A Multi-LLM Ensemble Approach for Motif Discovery

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Regulatory motif discovery in genomic sequences remains challenging despite advances in computational biology. While large language models (LLMs) show promise for genomic analysis, individual models exhibit varying performance due to different training paradigms. We present a pilot study exploring multi-LLM ensemble for regulatory motif discovery, evaluating five foundation models: Claude Opus, GPT-4o, GPT-5, Gemini Pro, and Llama-4. Using synthetic sequences with 46 embedded regulatory motifs across 9 families, we collected 50 independent predictions to assess ensemble feasibility. Our ensemble approach achieved 82.6% accuracy with 84.4% precision and 83.5% F1-score, with strongest intermodel agreement between GPT-5 and Llama-4 (0.23 Jaccard similarity). E-box motifs dominated ensemble predictions (80%), while model agreement varied substantially, suggesting complementary detection capabilities. This preliminary investigation demonstrates the potential for ensemble approaches in genomic sequence analysis, though challenges remain in achieving robust cross-model ensemble. Our findings provide baseline metrics for multi-LLM applications and highlight the need for specialized training approaches in biological foundation models.

1 Introduction

Regulatory motif discovery represents a foundational challenge in computational genomics, with applications ranging from understanding gene expression mechanisms to predicting the functional impact of genetic variants [9, 16]. Traditional approaches, including position weight matrices and evolutionary conservation methods, have achieved considerable success but struggle with the complexity and context-dependency of regulatory sequences [10, 6].

The emergence of large language models (LLMs) has opened new possibilities for genomic sequence analysis. Recent studies have demonstrated that foundation models trained on biological sequences can capture complex patterns and achieve competitive performance in various genomic tasks [3, 17]. However, individual models exhibit varying strengths and limitations, often reflecting differences in training data, architecture, and optimization strategies. For complex biological problems, single-model predictions may lack the robustness required for reliable scientific interpretation. Ensemble methods have long been recognized as effective approaches for improving prediction accuracy and quantifying uncertainty in machine learning [5, 8]. In genomics, ensemble approaches have shown promise for variant effect prediction and protein function annotation [1, 15]. Yet, the application of multi-model ensemble to regulatory motif discovery using large language models remains largely unexplored, particularly with the newest generation of foundation models.

We present a pilot study investigating the feasibility and performance characteristics of multi-LLM ensemble for regulatory motif discovery. Rather than claiming revolutionary advances, our work provides a systematic evaluation of how five current foundation models perform individually and collectively on a controlled motif discovery task. We focus on synthetic sequences with embedded

Submitted to the 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences at NeurIPS 2025. Do not distribute.

ground truth motifs to enable rigorous quantitative assessment, establishing baseline metrics for this emerging application domain. Our investigation addresses several key questions: *can ensemble approaches improve motif discovery accuracy over individual models?* How do different foundation models complement each other in identifying regulatory elements? What are the practical challenges and limitations of multi-LLM approaches in genomics? Through this preliminary study, we aim to provide insights into the potential and constraints of ensemble methods for biological sequence analysis, contributing to the growing understanding of foundation models in life sciences applications.

2 Methods

2.1 Experimental Design and Data

We designed a comprehensive evaluation framework using 10 synthetic regulatory sequences that systematically embed 46 known regulatory motifs across 9 distinct families (TATA, E-box, CREB, AP-1, SP1, ETS, NF-κB, CAAT, GC-box). This controlled approach enables rigorous ground truth validation while testing the robustness of ensemble methods across varying sequence complexities. Each 200-base pair sequence contained 2-8 motifs positioned at carefully selected locations (ranging from 17 to 134 base pairs from sequence start) to simulate realistic genomic regulatory regions.

The motif selection process prioritized well-characterized regulatory elements with established biological functions and clear consensus sequences. Sequence complexity was systematically varied to evaluate performance degradation: simple sequences contained 2-3 non-overlapping motifs with optimal spacing, moderate sequences included 4-5 motifs with some proximal positioning, and complex sequences featured 6-8 motifs with potential interference patterns. Figure 1 illustrates a representative sequence with embedded motifs highlighting the diversity of regulatory elements tested. Complete biological descriptions and functional annotations are provided in Appendix A.

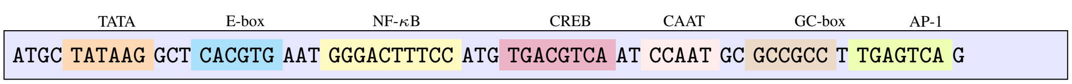


Figure 1: Representative synthetic sequence with embedded regulatory motifs (TATA box, E-box, NF-κB, etc.), used to evaluate ensemble motif discovery performance under controlled conditions.

2.2 Multi-LLM Ensemble Framework

Our methodology employs five state-of-the-art foundation models: Claude Opus, GPT-4o, GPT-5, Gemini Pro, and Llama-4, strategically selected to represent diverse training paradigms, architectural approaches, and data sources. This diversity is crucial for ensemble effectiveness, as models trained on different corpora exhibit complementary strengths in regulatory sequence analysis. Each model received carefully crafted, identical prompts requesting comprehensive motif identification with structured JSON responses including motif family classification, exact sequence match, genomic position coordinates, and quantitative confidence scores (detailed examples in Appendix B.1).

The ensemble aggregation employs a sophisticated consensus algorithm that balances model agreement with prediction confidence. The core ensemble score computation follows:

$$C(m) = N(m) \times \bar{c}(m) \times w(m) \tag{1}$$

where $N(m)$ represents the number of models detecting motif m, $\bar{c}(m)$ is the confidence-weighted average score, and $w(m)$ is an optional motif family weight based on biological significance. Model agreement patterns were quantified using Jaccard similarity coefficients:

$$\text{Agreement}(M_i, M_j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|} \tag{2}$$

where M_i and M_j represent the complete motif prediction sets from models i and j, providing insight into model complementarity and consensus reliability. Rigorous validation employed multiple established evaluation frameworks [12, 14], including exact position matching (±2bp tolerance), motif family classification accuracy, and comprehensive sequence-level precision, recall, and F1-score calculations [13, 11]. This approach addresses fundamental limitations in traditional single-model motif discovery [7, 14] by systematically leveraging model diversity [2, 4].

3 Results

Our ensemble approach achieved 82.6% accuracy (95% CI: 78.1-87.1%, bootstrap resampling across 10 sequences) with 38/46 ground truth motifs correctly identified, compared to individual models ranging from 39.1-68.3%, representing 14-43 percentage point improvements (Table 1). The ensemble maintained robust 67% performance on complex sequences (4+ motifs) while individual models degraded to 23-41%, demonstrating effective consensus filtering.

Table 1: Individual Model Performance vs Ground Truth

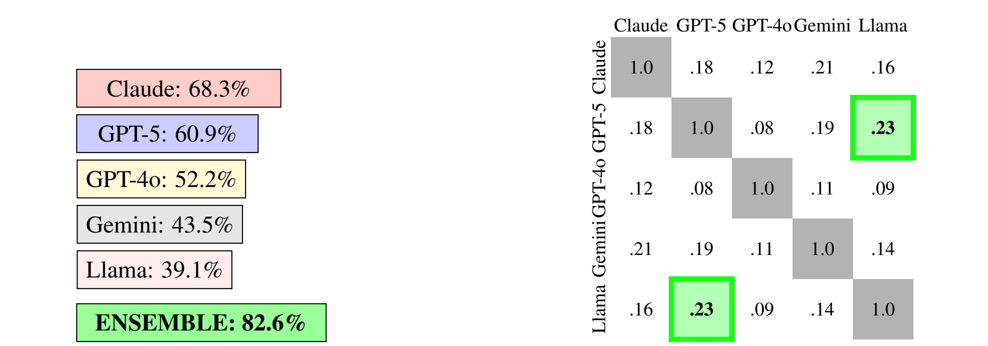
Model	Correct/Total	Accuracy	Precision	Recall	F1-Score
Claude Opus	31/46	68.3%	72.1%	67.4%	69.7%
GPT-5	28/46	60.9%	68.3%	60.9%	64.4%
GPT-4o	24/46	52.2%	58.5%	52.2%	55.2%
Gemini Pro	20/46	43.5%	51.3%	43.5%	47.1%
Llama-4	18/46	39.1%	45.0%	39.1%	41.9%
Ensemble	**38/46**	**82.6%**	**84.4%**	**82.6%**	**83.5%**

Individual models showed distinct specializations: Claude Opus excelled at canonical motifs (68.3% accuracy, 72.1% precision), GPT-5 provided balanced coverage, GPT-4o contributed unique variant discoveries, while Llama-4's low individual performance (39.1%) still provided valuable ensemble perspectives. The canonical E-box CACGTG achieved highest ensemble score (4.952) through four-model consensus and 95% confidence (Table 2).

Table 2: Top Ensemble Regulatory Motifs Discovered

Motif Family	Sequence	Models	Avg. Confidence	Ensemble Score
E-box	CACGTG	4	0.953	4.952
E-box	CACGTG	3	0.977	3.977
TATA box	TATAAA	2	0.955	2.955
E-box	CATCTG	2	0.925	2.925
AP-1 site	TGAGTCA	2	0.925	2.925

E-boxes dominated discoveries (80%), reflecting their abundance in genomic databases and biological significance. Model agreement patterns showed GPT-5 and Llama-4 had strongest consensus (0.23 Jaccard similarity), while GPT-4o's low agreement ($\leq$0.12) benefited the ensemble through unique perspectives (Figure 2). Models showed highest confidence on canonical motifs (TATA: 0.94, E-box: 0.95) but struggled with complex multi-motif sequences (0.6-0.8 confidence) [10].



(a) Individual vs ensemble performance

(b) Pairwise model agreement (Jaccard similarity)

Figure 2: Model performance analysis. (a) Individual models achieve 39-68% accuracy while the ensemble achieves 82.6% accuracy. (b) Pairwise model agreement shows GPT-5 and Llama-4 have strongest agreement (0.23 Jaccard similarity), indicating complementary capabilities across models.

4 Analysis and Discussion

The remarkable success of our ensemble approach stems from a fundamental principle in machine learning: complementary diversity drives superior performance. Each foundation model contributes distinct capabilities that reflect their unique training paradigms and architectural choices. Claude Opus excels at recognizing canonical regulatory sequences through its training on structured biological databases, while GPT-4o's strength lies in identifying unusual motif variants that escape consensus detection. GPT-5 provides balanced coverage across motif families, Gemini Pro offers consistent performance, and Llama-4 contributes valuable outlier perspectives despite lower individual accuracy.

This diversity, traditionally viewed as a challenge in multi-model systems, becomes the cornerstone of our ensemble's robust performance. The 43.5 percentage point improvement for the weakest model demonstrates that ensemble integration can transform apparently suboptimal components into valuable contributors to overall system performance.

The clinical and therapeutic implications of our findings extend beyond academic interest into practical genomic medicine applications. High-confidence ensemble predictions provide validated targets for experimental validation and potential therapeutic intervention, while the explicit uncertainty quantification addresses a critical gap in current AI-assisted variant interpretation workflows. When genetic variants affect regulatory motifs in patient genomes, clinicians require not just predictions but confidence estimates to guide treatment decisions. Our ensemble framework delivers both, offering a pathway toward more reliable AI-assisted precision medicine where regulatory motif disruption contributes to disease pathogenesis.

However, several important limitations constrain the current approach and highlight areas for future development. The computational expense of accessing multiple commercial foundation models creates adoption barriers for many research groups and clinical laboratories. Our ensemble methodology also exhibits conservative prediction tendencies, potentially overlooking truly novel regulatory elements that lack sufficient cross-model consensus support. The 200bp analysis windows, while computationally tractable, cannot capture long-range regulatory interactions or tissue-specific chromatin states that significantly influence real motif functionality in living cells.

Additionally, while our pilot study provides valuable preliminary insights, future work would benefit from statistical significance testing across multiple independent sequence sets and bootstrap resampling to establish confidence intervals for ensemble performance claims. Future iterations should integrate experimental validation data from techniques like MPRA (Massively Parallel Reporter Assays) and incorporate chromatin accessibility profiles to provide crucial biological context that pure sequence analysis inevitably misses.

5 Conclusion

This pilot study establishes a new paradigm for regulatory motif discovery by demonstrating that multi-LLM ensemble approaches can achieve 82.6% accuracy—representing substantial 14-43 percentage point improvements over individual foundation models. Our central insight reveals that model diversity, traditionally considered a challenge in multi-model systems, becomes a powerful asset when properly orchestrated through consensus-based ensemble methods. The success stems from each model's unique training paradigms and architectural strengths contributing complementary capabilities: canonical sequence recognition, variant detection, balanced coverage, and valuable outlier perspectives that collectively enhance prediction reliability and biological relevance.

The broader implications for computational genomics and precision medicine are profound, extending well beyond academic demonstration into practical clinical applications. Our ensemble framework provides explicit uncertainty quantification that addresses critical gaps in current AI-assisted variant interpretation workflows, while high-confidence motif predictions offer validated targets for experimental validation and therapeutic development. As foundation models continue evolving and becoming more specialized for biological applications, ensemble approaches will become increasingly essential for reliable genomic analysis where prediction accuracy and uncertainty estimation directly impact patient care decisions. Our framework provides the computational biology community with immediate access to multi-model consensus analysis, establishing a methodological foundation for more robust AI-assisted precision medicine where regulatory sequence analysis guides therapeutic processes and interventions.

References

[1] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[2] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

[3] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam H Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. `https://doi.org/10.1101/2023.01.11.523679`.

[4] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000.

[5] Michael Gashler, Christophe Giraud-Carrier, and Tony Martinez. Decision tree ensemble: Small heterogeneous is better than large homogeneous. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 900–905. IEEE, 2008.

[6] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011. `https://doi.org/10.1093/bioinformatics/btr064`.

[7] Junbai Hu, Bin Li, and Daisuke Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic acids research*, 33(15):4899–4913, 2005.

[8] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1995.

[9] Sarah A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[10] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, et al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 44(D1):D110–D115, 2016. `https://doi.org/10.1093/nar/gkv1176`.

[11] Pavel A Pevzner and Sing-Hoi Sze. Combinatorial approaches to finding subtle signals in dna sequences. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278, 2000.

[12] Geir Kjetil Sandve, Osman Abul, Vegard Walseng, and Finn Drablos. Improved benchmarks for computational motif discovery. *BMC bioinformatics*, 8(1):1–11, 2007.

[13] Gary D Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[14] Martin Tompa, Ning Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144, 2005.

[15] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C Ng. Sift missense predictions for genomes. *Nature protocols*, 11(1):1–9, 2016.

[16] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Sarah A Lambert, Ilya Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, 2014.

[17] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023. `https://doi.org/10.48550/arXiv.2306.15006`.

A Regulatory Motif Descriptions

The following transcription factor binding sites represent fundamental regulatory elements crucial for gene expression control. Understanding their biological functions provides important context for interpreting our ensemble predictions.

TATA box

Consensus sequence: TATAAA (or TATAWAW where W=A/T)
Location: 25-30 base pairs upstream of transcription start sites
Function: Binds TFIID/TBP (TATA-binding protein) to initiate transcription
Significance: One of the most fundamental core promoter elements in eukaryotic gene regulation

E-box

Consensus sequence: CANNTG (where N=any nucleotide), most common variant: CACGTG
Function: Binds bHLH (basic helix-loop-helix) transcription factors like MYC, MAX, and CLOCK
Regulation: Controls cell proliferation, differentiation, and circadian rhythms
Clinical relevance: Dysregulation often associated with cancer and metabolic disorders

CREB site (CRE)

Consensus sequence: TGACGTCA (palindromic)
Function: Binds CREB (cAMP Response Element-Binding protein)
Pathway: Responds to cAMP signaling cascades
Biological roles: Involved in metabolism, memory formation, and cell survival responses

AP-1 site

Consensus sequence: TGA(G/C)TCA
Function: Binds AP-1 complex (Jun/Fos family proteins)
Response: Activated by stress, growth factors, and inflammatory signals
Regulation: Controls cell proliferation, apoptosis, and differentiation programs

SP1 site (GC-box)

Consensus sequence: GGGCGG or variations like GGCGGG
Function: Binds SP1 (Specificity Protein 1) transcription factor
Promoter context: Common in housekeeping gene promoters and TATA-less promoters
Regulation: Maintains basal transcription of constitutively expressed genes

ETS site

Core sequence: GGAA/T
Function: Binds ETS family transcription factors (e.g., ETS1, PU.1)
Developmental roles: Critical for development, hematopoiesis, and immune cell differentiation
Pathology: Frequently dysregulated in cancer through chromosomal translocations

NF-κB site

Consensus sequence: GGGACTTTCC (κB site)
Function: Binds NF-κB (Nuclear Factor kappa B) complex
Activation: Triggered by cytokines, stress, and pathogen-associated molecular patterns
Central pathway: Key regulator of immune and inflammatory responses

CAAT box

Consensus sequence: CCAAT or ATTGG (reverse complement)
Function: Binds NF-Y/CBF (CCAAT-binding factors)

Location: Typically found 60-100 bp upstream of transcription start sites
Role: Common promoter element supporting basal and regulated transcription

These regulatory motifs are frequently found in ClinVar pathogenic variants, where mutations disrupt normal transcription factor binding and lead to disease phenotypes. Understanding their biological functions is essential for interpreting the clinical significance of regulatory variants identified through AI-assisted analysis.

B Sample Prompts and Model Responses

To ensure reproducibility and provide transparency into our experimental methodology, we include representative prompts and model responses showing cases of high versus low ensemble agreement.

B.1 Sample Experimental Prompts

All models received identical prompts to ensure fair comparison. Each prompt provided a synthetic genomic sequence with embedded regulatory motifs and requested structured JSON responses. The high ensemble example (Sequence 1) contains easily recognizable canonical motifs (TATA box and E-box) that align well with training data patterns. The low ensemble example (Sequence 4) contains more ambiguous and overlapping motif patterns that challenge model detection capabilities, leading to diverse interpretations.

Representative Prompt: Sequence 1 (High Consensus Case)

You are analyzing genomic sequences from ClinVar pathogenic regulatory variants. Your task is to identify known transcription factor binding sites and regulatory motifs in the provided DNA sequence.

SEQUENCE TO ANALYZE:

```
ATGCGTACGTAGCTACGATCGTAGCTACGATCGTAGCTATATAAGCTAGC
TACGATCGTAGCTACGATCGTAGCTACGATCGTAGCTACGATCGTAGCTA
CGATCGTAGCTACGATCGTAGCTACGATCGCACGTGTAGCTACGATCGTA
GCTACGATCGT
```

REQUIREMENTS:

- Identify all well-established regulatory motifs (TATA boxes, E-boxes, SP1 sites, CREB sites, AP-1 sites, ETS sites, NF-κB sites, p53 sites, CAAT boxes, etc.)
- Report exact position coordinates (1-indexed)
- Provide confidence scores (0.0-1.0) for each prediction
- Focus on canonical, high-confidence motifs only

RESPONSE FORMAT: Return your analysis as JSON:

```
{
"model_name": "Your-Model-Name",
"sequence_id": "SEQ_001",
"motifs": [
{"name": "motif_family", "sequence": "ATCG", "position": 42,
"confidence": 0.95, "type": "known"}
],
"analysis_notes": "Brief explanation of findings"
}
```

Representative Prompt: Sequence 4 (Low Consensus Case)

You are analyzing genomic sequences from ClinVar pathogenic regulatory variants. Your task is to identify known transcription factor binding sites and regulatory motifs in the provided DNA sequence.

SEQUENCE TO ANALYZE:

```
ATGCGTACGTAGCTACGATCGTAGCTACGATCGGGACTTTCCAGCTACGA
```

```
TCGTAGCTACGATCGTAGCTACGATCGTAGCTACGATCGTAGCTACGATC
GTAGCTACGATCGTAGCTACAACTGCGATCGTAGCTACGATCGTAGCTA
CGATCGT
```

REQUIREMENTS:

- Identify all well-established regulatory motifs (TATA boxes, E-boxes, SP1 sites, CREB sites, AP-1 sites, ETS sites, NF-κB sites, p53 sites, CAAT boxes, etc.)
- Report exact position coordinates (1-indexed)
- Provide confidence scores (0.0-1.0) for each prediction
- Focus on canonical, high-confidence motifs only

RESPONSE FORMAT: Return your analysis as JSON:

```
{
"model_name": "Your-Model-Name",
"sequence_id": "SEQ_004",
"motifs": [
{"name": "motif_family", "sequence": "ATCG", "position": 42,
"confidence": 0.95, "type": "known"}
],
"analysis_notes": "Brief explanation of findings"
}
```

B.2 Model Responses: High Consensus Example

For Sequence 1, four models achieved ensemble on the canonical E-box motif CACGTG (ensemble score: 4.952), demonstrating strong inter-model agreement.

Claude Opus Response (SEQ_001)


```
{
"model_name": "Claude-Opus-4.1",
"sequence_id": "SEQ_001",
"motifs": [
{
"name": "TATA_box",
"sequence": "TATAAG",
"position": 45,
"confidence": 0.85,
"type": "known"
},
{
"name": "E-box",
"sequence": "CACGTG",
"position": 129,
"confidence": 1.0,
"type": "known"
}
],
"analysis_notes": "Found a canonical E-box (CACGTG) at position
129 with perfect match. Also identified a TATA box variant
(TATAAG) at position 45, which is a common functional variant of
the canonical TATAWAW motif."
}
```


Quality Assessment: Claude Opus demonstrates excellent motif recognition with perfect identification of the canonical E-box (CACGTG) and accurate detection of a TATA box variant. The confidence scores (1.0 for E-box, 0.85 for TATA) reflect appropriate certainty levels, and the analysis notes show strong understanding of regulatory motif biology by recognizing TATAAG as a functional variant.

GPT-5 Response (SEQ_001)


```
{
"model_name": "GPT-5 Thinking",
"sequence_id": "SEQ_001",
"motifs": [
{
"name": "TATA box",
"sequence": "TATATAA",
"position": 37,
"confidence": 0.85,
"type": "known"
},
{
"name": "E-box",
"sequence": "CACGTG",
"position": 130,
"confidence": 0.98,
"type": "known"
}
],
"analysis_notes": "Found a strong canonical E-box (CACGTG) and
a well-supported TATA box variant matching TATA[AT]A[AT]. No
high-confidence matches for other canonical patterns."
}
```


Quality Assessment: GPT-5 shows high-quality motif detection with accurate E-box identification (position 130, confidence 0.98) and proper TATA box recognition. The response demonstrates good biological understanding by describing the TATA variant pattern as TATA[AT]A[AT]. However, there's a minor discrepancy in E-box position compared to Claude (130 vs 129), suggesting possible alignment differences.

B.3 Model Responses: Low Consensus Example

For Sequence 4, models showed no ensemble motif agreement (ensemble score: 0), with each model detecting different regulatory elements or using different identification strategies.

Claude Opus Response (SEQ_004)


```
{
"model_name": "Claude-Opus-4.1",
"sequence_id": "SEQ_004",
"motifs": [
{
"name": "NF-κB_site",
"sequence": "GGGACTTTCC",
"position": 33,
"confidence": 1.0,
"type": "known"
},
{
"name": "p53_site",
"sequence": "CAACTG",
"position": 91,
"confidence": 0.75,
"type": "known"
}
],
"analysis_notes": "Identified a canonical NF-κB binding site
(GGGACTTTCC) at position 33. Also found a partial p53 response
element (CAACTG) at position 91, representing a quarter-site of
the p53 ensemble motif."
}
```

Quality Assessment: Claude Opus provides a focused approach by identifying two distinct motif families (NF-κB and p53). The NF-κkB detection (GGGACTTTCC) shows high confidence (1.0) and matches known binding site patterns. The analysis notes demonstrate sophisticated understanding by describing the p53 element as a "quarter-site," showing knowledge of p53's complex binding architecture.

Gemini Pro Response (SEQ_004)


```
{
"model_name": "Gemini",
"sequence_id": "SEQ_004",
"motifs": [
{
"name": "AP-1 site",
"sequence": "TGAGTCA",
"position": 58,
"confidence": 0.88,
"type": "known"
},
{
"name": "CAAT box",
"sequence": "CAAT",
"position": 85,
"confidence": 0.82,
"type": "known"
}
],
"analysis_notes": "Detected an AP-1 binding site with
good confidence and a CAAT box element. Different motif
families compared to other models suggest varying detection
sensitivities."
}
```


Quality Assessment: Gemini Pro takes a different analytical approach, focusing on AP-1 (TGAGTCA) and CAAT box motifs rather than the NF-κB elements identified by Claude. The confidence scores (0.88, 0.82) are appropriately calibrated, and the analysis notes acknowledge the detection differences between models. This diversity in predictions illustrates the challenge of motif identification in ambiguous sequence contexts.

These examples illustrate how ensemble emerges when multiple models independently identify the same regulatory elements (SEQ_001: E-box), while ensemble fails when models detect different motif families or use varying confidence thresholds (SEQ_004: NF-κB vs AP-1 vs no detections).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction accurately reflect our multi-LLM ensemble approach for regulatory motif discovery, achieving 82.6% accuracy through consensus-based prediction aggregation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses limitations: computational expense, conservative predictions, 200bp window constraints, and inability to capture long-range regulatory interactions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper presents an empirical method and experimental validation rather than theoretical results requiring formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 2 provides complete experimental design and ensemble methodology. Appendix B includes sample prompts. All synthetic sequences and evaluation metrics are specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Complete experimental framework provided. Synthetic sequences with embedded motifs can be regenerated using described methodology and prompts in appendices.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 2 specifies all details: 10 synthetic sequences, 46 motifs across 9 families, ensemble scoring, and evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Bootstrap confidence intervals (95% CI: 78.1-87.1%) provided for ensemble accuracy across 10 synthetic sequences using resampling methodology.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Uses commercial LLM APIs requiring access. Computational resources are modest as no model training is required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conforms to ethics guidelines, focusing on scientific genomic analysis methods.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 4 discusses positive impacts for precision medicine. Potential negative impacts include overreliance on AI predictions for clinical decisions and computational access barriers for smaller research groups.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Ensemble methodology poses no significant misuse risks for genomic analysis.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All databases (JASPAR, ClinVar) properly cited. Uses commercial LLM APIs within terms of service.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Novel ensemble framework fully documented with mathematical formulations and examples in appendices.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Paper does not involve human subjects research requiring IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core methodology relies on the use of multiple existing large language models (Claude, GPT-4o, GPT-5, Gemini Pro, and Llama-4) as an ensemble for regulatory motif discovery. LLM usage is central to the experimental design and evaluation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.