Noise-Resistant Training Strategies for Sequence-to-Expression Prediction

Evaluation on CAGI5 Saturation Mutagenesis Benchmark

74 models · 7 methods · 4 regulatory elements

Methods: Rank Stability, Distributional, Noise-Gated, Contrastive,
Quantile Sampling, Curriculum Learning, Hard Negative Mining

Overview of Noise-Resistant Training Approaches

Motivation: LentiMPRA measurements contain inherent experimental noise (aleatoric uncertainty) that varies across sequences. Standard MSE training weights all samples equally, potentially overfitting to noisy measurements and hurting generalization to clinical variants.

Approach: We systematically evaluate seven noise-aware training strategies:

- Rank Stability (RS) — Down-weight pairwise comparisons involving high-noise samples
- Distributional (DH) — Jointly predict activity mean and variance, supervise with measured noise
- Noise-Gated (NG) — Combine heteroscedastic loss with noise-weighted ranking
- Contrastive (CA) — Learn representations where noise level determines similarity
- Quantile Sampling (QS) — Construct batches stratified by activity and noise levels
- Curriculum (QC) — Progressively increase quantile resolution during training
- Hard Negative (HN) — Focus on distinguishing similar sequences with reliable labels

Evaluation: CAGI5 benchmark with 4 K562-matched elements (GP1BB, HBB, HBG1, PKLR), reporting both Spearman and Pearson correlations stratified by variant confidence.

Figure 1. We developed seven noise-aware training strategies that leverage aleatoric uncertainty from lentiMPRA replicates. Each approach addresses the noise problem differently, from loss reweighting to sampling strategies.

Test vs CAGI5 Performance

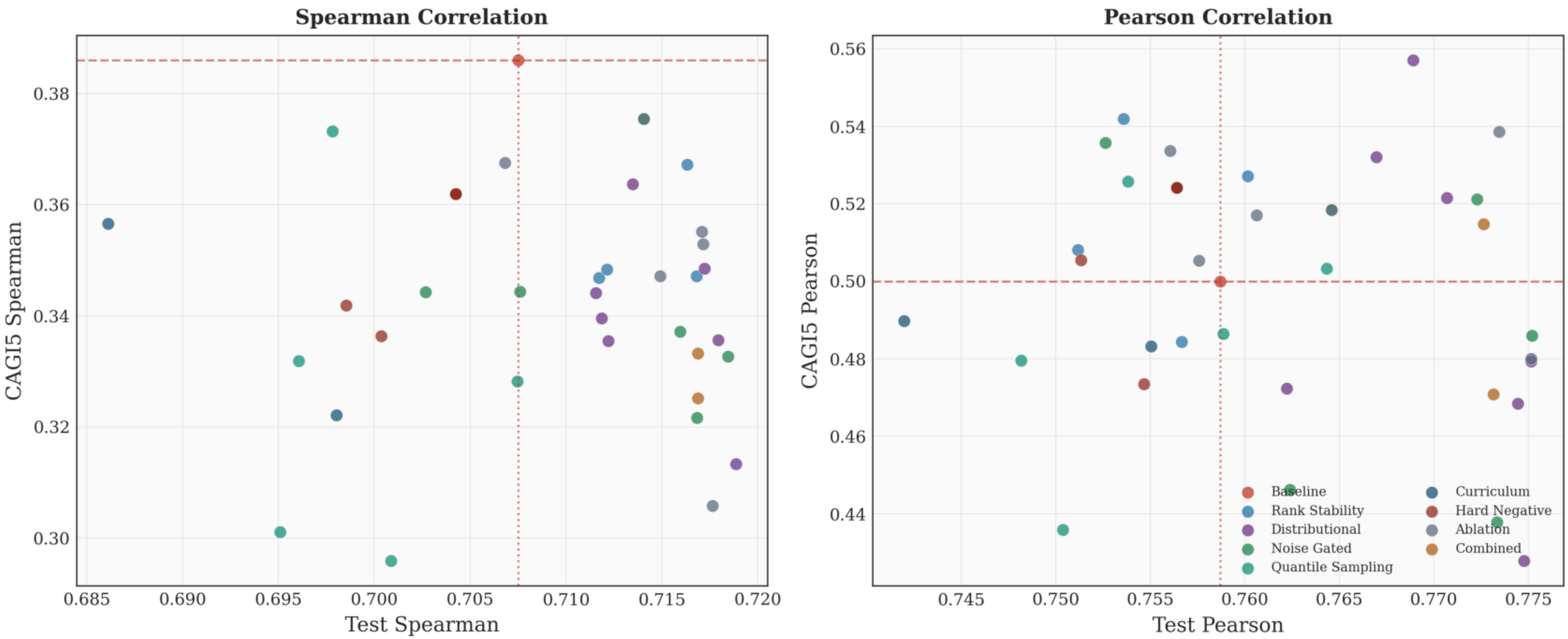


Figure 2. Comparison of held-out test performance with CAGI5 clinical variant prediction. Dashed lines show baseline. Test and CAGI5 metrics are only moderately correlated, highlighting the importance of evaluating on external benchmarks.

Rank Stability (RS) Results

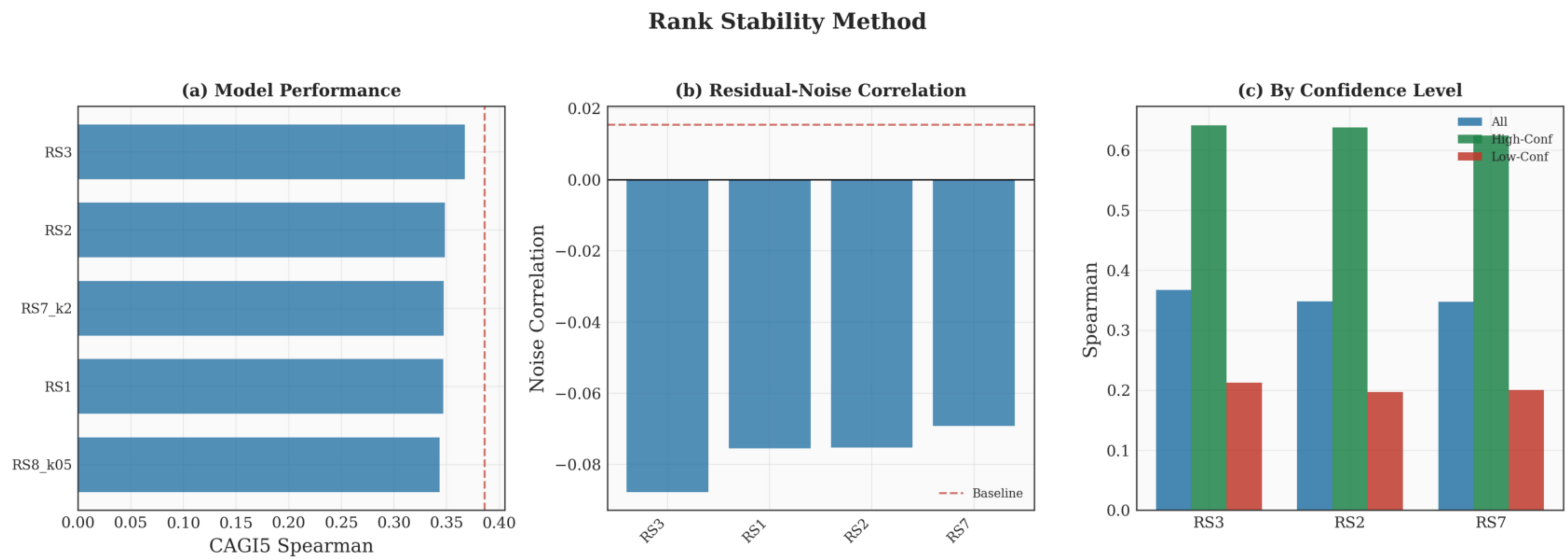


Figure 3. RS weights pairwise comparisons by inverse noise. (a) CAGI5 performance across RS variants. (b) Noise correlation—RS3 achieves -0.088, the only negative value observed. (c) Top models stratified by confidence.

Distributional (DH) Results

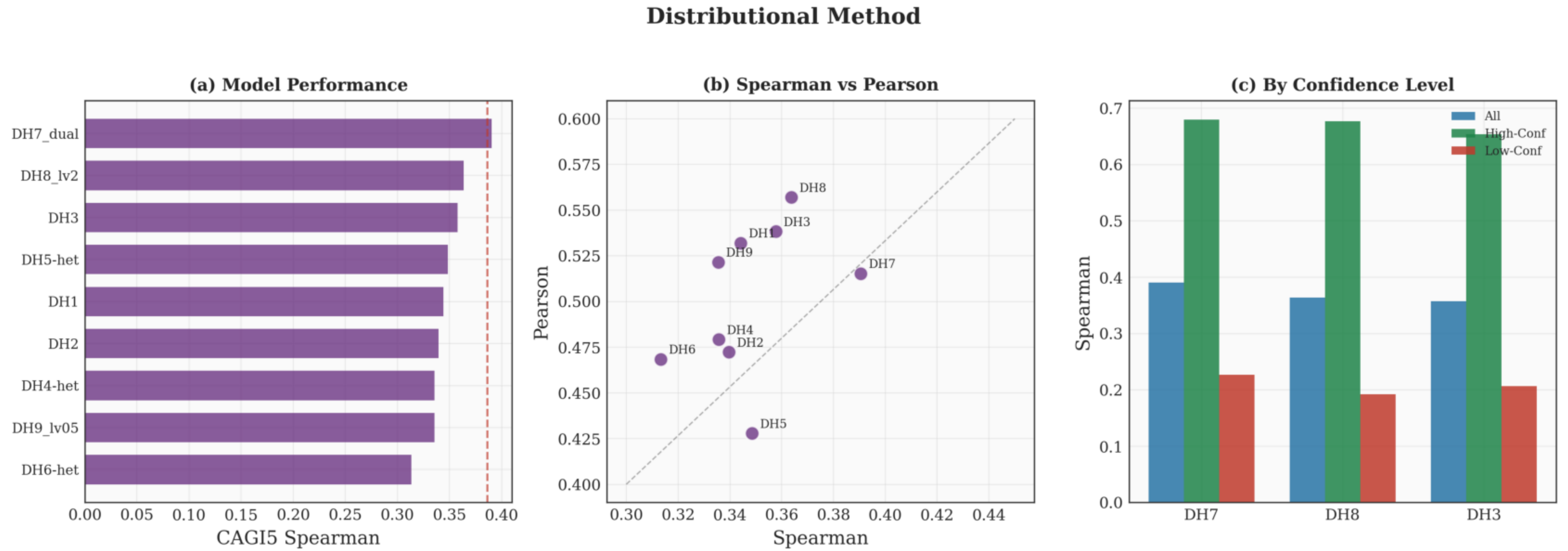


Figure 4. DH models predict both mean and variance. (a) DH7 (dual-head) achieves best Spearman at 0.391. (b) DH8 shows highest Pearson (0.557). (c) DH7 performs well across confidence levels.

Noise-Gated (NG) and Ablation Results

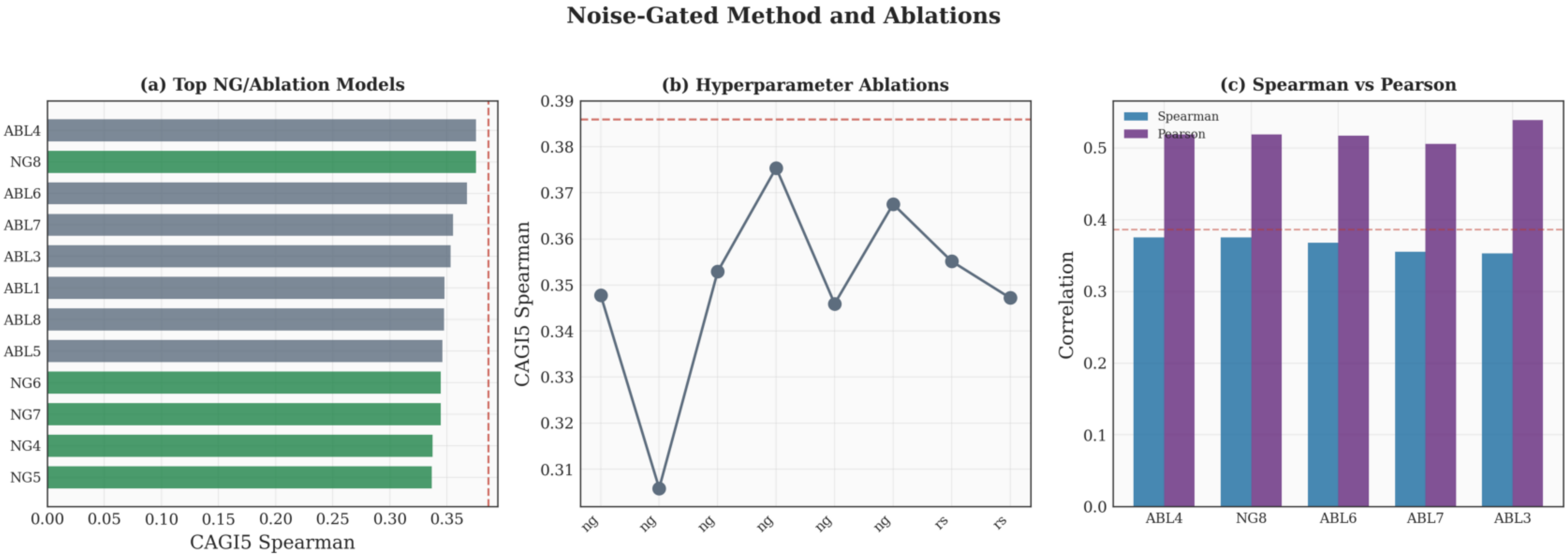


Figure 5. NG combines heteroscedastic loss with rank stability. (a) Top performers from NG and ablation experiments. (b) Ablation sweep showing optimal parameters. (c) Both metrics improve over baseline.

Contrastive (CA) Results

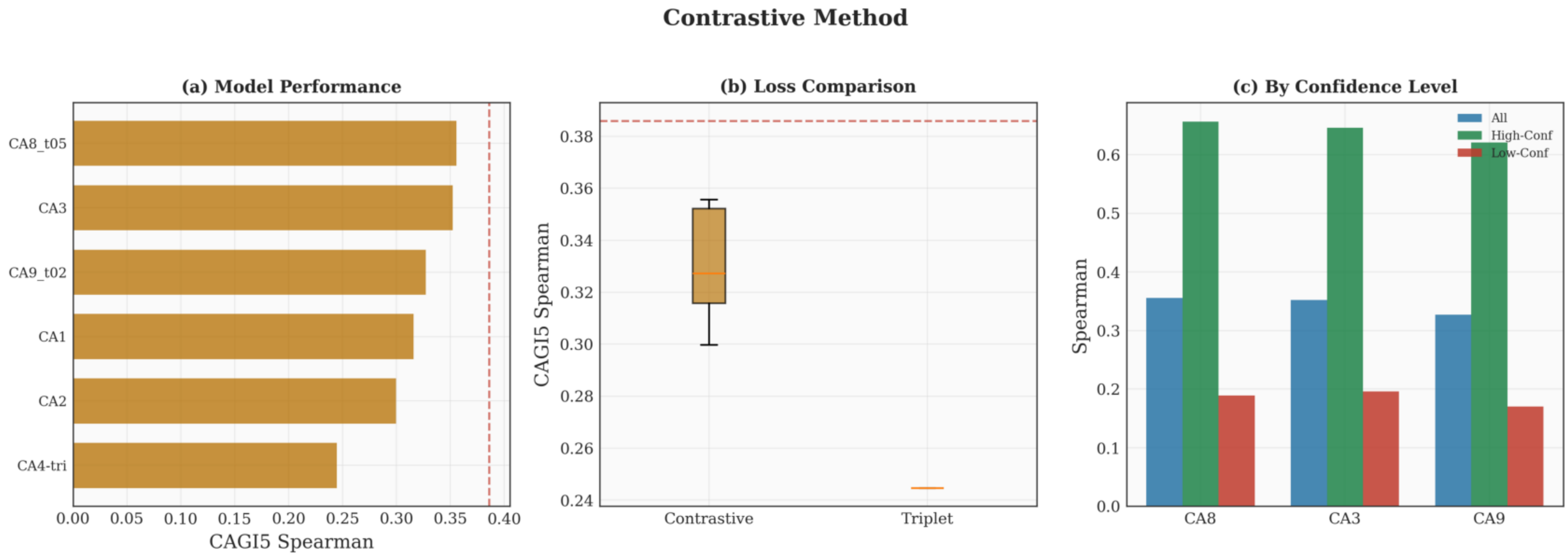


Figure 6. Contrastive learning uses noise-based similarity. (a) High variance across CA models. (b) Standard contrastive outperforms triplet loss. (c) Best CA models show reasonable stratified performance.

Sampling Strategy Results

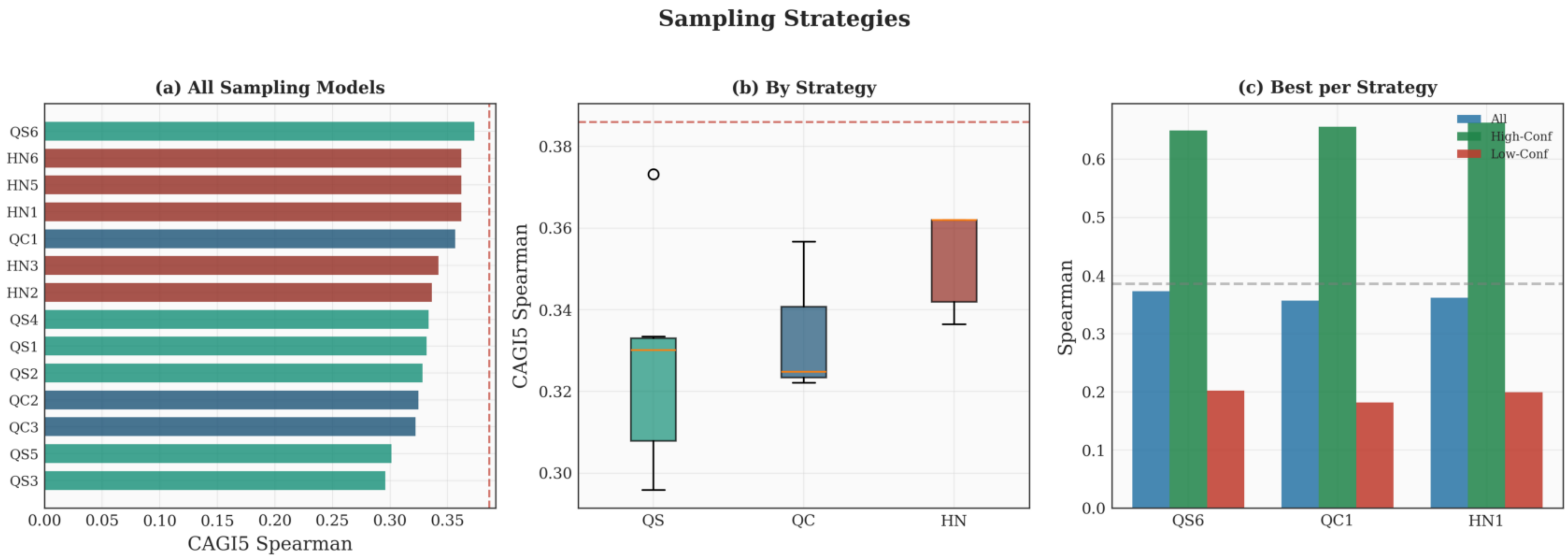


Figure 7. Sampling strategies modify batch construction. (a) QS6 achieves best overall. (b) Quantile sampling shows most consistent improvement. (c) Best model from each strategy.

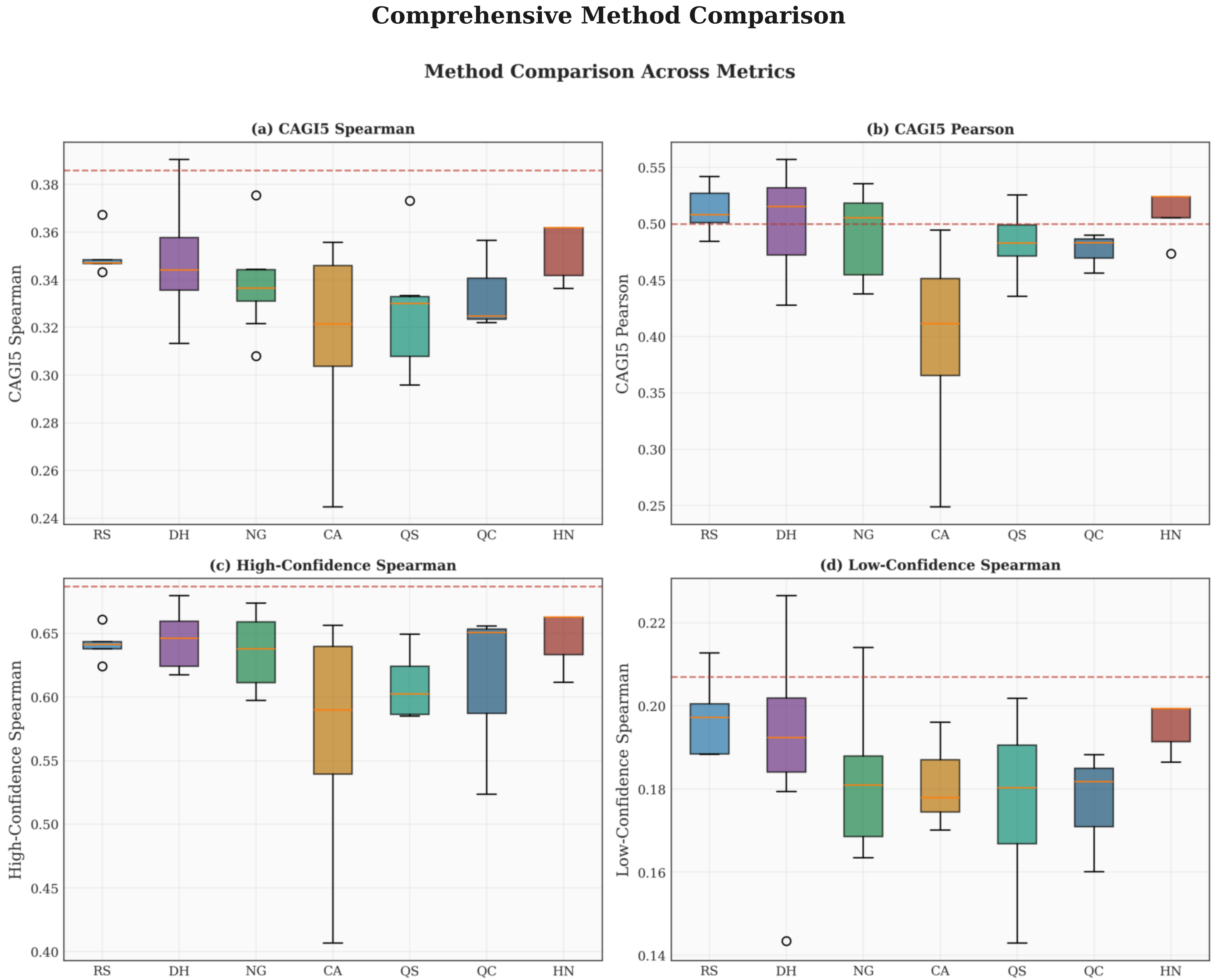


Figure 8. Comparison across all seven methods. (a) Overall Spearman—DH and NG lead. (b) Overall Pearson—similar pattern. (c) High-confidence performance. (d) Low-confidence—DH shows largest gains over baseline.

Top 15 Models

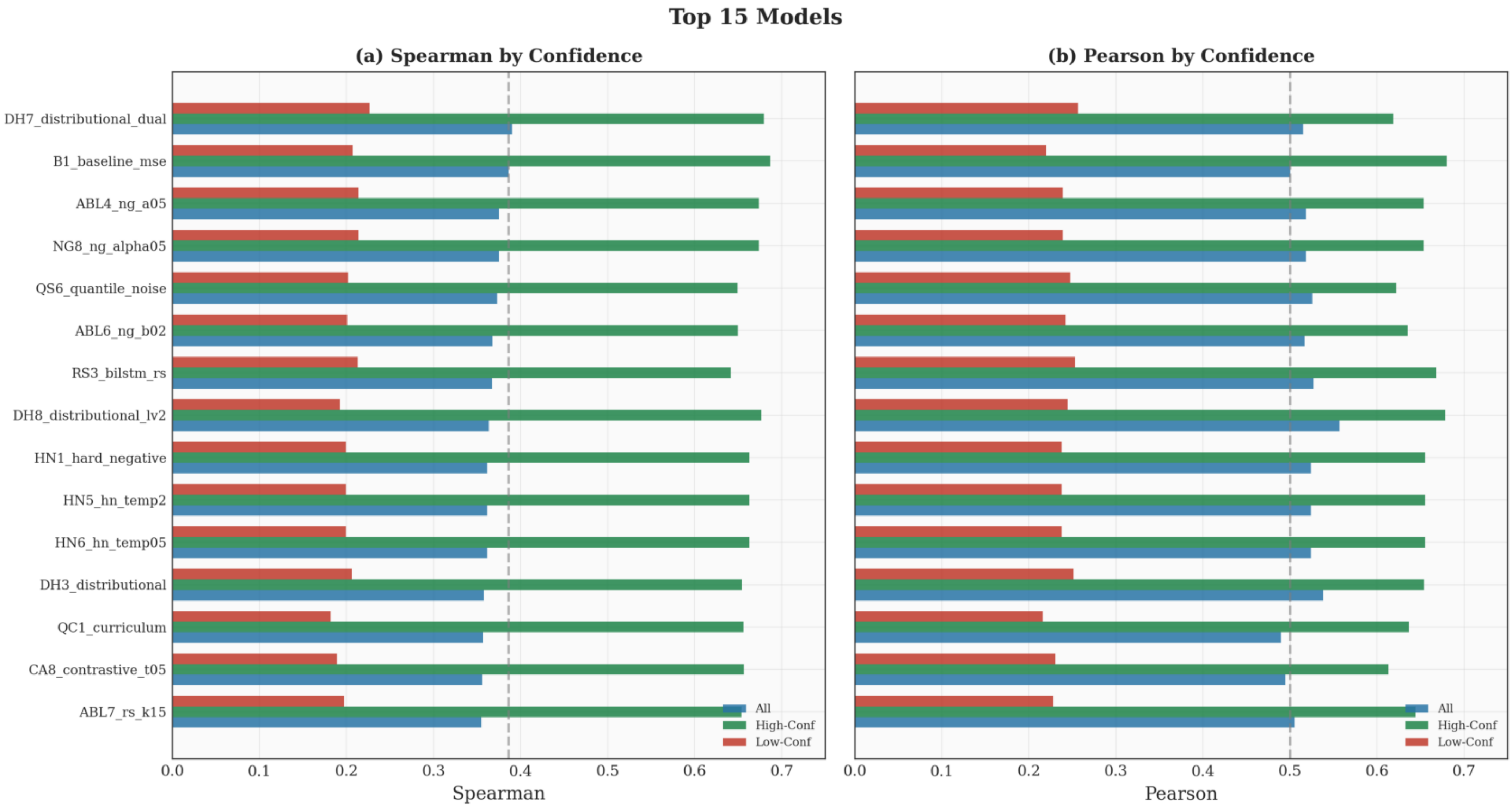


Figure 9. Best 15 models ranked by CAGI5 Spearman. DH7 leads overall with balanced performance across confidence levels.

Per-Element CAGI5 Performance

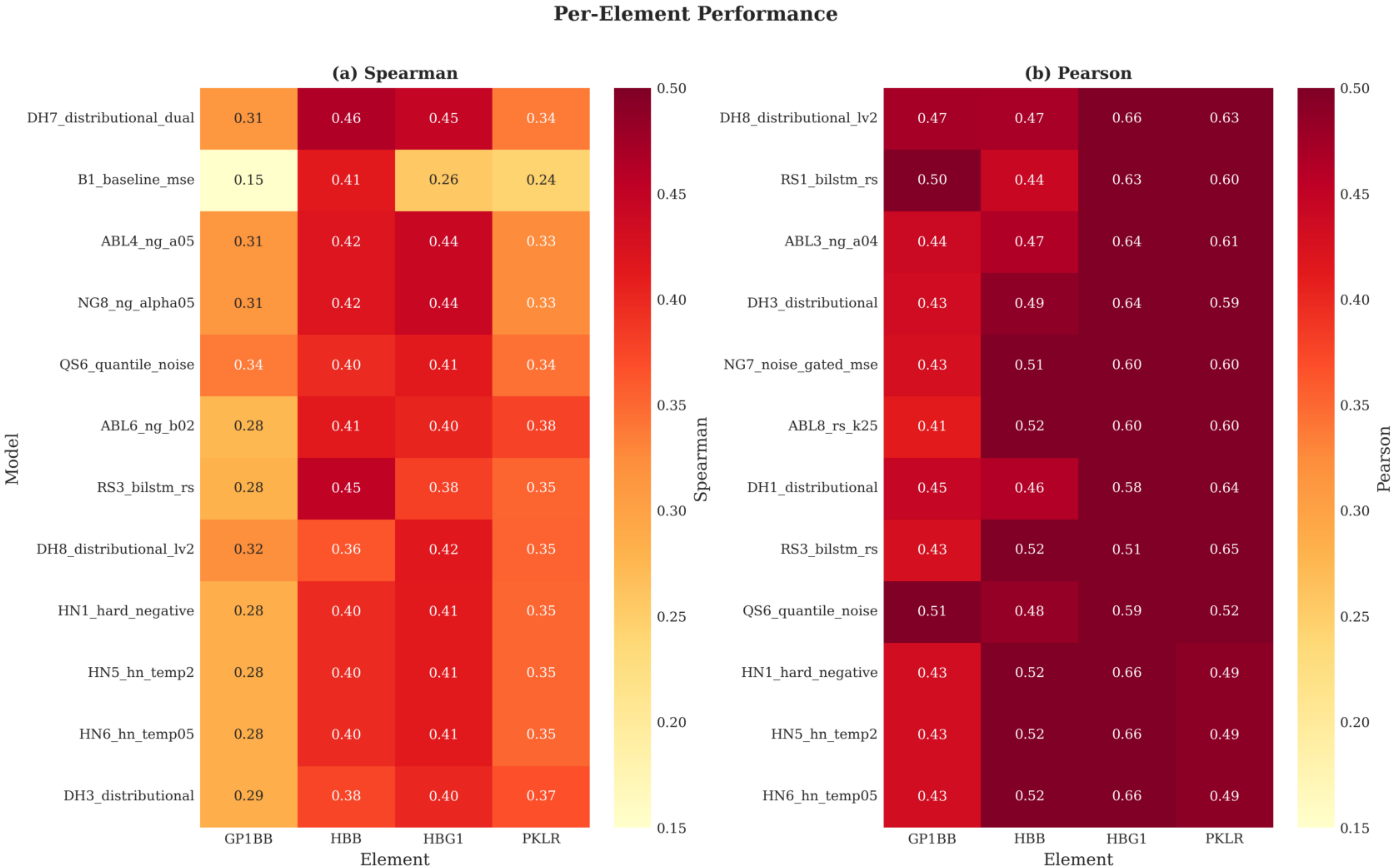


Figure 10. Performance on individual regulatory elements. HBB shows highest correlations across models. GP1BB shows largest improvement with noise-resistant methods (DH7 achieves 0.31 vs baseline 0.15).

Noise Correlation Analysis

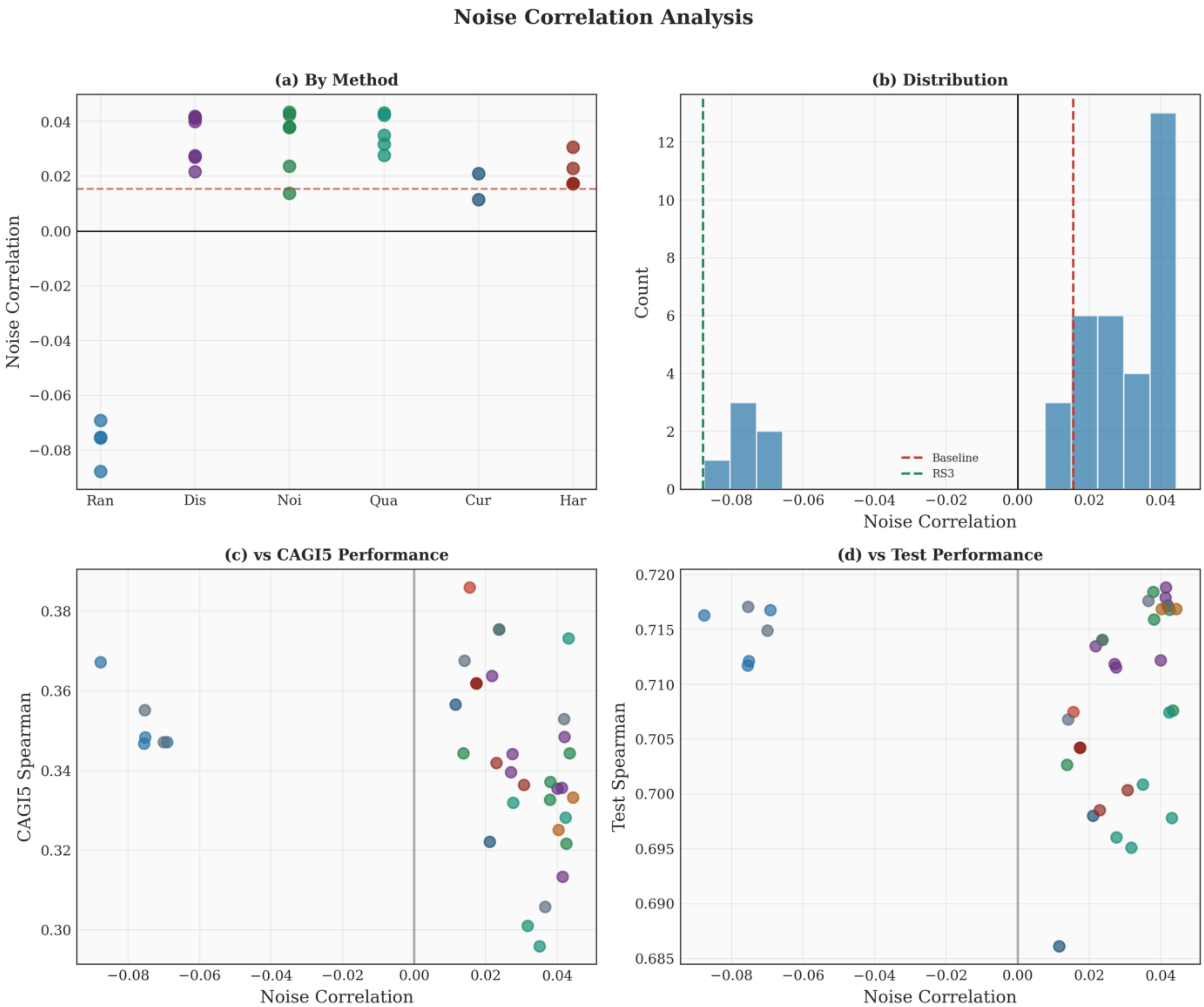


Figure 11. Residual-noise correlation measures whether prediction errors track with experimental noise (ideal=0). RS models achieve negative values, indicating they make smaller errors on noisy samples.

HC-LC Gap Analysis

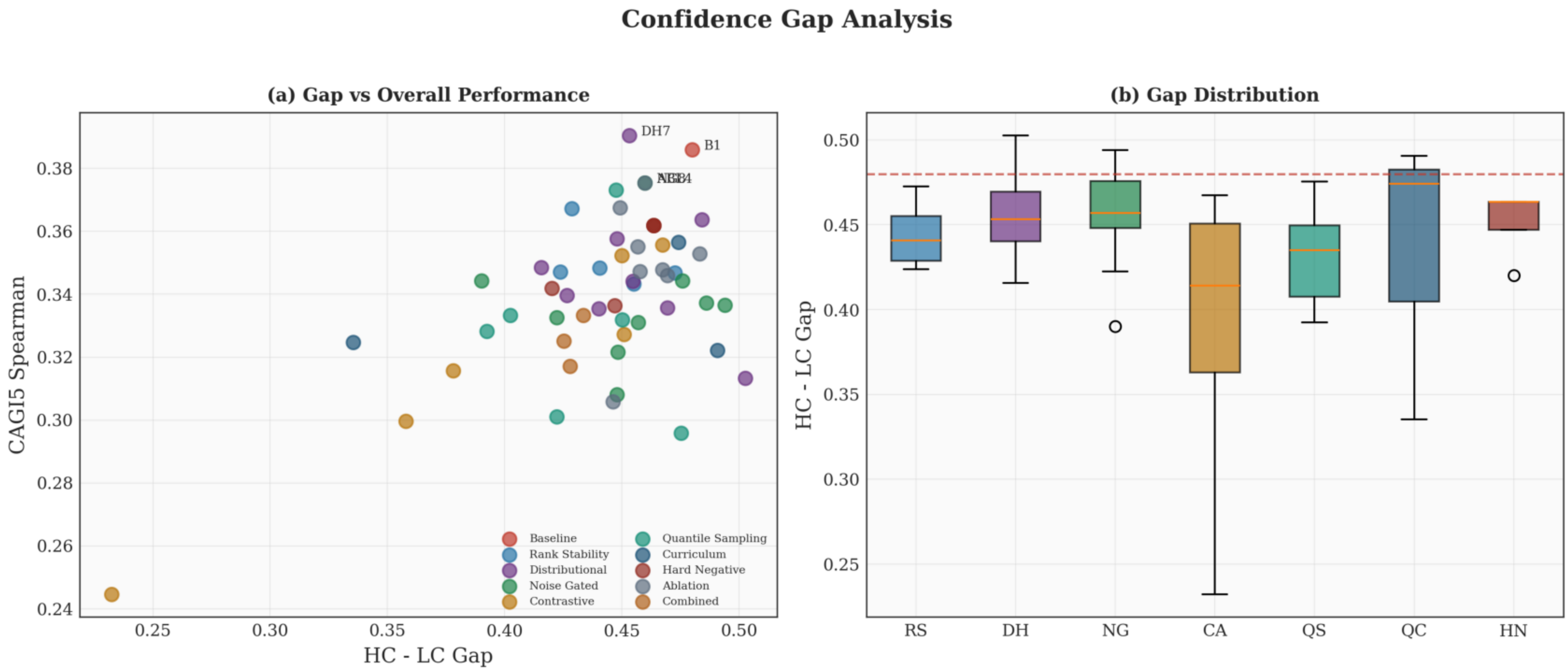


Figure 12. The gap between high-confidence and low-confidence performance indicates model balance. DH methods show smallest gaps among top performers.

Combined Methods (BEST)

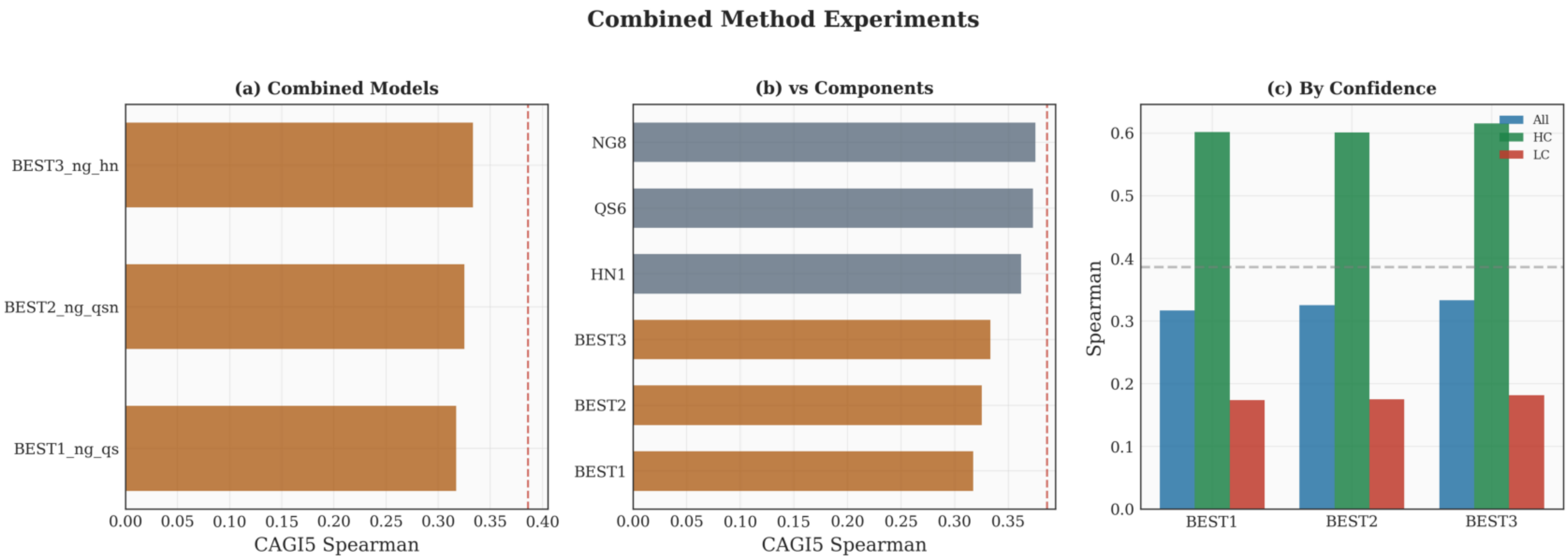
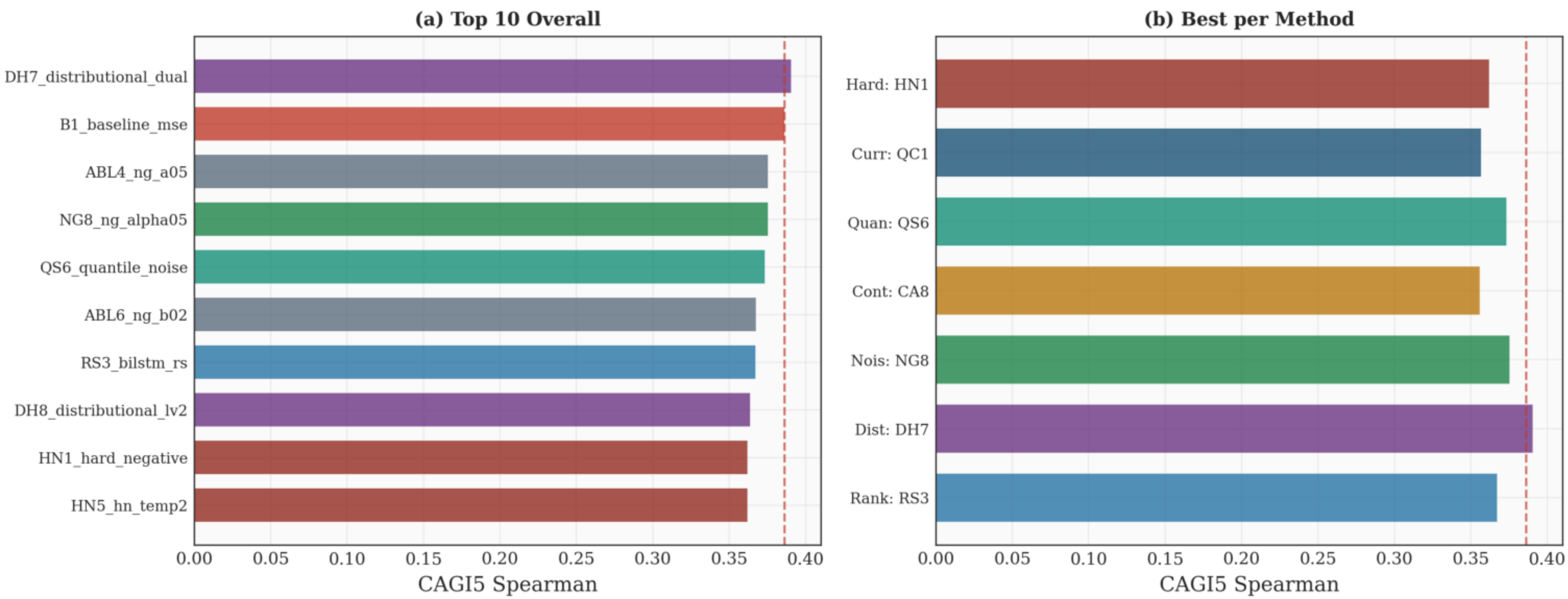


Figure 13. BEST models combine multiple strategies. Combinations do not consistently outperform the best single-method approaches.

Summary and Recommendations

Summary



(c) Key Results

Metric	Best	Value	Δ Baseline
CAGI5 Spearman	DH7	0.391	+1.3%
CAGI5 Pearson	DH8	0.557	+11.4%
LC Spearman	DH7	0.227	+9.7%
Noise Corr.	RS3	-0.088	Negative

(d) Recommendations

Recommendations:

- Best overall: DH7_distributional_dual
  Highest CAGI5 Spearman, balanced confidence
- Best noise resistance: RS3_bilstm_rs
  Only negative noise correlation
- Best Pearson: DH8_distributional_lv2
  Strongest linear relationship
- Robust choice: ABL4_ng_a05
  Good across all metrics

Figure 14. Overview of key findings. Distributional methods achieve best CAGI5, while rank stability provides best noise resistance. DH7 recommended for general use.