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.

Machine Learning-Based Prediction of Base Editor sgRNA fitness score

Alessandro Orro

alessandro.orro@cnr.it

National Research Council

Arianna Consiglio

National Research Council

Maria Ilaria Curci

National Research Council

Martina Scichilone

National Research Council

Faiza Hasin

University of Bari Aldo Moro

Michele Minervini

University of Bari Aldo Moro

Corrado Mencar

University of Bari Aldo Moro

Gianluca De Bellis

National Research Council

Cinzia Cocola

National Research Council

Paride Pelucchi

National Research Council

Tommaso Selmi

National Research Council

Article

Keywords:

Posted Date: April 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9224117/v1

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

Additional Declarations: No competing interests reported.

Machine Learning-Based Prediction of Base Editor sgRNA fitness score

Alessandro Orro[1] (*), Arianna Consiglio[2], Maria Ilaria Curci[1], Martina Scichilone[1,3], Faiza Hasin[2,4], Michele Minervini[4], Corrado Mencar[2,4], Gianluca De Bellis[1], Cinzia Cocola[1], Paride Pelucchi[1], Tommaso Selmi[1] (*)

1 Istituto di Tecnologie Biomediche Segrate, Consiglio Nazionale delle Ricerche, Segrate

2 Istituto di Tecnologie Biomediche Bari, Consiglio Nazionale delle Ricerche, Bari

3 Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano

4 Dipartimento di Informatica, Università di Bari Aldo Moro, Bari

(*) Corresponding authors

alessandro.orro@cnr.it

tommaso.selmi@cnr.it

Abstract

CRISPR Base Editors enable precise single-nucleotide modifications, offering advantages over CRISPR-Cas9 knock-out in programming the desired genetic effect. However, in pooled screens targeting essential genes, discrepancies between expected genetic and phenotypic outcomes are frequent: single guide RNAs (sgRNAs), expected to be disruptive, often appear “phenotypically silent” likely due to inefficient editing rather than absence of functional impact. Here, we investigate if Cas9-based gene-level sgRNA depletion data can help to predict the probability that an sgRNA used in base editing will yield the expected fitness effect in pooled proliferation screening.

We analysed proliferative effects (z-scores) from high-throughput CRISPR screens using cytosine Base Editors (BEs) and trained machine learning models to predict fitness effects. Our models integrate sequence features, edited strand, mutation type, predicted editing efficiencies and Cas9 gene essentiality scores.

Our models discriminate BE sgRNAs that generate a strong phenotypic effect (depletion) in pooled screening, with AUC-ROC greater than 93% in different cell lines. We provide exhaustive analysis of feature importance highlighting the significant impact of sequence features for predicting BE-associated fitness effects.

We found that editor-associated fitness predictions are primarily driven by sgRNA sequence features rather than predicted editing efficiency. Moreover, Cas9-derived gene essentiality partially contributes to predictions.

Introduction

Pooled CRISPR screens are a powerful reverse genetics approach to infer genotype–phenotype relationships at scale [1] [2] [3]. These approaches are increasingly used to study gene function and to discover molecular factors underlying disease. Originally established using CRISPR–Cas9–mediated gene knockout, pooled screens have enabled the systematic identification of essential genes and genetic dependencies based on the fitness effects of gene disruption and quantified using depletion-based metrics such as standardized z-scores. Large-scale efforts, including the DepMap project [4] [5], have leveraged Cas9 screening data across many cell lines to define gene essentiality as a reproducible, context-dependent fitness phenotype.

CRISPR Base Editors (BEs) represent a recent addition to the toolbox of pooled CRISPR screening technologies, enabling programmable single-nucleotide modifications without inducing double-strand breaks [6] [7] [Huang 2021]. This capability has opened new avenues for the functional interrogation of single-nucleotide variants through BE tiling screens, where guides are designed to introduce defined molecular changes along entire coding sequence, including splice site disruptions, nonsense mutations, missense substitutions, or synonymous edits. Annotation tools such as Variant Effect Predictor (VEP) [9] facilitate the design of base editing guide RNAs targeting specific variant classes. Pooled BE screens allow the investigation of fitness consequences associated with precise nucleotide perturbations, rather than complete gene knockout, providing a complementary view of genetic dependencies [11] [10] [12] [13].

However, the interpretation of BE screens is complicated by variable and context-dependent editing efficiency, which can result in “phenotypically silent” guides even when disruptive effects are expected. Compared to Cas9-mediated knockout, where double-strand breaks often produce strong and reproducible fitness phenotypes when essential genes or domains are targeted, BE operate within restricted editing windows and require precise modification of functionally critical nucleotides to elicit a phenotype. Therefore, BE–associated fitness effects might result weaker in phenotypic screening, and the absence of a phenotype may reflect inefficient editing rather than true absence of biological impact [Ryu2024] [15] [12].

To predict variability in BE outcomes, several sequence-based models have been developed to predict base editing efficiency from the local nucleotide context, including BE-Hive [16], DeepBE [17], and BE-DICT [18]. Trained primarily on surrogate target libraries, these approaches have demonstrated a good predictive power when applied to endogenous sites.

In this work, we use Machine Learning to integrate functional phenotypes from CRISPR Cas9 and cytosine base editing (CBE) screens with sequence-based descriptors to assess their combined ability to discriminate between BE guides associated with proliferative effects.

Material and methods

Dataset preparation

We used screening data from Hanna and colleagues, which tested a library of approximately 11,000 single guide RNAs (sgRNAs) [12] targeting 47 genes, including 10 pan-lethal essential genes (see Table 1). In details, the authors cloned sgRNA library into a lentiviral vector expressing BE3.9max [19] to enable base editing at endogenous loci. The resulting lentiviral library was then transduced into mammalian cells under conditions that ensure high coverage per sgRNA. Following transduction, the cells were cultured to allow BE editing or Cas9 mediated cleavage. The proliferative challenge was performed in independent experiments across five different cell lines (A375, HA1E, HAP1, MELJUSO and OVCAR8), each with two replicates. Sequencing data were then used to derive z-scores (fitness scores) reflecting sgRNA enrichment under the conditions tested.

Gene symbol	Ensembl gene ID	Gene symbol	Ensembl gene ID	Gene symbol	Ensembl gene ID
RNH1	ENSG00000023191	BCL2L1	ENSG00000171552	SLC25A28	ENSG00000155287
HRAS	ENSG00000174775	BRCA2	ENSG00000139618	YAP1	ENSG00000137693
CD81	ENSG00000110651	BAK1	ENSG00000030110	PAX8	ENSG00000125618
PEX10	ENSG00000157911	**PSMA6**	ENSG00000100902	NRAS	ENSG00000213281
EEF2	ENSG00000167658	SOX10	ENSG00000100146	STEAP3	ENSG00000115107
PELP1	ENSG00000141456	ERBB2	ENSG00000141736	HPRT1	ENSG00000165704
ICAM1	ENSG00000090339	PEX6	ENSG00000124587	MCL1	ENSG00000143384
ELOF1	ENSG00000130165	BRCA1	ENSG00000012048	ARNT	ENSG00000143437
NUDT5	ENSG00000165609	**POLR1C**	ENSG00000171453	TNFSF10	ENSG00000121858
FBXO42	ENSG00000037637	**KPNB1**	ENSG00000108424	NFE2L2	ENSG00000116044
SNRPD1	ENSG00000167088	BAX	ENSG00000087088	PIK3CA	ENSG00000121879
FOXA2	ENSG00000125798	CD33	ENSG00000105383	**TFRC**	ENSG00000072274
BCL2L2	ENSG00000129473	**RPS20**	ENSG00000008988	**SF3B1**	ENSG00000115524
KRAS	ENSG00000133703	PMAIP1	ENSG00000141682	CUL3	ENSG00000036257
NF2	ENSG00000186575	MED12	ENSG00000184634	**HNRNPU**	ENSG00000153187
NF1	ENSG00000196712	FAS	ENSG00000026103		

Table 1

The resulting data were further integrated with additional annotations. Gene essentiality was characterized using both functional domain information from PFAM [20] and proliferation z-scores from DepMap [4], a resource that integrates genome-scale CRISPR knockout screens across diverse cancer cell lines.

To assess the impact of base editing efficiency, predicted editing efficiencies were incorporated as additional features using three different tools: BE-Hive [16], a Gradient Boosting regressor model trained on large-scale datasets of cytosine and adenine base editors to predict editing efficiency with (Pearson's R = 0.7); DeepBE [17], a deep learning framework that integrates information from multiple Cas9 and BE variants to estimate editing outcomes and efficiencies (R

= 0.8-0.94) across diverse Protospacer adjacent motif (PAM) contexts; and BE-DICT [18], an attention-based neural network trained on tens of thousands of integrated targets to predict base editing outcomes for adenine and cytosine editors with high accuracy (accuracy 93.1-93.7% and AUC 93.8-94.7%).

Finally, sequence-based features were included (Table 2), such as the sgRNA sequence (see Figure 1), the strands of both sgRNA and target genes (-1 or +1), the on-target activity of sgRNAs estimated using the Rule Set 2 method [21], and the normalized distribution of temperatures (melting T) of the sgRNA sequence (range 0-1), calculated using the nearest-neighbour thermodynamic model (BioPython library [22]), which estimates the stability of the oligonucleotide, and the mutation type predicted by VEP, representing the functional consequence of the mutation on the gene or its encoded protein and categorized as 'No edits', 'Silent', 'Missense', 'Nonsense', 'Splice site', or 'Intron' and 'UTR'. The mutation type was encoded as a 7-length vector with one-shot encoding.

Feature	Description
RuleSet2	Prediction of on-target efficiency of sgRNAs (range 0-1).
DepMap	Gene essentiality defined by DepMap z-scores for cell lines A375, MELJUSO, OVCAR8, HAP1, HA1E)
BE efficiency prediction	BE efficiency prediction obtained by BE-DICT, BE-Hive and DeepBE tools.
BE3.9 z-score	Experimental BE z-scores for the five experiments with cell lines A375, MELJUSO, OVCAR8, HAP1 and HA1E
GC content (%)	Percent of GC in the target context sequence (the sequence containing the sgRNA)
Melting T (°C)	Sequence melting temperature
Mutation type	Classification of the target cytosine mutation induced by the BE (one of "Intron", "Missense", "No edits", "Nonsense", "Silent", "Splice site" and "UTR")
PFAM domain	Indicates if the target mutation is inside a known domain
sgRNA Strand	Indicates which DNA strand the guide RNA binds to (sense or antisense)
Gene Strand	Gene Strand annotation based on Ensembl (positive or negative)

Table 2

The model output is defined as the probability that a gene shows a proliferation z-score lower than −2 in the sequencing experiment. Accordingly, the target variable was encoded as a binary value, where 1 corresponds to z-scores < −2 and 0 otherwise.

Machine learning experiment setup

Sequence features of the sgRNAs were extracted from a context sequence of variable length (L) composed of the 20-nt sgRNA and the flanking regions consisting of upstream (U) and downstream (D) nucleotides, including the PAM (Figure 1).

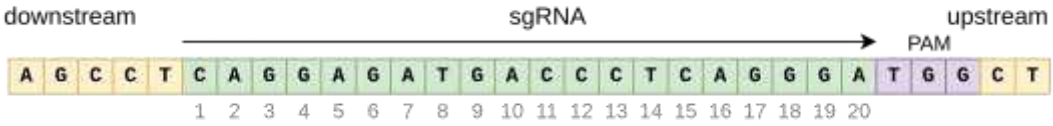


Figure 1

Sequence encoding tailed dinucleotide one-hot schema, where overlapping consecutive k-mers (k=2) were extracted from each guide sequence and encoded as a 16-dimensional binary vector, resulting in a (L-1) x 16 array, with L= (20, 30, 40 or 50) based on the sequence length. This encoding preserves local sequence context and captures compositional patterns. In addition, strand orientation of both gene and corresponding sgRNA was incorporated as categorical features. The combined representation encodes both sequence and strand-specific information, providing machine learning models with a rich input to detect subtle positional dependencies.

All features are, by definition, normalized to values in the range (-1, +1) or (0, +1), and therefore no further scaling or normalization was applied prior to model training. We evaluated several machine learning models to predict the effect of sgRNAs in terms of their ability to capture BE activity, including the python sklearn [23] implementation of Gaussian Naive Bayes (GaussianNB), standard Decision Tree (DT) and Random Forest (RF), XGBoost [24] implementation of Gradient Boosting over Decision Trees (XGB), CatBoost [25] and the LightGBM [26] implementation of Light Gradient Boosting Machine (LightGBM). All experiments were performed with default hyper parameters of the cited libraries.

Each model was independently trained and tested using a 5-fold cross-validation scheme to ensure robust estimation of predictive performance. To reduce the effect of variability in model training and data partitioning, each experiment was repeated 10 times with different random seeds. Performance metrics, including Accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), were computed for each fold and each replicate. Mean and standard deviations were reported to reflect variability across folds and replicates. Feature importance was computed for the best performing models to assess contributions of individual features to the predictive outcome using the SHapley Additive exPlanations (SHAP) method [27] [28] that computes, for each input feature, how much the overall prediction changes in response to a change in the feature value. It computes the SHAP values as the average contribution of each feature in shifting the model output toward positive or negative predictions. Alternatively, feature importance was also evaluated using a form of ablation analysis in which each feature is permuted in the validation dataset to assess its impact on model performance. The importance of a feature is then defined as the decrease in the evaluation metric compared to the baseline performance with AUC-ROC. The ablation analysis was performed on the best-performing model with AUC-ROC as evaluation metric.

ROC and precision-recall curves were plotted to visualize model discriminative ability. The cross-validation setup ensured that no data leakage occurred between training and test sets. All scripts and configurations were version-controlled to guarantee reproducibility of the results. The combined evaluation allowed us to identify models that consistently achieved high predictive

performance across folds and replicates, providing confidence in their applicability to variant effect prediction.

Although sequence information appeared to be the most informative feature, we conducted experiments using an increasing number of features, starting from sequence-only encoding and progressively incorporating all additional features.

Results

Motivations

By focusing on recently published pooled screening datasets comparing Cas9 and Base Editor (BE) disrupting activity on the same set of essential genes, we noticed that Cas9 and BE yielded significantly different fitness effects. In particular, we noticed that the distributions of proliferation scores (Z-scores) obtained with Cas9 are consistently shifted toward stronger negative values compared to those derived from BE from the same screen dataset (Hannah) (Figure 2). This was surprising as, given the high editing efficiency of both systems, we expected strong quantitative correlation between Cas9- and BE-derived fitness scores at single sgRNA level. We hypothesize that the observed discrepancies likely reflect fundamental differences between full gene knockout induced by Cas9-mediated double-strand breaks and the more subtle, strand-specific single-nucleotide changes and nicks introduced by BEs. In addition, variable and context-dependent editing efficiency of deaminase enzymes in BEs may further separate screen phenotypic readouts from the intended genetic perturbation [8] [11] [12]. Together, these observations motivate a systematic investigation of Cas9–BE fitness score discrepancies as a source of potential information to improve the predictive accuracy of machine learning models and the interpretation of the fitness readouts from BE pooled screens.

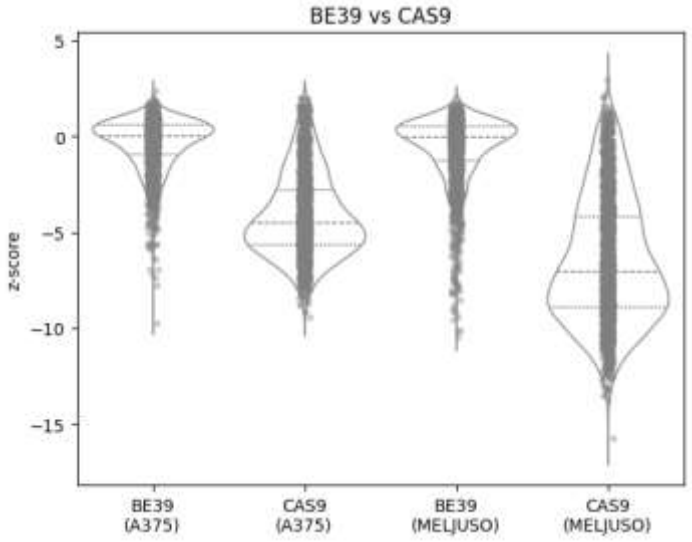


Figure 2

The experiments in this study were primarily designed to evaluate the ability of different machine learning models to predict the probability that an sgRNA produces a detectable fitness effect (Z-scores) in pooled BE screens. This setup enables a systematic comparison of established

machine learning approaches in terms of their predictive accuracy and AUC-ROC across different cell lines. As a secondary objective, we aimed at analysing the contribution of individual features to model performance. Although they do not show an evident correlation with the output variable from a purely statistical perspective, their combined effect may still encode complex relationships that are detectable by machine learning models.

Description of features

For input features we calculated mean, variance and Pearson R correlation for numerical feature and counts, percentage and Odd Ratio (OR) for categorical ones.

Table 3 summarizes statistics of input features. As expected, none of the features exhibited, by its own, a strong predictive relationship with the proliferation effect (z-scores) when considered alone. This reflects the inherent biological complexity underlying both BE activity and the pooled screen experimental protocol. The relatively low correlations (R and OR values) further confirm that no individual variable can account for the observed variability in proliferative outcomes. This justifies the adoption of more powerful statistical approaches, such as machine learning models, which are better suited to capture nonlinear dependencies and complex feature interactions.

Model performance across cell lines

We first evaluated the predictive performance on a set of relevant state-of-the-art Machine Learning models described in M&M section across the five cell lines (A375, HA1E, HAP1, MELJUSO, and OVCAR8) using proliferation z-scores binarized with threshold –2 so that positive sgRNAs are defined as those exhibiting a strong effect.

Overall, machine learning approaches achieved variable but consistent predictive performance, with AUC-ROC ranging from 83.8% to 94.8% and Accuracy ranging from 78.3% to 93.7% depending on the model and cell line (Figure 3 and Table 4). Among them, LightGBM, XGBoost and CatBoost showed the highest overall performance in all cell lines whereas Naive Bayes (GaussianNB) and Decision Tree showed limited performances. As our input data consists of fitness scores derived from a saturation mutagenesis experiment, where a small number of genes (N=47) was saturated by numerous sgRNAs (11,000), we reasoned that our model might be biased toward learning gene-identity features rather than more “transferable” sgRNA rules impacting on fitness scores. In order to check this effect, before proceeding with the interpretation of the results, we run a leave one out validation at gene level where each gene and the relative sgRNAs were removed from the training set at each fold. The AUC-ROC (Z-score<-2 or not) of the LightGBM prediction model dropped from 94.7 to 93.3 for A375 (1.4%), and from 93.5 to 91.3 (2.2%) for MELJUSO cell line. We therefore can infer that gene-identity features only marginally impact on accuracy predictions and sgRNA-centric fetures must provide a significant contribution to the model accuracy.

		Pearson R correlation				
Feature	**Avg ± Std**	**A375**	**MELJUSO**	**OVCAR8**	**HAP1**	**HA1E**
RuleSet2	0.49 ± 0.12	0.07	0.07	0.06	0.06	0.07

DepMap (A375)	-0.51± 0.72	-0.06	-0.09	-0.09	-0.11	-0.06
DepMap (MELJUSO)	-0.68 ± 0.76	-0.02	-0.06	-0.05	-0.06	-0.04
DepMap (OVCAR8)	-0.61 ± 0.77	-0.05	-0.09	-0.11	-0.11	-0.08
DepMap (HAP1)	-0.52 ± 0.69	-0.07	-0.10	-0.11	-0.14	-0.09
DepMap (HA1E)	-0.44 ± 0.76	0.00	-0.03	-0.04	-0.05	-0.04
Be-DICT efficiency	0.77 ± 0.13	0.08	0.07	0.07	0.07	0.09
Be-Hive efficiency	0.49 ± 0.15	-0.05	-0.02	-0.05	-0.02	-0.01
DeepBE efficiency	0.55 ± 0.13	0.02	0.03	0.02	0.00	0.02
BE3.9 z-score (A375)	-0.20 ± 1.22	-0.73	-0.65	-0.66	-0.63	-0.63
BE3.9 z-score (MELJUSO)	-0.29 ± 1.33	-0.64	-0.76	-0.70	-0.70	-0.66
BE3.9 z-score (OVCAR8)	-0.29 ± 1.29	-0.66	-0.71	-0.77	-0.72	-0.69
BE3.9 z-score (HAP1)	-0.37 ± 1.46	-0.61	-0.68	-0.69	-0.77	-0.67
BE3.9 z-score (HA1E)	-0.27 ± 1.37	-0.66	-0.70	-0.72	-0.72	-0.77
GC content (%)	55.44 ± 13.90	-0.34	-0.33	-0.36	-0.34	-0.30
Melting T (°C)	54.08 ± 6.18	-0.33	-0.32	-0.36	-0.34	-0.30

		OR				
	Count (%)	**A375**	**MELJUSO**	**OVCAR8**	**HAP1**	**HA1E**
Mutation type						
Intron	364 (3.30%)	1.07	0.83	1.10	0.84	0.82
Missense	4017 (36.43%)	0.91	0.89	0.90	0.90	0.83
No edits	4172 (37.83%)	0.49	0.44	0.49	0.46	0.39
Nonsense	489 (4.4%)	1.48	2.72	2.34	2.98	2.06
Silent	1794 (16.27%)	0.66	0.57	0.54	0.56	0.48
Splice site	150 (1.36%)	1.76	2.35	1.88	2.55	2.16
UTR	42 (0.38%)	0.17	0.14	0.15	0.13	0.30
PFAM domains	9468 (85.85%)	0.48	0.52	0.51	0.51	0.48
sgRNA Strand sense	5460 (49.51%)	0.52	0.56	0.55	0.56	0.52
sgRNA Strand antisense	5568 (50.49%)	0.43	0.43	0.44	0.45	0.49
Gene Strand positive	5041 (45.71%)	0.53	0.53	0.56	0.55	0.54
Gene Strand negative	5987 (54.29%)	0.43	0.47	0.45	0.49	0.43

Table 3

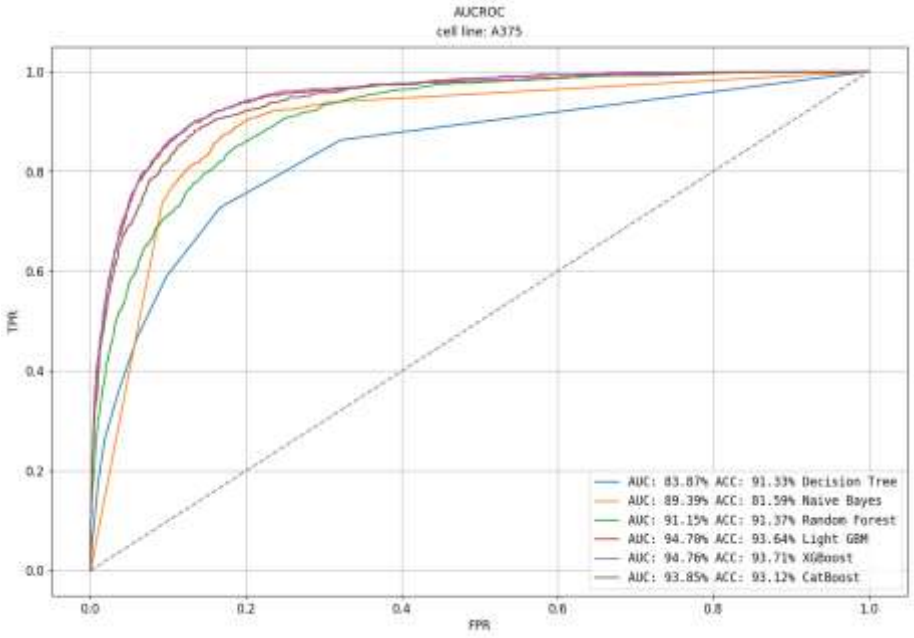


Figure 3

	AUC (%) / ACC (%)				
Feature	**A375**	**MELJUSO**	**OVCAR8**	**HAP1**	**HA1E**
Decision Tree	83.87 / 91.33	84.67 / 90.52	86.21 / 90.42	84.28 / 89.67	83.77 / 91.01
Naive Bayes	89.39 / 81.59	86.61 / 79.56	88.38 / 81.52	85.15 / 78.30	86.68 / 80.83
Random Forest	91.15 / 91.37	90.44 / 89.96	91.49 / 90.59	90.33 / 89.71	89.99 / 90.81
LightGBM	94.70 / 93.64	93.52 / **92.66**	**94.77** / **93.19**	93.21 / 92.05	**93.88** / **93.20**
XGBoost	**94.76** / **93.71**	**93.56** / 92.52	94.74 / 93.14	**93.43** / **92.10**	93.83 / 93.18
CatBoost	93.85 / 93.12	93.00 / 92.32	94.17 / 92.85	92.96 / 91.97	93.21 / 92.89

Table 4

Interpretability analysis

To quantify the contribution of individual features to model predictions, we employed SHAP (SHapley Additive exPlanations) values using the TreeExplainer, a model-agnostic approach specifically optimized for tree-based models such as gradient boosting. TreeExplainer computes Shapley values from cooperative game theory, estimating for each feature the marginal contribution to the predicted probability while accounting for interactions with other features. Positive SHAP values indicate a feature pushing the prediction toward the positive class, while negative values indicate a push toward the negative class. For each cell line, the model (LightGBM) was trained on multiple replicates and cross-validation folds, and SHAP values were calculated on held-out validation sets. Absolute SHAP values were averaged across folds and replicates to obtain a robust estimate of feature importance. Sequence-related features were subsequently aggregated to represent nucleotide-level contributions along the 20-nt sequence, enabling position-wise interpretation of model behavior. This procedure provides a quantitative measure of feature contributions at both the individual and grouped feature levels.

The results for A375 cell line (Figure 4) reveal a clear pattern in which sequence positions toward the 3' end of the sgRNA exhibit the most predictive signal. Specifically, early positions (pos:1–pos:5) contribute moderately (0.065–0.068), while importance steadily increases through the central positions, peaking in the last positions: pos:19 (0.237) and pos:20 (0.328). Mid-sequence positions (e.g., pos:13–pos:16) show substantial contributions (0.139–0.304), indicating that the model relies heavily on these positions for prediction.

Among non-sequence features, gene strand emerges as the most influential (0.493), followed by DepMap essentiality scores (0.179) and mutation type (0.101). Other features such as BE-Hive (0.089), sgRNA strand (0.056), and RuleSet2 score (0.027) show moderate contributions, while features like PFAM (0.003), BE-DICT (0.017), DeepBE (0.022), and melting temperature (0.011) are relatively minor. These results suggest that both the sequence composition of the sgRNA and certain biological context features (strand, essentiality, mutation type) are key determinants for model predictions. Similar results and considerations hold for other cell lines as reported in Figure 5 only for top 10 features.

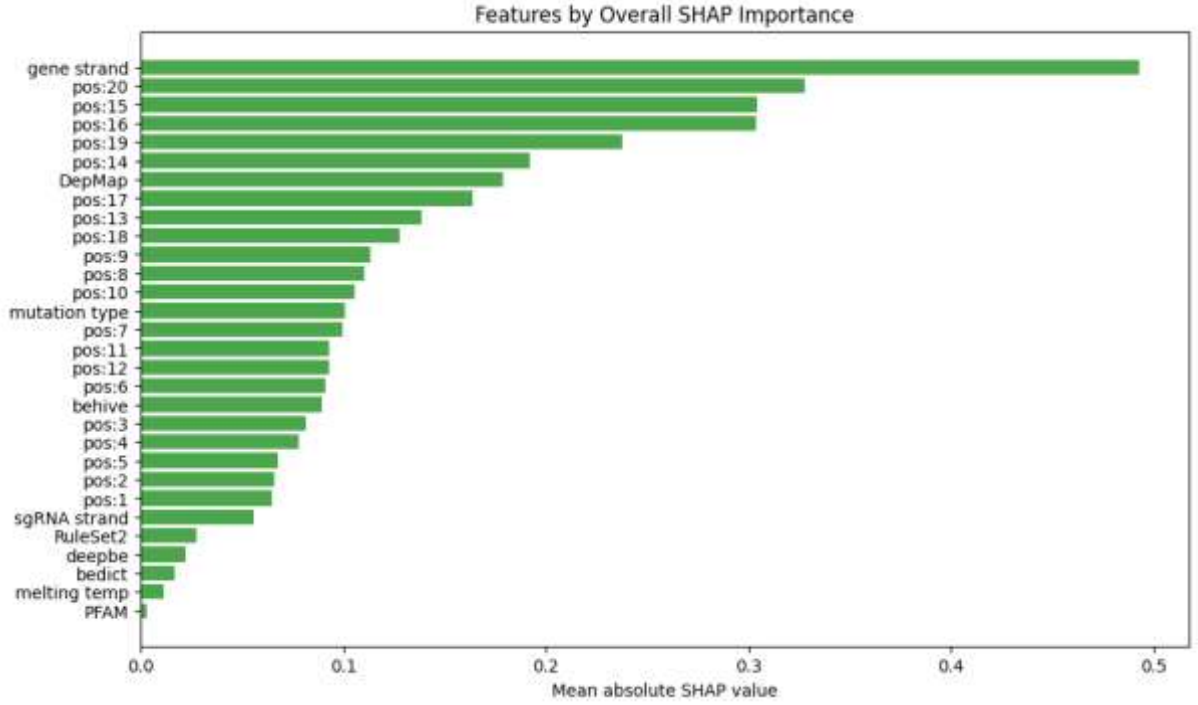


Figure 4

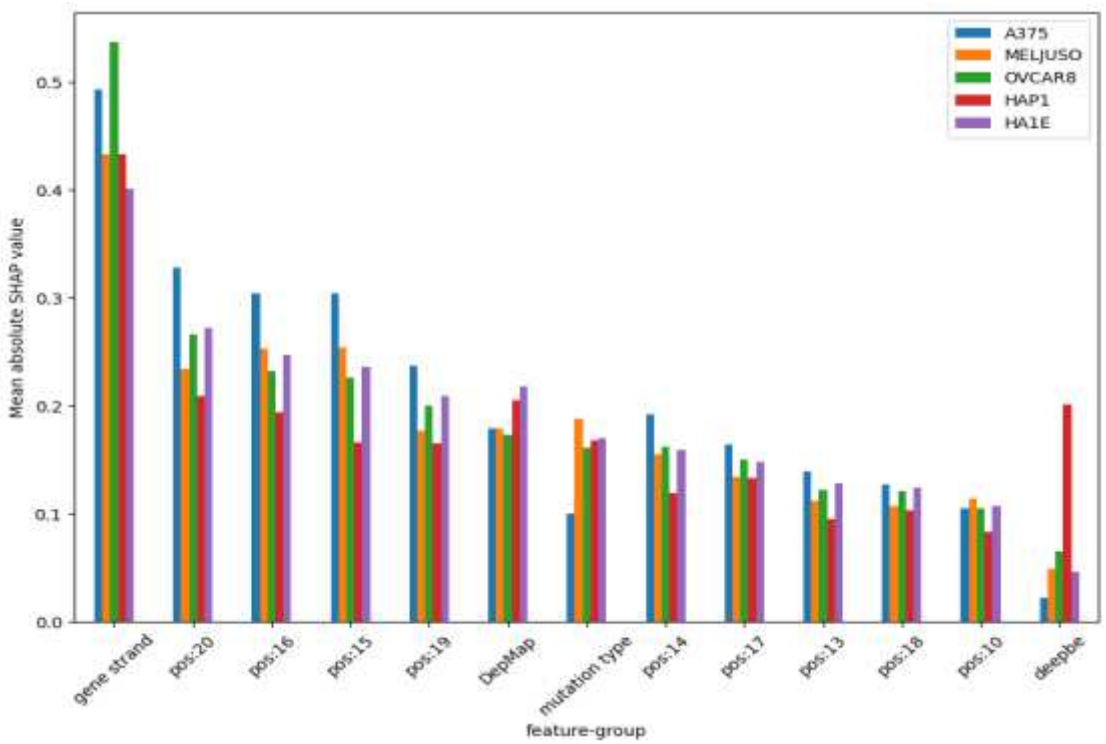


Figure 5

Complementary to the SHAP analysis, we investigated which sources of information most strongly influence the prediction of BE sgRNA effectiveness using a permutation-based feature importance approach. The analysis was conducted with a LightGBM classifier, selected for interpretability because it consistently ranked among the best-performing models across all cell lines, together with XGBoost and CatBoost, achieving comparable AUC–ROC values. For simplicity, we report results obtained with LightGBM only. Feature importance was quantified as the decrease of (drop) in AUC–ROC observed upon permuting each feature group on the validation set, providing a direct measure of the performance loss associated with the deletion of specific sources of information. Importance values were averaged across cross-validation folds and experimental replicates, separately for each cell line. Results of top five features are reported in Figure 6.

Across all five cell lines, sequence-derived features that include the contribution of all sequence positions (called “sequence” in Figure 6) consistently emerged as the most informative group, yielding the largest decrease in predictive performance when permuted, with an average AUC–ROC drop of approximately 15% (range: 11–19%). Gene strand information also showed a strong and stable contribution, ranking among the top two feature groups in every cell line and inducing an average performance drop of about 13% (range: 11–15%).

In contrast, mutation-related features (called “mutation type” in Figure 6) exhibited a moderate but non-negligible effect, with an average AUC–ROC decrease of ~2.0% (range: 0.5–2.8%), consistently exceeding the contribution of gene-essentiality information derived from DepMap, which showed a smaller yet reproducible impact (mean ~1.0%, range: 0.7–1.3%). BE-specific efficiency predictors such as BE-Hive contributed more modestly overall (mean ~0.3%), with

their relative importance varying across cell lines; notably, BE-Hive was the only base-editing efficiency predictor to appear among the top five feature groups.

While sequence and strand features dominated the importance rankings in all settings, the relative contributions of mutation type, essentiality (DepMap), and BE-specific predictors differed between cell lines, suggesting a degree of context dependency in how base-editing outcomes are determined. Overall, these results highlight that sgRNA efficiency prediction is primarily driven by local sequence and strand context, while mutation-level annotations and gene-level dependencies provide additional, cell line–dependent predictive signal. This feature importance landscape supports the use of integrative models that combine sequence, genomic context, and functional annotations to capture the multifactorial nature of base editor activity.

Finally, we assessed with a similar approach the feature importance along the nucleotide sequence of sgRNA for each cell line by considering only the contributions of sequence and separating the feature by sequence position (Figure 7). Overall, the feature importance profiles exhibited a consistent pattern across the five cell lines. For A375, the importance gradually increased along the sequence, with a peak at the last positions, reaching a maximum value of 0.0234 at position 20. Similarly, MELJUSO showed low importance at the initial positions, with a substantial increase from positions 15–20, culminating at 0.0162 at position 20. OVCAR8 displayed a more gradual increase along the sequence, with peaks at positions 16–20, reaching 0.0171 at the final position. HAP1 and HA1E showed analogous trends, with minor fluctuations in the middle positions but consistent high importance toward the 3' end of the sequence, reaching 0.0152 (HAP1) and 0.0212 (HA1E) at position 20.

Notably, positions in the middle of the sequence (around positions 10–16) generally contributed moderately to the overall predictive signal, while the very first positions exhibited low feature importance across all cell lines. The standard deviations across permutation repeats were generally small, indicating robust estimates of importance values.

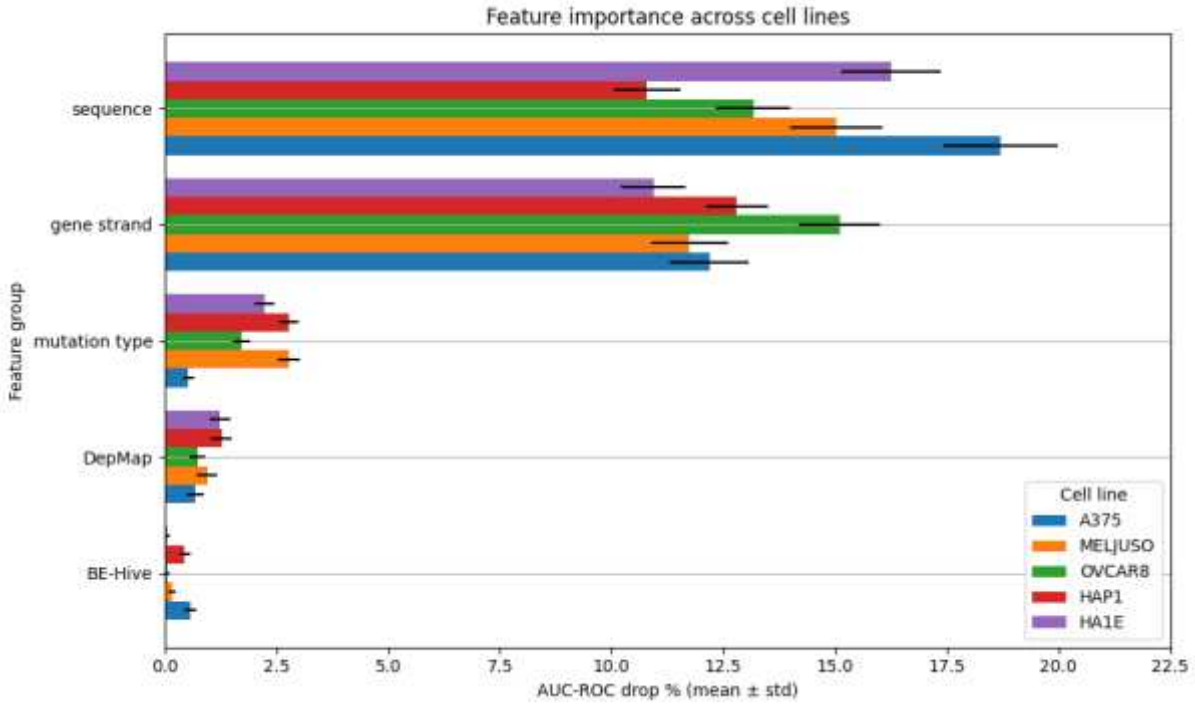


Figure 6

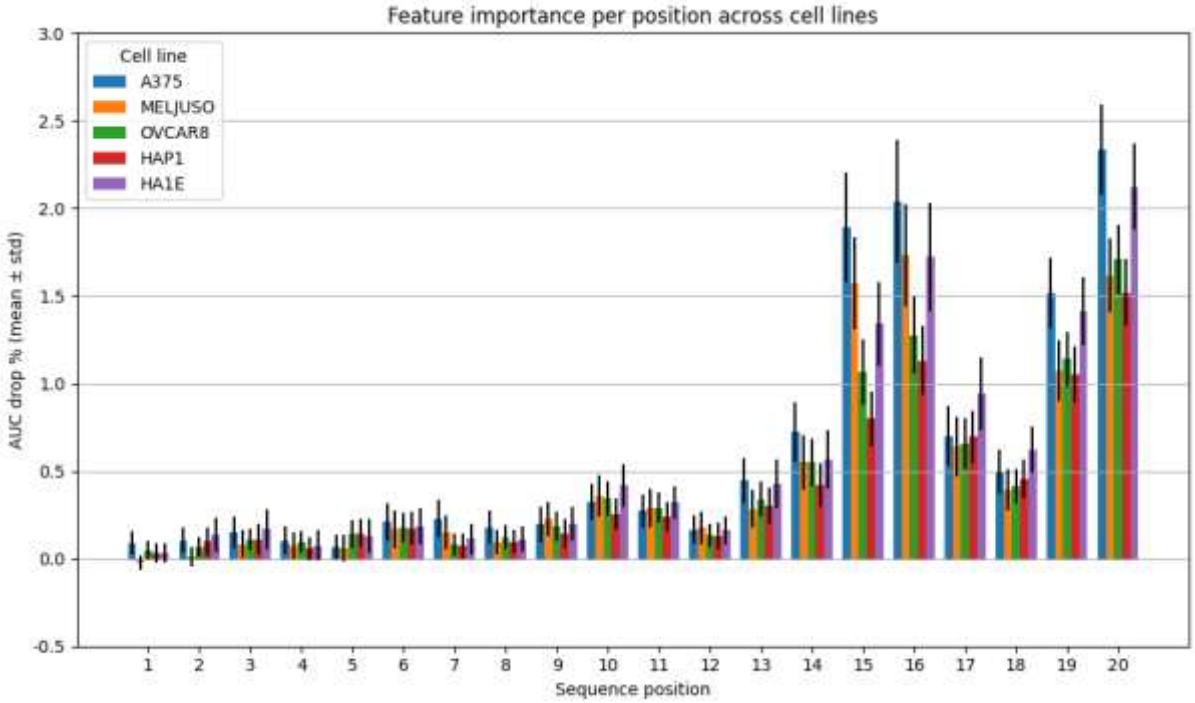


Figure 7

Relationship between BE and CAS9

In order to gain a deeper insight on the different mechanism of actions of BE and Cas9 we modelled the difference in z-scores between BE3.9 and Cas9 (Δz = BE3.9 z-score – Cas9 z-score) using a Linear Regression that included gene essentiality (DepMap score), gene and sgRNA strand, predicted editing efficiencies (Be-Dict, Be-Hive, DeepBE), RuleSet2 score, mutation type,

and the presence of editable cytosines at each position of the guide (positions 1–20). Results show that gene essentiality is the primary determinant of Δz, with higher DepMap scores associated with more negative Δz values (coef = -1.34), reflecting stronger fitness impact by BE3.9 relative to Cas9. Predicted efficiencies from Be-Hive and DeepBE positively correlate with Δz (coefs = 2.22 and 1.42, respectively), suggesting that guides predicted to be highly efficient partially counteract the general essentiality effect. Mutation types generally show small effects, although the absence of edits ("No edits") slightly increases Δz. Importantly, the presence of editable cytosines at specific positions within the guide modulates Δz in a position-dependent manner. For example, cytosines at positions 3, 6, 7, 14, 16, and 19 reduce Δz by 0.16–0.33 (in term of coeffs), highlighting these positions as key determinants of BE3.9 activity. Other positions have smaller or non-significant effects. Overall, while gene essentiality drives the global difference between BE3.9 and Cas9 activity, the precise positioning of editable cytosines and predicted guide efficiency further refines the observed differences. Figure 8 shows the grouped Δz coefficients of the regression, where features were aggregated into biologically meaningful categories, while Table 5 reports the contributions of individual features.

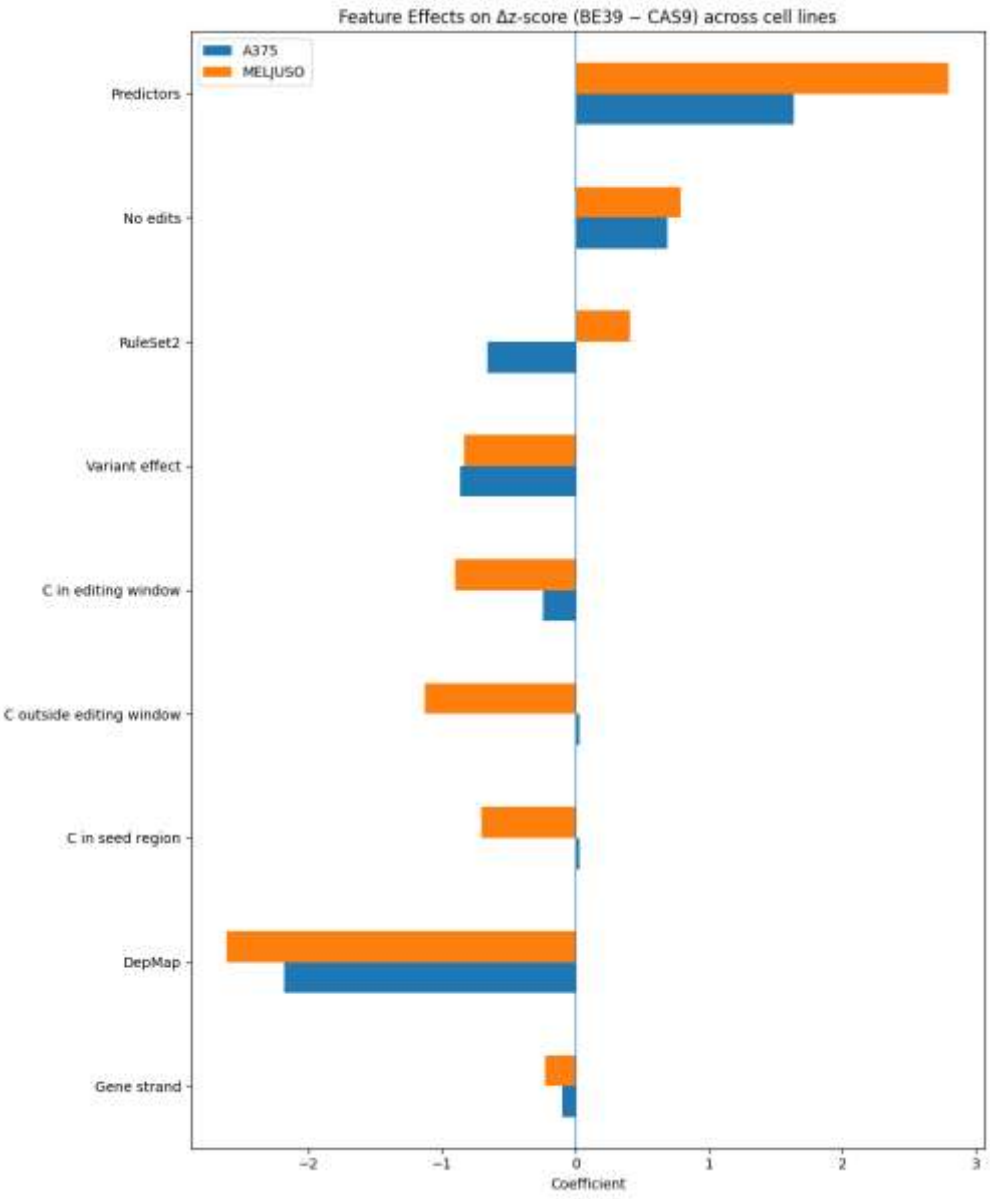


Figure 8

Feature	A375	MELJUSO	Group
DepMap	-2.18	-2.61	DepMap
Gene strand	-0.10	-0.23	Gene strand
Be-Dict	-1.27	-0.38	Predictors
Be-Hive	2.43	2.02	Predictors
DeepBE	0.47	1.16	Predictors
RuleSet2	-0.66	0.40	RuleSet2
No edits	0.68	0.79	No edits
Silent	0.09	0.30	Variant effect
Missense	-0.17	0.03	Variant effect
Nonsense	-0.21	-0.52	Variant effect
Splice site	-0.58	-0.65	Variant effect
Intron	0.01	0.48	

UTR	-0.24	-1.00	
pos1	-0.14	-0.14	C outside editing window
pos2	0.12	-0.05	C outside editing window
pos3	-0.13	-0.20	
pos4	0.11	-0.10	
pos5	-0.10	-0.24	
pos6	-0.15	-0.27	
pos7	-0.18	-0.24	
pos8	0.08	-0.05	C outside editing window
pos9	-0.17	-0.26	C outside editing window
pos10	-0.08	-0.22	C outside editing window
pos11	0.07	0.02	C outside editing window
pos12	0.08	-0.16	C outside editing window
pos13	0.17	-0.08	C outside editing window
pos14	-0.02	-0.25	C outside editing window
pos15	0.22	0.13	C in seed region
pos16	-0.29	-0.37	C in seed region
pos17	0.23	0.18	C in seed region
pos18	0.21	-0.01	C in seed region
pos19	-0.13	-0.34	C in seed region
pos20	-0.21	-0.30	C in seed region

Table 5

Discussion and Conclusion

The recent emergence of programmable genome editing tools such as CRISPR Base Editors enabled the precise single nucleotide modification of the genome, allowing the targeting of specific loci for the effective treatment of genetic diseases [Xu2025]. Therefore, base-editing efficiencies and outcomes are typically assessed experimentally before or during disease model generation and therapeutic correction of pathogenic variants. Nonetheless, such experimental evaluation is both time-consuming and costly [30]. That's why recent efforts have focused on developing computational models trained on empirical data to predict sgRNA efficiency in a cost-effect manner [31], such as DeepBE, BE-DICT and BE-Hive. In this work, we propose explainable machine learning models that integrate multiple features to predict the fitness score of base editors, thereby estimating the functional impact of BE sgRNAs on cell proliferation prior to experimental screening.

Our approach differs from previous attempts in the fact that each tested model was trained for prediction of sgRNA fitness effects (z-score) from pooled screening, rather than predicted editing efficiency experiments. Our models were integrated with information on gene essentiality, predicted editing efficiencies from DeepBE, BE-DICT and BE-Hive, sequence-based features, sgRNA and target-gene strand, the on-target activity scores from Rule Set 2, melting temperature and mutation type. Subsequently, we quantified the impact of selected biological features on the prediction of a complete biological experiment where the genomic editing must also correlate with a phenotypic (fitness) effect, which in pooled screening experiments is usually expressed as

a z-score. Depending on the cell line and model type, our model reached good predictive performance, with AUC-ROC values ranging from 83.8% to 94.8%, among which XGBoost showing the most robust results.

Sequence-based features predominantly contributed to classification performance. Combined, gene strand and sequence features accounted for more than 30% accuracy across cell lines. The strong influence of nucleotides proximal to the PAM position (19-20) likely reflects their key role in target recognition. Structural studies have demonstrated that "seed nucleation" happens at nucleotides 19-20 and initiates target DNA binding by the sgRNA. After that, nucleotides 15-18 are involved in the formation of a "partially active conformation" of the Cas9 complex which is key for the stabilisation of the R-loop and a key determinant of Cas9 activation and engagement, ultimately affecting editing efficiency [32]. Recent machine learning approaches [33] also confirmed the prominent role of PAM-proximal sequence features (15-20) in Cas9 efficiency. The consistent importance of target-gene strand might be viewed as a key complement to sequence features, but also reflect the known directional bias in CRISPR Cas efficiency. Indeed, CRISPR Cytosine Base editors operate through directional nicking of the DNA to guide mismatch repair on the edited strand [6], while directional replication activity impacts Cas9 editing outcomes [34]. Therefore, sequence and strand features might encode multiple key ground Cas9 processes which heavily influence fitness penetrance in the screening.

The third most influential feature contributing to model performance (about 2%) is gene essentiality, derived from DepMap Cas9 fitness scores (DepMap). This shows that gene essentiality is a feature that partially complement sequence-based features, possibly by providing "a priori" key fitness information for each sgRNA.

Other feature contributions were limited. In particular, the model showed minimal dependency on base editing predictors. Editing efficiency predictors perform robustly at modelling editing outcomes, however this process appears not limiting for fitness effect prediction in this dataset. Gene functional annotation (PFAM domain) also appeared not limiting, with their encoded binary information that could be redundant by VEP and gene essentiality features.

Lastly, with the Δz regression analysis we tried to asses whether features (mutation type, predicted editing efficiency) that appeared less important for the predictive task in the machine learning model were instead actually responsible for the divergence observed between Cas9 and CBE scores. Again our results provided evidence that gene-level information behaves as the main driver of the discrepancy between Cas9 and CBE scores. This again reinforces the idea that Cas9 phenotypic effect might be more penetrant compared to base editing, even when assessed on the same guide in the same screening experiment.

In conclusion, while our study aimed at identifying features that could explain the obvious difference in fitness scores recorded between Base Editor and Cas9 data, we found that the best performing ML model rely on broader, CRISPR-specific DNA-targeting constraints to achieve good prediction scores. The divergence between Cas9 and CBE fitness penetrance is only partially explained by the features we provided the model, highlighting limitations of current feature engineering and modelling approaches. This suggests that additional information layers such as "real" endogenous editing outcomes and DNA repair mechanisms might be required as inputs to recapitulate the base editor behavior.

Acknowledgements

This research, as well as F.H. and M.S. PhD fellowships were supported the European Union-Next Generation EU (PNRR National Center for Gene Therapy and Drugs based on RNA Technology CN00000041, CUP B83C22002860006).

Data availability

The dataset used in this work is available in the supplementary materials of [12].

References

[1] Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G. and Zhang, F., 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science, 343(6166), pp.84-87.

[2] Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B. and Stanley, G., 2022. High-content CRISPR screening. Nature Reviews Methods Primers, 2(1), p.8.

[3] Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S. and Mero, P., 2015. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell, 163(6), pp.1515-1526.

[4] Arafeh, Rand, Tsukasa Shibue, Joshua M. Dempster, William C. Hahn, and Francisca Vazquez. "The present and future of the Cancer Dependency Map." Nature Reviews Cancer 25, no. 1 (2025): 59-73.

[5] Dwane, L., Behan, F.M., Gonçalves, E., Lightfoot, H., Yang, W., Van Der Meer, D., Shepherd, R., Pignatelli, M., Iorio, F. and Garnett, M.J., 2021. Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. Nucleic Acids Research, 49(D1), pp. D1365-D1372.

[6] Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. and Liu, D.R., 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature, 533(7603), pp.420-424.

[7] Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I. and Liu, D.R., 2017. Programmable base editing of A• T to G• C in genomic DNA without DNA cleavage. Nature, 551(7681), pp.464-471.

[8] Huang, T.P., Newby, G.A. and Liu, D.R., 2021. Precision genome editing using cytosine and adenine base editors in mammalian cells. Nature protocols, 16(2), pp.1089-1128.

[9] McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F., 2016. The ensembl variant effect predictor. Genome biology, 17(1), p.122.

[10] Coelho, M.A., Strauss, M.E., Watterson, A., Cooper, S., Bhosle, S., Illuzzi, G., Karakoc, E., Dinçer, C., Vieira, S.F., Sharma, M. and Moullet, M., 2024. Base editing screens define the genetic landscape of cancer drug resistance mechanisms. Nature Genetics, 56(11), pp.2479-2492.

[11] Coelho, M.A., Cooper, S., Strauss, M.E., Karakoc, E., Bhosle, S., Goncalves, E., Picco, G., Burgold, T., Cattaneo, C.M., Veninga, V. and Consonni, S., 2023. Base editing screens map mutations affecting interferon-γ signaling in cancer. Cancer Cell, 41(2), pp.288-303.

[12] Hanna, Ruth E., Mudra Hegde, Christian R. Fagre, Peter C. DeWeirdt, Annabel K. Sangree, Zsofia Szegletes, Audrey Griffith et al. "Massively parallel assessment of human variants with base editor screens." Cell 184, no. 4 (2021): 1064-1080.

[13] Cuella-Martin, R., Hayward, S.B., Fan, X., Chen, X., Huang, J.W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C. and Shen, Y., 2021. Functional interrogation of DNA damage response variants with base editing screens. Cell, 184(4), pp.1081-1097.

[14] Ryu, J., Barkal, S., Yu, T., Jankowiak, M., Zhou, Y., Francoeur, M., Phan, Q.V., Li, Z., Tognon, M., Brown, L. and Love, M.I., 2024. Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification. Nature genetics, 56(5), pp.925-937.

[15] Sokirniy, I., Inam, H., Tomaszkiewicz, M., Reynolds, J., McCandlish, D. and Pritchard, J., 2025. A side-by-side comparison of variant function measurements using deep mutational scanning and base editing. Nucleic acids research, 53(14), p.gkaf738.

[16] Arbab, Mandana, 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, no. 2 (2020): 463-480.

[17] Kim, Nahye, 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, no. 3 (2024): 484-497.

[18] Marquart, Kim F., 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, no. 1 (2021): 5114.

[19] Koblan, Luke W., Jordan L. Doman, Christopher Wilson, Jonathan M. Levy, Tristan Tay, Gregory A. Newby, Juan Pablo Maianti, Aditya Raguram, and David R. Liu. "Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction." Nature biotechnology 36, no. 9 (2018): 843-846.

[20] Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna et al. "The Pfam protein families database." Nucleic acids research 32, no. suppl_1 (2004): D138-D141.

[21] Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith et al. "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9." Nature biotechnology 34, no. 2 (2016): 184-191.

[22] Cock, Peter JA, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics 25, no. 11 (2009): 1422.

[23] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[24] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

[25] Dorogush, A.V., Ershov, V. and Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

[26] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[27] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[28] Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." In International Conference on artificial intelligence and statistics, pp. 2907-2916. PMLR, 2020.

[29] Xu, W., Zhang, S., Qin, H. and Yao, K., 2024. From bench to bedside: cutting-edge applications of base editing and prime editing in precision medicine. Journal of Translational Medicine, 22(1), p.1133.

[30] Song, M., Kim, H.K., Lee, S., Kim, Y., Seo, S.Y., Park, J., Choi, J.W., Jang, H., Shin, J.H., Min, S. and Quan, Z., 2020. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. Nature biotechnology, 38(9), pp.1037-1043.

[31] Alipanahi, R., Safari, L. and Khanteymoori, A., 2025. Advancing CRISPR with deep learning: A comprehensive review of models and databases. Molecular Therapy Nucleic Acids, 36(4).

[32] Jiang, F., Zhou, K., Ma, L., Gressel, S. and Doudna, J.A., 2015. A Cas9–guide RNA complex preorganized for target DNA recognition. Science, 348(6242), pp.1477-1481.

[33] Li, J., Wu, P., Cao, Z., Huang, G., Lu, Z., Yan, J., Zhang, H., Zhou, Y., Liu, R., Chen, H. and Ma, L., 2024. Machine learning-based prediction models to guide the selection of Cas9 variants for efficient gene editing. Cell Reports, 43(2).

[34] Clarke, R., Heler, R., MacDougall, M.S., Yeo, N.C., Chavez, A., Regan, M., Hanakahi, L., Church, G.M., Marraffini, L.A. and Merrill, B.J., 2018. Enhanced bacterial immunity and mammalian genome editing via RNA-polymerase-mediated dislodging of Cas9 from double-strand DNA breaks. Molecular cell, 71(1), pp.42-55.

Legends

Figure 1: Schematic representation of the sgRNA sequence (green) and its downstream and upstream flanking regions (yellow) containing the PAM sequence (pink). Experiments were performed with U and D values of 0, 5, 10, and 15 nt corresponding to a full context length of 20, 30, 40 and 50 nt.

Figure 2: Violin plots showing the distribution of proliferation z-scores for individual sgRNAs targeting essential genes in Cas9 and Base Editor (BE3.9) across two cell lines (A375 and MELJUSO). Each point represents a single sgRNA.

Figure 3: Representative Receiver Operating Characteristic (ROC) curves for the different classifiers evaluated on the A375 cell line. Curves are obtained by varying the decision threshold on the predicted scores. AUC = Area Under the Curve; ACC = ACCuracy; TPR = True Positive Rate; FPR = False Positive Rate.

Figure 4. Absolute mean SHAP values for features in the LightGBM model (A375 cell line).

Figure 5. Absolute mean SHAP values (LightGBM model) for the top 10 features per cell line. The figure reports the union of these features across cell lines, resulting in 13 features overall.

Figure 6. Feature importance of the top 5 predictors (sequence, gene strand, mutation type, DepMap and BE-Hive) for each cell line. Bar plot showing the five most important features for each cell line (A375, MELJUSO, OVCAR8, HAP1, HA1E) as determined by permutation importance. Bars represent mean importance values, and error bars indicate the standard deviation across multiple permutation repeats. Feature importance was quantified as the drop in AUC–ROC (AUC drop %). The profiles highlight cell line–specific differences and identify the features with the strongest influence on model predictions.

Figure 7. Feature importance along the sgRNA sequence for each cell line. Bar plot showing feature importance by nucleotide position along the sgRNA sequence for each cell line. Bars indicate mean importance values, and error bars represent the standard deviation across permutation repeats. Overall, importance increases toward the 3' end of the sequence, while the first positions contribute minimally. Central positions show moderate contributions, with patterns consistent across the different cell lines.

Figure 8. Linear regression coefficients for Δz-score (BE3.9 − Cas9).

Table 1: The list of 47 genes including 10 essential genes (bold) used in the dataset.

Table 2: The list of features used in Machine Learning models.

Table 3. Summary statistics of the input features used in the Machine Learning models.

Table 4. Performance comparison of all evaluated classifiers in this study across the five analysed cell lines. For each ML machine learning model and cell line, the table reports the area under the ROC curve (AUC) and classification accuracy (ACC), providing a complete overview of both threshold-independent discrimination ability and threshold-dependent predictive performance.

Table 5. Linear regression coefficients for Δz-score (BE3.9 – Cas9).

Author contributions statements

Conceptualization: A.O., A.C., M.S., M.M., C.M., G.D.B., C.C., P.P., T.S.

Data curation: A.O., A.C., M.S., F.H., M.M., G.D.B., T.S.

Formal analysis: A.O., A.C., M.M.

Funding acquisition: G.D.B., T.S.

Investigation: A.O., A.C., M.M., G.D.B., C.C., P.P., T.S.

Methodology: A.O., A.C., M.I.C., M.S., F.H., M.M., C.M., G.D.B., T.S.

Project administration: G.D.B., T.S.

Software: A.O.

Supervision: A.O., A.C., M.M., G.D.B., T.S.

Validation: A.O., A.C., M.I.C., M.S., C.M., G.D.B., T.S.

Visualization: A.O., M.S., C.M., G.D.B., T.S.

Writing – original draft: A.O., T.S.

Writing – review & editing: A.O., A.C., M.I.C., M.S., F.H., M.M., C.M., G.D.B., C.C., P.P., T.S.