ell}$ are the left and right singular vectors, respectively. The diagonal entries of $\Sigma_{c,\ell}$ are the singular values in descending order. The first r right singular vectors form an orthonormal basis $Q_{c,\ell,r} \in \mathbb{R}^{d \times r}$ for a candidate component subspace. We select representation layers using paired perturbation discrimination, then determine the final subspace dimensions through functional assessment at the selected layers (Section 4.2). For the selected layer ℓ_c and dimension r_c, we denote the resulting basis by $Q_c = Q_{c,\ell_c,r_c}$.

For any input sequence x, the component coordinates at position p and the corresponding orthogonal projection operator are defined as

$$z_{c,p}(x) = Q_c^{\top} h_{\ell_c,p}(x), \qquad \Pi_c = Q_c Q_c^{\top}. \tag{4}$$

Here, $z_{c,p}(x) \in \mathbb{R}^{r_c}$ is the low-dimensional component representation at each position, and $\Pi_c \in \mathbb{R}^{d \times d}$ projects hidden representations onto the component subspace. In this study, we use subspaces spanned by the dominant directions of component-induced changes as the models' internal component representations. Shared biological components serve as semantic anchors, linking internal representations across models to evaluation against biological rules.

4 Discovering and Validating Biological Component Subspaces

In this section, we introduce the rule-driven perturbation dataset, subspace selection criteria, and then the validation of the subspaces' biological relevance and robustness.

4.1 Rule-driven Perturbation Dataset

To reduce non-target interference in component subspace discovery, we constructed background-matched positive–negative pairs through controlled local perturbations of native component sequences. Using GENCODE v49 annotations (Mudge et al., 2024) of GRCh38 and public branchpoint maps (Mercer et al., 2015; Taggart et al., 2017), we extracted 1,001-bp sequences covering donor sites, acceptor sites, candidate polypyrimidine tracts (PPTs) and branchpoints (BPs). For each component, we used local shuffling and constrained substitutions to disrupt known sequence features involved in its recognition (Kondo et al., 2015; Yoshida et al., 2020; Jenkins et al., 2013; Sickmier et al., 2006). Component positions and sequences outside the editing windows remained unchanged. Finally, we obtained 11,405 pairs, with 8,553 used for subspace discovery, 1,770 for validation-based selection and 1,082 for held-out evaluation. Detailed construction rules are provided in the appendix.

4.2 Subspace Selection and Robustness

Building on the subspace discovery procedure in Section 3.2, we next consider subspace selection across models. We examine six representative genomic foundation models with distinct architectures and pretraining strategies: DNABERT-2 (Zhou et al., 2024), NTv2 500M (Dalla-torre et al., 2024),

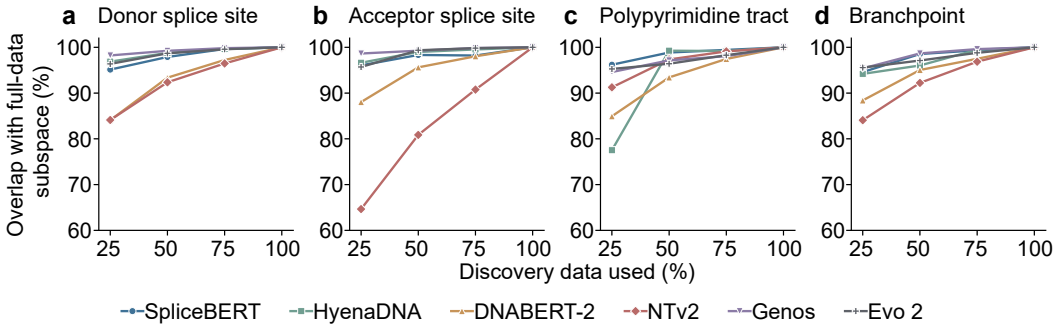


Figure 3: Subspace stability across discovery-data fractions. Subspaces discovered using 25%, 50%, 75% and 100% of the discovery data are compared with the full-data subspace for (a) donor sites, (b) acceptor sites, (c) polypyrimidine tracts and (d) branchpoints.

SpliceBERT (Chen et al., 2024), HyenaDNA small-32k (Nguyen et al., 2023), Evo 2 7B-1M (Brixi et al., 2026) and Genos 1.2B (Lin et al., 2025). With the paired dataset from Section 4.1 fixed, the key choices are the model layer and subspace dimension.

Layer selection. To balance computational cost and depth coverage, we used a fixed, model-specific set of candidate layers, listed in Appendix Table S4. At each candidate layer, we applied the method in Section 3.2 to the paired dataset in Section 4.1, extracting subspaces of 4, 8, 16, 32 and 64 dimensions. We determined component-specific thresholds empirically from the observed distributions of readout performance, as detailed in Appendix B. For each model–component combination, we selected the layer associated with the smallest subspace satisfying the thresholds for both perturbation families. Ties were resolved by validation performance and then layer depth. The selected layer was fixed for subsequent determination of the functional subspace dimension.

Dimension selection. We constructed component-recognition tasks using a separate dataset of natural target-component and non-target sequences to select the subspace dimension at each selected layer. For each component, we trained a classifier on final-layer representations and held its parameters fixed throughout the interventions. We erased candidate component subspaces and propagated the modified representations through the remaining network before classification. Random interventions matched in dimension and perturbation magnitude served as controls. Comparing accuracy under unmodified representations, component erasure and random interventions, we prioritized dimensions producing substantial component-erasure effects while limiting the effects of matched random interventions.

Subspace validation and robustness. Table 1 summarizes the selected component subspaces. In most cases, erasing these subspaces reduced classification accuracy towards chance level and had substantially larger effects than matched random subspace erasure (Figure 2). To assess the stability of the discovered subspaces, we repeated the discovery procedure using 25%, 50%, 75% and 100% of the discovery data. The resulting subspaces generally showed greater overlap with the full-data reference subspaces as more discovery data were used (Figure 3). The appendix provides full layer and dimension selection results with 95% confidence intervals.

Through these steps, we identified and validated stable, biologically meaningful component subspaces across different genomic foundation models. In the next section, we use these subspaces to assess the models' biological rule faithfulness.

5 Evaluating Biological Rule Faithfulness in Genomic Foundation Models

In this section, we introduce our approach to quantifying biological rule faithfulness using the component subspaces identified in each genomic foundation model.

Table 2: Experimental datasets and complete comparisons used to evaluate component relationships. Exon-spanning Acceptor–Donor in COMPASS follows the acceptor–target exon–donor structure; intron-spanning Donor–Acceptor in Schirman follows donor–target intron–acceptor. These represent distinct evaluation targets.

Dataset	Experimental readout	Component relationship	Complete comparisons	Comparisons used
COMPASS	Exon inclusion	Acceptor–PPT	111	51
		Acceptor–Donor (exon)	74	29
		Donor–PPT	221	91
Schirman	Intron removal	Donor–BP	458	240
		BP–Acceptor	575	303
		Donor–Acceptor (intron)	334	126
		Donor–BP–Acceptor	326	185

5.1 Experimental References for Component Interactions

Establishing biological reference relationships among splicing components requires direct experimental evidence. We therefore selected two public datasets with experimental perturbations of the components of interest. For two-component relationships, we assembled four-state comparisons using the same sequence background: reference, A-only, B-only and joint A–B perturbations. Three-component relationships were evaluated using eight-state comparisons. Table 2 summarizes the component relationships and comparison counts.

COMPASS measures exon inclusion following sequence perturbations in human reporter constructs (Koplik et al., 2025). **The Schirman yeast dataset** measures intron excision following combinatorial sequence perturbations (Schirman et al., 2021). In COMPASS, the acceptor–donor comparison spans the target exon; in Schirman, the donor–acceptor comparison spans the target intron. We therefore label these relationships as exon-spanning A–D and intron-spanning D–A, respectively, and do not pool their BRF scores or model rankings.

We define a component interaction by whether the effect of changing one component depends on the state of another. For components A and B, let y_{ab} denote the measured splicing readout at states $a, b \in \{0, 1\}$, respectively. We define J_{bio} as the signed component interaction contrast calculated from these experimental readouts:

$$J_{\mathrm{bio}} = (y_{11} - y_{01}) - (y_{10} - y_{00}). \tag{5}$$

For Acceptor–PPT, J_{bio} compares the effect of the same acceptor change between PPT states 0 and 1, quantifying non-additivity on the measured splicing scale. Here, 0 and 1 denote experimental sequence states, not necessarily component absence or presence. For Donor–BP–Acceptor, we use eight states to compare the Donor–BP interaction between the two acceptor states.

5.2 Measuring Rule Faithfulness through Subspace Interventions

We use state replacement to test whether the identified component subspaces encode relationships consistent with those measured in biological experiments. We keep the pretrained models, selected subspaces and task heads fixed throughout evaluation.

Joint component subspaces and state replacement. For the component set C under evaluation, let Q_c and ℓ_c denote each component's selected basis and layer, as defined in Section 3.2. At each intervention layer ℓ, we form the joint-subspace basis $Q_{C,\ell}$ and its orthogonal projection operator $\Pi_{C,\ell}$:

$$Q_{C,\ell} = \mathrm{orth}([Q_c]_{c \in C:\, \ell_c = \ell}), \qquad \Pi_{C,\ell} = Q_{C,\ell} Q_{C,\ell}^{\top}. \tag{6}$$

Brackets denote column concatenation of the bases assigned to layer ℓ. The operator orth returns an orthonormal basis for their combined span without double-counting shared directions.

Table 3: Biological rule faithfulness (BRF, %) for component relationships evaluated using wet-lab experimental datasets.

Dataset	Relationship	SpliceBERT	HyenaDNA	DNABERT-2	NTv2	Genos	Evo 2
COMPASS	Acceptor–PPT	54.53	48.53	48.65	43.50	54.04	54.29
	Acceptor–Donor (exon)	55.90	57.90	34.86	43.90	63.52	56.48
	Donor–PPT	60.40	40.03	51.61	60.34	48.95	56.05
Schirman	Donor–BP	51.67	54.17	55.42	49.17	44.58	52.92
	BP–Acceptor	48.84	46.86	60.40	48.84	52.15	48.84
	Donor–Acceptor (intron)	39.68	42.86	62.70	54.76	52.38	44.44
	Donor–BP–Acceptor	43.78	43.78	52.43	53.51	58.38	48.11

Let x_s denote the experimental sequence for state $s \in \{00, 10, 01, 11\}$, with the two indices specifying the states of components A and B, respectively. For every source state, we use the same reference sequence x_{00} as the recipient. At every DNA position, we replace its joint-subspace coordinates:

$$\widetilde{h}^{(s)}_{\ell,p} = h^{\mathrm{cur},(s)}_{\ell,p} + \Pi_{C,\ell}\left[h_{\ell,p}(x_s) - h^{\mathrm{cur},(s)}_{\ell,p}\right], \tag{7}$$

Here, $h_{\ell,p}(x_s)$ is the source representation from an unmodified forward pass, and $h^{\mathrm{cur},(s)}_{\ell,p}$ is the recipient's current representation before replacement. The updated representation $\widetilde{h}^{(s)}_{\ell,p}$ is indexed by layer ℓ, DNA position p and source state s. This replacement preserves the current representation outside the joint subspace. All four source states use the same joint subspace, with each state taken from its corresponding experimental sequence. Across layers, replacements follow network order and act on representations propagated from earlier interventions. The resulting final-layer representations are averaged across DNA positions and passed to a fixed task head. Each score f_s denotes the prediction for x_{00} after state-s replacement, rather than directly for x_s.

Interaction contrasts and rule faithfulness. We quantify rule faithfulness as directional agreement between subspace-mediated and experimentally measured interaction contrasts. We compute the model interaction contrast using the same four-state comparison as the biological reference:

$$J_{\mathrm{model}} = f_{11} - f_{10} - f_{01} + f_{00}. \tag{8}$$

For three components, we use eight source states and the corresponding third-order contrast. For comparison i, the agreement indicator $a_{m,i}$ equals 1 when model m's interaction direction matches the experimentally resolved direction. Opposite or numerically zero model responses receive 0.

We define $\mathrm{BRF}_{m,r}$ as the percentage of directional agreement for model m and component relationship r. For COMPASS, we average agreement within each gene and then equally across genes:

$$\mathrm{BRF}_{m,r} = \frac{100}{|\mathcal{G}_r|} \sum_{g \in \mathcal{G}_r} \frac{1}{|\mathcal{I}_{g,r}|} \sum_{i \in \mathcal{I}_{g,r}} a_{m,i}, \tag{9}$$

where $\mathcal{G}_r$ contains the evaluation genes for relationship r, and $\mathcal{I}_{g,r}$ contains comparisons with resolved experimental directions in gene g. For yeast, we average agreement across evaluable comparisons. The appendix details token alignment, intervention procedures and numerical-zero thresholds.

Table 3 summarizes the resulting biological rule faithfulness scores across models and component relationships for both experimental datasets.

6 BRF Reveals Task-Relevant Differences Beyond Benchmarks

We further investigated how BRF complements conventional benchmark evaluations. We focused on models for which BRF and benchmark performance yielded different assessments in specific

Table 4: Independent task evaluation of model pairs with conflicting BRF and benchmark rankings. A BRF win occurs when the model ranked higher by BRF performs better on the independent task.

Dataset	Task	Component relationship	Benchmark reference	Discordant model pairs	BRF wins	Win rate (%)
VARS-seq	Variant spliced fraction	Donor–BP	NT Donor	8	6	75.0
		Donor–BP	GUE splice	9	6	66.7
SPICE	Exon inclusion (PSI)	Acceptor–PPT	NT Acceptor	10	6	60.0
		Acceptor–PPT	GUE splice	7	4	57.1
FAS	Predicting non-additive effects of exon-boundary mutations	Acceptor–Donor (exon)	NT Donor	9	8	88.9
		Acceptor–Donor (exon)	GUE splice	8	7	87.5

settings, and evaluated these cases using measurements from independent published experimental studies.

We used four splicing-related benchmarks: donor recognition, acceptor recognition and combined splice-site classification from the Nucleotide Transformer benchmark (Dalla-torre et al., 2024), and splice-site classification from GUE (Zhou et al., 2024). Independent evaluation used three tasks: splicing-efficiency prediction on VARS-seq (Rangan et al., 2025), exon-inclusion prediction on SPICE (Chen et al., 2025), and prediction of non-additive effects of exon-boundary mutations on FAS (Julien et al., 2016). Appendices H and F provide the benchmark rankings and external evaluation protocols.

Table 4 summarizes the independent task results for model pairs ranked differently by BRF and the corresponding benchmark. These results suggest that, in some real-world biological settings, component-level BRF can better reflect how faithfully models represent biological relationships than conventional benchmark performance alone.

7 Discussion

Representational depth. Most selected component subspaces occurred in early layers, suggesting that core splicing information is accessible at shallow depths. This motivates testing whether deeper layers integrate these features into more complex biological representations, analogous to hierarchical feature composition in vision models (Yosinski et al., 2015).

Origins of BRF differences. Differences in BRF may reflect pretraining data, architecture, training objectives or their interactions. Controlled comparisons are needed to distinguish their contributions to learning biological relationships.

Fine-tuning and future training strategies. Appendix analyses suggest partial subspace stability after fine-tuning, alongside marked increases or decreases in component decodability and erasure effects. Such changes motivate investigating how fine-tuning alters component use, including possible reliance on dataset-specific shortcuts. Component-subspace interventions could help guide fine-tuning and pretraining towards representations consistent with biological rules.

8 Conclusion

We introduced biological rule faithfulness, a framework linking black-box model representations to experimentally established biological rules through component subspaces. Across six models, we found differences in rule faithfulness, with rankings that better matched independent task performance than benchmark rankings in selected settings. This probes one aspect of generalization to independent biological tasks using experimentally grounded rules. Our findings motivate evaluating the scientific basis of model predictions to advance reliable AI tools for scientific research.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, literature search and analytical discussions. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This study uses only publicly available data. We have identified no additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Experimental results and methodological details are provided in the appendix. We plan to release the dataset constructed in this study upon publication. During peer review, subsets of the dataset can be made available to reviewers upon request.

REFERENCES

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam N. Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary Amador, Euan A. Ashley, Stephen A. Baccus, Hao-Yu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Josephine or Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony Costa, Tina Hernandez-Boussard, Eric Dun Ho, Mingying Liu, Tom McGrath, Kimberly R. Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with evo 2. *Nature*, 652:1349 – 1361, 2026.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhi-Xiang Ren, and Yue-Dong Yang. Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction. *Briefings in Bioinformatics*, 25, 2024.

Xi Dawn Chen, Maile Jim, Mounica Vallurupalli, Kai Cao, Andrea Navarro Torres, Jing Wesley Leong, Yifan Zhang, David Wollensak, Qiyu Gong, Jing Sun, Mehdi Borji, Gail Schor, Sofia Mrowka, Margaret Hu, Anisha Laumas, Jennifer A. Roth, Todd R. Golub, and Fei Chen. Generative design of cell type-specific rna splicing elements for programmable gene regulation. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:282813302.

Zi-Rui Chen and Michael F. Bonner. Universal dimensions of visual representation. *Science Advances*, 11, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600, 2023.

Pedro Tomaz da Silva, Alexander Karollus, Johannes C. Hingerl, Gihanna Sta. Teresa Galindez, Nils Wagner, Xavier Hernández-Alias, Danny Incarnato, and Julien Gagneur. Nucleotide dependency analysis of genomic language models detects functional elements. *Nature Genetics*, 57:2589 – 2602, 2025. URL https://api.semanticscholar.org/CorpusID:282013369.

Hugo Dalla-torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo Pinto de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin J. Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287 – 297, 2024.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1934–1943, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Thomas Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy models of superposition. *ArXiv*, abs/2209.10652, 2022.

Veniamin S. Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, N. Chekanov, Olga L. Kardymon, and Mikhail E. Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53, 2024.

Leo Gao, Tom Dupr'e la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *ArXiv*, abs/2406.04093, 2024.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Simecek, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *ArXiv*, abs/2405.07987, 2024.

Jermaine L. Jenkins, Anant A. Agrawal, Ankit Gupta, Michael R. Green, and Clara L. Kielkopf. U2af65 adapts to diverse pre-mrna splice sites through conformational selection of specific and promiscuous rna recognition motifs. *Nucleic Acids Research*, 41:3859 – 3873, 2013.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *bioRxiv*, 2020.

Shi-Yu Jiang, Xuying Liu, and Zitong Jerry Wang. Evaluating dna function understanding in genomic language models using evolutionarily implausible sequences. *ACS Synthetic Biology*, 15:2256 – 2263, 2025. URL https://api.semanticscholar.org/CorpusID:279319117.

Philippe Julien, Belén Miñana, Pablo Baeza-Centurión, Juan Valcárcel, and Ben Lehner. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nature Communications*, 7, 2016. URL https://api.semanticscholar.org/CorpusID:288406953.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Jun Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2017.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. *ArXiv*, abs/2503.15699, 2025.

Yasushi Kondo, Chris Oubridge, A. M. M. van Roon, and Kiyoshi Nagai. Crystal structure of human u1 snrnp, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *eLife*, 4, 2015.

Samantha E. Koplik, Angela M. Yu, Madelyn R. Shelby, Gabriel C Fonseca, Charles M. Roco, Yue Zhang, Nicholas Bogard, Alex K. Sabo, Alexander B. Rosenberg, Johannes Linder, and Georg Seelig. Massively parallel assay of human splice variants reveals cis-regulatory drivers of disease-associated and cell type-specific splicing regulation. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:282310762.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. *ArXiv*, abs/1905.00414, 2019.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, 8:830 – 840, 2025.

Adi Lin, Bin Xie, Celina Dong Ye, Cheng Wang, Duoyuan Chen, Ercheng Wang, Fan Lu, Guirong Xue, Hai-Qiang Zhang, Jiajie Zhan, Jianfeng Zhang, Jiangshuan Pang, Jianqiang Liang, Jiawei Lin, Jiaxin Ma, Jie Hu, Jing Ma, Jin ling Dong, Jiong-Zheng Li, Junchen Liu, Junhong Chen, Junyou Li, Kai Ding, Kaiwen Deng, Kui Chen, Lihui Wang, Longqi Liu, Ling Guo, Liwen Xiong, Luhao Yang, Minghan Cheng, Natalie Chen, Ren-Zhong Chen, Shanxin Sun, Shaoshuai Li, Shi-Cheng Chen, Shiping Liu, Siwei Xie, Suyan Liu, Tao Zhou, Wangyang Tang, Weiqiang Zhang, Xianyue Jiang, Xianzhi Qi, Xin Jin, Xinjiang Tan, Xinyue Hu, Xun Xu, Xu-Yang Feng, Ya-Fei Lu, Yifan Gao, Yong Shang, Youzhe He, Yue Yuan, Yu-Fan Wang, Yuqi Liu, Zhan Xiao, Zhangyuan Meng, Zhaorong Li, Zhe Zhao, Zhen Yang, and Zi-Lin Wang. Genos: a human-centric genomic foundation model. *GigaScience*, 14, 2025.

Aaron Maiwald, Piotr Jedryszek, Florent Draye, Bernhard Schölkopf, Garrett M. Morris, and Oliver M. Crook. Decode-glm: Tools to interpret, audit, and steer genomic language models. *bioRxiv*, 2026.

Frederikke I. Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. Bend: Benchmarking dna language models on biologically meaningful tasks. *ArXiv*, abs/2311.12570, 2023.

Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. G. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars Keld Nielsen, M. Dinger, and John S. A. Mattick. Genome-wide discovery of human splicing branchpoints. *Genome Research*, 25:290 – 303, 2015.

Jonathan M. Mudge, Sílvia Carbonell-Sala, Mark E. Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E. Loveland, Carme Arnan, If H. A. Barnes, Ruth Bennett, Andrew E. Berry, Alexandra Bignell, Daniel Cerdán-Vélez, Kelly Cochran, Lucas T. Cortés, Claire Davidson, Sarah M. Donaldson, Cagatay Dursun, Reham Fatima, Matthew Hardy, Prajna Hebbar, Zoe Hollis, Benjamin T. James, Yunzhe Jiang, Rory Johnson, Gazaldeep Kaur, Mike P. Kay, Riley J. Mangan, Miguel Maquedano, Laura Martínez Gómez, Nourhen Mathlouthi, Ryan Merritt, Pengyu Ni, Emilio Palumbo, Tamara Perteghella, Fernando Pozo, Shriya Raj, Cristina Sisu, Emily Steed, Dulika Sumathipala, Marie-Marthe Suner, Barbara Uszczynska-Ratajczak, Elizabeth Wass, Yucheng T. Yang, Dingyao Zhang, Robert D. Finn, Marc H. Gerstein, Roderic Guigó, Tim Hubbard, Manolis Kellis, Anshul B. Kundaje, Benedict Paten, Michael L. Tress, Ewan Birney, Fergal J. Martin, and Adam Frankish. Gencode 2025: reference gene annotation for human and mouse. *Nucleic Acids Research*, 53:D966 – D975, 2024.

Akira A Nair, Jaehyun Joo, Jonghyun Lee, Lina Takemaru, Yidi Huang, Manu Shivakumar, Matthew Eric Lee, Jaesik Kim, Sokratis Apostolidis, and Dokyoon Kim. Interpreting genomic language models using sparse autoencoders. In *Forty-third International Conference on Machine Learning*, 2026.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton M. Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *ArXiv*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Adanna Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science (New York, N.Y.)*, 386:eado9336 – eado9336, 2024.

Aman Patel, Arpita Singhal, Austin T. Wang, Anusri Pampari, Maya M Kasowski, and Anshul B. Kundaje. Dart-eval: A comprehensive dna language model evaluation benchmark on regulatory dna. *ArXiv*, 2024.

Ramya Rangan, Rui Huang, Oarteze Hunter, Phillip Pham, Manny Ares, and Rhiju Das. Comprehensive analysis of saccharomyces cerevisiae intron structures in vivo. *Nature Structural & Molecular Biology*, 32:1488 – 1502, 2025. URL https://api.semanticscholar.org/CorpusID:279164872.

Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6:911 – 923, 2024.

Dvir Schirman, Zohar Yakhini, Yitzhak Pilpel, and Orna Dahan. A broad analysis of splicing regulation in yeast using a large library of synthetic introns. *PLoS Genetics*, 17, 2021. URL https://api.semanticscholar.org/CorpusID:238204610.

E. Allen Sickmier, Katherine E. Frato, Haihong Shen, Shanthi R. Paranawithana, Michael R. Green, and Clara L. Kielkopf. Structural basis for polypyrimidine tract recognition by the essential pre-mrna splicing factor u2af65. *Molecular cell*, 23 1:49–59, 2006.

Allison J. Taggart, Chien-Ling Lin, Barsha Shrestha, Claire Heintzelman, Seong Ki Kim, and William G. Fairbrother. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Research*, 27:639 – 649, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian K. Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte Stuart MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam S. Jermyn, Shan Carter, Chris Olah, and Thomas Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *ArXiv*, abs/2605.29358, 2026.

Harrish Thasarathan, Jonathon Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *ArXiv*, abs/2502.03714, 2025.

Jianyu Yang and Shaun Mahony. Tpcav: Interpreting deep learning genomics models via concept attribution. *bioRxiv*, 2026.

Hisashi Yoshida, Sam-Yong Park, Gyosuke Sakashita, Yuko Nariai, Kanako Kuwasako, Yutaka Muto, Takeshi Urano, and Eiji Obayashi. Elucidation of the aberrant 3′ splice site selection by cancer-associated mutations on the u2af1. *Nature Communications*, 11, 2020.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015. URL https://arxiv.org/abs/1506.06579.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations*, 2024.

Table S1: Sources of the native sequence backgrounds. All sequences were extracted from GRCh38; the annotations specify the component anchors.

Component	Anchor or region	Annotation source
Donor	Canonical intron start	GENCODE v49 exon-derived introns (Mudge et al., 2024)
Acceptor	Canonical intron end	GENCODE v49 exon-derived introns (Mudge et al., 2024)
PPT	Candidate region upstream of the acceptor	GENCODE v49 acceptor-anchored sequence; no separate experimental PPT annotation
BP	Published branch adenine	Mercer Data S2/S3 and Taggart Table S1, mapped from hg19 (Mercer et al., 2015; Taggart et al., 2017)

A Construction of the Rule-driven Perturbation Dataset

This section details the dataset construction underlying Section 4.1 of the main paper. We used native genomic sequences to construct background-matched pairs that differed in sequence features associated with a target splicing component. The final dataset contained 11,405 pairs spanning donor sites, acceptor sites, candidate polypyrimidine tracts (PPTs) and branchpoints (BPs).

A.1 Data Sources and Native Sequence Extraction

Reference sequences and splice-site annotations. We extracted human reference sequences from GRCh38 and obtained exon annotations from GENCODE release 49 (Mudge et al., 2024). For each transcript, we inferred introns from adjacent annotated exons and merged records sharing the same intron coordinates and strand. We excluded introns assigned to more than one gene. Eligible introns were at least 80 bp long and had canonical GT–AG boundaries in the strand-oriented reference sequence. Donor and acceptor anchors were defined by these annotated intron boundaries. Candidate PPT regions were defined upstream of the acceptor, rather than obtained from a separate collection of experimentally validated PPTs.

Branchpoint annotations. We combined the high-confidence BP annotations and intron associations of Mercer et al. (2015) with Supplemental Table S1 of Taggart et al. (2017). The Mercer records were drawn from Supplemental Data S2 and the associated intron records in Data S3. For Taggart records, we retained canonical calls with an adenine branch nucleotide, at least two unique supporting reads and a mutation in at least one read. We mapped the published hg19 coordinates to GRCh38 using the hg19-to-hg38 chain file. Accepted mappings were unique and preserved chromosome and orientation. Mercer sites additionally required agreement between the mapped donor–acceptor boundaries and the current intron annotation, with an adenine at the mapped BP. For Taggart sites, we required exact agreement of the reported local motif with the reference and a unique compatible annotated acceptor–intron match. Coordinate offsets for the Taggart table were resolved consistently within each source strand using motif and junction agreement, rather than adjusted independently for individual sites. We merged coincident BP coordinates across the two sources while retaining their source support.

Sequence extraction. Each background was a 1,001-bp reference window extending 500 bp on either side of its anchor. We reverse-complemented negative-strand windows so that every sequence followed the transcript's $5' \rightarrow 3'$ orientation. The donor anchor was the first intronic G; the acceptor anchor was the first exonic nucleotide immediately downstream of AG. PPT backgrounds used the acceptor anchor, whereas BP backgrounds were centred on the annotated branch adenine. We excluded windows outside the available reference or containing bases other than A, C, G and T.

A.2 Paired Perturbation Design

For each component, the native sequence served as the reference member of a pair. We generated its counterpart by changing a short component-associated window while retaining all bases outside

that window. The edits introduced no insertions or deletions and did not shift the coordinate anchor. Positive and negative states denote the intended recognition-feature contrast; they are not measured labels of complete component activity or inactivity.

We used two complementary perturbation families. *Composition-preserving shuffling* rearranged existing bases among editable positions, preserving the local A/C/G/T counts and therefore the counts of the complete sequence. *Constrained substitutions* changed base identities subject to the component-specific conditions below. This second family intentionally allowed changes in composition. For donor, acceptor and BP backgrounds with a valid shuffled counterpart, the substitution counterpart used the same number of changed positions. Otherwise, an admissible substitution counterpart could be retained on its own. PPT substitutions instead used the fixed six-pyrimidine depletion rule in Section A.3.

Each background contributed at most one pair from each family. If neither family produced an admissible counterpart, that selected background contributed no discovery pair and was not replaced by another background. The successful pairs from a background received equal weights summing to one: one for a single pair or one-half for each of two pairs. Thus, a background with two admissible edits did not receive twice the discovery weight of a background with only one.

We protected other BP nucleotides recorded in the available annotations wherever they overlapped an editing window. The target BP adenine was exempted when perturbing BP itself. For donor, acceptor and BP edits, we also rejected newly created AG or GT dinucleotides within the window or across its boundaries. PPT edits used the motif-preservation conditions specified below. These constraints preserve the shared sequence background without requiring biologically related component representations to be disjoint.

A.3 Component-specific Perturbation Rules

We expressed sequence rules in DNA notation, using T in place of RNA U. For splice sites, negative positions are upstream of the junction and positive positions downstream, with no position zero. For BP, position zero is the branch adenine. Sequence indices in Table S2 are zero-based and intervals are half-open. Index 500 corresponds to donor $+1$, acceptor $+1$, or BP 0, depending on the background. The component features were motivated by the cited biological studies; the numerical editing thresholds were choices of the perturbation design.

Donor. Recognition of the $5'$ splice site involves pairing between the splice-junction sequence and U1 snRNA (Kondo et al., 2015). We edited the nine-base window from positions -3 to $+6$, which includes the core GT at $+1, +2$. We counted matches at seven neighbouring positions: $-3 \in \{A, C\}$, $-2 = A$, $-1 = G$, $+3 \in \{A, G\}$, $+4 = A$, $+5 = G$ and $+6 = T$. An admissible counterpart replaced the core GT with a dinucleotide other than GT or GC and reduced the total neighbourhood match count by at least two. It also differed from the native sequence at four or more positions and introduced no new local AG or GT. Both shuffled and substituted counterparts satisfied these conditions. Observed edits changed four to nine bases.

Acceptor. The AG dinucleotide and adjacent bases contribute to $3'$ splice-site recognition by U2AF1 (Yoshida et al., 2020). We edited the six-base junction window from -3 to $+3$, containing AG at $-2, -1$. An eligible native background had at least one of two proximal features: a pyrimidine at -3 or G at $+1$. In an admissible counterpart, the core was no longer AG, position -3 was neither C nor T, and position $+1$ was not G. At least four bases had to change, without creating a new local AG or GT. Both perturbation families therefore altered the core together with proximal recognition features, rather than changing AG alone. Observed edits changed four to six bases.

Polypyrimidine tract. PPT recognition depends on pyrimidine identity and sequence organization, as shown by studies of U2AF65–RNA recognition (Jenkins et al., 2013; Sickmier et al., 2006). We edited the 17-base interval from acceptor positions -20 to -4, leaving the core AG unchanged. The longest consecutive run of C/T bases was evaluated over -20 to -3, including the unchanged base immediately adjacent to the editing window. For the shuffling family, we preserved the A/C/G/T counts and required this run to shorten by at least two bases. We additionally preserved the positions and identities of AG/GT dinucleotides whose start indices lay between 470 and 500, inclusive. For the depletion family, we replaced exactly six editable C/T bases with A or G, reduc-

Table S2: Editing windows and required changes. Bounds in the index column are zero-based and half-open. All edits preserve sequence length and bases outside the editing window.

Component	Relative positions	Sequence indices	Required feature change
Donor	-3 to $+6$	$[497, 506)$	Core neither GT nor GC; neighbourhood match count decreases by ≥ 2; ≥ 4 changed positions.
Acceptor	-3 to $+3$	$[497, 503)$	Core no longer AG; loss of -3 pyrimidine and $+1$ G features; ≥ 4 changed positions.
PPT	-20 to -4	$[480, 497)$	Composition-preserving rearrangement or depletion of six pyrimidines; consecutive C/T run shortens by ≥ 2.
BP	-5 to $+1$	$[495, 502)$	Central A becomes C/T; neighbourhood match count decreases by ≥ 2; native -2 T is changed; ≥ 4 changed positions.

ing the pyrimidine count by six. The same minimum run-shortening condition applied, and no new AG or GT could appear within the editing window or across its boundaries. Shuffled counterparts changed two to fifteen positions in the resulting dataset; depletion counterparts changed exactly six.

Branchpoint. Published BP maps establish the predominance of branch adenine and the contribution of surrounding sequence features to branch-site recognition (Mercer et al., 2015; Taggart et al., 2017). We edited positions -5 to $+1$ around the target BP. The neighbourhood match count used five conditions: $-5 \in \{C, T\}$, $-3 \in \{C, T\}$, $-2 = T$, $-1 \in \{A, G\}$ and $+1 \in \{C, T\}$. An admissible counterpart changed the central A to C or T and reduced this match count by at least two. If the native -2 base was T, that base also had to change. At least four positions were modified, without creating a new local AG or GT or altering another protected BP. Both shuffled and substituted counterparts satisfied these conditions; observed edits changed four to seven bases.

A.4 Dataset Splits and Statistics

Backgrounds retained their chromosome-based partition throughout dataset construction. The training chromosomes were chr3, chr5, chr7, chr11, chr15, chr17 and chr19. The validation chromosomes were chr8 and chr20; the test chromosomes were chr1, chr2 and chr22. All variants derived from a background remained in its original partition. Gene identifiers and the inherited sequence-cluster assignments were disjoint across partitions. We also checked that neither complete sequences nor their reverse complements occurred in different partitions.

We selected 1,200 training, 250 validation and 150 test backgrounds per component before applying the final perturbation constraints. Some backgrounds yielded only one perturbation family, whereas others yielded neither. The resulting collection comprised 11,405 pairs from 6,083 component-specific backgrounds (Table S3). Across components, these backgrounds represented 5,034 distinct genes: 3,706 in training, 771 in validation and 557 in testing. A gene could contribute different component backgrounds within the same partition, so component-background counts should not be interpreted as globally distinct gene counts.

The BP subset contained 1,402 successful backgrounds. Of these, 1,362 were supported only by Mercer, 32 by both studies, and eight only by Taggart. The combined collection contained 5,328 shuffled pairs and 6,077 substitution or depletion pairs. Only the 8,553 training pairs entered subspace discovery; the 1,770 validation pairs were reserved for selection and the 1,082 test pairs for held-out evaluation. The separate matched-control and core-preserving validation panels were not included in these counts or used for discovery.

B Candidate Subspaces and Layer Selection

This section details the paired-validation procedure used to select a representation layer for each model and component. The procedure jointly identifies a layer and the smallest candidate dimension that meets the paired-validation criterion. The selected layer is then fixed for subsequent func-

Table S3: Discovery dataset statistics. Parentheses indicate successful native backgrounds, counted separately for each component. Family totals include all three partitions.

Component	Training	Validation	Test	Total pairs
Donor	2,377 (1,192)	499 (250)	295 (149)	3,171
Acceptor	2,230 (1,140)	457 (235)	275 (143)	2,962
PPT	1,953 (1,177)	412 (245)	253 (150)	2,618
BP	1,993 (1,057)	402 (208)	259 (137)	2,654
Total	8,553 (4,566)	1,770 (938)	1,082 (579)	11,405

Component	Composition-preserving shuffling	Substitution / depletion
Donor	1,580	1,591
Acceptor	1,444	1,518
PPT	1,052	1,566
BP	1,252	1,402
Total	5,328	6,077

Table S4: Representation dimensions and candidate layer indices used for paired-validation selection. The same layer grid was evaluated for all four components within each model.

Model	Hidden dimension d	Candidate layers
SpliceBERT	512	1, 3, 6
HyenaDNA	256	1, 3, 4
DNABERT-2	768	1, 3, 6, 9, 12
NTv2	1,024	1, 3, 5, 6, 9, 12, 15, 18, 21, 24, 27, 29
Genos	1,024	1, 2, 3, 6, 9, 12
Evo 2	4,096	1, 3, 5, 6, 9, 12, 15, 16, 18, 21, 24, 27, 30, 32

tional comparisons across dimensions. The dimension selected here measures retention of paired perturbation discrimination, rather than defining the final dimension for functional interventions.

B.1 Candidate Representations and Subspaces

We applied the discovery procedure in Section 3 of the main paper to six genomic foundation models. Table S4 lists the evaluated layers and representation dimensions. All backbone parameters remained frozen. For each reference–edited pair, we obtained representations aligned to all 1,001 nucleotide positions and computed their position-wise difference at each candidate layer. We did not average positions before constructing the discovery matrix.

Writing this difference as $\Delta H_i^{(\ell)} \in \mathbb{R}^{1001\times d}$, we accumulated the uncentred second moment

$$C_{c,\ell} = \frac{1}{\sum_i w_i} \sum_i \frac{w_i}{1001} \big(\Delta H_i^{(\ell)}\big)^\top \Delta H_i^{(\ell)}. \tag{S1}$$

Only training pairs contributed to this matrix. The weights of all admissible pairs from one background summed to one, as described in Section A.2. The leading eigenvectors of $C_{c,\ell}$ are equivalent to the right singular vectors of the weighted, stacked position-wise differences. We retained 64 directions per layer and formed nested bases $Q_{c,\ell,r}$ with $r \in \{4, 8, 16, 32, 64\}$. Thus, r denotes the number of retained representation directions at each position, not the number of scalar features for the entire sequence.

B.2 Paired Direction Readout

We evaluated each candidate using a regularized linear readout of the projected pair difference,

$$v_i = \mathrm{vec}\Big(\Delta H_i^{(\ell)} Q_{c,\ell,r}\Big) \in \mathbb{R}^{1001r}. \tag{S2}$$

Separate readouts were fitted for the two perturbation families of each component. For donor, acceptor and BP, these were shuffling and constrained substitution; for PPT, they were shuffling and pyrimidine depletion. The readout assessed whether unseen reference-minus-edited differences followed the direction learned from training pairs. It was not a classifier of component presence in individual natural sequences.

For each family and candidate, let X contain the training vectors $v_i^\top$ as rows. We fitted a ridge readout without an intercept, assigning the target value $+1$ to each reference-minus-edited difference:

$$\widehat{\beta}_\lambda = \operatorname*{argmin}_\beta \|X\beta - \mathbf{1}\|_2^2 + \lambda\eta\|\beta\|_2^2, \qquad \eta = \max\left\{\frac{\operatorname{tr}(XX^\top)}{n}, 10^{-24}\right\}, \tag{S3}$$

where n is the number of training pairs in that family. The readout used equal weight per pair within a family; background weighting applied to the discovery matrix in Equation S1. Neither the backbone nor the subspace basis was updated when fitting the readout.

We selected λ from $\{0.01, 0.1, 1, 10, 100\}$ by validation direction accuracy, preferring the larger value when accuracies tied. A validation score $v_i^\top \widehat{\beta}_\lambda > 10^{-10}$ received one point; an absolute score at most 10^{-10} received half a point. All remaining scores received zero, and accuracy was the mean across validation pairs. We fitted the same type of readout to the uncompressed, flattened differences at every candidate layer. These full-dimensional readouts provided an empirical comparison for compression, rather than a theoretical upper bound on component information.

B.3 Eligibility and Selection Rule

For each model and component, let $a_{\ell,r,f}$ denote validation direction accuracy for layer ℓ, dimension r and perturbation family f. We first obtained the best full-dimensional validation accuracy for each family across candidate layers:

$$a_f^{\text{full},*} = \max_\ell a_{\ell,\text{full},f}. \tag{S4}$$

A candidate was eligible only if both families met their respective thresholds:

$$a_{\ell,r,f} \geq \tau_{c,f} \quad \text{for every } f, \qquad \tau_{c,f} = \begin{cases} a_f^{\text{full},*} - 0.01, & c \in \{\text{Donor, Acceptor, BP}\}, \\ 0.95\, a_f^{\text{full},*}, & c = \text{PPT}. \end{cases} \tag{S5}$$

The first condition permits an absolute loss of one percentage point. The PPT condition retains at least 95% of the corresponding best full-dimensional accuracy; it does not require an absolute accuracy of 95%.

Among eligible candidates, we first selected the smallest dimension. Ties were resolved by the higher minimum accuracy across families, then the higher mean accuracy, and finally the shallower layer. Subspace discovery used the training partition, whereas regularization and layer–dimension selection used the validation partition. Neither the paired test partition nor the separate semantic or functional evaluations entered this selection rule.

B.4 Layer-selection Results

All 24 model–component combinations had an eligible candidate within the evaluated grids (Table S5). The smallest eligible dimension was four for 20 combinations, eight for three, and sixteen for one. DNABERT-2 selected eight dimensions for acceptor and BP; NTv2 selected eight for PPT, and Genos selected sixteen for PPT. Selected layers ranged from layer 1 to layer 9. The complete layer–dimension comparisons are reported in Tables S6–S11, rather than only at the selected layers. These results identify compact spaces that retain the specified paired direction discrimination within the candidate grids. They do not imply that component information is confined to the selected layer or exhausted by the paired-validation dimension.

Component-associated differences were accessible in shallow layers. Layer 1 was selected for 19 of the 24 model–component combinations (Table S5). The full candidate grids also showed strong early-layer discrimination at a fixed dimension, providing evidence beyond the distribution of selected layers. For example, DNABERT-2 acceptor accuracy at rank four decreased from 97.45%

Table S5: Layer and smallest eligible dimension selected by paired validation. Each entry is layer ℓ / dimension r. These dimensions are not the final functional-intervention dimensions.

Model	Donor	Acceptor	PPT	BP
SpliceBERT	1 / 4	1 / 4	1 / 4	1 / 4
HyenaDNA	1 / 4	1 / 4	1 / 4	1 / 4
DNABERT-2	3 / 4	1 / 8	3 / 4	1 / 8
NTv2	3 / 4	1 / 4	9 / 8	5 / 4
Genos	1 / 4	1 / 4	1 / 16	1 / 4
Evo 2	1 / 4	1 / 4	1 / 4	1 / 4

at layer 1 to 72.97% at layer 12, taking the lower accuracy across families (Table S8). For Genos donor, the corresponding values were 100% at layer 1 and 82.73% at layer 12 (Table S10). For PPT, rank-four accuracy was lower at the deepest evaluated layer than at layer 1 in all six models (Tables S6–S11). These comparisons indicate that early layers often exposed the tested component-associated changes in a compact, linearly readable form.

Depth effects depended on the model and component. Strong paired discrimination was not restricted to shallow layers. SpliceBERT retained 100% accuracy for both families at rank four across all evaluated layers for donor, acceptor and BP (Table S6). HyenaDNA had eligible candidates at every evaluated layer for all four components, although some deeper layers required larger dimensions (Table S7). NTv2 retained eligible donor and acceptor candidates throughout its layer grid, while its selected PPT and BP layers were 9 and 5 (Table S9). In DNABERT-2, the full-dimensional acceptor readout at layer 12 reached 99.15% for the weaker family, despite the lower rank-four accuracy (Table S8). Thus, reduced low-rank discrimination did not necessarily indicate loss of the underlying perturbation information. Evo 2 showed a narrower range of eligible layers for PPT than for donor, acceptor or BP (Table S11).

Together, these results support early availability of compact component-associated representations, rather than an exclusive concentration of component information in shallow layers. The final shallow-layer tie-break also contributed to the selection pattern when candidates had identical dimension and validation accuracy. The layer comparisons concern linear discrimination of the specified perturbations; they do not measure the total component information encoded at each depth.

B.5 Complete Candidate-layer Validation Results

Tables S6–S11 report every evaluated layer and dimension, including the full-dimensional comparison. Each cell gives validation direction accuracy (%) for shuffling / substitution; for PPT, the second value denotes pyrimidine depletion. Each value uses the regularization coefficient selected for that candidate and family. Bold cells identify the selected layer–dimension pair; a superscript + marks other candidates meeting both family-specific thresholds. Eligibility and tie-breaking use unrounded values, whereas displayed accuracies are rounded to two decimal places. The full-dimensional column is a comparison, not an eligible low-dimensional candidate. These are validation results used for selection, not independent test results.

Table S6: Complete paired-validation results for SpliceBERT. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
Acceptor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
PPT; thresholds: 93.87 / 95.00%						
1	**94.64 / 100.00**	98.21 / 100.00$^+$	97.62 / 100.00$^+$	97.02 / 100.00$^+$	98.81 / 100.00$^+$	98.81 / 100.00

Continued on next page

Table S6 (continued)

Layer	4	8	16	32	64	Full
3	89.29 / 100.00	91.07 / 100.00	91.67 / 100.00	94.64 / 100.00$^+$	96.43 / 100.00$^+$	98.81 / 100.00
6	88.69 / 100.00	90.48 / 100.00	91.07 / 100.00	91.07 / 100.00	92.26 / 100.00	94.05 / 100.00
BP; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00

Table S7: Complete paired-validation results for HyenaDNA. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
4	97.59 / 98.80	99.20 / 100.00$^+$	99.60 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
Acceptor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
4	98.65 / 98.72	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
PPT; thresholds: 92.17 / 95.00%						
1	**92.86 / 100.00**	94.64 / 100.00$^+$	95.24 / 100.00$^+$	95.83 / 100.00$^+$	94.64 / 100.00$^+$	94.05 / 100.00
3	86.31 / 100.00	90.48 / 100.00	94.05 / 100.00$^+$	94.64 / 100.00$^+$	95.83 / 100.00$^+$	97.02 / 100.00
4	80.36 / 99.18	88.69 / 100.00	92.26 / 100.00$^+$	92.26 / 100.00$^+$	94.05 / 100.00$^+$	95.83 / 100.00
BP; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
4	99.48 / 99.52$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00

Table S8: Complete paired-validation results for DNABERT-2. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	98.39 / 99.20	100.00 / 100.00$^+$	99.60 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	99.20 / 99.20$^+$	99.60 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	96.39 / 92.40	95.58 / 92.00	97.99 / 99.20	99.20 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	90.36 / 91.20	93.57 / 96.00	95.58 / 97.20	97.19 / 98.40	98.80 / 99.20	100.00 / 99.60
Acceptor; thresholds: 99.00 / 99.00%						
1	100.00 / 97.45	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	99.10 / 98.72	99.55 / 99.57$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	96.40 / 96.60	98.65 / 98.30	100.00 / 99.57$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	92.34 / 92.34	96.40 / 95.74	98.20 / 97.45	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	72.97 / 77.87	85.59 / 85.11	90.54 / 91.06	92.79 / 91.91	96.85 / 94.04	99.55 / 99.15
PPT; thresholds: 91.61 / 95.00%						
1	84.52 / 100.00	87.50 / 100.00	89.29 / 100.00	88.69 / 100.00	88.10 / 100.00	88.10 / 100.00
3	**92.86 / 100.00**	94.05 / 100.00$^+$	97.02 / 100.00$^+$	97.62 / 100.00$^+$	97.62 / 100.00$^+$	96.43 / 100.00
6	85.12 / 100.00	87.50 / 100.00	89.29 / 100.00	91.07 / 100.00	92.26 / 100.00$^+$	93.45 / 100.00
9	73.21 / 100.00	83.93 / 100.00	84.52 / 100.00	86.90 / 100.00	89.29 / 100.00	91.67 / 100.00
12	64.29 / 79.51	66.67 / 89.75	71.43 / 94.26	69.64 / 96.31	70.83 / 99.18	78.57 / 100.00
BP; thresholds: 99.00 / 98.52%						
1	94.85 / 95.67	**100.00 / 99.04**	100.00 / 100.00$^+$	100.00 / 99.04$^+$	100.00 / 99.52$^+$	100.00 / 99.52
3	75.26 / 86.06	90.72 / 94.71	94.33 / 97.60	98.97 / 98.56	99.48 / 98.08	96.39 / 98.56
6	81.44 / 82.69	91.24 / 95.67	99.48 / 98.08	99.48 / 99.52$^+$	100.00 / 99.04$^+$	97.94 / 98.56
9	72.16 / 74.04	89.69 / 90.38	97.94 / 98.08	100.00 / 100.00$^+$	98.97 / 99.52	97.94 / 99.52
12	74.23 / 75.00	83.51 / 83.17	86.60 / 87.98	89.18 / 86.06	89.69 / 88.94	91.75 / 90.87

Table S9: Complete paired-validation results for NTv2. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	100.00 / 98.80	100.00 / 99.60^{+}	100.00 / 99.60^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
3	**100.00 / 100.00**	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
5	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
6	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
9	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
12	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
15	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
18	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
21	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
24	99.20 / 99.20^{+}	100.00 / 98.80	100.00 / 99.60^{+}	100.00 / 99.20^{+}	100.00 / 99.60^{+}	100.00 / 100.00
27	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
29	100.00 / 99.20^{+}	99.60 / 99.20^{+}	100.00 / 99.60^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
Acceptor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
3	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
5	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
6	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
9	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
12	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
15	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
18	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
21	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
24	97.75 / 100.00	99.55 / 99.57^{+}	99.55 / 100.00^{+}	99.55 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
27	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
29	98.20 / 98.30	98.65 / 99.15	99.10 / 98.72	99.10 / 99.15^{+}	99.10 / 99.15^{+}	100.00 / 100.00
PPT; thresholds: 93.87 / 95.00%						
1	79.76 / 92.62	83.93 / 97.13	84.52 / 99.18	86.31 / 100.00	86.31 / 100.00	87.50 / 100.00
3	86.90 / 100.00	85.71 / 100.00	90.48 / 100.00	92.26 / 100.00	91.67 / 100.00	92.26 / 100.00
5	83.33 / 100.00	91.67 / 100.00	92.86 / 100.00	94.64 / 100.00^{+}	97.02 / 100.00^{+}	97.02 / 100.00
6	83.33 / 100.00	91.07 / 100.00	92.26 / 100.00	93.45 / 100.00	94.64 / 100.00^{+}	96.43 / 100.00
9	92.26 / 100.00	**95.83 / 100.00**	97.62 / 100.00^{+}	97.02 / 100.00^{+}	96.43 / 100.00^{+}	98.21 / 100.00
12	86.90 / 100.00	92.86 / 100.00	93.45 / 100.00	97.02 / 100.00^{+}	96.43 / 100.00^{+}	98.81 / 100.00
15	85.12 / 100.00	86.90 / 100.00	91.07 / 100.00	95.24 / 100.00^{+}	92.86 / 100.00	97.02 / 100.00
18	72.62 / 100.00	78.57 / 100.00	85.12 / 100.00	89.88 / 100.00	94.05 / 100.00^{+}	95.24 / 100.00
21	73.81 / 100.00	84.52 / 100.00	92.26 / 100.00	92.86 / 100.00	94.64 / 100.00^{+}	96.43 / 100.00
24	85.12 / 99.59	89.88 / 100.00	89.88 / 100.00	90.48 / 100.00	90.48 / 100.00	93.45 / 100.00
27	80.36 / 100.00	82.14 / 100.00	91.67 / 100.00	95.83 / 100.00^{+}	95.83 / 100.00^{+}	95.24 / 100.00
29	78.57 / 97.95	79.76 / 98.77	82.74 / 99.59	88.10 / 99.59	87.50 / 99.59	90.48 / 100.00
BP; thresholds: 99.00 / 99.00%						
1	97.42 / 95.67	96.91 / 97.60	98.97 / 99.52	98.97 / 99.52	98.97 / 100.00	99.48 / 100.00
3	99.48 / 99.52^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
5	**100.00 / 100.00**	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
6	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
9	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
12	88.66 / 93.75	100.00 / 100.00^{+}	99.48 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
15	98.97 / 99.04	99.48 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
18	79.38 / 79.33	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
21	79.38 / 82.21	96.91 / 96.15	97.94 / 98.56	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
24	82.99 / 88.94	84.54 / 88.94	88.66 / 91.35	94.85 / 96.15	100.00 / 100.00^{+}	100.00 / 100.00
27	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
29	98.97 / 97.60	99.48 / 97.60	98.97 / 98.56	99.48 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00

Table S10: Complete paired-validation results for Genos. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
2	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
3	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
6	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
9	79.92 / 76.40	87.55 / 85.60	99.60 / 99.20^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
12	82.73 / 91.20	87.95 / 92.80	91.57 / 97.20	93.17 / 98.00	94.38 / 98.40	95.58 / 99.60

Continued on next page

Table S10 (continued)

Layer	4	8	16	32	64	Full
Acceptor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
2	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	81.53 / 80.85	90.09 / 91.06	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	90.99 / 94.47	90.54 / 95.32	92.34 / 97.02	94.59 / 97.45	97.75 / 97.45	97.30 / 97.45
PPT; thresholds: 92.17 / 95.00%						
1	89.29 / 100.00	91.67 / 100.00	**93.45 / 100.00**	94.64 / 100.00$^+$	95.24 / 100.00$^+$	95.83 / 100.00
2	89.29 / 100.00	89.88 / 100.00	91.07 / 100.00	92.86 / 100.00$^+$	96.43 / 100.00$^+$	97.02 / 100.00
3	86.90 / 100.00	89.88 / 100.00	89.88 / 100.00	94.05 / 100.00$^+$	94.05 / 100.00$^+$	97.02 / 100.00
6	83.93 / 100.00	86.31 / 100.00	89.88 / 100.00	92.26 / 100.00$^+$	93.45 / 100.00$^+$	96.43 / 100.00
9	65.48 / 88.52	70.83 / 96.31	79.17 / 100.00	81.55 / 100.00	83.93 / 100.00	94.64 / 100.00
12	78.57 / 95.90	77.38 / 96.72	77.38 / 98.36	76.79 / 98.77	77.98 / 98.77	78.57 / 99.18
BP; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
2	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	77.32 / 75.96	89.69 / 88.94	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	89.69 / 89.42	92.78 / 90.87	95.36 / 93.27	96.39 / 94.71	96.91 / 95.67	97.42 / 97.12

Table S11: Complete paired-validation results for Evo 2. Values are shuffling / substitution (depletion for PPT), in percent.

Layer	4	8	16	32	64	Full
Donor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
5	99.60 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	98.39 / 98.80	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	96.39 / 98.00	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	88.35 / 88.40	99.60 / 99.60$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
15	92.37 / 94.00	100.00 / 99.60$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
16	90.76 / 92.80	100.00 / 99.60$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
18	86.75 / 90.00	100.00 / 99.20$^+$	99.60 / 99.60$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
21	91.97 / 92.00	99.60 / 99.60$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
24	92.37 / 88.00	95.98 / 94.40	98.80 / 99.60	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
27	91.97 / 90.80	98.80 / 98.00	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
30	93.57 / 94.80	96.39 / 96.80	96.39 / 96.80	96.39 / 96.80	96.39 / 96.80	96.39 / 96.80
32	93.57 / 95.60	94.78 / 98.40	94.78 / 98.40	94.78 / 98.40	94.78 / 98.40	94.78 / 98.40
Acceptor; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
3	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
5	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
6	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
9	95.50 / 97.87	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
12	92.34 / 89.36	99.10 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
15	95.50 / 92.77	99.10 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
16	93.69 / 89.36	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
18	95.95 / 96.60	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
21	96.85 / 95.74	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
24	88.74 / 91.91	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
27	95.50 / 98.72	99.55 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00$^+$	100.00 / 100.00
30	88.74 / 94.47	89.19 / 95.32	89.19 / 95.32	89.19 / 95.74	89.19 / 95.74	89.19 / 95.74
32	87.84 / 94.47	88.29 / 96.60	88.29 / 96.60	88.29 / 96.60	88.29 / 96.60	88.29 / 96.60
PPT; thresholds: 93.30 / 95.00%						
1	**95.24 / 100.00**	97.02 / 100.00$^+$	97.02 / 100.00$^+$	98.21 / 100.00$^+$	99.40 / 100.00$^+$	98.21 / 100.00
3	85.71 / 100.00	87.50 / 100.00	87.50 / 100.00	89.88 / 100.00	94.64 / 100.00$^+$	95.83 / 100.00
5	79.76 / 100.00	82.14 / 100.00	85.71 / 100.00	86.31 / 100.00	94.05 / 100.00$^+$	97.02 / 100.00
6	80.36 / 100.00	85.71 / 100.00	84.52 / 100.00	86.31 / 100.00	91.67 / 100.00	94.05 / 100.00
9	82.14 / 100.00	83.33 / 100.00	83.93 / 100.00	84.52 / 100.00	86.31 / 100.00	91.67 / 100.00
12	71.43 / 93.03	76.79 / 100.00	82.14 / 100.00	82.74 / 100.00	85.71 / 100.00	88.69 / 100.00
15	72.62 / 94.67	74.40 / 100.00	80.36 / 100.00	81.55 / 100.00	85.12 / 100.00	88.10 / 100.00
16	71.43 / 93.85	75.00 / 100.00	79.17 / 100.00	82.74 / 100.00	83.93 / 100.00	86.90 / 100.00
18	73.21 / 97.54	77.38 / 100.00	83.33 / 100.00	82.74 / 100.00	86.31 / 100.00	88.69 / 100.00

Continued on next page

Table S11 (continued)

Layer	4	8	16	32	64	Full
21	72.02 / 100.00	77.98 / 100.00	79.76 / 100.00	82.74 / 100.00	85.71 / 100.00	88.10 / 100.00
24	69.64 / 92.21	82.74 / 100.00	88.10 / 100.00	87.50 / 100.00	86.90 / 100.00	89.88 / 100.00
27	81.55 / 100.00	85.71 / 100.00	87.50 / 100.00	90.48 / 100.00	91.67 / 100.00	91.67 / 100.00
30	77.38 / 99.18	77.38 / 99.59	77.38 / 99.59	77.38 / 99.59	77.38 / 99.59	77.38 / 99.59
32	73.21 / 98.36	77.38 / 98.77	77.38 / 98.77	77.38 / 98.77	77.38 / 98.77	77.38 / 98.77
BP; thresholds: 99.00 / 99.00%						
1	**100.00 / 100.00**	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
3	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
5	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
6	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
9	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
12	98.97 / 100.00	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
15	100.00 / 99.52^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
16	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
18	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
21	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
24	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
27	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00^{+}	100.00 / 100.00
30	96.91 / 97.60	97.42 / 98.08	97.42 / 98.08	97.42 / 98.08	97.42 / 98.08	97.42 / 98.08
32	97.42 / 97.12	97.42 / 96.63	97.42 / 96.63	97.42 / 96.63	97.42 / 96.63	97.42 / 96.63

C Natural Component Recognition and Functional Dimension Selection

We evaluated whether the component-associated directions identified by paired perturbations contributed to recognition of natural sequences. For each of the six models and four components, we fixed the layer selected in Section B and compared dimensions $r \in \{4, 8, 16, 32, 64\}$. This yielded 24 model–component curves and 120 rank-specific comparisons. All comparisons used a classifier fitted to unmodified representations and held fixed during intervention.

C.1 Dataset Sizes and Independence from Subspace Discovery

Table S12: Natural recognition dataset sizes after overlap removal. Entries give total sequences (positive/negative). The same sequences were used for all six models.

Task	Training	Validation	Test
Donor	1,966 (966/1,000)	486 (236/250)	2,957 (1,441/1,516)
Acceptor	1,974 (975/999)	484 (235/249)	2,934 (1,411/1,523)
PPT	1,966 (983/983)	484 (242/242)	1,306 (653/653)
BP	1,998 (999/999)	498 (249/249)	1,122 (561/561)

The recognition datasets were separate from the reference–edited pairs used to discover subspaces. All recognition inputs were 600-bp natural genomic sequences; no artificial sequence edits were used to define the recognition test inputs. Before fitting these classifiers, we removed positive and negative examples whose genomic windows overlapped any 1,001-bp window in the current discovery, matched-control validation or core-preserving context-validation datasets. The exclusion covered all four components and all training, validation and test partitions of those datasets. Windows were compared on the same chromosome irrespective of strand. For PPT and BP, we removed the entire matched pair if either member overlapped an excluded window. This removed 249 sequences across the four recognition datasets; the remaining sequences had zero overlap with the specified exclusion windows. We preserved the original recognition splits without drawing replacement examples.

Thus, the recognition data were independent of the current subspace-discovery and semantic-validation inputs at the genomic-window level. This does not imply exclusion from foundation-model pretraining, gene-level separation across every task, or independence from all earlier project analyses. Moreover, these recognition test curves were subsequently used to choose functional dimensions (Section C.3); they are not a second independent confirmation of that choice.

Donor and acceptor. We retained the binary annotation labels and natural negative sequences from the revised Nucleotide Transformer downstream donor and acceptor datasets.[1] The original sampling used chromosome 8 for validation and the official chromosome 20/21 test partition, with training examples taken from the remaining training chromosomes. Exact-sequence and reverse-complement conflicts across partitions were excluded in the original construction; the additional window exclusion above was applied to all partitions. These are annotation-based recognition labels rather than experimental confirmation that a component is active, or absent, in every cellular context.

Polypyrimidine tract. We used natural sequences centred on GENCODE v49 acceptor sites, with AG retained at the same central position in both classes. The target was the 18-bp tract at offsets -20 through -3 relative to the acceptor anchor. A positive tract contained at least 13 C/T bases and a longest consecutive C/T run of at least five; a weak-tract control contained at most nine C/T bases and a longest run of at most four. Pairs were matched by chromosome, strand and log-two intron-length bin, with GC content outside the target tract differing by at most three percentage points. These labels distinguish rule-defined strong and weak natural tracts, not experimentally established presence and complete absence of PPT function.

[1] Source: Nucleotide Transformer revised downstream datasets.

Branch point. Positive sites were high-confidence branch points from the experimental catalogue of Mercer et al. (2015), uniquely mapped to GRCh38 and matched to introns. Both positive and control sequences were centred on A. Controls were unannotated natural A sites in the same intron, offset by 4–8 bases from the positive site, located 15–60 bases upstream of the acceptor and more than three bases from loaded high-confidence branch-point annotations. The task concerns the annotation of the central site, not whether any branch point occurs anywhere in the 600-bp sequence. Because annotations are incomplete, these controls are not experimentally verified non-branch-points. PPT and BP train/validation/test partitions were separated by gene, and matched pairs were retained or removed together.

C.2 Frozen Recognition Heads and Propagated Erasure

For each model and task, we extracted the final-layer hidden states aligned to all 600 DNA positions and flattened the resulting $600 \times d$ array. We fitted a linear ridge classifier to the cleaned training set, encoding binary labels as -1 and $+1$. Training features and labels were centred; regularization was selected from $\{0.01, 0.1, 1, 10, 100\}$ using validation accuracy, with the larger value preferred in a tie. The ridge penalty was scaled by the mean diagonal of the training Gram matrix. A test score greater than zero was classified as positive. The head, feature centre, intercept and backbone parameters were then fixed for every rank and intervention condition. In particular, classifiers were not refitted after erasure.

Let $H^{(\ell)}(x)$ contain the hidden-state row vectors at the selected layer, $Q_{c,r} \in \mathbb{R}^{d\times r}$ the first r columns of the discovered orthonormal basis, and μ_ℓ the fixed reference mean. Component erasure replaced the DNA-position states by

$$\widetilde{H}^{(\ell)}(x) = H^{(\ell)}(x) - \big(H^{(\ell)}(x) - \mathbf{1}\mu_\ell^\top\big)Q_{c,r}Q_{c,r}^\top. \tag{S6}$$

The reference mean was held constant across the rank sweep. The intervention was applied at the component's selected layer to the DNA positions, not only to a pooled local window. We then recomputed the remaining model layers and passed the resulting final-layer representation to the same recognition head. This experiment measures the functional effect of a single component-associated subspace; it is not a cross-component or joint-erasure experiment.

Matched random interventions. For each condition, three fixed random orthonormal bases were evaluated at the same layer and rank. In addition to matching rank, we rescaled each sample's random projection displacement to match the requested component-erasure displacement energy. Writing these pre-rescaling displacements as $D_c(x)$ and $D_j(x)$, the random intervention used

$$\widetilde{H}_j^{(\ell)}(x) = H^{(\ell)}(x) - \alpha_j(x)D_j(x), \qquad \alpha_j(x) = \frac{\|D_c(x)\|_F}{\|D_j(x)\|_F}, \tag{S7}$$

with numerical protection for vanishing norms in the implementation. Matching refers to the requested hidden-state displacement before model-dtype rounding. These scaled random interventions need not be exact orthogonal erasures when $\alpha_j \neq 1$. The random curve is the mean accuracy of the three intervened classifiers, not the accuracy of an ensemble formed by averaging their scores.

Accuracy and uncertainty. We report sequence-weighted accuracy and paired accuracy differences. Intervals were computed using 2,000 paired cluster-bootstrap draws with seed 20260923, resampling whole groups with replacement and retaining their sequences across all conditions. Donor and acceptor used the source-cluster identifiers (2,915 and 2,898 test groups); PPT and BP used gene identifiers (937 and 281 test groups). The plotted accuracy intervals were computed directly from per-sequence predictions, rather than obtained by subtracting a drop interval from the baseline. Intervals are pointwise 95% percentile intervals, without multiple-comparison adjustment; they condition on the fitted head, chosen rank and three fixed random bases. They do not include retraining, dimension-selection or random-basis-resampling uncertainty.

C.3 Functional Dimension Selection

For each fixed model, component and layer, let A_0 denote unmodified accuracy, $A_c(r)$ accuracy after component erasure, and $\overline{A}_{\mathrm{rnd}}(r)$ mean accuracy after random intervention, all expressed as

percentages. We defined the component decrease, random decrease and excess functional effect in percentage points as

$$C(r) = A_0 - A_c(r), \qquad R(r) = A_0 - \overline{A}_{\mathrm{rnd}}(r), \qquad S(r) = C(r) - R(r). \tag{S8}$$

Negative values of R indicate a small accuracy improvement under the random intervention and were not truncated to zero. We set the effective random-cost budget and eligible rank set to

$$b = \max\{3, \min_r R(r)\}, \qquad \mathcal{E} = \{r : R(r) \le b\}. \tag{S9}$$

Thus, ranks with random decreases no greater than three percentage points were preferred; if none satisfied that budget, only the lowest-cost candidates remained eligible. We then selected

$$r^* = \min\left\{ r \in \mathcal{E} : S(r) \ge 0.95 \max_{u \in \mathcal{E}} S(u) \right\}. \tag{S10}$$

All 24 combinations had a positive maximum eligible S. The three-percentage-point budget and 95% retention fraction are operational choices, not biological constants. Confidence intervals were reported but were not used as an additional exclusion rule.

The final configurations are listed in Table S13. There was one explicit compactness adjustment: HyenaDNA donor used $r = 32$ rather than the automated $r = 64$. Its excess decrease was 18.93 rather than 23.01 percentage points, retaining approximately 82.3% rather than 95% of the eligible maximum. This exception is marked separately and is not attributed to Equation S10. Dimension selection was exploratory after inspection of the recognition curves; it was fixed before the subsequent downstream experiments and was not optimized on their outcomes.

Table S13: Final functional-intervention configurations. Entries are dimension r / layer ℓ, not the paired-validation minimum dimensions in Table S5.

Model	Donor	Acceptor	PPT	BP
SpliceBERT	32 / 1	64 / 1	64 / 1	32 / 1
HyenaDNA	32 / 1[†]	32 / 1	4 / 1	32 / 1
DNABERT-2	64 / 3	32 / 1	64 / 3	32 / 1
NTv2	64 / 3	64 / 1	32 / 9	64 / 5
Genos	4 / 1	4 / 1	64 / 1	32 / 1
Evo 2	32 / 1	16 / 1	32 / 1	32 / 1

[†]HyenaDNA donor uses the separate 32-dimensional compactness adjustment; the automated criterion selected 64 dimensions.

C.4 Results Across Models and Components

The selected spaces often had greater functional effects than matched random interventions, but the size of this difference depended on the model and component. For example, SpliceBERT donor erasure at $r = 32$ reduced accuracy by 29.35 percentage points, compared with 0.70 for the random intervention; the excess was 28.66 [26.72, 30.57]. HyenaDNA PPT showed an excess of 43.80 [40.78, 46.73] at only four dimensions. Evo 2 donor showed an excess of 25.59 [23.43, 27.82] at 32 dimensions. These comparisons quantify dependence of the frozen recognition head after propagation, rather than the total amount of component information in the model.

Two qualifications are visible in the full curves. First, the lowest random costs for Genos donor and acceptor were already 6.66 and 4.75 percentage points at $r = 4$; both therefore used the minimum-cost fallback rather than meeting the nominal three-point budget. Second, the selected NTv2 acceptor and Genos BP excess effects were 0.37 [-0.30, 1.08] and 2.82 [-0.06, 5.77], respectively. These spaces remain in the final configuration, but their excess functional effects were not clearly distinguished from zero in this recognition experiment. Increasing dimension did not uniformly improve the excess effect, because the random intervention could also become more disruptive.

Table S15 reports all unmodified accuracies and AUROCs. Figures S2–S7 show every model–component curve; Tables S16–S21 report all 120 component, random and excess decreases with their intervals. Each panel marks the final configuration, including the HyenaDNA donor exception.

Overview of the dimension sweep. Figure S1 places all 120 excess effects on a shared colour scale, with the final configurations outlined. Several curves increased with dimension, whereas others saturated or declined as additional directions were removed. For example, HyenaDNA donor excess effects increased from 2.57 to 23.01 percentage points across the sweep, while Genos donor decreased from 23.03 to 0.23. The marked dimensions therefore reflect the random-cost and compactness criteria in Section C.3, rather than simply the darkest cell. Table S14 gives the corresponding unmodified, component-erased and random-intervention accuracies for all 24 final configurations.

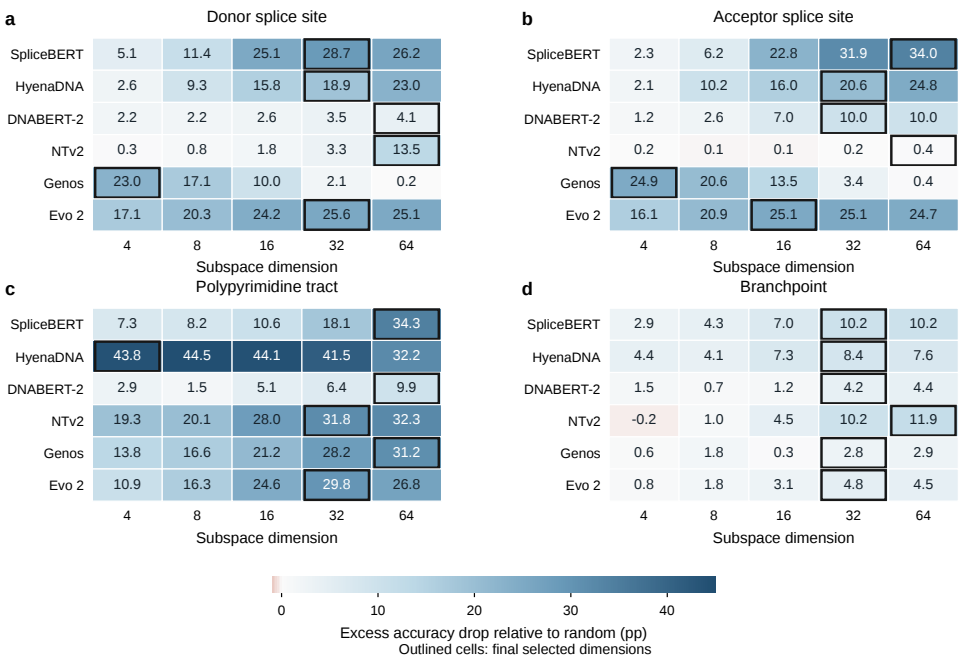


Figure S1: Excess functional effects across candidate dimensions. Panels correspond to donor splice site, acceptor splice site, polypyrimidine tract and branchpoint recognition. Cells show $S(r) = C(r) - R(r)$ in percentage points, using one shared linear colour scale. Rows are models and columns are dimensions. Black outlines mark final configurations in Table S13, including the separate HyenaDNA donor adjustment. Colour indicates the point estimate, not statistical significance; pointwise intervals appear in Tables S16–S21. These recognition curves were used in exploratory dimension selection, not as independent post-selection tests.

Table S14: Recognition accuracy at the final configurations (%). A_0: unmodified; A_c: component erasure; $\overline{A}_{\mathrm{rnd}}$: mean over random interventions. Excess effects and pointwise 95% intervals are in percentage points.

Model	Component	r	A_0	A_c	$\overline{A}_{\mathrm{rnd}}$	S [95% CI]
SpliceBERT	Donor	32	88.30	58.94	87.60	28.66 [26.72, 30.57]
SpliceBERT	Acceptor	64	89.71	53.72	87.70	33.98 [32.07, 35.98]
SpliceBERT	PPT	64	96.32	63.25	97.52	34.28 [31.33, 37.00]
SpliceBERT	BP	32	68.98	57.31	67.47	10.16 [7.24, 13.02]
HyenaDNA	Donor	32	79.91	60.87	79.80	18.93 [16.86, 21.07]
HyenaDNA	Acceptor	32	84.53	62.75	83.31	20.56 [18.77, 22.46]
HyenaDNA	PPT	4	96.25	52.22	96.02	43.80 [40.78, 46.73]
HyenaDNA	BP	32	61.05	51.16	59.60	8.44 [5.28, 11.71]
DNABERT-2	Donor	64	81.23	76.09	80.16	4.07 [2.87, 5.29]
DNABERT-2	Acceptor	32	80.54	70.04	80.02	9.98 [8.41, 11.54]
DNABERT-2	PPT	64	83.54	72.89	82.80	9.90 [7.57, 12.47]
DNABERT-2	BP	32	57.84	53.83	58.08	4.25 [1.06, 7.46]
NTv2	Donor	64	92.26	78.76	92.28	13.52 [11.97, 15.01]
NTv2	Acceptor	64	88.31	87.87	88.24	0.37 [-0.30, 1.08]
NTv2	PPT	32	94.41	62.25	94.00	31.75 [29.04, 34.52]
NTv2	BP	64	64.53	53.30	65.21	11.91 [8.74, 15.02]
Genos	Donor	4	80.99	51.30	74.33	23.03 [21.26, 24.88]
Genos	Acceptor	4	81.60	51.94	76.85	24.90 [22.99, 26.84]
Genos	PPT	64	84.15	50.31	81.55	31.24 [27.73, 34.75]
Genos	BP	32	55.61	51.07	53.89	2.82 [-0.06, 5.77]
Evo 2	Donor	32	76.97	51.44	77.03	25.59 [23.43, 27.82]
Evo 2	Acceptor	16	77.61	51.77	76.83	25.06 [23.02, 27.11]
Evo 2	PPT	32	79.71	48.77	78.56	29.79 [26.17, 33.13]
Evo 2	BP	32	53.12	48.93	53.77	4.84 [1.41, 8.50]

Table S15: Unmodified recognition performance. Accuracy is reported as percent with a pointwise 95% cluster-bootstrap interval; AUROC is computed from the continuous classifier score.

Model	Task	Accuracy [95% CI]	AUROC
SpliceBERT	Donor	88.30 [87.11, 89.48]	0.9522
	Acceptor	89.71 [88.60, 90.84]	0.9576
	PPT	96.32 [95.23, 97.29]	0.9952
	BP	68.98 [66.25, 71.72]	0.7490
HyenaDNA	Donor	79.91 [78.47, 81.35]	0.8936
	Acceptor	84.53 [83.21, 85.95]	0.9238
	PPT	96.25 [95.21, 97.25]	0.9946
	BP	61.05 [58.19, 63.81]	0.6559
DNABERT-2	Donor	81.23 [79.80, 82.69]	0.9038
	Acceptor	80.54 [79.13, 81.96]	0.8954
	PPT	83.54 [81.42, 85.57]	0.9179
	BP	57.84 [54.90, 60.79]	0.6150
NTv2	Donor	92.26 [91.25, 93.20]	0.9748
	Acceptor	88.31 [87.14, 89.54]	0.9524
	PPT	94.41 [93.13, 95.60]	0.9881
	BP	64.53 [61.64, 67.41]	0.7026
Genos	Donor	80.99 [79.59, 82.45]	0.8941
	Acceptor	81.60 [80.16, 82.98]	0.8932
	PPT	84.15 [82.09, 86.15]	0.9267
	BP	55.61 [52.41, 58.79]	0.5660
Evo 2	Donor	76.97 [75.42, 78.50]	0.8514
	Acceptor	77.61 [76.12, 79.10]	0.8561
	PPT	79.71 [77.55, 82.01]	0.8721
	BP	53.12 [50.45, 56.00]	0.5370

C.4.1 SPLICEBERT

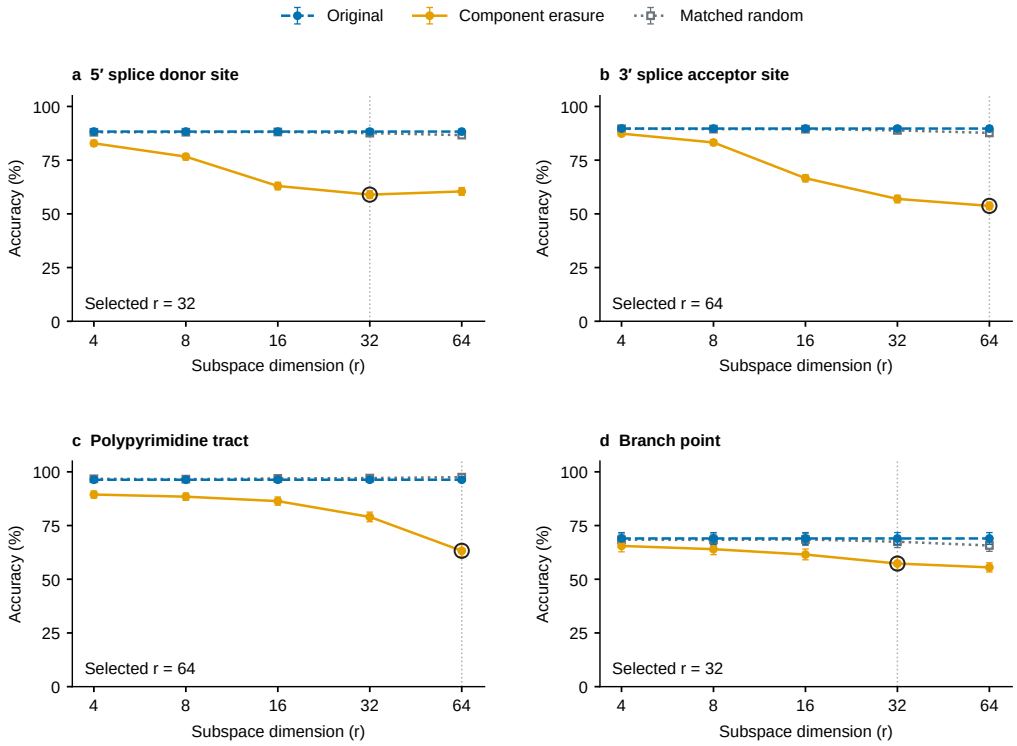


Figure S2: SpliceBERT: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend.

Table S16: SpliceBERT: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S2.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	4	5.44 [4.39, 6.50]	0.32 [0.00, 0.62]	5.13 [4.10, 6.15]
	8	11.67 [10.25, 13.18]	0.29 [-0.06, 0.64]	11.37 [10.00, 12.82]
	16	25.33 [23.39, 27.34]	0.18 [-0.28, 0.62]	25.15 [23.33, 27.07]
	32	29.35 [27.36, 31.44]	0.70 [0.12, 1.29]	28.66 [26.72, 30.57]
	64	27.87 [25.86, 29.93]	1.63 [0.93, 2.39]	26.23 [24.43, 28.07]
Acceptor	4	2.35 [1.57, 3.17]	0.08 [-0.18, 0.33]	2.27 [1.53, 3.04]
	8	6.48 [5.19, 7.81]	0.26 [-0.08, 0.60]	6.21 [4.96, 7.55]
	16	23.11 [21.23, 25.09]	0.30 [-0.10, 0.67]	22.81 [21.09, 24.78]
	32	32.75 [30.75, 34.83]	0.85 [0.32, 1.36]	31.90 [29.95, 34.03]
	64	35.99 [33.87, 38.15]	2.01 [1.29, 2.70]	33.98 [32.07, 35.98]
PPT	4	6.89 [5.40, 8.42]	-0.43 [-0.80, -0.13]	7.33 [5.88, 8.82]
	8	7.89 [6.33, 9.50]	-0.28 [-0.63, 0.03]	8.17 [6.69, 9.70]
	16	9.95 [8.08, 11.94]	-0.61 [-1.06, -0.21]	10.57 [8.71, 12.47]
	32	17.30 [14.91, 19.72]	-0.77 [-1.37, -0.21]	18.07 [15.74, 20.29]
	64	33.08 [30.08, 35.79]	-1.20 [-2.04, -0.46]	34.28 [31.33, 37.00]
BP	4	3.48 [1.46, 5.48]	0.59 [-0.03, 1.25]	2.88 [0.95, 4.88]
	8	4.99 [2.70, 7.27]	0.68 [-0.00, 1.40]	4.31 [2.11, 6.59]
	16	7.49 [4.87, 10.09]	0.48 [-0.41, 1.34]	7.01 [4.43, 9.50]
	32	11.68 [8.60, 14.60]	1.52 [0.46, 2.55]	10.16 [7.24, 13.02]
	64	13.46 [10.37, 16.40]	3.27 [1.72, 4.75]	10.19 [7.10, 13.14]

C.4.2 HYENADNA

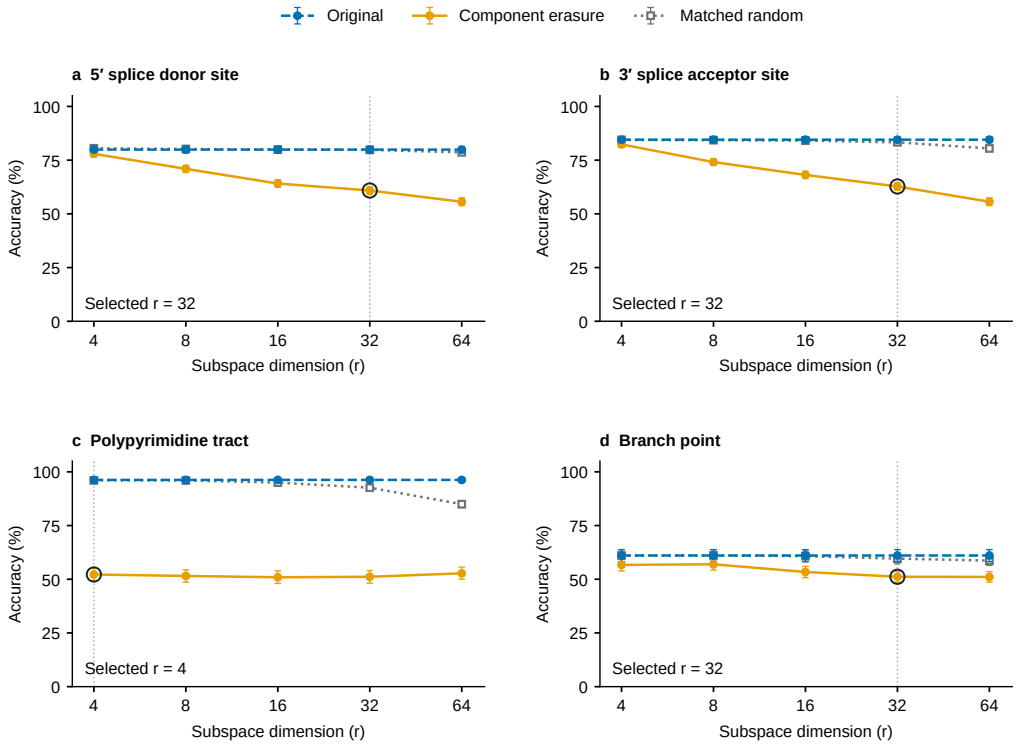


Figure S3: HyenaDNA: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend. The donor marker denotes the separately chosen 32-dimensional configuration rather than the automated 64-dimensional choice.

Table S17: HyenaDNA: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S3.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	4	1.96 [0.64, 3.21]	-0.61 [-1.22, -0.02]	2.57 [1.35, 3.68]
	8	8.96 [7.15, 10.68]	-0.29 [-1.08, 0.45]	9.25 [7.62, 10.90]
	16	15.79 [13.70, 17.82]	-0.02 [-0.89, 0.84]	15.82 [13.78, 17.75]
	32	19.04 [16.80, 21.41]	0.11 [-0.81, 0.99]	18.93 [16.86, 21.07]
	64	24.28 [21.85, 26.69]	1.27 [0.14, 2.34]	23.01 [20.86, 25.25]
Acceptor	4	2.18 [1.02, 3.37]	0.11 [-0.45, 0.66]	2.07 [1.06, 3.10]
	8	10.43 [8.58, 12.24]	0.22 [-0.48, 0.88]	10.21 [8.62, 11.84]
	16	16.39 [14.35, 18.39]	0.44 [-0.35, 1.22]	15.95 [14.11, 17.76]
	32	21.78 [19.75, 23.96]	1.22 [0.25, 2.23]	20.56 [18.77, 22.46]
	64	28.87 [26.59, 31.23]	4.07 [2.84, 5.42]	24.80 [22.97, 26.73]
PPT	**4**	44.03 [41.03, 47.10]	0.23 [-0.34, 0.82]	43.80 [40.78, 46.73]
	8	44.72 [41.67, 47.60]	0.26 [-0.47, 0.96]	44.46 [41.57, 47.24]
	16	45.33 [42.34, 48.37]	1.28 [0.44, 2.10]	44.05 [41.20, 46.99]
	32	45.10 [42.11, 48.14]	3.62 [2.52, 4.73]	41.48 [38.71, 44.19]
	64	43.49 [40.50, 46.37]	11.31 [9.71, 12.88]	32.18 [29.49, 34.49]
BP	4	4.37 [1.12, 7.78]	0.00 [-1.12, 1.17]	4.37 [0.99, 7.83]
	8	4.10 [0.73, 7.59]	-0.03 [-1.28, 1.22]	4.13 [0.83, 7.50]
	16	7.66 [4.24, 11.10]	0.33 [-1.08, 1.70]	7.34 [4.19, 10.67]
	32	9.89 [6.30, 13.61]	1.46 [-0.23, 3.13]	8.44 [5.28, 11.71]
	64	9.98 [6.43, 13.58]	2.38 [0.39, 4.48]	7.61 [4.44, 10.71]

C.4.3 DNABERT-2

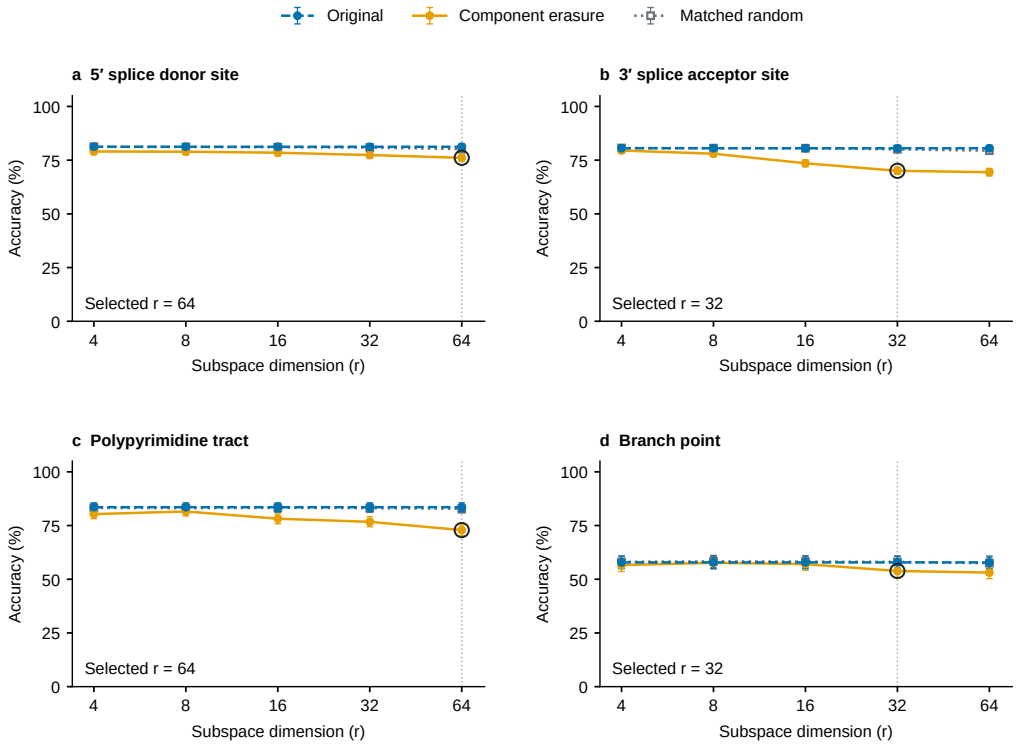


Figure S4: DNABERT-2: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend.

Table S18: DNABERT-2: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S4.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	4	2.20 [1.43, 3.01]	-0.02 [-0.26, 0.21]	2.22 [1.50, 3.01]
	8	2.30 [1.42, 3.19]	0.09 [-0.23, 0.41]	2.21 [1.39, 3.06]
	16	2.77 [1.79, 3.76]	0.20 [-0.19, 0.59]	2.57 [1.71, 3.48]
	32	3.86 [2.74, 4.92]	0.39 [-0.12, 0.90]	3.46 [2.47, 4.50]
	64	5.14 [3.91, 6.47]	1.07 [0.52, 1.65]	4.07 [2.87, 5.29]
Acceptor	4	1.06 [0.27, 1.85]	-0.12 [-0.43, 0.18]	1.18 [0.48, 1.91]
	8	2.52 [1.50, 3.54]	-0.06 [-0.43, 0.33]	2.58 [1.63, 3.50]
	16	6.99 [5.53, 8.43]	0.00 [-0.47, 0.49]	6.99 [5.60, 8.35]
	32	10.50 [8.87, 12.21]	0.52 [-0.05, 1.09]	9.98 [8.41, 11.54]
	64	11.15 [9.29, 13.00]	1.12 [0.47, 1.82]	10.02 [8.25, 11.82]
PPT	4	3.22 [1.98, 4.46]	0.33 [-0.08, 0.75]	2.88 [1.74, 4.04]
	8	1.99 [0.53, 3.48]	0.46 [0.00, 0.96]	1.53 [0.12, 2.93]
	16	5.36 [3.62, 7.20]	0.31 [-0.30, 0.93]	5.05 [3.43, 6.79]
	32	6.81 [4.70, 9.11]	0.43 [-0.29, 1.15]	6.38 [4.39, 8.56]
	64	10.64 [8.20, 13.34]	0.74 [-0.16, 1.59]	9.90 [7.57, 12.47]
BP	4	1.25 [-0.52, 2.77]	-0.27 [-0.96, 0.37]	1.52 [-0.12, 3.02]
	8	0.18 [-1.95, 2.34]	-0.48 [-1.23, 0.24]	0.65 [-1.39, 2.72]
	16	0.80 [-1.92, 3.57]	-0.36 [-1.33, 0.59]	1.16 [-1.51, 3.89]
	32	4.01 [0.52, 7.50]	-0.24 [-1.24, 0.78]	4.25 [1.06, 7.46]
	64	4.72 [0.91, 8.39]	0.36 [-0.92, 1.58]	4.37 [0.74, 7.95]

C.4.4 NTv2

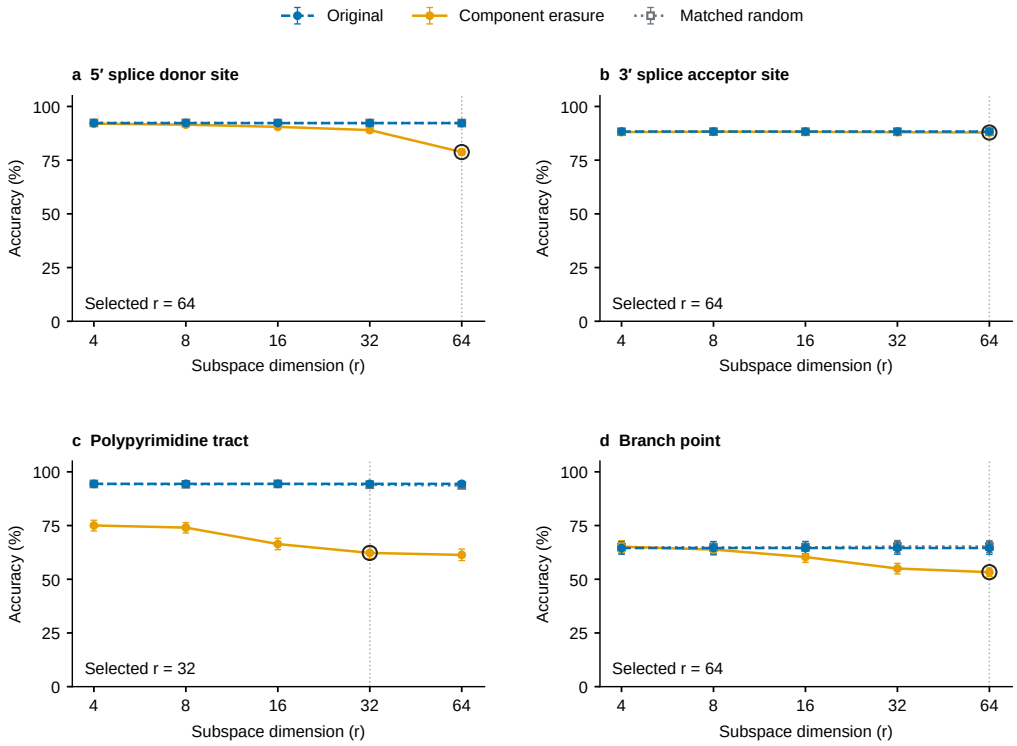


Figure S5: NTv2: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend.

Table S19: NTv2: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S5.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	4	0.24 [-0.17, 0.64]	-0.10 [-0.28, 0.07]	0.34 [-0.00, 0.69]
	8	0.74 [0.20, 1.35]	-0.10 [-0.33, 0.14]	0.85 [0.29, 1.42]
	16	1.76 [0.95, 2.54]	-0.02 [-0.32, 0.26]	1.78 [0.98, 2.57]
	32	3.25 [2.19, 4.25]	-0.03 [-0.39, 0.30]	3.28 [2.30, 4.25]
	64	13.49 [11.95, 15.01]	-0.02 [-0.38, 0.35]	13.52 [11.97, 15.01]
Acceptor	4	0.24 [-0.14, 0.61]	0.07 [-0.10, 0.24]	0.17 [-0.18, 0.53]
	8	0.10 [-0.34, 0.58]	0.03 [-0.18, 0.26]	0.07 [-0.34, 0.52]
	16	0.14 [-0.41, 0.71]	0.08 [-0.16, 0.35]	0.06 [-0.46, 0.59]
	32	0.31 [-0.34, 0.99]	0.07 [-0.20, 0.39]	0.24 [-0.40, 0.88]
	64	0.44 [-0.24, 1.26]	0.07 [-0.27, 0.47]	0.37 [-0.30, 1.08]
PPT	4	19.37 [17.09, 21.87]	0.05 [-0.41, 0.49]	19.32 [17.08, 21.72]
	8	20.37 [18.00, 22.92]	0.23 [-0.29, 0.78]	20.14 [17.80, 22.61]
	16	28.02 [25.34, 30.72]	0.03 [-0.62, 0.72]	28.00 [25.26, 30.74]
	32	32.16 [29.36, 34.92]	0.41 [-0.33, 1.17]	31.75 [29.04, 34.52]
	64	33.08 [30.21, 35.94]	0.74 [-0.08, 1.58]	32.34 [29.56, 35.13]
BP	4	-0.62 [-2.21, 1.01]	-0.42 [-1.00, 0.15]	-0.21 [-1.65, 1.28]
	8	0.71 [-1.40, 2.72]	-0.27 [-0.97, 0.47]	0.98 [-1.10, 2.83]
	16	4.19 [1.55, 6.71]	-0.33 [-1.15, 0.54]	4.52 [1.95, 6.98]
	32	9.54 [6.55, 12.63]	-0.68 [-1.63, 0.26]	10.22 [7.28, 13.23]
	64	11.23 [8.03, 14.44]	-0.68 [-1.69, 0.30]	11.91 [8.74, 15.02]

C.4.5 GENOS

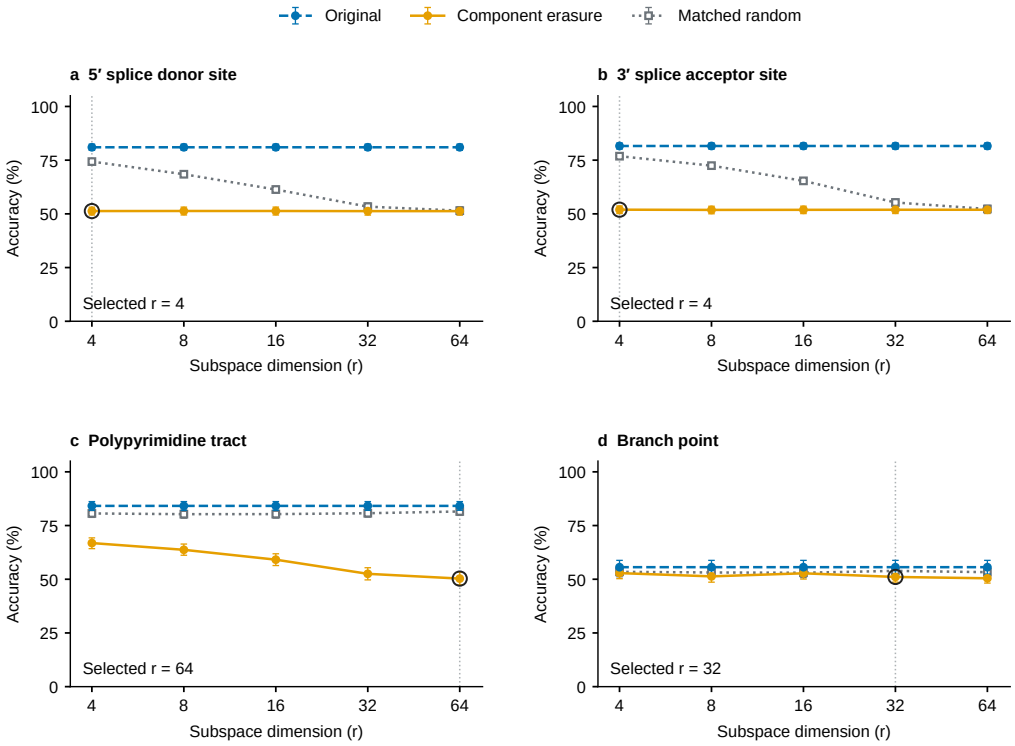


Figure S6: Genos: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend.

Table S20: Genos: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S6.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	**4**	29.69 [27.35, 32.07]	6.66 [5.31, 8.02]	23.03 [21.26, 24.88]
	8	29.66 [27.28, 32.01]	12.55 [10.95, 14.26]	17.11 [15.61, 18.68]
	16	29.66 [27.30, 32.04]	19.67 [17.69, 21.64]	9.99 [8.98, 10.98]
	32	29.73 [27.33, 32.12]	27.63 [25.33, 29.84]	2.10 [1.61, 2.59]
	64	29.73 [27.33, 32.12]	29.50 [27.14, 31.84]	0.23 [-0.07, 0.51]
Acceptor	**4**	29.65 [27.37, 32.04]	4.75 [3.57, 5.90]	24.90 [22.99, 26.84]
	8	29.75 [27.47, 32.15]	9.15 [7.70, 10.61]	20.61 [18.98, 22.31]
	16	29.72 [27.43, 32.13]	16.23 [14.40, 18.11]	13.49 [12.31, 14.78]
	32	29.69 [27.41, 32.09]	26.30 [24.15, 28.56]	3.39 [2.84, 3.97]
	64	29.69 [27.41, 32.09]	29.32 [27.06, 31.67]	0.36 [0.12, 0.60]
PPT	4	17.30 [14.51, 20.11]	3.52 [1.86, 5.23]	13.78 [11.26, 16.33]
	8	20.44 [17.19, 23.56]	3.85 [2.07, 5.68]	16.59 [13.71, 19.38]
	16	25.04 [21.45, 28.50]	3.83 [2.05, 5.60]	21.21 [18.10, 24.39]
	32	31.62 [27.87, 35.27]	3.42 [1.51, 5.40]	28.20 [24.80, 31.65]
	64	33.84 [29.89, 37.52]	2.60 [0.70, 4.54]	31.24 [27.73, 34.75]
BP	4	2.76 [-0.93, 6.32]	2.17 [-0.12, 4.56]	0.59 [-2.43, 3.45]
	8	4.28 [0.69, 7.73]	2.47 [0.09, 4.90]	1.81 [-1.07, 4.58]
	16	2.85 [-0.86, 6.74]	2.58 [0.06, 5.07]	0.27 [-2.65, 3.36]
	32	4.55 [0.70, 8.29]	1.72 [-1.15, 4.55]	2.82 [-0.06, 5.77]
	64	5.17 [1.28, 9.12]	2.32 [-0.61, 5.29]	2.85 [-0.12, 6.06]

C.4.6 Evo 2

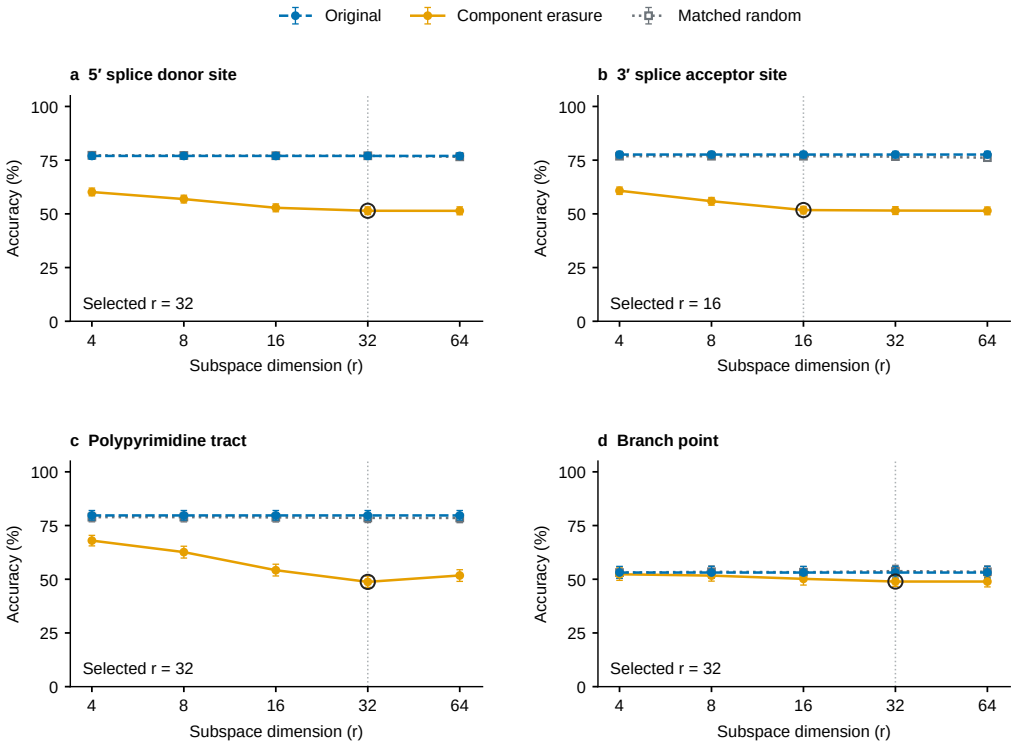


Figure S7: Evo 2: accuracy across subspace dimensions for (a) donor, (b) acceptor, (c) polypyrimidine tract and (d) branch point. Blue: unmodified classifier; orange: component erasure; grey: mean of three matched random interventions. Error bars are pointwise 95% paired cluster-bootstrap intervals. The vertical dotted marker and black ring identify the final dimension. The dimension axis is logarithmic (base two); lines connect evaluated ranks without fitting a trend.

Table S21: Evo 2: all rank-dependent accuracy decreases in percentage points. C is the component decrease, R the mean random decrease and $S = C - R$. Brackets denote pointwise 95% paired intervals. Bold ranks are the final configurations shown in Figure S7.

Task	r	C [95% CI]	R [95% CI]	S [95% CI]
Donor	4	16.81 [14.90, 18.71]	-0.30 [-0.73, 0.15]	17.11 [15.20, 18.98]
	8	20.09 [18.05, 22.20]	-0.20 [-0.65, 0.27]	20.29 [18.31, 22.28]
	16	24.15 [21.99, 26.45]	-0.09 [-0.61, 0.44]	24.24 [22.05, 26.55]
	32	25.53 [23.34, 27.78]	-0.06 [-0.59, 0.54]	25.59 [23.43, 27.82]
	64	25.57 [23.37, 27.79]	0.42 [-0.18, 1.05]	25.15 [23.02, 27.31]
Acceptor	4	16.84 [15.00, 18.68]	0.76 [0.29, 1.24]	16.08 [14.24, 17.90]
	8	21.71 [19.71, 23.64]	0.76 [0.27, 1.26]	20.95 [19.00, 22.87]
	16	25.84 [23.71, 27.86]	0.77 [0.22, 1.33]	25.06 [23.02, 27.11]
	32	26.07 [23.99, 28.17]	0.95 [0.38, 1.57]	25.12 [23.12, 27.19]
	64	26.18 [24.03, 28.27]	1.44 [0.83, 2.07]	24.73 [22.71, 26.87]
PPT	4	11.72 [8.89, 14.56]	0.79 [-0.10, 1.69]	10.92 [8.18, 13.79]
	8	17.08 [13.83, 20.18]	0.79 [-0.11, 1.71]	16.28 [13.15, 19.36]
	16	25.50 [22.23, 28.90]	0.89 [-0.13, 1.94]	24.60 [21.28, 27.96]
	32	30.93 [27.29, 34.42]	1.15 [-0.03, 2.29]	29.79 [26.17, 33.13]
	64	27.95 [24.53, 31.51]	1.17 [0.03, 2.37]	26.77 [23.47, 30.17]
BP	4	0.80 [-2.46, 4.06]	0.03 [-1.19, 1.22]	0.77 [-2.52, 3.95]
	8	1.43 [-1.95, 4.88]	-0.36 [-1.54, 0.91]	1.78 [-1.50, 5.20]
	16	2.94 [-0.60, 6.73]	-0.15 [-1.48, 1.22]	3.09 [-0.43, 6.74]
	32	4.19 [0.70, 7.95]	-0.65 [-2.04, 0.71]	4.84 [1.41, 8.50]
	64	4.19 [0.27, 8.09]	-0.36 [-1.86, 1.16]	4.55 [0.60, 8.51]

D Robustness to Discovery-set Size

We examined whether component subspaces and their functional effects were sensitive to the number of discovery backgrounds. The analysis compared spaces re-estimated from 25%, 50% and 75% of the discovery training backgrounds with their full-data counterparts. This section provides the numerical evidence underlying the main-text subspace-overlap figure and complements it with functional comparisons.

D.1 Nested Background Subsampling

Subsampling operated on native backgrounds, rather than on individual perturbation pairs or recognition test examples. All admissible edits from a retained background were included, with their original weights summing to one. We retained $\lfloor pN_c \rfloor$ backgrounds for component c and fraction p, where N_c is its full discovery-training background count. The 25% set was contained in the 50% set, which was contained in the 75% set. Each fraction used the same selected backgrounds across all six models. The original validation and test partitions were not subsampled or used to discover the new bases. Table S22 reports the resulting background and pair counts.

Table S22: Discovery training backgrounds / paired sequences at each data fraction. Background counts are component-specific, not counts of globally distinct genes.

Component	25%	50%	75%	100%
Donor	298 / 593	596 / 1,189	894 / 1,785	1,192 / 2,377
Acceptor	285 / 554	570 / 1,114	855 / 1,674	1,140 / 2,230
PPT	294 / 488	588 / 976	882 / 1,467	1,177 / 1,953
BP	264 / 497	528 / 985	792 / 1,485	1,057 / 1,993

For each fraction, we re-estimated the full-sequence difference-based subspace using the procedure in Section B.1. We did not obtain smaller-data spaces by truncating or resampling the full-data basis. Layers remained fixed at the full-data choices; nested dimensions of 4, 8, 16, 32 and 64 were evaluated at each fraction. This yielded 360 fraction-specific functional comparisons across six models, four components and five dimensions. The primary summaries used the 24 final configurations in Table S13, including HyenaDNA donor at 32 dimensions. No layer or dimension was reselected using the subsampling results.

The natural recognition inputs, fitted classifiers, centring vectors and classification thresholds were unchanged from Section C. Random bases were also held fixed; their displacement magnitudes were rematched to the component intervention for each fraction and sample. The experiment therefore varied discovery-set size while keeping the downstream measurement interface fixed. It comprised one fixed nested subsample per fraction, rather than repeated independent draws of discovery backgrounds.

D.2 Geometric Stability of the Discovered Spaces

Let $Q_{c,p}$ and $Q_{c,1}$ be orthonormal bases estimated from fraction p and the full discovery set at the same layer and dimension r. We quantified their overlap as

$$O_c(p) = \frac{100}{r} \left\| Q_{c,p}^\top Q_{c,1} \right\|_F^2 . \tag{S11}$$

This quantity is the mean squared cosine of the principal angles, expressed as a percentage. It equals 100% for identical subspaces and zero for orthogonal subspaces, irrespective of basis signs or within-subspace orthogonal rotations. It measures geometric agreement, not the percentage of biological information retained. The full-data self-comparison is 100% by definition.

Mean overlap across the final configurations increased from 91.12% at 25% discovery data to 96.38% at 50% and 98.40% at 75%. Table S23 reports each model–component value, rather than only the aggregate trend. NTv2 acceptor showed appreciable data sensitivity: overlap was 64.65%, 80.86% and 90.76% at the three fractions. Thus, compact component-associated spaces were generally reproducible with fewer backgrounds, but convergence to the full-data space varied across configurations.

Table S23: Geometric overlap (%) with the full-data subspace at the final fixed layer and dimension. The 100% self-comparison equals 100% throughout.

Model	Component	ℓ	r	25%	50%	75%
SpliceBERT	Donor	1	32	95.13	97.88	99.63
SpliceBERT	Acceptor	1	64	96.05	98.33	98.20
SpliceBERT	PPT	1	64	96.19	98.86	99.43
SpliceBERT	BP	1	32	94.58	98.53	99.34
HyenaDNA	Donor	1	32	97.30	99.13	99.68
HyenaDNA	Acceptor	1	32	96.62	98.98	99.53
HyenaDNA	PPT	1	4	77.52	99.26	99.18
HyenaDNA	BP	1	32	94.19	96.03	99.47
DNABERT-2	Donor	3	64	84.08	93.36	97.23
DNABERT-2	Acceptor	1	32	88.04	95.62	98.04
DNABERT-2	PPT	3	64	84.93	93.42	97.45
DNABERT-2	BP	1	32	88.39	95.04	97.49
NTv2	Donor	3	64	84.12	92.35	96.47
NTv2	Acceptor	1	64	64.65	80.86	90.76
NTv2	PPT	9	32	91.25	97.34	99.10
NTv2	BP	5	64	84.07	92.23	96.86
Genos	Donor	1	4	98.22	99.23	99.82
Genos	Acceptor	1	4	98.64	99.26	99.85
Genos	PPT	1	64	94.60	97.14	98.06
Genos	BP	1	32	95.39	98.66	99.60
Evo 2	Donor	1	32	96.38	98.68	99.58
Evo 2	Acceptor	1	16	95.70	99.35	99.82
Evo 2	PPT	1	32	95.27	96.41	98.27
Evo 2	BP	1	32	95.57	97.11	98.79

D.3 Stability of Functional Effects

Geometric agreement alone does not establish that two spaces have the same effect on predictions. We therefore repeated the propagated component erasures and matched random interventions using each re-estimated space, without refitting the recognition head. Writing C_p and S_p for the component decrease and excess effect at fraction p, we compared $C_p - C_1$ and $S_p - S_1$ on identical recognition examples. Intervals used 2,000 paired cluster-bootstrap draws with the grouping and percentile procedure in Section C.2. These pointwise 95% intervals describe variation across recognition backgrounds, conditional on the fixed discovery subsamples and fitted classifiers. They do not estimate uncertainty across alternative discovery subsamples or repeat the layer–dimension selection procedure.

For the 24 final configurations, mean absolute changes in component-erasure effects were 1.14, 0.74 and 0.45 percentage points at 25%, 50% and 75%, respectively. The corresponding changes were at most two percentage points for 22, 23 and 23 configurations (Table S24). Excess effects retained their full-data sign in all 24 configurations at each fraction; this sign agreement does not imply that every individual effect was distinguishable from zero. Across all 120 layer-fixed model–component–dimension combinations, mean absolute component-effect changes were 0.62, 0.50 and 0.35 percentage points, respectively.

Table S24: Unweighted summary across the 24 final model–component configurations. Absolute effect changes are in percentage points; sign agreement concerns the excess effect relative to random intervention.

Data	Overlap (%)	$\lvert\Delta C\rvert$	$\lvert\Delta S\rvert$	$\lvert\Delta C\rvert \le 2$	Sign agreement
25%	91.12	1.14	1.15	22/24	24/24
50%	96.38	0.74	0.73	23/24	24/24
75%	98.40	0.45	0.48	23/24	24/24

Figure S8 shows the excess effects for each final configuration, while Table S25 reports paired changes in component-erasure effects with intervals. Most curves remained close to their full-data

values, but not all configurations were insensitive to discovery-set size. NTv2 donor at 64 dimensions produced component decreases of 5.61 and 10.01 percentage points with 25% and 50% data, compared with 13.49 using all backgrounds. Its 25%-minus-full difference was -7.88 [-9.25, -6.45] percentage points. HyenaDNA polypyrimidine tract at four dimensions decreased from a full-data effect of 44.03 to 38.51 percentage points with 25% data, but reached 43.87 with 50%. These results support overall stability while identifying configurations that benefit from more discovery backgrounds.

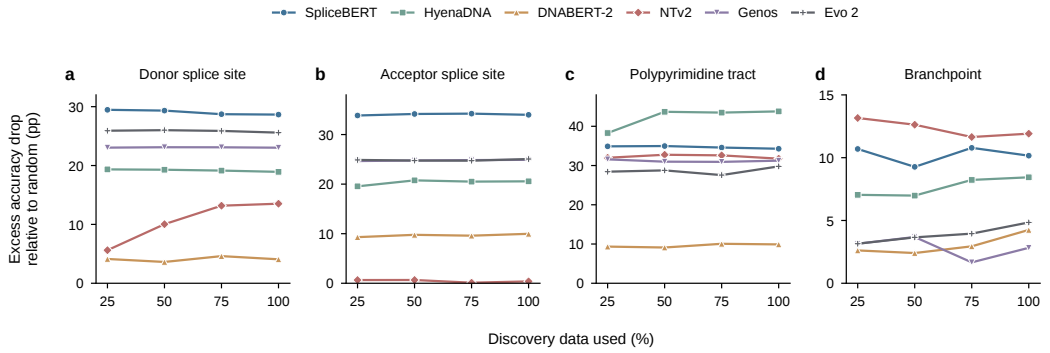


Figure S8: Functional effects across discovery-set sizes at fixed final configurations. Each curve shows the component-erasure accuracy reduction minus the mean reduction under three matched random interventions. Layers, dimensions, classifiers and recognition examples remain fixed; only the discovery backgrounds change. Model colours and markers are consistent across panels. Vertical axes use component-specific ranges, explicitly labelled, to resolve within-task changes. Lines connect observed values without smoothing. The final HyenaDNA donor dimension is 32, as in Table S13. The figure reports one nested discovery subsample per fraction, not a distribution over sampling seeds; paired effect-change intervals are provided in Table S25.

Table S25: Change in component-erasure accuracy reduction relative to full discovery data, in percentage points. Positive values denote a larger reduction with the subset. Brackets are pointwise 95% paired cluster-bootstrap intervals.

Model	Component	25% minus full	50% minus full	75% minus full
SpliceBERT	Donor	+0.81 [+0.37, +1.28]	+0.68 [+0.30, +1.12]	+0.07 [-0.20, +0.34]
SpliceBERT	Acceptor	-0.14 [-0.44, +0.20]	+0.17 [-0.03, +0.38]	+0.24 [+0.07, +0.44]
SpliceBERT	PPT	+0.61 [-0.15, +1.44]	+0.69 [+0.08, +1.32]	+0.31 [-0.23, +0.84]
SpliceBERT	BP	+0.53 [-0.97, +1.98]	-0.89 [-1.92, +0.09]	+0.62 [-0.26, +1.45]
HyenaDNA	Donor	+0.41 [-0.41, +1.15]	+0.34 [-0.30, +0.95]	+0.20 [-0.30, +0.68]
HyenaDNA	Acceptor	-0.99 [-1.74, -0.27]	+0.20 [-0.31, +0.75]	-0.03 [-0.51, +0.44]
HyenaDNA	PPT	-5.51 [-6.89, -4.19]	-0.15 [-0.39, +0.00]	-0.31 [-0.63, -0.08]
HyenaDNA	BP	-1.43 [-3.51, +0.61]	-1.43 [-3.19, +0.28]	-0.09 [-1.27, +1.05]
DNABERT-2	Donor	-0.03 [-1.09, +1.02]	-0.51 [-1.49, +0.51]	+0.51 [-0.31, +1.32]
DNABERT-2	Acceptor	-0.75 [-1.57, +0.07]	-0.20 [-0.85, +0.41]	-0.37 [-0.89, +0.10]
DNABERT-2	PPT	-0.54 [-2.56, +1.51]	-0.77 [-2.75, +1.07]	+0.15 [-1.61, +1.80]
DNABERT-2	BP	-1.78 [-4.52, +0.86]	-1.87 [-4.26, +0.62]	-1.34 [-3.33, +0.75]
NTv2	Donor	-7.88 [-9.25, -6.45]	-3.48 [-4.58, -2.37]	-0.34 [-1.25, +0.57]
NTv2	Acceptor	+0.34 [-0.27, +0.99]	+0.31 [-0.20, +0.85]	-0.24 [-0.61, +0.17]
NTv2	PPT	+0.23 [-1.20, +1.76]	+1.00 [-0.08, +2.14]	+0.84 [-0.07, +1.71]
NTv2	BP	+1.25 [-0.76, +3.30]	+0.71 [-1.04, +2.53]	-0.27 [-1.86, +1.37]
Genos	Donor	-0.03 [-0.10, +0.00]	+0.00 [-0.10, +0.10]	+0.00 [-0.10, +0.10]
Genos	Acceptor	+0.00 [-0.10, +0.10]	-0.03 [-0.17, +0.07]	+0.00 [-0.14, +0.14]
Genos	PPT	+0.23 [-0.39, +0.87]	-0.31 [-1.02, +0.46]	-0.31 [-1.02, +0.39]
Genos	BP	+0.45 [-2.50, +3.52]	+0.98 [-1.67, +3.75]	-0.98 [-3.54, +1.44]
Evo 2	Donor	+0.30 [-0.27, +0.84]	+0.51 [-0.07, +1.11]	+0.30 [-0.17, +0.81]
Evo 2	Acceptor	-0.17 [-0.61, +0.24]	-0.31 [-0.65, +0.00]	-0.24 [-0.48, +0.00]
Evo 2	PPT	-1.53 [-3.45, +0.45]	-1.15 [-2.79, +0.62]	-2.30 [-3.93, -0.62]
Evo 2	BP	-1.43 [-4.01, +1.31]	-1.07 [-3.63, +1.39]	-0.80 [-3.11, +1.49]

Together, the geometric and functional comparisons assess stability of subspace discovery conditional on fixed layers and dimensions. They do not establish invariance to the choice of discovery

backgrounds, recovery of all component information, or stability of rerunning the complete selection procedure.

E Implementation Details of Subspace-Based Rule Faithfulness

This section details the state-replacement procedure and directional evaluation in main-text Section 5.2. We use the same component bases Q_c, selected layers ℓ_c, joint bases $Q_{C,\ell}$ and projectors $\Pi_{C,\ell}$ as in the main text. The pretrained models, component spaces and task heads remain fixed. The source representations are obtained from unmodified forward passes of the experimentally tested sequences; experimental measurements are used for evaluation, not to fit the replacement operation.

E.1 Joint bases and DNA-coordinate alignment

At each intervention layer, we concatenate the bases assigned to that layer and compute their singular value decomposition. We retain left singular vectors with singular values greater than 10^{-6} times the largest singular value, followed by reduced QR orthogonalization for numerical stability. This constructs a basis for the numerical joint span, rather than adding the component projectors. Closely aligned but linearly independent directions can remain in the joint span; geometric similarity alone does not imply that their dimensions are removed. Components assigned to different layers are handled separately.

All sequences within a comparison have the same DNA length and corresponding nucleotide positions, but their token boundaries can differ. Let $\tau_x(p)$ identify the token covering DNA position p in sequence x. We expand source token states to DNA coordinates using

$$h_{\ell,p}(x_s) = v_{\ell,\tau_{x_s}(p)}(x_s), \tag{S12}$$

where $v_{\ell,t}(x_s)$ is the native hidden state of source token t. Expansion is performed separately for each source sequence; token indices are not assumed to align across sequences.

The recipient retains the tokenization of the reference sequence $x_{\mathbf{0}}$, where $\mathbf{0} = 00$ for two components and 000 for three. Let $\mathcal{P}_t$ be the DNA positions covered by recipient token t, and let $w_t = |\mathcal{P}_t|$. For $w_t > 0$, the target coordinates for that token are

$$\bar{z}^{(s)}_{\ell,t} = \frac{1}{w_t} \sum_{p \in \mathcal{P}_t} Q_{C,\ell}^{\top} h_{\ell,p}(x_s). \tag{S13}$$

The native-token implementation of state replacement is

$$\widetilde{v}^{(s)}_{\ell,t} = v^{\mathrm{cur},(s)}_{\ell,t} + Q_{C,\ell} \left(\bar{z}^{(s)}_{\ell,t} - Q_{C,\ell}^{\top} v^{\mathrm{cur},(s)}_{\ell,t} \right). \tag{S14}$$

Tokens with no DNA coverage, including special and padding tokens, are not directly modified. When source and recipient token boundaries coincide, this is ordinary coordinate replacement. When they differ, the target coordinates are averaged over the recipient token's DNA span; the implementation does not independently assign incompatible states to nucleotides sharing one recipient token. This procedure applies across the entire input, not only a local component window.

E.2 Execution across layers and task readout

For each comparison, we cache unmodified source representations at all required intervention layers. We then perform a recipient forward pass for each non-reference source state, always using $x_{\mathbf{0}}$ as the input. The same layer-specific joint spaces are used for every state. At each selected layer, the replacement in Eq. (S14) is applied to the hidden states currently present in that forward pass. Thus, a later replacement acts on states already affected by earlier interventions. It does not add a precomputed source-minus-reference difference to an unchanged hidden state. The double- and triple-change source states are taken from their actual experimental sequences, not constructed by adding single-change differences.

Immediately at replacement, the current representation outside the joint span is preserved up to numerical precision. All subsequent network computations proceed normally, so other directions and positions may change downstream. The reference-state score is reused from its unmodified forward pass, since replacing its coordinates with its own coordinates is the identity operation.

At the final layer L_{final}, we average over the L DNA positions and apply the existing standardized linear task head:

$$u_s = \frac{1}{L}\sum_{p=1}^{L} \widetilde{h}^{(s)}_{L_{\text{final}},p}, \qquad f_s = \beta^\top \left[(u_s - \mu) \oslash \sigma\right] + b, \tag{S15}$$

where $\oslash$ denotes element-wise division. The standardization parameters and coefficients are those of the previously fitted task head and are not re-estimated on the relationship-evaluation set. DNA-position averaging is equivalent to weighting each native token by its DNA coverage, excluding special tokens. COMPASS and yeast use their respective task heads. Outputs are task scores; they are not clipped or recalibrated to the interval $[0, 100]$ during relation evaluation.

E.3 Four- and eight-state interaction contrasts

State indices refer to the experimental sequence configurations, not necessarily component absence and presence. We preserve the same component identities and state ordering in the biological and model contrasts. For two components, the contrast is $J = f_{11} - f_{10} - f_{01} + f_{00}$, with experimental measurements substituted for f to obtain J_{bio}.

For the yeast three-component experiment, the indices (a, b, c) refer to Donor, BP and Acceptor, respectively. Define the Donor–BP contrast at fixed Acceptor state c as

$$J_{AB|c} = f_{11c} - f_{10c} - f_{01c} + f_{00c}. \tag{S16}$$

The third-order contrast is

$$J^{(3)}_{\text{model}} = J_{AB|1} - J_{AB|0} \tag{S17}$$

$$= f_{111} - f_{101} - f_{011} + f_{001} - f_{110} + f_{100} + f_{010} - f_{000}. \tag{S18}$$

The biological third-order contrast uses the corresponding eight measured values. Equivalently, for $k \in \{2, 3\}$ components,

$$J^{(k)} = \sum_{s\in\{0,1\}^k} (-1)^{k-\sum_j s_j} f_s. \tag{S19}$$

This is a contrast of one set of joint-state outputs, not an average of single-component path scores. Contrasts are evaluated on the specified measurement and task-score scales; changing to a nonlinear output transformation would define a different contrast.

E.4 Experimental directions, zero responses and aggregation

For COMPASS, direction evaluation retains comparisons passing the existing quality and exclusion criteria. Each of the four states must have at least ten total reads in each of two experimental replicates. We calculate the four-state PSI contrast separately for each replicate and require matching signs, with both absolute contrasts exceeding 10^{-5} PSI percentage points. Their common sign defines the reference direction $d_i \in \{-1, +1\}$. A nonzero contrast after pooling reads does not override disagreement between replicate directions. K562 is the primary presentation; other cell lines are evaluated separately rather than counted as independent gene backgrounds.

For yeast, the reference direction is the sign of the measured contrast, retaining values with absolute magnitude greater than 10^{-8} percentage points to exclude numerical zeros. Unlike the COMPASS replicate criterion, this is a nonzero point-estimate criterion, not a replicate-based significance test. All tested yeast combinations derive from one MUD1-based background and can share sequence states.

For model m and an evaluable comparison i, we use

$$a_{m,i} = \mathbf{1}[|J_{\text{model},m,i}| > \epsilon_{\text{model}} \ \wedge \ \text{sign}(J_{\text{model},m,i}) = d_i]\,. \tag{S20}$$

The model zero-response tolerances are 10^{-5} task-score units for COMPASS and 0.01 task-score units for yeast, applied to both second- and third-order model contrasts where applicable. These are numerical response tolerances, not biological significance thresholds. Responses at or below the tolerance receive zero and remain in the evaluation denominator. Comparisons lacking an eligible experimental direction are excluded from directional evaluation, irrespective of model predictions.

COMPASS BRF follows the gene-equal aggregation in main-text Section 5.2. For yeast, letting $\mathcal{I}_r$ denote the evaluable combinations for relationship r, we report

$$\mathrm{BRF}_{m,r}^{\mathrm{yeast}} = \frac{100}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} a_{m,i}. \tag{S21}$$

Every model is evaluated on the same eligible set for a given relationship and cell line. Rankings, where shown, use BRF itself and are relationship-specific; no score is pooled across relationships or species.

For human results, descriptive 95% intervals are obtained from 2,000 bootstrap samples of genes with replacement, preserving within-gene comparisons and using seed 20260925. Component-versus-random differences use the same sampled genes. These intervals are pointwise and are not adjusted for multiple comparisons. We do not report cross-gene bootstrap intervals for the single-background yeast library. Constant-positive and constant-negative direction references and model zero-response rates are retained in the detailed results; 50% is not assumed to be a universal no-information baseline.

E.5 Matched random-subspace controls

For each model, dataset and relationship, we generate three random controls. At every intervention layer, a Gaussian matrix with the same dimensions as the numerical joint basis is orthonormalized by QR. Each random basis is fixed across all comparisons and source states in that panel. Reproducible seeds are derived from the model, dataset, panel and control index. The controls use the same source states, recipient sequence, intervention layers and fixed task head as the component intervention.

Perturbation magnitude is matched per comparison and layer, using the sum of squared requested native-token updates weighted by DNA coverage. If $\delta_{\ell,t}^{(s)}$ denotes the update before casting to the model's native dtype, define

$$E_{i,\ell} = \sum_{s \neq \mathbf{0}} \sum_{t : w_t > 0} w_t \|\delta_{\ell,t}^{(s)}\|_2^2. \tag{S22}$$

For random control j, its raw coordinate-replacement update is multiplied by

$$\alpha_{i,\ell}^{(j)} = \sqrt{E_{i,\ell}^{\mathrm{component}} / E_{i,\ell}^{\mathrm{random},j,\mathrm{raw}}}. \tag{S23}$$

The same multiplier is used for every non-reference state in that comparison at that layer; states are not normalized independently. A zero component target energy produces a zero multiplier, and a positive target requires nonzero raw random energy. Because the multiplier need not equal one, the matched random intervention is a scaled displacement toward the source coordinates, not necessarily an exact coordinate replacement.

For multiple intervention layers, multipliers are determined in network order. The raw energy at a later layer is measured after propagation through the earlier scaled random interventions. The final forward pass verifies the per-comparison, per-layer energy match, using a relative tolerance of 2×10^{-4}. This matches requested FP32 update energy, not necessarily the realized energy after casting to a lower-precision network dtype.

We calculate directional agreement separately for each random control and average the three agreement rates. This is not the agreement rate of their averaged interaction contrast. Random-control agreement and component-minus-random differences are reported separately: they do not redefine BRF, alter its ranking, or determine whether a model is included in evaluation.

E.6 Numerical checks and supplementary correspondence analysis

Projection and coordinate-update calculations use FP32 while preserving each backbone's existing execution precision. Source representations are cached in DNA coordinates, with projected-coordinate caches checked against the basis used by each worker. This avoids reusing coefficients under a numerically rotated basis with the same span.

Numerical checks verify that source-equals-recipient replacement reproduces the unmodified output, that the joint bases are orthonormal within tolerance, and that the first-layer update energy agrees

with an explicit coordinate calculation. Synthetic additive and non-additive state arrays test the signs of both interaction formulas. Result checks ensure that every combination has one complete component result and three random results, that outputs are finite, and that each stored interaction equals the contrast reconstructed from its state scores. An independent implementation for SpliceBERT and HyenaDNA reproduced the COMPASS interaction-direction assignments for all 406 combinations. This is numerical replication, not additional biological data.

As a supplementary analysis, we compute Spearman correspondence between signed experimental and joint-subspace interaction contrasts. For COMPASS, each gene receives total weight one and correlations are calculated from weighted midranks, with gene-bootstrap intervals. This analysis includes all quality-eligible continuous measurements, so its sample size can exceed that of directional evaluation. For the single-background yeast library, it reduces to ordinary within-library rank correlation. Undefined correlations for constant inputs are left undefined rather than assigned zero. Spearman is not used to define BRF or replace its relationship-specific rankings.

F External Evaluation of Exon-Spanning Relationships with FAS

F.1 Biological assay and relationship being evaluated

We used the FAS exon 6 mutational dataset of Julien et al. (2016) to test whether the COMPASS-derived exon-spanning Acceptor–Donor (A–D) ranking recommended models that better predicted experimentally measured non-additive effects in another reporter assay. The original study introduced single and double substitutions into the 63-nucleotide exon, expressed the reporter library in HEK293 cells and quantified exon inclusion by sequencing. Its enrichment scores compare variant abundance in exon-including RNA with abundance in the input library. We used the reported wild-type (WT)-normalized scores, rather than treating them as absolute percent-spliced-in values.

The relationship here spans one exon: Acceptor–target exon–Donor. Its BRF ranking was obtained from the human COMPASS K562 comparisons and was held fixed throughout the FAS evaluation. It is distinct from the intron-spanning Donor–Acceptor relationship measured in the yeast Schirman library. We did not substitute the yeast ranking or pool the two relationships.

F.2 Sequence construction and held-out comparisons

We selected measured double mutants with one substitution at exon positions 1–3 and the other at positions 61–63, using one-based coordinates. For each double mutant, we retrieved the corresponding two measured single mutants and used WT as the reference, yielding 70 complete four-state comparisons and 89 unique sequence states. The states were ordered as WT, acceptor-side single mutant, donor-side single mutant and double mutant. These substitutions altered the exonic contexts of the two boundaries; the intronic AG and GT dinucleotides were unchanged.

Each model received the published 135-nucleotide local fragment: 22 upstream intronic nucleotides, the 63-nucleotide exon and 50 downstream intronic nucleotides. The flanking sequences were `TGTCCAATGTTCCAACCTACAG` and

`GTAAGTTCTTGCTTTGTTCAAACTGCAGATTGAAATAACTTGGGAAGTAG`.

This input represents the local insert, not the full reporter plasmid.

All 89 test states were excluded from readout training and validation. We additionally excluded every development variant containing a substitution at any of the six boundary positions, removing 2,966 variants and leaving 13,951 eligible variants. We sorted the eligible variants by the SHA-256 hash of the string `fas-heldends-20260926:` concatenated with the original variant identifier. The first 2,000 variants formed the training set and the next 500 formed the validation set; the remaining eligible variants were unused. Thus, evaluation tested held-out mutation positions and combinations within FAS, rather than transfer to a held-out gene. The 70 comparisons share WT and single-mutant states and are not 70 independent biological backgrounds.

F.3 Frozen representations and task-specific readouts

For each of the six foundation models, we froze the backbone and averaged final-layer hidden states over all DNA positions in the 135-nucleotide input, excluding special tokens. Token states covering multiple nucleotides were weighted by their DNA coverage. We standardized each feature using its training-set mean and standard deviation; standard deviations below 10^{-6} were replaced by one.

Let $e(x)$ denote the reported $\log_2$ WT-normalized enrichment. We trained a separate ridge-regression readout for each model to predict

$$y(x) = 2^{e(x)}, \qquad y(x_{00}) = 1, \tag{S24}$$

using the same training and validation sequences for all models. Regularization was selected from $\{1000, 100, 10, 1, 0.1\}$ by validation mean squared error. The solver was LSQR with tolerance 10^{-5} and at most 5,000 iterations. Feature storage, standardization and readout inputs used FP32. The selected regularization was 100 for SpliceBERT and 1,000 for each other model. Predictions were not clipped or calibrated on the test set, and the predicted WT value was obtained from the readout rather than forced to one.

These were newly fitted FAS readouts on frozen representations, not the previously fitted natural-PSI heads. No subspace intervention was performed in this external prediction experiment: the purpose was to test the external relevance of an existing internal BRF ranking, not to estimate a new BRF ranking from FAS labels.

F.4 Interaction prediction and benchmark disagreements

For comparison i, we calculated the experimental and predicted contrasts on the same WT-normalized relative-enrichment scale:

$$J_i^{\mathrm{bio}} = y_{11,i} - y_{10,i} - y_{01,i} + y_{00,i}, \tag{S25}$$

$$\widehat{J}_{m,i} = \widehat{y}_{m,11,i} - \widehat{y}_{m,10,i} - \widehat{y}_{m,01,i} + \widehat{y}_{m,00,i}. \tag{S26}$$

The primary external metric was interaction-direction accuracy. Comparisons with $|J_i^{\mathrm{bio}}| \leq 10^{-8}$ were excluded; all 70 comparisons remained. A matching predicted sign received one, an opposite sign zero, and $|\widehat{J}_{m,i}| \leq 10^{-8}$ received one-half. The mean was reported as a percentage. This tie convention belongs to the external task metric and does not modify the internal BRF zero-response convention in Appendix E.4. Secondary metrics were Spearman correlation between the 70 predicted and measured contrasts and between predictions and measurements for the 89 unique test states. There were 35 positive and 35 negative experimental contrasts, giving either constant-direction prediction an accuracy of 50%.

We compared the fixed COMPASS exon-spanning A–D ranking with two benchmark rankings: NT Donor recognition and GUE splice-site classification, using our measured benchmark MCC values. These are conventional splice-site benchmarks, not dedicated A–D interaction benchmarks. For each unordered model pair with opposite BRF and benchmark preferences, we compared the two models' FAS interaction-direction accuracies. The recommendation with the higher point estimate was counted as the winner; equality was a tie. Confidence intervals were not used to determine wins. All discordant pairs were retained. Pairwise outcomes share models and test sequences, so win fractions are descriptive counts rather than independent-trial estimates.

F.5 Six-model results

Table S26 reports absolute external performance. Across six models, BRF selected the higher-accuracy model in 8 of 9 disagreements with NT Donor and 7 of 8 disagreements with GUE (Table S27). The sole loss under both references was the recommendation of Evo2 over DNABERT-2. Restricting the comparison to SpliceBERT, NTv2, HyenaDNA and DNABERT-2 gave 2/2 wins against NT Donor and 3/3 against GUE. Thus, the positive recommendation result was also present without Genos and Evo2.

Table S26: FAS held-out performance. Direction accuracy and interaction Spearman use 70 four-state comparisons; state Spearman uses 89 unique sequences. All models use the same split and frozen-backbone readout protocol.

Model	Direction accuracy (%)	Interaction Spearman	State Spearman
SpliceBERT	47.14	0.014	0.512
NTv2	41.43	−0.291	0.323
HyenaDNA	58.57	0.213	0.338
DNABERT-2	57.14	0.147	0.517
Genos	65.71	0.401	0.416
Evo2	51.43	−0.085	0.084

This additional external evaluation supports the relevance of the COMPASS-derived ranking for predicting held-out exon-boundary combinations in FAS. It was an exploratory extension of the external evaluations, not a new independent-gene replication or evidence that the same ranking applies to intron-spanning relationships. The measured target is relative enrichment, and its four-state contrast is specific to that scale; it is not an interaction contrast in absolute PSI or the original study's fitted epistasis score.

Table S27: All six-model FAS ranking disagreements. The difference is FAS direction accuracy of the BRF-recommended model minus that of the benchmark-recommended model, in percentage points. "Both" denotes NT Donor and GUE. Positive differences favour BRF. Rows shared by two references are shown once.

Benchmark reference	BRF recommends	Benchmark recommends	Difference
Both	Evo2	DNABERT-2	-5.71
Both	Evo2	NTv2	$+10.00$
Both	Evo2	SpliceBERT	$+4.29$
Both	Genos	NTv2	$+24.29$
Both	Genos	SpliceBERT	$+18.57$
Both	HyenaDNA	NTv2	$+17.14$
Both	HyenaDNA	SpliceBERT	$+11.43$
NT Donor	Genos	DNABERT-2	$+8.57$
NT Donor	Genos	HyenaDNA	$+7.14$
GUE	HyenaDNA	DNABERT-2	$+1.43$

G Component Subspaces Before and After Fine-tuning

We examined whether task fine-tuning changed the geometry and functional contribution of previously identified component subspaces. Two complementary experiments were conducted: a small-sample geometric comparison in DNABERT-2, SpliceBERT and HyenaDNA, and a matched-readout erasure comparison in these three models plus NTv2 and Genos. The erasure comparison covered two recognition tasks and four individually intervened components, giving 40 paired before–after comparisons. Evo 2 is not included in these comparisons because the correspondence between its subspace-extraction and fine-tuning implementations had not been resolved. The original component layers, dimensions and bases were not reselected using these results.

G.1 Task Checkpoints and Matched Readouts

For each model, we compared the pretrained checkpoint with two separately adapted checkpoints: one fine-tuned for donor recognition and one for acceptor recognition. DNABERT-2, SpliceBERT, HyenaDNA and Genos used full-backbone fine-tuning; NTv2 used an IA3 parameter-efficient reconstruction with learned key, value and feed-forward scaling in its 29 layers. Fine-tuning protocols were model-specific rather than a controlled comparison of adaptation algorithms. All subsequent representation measurements froze the checkpoint parameters.

Both recognition tasks used the revised Nucleotide Transformer downstream datasets described in Section C.1. Here we used the full 30,000-sequence training partition and its unchanged 3,000-sequence official test partition, with 600 bases per sequence. A fixed random 90:10 split of the training partition (seed 42) provided 27,000 training and 3,000 validation examples. These partitions were identical across models and checkpoint states; they were not the overlap-filtered recognition subsets used for functional dimension selection in Section C. This comparison consequently measures changes under a fixed benchmark interface, rather than providing an additional independent confirmation of subspace discovery.

We did not reuse the original fine-tuning classification heads. Instead, for each model, task and checkpoint state, we fitted an identically specified linear readout to unmodified final-layer representations. Token states were averaged over the entire sequence with weights equal to the number of DNA bases covered by each token; special and padding tokens had zero weight. Training features were standardized coordinate-wise using training-set means and standard deviations. A binary logistic-regression readout used $C = 1$, the L-BFGS solver, tolerance 10^{-6}, a maximum of 2,000 iterations and seed 42; a logit greater than zero was classified as positive. Validation data were used only for monitoring, without selecting new readout hyperparameters, component layers or dimensions. Each readout was trained once on its own checkpoint's normal training representations and remained fixed for every subsequent intervention. Thus, the two checkpoint states shared the readout architecture and fitting procedure, but not the fitted readout coefficients. The normal accuracies and Matthews correlation coefficients (MCC) in Table S28 describe this matched-readout interface, not the original fine-tuning heads' benchmark performance.

Table S28: Normal test performance under matched linear-readout training. Each task has 3,000 test sequences. Pre and Post refer to pretrained and task-fine-tuned backbones; these are not the original fine-tuning-head benchmark scores.

Model	Task	Pre Acc. (%)	Post Acc. (%)	Pre MCC	Post MCC
DNABERT-2	Donor	82.03	91.40	0.6416	0.8280
DNABERT-2	Acceptor	80.70	90.87	0.6161	0.8173
SpliceBERT	Donor	80.73	93.93	0.6146	0.8792
SpliceBERT	Acceptor	79.80	95.63	0.5961	0.9127
HyenaDNA	Donor	73.00	94.67	0.4672	0.8934
HyenaDNA	Acceptor	73.03	90.90	0.4684	0.8182
NTv2	Donor	76.43	97.47	0.5304	0.9493
NTv2	Acceptor	74.13	96.80	0.4841	0.9360
Genos	Donor	79.00	82.40	0.5810	0.6491
Genos	Acceptor	79.20	79.93	0.5845	0.5997

G.2 Fixed-space Erasure and Random Controls

We retained the full-discovery component bases and the selected layers and dimensions in Table S13. For each task and checkpoint state, the intervention centre was estimated at each selected layer using the same 256 training examples, sampled without replacement with seed 42. The centre was computed from that checkpoint's own DNA-base-weighted hidden states. Each of Donor, Acceptor, PPT and BP was erased separately, including non-target spaces: for example, the donor-recognition head was also evaluated after Acceptor, PPT or BP erasure. This was not simultaneous four-component erasure.

Writing $s \in \{\mathrm{pre}, \mathrm{post}\}$ for checkpoint state, the intervention on DNA-position row vectors was

$$\widetilde{H}_s^{(\ell)}(x) = H_s^{(\ell)}(x) - \left(H_s^{(\ell)}(x) - \mathbf{1}\mu_{s,\ell}^\top\right) Q_c Q_c^\top . \tag{S27}$$

After intervention, all remaining backbone layers were recomputed, followed by final-layer pooling and the fixed state-specific readout. The operation covered all DNA positions at the selected layer, rather than only the original component window. Three random orthonormal bases were drawn at the same layer and dimension, using the same seeds in the two checkpoint states. For each sequence and checkpoint state, each random displacement was rescaled to match the component displacement's DNA-base-weighted squared norm. The same control procedure was therefore used before and after fine-tuning, while its displacement magnitude was matched within each state. Projection arithmetic used FP32; forward evaluation used FP16 autocast for DNABERT-2 and SpliceBERT, BF16 autocast for HyenaDNA and Genos, and FP32 for NTv2, identically on both sides of each model's comparison.

Let a_s^0, a_s^c and $a_s^{R_k}$ denote normal, component-erased and random-intervention test accuracies as proportions. We report the two direct decreases separately, in percentage points:

$$D_s^c = 100\left(a_s^0 - a_s^c\right), \qquad D_s^R = \frac{100}{3}\sum_{k=1}^{3}\left(a_s^0 - a_s^{R_k}\right). \tag{S28}$$

Neither column subtracts the other; negative values indicate a small improvement over normal accuracy and are not clipped. Pointwise 95% percentile intervals used 1,000 paired bootstrap resamples of the same 3,000 test rows, with identical sampled indices across conditions and checkpoint states. These intervals are conditional on the fitted checkpoints, readouts and three fixed random bases; they do not quantify variation across fine-tuning seeds or alternative random bases, and they are not gene-cluster intervals. They are descriptive and are not adjusted for multiple comparisons. Tables S29–S33 report every completed model–task–component condition, including the original accuracy for each checkpoint state.

Table S29: DNABERT-2: direct component-erasure and matched-random accuracy decreases before and after fine-tuning. Entries are percentage points with pointwise 95% paired-bootstrap intervals. Normal accuracy is in percent. Each component is erased separately, and each random value averages three fixed directions; negative decreases are retained.

Task	Space	State	Normal (%)	Component drop	Random drop
Donor	Donor	Pre	82.03	12.60 [10.83, 14.33]	2.54 [1.78, 3.34]
Donor	Donor	Post	91.40	23.67 [21.87, 25.57]	3.04 [2.14, 4.00]
Donor	Acceptor	Pre	82.03	17.53 [15.67, 19.50]	0.64 [−0.09, 1.32]
Donor	Acceptor	Post	91.40	15.67 [14.00, 17.17]	0.40 [−0.20, 1.03]
Donor	PPT	Pre	82.03	11.47 [9.67, 13.17]	2.69 [1.88, 3.47]
Donor	PPT	Post	91.40	10.27 [8.87, 11.77]	2.98 [2.09, 3.94]
Donor	BP	Pre	82.03	14.97 [13.10, 16.80]	0.61 [−0.12, 1.29]
Donor	BP	Post	91.40	16.50 [14.67, 18.17]	0.39 [−0.22, 1.02]
Acceptor	Donor	Pre	80.70	14.00 [12.30, 15.73]	3.19 [2.32, 4.08]
Acceptor	Donor	Post	90.87	12.10 [10.67, 13.57]	4.90 [3.81, 5.93]
Acceptor	Acceptor	Pre	80.70	14.97 [13.07, 16.87]	1.04 [0.26, 1.87]
Acceptor	Acceptor	Post	90.87	10.83 [9.27, 12.37]	−0.21 [−0.72, 0.34]

Continued on next page

Table S29 (continued)

Task	Space	State	Normal (%)	Component drop	Random drop
Acceptor	PPT	Pre	80.70	14.63 [12.87, 16.33]	3.18 [2.29, 4.07]
Acceptor	PPT	Post	90.87	11.37 [9.87, 12.93]	4.98 [3.93, 6.01]
Acceptor	BP	Pre	80.70	13.23 [11.53, 15.00]	0.91 [0.17, 1.69]
Acceptor	BP	Post	90.87	8.33 [6.87, 9.80]	−0.26 [−0.77, 0.27]

Table S30: SpliceBERT: direct component-erasure and matched-random accuracy decreases before and after fine-tuning. Entries are percentage points with pointwise 95% paired-bootstrap intervals. Normal accuracy is in percent. Each component is erased separately, and each random value averages three fixed directions; negative decreases are retained.

Task	Space	State	Normal (%)	Component drop	Random drop
Donor	Donor	Pre	80.73	26.63 [24.33, 28.80]	1.11 [0.30, 1.92]
Donor	Donor	Post	93.93	43.27 [41.33, 45.17]	1.02 [0.48, 1.53]
Donor	Acceptor	Pre	80.73	18.70 [16.77, 20.67]	3.97 [2.92, 4.97]
Donor	Acceptor	Post	93.93	43.27 [41.33, 45.17]	2.99 [2.24, 3.71]
Donor	PPT	Pre	80.73	15.00 [13.03, 16.87]	3.94 [2.89, 4.98]
Donor	PPT	Post	93.93	43.27 [41.33, 45.17]	2.92 [2.19, 3.63]
Donor	BP	Pre	80.73	22.93 [20.53, 24.90]	1.03 [0.21, 1.90]
Donor	BP	Post	93.93	42.37 [40.47, 44.27]	0.96 [0.40, 1.49]
Acceptor	Donor	Pre	79.80	28.97 [26.60, 31.00]	2.12 [1.18, 3.07]
Acceptor	Donor	Post	95.63	43.50 [41.57, 45.43]	2.81 [2.02, 3.52]
Acceptor	Acceptor	Pre	79.80	27.23 [24.90, 29.40]	4.59 [3.37, 5.84]
Acceptor	Acceptor	Post	95.63	44.67 [42.77, 46.57]	8.40 [7.39, 9.44]
Acceptor	PPT	Pre	79.80	28.17 [25.87, 30.27]	4.56 [3.31, 5.86]
Acceptor	PPT	Post	95.63	40.50 [38.50, 42.30]	8.29 [7.28, 9.32]
Acceptor	BP	Pre	79.80	28.67 [26.30, 30.67]	2.22 [1.26, 3.16]
Acceptor	BP	Post	95.63	29.13 [27.40, 31.10]	2.81 [2.07, 3.56]

Table S31: HyenaDNA: direct component-erasure and matched-random accuracy decreases before and after fine-tuning. Entries are percentage points with pointwise 95% paired-bootstrap intervals. Normal accuracy is in percent. Each component is erased separately, and each random value averages three fixed directions; negative decreases are retained.

Task	Space	State	Normal (%)	Component drop	Random drop
Donor	Donor	Pre	73.00	16.80 [14.73, 18.80]	8.46 [7.03, 9.90]
Donor	Donor	Post	94.67	17.20 [15.77, 18.73]	4.26 [3.59, 4.93]
Donor	Acceptor	Pre	73.00	16.80 [14.77, 18.77]	7.57 [6.14, 8.93]
Donor	Acceptor	Post	94.67	9.27 [8.03, 10.53]	3.31 [2.69, 3.93]
Donor	PPT	Pre	73.00	4.83 [3.03, 6.60]	2.01 [0.96, 2.96]
Donor	PPT	Post	94.67	0.30 [−0.13, 0.73]	0.54 [0.18, 0.89]
Donor	BP	Pre	73.00	17.43 [15.27, 19.60]	8.49 [7.09, 9.95]
Donor	BP	Post	94.67	12.77 [11.40, 14.23]	4.36 [3.70, 5.01]
Acceptor	Donor	Pre	73.03	17.83 [15.70, 20.07]	9.06 [7.68, 10.47]
Acceptor	Donor	Post	90.90	1.57 [0.50, 2.60]	0.67 [0.12, 1.23]
Acceptor	Acceptor	Pre	73.03	17.37 [15.33, 19.40]	8.14 [6.81, 9.49]
Acceptor	Acceptor	Post	90.90	3.30 [2.20, 4.47]	0.60 [0.08, 1.16]
Acceptor	PPT	Pre	73.03	6.40 [4.87, 8.10]	2.24 [1.27, 3.27]
Acceptor	PPT	Post	90.90	1.53 [0.77, 2.30]	0.09 [−0.30, 0.47]
Acceptor	BP	Pre	73.03	17.90 [15.90, 19.90]	9.23 [7.82, 10.71]
Acceptor	BP	Post	90.90	4.17 [3.07, 5.40]	0.73 [0.18, 1.31]

Table S32: NTv2: direct component-erasure and matched-random accuracy decreases before and after fine-tuning. Entries are percentage points with pointwise 95% paired-bootstrap intervals. Normal accuracy is in percent. Each component is erased separately, and each random value averages three fixed directions; negative decreases are retained.

Task	Space	State	Normal (%)	Component drop	Random drop
Donor	Donor	Pre	76.43	16.77 [14.57, 19.03]	0.07 [−0.66, 0.89]
Donor	Donor	Post	97.47	28.70 [27.07, 30.20]	0.08 [−0.23, 0.38]
Donor	Acceptor	Pre	76.43	0.33 [−0.77, 1.43]	0.20 [−0.41, 0.83]
Donor	Acceptor	Post	97.47	0.30 [−0.10, 0.70]	−0.01 [−0.29, 0.23]
Donor	PPT	Pre	76.43	16.93 [14.67, 19.17]	0.68 [−0.26, 1.62]
Donor	PPT	Post	97.47	0.23 [−0.20, 0.70]	−0.12 [−0.43, 0.17]
Donor	BP	Pre	76.43	24.70 [22.60, 26.87]	1.37 [0.56, 2.19]
Donor	BP	Post	97.47	23.40 [21.83, 24.93]	0.11 [−0.23, 0.47]
Acceptor	Donor	Pre	74.13	18.87 [16.80, 21.10]	−0.46 [−1.26, 0.46]
Acceptor	Donor	Post	96.80	10.67 [9.47, 11.97]	0.09 [−0.21, 0.40]
Acceptor	Acceptor	Pre	74.13	0.47 [−0.63, 1.67]	0.09 [−0.61, 0.83]
Acceptor	Acceptor	Post	96.80	0.17 [−0.23, 0.63]	0.02 [−0.21, 0.26]
Acceptor	PPT	Pre	74.13	15.97 [13.93, 18.07]	1.61 [0.44, 2.89]
Acceptor	PPT	Post	96.80	6.43 [5.50, 7.47]	0.04 [−0.34, 0.49]
Acceptor	BP	Pre	74.13	22.87 [20.60, 25.20]	0.92 [0.08, 1.78]
Acceptor	BP	Post	96.80	8.47 [7.37, 9.60]	0.21 [−0.17, 0.56]

Table S33: Genos: direct component-erasure and matched-random accuracy decreases before and after fine-tuning. Entries are percentage points with pointwise 95% paired-bootstrap intervals. Normal accuracy is in percent. Each component is erased separately, and each random value averages three fixed directions; negative decreases are retained.

Task	Space	State	Normal (%)	Component drop	Random drop
Donor	Donor	Pre	79.00	28.33 [25.90, 30.80]	0.73 [−0.22, 1.66]
Donor	Donor	Post	82.40	28.30 [25.97, 30.60]	0.19 [−0.61, 0.99]
Donor	Acceptor	Pre	79.00	25.80 [23.47, 28.17]	0.41 [−0.51, 1.26]
Donor	Acceptor	Post	82.40	13.60 [11.60, 15.60]	0.03 [−0.69, 0.79]
Donor	PPT	Pre	79.00	28.33 [25.90, 30.80]	22.97 [20.64, 25.33]
Donor	PPT	Post	82.40	31.73 [29.43, 34.03]	8.89 [7.27, 10.61]
Donor	BP	Pre	79.00	28.33 [25.90, 30.80]	13.46 [11.53, 15.37]
Donor	BP	Post	82.40	31.73 [29.43, 34.03]	4.30 [2.87, 5.83]
Acceptor	Donor	Pre	79.20	28.37 [25.93, 30.77]	1.09 [0.19, 2.03]
Acceptor	Donor	Post	79.93	26.77 [24.13, 29.03]	−0.02 [−0.91, 0.81]
Acceptor	Acceptor	Pre	79.20	28.13 [25.70, 30.57]	0.91 [0.06, 1.78]
Acceptor	Acceptor	Post	79.93	15.70 [13.40, 17.80]	−0.16 [−0.94, 0.58]
Acceptor	PPT	Pre	79.20	28.37 [25.93, 30.77]	27.18 [24.73, 29.60]
Acceptor	PPT	Post	79.93	29.10 [26.47, 31.33]	12.00 [10.12, 13.87]
Acceptor	BP	Pre	79.20	28.37 [25.93, 30.77]	19.72 [17.67, 21.86]
Acceptor	BP	Post	79.93	29.10 [26.47, 31.33]	6.26 [4.63, 7.82]

G.3 Exploratory Geometric Comparison

For DNABERT-2, SpliceBERT and HyenaDNA, we additionally re-estimated component spaces at the fixed original layer and dimension using a small, identical set of pairs in the pretrained and fine-tuned checkpoints. For each component, 16 discovery-training pairs were used to estimate the space and 16 discovery-test pairs to measure held-out response retention. Each set contained eight pairs from each of the two discovery editing families; backgrounds were distinct within each component's selected train/test pairs. The pairs and their original split identities were unchanged across checkpoint states. We aligned hidden states to all DNA positions, formed reference-minus-variant

differences and concatenated the 16 training-pair difference matrices by rows. The leading eigenvectors of the uncentred difference second-moment matrix gave the fixed-rank basis, equivalently the right singular vectors of the concatenated difference matrix.

Let $\widehat{Q}_{c,\mathrm{pre}}$ and $\widehat{Q}_{c,\mathrm{post}}$ denote these two small-sample bases. Their geometric overlap was

$$O_c = \frac{100}{r_c} \left\| \widehat{Q}_{c,\mathrm{pre}}^{\top} \widehat{Q}_{c,\mathrm{post}} \right\|_F^2 . \tag{S29}$$

This is the mean squared cosine of the principal angles, expressed as a percentage. Separately, for each held-out pair i, we computed the fraction of squared response length captured by the original full-discovery basis Q_c and averaged over the 16 pairs:

$$E_{c,s} = \frac{100}{16} \sum_{i=1}^{16} \frac{\|\Delta H_{c,i,s} Q_c\|_F^2}{\|\Delta H_{c,i,s}\|_F^2}. \tag{S30}$$

Overlap compares the two re-estimated bases; retention uses the original basis. Neither quantity is classification accuracy or a fraction of all biological information recovered. The small-sample re-estimated bases were used only for geometric exploration and did not replace the full-discovery bases in the erasure experiments. Table S34 includes all 24 completed geometric comparisons; no corresponding geometric estimate is asserted for NTv2 or Genos.

Table S34: Exploratory geometry using the same 16 discovery pairs per component before and after fine-tuning and 16 separate held-out pairs for retention. Retention measures the fraction of squared perturbation-response length captured by the original full-discovery basis. Overlap compares the two small-sample re-estimated bases, not either basis against the original full-discovery space. All values are percentages.

Model	Space	Fine-tuning task	Pre retention	Post retention	Overlap
DNABERT-2	Donor	Donor	39.13	39.63	86.11
DNABERT-2	Donor	Acceptor	39.13	39.55	87.95
DNABERT-2	Acceptor	Donor	38.91	37.73	98.97
DNABERT-2	Acceptor	Acceptor	38.91	37.79	98.73
DNABERT-2	PPT	Donor	39.04	38.18	85.76
DNABERT-2	PPT	Acceptor	39.04	38.84	88.42
DNABERT-2	BP	Donor	35.90	34.59	99.01
DNABERT-2	BP	Acceptor	35.90	34.85	98.87
SpliceBERT	Donor	Donor	52.38	51.73	92.95
SpliceBERT	Donor	Acceptor	52.38	53.47	96.43
SpliceBERT	Acceptor	Donor	68.27	68.12	93.95
SpliceBERT	Acceptor	Acceptor	68.27	68.25	96.98
SpliceBERT	PPT	Donor	68.37	68.18	94.47
SpliceBERT	PPT	Acceptor	68.37	68.11	96.38
SpliceBERT	BP	Donor	52.20	52.63	93.68
SpliceBERT	BP	Acceptor	52.20	52.15	97.40
HyenaDNA	Donor	Donor	53.66	20.87	18.68
HyenaDNA	Donor	Acceptor	53.66	40.42	27.14
HyenaDNA	Acceptor	Donor	57.51	19.41	16.86
HyenaDNA	Acceptor	Acceptor	57.51	36.60	25.49
HyenaDNA	PPT	Donor	21.42	6.72	7.72
HyenaDNA	PPT	Acceptor	21.42	17.45	26.83
HyenaDNA	BP	Donor	57.05	21.93	17.98
HyenaDNA	BP	Acceptor	57.05	42.96	27.19

G.4 Discussion: Stable Directions and Shared Component Spaces

Geometric stability need not imply unchanged functional dependence. After acceptor-task fine-tuning, DNABERT-2's two small-sample Acceptor bases overlapped by 98.73%, and response retention in the original Acceptor space changed from 38.91% to 37.79%. Under the matched-readout comparison, normal acceptor accuracy increased from 80.70% to 90.87%, whereas erasing the original Acceptor space reduced accuracy by 14.97 percentage points before fine-tuning

and 10.83 after it; the corresponding random decreases were 1.04 and -0.21 percentage points. Together, these exploratory observations suggest that similar component-associated directions can accompany different functional dependence after adaptation. The geometric and functional measurements address different aspects of the representation; geometric preservation alone does not determine how much a task depends on the original space.

Cross-component effects may reflect shared subspaces. After donor-task fine-tuning of SpliceBERT, separately erasing Donor, Acceptor, PPT or BP reduced donor-recognition accuracy by 43.27, 43.27, 43.27 and 42.37 percentage points, respectively. The corresponding random decreases were 1.02, 2.99, 2.92 and 0.96 percentage points. These cross-component effects may reflect shared subspaces among components, such that different component-associated bases cover overlapping task-relevant directions. The individual decreases should therefore not be added as independent component contributions. This interpretation concerns shared model representations; the erasure comparison itself is not a test of biological interactions or of improved biological rule faithfulness.

H External Ranking Evaluation with SPICE and VARS-seq

This section documents whether relation-specific biological rule faithfulness (BRF) rankings recommended models that performed better on external splicing measurements when conventional benchmark rankings recommended otherwise. SPICE and VARS-seq supplied external prediction targets; they were not used to fit new component spaces or to construct additional four-state interaction references. In contrast to the FAS experiment in Appendix F, these two evaluations transferred existing frozen regression heads without training a new readout on the external data.

H.1 Fixed Rankings, Checkpoints and Readouts

For SPICE, the fixed BRF rankings came from the COMPASS K562 joint-subspace directional agreements described in Appendix E. The evaluated relationships were Acceptor–PPT, exon-spanning Acceptor–Donor and Donor–PPT. For VARS-seq, the fixed rankings came from the Schirman yeast relationships Donor–BP, BP–Acceptor, intron-spanning Donor–Acceptor and Donor–BP–Acceptor. The exon-spanning human and intron-spanning yeast relationships were kept distinct. For compact tables, A, D, P and B denote Acceptor, Donor, PPT and BP, respectively. The source BRF values were used directly, without subtracting random-control agreement or combining different relationships into a total score.

Conventional rankings used our measured Matthews correlation coefficient (MCC) on the revised Nucleotide Transformer (NT) splice-site recognition tasks and, in the expanded comparison, GUE splice-site classification. The NT Acceptor task was the reference for Acceptor–PPT and BP–Acceptor; NT Donor was the reference for the other relationships. GUE supplied a second reference for every relationship. The four-model NT comparisons initially used SpliceBERT, NTv2, HyenaDNA and DNABERT-2. Genos and Evo 2 were subsequently added using the same external inputs and frozen regression interfaces, and both the original-four and all-six subsets were tabulated against the expanded NT/GUE results. The six-model extension followed inspection of the four-model results; it was not an originally specified six-model confirmatory analysis. Tables S35, S36 and S37 provide the numerical rankings needed to reproduce every recommendation.

The external forward passes used the pretrained backbones and the same human-PSI or yeast-splicing heads used in the corresponding relation experiment, not the benchmark-fine-tuned classification checkpoints. We averaged final-layer hidden states over all DNA positions, equivalently weighting native tokens by their DNA coverage and excluding special tokens, and then applied the existing standardization and linear coefficients in Equation S15. SPICE used the frozen human-PSI head; VARS-seq used the frozen yeast head. No component replacement, component erasure, new readout training, external-label calibration or prediction clipping was performed. The comparison therefore concerns the external predictive relevance of model recommendations across task interfaces, not the out-of-domain accuracy of the same fine-tuned classification checkpoint. The input construction and initial four-model rankings were fixed before evaluating the new predictions; original predictions were saved before the external performance comparison.

Table S35: Measured benchmark MCC values used by the six-model ranking comparison. These are the classification-adaptation results, not the frozen external-regression readout scores.

Model	NT Donor	NT Acceptor	GUE splice
SpliceBERT	0.9713	0.9613	0.9182
NTv2	0.9621	0.9581	0.8967
HyenaDNA	0.8854	0.8269	0.8584
DNABERT-2	0.8230	0.8061	0.8639
Genos	0.6208	0.6098	0.8666
Evo 2	0.4469	0.2922	0.1918

Table S36: COMPASS K562 relation-specific BRF values used to form the fixed recommendations (percent directional agreement). Random-control scores are not subtracted. A, D, P and B denote Acceptor, Donor, PPT and BP.

Model	A–P	A–D	D–P
SpliceBERT	54.53	55.90	60.40
NTv2	43.50	43.90	60.34
HyenaDNA	48.53	57.90	40.03
DNABERT-2	48.65	34.86	51.61
Genos	54.04	63.52	48.95
Evo 2	54.29	56.48	56.05

Table S37: Schirman yeast relation-specific BRF values used to form the fixed recommendations (percent directional agreement). Random-control scores are not subtracted. A, D, P and B denote Acceptor, Donor, PPT and BP.

Model	D–B	B–A	D–A	D–B–A
SpliceBERT	51.67	48.84	39.68	43.78
NTv2	49.17	48.84	54.76	53.51
HyenaDNA	54.17	46.86	42.86	43.78
DNABERT-2	55.42	60.40	62.70	52.43
Genos	44.58	52.15	52.38	58.38
Evo 2	52.92	48.84	44.44	48.11

H.2 SPICE: Human Reporter Exon Inclusion

Source and measured target. We used the author-released SPICE design library and RNA-derived K562 PSI measurements.[2] Design sequences were matched to the authors' K562 measurement rows by their event identifiers. The reported inclusion fraction was multiplied by 100 to obtain PSI on a percentage scale. We used experimental RNA measurements, not labels predicted by the SPICE design model. Table S38 distinguishes the 7,772 candidate records from the 7,657 records with a K562 measurement and the 5,368 final evaluated events.

Reporter sequence reconstruction. Each model input contained the target exon and 256 bases on each side from the reporter construct, yielding 532–631 bases. We reconstructed the construct using the authors' library design fields, the DC297 backbone and the upstream/downstream reference constants in the RNA-counting script. No genomic, random or N-filled flanks were substituted for the reporter context. The illustrative final plasmid and the RNA-mapping reference differed by one base at the variable-insert boundary; the input definition followed the published RNA-mapping reference used to obtain the measurements. This definition was fixed before model evaluation and represents the specified local transcribed context, not the entire plasmid.

Exclusions and independence. Candidate eligibility incorporated a previously recorded exclusion list of used experimental genes and sequences. We additionally excluded identities present in the current natural-task and relationship inputs, checking gene identifiers, symbols, exact sequences and reverse complements. Because the NT benchmark records lacked gene identifiers, we screened all 66,000 donor/acceptor training, validation and test sequences for exact 61-base matches on both strands against candidate inserts, and excluded every candidate belonging to a matched gene. Among records with K562 measurements, 2,289 further records were excluded, leaving 5,368 unique evaluated sequences from 3,144 genes. Exclusions used input identity and measurement availability, not PSI magnitude or model predictions. The 61-base screen does not establish complete gene-level disjointness from every benchmark, including the later GUE reference, or absence from foundation-model pretraining.

[2] Author repository: https://github.com/chen-dawn/SPICE; the archived study is *Generative Design of Cell Type-Specific RNA Splicing Elements for Programmable Gene Regulation*, https://doi.org/10.1101/2025.11.05.686847.

Table S38: SPICE selection flow. Missing K562 measurements removed 115 candidate records. Previously used identities and benchmark-overlapping genes removed a further 2,289 measured candidates. The 5,368 final sequences were unique.

Stage	Events	Genes
Initial candidate records	7,772	4,030
Candidates with K562 measurement	7,657	3,991
Final retained K562 events	5,368	3,144

External metric. The primary SPICE metric was event-level Spearman correlation between the 5,368 measured PSI values and model predictions:

$$M_m^{\mathrm{SPICE}} = \rho_{\mathrm{S}}\big((y_i)_{i=1}^{5368}, (\widehat{y}_{m,i})_{i=1}^{5368}\big). \tag{S31}$$

Events were not first averaged by gene. Genes instead defined the resampling units for uncertainty estimation below.

H.3 VARS-SEQ: WITHIN-GENE VARIANT SPLICING

Source and reconstruction. We used the author-released VARS-seq variant designs and barcode-level spliced, unspliced and other RNA counts for seven yeast genes.[3] Full wild-type introns were matched exactly to SacCer3 annotations to establish systematic gene identifiers. The authors' prefix/suffix operations and variant coordinates were used to restore each variant into its full intron, including the native tail not covered by the synthesized fragment. The final model input was the complete target intron plus ten real bases on either side, with lengths of 233–531 bases (Table S39). These windows excluded the design and PCR barcodes and were not treated as full transcripts. Previously used yeast training, validation and relationship-test genes and exact input sequences were excluded according to the recorded usage list; all seven retained genes passed these checks. This exclusion does not establish absence from pretraining data.

Barcode aggregation and quality filtering. Within each gene, repeated source-sequence blocks were combined by barcode, with identical barcode/count records counted once. Barcodes assigned to multiple source sequences or carrying conflicting count records were designated ambiguous and removed; no such ambiguity remained in the seven retained genes. The reconstructed input windows were also checked to be unique within each gene. For variant i, let S_i and U_i be the summed spliced and unspliced counts from non-ambiguous barcodes. The measured target was

$$y_i = 100\frac{S_i}{S_i + U_i}. \tag{S32}$$

Counts classified as other were excluded from this denominator; the target was not silently equated with one minus a differently defined intron-retention score. Of 1,324 unique source variants, 57 had no canonical counts, leaving 1,267 usable variants for prediction. The primary analysis required $S_i + U_i \geq 20$ and at least three barcodes with nonzero canonical counts, retaining 1,102 variants. The remaining 165 low-count variants were not used for the primary correlation metric. Table S39 reports the counts for every gene rather than treating all variants as independent biological backgrounds.

[3] Author repository: https://github.com/ramyarangan/VARSseq; the study is *Comprehensive analysis of Saccharomyces cerevisiae intron structures in vivo*, https://doi.org/10.1038/s41594-025-01565-x.

Table S39: VARS-seq reconstruction and quality filtering by gene. Length is the complete intron plus 10 bases on either side. Usable variants have nonzero canonical RNA counts; primary variants additionally have at least 20 canonical counts and three informative barcodes.

Gene	Systematic ID	Length (nt)	Unique variants	Usable	Primary
QCR9	YGR183C	233	180	171	156
RPL28	YGL103W	531	192	189	182
RPL36B	YPL249C-A	258	192	191	190
RPL7A	YGL076C	479	192	190	186
RPS14B	YJL191W	428	184	162	102
RPS9A	YPL081W	521	192	174	115
RPS9B	YBR189W	433	192	190	171
Total	7 genes	233–531	1,324	1,267	1,102

External metric. The primary metric was the equal-weight mean of within-gene Spearman correlations:

$$M_m^{\mathrm{VARS}} = \frac{1}{7}\sum_{g=1}^{7} \rho_{\mathrm{S}}\big((y_i)_{i\in\mathcal{I}_g}, (\widehat{y}_{m,i})_{i\in\mathcal{I}_g}\big), \tag{S33}$$

where $\mathcal{I}_g$ contains primary-analysis variants for gene g. This prevents differences between gene baselines from substituting for within-gene variant discrimination. The pooled correlation across all variants is a different quantity and was not substituted for this primary metric. Undefined correlations would not be set to zero; the reported primary correlations were defined for every retained gene and model.

H.4 Discordant Recommendations and Uncertainty

For a given relation, benchmark reference and model subset, we considered every unordered model pair. Pairs tied in the source BRF ranking within 10^{-10} were not forced into an ordering. Among the remaining pairs, only those with opposite BRF and benchmark recommendations entered the discordance comparison; the benchmark values in Table S35 contained no relevant ties. For discordant pair p, with recommended models m_p^{BRF} and m_p^{bench}, the external difference was

$$d_p = M_{m_p^{\mathrm{BRF}}} - M_{m_p^{\mathrm{bench}}}. \tag{S34}$$

A difference greater than 10^{-12} counted as a BRF win, less than -10^{-12} as a benchmark win, and otherwise as a tie. Win fractions, when reported, counted ties as one-half and retained them in the denominator. No discordant pair meant an undefined win fraction, not a zero success rate. These are comparisons of dataset-level performance, not counts of individual variants whose effects were predicted correctly. All discordant pairs were retained; confidence intervals were not used as a gate for declaring a point-estimate winner.

The original four-model analyses used 5,000 paired gene-bootstrap draws with seed 20260926. For SPICE, each draw sampled 3,144 genes with replacement, retained all events of each sampled gene and recomputed event-level Spearman. For VARS-seq, each draw sampled seven complete genes with replacement and averaged their within-gene correlations. The same draw was used across models, so external differences and their relation-wise means preserved pairing. Pointwise 95% intervals used the 2.5th and 97.5th percentiles. For the mean difference across discordant pairs, SPICE additionally retained the original three-relation Bonferroni percentile intervals (98.33%). The archived VARS-seq intervals retained their original eight-comparison adjustment (99.375%): four VARS-seq relations and four parallel natural-yeast comparisons using GSE94404 belonged to that analysis family. That correction was not redefined after restricting the present tables to VARS-seq.

The added Genos/Evo 2 predictions and GUE-reference comparisons were summarized by point estimates. We did not transfer the original four-model confidence intervals to the six-model or GUE extensions. Source BRF and benchmark rankings were treated as fixed: the external intervals do not include source-ranking re-estimation or training-seed uncertainty. Model pairs share models and external observations; neither their win counts nor multiple relations evaluated on one dataset

are independent biological replications. The yeast evaluation also compares a yeast-derived BRF ranking with human splice-site benchmark rankings, so it is a cross-species recommendation test.

H.5 Absolute Performance and Complete Recommendation Results

Table S40 reports the six models' absolute performance on both external tasks. The different aggregation rules in Equations S31 and S33 are retained; the two columns are not pooled into a common model score. Table S41 gives all relation/reference-specific win, loss and tie counts, separately for the original four and all six models. Table S42 retains the original four-model mean differences and their intervals against NT. Tables S43 and S44 enumerate every six-model discordant recommendation; entries shared by NT and GUE are displayed once as Both, without removing them from either reference's count.

Table S40: Absolute external performance for all six models. SPICE uses event-level Spearman on 5,368 events; VARS-seq uses the mean of within-gene Spearman correlations on 1,102 variants in seven genes. Intervals are the archived original-four-model 95% gene-bootstrap intervals. Dashes for Genos and Evo 2 mean not estimated in the extension, not zero uncertainty.

Model	SPICE ρ	95% interval	VARS-seq mean ρ	95% interval
SpliceBERT	0.1619	[0.1346, 0.1886]	−0.0321	[−0.1869, 0.1363]
NTv2	0.0157	[−0.0109, 0.0432]	−0.0119	[−0.1053, 0.1020]
HyenaDNA	−0.0746	[−0.1019, −0.0460]	−0.0861	[−0.2326, 0.0337]
DNABERT-2	0.1343	[0.1072, 0.1623]	0.0221	[−0.0285, 0.0724]
Genos	−0.0323	—	−0.0724	—
Evo 2	0.0843	—	0.0043	—

Table S41: All relationship-specific ranking disagreements. W/L/T are BRF wins, benchmark wins and ties, determined from external point estimates; N is the number of discordant model pairs. The four-model column is a subset of the six-model comparison, not an additional replication. A zero N means no disagreement, not a loss.

External data	Relation	Reference	Four: W/L/T	N	Six: W/L/T	N
SPICE	A–P	NT	2 / 1 / 0	3	6 / 4 / 0	10
SPICE	A–P	GUE	1 / 1 / 0	2	4 / 3 / 0	7
SPICE	A–D	NT	0 / 2 / 0	2	2 / 7 / 0	9
SPICE	A–D	GUE	0 / 3 / 0	3	1 / 7 / 0	8
SPICE	D–P	NT	1 / 0 / 0	1	4 / 1 / 0	5
SPICE	D–P	GUE	0 / 0 / 0	0	3 / 1 / 0	4
VARS-seq	D–B	NT	3 / 2 / 0	5	6 / 2 / 0	8
VARS-seq	D–B	GUE	2 / 2 / 0	4	6 / 3 / 0	9
VARS-seq	B–A	NT	3 / 0 / 0	3	5 / 2 / 0	7
VARS-seq	B–A	GUE	2 / 0 / 0	2	4 / 2 / 0	6
VARS-seq	D–A	NT	4 / 1 / 0	5	7 / 2 / 0	9
VARS-seq	D–A	GUE	3 / 1 / 0	4	6 / 2 / 0	8
VARS-seq	D–B–A	NT	3 / 0 / 0	3	6 / 3 / 0	9
VARS-seq	D–B–A	GUE	2 / 0 / 0	2	4 / 2 / 0	6

Table S42: Archived four-model comparisons against NT. The advantage is BRF-recommended minus benchmark-recommended external correlation, averaged over discordant pairs. Adjusted intervals are 98.33% for SPICE (three relations) and 99.375% for VARS-seq (the original eight-comparison family). These intervals do not apply to the six-model or GUE extensions.

Data	Relation	N	Mean advantage	95% interval	Adjusted interval
SPICE	A–P	3	0.0790	[0.0580, 0.0995]	[0.0533, 0.1046]
SPICE	A–D	2	−0.1635	[−0.1893, −0.1367]	[−0.1952, −0.1312]
SPICE	D–P	1	0.2089	[0.1804, 0.2392]	[0.1745, 0.2451]
VARS-seq	D–B	5	0.0136	[−0.0538, 0.0710]	[−0.0881, 0.0892]
VARS-seq	B–A	3	0.0654	[−0.0379, 0.1591]	[−0.0772, 0.1840]
VARS-seq	D–A	5	0.0325	[−0.0872, 0.1495]	[−0.1318, 0.1962]
VARS-seq	D–B–A	3	0.0609	[−0.0936, 0.2152]	[−0.1536, 0.2874]

The results were relationship- and task-dependent. For example, on SPICE the six-model Donor–PPT ranking won four of five disagreements with NT and three of four with GUE, whereas the exon-spanning Acceptor–Donor ranking won two of nine and one of eight, respectively. On VARS-seq, each of the four relation rankings won a majority of its six-model disagreements under both references, but the absolute within-gene correlations were weak, and the original four-model mean-difference intervals crossed zero. These observations concern the relative usefulness of fixed recommendations on the stated external tasks; they do not establish high-accuracy reporter prediction or new interaction ground truth for every source relationship.

Table S43: SPICE: every six-model discordant recommendation. Both denotes identical recommendations under NT and GUE, printed once but counted separately in their respective reference columns. The advantage uses the dataset-specific correlation metric in Table S40; it is not a percentage-point difference.

Relation	Reference	BRF recommends	Benchmark recommends	Advantage	Winner
A–P	Both	HyenaDNA	NTv2	−0.0904	Benchmark
A–P	Both	DNABERT-2	NTv2	0.1186	BRF
A–P	NT	DNABERT-2	HyenaDNA	0.2089	BRF
A–P	Both	Genos	NTv2	−0.0480	Benchmark
A–P	NT	Genos	HyenaDNA	0.0424	BRF
A–P	NT	Genos	DNABERT-2	−0.1665	Benchmark
A–P	Both	Evo 2	NTv2	0.0685	BRF
A–P	Both	Evo 2	HyenaDNA	0.1589	BRF
A–P	Both	Evo 2	DNABERT-2	−0.0500	Benchmark
A–P	Both	Evo 2	Genos	0.1165	BRF
A–D	Both	HyenaDNA	SpliceBERT	−0.2366	Benchmark
A–D	Both	HyenaDNA	NTv2	−0.0904	Benchmark
A–D	GUE	HyenaDNA	DNABERT-2	−0.2089	Benchmark
A–D	Both	Genos	SpliceBERT	−0.1942	Benchmark
A–D	Both	Genos	NTv2	−0.0480	Benchmark
A–D	NT	Genos	HyenaDNA	0.0424	BRF
A–D	NT	Genos	DNABERT-2	−0.1665	Benchmark
A–D	Both	Evo 2	SpliceBERT	−0.0777	Benchmark
A–D	Both	Evo 2	NTv2	0.0685	BRF
A–D	Both	Evo 2	DNABERT-2	−0.0500	Benchmark
D–P	NT	DNABERT-2	HyenaDNA	0.2089	BRF
D–P	GUE	DNABERT-2	Genos	0.1665	BRF
D–P	NT	Genos	HyenaDNA	0.0424	BRF
D–P	Both	Evo 2	HyenaDNA	0.1589	BRF
D–P	Both	Evo 2	DNABERT-2	−0.0500	Benchmark
D–P	Both	Evo 2	Genos	0.1165	BRF

Table S44: VARS-seq: every six-model discordant recommendation. Both denotes identical recommendations under NT and GUE, printed once but counted separately in their respective reference columns. The advantage uses the dataset-specific correlation metric in Table S40; it is not a percentage-point difference.

Relation	Reference	BRF recommends	Benchmark recommends	Advantage	Winner
D–B	Both	HyenaDNA	SpliceBERT	−0.0541	Benchmark
D–B	Both	HyenaDNA	NTv2	−0.0743	Benchmark
D–B	GUE	HyenaDNA	Genos	−0.0138	Benchmark
D–B	Both	DNABERT-2	SpliceBERT	0.0541	BRF
D–B	Both	DNABERT-2	NTv2	0.0339	BRF
D–B	NT	DNABERT-2	HyenaDNA	0.1082	BRF
D–B	GUE	DNABERT-2	Genos	0.0944	BRF
D–B	Both	Evo 2	SpliceBERT	0.0363	BRF
D–B	Both	Evo 2	NTv2	0.0161	BRF
D–B	Both	Evo 2	Genos	0.0766	BRF
B–A	Both	DNABERT-2	SpliceBERT	0.0541	BRF
B–A	Both	DNABERT-2	NTv2	0.0339	BRF
B–A	NT	DNABERT-2	HyenaDNA	0.1082	BRF
B–A	GUE	DNABERT-2	Genos	0.0944	BRF
B–A	Both	Genos	SpliceBERT	−0.0403	Benchmark
B–A	Both	Genos	NTv2	−0.0605	Benchmark
B–A	NT	Genos	HyenaDNA	0.0138	BRF
B–A	Both	Evo 2	HyenaDNA	0.0904	BRF
D–A	Both	NTv2	SpliceBERT	0.0202	BRF
D–A	Both	HyenaDNA	SpliceBERT	−0.0541	Benchmark
D–A	Both	DNABERT-2	SpliceBERT	0.0541	BRF
D–A	Both	DNABERT-2	NTv2	0.0339	BRF
D–A	NT	DNABERT-2	HyenaDNA	0.1082	BRF
D–A	GUE	DNABERT-2	Genos	0.0944	BRF
D–A	Both	Genos	SpliceBERT	−0.0403	Benchmark
D–A	NT	Genos	HyenaDNA	0.0138	BRF
D–A	Both	Evo 2	SpliceBERT	0.0363	BRF
D–A	Both	Evo 2	HyenaDNA	0.0904	BRF
D–B–A	Both	NTv2	SpliceBERT	0.0202	BRF
D–B–A	Both	DNABERT-2	SpliceBERT	0.0541	BRF
D–B–A	NT	DNABERT-2	HyenaDNA	0.1082	BRF
D–B–A	Both	Genos	SpliceBERT	−0.0403	Benchmark
D–B–A	Both	Genos	NTv2	−0.0605	Benchmark
D–B–A	NT	Genos	HyenaDNA	0.0138	BRF
D–B–A	NT	Genos	DNABERT-2	−0.0944	Benchmark
D–B–A	Both	Evo 2	SpliceBERT	0.0363	BRF
D–B–A	Both	Evo 2	HyenaDNA	0.0904	BRF