CRISFitFormer - User Manual

Introduction

CRISFitFormer is a deep learning-based platform for predicting and calculating bacterial gene fitness scores using CRISPR interference (CRISPRi) data. It was developed to help biologists overcome the challenges of genome-wide CRISPRi screens, which can be expensive, time-consuming, and data-intensive. The platform provides two complementary modules: a Fitness Prediction module that predicts gene fitness from guide RNA sequences and related features and a Fitness Calculation module that computes fitness scores and identifies essential genes from experimental CRISPRi screening data. CRISFitFormer currently supports five bacteria species (e.g., ***E. coli***, ***Synechocystis sp.PCC 6803***, ***E. limosum, B. subtilis***, **and** ***S. aureus***), with the possibility to extend to other species upon request. Overall, the platform aims to streamline gene fitness analysis and essential gene identification through an easy-to-use interface and robust computational methods. For expansion to other strains, users may contact 202410194277@mail.scut.edu.cn for model retraining services. By continuously integrating research data, this platform aims to enhance the universality of its prediction models and create a collaborative and shared platform for the prediction of cell fitness.

Key Features

CRISFitFormer offers several features to facilitate CRISPRi fitness analysis:

- **Accurate Prediction:** Accurately estimate cell fitness values following CRISPRi-mediated gene editing using only the user-provided sgRNA sequences and strain-specific information.
- **Cell Fitness Prediction:** Predict gene fitness values from CRISPRi knockdown data using a pre-trained deep learning model.
- **Detrimental Subsequence Identification:** Identify potentially harmful “detrimental seed” guide RNA subsequences (short PAM-proximal sequences) that consistently cause reduced fitness in experiments.
- **Cell Fitness Calculation:** Calculate gene fitness scores automatically from user-provided cell growth screening data, helping to pinpoint essential genes crucial for survival.
- **Quick Prediction:** Provides fast estimation of fitness effects for a small set of sgRNAs. Designed for quick exploration or small-scale experiment

planning.

- **User-Friendly Interface:** Interact via an intuitive web interface supporting data upload (via file import or copy-paste), species selection, and visualization of results. Sample datasets are provided for practice, allowing users to try the full workflow even without their own data.
- **Data Visualization:** Explore results with built-in plots (e.g., scatter plots and density plots) and multi-dimensional filters (by gene attributes or conditions). One-click download options allow users to easily export prediction or calculation results for further analysis.

Glossary

Detrimental subsequence: A short guide RNA subsequence (typically near the PAM site) that has a consistently negative effect on cell fitness when present. Identifying detrimental subsequences helps avoid using guide RNAs that might nonspecifically reduce cell viability.

CRISPRi: Abbreviation for CRISPR interference, a technology using a catalytically inactive Cas9 (dCas9) and a guide RNA to repress (knock down) gene expression without cutting DNA.

Essential gene: A gene required for an organism's survival or normal growth. If a gene is essential, strongly reducing its expression (or deleting it) causes lethality or a severe fitness defect.

Fitness value: A quantitative measure of cell fitness (growth or survival) associated with a particular gene perturbation. In CRISFitFormer, a lower fitness score indicates a larger negative impact on cell growth when the gene is knocked down.

Guide RNA (gRNA): The 20-nucleotide RNA sequence that directs the dCas9 to a specific DNA target. The guide RNA determines which gene is targeted by the CRISPRi system.

PAM: The Protospacer Adjacent Motif, a short DNA sequence (e.g., "NGG" for S. pyogenes Cas9) immediately following the target sequence. A proper PAM is required for Cas9 binding; "PAM-proximal" refers to nucleotides near this motif on the target DNA.

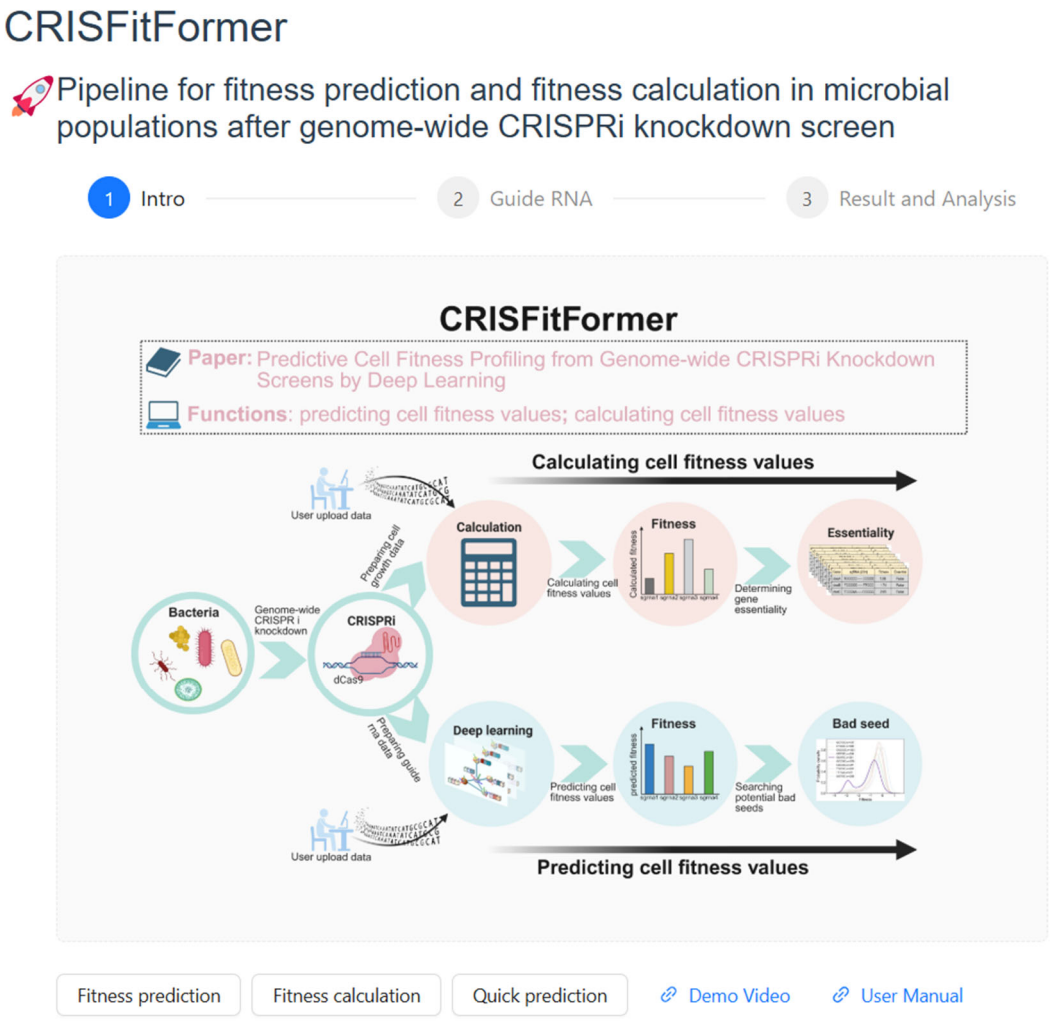


Figure 1: Overview of the platform.

Cover information

Figure 1 schematic illustrates the CRISFitFormer platform, a comprehensive pipeline designed for the prediction and calculation of microbial cell fitness following genome-wide CRISPR interference (CRISPRi) knockdown screens. The upper branch outlines the process for calculating empirical fitness values. Users upload processed CRISPRi screen data, which are subjected to computational analysis to derive quantitative fitness scores for individual sgRNAs. These values are then used to infer gene essentiality, identifying genes whose knockdown significantly impairs cell viability.

The lower branch depicts the predictive workflow. Upon input of sgRNA sequences and associated metadata, the platform employs a transformer-based deep learning model to estimate the expected impact of gene knockdown on cell fitness. The predicted fitness values are subsequently analyzed to detect “detrimental seed” motifs—specific sgRNA subsequences associated with deleterious effects. This dual-

function platform enables both retrospective analysis of experimental screens and prospective guide design by providing accurate, interpretable fitness estimates for microbial genomes.

The bottom panel of the CRISFitFormer interface provides five interactive buttons that allow users to access core platform functionalities, including fitness prediction, empirical fitness calculation, and user support tools. Each function is described in detail in the **Table 1**.

Table 1 interactive buttons definition

Button	Function Description
Fitness prediction	Allows users to input sgRNA sequences and target gene information. The platform then predicts the impact on cell growth using a deep learning model.
Fitness calculation	Enables users to upload raw CRISPRi screen data for calculating empirical cell fitness scores using internal algorithms.
Quick prediction	Provides fast estimation of fitness effects for a small set of sgRNAs. Designed for quick exploration or small-scale experiment planning.
Demo Video	A video tutorial demonstrating how to use the platform. Useful for new users to get started quickly.
User Manual	A comprehensive manual covering usage instructions, input formats, model parameters, and result interpretation.

Fitness Prediction Module

This module predicts the fitness impact of guide RNAs on a gene of interest using a trained model. Follow the steps below to use the fitness prediction function:

- **Navigate to the Prediction Module:** Log in to the CRISFitFormer platform. After accessing the main interface (see **Figure 1**), click the **'Fitness Prediction'** button (typically located at the bottom-left of the dashboard) to enter the prediction module (**Figure 2**).
- **Species Selection (Top "Precise" Dropdown)：** Users can select the target bacterial species from a dropdown menu (default: *E. coli*). The platform currently supports five species. For prediction on additional species, please contact the technical support team at **202410194277@mail.scut.edu.cn**.
- **Training Data Notification：** Once a species is selected, the interface displays the source of the model's training dataset to ensure transparency and model relevance.
- **Data Upload Options：** Click "**Import CSV**" to upload a local file**;** Click "**Paste CSV**"

to directly enter small-scale datasets**;** Use “**Download Samples**” to access a template file and check data formatting requirements.(refer to the "Data Format" section for detailed requirements).

- **Data Preview and Validation：** Uploaded or pasted data are displayed in a structured **table** for review and verification, helping users to ensure correct formatting before proceeding.
- **Prediction Execution:** After data entry is complete, click the “**Next**” button at the bottom to initiate the prediction. The platform will automatically process the input and direct the user to the results and analysis interface **(Figure 3 - 6**).
- **Correlation Analysis Between Predicted and Measured Fitness Values (Figure 3)**: Figure 3 shows the correlation between actual and predicted values: The x-axis represents the model's predicted values, while the y-axis represents the experimentally measured actual fitness scores. Hovering the cursor over any data point will display detailed information (gRNA sequence, actual value, and predicted value). The Pearson correlation coefficient (PCC) is labeled in the lower-left corner, indicating the correlation between predicted and actual values, with a range of [-1,1], corresponding to perfect negative correlation, no correlation, and perfect positive correlation, respectively.
- **Probability Density Distribution of Predicted Fitness Values (Figure 4)**: **Figure 4** shows the probability density distribution of predicted values: The x-axis represents the predicted fitness score range, and the y-axis represents the probability density. The red box displays the maximum, minimum, and average predicted values. The yellow filtering area allows data to be filtered by ori (direction), coding (coding status), and essential (essential gene). The green box provides a "**Download Result**" option for downloading the data.
- **Prediction Results in a Tabular Format (Figure 5): Figure 5** presents the prediction results in a tabular format, with a yellow control in the top-right corner for switching between ascending and descending order modes.
- **Unique Subsequences (Figure 6): Figure 6** displays the fitness value distribution of different guide RNA subsequences, used to identify potential detrimental seed sequences. Detrimental seeds are specific subsequences in the gRNA that negatively affect cellular fitness value after CRISPRi editing. "Subsequence length" defines the analysis segment length, and the "top" parameter controls the number of high-frequency subsequences (corresponding to the number of curves). After setting the parameters, click "Generate" to generate the analysis plot, where each curve represents the fitness distribution of a different subsequence. A left-skewed distribution suggests the presence of potential harmful seeds. The curve label "n" indicates the counts of the subsequence in the dataset.

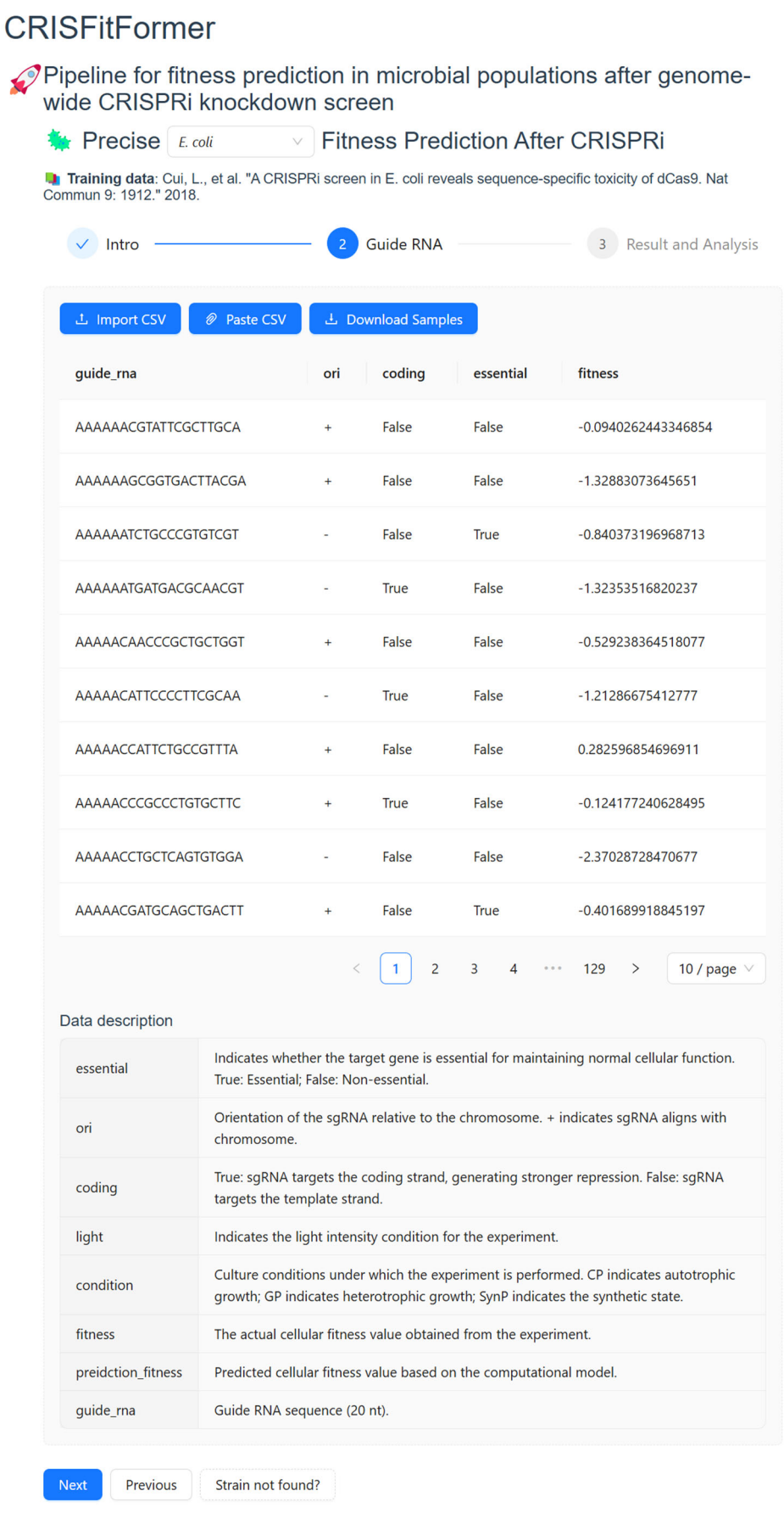


Figure 2: Species selection and data in the prediction module.

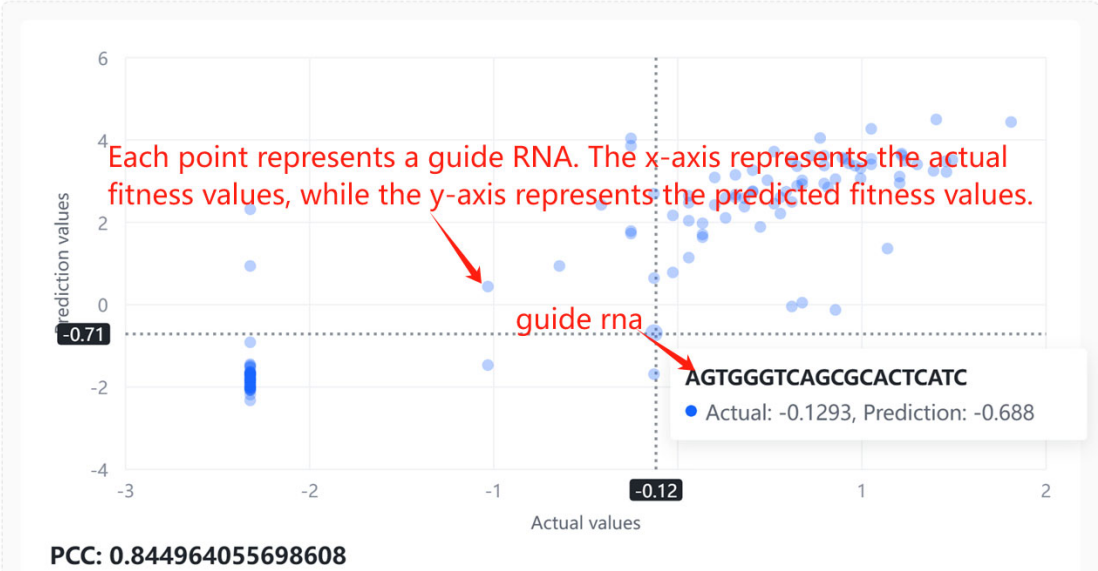


Figure 3: Results and analysis of the prediction module 1.

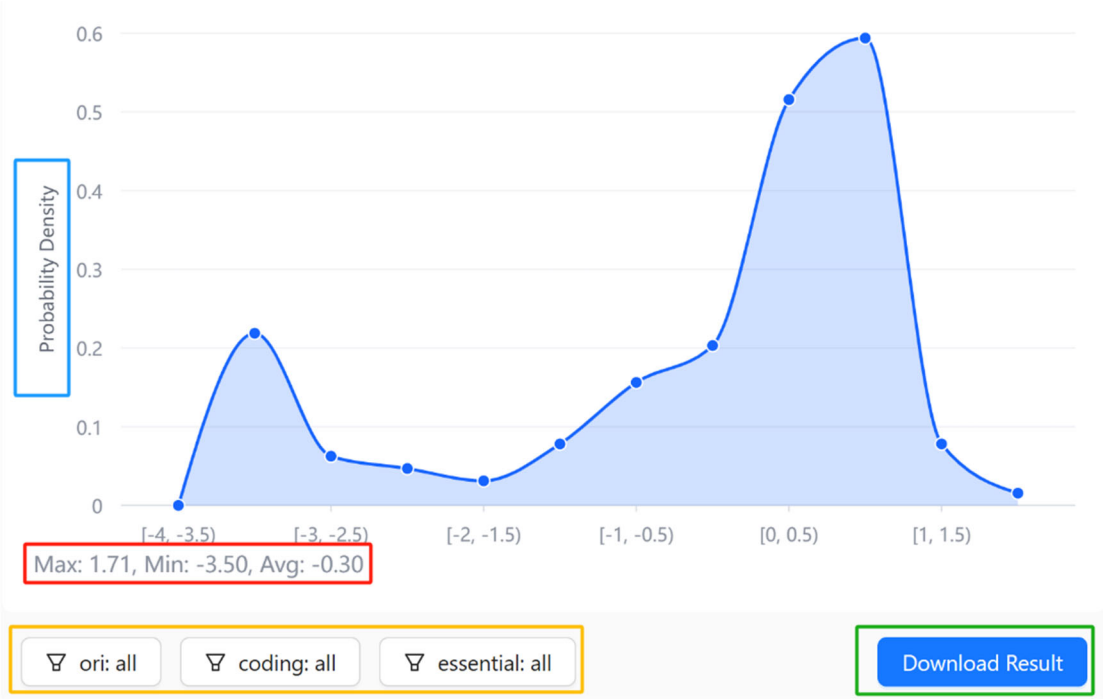
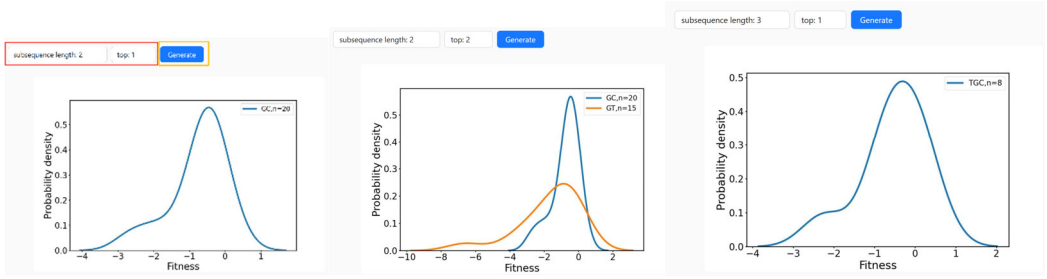


Figure 4: Results and analysis of the prediction module 2.

guide_rna	ori	coding	essential	prediction_fitness
AAAAAACGTATTCGCTTGCA	+	FALSE	FALSE	1.7100000381469727
AAAAAAGCGGTGACTTACGA	+	FALSE	FALSE	-1.1180450916290283
AAAAAATCTGCCCGTGTCGT	-	FALSE	TRUE	-0.12368369102478027
AAAAAATGATGACGCAACGT	-	TRUE	FALSE	-0.8992371559143066
AAAAACAACCCGCTGCTGGT	+	FALSE	FALSE	0.7227859497070312
AAAAACATTCCCCTTCGCAA	-	TRUE	FALSE	-0.12367892265319824
AAAAACCATTCTGCCGTTTA	+	FALSE	FALSE	0.520871639251709
AAAAACCCGCCCTGTGCTTC	+	TRUE	FALSE	0.6621608734130859
AAAAACCTGCTCAGTGTGGA	-	FALSE	FALSE	-3.1152377128601074
AAAAACGATGCAGCTGACTT	+	FALSE	TRUE	0.4465620517730713

1 2 3 4 ••• 13 > 10 / page

Figure 5: Results and analysis of the prediction module 3.



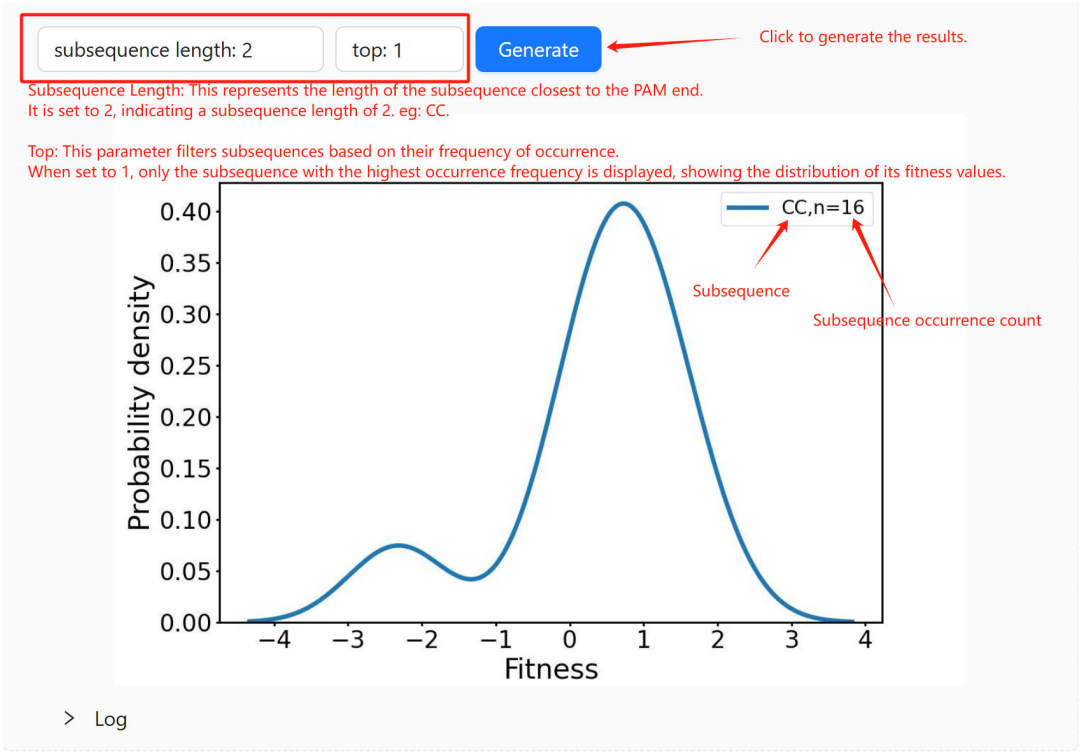


Figure 6: Results and analysis of the prediction module 4.

Fitness Value Calculation Module

Figure 7 illustrates the second step of the CRISFitFormer pipeline, where users perform **cell fitness score calculation** from CRISPRi screening data. The interface supports the upload and parsing of sequencing read count data and allows the user to define key parameters for fitness estimation.

1. **Upload Data：** Users may upload data in TSV format using either the “**Import TSV**” button for local files or the “**Paste TSV**” option for manual entry. The Growth parameter represents the number of bacterial generations between the start and end of the experiment and is critical for accurate normalization. Example: if samples were collected at generations 2 and 11, set Growth = 10 (i.e., 11 - 2 + 1). **Download Samples**: Button for downloading a template TSV file to guide proper data formatting.
2. **Data Description Panel (Bottom Section):** A well-structured reference guide helps users interpret each data column
3. **Computing Fitness Values:** After uploading the data and setting the parameters, click the "Next" button at the bottom-left to start the calculation. The platform will calculate the fitness score and redirect to the results interface in **Figure** 8 and **Figure** 9.

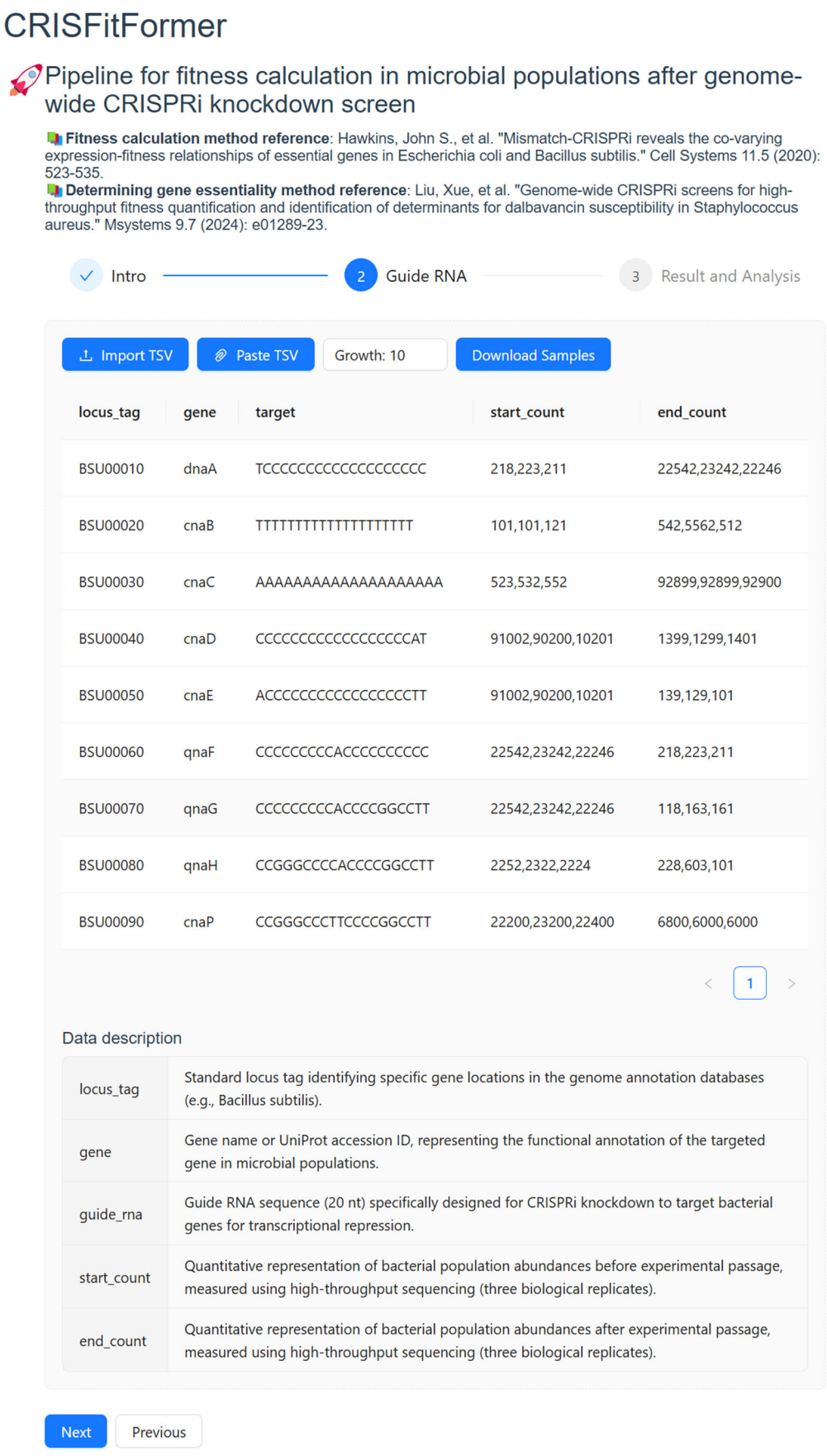


Figure 7 Data upload of the computing module.

Result Analysis and Visualization

- **Figure 8, 9 and 10** describe the results of the fitness value calculation module.
- **Figure 8** displays the distribution of the calculated fitness values along with statistical metrics (maximum, minimum, and average values).
- **Figure 9** evaluates gene fitness based on the average fitness values of multiple gRNAs (as the same gene can have multiple gRNAs with different fitness values). Each data point represents a gene, where the size of the point reflects the absolute fitness value, and the color distinguishes positive and negative values (blue for positive, green for negative). Hovering the mouse over a point will display the gene name and the specific fitness value.

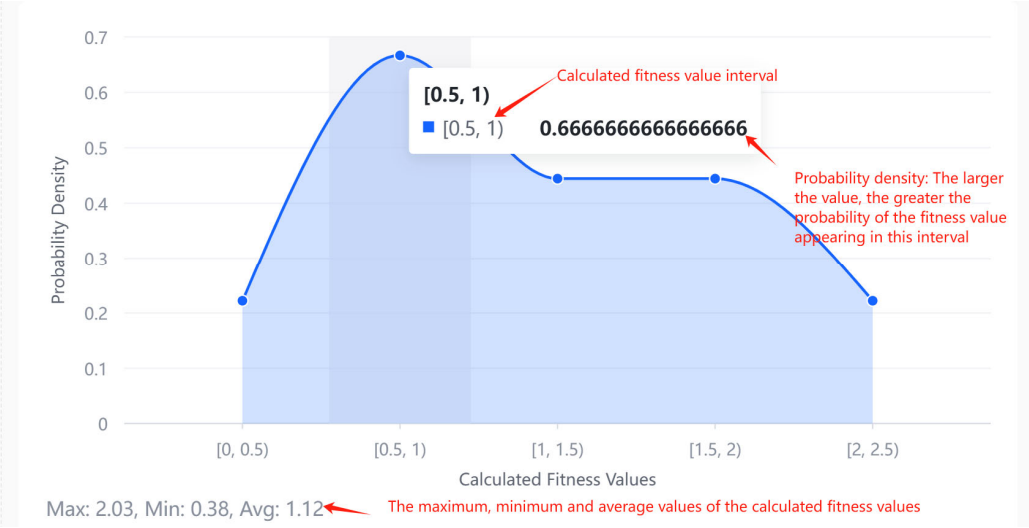


Figure 8: Results and analysis of calculation module 1.

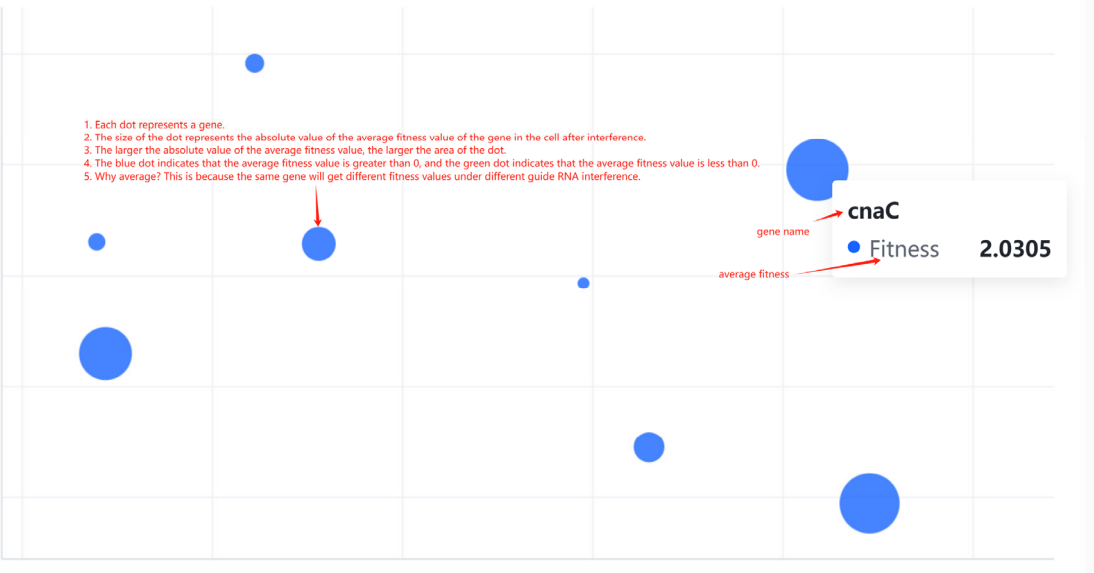


Figure 9 Results and analysis of calculation module 2.

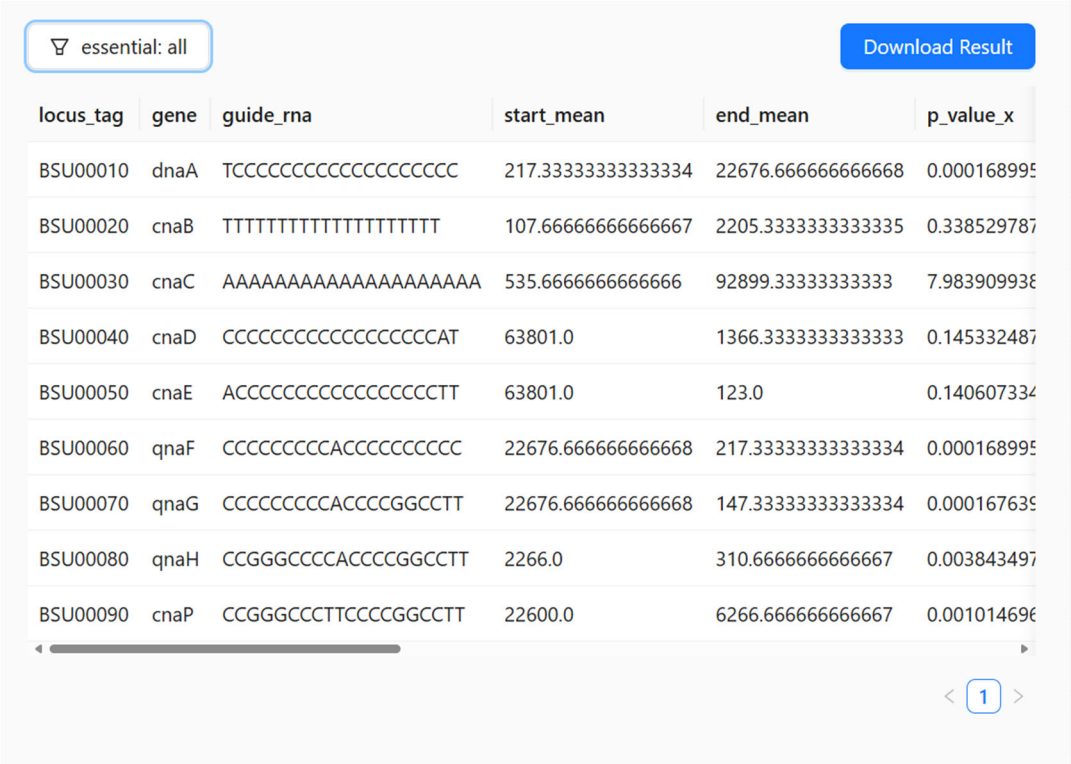


Figure 10 Results and analysis of calculation module 3.

- **Figure 10** displays the output of the CRISFitFormer fitness calculation module in a structured tabular format. Each row corresponds to a gene targeted by CRISPRi. **Essentiality filter (top-left corner)**: The essential: all dropdown allows users to filter results based on predicted gene essentiality (e.g., show only essential or non-essential genes); **Download function (top-right corner):** The “Download Result” button enables users to export the computed fitness results in file format for downstream analysis, such as statistical comparison, visualization, or gene prioritization.

Quick prediction

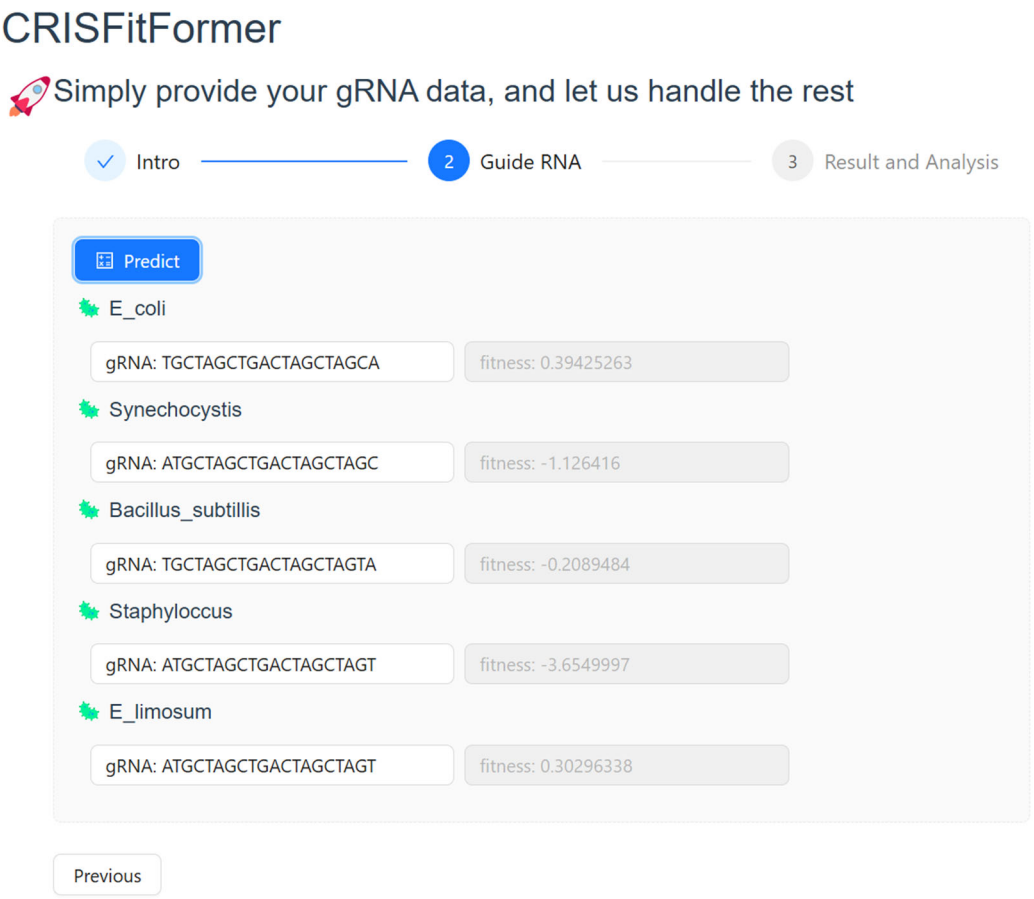


Figure 11 Results and inputs of quick prediction module

Figure 11 shows the Quick Prediction Module in the CRISFitFormer platform, designed for fast, user-friendly prediction of sgRNA-induced cellular fitness effects across multiple microbial species. The interface simplifies the user experience by allowing single-guide RNA (sgRNA) input per species.

1. **Input Fields**: For each supported species, users can input a 20-nt sgRNA sequence in the corresponding text box labeled “gRNA:”. The platform currently supports **five species**: *E. coli, Synechocystis, Bacillus subtilis, Staphylococcus, E. limosum*.
2. **Prediction Output:** Upon clicking the “**Predict**” button, the platform processes each input sequence using a species-specific deep learning model. The corresponding “fitness” fields will display the predicted cellular fitness value for each sgRNA.

This module is optimized for speed and ease of use, making it ideal for small-scale testing, validation of candidate sgRNAs, or educational demonstrations.

Data Format

Proper data formatting is important for CRISFitFormer to function correctly. Below we outline the required input and output data formats for both modules. Users should ensure their files follow these specifications.

Fitness Prediction Module — Feature Definitions

Feature	Definition
guide_rna	The sequence of the guide RNA, with a length of 20 nucleotides.
essential (True, False)	Whether the target gene is essential. TRUE indicates the target gene is essential.
ori (+,-)	The direction of the guide RNA relative to the chromosome. '+' indicates the guide RNA aligns with the chromosome in the same direction [1].
coding (True, False)	Whether the target site is located on the coding or template strand. TRUE indicates the target site is on the coding strand.
light (100,300,0)	The light intensity of the external environment[3]。100：100μmol$m^{-2}s^{-1}$；300：300μmol$m^{-2}s^{-1}$；0：300μmol$m^{-2}s^{-1}$-dark cycle.
condition (CP，GP，SynP)	The growth condition of the strain. CP indicates autotrophic growth, GP indicates heterotrophic growth, and SynP indicates the synthetic state[5].
fitness	The experimentally measured cell fitness value.
prediction_fitness	The predicted cell fitness value.

Fitness prediction module-input data——E_coli[1]: csv format

```
guide_rna,ori,coding,essential,fitness
AAAAAACGTATTCGCTTGCA,+,False,False,-0.0940262443346854
AAAAAAGCGGTGACTTACGA,+,False,False,-1.32883073645651
AAAAAATCTGCCCGTGTCGT,-,False,True,-0.840373196968713
AAAAAATGATGACGCAACGT,-,True,False,-1.32353516820237
AAAAACAACCCGCTGCTGGT,+,False,False,-0.529238364518077
AAAAACATTCCCCTTCGCAA,-,True,False,-1.21286675412777
AAAAACCATTCTGCCGTTTA,+,False,False,0.282596854696911
AAAAACCCGCCCTGTGCTTC,+,True,False,-0.124177240628495
AAAAACCTGCTCAGTGTGGA,-,False,False,-2.37028728470677
AAAAACGATGCAGCTGACTT,+,False,True,-0.401689918845197
AAAAACGCAGTGACATAGTC,-,True,False,-0.630161455719922
AAAAACGCCCTGCTGCTGGC,-,False,False,-1.00528541045869
AAAAACGGTGTCACGCTGTT,-,False,False,-0.253844166227893
AAAAACGTGATTAACCGTAC,+,False,False,-1.98195012091204
AAAAACGTTGTCGCCGCTGG,-,False,False,0.108692685922416
```

Fitness prediction module-input data——Bacillus[2]: csv format

```
guide_rna,fitness,essential
AAAAAAACATTATCCTTAGC,0.323869184,True
AAAAAACCATTATCATTAGC,0.986884716,True
AAAAAACCATTATCCTTAGC,0.333269408,True
AAAAAACCATTATCCTTATC,0.991810631,True
AAAAAACCATTATCCTTCGC,0.865442119,True
AAAAAACCATTATCGTTAGC,0.978131895,True
AAAAAACCATTATGCTTAGC,0.9680466,True
AAAAAACCATTTTCCTTAGC,0.913887612,True
AAAAAAGATGCAGCGATAAC,0.041419993,True
AAAAAAGGAACTGAAAAATC,0.719910623,True
AAAAAATAGACCTCCATTTT,-0.018795806,True
AAAAAATCAACCTCCATTTT,0.08800205,True
AAAAAATCGACCCCCATTTT,0.485983514,True
AAAAAATCGACCTACATTTT,1.007292254,True
AAAAAATCGACCTCCATGTT,0.102543241,True
```

Fitness prediction module-input data——Synechocystis[3]: csv format

```
guide_rna,fitness,ori,light
GTTCTGGAGTGCGTTCTTGA,0.116852411,-,100
GTTCTTGACGGAGTCCAGTG,0.1943717,-,100
AACAGCGGGCTGACAGCCTA,0.020383615,-,100
TTGGACAAAGCGGGGAACAG,0.35421306,-,100
CCTTTCGTAAACCGCTTCTG,0.048713523,+,100
ATTCATCGGTGCAACAAAGA,0.28277602,+,100
TGTAGGGCTTCTATGGCCTC,-0.268264018,+,100
GGCATCGAGGGCAATCTGTA,0.023791293,+,100
AATAAGAATGTCGCGCATAG,0.046884456,-,100
TGTTGGCCTGATTATCCCAT,0.205190492,-,100
AGCGTTAGGAGCCACTGTTG,-0.508871457,-,100
GTTTTTTCATAGGGCAAACA,0.003416015,-,100
TAGGATGACCAGAACCACCG,-4.860002404,-,100
GTCCCTTGACGTACCCCAAA,-0.250744315,-,100
AGACGACTTTTATAGACTCC,-0.968431855,-,100
```

Fitness prediction module-input data——Staphyloccus[4]: csv format

```
guide_rna,fitness,essential
TGTAAAGCTCAGTATCTTTT,-5.646463856,True
AATATAGGTAATGTTGTTCT,-1.639992287,True
CAGCATCTTCAGGTAATTCT,0.19323969,False
TGTAATGCTGAATGATTTGT,0.159712655,False
GTAATTCATTATATTGCGTC,0.000738166,False
ATTTTGCTCTTAACTGTGTC,-6.198805533,True
CGGCATAACCCAATAATGTT,0.136531628,False
GCATGACAAACTACAACTAA,0.086525972,False
GTTGCAATTATTGTTGCTTT,-0.128787996,False
CTCCGCCAAATTGAATGGTA,-0.05263188,False
AATTCTAAAATATCAGCAAT,-5.090290883,True
ATACACTCATGCGTATCAAG,0.095095216,False
GTTACTGCATGTGCTTTCAT,-0.003135502,False
TTCGTCATTGGCGGATCAAA,0.07920728,False
TAGTTTCCCTACAGCTAAAA,-0.087039587,False
```

Fitness prediction module-input data——E_limosum[5]: csv format

```
guide_rna,fitness,condition
CAATCGCATACCGAATTGTT,-1.031597256,GP
TTGCGGGCTTGTGAAAAATT,0.200247261,GP
GTAGTCCATGTGTGGCTGCA,1.050367163,GP
CCCAAGGCGCTCCTGCGCGA,-2.322510734,GP
GGCCGAAAATTTAACGGCAT,-2.322510734,GP
CCACAACGTCGCCGGACTTC,0.312750598,GP
AGTGCGTCACCGCGCCCTCC,1.452480811,GP
CACATCTGGGTACAGTCCGC,0.675273453,GP
CTCCCCGGGCGCCTTTCTGC,-2.322510734,GP
TTACACTCCACATGGACACC,0.312750598,GP
CCAGAATGTGCTGTTTAGCC,0.200247261,GP
TTTAGCGTCAATATATGCCA,0.059351195,GP
CGACGGGAACTTTGAGAACC,0.928672856,GP
CCAGCGCTCCCCTTTTTTGA,-0.254389898,GP
GCTCAATCCGCCCCTCTGCG,-2.322510734,GP
```

Fitness prediction module-output data (csv format)

The output data of the prediction module has a new "prediction_fitness" column compared to the input data, which corresponds to the cell fitness value predicted by the model.

guide_rna	fitness	condition	prediction_fitness
CAATCGCATACCGAATTGTT	-1.031597256	GP	0.4411046504974365
TTGCGGGCTTGTGAAAAATT	0.200247261	GP	2.4240238666534424
GTAGTCCATGTGTGGCTGCA	1.050367163	GP	3.4087603092193604
CCCAAGGCGCTCCTGCGCGA	-2.322510734	GP	-1.7045824527740479
GGCCGAAAATTTAACGGCAT	-2.322510734	GP	-2.0579476356506348
CCACAACGTCGCCGGACTTC	0.312750598	GP	3.1535704135894775
AGTGCGTCACCGCGCCCTCC	1.452480811	GP	3.4841063022613525
CACATCTGGGTACAGTCCGC	0.675273453	GP	0.049986839294433594
CTCCCCGGGCGCCTTTCTGC	-2.322510734	GP	-2.1816251277923584
TTACACTCCACATGGACACC	0.312750598	GP	2.673779249191284

Fitness Calculation Module — Feature Definitions[4,6]

Feature	Definition
locus_tag	Gene name
gene	The alternative gene name (Uniprot designation)
guide_rna	The sequence of the guide RNA (20 nt)
start_count	The CRISPRi sequencing results of the strain before passage (Each strain's CRISPRi sequencing results include data from three replicate experiments, which are stored as comma-separated values for downstream analysis.)
end_count	The CRISPRi sequencing results of the strain after passage (Each strain's CRISPRi sequencing results include data from three replicate experiments, which are stored as comma-separated values for downstream analysis.)
start_mean	The average CRISPRi sequencing result of the strain before passage
end_mean	The average CRISPRi sequencing result of the strain after passage
p_value_x	The p-value of the sample[4]
gamma	This is a parameter representing the consistency of gene behavior across experimental replicates, which contributes to fitness calculations based on coverage[4]
start_mask	Whether the target sgRNA sequence exists, which affects downstream analyses like fold-change calculations[4]
p_value_y	The p-value of the batch[4]
P_{adj}	The adjusted p-value, corrected for false positives[6]
fitness	The calculated cell fitness value
FoldChange	The fold change in gene relative abundance before and after passage
log2FoldChange	The log2 transformation of the fold change in gene relative abundance before and after passage
essential	The essentiality of the gene for the strain

Fitness value calculation module-input data (tsv format)

locus_tag	gene	guide_rna	start_count	end_count
BSU00010	dnaA	TCCCCCCCCCCCCCCCCCCC	218,223,211	22542,23242,22246
BSU00020	cnaB	TTTTTTTTTTTTTTTTTTTT	101,101,121	542,5562,512
BSU00030	cnaC	AAAAAAAAAAAAAAAAAAAA	523,532,552	92899,92899,92900
BSU00040	cnaD	CCCCCCCCCCCCCCCCCCAT	91002,90200,10201	1399,1299,1401
BSU00050	cnaE	ACCCCCCCCCCCCCCCCCTT	91002,90200,10201	139,129,101
BSU00060	qnaF	CCCCCCCCCACCCCCCCCCC	22542,23242,22246	218,223,211
BSU00070	qnaG	CCCCCCCCCACCCCGGCCTT	22542,23242,22246	118,163,161
BSU00080	qnaH	CCGGGCCCCACCCCGGCCTT	2252,2322,2224	228,603,101
BSU00090	cnaP	CCGGGCCCTTCCCCGGCCTT	22200,23200,22400	6800,6000,6000

Fitness value calculation module-output data (csv format)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	locus_tag	gene	guide_rna	start_mean	end_mean	p_value_x	gamma	start_mask	p_value_y	padj	fitness	FoldChange	log2FoldChange	essential
2	BSU00010	dnaA	TCCCCCCCCCCCCCCCCCCC	217.333333333	22676.66666	0.000169	0.9571863	TRUE	0.000169	0.00038	1.9571863	104.34049079	6.70515531442821	FALSE
3	BSU00020	cnaB	TTTTTTTTTTTTTTTTTTTT	107.666666666	2205.333333	0.33852978	0.7223061	TRUE	0.3385297	0.3385297	1.7223061	20.482972136	4.356353164	FALSE
4	BSU00030	cnaC	AAAAAAAAAAAAAAAAAAAA	535.666666666	92899.33333	7.98390993	1.0304897	TRUE	7.9839099	7.1855189	2.0304897	173.42750466	7.43818891013406	FALSE
5	BSU00040	cnaD	CCCCCCCCCCCCCCCCCCAT	63801	1366.333333	0.14533248	-0.26785	TRUE	0.1453324	0.1634990	0.7321510	0.0214155473	-5.545197642	FALSE
6	BSU00050	cnaE	ACCCCCCCCCCCCCCCCCTT	63801	123	0.14060733	-0.61521	TRUE	0.1406073	0.1634990	0.3847931	0.001927869	-9.018776911	FALSE
7	BSU00060	qnaF	CCCCCCCCCACCCCCCCCC	22676.6666666	217.3333333	0.000169	-0.38384	TRUE	0.000169	0.00038	0.6161553	0.009584007	-6.705155314	TRUE
8	BSU00070	qnaG	CCCCCCCCCACCCCGGCCTT	22676.6666666	147.3333333	0.0001676	-0.43993	TRUE	0.000168	0.00038	0.560073	0.006497134	-7.265980909	TRUE
9	BSU00080	qnaH	CCGGGCCCCACCCCGGCCTT	2266	310.6666666	0.0038435	0	TRUE	0.003843	0.005765	1	0.1370991468	-2.866708502	TRUE
10	BSU00090	cnaP	CCGGGCCCTTCCCCGGCCTT	22600	6266.666666	0.0010147	0.1016155	TRUE	0.001015	0.001826	1.1016155	0.2772861356	-1.850552611	TRUE

Troubleshooting and FAQ

Here we address some common issues and questions users might encounter when using CRISFitFormer:

Q1: My data file won't upload or I get an error after clicking Next. What should I do?

A1: First, verify that your input file is in the correct format (CSV for prediction, TSV for calculation) and that all required columns are present with the correct header names. Check for typos or extra spaces in the headers (they must match exactly the expected names like guide_rna, fitness, etc.). If the file is very large and the upload method isn't responding, try the copy-paste method or ensure your internet connection is stable. If the problem persists, it may be an issue with the server—try again later or contact the support email provided.

Q2: I don't see my bacterial species in the dropdown list. Can I still use the prediction module?

A2: Currently, CRISFitFormer supports five species in the prediction module. If your species is not listed, the model is not yet trained for it. You can contact the development team (via the provided email) to discuss adding support for a new species. They may be able to retrain or extend the model with new data. For the calculation module (which uses your own experimental data), you can analyze any species as long as you have the required input data.

Q3: How should I choose the "Growth" parameter for the calculation module?

A3: The Growth parameter should reflect the number of generations that passed between your starting sample and ending sample in the CRISPRi experiment. If you know the exact generation numbers at which samples were taken, use the formula end_generation − start_generation + 1. If you are unsure, estimate based on how many doublings occurred during the experiment. This parameter is critical: if set too high or too low, the computed fitness values will be scaled incorrectly. Always double-check your experiment log or notes to set this value.

Q4: The prediction results show some negative fitness values. What does a

negative fitness score indicate?

A4: A negative fitness score means that knocking down that gene had a negative impact on cell growth in the conditions tested (i.e., fewer cells survived or grew compared to control). In other words, the gene is likely important for growth under those conditions (potentially an essential gene). Positive fitness values indicate the knockdown had little negative effect (or even a slight positive effect) on growth. The distribution of fitness scores can help highlight which genes are likely essential (significantly negative scores) versus non-essential.

Q5: What exactly are "Detrimental subsequences" and how can I avoid them?

A5: "Detrimental subsequence" are short sequences within a guide RNA that tend to cause unintended toxicity or reduced fitness, regardless of the target gene. In the prediction module's detrimental seed analysis, if certain 4–5 base motifs are associated with strong fitness defects across many guides, you should try to avoid designing new guide RNAs that contain those motifs. The detrimental subsequence analysis tool can guide the design of gRNAs by highlighting sequences to be cautious about.

Q6: The results indicate some genes as essential. How are essential genes determined here?

A6: CRISFitFormer can mark genes as "essential" based on criteria like having a strongly negative fitness score and meeting a significance cutoff (for example, an adjusted p-value below 0.05). If a gene is marked essential, it means that the data suggest suppressing that gene severely hinders cell growth. You can refer to the output's fitness[6] and P_{adj}[4] values to evaluate the confidence of each essential gene call.

Reference

[1] Cui, Lun, et al. "A CRISPRi screen in E. coli reveals sequence-specific toxicity of dCas9." Nature communications 9.1 (2018): 1912.

[2] Hawkins, John S., et al. "Mismatch-CRISPRi reveals the co-varying expression-fitness relationships of essential genes in Escherichia coli and Bacillus subtilis." Cell Systems 11.5 (2020): 523-535.

[3] Yao, Lun, et al. "Pooled CRISPRi screening of the cyanobacterium Synechocystis sp PCC 6803 for enhanced industrial phenotypes." Nature Communications 11.1 (2020): 1666.

[4] Liu, Xue, et al. "Genome-wide CRISPRi screens for high-throughput fitness quantification and identification of determinants for dalbavancin susceptibility in Staphylococcus aureus." Msystems 9.7 (2024): e01289-23.

[5] Shin, Jongoh, et al. "Genome-wide CRISPRi screen identifies enhanced autolithotrophic phenotypes in acetogenic bacterium Eubacterium limosum." Proceedings of the National Academy of Sciences 120.6 (2023): e2216244120.

[6] Hawkins, John S., et al. "Mismatch-CRISPRi reveals the co-varying expression-fitness relationships of essential genes in Escherichia coli and Bacillus subtilis." Cell Systems 11.5 (2020): 523-535.