Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Target-label design, not feature scarcity, drives cross-library transfer failure in CRISPRi/a guide-efficacy models

**Daniel Albuja**

albujadaniel@hotmail.com

Instituto Nacional de Biodiversidad https://orcid.org/0009-0000-5777-1365

**Research Article**

**Keywords:** CRISPRi, guide RNA efficacy, chromatin accessibility, transfer learning, target-label design, leave-one-chromosome-out, multi-modal prediction, K562

**Posted Date:** October 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11460851/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Abstract

CRISPR interference and activation (CRISPRi/a) allow reversible control over endogenous gene expression, but the efficacy of guide RNA depends strongly on the chromatin status of the target locus, which is not taken into account by guide-design programs that rely purely on the guide sequence. Thus, a 101-feature, multi-modal predictor of CRISPRi/a efficacy which relies on spanning guide sequence, chromatin accessibility, histone modification state, three-dimensional genome contacts, and local genomic context has been trained based on data from two published K562 CRISPRi studies (consisting of a total of 19446 guide RNAs) and evaluated under leave-one-chromosome-out cross-validation. Through complementary and sequential ablation, coupled with an independent interpretability study based on SHAP, it was determined that chromatin accessibility and local genomic context, but not histone marks or three-dimensional contacts, are the most important components contributing to prediction efficacy apart from sequence. The model did not succeed when it was tested, without retraining, using a new genome-wide guide library targeting mainly non-overlapping genes, in two cell lines, both of which were not part of the training, as evidenced by the low transfer performance (Spearman $\rho \approx 0.01$). In contrast, the performance of the model on the training data had been substantially better. Through fixing the model architecture, features, and hyperparameters, while retraining solely based on the continuous, cross-study-harmonized label (as opposed to the original within-gene percentile labels), most of the loss in transfer signal was retrieved ($\rho \approx 0.14–0.16$; over ten-fold increase, p-value of $\approx 0.0001$), without any negative impact on performance within the distribution, which instead improved. Control using the dataset of origin suggested that this gain was not due to artifacts. Moreover, it was concluded that the limited generalizability of CRISPRi efficacy models to guide libraries and genes outside the training set is driven primarily by within-gene label normalization rather than by a deficiency of chromatin features.


# Introduction

Most of the human genome excluding protein-coding exons contains regulatory sequence — enhancers, silencers, promoters, and insulators — that is involved in regulation of expression of particular genes in each particular cell type and context. The results of genome-wide association studies indicate that genetic variants associated with disease are located mainly in these non-coding regulatory regions and not in the coding regions (Maurano et al., 2012; Pickrell, 2014), hence, it paves the way for innovative therapeutic and scientific methods aimed at direct alteration of the activity of regulatory element, rather than affecting proteins synthesized by the respective genes.

The use of transcriptional regulation using CRISPR — including CRISPR interference (CRISPRi) and activation (CRISPRa) — provides this capability. In this method, a catalytically inactive Cas9, known as dCas9, is linked with a repressive (Gilbert et al., 2013) or activating (Chavez et al., 2015; Konermann et al., 2015; Qi et al., 2013) effector domain and is programmed using a single guide RNA (sgRNA). This approach enables reversible, titratable control over the expression of genes found in the cell without having to change any of the gene's DNA sequences. Using this method, as opposed to using nuclease-

based editing, makes CRISPRi/a well suited to both functional genomic screening and, prospectively, to therapeutic applications targeting genes that are difficult to modulate with small molecules.

Nevertheless, efficacy of guide RNA does not depend merely on its sequence. Rather, the dCas9 protein needs to find its way into its target sequence in the genome that is substantially occluded by nucleosomes, and many independent studies proved that local chromatin accessibility is one of the best indicators of whether a particular genomic site is bound and productively engaged by Cas9-based effectors, independent of guide sequence (Isaac et al., 2016; Kuscu et al., 2014; Yarrington et al., 2018). This shows that chromatin status is a major factor in terms of CRISPRi/a efficacy, and it cannot simply be considered through the guide design systems that rely solely on sequence information.

Numerous computational models have integrated levels of epigenomic knowledge into prediction of guide efficacy. The model DeepCRISPR(Chuai et al., 2018) incorporated both sequence and chromatin-accessibility data into a convolutional autoencoder, showing a small improvement of the epigenomic component over a sequence-only ablation of the same model. Later developments have taken this integration further: a simple preprint tool trained the models with epigenomic features from ENCODE and CRISPRa/i screening data (Zheng et al., 2021); EpiCas-DL integrated DNA methylation, chromatin accessibility, and nucleosome positioning information, building the model on 769918 guides from 8 published screens (Yang et al., 2023); and, finally, the very latest development, launch dCas9 combined sequence and chromatin annotation data for over a million guides, and showed a modest, though not consistent generalization of its trained models across cell lines when using a shared panel of genomic loci for evaluation (Mu et al., 2024).

Within this entire body of work, model evaluation has always been one of two types: one based on performance on guides sampled from the same distribution as training (by means of random or, less frequently, chromosome-held-out splits), or, uniquely in a case that examines cross-cell-type transfer directly (Mu et al., 2024), performance measured on a pre-defined set of loci uniformly targeted across examined cell lines. None of these evaluation regimes is suited for the situation that is the most relevant for design of novel guides: making predictions by means of trained models for newly designed guide libraries targeting largely dissimilar genes, possibly in a different cellular context. Notably, the three aspects of transfer — genome-wide, cross-library, and cross-cell-type — have been identified as the most demanding and practically significant in the context of predicting transcriptional response to perturbations, and numerous benchmarking activities became established around it (Shoeibi & Yousefi, 2026). Whether guide-efficacy prediction models transfer under this harder regime, and if not, why, has not to our knowledge been directly tested or diagnosed for CRISPRi/a on-target efficacy prediction specifically.

This project seeks to understand whether a multi-modal, chromatin-aware model that predicts the efficacy of CRISPRi/a guides based on the guide sequence, chromatin accessibility, histone modification status, three-dimensional genome contacts, and local genomic context — in addition to extensive validation inside its training distribution — can be transferred to a completely independent designed,

genome-wide guide library targeting largely non-overlapping genes in cell types outside its training distribution. In the events of transfer failure, this project will seek to locate where it originates, and whether the formulation of the training label itself, independent of feature availability, can account for and remedy the failure.

## Material and methods

# Training data sources and label construction

The training set was developed from two previously published CRISPRi/a screens that were performed in K562 cells. Horlbeck et al. (2016) provided the primary dataset in the form of a genome-scale dCas9-KRAB CRISPRi screen with the efficacy at the guide level recorded in terms of the growth-phenotype (γ) score. After the internal quality-control flag from the dataset was applied, a total of 18318 guides with resolver coordinates in hg38 were obtained.

Gasperini et al. (2019) published a second, complementary dataset that consisted of CRISPRi perturbations aimed at candidate distal enhancer elements, producing efficacies measured as the proportion of target-gene transcript remaining after perturbation (i.e., a single-cell RNA-seq readout). The analysis involved only high-confidence enhancer-gene pairs (i.e. 470 enhancer-gene pairs) identified by the original study and included 1128 guides after eliminating identical sequences within the dataset. The metric is presented inversely to Horlbeck's efficacy measure (i.e. the higher the percentage of remaining transcript, the lower the knockdown).

Therefore, the metric was transformed as effect_size = $-\log_2$(proportion_remaining + $\epsilon$), where $\epsilon$ = 0.001, which resulted in a monotonic continuous score that is correctly oriented (Spearman correlation ρ = −1.000 with regards to untransformed data).

The third dataset assessed for inclusion (Replogle et al., 2022, genome-scale Perturb-seq) was ruled out: the supplementary public materials contain only the guide-design library and no concise, distributed guide-level effectiveness chart, while the whole single-cell expression objects processed (≈ 97–160 GB) were determined to be infeasible for processing in the current project.

A fourth dataset was set aside specifically for external validation and was not involved in training at any point (Sanson et al., 2018, Dolcetto CRISPRi library, HT29 and A375 cell lines, 100,942 guides with raw read counts from plasmid and post-selection replicates). From this dataset, depletion scores at the guide level were calculated as $\log_2$(CPM-normalized post-selection counts / CPM-normalized plasmid counts), averaging over three biological replicates for each cell line.

## Label harmonization

The two label formulations were derived from the Horlbeck and Gasperini training dataset (19446 guides that target 2151 target genes) and utilized the same first step: z-scoring of the efficacy metrics of each dataset, computed independently before any cross-study combination (there were no overlaps between the guide sequences/coordinates in the two source datasets).

- The EEP (ensemble efficacy percentile) was computed by transforming z-score obtained in each study to a percentile rank (0–100) within each target gene, using rank-based approaches adopted for aggregating heterogeneous CRISPR screen readouts. Genes represented by a single guide (n = 75, all from Horlbeck) were assigned a fixed percentile of 50 by convention.
- The Continuous harmonized z-score made use of the within-study z-score as is, without the within-gene percentile collapse. This construct was used specifically to determine whether the within-gene normalization — which eliminates any signal that separates between efficacy in different target genes — was responsible for a transfer-learning failure observed during development (see Results)

Both label columns were retained in the final training matrix for direct comparison that was done using the same feature set, model architecture, and configuration of hyperparameters.

## Feature engineering

The predictive features were divided into five respective classes, in accordance with a multi-modal approach, which includes guide sequences, chromatin accessibility, statuses of histone modification, 3D genome connections, as well as local genomic context. A total of 101 features were retained after the process; Table S1 elucidates the total coverage relative to the initial multi-group specification, and also full detail about the two internal inconsistencies found during implementation (see “Transparency notes”).

## Sequence features (F1)

The inputs variables included GC content, maximum homopolymer run length, poly-T run count, predicted minimum free energy of the guide secondary structure (using RNAfold, ViennaRNA) as well as calculated melting temperature [using the nearest-neighbor model by Biopython Tm_NN and the parameter table of Santalucia (1998)] and a score produced by using unsupervised position-weight-matrix for 8-nt seed region (in which frequencies of bases were analyzed for the whole training data set without being affected by the efficacy conditioning to prevent leakage of labeling in the future as it was confirmed later using Spearman’s $\rho = 0.075$ when compared to the training label). The information about the protospacer-adjacent motif (PAM) emerged from comparing each guide (forward and reverse) with a ± 80bp genomic window and taking three nucleotide sequences 3' downstream from the protospacer on the aligned strand (the PAM was identified for 18433 out of 19446 guides, and 99.99% of them had a correct NGG motif, proved with three different tests). Furthermore, a 16-column one-hot encoding of the

four 3' terminal guide positions was preserved as a documented sequence proxy, different from the actual PAM context, which lies outside the guide sequence itself.

## Chromatin accessibility features (F2)

The ATAC-seq fold-change-over-background signal (K562; ENCODE accession ENCSR868FGK) was obtained from six symmetrical windows surrounding each genomic position of the guide (± 100 bp, ± 200 bp, ± 500 bp, ± 1 kb, ± 2 kb, ± 5 kb), and mean, maximum, and 90th-percentile signal were calculated for each window (18 features total). Also, the DNase-seq signal (ENCSR000EKS) at ± 200 bp, a binary variable of overlap with an ATAC-seq peak summit, and distance to the nearest peak center were included as three additional features.

# Histone modification and chromatin-state features (F3)

ChIP-seq fold-enrichment signal for five histone marks — H3K27ac (ENCSR000AKP), H3K4me3 (ENCSR000AKQ), H3K4me1 (ENCSR000EWC), H3K27me3 (ENCSR000AKS), and H3K9me3 (ENCSR000DWD) — was obtained using three window sizes (± 500 bp, ± 1 kb, ± 2 kb; with mean signal per window, 15 features). CTCF ChIP-seq signal at ± 1 kb (ENCSR000AKO) and a one-hot-encoded 15-state ChromHMM chromatin-state segmentation (Roadmap Epigenomics, K562/E123; ENCSR365YNI) were also included as 16 additional features (31 total for this group). During ENCODE accession verification, the H3K4me1 and CTCF accessions were found to correspond, respectively, to a control experiment and to H3K36me3 signal. Thus, corrected accessions are reported above (identified via ENCODE experiment metadata).

## Three-dimensional genome contact features (F4)

The in-situ data on Hi-C for the K562 cell line (Rao et al., 2014; 4D Nucleome database accession number 4DNFITUOMFUQ) were processed to produce a balanced contact matrix with a spatial resolution of 10 kb. For each corresponding guide, the data on the normalized Knight–Ruiz contact frequency between the gene, its corresponding genomic bin and the transcription start site (TSS) bin was extracted and log-transformed. Guides within 5 kb of the TSS were zero-padded to account for the resolution limit of the assay (this proved to be successful with approximately 90% of all guides utilized, in line with the promoter-proximal nature of the Horlbeck library). The data for A/B compartment assignment computed on the basis of the first eigenvector of the contact matrix (calculated at the resolution of 100 kb) were obtained but failed to meet the independent biological criterion (correlation not observed with ATAC-seq accessibility; $p = 1.0$ using Mann–Whitney test) and were therefore excluded from the building blocks of the model even though it was present in the output dataset as an unvalidated exploratory column.

## Local genomic context features (F5)

In total, seven variables constitute the extracted features: the log-transformed distance between each guide and the TSS of its target-gene and to the nearest TSS of a gene expressed in K562 (RNA-seq TPM > 1; ENCSR615EEK), the overlaps of ENCODE candidate cis-regulatory elements across four types of regulatory features (promoter-like, proximal enhancer-like, distal enhancer-like, CTCF-only; ENCSR800VNX), and the mean PhyloP 100-way vertebrate conservation score within ± 100 bp (UCSC hg38 track).

# Missing data handling

The missing values caused by incomplete signal-track coverage (74 guides for H3K4me3/H3K27me3; 60 guides for target-TSS distance; 2 guides for PhyloP) were filled in with the training-set column median with a binary missingness variable kept as another feature for every imputed column, following standard procedure for continuous genomic variables with low levels of missingness.

# Model architecture and selection

In this research, six different types of model architectures were utilized - ridge regression, LASSO, random forest, gradient-boosted trees (under the XGBoost system), multi-layer perceptron with batch normalization and dropout of 0.3, and one-dimensional convolutional network with a one-hot encoding of the sequences of guidance in ± 10-nt flanking regions. The aforementioned models were trained on a 101-feature matrix and test results were evaluated according to a predefined four-component evaluation system, including prediction accuracy of a reserved validation chromosome (40%), calibration of the 90% prediction intervals (20%), presence of exact vs. approximate interpretability using SHAP values (30%), and stability of cross-fold results performance (10%). Ridge regression had the best score (S = 0.727) according to the assessed criteria, considering its interpretability weight was kept to the maximum level, while XGBoost produced the best raw validation outcome (ρ = 0.403 instead of ρ = 0.335 for ridge regression) with lower overall score (S = 0.662) according to the same criteria.

XGBoost was utilized for all further analysis (cross-validation, ablation, SHAP interpretation, external transfer evaluation, benchmarking). The rationale behind this choice is that any precise computation of the Shapley values can only be done on tree-based models through Tree-SHAP, thus it is essential for the current study whose core method is just based on interpretability analysis; this is a deliberate methodological decision, made explicit here, rather than a rejection of the pre-specified selection protocol’s outcome, which is reported in full in the Supplementary Materials.

XGBoost hyperparameters (maximum tree depth = 4, learning rate ≈ 0.0297, n_estimators = 289, minimum child weight = 4, subsample ≈ 0.766, column subsample ≈ 0.567, L2 regularization ≈ 0.318) were selected through Bayesian optimization (Optuna, 100 iterations) on the training and validation process described below and held constant throughout the rest of the analysis (cross-validation folds, ablation steps, and the two label formulations), which is necessary to make comparisons between the effect of the experimental variable under study rather than confounding it with re-tuning.

# Cross-validation protocol

The way the model performed was evaluated using the leave-one-chromosome-out (LOCO) method of cross-validation. In all, there were 23 folds, including autosomes 1–22 and chromosome X (chromosome Y was absent from the feature matrix). All the guides from a particular chromosome were omitted for testing, while the rest were used for training the model. The hyperparameters were fixed in the same manner as mentioned above. Then, fold-level Spearman correlations between predicted and observed labels were obtained as the primary performance metric (mean ± SD across folds), in preference to a single pooled out-of-fold correlation, to appropriately reflect between-chromosome variability.

# Feature-ablation analysis

Sequential ablation was performed under the same leave-one-chromosome-out (LOCO) protocol by cumulatively adding feature groups in five steps: sequence only (F1), sequence + accessibility (F1 + F2), + histone/chromatin-state (F1 + F2+F3), + 3D contacts (F1 + F2+F3 + F4), and finally, the complete 101-feature set. Hyperparameters were kept constant in all steps to determine the unique contribution from each set of features due to the complexity of the model. Finally, a leave-one-group-out approach was also performed, in addition, to carry out the marginal contribution analysis considering the participation of every other feature group to the ultimate result, particularly in cases of dealing with groups of highly correlated features like ATAC-seq and H3K27ac signals. The changes from the different performances were confirmed statistically using a paired permutation test over 10000 iterations based on fold-level Spearman ρ differences.

# Model interpretation

The SHAP (SHapley Additive exPlanations) scores were calculated for the XGBoost model, using its own pred_contribs feature (equivalent to TreeSHAP; the shap package's TreeExplainer implementation was incompatible with this model's base_score configuration). The global importance for each feature was determined by taking an average of the absolute SHAP value of each feature, aggregated over the out-of-fold LOCO predictions using the entire training set (totaling 19446 training guides, chosen over a smaller held-out test partition, n = 490, after confirming high rank concordance between the two, Spearman ρ = 0.982, with 95% confidence intervals obtained by bootstrap resampling, 200 iterations). The biological validation of the features was carried out through designed experiment including controls of positive/negative testing. For the 100 best-performing training guides, the proportion of those that gave SHAP-positive values with respect to chromatin-accessibility and genomic-context features was calculated. For guides in overlapping constitutive heterochromatin function (those containing H3K9me3-marked ChromHMM states) it was determined whether their SHAP value behaved according to the expected negative-repressive pattern.

# External transfer evaluation

Cross-cell-type, cross-library transfer performance was tested using the K562 trained model without retraining, to the independent Sanson et al. (2018) Dolcetto guide library (100942 guides, HT29 and A375 cell lines), which targets a largely non-overlapping set of genes relative to the training set (exact-

sequence overlap with the combined training set consisted of 526 guides). All the 101 features were obtained for this external guide set using K562 chromatin tracks (i.e., no matched HT29/A375 Epigenomic data was available, which was confirmed due to a low signal coverage in ENCODE for the required histone marks — accessibility data alone was obtained for HT29 via DNase-seq only). Transfer performance was evaluated as the Spearman correlation between the model predictions and an observed depletion score for each of the cell lines, under both label formulations (EEP-trained and continuous-z-score-trained models, with the same architecture, features, and hyperparameters), with significance of the difference determined with a paired permutation test (10000 iterations).

# Benchmarking against existing scoring approaches

The performance of the model was assessed against the following two conditions under the same LOCO protocol: (i) F1 features only, which represents the sequence-only ablation condition of the same model; (ii) Rule Set 3 (an implementation based on Doench et al.'s Azimuth), which is known to replace the documented Rule Set 2 due to incompatibility issues with the library (see Transparency notes), was evaluated on the sub- set of Horlbeck guides with a canonical NGG PAM context (n = 5056); and (iii) a genomic-distance-only baseline (guide-to-target-TSS distance, evaluated under both a linear model and a non-linear gradient-boosted-stump model). Attempts to compare the model performance against CRISPOR's native Rule Set 2 and DeepCRISPR were made but were unsuccessful; the reasons were of infrastructure nature, rather than data or methodological limitations (see Transparency notes).

# Statistical analysis

Confidence intervals at the 95% level were computed for all main metrics analyzed in this study (i.e., LOCO mean ρ under each label formulation, external-transfer ρ under each label formulation, ablation-step ρ, and benchmarking comparators), using a non-parametric bootstrap scheme (1000 iterations), performing resampling of fold-level correlations for the LOCO and ablation metric calculations and resampling of prediction–label pairs for the external-transfer metrics. Statistical tests were two-sided, except when otherwise stated; permutation tests were always performed with 10000 iterations.

# Transparency notes

There were a few discrepancies noticed in the initial multi-group feature characterization during implementation and are reported here rather than silently resolved: (i) the dimensionality claimed by the accessibility feature group (18) was found to be attainable through symmetric interpretation based on six window and three statistics, and the achieved dimensionality and claimed dimensionality matched perfectly; (ii) the dimensionality reported by the histone/chromatin-state group (42) could not be matched with credibility to any proper interpretation of the text (working out to 31), the literal reading (31) was implemented, and the discrepancy is reported unresolved; (iii) the dimensionality declared for the 3D-contact group (8) could not be counted on to establish reliability against its text description (≈ 5–6); two features passing independent validation were retained. A total of two ENCODE experiment accessions in the original specification of features (H3K4me1 and CTCF) have turned out not to refer the

expected experiments after metadata verification, they were found to reference different assays than intended (a control experiment and H3K36me3, respectively) and were corrected as described above.

# Results

### Chromatin accessibility and local genomic context, but not histone marks or 3D contacts, improve prediction beyond sequence

The use of sequential feature-group ablation (EEP-trained model, LOCO) indicated that using sequence features alone (F1) yielded a mean Spearman's ρ of 0.329 [0.315, 0.343]. The inclusion of chromatin-accessibility features (F1 + F2) led to a substantial increase to 0.360 [0.345, 0.375] (with a change in Spearman's correlation coefficient of + 0.032, and permutation p-value of 0). Adding histone-modification and chromatin-state features (F1 + F2+F3) did not yield further improvements (0.360 [0.344, 0.375]; with a change of − 0.001, and a p-value of 0.70), and including 3D-contact features (F1 + F2+F3 + F4) likewise produced no significant change (0.358 [0.343, 0.373] with a change of − 0.002, and a p-value of 0.18). The complete 101-feature model, including local genomic-context features (F5), produced 0.375 [0.360, 0.389] (with a change of + 0.018 over F1 + F2+F3 + F4, and a p-value of 0). Confirmation of these results was provided with leave-one-group-out ablation: while removing sequence-related features F1 resulted in collapsed performance to 0.223 ($\Delta = -0.153$ compared to the complete model, $p \approx 0$), discarding chromatin-accessibility features F2 produced lost performance by − 0.021 ($p \approx 0$), and omitting local genomic-context features F5 caused decrease of performance by − 0.018 ($p = 0.0001$), but removing F3 or F4 separately, didn't create any significant change ($\Delta \approx 0$, $p = 0.89$ and $p = 0.60$, respectively) (Fig. 1).

Global SHAP feature importance, which was calculated using out-of-fold LOCO predictions obtained from training data of 19446 training guides (rank-consistent with a smaller held-out test partition, Spearman $\rho = 0.982$), revealed the same ordering of feature categories as found in ablation experiments: sequence (F1) yielded the maximal total contribution, followed by accessibility (F2) and genomic context (F5), which contributed also considerably while histones/chromatin state (F3) and 3D-contact (F4) had negligible effects (Fig. 2). For guides overlapping onto constitutive heterochromatin marked with H3K9me3, the contributions of SHAP were close to zero instead of following the anticipated negative-repressive, consistent with the null marginal contribution of the feature group in the ablation experiments.

The biological-validation analysis of the 100 training guides with the highest efficacy established that 57% of them received a net-positive SHAP contribution from chromatin-accessibility features and 60% from the features of genomic context, while 61% turned out to be sequence-dominated (net SHAP contribution from sequence features was larger than the combined contribution of all chromatin and genomic-context features). Only 26% of these best guides were simultaneously chromatin-positive and not sequence-dominated. This states that, while chromatin accessibility gives a significant amount of predictive information in general terms, the mechanisms responsible for high efficacy are heterogeneous

across each guide, with a majority of top-performing guides in this dataset explained primarily by sequence composition.

# Comparison against sequence-based and position-based scoring approaches

With an identical LOCO protocol, the entire EEP-trained model with 101-features (ρ = 0.375 [0.360, 0.389]) was better than Rule Set 3 (ρ = 0.259 [0.232, 0.286]; using 5056 Horlbeck guides with canonical NGG PAM contexts) and a baseline genomic-distance-only approach (linear model, ρ = 0.021 [0.006, 0.034]; non-linear gradient-boosted-stump model, ρ ≈ 0.12, point estimate without bootstrap resampling), having non-overlapping 95% confidence intervals compared to the linear model (see Table 1). The comparison of our model versus the Rule Set 2 system that CRISPOR employs as well as DeepCRISPR could not be completed due to the specific limitations (see Methods, Transparency notes).

Table 1
Benchmarking against sequence-based and position-based scoring approaches (LOCO, K562, EEP label).

| Method | n | Spearman ρ [95% CI] |
|---|---|---|
| **Full model (101 features)** | 19446 | 0.375 [0.360, 0.389] |
| **Sequence-only (F1)** | 19446 | 0.329 [0.315, 0.343] |
| **Rule Set 3** | 5056 | 0.259 [0.232, 0.286] |
| **Distance-to-TSS, linear** | 19446 | 0.021 [0.006, 0.034] |
| **Distance-to-TSS, non-linear (GBR)** | 19446 | ~ 0.12 (point estimate) |

# The trained model does not transfer to an independent, genome-wide guide library

When the K562-trained, EEP-labeled model was used without retraining on the independent Dolcetto library from Sanson et al. (2018) (HT29 and A375 cell lines with a total of 100942 guides and largely non-overlapping target genes), its performance resulted in near-null transfer: Spearman ρ = 0.008 [0.002, 0.014] in HT29 and 0.012 [0.006, 0.018] in A375 (Table 2, Fig. 3). Such results sharply contrast with the higher performance level of the model while it was trained on K562 data (LOCO ρ = 0.375) and to the magnitude of chromatin contribution demonstrated by the ablation analysis above, motivating a direct investigation of its cause.

# Target-label design, not feature availability, is the dominant cause of transfer failure

An identical model — same architecture, same number of 101 features, same set hyperparameters — was retrained with the help of a continuous, within-study z-score label (given there was no collapse within-gene percentile), and the same procedure was used to evaluate it. The model built on training K562 data, the continuous-target model displayed considerably better LOCO performance compared to the EEP-trained model (0.429 [0.412, 0.447] vs. 0.375 [0.360, 0.389]; non-overlapping 95% confidence intervals). As for the external library by Sanson et al., the model using a continuous target achieved ρ = 0.136 [0.129, 0.142] in HT29 and 0.158 [0.152, 0.164] in A375, which puts it more than an order of magnitude above the EEP-trained model, with non-overlapping confidence intervals in both cell lines and a paired permutation test confirming the difference ($p \approx 0.0001$ in both HT29 and A375). It is worth mentioning that the improvement in external transfer was not accompanied by any degradation in K562 performance; the continuous-target model was simultaneously superior in both the source and target evaluation regimes (Table 2, Fig. 3).

Table 2
Effect of target-label design on within-distribution (LOCO) and cross-library, cross-cell-type transfer performance.

| **Label formulation** | **LOCO K562 ρ [95% CI]** | **Transfer ρ HT29 [95% CI]** | **Transfer ρ A375 [95% CI]** |
|---|---|---|---|
| **EEP (within-gene percentile)** | 0.375 [0.360, 0.389] | 0.008 [0.002, 0.014] | 0.012 [0.006, 0.018] |
| **Continuous z-score (harmonized)** | 0.429 [0.412, 0.447] | 0.136 [0.129, 0.142] | 0.158 [0.152, 0.164] |

Note: n = 19446 for LOCO (23 folds, chromosome-held-out); n = 100942 for external transfer (Sanson et al., 2018, Dolcetto library). All confidence intervals from 1000-iteration bootstrap resampling. Permutation test (10000 iterations) for the continuous-vs-EEP difference: $p \approx 0.0001$ in both HT29 and A375.

In order to demonstrate that the increase in the model's performance was a result of true efficacy signal, rather than simply being based on the ability of the model to differentiate between sources from different datasets, two additional steps were taken. First, the distribution of z-scores for the two source studies within each study was checked for similarity in location and scale (z-scores are centered at 0 and with SD = 1), thereby eliminating a scale-based confound. Second, the Random Forest model, trained on the 101 features trained in order to predict the dataset of origin had good separation performance (AUC ≈ 1.0). This was valid considering that the Gasperini guides have the goal of targeting distal enhancers, whereas the Horlbeck guides are primarily promoter-proximal, hence there is a real biological difference in terms of the features (not just the labels). It is important that the LOCO performance of the continuous target when using only Horlbeck data, and not combining the Gasperini dataset, was 0.448 ± 0.041, which is better than the LOCO performance of the combined model, which was 0.429 ± 0.044. In case the superior performance of the combined Horlbeck + Gasperini model compared to EEP was due to the usage of the dataset-of-origin separability, the inclusion of one more dataset would increase this parameter. This is not observed in the results obtained, thus demonstrating that the increase in transfer performance is due to the target design effects and not study heterogeneity artifacts.

## Discussion

# Target normalization as an overlooked determinant of transfer performance

The main conclusion reached through this research is more mechanistic as opposed to descriptive because the failures encountered as a result of cross-library transfer when using the CRISPRi guide-efficacy models occur not as a result of the lack of chromatin information, but because of the way the training labels were designed. The in-gene percentile normalization appears to be the reasonable technique when considering a single training distribution since it controls for variations in the base level of gene expression or essentiality of the target genes despite the fact that it excludes signals that could help in ranking the guides targeting different genes in relation to one another. An external guide library composed mostly of genes outside the training set (which is usually the case in genome-wide transfer) is where the signal discarded would be useful. The magnitude of the effect observed here (in a more than ten-fold improvement in the transfer Spearman ρ from changing the format of the label definition without changing the model architecture, features, and hyperparameters) shows that the design of target-label should be scrutinized explicitly in this area as it is done for feature selection and model architecture.

# Relationship with prior evidence on cross-cell-type generalization

This finding does not conflict with previous evidence which is related to direct comparisons on cross-cell-type transfer of CRISPR perturbation-effect models. According to Mu et al. (2024), both XGBoost and CNN based models demonstrated reasonable performance across different types of cell lines, such as K562, induced pluripotent stem cells (iPSCs), and iPSC-derived neural progenitor cells (NPCs), by evaluating these models on a common dataset of 581 regulatory regions within the Major Histocompatibility Complex locus, where the target remained the same across all three cell lines (5-fold cross-validation by target region). Nevertheless, even though the same-locus approach was used, the accuracy of predictions for the more differentiated NPCs was lower than for K562 and iPSCs. Moreover, the study design — same genomic loci, but different cellular context — is not as demanding as the regime tested here with an external library targeting a largely disjoint collection of genes and a different cell line. Hence, both of the outcomes should be interpreted as complementary and not contradictory: past research provides evidence for generalization across cell types when the target loci are held constant; however, there is visible degradation even in that case as the cellular context shifts farther away from the training line; and the present research demonstrates that generalization also requires the appropriate target-label formulation when the target loci are permitted to vary. All three other multi-modal efficacy-prediction frameworks discussed in this paper — DeepCRISPR (Chuai et al., 2018), CRSIPR-A-I (Zheng et al., 2021), and EpiCas-DL (Yang et al., 2023) — do not perform evaluation on an autonomous, genome-wide guide library following the chromosome- or gene held-out strategy used in

the current study; none of their validation designs (random splits in all cases) allow for the verification of the failure mode explained in the present study.

In a broader sense, the challenge of applying machine-learning models of CRISPR perturbation interference impacts to genes that did not undergo any training has lately been underscored as the toughest and most practically relevant evaluation regime for the closely related task of transcription-response prediction, for which a dedicated community benchmark has been established (Shoeibi & Yousefi, 2026). The current findings bring this issue down to the less general and, in practical applications of guide-design, more immediately applicable matter of predicting on-target efficacy, and provide the concrete, experimentally isolated mechanism — rather than a purely empirical observation — for one contributor to that difficulty.

# Practical implications

When considering groups designing models for the efficacy of CRISPRi/a that must handle genes or gene sets outside their training data — and this is typically the case for prospective use of a guide-design system — the results of the present work suggest that the process of auditing the formulation of labels for within-group normalization should be seen as an indispensable diagnostic procedure to undertake prior to concluding that extra features, increased size of the training data set, or a more expressive architectures is necessary for gaining any external performance improvement. Specifically, the label fix performed in this study was able to retrieve most of the lost transfer signal without adding any additional data or features, while also maintaining the performance within the original training distribution, showing that in those cases where this fix is applicable, it is close to a strictly dominant choice rather than a trade-off to be weighed against within-distribution accuracy.

# Limitations

There are certain limitations in relation to the current research which need to be taken into account during the interpretation of the findings. First, despite the label alterations, the level of transfer performance ($\rho \approx 0.14–0.16$) remains relatively low and below the model within-distribution performance; the introduced target-design modification reduces, but does not close the cross-library, cross-cell-type generalization gap. Second, a third published CRISPRi K562 dataset(Replogle et al., 2022) could not be used in the training process due to practical impossibility of processing its full single-cell expression objects within the scope of this study; nevertheless, this dataset could probably contribute to improving the between-gene signal in the training label. Third, it was not possible to conduct cell-type-matched epigenomic validation that would have used the model with HT29 or A375 specific chromatin tracks, instead of using K562 tracks since an insufficient number of histone markers, required for this validation, was provided by ENCODE for both cell lines at the time of this study. This would have meant that the transfer results provided in this study reveal a cell-type mismatch in both labeling and feature domains simultaneously, and the present data does not allow determine how much of each of these aspects is responsible for the residual performance gap. Fourth, two of the feature groups (three-dimensional genome contacts and part of the local genomic-context group) are somewhat

underdeveloped compared to what they are intended to be, but ablation analysis suggest that these will be unlikely to affect the main result, as neither feature has substantially changed the result even in its current partial form. Finally, due to the limitations of the computing environment, a complete comparison against CRISPOR's native Rule Set 2 implementation and DeepCRISPR could not be done; while Rule Set 3 and a distance-based baseline provide informative points of comparison, a direct comparison against these two widely used tools would strengthen the results presented in this study.

# Future directions

The specific result of target redesign that has been presented in this paper was derived by taking a combination of two dataset for training. The natural next step is to see if the same correction would be beneficial in terms of transfer when applied to a training set containing a third or fourth published CRISPRi/a screen. Moreover, it could be checked if the size of the improvement scales with the number and diversity of target genes used in training. The findings outlined above, in a broader context, highlight the need for a more rigorous comparison of approaches to label-harmonization for the prediction of CRISPR efficacy. For instance, comparing the within-gene percentile normalization method with the continuous cross-study harmonization technique, using a broader selection of already published screens and perturbation modalities (CRISPRi, CRISPRa, and base/prime editing efficacy prediction), in order to formulate general guidelines for training-label design in this domain.

## Conclusion

Chromatin accessibility and local genomic context offer independent and strong predictive signals along with the guide sequence alone in predicting CRISPRi guide-efficacy models; however, it is not sufficient to guarantee that the trained model can generalize on an independent guide library designed. It has been shown that nearly complete failure of the transferability of the trained model to the independent library can be mainly attributed to within-gene normalization of the label used in the training process instead of any intrinsic problem with the chromatin information, and the simple correction to the labeling process can regain most of the signal without sacrificing the within-distribution performance. The separation of failure mechanisms according to labels vs. features has, to the best of current knowledge, not been performed explicitly for the task of CRISPRi/a efficacy prediction and offers a practical and low-cost diagnostic step that can be taken in future studies involving training and deployment of guide-efficacy models outside of its training distribution.

## Declarations

# Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

# Data, scripts, code, and supplementary information availability

Data are available online: The Horlbeck et al. (2016), Gasperini et al. (2019), and Sanson et al. (2018) sgRNA efficacy datasets used for training and external validation are publicly available at their original sources (eLife, Cell, and Nature Communications supplementary materials, respectively). ENCODE, 4D Nucleome, and Roadmap Epigenomics accessions used for feature extraction are listed in Methods. The compiled 101-feature training matrix (both label formulations) is available online: https://doi.org/10.5281/zenodo.22820926 (Albuja, 2026).

Scripts and code are available online: All feature-extraction, model-training, and analysis code is available online: DOI: https://doi.org/10.5281/zenodo.22820678 (Albuja, 2026); https://github.com/evanalbuja-png/ChromaCRISPR

Supplementary information is available online: https://doi.org/10.5281/zenodo.22820991 (Albuja, 2026).

# Funding

No specific funding was received for this study.

# Acknowledgements

Thanks to the maintainers of ENCODE, 4D Nucleome, and the original screen authors for public data. No institutional laboratory resources were used for the analyses reported here.

## References

1. Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, Iyer PR, Lin E, Kiani S, Guzman S, Wiegand CD, Ter-Ovanesyan DJ, Braff D, Davidsohn JL, Housden N, Perrimon BE, Weiss N, Aach R, Collins J, J. J., Church GM (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12(4):326–328. https://doi.org/10.1038/nmeth.3312
2. Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, Zhou C, Zhu C, Chen K, Duan B, Gu F, Qu S, Huang D, Wei J, Liu Q (2018) DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. Genome Biol 19(1). https://doi.org/10.1186/s13059-018-1459-4
3. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 34(2):184–191. https://doi.org/10.1038/nbt.3437

4. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, Trapnell C, Ahituv N, Shendure J (2019) A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176(1–2):377–390e19. https://doi.org/10.1016/j.cell.2018.11.029
5. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154(2):442. https://doi.org/10.1016/j.cell.2013.06.044
6. Horlbeck M, Gilbert L, Villalta J, Adamson B, Pak R, Chen Y, Fields A, Park C, Corn J, Kampmann, Weissman J (2016) Compact and highly active next- generation libraries for CRISPR-mediated gene repression and activation. ELife. https://doi.org/10.7554/eLife.19760.001
7. Isaac R, Jiang F, Doudna J, Lim W, Narlikar G, Almeida R (2016) Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. ELife. https://doi.org/10.7554/eLife.13450.001
8. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517(7536):583–588. https://doi.org/10.1038/nature14136
9. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol 32(7):677–683. https://doi.org/10.1038/nbt.2916
10. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Stamatoyannopoulos JA (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337(6099):1190–1195. https://doi.org/10.1126/science.1222794
11. Mu W, Luo T, Barrera A, Bounds LR, Klann TS, Weele Mter, Bryois J, Crawford GE, Sullivan PF, Gersbach CA, Love MI, Li Y (2024) *Machine learning methods for predicting guide RNA effects in CRISPR epigenome editing experiments*. https://doi.org/10.1101/2024.04.18.590188
12. Pickrell JK (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet 94(4):559–573. https://doi.org/10.1016/j.ajhg.2014.03.004
13. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152(5):1173–1183. https://doi.org/10.1016/j.cell.2013.02.022
14. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665–1680. https://doi.org/10.1016/j.cell.2014.11.021
15. Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, Mascibroda L, Wagner EJ, Adelman K, Lithwick-Yanai G, Iremadze N, Oberstrass F, Lipson D, Bonnar JL, Jost M, Norman TM, Weissman JS (2022) Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell 185(14):2559–2575e28. https://doi.org/10.1016/j.cell.2022.05.013

16. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, Vaimberg EW, Goodale A, Root DE, Piccioni F, Doench JG (2018) Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun 9(1). https://doi.org/10.1038/s41467-018-07901-8
17. Santalucia J (1998) *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics* (Vol. 95). https://doi.org/10.1073/pnas.95.4.1460
18. Shoeibi M, Yousefı N (2026) Can We Trust In-Distribution Success? Locked Evaluation Reveals Transfer Failure and Sampling-Depth Entanglement in CRISPRi Perturbation Prediction. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2608.00152
19. Yang Q, Wu L, Meng J, Ma L, Zuo E, Sun Y (2023) EpiCas-DL: Predicting sgRNA activity for CRISPR-mediated epigenome editing by deep learning. Comput Struct Biotechnol J 21:202–211. https://doi.org/10.1016/j.csbj.2022.11.034
20. Yarrington RM, Verma S, Schwartz S, Trautman JK, Carroll D (2018) Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. Proc Natl Acad Sci USA 115(38):9351–9358. https://doi.org/10.1073/pnas.1810062115
21. Zheng X, Cui J, Wang Y, Zhang J, Wang C (2021) *CRSIPR-A-I: a webtool for the efficacy prediction of CRISPR activation and interference*. https://doi.org/10.1101/2021.12.02.470943

# Figures

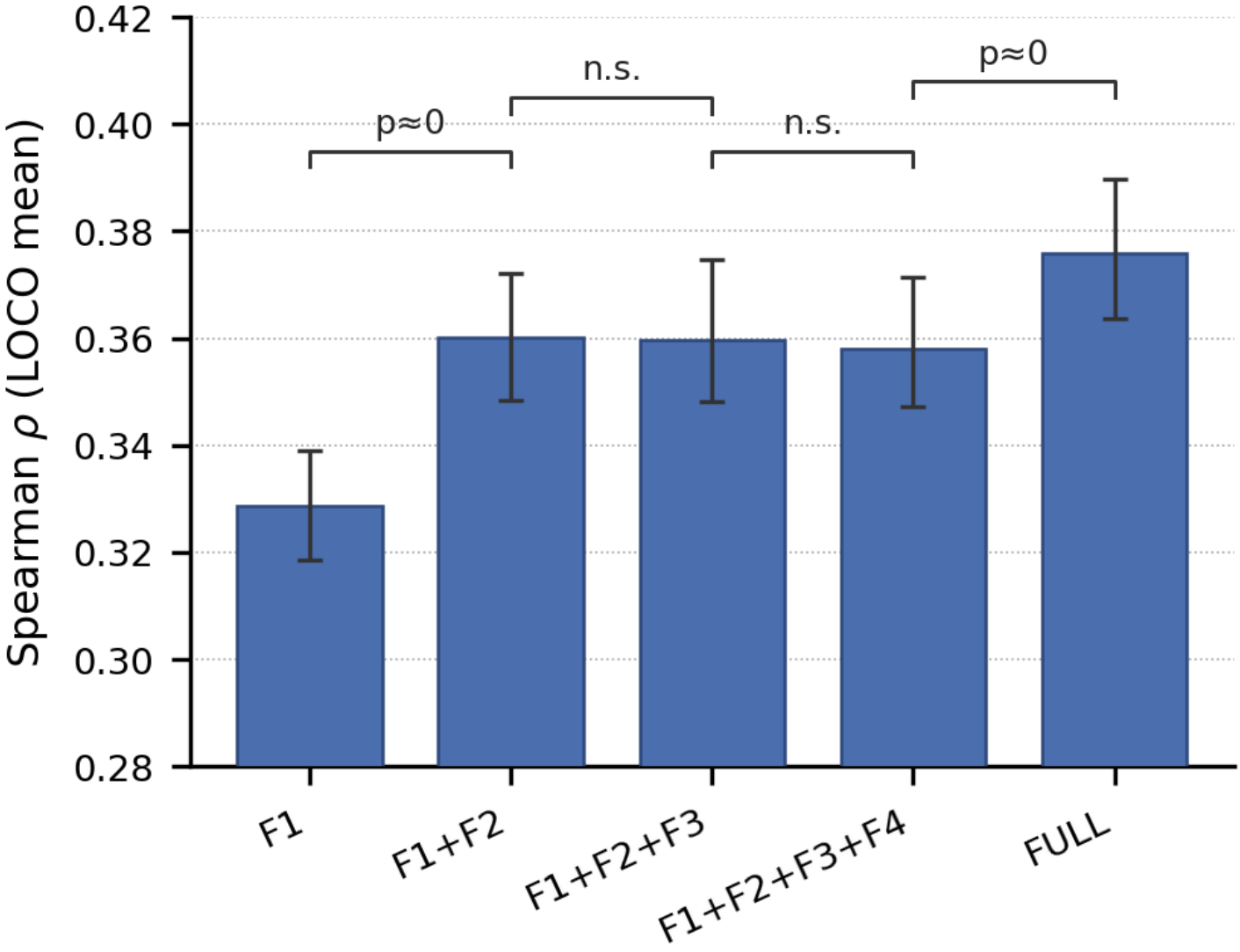


**Figure 1**

Sequential feature-group ablation (LOCO, EEP label). Mean Spearman ρ (± 95% bootstrap CI, n = 1,000 resamples) across 23 chromosome-held-out folds. Sequence alone (F1) is significantly improved by adding accessibility (F2, p ≈ 0) and, in the full model, by genomic context (F5, p ≈ 0); histone/chromatin-state (F3) and 3D-contact (F4) additions are not significant (p = 0.70 and p = 0.18, respectively)

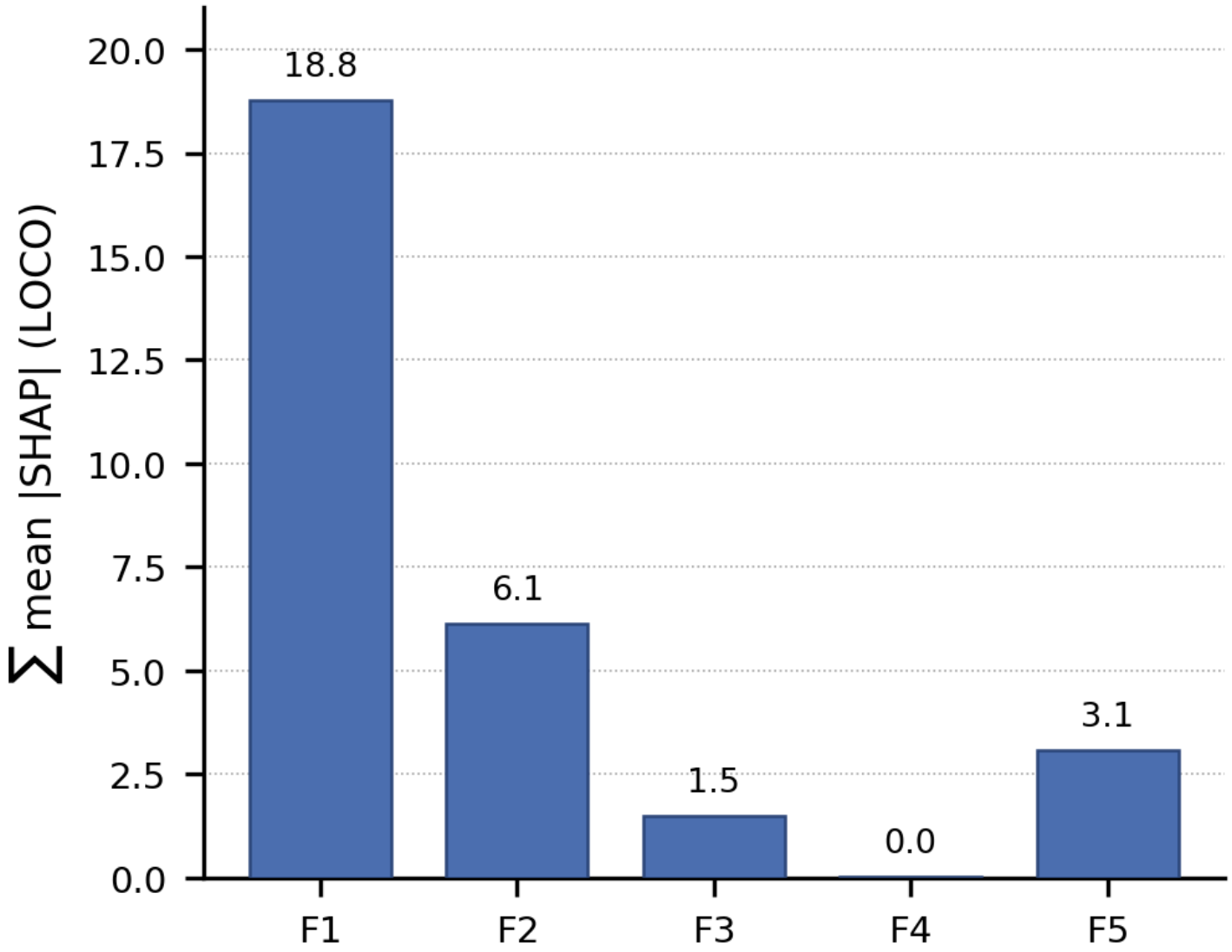


**Figure 2**

Global SHAP feature importance by modality. Sum of mean |SHAP value| per feature group, computed over out-of-fold LOCO predictions for all 19,446 training guides. The ranking (F1 ≫ F2 ≈ F5 ≫ F3 ≈ F4) closely mirrors the ablation-derived contribution pattern in panel A, despite being computed by an entirely independent method

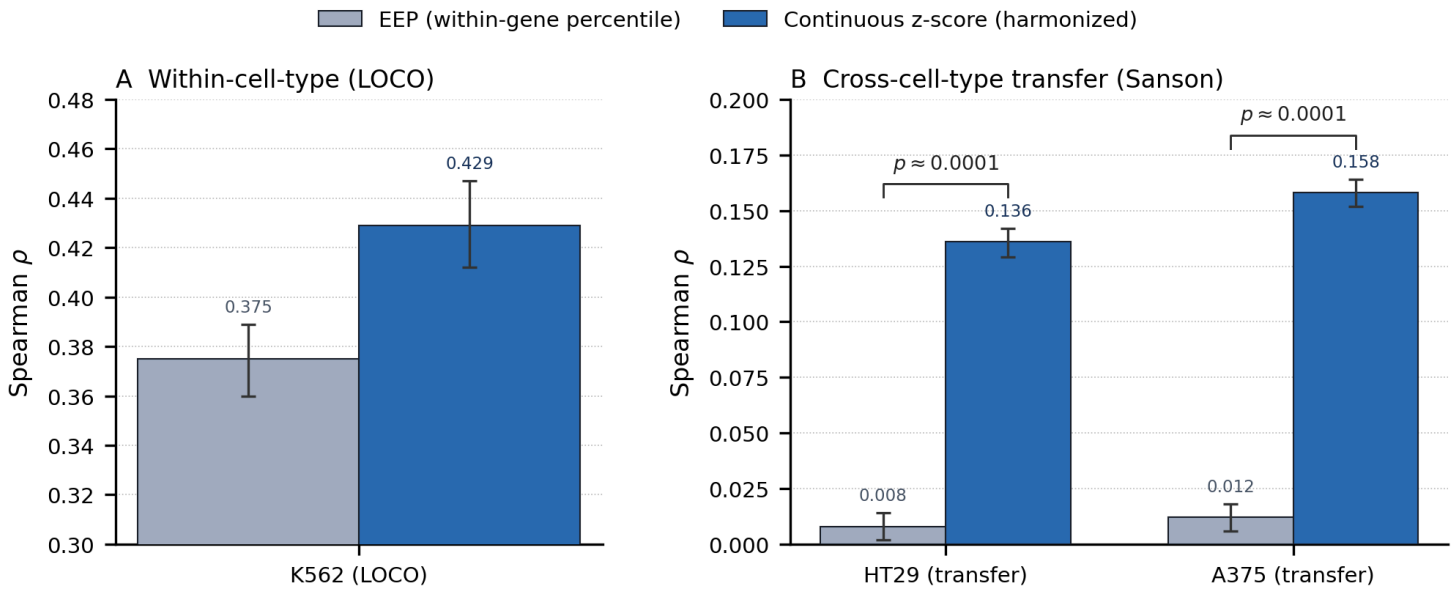


**Figure 3**

Target-label design determines cross-library transfer performance. Spearman ρ (± 95% bootstrap CI) for the EEP-labeled and continuous-z-score-labeled models, evaluated within the K562 training distribution (LOCO) and on external transfer (HT29, A375). The continuous-target model is superior in all three evaluation regimes; the improvement is modest within K562 but transformative in external transfer.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx