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)