# Process gcPBM: End‑to‑End Tutorial & Logic

This tutorial walks through **what the script does and why each step matters**, so you can explain the reasoning behind the pipeline clearly.

## 0) Purpose & Outputs

The script converts raw gcPBM binding data into **balanced, labeled datasets** for machine learning.

* **Why:** Raw experimental data are noisy, unbalanced, and contain confounding motifs. We want clean inputs for fair training/testing.
* **Outputs:**
  + dataset.csv → balanced sample of probes for experiments.
  + exp\_data\_all.csv → complete labeled dataset for ML.

## 1) Loading Data

We load three key inputs:

1. **Universal 8‑mer E‑scores** → quantitative binding preferences for every possible 8‑base sequence.
2. **Probe sequences** → 36‑bp DNA probes used in gcPBM arrays.
3. **Normalized intensities** → measured fluorescence signal (binding strength) from the Myc/Max experiment.

Go to <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2579597> and then click on "Platform ID" (you get the table with sequences:  [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL17173](https://urldefense.com/v3/__https:/www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL17173__;!!Mih3wA!GwpLew34CdzWU5fZ4M1e2DC6HL3Jc9C2flN1u27tw3qMe_phMMY6yTw5ke7V1ZCMaO3gESZ5Me6qnzCxeg$" \t "_blank))

**Why:** These files together let us connect DNA sequence (probe) with experimental binding signal and Enrichment scores (E‑scores).

## 2) Annotating Probes with Motifs

Each 36‑bp probe contains a central 8‑mer candidate binding site and flanking 8‑mers.

* Extract central 8‑mer and neighboring 6‑, 10‑, and 12‑mers.
* Map their **E‑scores** for quantitative comparison.

**Why:**

* Central 8‑mer should drive binding.
* Flanks and other overlapping motifs can interfere.
* Having multiple k‑mers allows later filtering and ensures motif diversity when sampling.

## 3) Applying Biological Filters

1. **Remove strong flanks:** If either flank has an E‑score ≥ 0.3, discard the probe.
   * *Why:* To avoid probes where binding could be driven by side motifs.
2. **Central dominance check:** Keep only probes where the central 8‑mer’s E‑score is highest compared to immediate neighbors.
   * *Why:* Ensures we measure binding attributable to the intended motif, not a stronger overlapping one.

## 4) Transforming Intensities

* Compute **log‑transformed intensities.**
  + *Why:* Binding intensities are skewed; log scaling normalizes the distribution and highlights differences between weak and strong binders.

## 5) Labeling Binding Strength

Classify probes into categories:

* **Unbound:** log intensity below threshold.
* **Weak:** log intensity between thresholds.
* **Strong:** log intensity above higher threshold.

**Why:** Different ML tasks require different labels:

* Regression → use continuous log intensity.
* Binary classification → binder vs. non‑binder.
* Multiclass classification → unbound/weak/strong.

## 6) Visualizing the Data

Histograms of log intensity with thresholds drawn:

* Green = unbound, Orange = weak, Red = strong.

**Why:** Visualization validates threshold choice and shows class balance.

## 7) Stratified Sampling with Motif Uniqueness

We sample balanced subsets of probes for ML training:

* Bin intensities into 0.1‑wide bins for even coverage.
* Select a fixed number of probes per class.
* Ensure no repeated motifs using tie‑breaker k‑mers (8‑, 10‑, 12‑, 6‑mers).

**Why:**

* Prevents class imbalance (e.g., too many weak binders).
* Maintains motif diversity, so ML doesn’t just memorize a few recurring sequences.
* Produces reproducible, fair datasets.

## 8) Dataset Augmentation

After the initial sample (e.g., 33 per class), the script expands to larger numbers (e.g., 78 unbound, 45 weak, 45 strong).

**Why:** More data improves generalization while still keeping the dataset balanced.

## 9) Final Exports

Two main exports:

1. **exp\_data\_all.csv** → all probes with sequences, log intensities, and labels.
2. **dataset.csv** → balanced sampled subset for modeling.

**Why:**

* Full dataset is useful for exploratory analysis or regression.
* Balanced dataset is optimized for classification training/testing.

## 10) Sanity Checks & Reproducibility

The script performs checks like:

* Probe length = 36 bp.
* No missing values.
* Correct class counts.
* Uniqueness of motifs.

Random seeds are fixed for reproducibility.

**Why:** Ensures results are valid and repeatable across runs.

## TLDR;

* **Load raw data** → combine probe sequences, intensities, and E‑scores.
* **Filter biologically** → ensure clean attribution of binding.
* **Normalize & label** → log transform and categorize.
* **Visualize** → confirm distributions and thresholds.
* **Sample** → balanced, stratified, and motif‑unique subsets.
* **Export** → save full labeled and balanced datasets.

This creates a robust pipeline from raw gcPBM output to ML‑ready inputs.

We convert raw gcPBM measurements into trustworthy labels tied to the center motif, then craft a balanced, motif‑diverse sample so downstream ML learns general sequence–affinity relationships rather than memorizing repeated motifs or artifacts.