Unsupervised Detection of Transcription Factor Communities

Gregory Orme

October 23, 2025

1 Introduction

Traditional approaches to transcription factor (TF) taxonomy have relied on structural analysis of DNA binding domains (DBD), yielding familiar shape-based family names like zinc finger, basic leucine zipper, and basic helix-loop-helix. My goal in this paper is to apply unsupervised methods to produce alternative functional modules directly from chromatin accessibility data. Using ATAC-seq from the K562 cell line, I scan accessible peaks with TF position weight matrices (PWMs) and link matched TFs to nearby genes. These scored TF→gene pairs become edges in a bipartite directed graph, where edges represent putative regulatory influence. I then apply Louvain community detection to this graph and interpret each module using gene-set enrichment of its downstream genes, comparing the results with known family biology. This unsupervised approach proposes novel functional groupings of TFs and putative regulatory programs which complement the traditional DBD-based taxonomy.

2 Data/Provenance

2.1 ATAC-seq peaks (K562)

I sourced IDR ranked ATAC-seq peaks from the K562 biosample from **20 ENCODE experiments**, using **80 files** (assembly=`GRCh38`). After merging peaks, I retained a final set of **150,000** intervals for downstream analysis.

- **Biosample:** K562 **Organism:** Human

- **Assembly:** `GRCh38` **Output type:** IDR ranked peaks
- **Labs:** Will Greenleaf (Stanford), *et al.* (see Appendix for full provenance)
- **Experiments:** 20 **Files used:** 80 **Peaks retained:** 150,000

2.2 Reference genome

GENCODE v43 (Human, GRCh38.p13) was downloaded and used to derive TSS coordinates (FASTA; indexed with `samtools faidx`). See appendix for full provenance.

2.3 Gene annotation and TSS

I defined each gene's TSS as the 5' end of its representative transcript—i.e., the transcript start for '+' strand genes and the transcript end for '–' strand genes. I then converted from 1-based (GTF) to 0-based (BED) to write a 1-bp interval $[tss, tss{+}1)$. This yielded **20,029** TSS records on canonical chromosomes (chr1–22, X, Y). To reduce low-specificity matches in downstream enrichment, I excluded olfactory genes (regex pattern in Appendix). **19,591** records remained after this screen, corresponding to **19,562** distinct gene symbols.

- **File:** `gencode_tss.bed`
- **Window size (mode):** 1 bp
- **Chromosomes:** chr1–22, X, Y
- **Distinct gene symbols (post-filter):** 19,562

Provenance: see appendix.

2.4 Motif database

Position frequency matrices (PFM) from *JASPAR 2024 CORE (vertebrates)* were used for TF motif modeling.

3 Methods

3.1 Peak set construction

I included only IDR-ranked peaks from canonical chromosomes when querying experiments from ENCODE to maximize the stability and reproducibility of results.

3.2 Background sequence construction

For each peak, I sampled a non-overlapping, non-peak interval of identical length from the same chromosome using `bedtools shuffle` (`-chrom -excl peaks -noOverlapping`, seeded) to control for chromosome- and length-specific biases. Then, for each 6-mer, I summed frequencies across these background sequences to produce an empirical distribution for global kmer frequencies. The goal is to measure sequence features that are enriched in peaks instead of features that are common in the genome at large. A matched background provides an empirical null model for the distribution of 6-mer frequencies in a global sample with similar length and chromosome composition to our chosen peaks.

3.3 FASTA extraction and 6-mer counting

I used `samtools faidx` and `bedtools getfasta` to extract sequences for foreground (peaks) and background from `hg38`. Canonical 6-mer counts were computed with Jellyfish (`-m 6 -C`); enrichment was defined as $\log_2((\text{fg}+1)/(\text{bg}+1))$, where fg denotes the total 6-mer count in the foreground and bg in the background.

3.4 Motif selection and scoring

For TFs with multiple JASPAR PFMs, I kept the motif with the highest information content and converted that motif to a Position-Specific Scoring Matrix (PSSM) using Biopython. For each (6-mer, motif) pair, I calculated the maximal log-odds score by sliding the 6-mer across both strands of the motif and selecting the best window. Background 6-mer frequencies defined an empirical distribution of scores per motif, which we later use to weight (TF, gene) pairs.

More precisely, let m be a motif of length L with PFM counts $C_{b,i}$ for base $b \in \{A, C, G, T\}$ at position $i \in \{1, \ldots, L\}$. With a pseudocount $\alpha = 0.1$ and a uniform background $q_b = \frac{1}{4}$, define

$$p_{b,i} \;=\; \frac{C_{b,i} + \alpha}{\sum_{b'} (C_{b',i} + \alpha)}, \qquad \ell_{b,i} \;=\; \log \frac{p_{b,i}}{q_b}\,,$$

where $\ell_{b,i}$ is the PSSM log-odds (Biopython). The information content is

$$\mathrm{IC}(m) \;=\; \sum_{i=1}^{L} \sum_{b} p_{b,i}\, \ell_{b,i}.$$

For a TF with multiple JASPAR PFMs, I select the motif $\hat{m}$ with maximal IC and use its PSSM ℓ.

6-mer scoring against a motif. Let $k = b_1 b_2 \cdots b_6$ be a canonical 6-mer (lexicographical minimum of the 6-mer and its reverse complement). The motif score is the best 6-bp window on either strand:

$$s_{\hat{m}}(k) \;=\; \max \left\{ \max_{1 \le i \le L-5} \sum_{j=1}^{6} \ell_{b_j,\, i+j-1}, \;\; \max_{1 \le i \le L-5} \sum_{j=1}^{6} \ell_{b'_j,\, i+j-1} \right\},$$

where $b'_1 \cdots b'_6$ is the reverse complement of k.

Empirical background and z-scores. Let $\mathcal{K}$ be the set of canonical 6-mers and let $\mathrm{BG}(k)$ be the frequency of k in the background (a set of shuffled, non-peak intervals). Then define the normalized background frequency $w(k)$ as:

$$w(k) \;=\; \frac{\mathrm{BG}(k)}{\sum_{k' \in \mathcal{K}} \mathrm{BG}(k')}\,.$$

For each TF/motif m, compute the frequency-weighted mean and second moment

$$\mu \;=\; \sum_{k \in \mathcal{K}} w(k)\, s_{\hat{m}}(k), \qquad \mu_2 \;=\; \sum_{k \in \mathcal{K}} w(k)\, s_{\hat{m}}(k)^2,$$

and the standard deviation $\sigma = \sqrt{\max(\mu_2 - \mu^2,\, \varepsilon)}$ (small ε prevents zero variance). For each k and motif m, the standardized score is

$$Z_{\hat{m}}(k) \;=\; \frac{s_{\hat{m}}(k) - \mu}{\sigma}.$$

Filtering and fan-out cap. I filter for strong (motif, 6-mer) links by removing (motif, 6-mer) pairs with motif-specific z-scores below 2.5. I also cap the number of TFs each 6-mer can match at 32:

keep $\{k, \hat{m}\}$ if $Z_{\hat{m}}(k) \geq z_{\min}$ ($z_{\min} = 2.5$), keep at most $M = 32$ motifs per k with largest Z.

(N.B. each motif $\hat{m}$ corresponds to a unique TF, since we previously chose the motif with the highest IC)

3.5 TF–gene scoring

Peak-level TF evidence. For peak p with sequence S_p, let $c_p(k)$ be the count of canonical 6-mers k in S_p. For TF t (with selected motif $\hat{m}_t$), define the peak-level score as the Z-weighted sum across the peak's 6-mers, multiplied by a distance weight:

$$S_p(t) \;=\; w_{\text{dist}}(d_{p\to g}) \sum_{k\in\mathcal{K}} c_p(k)\,\big[Z_{\hat{m}_t}(k)\big]_+, \qquad [x]_+ \equiv \max(x, 0).$$

The peak–gene distance $d_{p\to g}$ is the distance from p to the TSS of the gene g with nearest TSS (capped at 50 kb), and

$$w_{\text{dist}}(d) \;=\; \exp\Big(-\frac{d}{\tau}\Big) \;=\; 2^{-d/H}, \qquad H = 10 \text{ kb},\ \tau = \frac{H}{\ln 2}.$$

For sparsity, I retain only the top $P = 20$ TFs per peak by $S_p(t)$.

Gene-level edge weights. Let $\mathcal{P}(g)$ be the set of peaks assigned to gene g (nearest TSS within 50 kb). The TF–gene edge weight is the sum of peak contributions:

$$W(t, g) \;=\; \sum_{p\in\mathcal{P}(g)} S_p(t).$$

Edges with $W(t, g) = 0$ are dropped.

3.6 Edge normalization

Our method above for calculating weights introduces two sources of skew to the distribution of weights. First, hub genes with large numbers of incoming TFs have inflated weights across all linked TFs, and second, because we clipped z-scores at a 2.5 standard deviation threshold, edge weights have

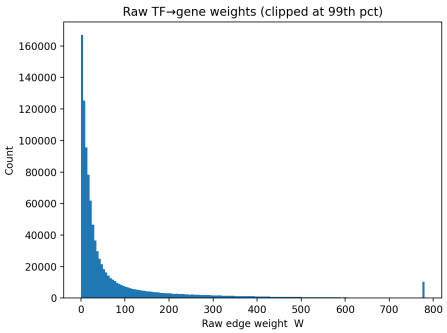

Figure 1: Raw TF–gene weights W (linear scale).

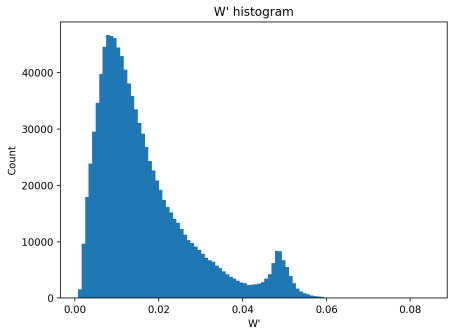

Figure 2: Normalized weights W' post log compression+row L1.

a long right tail to higher weights. To stabilize the distribution, we log compress the weights and then L1 normalize per gene so that each gene's weights sum to 1. This normalization reduces the heaviness of the tail and the bias toward hub genes:

$$W^{'}(t,g) \;=\; \frac{\log\left(1+W(t,g)\right)}{\sum_{t'\in T_g}\log\left(1+W(t',g)\right)}.$$

Where T_g is the set of all incoming TFs to gene g. Notice in figure 1 the concentration of edge values near 0, followed by a long tail of edges to very high values, with 1% of edges exceeding ~ 750 in edge weight. Figure 2 shows the distribution of edges after the log compression and gene normalization, showing a much more balanced though slightly skewed distribution.

The hump in normalized edge weights around 0.05 is suspicious but easily explained. Recall from 3.5 that we capped the number of (TF, gene) links per peak at 20. As a result, when we plot the distribution of TF in-degree to each gene (figure 3), we see a large hump of genes with exactly 20 incoming TFs. These genes' TSS were the nearest TSS to exactly 1 peak with 20 TF connections. When normalizing these genes, we expect the edge values to be about 0.05, because we enforce that each gene's incoming weight sum to 1. Therefore, we expect a hump in the distribution of edge values around 0.05, which figure 4 confirms by superimposing the histogram of $\frac{1}{in-degree}$ onto the histogram of W'.

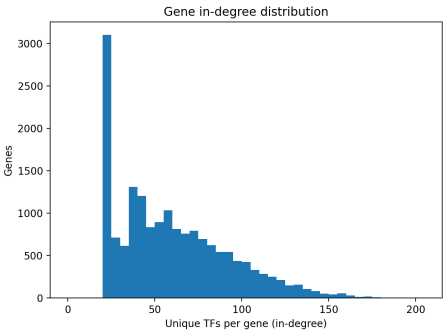

Figure 3: Distribution of gene in-degree

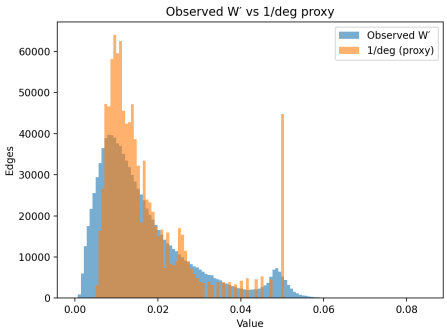

Figure 4: Inverse degree overlayed on histogram of W'

3.7 Graph construction and community detection

The TF-gene scoring and normalization process produces a bipartite graph with (TF, gene) edges. While it is possible for TFs to appear as targeted genes in this bipartite graph, for the most part the TFs are all on the same side of the graph. To cluster TFs, we need to construct a new, TF-only network where edges succinctly capture the regulatory similarity of TFs in the bipartite graph. That is, we want TFs with similar downstream gene profiles to be connected with a relatively high edge weight in this new graph. We adopt a cosine similarity approach, starting from the (TF, Gene) weights derived in the last section.

TF–gene weight matrix. Let T be the set of transcription factors and G the set of genes. We assemble the weight matrix

$$M \in \mathbb{R}^{|T|\times|G|}, \qquad M_{t,g} = W'(t,g).$$

In our case, $|T| = 584$, $|G| = 17,161$

Row L2 normalization. For each TF t, let $\mathbf{m}_t$ denote the t-th row of M. We form the L2-normalized matrix $\tilde{M}$ by

$$\tilde{M}_{t,g} \;=\; \frac{M_{t,g}}{\sqrt{\sum\limits_{g'\in G} M_{t,g'}^2 + \varepsilon}}, \qquad \text{with } \varepsilon = 10^{-12}.$$

Just as L1 normalization for genes rescaled weights to correct for hub genes, L2 normalization for TFs corrects for magnitude mismatches across TFs.

Cosine projection (TF–TF similarity). We compute the TF–TF cosine similarity matrix

$$S \;=\; \tilde{M}\,\tilde{M}^{\top}, \qquad S_{t,t'} \;=\; \sum_{g\in G} \tilde{M}_{t,g}\,\tilde{M}_{t',g},$$

and set the diagonal to zero to ignore self-similarities:

$$\mathrm{diag}(S) \leftarrow \mathbf{0}.$$

Null floor by permutation. Let $\tilde{M} \in \mathbb{R}^{|T|\times|G|}$ be row-L^2–normalized and let $\tilde{M}^{\pi}$ be the column–permuted version (same rows/columns, genes shuffled). Write the tth row as a vector $\mathbf{m}_t \in \mathbb{R}^{|G|}$ and its permuted counterpart as $\mathbf{m}_{t'}^{\pi}$ for TF t'. We form an empirical null model of cosine scores by taking inner products

$$s_{t,t'} \;=\; \langle \mathbf{m}_t,\; \mathbf{m}_{t'}^{\pi} \rangle, \qquad t \in I,\; t' \in T,\; t' \neq t,$$

and collecting them into the set $\mathcal{N} = \{s_{t,t'}\}$. We set a global floor

$$\mathrm{MIN\_SIM} \;=\; \mathrm{Quantile}_p(\mathcal{N})$$

For some $0 < p < 1$ that we will determine in the next section. MIN_SIM reflects the distribution of similarities expected under broken TF–gene alignment but preserved distribution of gene weights per TF. We will use this floor as a filter to pluck out entries from the cosine similarity matrix into our new TF-only network.

Per-node adaptive filtering and degree control. For TF t with neighbor set $\mathcal{N}(t) = \{(t', s_{t,t'})\}$ in the cosine matrix S, we compute the per-row quantile $q_t = \mathrm{Quantile}_{\alpha}\{s_{t,t'} : (t', s_{t,t'}) \in \mathcal{N}(t)\}$ for some $0 < \alpha < 1$ that we will determine in the next section. We keep edges satisfying

$$s_{t,t'} \;\geq\; \theta_t \;\equiv\; \max\big(\mathrm{MIN\_SIM},\, q_t\big),$$

then cap degree by retaining the top-K similarities per t. Optionally, we can enforce reciprocity (mutual kNN): an undirected edge $\{t, t'\}$ is kept only if

$t \in \mathrm{TopK}(t')$ and $t' \in \mathrm{TopK}(t)$. In this paper we do not enforce reciprocity, having noted similar community sizes and only slightly lower modularity scores on graphs where reciprocity is not enforced.

Rationale and tuning. The permuted-column null preserves row norms and column frequencies while destroying genuine co-targeting, giving a data-specific noise floor. The per-row quantile makes the filter flexible to differing cosine distributions in each row.'TOPK' controls graph sparsity.

3.8 Graph hyperparameter search and selection

We tune three construction parameters for the TF–TF graph: (i) a per-row adaptive quantile q that keeps, for each TF, neighbors with similarity above the q-percentile (where q is specific to the TF, not global); (ii) a global cosine floor MIN_SIM; and (iii) a degree cap K (top-K neighbors per TF). We sweep

$$K \in \{5, 10, 15, 20, 25, 30\},$$

$$\text{MIN\_SIM} \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\},$$

$$q \in \{\text{None}, 0.80, 0.90, 0.95\}.$$

To anchor MIN_SIM, we estimate a null cosine floor by permuting gene columns of $\tilde{M}$ once and taking the 99.5th percentile of the cosines between original rows and permuted rows; we include that value (0.35) and that value+0.05 in the sweep.

For each $(K, \text{MIN\_SIM}, q)$ triple we build the graph and run Louvain over $r \in \{0.75, 1.0, 1.25, 1.5\}$, keeping the best-modularity resolution for that triple. We then retain all triples whose modularity is within 0.01 of the global maximum and apply structural filters: giant-component fraction ≥ 0.40, median community size in $[5, 40]$, and average degree in $[3, 10]$. Remaining ties are broken by (in order): higher seed stability (mean AMI across random seeds), larger giant component, lower q (more permissive pruning), and higher K.

Chosen parameters and graph summary. On our dataset the chosen parameters were $K = 5$, $q = 0.95$, MIN_SIM $= 0.2$, $r = 1.25$, yielding modularity ≈ 0.891 and a stable partition (seed AMI ≈ 0.953). The resulting

graph contains $|V| = 545$ TFs and $|E| = 1033$ edges, with $n_{\text{comms}} = 57$, median community size = 5, maximum community size = 48, giant-component fraction ≈ 0.783, and average degree ≈ 3.79.

Seed stability (AMI). We report the mean adjusted mutual information across pairs of Louvain runs with different random seeds; values near 1 indicate highly reproducible community assignments, so we are confident in the reproducibility of our node partition.

3.9 Functional enrichment analysis

TF communities identified by Louvain clustering are not inherently interpretable, so I returned to the bipartite TF–gene graph and examined the downstream genes associated with each community using functional enrichment analysis. I employed two complementary approaches in sequence: over-representation analysis (ORA) and preranked gene set enrichment analysis (GSEA). ORA tests whether specific biological categories are statistically overrepresented among a community's target genes compared to a background universe, while GSEA evaluates whether predefined functional modules tend to occur near the top of a ranked gene list derived from the same community. In our case, the total edge weight entering each gene from the community being tested provides the score for preranking.

Gene sets and background. Each TF community was associated with the set of genes receiving at least one incoming edge from a TF within that community. To limit the size of enrichment queries, genes were ranked by their total incoming weight from that community, and sets were truncated adaptively: if the cumulative weight of the top genes exceeded 80% of the total, I kept only those genes (but no fewer than 150 and no more than 600). The background universe for ORA consisted of all genes represented in the bipartite TF–gene network.

Over-representation analysis (ORA). I performed ORA using `g:Profiler` with annotation sources `GO:BP`, `REAC`, `KEGG`, and `CORUM`. Communities with at least ten target genes were tested, and enriched terms with very small or very large background sizes (term size $K < 15$ or $K > 1500$) were excluded. I applied an adaptive false-discovery-rate threshold: $q \leq 0.05$ for queries smaller than 250 genes and $q \leq 0.03$ otherwise. To ensure specificity, matched

genes were required to comprise at least 5% of the annotated term for small queries and 6% for large ones.

Preranked GSEA. For communities without significant ORA results, I applied GSEA using `gseapy` on the same gene sets ranked by total incoming community weight. I used the libraries `GO_Biological_Process_2023`, `Reactome_2022`, and `KEGG_2021_Human`, with 1,000 permutations. Term size limits were relaxed to $5 \leq K \leq 5000$ to capture broader functional categories, because the communities being tested did not match any terms under our stricter ORA filters.

Outputs and reproducibility. All enrichment results were merged into a unified table (`enrichment_combined.tsv`) mapping community identifiers to enriched terms, including the database source, adjusted p-value (FDR) or normalized enrichment score (NES), term size, and overlap fraction. I used fixed random seeds and versioned annotation libraries to ensure reproducibility.

3.10 Transcription factor family annotation

To provide higher-level biological context for TF communities, I annotated each transcription factor with its canonical family and structural class. Annotations were derived primarily from the JASPAR 2024 database via its REST API, supplemented by curated pattern-based heuristics.

For each transcription factor t, including heterodimeric forms (e.g., `MAX::MYC`), I queried the JASPAR API (`https://jaspar.elixir.no/api/v1/`) for family and class metadata associated with its motif matrix. When JASPAR returned multiple hits, I selected the best match by exact name alignment. If no entry was found, a set of regex-based fallback rules was applied to infer likely families (e.g., `FOXO` → Forkhead, `GATA` → Zinc-coordinating, `NR*` → Nuclear receptor, `KLF/SP` → Zinc-coordinating). When multiple subunits or conflicting assignments occurred, a majority vote was taken across labels, with ties broken by alphabetical order.

Each record in the resulting table contains the transcription factor name, assigned family, structural class, and source provenance string (e.g., `JASPAR:1`, `Fallback`, or combined). This file was saved as `tf_families_revised.csv` and used in downstream community summaries to characterize the structural

composition of TF modules. Of the 579 TFs in the bipartite graph, I annotated 558 with matches from the JASPAR database and resorted to regex fallback for the remainder (Table 1). The regex rules are available in the appendix.

Table 1: Provenance of TF family/class labels

Source	**n TFs**	**%**
JASPAR:1	558	96.4%
Fallback	20	3.5%
Fallback+JASPAR:1	1	0.2%
Total	579	100%

3.11 TF family/class summaries for communities

I summarized the structural composition of each TF community by joining TFs to a curated family/class table (`tf_families_revised.csv`; derived in the previous section). I normalized TF identifiers (uppercase, trimmed, and with heterodimer subunits handled consistently) and mapped each TF to its family and structural class. For a given community c, I computed: (i) the dominant family and class (highest-frequency labels) with their fractions, (ii) a normalized Shannon entropy to quantify label diversity,

$$H_{\text{norm}}(c) \;=\; \frac{-\sum_i p_i \log_2 p_i}{\log_2 k},$$

where p_i are family (or class) frequencies and k is the number of distinct labels observed in c, and (iii) up to three example TFs from the dominant family for quick inspection. I output one row per community containing `n_TFs`, `dom_family`, `dom_family_frac`, `dom_class`, `dom_class_frac`, `family_entropy`, `class_entropy`, and `example_TFs`. This table (`community_family_summary.tsv`) is later merged with the enrichment results and used to identify enrichment terms that align with the community's dominant family.

4 Results

4.1 Global overview of TF communities

I constructed a TF–TF similarity graph from the bipartite TF–gene weights and detected communities via Louvain (Sections 3.7/3.8). The final graph contained **545** TFs and **1033** edges, yielding **57** communities (median size **5**, maximum **48**). The graph exhibited a modularity close to 1 (**0.89**), indicating highly segregated TF clusters with sparse connections between communities. Despite the high modularity, the graph remained well connected, with the giant component covering **78.3%** of TFs. TFs had an average degree of **3.79**. These features suggest a modular and sparsely connected network, where most TFs interact within compact neighborhoods rather than forming large diffuse hubs.

Figure 5a shows the distribution of community sizes, with about half (27) of communities having 1-5 members, 7 communities having 5-10, and a steady decrease out to 30-35. The largest community has 48 members. Figure 5b shows the distribution of degree per TF in the similarity graph, showing near uniform density out to degree 6 followed by rapid decay to 0 density by degree 15. The highest degree TF stands all the way on the right with 35 neighbors. Figure 5c plots dominant family fraction, or the percent of each TF community accounted for by a single family, as a function of community size, showing only a modest inverse correlation. From this we conclude that large communities do not represent diluted amalgams, but rather retain salient family-specific information.

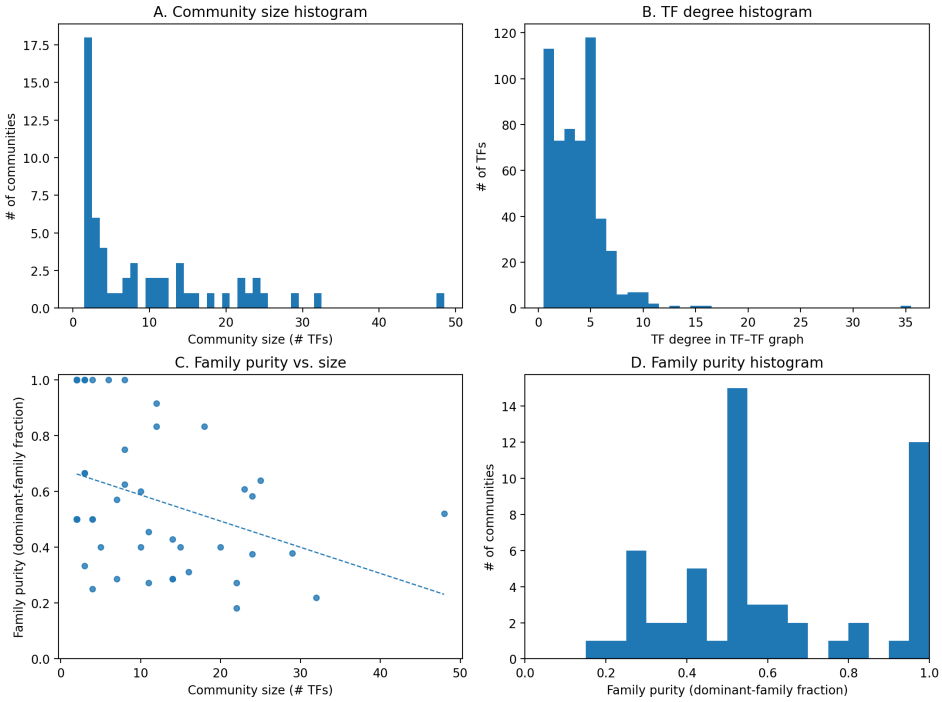

Figure 5: Global TF–TF graph overview. (A) Community size distribution. (B) TF degree distribution. (C) Family purity vs. community size. (D) Histogram of family purity.

4.2 Composition of TF families and classes

Communities were often biochemically coherent. 28% (**16/57**) were "pure", with dominant-family fraction ≥ 0.7, while 43% (**25/57**) had dominant family fraction ≥ 0.5. Family entropy had (mean **0.69** $\pm$ sd **0.38**) on a 0–1 normalized scale (0 = single-family, 1 = maximally mixed), implying moderate-to-high biochemical coherence. The most frequent dominant families were **More than 3 adjacent zinc fingers** (dominated eight communities), **ETS-related** (dominated four communities), **CREB-related factors** (dominated three communities), and **Tal-related** (dominated three communities). In total, 35 TF families dominated at least one community. Figure 6 shows the distribution of TFs in the similarity network over the different TF families; the top family alone (**More than 3 adjacent zinc fingers**) accounts for almost one fifth of all TFs, while the top eight account for half.

Taken together, the sparsity/modularity of the graph and family con-

sistency of the communities indicate that the community structure is both statistically robust and biologically interpretable, motivating a functional characterization of each community's downstream targets via enrichment analysis.

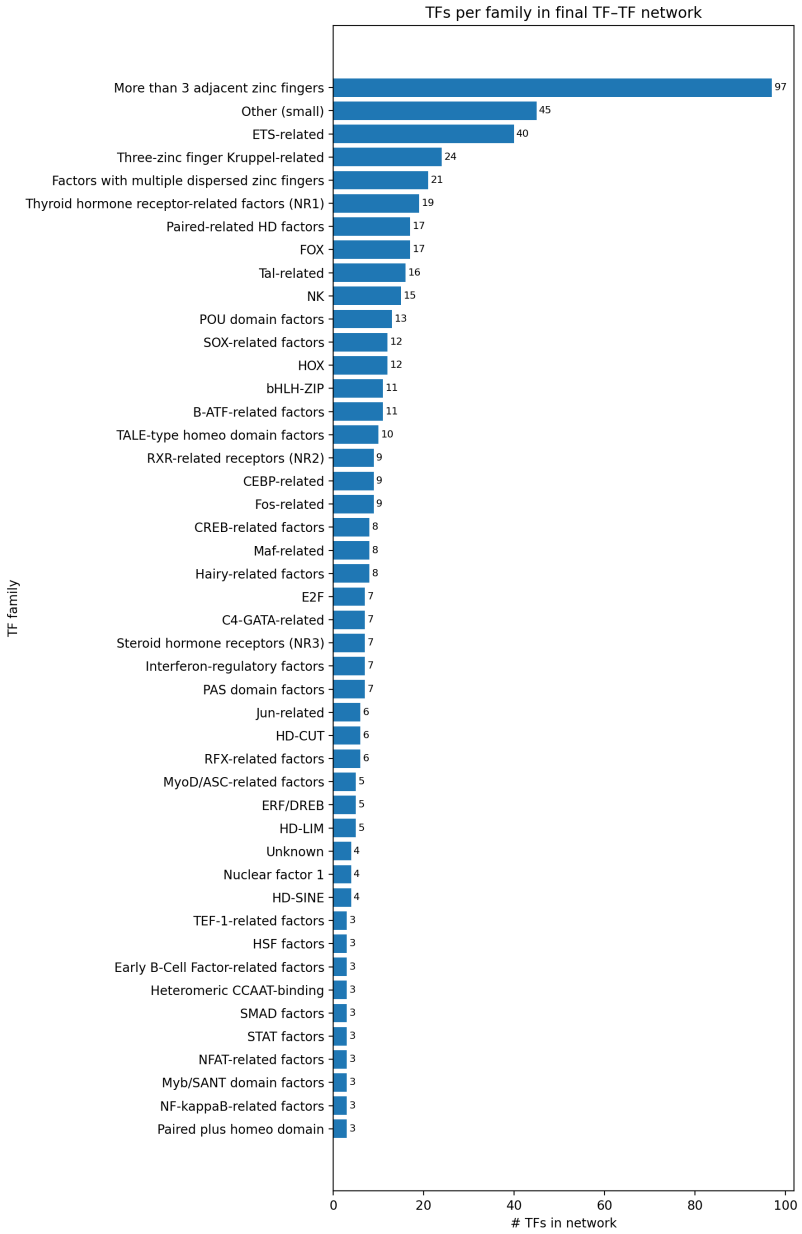

Figure 6: Counts of transcription factors per family among TFs present in the final TF–TF network. Small families (<3 TFs) are grouped as "Other (small)".

4.3 Functional enrichment overview

I assessed downstream genes per community via ORA, falling back to preranked GSEA where ORA returned no matches (Section 3.9). I report ORA terms at FDR ≤ 0.05 and GSEA terms at FDR ≤ 0.25 (the common preranked threshold), and I flag especially strong matches at FDR ≤ 0.05 regardless of method. ORA yielded at least one significant term in **13** communities and GSEA in **17**, so **30/57 (52.6%)** of communities showed at least one functional annotation. The number of significant terms per hit community was skewed (*mean* **5.5**, *median* **2**), driven largely by community 32, which contributed **76/164** (**46.3%**) total terms.

By source, most terms came from Reactome (REAC, **67%**) and GO:BP (**22.5%**), with smaller contributions from KEGG and CORUM. Among the most frequent processes were **Keratinization** (5 matches) and **Formation of the cornified envelope** (3 matches); beyond these, individual terms occurred $\leq$2 times.

To collapse near-identical terms within communities, I clustered terms using (i) token-level Jaccard similarity of normalized names (whitespace/punctuation split, stopwords removed) and (ii) Jaccard overlap of reported gene intersections when available. Pairs exceeding either threshold (name ≥ 0.80 or gene-overlap ≥ 0.70) were merged, retaining the most significant representative. This reduced the total count of terms modestly ($\sim$4%).

Figure 7 summarizes enrichment signal across communities. I display up to the top 15 terms globally (rows) against communities (columns), limiting each community to contribute at most two novel terms to prevent single communities from dominating the plot; color encodes $-\log_{10}$(FDR) for ORA and NES for GSEA, with the ORA scale clipped at the 99th percentile to stabilize contrast. Despite these caps, the matrices remain sparse, consistent with enrichment being concentrated in a subset of communities. Within those communities, the matched terms show internal coherence—for example, communities 2 and 13 display epidermal programs (**Keratinization**, **Formation of the cornified envelope**), community 14 shows immune signaling (**Rheumatoid arthritis**, **Humoral immune response**), community 32 highlights RNA processing/metabolism, and community 56 enriches for **Phagosome/Immune effector process**.

Figure 13 provides a compact "one-line-per-community" view: for each community with at least one hit (30 total), I list the dominant TF family,

the single best term by FDR, the FDR itself, and the term's source (ORA vs. GSEA). Family-level patterns are apparent: ETS-majority communities 2, 18, and 23 tend toward immune pathways (**Phagosome**, **Rheumatoid arthritis**); zinc-finger-dominated communities split between cell-cycle/RNA regulation (e.g., **Positive regulation of RNA metabolic process**) and epidermal programs (**Formation of the cornified envelope**; **Regulation of hair follicle development**). For smaller or more heterogeneous families (e.g., FOX), the leading terms are plausible (e.g., **Cellular lipid catabolic process**) but less definitive. Given that the median community yields only two significant terms, I focus deeper interpretation on a few exemplar communities with richer signal (Section 4.4).

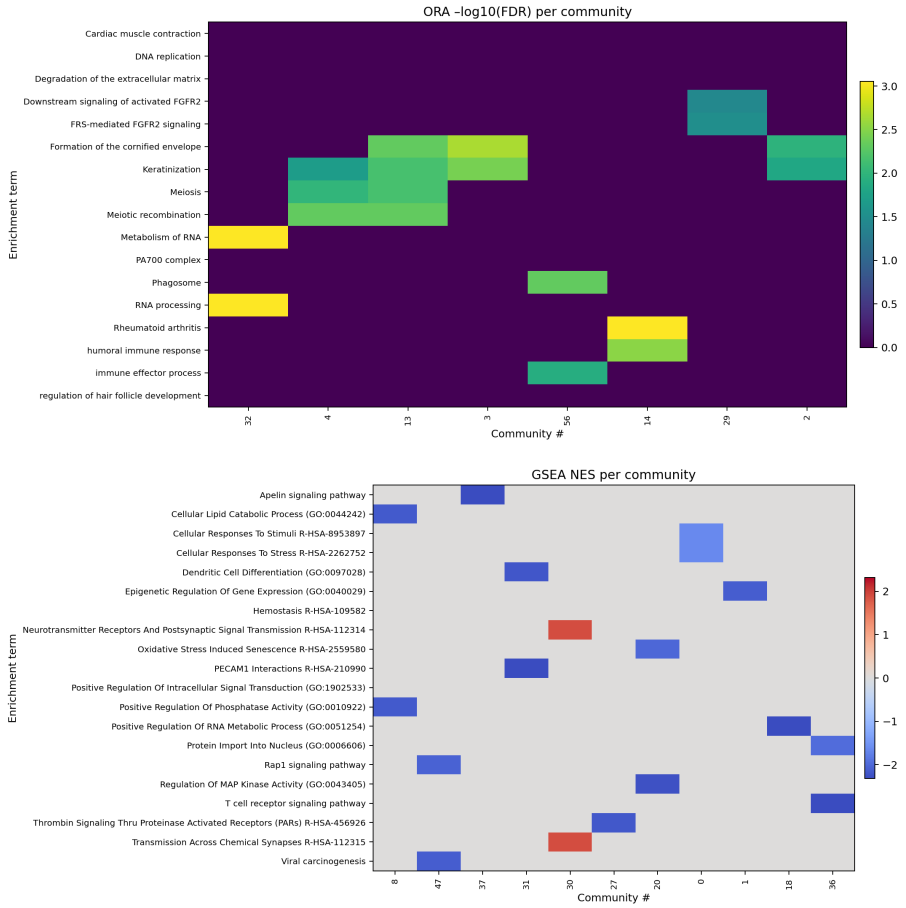

Figure 7: Enrichment summary. **Top:** ORA $-\log_{10}(\text{FDR})$ for selected terms per community (rows: terms; columns: communities; color clipped at the 99th percentile). **Bottom:** GSEA NES for selected terms per community. In both panels, each community contributes at most three terms to the visualization.

community	n_TFs	dom family	dom family frac	Top term	FDR	Source
36	6	C4-GATA-related	1.0	T cell receptor signaling pathway	6.5E-03	GSEA
27	32	CEBP-related	0.22	Thrombin Signaling Thru Proteinase Activated Receptors (PARs) R-HSA-456926	1.1E-01	GSEA
37	3	CP2-related factors	0.67	Apelin signaling pathway	1.6E-02	GSEA
4	22	CREB-related factors	0.27	Meiotic recombination	4.9E-03	ORA
30	4	E2F	1.0	Transmission Across Chemical Synapses R-HSA-112315	4.3E-03	GSEA
56	2	ETS-related	1.0	Phagosome	4.7E-03	ORA
14	23	ETS-related	0.61	Rheumatoid arthritis	8.7E-04	ORA
32	18	ETS-related	0.83	Metabolism of RNA	3.6E-12	ORA
8	24	FOX	0.58	Cellular Lipid Catabolic Process (GO:0044242)	4.7E-02	GSEA
38	2	Factors with multiple dispersed zinc fingers	0.5	Vesicle-mediated Transport R-HSA-5653656	2.2E-01	GSEA
2	11	HD-SINE	0.27	Formation of the cornified envelope	1.1E-02	ORA
39	14	HOX	0.29	Degradation of the extracellular matrix	9.6E-04	ORA
1	11	Hairy-related factors	0.46	Epigenetic Regulation Of Gene Expression (GO:0040029)	6.6E-02	GSEA
17	3	Jun-related	0.67	Colorectal cancer	2.1E-01	GSEA
40	22	More than 3 adjacent zinc fingers	0.18	Positive Regulation Of Mononuclear Cell Migration (GO:0071677)	2.4E-01	GSEA
34	8	More than 3 adjacent zinc fingers	0.75	Cardiac muscle contraction	5.5E-03	ORA
23	25	More than 3 adjacent zinc fingers	0.64	regulation of hair follicle development	1.5E-02	ORA
20	2	More than 3 adjacent zinc fingers	1.0	Regulation Of MAP Kinase Activity (GO:0043405)	3.7E-02	GSEA
18	14	More than 3 adjacent zinc fingers	0.29	Positive Regulation Of RNA Metabolic Process (GO:0051254)	0.0E+00	GSEA
3	48	More than 3 adjacent zinc fingers	0.52	Formation of the cornified envelope	2.2E-03	ORA
9	2	PAS domain factors	0.5	PA700 complex	2.1E-02	ORA
0	14	Paired-related HD factors	0.43	Cellular Responses To Stress R-HSA-2262752	4.7E-02	GSEA
13	29	RXR-related receptors (NR2)	0.38	Meiotic recombination	4.8E-03	ORA
29	5	SOX-related factors	0.4	FRS-mediated FGFR2 signaling	3.1E-02	ORA
12	15	TALE-type homeo domain factors	0.4	Regulation Of Lipid Transport (GO:0032368)	1.4E-01	GSEA
47	7	TBX2-related factors	0.29	Viral carcinogenesis	7.9E-02	GSEA
48	2	Tal-related	0.5	Positive Regulation Of Intracellular Signal Transduction (GO:1902533)	5.6E-02	GSEA
31	12	Three-zinc finger Kruppel-related	0.92	PECAM1 Interactions R-HSA-210990	8.8E-03	GSEA
24	24	Three-zinc finger Kruppel-related	0.38	DNA replication	4.0E-03	ORA
22	4	p53-related factors	0.5	Hemostasis R-HSA-109582	4.3E-02	GSEA

Figure 8: One-line summary for the 30 annotated communities: dominant TF family, top term (by FDR), FDR, and source (ORA vs. GSEA). Communities are ordered alphabetically by dominant family.

4.4 Exemplary communities

I highlight five representative modules with the most uniform family composition and coherent enrichment results. For each, I report size, dominant family/class, representative TFs, and top enrichment terms.

4.4.1 Community 3

Dominant family: More than 3 adjacent zinc fingers (52%)
Dominant class: C2H2 zinc finger factors (56%)
Family entropy: 0.719
 # TFs: 48
Gene set size: 600
Top enrichment terms (ORA/GSEA):

Community 3 — More than 3 adjacent zinc fingers (frac=0.52)

Source	DB	Term	FDR	NES	k/K	Overlap
ORA	REAC	Formation of the cornified envelope	2.18e-03		14/89	0.16
ORA	REAC	Keratinization	3.73e-03		15/112	0.13
ORA	KEGG	Neuroactive ligand-receptor interaction	6.37e-03		26/306	0.08
ORA	REAC	GPCR ligand binding	8.26e-03		30/380	0.08
ORA	GO:BP	keratinocyte differentiation	0.02		17/141	0.12

Figure 9: Enrichment terms ranked by FDR, community 3

Community 3 is the largest module (48 TFs), dominated by the family More than 3 adjacent zinc fingers (52%, entropy 0.72). Downstream genes split into two coherent clusters: epidermal differentiation (Keratinization, Formation of the cornified envelope, Keratinocyte differentiation) and ligand signaling (GPCR ligand binding, Neuroactive ligand–receptor interaction). Overlaps with the term sets are modest (k/K 0.08–0.16), but the thematic consistency and family homogeneity support a biologically coherent ZNF program.

4.4.2 Community 4

Dominant family: CREB-related factors (27%)
Dominant class: Basic leucine zipper factors (bZIP) (77%)
Family entropy: 0.905
 # TFs: 22
Gene set size: 600
Top enrichment terms (ORA/GSEA):

Community 4 — CREB-related factors (frac=0.27)

Source	DB	Term	FDR	NES	k/K	Overlap
ORA	REAC	Meiotic recombination	4.86e-03		12/73	0.16
ORA	REAC	Meiosis	9.91e-03		13/97	0.13
ORA	REAC	B-WICH complex positively regulates rRNA expression	0.01		11/77	0.14
ORA	REAC	RNA Polymerase I Promoter Escape	0.01		11/78	0.14
ORA	REAC	RNA Polymerase I Promoter Opening	0.02		8/51	0.16
ORA	REAC	Assembly of the ORC complex at the origin of replication	0.02		8/51	0.16
ORA	REAC	DNA methylation	0.02		8/53	0.15
ORA	REAC	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	0.02		8/54	0.15
ORA	REAC	HDACs deacetylate histones	0.02		10/80	0.12
ORA	REAC	Positive epigenetic regulation of rRNA expression	0.02		11/91	0.12
ORA	REAC	NoRC negatively regulates rRNA expression	0.02		11/92	0.12
ORA	REAC	HCMV Late Events	0.02		12/102	0.12
ORA	REAC	RNA Polymerase I Promoter Clearance	0.02		11/96	0.11
ORA	REAC	RNA Polymerase I Transcription	0.02		11/97	0.11
ORA	REAC	Keratinization	0.02		12/112	0.11

Figure 10: Enrichment terms by FDR, community 4 (truncated at 15 rows)

Community 4 (22 TFs) has relatively high family entropy (0.91) with

CREB-related as the largest family (27%). Despite this heterogeneity, enrichment is focused on transcriptional and chromatin programs (e.g., RNA polymerase I promoter escape/opening/transcription, HDACs deacetylate histones, positive epigenetic regulation of rRNA), consistent with bZIP-associated regulation of transcriptional machinery. Overlaps are again modest (typ. k/K 0.11–0.16).

4.4.3 Community 13

Dominant family: RXR-related receptors (NR2) (38%)
Dominant class: Nuclear receptors with C4 zinc fingers (93%)
Family entropy: 0.776
 # TFs: 29
Gene set size: 600
Top enrichment terms (ORA/GSEA):

Community 13 — RXR-related receptors (NR2) (frac=0.38)

Source	DB	Term	FDR	NES	k/K	Overlap
ORA	REAC	Meiotic recombination	4.85e-03		12/73	0.16
ORA	REAC	Formation of the cornified envelope	4.85e-03		13/89	0.15
ORA	REAC	Meiosis	6.88e-03		13/97	0.13
ORA	REAC	Keratinization	6.88e-03		14/112	0.12
ORA	REAC	Nonhomologous End-Joining (NHEJ)	0.01		10/64	0.16
ORA	REAC	G2/M DNA damage checkpoint	0.03		11/87	0.13
ORA	REAC	Reproduction	0.03		14/131	0.11
ORA	KEGG	Systemic lupus erythematosus	0.03		13/110	0.12
ORA	REAC	Processing of DNA double-strand break ends	0.03		11/90	0.12

Figure 11: Enrichment terms ranked by FDR, community 13

Community 13 (29 TFs) shows low family dominance (38% RXR-related) but high class dominance (93% Nuclear receptors with C4 zinc fingers; class entropy 0.27). Terms span epidermal processes (Keratinization, Formation of the cornified envelope) as well as genome maintenance (Nonhomologous end joining, G2/M DNA damage checkpoint, processing of DNA double-strand break ends), suggesting a nuclear-receptor–centered program targeting RNA splicing and DNA repair.

4.4.4 Community 14

Dominant family: ETS-related (61%)
Dominant class: Winged helix-turn-helix (61%)
Family entropy: 0.693

 # TFs: 23
Gene set size: 600
Top enrichment terms (ORA/GSEA):

Community 14 — ETS-related (frac=0.61)

Source	DB	Term	FDR	NES	k/K	Overlap
ORA	KEGG	Rheumatoid arthritis	8.69e-04		13/80	0.16
ORA	GO:BP	humoral immune response	2.99e-03		24/221	0.11

Figure 12: Enrichment terms ranked by FDR, community 14

Community 14 is an ETS-related module (61% dominance; entropy 0.69) with immune-focused terms (Rheumatoid arthritis, humoral immune response), consistent with ETS roles in hematopoietic/immune regulation.

4.4.5 Community 32

Dominant family: ETS-related (83%)
Dominant class: Winged helix-turn-helix (83%)
Family entropy: 0.65
 # TFs: 18
Gene set size: 600
Top enrichment terms (ORA/GSEA):

Community 32 — ETS-related (frac=0.83)

Source	DB	Term	FDR	NES	k/K	Overlap
ORA	REAC	Metabolism of RNA	3.57e-12		65/652	0.10
ORA	GO:BP	RNA processing	5.47e-07		68/897	0.08
ORA	KEGG	Herpes simplex virus 1 infection	7.92e-07		41/437	0.09
ORA	GO:BP	ribonucleoprotein complex biogenesis	7.95e-07		42/430	0.10
ORA	REAC	Translation	1.60e-06		31/271	0.11
ORA	REAC	RNA Polymerase II Transcription	7.42e-06		78/1218	0.06
ORA	REAC	Gene expression (Transcription)	7.42e-06		85/1374	0.06
ORA	REAC	Processing of Capped Intron-Containing Pre-mRNA	3.42e-05		27/255	0.11
ORA	KEGG	Ribosome	8.67e-05		17/122	0.14
ORA	REAC	Generic Transcription Pathway	9.16e-05		69/1107	0.06
ORA	GO:BP	mitochondrial gene expression	1.35e-04		20/159	0.13
ORA	GO:BP	mRNA processing	1.70e-04		37/450	0.08
ORA	REAC	rRNA processing	1.75e-04		21/184	0.11
ORA	REAC	Major pathway of rRNA processing in the nucleolus and cytosol	3.98e-04		19/164	0.12
ORA	REAC	Cap-dependent Translation Initiation	4.88e-04		15/111	0.14

Figure 13: Enrichment terms by FDR, community 32 (truncated at 15 rows)

Community 32 (18 TFs) is the most enrichment-dense module, dominated

by ETS-related (83%; entropy 0.65). Enrichment terms gravitate toward RNA metabolism (Metabolism of RNA, RNA polymerase II transcription, processing of capped intron-containing pre-mRNA, rRNA processing). Despite modest per-term overlap (k/K up to 0.14), the repeated, high-confidence hits within RNA processing/transcription strongly suggest a core ETS regulatory subnetwork.

Supplementary: Community atlas and cards (summary)

A per-community atlas summarizing family/class composition and top enrichments is provided in `report_bundle/community_atlas.tsv`, available at the GitHub link in the appendix.

5 Limitations and Improvements

5.1 Design choices and hyperparameters

- **K-mer length and motif mapping.** I fixed $k = 6$, chose a single PFM per TF by IC, capped TF matches per 6-mer (32) and per peak (20), and thresholded 6-mer→TF pairs at $z \geq 2.5$. These choices bias sensitivity/specificity.
  - *Sensitivity tests:* we can sweep across $k \in 5, 6, 7, 8$, caps $\in 16, 32, 64$, and $z \in 2.0, 2.5, 3.0$. then quantify changes in TF–gene edge by Jaccard, degree distributions, and downstream modularity.
  - *Consensus scoring:* We can average edges across settings (or take intersection) and re-cluster, then report AMI of communities versus the baseline derived in this paper.
- **Background construction.** I used length- and chromosome-matched non-peak background; this controls gross composition but not local sequence quirks.
  - *Alternatives:* We can take GC-matched shuffles (dinucleotide-preserving), local flanks as background, or construct per-chromosome background models. Then we can compare k-mer log-enrichments and see if family composition shifts.

5.2 Graph construction and selection bias

- **Tendency toward sparsity.** Our grid search for thresholds (MIN_SIM, TOPK, ADAPTIVE_Q) to prune the TF-TF cosine matrix, with graph modularity as the objective function, favors sparser graphs that inflate modularity.
  - *Stability analysis:* sweep (TOPK, MIN_SIM, ADAPTIVE_Q) on a lattice; compute AMI/NMI and edge set Jaccard across partitions. We can plot a "robustness map" (modularity vs. stability vs. density) and prefer solutions on a Pareto frontier rather than single best modularity.
  - *Consensus communities:* build a co-assignment matrix over sweeps and cluster it; report consensus scores per community.
  - *Alternatives:* operate directly on the bipartite graph (bipartite modularity / spectral co-clustering) and compare to cosine-projected results.

5.3 Motifs and scoring

- **PFM/PSSM uncertainty.** One motif per TF ignores paralog variability and condition-specific motifs, and heterodimers are simplified.
  - *Alternatives:* retain multiple PFMs per TF and aggregate; compare with direct PWM scanning (FIMO/MOODS) on peaks; evaluate concordance with our 6-mer proxy.
- **Z-scoring with global background.** Our per-motif background uses pooled k-mer frequencies; a motif's "expected" score may vary across chromatin contexts.
  - *Alternatives:* stratify background by chromosome, GC composition, or peak length bins and recompute μ, σ; check whether community composition/enrichment changes.

5.4 Peak–gene linking

- **Distance-only linkage.** Nearest-TSS within 50 kb is a simplification, ignoring distal enhancers and promoter–enhancer wiring.

- *Alternatives:* integrate ABC scores, promoter-capture Hi-C, eQTLs, or enhancer–gene maps to re-score TF–gene edges, then compare communities/enrichments.

5.5 Enrichment analysis

- **Generic terms.** ORA/GSEA often return broad processes.
  - *Data-driven ontology:* build a gene–gene similarity network from shared TF input (identically to how we constructed the TF-TF network), detect gene communities, and use them as custom gene sets. Guard against circularity by holding out the community's own edges when scoring its enrichment (or use cross-validation across chromosomes).
  - *Specificity filters:* keep terms with high overlap fraction within size bounds, and prioritize terms whose keywords align with the community's dominant TF family. This approach faces a structural challenge from the low overlap fractions observed across communities.

5.6 Cell/context dependence

- **Single cell line (K562).** Communities are context-specific.
  - *Multi-cell benchmarking:* replicate pipeline on other ENCODE cell types; align TF communities across cell types and report conserved vs. cell-specific modules.

5.7 External validation and calibration

- **Comparative ground truth.** Compare edges/communities against TRRUST, DoRothEA/ChEA, and literature regulons.
- **Multiple testing / duplication.** Our intra-community deduplication handles near-identical terms within commuinities, resulting in a small reduction in the number of terms (4%). A global FDR across all communities/terms may collapse more terms.

6 Conclusion

In this paper, I designed and implemented an end-to-end pipeline for parsing ATAC seq peaks to infer transcriptional regulatory structure in a K562 cell. Starting from peak calls, I constructed a bipartite graph linking transcription factors (TFs) to genes based on motif occurrence near gene TSS. Using cosine similarity on edge weights, I projected this bipartite graph into a TF–TF similarity network whose edges capture co-binding patterns inferred from shared target genes. Applying Louvain community detection revealed cohesive TF modules with moderate family entropy, indicating that DNA-binding domain families are partly recapitulated in their inferred regulatory connectivity.

Functional enrichment of genes downstream of each TF community showed overall sparsity—only about half of the 57 modules yielded significant terms—but several communities displayed clear thematic coherence, aligning with known biological functions of their dominant families (e.g., zinc-finger clusters in keratinization, ETS modules in immune and RNA-processing programs).

I hope to extend this modeling framework by systematically investigating hyperparameter sensitivity (motif background, peak–gene distance thresholds, community resolution) and integrating multiple cell types to test cross-context reproducibility. My ultimate goal is to quantify how known TF family structure influences, or is complicated by, modular gene-regulatory organization derived using fully unsupervised approaches.

A Appendix

A.1 Input and Provenance

All input data and intermediate results are available in the project's companion GitHub repo available at `https://github.com/gregorme2001/Transcription-Factor-Clustering`. Table 2 lists the key files.

Table 2: Summary of input and intermediate files.

File	Type	Description
encode_atac_provenance.tsv	Metadata	ENCODE ATAC-seq provenance and accession info.
gencode_tss.bed	Data	Gene names, chromosome, and start/end coordinates from GENCODE.
JASPAR2024_CORE_vertebrates.pfm.txt	Data	TF PFMs from JASPAR
bipartite_edges.tsv	Data	TF→Gene bipartite edges with weights.
tf_families_revised.tsv	Data	JASPAR family/class labels for TFs.
tf_graph.edgelist	Data	Projected TF–TF similarity network.
tf_partition.tsv	Data	Community assignments for each TF.
df_com_gene.tsv	Data	Gene sets downstream of each TF community.
communities.tsv	Results	Louvain community assignments.
community_family_summary.tsv	Results	Family composition summary per community.
community_atlas.tsv	Results	Per-community enrichment summary.
top_terms_compact.tsv	Results	Compact list of top enrichment terms.
enrichment_combined.tsv	Results	Enrichment results combined for GSEA + ORA.

A.2 Regex filter for olfactory/taste receptor symbols

To reduce low-specificity signal from very large chemosensory receptor families, I filtered gene symbols through a case-insensitive, fully anchored regular expression that matches canonical olfactory (`OR`) and taste receptor (`TAS1R`, `TAS2R`) names.

```
OLFA_TASTE_RX = re.compile(
    r"""^(
            OR\d+[A-Z]*\d*      # OR1, OR2A1, OR10G7, OR52E4, ...
          | OR[A-Z]{1,2}\d+     # ORA1, ORAA2  (rare formats)
          | TAS1R[123]          # TAS1R1, TAS1R2, TAS1R3
          | TAS2R\d+            # TAS2R38, TAS2R13, ...
        )$""",
    re.IGNORECASE | re.VERBOSE
)
```

A.3 Fallback symbols for TF families/classes

The following table shows my guess of family/class assignments when JASPAR failed to return a match.

Regex pattern	Examples	Family	Structural class
$^(ETS\|ELF\|ELK\|ERG\|ETV\|FLI)$	ETS1, ELF1, ERG	ETS	Winged helix–turn–helix
$^(FOXO?\|FOX[A-Z0-9]+)$	FOXO3, FOXA1	FOX	Forkhead
$^(HOX\|HOXA\|HOXB\|HOXC\|HOXD)$	HOXA1, HOXD13	HOX	Homeobox
$^(POU\|OCT\|POU[0-9A-Z]+)$	POU5F1/OCT4	POU	Homeobox
$^(GATA)$	GATA1, GATA3	GATA	Zinc-coordinating
$^(KLF\|SP[1-9])$	KLF4, SP1	KLF/SP	Zinc-coordinating
$^(NR[0-9A-Z]\|RARA\|RARB\|RARG\|RXR\|PPAR\|ESR\|THR\|VDR)$	ESR1, NR2F2, PPARA	Nuclear Receptor	Nuclear receptor
$^(IRF)$	IRF1, IRF8	IRF	Winged helix–turn–helix
$^(SMAD)$	SMAD2, SMAD4	SMAD	MH1/SMAD
$^(MAF\|ATF\|CREB\|JUN\|FOS\|CEBP)$	JUN, FOSL1, ATF3	bZIP	Leucine zipper
$^(MYC\|MAX\|MLX\|USF)$	MYC, MAX, USF1	bHLH	Helix–loop–helix
$^(TFEB\|TFEC\|MITF\|TFE3)$	MITF, TFE3	MiT/TFE	bHLH-Zip
$^(SOX)$	SOX2, SOX9	SOX	HMG box
$^(RFX)$	RFX5	RFX	Winged helix
$^(E2F)$	E2F1, E2F7	E2F	Winged helix
$^(NFY)$	NFYA, NFYB	NFY	CCAAT-binding
$^(CTCF\|CTCFL)$	CTCF, CTCFL	CTCF	Zinc-coordinating
$^(ZNF\|ZBTB)$	ZNF143, ZBTB7A	C2H2 ZNF	Zinc-coordinating

Table 3: Regex fallbacks used when JASPAR metadata are unavailable.

References

Blondel, Vincent D. et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

Buenrostro, Jason D. et al. (2013). "Transposition of native chromatin for fast and sensitive profiling of open chromatin". In: *Nature Methods* 10.12, pp. 1213–1218. DOI: 10.1038/nmeth.2688.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11, pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

Consortium, GENCODE (2025). *GENCODE Human Release v43*. https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_43/gencode.v43.annotation.gtf.gz. GTF release used in this study. Accessed 2025-10-15.

Dale, Ryan K., Brent S. Pedersen, and Aaron R. Quinlan (2011). "Pybedtools: a flexible Python library for manipulating genomic datasets and annotations". In: *Bioinformatics* 27.24, pp. 3423–3424. DOI: 10.1093/bioinformatics/btr539.

Davis, Carrie A. et al. (2018). "The Encyclopedia of DNA Elements (ENCODE): data portal update". In: *Nucleic Acids Research* 46.D1, pp. D794–D801. DOI: 10.1093/nar/gkx1081.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, et al. (2019). "GENCODE reference annotation for the human and mouse genomes". In: *Nucleic Acids Research* 47.D1, pp. D766–D773. DOI: 10.1093/nar/gky955.

Li, Heng, Bob Handsaker, Alec Wysoker, et al. (2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li, Qunhua et al. (2011). "Measuring reproducibility of high-throughput experiments". In: *The Annals of Applied Statistics* 5.3, pp. 1752–1779. DOI: 10.1214/11-AOAS466.

Marçais, Guillaume and Carl Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6, pp. 764–770. DOI: 10.1093/bioinformatics/btr011.

Quinlan, Aaron R. and Ira M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, et al. (2019). "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". In: *Nucleic Acids Research* 47.W1, W191–W198. DOI: 10.1093/nar/gkz369.

Rauluseviciute, Ieva et al. (2024). "JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 52.D1, pp. D174–D182. DOI: 10.1093/nar/gkad1059.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *PNAS* 102.43, pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

Zhang, Yong et al. (2008). "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9, R137. DOI: 10.1186/gb-2008-9-9-r137.