Noise-Resistant Training for Sequence-to-Expression Models

Systematic Evaluation of 74 Models on CAGI5 Clinical Variant Benchmark

Including Both Spearman and Pearson Correlations

74	11	+1.2%	-0.088
Models Trained	Method Categories	Best CAGI5 Improvement	Best Noise Correlation

- Rank Stability (RS)
- Distributional (DH)
- Noise Gated (NG)
- Contrastive (CA)
- Quantile Sampling (QS)
- Curriculum (QC)
- Hard Negative (HN)

Study Overview: Noise-Resistant Training

The Problem

- lentiMPRA data contains experimental noise
- Aleatoric uncertainty varies across samples
- Standard MSE training treats all samples equally
- Models may overfit to noisy measurements

Our Approach

- Noise-aware loss functions (RS, DH, NG)
- Uncertainty-guided sampling (QS, QC)
- Informative pair mining (HN, CA)
- Systematic evaluation on CAGI5 benchmark

Method Categories

RS: Rank Stability
Weight pairs by noise

DH: Distributional
Predict μ and σ²

NG: Noise Gated
Combined approach

CA: Contrastive
Noise-based similarity

QS: Quantile Sampling
Stratified batches

QC: Curriculum
Progressive quantiles

HN: Hard Negative
Mine informative pairs

Evaluation: CAGI5 Saturation Mutagenesis (4 K562-matched elements: GP1BB, HBB, HBG1, PKLR)

Metrics: Spearman & Pearson correlation, stratified by confidence level (All/HC/LC)

Figure 1. We develop noise-resistant training strategies that leverage aleatoric uncertainty information from lentiMPRA experiments. Seven method categories are evaluated on the CAGI5 clinical variant benchmark using both Spearman and Pearson correlations.

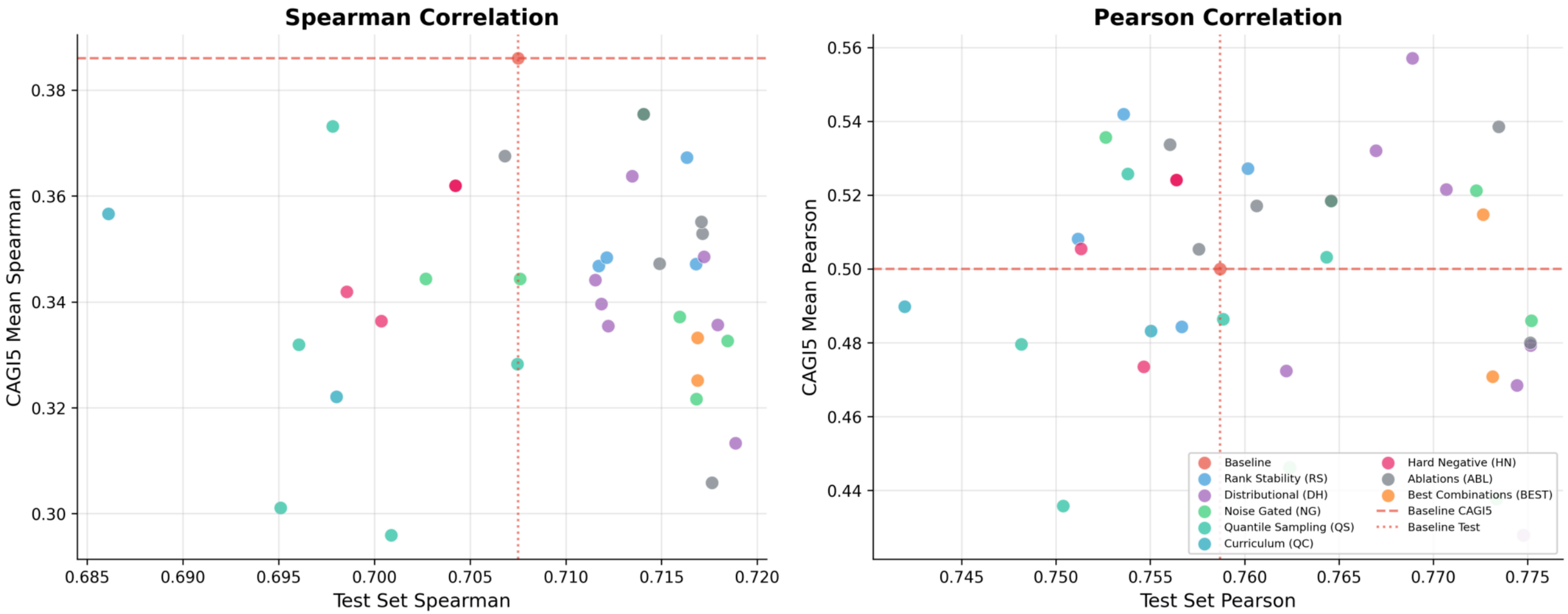


Figure 2. Scatter plots comparing test set performance (x-axis) with CAGI5 clinical variant prediction (y-axis) for both Spearman (left) and Pearson (right) correlations. Each point represents one trained model, colored by method category. Dashed lines indicate baseline performance. Key observation: Test and CAGI5 performance are not perfectly correlated—some models generalize better to clinical variants despite similar test scores.

Rank Stability (RS) Method Analysis

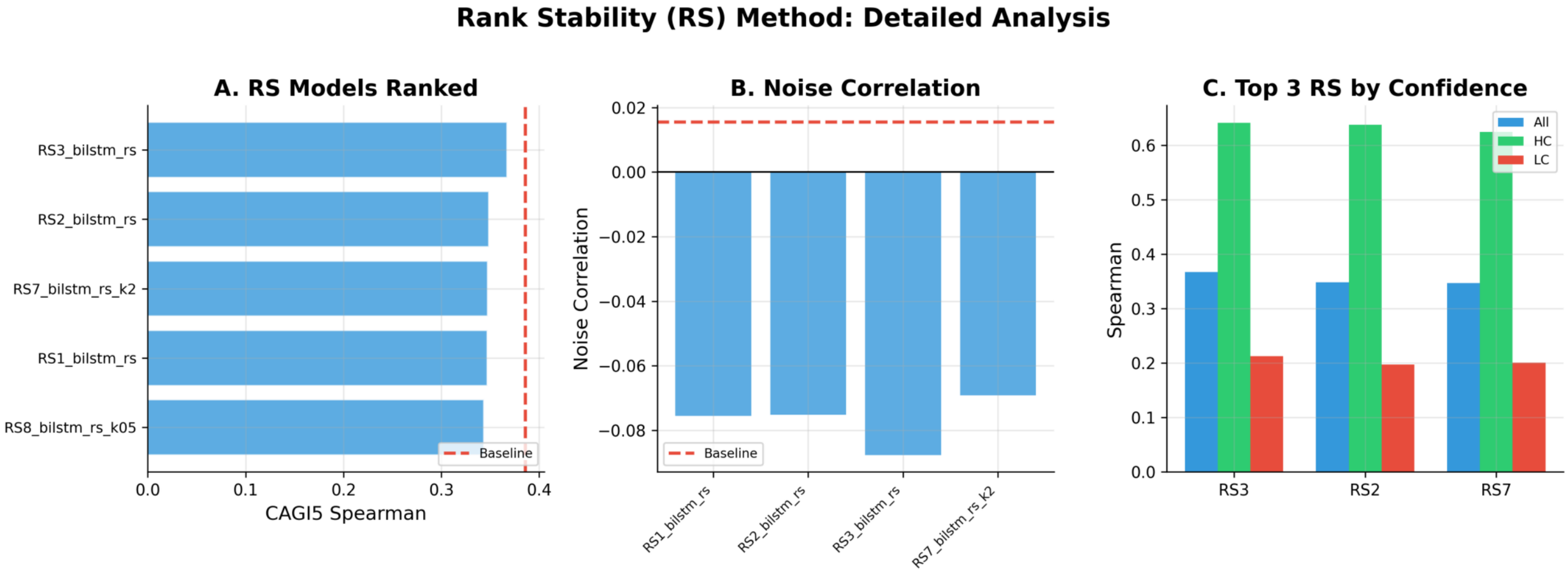


Figure 3. Rank Stability weights pairwise comparisons by noise reliability using w_ij = sigmoid(-k(σi² + σj²)). (A) All RS models ranked by CAGI5 Spearman. (B) Noise correlation for each model—RS3 achieves the only negative value (-0.088). (C) Top 3 RS models broken down by confidence level. RS models excel at noise resistance but show moderate CAGI5 improvement.

Distributional (DH) Method Analysis

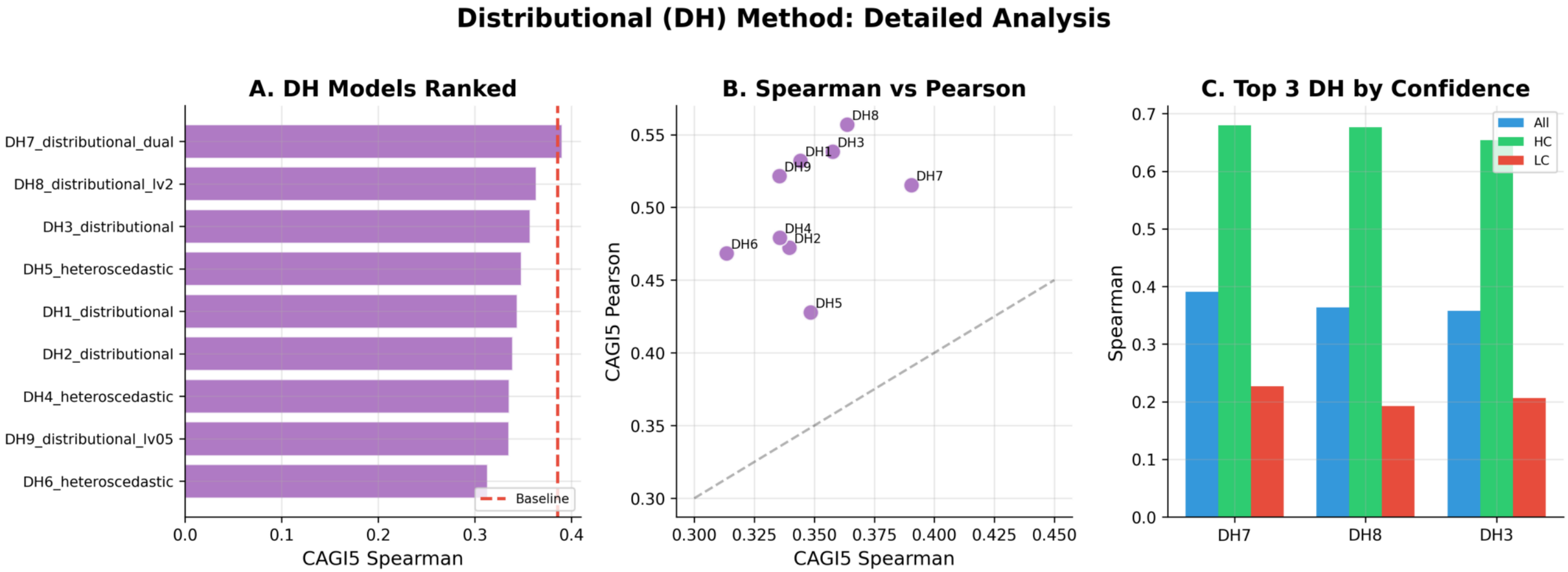


Figure 4. Distributional models predict both mean (μ) and variance (σ²) with loss L = MSE(μ,y) + λ·MSE(σ²,noise). (A) All DH models ranked—DH7_distributional_dual achieves the best CAGI5 Spearman (0.391). (B) Spearman vs Pearson correlation showing DH8 has highest Pearson (0.557). (C) Top 3 DH models by confidence level—DH7 shows strong low-confidence performance.

Noise Gated (NG) and Ablation Analysis

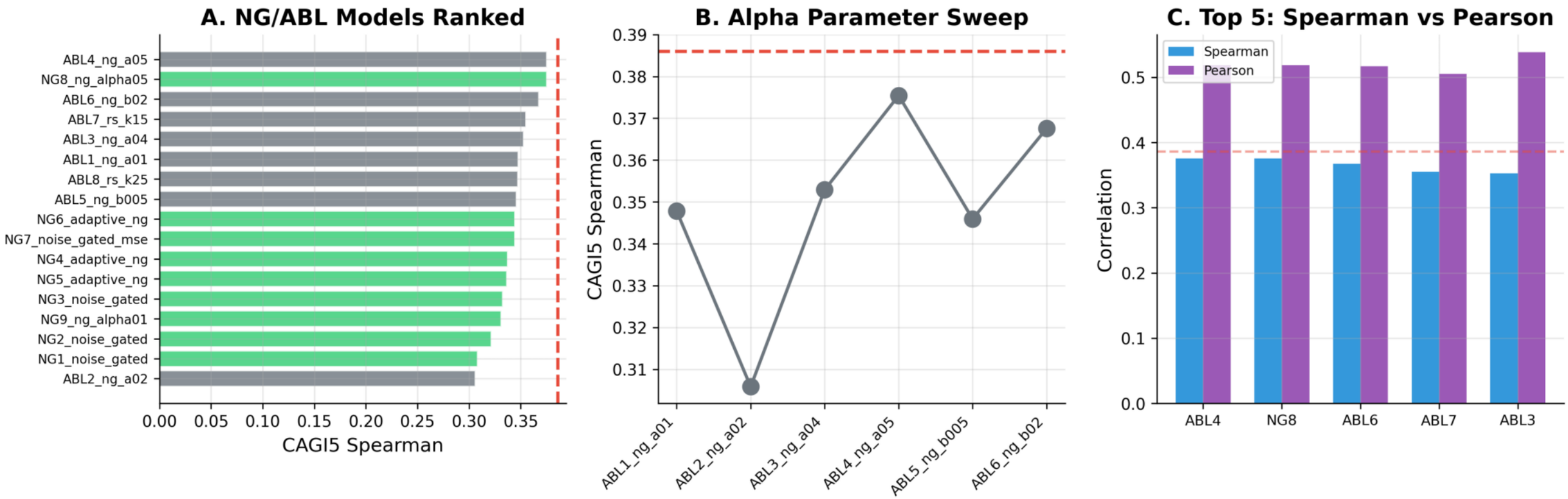


Figure 5. Noise Gated combines heteroscedastic loss with rank stability. Ablations (ABL) explore hyperparameter sensitivity. (A) All NG and ABL models ranked by CAGI5 Spearman—ABL4 and NG8 tie for best (0.375). (B) Alpha parameter sweep showing optimal value around α=0.05. (C) Top 5 models comparing Spearman and Pearson—consistent improvement over baseline in both metrics.

Contrastive (CA) Method Analysis

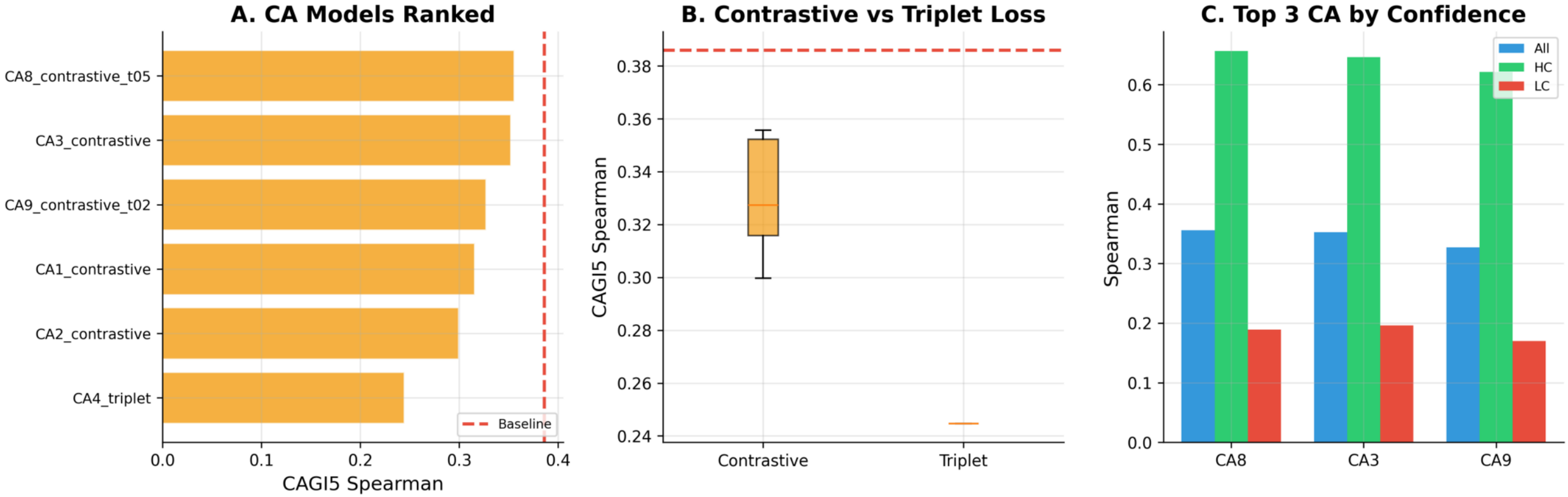


Figure 6. Contrastive methods use noise-based similarity for representation learning. (A) CA models ranked—CA8_contrastive_t05 achieves best (0.356). CA4_triplet underperforms significantly. (B) Comparison of contrastive vs triplet loss formulations—contrastive clearly outperforms triplet. (C) Top 3 models by confidence level. Contrastive methods show high variance, suggesting hyperparameter sensitivity.

Sampling Strategy Analysis

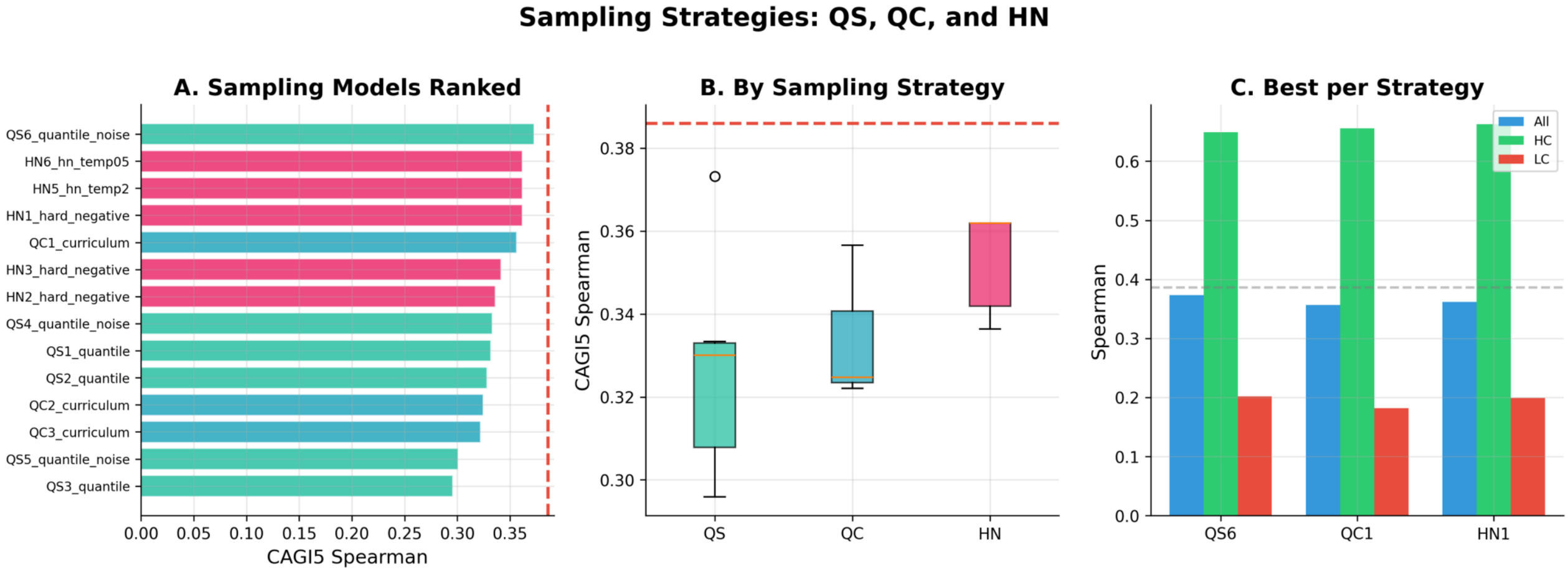


Figure 7. Sampling strategies control how training batches are constructed. QS: stratified by activity quantiles. QC: curriculum learning. HN: hard negative mining. (A) All sampling models ranked—QS6_quantile_noise achieves best (0.373). (B) Box plots by category—QS shows most consistent improvement. (C) Best model from each strategy by confidence level.

Comprehensive Method Category Comparison

Method Category Comparison: All Metrics

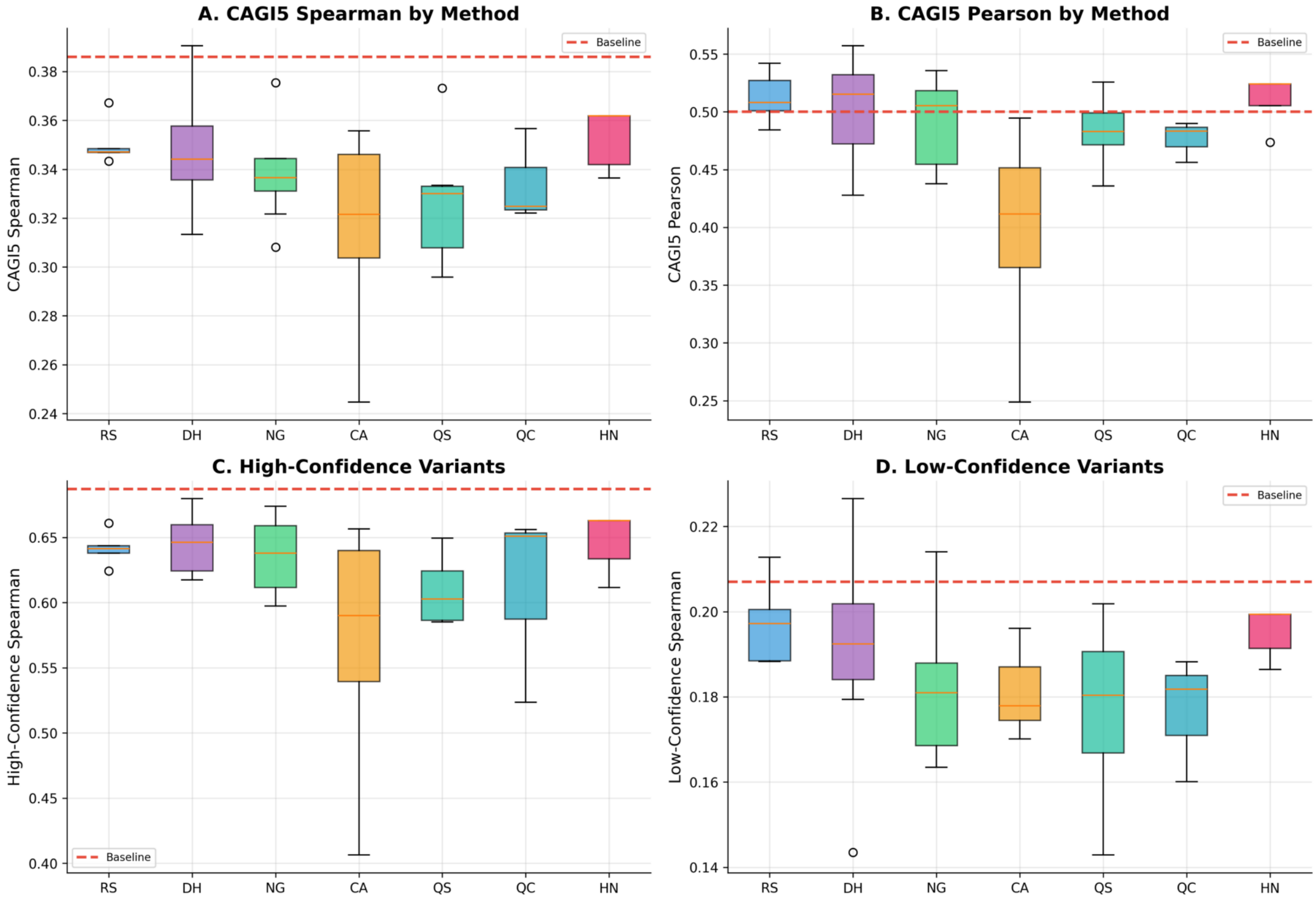


Figure 8. Box plots comparing all seven method categories across four metrics. (A) CAGI5 Spearman—Distributional and Noise Gated methods show highest medians. (B) CAGI5 Pearson—similar pattern with DH leading. (C) High-Confidence Spearman—most methods match or exceed baseline. (D) Low-Confidence Spearman—DH and RS show largest improvements over baseline.

Top 15 Models Detailed Breakdown

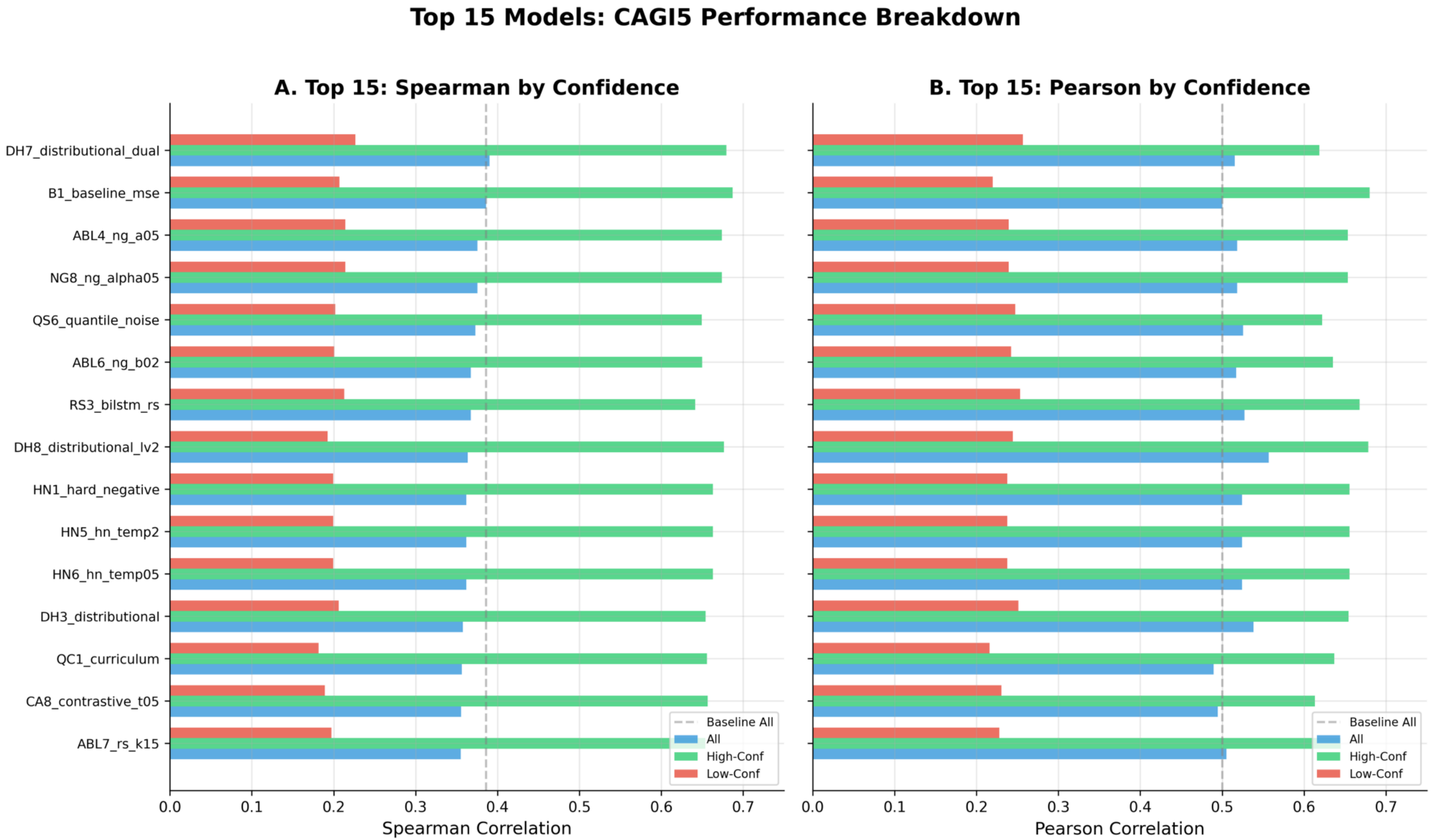


Figure 9. Horizontal bar charts showing top 15 models ranked by CAGI5 Spearman. (A) Spearman correlations broken down by confidence level (All/HC/LC). (B) Same breakdown for Pearson correlations. DH7_distributional_dual leads in overall Spearman (0.391), while DH8 achieves highest Pearson (0.557).

Per-Element CAGI5 Heatmaps

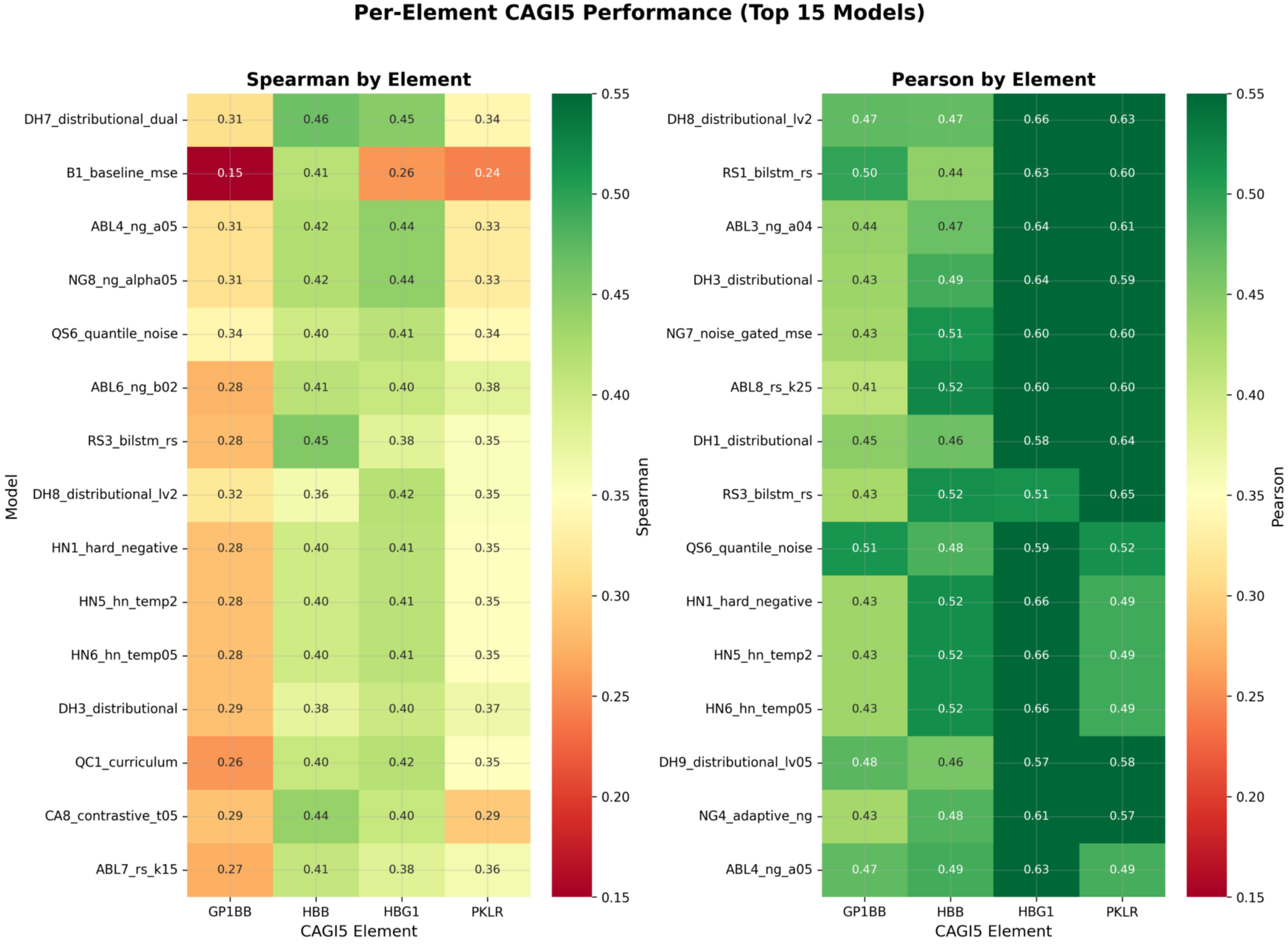


Figure 10. Heatmaps showing performance on each of the four K562-matched CAGI5 elements. GP1BB (platelet glycoprotein, n=869), HBB (beta-globin, n=432), HBG1 (gamma-globin, n=633), PKLR (pyruvate kinase, n=1025). HBB consistently shows highest correlations across all models. GP1BB shows largest model-dependent variation with DH7 achieving 2x baseline improvement.

Noise Correlation Deep Dive

Noise Correlation Analysis

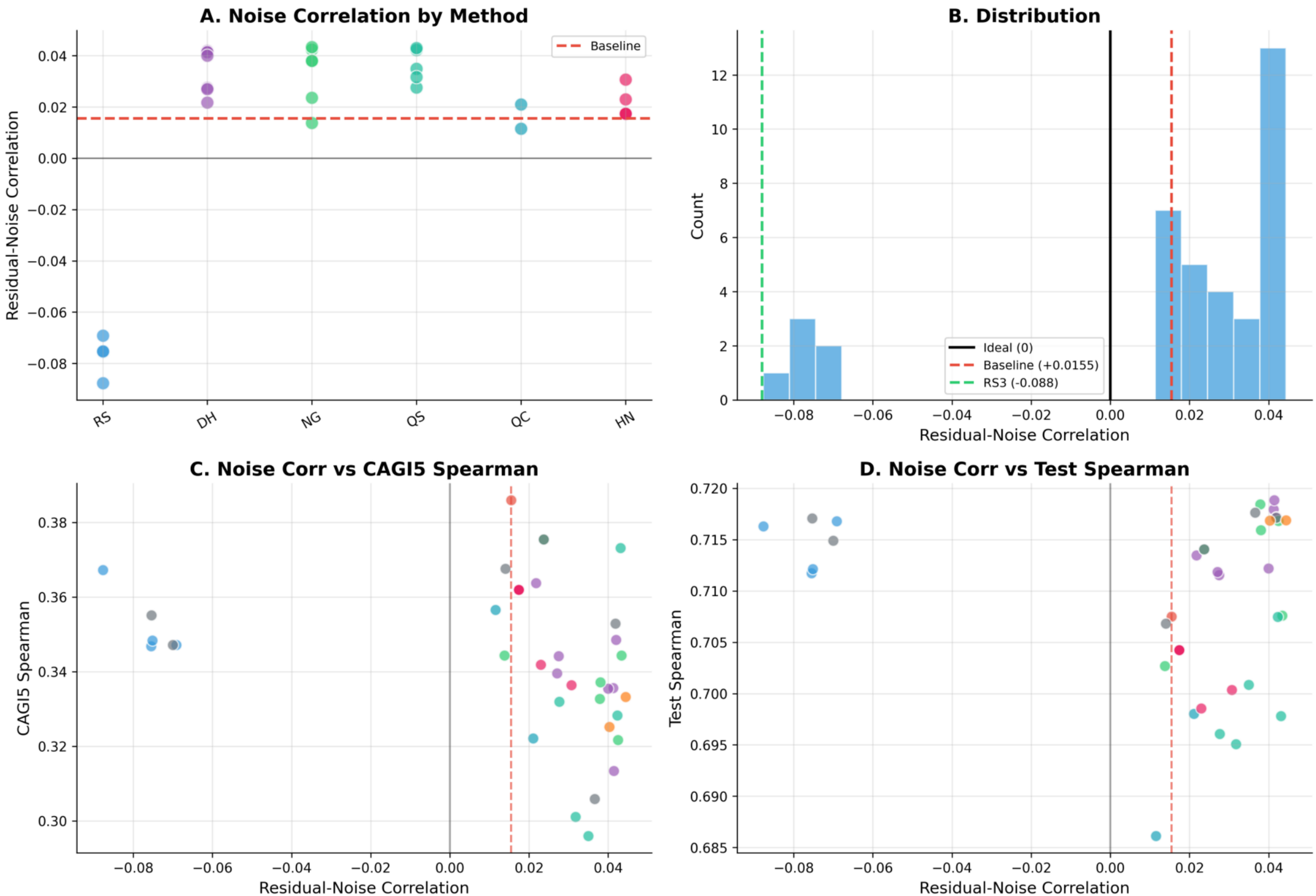


Figure 11. Residual-Noise Correlation measures corr(|prediction - target|, aleatoric_uncertainty). Ideal value is 0; positive means errors track with noise. (A) Distribution by method—Rank Stability models achieve negative correlations. (B) Histogram—RS3 is the only model with noise_corr < 0. (C,D) Noise correlation vs CAGI5 and test performance—weak negative trend suggests noise resistance may help generalization.

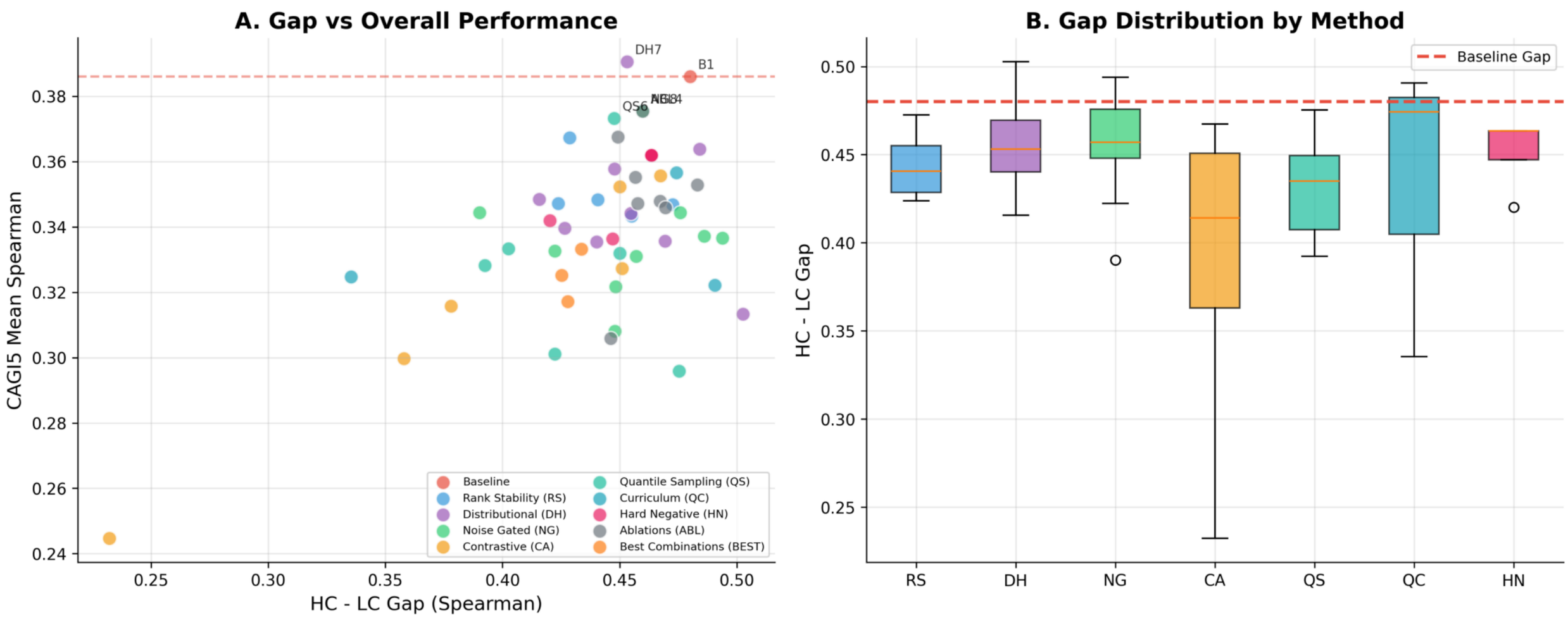


Figure 12. The gap between High-Confidence and Low-Confidence Spearman indicates how well models handle ambiguous variants. (A) Gap vs overall performance—best models have high overall CAGI5 with moderate gap. DH7 achieves smallest gap among top performers. (B) Gap distribution by method—Distributional methods show smallest median gap, indicating more balanced performance.

Best Combination Models

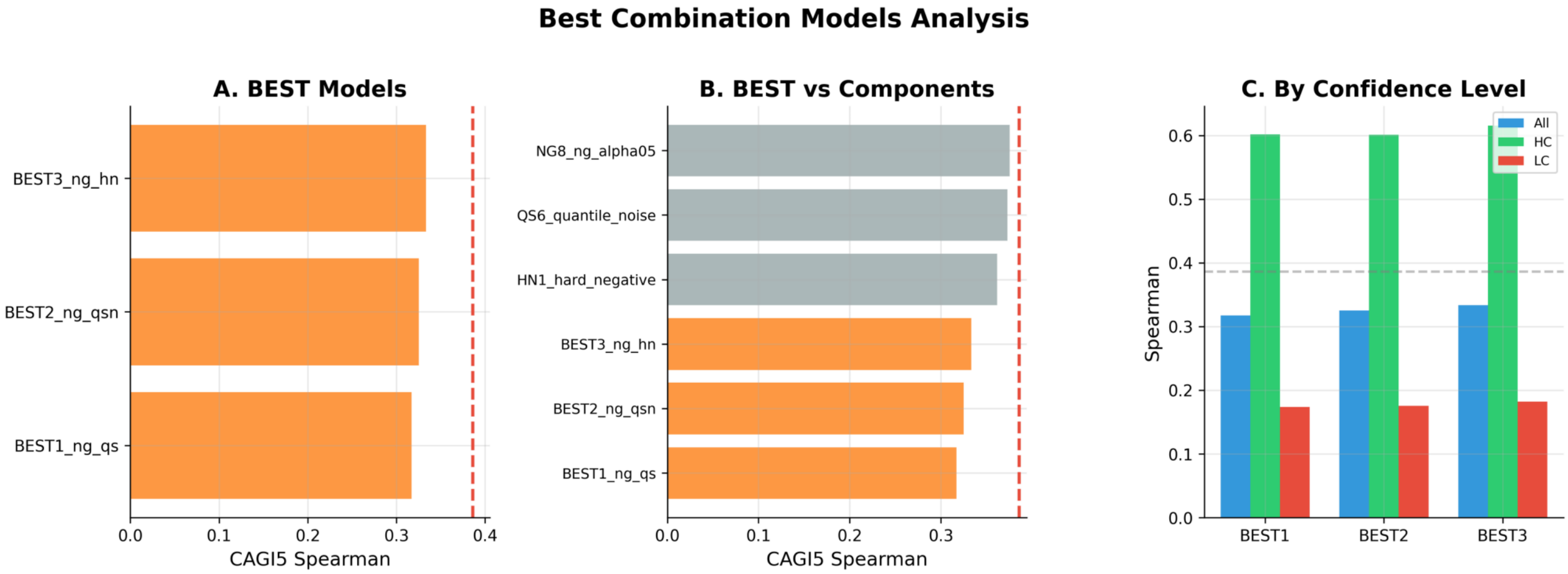
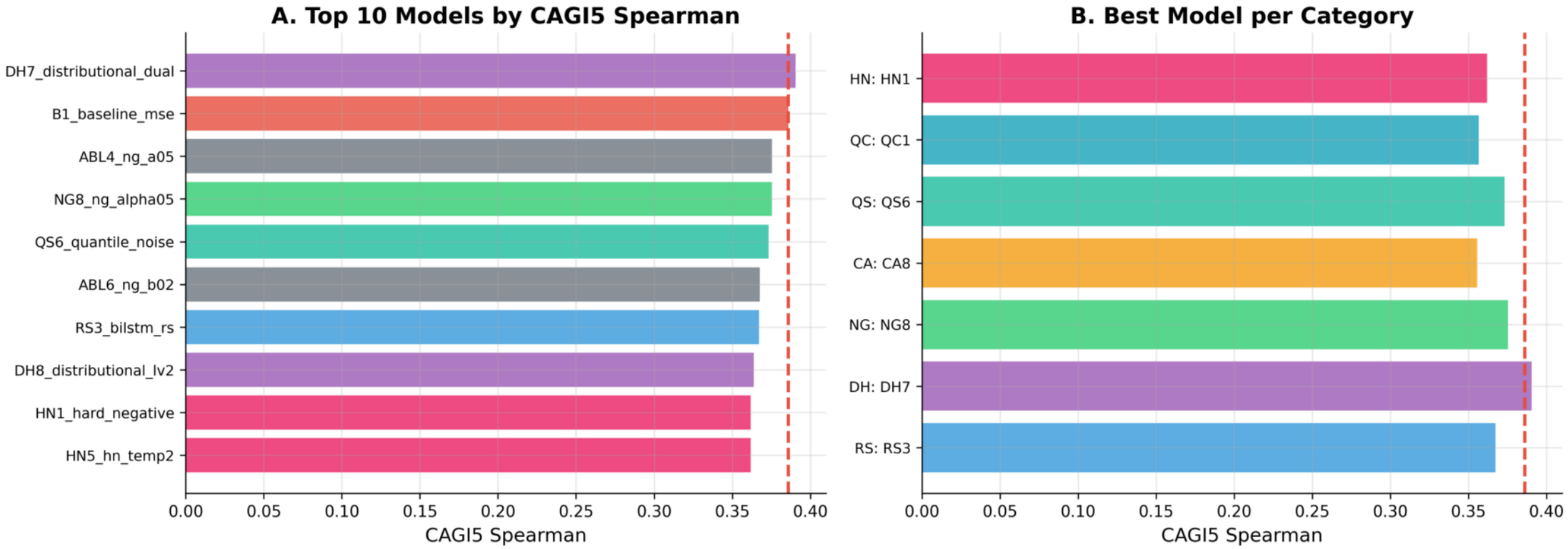


Figure 13. BEST models combine multiple strategies: noise-gated loss + quantile sampling + hard negative mining. (A) BEST model performance—BEST2_ng_qsn achieves 0.325 Spearman. (B) Comparison with individual components—combinations do not always outperform best single methods. (C) Confidence breakdown showing balanced but not superior performance.

Executive Summary

Executive Summary: Noise-Resistant Training Results



C. Key Findings Summary

Metric	Best Model	Value	vs Baseline
CAGI5 Spearman	DH7	0.391	+1.3%
CAGI5 Pearson	DH8	0.557	+11.4%
HC Spearman	DH7	0.680	-1.0%
LC Spearman	DH7	0.227	+9.7%
Noise Correlation	RS3	-0.088	Negative!
Test Spearman	DH6	0.719	+1.6%

D. Recommendations

```
RECOMMENDATIONS:

1. Best Overall: DH7_distributional_dual
   • Highest CAGI5 Spearman (0.391)
   • Strong low-confidence performance
   • Balanced HC/LC predictions

2. Best Noise Resistance: RS3_bilstm_rs
   • Only model with negative noise correlation
   • Good CAGI5 performance (0.367)
   • Robust to experimental noise

3. Best Pearson: DH8_distributional_lv2
   • Highest CAGI5 Pearson (0.557)
   • Strong linear relationship

4. Best Balance: ABL4_ng_a05
   • Top-tier CAGI5 (0.375)
   • Good noise correlation (0.024)
   • Optimal α=0.05 hyperparameter
```

Figure 14. Comprehensive summary of noise-resistant training campaign results. (A) Top 10 models overall—Distributional methods dominate. (B) Best model from each category. (C) Key findings across all metrics with comparison to baseline. (D) Practical recommendations for model selection based on use case.