

Supplementary Information

Supplementary Methods: Evaluation metrics

Let $\mathcal{O}$ denote the set of test guide (oligo) sequences. For a given guide $o \in \mathcal{O}$, let $\mathcal{I}_o = \{1, \dots, N_o\}$ index the indel events enumerated by the `indelgentarget` universe plus, where applicable, the outer-merge events carried over from observed data (e.g. TREX2 one-sided deletions). For event $i \in \mathcal{I}_o$, we write $p_{o,i} \geq 0$ for the observed (replicate 1) frequency and $\hat{p}_{o,i} \geq 0$ for the predicted probability, both normalized so that $\sum_i p_{o,i} = \sum_i \hat{p}_{o,i} = 1$. We write $\ell_i \in \mathbb{Z}$ for the signed indel length of event i (positive for insertions, negative for deletions; the reading-frame index uses $|\ell_i| \bmod 3$, but because ℓ_i and $-\ell_i$ agree modulo 3 we use $\ell_i \bmod 3$ throughout). Unless stated otherwise, a metric $m(\hat{p}_o, p_o)$ defined for a single guide is aggregated across the test set by $\bar{m} = |\mathcal{O}|^{-1} \sum_{o \in \mathcal{O}} m(\hat{p}_o, p_o)$.

1. **Kullback–Leibler divergence of the full indel-frequency distribution** ($D_{\text{KL}}^{\text{full}}: [0, \infty)$; lower is better). For each guide,

$$D_{\text{KL}}^{\text{full}}(\hat{p}_o \| p_o) = \sum_{i \in \mathcal{I}_o} (\hat{p}_{o,i} + \varepsilon) \log \frac{\hat{p}_{o,i} + \varepsilon}{p_{o,i} + \varepsilon}, \quad \varepsilon = 10^{-8}, \quad (\text{S1})$$

implemented with `inDecay.analysis_fn.kld_fn`. The additive ε prevents $\log 0$ for events that are predicted or observed at exactly zero.

Biological meaning: how well the model reproduces the relative abundances of *all* identifier-level outcomes (e.g. distinguishing a -3 bp MH-deletion from a -3 bp non-MH deletion at the same cut site).

2. **Kullback–Leibler divergence of the indel-length distribution** ($D_{\text{KL}}^{\text{len}}$, reported as `KLD_IDL`). Events are first collapsed to an indel-size representation $L_o \in \mathbb{R}^{41}$ spanning deletions of size 1–38 bp and insertions of size 1–3 bp:

$$L_{o,k} = \sum_{i \in \mathcal{I}_o: \ell_i = k} p_{o,i}, \quad \hat{L}_{o,k} = \sum_{i \in \mathcal{I}_o: \ell_i = k} \hat{p}_{o,i}, \quad (\text{S2})$$

and the divergence is then computed on the length-binned vectors:

$$D_{\text{KL}}^{\text{len}}(\hat{L}_o \| L_o) = \sum_k (\hat{L}_{o,k} + \varepsilon) \log \frac{\hat{L}_{o,k} + \varepsilon}{L_{o,k} + \varepsilon}. \quad (\text{S3})$$

Biological meaning: how well the model reproduces the spectrum of indel sizes irrespective of which microhomology variant produced them; this is the quantity most directly comparable to published indel-size distributions. *Range and direction:* $[0, \infty)$; lower is better.

3. **Kendall's τ on the indel-length distribution** (`Kendall_tau_IDL`: $[-1, +1]$; higher is better). For each guide, the Kendall rank correlation coefficient (the “b” variant, which corrects for ties) between $\hat{L}_o$ and L_o is computed using `scipy.stats.kendalltau`:

$$\tau_b(\hat{L}_o, L_o) = \frac{C - D}{\sqrt{((\binom{n}{2} - T_{\hat{L}})(\binom{n}{2} - T_L))}}, \quad (\text{S4})$$

where C and D are the numbers of concordant and discordant length-bin pairs across the $n = 41$ bins, and $T_{\hat{L}}$ and T_L are the tie-correction terms for $\hat{L}_o$ and L_o , respectively.

Biological meaning: whether the model gets the relative *ordering* of indel sizes right (e.g. that -3 bp is more frequent than -7 bp) even when absolute probabilities are miscalibrated.

4. **Top- K event recall** ($K \in \{1, 5, 10\}$; range: $[0, K]$; higher is better). For each guide, let $\text{TopK}(\hat{p}_o) \subset \mathcal{I}_o$ denote the set of K indel events with the largest predicted probability, and analogously $\text{TopK}(p_o)$ for the observed

frequencies. The per-guide overlap is

$$\text{TopK-recall}(\hat{p}_o, p_o) = |\text{TopK}(\hat{p}_o) \cap \text{TopK}(p_o)|, \quad (\text{S5})$$

implemented in `analysis_fn.py.top_k_overlap` and averaged over guides.

Biological meaning: whether the model correctly identifies the dominant editing outcome(s) that an experimental user would observe by amplicon sequencing.

5. **Major events recall** (range: $[0, 1]$; higher is better). The major events $M_y(\beta)$, with $\beta < 1$, are defined as those with a true ratio larger than the threshold β . For each oligo,

$$\text{Major events recall}(\beta) = \frac{1}{n} \sum_{i=1}^n \frac{\text{Count}[M_y(\beta) \cap M_p(\beta)]}{\text{Count}[M_y(\beta)]}, \quad (\text{S6})$$

where $M_p(\beta)$ are the events whose predicted ratio is larger than the threshold β , and n is the number of available gRNAs. This metric is omitted for a gRNA if $\text{Count}(M_y(\beta)) = 0$.

Biological meaning: how well the prediction recalls prominent events for samples with skewed distributions, such as those with only one dominant event.

6. **Frameshift ratio** R^2 (R2 of Frameshift ratio: $[0, 1]$; higher is better). For each guide, the observed and predicted frameshift ratios are the total probability mass on events whose signed indel length is not a multiple of three:

$$f(p_o) = \sum_{i \in \mathcal{I}_o} p_{o,i} \cdot \mathbf{1}[\ell_i \bmod 3 \neq 0], \quad f(\hat{p}_o) = \sum_{i \in \mathcal{I}_o} \hat{p}_{o,i} \cdot \mathbf{1}[\ell_i \bmod 3 \neq 0], \quad (\text{S7})$$

(`analysis_fn.forecast_frameshift`). The reported metric is the squared Pearson correlation between the vectors of per-guide frameshift ratios across the full test set:

$$R_{\text{fs}}^2 = [\text{corr}(\{f(\hat{p}_o)\}_{o \in \mathcal{O}}, \{f(p_o)\}_{o \in \mathcal{O}})]^2. \quad (\text{S8})$$

Biological meaning: how well the model ranks guides by their probability of producing a functional knockout via a protein-coding frameshift.

7. **Per-reading-frame** R^2 (`frame0_r2`, `frame1_r2`, `frame2_r2`; $[0, 1]$; higher is better). For each reading frame $k \in \{0, 1, 2\}$ and each guide, the observed and predicted in-frame- k ratios are

$$f_k(p_o) = \sum_{i \in \mathcal{I}_o} p_{o,i} \cdot \mathbf{1}[\ell_i \bmod 3 = k], \quad f_k(\hat{p}_o) = \sum_{i \in \mathcal{I}_o} \hat{p}_{o,i} \cdot \mathbf{1}[\ell_i \bmod 3 = k], \quad (\text{S9})$$

(`analysis_fn.assessment_recipe_frame_breakdown`). Frame 0 is in-frame (no net frameshift, including 0 bp wild-type); frames 1 and 2 correspond to +1 and +2 net frameshifts, respectively. The reported coefficient is

$$R_k^2 = [\text{corr}(\{f_k(\hat{p}_o)\}_{o \in \mathcal{O}}, \{f_k(p_o)\}_{o \in \mathcal{O}})]^2, \quad k \in \{0, 1, 2\}. \quad (\text{S10})$$

By construction, $f_0(p_o) + f_1(p_o) + f_2(p_o) = 1$ and $f_1(p_o) + f_2(p_o) = f(p_o)$, so the per-frame decomposition refines the aggregate frameshift R^2 of metric (6).

Biological meaning: the relative contributions of +1- and +2-shift repair products are the quantities of direct interest for knockout-pipeline design, because they determine which protein-level consequence (premature stop, altered C-terminus, etc.) will dominate at a given guide.

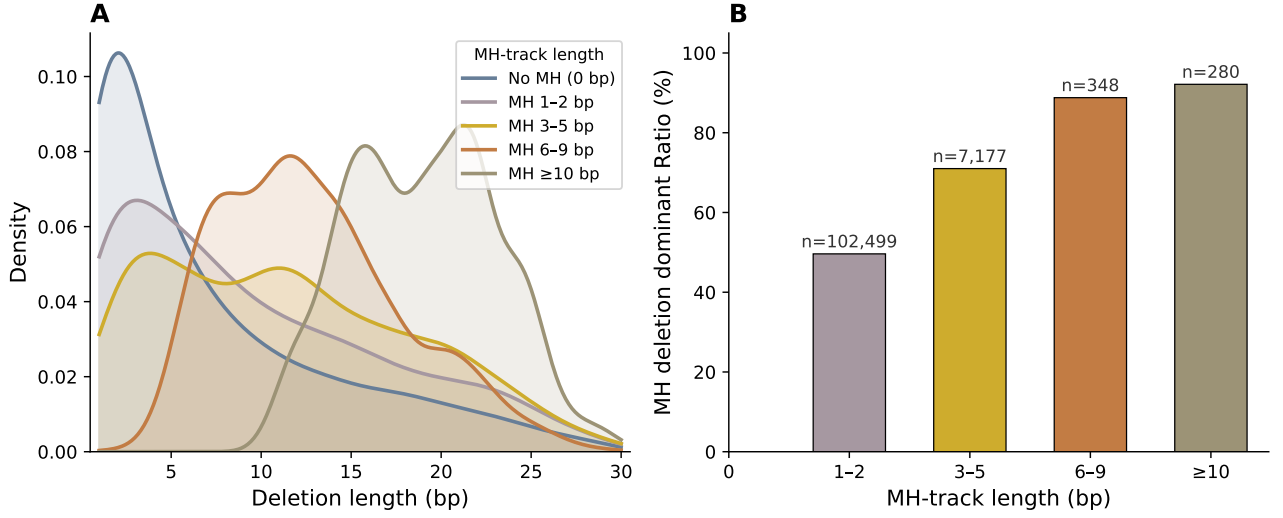


Fig S1. Frequency decay of deletion events by deletion length and microhomology length.

(A) Probability density of deletion events as a function of deletion length, stratified by MH-track length (color). Deletions without microhomology (blue) show the steepest frequency decay with increasing length, whereas progressively longer MH tracks attenuate this decay and shift probability mass toward longer deletions. (B) Proportion of deletion events at each indel size for which MH-mediated deletions exceeded non-MH deletions in frequency.

8. **Mean absolute Δ frameshift ratio** ($[0,1]$; lower is better) (Mean Abs Delta Frameshift). Using the per-guide frameshift ratios from metric (6),

$$\text{MA}\Delta f = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |f(\hat{p}_o) - f(p_o)|. \quad (\text{S11})$$

Unlike R_{fs}^2 , which is invariant to affine rescaling of the predictions, $\text{MA}\Delta f$ penalises absolute-scale miscalibration and is therefore reported alongside the R^2 to expose systematic over- or under-prediction of frameshift rates.

Biological meaning: the typical absolute error, in units of probability, with which the model predicts the frameshift rate of an arbitrary guide.

Throughout, Pearson correlations use `scipy.stats.pearsonr`, Kendall rank correlations use `scipy.stats.kendalltau`, and KL divergences use `torchmetrics.KLDivergence` with the $\varepsilon = 10^{-8}$ smoothing described above. Source code for all metrics is provided in `inDecay.analysis_fn.py` at the GitHub repository.

Supplementary Figures

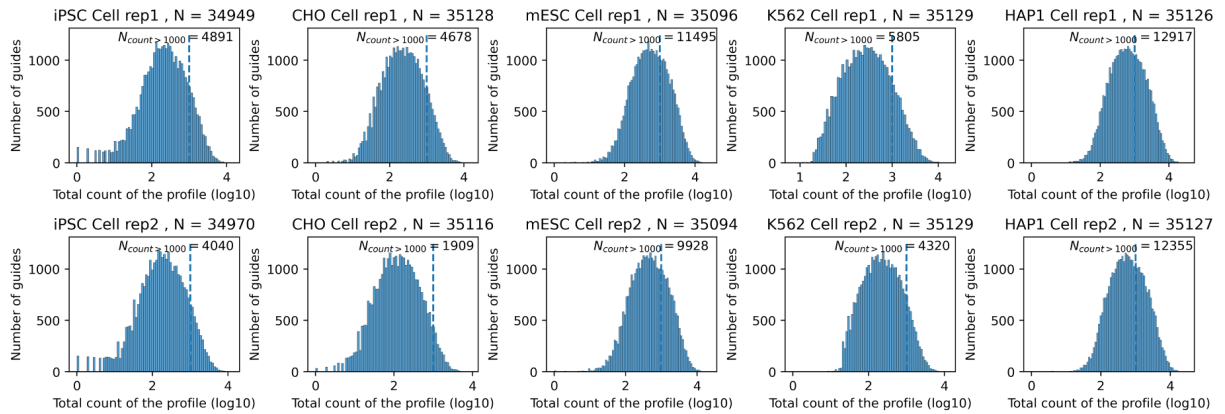


Fig S2. Total read count distribution of FORECastT's multi-cell-line dataset.

This figure shows the distribution of total read counts in FORECastT's multi-cell-line dataset. Each panel displays the density of target sequences based on their total number of edited reads. The x-axis uses a log10 scale, with a dashed line marking the 1,000 count position. Each cell line includes two repeats, with the total sample size mentioned in the subtitle. The number of samples exceeding 1,000 reads is also noted for each cell line and experimental repeat.

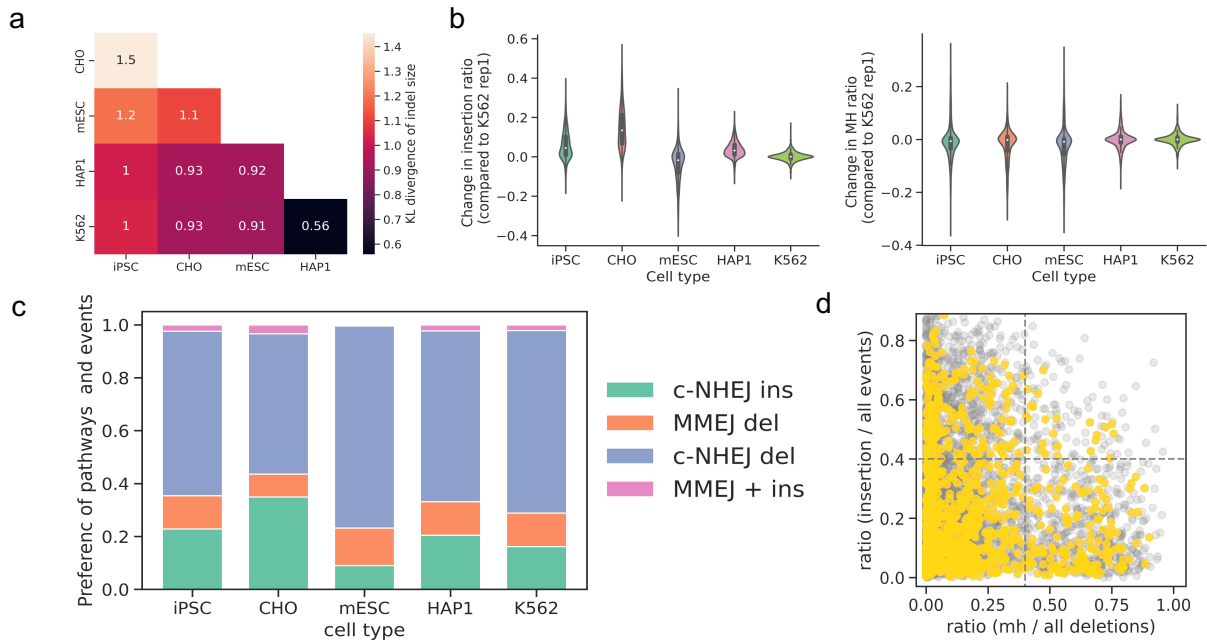


Fig S3. Cell type variability of editing profiles.

(a) The pairwise KL divergence of indel-size distribution. (b) Violin plot showing the changes in insertion ratio (left) and MH strength (right). The ratio changes were calculated relative to replicate 1 of K562 cells between matched guides (replicate 2 to replicate 1 for K562). (c) The proportion of samples dominated by each repair pathway. In the stacked bar plot, color showed different repair pathways. Guides with both MH ratio and insertion ratio higher than 0.4 are noted with "MMEJ + ins". (d) Scatter plot showing MH ratio (x-axis) and the ratio ratio (y-axis) of each guide from K562 cells. The gray dots are background generated by pooling samples from all cell lines.

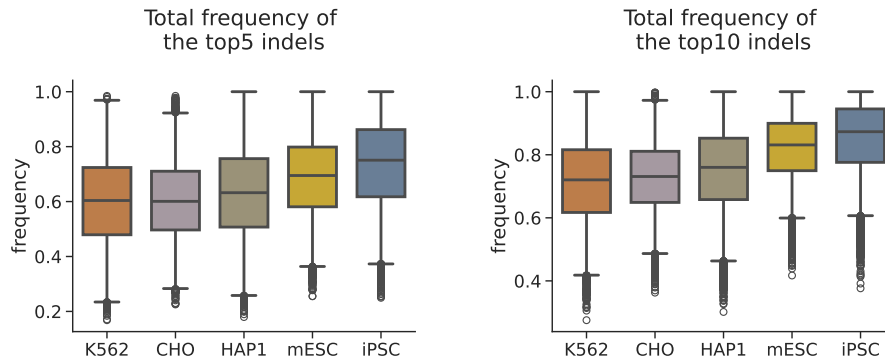


Fig S4. The frequency of the most frequent indels

Box plots showing the total frequency of the top 5 and top 10 indels across different cell lines (K562 $n=10,174$, CHO $n=8,575$, HAP1 $n=17,225$, mESC $n=15,532$, iPSC $n=7,831$). The left plot represents the top 5 indels, and the right plot represents the top 10 indels. Each box plot uses a whisker length of 1.

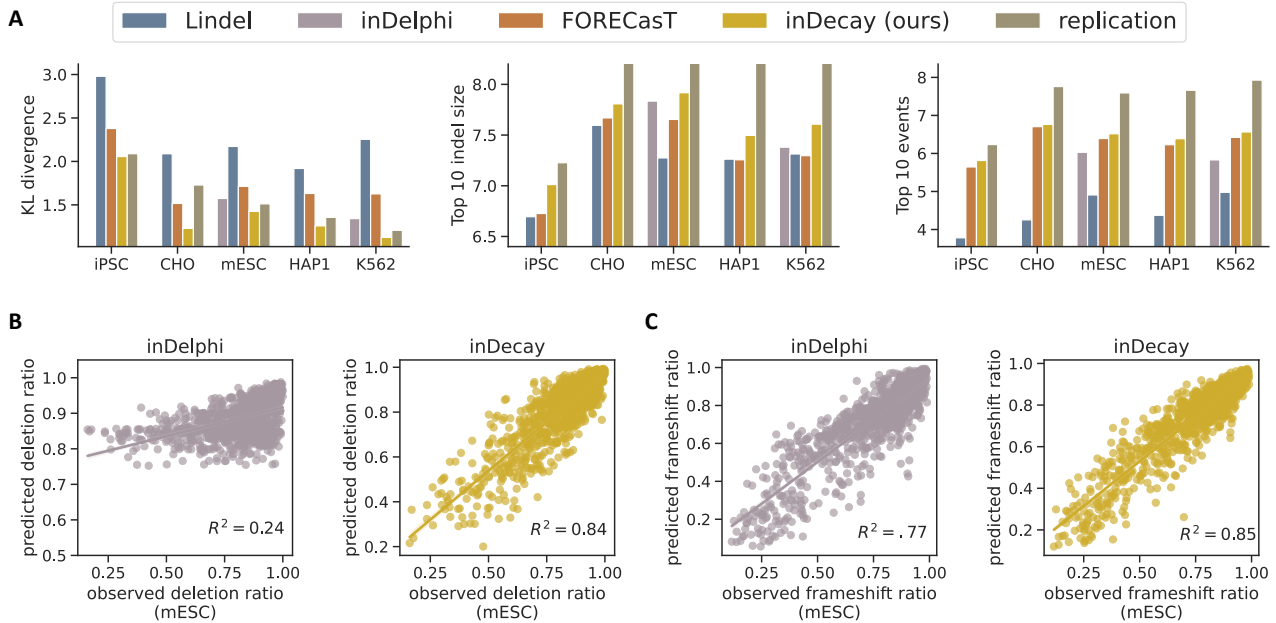
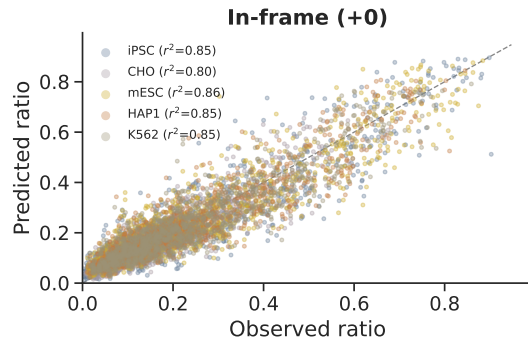


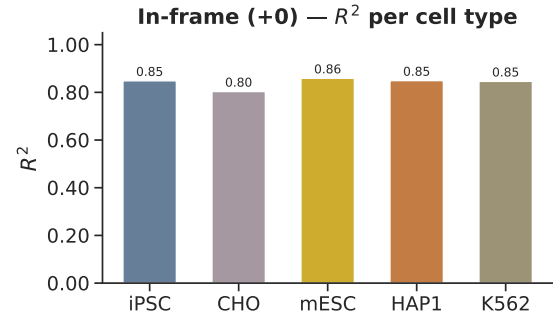
Fig S5. Comparison of inDecay with other method and ratios

(A) Three additional evaluation metrics for benchmarking different repair outcome prediction methods. The left panel shows the KL divergence between the observed and predicted outcome probabilities. The middle and right panels show the recall of the top 10 indel sizes and top 10 events. These methods are trained, evaluated, and visualized in the same way as in Figure 2. (B) The deletion ratio for each test set sample is shown, compared with predicted ratios from inDelphi (gray) and our inDecay (yellow). The R^2 score is calculated and also presented as a bar plot in Figure 2. (C) Scatter plots illustrate the frameshift ratio of each sample, using the same color scheme as in (B).

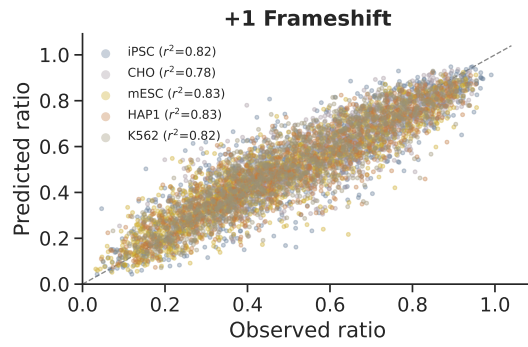
A



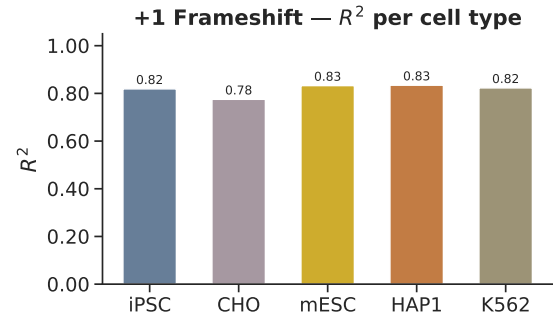
B



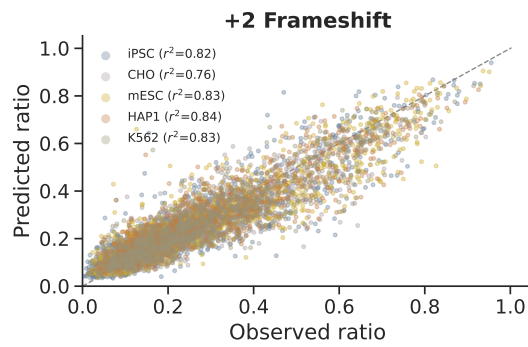
C



D



E



F

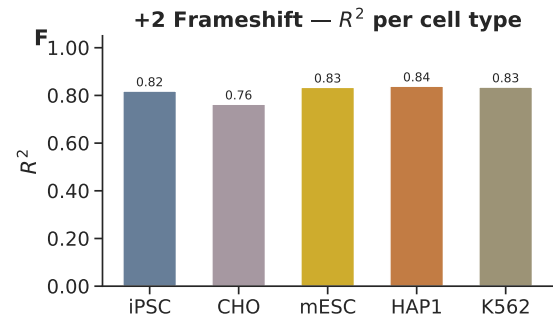


Fig S6. Reading-frame-specific evaluation of inDecay performance across somatic cell lines.

(A) Correlation between predicted and observed in-frame ratios. Each point represents a unique target sequence ($n = 1,133$ sample), color-coded by cell line. (B) Accuracy of in-frame ratio predictions per cell line, measured by the coefficient of determination (R^2). (C–D) Scatter plots and R^2 values for +1 frameshift predictions. (E–F) Scatter plots and R^2 values for +2 frameshift predictions. All scatter plots display predicted values on the y-axis against observed experimental values on the x-axis.

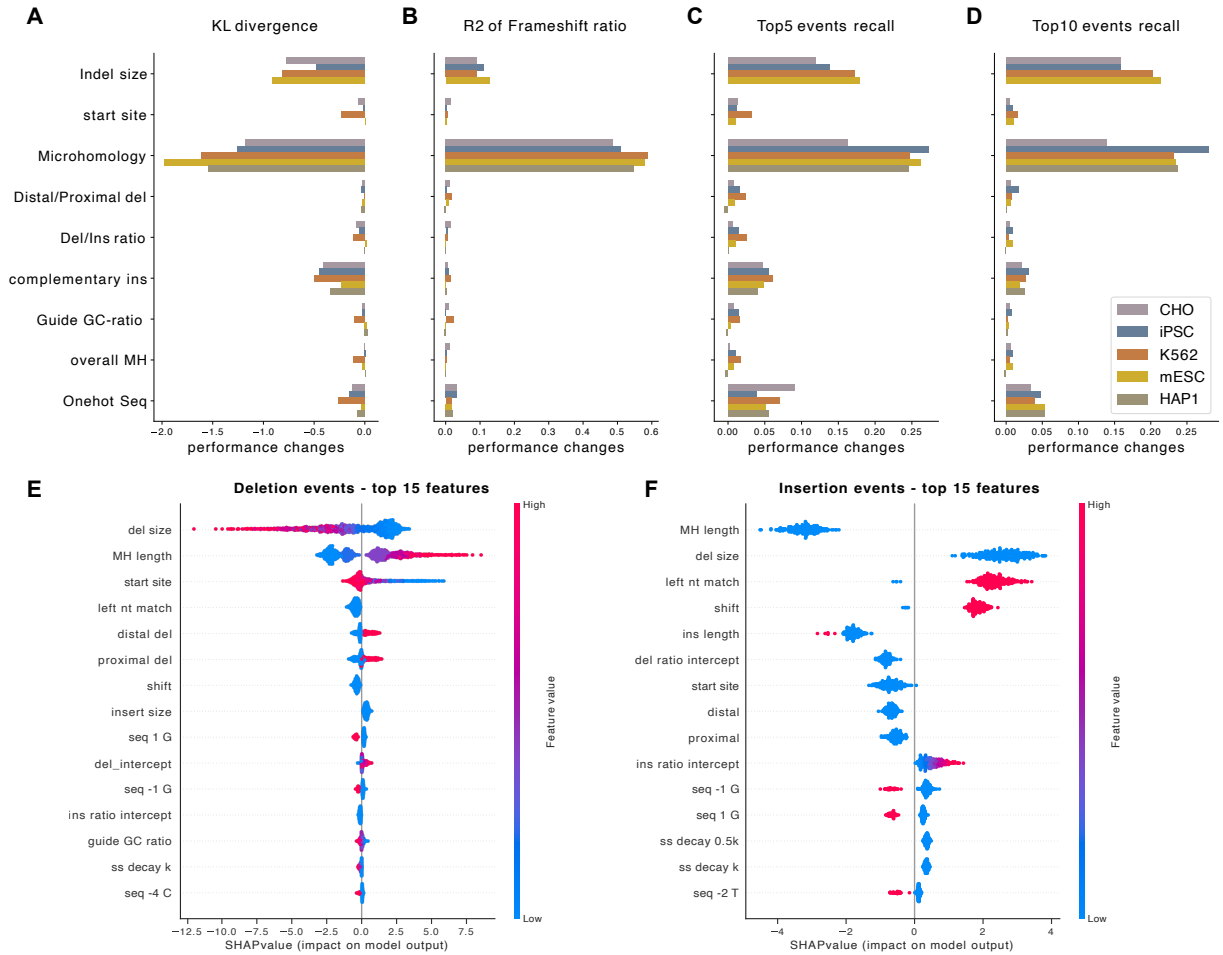


Fig S7. Feature importance of the ablation study

(A-D) Performance sensitivity following feature group ablation. In our ablation study, we retrained the inDecay model, selectively masking certain input dimensions from the same feature set. The Y-axis indicates the feature set being masked, while the X-axis shows the performance difference, which was obtained by the ratio of difference (the original model minus the masked model) over the original value. The color of the bar denotes the cell-line of the training data. Performance sensitivity following feature group ablation. (E-F) SHAP beeswarm plots displaying the top 15 most influential features for deletion (E) and insertion (F) events. Each point represents an individual indel; the x-axis indicates the SHAP value (impact on model output), while the color gradient represents the feature magnitude (e.g., low to high). SHAP values were calculated using the GradientExplainer from the shap package.

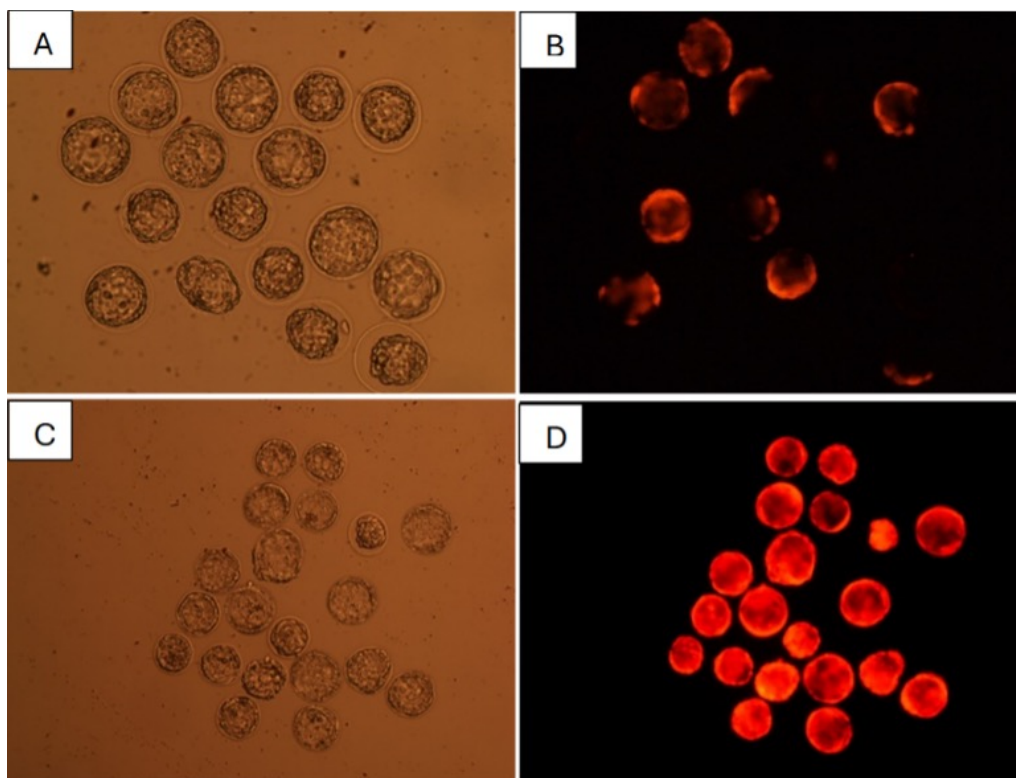


Fig S8. Microscopic images of for td-Tomato gene KO (WT oocytes and td-Tomato positive sperm). (A): light field (edited); (B): red fluorescence (edited); (C): light field (non-edited); (D): red fluorescence (non-edited).

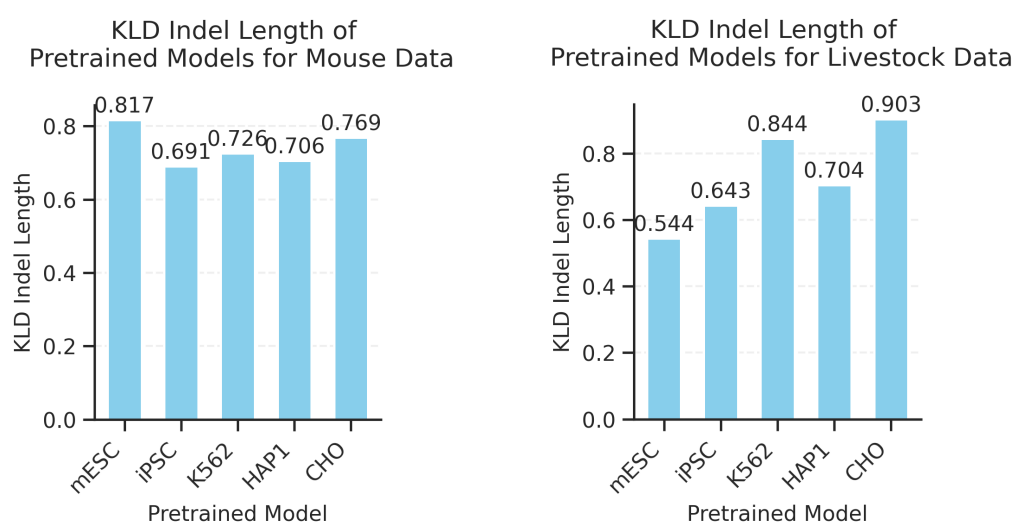


Fig S9. KLD of indel length for different pretrained model

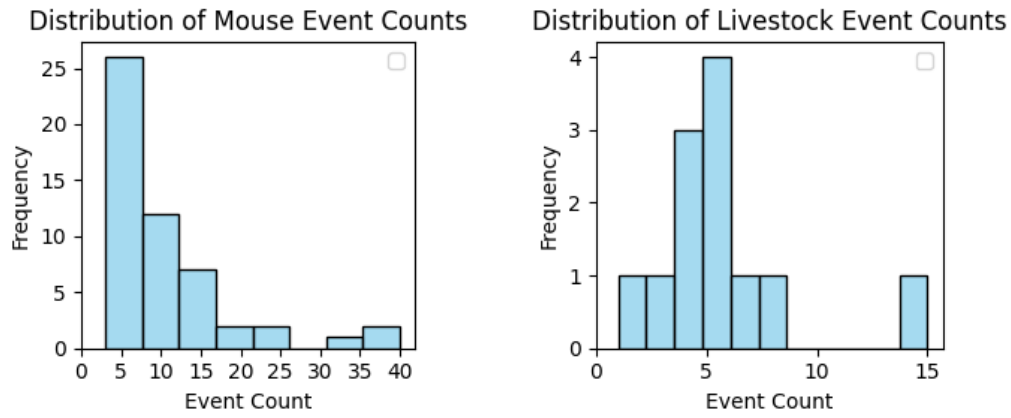


Fig S10. Total event count distribution of in-house dataset.

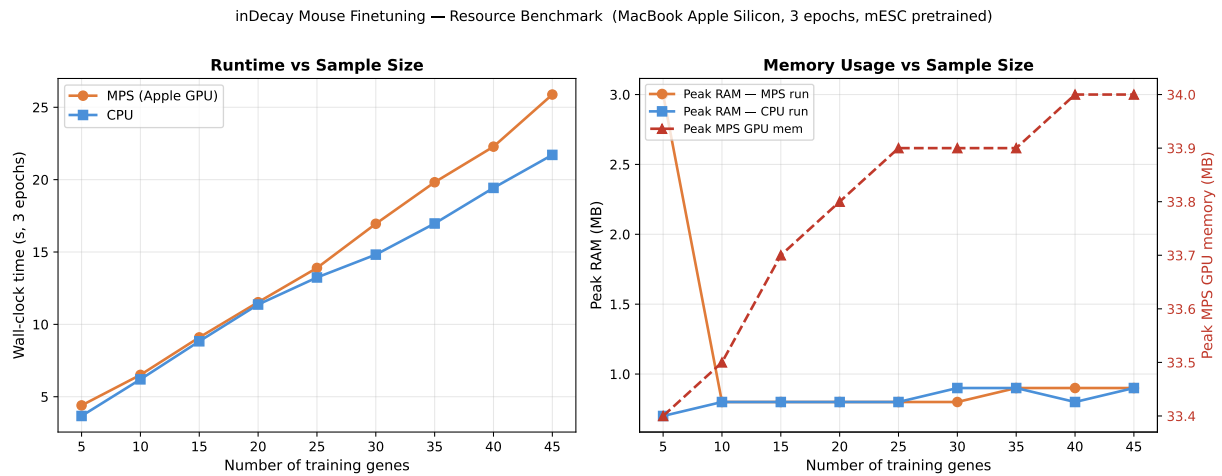


Fig S11. inDecay finetuning resource benchmark. Wall-clock runtime (left) and peak memory usage (right) of inDecay mouse finetuning as a function of the number of training genes, measured on a MacBook with Apple Silicon (3 epochs, mESC-pretrained model). Runtime is shown for CPU and Apple GPU (MPS) backends; memory panel shows peak RAM for both runs and peak MPS GPU memory. The lightweight, parameter-efficient design keeps both runtime and memory low across sample sizes.