# Run Report | Iteration 43

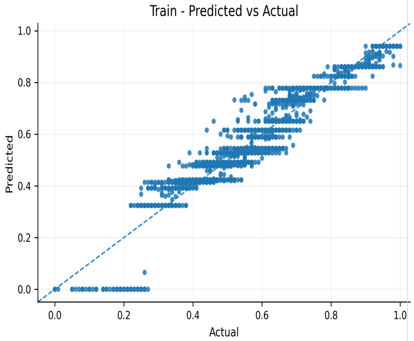
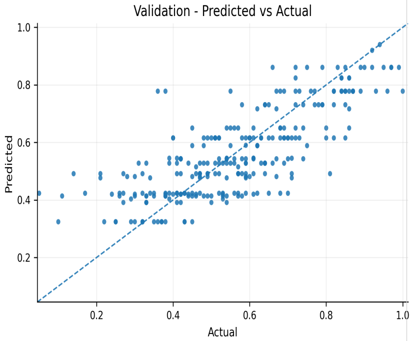
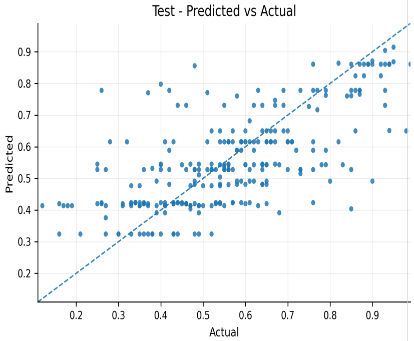
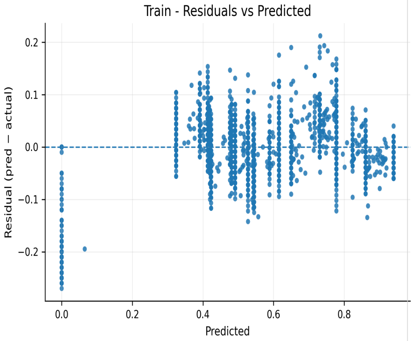
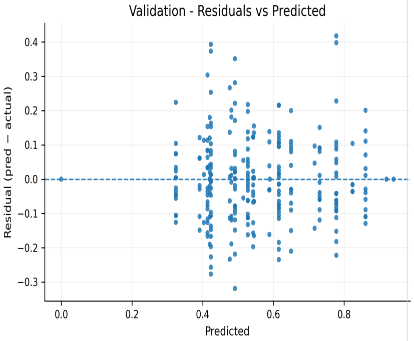
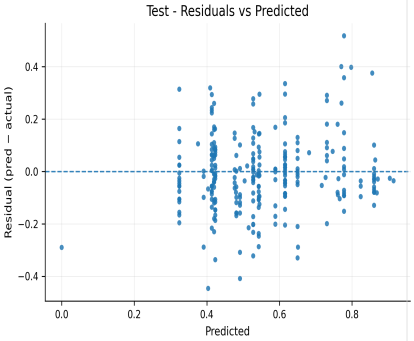
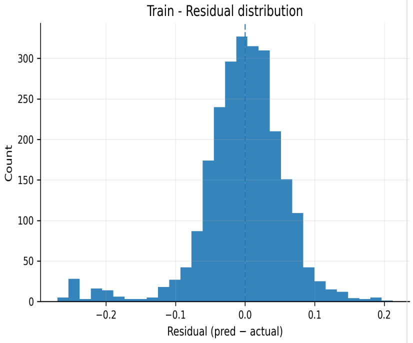
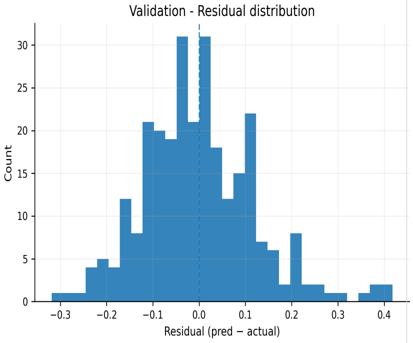
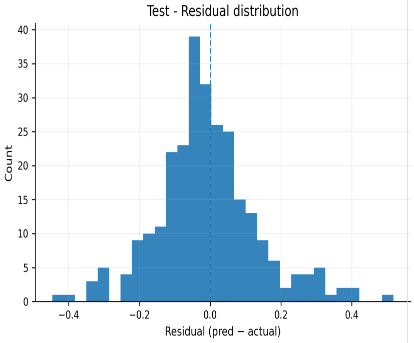
| | |
|---|---|
| **Agent ID** | talkative_building_crack |
| **Model** | openai/gpt-5.1-codex-max |
| **Dataset** | siRBench |
| **Task** | regression |
| **Optimized Metric** | R2 |
| **Split Allowed Iterations** | 0 |
| **Exploration Iterations** | 10 |

## Metrics

| Metric | Train | Validation | Test |
|---|---:|---:|---:|
| MAE | 0.0432422 | 0.0929838 | 0.105055 |
| MAPE | 0.109942 | 0.281047 | 0.356507 |
| MSE | 0.00376003 | 0.0146286 | 0.0200848 |
| PEARSON | 0.952953 | 0.782782 | 0.688225 |
| R2 | 0.898965 | 0.612747 | 0.461247 |
| RMSE | 0.0613191 | 0.120949 | 0.141721 |
| SPEARMAN | 0.974041 | 0.764223 | 0.676664 |

# Plots comparison (Iteration 43)

Columns are dataset splits (train / validation / test if exists). Rows correspond to the same plot type across splits.

| Train | Validation | Test |
|---|---|---|
|  |  |  |
|  |  |  |
|  |  |  |

## 1. Summary

- Preprocessed siRNA/mRNA by uppercasing, T→U, length 19 trunc/pad; built one-hot encodings, per-position pairing features (WC/ wobble/mismatch), seed and global interaction summaries, k■mer counts, numeric thermodynamic columns, and one■hot encodings for source/cell_line; saved encoder/schema artifacts for reproducible inference.

- Trained a two■model ensemble: XGBoostRegressor (DART, GPU, 2600 trees, depth 7, rate_drop 0.1) and LightGBMRegressor (GBDT, GPU, 7000 trees, depth 9, 640 leaves) with sample weights emphasizing deviations from 0.5; early stopping on validation.

- Averaged the two models' validation outputs and fitted an IsotonicRegression calibrator; inference reuses artifacts, averages model predictions, applies calibration, and clips to [0,1].

- Validation performance: RMSE ~0.121, MAE ~0.093, $R^2$ ~0.613, Pearson ~0.783 (train $R^2$ ~0.899), indicating moderate generalization with some overfit gap.

- Bias analysis: overestimates lowest quintile and underestimates higher true efficiencies (Q3–Q4); small cell line/source cohorts exhibit higher variance; calibration reduces but does not eliminate slope.

- Suggests future work on steeper mapping or loss adjustments for tails and addressing data scarcity in minor cohorts.

## 2. Data Split

**Train Path:**

datasets/siRBench/train.csv

**Val Path:**

datasets/siRBench/validation.csv

**Splitting Strategy:**

provided

## 3. Data Representation

Representation:

- *Inputs: columns from train.csv/validation.csv. Target: numeric_label.*

- *Sequence cleaning: siRNA and mRNA uppercased, T→U, truncate/pad to 19; invalid bases→N.*

- *One-hot: per-position A,C,G,U (order) for siRNA (19×4) then mRNA (19×4); N is all-zero.*

- *Interaction per-position (len 19 each): wc_match, wobble_GU, mismatch_other derived from aligned siRNA–mRNA pairs.*

- *Interaction summaries: total_wc, total_wobble, total_mismatch over 19; seed_wc and seed_wobble over positions 2–8 (1-indexed).*

- *K-mer counts: normalized mono (4) and di (16) counts for siRNA; same for mRNA. Mono divided by len (19); di by (len-1).*

*- Numeric thermodynamic/accessibility: all remaining numeric cols after dropping ['siRNA','mRNA','extended_mRNA','efficiency','numeric_label','id','source','cell_line'], preserving CSV order, cast float32.*

*- Categorical: OneHotEncoder(handle_unknown='ignore', sparse_output=False) on source and cell_line fitted on train; categories saved.*

*- Feature concat order: seq one-hot (siRNA then mRNA) → interaction per-position → interaction summaries → k-mer counts (siRNA then mRNA) → numeric thermodynamic → one-hot categorical. No scaling.*

*- Artifacts saved: feature_artifacts.json with encoder categories, numeric column list, feature names for inference alignment.*

**Files created:**

- feature_builder.py

## 4. Model Architecture

Architecture:

Ensemble: XGBoostRegressor (booster='dart') + LightGBMRegressor (GBDT) with isotonic calibration on averaged outputs

Hyperparameters:

*XGB: n_estimators=2600, learning_rate=0.028, max_depth=7, min_child_weight=1.2, subsample=0.9, colsample_bytree=0.9, gamma=0.0, reg_alpha=0.0, reg_lambda=1.0, booster='dart', sample_type='uniform', normalize_type='tree', rate_drop=0.1, skip_drop=0.5, objective='reg:squarederror', eval_metric='rmse', tree_method='gpu_hist' (fallback 'hist'), predictor='gpu_predictor' (fallback 'auto'), early_stopping_rounds=250.*

*LightGBM: objective='regression', metric='rmse', boosting_type='gbdt', device='gpu' else 'cpu', n_estimators=7000, learning_rate=0.01, max_depth=9, num_leaves=640, min_child_samples=10, min_child_weight=1e-3, subsample=0.85, subsample_freq=1, colsample_bytree=0.85, reg_alpha=0.0, reg_lambda=0.7, early_stopping_rounds=500.*

*Ensemble: simple average of model outputs followed by IsotonicRegression(out_of_bounds='clip'). Sample weights: w = 1 + 3.5 * abs(y - 0.5).*

## 5. Model Training

### Path To Train File:

```
run_files/iteration_43/train.py
```

### Path To Model File:

```
run_files/iteration_43/training_artifacts/xgb_model.json
```

### Path To Artifacts Dir:

```
run_files/iteration_43/training_artifacts
```

**Training Summary:**

> *Built augmented sequence+interaction+thermo+categorical features with encoder artifacts saved. Trained XGBoost DART (GPU fallback) and LightGBM GBDT (GPU) with sample-weighting, averaged their validation predictions, and fit an isotonic calibrator. Saved models, feature artifacts, calibration, metrics, and bias analyses.*

**Files created:**

- training_artifacts
- __pycache__
- train.py
- feature_builder.py

## 6. Model Inference

**Path To Inference File:**

```
run_files/iteration_43/inference.py
```

**Inference Summary:**

> *Inference loads artifacts (feature schema, encoder, XGB DART, LightGBM, isotonic calibrator), rebuilds the augmented sequence/interaction/thermo/categorical features, predicts with both models (GPU preferred), averages, calibrates, clips to [0,1], and writes CSV preserving id when present.*

**Files created:**

- __pycache__
- inference.py
- dry_run_metrics.txt

## 7. Prediction Exploration

Statistics:

> *Validation metrics: RMSE 0.12095, MAE 0.09298, R2 0.61275, Pearson 0.78278, Spearman 0.76422. Quintile biases (true_mean → pred_mean, bias, RMSE, count): Q0 0.2925→0.4156 (+0.1232, 0.1707, n=57); Q1 0.4485→0.4807 (+0.0321, 0.0885, n=55); Q2 0.5553→0.5323 (-0.0229, 0.0907, n=59); Q3 0.6632→0.6068 (-0.0564, 0.1211, n=56); Q4 0.8447→0.7641 (-0.0806, 0.1147, n=53). Cell-line biases: h1299 n=213 true_mean 0.5131 pred_mean 0.5075 bias -0.0056 RMSE 0.1089; hek293 n=19 bias +0.0351 RMSE 0.1400; hek293t n=6 bias +0.0669 RMSE 0.2052; hep3b n=30 bias +0.0076 RMSE 0.1483; t24 n=8 bias -0.0075 RMSE 0.1785; halacat n=4 bias -0.0107 RMSE 0.0871. Source biases mirror cell lines; notable outlier khvorova (n=1) bias +0.3732 RMSE 0.3732.*

Insights:

*Strong performance overall but systematic underestimation at higher true efficiencies (Q3–Q4) and overestimation at the lowest quintile; calibration still leaves residual slope. Seed/higher activity regimes likely need steeper mapping; consider feature-target interactions or alternative loss to reduce tails. Small cohorts (hek293t, t24, khvorova) show larger errors; variance likely due to data scarcity rather than model bias.*

### Files created:

- validation_quintile_bias.csv
- val_predictions.csv
- validation_cell_bias.csv
- validation_with_predictions.csv
- validation_source_bias.csv
- analyze_predictions.py
- dry_run_metrics.txt
- validation_metrics.txt