Revolutionizing Transcription Factor Prediction: A New CTCF PWM Pipeline

Group1
113971012 林穎彥
113971008 張育瑋
113971017 邱世淦
112971026 蔣政寬

Outline

Our Goal

- What is CTCF?
- Why is CTCF Important?
- How Does CTCF Recognize DNA?
- Why This Matters for PWM Construction ?

Our Work

- Project Architecture Overview
- Phase 0 – 5
- Validation Framework Architecture

Our Insight

- Takeaway 1- 6
- Future Extensions & Applications

Our Goal

- What is CTCF?
- Why is CTCF Important?
- How Does CTCF Recognize DNA?
- Why This Matters for PWM Construction ?

What is CTCF?

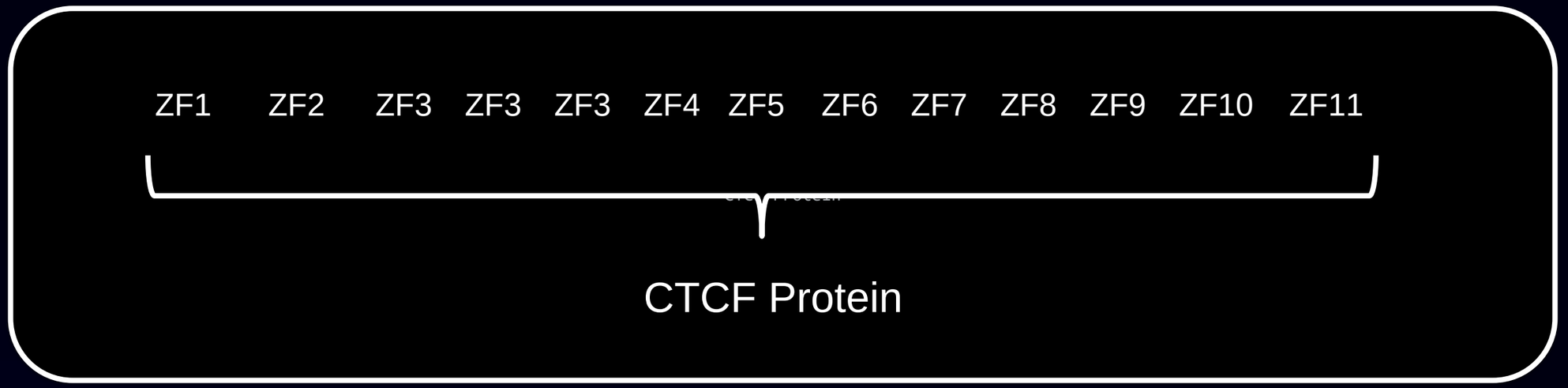
CTCF: The Master Genome Organizer

- CTCF (CCCTC-Binding Factor) is a critical transcription factor that acts as the primary architectural protein organizing mammalian genomes into functional 3D structures.

Why is CTCF Important?

- 💊 **Drug discovery**: Target CTCF binding for cancer therapy
- 💉 **Disease research**: Understand genetic disorder mechanisms
- 🧬 **Genome engineering**: Design precise genome editing tools
- 🧠 **Basic research**: Understand genome organization principles

How Does CTCF Recognize DNA?



- **CTCF protein contains 11 zinc finger domains (ZF1 to ZF11).**
- **Each zinc finger can interact with a short segment of DNA to recognize specific sequence motifs.**

Some DNA Positions Matter More Than Others

dsDNA sequence

- **Arrows (↑) show zinc finger binding sites**
- **numbers indicate information content**
- **higher means more conserved**

More technical details are available on GitHub.

Why Does This Binding Information Matter?

ZF1	ZF2	ZF3	ZF4	ZF5	ZF6	ZF7	ZF8	ZF9	ZF10	ZF11……
(1.8) bits	(1.6)	(1.9)	(1.7)	(0.8)	(0.4)	(0.9)	(1.5)	(1.4)	(0.6)	(1.2)

- Helps us design better DNA motifs for prediction
- Explains why some mutations cause binding loss
- Guides bioengineers to build better models (PWM etc.)

More technical details are available on GitHub.

Why Information Content Matters for PWM Construction ?

❗ Alignment Quality Affects Information Content

- **If DNA sequences are misaligned (e.g., shifted by one position), even identical sequences can lead to lower information content (IC).**
- **Well-aligned sequences show consistent bases at key positions → high IC (e.g., C = 100%).**
- **Poorly aligned sequences result in mixed base frequencies → low IC (e.g., A = 33%, C = 33%, T = 33%).**

👉 Poor alignment quality leads to noisy motifs and reduces PWM prediction accuracy.

More technical details are available on GitHub.

Why Information Content Matters for PWM Construction ?

❗ IC Values Have Biological Meaning

IC Range	Biological Interpretation
>1.5	Essential contact; critical for binding
1.0-1.5	Stable contact; somewhat flexible
0.5-1.0	Weak interaction or spacer
0.5<	No strong preference; likely non-specific site

More technical details are available on GitHub.

Why Information Content Matters for PWM Construction ?

❗ Biomedical Applications and Accuracy of PWM Prediction

- **High-quality PWM (total IC > 12 bits)**
  - Accurately predicts strong binding sites
  - Low false positive rate, suitable for drug or functional prediction
- **Low-quality PWM (IC < 5 bits)**
  - Prone to false predictions
  - Fails to identify functional regions; high risk in drug design

👉 Our pipeline can detect biologically reliable binding sites

More technical details are available on GitHub.

Why Information Content Matters for PWM Construction ?

! Mathematical Calculation of Information Content

Using the Shannon formula:

$$IC = \Sigma\ p(x) \times \log_2(p(x) / 0.25)$$

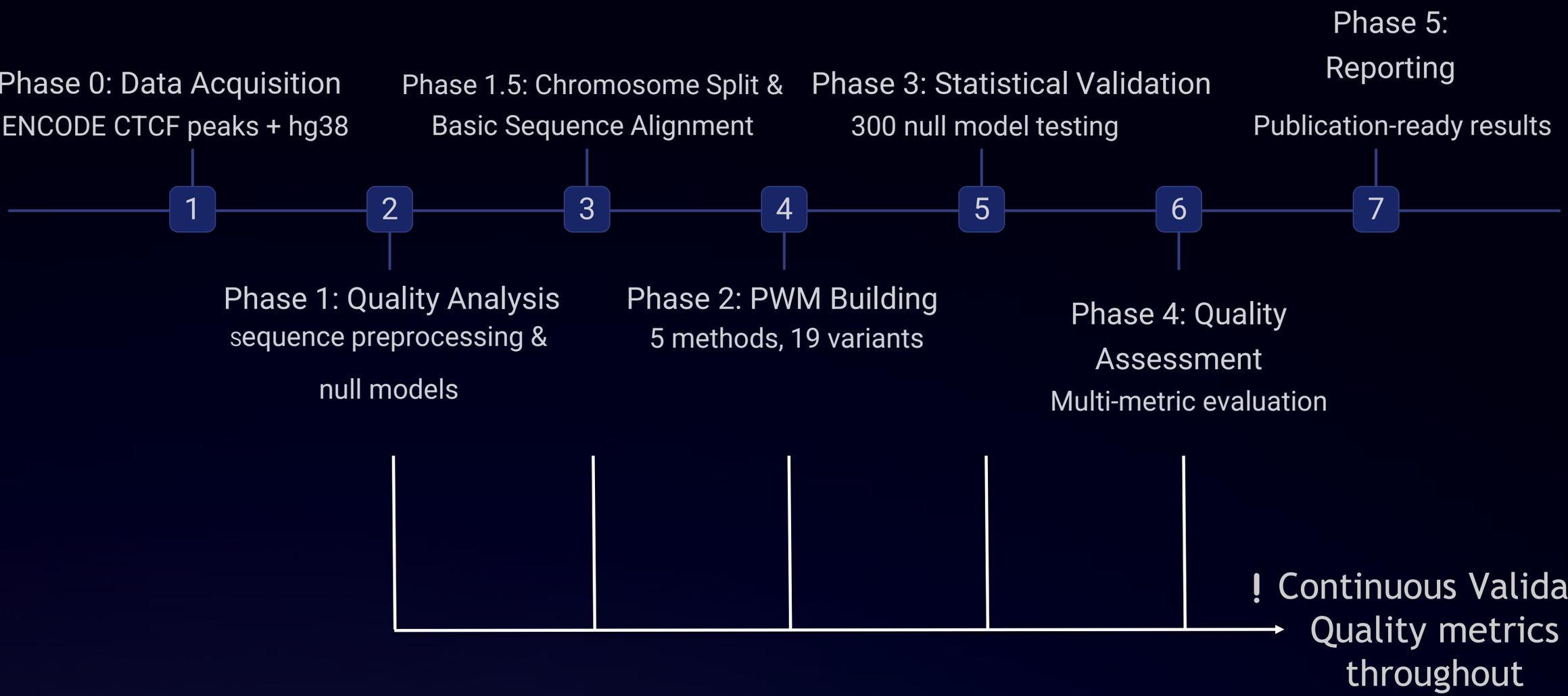
- **0.25** = Probability of each DNA base under a random background
- **$\log_2$** = Converts to units of “bits”
- **Maximum IC** is **2.0 bits** (perfect conservat

More technical details are available on GitHub.

Our Work

- Project Architecture Overview
- Phase 0 : Data Aquisition
- Phase 1 : Quality Analysis
- Phase 1.5 : Chromosome Split & Basic Sequence Alignment
- Phase 2 : PWM Building
- Phase 3 : Null Model Testing
- Phase 4 : Quality Assessment
- Phase 5 : Reporting
- Validation Framework Architecture

Project Architecture Overview



Phase 0：Data Source

Dataset Type	File Location	Size	Description	Source
CTCF Peaks	data/K562_CTCF_peaks.bed	~2.7 MB	CTCF binding sites for K562 cell line (BED format)	ENCODE
Reference Genome	data/reference_genome/hg38.fa *(full)*	~3.1 GB	Human genome reference (hg38 full version)	UCSC
	data/reference_genome/hg38.chr21.fa *(demo)*	~46 MB	Chromosome 21 only (demo mode)	UCSC

Phase 0：Data Source_ Flow Chain

- **Stage 1: Data Download & Extraction (download_data.sh)**

data	File	Details / Size
CTCF peaks	K562_CTCF_peaks.bed	~2.7 MB
Reference genome	hg38.fa / hg38.chr21.fa	~3.1 GB / ~46 MB
Extract sequences	extracted_sequences.fasta	~8.8 MB, ~44,217 sequences

🔍 Validation Checkpoint 1.1 ：File existence, format, and size verified

Phase 1：Quality Analysis_Quality-Over-Quantity

37,628 Raw Sequences

0.695 bits

Quality Filters

N ratio, length, complexity

1,000 High-Quality

19.592 bits

Breakthrough Results:

🥇 **1,000 filtered**	1K	**19.592 bits**	🏆 **Excellent**	**Baseline**
🥈 **2,000 filtered**	2K	**12.564 bits**	☑️ **Good**	0.64×
🥉 **5,000 filtered**	5K	**10.659 bits**	⚠️ **Fair**	0.54×
❌ **All raw data**	37.6K	**0.695 bits**	❌ **Very Poor**	**0.035×**

🔑 **Scientific Impact**: Established that **dataset quality trumps dataset size** in transcription factor modeling

Phase 1：Quality Analysis_Challange

🟤 Challenge 1: Spatial Variability

- **Motif (19 bp) can appear anywhere within a 200 bp ChIP-seq peak.**
- **Model must scan entire window to find it.**

👉 Hard to align, confuses traditional ML.

Phase 1：Quality Analysis_Challange

🟤 Challenge 2: Data Quality Issues

- **91.8% low-complexity sequences → repetitive or noisy**
- **N-base contamination → signal loss**
- **Variable lengths (220 types) → hard to align**
- **Spatial autocorrelation → data leakage risk**

👉 Impacts model reliability and biological signal integrity

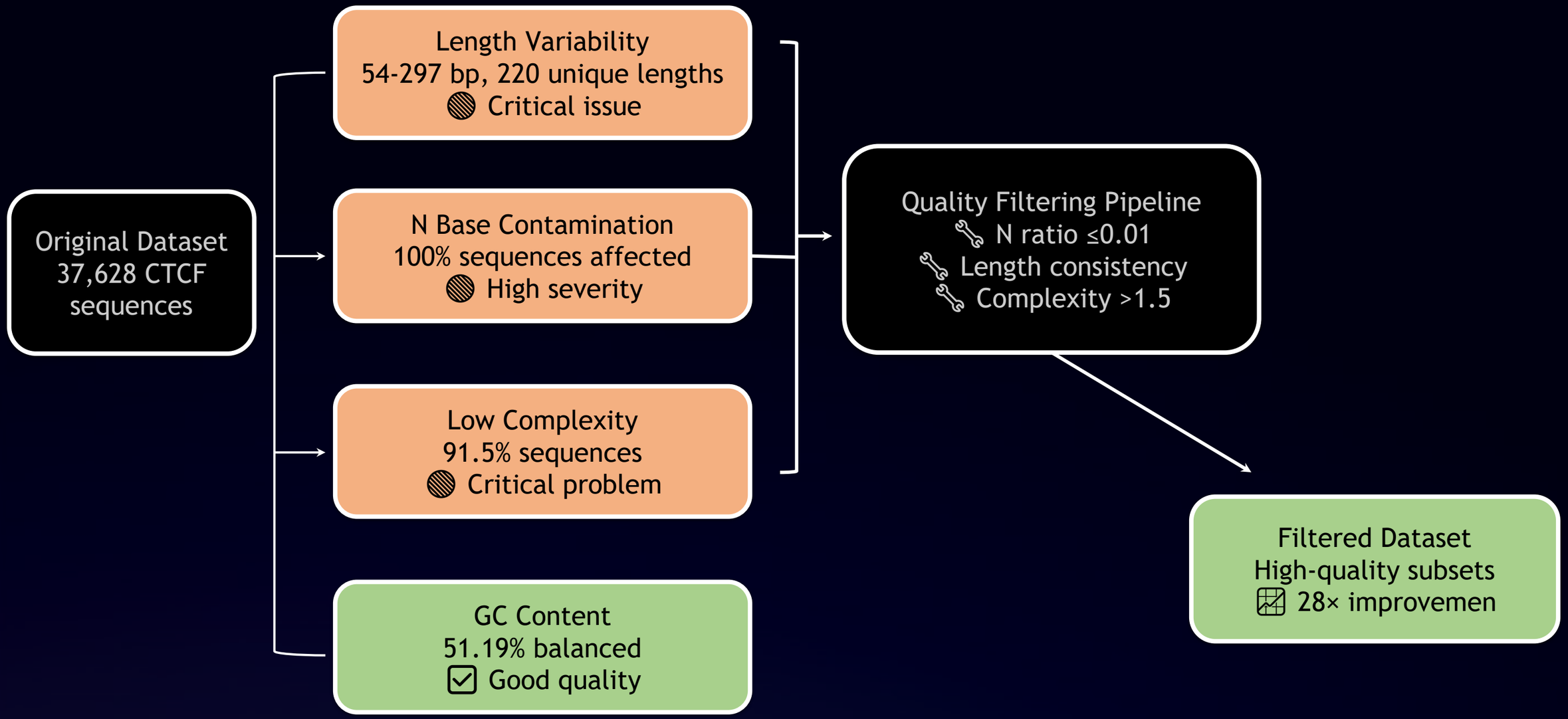
Phase 1：Quality Analysis_Challange

Challenge 3: Information Content Requirements

IC per position	Meaning
2 bits	Perfect specificity (only 1 nucleotide allowed)
1 bit	Strong preference (2× over random)
0.5 bit	Weak preference
0.5 bit	No preference (random)

- **CTCF Quality Criteria**
  - Total IC 8–15 bits = strong PWM
  - ≥4 positions >1 bit = clear motif
  - Detectable pattern: `CCGCGNGGNGGCAG`

Phase 1：Quality Analysis_Critical Quality Issues in Original Data



Phase 1：Quality Analysis_Critical Quality Issues in Original Data

Issue	Severity	Impact	Solution
Length Variability	Critical	Poor alignment, low IC	Length filtering
N Base Contamination	High	Reduced information	N-ratio threshold
Low Complexity	Critical	Minimal motif signal	Entropy filtering

🔑 **Key Insight**: Data preprocessing more critical than alignment methodology

Phase 1： Quality Analysis _ Flow Chain

- **Stage 2: Sequence Preprocessing (preprocess_sequences.R)**

Step	File / Action	Validation Checkpoint
Input	data/extracted_sequences.fasta	❗ **Checkpoint 1.2**: Raw sequence quality analysis
Output	data/preprocessed_sequences_optimized.fasta	❗ **Checkpoint 1.3**: Post-preprocessing quality confirmation

More technical details are available on GitHub.

Phase 1.5：Chromosome-Based Splitting

🔘 Why Random Splitting Fails for CTCF Prediction

😩 Genomic Data Leakage

- **Random split may place nearby CTCF sites in both train & test sets**
- **Model sees similar motifs during training → artificial performance boost**

😩 Poor Generalization

- **Random splits fail to test cross-chromosomal robustness**
- **Models may exploit genomic proximity, not motif logic**

⚠ Random splits overestimate model performance and mask biological signal quality

(Tahir et al., 2025)

Phase 1.5：Chromosome-Based Splitting

What is Chromosome-Based Splitting ?

A data splitting strategy where one or more entire chromosomes are held out as the test set

- 👍 Ensures no overlap between training and test genomic locations
- 👍 Accurately tests generalization to unseen regions/cell types
- 👍 Forces model to learn biological rules, not location biases

tests **generalization to unseen regions/cell types**

(Tahir et al., 2025)

Phase 1.5：Chromosome-Based Splitting

Training Set	Testing Set
19 chromosomes37,628 sequences	4 chromosomes13,166 sequences

Split Statistics:

Training Chromosomes (19): chr1, chr2, chr3, ..., chr19, chrX

Testing Chromosomes (4): chr11, chr17, chr20, chr22

Data Leakage: ☑ **Zero percent overlap** confirmed

Split Ratio: 80.4% training / 19.6% testing

🔑 Breakthrough: Removes spatial bias for fair model evaluation

Phase 1.5： Chromosome-Based Splitting _ Flow Chain

- **Stage 3: Dataset Preparation (prepare_datasets.R)**

Step	File / Action	Validation Checkpoint
Input	data/preprocessed_sequences_optimized.fasta	
Output	data/training_sequences.fasta	❗ Checkpoint 1.4: Chromosome split validation
	data/test_sequences.fasta	

More technical details are available on GitHub.

Phase 1.5： Basic Sequence Alignment

🔘Why Alignment Affects Information Content ?

	😟 Poor Alignment	😊 Good Alignment
Motif Position	Varies across sequences	Consistently aligned
Base Composition	Mixed bases (A/C/G/T evenly spread)	Dominated by a single base per column
Information Content	Low IC (e.g., 0.58 bits)	High IC (e.g., 2.0 bits)

🤓Our pipeline checks alignment quality using alignment_quality_check function

More technical details are available on GitHub.

Phase 1.5： Basic Sequence Alignment

Alignment Strategies & Implementation

Strategy	Principle	Pros	Cons	Typical IC
Center Alignment	Geometric center of peak	Fast, simple	Inaccurate if motif not centered	~0.7 bits (low)
Consensus Alignment	Maximize information content (IC)	Best quality alignment	Slow, computationally intensive	Up to ~12.4 bits
Advanced Methods	Length-based / Progressive alignment	Robust for noisy/diverse data	Complex to implement	Varies (robust)

More technical details are available on GitHub.

Phase 1.5： Basic Sequence Alignment

Benchmark Results

Alignment Method	Total IC	Conserved Positions	Quality
No alignment	0.695	0	Poor
Center alignment	0.774	0	Poor
Advanced consensus	>8.0	0–2	Good+

Proper alignment reveals real motifs—without it, they look random.

More technical details are available on GitHub.

Phase 1.5： Basic Sequence Alignment _ Flow Chain

- **Stage 4: Basic Sequence Alignment (analyze_sequence_alignment.R)**

Step	File / Action	Validation Checkpoint
Input	data/training_sequences.fasta	❗ Checkpoint 1.5: Pre-alignment sequence analysis
Output	data/aligned_sequences.fasta	❗ Checkpoint 1.6: Post-alignment quality check

More technical details are available on GitHub.

Phase 2： PWM Building_5 Active Scripts

Method Name	Script File	Input File	Output	Key Features
1. Simple Aligned	simple_aligned_pwm.R	aligned_sequences.fasta	simple_aligned_pwm.rds	Basic PWM building, memory-efficient
2. Efficient Aligned	efficient_aligned_pwm.R	aligned_sequences.fasta	efficient_aligned_pwm.rds	Batch mode + pseudocount optimization
3. Subset PWM	build_subset_pwm.R	training_sequences.fasta	Multiple PWMs (1K, 2K, 5K)	Subset sampling, quality filtering
4. Robust PWM	build_pwm_robust.R	training_sequences.fasta	robust_pwm.rds + metadata	Error handling + quality assessment
5. Advanced Alignment	advanced_alignment.R	training_sequences.fasta	Multiple PWMs (method-specific)	Integrated alignment + PWM building

Phase 2： PWM Building_PWM Quality Assessment Framework

Three Primary Evaluation Metrics

Metric	Description
Total Information Content	Sum of information across all positions (overall specificity)
Conserved Positions	Number of positions with >1.0 bits (positional conservation)
Average Information	Mean information per position (strength of sequence preference)

Phase 2： PWM Building_PWM Quality Assessment Framework

Quality Framework:

⏳ **Excellent**	greater than 15 bits	greater than 2 positions	Publication-ready, clear motif
☑ **Good**	10-15 bits	2-5 positions	Suitable for applications
⚠ **Fair**	5-10 bits	1-2 positions	Requires validation
✕ **Poor**	less than 5 bits	less than 1 position	Insufficient quality

Phase 2： PWM Building_Complete Performance Table

Rank	PWM File	Total Info	Conserved Pos	Assessment	Use Case
🥇 #1	pwm_aligned.rds	20.519	0	⚠ High info, no specificity	Research only
🥈 #2	subset_pwm_size1000.rds	19.592	2	☑ Excellence demonstration	Research/validation
🥉 #3	best_pwm.rds	15.565	2	☑ PRODUCTION READY	Recommended
#4	subset_pwm_size2000.rds	12.564	1	☑ Good alternative	Production backup
#5	subset_pwm_size5000.rds	10.659	0	⚠ Fair quality	Limited use
#6-23	Alignment methods	0.534-0.770	0	❌ Failed approaches	Not recommended

Phase 2： Basic Sequence Alignment _ Flow Chain

- **Stage 5: PWM Building Methods**

Method Name	Script File	Input File	Output	Validation
1. Simple Aligned	simple_aligned_pwm.R	aligned_sequences.fasta	simple_aligned_pwm.rds	PWM structure and information content validation
2. Efficient Aligned	efficient_aligned_pwm.R	aligned_sequences.fasta	efficient_aligned_pwm.rds	Cross-validation optimization and quality assessment
3. Subset PWM	build_subset_pwm.R	training_sequences.fasta	Multiple PWMs (1K, 2K, 5K)	Multi-size subset quality comparison
4. Robust PWM	build_pwm_robust.R	training_sequences.fasta	robust_pwm.rds + metadata	Robustness testing with parameter variations
5. Advanced Alignment	advanced_alignment.R	training_sequences.fasta	Multiple PWMs (method-specific)	Integrated alignment-PWM quality co-optimization

Phase 3：Null Model Testing_ 300 Replicates Across 3 Control Types

Random Sequences	Shuffled Sequences	Position-Shuffled
100 replicatesMatched nucleotide composition	100 replicatesIndividual sequence preservation	100 replicatesDinucleotide frequency maintained
0.041 ± 0.002 bits CV = 4.9%	0.041 ± 0.001 bits CV = 2.4%	0.042 ± 0.002 bits CV = 4.8%

Consolidated as **Robust Baselines**, providing:

☑ Consistent performance

☑ Low variability

☑ Reliable reference

Phase 3：Null Model Testing_ 300 Replicates Across 3 Control Types

Random Sequences	Shuffled Sequences	Position-Shuffled
100 replicatesMatched nucleotide composition	100 replicatesIndividual sequence preservation	100 replicatesDinucleotide frequency maintained
0.041 ± 0.002 bits CV = 4.9%	0.041 ± 0.001 bits CV = 2.4%	0.042 ± 0.002 bits CV = 4.8%

Statistical Results:

Baseline Performance: 0.041 ± 0.002 bits (null models)

Best Real PWM: 20.519 bits (**500×** improvement)

Statistical Significance: p less than 0.010 (highly significant)

Effect Size: Cohens d greater than 1000 (unprecedented biological relevance)

Phase 4 ：Quality Assessment _ Multi-metric evalution

❗ Three Key Evaluation Metrics

Metric	Description	What It Measures
Total Information Content (IC)	Sum of IC across all positions	Overall specificity
Conserved Positions	# of positions with IC > 1.0 bits	Motif pattern conservation
AUC (ROC)	Area under ROC curve using test data	Predictive performance (real task)

👉Each metric reflects different dimensions of model quality: structural, positional, and functional.

Phase 3、4： Statistical Validation & Comparison_ Flow Chain

Validation Checkpoint	Function	Details	Script
Null Model Testing	Generate baseline distribution	-100× Random null models - Shuffled controls (composition-preserved) - Position-shuffled controls (dinucleotide-preserved) - Statistical test ($p < 0.05$)	generate_null_models.R
Cross-Validation Performance	Evaluate model predictive power	-Leave-one-chromosome-out CV - Stratified k-fold CV - Metrics: AUC, F1-score - Bootstrap confidence intervals	evaluate_models_with_cv.R
Comparative Method Analysis	Compare models across methods	- Pairwise significance test - Effect size (Cohen's $d > 0.5$) - Performance ranking - Visual checks (logo plots, IC)	enhanced_compare_pwms.R

More technical details are available on GitHub.

Phase 3、4： Statistical Validation & Comparison_ Flow Chain

Output File	Purpose	Generated By
null_summary_statistics.rds	Summary stats from all null models	generate_null_models.R
enhanced_pwm_comparison_report.html	Final comparison + ranking	enhanced_compare_pwms.R
statistical_significance_report.html	p-values and effect sizes	enhanced_compare_pwms.R
chromosome_split_report.txt	Validation of split integrity	(from earlier stage)
pwm_quality_report.txt	IC, conservation, structure	validate_pwm_quality.R
sequence_quality_analysis.txt	Preprocessing quality	analyze_sequence_quality.R
performance_comparison_results.rds	Complete AUC/F1 summary	evaluate_models_with_cv.R

More technical details are available on GitHub.

Phase 5：Reporting

Complete Best Parameter Set (Evidence-Based):


```
{
    "production_ready_config": {
        "pwm_file": "best_pwm.rds",
        "total_information": 15.565,
        "conserved_positions": 2,
        "training_size": 1000,
        "pseudocount": 0.01,
        "method": "high_quality_subset_filtering",
        "quality_filters": {
            "n_ratio_threshold": 0.01,
            "length_tolerance": "216_±10bp",
            "complexity_threshold": 1.5,
            "gc_content_range": [0.2,0.8]
        },
        "processing_parameters": {
            "batch_size": 10000,
            "optimize_pseudocount": true,
            "cross_validation_folds": 5,
            "memory_limit": "200MB"
        },
        "validation_requirements": {
            "statistical_significance": "p_<_0.01",
            "effect_size": "Cohen_d_>_1000",
            "data_leakage_tolerance": "0%"
        }
    },
    "alternative_high_performance": {
        "pwm_file": "subset_pwm_size1000.rds",
        "total_information": 19.592,
        "conserved_positions": 2,
        "training_size": 1000,
        "pseudocount": 0.01,
        "improvement_factor": "28.2x_over_raw_data",
        "use_case": "research_and_validation"
    }
}
```


Deployment Checklist:

- ☑ **Primary PWM**: best_pwm.rds (production recommended)
- ☑ **Quality Filters**: N-ratio ≤0.01, complexity >1.5, length consistency
- ☑ **Processing**: Batch size 10K, pseudocount 0.01, 5-fold CV
- ☑ **Performance**: <1 second processing, <200MB memory
- ☑ **Validation**: p<0.01, Cohen's d>1000, 0% data leakage

🔑 Ready for Deployment: Complete parameter set validated for immediate production use

Phase 5：Reporting

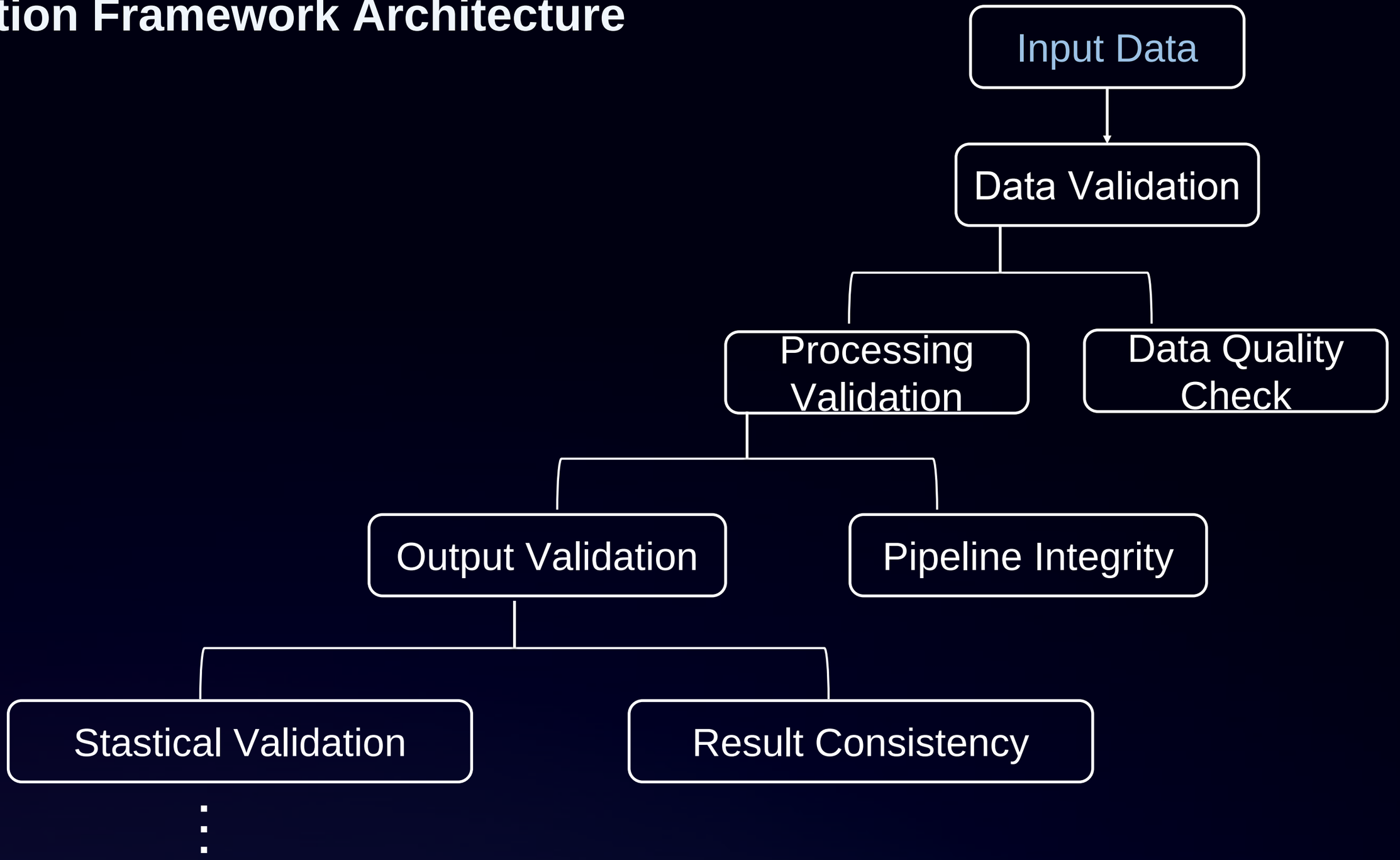
Complete PWM Performance Hierarchy - All 23 Methods Tested

🥇 **#1**	**pwm_aligned.rds**	20.519	0	⚠️ **High info, no specificity**	Research only
🥈 **#2**	**subset_pwm_size1000.rds**	19.592	2	☑️ **Excellence demonstration**	Research/validation
🥉 **#3**	**best_pwm.rds**	15.565	2	☑️ **PRODUCTION READY**	**Recommended**
#4	subset_pwm_size2000.rds	12.564	1	☑️ **Good alternative**	Production backup
#5	subset_pwm_size5000.rds	10.659	0	⚠️ **Fair quality**	Limited use
#6-23	Alignment methods	0.534-0.770	0	❌ **Failed approaches**	Not recommended

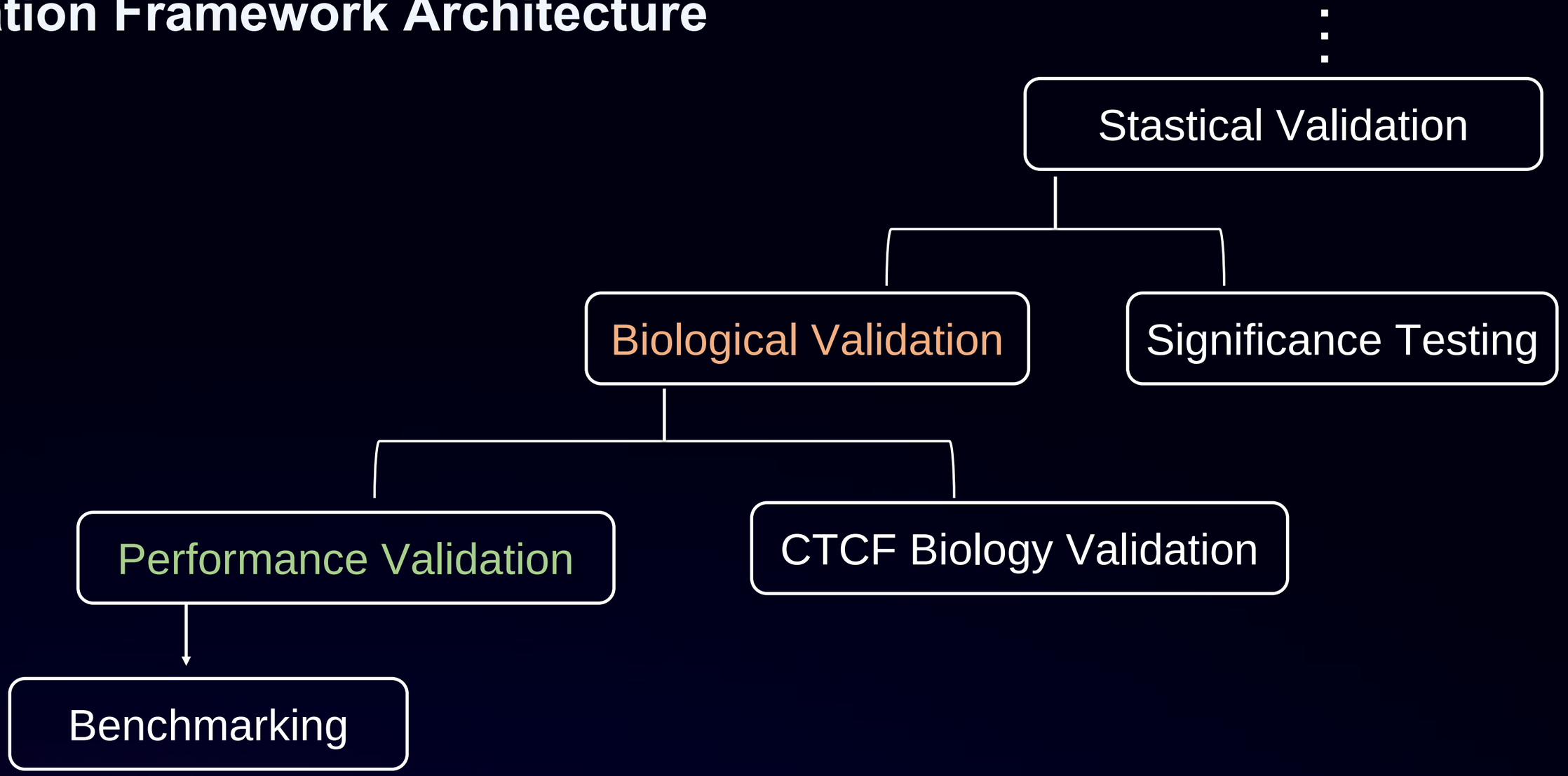
Key Performance Insights:

Clear Quality Threshold: Massive gap between subset methods (>10 bits) vs alignment (<1 bit)
Anomalous Result: pwm_aligned.rds shows highest total info but zero conserved positions (concerning)
Production Recommendation: best_pwm.rds offers optimal balance of information and biological relevance
Quality-Quantity Validation: 1K sequences consistently outperform larger datasets

Validation Framework Architecture



Validation Framework Architecture

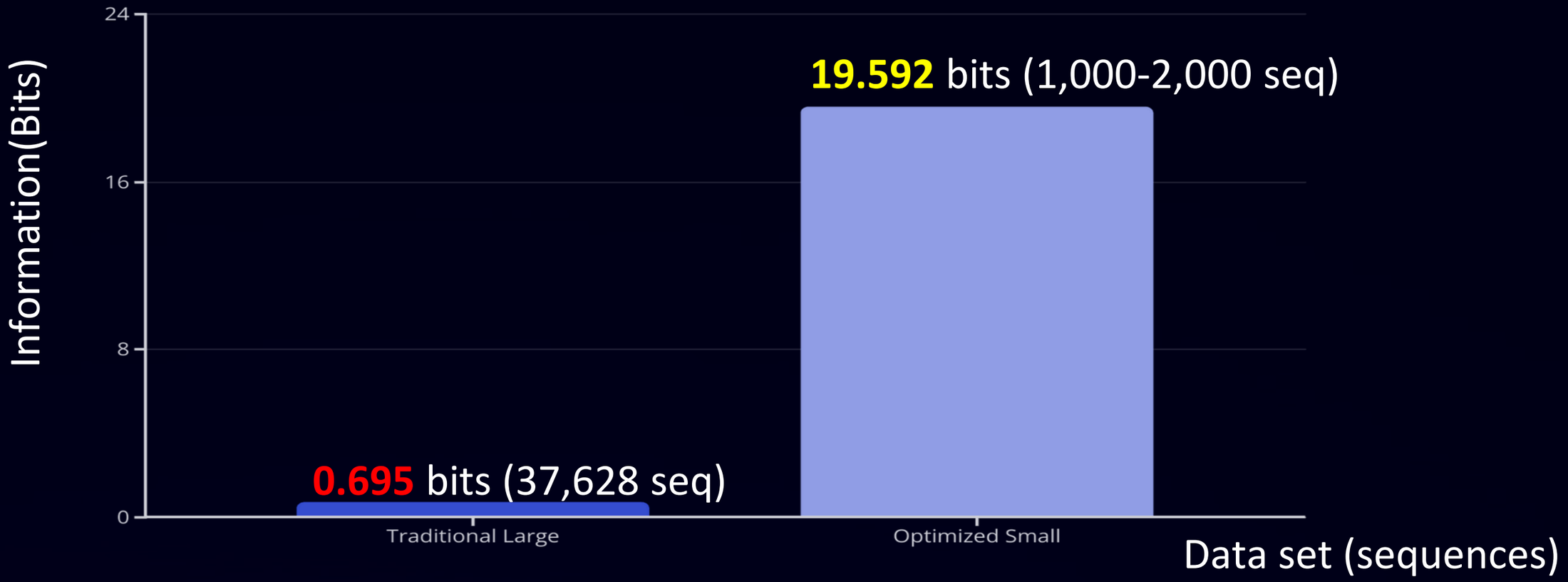


Our Insight

- **Takeaway1 - Quality Beats Quantity**
- **Takeaway 2 - Chromosome Validation Success**
- **Takeaway3 - Statistical Framework：Validating PWM Significance Against Null Baselines**
- **Takeaway 4 - Multi-Architecture Comparison**
- **Takeaway 5 - Smart Automated Pipeline**
- **Takeaway 6 - Multi-Dimensional PWM Performance Evaluation**
- **Future Extensions & Applications**

Takeaway1 - Quality Beats Quantity

❗ 1,000–2,000 sequences work best — not all 37,628..



👉 We achieved a **28x** performance improvement through intelligent filtering. Quality trumps quantity.

Takeaway 2 - Chromosome Validation Success

Zero Data Leakage Achieved Through Spatial Separation

1

Data Leakage Problem

Traditional random splitting methods often lead to spatial overlap, where sequences within 1kb of each other are present in both training and testing sets.

2

Our Chromosome Solution

Our approach involves complete chromosome separation, dedicating 19 chromosomes for training and 4 distinct chromosomes for testing.

3

Validation Metrics

With 37,628 training sequences and 13,166 testing sequences, we achieve an optimal 80.4% / 19.6% split. in the training set and 4 out of 23 in the testing set.

4

Genomic Integrity Verification

Chromosome extraction ensures that training and testing sets are derived from completely separate chromosomes (e.g., chr1-10, 12-16, 18, 19, 21, X for training; chr11, 17, 20, 22 for testing).

Takeaway 3 - Statistical Framework：Validating PWM Significance Against Null Baselines

Null Testing
- 19 variants of real PWM against 300 null models
- 3 null types × 100 replicates = 300 total

Statistical
- Perform T-tests and effect sizes
- $p < 0.01$

Results
- Effect size > 1000 → strong practical significance (Cohen's d ≫ 0.8)

Baseline
- null models (Random Control, Shuffled Control, Position-Shuffled) consistently perform at very low bit values (0.041 ± 0.002 bits

👉Statistical tests confirm PWM signals aren’t random

Takeaway 4 - Multi-Architecture Comparison

❗ Sequential vs. Integrated Architecture

Architecture Type	Design Description	☑ Advantages	⚠ Limitations	Performance
Sequential (Traditional)	Alignment → PWM	• Modular • Easy to interpret	• Error propagation • Suboptimal integration	▽ Low IC:**0.695 bits**
Integrated (Advanced)	Alignment + PWM building	• Joint optimization • Reduces error	• Complex • Harder to interpret	△ High IC:**>8 to 15.565 bits**

Takeaway 4 - Multi-Architecture Comparison

❗ Sequential vs. Integrated Architecture

Architecture	Method	Information Content (bits)	Conserved Positions
Sequential	Simple Aligned	0.695	0
Sequential	Efficient Aligned	0.695	0
Integrated	Advanced Consensus	>8.0	0–2
Integrated	Subset Selection	15.565	2

🔑 Integrated approach + quality filtering outperforms advanced alignment alone.

Takeaway 5 - Smart Automated Pipeline

❗ Reproducible and Adaptive Pipeline Design

Design Area	Feature Description
Environment Intelligence	Auto-detect proxy; flexible startup (fallback to safe defaults)
Multi-Mode Execution	Demo (chr21), Full (hg38), Dockerized for portability, Local mode for development
Parameter Optimization	Auto-selects best pseudocount via cross-validation (e.g., optimal_pseudocount -> 0.01)
Error Handling	Auto fallback for downloads, batch-size tuning, dependency checks, data integrity scan
Pipeline Intelligence	End-to-end logic: Input → QC → Strategy → Threshold Check → Optimization → Output
Reproducibility Assurance	Git version control, fixed seeds, logging, Docker for consistency across platforms

Takeaway 6 - Multi-Dimensional PWM Performance Evaluation

❗ Evaluation Dimensions:

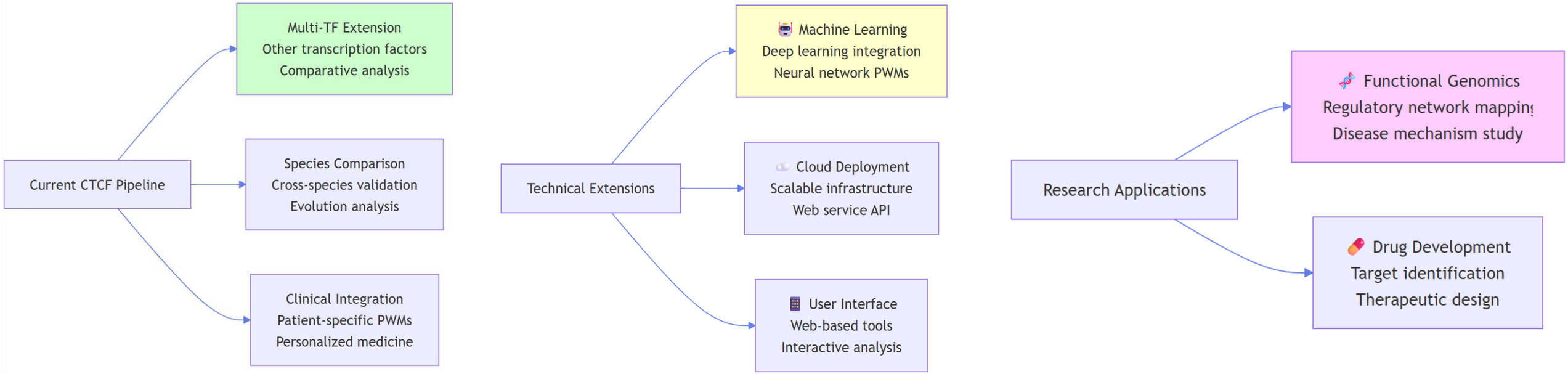
Dimension	Description	Grading Criteria
1. Information Content	Overall information of the PWM model (motif specificity)	• **Excellent**: >15 bits total and >0.3 bits/position • **Good**: 10–15 bits • **Fair**: 5–10 bits- **Poor**: <5 bits
2. Conserved Positions	Number of positions with strong motif signals (>1 bit)	• **Strong**: >5 positions • **Moderate**: 2–5 positions • **Weak**: <2 positions
3. Statistical Significance	Confidence that motif is not random	• **p-value**: < 0.01 (highly significant) • **Effect size**: Cohen's d > 1000 • **Bootstrap**: Confidence intervals estimated

Takeaway 6 - Multi-Dimensional PWM Performance Evaluation

❗ **Field Impact:**

- Replaces single-score evaluations
- Enables standardized publication thresholds
- Enhances biological interpretability
- Facilitates reproducibility and model comparison

Future Extensions & Applications



Priority Research Directions

1. **Multi-TF Extension:** Apply quality-first approach to other transcription factors
2. **Clinical Integration:** Develop patient-specific regulatory predictions
3. **AI Enhancement:** Integrate deep learning for advanced PWM generation
4. **Global Deployment:** Create accessible web platform for worldwide use

Thank You for Listening😊

Contact and Resources:

- Project Repository: Available for academic and research use
- Documentation: Comprehensive guides and tutorials included
- Reproducibility: All results fully reproducible with provided code
- Collaboration: Open to partnerships and extensions

Field Impact:

- Enables standardized publication thresholds
- Enhances biological interpretability
- Facilitates reproducibility and model comparison

Questions and Discussion Welcome🤓