# Accelerating Curation of CRISPR Screen Data from Public Repositories



Contact Shantanu Bafna Bioinformatics Engineer III

shantanu.bafna@strandls.com

+91 96298 70197







Radhakrishna Bettadapura<sup>1</sup>, Aastha Tripathi<sup>1</sup>, Manjushree Sahoo<sup>1</sup>, Niharika A<sup>1</sup>, Suraj Kumar Sharma<sup>1</sup>, Anna Zacharia<sup>1</sup>, Hima Varghese<sup>1</sup>, Charles Lu<sup>2</sup>, Neil Kuehnle<sup>2</sup>, Swaraj Basu<sup>1</sup>, Shantanu Bafna<sup>1</sup>

- <sup>1</sup>Strand Life Sciences, Bengaluru, India
- <sup>2</sup> AbbVie, North Chicago, IL, USA 600

## Introduction

CRISPR (clustered regularly interspaced short palindromic repeats) has revolutionized functional genomics by enabling systematic, large-scale studies of gene function.

## CRISPR screening approaches

Pooled screens: perturbations in mixed cell populations; gene effects read out via selection/sequencing.

Arrayed screens: perturbations delivered well-by-well; suited for high-resolution phenotypic assays.

### Current challenge

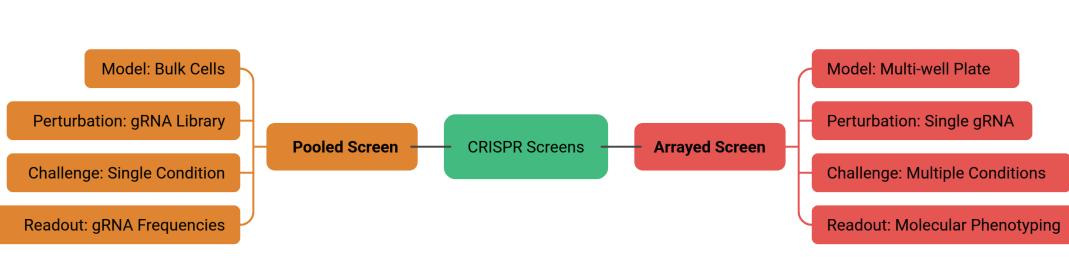
Rapidly growing datasets, but finding the right CRISPR data requires manual, expert led curation. This process is:

- 1. Labor-intensive
- 2. Hard to scale
- 3. Not easily reproducible

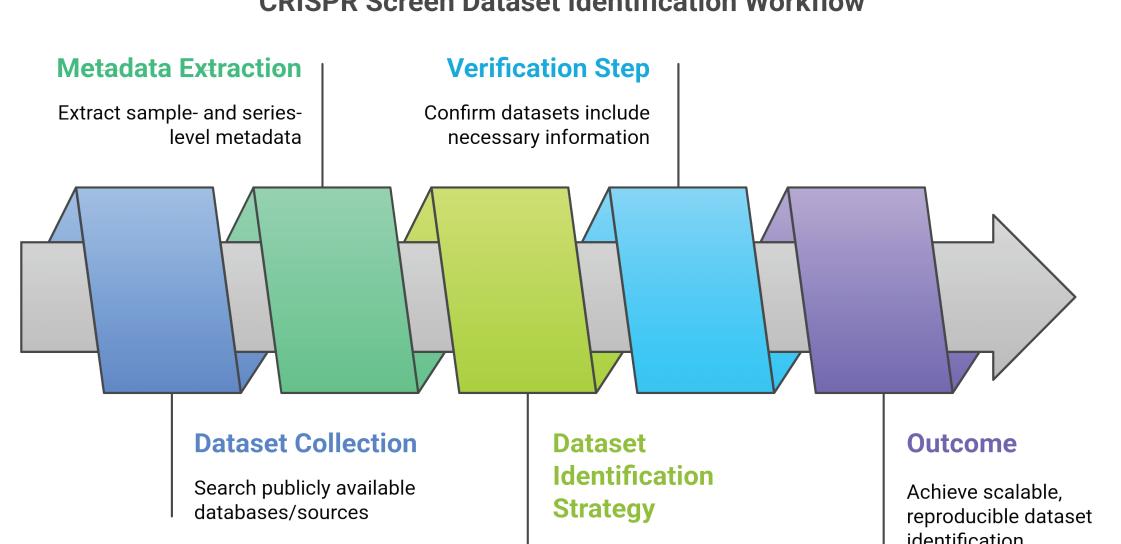
## Why curation matters

- Ensures dataset discoverability
- Supports cross-study comparisons Accelerates target discovery and therapeutic development

We explored methods to streamline curation using NLP, Al-driven agents, and automated pipelines, making CRISPR datasets easier to organize, search, and apply.

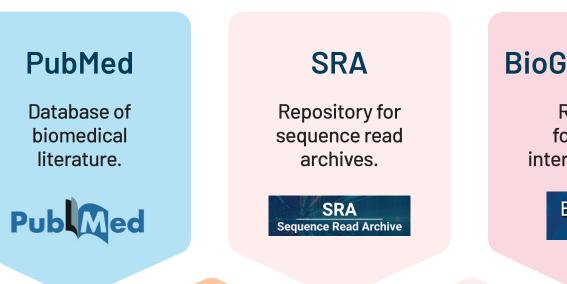


## **CRISPR Screen Dataset Identification Workflow**



Materials And Methods

## **Data Sources**



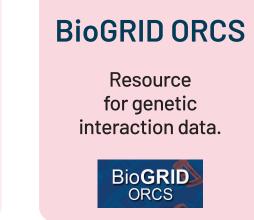
SSO

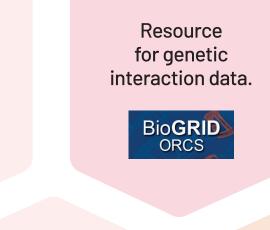
GEO

Database for

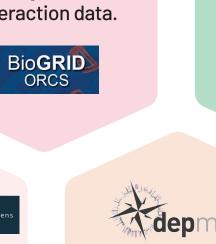
gene expression

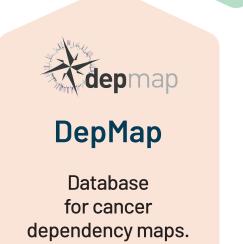
data.





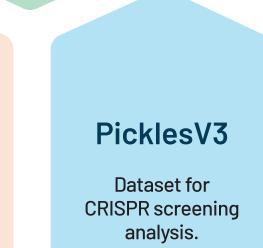




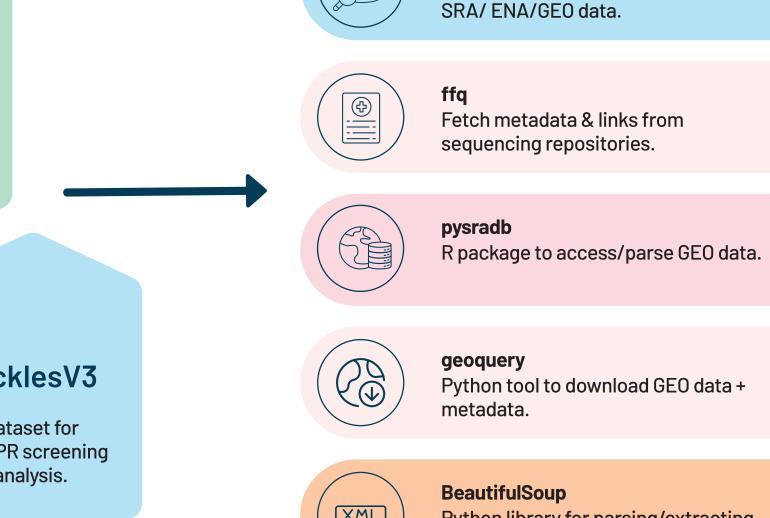


# Tool for visualizing CRISPR data.

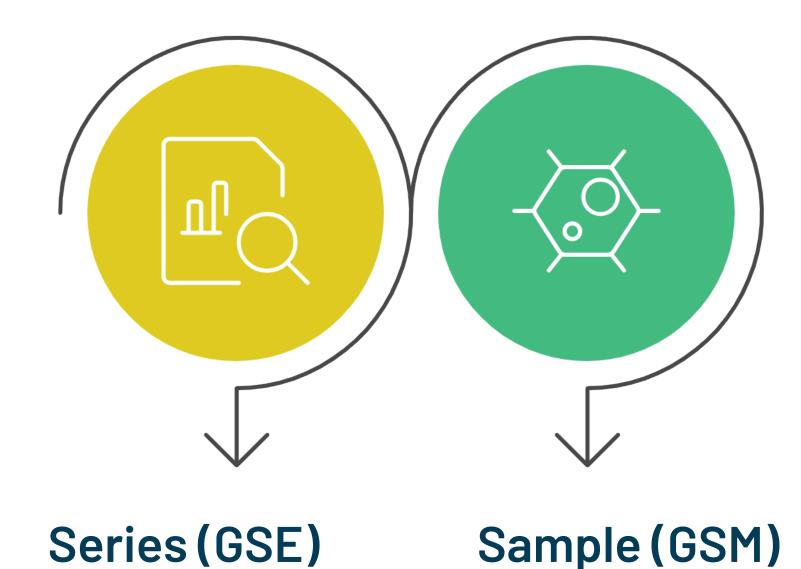
**CRISPView** 



## Metadata extraction tools



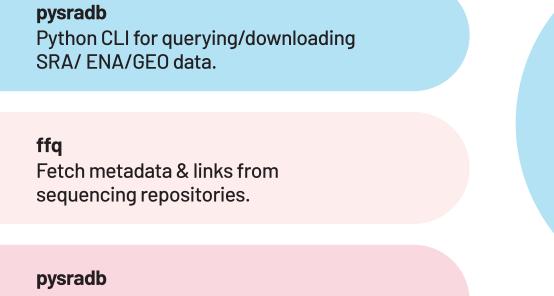
- GEO serves as the main resource: Among all databases surveyed, GEO was the main resource for data extraction and curation, since most other repositories were directly linked or cross-referenced to GEO.
- Metadata extraction tools: Combination of tools was used for reproducible data retrieval, including pysradb, ffq, GEOquery, geofetch, and BeautifulSoup.
- Series-level (GSE): Represents entire studies or experiments, containing high-level details such as library strategy, overall design, and data processing pipelines.
- Sample-level (GSM): Contains detailed metadata for each biological sample, including sample characteristics, library preparation, sequencing platform, and associated processed/raw data.



## Series (GSE)

Represents the overall study/experiment, providing design and context. It groups and organizes multiple samples.

Represents an individual biological sample with specific details. It is linked to raw and processed data files.



Python library for parsing/extracting

### Developed a Chrome plugin to classify GEO datasets as CRISPR screen or non-screen. **NLP Analysis** The plugin scrapes metadata from GEO pages and sends it as contextual input to GPT-40. Classification criteria:

Conclusion

precision and interpretability.

downstream analysis.

Streamlining Data Solutions

O1 Automatically extract datasets from unstructured metadata.

Use Agentic Al for client-specific data standardization.

Enhance dynamic data processing through improved contextual understanding.

• The identification and curation of CRISPR screen datasets remain a critical yet

strategy manual review, pattern matching, semantic search, and LLM-based

challenging step for large-scale functional genomics. Our multi-pronged

approaches showed that while accurate identification is possible, it often

requires significant time, labor, or computational resources. Among these,

Although the accuracy of GPT-4o (~82%) is considered good, it is expected to

improve as newer models gain better contextual understanding. Accuracy will

also benefit from reinforcement learning and iterative input refinement by data

scientists and bioinformaticians, further strengthening identification reliability.

Looking ahead, the challenge shifts from dataset identification to making them

2) processing harmonization to ensure uniform pipelines for raw data and

Establishing such a framework will enable a unified repository of high-quality

such as predictive modeling, network inference, and therapeutic discovery.

**Dataset** 

Identification

• Utilized multiple open-source zero-shot classification and NLP

models from the Hugging Face repository for dataset analysis

• Generated embeddings from data processing descriptions

• Compared these embeddings to CRISPR screen-specific

• Focused analysis on sentences containing CRISPR-related

terms (e.g., sgRNA, pooled screen, knockout) to enhance

using advanced sentence transformer models.

reference texts based on semantic similarity.

CRISPR screens, accelerating integrative analyses and downstream applications

analysis-ready. This requires harmonization on two fronts:

(1) ontology harmonization to standardize metadata; and

LLM-based methods emerged as the most scalable and efficient, offering both

 Positive-Presence of sgRNA count matrices, gene-level results, or raw CRISPR screen files. Negative-Only RNA-seq/scRNA-seq/ATAC-seq data, CRISPR editing without screen results, or CRISPR mentioned without data.

NLP classifies text

based on patterns in

training data, while

LLMs understand

context and can adapt

classification to new

situations without

retraining.

 GPT 4o applies inclusion/exclusion logic via API, returning: Decision (screen vs. non-screen) Confidence score & reasoning Relevant keywords supporting the classification

Find datasets based on meaning LLMS Search Multi-Layered Automate Pattern based complex Strategy curation

**Manual Curation Automated Curation** Fast, scalable, Slow, unscalable reproducible curation expert curation

To identify CRISPR pooled screen datasets, we first examined the series-level metadata. This included the dataset Overall design

 Supplementary files If the dataset type could not be determined directly from this level, we then moved on to the sample-level metadata.

At the sample level, we reviewed fields such as CharacteristicsDescription • Treatment protocol • Data processing

that were not clearly labeled at the series level.

 Library strategy These details were used both to confirm the dataset type and, when necessary, to identify CRISPR pooled screen datasets

Results

## Challenges with Databases



**iCSDB** 

Based on DepMap & BioGRID ORCS

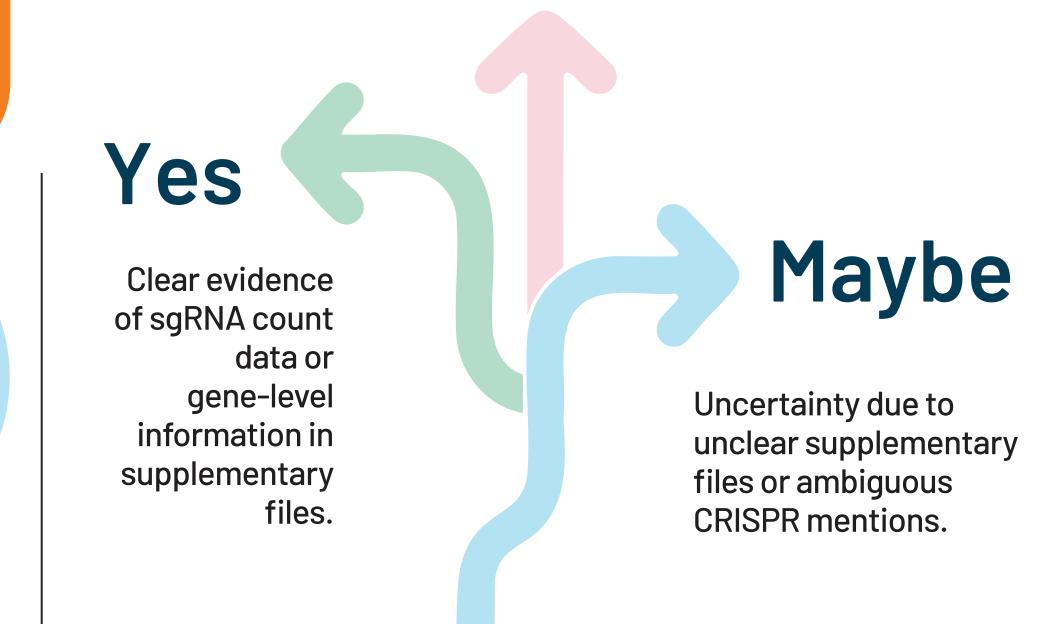
last updated in

PicklesV3

PubMed via

**CRISPR** screen Dataset Classification

Data is irrelevant or unavailable (e.g., single-cell data or non-CRISPR GEO IDs).



## **Text Search**

Where text search is

limited to exact string

matches, NLP models

leverage semantic

understanding to capture

sentiment and context

across varied wording.

Applied keyword-based matching

to identify CRISPR-related terms

in the data processing section.

• Datasets were assigned weights

related terms to reflect their

Automated approaches such as text-search,

faster and more scalable alternatives.

semantic-search, and LLM-based methods offered

• Together, these automated methods compensated for

the time cost of manual curation. When combined they

provided an effective triaging system, rapidly narrowing

down candidate datasets for closer manual validation.

based on the presence of CRISPR

Used a predefined list of

CRISPR-specific keywords

(CAS9, CRISPRi, CRISPRa).

- Utilized string-matching approaches to detect CRISPR-related terms within dataset metadata (esp. data processing fields).
- Constructed a predefined keyword list covering multiple CRISPR screen terminologies (e.g., CRISPR screen experiment, sgRNA-based study, CRISPR perturbation, sgRNA counts, raw CRISPR screen
- Expanded search to include different forms of labeling within Library Strategy fields (e.g., "CRISPR screen").
- Applied keyword search across both Series (GSE) and Sample (GSM) metadata levels.
- Identified candidate datasets for further validation by manual review, NLP-based semantic search, and Al classification.
- Helped capture datasets that were correctly labeled but otherwise missed due to inconsistent annotation across repositories.

## Semantic Search

- Implemented semantic search to identify CRISPR screen studies across multiple metadata fields (e.g., summary, overall design, data processing). FremyCompany/BioLORD-2023
- Sentence Transformer performed relatively better for this task (10-15% increased sensitivity at similar specificity), as it is heavily trained on clinical sentences and biomedical concepts. Generated sentence embeddings from each dataset's processing-related text using the BioLORD-2023 model.
- Compared embeddings against predefined CRISPR-specific reference texts using cosine similarity.
- Flagged datasets exceeding a similarity threshold as potential CRISPR screen studies. This approach focused on conceptual understanding rather than exact keyword
- However, often misclassified datasets where CRISPR-related terms appeared in the description as part of the analysis, but the dataset was not a true CRISPR screen (e.g., different library strategy also used in the experiment).

## **GEO LLM Plugin**

## CRISPR GEO Analyzer

This page contains CRISPR screen data! Confidence: High

The webpage for GEO Series GSE125507 describes a genome-wide CRISPR-Cas9 screening performed on KSHV-transformed cells to identify key genes responsible for cell survival, specifically targeting the exportin XP01. The presence of supplementary files named with 'reads\_count followed by sample identifiers and time points (e.g., 'GSE125507 \_reads\_count\_KMMd1.txt.gz') indicates the availability of sgRNA count data from the screening. These files likely contain sgRNA count matrices or gene-level screen data, fulfilling the criteria for a true positive CRISPR experiment. Relevant terms found: CRISPR-Cas9 screening, sgRNAs,

Casg-expressing, genome-wide CRISPR-Cas9, CRISPR pooled

Samples (24) <u>H More</u>	GSM3576172 MMd1_rep1					
	GSM3576173 MMd1 rep2					
	GSM3576174 MMd1_rep3					
Relations						
BioProject	PRJNA516644					
SRA	SRP181626					
Download fam	ily		Format			
SOFT formatted family file(s)			SOFT ?			
MINIML formatted family file(s)			MINIML ?			
MINIML formatt	ed family file(s)		MIN	liML 🛽		
MINIML formatt Series Matrix Fi				liML ? · ?		
		Size		?		
Series Matrix Fi	ile(s)	Size 2.1 Mb	TXT	?		
Series Matrix Fi	Supplementary file		Download	File type/resource		
Series Matrix Fi	Supplementary file ads_count_KMMd1.txt.gz	2.1 Mb	Download (ftp)(http)	File type/resource		
GSE125507_re GSE125507_re GSE125507_re	Supplementary file ads_count_KMMd1.txt.gz ads_count_KMMd11.txt.gz	2.1 Mb 2.1 Mb	Download (ftp)(http) (ftp)(http)	File type/resource TXT TXT		
GSE125507_re GSE125507_re GSE125507_re GSE125507_re	Supplementary file ads_count_KMMd1.txt.gz ads_count_KMMd21.txt.gz ads_count_KMMd21.txt.gz	2.1 Mb 2.1 Mb 2.1 Mb	Download (ftp)(http) (ftp)(http) (ftp)(http)	File type/resource TXT TXT TXT		
GSE125507_re GSE125507_re GSE125507_re GSE125507_re GSE125507_re	Supplementary file ads_count_KMMd1.txt.gz ads_count_KMMd21.txt.gz ads_count_KMMd21.txt.gz ads_count_KMMd21.txt.gz	2.1 Mb 2.1 Mb 2.1 Mb 2.1 Mb	Download (ftp)(http) (ftp)(http) (ftp)(http) (ftp)(http)	File type/resource TXT TXT TXT TXT		
GSE125507_re GSE125507_re GSE125507_re GSE125507_re GSE125507_re GSE125507_re	Supplementary file ads_count_KMMd1.txt.gz ads_count_KMMd21.txt.gz ads_count_KMMd21.txt.gz ads_count_KMMd4.txt.gz ads_count_KMMd4.txt.gz	2.1 Mb 2.1 Mb 2.1 Mb 2.1 Mb 2.1 Mb	Download (ftp)(http) (ftp)(http) (ftp)(http) (ftp)(http) (ftp)(http)	File type/resource TXT TXT TXT TXT TXT TXT		

- Developed a Chrome browser plugin to identify whether dataset is a CRISPR screen based on its GEO (GSE/GSM) page.
- The plugin scrapes GEO page content and compiles it as contextual input. Content is passed to GPT-40, which evaluates datasets using predefined inclusion/exclusion criteria. Furthermore, bulk dataset classification can be performed by executing GPT API calls with the specified context on the backend via a Python script.
- Positive classification: presence of CRISPR screen outputs (e.g., sgRNA count matrices, gene-level results, raw screening files).

## Performance Comparison of Classification Approaches

- The performance of separate approaches was based on 100 randomly selected datasets.
- Manual curation ensures high sensitivity and specificity but requires substantial manual effort and deep understanding of experimental design and biological context to accurately review each series/experiment ID and its samples.
- Pattern matching (text-based search) is highly sensitive but lacks specificity, leading to more false
- Semantic analysis offers medium to low sensitivity and specificity, and requires extensive tuning for optimal performance.
- Large Language Models (LLMs) provide both high sensitivity and high specificity, and are efficient and scalable with proper tuning.

Approach	Sensitivi	ty Spec	ificity	Notes
Manual Curatio	on High	Hi	bu ma ea ID	sures accurate data t require significant nual effort to review ch series/experimen and its associated mples.
Pattern Match (Text-Based)	ing High	Lo	J VV	nsitive but lacks ecificity.
Semantic Analysis	Med to L		to Low tur	quires extensive ning for optimal sults.
Large Languag Models(LLMs)	ge High	า	riigii	ricient and alable with tuning
Approach		Precision		<b>Accuracy</b>
	Yes	No	Maybe	Accuracy

70.70% 60.00% 15.00%

58.30% 33.30%

82%

77.40% 50%

- Negative classification: datasets limited to single-cell, RNA-seq, ATAC-seq; CRISPR editing/knockout without screening results; or CRISPR mentioned without relevant
- The plugin makes a backend API call, sending scraped content with structured prompts encoding the rules.

87.30%

GPT-40 returns: 1) Classification (CRISPR screen or not) 2) Confidence score 3) Reasoning 4) Extracted keywords supporting the decision