BE-CORR: LEARNING TO CORRECT SOURCE-LIBRARY OUTCOMES FOR ENDOGENOUS BASE-EDITING PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

DNA base editors enable targeted single-nucleotide changes and are promising tools for functional genomics and therapeutic genome engineering. Machine learning has become an effective approach for predicting base-editing outcomes, helping prioritize editor–guide designs early in development. Many predictors are trained on integrated or synthetic target-library assays (*source libraries*), where many targets are measured in controlled, non-native contexts. However, practical editing occurs at endogenous genomic loci, where chromatin and regulatory state can substantially shift outcomes relative to source libraries. Existing approaches typically target either source-library prediction or direct endogenous prediction: the former does not directly transfer to native locus behavior, while the latter is constrained by scarce endogenous training data. We therefore reformulate endogenous outcome prediction as *source-to-endogenous correction*: obtain a source outcome profile as a source-library prior, predict a context-induced shift from endogenous features (e.g., chromatin accessibility, transcriptional activity), and recover the endogenous profile as the source profile plus the shift. We introduce BE-CORR (*Source-to-Endogenous Correction of Base-Editing Outcomes*), an interpretable and robust correction model that represents the source profile as nonnegative editing hazard, rescales overall editing opportunity, and redistributes per-position opportunity via constrained hazard transport with reliability-gated blending. Experiments show that BE-CORR consistently improves endogenous endpoint accuracy and ranking and matched source-to-endogenous transfer over strong sequence-only, context-aware, and source-using baselines. Code and data are available in our anonymous code repository.

1 INTRODUCTION

Base editors enable targeted single-nucleotide substitutions without creating a double-strand break, making them powerful tools for functional genomics and the therapeutic correction of pathogenic variants (Komor et al., 2016; Gaudelli et al., 2017; Rees & Liu, 2018; Mawson et al., 2026). Yet base editing is inherently *multi-outcome*: a given editor–guide pair can modify multiple nearby positions, producing a characteristic per-position editing profile together with undesired bystander edits. Here, the editor determines which base conversion is possible, and the guide RNA specifies the target sequence. This has motivated a large literature on data-driven outcome prediction trained on high-throughput target-library assays, spanning feature-based models such as BE-Hive (Arbab et al., 2020) and neural sequence models (Marquart et al., 2021; Kim et al., 2024; Pallaseni et al., 2022; Kissling et al., 2025; Sun et al., 2025). Beyond task-specific outcome predictors, recent genomic foundation models also study general-purpose DNA representation learning at scale (Brixi et al., 2026; Del Vecchio et al., 2026; Vishniakov et al., 2026).

A central obstacle is that most large-scale training data are collected in **source-library** settings—integrated, plasmid, or otherwise synthetic targets measured outside their native genomic loci—whereas most applications edit **endogenous** sites in their native chromatin context. Endogenous loci vary in chromatin accessibility, transcriptional activity, epigenetic marks, and protein occupancy, all of which can modulate editor access and local processing (Daer et al., 2017; Jensen et al., 2017). As a result, even with the same editor and the same target sequence, editing outcomes can shift

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.

substantially between source-library measurements and endogenous loci (Yuan et al., 2024). This creates a dataset shift between source-library and endogenous measurements (Quiñonero-Candela et al., 2008) and, more specifically, a source-to-target transfer problem (Ben-David et al., 2010; Pan & Yang, 2010). Recent endogenous models incorporate locus-level features to improve accuracy (Yuan et al., 2024), but they largely treat endogenous prediction as a *direct* supervised mapping from sequence and genomic features to outcomes.

We propose to recast endogenous prediction as *source-to-endogenous correction*. Instead of predicting endogenous outcomes from scratch, we **first** obtain a source-library outcome profile as a source-library prior learned from large-scale source-library measurements such as CRISPRbase (Fan et al., 2023). We **then** predict a context-dependent correction term conditioned on endogenous-context features, and form **the endogenous prediction as the source profile plus the correction** (Problem 1). This decomposition separates the source-library profile from context-dependent native-locus effects, enabling data-efficient transfer while making full use of abundant source-library data.

To solve this correction problem, we introduce BE-CORR (*Source-to-Endogenous Correction of Base-Editing Outcomes*). BE-CORR is designed to be both robust and interpretable: it represents the source profile as a source-derived *editing hazard* (editing opportunity) along the protospacer (20-nt target sequence), then uses endogenous features to (i) rescale overall editing opportunity and (ii) *transport* hazard mass (reallocate per-position opportunity) locally under biological constraints. Because source profiles may be noisy or imputed, BE-CORR includes a reliability gate that falls back to a direct endogenous predictor when the source signal is unreliable, thereby limiting the influence of an unreliable source profile. It also provides interpretable learned factors that explain how overall editing changes, where editing opportunity is reallocated, and how strongly the source profile is trusted.

Empirically, we use the BE_Endo endogenous benchmark (Yuan et al., 2024) and reconstruct 8,450 measured integrated–endogenous pairs from its source-library experiments. We compare against sequence-only predictors that use editor and local target-sequence information and context-aware predictors that also use native-locus features, together with adapted genomic foundation models. Overall, BE-CORR outperforms these model families in endpoint prediction and matched source-to-endogenous transfer tasks (Tables 1 and 2).

Our main contributions are:

- **Problem formulation:** recast endogenous prediction as *source-to-endogenous correction*, decomposing the endogenous profile into a source-library prior and a context-induced shift, thereby leveraging abundant source-library data.
- **Method:** introduce BE-CORR, a principled framework that maps a source-library outcome to its endogenous counterpart for a given editor–DNA target pair, with robustness and interpretability built into the correction process.
- **Empirical evidence:** demonstrate consistent improvements over strong sequence-only and context-aware baselines on both large-scale endogenous endpoint prediction and matched source-to-endogenous transfer benchmarks.

2 Background and Problem Setup

Base editors make targeted DNA base substitutions at a guide-specified site. Predicting editing outcomes is essential for selecting target sites and estimating efficiency and bystander editing before running experiments. Most large-scale training data come from **source-library** assays that measure many targets in controlled settings, but real applications typically edit **endogenous** sites at their native genomic loci, where chromatin and other cellular factors can shift outcomes. This section formalizes these settings and introduces our task: correcting a source-library outcome into an endogenous outcome.

2.1 DNA Editing Outcome Prediction and Source-to-Endogenous Shift

Base-editing outcome prediction. This paper studies outcome prediction for base editing, a genome editing modality that makes targeted DNA base substitutions near a guide-specified tar-

get site. The unit of prediction is an *editing instance*: one **measurement** of applying an editor (ABE or CBE in our setting) to one **local DNA target site** under one **context**. We represent each instance by the editor e_i, a 20-nt guide-matched protospacer g_i, an adjacent PAM motif p_i, and a 40-nt local sequence s_i containing the protospacer, PAM, and flanking bases. Because an editor can modify multiple nearby positions (including unintended "bystander" edits), the label is a 20-position editing-rate vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i20}) \in [0,1]^{20}$, where y_{ij} is the marginal editing frequency at protospacer position j. The standard sequence-to-outcome problem can be written as

$$f_\theta(e_i, g_i, p_i, s_i) \to \hat{\boldsymbol{y}}_i. \tag{1}$$

This prediction targets the full per-position editing profile from editor and local sequence. Readers unfamiliar with DNA base editing may refer to Appendix A for Table 4 (notation) and detailed explanations of biological terms.

Source-library and endogenous context as distribution shift. Most high-throughput data for base-editing prediction come from **source-library context**: controlled library assays in which many target sequences are measured outside their native genomic loci, for example in artificial, plasmid, or integrated settings. These data are useful for learning sequence-level editing rules, but the clinically or biologically relevant use case is usually **endogenous context**: editing the target at its native genomic locus inside a cell. Endogenous loci differ in chromatin accessibility, transcriptional activity, DNA and histone modifications, and protein occupancy, all of which can affect whether an editor reaches the site and what outcomes are realized. Thus, matching the editor and target sequence alone does not guarantee the same editing outcome in source-library and endogenous settings. Treating this shift as residual noise limits how well sequence-trained models transfer to the setting users ultimately care about. A direct endogenous predictor can incorporate native-locus features,

$$f_\theta(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}) \to \hat{\boldsymbol{y}}_i^{\mathrm{endo}}. \tag{2}$$

However, this still frames endogenous prediction as a direct supervised mapping and does not explicitly model how a source-library outcome should change when the same target is edited in its native genomic context.

2.2 Source-to-Endogenous Correction

Source-to-endogenous transfer for endogenous outcome prediction. One can frame endogenous outcome prediction directly as learning $f_\theta(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}) \to \hat{\boldsymbol{y}}_i^{\mathrm{endo}}$. However, endogenous measurements are expensive and limited, while source-library assays can produce much larger datasets that capture sequence- and editor-associated editing patterns under controlled assay conditions. We therefore reformulate endogenous prediction as *source-to-endogenous correction*: given a source-library outcome $\boldsymbol{y}_i^{\mathrm{src}}$ (measured or predicted) as a baseline summary of how the editor behaves on the target sequence, the model focuses on learning how endogenous context shifts this baseline to the native locus. This formulation is well-grounded and practically important: it turns the source–endogenous mismatch into an explicit object of prediction, enabling data-efficient transfer from source libraries. We formalize this learning problem below.

Problem 1 (Source-to-endogenous correction)**.** For an editing instance i, let $\boldsymbol{y}_i^{\mathrm{src}} \in [0,1]^{20}$ denote the source-library outcome and $\boldsymbol{y}_i^{\mathrm{endo}} \in [0,1]^{20}$ denote the endogenous outcome. Let $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ be the available source profile, equal to $\boldsymbol{y}_i^{\mathrm{src}}$ when measured and otherwise imputed. Given the target information

$$\mathbf{x}_i = (e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}), \tag{3}$$

which contains the editor, guide-matched protospacer, PAM, local sequence, and endogenous context, learn a source-conditioned correction and the corresponding endogenous outcome:

$$\hat{\Delta}_i = h_\theta(\mathbf{x}_i, \tilde{\boldsymbol{y}}_i^{\mathrm{src}}), \qquad \hat{\boldsymbol{y}}_i^{\mathrm{endo}} = \tilde{\boldsymbol{y}}_i^{\mathrm{src}} + \hat{\Delta}_i \in [0,1]^{20}. \tag{4}$$

Here h_θ predicts how the available source profile should be corrected using the target information. When the source outcome is measured, the ground-truth shift is $\Delta_i = \boldsymbol{y}_i^{\mathrm{endo}} - \boldsymbol{y}_i^{\mathrm{src}}$.

This differs from both sequence-only prediction (Eq. 1) and direct context-aware prediction (Eq. 2): the predictor receives an explicit source-library outcome and learns how to correct it for the endogenous context. This encourages a decomposition into (i) a source-library prior and (ii) a context-dependent correction informed by the target features.

Practical transfer formulation. We consider two practical regimes. When a matched source measurement is available, the model receives experimental $\boldsymbol{y}_i^{\mathrm{src}}$; otherwise it receives an imputed prior $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ from a predictor q_ϕ trained on large-scale source-library data. In both cases, $\boldsymbol{y}_i^{\mathrm{endo}}$ is an experimental endogenous target. Formal dataset definitions and pairing criteria are given in Appendix B. **Terminology.** If $\boldsymbol{y}_i^{\mathrm{src}}$ is experimentally measured, we refer to this setting as *matched source-to-endogenous transfer*; if only $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ is available, we refer to *endogenous endpoint prediction* because the measured shift is unobserved.

3 METHOD

3.1 MOTIVATION AND DESIGN PRINCIPLES

Source-library outcomes capture sequence- and editor-associated editability in a controlled, non-native assay context. At an endogenous locus, chromatin and regulatory context can change both the overall editing opportunity and its allocation across positions. BE-CORR therefore represents the source profile as nonnegative hazard, rescales its total mass, and locally redistributes that mass under biological constraints. Because the source profile may be imputed or noisy, a reliability gate blends the transported source signal with a direct endogenous prediction. This separates three questions—how much editing occurs, where it occurs, and how much the source should be trusted—whose mathematical properties are studied in Section 4 and Appendix E.

3.2 BE-CORR MODEL

We now describe these operations in three modules: source-hazard construction and activity scaling, local hazard reallocation, and reliability-gated endogenous prediction.

Transforming Source Probabilities into Hazard and Scaling Total Activity. We first transform the bounded source editing probabilities into a nonnegative hazard representation, in which editing opportunity can be naturally scaled and redistributed, and then calibrate its overall activity to the endogenous locus. For instance i, let $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ denote the measured source editing profile when available, or the prior predicted by q_ϕ otherwise. Let $\mathbf{m}_i = \mathbf{m}(e_i, s_i) \in \{0,1\}^{20}$ denote the corresponding chemically editable mask, with m_{ik} indicating whether position k is editable. A context encoder $F_\theta(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}, \tilde{\boldsymbol{y}}_i^{\mathrm{src}})$ integrates the local DNA sequence, editor and PAM information, endogenous context features, and summaries of the source profile, and outputs a raw activity shift $\tilde{a}_i$, a position-specific exposure potential $\mathbf{b}_i$, and a learned reliability representation $\mathbf{z}_i^{\mathrm{rel}}$. After bounding $\tilde{a}_i$ to obtain a_i, we define

$$\mathcal{H}(y) = -\log(1-y), \qquad \boldsymbol{\mu}_i = \mathbf{m}_i \odot \mathcal{H}(\tilde{\boldsymbol{y}}_i^{\mathrm{src}}), \qquad \bar{\boldsymbol{\mu}}_i = \exp(a_i)\boldsymbol{\mu}_i. \tag{5}$$

Here, $\boldsymbol{\mu}_i$ represents the source-derived editing opportunity across positions. Its total mass, $\sum_j \mu_{ij}$, captures the overall editing activity, whereas its relative coordinates characterize how that activity is distributed across the editing window. The multiplicative factor $\exp(a_i)$ therefore adjusts the total editing opportunity while preserving its source-derived positional pattern, providing a parsimonious representation of locus-level changes in editor accessibility and activity.

The exposure potential $\mathbf{b}_i$ modulates positional redistribution through the transport cost in Eq. 6, whereas the reliability representation $\mathbf{z}_i^{\mathrm{rel}}$ determines how strongly the model should trust the source-derived profile through the reliability gate in Eq. 8. Details of the bounded activity parameterization and encoder features are provided in Appendix C.

Locally Reallocating Hazard. We next reallocate the scaled hazard locally among chemically editable positions while preserving its total mass. Let $\mathcal{A}_i = \{k : m_{ik} = 1\}$ be the editable support and W_{e_i} the canonical window of editor e_i. Moving hazard from source position j to destination $k \in \mathcal{A}_i$ incurs cost

$$C_{i,jk} = \frac{(j-k)^2}{2\ell_{e_i}^2} + \lambda_{\mathrm{win}} \mathbf{1}\{k \notin W_{e_i}\} - b_{ik}. \tag{6}$$

The three terms respectively discourage long-range movement, penalize destinations outside the editor window, and favor positions supported by the endogenous exposure potential. In particular,

the position-wise encoder head constructs $\mathbf{b}_i$ from sequence, editor/PAM, endogenous features, and source summaries, and $-b_{ik}$ lowers the cost of context-supported destinations. Here ℓ_e is the learned movement scale and λ_{win} the window penalty. A row-constrained entropy-regularized transport with temperature τ_{ot} then gives

$$\Pi_{i,jk} = \frac{\exp(-C_{i,jk}/\tau_{\text{ot}})}{\sum_{u\in\mathcal{A}_i}\exp(-C_{i,ju}/\tau_{\text{ot}})}\mathbf{1}\{k\in\mathcal{A}_i\}, \qquad \nu^{\text{tr}}_{ik} = \sum_{j=1}^{20}\bar{\mu}_{ij}\Pi_{i,jk}. \tag{7}$$

Here $\Pi_{i,jk}$ is the fraction of hazard at source position j assigned to destination k, and $\boldsymbol{\nu}^{\text{tr}}_i$ is the resulting transported hazard profile. By construction, it preserves the scaled total hazard, assigns zero mass to non-editable destinations, and respects locality, editor-window, and exposure constraints. Unlike balanced optimal transport, it leaves destination mass unconstrained because endogenous opportunity is context-dependent; column constraints would impose an unsupported prior. Proposition 1 bounds error propagation and source-noise fallback, and characterizes when complementary source information improves the direct hazard estimate.

Choosing How Much to Trust the Transported Source. Finally, we choose how strongly to trust the transported source estimate from Eq. 7 relative to a source-independent direct endogenous estimate, and map the resulting hazard back to editing probabilities. An independently trained direct predictor $f_\psi(e_i, g_i, p_i, s_i, c^{\text{endo}}_i)$ provides the masked direct prediction $\hat{\boldsymbol{y}}^{\text{dir}}_i = \mathbf{m}_i \odot f_\psi(e_i, g_i, p_i, s_i, c^{\text{endo}}_i)$ and direct endogenous hazard $\boldsymbol{\nu}^{\text{dir}}_i = \mathbf{m}_i \odot \mathcal{H}(\hat{\boldsymbol{y}}^{\text{dir}}_i)$ without using the source profile. Let $G_i = \|\boldsymbol{\nu}^{\text{tr}}_i - \boldsymbol{\nu}^{\text{dir}}_i\|_1$ denote their hazard-space disagreement and let $M^{\text{src}}_i = \sum_j \mu_{ij}$ be the available source hazard. A learned score ρ_i combines this disagreement with transport difficulty and the learned reliability representation. We then attenuate that score explicitly when the two branches strongly disagree:

$$\rho_i = \rho_{\min} + (1-\rho_{\min})\sigma\big(r_\omega[\text{cost}(\Pi_i), G_i, \mathbf{z}^{\text{rel}}_i, M^{\text{src}}_i]\big)\,,\ \eta_i = \alpha^\star\rho_i\exp(-G_i/\kappa), \quad \kappa = 1. \tag{8}$$

Here $\sigma(t) = (1+e^{-t})^{-1}$ is the logistic sigmoid and r_ω is a learned scalar-valued reliability network; ρ_i is the learned source-reliability score, whereas η_i is the effective source weight after validation-selected global shrinkage $\alpha^\star$ and fixed disagreement attenuation κ. The corrected hazard and probability profile are

$$\hat{\boldsymbol{\nu}}^{\text{endo}}_i = \eta_i\boldsymbol{\nu}^{\text{tr}}_i + (1-\eta_i)\boldsymbol{\nu}^{\text{dir}}_i, \qquad \hat{\boldsymbol{y}}^{\text{endo}}_i = \mathbf{m}_i \odot \mathcal{H}^{-1}(\hat{\boldsymbol{\nu}}^{\text{endo}}_i). \tag{9}$$

Thus a reliable, compatible source profile can influence the endogenous prediction, whereas a large source–direct disagreement forces a conservative fallback toward the direct predictor. The complete gate, objective, hyperparameters, and pseudocode are given in Appendices C and D. The activity shift a_i, exposure potential $\mathbf{b}_i$, transport plan Π_i, and reliability score ρ_i remain available as interpretable correction factors; Appendix H compares them with measured source-to-endogenous activity and positional changes.

4 Theoretical Analysis

We bound error propagation and source-induced degradation, and characterize when complementary transported source information improves the direct hazard estimate. See Appendix E for the proof.

Proposition 1 (Prediction-error bound, source-noise robustness, and source utility)**.** *For one record, we suppress the instance index i. Let $\boldsymbol{\nu}^\star$ be the true endogenous hazard and let $\boldsymbol{\mu}^\star, a^\star, C^\star$ be ideal source, activity, and cost quantities on a common nonempty editable support. BE-CORR uses $\tilde{\boldsymbol{\mu}}, a, C$. Assume $\|\boldsymbol{\mu}^\star\|_1 \le M$ and $\max\{|a|, |a^\star|\} \le A$. Let ϵ_{str} be the ideal transport approximation error, and let δ_μ, δ_a, and δ_C be the ℓ_1 source, absolute activity, and uniform cost errors (defined explicitly in Appendix E). Then*

$$\begin{aligned}\|\boldsymbol{\nu}^{\text{tr}} - \boldsymbol{\nu}^\star\|_1 &\le B_{\text{tr}} := \epsilon_{\text{str}} + e^A[\delta_\mu + M\delta_a + 2M\tanh(\delta_C/\tau_{\text{ot}})]\,,\\ \|\hat{\boldsymbol{y}}^{\text{endo}} - \boldsymbol{y}^\star\|_1 &\le (1-\eta)\epsilon_{\text{dir}} + \eta B_{\text{tr}}.\end{aligned} \tag{10}$$

Here $\boldsymbol{y}^\star = \mathbf{m}\odot\mathcal{H}^{-1}(\boldsymbol{\nu}^\star)$, $\hat{\boldsymbol{y}}^{\text{dir}} = \mathbf{m}\odot\mathcal{H}^{-1}(\boldsymbol{\nu}^{\text{dir}})$, $\epsilon_{\text{dir}} = \|\boldsymbol{\nu}^{\text{dir}} - \boldsymbol{\nu}^\star\|_1$, and $\eta = \alpha^\star\rho e^{-G/\kappa}$, where $G = \|\boldsymbol{\nu}^{\text{tr}} - \boldsymbol{\nu}^{\text{dir}}\|_1$ and $\alpha^\star, \rho \in [0,1]$, $\kappa > 0$. Moreover,

$$\|\hat{\boldsymbol{y}}^{\text{endo}} - \hat{\boldsymbol{y}}^{\text{dir}}\|_1 \le \alpha^\star\rho G e^{-G/\kappa} \le \frac{\alpha^\star\kappa}{e}, \qquad \|\hat{\boldsymbol{y}}^{\text{endo}} - \boldsymbol{y}^\star\|_1 - \|\hat{\boldsymbol{y}}^{\text{dir}} - \boldsymbol{y}^\star\|_1 \le \frac{\alpha^\star\kappa}{e}. \tag{11}$$

For squared hazard error, set $\mathbf{e}_b = \boldsymbol{\nu}^b - \boldsymbol{\nu}^\star$ *for* $b \in \{\mathrm{dir}, \mathrm{tr}\}$. *Then*

$$\|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2 = \eta\big(\|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2\big) - \eta(1-\eta)\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2. \tag{12}$$

For $\eta > 0$*, source improves squared hazard error over the direct branch iff* $(1-\eta)\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2 > \|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2$*; this can hold for a less accurate but complementary source branch (Appendix E).*

Proposition 1 quantifies how errors in the source profile, activity scaling, and transport cost propagate to endogenous prediction. It further shows that large source–direct disagreement drives BE-CORR toward direct prediction, while a transported source need not be the more accurate standalone branch to help: sufficiently complementary errors can still reduce squared hazard error.

5 EXPERIMENTS

Research questions addressed in Section 5 and key findings.

- **RQ1 (predictive performance):** How does BE-CORR perform with imputed versus experimentally measured source profiles? **Finding:** It attains average ranks of 1.10 and 1.00, respectively (Tables 1 and 2).
- **RQ2 (source-noise robustness):** How sensitive is BE-CORR to an inaccurate source profile? **Finding:** As source corruption increases, reliability gating keeps BE-CORR closer to direct endogenous prediction, consistent with the gate's intended fallback behavior (Figure 1A).
- **RQ3 (source utility):** Does the source profile provide useful information for endogenous prediction? **Finding:** Source-using models achieve positive correction gain with measured source profiles, and BE-CORR gains more where it makes greater effective use of source information (Table 2; Figure 1B).
- **RQ4 (ablation):** Do BE-CORR's source and correction design choices contribute to performance? **Finding:** BE-CORR improves over direct endogenous prediction (Eq. 2), using the source profile alone without endogenous context, and a simple source plus linear correction in both settings (Table 3).
- **RQ5 (interpretability):** Do the exposed correction factors track measured source-to-endogenous changes? **Finding:** Activity, exposure, transport, and reliability factors show empirical correspondence with total-activity changes, positional reallocation, and source utility (Appendix H; Figure 3).

Appendix H reports additional results on low-data behavior, correction-design ablations, and correction-factor diagnostics.

5.1 BENCHMARKS

Main endogenous endpoint benchmark. We use BE_Endo (Yuan et al., 2024) as our primary benchmark for endogenous editing prediction. After curation, the dataset contains $N_{\mathrm{endo}} = 9{,}116$ records, comprising 4,529 ABE and 4,587 CBE examples. Each record provides an experimentally measured 20-position endogenous editing-rate vector together with nine native-locus features spanning expression, methylation, DNase accessibility, histone marks, CTCF, and Pol II (Section 2.1). We evaluate all methods over five splits, each containing 6,381/1,367/1,368 records for training, validation, and testing, respectively. No test protospacer exactly matches a same-editor training protospacer in any split (Appendix B). Endpoint experiments do not expose experimentally measured source profiles. Only the source-prior augmented family additionally receives an *imputed* source profile from a model trained on measured integrated-library outcomes in the corresponding training fold; held-out source outcomes are never used. Dataset definitions, filtering procedures, source-profile augmentation, and split construction are detailed in Appendix B.

Matched source-to-endogenous transfer benchmark. We reconstruct a matched transfer benchmark by pairing BE_Endo integrated-library measurements with the corresponding endogenous experiments. The resulting dataset contains 8,450 exact source–endogenous pairs, including 4,207 ABE and 4,243 CBE examples, matched by both the concrete editor construct and the 40-nt local target context. We use 70/15/15 train/validation/test splits. Unlike the endpoint setting above, this transfer benchmark provides experimentally measured source outcomes; the endpoint benchmark uses imputed source profiles. Details of read processing, source–endogenous matching, and split construction are provided in Appendix B.

5.2 BASELINES

Baseline families and protocol. We compare models according to the information available at prediction time: sequence/editor-using models, models that additionally use endogenous context, and models that also receive a target-matched source profile. Only source-prior augmented models receive the imputed source prior $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ on the endpoint benchmark or measured $\boldsymbol{y}_i^{\mathrm{src}}$ on the transfer benchmark; other methods retain their native input families.

Our sequence-only baselines include BE-Hive, DeepBE, BE-DICT, BEDICT2.0, and CRISPRon-BE (Arbab et al., 2020; Kim et al., 2024; Marquart et al., 2021; Kissling et al., 2025; Sun et al., 2025), and our sequence+endogenous-context baselines follow the BE_Endo models (Yuan et al., 2024). We also adapt four methods from the broader machine-learning literature as baselines that incorporate source editing profiles: HyenaDNA (Nguyen et al., 2023), Caduceus (Schiff et al., 2024), FT-Transformer (Gorishniy et al., 2021), and Deep Adaptation Networks (DAN) (Long et al., 2015). Because public baselines differ in their native outputs (scalar efficiency, window summaries, or edited-product distributions), we keep each method's released sequence encoder or feature recipe wherever the BE_Endo records provide the required inputs, and replace only the task-specific prediction head to match the common 20-position endogenous target. Appendix G lists the baseline-specific inputs, output harmonization, and training details.

5.3 METRICS

Evaluation protocol. We report metrics under two evaluation regimes. On the main endogenous benchmark, measured source outcomes are not available for each BE_Endo target; therefore, all reported metrics are endpoint endogenous metrics. Position-level MAEs evaluate the 20-position outcome vector directly, scalar summaries evaluate window marginal-edit burden, dominant-position editing, and non-dominant burden, and ranking/calibration metrics evaluate target prioritization and burden calibration.

On the matched source-to-endogenous benchmark, source outcomes are experimentally measured for the same editor–target pairs as the endogenous outcomes. We report endpoint endogenous errors and shift-focused metrics on $\Delta_i = \boldsymbol{y}_i^{\mathrm{endo}} - \boldsymbol{y}_i^{\mathrm{src}}$, including shift rank correlation, large-shift discovery, and correction gain over the measured source outcome. Because shift-space MAEs are algebraically equal to the corresponding endpoint MAEs, we report each error only once and reserve the shift-specific columns for ranking, discovery, and gain. Appendix F gives all metric definitions and formulas.

5.4 RESULTS

Endogenous endpoint prediction. Table 1 reports mean performance over the five seeds on the main BE_Endo benchmark. This table evaluates endpoint endogenous prediction for all methods. BE-CORR ranks first on nine of ten endpoint metrics and has the best average rank. FT-Transformer is the strongest competing method; relative to it, BE-CORR reduces window, burden, and dominant-edit MAE by 11.6%, 10.5%, and 10.9%, and raises window correlation from 0.8868 to 0.9075. It also lowers window MAE by 17.4% relative to BE-Hive. BE-Hive attains the lowest ECE, with BE-CORR ranking second on calibration.

Matched source-to-endogenous transfer. Table 2 evaluates matched source-to-endogenous transfer on the paired test folds. BE-CORR ranks first on every reported transfer metric. FT-Transformer is the strongest competing method and the overall runner-up by average rank; relative to it, BE-CORR reduces window, burden, and dominant-edit MAE by 14.8%, 13.7%, and 13.3%, raises burden-shift Spearman correlation from 0.5860 to 0.6427, and raises correction gain from 0.2340 to 0.3466. Because measured-source copy has access to an experimentally measured source outcome that source-free models do not receive, it is an information-advantaged reference for those models; accordingly, most source-free models have negative Gain.

Robustness to Corrupted Source Profiles and Transport Validity. Figure 1 probes BE-CORR internals on the full BE_Endo test splits. Panel A tests robustness to predicted-source quality: when the imputed source prior is corrupted, BE-CORR attenuates the resulting degradation and moves to-

Table 1: **Main endogenous endpoint benchmark over five seeds.** Errors and ECE are percentage-point (pp) values. Win./Full: window/20-position MAE; Burden/Dom./Non-dom.: MAE of marginal-edit burden, dominant editing, and non-dominant burden; Win. r/Burden ρ: window Pearson/burden Spearman; AUPRC/NDCG@10%: high-burden discovery/top-decile ranking. Avg. rank averages metric-wise ranks. **Bold** marks the best and underlining marks the runner-up; arrows indicate direction.

		Error metrics (pp) ↓					Ranking and correlation ↑				Calibration (pp) ↓	Overall ↓
Model	N	**Win.**	**Full**	**Burden**	**Dom.**	**Non-dom.**	**Win.** r	**Burden** ρ	**AUPRC**	**NDCG@10%**	**ECE**	**Avg. rank**
Sequence + editor information												
kNN sequence	5	6.71 ± 0.11	2.05 ± 0.02	24.62 ± 0.56	17.22 ± 0.35	13.95 ± 0.19	0.6708 ± 0.0084	0.6073 ± 0.0305	0.4042 ± 0.0261	0.7171 ± 0.0233	3.25 ± 0.76	10.70
BE-Hive	5	2.95 ± 0.04	0.94 ± 0.01	14.46 ± 0.31	11.17 ± 0.17	4.33 ± 0.13	0.8839 ± 0.0027	0.8660 ± 0.0073	0.5840 ± 0.0236	0.8841 ± 0.0294	**1.72 ± 0.32**	4.40
DeepBE	5	6.87 ± 0.12	2.06 ± 0.03	37.83 ± 0.65	27.90 ± 0.25	9.93 ± 0.41	0.5076 ± 0.0535	0.5701 ± 0.0446	0.3784 ± 0.0572	0.5788 ± 0.0650	37.83 ± 0.65	12.90
BE-DICT	5	6.10 ± 0.12	2.01 ± 0.08	33.11 ± 0.65	24.67 ± 0.37	8.70 ± 0.35	0.6897 ± 0.0758	0.6797 ± 0.0650	0.5180 ± 0.0809	0.6965 ± 0.0356	32.07 ± 0.66	11.10
BEDICT2.0	5	5.54 ± 0.14	1.66 ± 0.03	30.12 ± 0.80	21.93 ± 0.40	8.42 ± 0.45	0.7332 ± 0.0045	0.7011 ± 0.0132	0.5657 ± 0.0112	0.6790 ± 0.0399	27.65 ± 0.90	10.30
CRISPRon-BE	5	3.91 ± 0.19	1.23 ± 0.05	20.26 ± 1.15	15.29 ± 1.04	5.59 ± 0.25	0.7805 ± 0.0246	0.7460 ± 0.0401	0.5162 ± 0.0309	0.8722 ± 0.0188	14.66 ± 1.53	8.50
BE_Endo Seq	5	3.05 ± 0.05	1.04 ± 0.02	15.07 ± 0.26	11.59 ± 0.22	4.40 ± 0.12	0.8766 ± 0.0054	0.8594 ± 0.0087	0.5906 ± 0.0236	0.8880 ± 0.0117	4.27 ± 0.97	6.20
Sequence + editor information + endogenous context												
BE_Endo Endo	5	3.04 ± 0.05	1.03 ± 0.02	15.02 ± 0.35	11.56 ± 0.21	4.39 ± 0.19	0.8776 ± 0.0059	0.8602 ± 0.0098	0.5896 ± 0.0228	0.8930 ± 0.0122	4.59 ± 0.64	5.60
DAN	5	4.99 ± 0.04	1.59 ± 0.04	26.61 ± 0.25	19.98 ± 0.22	7.06 ± 0.16	0.7780 ± 0.0263	0.7874 ± 0.0131	0.5165 ± 0.0343	0.8105 ± 0.0268	23.88 ± 0.37	9.30
Sequence + editor information + endogenous context + source profile												
HyenaDNA	5	2.90 ± 0.08	1.15 ± 0.08	14.30 ± 0.49	10.94 ± 0.37	4.31 ± 0.12	0.8777 ± 0.0071	0.8725 ± 0.0085	0.6132 ± 0.0307	0.8823 ± 0.0198	3.03 ± 0.89	4.50
Caduceus	5	2.90 ± 0.04	1.09 ± 0.03	14.27 ± 0.27	10.98 ± 0.25	4.24 ± 0.09	0.8789 ± 0.0050	0.8722 ± 0.0084	0.6208 ± 0.0266	0.8874 ± 0.0201	2.90 ± 0.44	3.70
FT-Transformer	5	2.76 ± 0.06	0.97 ± 0.02	13.61 ± 0.29	10.27 ± 0.35	4.22 ± 0.12	0.8868 ± 0.0056	0.8793 ± 0.0078	0.6308 ± 0.0346	0.8859 ± 0.0101	3.03 ± 0.74	2.70
BE-CORR	5	**2.44 ± 0.05**	**0.81 ± 0.02**	**12.19 ± 0.30**	**9.15 ± 0.22**	**3.74 ± 0.09**	**0.9075 ± 0.0044**	**0.8944 ± 0.0076**	**0.6524 ± 0.0315**	**0.9065 ± 0.0146**	2.30 ± 0.61	**1.10**

Table 2: **Matched source-to-endogenous transfer over five seeds.** Results use 8,450 exact editor-construct/40-nt-context pairs with measured source and endogenous outcomes. Burden/Dom.: window marginal-edit burden and dominant-position rate; $\rho_{\Delta T}$: burden-shift Spearman; AUPRC: large-shift discovery; Gain: relative window-MAE improvement over source copy. **Bold** marks the best and underlining marks the runner-up.

		Endogenous error (pp) ↓			Shift recovery / gain ↑			Overall ↓
Model	N	**Win.**	**Burden**	**Dom.**	$\rho_{\Delta T}$	**AUPRC**	**Gain**	**Avg. rank**
Measured-source reference								
Source measured outcome	5	3.31 ± 0.08	17.07 ± 0.50	12.95 ± 0.29	–	0.2003 ± 0.0002	0.0000 ± 0.0000	10.00
Sequence + editor information								
kNN sequence	5	6.76 ± 0.11	24.58 ± 0.40	17.33 ± 0.35	0.4129 ± 0.0251	0.3584 ± 0.0153	−1.0423 ± 0.0502	11.50
BE-Hive	5	2.97 ± 0.06	14.52 ± 0.35	11.21 ± 0.26	0.5225 ± 0.0213	0.4818 ± 0.0307	0.1026 ± 0.0170	4.83
DeepBE	5	6.93 ± 0.13	38.18 ± 0.65	28.12 ± 0.34	0.3898 ± 0.0290	0.4268 ± 0.0203	−1.0926 ± 0.0824	13.00
BE-DICT	5	6.18 ± 0.13	33.62 ± 0.76	25.07 ± 0.77	0.3875 ± 0.0310	0.4259 ± 0.0212	−0.8675 ± 0.0491	12.33
BEDICT2.0	5	5.67 ± 0.16	30.89 ± 0.90	22.50 ± 0.55	0.3824 ± 0.0299	0.4294 ± 0.0211	−0.7122 ± 0.0804	11.33
CRISPRon-BE	5	3.93 ± 0.20	20.44 ± 1.32	15.36 ± 1.10	0.5546 ± 0.0238	0.4375 ± 0.0223	−0.1861 ± 0.0724	7.67
BE_Endo Seq	5	3.05 ± 0.07	15.05 ± 0.30	11.59 ± 0.33	0.5323 ± 0.0248	0.4798 ± 0.0318	0.0799 ± 0.0175	6.17
Sequence + editor information + endogenous context								
BE_Endo Endo	5	3.05 ± 0.07	15.01 ± 0.31	11.57 ± 0.32	0.5387 ± 0.0242	0.4834 ± 0.0279	0.0803 ± 0.0237	5.00
DAN	5	4.97 ± 0.14	26.56 ± 0.82	19.83 ± 0.60	0.4200 ± 0.0350	0.4401 ± 0.0277	−0.5023 ± 0.0792	9.50
Sequence + editor information + endogenous context + source profile								
HyenaDNA	5	2.67 ± 0.04	13.45 ± 0.25	10.25 ± 0.22	0.5036 ± 0.0278	0.4289 ± 0.0321	0.1926 ± 0.0150	5.67
Caduceus	5	2.60 ± 0.07	13.02 ± 0.42	10.01 ± 0.28	0.5200 ± 0.0196	0.4346 ± 0.0265	0.2156 ± 0.0153	4.50
FT-Transformer	5	2.54 ± 0.07	12.67 ± 0.43	9.54 ± 0.38	0.5860 ± 0.0253	0.4732 ± 0.0238	0.2340 ± 0.0197	2.50
BE-CORR	5	**2.16 ± 0.06**	**10.94 ± 0.39**	**8.27 ± 0.26**	**0.6427 ± 0.0181**	**0.5491 ± 0.0513**	**0.3466 ± 0.0154**	**1.00**

ward the direct endogenous prediction, supporting the reliability-gated correction in Eq. 9. Panel B asks where the source signal helps: records with larger effective source use have larger gains over direct prediction, showing that BE-CORR can exploit source information when it is reliable. Panel C checks transport validity: no hazard mass is assigned to non-editable positions, and transport is highly local, with 94.3% of mass moving by at most one position and 99.2% by at most two positions.

5.5 Ablation Study

Table 3 contrasts direct endogenous prediction, source copying, and a linear source–context correction. The direct predictor is competitive but does not match the full model: BE-CORR reduces endpoint window MAE by 14.8% and matched-transfer window and burden MAE by 27.7% and 16.5%, respectively, relative to direct prediction. Correction gain rises from 0.096 to 0.3466, sup-

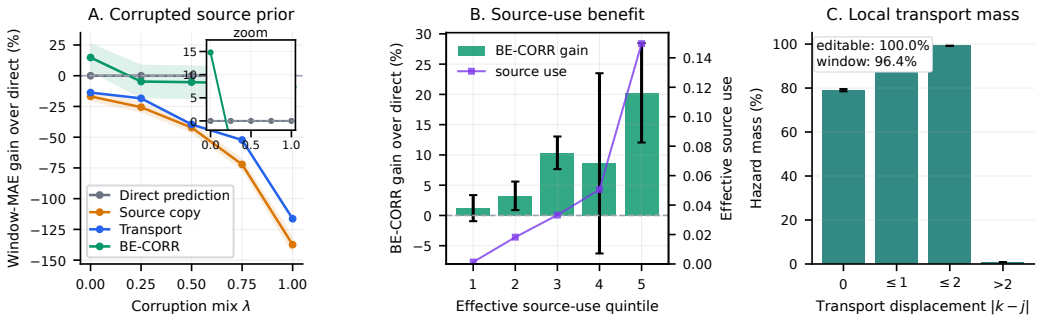


Figure 1: **BE-CORR robustness and transport diagnostics over five BE_Endo test seeds.** (A) Source corruption interpolates in hazard space between the imputed prior ($\lambda = 0$) and a similar target's prior ($\lambda = 1$). Curves report window-MAE gain over direct prediction; BE-CORR remains substantially closer to direct prediction than source copy or transport as corruption increases. (B) Records are grouped into seed-specific quintiles of effective source use, $U_i = \|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \hat{\boldsymbol{y}}^{\mathrm{dir}}\|_{1,W} / (\|\hat{\boldsymbol{y}}^{\mathrm{tr}} - \hat{\boldsymbol{y}}^{\mathrm{dir}}\|_{1,W} + \epsilon)$. Bars show BE-CORR gain over direct prediction; the purple line shows mean U_i. (C) Hazard-mass-weighted displacement (j: source; k: destination), with editable-support and canonical-window mass fractions annotated. Transport is confined to editable positions and is predominantly local ($|k - j| \leq 2$).

Table 3: **Source and correction ablations over five seeds.** Endpoint uses an imputed source prior; transfer uses the measured source profile. Controls use no source, source only, or a linear source–context correction. **Bold** marks the best mean.

	Endogenous endpoint benchmark			Matched source-to-endogenous transfer			
Variant	**Win. MAE (pp)**	**Burden MAE (pp)**	**Win. r**	**Win. MAE (pp)**	**Burden MAE (pp)**	$\rho_{\Delta T}$	**Gain**
Direct endogenous predictor	2.86 ± 0.07	13.09 ± 0.44	0.8926 ± 0.0034	2.99 ± 0.09	13.10 ± 0.47	0.5539 ± 0.0267	0.0961 ± 0.0243
Source profile only	3.34 ± 0.04	16.47 ± 0.21	0.8780 ± 0.0018	3.31 ± 0.08	17.07 ± 0.50	–	0.0000 ± 0.0000
Source + linear correction	3.04 ± 0.05	13.82 ± 0.34	0.8898 ± 0.0021	2.71 ± 0.07	12.54 ± 0.31	0.5283 ± 0.0258	0.1813 ± 0.0098
BE-CORR	$\mathbf{2.44 \pm 0.05}$	$\mathbf{12.19 \pm 0.30}$	$\mathbf{0.9075 \pm 0.0044}$	$\mathbf{2.16 \pm 0.06}$	$\mathbf{10.94 \pm 0.39}$	$\mathbf{0.6427 \pm 0.0181}$	$\mathbf{0.3466 \pm 0.0154}$

porting hazard scaling, local reallocation, and reliability weighting as a source-to-endogenous correction mechanism.

6 Conclusion and Limitations

We study endogenous base-editing outcome prediction under assay-to-native distribution shift and recast it as source-to-endogenous correction. We introduced BE-CORR, which represents source-library outcomes as editing hazards, uses endogenous context to rescale and locally transport editing opportunity, and reliability-gates the transported signal against direct endogenous prediction. Across endpoint and matched-transfer benchmarks, BE-CORR improves accuracy and ranking while exposing interpretable activity, transport, and reliability factors.

The low-data study further identifies where source correction is most useful: with only 10% of the endogenous training labels, BE-CORR reduces window MAE by 21.7% relative to direct prediction, and this advantage decreases as endogenous supervision grows. Together with the matched-transfer results, this finding shows how large source-library datasets can provide a useful prior when native-locus measurements are limited.

Limitations. The available matched data cover one ABE and one CBE construct in a single cellular setting. Generalization across editor variants, cell types, assay platforms, and laboratories therefore remains to be established. Several public baselines were originally designed for scalar efficiency or product-distribution prediction and therefore require output harmonization to the 20-position marginal-rate target. Finally, our study is computational. Prospective wet-lab validation is needed to establish whether BE-CORR improves experimental target selection.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, literature searches, research discussions, and code development. All AI-assisted text, code, and analyses were critically reviewed and verified against the underlying sources, data, and implementations. All methodological and interpretive decisions were made by the authors, who take full responsibility for the final content of the work.

ETHICS STATEMENT

This study uses publicly available, aggregate base-editing datasets and does not involve new human-subject, animal, or wet-lab experiments. The work concerns prediction of genome-editing outcomes and may therefore have dual-use implications. We present BE-CORR as a computational research tool, not a clinical decision system; any experimental or therapeutic use requires appropriate biosafety, ethical, and regulatory review. We adhere to the ICLR Code of Ethics and take responsibility for the integrity and reporting of this work.

REPRODUCIBILITY STATEMENT

We provide details on dataset construction, filtering, matching, and split protocols in Appendix B; model implementation, pseudocode, and validation-selected hyperparameters in Appendices C and D; complete metric definitions in Appendix F; baseline adaptations in Appendix G; and full proofs in Appendix E. Results are reported over five shared split-and-initialization seeds. All code and data required to reproduce our method and baseline results are available in our anonymous code repository.

REFERENCES

Mandana Arbab, Max W. Shen, Beverly Mok, Christopher Wilson, Żaneta Matuszek, Christopher A. Cassa, and David R. Liu. Determinants of base editing outcomes from target library analysis and machine learning. *Cell*, 182(2):463–480.e30, 2020. doi: 10.1016/j.cell.2020.05.037.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652:1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

René M. Daer, Josh P. Cutts, David A. Brafman, and Karmella A. Haynes. The impact of chromatin dynamics on Cas9-mediated genome editing in human cells. *ACS Synthetic Biology*, 6(3):428–438, 2017. doi: 10.1021/acssynbio.5b00299.

Alice Del Vecchio, Chantriolnt-Andreas Kapourani, Abdullah M. Athar, Agnieszka Dobrowolska, Andrew Anighoro, Benjamin Tenmann, Lindsay Edwards, and Cristian Regep. PatchDNA: A flexible and biologically-informed alternative to tokenization for DNA. In *International Conference on Learning Representations*, 2026.

Jibiao Fan, Leisheng Shi, Qi Liu, Zhipeng Zhu, Fan Wang, Runxian Song, Jimeng Su, Degui Zhou, Xiao Chen, Kailong Li, Lixiang Xue, Lichao Sun, and Fengbiao Mao. Annotation and evaluation of base editing outcomes in multiple cell types using CRISPRbase. *Nucleic Acids Research*, 51 (D1):D1249–D1256, 2023. doi: 10.1093/nar/gkac967.

Nicole M. Gaudelli, Alexis C. Komor, Holly A. Rees, Michael S. Packer, Ahmed H. Badran, David I. Bryson, and David R. Liu. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, 551(7681):464–471, 2017. doi: 10.1038/nature24644.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Kristopher T. Jensen, Lasse Fløe, Trine Skov Petersen, Jinrong Huang, Fengping Xu, Lars Bolund, Yonglun Luo, and Lin Lin. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Letters*, 591(13):1892–1901, 2017. doi: 10.1002/1873-3468.12707.

Nahye Kim, Sungchul Choi, Sungjae Kim, Myungjae Song, Jung Hwa Seo, Seonwoo Min, Jinman Park, Sung-Rae Cho, and Hyongbum Henry Kim. Deep learning models to predict the editing efficiencies and outcomes of diverse base editors. *Nature Biotechnology*, 42(3):484–497, 2024. doi: 10.1038/s41587-023-01792-x.

Lucas Kissling, Amina Mollaysa, Sharan Janjuha, Nicolas Mathis, Kim F. Marquart, Yanik Weber, Woohyun J. Moon, Paulo J. C. Lin, Steven H. Y. Fan, Hiromi Muramatsu, Máté Vadovics, Ahmed Allam, Norbert Pardi, Ying K. Tam, Michael Krauthammer, and Gerald Schwank. Predicting adenine base editing efficiencies in different cellular contexts by deep learning. *Genome Biology*, 26:115, 2025. doi: 10.1186/s13059-025-03586-7.

Alexis C. Komor, Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533(7603):420–424, 2016. doi: 10.1038/nature17946.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105. PMLR, 2015.

Kim F. Marquart, Ahmed Allam, Sharan Janjuha, Anna Sintsova, Lukas Villiger, Nina Frey, Michael Krauthammer, and Gerald Schwank. Predicting base editing outcomes with an attention-based deep learning algorithm trained on high-throughput target library screens. *Nature Communications*, 12:5114, 2021. doi: 10.1038/s41467-021-25375-z.

S. J. Mawson, Kara A. Dunne-Dombrink, Benjamin R. Pollak, and Alexis C. Komor. Precision genome editing with DNA base editors. *Nature Reviews Methods Primers*, 6:23, 2026. doi: 10.1038/s43586-026-00478-3.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43224, 2023.

Ananth Pallaseni, Elin Madli Peets, Jonas Koeppel, Juliane Weller, Thomas Vanderstichele, Uyen Linh Ho, Luca Crepaldi, Jolanda van Leeuwen, Felicity Allen, and Leopold Parts. Predicting base editing outcomes using position-specific sequence determinants. *Nucleic Acids Research*, 50(6):3551–3564, 2022. doi: 10.1093/nar/gkac161.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, 2008. ISBN 9780262170055. doi: 10.7551/mitpress/9780262170055.001.0001.

Holly A. Rees and David R. Liu. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature Reviews Genetics*, 19(12):770–788, 2018. doi: 10.1038/s41576-018-0059-1.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024.

Ying Sun, Kunli Qu, Giulia I. Corsi, Christian Anthon, Xiaoguang Pan, Xi Xiang, Lars Juhl Jensen, Lin Lin, Yonglun Luo, and Jan Gorodkin. Deep learning models simultaneously trained on multiple datasets improve base-editing activity prediction. *Nature Communications*, 16:9821, 2025. doi: 10.1038/s41467-025-65200-5.

Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveenkumar Kanithi, Marco A. F. Pimentel, Ronnie Rajan, and Shadab Khan. Tokenization to transfer: Do genomic foundation models learn good representations? In *International Conference on Learning Representations*, 2026.

Tanglong Yuan, Leilei Wu, Shiyan Li, Jitan Zheng, Nana Li, Xiao Xiao, Haihang Zhang, Tianyi Fei, Long Xie, Zhenrui Zuo, Di Li, Pinzheng Huang, Hu Feng, Yaqi Cao, Nana Yan, Xinming Wei, Lei Shi, Yongsen Sun, Wu Wei, Yidi Sun, and Erwei Zuo. Deep learning models incorporating endogenous factors beyond DNA sequences improve the prediction accuracy of base editing outcomes. *Cell Discovery*, 10:20, 2024. doi: 10.1038/s41421-023-00624-1.

A CONCEPT PRIMER

Notation. For quick reference, Table 4 summarizes notation used throughout the paper.

Table 4: **Notation.** A gene may contain multiple possible target sites; the unit of prediction here is the local editor–target-site measurement.

Symbol	Meaning
i	Editing instance: one editor–target-site measurement under one context
e_i	Broad editor class in this paper: adenine base editor (ABE) or cytosine base editor (CBE)
g_i	20-nt protospacer sequence matched by the guide RNA
p_i	PAM, the short DNA motif adjacent to the protospacer that is recognized by the Cas targeting module, such as Cas9, and required for target binding
s_i	40-nt local sequence containing the protospacer, PAM, and flanks
c_i^{endo}	Endogenous genomic context
$\boldsymbol{y}_i^{\mathrm{src}}, \boldsymbol{y}_i^{\mathrm{endo}}$	20-position editing-rate vectors in source and endogenous contexts
Δ_i	Source-to-endogenous outcome shift, $\boldsymbol{y}_i^{\mathrm{endo}} - \boldsymbol{y}_i^{\mathrm{src}}$

DNA bases and nucleotides. DNA is a polymer made of nucleotide positions. Each nucleotide carries one of four bases: adenine (A), cytosine (C), guanine (G), or thymine (T). In genomics papers, "base" is often used as shorthand for the identity of a nucleotide at a position in the DNA sequence. A single-base change therefore means changing the letter at one DNA position, for example from A to G or from C to T.

Base editors. A base editor is a programmable molecular system that can change the identity of one or more DNA bases within a local target site; each edited base can ultimately convert one base pair into another. DNA can be viewed as a sequence of four letters, A, C, G, and T, and base editing aims to install or correct such letters without replacing a long DNA segment. Unlike nuclease-based genome editing, which often cuts both DNA strands and relies on cellular repair, base editing is designed to make base substitutions without intentionally generating a double-strand break. A typical base editor combines a DNA-targeting module with a base-modifying enzyme domain. The targeting module, usually a Cas protein guided by a guide RNA, brings the editor to a local DNA sequence by recognizing a nearby PAM and matching the guide RNA to the protospacer. The enzyme domain then chemically modifies editable bases within a small nearby editing window.

Editing instance and target site. An editing instance refers to one measured case of applying a particular editor to a particular target site under a specific experimental or genomic context. The target site is a short local DNA sequence around the intended editing region. It is usually represented by a 20-nt protospacer, i.e., the 20 nucleotides recognized by the guide RNA, together with a nearby PAM and flanking bases on both sides. The PAM is required for Cas binding and positioning, while the flanking bases are included because the local sequence context can affect editing efficiency and possible nearby edits. A target site is different from a gene: one gene can contain many candidate target sites, and the same local target sequence can be tested with different editors or under different context conditions, producing different editing instances.

Adenine and cytosine base editors. Adenine base editors (ABEs) are commonly used to produce A→G edits, while cytosine base editors (CBEs) are commonly used to produce C→T edits. These are editor classes: they describe broad families of editors with different editable bases, different preferred editing positions, and different outcome distributions. Specific molecular editor systems can exist within each class, but the curated benchmark used here records the broad class only. Thus, in this paper, $e_i \in \{\mathrm{ABE}, \mathrm{CBE}\}$ denotes the editor class for editing instance i.

Guide, protospacer, and PAM. The guide RNA directs the editor to a DNA site through sequence complementarity. The 20-nt protospacer is the target segment matched by the guide RNA. The PAM, or protospacer-adjacent motif, is a short neighboring DNA motif recognized by the Cas targeting

module, such as Cas9, and is required for target binding and positioning. In our notation, g_i denotes the protospacer, p_i denotes the PAM, and s_i denotes a wider local sequence window containing the protospacer, PAM, and flanking bases.

Editing outcome vector. An editor does not necessarily edit every editable base equally. We represent the outcome for a target site as a 20-position vector, one entry per protospacer position. Entry y_{ij} is the marginal editing frequency at position j for editing instance i. This vector supports per-position and window-level summaries, including dominant and non-dominant marginal editing.

Editing window and bystander editing. Editors usually act most strongly within a local editing window inside the protospacer. If multiple editable bases fall in this window, the editor may modify more than one of them. In our metrics, the dominant edit is the largest edited position inside the canonical editing window, not necessarily a pre-specified intended therapeutic edit. Additional edits at nearby positions are bystander edits. Bystander editing matters because a target can have high window marginal-edit burden while still producing undesired neighboring edits.

Endogenous context (native genomic context). An endogenous site is the target's native genomic location inside the cell. Endogenous context refers to local genomic and cellular features around that site, such as chromatin accessibility, transcriptional activity, DNA methylation, histone marks, and protein occupancy. These features can affect whether an editor can access and edit the target. In our notation, c_i^{endo} denotes these endogenous-context features.

Source-library context (controlled library context). A source-library assay measures many target sequences in an artificial, integrated, plasmid, or otherwise controlled setting. Such assays are scalable and useful for learning sequence-level editing rules, but they may not reproduce the endogenous genomic context of each target. In this paper, $\boldsymbol{y}_i^{\mathrm{src}}$ denotes the editing outcome measured or predicted in this source-library context, while $\boldsymbol{y}_i^{\mathrm{endo}}$ denotes the outcome at the endogenous site.

B Benchmarks and Data Processing

BE_Endo endogenous benchmark. Our main endogenous endpoint benchmark is BE_Endo (Yuan et al., 2024), which instantiates

$$\mathcal{D}^{\mathrm{endo}} = \{(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}, \boldsymbol{y}_i^{\mathrm{endo}})\}_{i=1}^{N_{\mathrm{endo}}}. \tag{13}$$

Here $\boldsymbol{y}_i^{\mathrm{endo}}$ is an experimentally measured endogenous 20-position editing-rate vector, and c_i^{endo} denotes endogenous-context features at the native locus (Section 2.1). After curation, $N_{\mathrm{endo}} = 9{,}116$ (4,529 ABE; 4,587 CBE). For the main reported results, we use five repeated grouped splits, where each seed defines both the (editor, protospacer, PAM)-level train/validation/test split and the model initialization. Each split contains 6,381/1,367/1,368 train/validation/test records.

Endogenous-context features. We use the nine locus-level covariates released with the HEK293T BE_Endo records: gene expression, DNA methylation, DNase accessibility, H3K27ac, H3K4me3, RNA polymerase II (Pol II), H3K4me1, CTCF, and H3K36me3. Their construction and genomic-overlap processing follow BE_Endo; we retain the released numerical values without recomputing or substituting genomic tracks.

Sequence-overlap audit. For each test record and split, we computed the minimum Hamming distance to every same-editor training record. Across 6,840 test assignments, neither the 20-nt protospacer nor the 40-nt local context had an exact training match. Only 18 protospacers (0.26%) and nine 40-nt contexts (0.13%) were within two substitutions of a training sequence; the median nearest distances were seven and 17 substitutions, respectively.

Integrated-library records for training q_ϕ. For each endpoint split, we align the integrated-library source records to the same grouped train/validation/test assignment as their endogenous targets. We fit the source-outcome predictor q_ϕ only on measured source outcomes from the training fold,

$$\mathcal{D}^{\mathrm{src}} = \{(e_i, g_i, p_i, s_i, \boldsymbol{y}_i^{\mathrm{src}}) : i \in \mathcal{I}_{\mathrm{train}}\}, \tag{14}$$

using the source-validation fold for model selection. Across the five splits, the source train/validation folds contain 5,904–5,933 and 1,261–1,270 records, respectively. We then obtain an imputed source prior for every endpoint record,

$$\tilde{\boldsymbol{y}}_i^{\mathrm{src}} = q_\phi(e_i, g_i, p_i, s_i). \tag{15}$$

Experimental source outcomes assigned to the endpoint test fold are hidden from both source-model fitting and downstream endpoint prediction.

Source-prior augmentation for endpoint prediction. Attaching $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ yields the augmented endpoint-prediction dataset

$$\tilde{\mathcal{D}}^{\mathrm{aug}} = \{(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}, \tilde{\boldsymbol{y}}_i^{\mathrm{src}}, \boldsymbol{y}_i^{\mathrm{endo}})\}_{i=1}^{N_{\mathrm{endo}}}. \tag{16}$$

All methods are trained and evaluated on the same endogenous target $\boldsymbol{y}_i^{\mathrm{endo}}$; $\tilde{\boldsymbol{y}}_i^{\mathrm{src}}$ is an additional noisy source-prior feature used only by the source-prior augmented models.

Matched source-to-endogenous transfer set. We reconstruct source outcomes from BE_Endo's paired integrated-library experiments and match them to endogenous records with the same concrete editor construct and exact 40-nt local target context. This gives

$$\mathcal{D}^{\mathrm{match}} = \{(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}, \boldsymbol{y}_i^{\mathrm{src}}, \boldsymbol{y}_i^{\mathrm{endo}})\}_{i=1}^{8,450}. \tag{17}$$

The pairs comprise 4,207 ABEmaxF148A and 4,243 YE1-BE3-FNLS records. Following the BE_Endo integrated-library assay design, we trim the redundant barcode, merge clean paired reads, and globally align each read to its 170-bp target amplicon. We compute the intended A-to-G or C-to-T marginal rate at every protospacer position and average two biological replicates after requiring more than 100 target-aligned reads in each replicate. Replicate profiles are highly reproducible (window-position Pearson $r = 0.985$ for ABE and 0.988 for CBE). Five editor–protospacer–PAM-grouped 70/15/15 splits define paired training, validation, and test sets. All methods use the same split for a seed; source-conditioned models receive measured $\boldsymbol{y}_i^{\mathrm{src}}$, while every method predicts measured $\boldsymbol{y}_i^{\mathrm{endo}}$.

C BE-CORR Implementation Details

This appendix gives implementation-level details and pseudocode for the training and inference phases of BE-CORR. Endpoint and matched-transfer experiments fit and select models on their respective training and validation partitions.

Hazard-mass interpretation. The source-to-hazard mapping in Eq. 5 converts a bounded per-position probability profile into a nonnegative *editing-opportunity* profile. In this hazard space, each coordinate $\mu_{ij} \geq 0$ represents the amount of editable opportunity assigned to protospacer position j, and the total hazard mass $\sum_j \mu_{ij}$ summarizes the overall amount of source-derived editing opportunity for the target. The editable mask $\mathbf{m}_i = \mathbf{m}(e_i, s_i)$ is applied before any correction so that chemically non-editable positions start with exactly zero hazard and remain excluded from subsequent transport. Numerically, we use $\mathcal{H}(y) = -\log(1-\mathrm{clip}(y, \epsilon, 1-\epsilon))$ and $\mathcal{H}^{-1}(\lambda) = 1-\exp(-\lambda)$. The context encoder returns $(\tilde{a}_i, \mathbf{b}_i, \mathbf{z}_i^{\mathrm{rel}})$, and we bound its activity output as $a_i = a_{\max}\tanh(\tilde{a}_i/a_{\max})$ before computing $\bar{\boldsymbol{\mu}}_i = \exp(a_i)\boldsymbol{\mu}_i$.

Exposure-potential features. The exposure potential $\mathbf{b}_i$ is constructed from compact position-wise features that combine (i) local sequence information, (ii) editor-window and editability indicators, (iii) endogenous-context features at the native locus, and (iv) source-profile summary signals. Concretely, these include the protospacer coordinate, membership in the canonical editor window W_{e_i}, editable-base indicators, local bases and PAM features, endogenous tracks, and the predicted source values at each position. This design allows $\mathbf{b}_i$ to favor exposed editable positions without enabling arbitrary long-range effects. Under 1-based protospacer indexing, we use $W_e = \{5, \ldots, 8\}$ for ABE and $W_e = \{3, \ldots, 9\}$ for CBE.

Transport-difficulty summary. For reliability estimation, we summarize the structural difficulty of a transport plan by its source-mass-weighted local movement cost:

$$\mathrm{cost}(\Pi_i) = \frac{\sum_{j=1}^{20} \bar{\mu}_{ij} \sum_{k \in \mathcal{A}_i} \Pi_{i,jk} \frac{(j-k)^2}{2\ell_{e_i}^2}}{\sum_{j=1}^{20} \bar{\mu}_{ij} + \epsilon}. \tag{18}$$

Reliability features, disagreement attenuation, and floor. The reliability network asks whether transport was structurally costly, how strongly the transported hazard disagrees with the direct endogenous hazard, what source/context quality is encoded by $\mathbf{z}_i^{\mathrm{rel}}$, and how much source-derived hazard mass is available. Define $G_i = \|\boldsymbol{\nu}_i^{\mathrm{tr}} - \boldsymbol{\nu}_i^{\mathrm{dir}}\|_1$. Its full parameterization is

$$\begin{aligned} \rho_i &= \rho_{\min} + (1 - \rho_{\min})\sigma\left(r_\omega[\mathrm{cost}(\Pi_i), G_i, \mathbf{z}_i^{\mathrm{rel}}, \sum_j \mu_{ij}] \right), \\ \eta_i &= \alpha^\star \rho_i \exp(-G_i/\kappa), \qquad \kappa = 1. \end{aligned} \tag{19}$$

Here $\sigma(\cdot)$ is the logistic sigmoid. The lower bound $\rho_{\min}$ prevents the learned score from collapsing to a direct-only solution early in training, whereas the exponential factor can still drive the effective weight η_i to zero when the branches strongly disagree. During network training we set $\alpha = 1$; after convergence, $\alpha^\star$ is selected on the validation split and used at inference.

Training objective. We optimize a window-weighted regression loss on the endogenous 20-position vector with locality, smoothness, and reliability regularization:

$$\mathcal{L}_{\mathrm{endo}} = \frac{1}{20N} \sum_{i=1}^{N} \sum_{j=1}^{20} w_{ij} \left(\hat{y}_{ij}^{\mathrm{endo}} - y_{ij}^{\mathrm{endo}} \right)^2, \qquad w_{ij} = 1 + \lambda_{\mathrm{W}} \mathbf{1}\{j \in W_{e_i}\}. \tag{20}$$

$$\begin{aligned} \mathcal{R}_{\mathrm{local}} &= \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{20} \sum_{k=1}^{20} \bar{\mu}_{ij} \Pi_{i,jk} |j-k|}{\sum_{j=1}^{20} \bar{\mu}_{ij} + \epsilon}, \\ \mathcal{R}_{\mathrm{smooth}} &= \frac{1}{19N} \sum_{i=1}^{N} \sum_{j=1}^{19} (b_{i,j+1} - b_{ij})^2, \qquad \mathcal{R}_\rho = \frac{1}{N} \sum_{i=1}^{N} (\rho_i - \rho_0)^2, \\ \mathcal{L} &= \mathcal{L}_{\mathrm{endo}} + \lambda_{\mathrm{local}} \mathcal{R}_{\mathrm{local}} + \lambda_{\mathrm{smooth}} \mathcal{R}_{\mathrm{smooth}} + \lambda_\rho \mathcal{R}_\rho. \end{aligned} \tag{21}$$

Here $\epsilon > 0$ is a numerical stabilizer.

Algorithm 1 BE-CORR Training

Input: Source records $\mathcal{D}^{\mathrm{src}}$; endogenous records $\mathcal{D}^{\mathrm{endo}}$; editable masks $\mathbf{m}(e, s)$; windows W_e; transport temperature τ_{ot}; fixed disagreement scale $\kappa = 1$; regularization weights
Output: Source predictor q_ϕ, BE-CORR parameters (θ, ψ, ω), and validation-selected shrinkage $\alpha^\star$
Fit $q_\phi(e, g, p, s) \to \tilde{\boldsymbol{y}}^{\mathrm{src}}$ on $\mathcal{D}^{\mathrm{src}}$ using source-library outcomes
Independently fit $f_\psi(e, g, p, s, c^{\mathrm{endo}}) \to \hat{\boldsymbol{y}}^{\mathrm{dir}}$ on $\mathcal{D}^{\mathrm{endo}}$
for each endogenous record i **do**
    Attach predicted source prior $\tilde{\boldsymbol{y}}_i^{\mathrm{src}} \leftarrow q_\phi(e_i, g_i, p_i, s_i)$
    Construct editable mask $\mathbf{m}_i \leftarrow \mathbf{m}(e_i, s_i)$ and admissible set $\mathcal{A}_i \leftarrow \{k : m_{ik} = 1\}$
    Convert source probability to hazard $\boldsymbol{\mu}_i \leftarrow \mathbf{m}_i \odot \mathcal{H}(\tilde{\boldsymbol{y}}_i^{\mathrm{src}})$
end for
while validation loss has not converged **do**
    Sample a minibatch $\mathcal{B} \subset \mathcal{D}^{\mathrm{endo}}$
    for each $i \in \mathcal{B}$ **do**
        Encode correction fields: $(\tilde{a}_i, \mathbf{b}_i, \mathbf{z}_i^{\mathrm{rel}}) \leftarrow F_\theta(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}, \tilde{\boldsymbol{y}}_i^{\mathrm{src}})$
        Compute bounded activity $a_i \leftarrow a_{\max} \tanh(\tilde{a}_i / a_{\max})$ and scaled source hazard $\bar{\boldsymbol{\mu}}_i \leftarrow \exp(a_i)\boldsymbol{\mu}_i$
        Obtain the masked direct profile $\hat{\boldsymbol{y}}_i^{\mathrm{dir}} \leftarrow \mathbf{m}_i \odot f_\psi(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}})$ and hazard $\boldsymbol{\nu}_i^{\mathrm{dir}} \leftarrow \mathbf{m}_i \odot \mathcal{H}(\hat{\boldsymbol{y}}_i^{\mathrm{dir}})$
        for $j = 1, \ldots, 20$ **do**
            for each $k \in \mathcal{A}_i$ **do**
                Build editable-support cost $C_{i,jk}$ from distance, window membership, and exposure potential $\mathbf{b}_i$
            end for
            Compute row transport $\Pi_{i,jk} \leftarrow \exp(-C_{i,jk}/\tau_{\mathrm{ot}}) / \sum_{u \in \mathcal{A}_i} \exp(-C_{i,ju}/\tau_{\mathrm{ot}})$ for $k \in \mathcal{A}_i$, and $\Pi_{i,jk} \leftarrow 0$ otherwise
        end for
        Transport source hazard: $\boldsymbol{\nu}_i^{\mathrm{tr}} \leftarrow \sum_j \bar{\mu}_{ij} \Pi_{i,j:}$
        Compute hazard disagreement $G_i \leftarrow \|\boldsymbol{\nu}_i^{\mathrm{tr}} - \boldsymbol{\nu}_i^{\mathrm{dir}}\|_1$ and reliability features $[\mathrm{cost}(\Pi_i), G_i, \mathbf{z}_i^{\mathrm{rel}}, \sum_j \mu_{ij}]$
        Predict reliability $\rho_i \leftarrow \rho_{\min} + (1 - \rho_{\min})\sigma(r_\omega[\mathrm{cost}(\Pi_i), G_i, \ldots, \sum_j \mu_{ij}])$
        Set training-time source weight $\eta_i \leftarrow \rho_i \exp(-G_i/\kappa)$
        Blend transported and direct hazards: $\hat{\boldsymbol{\nu}}_i^{\mathrm{endo}} \leftarrow \eta_i \boldsymbol{\nu}_i^{\mathrm{tr}} + (1 - \eta_i)\boldsymbol{\nu}_i^{\mathrm{dir}}$
        Convert to editing rates: $\hat{\boldsymbol{y}}_i^{\mathrm{endo}} \leftarrow \mathbf{m}_i \odot (1 - \exp(-\hat{\boldsymbol{\nu}}_i^{\mathrm{endo}}))$
    end for
    Optimize the correction parameters in Eq. 21 on endogenous labels $\{\boldsymbol{y}_i^{\mathrm{endo}} : i \in \mathcal{B}\}$
end while
Select early stopping and $\alpha^\star$ on the validation split

Algorithm 2 BE-CORR Inference

Input: Trained q_ϕ and BE-CORR parameters; test record $(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}})$; optional measured source outcome $\boldsymbol{y}_i^{\mathrm{src}}$
Output: Endogenous prediction $\hat{\boldsymbol{y}}_i^{\mathrm{endo}}$, shift $\hat{\Delta}_i$, and interpretable summaries $(a_i, \mathbf{b}_i, \Pi_i, \rho_i, \eta_i)$
if measured source outcome $\boldsymbol{y}_i^{\mathrm{src}}$ is available **then**
    Set source input $\tilde{\boldsymbol{y}}_i^{\mathrm{src}} \leftarrow \boldsymbol{y}_i^{\mathrm{src}}$
else
    Predict source prior $\tilde{\boldsymbol{y}}_i^{\mathrm{src}} \leftarrow q_\phi(e_i, g_i, p_i, s_i)$
end if
Construct $\mathbf{m}_i$, $\mathcal{A}_i$, and $\boldsymbol{\mu}_i \leftarrow \mathbf{m}_i \odot \mathcal{H}(\tilde{\boldsymbol{y}}_i^{\mathrm{src}})$
Compute $(\tilde{a}_i, \mathbf{b}_i, \mathbf{z}_i^{\mathrm{rel}})$ with F_θ, set $a_i \leftarrow a_{\max} \tanh(\tilde{a}_i / a_{\max})$, and compute the masked direct hazard $\boldsymbol{\nu}_i^{\mathrm{dir}} \leftarrow \mathbf{m}_i \odot \mathcal{H}(\mathbf{m}_i \odot f_\psi(e_i, g_i, p_i, s_i, c_i^{\mathrm{endo}}))$
Scale source hazard by a_i, build Π_i on editable positions, and compute $\boldsymbol{\nu}_i^{\mathrm{tr}}$ and $\mathrm{cost}(\Pi_i)$
Compute $G_i \leftarrow \|\boldsymbol{\nu}_i^{\mathrm{tr}} - \boldsymbol{\nu}_i^{\mathrm{dir}}\|_1$, predict ρ_i with r_ω, and set $\eta_i \leftarrow \alpha^\star \rho_i \exp(-G_i/\kappa)$
Blend $\boldsymbol{\nu}_i^{\mathrm{tr}}$ and $\boldsymbol{\nu}_i^{\mathrm{dir}}$ using η_i, then convert the resulting hazard to $\hat{\boldsymbol{y}}_i^{\mathrm{endo}}$
Compute shift estimate $\hat{\Delta}_i \leftarrow \hat{\boldsymbol{y}}_i^{\mathrm{endo}} - \tilde{\boldsymbol{y}}_i^{\mathrm{src}}$
return $\hat{\boldsymbol{y}}_i^{\mathrm{endo}}$, $\hat{\Delta}_i$, $(a_i, \mathbf{b}_i, \Pi_i, \rho_i, \eta_i)$

D HYPERPARAMETERS AND VALIDATION SELECTION

All tuned hyperparameters are selected using the validation partition of each split. Table 5 reports the selected values and search ranges.

Table 5: **BE-CORR hyperparameters and validation search ranges.** Tuned values are selected by validation performance within each split.

Hyperparameter	Selected value	Validation search range	Role
Window loss weight λ_{W}	5	$\{3, 5, 7\}$	Upweights canonical editing-window positions in $\mathcal{L}_{\mathrm{endo}}$
Activity bound $a_{\max}$	3.2	$\{2.5, 3.0, 3.2, 3.5\}$	Bounds global endogenous activity scaling
Transport temperature τ_{ot}	0.42	$\{0.35, 0.42, 0.50, 0.55\}$	Controls sharpness of row-constrained transport
Window penalty λ_{win}	3.2	$\{2.0, 3.0, 3.2, 4.0\}$	Penalizes transport outside editor activity windows
Reliability floor $\rho_{\min}$	0.02	$\{0, 0.02, 0.05\}$	Prevents complete collapse of the source-transport branch
Global shrinkage $\alpha^\star$	0.25	$\{0, 0.25, 0.5, 1.0\}$	Global shrinkage on ρ_i; includes strict direct-prediction fallback
Disagreement scale κ	1.0	Fixed (not tuned)	Attenuates source use when transported and direct hazards disagree
Locality penalty λ_{local}	5×10^{-3}	Fixed (not tuned)	Penalizes long-distance hazard reallocation
Exposure smoothness $\lambda_{\mathrm{smooth}}$	5×10^{-3}	Fixed (not tuned)	Regularizes adjacent exposure potentials
Reliability penalty λ_ρ	5×10^{-4}	Fixed (not tuned)	Prevents degenerate always-source or always-direct reliability
Learning rate	5×10^{-4}	Fixed (not tuned)	AdamW optimizer learning rate

E PROOFS

E.1 PROOF OF PROPOSITION 1

Proposition 1 (Prediction-error bound, source-noise robustness, and source utility)**.** *Suppress the instance index and let $\boldsymbol{\nu}^\star$ be the true endogenous hazard, with $\boldsymbol{y}^\star = \mathbf{m} \odot \mathcal{H}^{-1}(\boldsymbol{\nu}^\star)$. Let $\boldsymbol{\mu}^\star, a^\star, C^\star$ denote ideal source hazard, activity shift, and transport cost on a common nonempty editable support, and define $\boldsymbol{\nu}_0^\star = e^{a^\star}\boldsymbol{\mu}^\star\Pi(C^\star)$. BE-CORR instead uses $\tilde{\boldsymbol{\mu}}, a, C$ and produces $\boldsymbol{\nu}^{\mathrm{tr}} = e^a\tilde{\boldsymbol{\mu}}\Pi(C)$. Assume $\|\boldsymbol{\mu}^\star\|_1 \le M$ and $\max\{|a|, |a^\star|\} \le A$, and define*

$$\epsilon_{\mathrm{str}} = \|\boldsymbol{\nu}^\star - \boldsymbol{\nu}_0^\star\|_1, \quad \delta_\mu = \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\|_1, \quad \delta_a = |a - a^\star|, \quad \delta_C = \max_{j\in\{1,\dots,20\},\, k\in\mathcal{A}} |C_{jk} - C^\star_{jk}|.$$

Then

$$\|\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}^\star\|_1 \le B_{\mathrm{tr}} := \epsilon_{\mathrm{str}} + e^A[\delta_\mu + M\delta_a + 2M\tanh(\delta_C/\tau_{\mathrm{ot}})]\,.$$

Let $\epsilon_{\mathrm{dir}} = \|\boldsymbol{\nu}^{\mathrm{dir}} - \boldsymbol{\nu}^\star\|_1$, $G = \|\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}^{\mathrm{dir}}\|_1$, and $\eta = \alpha^\star\rho e^{-G/\kappa}$ for $\alpha^\star, \rho \in [0, 1]$ and $\kappa > 0$. Then

$$\|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^\star\|_1 \le (1 - \eta)\epsilon_{\mathrm{dir}} + \eta B_{\mathrm{tr}}.$$

Moreover, for $\hat{\boldsymbol{y}}^{\mathrm{dir}} = \mathbf{m} \odot \mathcal{H}^{-1}(\boldsymbol{\nu}^{\mathrm{dir}})$,

$$\|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \hat{\boldsymbol{y}}^{\mathrm{dir}}\|_1 \le \alpha^\star\rho G e^{-G/\kappa} \le \frac{\alpha^\star\kappa}{e}, \qquad \|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^\star\|_1 - \|\hat{\boldsymbol{y}}^{\mathrm{dir}} - \boldsymbol{y}^\star\|_1 \le \frac{\alpha^\star\kappa}{e}.$$

The final prediction converges to the direct prediction as $G \to \infty$. For squared hazard error, let $\mathbf{e}_{\mathrm{dir}} = \boldsymbol{\nu}^{\mathrm{dir}} - \boldsymbol{\nu}^{\star}$ and $\mathbf{e}_{\mathrm{tr}} = \boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}^{\star}$. Then

$$\|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^{\star}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2 = \eta\big(\|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2\big) - \eta(1-\eta)\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2.$$

For $\eta > 0$, the transported source improves squared hazard error over the direct branch if and only if

$$(1-\eta)\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2 > \|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2.$$

If $\mathbf{e}_{\mathrm{dir}} \neq \mathbf{0}$, $\langle \mathbf{e}_{\mathrm{dir}}, \mathbf{e}_{\mathrm{tr}} \rangle \leq 0$, and $0 < \eta < \min\{1, 2\|\mathbf{e}_{\mathrm{dir}}\|_2^2 / \|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2\}$, then

$$\|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^{\star}\|_2^2 < \|\mathbf{e}_{\mathrm{dir}}\|_2^2,$$

even though the condition permits $\|\mathbf{e}_{\mathrm{tr}}\|_2 > \|\mathbf{e}_{\mathrm{dir}}\|_2$.

Proof. Write $\Pi = \Pi(C)$ and $\Pi^{\star} = \Pi(C^{\star})$. Both plans are row-stochastic on the same editable support. Hence, for every signed row vector $\mathbf{x}$,

$$\|\mathbf{x}\Pi\|_1 \leq \sum_{j,k} |x_j| \Pi_{jk} = \sum_j |x_j| = \|\mathbf{x}\|_1. \tag{22}$$

We first control how a cost error changes the Gibbs transport plan. Fix row j, and write

$$w_k = e^{-C_{jk}/\tau_{\mathrm{ot}}}, \quad w_k^{\star} = e^{-C_{jk}^{\star}/\tau_{\mathrm{ot}}}, \quad Z = \sum_{u \in \mathcal{A}} w_u, \quad Z^{\star} = \sum_{u \in \mathcal{A}} w_u^{\star}.$$

The uniform cost bound gives

$$e^{-\delta_C/\tau_{\mathrm{ot}}} \leq \frac{w_k}{w_k^{\star}} \leq e^{\delta_C/\tau_{\mathrm{ot}}} \quad \text{for every } k.$$

Summing the inequalities over the common support yields the same bounds for $Z/Z^{\star}$. Thus, for $p_k = w_k/Z$ and $q_k = w_k^{\star}/Z^{\star}$,

$$e^{-2\delta_C/\tau_{\mathrm{ot}}} \leq \frac{p_k}{q_k} \leq e^{2\delta_C/\tau_{\mathrm{ot}}} \quad \text{for every } k \in \mathcal{A}.$$

To make the resulting ℓ_1 bound explicit, let $L = e^{2\delta_C/\tau_{\mathrm{ot}}}$. From $L^{-1} \leq p_k/q_k \leq L$, one obtains $(L+1)|p_k - q_k| \leq (L-1)(p_k + q_k)$, regardless of whether $p_k \geq q_k$ or $p_k < q_k$. Summing over k and using $\sum_k p_k = \sum_k q_k = 1$ gives

$$\|\Pi_{j\cdot} - \Pi_{j\cdot}^{\star}\|_1 \leq 2\frac{L-1}{L+1} = 2\tanh(\delta_C/\tau_{\mathrm{ot}}). \tag{23}$$

Add and subtract intermediate terms to decompose the transported-hazard error:

$$\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}_0^{\star} = e^a(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\star})\Pi + (e^a - e^{a^{\star}})\boldsymbol{\mu}^{\star}\Pi + e^{a^{\star}}\boldsymbol{\mu}^{\star}(\Pi - \Pi^{\star}). \tag{24}$$

For the first term, Eq. 22 and $e^a \leq e^A$ give

$$\|e^a(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\star})\Pi\|_1 \leq e^A \delta_{\mu}.$$

For the second, the mean-value theorem gives $|e^a - e^{a^{\star}}| = e^{\xi}|a - a^{\star}| \leq e^A \delta_a$ for some ξ between a and $a^{\star}$. Applying Eq. 22 again,

$$\|(e^a - e^{a^{\star}})\boldsymbol{\mu}^{\star}\Pi\|_1 \leq e^A M \delta_a.$$

For the third term, nonnegativity of $\boldsymbol{\mu}^{\star}$ and Eq. 23 imply

$$\|e^{a^{\star}}\boldsymbol{\mu}^{\star}(\Pi - \Pi^{\star})\|_1 \leq e^A \sum_j \mu_j^{\star} \|\Pi_{j\cdot} - \Pi_{j\cdot}^{\star}\|_1 \leq 2e^A M \tanh(\delta_C/\tau_{\mathrm{ot}}).$$

Combining these three inequalities with the triangle inequality gives

$$\|\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}_0^\star\|_1 \le e^A[\delta_\mu + M\delta_a + 2M\tanh(\delta_C/\tau_{\mathrm{ot}})]\,.$$

Adding $\|\boldsymbol{\nu}_0^\star - \boldsymbol{\nu}^\star\|_1 = \epsilon_{\mathrm{str}}$ proves the first line of Eq. 10.

The final hazard satisfies

$$\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star = (1-\eta)(\boldsymbol{\nu}^{\mathrm{dir}} - \boldsymbol{\nu}^\star) + \eta(\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}^\star).$$

Because $0 \le \eta \le 1$, the triangle inequality gives the corresponding convex combination explicitly as

$$\|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star\|_1 \le (1-\eta)\epsilon_{\mathrm{dir}} + \eta\|\boldsymbol{\nu}^{\mathrm{tr}} - \boldsymbol{\nu}^\star\|_1 \le (1-\eta)\epsilon_{\mathrm{dir}} + \eta B_{\mathrm{tr}}.$$

Moreover, $\mathcal{H}^{-1}(x) = 1 - e^{-x}$ has derivative in $[0,1]$ for $x \ge 0$ and is coordinatewise 1-Lipschitz. Therefore summing the coordinatewise inequalities shows that $\mathcal{H}^{-1}$ is nonexpansive in ℓ_1; applying the binary editable mask cannot increase this norm either. Hence

$$\|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^\star\|_1 \le \|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star\|_1,$$

which proves the second line of Eq. 10.

Finally,

$$\|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \hat{\boldsymbol{y}}^{\mathrm{dir}}\|_1 \le \|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^{\mathrm{dir}}\|_1 = \eta G = \alpha^\star \rho G e^{-G/\kappa}.$$

The scalar function $f(G) = Ge^{-G/\kappa}$ has derivative $f'(G) = e^{-G/\kappa}(1 - G/\kappa)$, so its maximum on $G \ge 0$ is $f(\kappa) = \kappa/e$. Since $\rho \le 1$, this proves the first inequality in Eq. 11. The second follows from

$$\|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^\star\|_1 \le \|\hat{\boldsymbol{y}}^{\mathrm{dir}} - \boldsymbol{y}^\star\|_1 + \|\hat{\boldsymbol{y}}^{\mathrm{endo}} - \hat{\boldsymbol{y}}^{\mathrm{dir}}\|_1.$$

Because $Ge^{-G/\kappa} \to 0$ as $G \to \infty$, the final prediction converges to the direct prediction under severe branch disagreement.

For the source-utility statement, the hazard-space interpolation gives

$$\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star = (1-\eta)\mathbf{e}_{\mathrm{dir}} + \eta\mathbf{e}_{\mathrm{tr}}.$$

Expanding the squared norm gives

$$\begin{aligned}\|(1-\eta)\mathbf{e}_{\mathrm{dir}} + \eta\mathbf{e}_{\mathrm{tr}}\|_2^2 = (1-\eta)^2\|\mathbf{e}_{\mathrm{dir}}\|_2^2 + \eta^2\|\mathbf{e}_{\mathrm{tr}}\|_2^2 \\ + 2\eta(1-\eta)\langle\mathbf{e}_{\mathrm{dir}}, \mathbf{e}_{\mathrm{tr}}\rangle.\end{aligned}$$

Using $2\langle\mathbf{e}_{\mathrm{dir}}, \mathbf{e}_{\mathrm{tr}}\rangle = \|\mathbf{e}_{\mathrm{dir}}\|_2^2 + \|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2$ and collecting the coefficients of the two squared errors yields Eq. 12. For $\eta > 0$, its right-hand side is negative exactly when

$$(1-\eta)\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2 > \|\mathbf{e}_{\mathrm{tr}}\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2,$$

which proves the stated if-and-only-if criterion. An equivalent expansion, useful for proving the sufficient condition, is

$$\begin{aligned}\Delta_\eta &= \|\hat{\boldsymbol{\nu}}^{\mathrm{endo}} - \boldsymbol{\nu}^\star\|_2^2 - \|\mathbf{e}_{\mathrm{dir}}\|_2^2 \\ &= 2\eta\big(\langle\mathbf{e}_{\mathrm{dir}}, \mathbf{e}_{\mathrm{tr}}\rangle - \|\mathbf{e}_{\mathrm{dir}}\|_2^2\big) + \eta^2\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2.\end{aligned}$$

When $\langle\mathbf{e}_{\mathrm{dir}}, \mathbf{e}_{\mathrm{tr}}\rangle \le 0$,

$$\Delta_\eta \le -2\eta\|\mathbf{e}_{\mathrm{dir}}\|_2^2 + \eta^2\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2 < 0$$

under the stated upper bound on η, proving strict improvement. This result does not require the transported branch to be more accurate by itself: taking $\mathbf{e}_{\mathrm{tr}} = -c\mathbf{e}_{\mathrm{dir}}$ for any $c > 1$ gives $\|\mathbf{e}_{\mathrm{tr}}\|_2 > \|\mathbf{e}_{\mathrm{dir}}\|_2$ while satisfying the alignment condition. Here $\|\mathbf{e}_{\mathrm{tr}} - \mathbf{e}_{\mathrm{dir}}\|_2^2 = (c+1)^2\|\mathbf{e}_{\mathrm{dir}}\|_2^2$, so every $0 < \eta < 2/(c+1)^2$ satisfies the sufficient condition and improves the mixture. □

F Evaluation Protocols

F.1 Evaluation Scope

We evaluate two related but distinct questions. The first is endogenous endpoint prediction: given the inputs available to a method, predict the endogenous 20-position editing-rate vector. This is the task evaluated on the main BE_Endo benchmark. Standard baselines use sequence-only or sequence-plus-endogenous-context inputs, while source-prior augmented models additionally receive an imputed source-library prior $\tilde{\boldsymbol{y}}^{\mathrm{src}}$; no method receives an experimental source profile in this regime. The main benchmark therefore does not directly evaluate true source-to-endogenous shift recovery.

The second question is source-to-endogenous transfer: given a measured source-library outcome for the same editor construct and 40-nt target context, predict how that profile changes at the native endogenous locus. This task uses the paired benchmark, where both $\boldsymbol{y}_i^{\mathrm{src}}$ and $\boldsymbol{y}_i^{\mathrm{endo}}$ are measured and models are trained, selected, and tested on grouped paired partitions. Thus the endpoint benchmark tests prediction with an imputed prior, whereas the paired benchmark directly tests learning the measured context shift.

F.2 Common Notation and Editing Summaries

For each evaluation record $i \in \{1, \ldots, N\}$, the ground-truth endogenous outcome is a 20-position vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i20}) \in [0,1]^{20}$ and the model prediction is $\hat{\boldsymbol{y}}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{i20})$. Predictions are clipped to $[0,1]$. Let W_i be the canonical editing window for the editor of record i: positions 5–8 for ABE and positions 3–9 for CBE under 1-based protospacer indexing. Let $M = \sum_i |W_i|$ be the number of evaluated window positions.

Scalar editing summaries. For any 20-position vector $\boldsymbol{y}_i$, we define three scalar summaries inside the editing window:

$$T_i(\boldsymbol{y}) = \sum_{j \in W_i} y_{ij}, \qquad D_i(\boldsymbol{y}) = \max_{j \in W_i} y_{ij}, \qquad B_i(\boldsymbol{y}) = T_i(\boldsymbol{y}) - D_i(\boldsymbol{y}). \tag{25}$$

Here T_i is the window marginal-edit burden, D_i is the dominant-position editing rate, and B_i is the non-dominant marginal-edit burden. The dominant position is defined by the largest rate in the window, not by an externally specified intended edit position. The same definitions apply to predictions, yielding $\hat{T}_i = T_i(\hat{\boldsymbol{y}})$, $\hat{D}_i = D_i(\hat{\boldsymbol{y}})$, and $\hat{B}_i = B_i(\hat{\boldsymbol{y}})$.

F.3 Main Endogenous Benchmark Metrics

The main BE_Endo benchmark evaluates endpoint endogenous prediction. Since measured source outcomes are unavailable for these records, all metrics in this section compare predicted and measured endogenous outcomes. Source-prior augmented models may use $\tilde{\boldsymbol{y}}^{\mathrm{src}}$ as an input, but no true source-to-endogenous shift is computed on this benchmark.

Position-level errors. Window MAE measures per-position error only at positions where the editor is expected to act:

$$\mathrm{MAE}_W = \frac{1}{M} \sum_{i=1}^{N} \sum_{j \in W_i} |\hat{y}_{ij} - y_{ij}|. \tag{26}$$

Full-vector MAE averages over all 20 protospacer positions:

$$\mathrm{MAE}_{20} = \frac{1}{20N} \sum_{i=1}^{N} \sum_{j=1}^{20} |\hat{y}_{ij} - y_{ij}|. \tag{27}$$

Lower values indicate more accurate marginal editing-rate predictions.

Scalar editing errors. Burden, dominant, and non-dominant MAE compare the scalar summaries above:

$$\mathrm{MAE}_{\mathrm{burden}} = \frac{1}{N}\sum_{i=1}^{N}|\hat{T}_i - T_i|, \qquad \mathrm{MAE}_{\mathrm{dom}} = \frac{1}{N}\sum_{i=1}^{N}|\hat{D}_i - D_i|, \tag{28}$$

$$\mathrm{MAE}_{\mathrm{non-dom}} = \frac{1}{N}\sum_{i=1}^{N}|\hat{B}_i - B_i|. \tag{29}$$

These metrics distinguish aggregate window burden from dominant and non-dominant marginal editing.

Correlation and ranking metrics. Window Pearson correlation is the Pearson correlation between flattened predicted and true editing rates over all window positions:

$$r_W = \mathrm{Pearson}\big(\{\hat{y}_{ij} : j \in W_i\}_{i=1}^{N}, \{y_{ij} : j \in W_i\}_{i=1}^{N}\big). \tag{30}$$

Burden Spearman correlation ranks predicted and true window burden:

$$\rho_T = \mathrm{Spearman}\Big(\{\hat{T}_i\}_{i=1}^{N}, \{T_i\}_{i=1}^{N}\Big). \tag{31}$$

For high-burden AUPRC, we compute editor-specific thresholds from the training split: q_e is the 80th percentile of training-set window burden for editor e. Test label $\ell_i = \mathbf{1}\{T_i \geq q_{e_i}\}$ marks high-burden targets, and the score is $\hat{T}_i$. Sorting records by score in descending order gives indices $\pi_1, \ldots, \pi_N$. Average precision is

$$\mathrm{AP} = \frac{1}{\sum_i \ell_i}\sum_{r=1}^{N}\left(\frac{\sum_{u=1}^{r}\ell_{\pi_u}}{r}\right)\ell_{\pi_r}, \tag{32}$$

which is the area under the precision–recall curve for high-burden target discovery. NDCG@10% ranks records by $\hat{T}_i$ and uses T_i as graded relevance at $k = \lceil 0.1N \rceil$:

$$\mathrm{NDCG@10\%} = \frac{\mathrm{DCG@}k(\hat{T}; T)}{\mathrm{IDCG@}k(T)}. \tag{33}$$

Here

$$\mathrm{DCG@}k(\hat{T}; T) = \sum_{r=1}^{k}\frac{2^{T_{\pi_r}} - 1}{\log_2(r+1)}, \tag{34}$$

and $\mathrm{IDCG@}k(T)$ is the maximum possible DCG@k obtained by ranking records by the true window burden T_i. Higher values indicate better prioritization of high-burden endogenous targets.

Calibration. We measure calibration of window burden using equal-frequency bins over predicted burden. Let $\{\mathcal{B}_b\}_{b=1}^{10}$ be ten quantile bins after sorting by $\hat{T}_i$. The expected calibration error is

$$\mathrm{ECE} = \sum_{b=1}^{10}\frac{|\mathcal{B}_b|}{N}\left|\frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b}T_i - \frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b}\hat{T}_i\right|. \tag{35}$$

Lower ECE indicates better calibration of predicted to observed window burden.

Average metric-wise rank. For compact table summaries, we also report average rank across the metrics in that table. Let $\mathcal{H}$ be the evaluated methods and $\mathcal{M}$ the reported metrics. For method h and metric m, let R_{hm} be the metric-wise rank after orienting the metric so that rank 1 is best. The average rank is

$$\mathrm{AvgRank}(h) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}R_{hm}. \tag{36}$$

Undefined metrics, such as correlation for a constant shift-control baseline, receive the worst rank $|\mathcal{H}|$. Average rank is descriptive and is not used for model selection.

F.4 Matched Source-to-Endogenous Transfer Metrics

For each paired record, we observe a measured source-library vector $\boldsymbol{y}_i^{\mathrm{src}}$ for the same concrete editor construct and exact 40-nt target context as its endogenous outcome. This benchmark directly evaluates source-to-endogenous transfer. Let $\boldsymbol{y}_i^{\mathrm{endo}}$ be the endogenous vector, $\hat{\boldsymbol{y}}_i^{\mathrm{endo}}$ the predicted endogenous vector, and define the true and predicted source-to-endogenous shifts

$$\Delta_i = \boldsymbol{y}_i^{\mathrm{endo}} - \boldsymbol{y}_i^{\mathrm{src}}, \quad \hat{\Delta}_i = \hat{\boldsymbol{y}}_i^{\mathrm{endo}} - \boldsymbol{y}_i^{\mathrm{src}}. \tag{37}$$

Let $T_i^{\mathrm{src}} = T_i(\boldsymbol{y}^{\mathrm{src}})$, $T_i^{\mathrm{endo}} = T_i(\boldsymbol{y}^{\mathrm{endo}})$, and $\hat{T}_i^{\mathrm{endo}} = T_i(\hat{\boldsymbol{y}}^{\mathrm{endo}})$.

Endpoint endogenous errors. Endo Window MAE, Endo Burden MAE, and Endo Dominant Edit MAE use the same definitions as Appendix F.3, but are computed only on the matched source-endogenous evaluation records. They measure the final endogenous prediction quality independent of how the model uses source-context information.

Shift-space errors and ranking. The window-level shift error is

$$\mathrm{MAE}_{\Delta,W} = \frac{1}{M}\sum_{i=1}^{N}\sum_{j\in W_i} |\hat{\Delta}_{ij} - \Delta_{ij}|. \tag{38}$$

The burden-shift error compares scalar source-to-endogenous changes:

$$\mathrm{MAE}_{\Delta,T} = \frac{1}{N}\sum_{i=1}^{N}\left|(\hat{T}_i^{\mathrm{endo}} - T_i^{\mathrm{src}}) - (T_i^{\mathrm{endo}} - T_i^{\mathrm{src}})\right|. \tag{39}$$

We also report Spearman correlation between predicted and true burden shifts:

$$\rho_{\Delta T} = \mathrm{Spearman}\Big(\{\hat{T}_i^{\mathrm{endo}} - T_i^{\mathrm{src}}\}_{i=1}^{N}, \{T_i^{\mathrm{endo}} - T_i^{\mathrm{src}}\}_{i=1}^{N}\Big). \tag{40}$$

The shift-error metrics are algebraically equal to the corresponding endpoint errors when the same measured source vector is subtracted from prediction and truth, since $|(\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^{\mathrm{src}}) - (\boldsymbol{y}^{\mathrm{endo}} - \boldsymbol{y}^{\mathrm{src}})| = |\hat{\boldsymbol{y}}^{\mathrm{endo}} - \boldsymbol{y}^{\mathrm{endo}}|$. We therefore show these errors once as endpoint columns in Table 2; the formulas here make their shift-space interpretation explicit.

Large-shift discovery and correction gain. Let $d_i = T_i^{\mathrm{endo}} - T_i^{\mathrm{src}}$ be the true burden shift. We mark large-shift records using the top quintile of $|d_i|$ within the matched source-endogenous set:

$$\ell_i^{\Delta} = \mathbf{1}\{|d_i| \geq \mathrm{Quantile}_{0.8}(|d_1|, \ldots, |d_N|)\}. \tag{41}$$

Large-shift AUPRC is average precision, using the AP formula above, for labels ℓ_i^{Δ} and score $|\hat{T}_i^{\mathrm{endo}} - T_i^{\mathrm{src}}|$. Finally, correction gain measures improvement in window MAE over directly using the measured source outcome as the endogenous prediction:

$$\mathrm{Gain} = 1 - \frac{\mathrm{MAE}_W(\hat{\boldsymbol{y}}^{\mathrm{endo}}, \boldsymbol{y}^{\mathrm{endo}})}{\mathrm{MAE}_W(\boldsymbol{y}^{\mathrm{src}}, \boldsymbol{y}^{\mathrm{endo}})}. \tag{42}$$

Positive gain means that the model improves on the source-context measurement; zero means no improvement over source; negative gain means the correction made the endpoint endogenous prediction worse.

G BASELINE DETAILS

Table 6: **Baseline inputs and output harmonization.** "Released" means that the public repository's backbone, feature recipe, or architecture family is reused where possible; only the prediction target is harmonized to BE_Endo's 20-position marginal-rate vector.

Baseline	Input setting	BE_Endo output harmonization
kNN sequence	Editor class, protospacer, PAM	Retrieves nearest training records in one-hot sequence space within editor class and averages their 20-position vectors.
BE-Hive	Released sequence features over the available local target window	Keeps the BE-Hive hand-crafted feature family and tree-regression style; trains a multi-output marginal-rate head because BE_Endo does not provide BE-Hive's native scalar/product labels and supplies a 40-nt instead of 50-nt window.
DeepBE	30-nt local target sequence	Keeps the released DeepBE Conv1D recipe; replaces the task-specific output with a 20-dimensional sigmoid head.
BE-DICT	20-nt protospacer sequence	Uses the released per-base transformer backbone; trains a two-class edited/not-edited objective per protospacer position and uses the edited probability as the marginal-rate prediction.
BEDICT2.0	Protospacer plus PAM	Uses the released transformer backbone with a 20-position marginal-rate head because full edited-product labels are unavailable in BE_Endo.
CRISPRon-BE	30-nt local target sequence plus auxiliary channels	Keeps the released CNN recipe; CRISPRoff energy and Cas9 activity channels are unavailable in BE_Endo and are zeroed, and the output is changed to the 20-position target.
BE_Endo Seq	40-nt local target sequence	Follows the released BE_Endo sequence CNN and predicts the 20-position endogenous vector.
BE_Endo Endo	40-nt sequence plus nine endogenous factors	Follows the released BE_Endo sequence-plus-endogenous-factor CNN.
HyenaDNA	Pretrained DNA representation, editor, endogenous factors, source profile	Fine-tunes the official HyenaDNA backbone and fuses its pooled 40-nt sequence representation with the numerical inputs before a 20-position sigmoid head.
Caduceus	Pretrained reverse-complement-aware DNA representation, editor, endogenous factors, source profile	Fine-tunes the official Caduceus-PH backbone with the same context/source fusion head used for HyenaDNA.
FT-Transformer	Per-base sequence tokens, editor, endogenous factors, source profile	Applies feature tokenization and a Transformer to the complete information-matched input, followed by a 20-position sigmoid head.
DAN	Source and endogenous training records; sequence, editor, endogenous factors	Learns a shared source/endogenous representation with multi-kernel maximum mean discrepancy and predicts the endogenous vector without a target-matched source profile at inference.
Source-prior controls and BE-CORR	Predicted or measured source outcome, sequence, editor class, endogenous factors	Use source priors predicted from integrated-library training folds on the endpoint benchmark and measured integrated-library outcomes on the transfer benchmark; BE-CORR converts the source profile to editing hazard, applies chromatin-guided local transport, and combines correction candidates by learned source reliability.

Common target and split protocol. All baselines are evaluated on the same curated records and predict the same target: a 20-position vector of endogenous marginal editing rates. Public base

editing models differ in their native outputs, ranging from scalar efficiency to edited-product distributions. We therefore keep the released sequence encoder, feature recipe, or model family wherever the BE_Endo data provide the required inputs, and replace only the final task-specific head when the native output is not available. All learned models use the five repeated grouped split seeds described in Section 5.1; for each benchmark and seed, every baseline and BE-CORR use the same train/validation/test records and initialization seed.

Baseline implementations. Table 6 summarizes the input setting and output target for each baseline.

Nearest-neighbor control. The nearest-neighbor baseline encodes the protospacer and PAM as one-hot sequence features, retrieves the closest training targets within the same editor class, and averages their endogenous outcome vectors. This control tests how much of the benchmark can be solved from local sequence similarity alone.

BE-Hive (Arbab et al., 2020). BE-Hive was introduced for learning determinants of base-editing outcomes from target-library analysis. Its public tools include an efficiency model and a bystander model. The native inputs are base editor, cell type, and target sequence; the native outputs include edited-read fraction and distributions over edited products. For our benchmark, we retain the released hand-crafted sequence feature family and scikit-learn tree-regressor style, but train a multi-output head for the 20 marginal editing rates. BE_Endo provides a 40-nt local window, while the BE-Hive release uses a full 50-nt flank, so we use the overlapping feature window available in BE_Endo.

DeepBE (Kim et al., 2024). DeepBE is a sequence-based predictor for base-editing outcomes across base editors and Cas9 variants. It is designed for target-sequence inputs and returns editor-specific editing predictions, often in terms of edited output sequences or activity scores. We keep the released DeepBE-style Conv1D recipe and replace only the final task-specific output with a 20-dimensional sigmoid head. The adaptation keeps the sequence-only setting while changing the output layer to match the endogenous per-position target.

BE-DICT (Marquart et al., 2021). BE-DICT uses attention-based neural networks trained on high-throughput target library screens. It includes both per-base and bystander-oriented components and is intended to model editing outcomes from target sequence and editor identity. We instantiate the released per-base transformer backbone and train it with a two-class edited/not-edited target at each protospacer position; the edited probability is then used as the marginal editing-rate prediction.

BEDICT2.0 (Kissling et al., 2025). BEDICT2.0 extends the BE-DICT line with updated base-editing outcome models and supports user-provided target sequences and PAMs. It is designed as a general base-editing outcome predictor, not an endogenous-context correction model.

CRISPRon-BE (Sun et al., 2025). CRISPRon-BE predicts and optimizes base-editing outcomes using target sequence and auxiliary sequence-derived properties. Its native outputs emphasize editing efficiency and outcome frequency. We keep the released convolutional sequence recipe and append the auxiliary input block expected by the architecture. The CRISPRoff energy and Cas9 activity features are not available in BE_Endo, so those channels are zeroed and the output is replaced with the common 20-dimensional marginal-rate head.

BE_Endo Seq and BE_Endo Endo (Yuan et al., 2024). BE_Endo directly studies endogenous base editing by incorporating factors beyond DNA sequence. The sequence-only setting uses local target sequence and editor class, while the endogenous-context setting additionally uses genomic-context features such as expression, methylation, DNase accessibility, histone marks, CTCF, and Pol II. We implement both settings under the same output protocol: BE_Endo Seq receives editor and local sequence, and BE_Endo Endo receives those features plus the nine endogenous-context covariates. Both models follow the released BE_Endo convolutional architecture and training recipe.

HyenaDNA and Caduceus (Nguyen et al., 2023; Schiff et al., 2024). These baselines are genomic foundation models. We fine-tune the official `hyenadna-tiny-1k-seqlen` and

`caduceus-ph-seqlen-1k-d_model-118-n_layer-4` checkpoints on the 40-nt local sequence. Each pooled DNA representation is fused with editor identity, endogenous factors, and the same predicted or measured source profile available to BE-CORR, then mapped directly to the 20-position endogenous target.

FT-Transformer (Gorishniy et al., 2021). FT-Transformer provides a modern information-matched neural control without a pretrained genomic backbone. It treats each nucleotide and editor identity as a categorical token and each endogenous/source feature as a numerical token. A three-layer Transformer aggregates these tokens and directly predicts the 20-position endogenous profile.

Deep Adaptation Network (Long et al., 2015). DAN tests generic domain alignment without target-matched source conditioning. A shared convolutional sequence encoder is trained on source and endogenous records, while a multi-kernel maximum mean discrepancy penalty aligns their hidden representations. We adapt the original classification objective to multi-output regression and apply the resulting endogenous head without a source profile at test time.

Source-prior augmented baselines. The predicted-source-only baseline trains q_ϕ on the integrated-library source outcomes in each training fold and applies it to endogenous targets. The linear correction baseline uses the predicted source outcome and endogenous features in a ridge regression correction model.

Baseline hyperparameter search spaces. Table 7 summarizes the configuration search spaces used for four adapted base-editing baselines.

Table 7: **Hyperparameter search spaces for selected baselines.**

Baseline	Number of configurations	Search dimensions
BE-DICT	36	Transformer units, dropout, learning rate, batch size
BEDICT2.0	36	Transformer units, dropout, learning rate, loss
DeepBE	32	Dropout, learning rate, loss
CRISPRon-BE	24	Dropout, learning rate, loss

Inference time. Table 8 reports wall-clock endpoint inference for three representative models on an NVIDIA H100. Timings use 1,368 test records, three warm-up passes, and the median of 15 repeats; they include model-specific feature construction and source-prior prediction where applicable, but exclude checkpoint loading and input-file I/O.

Table 8: **Wall-clock endpoint inference time for representative models.**

Model	Seconds per test fold	Records/s
BE-Hive	0.41	3,341
FT-Transformer	0.81	1,683
BE-CORR	1.80	760

H Additional Experimental Results

H.1 Low-Data Learning Curve

We vary the amount of endogenous training data while retaining the complete source-library training set. Within each seed, the source-profile predictor, validation set, and test set are fixed, and the endogenous training records are subsampled into nested 10%, 25%, 50%, and 100% subsets stratified by editor class and editing burden. This experiment uses the endpoint setting, so the source profiles supplied to source-conditioned models are imputed.

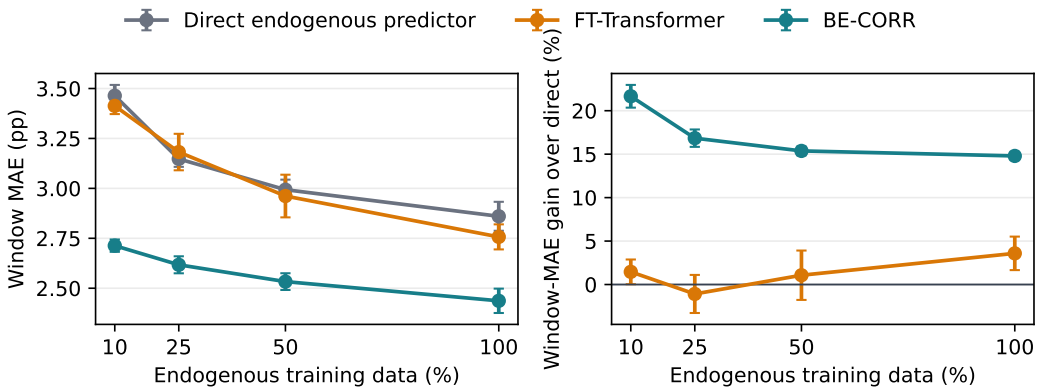


Figure 2: **Low-data endogenous endpoint prediction over five seeds.** Left: window MAE as the fraction of endogenous training data increases. Right: relative window-MAE gain over the direct endogenous predictor trained on the same subset. Points and error bars report mean ± standard deviation. BE-CORR benefits most from the source prior when endogenous labels are scarce, improving over direct prediction by 21.7% at 10% data.

Figure 2 shows that BE-CORR improves over direct endogenous prediction at every training fraction and in every seed. The relative gain is largest in the lowest-data regime and decreases as more endogenous labels become available. FT-Transformer, which receives the same imputed source profile as an input feature, does not exhibit the same low-data benefit.

H.2 Correction-Design Ablations

We compare BE-CORR with two nonlinear controls in addition to direct endogenous prediction. The unconstrained source-context MLP receives the same sequence, editor, endogenous-context, and source-profile inputs but predicts the endogenous profile directly. The no-transport variant retains hazard conversion, site-level activity scaling, and reliability gating, but replaces learned local reallocation with the identity transport. The no-correction-context variant removes endogenous-context features from the correction encoder, whereas the context-permutation variant randomly permutes these features across records. Table 9 reports results under both imputed-source endpoint prediction and measured-source matched transfer.

Table 9: **Additional correction-design ablations over five seeds.** The endpoint benchmark uses imputed source profiles; the matched-transfer benchmark uses experimentally measured source outcomes. Gain_{dir} is relative window-MAE improvement over direct endogenous prediction, whereas Gain_{src} is improvement over copying the measured source outcome. **Bold** marks the best mean in each column.

Endogenous endpoint benchmark

Variant	Win. MAE (pp)	Burden MAE (pp)	Gain_{dir}
Direct endogenous predictor	2.860 ± 0.072	13.087 ± 0.438	0.0000 ± 0.0000
Unconstrained source-context MLP	2.743 ± 0.065	13.709 ± 0.392	0.0409 ± 0.0098
BE-CORR without transport	2.682 ± 0.057	12.868 ± 0.343	0.0622 ± 0.0067
BE-CORR without correction context	2.559 ± 0.062	12.649 ± 0.354	0.1052 ± 0.0072
BE-CORR with permuted context	2.601 ± 0.067	12.832 ± 0.360	0.0906 ± 0.0102
BE-CORR	$\mathbf{2.437 \pm 0.048}$	$\mathbf{12.186 \pm 0.302}$	$\mathbf{0.1480 \pm 0.0095}$

Matched source-to-endogenous transfer

Variant	Win. MAE (pp)	Burden MAE (pp)	Gain_{src}
Direct endogenous predictor	2.992 ± 0.085	13.099 ± 0.469	0.0961 ± 0.0243
Unconstrained source-context MLP	2.544 ± 0.053	12.781 ± 0.283	0.2318 ± 0.0151
BE-CORR without transport	2.457 ± 0.077	11.241 ± 0.339	0.2577 ± 0.0196
BE-CORR	$\mathbf{2.164 \pm 0.064}$	$\mathbf{10.939 \pm 0.389}$	$\mathbf{0.3466 \pm 0.0154}$

The unconstrained nonlinear control does not match BE-CORR in either setting. Removing transport weakens window-level accuracy and gain on both benchmarks, with a larger difference when measured source outcomes are available. Thus, the improvement is not explained solely by adding source features or by global activity scaling; constrained local reallocation provides additional predictive value.

H.3 Correction-Factor Diagnostics

We evaluate whether BE-CORR's exposed factors correspond to measured source-to-endogenous changes using all held-out matched records from the five test folds. The activity target is the measured log ratio of total endogenous to source hazard. Because $\mathbf{b}_i$ is a relative potential, not an editing rate, we isolate its effect through a counterfactual transport plan in which $\mathbf{b}_i$ is removed while source hazard, editable support, window cost, and locality are fixed. We compare the changes induced by $\mathbf{b}_i$ and Π_i with measured positional changes, and stratify transport utility by ρ_i.

The factors capture distinct aspects of the observed shift. The activity shift has a seed-wise Spearman correlation of 0.35 ± 0.03 with measured total-hazard change. The median per-record Pearson correlation is 0.67 between the exposure-induced and required measured effects and 0.51 between the complete transport reallocation and measured positional change. Transport wins on 69.3% of records in the highest-ρ_i quintile versus 48.0% in the lowest, and its mean gain changes from -0.67 to $+1.34$ percentage points. Together with Figure 1C, these results support the interpretation of a_i, $\mathbf{b}_i$, Π_i, and ρ_i as activity, exposure, reallocation, and source-reliability summaries, respectively.

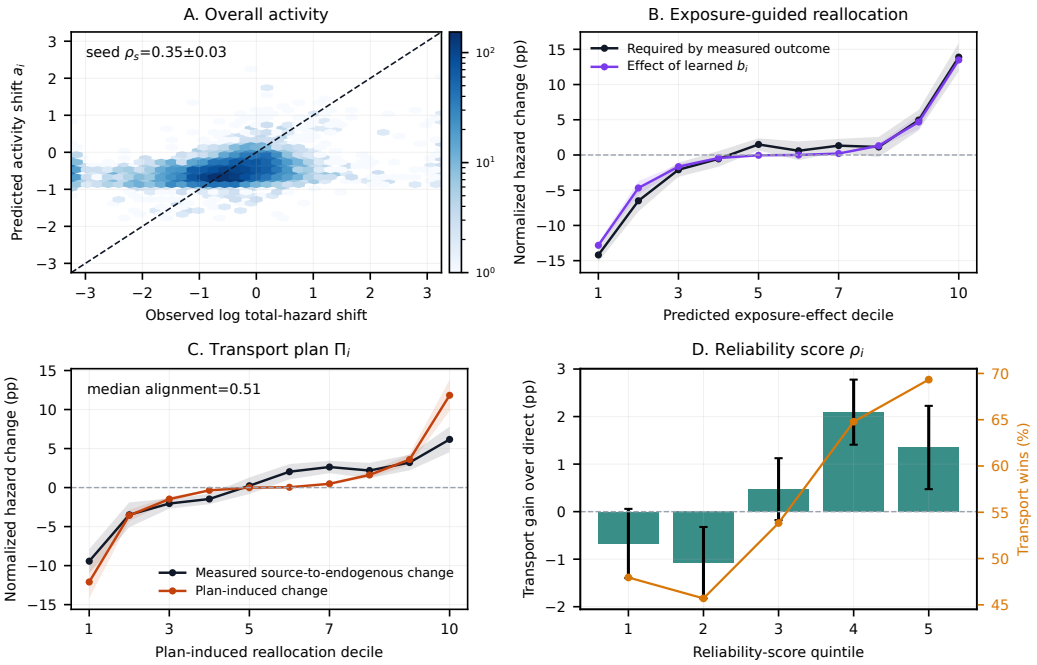


Figure 3: **Empirical correspondence between BE-CORR correction factors and measured source-to-endogenous changes over five held-out test folds.** (A) Activity shift a_i versus the measured log total-hazard ratio; the annotation reports the mean ± standard deviation of seed-wise Spearman correlation. (B) Removing $\mathbf{b}_i$ defines a counterfactual transport; deciles of its learned exposure effect are compared with the additional change required by the measured endogenous outcome. (C) Deciles of the reallocation induced by Π_i are compared with the measured positional change. (D) Records are stratified by ρ_i; bars report transport window-MAE gain over direct endogenous prediction and the line reports how often transport wins. Shading and error bars summarize variation across seeds.