

# Supplementary Note: ATAC-seq to Gene Programs and Gene Sets in dig-gene-set-extractors

## Overview and goals

The `dig-gene-set-extractors` repository provides a suite of converters that map diverse omics data types to: (i) *gene programs* (scores or weights over genes) and (ii) compact *gene sets* (100–500 genes) suitable for enrichment and mechanism-level interpretation.

This note focuses on ATAC-seq converters. ATAC-seq measures chromatin accessibility at genomic loci, so any gene-level representation is a model-based projection from locus space to gene space. In practice, users rarely want “genes with the most open chromatin” in an absolute sense; they typically want gene sets that reflect the *biological question that motivated generating ATAC-seq*, such as:

- cell type or cell state identity programs (scATAC clusters),
- condition or phenotype-associated regulatory changes (opening vs closing),
- enhancer-driven (distal) regulation rather than broadly open promoters,
- upstream regulator (TF) programs,
- and (optionally) reference-calibrated tissue specificity or atlas matching.

Accordingly, the ATAC-seq pipeline is designed as a *program factory*: for a given dataset (and optionally for each group/cluster), it can emit multiple named program instances. Each program instance is then converted to one or more gene sets and exported as `.gmt`.

## Default behavior and extensibility

The implementation supports two usage modes:

- **Sensible defaults:** run a small number of high-value, low-dependency program methods that work broadly across datasets.
- **All available methods:** run all program methods that are requested, skipping or failing on methods whose auxiliary resources are missing according to user preference.

This note is organized to match that workflow: we define a unified scoring framework and then describe a catalog of program methods, grouped by motivation and data requirements, including the auxiliary resources needed for each method.

# 1 Notation

- Peaks (regions) indexed by  $p \in \{1, \dots, P\}$  with coordinates  $(c_p, s_p, e_p)$  and midpoint  $m_p = (s_p + e_p)/2$ .
- Genes indexed by  $g \in \{1, \dots, G\}$  with chromosome  $c_g$ , strand  $\sigma_g \in \{+, -\}$ , TSS position  $t_g$ , promoter interval  $\mathcal{P}_g$ , and gene locus interval  $\mathcal{G}_g$ .
- In bulk ATAC, samples indexed by  $i \in \{1, \dots, n\}$ .
- In scATAC, cells indexed by  $c \in \{1, \dots, C\}$ , with peak-by-cell matrix  $X_{pc} \geq 0$  (counts).
- A group/cluster label function  $\gamma(c) \in \{1, \dots, K\}$ , where  $\mathcal{C}_k = \{c : \gamma(c) = k\}$ .
- A peak-level statistic  $x_p$  (absolute accessibility, differential statistic, association statistic, or other).
- A peak weight  $\alpha_p \geq 0$  derived from  $x_p$  via a transform  $\phi$  (Sections ?? and ??).
- A peak-to-gene linkage weight  $L_{pg} \geq 0$  (Section ??).
- A gene score  $s_g$  (real-valued or nonnegative) and optionally a normalized distribution  $w_g$ .

## 2 Unified scoring framework

### 2.1 Peak weights

Define a transform  $\phi$  that maps a peak statistic  $x_p$  to a peak weight  $\alpha_p$ :

$$\alpha_p = \phi(x_p).$$

Common transforms:

$$\begin{aligned}\phi_{\text{signed}}(x) &= x, \\ \phi_{\text{abs}}(x) &= |x|, \\ \phi_{\text{pos}}(x) &= \max(x, 0), \\ \phi_{\text{neg}}(x) &= \max(-x, 0).\end{aligned}$$

For directional programs (opening vs closing), apply  $\phi_{\text{pos}}$  and  $\phi_{\text{neg}}$  separately to obtain two non-negative peak-weight vectors.

### 2.2 Peak-to-gene projection

Given peak weights  $\alpha_p$  and link weights  $L_{pg}$ , define a raw gene score:

$$s_g = \sum_{p=1}^P \alpha_p L_{pg}. \tag{1}$$

Equation (1) is the core projection used throughout. Biological meaning is determined by: (i) how  $x_p$  is defined (absolute vs contrast vs association), and (ii) how  $L_{pg}$  is defined (promoter vs distal vs learned wiring).

### 2.3 From scores to distributions (optional)

If  $s_g \geq 0$  and not all zero, a global L1-normalized distribution is:

$$w_g = \frac{s_g}{\sum_{h=1}^G s_h}, \quad \sum_{g=1}^G w_g = 1. \quad (2)$$

This is a *relative allocation* across genes, not a per-gene probability of activity. When the goal is compact programs (100–500 genes), global normalization can make weights appear numerically small and visually uniform if many genes receive nonzero score. Therefore, gene set extraction (Section ??) is an explicit step, and within-program normalization is preferred for interpretability.

## 3 Peak-to-gene linkage models

The linkage model encodes biological assumptions about cis-regulatory wiring. All linkage models can be combined with any definition of  $x_p$  and any gene set extraction operator.

### 3.1 Link A: promoter overlap

**Biology.** Promoter accessibility is directly connected to transcriptional competence and is minimally ambiguous to assign.

**Definition.**

$$L_{pg} = \mathbf{1}\{c_p = c_g \wedge [s_p, e_p) \cap \mathcal{P}_g \neq \emptyset\}.$$

**Heuristics.** Promoter window  $(u, d)$  around the TSS (e.g.,  $u = 2000$  bp upstream,  $d = 500$  bp downstream). Record whether transcript union or a canonical TSS is used.

### 3.2 Link B: nearest TSS assignment

**Biology.** A simple distal heuristic: regulatory elements often regulate a nearby gene.

**Definition.** With distance  $d(p, g) = |m_p - t_g|$  and max distance  $D$ :

$$g^*(p) = \arg \min_{g: d(p,g) \leq D} d(p, g), \quad L_{pg} = \mathbf{1}\{g = g^*(p)\}.$$

**Heuristics.** Use per-chromosome sorted TSS arrays for  $O(\log G)$  nearest lookup. Tie-break deterministically by gene ID.

### 3.3 Link C: distance-decay soft assignment

**Biology.** Contact probability and average regulatory influence decrease with genomic distance.

**Definition.** For decay length  $\lambda$  and max distance  $D$ :

$$A_{pg} = \begin{cases} \exp(-d(p, g)/\lambda), & d(p, g) \leq D \\ 0, & \text{otherwise.} \end{cases}$$

Either use  $L_{pg} = A_{pg}$  (unnormalized), or normalize per peak:

$$L_{pg} = \frac{A_{pg}}{\sum_h A_{ph} + \epsilon}.$$

**Heuristics.** Defaults such as  $D = 500$  kb,  $\lambda = 50$  kb; optionally cap to the top  $K_{\max}$  genes per peak by  $A_{pg}$ .

### 3.4 Link D: gene locus activity (gene body plus promoter)

**Biology.** Aggregating accessibility across promoter and gene body yields a gene-activity proxy useful in sparse scATAC.

**Definition.** Define a gene locus interval  $\mathcal{G}_g$  and compute:

$$s_g = \sum_{p:[s_p, e_p) \cap \mathcal{G}_g \neq \emptyset} \alpha_p.$$

**Heuristics.** This can induce gene-length bias. Mitigate by standardized windows, length normalization, or by using this as a complementary program rather than the sole output.

### 3.5 Link E: co-accessibility linkage

**Biology.** Peaks that covary across cells can reflect enhancer-promoter coupling.

**Definition.** Let  $C_{pq}$  be a peak-peak co-accessibility score. For each gene  $g$  with promoter peak(s)  $q \in \mathcal{Q}(g)$ :

$$L_{pg} = \max \left( \max_{q \in \mathcal{Q}(g)} C_{pq}, 0 \right) \cdot \mathbf{1}\{d(p, g) \leq D\}.$$

**Heuristics.** Requires many cells. Restrict candidate pairs to windows (e.g., 250–500 kb), threshold weak edges, and record co-accessibility method.

### 3.6 Link F: external enhancer-gene priors

**Biology.** External enhancer-gene maps integrate multi-assay evidence and can improve specificity.

**Definition.** Overlap peaks to external elements  $e$  with a link matrix  $M_{eg}$ :

$$L_{pg} = \sum_e \mathbf{1}\{p \cap e \neq \emptyset\} M_{eg}.$$

**Heuristics.** Choose a default map per genome build; record version and provenance. Handle biosample mismatch explicitly (closest match vs union vs average).

### 3.7 Link G: activity-by-contact style linkage (optional)

**Biology.** Enhancer effect is proportional to enhancer activity and contact with a promoter.

**Definition.** With an activity estimate  $A_p$  (from ATAC or multi-assay) and a contact kernel  $K_{pg}$ :

$$S_{pg} = A_p K_{pg}, \quad L_{pg} = \frac{S_{pg}}{\sum_h S_{ph} + \epsilon}.$$

**Heuristics.** If no contact map is available, approximate  $K_{pg}$  by a distance kernel (exponential or power-law), and record the approximation.

## 4 From gene scores to gene sets

This section specifies how a gene-score vector (or distribution) is converted into one or more gene sets of size 100–500.

### 4.1 Selection operators

Let  $\mathbf{s} = (s_1, \dots, s_G)$  be gene scores for a program instance. The default operator is fixed-size top- $K$ .

**S1: top- $K$  selection (default).** Sort genes by decreasing score  $s_{(1)} \geq \dots \geq s_{(G)}$  with genes  $g_{(1)}, \dots, g_{(G)}$ , tie-breaking deterministically by gene ID. Define:

$$\hat{\mathcal{G}}_K = \{g_{(1)}, \dots, g_{(K)}\}, \quad K \in [100, 500].$$

Within-program weights (optional) are:

$$\tilde{w}_g = \begin{cases} \frac{\max(s_g, 0)}{\sum_{h \in \hat{\mathcal{G}}_K} \max(s_h, 0)}, & g \in \hat{\mathcal{G}}_K \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

If the denominator is zero, fall back to uniform weights within  $\hat{\mathcal{G}}_K$ .

**S2: HPD mass set.** If scores are converted to a nonnegative distribution  $w_g$  (Eq. 2), an HPD mass set for target mass  $\tau \in (0, 1)$  is:

$$K^*(\tau) = \min \left\{ K : \sum_{i=1}^K w_{(i)} \geq \tau \right\}, \quad \hat{\mathcal{G}}_\tau = \{g_{(1)}, \dots, g_{(K^*(\tau))}\}.$$

To enforce 100–500 genes, clamp  $K$  to  $[K_{\min}, K_{\max}]$  and use  $\hat{\mathcal{G}}_K$ .

**S3: multi-resolution gene sets.** For a fixed score vector, optionally emit multiple nested sets (e.g.,  $K \in \{100, 200, 500\}$ ) and/or disjoint tiers (e.g., ranks 1–100, 101–200, etc.) for sensitivity analysis.

## 4.2 Signed programs

If a program produces signed gene scores, define nonnegative components:

$$s_g^+ = \max(s_g, 0), \quad s_g^- = \max(-s_g, 0),$$

and apply selection operators separately to obtain two gene sets (e.g., opening vs closing).

## 5 GMT export

We export gene sets as `.gmt` files using a simplified one-tab, space-delimited convention:

```
<gene_set_name>\t<gene1> <gene2> ... <geneN>.
```

Gene tokens are gene symbols when available and non-empty; otherwise stable gene IDs. Within a set, duplicates are removed while preserving order.

**Gene set names.** A robust default concatenates provenance fields, for example:

```
<dataset_id>_<converter>_<group>_<program_method>_<link>_<select>.
```

Names are sanitized to remove whitespace and path separators.

## 6 ATAC program methods catalog

This section defines a catalog of program methods. Each method corresponds to a biological motivation and specifies: (i) how peak weights  $\alpha_p$  are constructed, (ii) which peaks are included (promoter vs distal), (iii) which linkage model  $L_{pg}$  is used, and (iv) what program instances are emitted (one per sample, per group, per direction, per module, etc.).

### 6.1 Summary table

Table ?? summarizes the most important methods, their typical use cases, and resource requirements.

Table 1: ATAC program methods (non-exhaustive). Defaults are intended to be high-value and low-dependency.

Method ID	What it captures	Requires	Default?
promoter_activity	Broad promoter accessibility (QC / global activity)	Peaks + GTF	Yes
distal_activity	Enhancer-driven accessibility (more tissue-specific)	Peaks + GTF	Yes
enhancer_bias	Distal vs promoter residual (enhancer emphasis)	Peaks + GTF	Yes
group_vs_rest_open	Cluster-specific opening (scATAC)	scATAC + groups	Yes
group_vs_rest_close	Cluster-specific closing (scATAC)	scATAC + groups	Optional
tfidf_group_open	TF-IDF specificity-weighted cluster program	scATAC matrix	Yes (sc)

Method ID	What it captures	Requires	Default?
ref_ubiquity_penalty	Penalize elements open in many tissues/cells	Reference cCRE/atlas	Optional
atlas_residual	Sample-specific vs atlas baseline	Reference atlas	Optional
motif_regulons	TF driver programs via motif hits	Motifs + genome/hits	Optional
coaccess_modules	Peak communities - $i$ gene modules	Many cells	Optional
topic_programs	Latent topics (NMF/LDA) - $i$ gene programs	Many cells	Optional
phenotype_open/cls	Condition/phenotype differential programs	Multi-sample labels	Optional
variant_anchored	Variant-overlap anchored gene programs	Variant set	Optional

## 6.2 Method family 1: absolute activity programs

These methods operate on an absolute accessibility statistic per peak and are most appropriate for: (i) multi-sample studies (where differential methods can be used) or (ii) quality control summaries. For single-sample bulk ATAC, absolute promoter activity tends to yield broadly open chromatin genes; distal and specificity-weighted programs are preferred.

### 6.2.1 M1.1: promoter\_activity

**Peak weights.** Choose an absolute peak statistic  $x_p$  (e.g., normalized coverage, peak score, or group summary). Set  $\alpha_p = \phi_{\text{pos}}(x_p)$  to ensure nonnegativity.

**Peak filter.** Keep only peaks overlapping any promoter interval  $\mathcal{P}_g$ .

**Link.** Use Link A (promoter overlap).

**Gene score.** Compute  $s_g$  by Eq. 1.

**Gene sets.** Select top- $K$  genes (Section ??).

### 6.2.2 M1.2: distal\_activity

**Motivation.** Distal regulatory elements are often more tissue- and state-specific than promoters.

**Peak weights.** Same as promoter\_activity.

**Peak filter.** Keep only distal peaks:

$$\mathcal{D} = \{p : \forall g, [s_p, e_p) \cap \mathcal{P}_g = \emptyset\},$$

or equivalently peaks with distance to nearest TSS greater than a chosen promoter radius.

**Link.** Use Link C (distance-decay) or Link B (nearest TSS) for distal assignment.

**Gene score and sets.** As above.

### 6.2.3 M1.3: enhancer\_bias (distal vs promoter residual)

**Motivation.** A gene may be broadly expressed (open promoter) but not enhancer-driven in the sample. This method prioritizes genes whose distal accessibility is high relative to promoter accessibility.

**Definition.** Compute two nonnegative scores:

$$s_g^{\text{prom}} = \sum_{p \in \text{promoters}} \alpha_p L_{pg}, \quad s_g^{\text{dist}} = \sum_{p \in \text{distal}} \alpha_p L_{pg}.$$

Define an enhancer-bias score:

$$b_g = \log \left( \frac{s_g^{\text{dist}} + \epsilon}{s_g^{\text{prom}} + \epsilon} \right). \quad (4)$$

Alternatively, normalize each component across genes before differencing:

$$b_g = \log(s_g^{\text{dist}} + \epsilon) - \log(s_g^{\text{prom}} + \epsilon), \quad \text{or} \quad b_g = z(s_g^{\text{dist}}) - z(s_g^{\text{prom}}).$$

**Gene sets.** Select top- $K$  genes by  $b_g$ .

## 6.3 Method family 2: contrasts and specificity (within-study)

These methods are designed for the most common scientific question: *what differs across groups, conditions, or cell states?*

## 6.4 scATAC peak summaries

For scATAC, define a group-level peak summary for group  $\mathcal{C}_k$ :

$$\begin{aligned} \text{sum counts: } \bar{x}_p^{(k)} &= \sum_{c \in \mathcal{C}_k} X_{pc}, \\ \text{mean counts: } \bar{x}_p^{(k)} &= \frac{1}{|\mathcal{C}_k|} \sum_{c \in \mathcal{C}_k} X_{pc}, \\ \text{fraction nonzero: } \bar{x}_p^{(k)} &= \frac{1}{|\mathcal{C}_k|} \sum_{c \in \mathcal{C}_k} \mathbf{1}\{X_{pc} > 0\}. \end{aligned}$$

### 6.4.1 M2.1: group\_vs\_rest\_open (scATAC default)

**Motivation.** Cell identity and state are most naturally described by accessibility that is *specific* to a group vs background.

**Peak statistic.** For group  $k$ , let  $\bar{x}_p^{(k)}$  be the group summary and  $\bar{x}_p^{(\neg k)}$  be the background (all other cells) summary. Define:

$$x_p^{(k)} = \log_2 \left( \frac{\bar{x}_p^{(k)} + a}{\bar{x}_p^{(\neg k)} + a} \right), \quad (5)$$

with pseudocount  $a > 0$  (e.g.,  $a = 10^{-3}$  for fractions,  $a = 1$  for counts).

**Peak weights.**  $\alpha_p^{(k)} = \max(x_p^{(k)}, 0)$  (opening program).

**Link.** Use Link A for promoter-focused cluster markers and Link C for distal-aware markers; both may be emitted.

**Gene sets.** Select top- $K$  genes.

#### 6.4.2 M2.2: group\_vs\_rest\_close (optional)

Same as M2.1 but with  $\alpha_p^{(k)} = \max(-x_p^{(k)}, 0)$  to capture peaks that are less accessible in the group than background.

#### 6.4.3 M2.3: tfidf\_group\_open (scATAC default)

**Motivation.** Peaks that are accessible in many cells are less informative for defining cell identity. A TF-IDF style reweighting emphasizes more specific peaks.

**Definition.** Let  $N = C$  be the number of cells in the dataset and define the document frequency:

$$df_p = \sum_{c=1}^C \mathbf{1}\{X_{pc} > 0\}.$$

For group  $k$ , define a term frequency:

$$tf_p^{(k)} = \sum_{c \in \mathcal{C}_k} X_{pc} \quad \text{or} \quad tf_p^{(k)} = \sum_{c \in \mathcal{C}_k} \mathbf{1}\{X_{pc} > 0\}.$$

Define:

$$idf_p = \log\left(\frac{N+1}{df_p + 1}\right) + 1, \quad \alpha_p^{(k)} = \log(1 + tf_p^{(k)}) \cdot idf_p. \quad (6)$$

**Gene projection and sets.** Apply Eq. 1 and top- $K$  selection. As with M2.1, both promoter-only and distal-only variants may be emitted.

#### 6.4.4 M2.4: phenotype\_open / phenotype\_close (bulk or sc, optional)

**Motivation.** Many ATAC studies compare donors/conditions (case vs control, treatment, time, severity).

**Peak statistic.** Suppose we have per-sample peak measurements  $x_{p,i}$  and a phenotype  $y_i$ . Fit a regression model per peak:

$$x_{p,i} = \beta_{p,0} + \beta_{p,1}y_i + \mathbf{z}_i^\top \boldsymbol{\gamma}_p + \varepsilon_{p,i},$$

where  $\mathbf{z}_i$  are covariates. Let  $t_p$  or  $z_p$  be a signed test statistic for  $\beta_{p,1}$ .

**Peak weights.** Opening:  $\alpha_p^+ = \max(z_p, 0)$ ; closing:  $\alpha_p^- = \max(-z_p, 0)$ .

**Gene projection.** Use any linkage model (promoter or distal-aware).

**Gene sets.** Emit top- $K$  genes for opening and closing.

## 6.5 Method family 3: reference-calibrated specificity (optional resources)

Absolute accessibility and within-study contrasts do not always provide tissue specificity for single-sample bulk ATAC. Reference-based calibration addresses this.

### 6.5.1 M3.1: ref\_ubiquity\_penalty (reference IDF)

**Motivation.** Penalize peaks that are open in many tissues/cell types; emphasize tissue- or state-specific regulatory DNA.

**Inputs.** A reference registry of regulatory elements (e.g., a cCRE catalog) with an “accessibility ubiquity” statistic across reference biosamples.

**Definition.** Let  $df^{\text{ref}}(e)$  be the number of reference biosamples in which element  $e$  is accessible, and  $N_{\text{ref}}$  the number of reference biosamples. Define:

$$idf^{\text{ref}}(e) = \log \left( \frac{N_{\text{ref}} + 1}{df^{\text{ref}}(e) + 1} \right) + 1.$$

If a peak  $p$  overlaps element  $e$ , modify its weight:

$$\alpha_p^{\text{adj}} = \alpha_p \cdot idf^{\text{ref}}(e), \quad (7)$$

or sum over multiple overlapping elements. Then project to genes by Eq. 1.

**Heuristics.** If no overlap is found, use  $idf^{\text{ref}} = 1$ . Record the reference catalog version and the exact definition of  $df^{\text{ref}}$ .

### 6.5.2 M3.2: atlas\_residual (gene-level reference residualization)

**Motivation.** Rather than reweight peaks, calibrate at the gene-score level relative to an atlas of reference tissues/cell types.

**Inputs.** A reference panel of gene scores  $s_g^{(t)}$  for reference context  $t$  (precomputed using the same projection model).

**Definitions.** Let  $m_g = \text{median}_t(s_g^{(t)})$  and  $d_g = \text{MAD}_t(s_g^{(t)})$  (median absolute deviation). Define either:

$$r_g = \log \left( \frac{s_g + \epsilon}{m_g + \epsilon} \right), \quad (8)$$

$$z_g = \frac{s_g - m_g}{d_g + \epsilon}. \quad (9)$$

**Gene sets.** Select top- $K$  genes by  $r_g$  or  $z_g$  to obtain a reference-calibrated specificity program.

## 6.6 Method family 4: regulator-centric programs (optional resources)

### 6.6.1 M4.1: motif\_regulons

**Motivation.** A common ATAC question is “which TFs are driving the regulatory program?” Motif-linked programs provide a gene-level readout of TF activity hypotheses.

**Inputs.** A motif database and motif occurrences in peaks (computed or precomputed).

**Definition.** For motif  $m$ , let  $\mathcal{P}(m)$  be the peaks containing that motif (optionally weighted by match score). Define a motif-specific gene score:

$$s_g^{(m)} = \sum_{p \in \mathcal{P}(m)} \alpha_p L_{pg}. \quad (10)$$

**Program selection.** To avoid producing hundreds of sets, restrict to motifs with high total mass  $\sum_g s_g^{(m)}$  or motifs enriched among top peaks. Emit top- $K$  genes per selected motif.

## 6.7 Method family 5: module discovery (optional, heavier compute)

These methods produce *multiple* programs per dataset by discovering regulatory modules in peak space and projecting them to genes.

### 6.7.1 M5.1: coaccess\_modules

**Inputs.** scATAC peak-by-cell matrix with sufficient cell count.

**Definition.** Compute a co-accessibility graph  $G = (V, E)$  where nodes are peaks and edges connect peaks with  $C_{pq} > \theta$  within a distance window. Cluster  $G$  into modules  $\mathcal{M}_1, \dots, \mathcal{M}_M$ . For module  $k$ , define peak weights:

$$\alpha_p^{(k)} = \alpha_p \cdot \mathbf{1}\{p \in \mathcal{M}_k\}.$$

Project to genes by Eq. 1 to obtain  $s_g^{(k)}$ , then select top- $K$  genes to define a gene program per module.

### 6.7.2 M5.2: topic\_programs (NMF/LDA)

**Inputs.** scATAC peak-by-cell matrix (typically TF-IDF normalized).

**Definition.** Fit a nonnegative factorization:

$$X \approx WH,$$

where  $W$  is peak-by-topic and  $H$  is topic-by-cell. For topic  $k$ , define peak weights  $\alpha_p^{(k)} = W_{pk}$  and project to genes:

$$s_g^{(k)} = \sum_p W_{pk} L_{pg}.$$

Select top- $K$  genes per topic to obtain topic gene sets.

## 6.8 Method family 6: variant-anchored programs (optional)

### 6.8.1 M6.1: variant\_anchored

**Motivation.** ATAC-seq is often used to interpret noncoding variants (GWAS fine-mapping, QTL) by intersecting variants with accessible chromatin.

**Inputs.** A set of variants with optional weights (e.g., posterior inclusion probabilities, PIPs) and genome build matching the peaks.

**Definition.** Let  $\mathcal{V}(p)$  be variants overlapping peak  $p$ . Define:

$$\alpha_p = \sum_{v \in \mathcal{V}(p)} w_v,$$

with  $w_v = 1$  if unweighted. Project to genes by Eq. 1 using a distal-aware linkage (Link C/F/G). Select top- $K$  genes to obtain a variant-anchored gene program.

## 7 Program presets and execution policy

To balance usability and flexibility, program methods are grouped into presets.

### 7.1 Default preset (high-value, low-dependency)

A recommended default preset emits:

- **bulk:** promoter\_activity, distal\_activity, enhancer\_bias (and opening/closing if signed peak statistics are provided).
- **scATAC (grouped):** group\_vs\_rest\_open (and optionally close), tfidf\_group\_open, plus distal variants.

These methods require only the input peaks/matrix and a gene annotation (GTF).

### 7.2 All-methods preset (resource-aware)

An “all” preset runs all requested methods for which required resources are available in the configured resources directory. Missing resources can be handled by: (i) skipping the method while recording that it was skipped, or (ii) failing the run (fail-fast). The chosen policy is recorded in metadata.

## 8 Auxiliary resources and reproducibility

Many optional methods require external resources (reference catalogs, atlases, motifs). The repository adopts best practices:

- Resources are identified by stable IDs and versioned.
- Downloads are handled by a resource manager to a single directory.
- Each resource is verified by checksum (e.g., SHA256) and recorded in output metadata.
- Outputs record genome build, annotation version, and all method parameters needed for reproducibility.

**Bundled vs external.** Small, permissively licensed resources may be bundled in the repository. Larger resources are hosted externally (e.g., on Zenodo) and fetched on demand.

## 9 Output artifacts and metadata

Each program instance is written to its own output directory containing:

- `geneset.full.tsv`: all genes with nonzero score (or all genes, depending on implementation),
- `geneset.tsv`: the selected program-sized gene list (100–500 genes) with scores and optional within-program weights,
- `geneset.meta.json`: machine-readable metadata describing inputs, methods, parameters, and resources,
- `genesets.gmt`: one or more gene sets derived from this program instance.

For runs producing multiple program instances (e.g., many scATAC clusters, or multiple method families), a root manifest enumerates all program directories and enables validation of the full output tree.

## 10 Practical heuristics and QC checks

- **Genome build consistency.** Ensure peaks, variants, and reference resources use the same build (e.g., hg38 vs hg19). Mismatches often lead to low assignment rates.
- **Chromosome naming.** Handle `chr1` vs `1` consistently; record any normalization performed.
- **Promoter dominance.** If `promoter_activity` yields broad housekeeping programs, prefer `distal_activity`, `enhancer_bias`, TF-IDF, or reference-calibrated methods.
- **Assignment diagnostics.** Record the fraction of peaks assigned to at least one gene and the distribution of links per peak.
- **Determinism.** Ensure sorting tie-breakers are deterministic to make runs reproducible.

## 11 Computational complexity (high level)

For a single program instance with  $P$  peaks and  $G$  genes, projection via Eq. 1 is  $O(\#links)$ . In distance-based models,  $\#links$  is controlled by max distance and link caps per peak. Gene set extraction is dominated by sorting gene scores:  $O(G \log G)$ .