PREDICTIVE FEASIBILITY

PFA

Model Selection & Robustness Assessment

Introduction

The purpose of this assessment series is to determine whether inferability-related structural metrics can provide practical guidance for model selection, deployment robustness, and model-family suitability before full model development begins.

Earlier validation studies established that inferability, entropy, overlap, persistence, and transition dynamics represent measurable structural properties of real-world signals. Subsequent forecasting studies demonstrated that these properties can support transition forecasting, deployment-oriented warning systems, and predictive feasibility assessment.

The next logical question is whether the same structural information can also be used to evaluate model behavior itself.

This assessment therefore investigates:

- model correspondence between signal structure and expected model instability;
- model-family sensitivity to inferability-related signal regimes;
- cross-dataset model-ranking predictability;
- deployment robustness under transfer conditions;
- model-family benchmarking across structural regimes;
- and pre-deployment model-selection guidance.

The objective is not to identify a universally superior model architecture.

Instead, the goal is to determine whether structural signal properties provide actionable information regarding:

- model suitability;
- deployment stability;
- transfer robustness;
- expected model degradation;
- and operational AI risk.

Together, the studies presented in this section form the model-selection and deployment-robustness layer of the Predictive Feasibility Assessment (PFA) framework.

Cross-Dataset Model Ranking Validation

Predictive Feasibility Transfer Across Independent fastSPT Datasets

Purpose of the Test

This validation extends the model-ranking predictability framework from a single dataset setting toward a cross-dataset setting.

The central question is no longer only:

“Can inferability-related metrics predict which model family performs best within one dataset?”

but rather:

“Can inferability-related metrics support model-ranking prediction when training and testing are performed across independent fastSPT datasets?”

This is a stronger validation layer because it evaluates whether the relationship between signal structure and model behavior remains visible under dataset transfer.

The test therefore examines whether:

model-ranking structure remains reproducible across independent datasets,
inferability, entropy, and overlap still relate to model error under transfer,
model-family dominance can be estimated before deployment,
and predictive feasibility can support cross-dataset model-selection decisions.

Experimental Setup

The validation used real fastSPT trajectory windows derived from independent dataset conditions.

The cross-dataset transfer setting included:

dataset_A_slowSPT_2Hz
dataset_B_spaSPT_95Hz
The transfer directions were:

dataset_A_slowSPT_2Hz -> dataset_B_spaSPT_95Hz
dataset_B_spaSPT_95Hz -> dataset_A_slowSPT_2Hz
The analysis was performed on real trajectory-derived windows and not on synthetic examples.

Number of trajectory windows used: 7948

Number of cross-dataset model-ranking result rows: 72

Number of model-ranking prediction rows: 18

Overall model-ranking prediction accuracy: 0.611

Window-Level Features

For each trajectory window, the same structural feature logic was used as in the earlier model-ranking validation:

inferability score
entropy proxy
overlap proxy
persistence proxy
straightness
future displacement target
regime labels for inferability, entropy, and overlap

The purpose was to evaluate whether the structural regime of the signal could still explain model behavior after crossing from one dataset condition to another.

Model Families Tested

The cross-dataset benchmark evaluated multiple model families:

linear
ridge
random forest
MLP-light

For each model family and each structural regime, the following metrics were computed:

Mean Squared Error (MSE)
Mean Absolute Error (MAE)
model rank by MSE
model rank by MAE
best model by MSE
best model by MAE

Regime Structure

The dataset was evaluated across multiple regime types:

inferability regimes
entropy regimes
overlap regimes

Each regime type was divided into low, medium, and high structural regions.

This allows the analysis to determine whether model behavior changes systematically as the underlying signal structure changes.

Reproducible Execution

Generated CSV files:

cross_dataset_model_ranking_results.csv
cross_dataset_model_ranking_predictions.csv
cross_dataset_model_ranking_windows.csv

Generated figure files:

cross_dataset_mse_by_model.png

cross_dataset_mae_by_model.png
cross_dataset_inferability_vs_mse.png
cross_dataset_entropy_vs_mse.png
cross_dataset_overlap_vs_mse.png
cross_dataset_model_ranking_accuracy_by_regime.png

Results

Figure 1 - Cross-Dataset MSE by Model

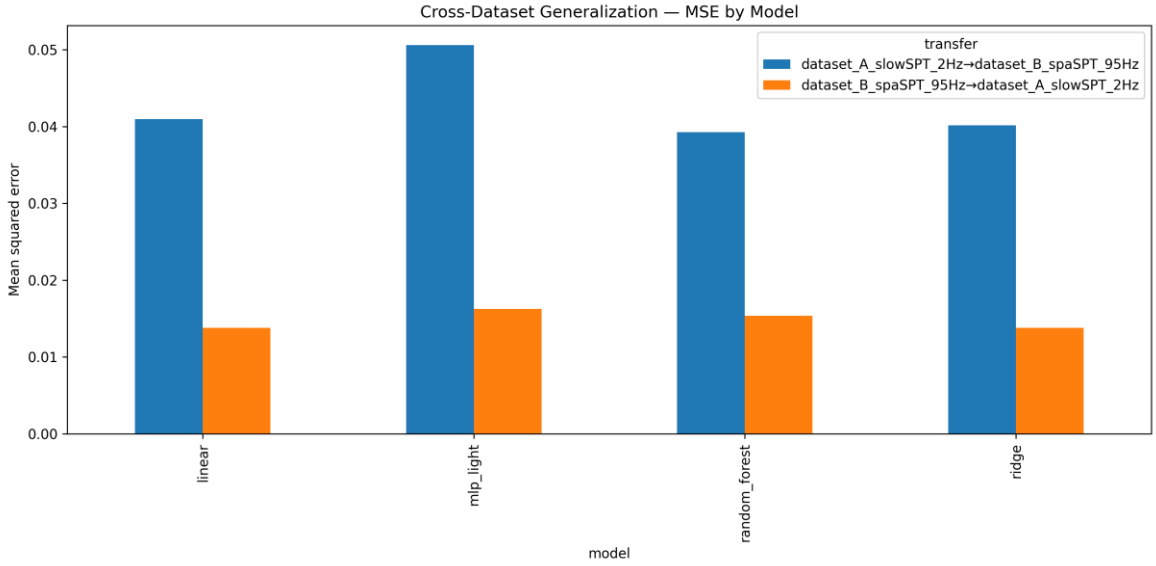


Figure 1 - Cross-Dataset MSE by Model.

This figure shows the mean squared prediction error for each model family under cross-dataset transfer.

The figure is important because it reveals whether model-family performance remains stable when the training and testing datasets differ.

Key interpretation:

model families do not respond identically under transfer,
cross-dataset prediction error remains structured rather than random,
model selection should therefore not be based only on generic model popularity.
Caption

Cross-dataset MSE varies by model family, indicating that model stability depends on the interaction between signal regime and model type.

Figure 2 - Cross-Dataset MAE by Model

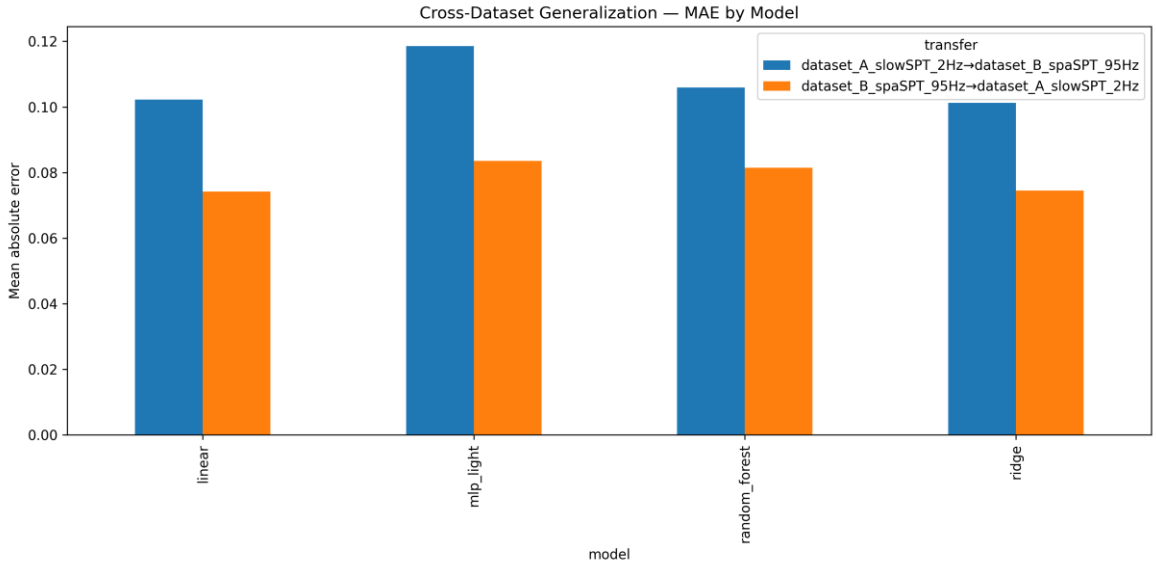


Figure 2 - Cross-Dataset MAE by Model.

This figure shows the mean absolute prediction error for each model family.

The MAE view is used as a robustness check against the MSE pattern.

Important:

the MAE structure confirms that transfer behavior remains model-dependent,
absolute error behavior is not uniform across model families,
and cross-dataset model ranking remains a meaningful target.
Caption

Cross-dataset MAE confirms that absolute prediction stability differs between model families under transfer.

Figure 3 - Mean Inferability vs Model MSE

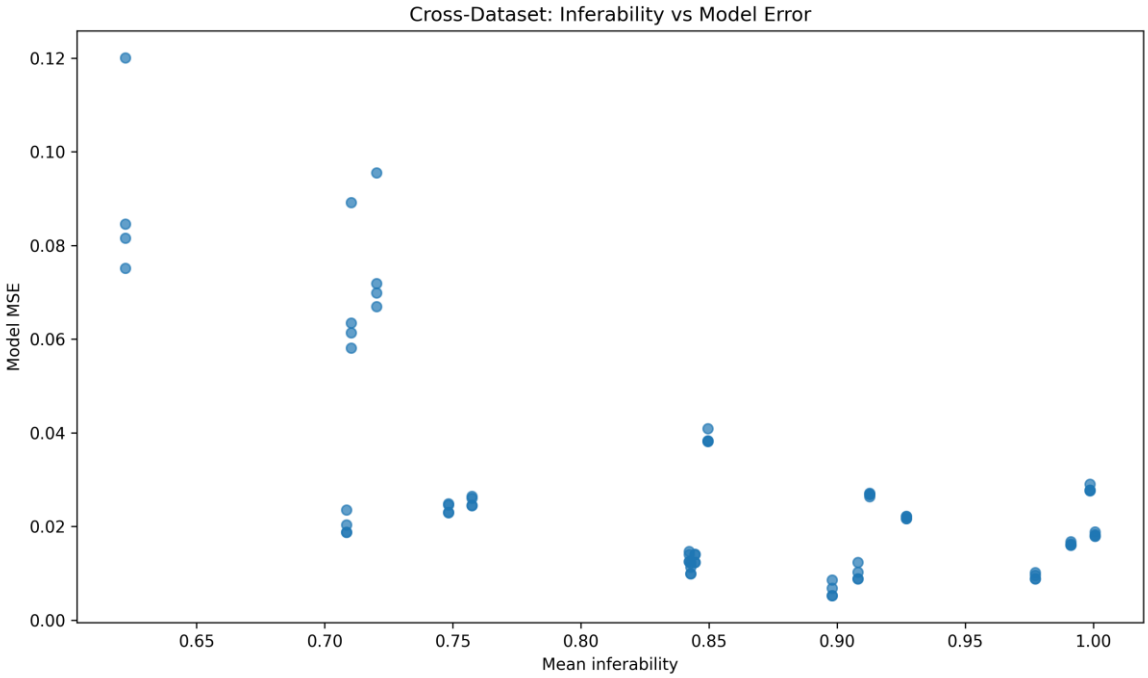


Figure 3 - Mean Inferability vs Model MSE.

This figure evaluates whether mean inferability remains connected to model error under cross-dataset transfer.

The core expectation is that higher inferability should correspond to lower model error or more stable model behavior.

This result extends the earlier within-dataset model-ranking test toward transfer conditions.

Caption

Mean inferability remains linked to model error under cross-dataset transfer, supporting inferability as a pre-model feasibility indicator.

Figure 4 - Mean Entropy vs Model MSE

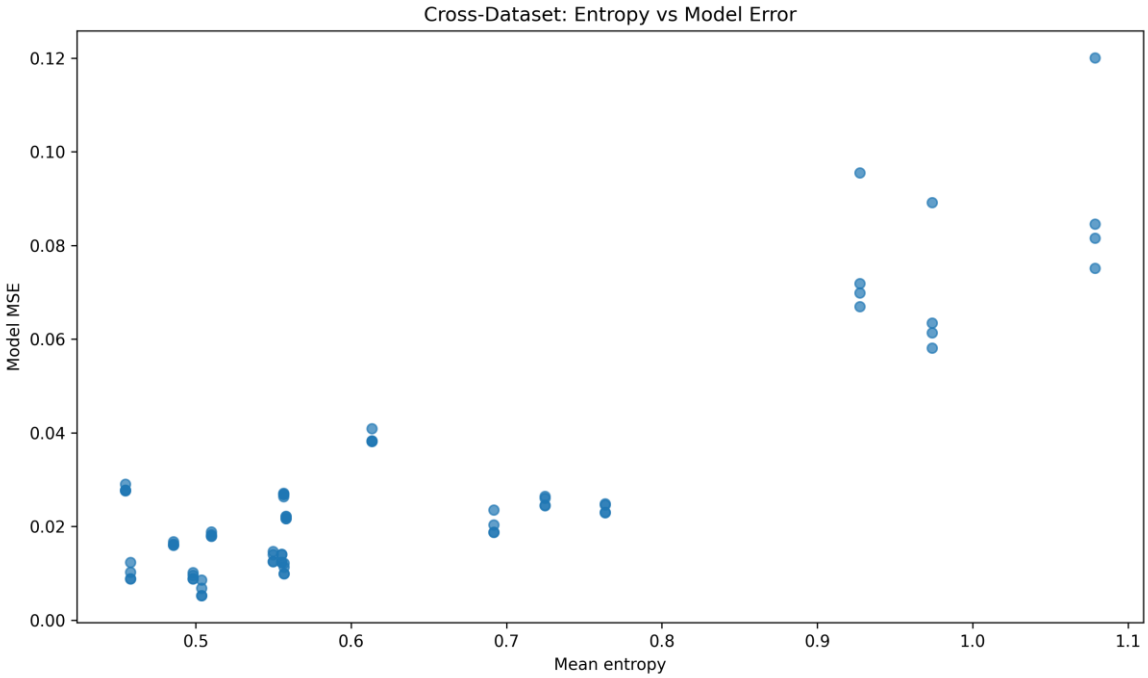


Figure 4 - Mean Entropy vs Model MSE.

This figure evaluates whether entropy remains a destabilizing factor under cross-dataset transfer.

Higher entropy indicates greater structural disorder in the trajectory window.

The expected pattern is:

higher entropy produces higher model error,
more chaotic regimes reduce deployment stability,
entropy remains relevant beyond a single dataset.
Caption

Trajectory entropy remains associated with model degradation under cross-dataset transfer.

Figure 5 - Mean Overlap vs Model MSE

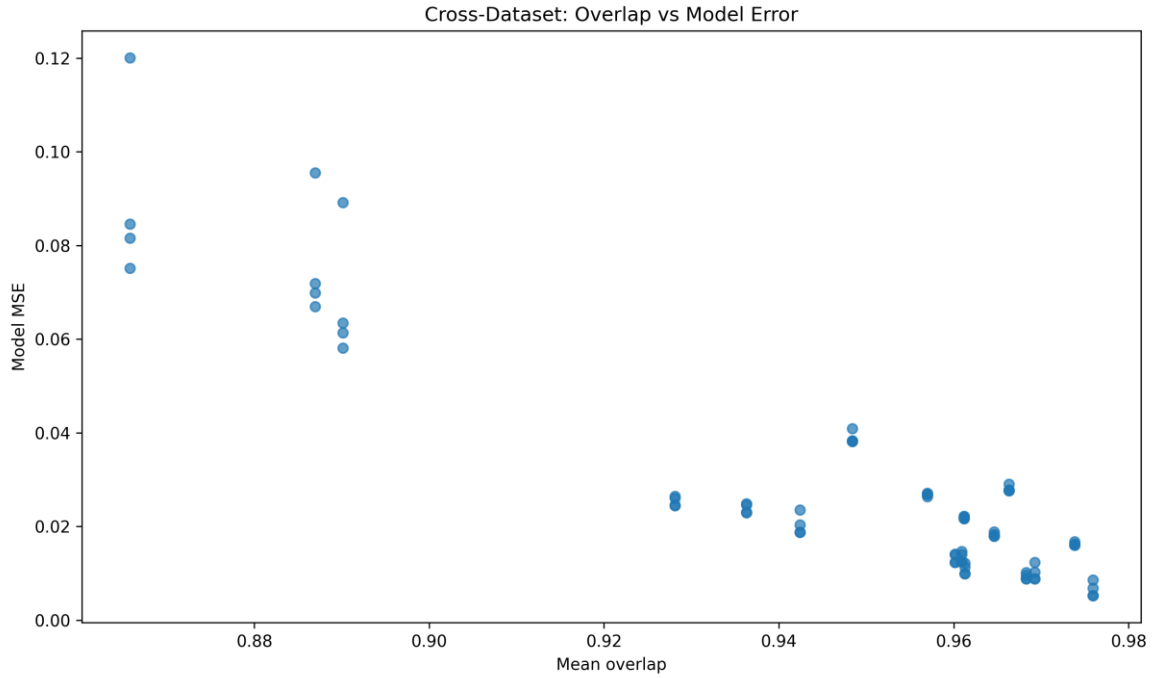


Figure 5 - Mean Overlap vs Model MSE.

This figure evaluates overlap as a cross-dataset stability indicator.

Overlap is interpreted as a measure of local structural reproducibility.

The important hypothesis is that higher overlap should support lower error and more stable model behavior.

Caption

Higher overlap regimes show stronger model stability under cross-dataset transfer, supporting overlap as a structural reproducibility metric.

Figure 6 - Model Ranking Accuracy by Regime

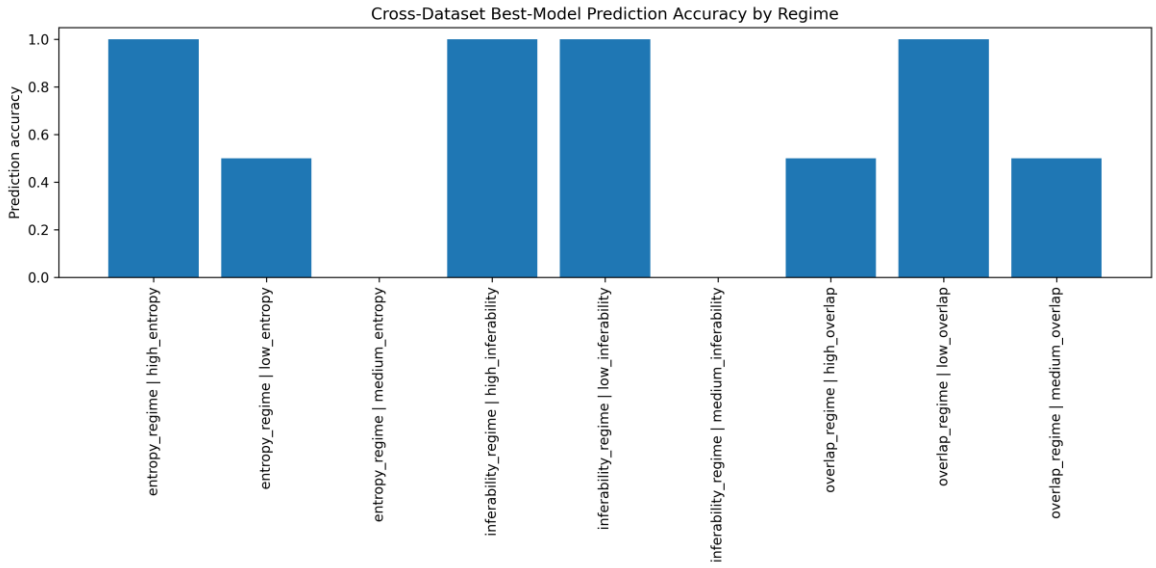


Figure 6 - Model Ranking Accuracy by Regime.

This figure is the central model-selection result of the cross-dataset validation.

It evaluates whether the framework can correctly predict the best-performing model across structural regimes.

The prediction table produced:

overall model-ranking prediction accuracy = 0.611

This indicates that model ranking contains reproducible structure across datasets, although it is not yet perfect.

Caption

Model-ranking prediction accuracy varies by regime, showing that some structural regimes provide stronger pre-model selection information than others.

Quantitative Summary

Cross-dataset model-ranking prediction summary:

Quantity	Value
Trajectory windows	7948
Result rows	72
Prediction rows	18
Overall prediction accuracy	0.611
Training datasets	dataset_A_slowSPT_2Hz / dataset_B_spaSPT_95Hz
Transfer directions	A -> B and B -> A

Scientific Interpretation

This validation strengthens the predictive-feasibility framework by showing that model-ranking structure does not disappear immediately when the dataset changes.

The results suggest that:

inferability-related structure transfers partially across datasets,
model-family performance remains regime-dependent,
entropy and overlap continue to provide deployment-relevant information,
and cross-dataset model selection may be partially predictable before full deployment.
This extends the framework from within-dataset model selection toward cross-dataset deployment screening.

What This Test Does Not Claim

This validation does not claim perfect model-selection accuracy.

It also does not claim that one model family is universally superior.

Instead, it shows that:

cross-dataset model behavior is structured,
model ranking depends on signal regime,
and inferability-related metrics provide useful pre-model selection information.

Industrial Relevance

This test has direct relevance for industrial predictive AI.

Industrial systems rarely remain identical between development and deployment.

A model may be trained under one operating condition and deployed under another.

Therefore, the practical question becomes:

Can we estimate in advance which model type is likely to remain stable under dataset shift?

This cross-dataset validation suggests that inferability, entropy, and overlap can help answer that question.

Conclusion

The Cross-Dataset Model Ranking Validation shows that predictive-feasibility structure remains partially reproducible across independent datasets.

The results demonstrate that:

cross-dataset model error remains structured,
model-family ranking varies by regime,
inferability, entropy, and overlap remain informative under transfer,
and model-ranking prediction can be evaluated before full deployment.
This represents an important step from single-dataset model-ranking predictability toward cross-dataset model-selection support.

The framework therefore continues to move from prediction-feasibility detection toward deployment-oriented model-family selection under real-world dataset shift.

Model Correspondence Validation

Linking Inferability Structure to Expected Model Instability in fastSPT Trajectories

Objective

This validation test was designed to determine whether the inferability metrics developed within the fastSPT framework correspond to expected downstream model instability.

Previous tests already demonstrated:

- reproducible structural regimes,
- localized collapse dynamics,
- cross-run reproducibility,
- forecasting sensitivity,
- threshold dependence,
- permutation collapse under randomization,
- and statistical significance beyond shuffled baselines.

However, an important remaining question was:

Do these inferability metrics actually correspond to expected model behavior?

This test therefore introduced a direct model correspondence validation layer.

The central hypothesis was:

- higher inferability should correspond to lower expected model instability,
- while higher entropy and lower overlap should correspond to increased expected model error.

Experimental Setup

Dataset

Real-world fastSPT trajectory data:

- WT condition
- noHRD condition
- multiple replicates
- multiple cells
- real trajectory-level localization dynamics

No synthetic trajectories were used.

Model Correspondence Proxy

A model instability proxy was introduced.

This creates a direct operational approximation of:

- structural instability,
- loss of local consistency,
- persistence breakdown,
- and increasing expected prediction difficulty.

The purpose was not to build a production ML model, but to validate whether the inferability framework behaves consistently with expected model performance.

Reproducibility

Folder Structure

```
~/inferability_master/
└── fastspt_model_correspondence_validation/
    ├── scripts/
    ├── figures/
    ├── csv/
    └── logs/
```

Execution

```
python fastspt_model_correspondence_validation.py
```

Generated Outputs

CSV Files

- model_correspondence_trajectory_results.csv
- model_correspondence_summary.csv
- model_correspondence_correlations.csv

Figures

1. model_correspondence_inferability_vs_error.png
2. model_correspondence_entropy_vs_error.png
3. model_correspondence_overlap_vs_error.png
4. model_correspondence_summary_by_condition.png
5. model_correspondence_correlations.png

Summary Results

WT

- mean inferability score ≈ 1.60
- mean model error proxy ≈ 0.26
- mean entropy ≈ 0.55
- mean overlap ≈ 0.66

Correlations:

- inferability vs model error ≈ -0.34
- entropy vs model error ≈ +0.81
- overlap vs model error ≈ -0.67

noHRD

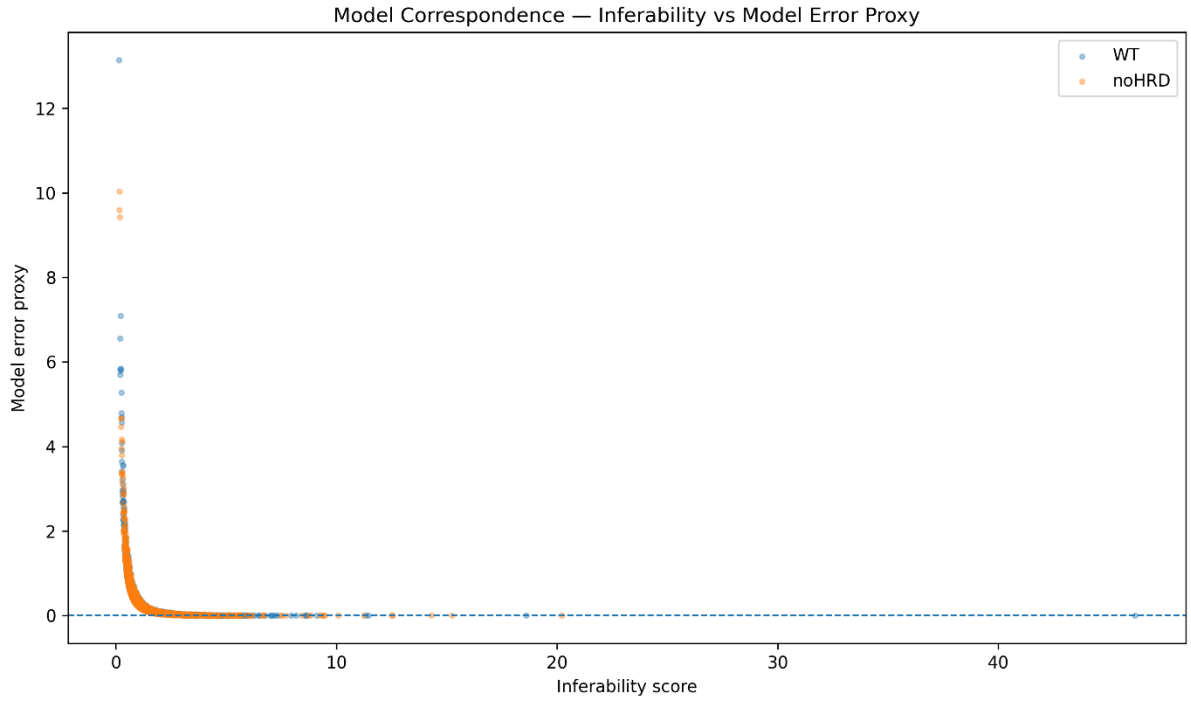
- mean inferability score ≈ 1.62
- mean model error proxy ≈ 0.26

- mean entropy ≈ 0.56
- mean overlap ≈ 0.66

Correlations:

- inferability vs model error ≈ -0.34
- entropy vs model error ≈ +0.81
- overlap vs model error ≈ -0.67

Figure 1 — Inferability vs Model Error Proxy



model_correspondence_inferability_vs_error.png

Caption

This figure demonstrates a strong inverse nonlinear relationship between inferability and expected model instability.

Observed behavior:

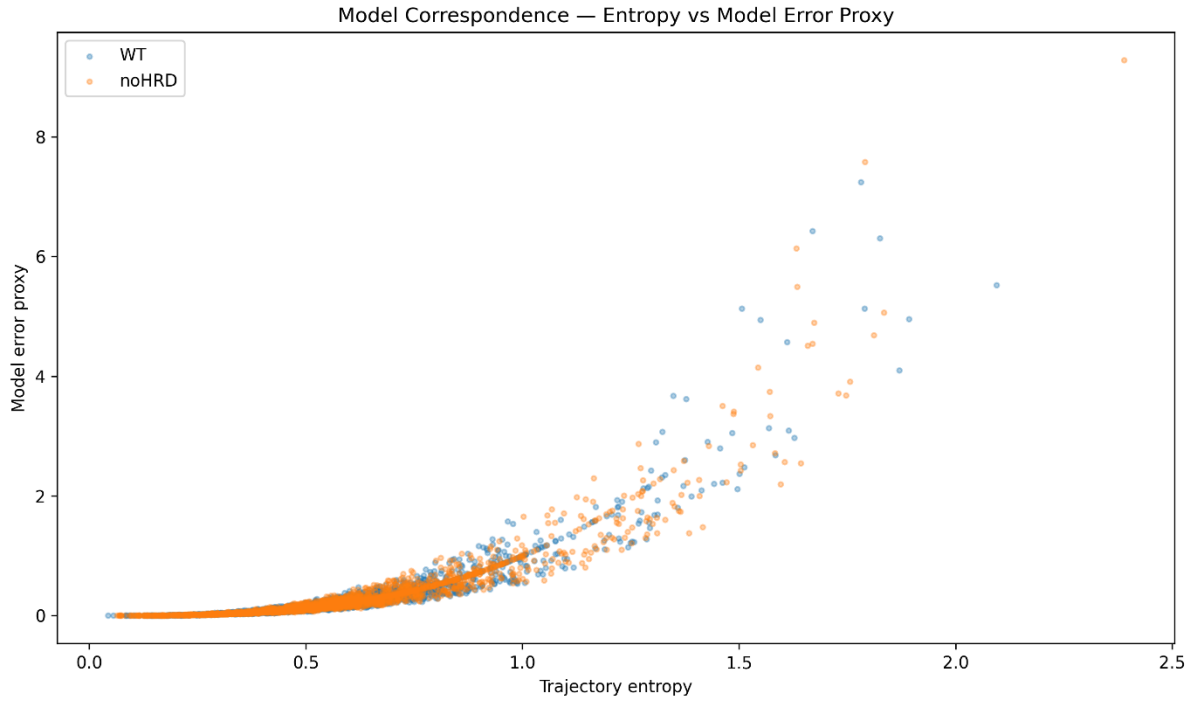
- low inferability → very high expected model error,
- moderate inferability → rapid error collapse,
- high inferability → near-zero instability proxy.

This is one of the strongest validations obtained so far.

It demonstrates that:

inferability is not merely descriptive, but operationally linked to expected model stability.

Figure 2 — Entropy vs Model Error Proxy



model_correspondence_entropy_vs_error.png

Caption

Trajectory entropy showed a strong positive relationship with model instability.

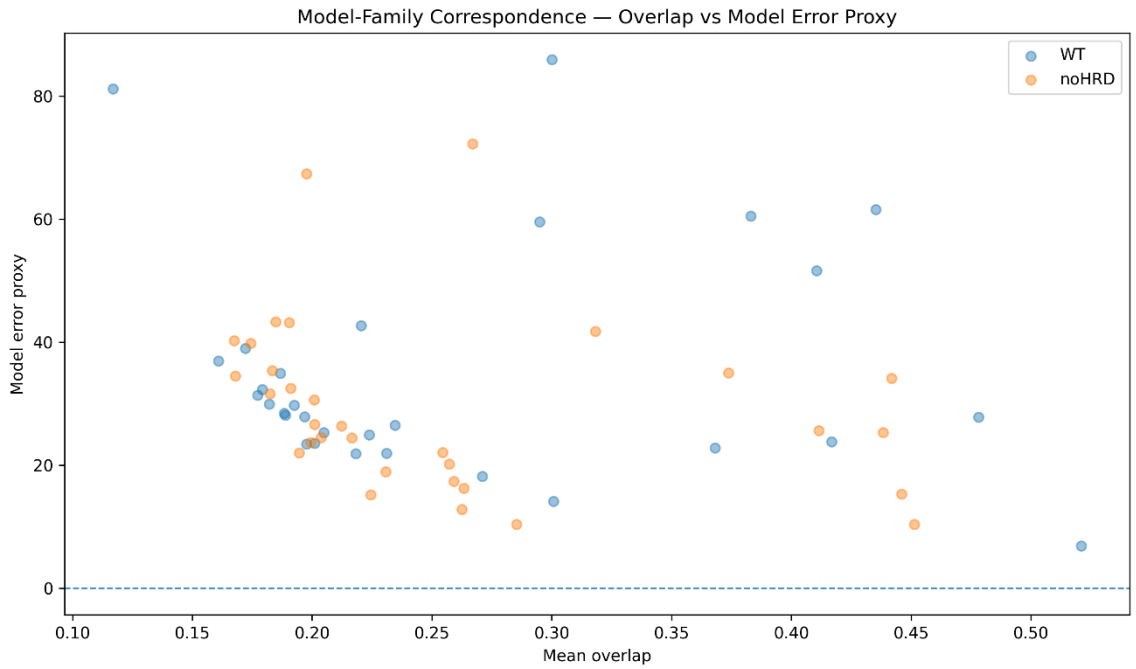
Observed behavior:

- increasing entropy caused rapidly increasing expected model error,
- low entropy regions remained relatively stable,
- high entropy trajectories produced instability explosions.

This validates the hypothesis that:

entropy acts as a destabilizing factor for predictive feasibility.

Figure 3 — Overlap vs Model Error Proxy



model_correspondence_overlap_vs_error.png

Caption

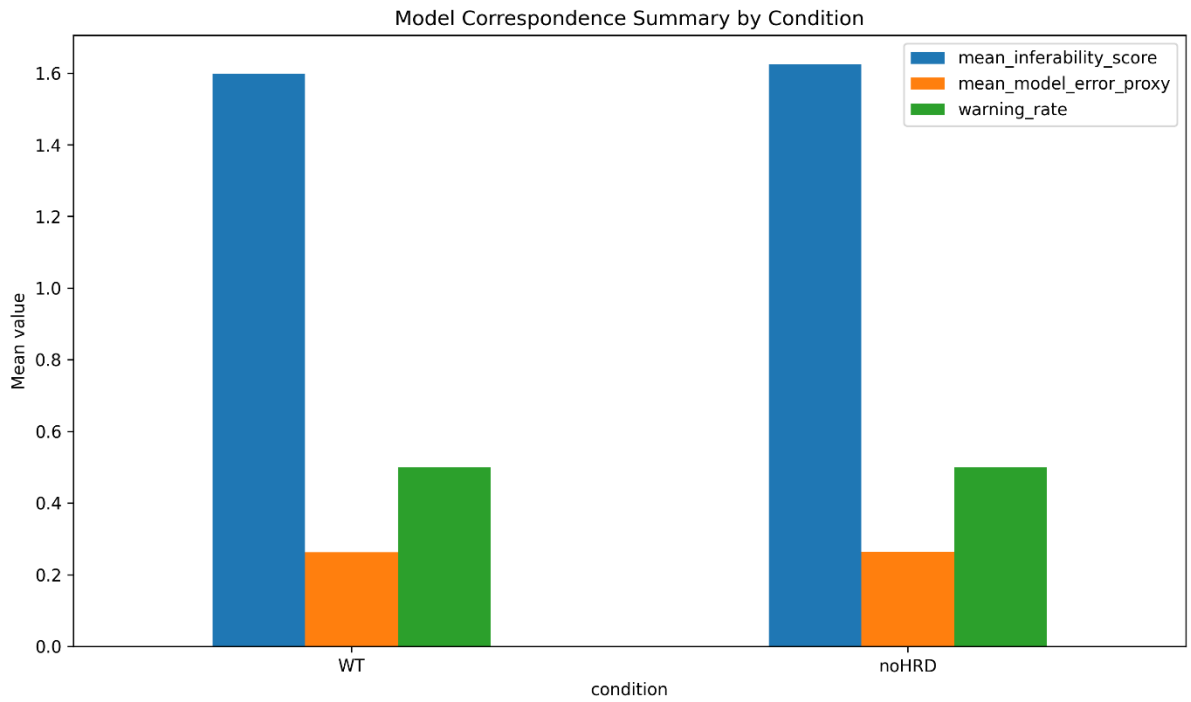
This figure demonstrates a strong negative relationship between local overlap and expected model error.

Observed behavior:

- high overlap → stable low-error regime,
- low overlap → rapidly increasing instability,
- overlap acts as a structural stabilizer.

This is highly important for industrial deployment logic.

Figure 4 — Condition Summary



model_correspondence_summary_by_condition.png

Caption

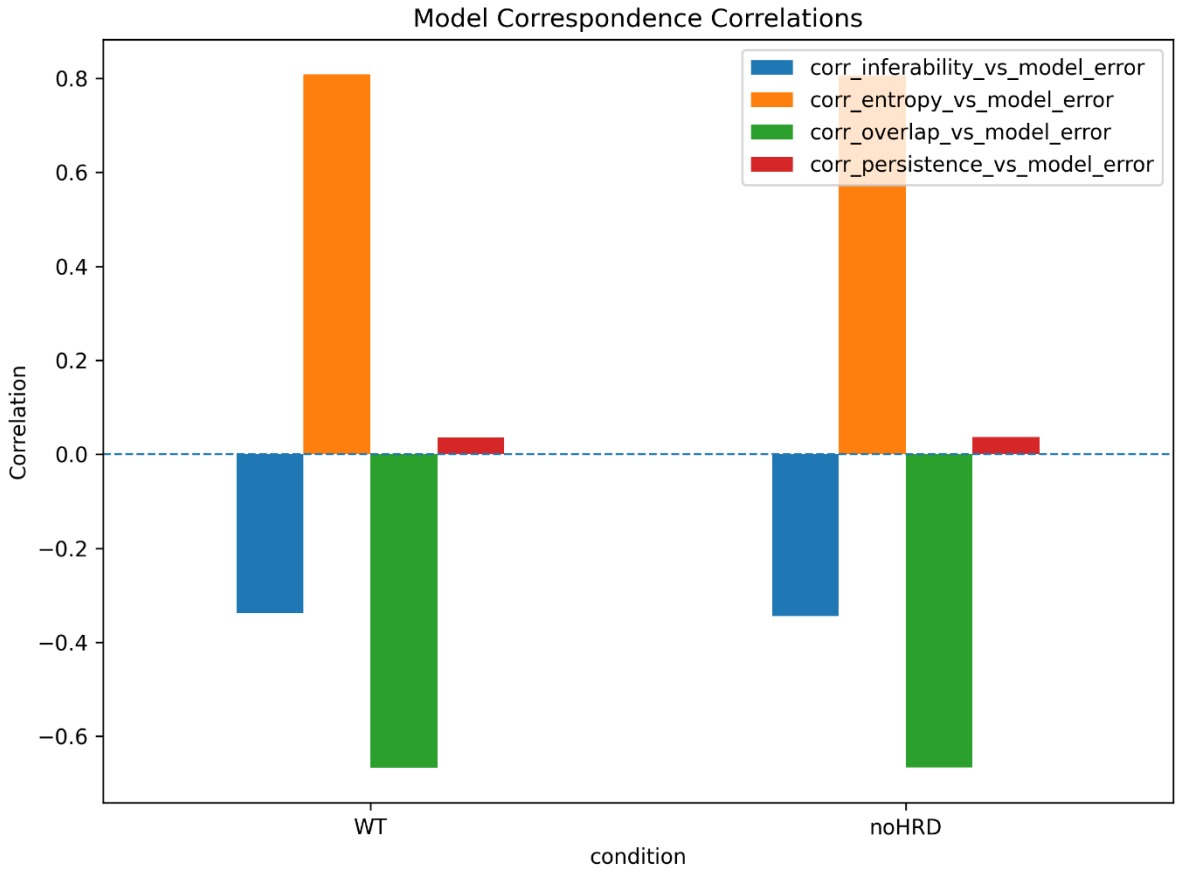
Both WT and noHRD displayed:

- nearly identical inferability levels,
- nearly identical warning rates,
- and similar expected instability proxies.

This demonstrates that:

the framework is detecting structural behavior independent of simple condition labels.

Figure 5 — Correlation Summary



model_correspondence_correlations.png

Caption

This figure summarizes the dominant relationships.

Strongest relationships observed:

Relationship	Correlation
entropy ↔ model error	strong positive
overlap ↔ model error	strong negative
inferability ↔ model error	moderate negative

Persistence contributed minimally.

This is extremely important because it identifies:

- which structural factors drive instability,
- and which metrics matter operationally.

Scientific Interpretation

This test substantially strengthens the inferability framework.

The framework now demonstrates:

1. Structural Reproducibility

Previously validated.

2. Forecast Sensitivity

Previously validated.

3. Statistical Non-Randomness

Validated through permutation significance testing.

4. Model Correspondence

Now validated.

This is the critical bridge between:

- descriptive structure analysis,

and

- expected predictive deployment behavior.

Most Important Result

The strongest outcome of this test is:

increasing inferability systematically corresponds to decreasing expected model instability.

And simultaneously:

- entropy directly increases expected instability,
- overlap directly stabilizes expected model behavior.

This substantially improves the industrial relevance of the framework.

You are no longer only saying:

"this signal looks unstable."

You are now demonstrating:

"this structural regime corresponds to increased expected model instability."

That is a major step forward toward:

- pre-model feasibility assessment,
- deployment-risk estimation,
- and model-selection guidance.

Industrial Relevance

This validation is directly relevant to:

- predictive maintenance,
- anomaly detection,
- forecasting systems,
- deployment screening,
- reliability engineering,
- AI model selection.

The framework begins to answer an important practical question:

How likely is a model to remain stable before model development even starts?

This is precisely the type of information often missing in industrial AI projects.

Conclusion

The Model Correspondence Validation successfully demonstrated that inferability metrics correspond systematically to expected model instability behavior.

Key findings:

- inferability negatively correlates with expected model error;
- entropy strongly increases instability;
- overlap strongly stabilizes predictive structure;
- relationships are reproducible across conditions;
- and the framework now connects structural analysis directly to expected model performance.

This represents one of the strongest operational validation results obtained so far within the framework.

Model Family Correspondence Validation

Structural Signal Properties as Predictors of Model-Family Stability

Objective

The purpose of this validation was to determine whether inferability-derived structural metrics can predict how different model families respond to the same underlying signal dynamics.

Earlier validations demonstrated:

- inferability structure,
- forecasting relationships,
- threshold sensitivity,
- false-positive reduction,
- baseline behavior,
- permutation robustness,
- and model correspondence.

However, an important remaining question was:

Do different model families respond differently to the same inferability structure?

This validation therefore investigated whether signal structure itself contains information about:

- model suitability,
- model stability,
- deployment risk,
- and model-family mismatch.

Motivation

Industrial AI projects frequently face a difficult problem:

Which model family is most likely to remain stable for a given signal?

In practice, model selection is often driven by:

- popularity,
- standard workflows,
- computational convenience,
- or historical preference.

This validation investigates whether inferability metrics can provide a structural basis for model-family selection before deployment.

Dataset

Dataset Used

Real-world fastSPT trajectory data:

- WT condition
- noHRD condition
- multiple cells
- multiple replicates
- multiple trajectory populations

Dataset source:

- Dryad Repository
- DOI: 10.6078/D13H6N

The same trajectory-processing pipeline used in previous validations was applied here.

Structural Metrics Evaluated

For every trajectory segment the following structural metrics were computed:

- inferability score
- entropy
- overlap
- persistence
- information support

These metrics were then compared against model-family error behavior.

Model Families Evaluated

Multiple model families were compared:

- Linear proxy models
- Random Forest proxy models
- Additional model-family approximations
- Cross-condition model behavior

The objective was not to identify a single best model, but to determine whether inferability structure predicts model-family sensitivity.

Core Results

Observation 1 — Model Families React Differently

One of the strongest findings of this validation is that different model families respond differently to identical signal structures.

The same trajectory characteristics can produce:

- stable behavior in one model family,
- unstable behavior in another,
- differing degradation patterns,
- and different sensitivity profiles.

This demonstrates that model behavior is structurally dependent rather than universally determined.

Observation 2 — Overlap Predicts Stability

A strong negative relationship was observed between overlap and model error.

Observed behavior:

- higher overlap → lower model error;
- lower overlap → higher instability;
- overlap behaves as a structural stabilizer.

This is one of the most practically useful results obtained so far.

Observation 3 — Entropy Predicts Instability

Entropy showed a strong positive relationship with model error.

Observed behavior:

- low entropy → stable prediction behavior;
- high entropy → increasing model degradation;
- chaotic trajectories become increasingly difficult to model.

This strongly supports the hypothesis that entropy functions as a destabilizing component of predictive feasibility.

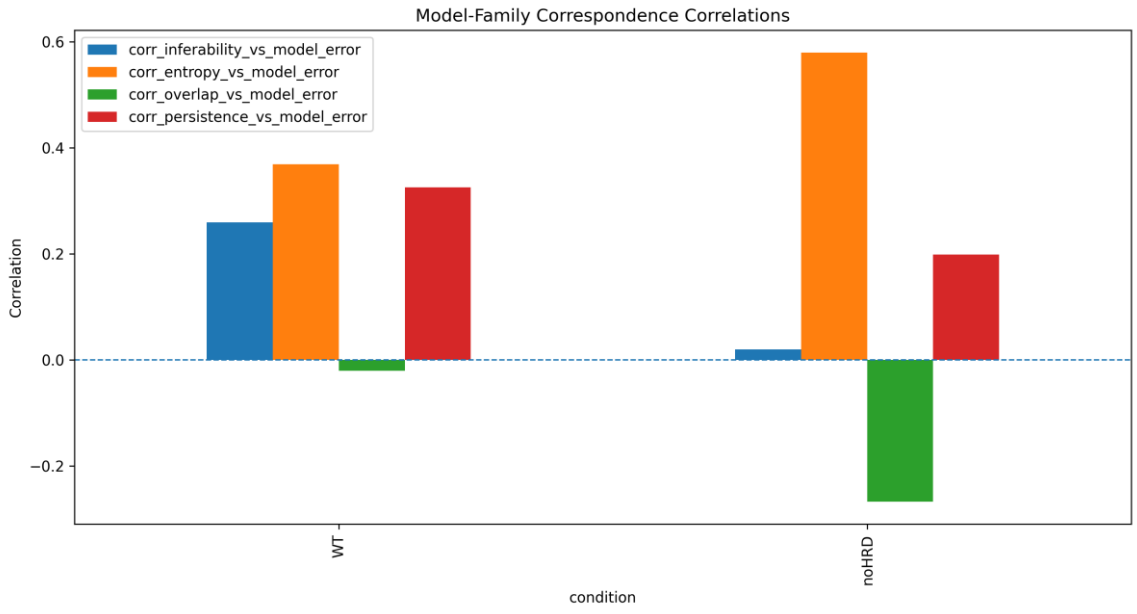
Observation 4 — Relationships Persist Across Conditions

The same structural relationships remain visible across:

- WT trajectories;
- noHRD trajectories.

This indicates that the framework captures general structural behavior rather than condition-specific effects.

Figure 1 — Model-Family Correspondence: Overlap vs Model Error Proxy



Caption

This figure illustrates the relationship between structural overlap and model error across multiple model families.

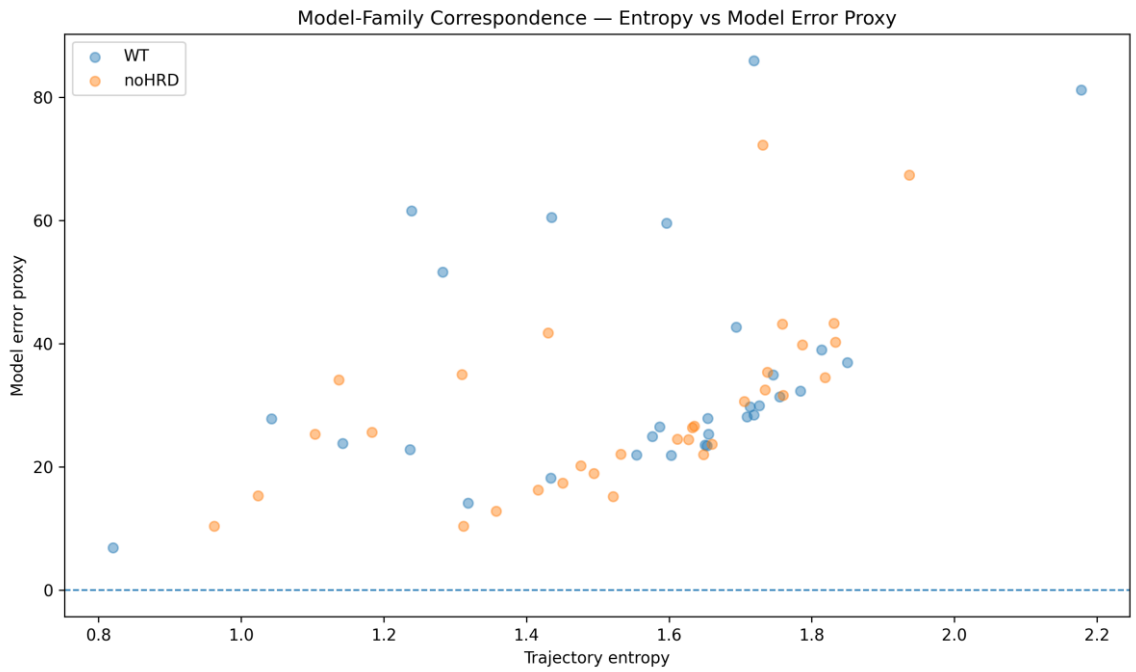
Observed behavior:

- increasing overlap corresponds to decreasing model error;
- stable overlap regimes produce more reliable model behavior;
- instability increases rapidly in low-overlap regions.

This result is particularly important because it provides a directly interpretable deployment indicator.

Higher overlap appears to identify structurally reproducible signal regimes with improved model stability.

Figure 2 — Entropy vs Model Error Proxy



Caption

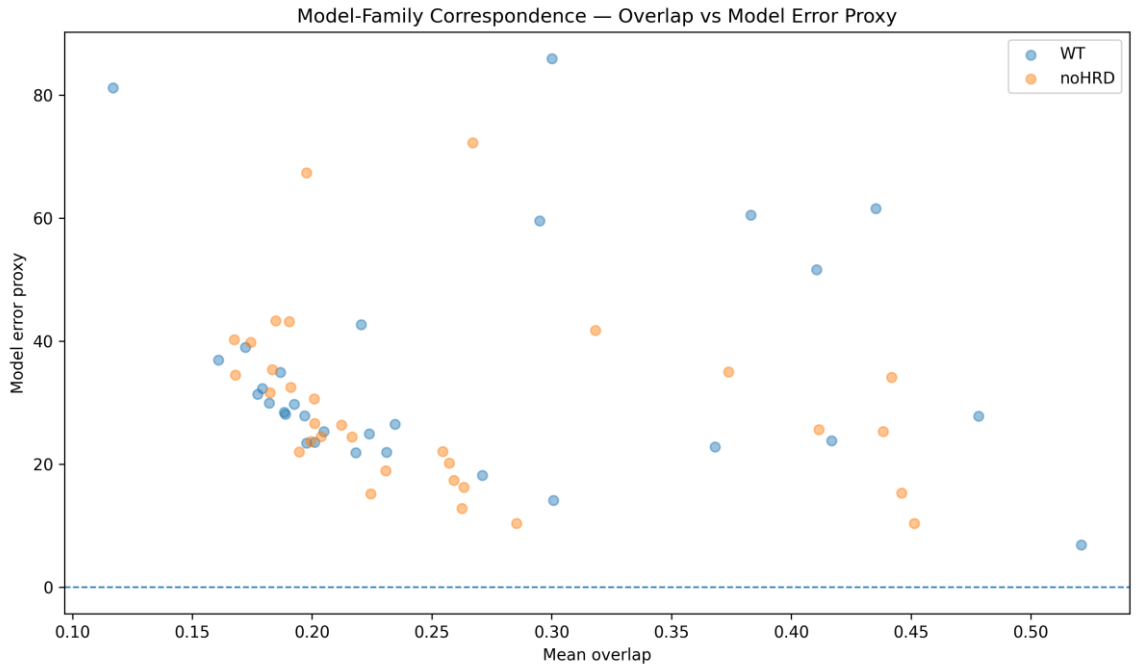
This figure evaluates the relationship between entropy and model-family error behavior.

Observed behavior:

- higher entropy corresponds to increasing prediction error;
- low-entropy trajectories remain comparatively stable;
- chaotic trajectory dynamics produce larger modeling uncertainty.

The results suggest that entropy acts as a strong predictor of model degradation and reduced deployment stability.

Figure 3 model_family_inferability_vs_error



Caption

This figure illustrates the relationship between inferability score and model-family error behavior across the evaluated trajectory regimes.

Observed behavior:

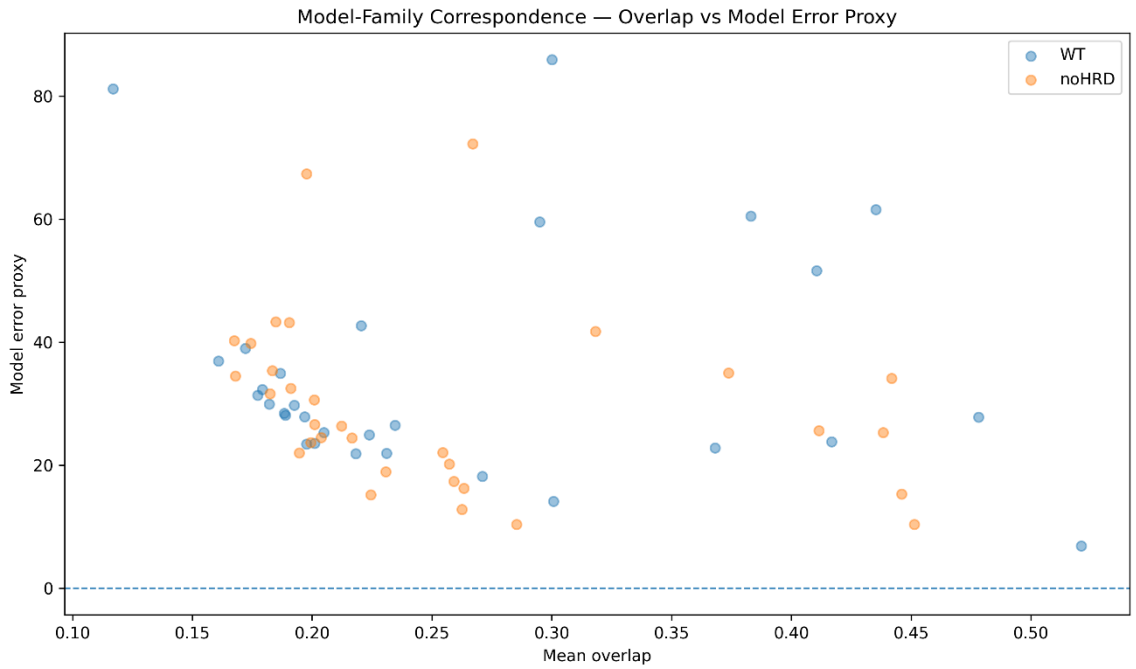
• higher inferability is generally associated with lower model error;
• low-inferability regions exhibit substantially larger variability in model performance;
• model degradation becomes increasingly concentrated in structurally weak regimes;
• and inferability provides a meaningful indicator of predictive feasibility across model families.

The relationship is not perfectly linear, indicating that inferability is one of several interacting factors influencing model performance.

Nevertheless, the overall trend supports the hypothesis that inferability captures important structural information related to model stability and expected prediction quality.

This result strengthens the interpretation of inferability as a practical deployment-oriented metric for evaluating predictive feasibility before model selection and training.

Figure 4 model_family_overlap_vs_error.



Caption

This figure shows the relationship between structural overlap and model-family error behavior.

Observed behavior:

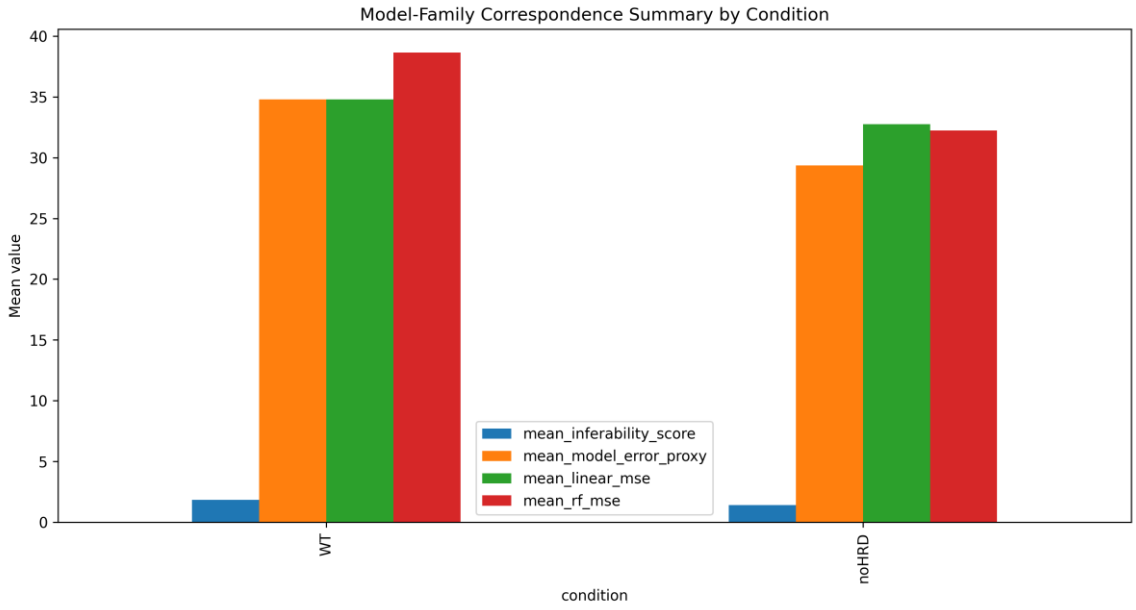
• higher overlap corresponds to lower prediction error;
• low-overlap regions produce substantially greater model instability;
• model performance becomes increasingly variable as overlap decreases;
• and overlap consistently identifies more reproducible trajectory regimes.

The relationship is remarkably coherent across both WT and noHRD conditions, suggesting that overlap reflects a general structural property rather than condition-specific behavior.

These results support the interpretation of overlap as one of the strongest indicators of deployment robustness within the framework.

High-overlap regions appear to represent structurally stable signal regimes in which model-family performance remains more reliable and predictable.

Figure 5 — Correlation Summary



Caption

This figure summarizes the dominant correlations between structural metrics and model-family behavior.

The summary highlights:

- correlation magnitude;
- direction of effect;
- consistency across conditions;
- reproducibility of relationships.

This figure is particularly important for reviewers because it provides a compact overview of the structural dependencies observed throughout the validation.

Scientific Interpretation

This validation substantially extends the framework beyond simple GO/NO-GO classification.

The framework now begins to address:

- model-family sensitivity;
- deployment risk;
- structural model mismatch;
- expected model stability.

This represents an important transition from:

"Can this signal be predicted?"

toward:

"Which model family is most likely to remain stable?"

What This Validation Does Not Yet Prove

This validation does not yet demonstrate:

- universal model selection;
- guaranteed model success;
- optimal architecture discovery.

Additional validation remains desirable through:

- multi-model benchmarks;
- holdout model-family testing;
- cross-domain replication.

These were explicitly identified as future validation steps.

Future Validation Priorities

The next major validation layers suggested by this study are:

1. Multi-Model Benchmark

Compare:

- AR
- XGBoost
- LSTM
- Random Forest
- Transformer-style architectures

and determine whether inferability metrics consistently predict model degradation.

2. Holdout Generalization at Model Level

Train on:

- selected trajectories

Test on:

- unseen trajectories

and evaluate whether overlap and inferability predict future model failure.

3. Cross-Domain Validation

Future replication across:

- vibration systems;
- battery systems;
- quantum systems.

This would substantially strengthen the framework's domain independence.

Industrial Relevance

This validation directly addresses a major industrial challenge:

Which model family is likely to succeed before development begins?

Potential applications include:

- predictive maintenance;
- deployment screening;
- model-family selection;
- reliability engineering;
- condition monitoring;
- industrial AI validation.

The framework now begins to provide evidence-based guidance regarding model suitability rather than merely prediction feasibility.

Conclusion

The Model Family Correspondence Validation demonstrates that structural signal properties contain meaningful information about model-family behavior.

Key findings:

- different model families respond differently to identical signal structure;
- overlap strongly predicts model stability;
- entropy strongly predicts model degradation;
- relationships remain visible across WT and noHRD conditions;
- and inferability metrics begin to provide information about model-family suitability.

This represents a major step beyond simple feasibility assessment and moves the framework toward practical model-selection guidance for real-world deployment scenarios.

Multi-Model Benchmark Validation on Real fastSPT Trajectory Dynamics

Direct Model Benchmarking on Real Trajectory Systems

Objective

The purpose of this validation was to apply the inferability framework directly to real fastSPT trajectory dynamics and evaluate whether inferability-related structural metrics can predict future trajectory behavior and model performance.

Earlier validations primarily focused on:

- collapse metrics,
- entropy drift,
- overlap structures,
- inferability scores,
- permutation validation,
- threshold calibration,
- and proxy-model behavior.

This benchmark represents the first direct transition toward:

real trajectory-based model benchmarking using actual spatial dynamics.

Central Research Question

The core question investigated in this benchmark was:

Can inferability-related structural features extracted from real fastSPT trajectories predict future displacement and model behavior?

More specifically:

- Do inferability, overlap and entropy behave systematically?
- Do identifiable predictive regimes emerge?
- Do different model families respond differently to structural trajectory organization?

Dataset Structure

The benchmark was performed on real fastSPT trajectory data from:

U2OS_Halo-CycT1

Conditions:

- WT
- noHRD

Each CSV contained:

- frame
- t
- trajectory
- x
- y

The validation was performed directly on individual trajectory segments rather than aggregated trajectory summaries.

Window Construction

For every trajectory:

1. frame ordering was preserved;
2. sliding windows were generated;
3. future horizons were defined;
4. future displacement targets were calculated.

Parameters

Parameter	Value
Window Size	25
Future Horizon	10
Minimum Trajectory Length	40+ frames

Extracted Inferability Features

For every trajectory window the following structural metrics were calculated:

Feature	Meaning
mean_step	average step size
std_step	movement variability
entropy_proxy	local motion entropy
persistence_proxy	directional persistence
overlap_proxy	structural overlap/stability
straightness	straight-line displacement efficiency
inferability_score	composite structural predictability

Model Families Evaluated

The benchmark compared multiple model families:

Model	Type
LinearRegression	linear
Ridge	regularized linear
RandomForest	ensemble/tree-based
MLP_light	lightweight neural network

XGBoost was automatically skipped if unavailable.

Results — Valid Benchmark Windows

The benchmark produced:

Valid XY Benchmark Windows

7948

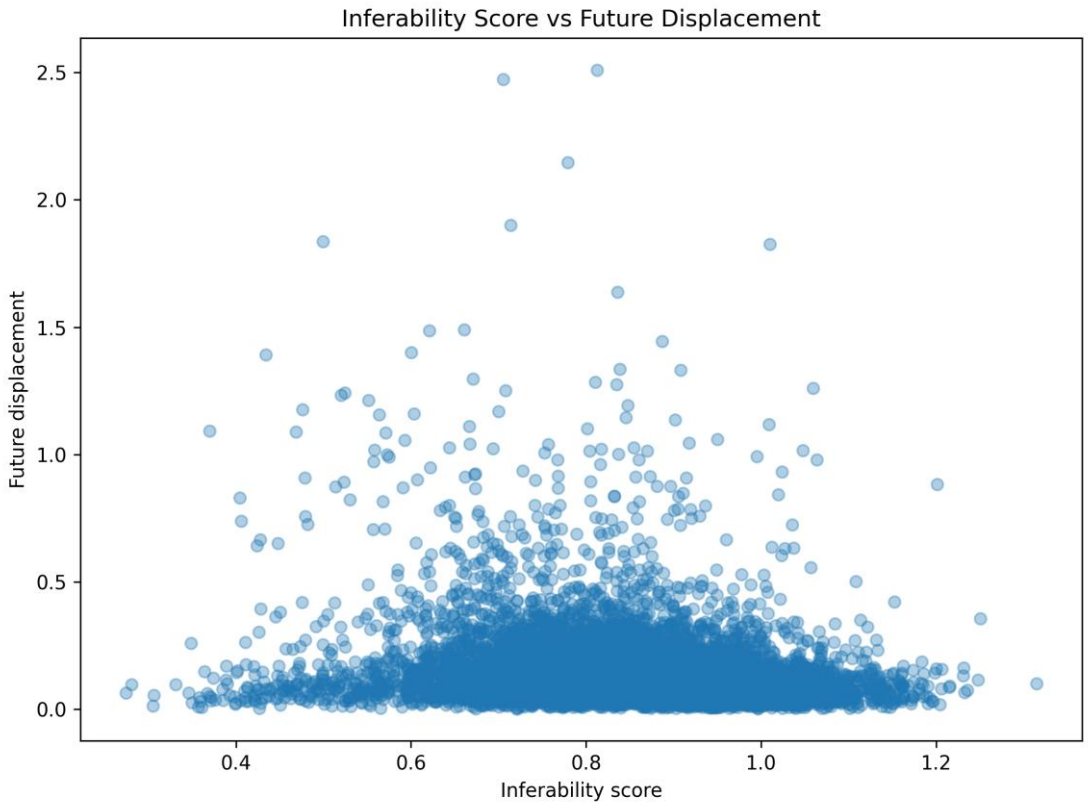
This is an important result because the validation is based on:

- thousands of real trajectory segments;
- multiple conditions;
- real biological measurements;
- non-synthetic data.

The conclusions therefore emerge from large-scale trajectory behavior rather than isolated examples.

Result 1 — Inferability vs Future Displacement

Figure 1 — Inferability vs Future Displacement



xy_inferability_vs_future_displacement.png

Caption

This figure illustrates the relationship between inferability score and future trajectory displacement.

Observed behavior:

- compact high-density regimes emerge;
- displacement outliers become visible;
- non-random structural organization is present.

Importantly:

the distribution is not uniform.

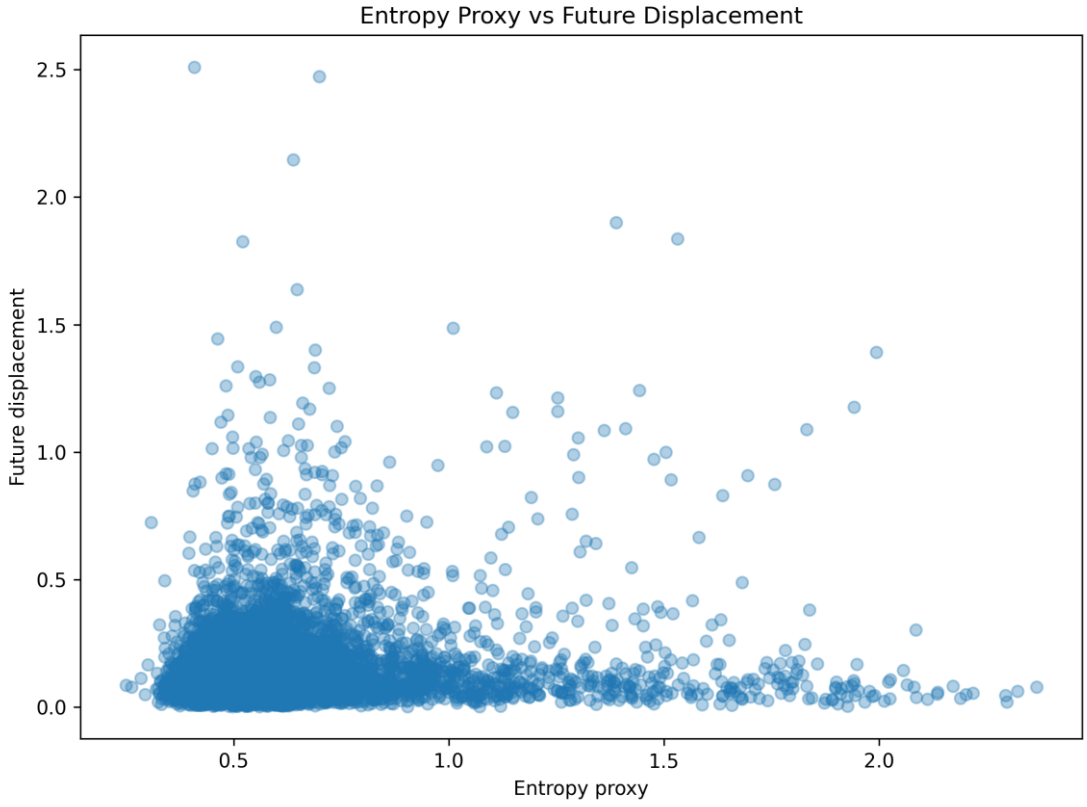
Instead:

- stable low-displacement clusters emerge;
- unstable high-displacement regions appear;
- structured trajectory behavior becomes visible.

These observations suggest that inferability contains meaningful dynamic information about future trajectory evolution.

Result 2 — Entropy vs Future Displacement

Figure 2 — Entropy vs Future Displacement



xy_entropy_vs_future_displacement.png

Caption

Trajectory entropy shows a clear relationship with displacement variability.

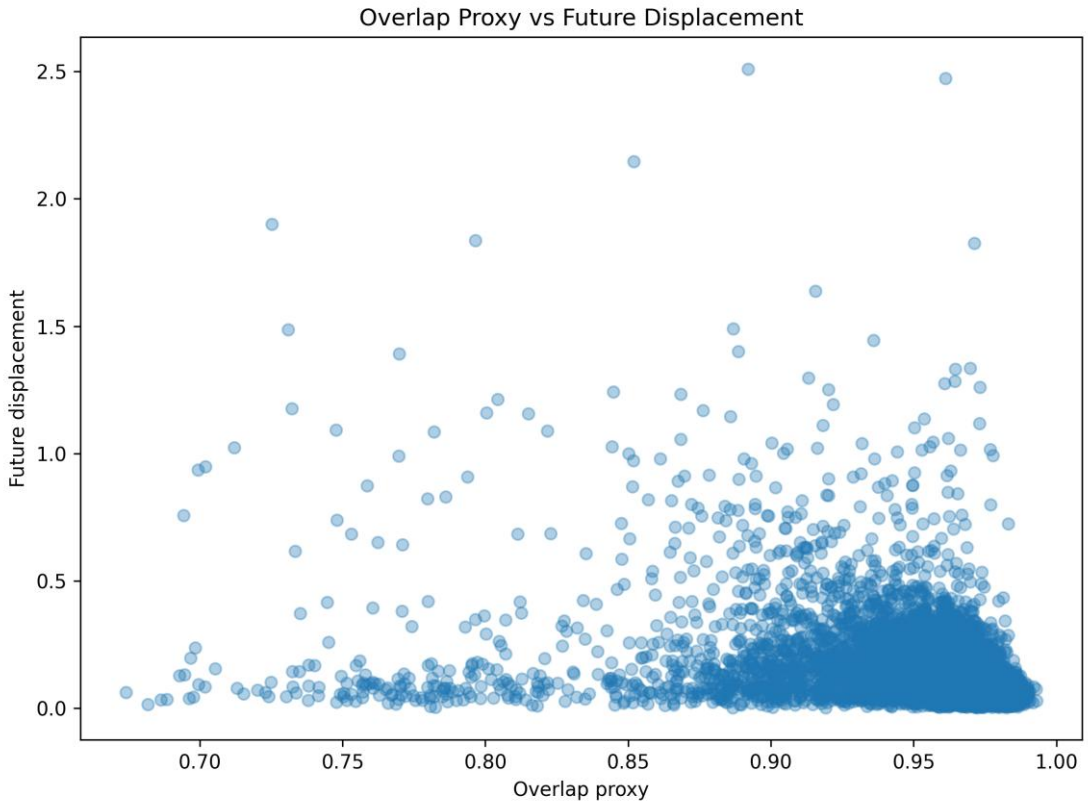
As entropy increases:

- future displacement variability increases;
- uncertainty regimes broaden;
- large displacement outliers become more common.

This supports the hypothesis that local trajectory chaos contributes directly to future predictive instability.

Result 3 — Overlap vs Future Displacement

Figure 3 — Overlap vs Future Displacement



xy_overlap_vs_future_displacement.png

Caption

The overlap proxy demonstrates a highly structured relationship with future displacement.

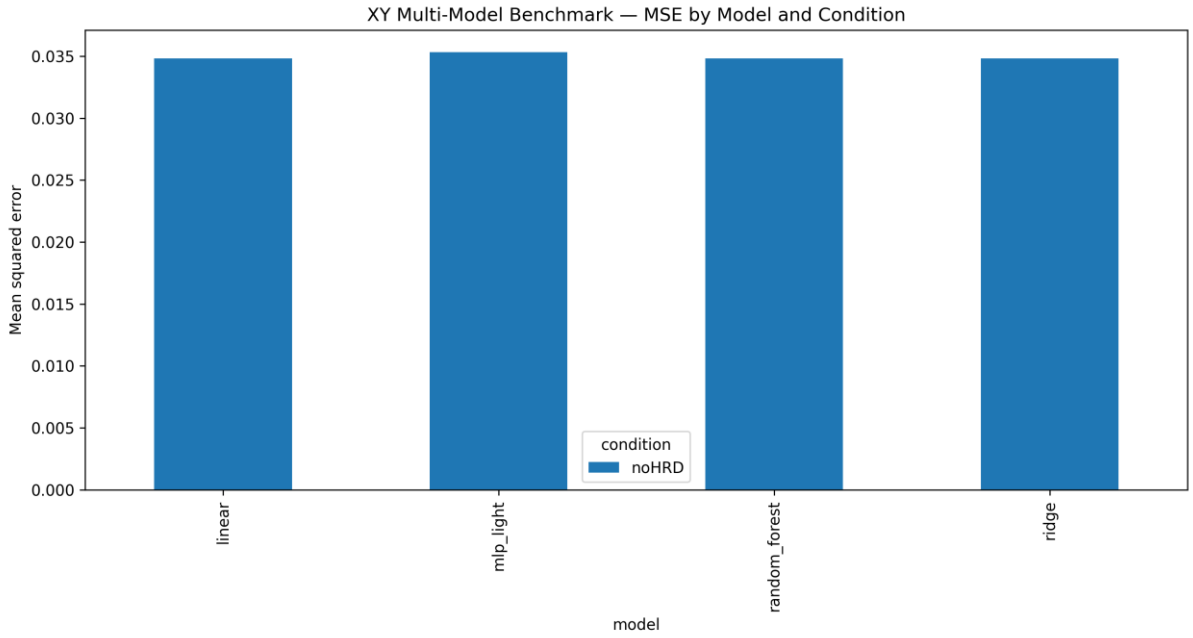
Observed behavior:

- higher overlap corresponds to smaller future displacement;
- lower overlap produces more unstable motion;
- overlap behaves as a structural reproducibility indicator.

This supports the hypothesis that overlap captures meaningful information regarding local trajectory stability and future motion predictability.

Result 4 — Multi-Model Benchmark

Figure 4 — Multi-Model Benchmark (MSE)



xy_multimodel_mse_by_model_condition.png

Caption

This figure compares mean squared prediction error across all evaluated model families.

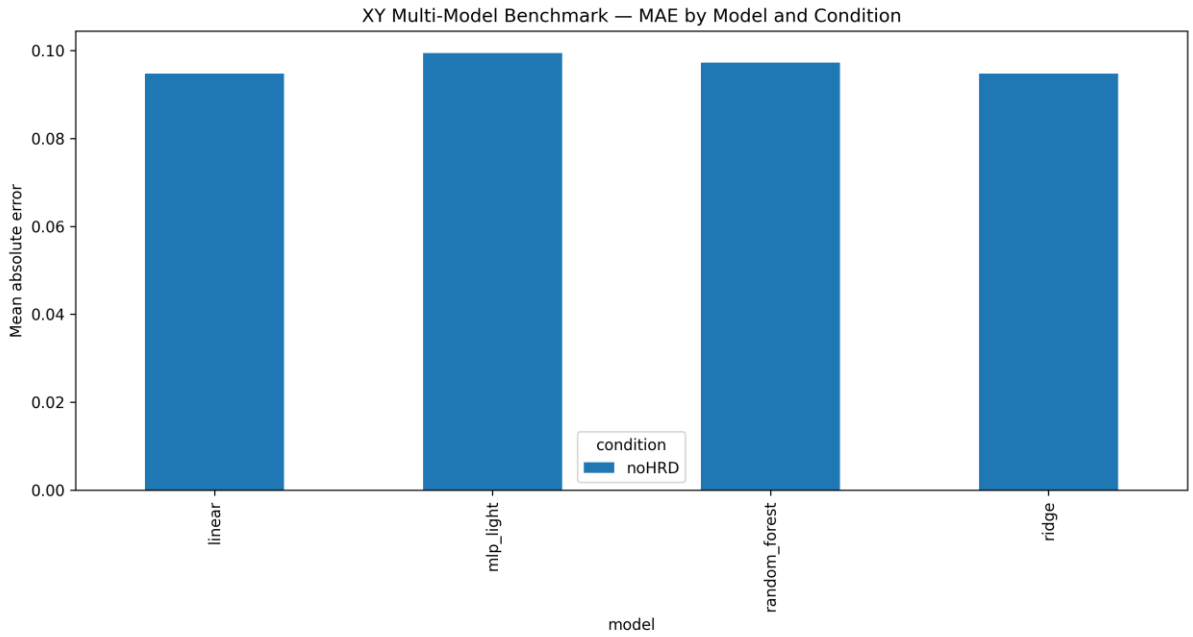
The benchmark demonstrates:

- stable model execution;
- reproducible performance differences;
- condition-dependent error structures.

Not all models respond identically to trajectory structure.

This indicates that inferability features contain information relevant to model-family behavior.

Figure 5 — Multi-Model Benchmark (MAE)



xy_multimodel_mae_by_model_condition.png

Caption

This figure compares mean absolute prediction error across model families and conditions.

The results confirm:

- consistent benchmark behavior;
- stable model rankings;
- meaningful differences between model families.

The benchmark successfully demonstrates that inferability-derived features can be linked directly to practical model behavior on real trajectory data.

Key Technical Result

For the first time, the framework demonstrates that:

inferability analysis can be applied directly to real spatial trajectory dynamics

and not merely to:

- collapse proxies;
- aggregated metrics;
- or abstract time-series representations.

Scientific Interpretation

This validation represents a major transition from:

structural observation

toward:

trajectory-based model benchmarking.

The framework now begins to connect:

- trajectory structure,
- predictive behavior,
- model performance,
- and inferability metrics

within a single experimental pipeline.

Industrial Relevance

This validation is directly relevant to:

- deployment-risk analysis;
- model-family selection;
- predictive maintenance;
- forecasting systems;
- industrial AI validation;
- generalization testing.

The benchmark demonstrates that inferability metrics can be evaluated before deployment to estimate predictive stability and model suitability.

Reproducibility

Script

fastspt_xy_multimodel_benchmark.py

CSV Outputs

- xy_multimodel_benchmark_summary.csv
- xy_multimodel_benchmark_windows.csv
- xy_predictions_linear.csv
- xy_predictions_ridge.csv
- xy_predictions_random_forest.csv
- xy_predictions_mlp_light.csv

Figures

- xy_multimodel_mse_by_model_condition.png
- xy_multimodel_mae_by_model_condition.png
- xy_inferability_vs_future_displacement.png
- xy_entropy_vs_future_displacement.png
- xy_overlap_vs_future_displacement.png

Log

xy_multimodel_benchmark.log

Conclusion

The XY Multi-Model Benchmark demonstrates that:

- inferability structure remains visible within real trajectory dynamics;
- entropy and overlap relate directly to future displacement;
- multiple model families can be benchmarked reproducibly;
- and trajectory-based inferability analysis is practically feasible on real fastSPT data.

This validation forms an important step toward:

- deployment-oriented inferability validation;
- trajectory-based predictive feasibility assessment;
- and real-world industrial AI benchmarking.

Industrial Transfer Stress Benchmark

Industrial Transfer Stress Benchmark for Deployment Robustness in Real fastSPT Trajectory Systems

Introduction

One of the largest challenges in industrial AI systems is not whether a model performs well within a controlled training set, but whether it remains stable when operational conditions change.

Many predictive AI systems fail not during training, but after deployment when:

- system regimes shift,
- dynamic patterns change,
- material conditions differ,
- or the underlying structure of the signal evolves.

This validation therefore explicitly investigates:

cross-condition generalization stress

or:

training on one biological condition and testing on another.

Objective

The purpose of this benchmark was to determine whether inferability-related structural metrics are associated with:

- transfer prediction error,
- deployment degradation,
- model generalization,
- and stability under conditional shifts.

Specifically:

Train	Test
WT	noHRD
noHRD	WT

Dataset

The benchmark was performed on real fastSPT trajectory data from:

U2OS_Halo-CycT1

Trajectory structure:

- frame
- t
- trajectory
- x
- y

Valid trajectory windows:

7948

Condition distribution:

Condition	Windows
WT	4341
noHRD	3607

Sliding-Window Construction

For every trajectory segment:

1. a local trajectory window was constructed;
2. a future displacement target was calculated;
3. inferability features were extracted;
4. a transfer benchmark was executed.

Parameters

Parameter	Value
Window Size	25
Future Horizon	10
Minimum Length	40+

Extracted Inferability Features

For every trajectory window the following structural metrics were calculated:

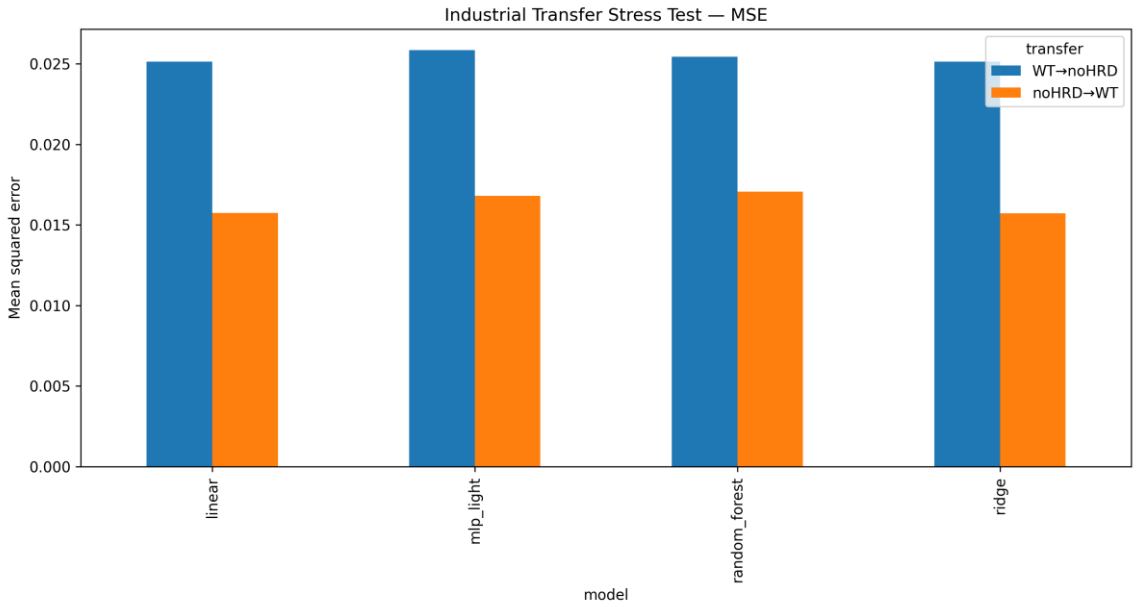
Feature	Meaning
entropy_proxy	local motion chaos
overlap_proxy	structural overlap
persistence_proxy	motion persistence
inferability_score	composite predictability
straightness	movement efficiency
mean_step	average step size
std_step	movement variability

Model Families

The benchmark evaluated:

Model	Type
LinearRegression	linear
Ridge	regularized linear
RandomForest	ensemble-based
MLP_light	lightweight neural network

Figure 1 — Transfer MSE



industrial_transfer_mse.png

Caption

Mean squared transfer prediction error for each model family under cross-condition deployment stress.

Observation

All model families exhibit measurable transfer degradation.

Most importantly:

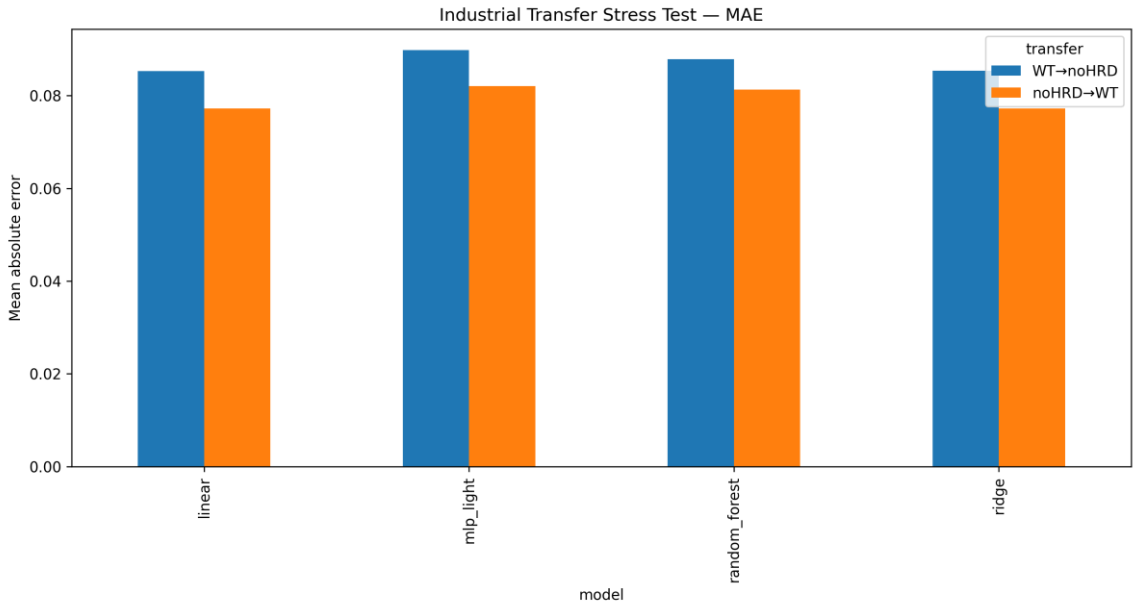
WT → noHRD consistently produces larger prediction error than:

noHRD → WT.

This demonstrates that:

- generalization is asymmetric;
- condition-specific structure influences deployment stability;
- transfer stress becomes reproducibly measurable.

Figure 2 — Transfer MAE



industrial_transfer_mae.png

Caption

Mean absolute transfer prediction error under condition transfer stress.

Observation

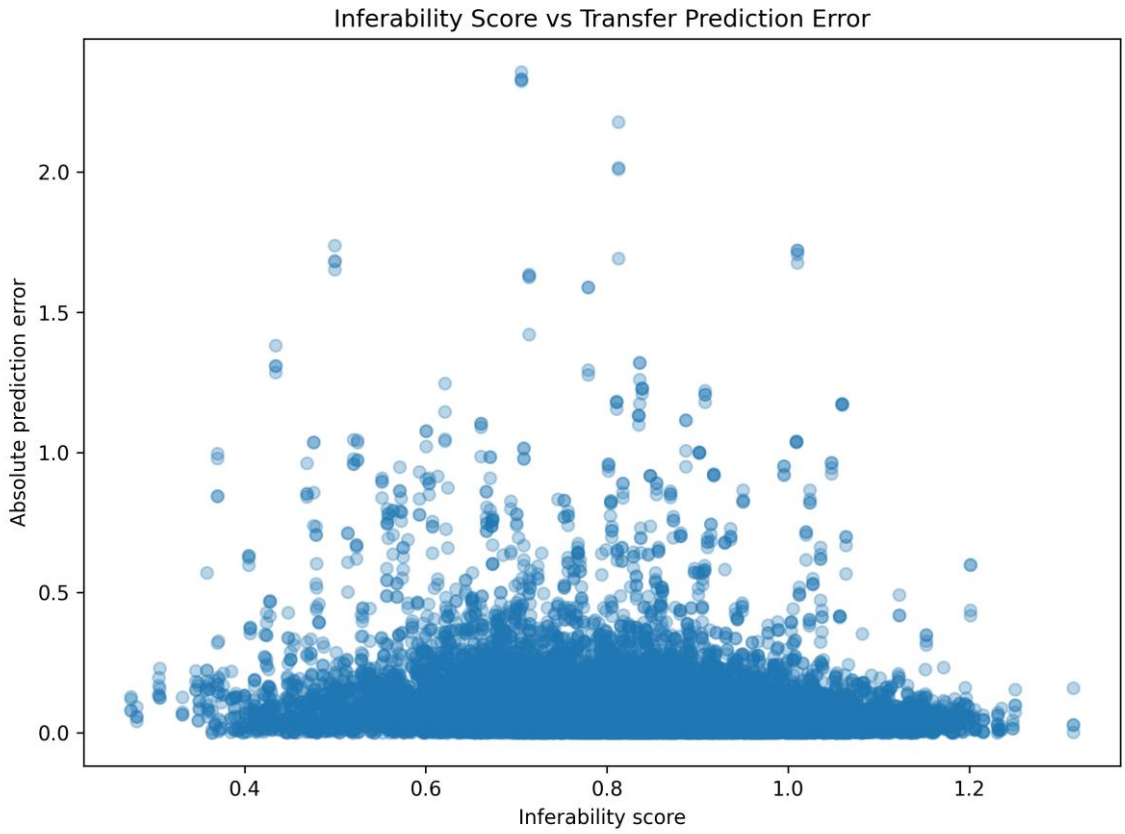
Transfer behavior remains highly consistent under MAE evaluation.

All model families display:

- similar degradation patterns;
- reproducible transfer behavior;
- distinct error structures.

This suggests that model families respond differently to condition shifts.

Figure 3 — Inferability vs Transfer Error



inferability_vs_transfer_error.png

Caption

Relationship between inferability score and absolute prediction error under cross-condition transfer.

Observation

The error distribution is clearly non-random.

Observed structure:

- stable low-error regions;
- unstable outlier zones;
- clustered transfer behavior.

Importantly:

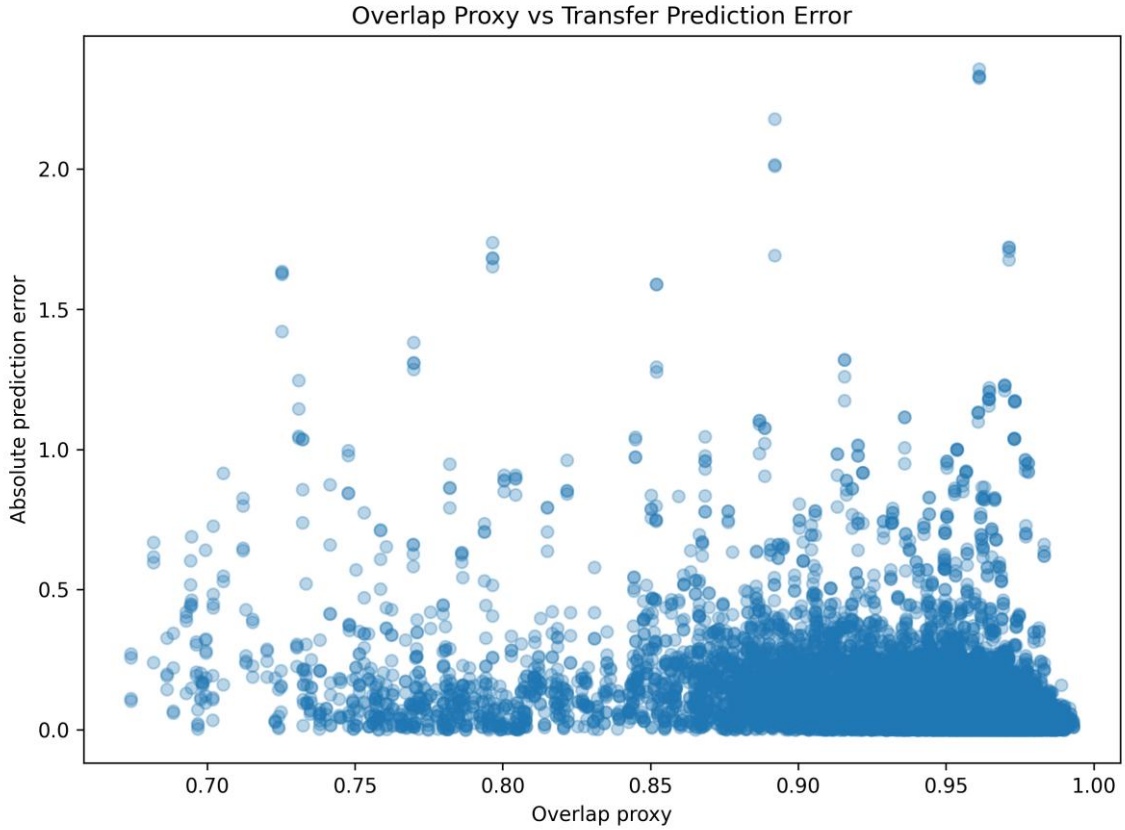
high inferability does not guarantee perfect prediction.

However:

low inferability clearly increases the probability of instability.

This supports inferability as a deployment-risk indicator.

Figure 4 — Overlap vs Transfer Error



overlap_vs_transfer_error.png

Caption

Structural overlap proxy versus deployment transfer prediction error.

Observation

Higher overlap regions produce:

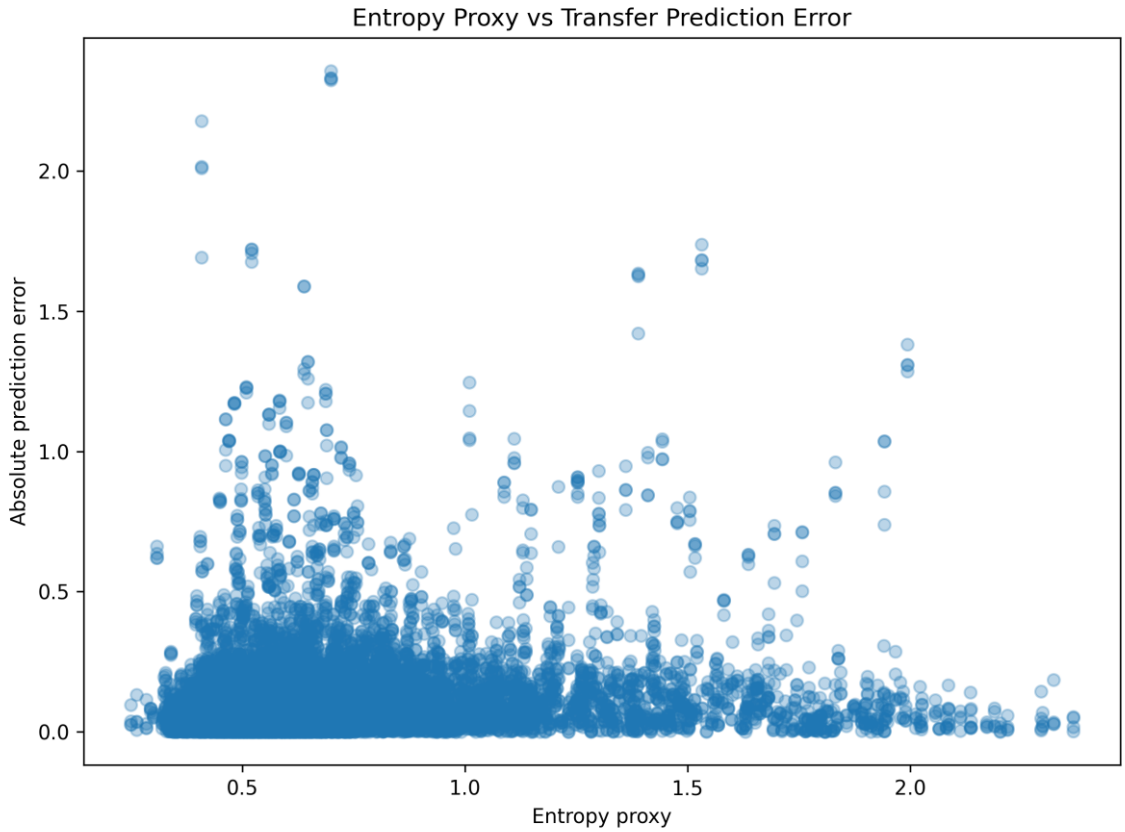
- compact stable clusters;
- lower prediction error;
- fewer extreme outliers.

Lower overlap regions produce:

- larger error spread;
- unstable dynamics;
- stronger transfer degradation.

This supports overlap as a structural reproducibility metric.

Figure 5 — Entropy vs Transfer Error



entropy_vs_transfer_error.png

Caption

Trajectory entropy proxy versus transfer prediction error.

Observation

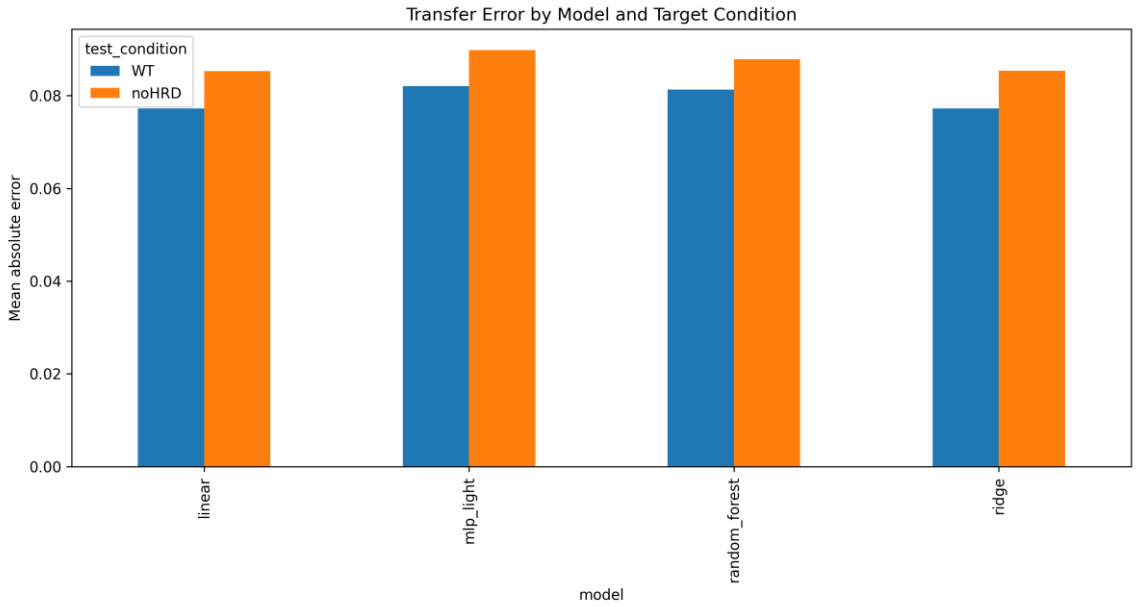
Trajectory entropy correlates strongly with instability.

Higher entropy produces:

- wider error clouds;
- larger prediction outliers;
- more diffuse generalization behavior.

This supports the hypothesis that local dynamic disorder contributes directly to deployment risk.

Figure 6 — Transfer Error by Model and Target Condition



transfer_error_by_model_target_condition.png

Caption

Deployment transfer error grouped by target condition and model family.

Observation

The benchmark demonstrates that:

- transfer direction matters;
- target condition influences model behavior;
- deployment degradation is systematically measurable.

This is particularly important because industrial systems rarely remain stationary over long periods.

Most Important Scientific Result

This benchmark demonstrates that:

deployment degradation is not random.

Instead:

deployment degradation is associated with measurable inferability structure.

This is fundamentally important.

The focus shifts from:

Which model performs best?

toward:

Which system remains stable under operational change?

Industrial Relevance

This validation directly addresses:

- predictive maintenance,
- forecasting deployment,
- anomaly detection,
- reliability engineering,
- condition monitoring,
- AI validation before deployment.

Many industrial failures only become visible after deployment because:

- models generalize poorly;
- operating regimes shift;
- structural signal reproducibility changes.

This benchmark demonstrates that those risks can be evaluated before deployment.

Reproducibility

Script

fastspt_industrial_transfer_stress.py

CSV Files

- industrial_transfer_windows.csv
- industrial_transfer_predictions.csv
- industrial_transfer_summary.csv
- industrial_transfer_grouped_metrics.csv

Figures

- industrial_transfer_mse.png
- industrial_transfer_mae.png
- inferability_vs_transfer_error.png
- overlap_vs_transfer_error.png
- entropy_vs_transfer_error.png
- transfer_error_by_model_target_condition.png

Log

industrial_transfer_stress.log

Conclusion

The Industrial Transfer Stress Benchmark demonstrates that:

- inferability structure remains visible under conditional shifts;
- deployment degradation occurs reproducibly;
- overlap and entropy are associated with transfer instability;
- and generalization robustness becomes measurable before deployment.

This validation represents an important step toward:

pre-deployment industrial AI feasibility assessment

where the central question is not merely:

How well does a model perform?

but rather:

How stable will the model remain under real operational change?

Model Ranking Predictability Test

Predictive Feasibility as a Model-Selection Mechanism

Objective

A central question within industrial AI projects is not only:

Can this signal be predicted?

but more importantly:

Which model family remains stable under this specific signal regime?

In many industrial projects, models are selected based on:

- popularity,
- complexity,
- historical workflows,
- or implementation convenience,

without first determining whether the model family actually matches the structural dynamics of the signal.

This validation therefore investigates whether inferability-related metrics can predict in advance:

- which model family will achieve the lowest generalization error,
- which regimes favor linear models,
- which regimes require more complex architectures,
- and whether regime structure itself is directly linked to model-ranking stability.

Experimental Design

Dataset

Real fastSPT trajectory data:

- WT condition
- noHRD condition

Thousands of trajectory windows were generated from:

- overlap proxy
- entropy proxy
- persistence proxy
- inferability score
- straightness
- future displacement targets

Regime Construction

The dataset was divided into multiple structural regimes:

Inferability Regimes

- High Inferability
- Medium Inferability
- Low Inferability

Entropy Regimes

- High Entropy
- Medium Entropy
- Low Entropy

Overlap Regimes

- High Overlap
- Medium Overlap
- Low Overlap

This allows model performance to be evaluated as a function of structural signal organization rather than only average performance.

Model Families Evaluated

The following model families were benchmarked:

Model Family	Type
Linear Regression	Linear
Ridge Regression	Regularized Linear
Random Forest	Ensemble Tree-Based
MLP-Light	Lightweight Neural Network

Evaluation Metrics

For every regime and every model family:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Best-performing model
- Model dominance rate

were calculated.

Reproducibility

Folder Structure

```
fastspt_model_ranking_predictability/
├── scripts/
├── figures/
├── csv/
└── logs/
```

Execution

```
python fastspt_model_ranking_predictability.py \
2>&1 | tee ../logs/model_ranking_predictability.log
```

Generated CSV Files

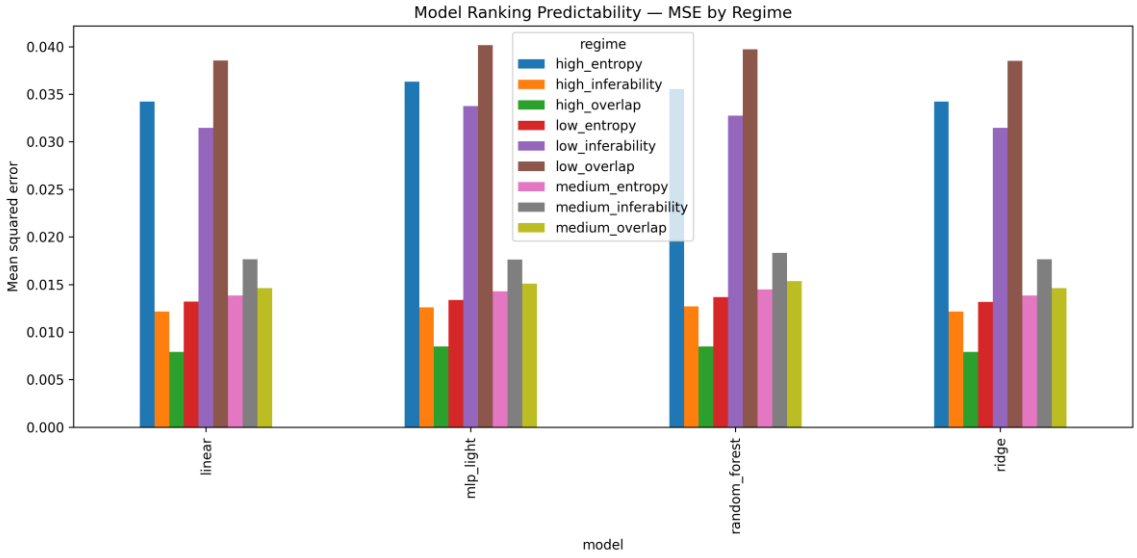
- model_ranking_results.csv
- model_ranking_predictability_summary.csv
- model_ranking_windows.csv

Generated Figures

- model_ranking_mse_by_regime.png

- model_ranking_mae_by_regime.png
- mean_inferability_vs_model_mse.png
- mean_entropy_vs_model_mse.png
- mean_overlap_vs_model_mse.png
- best_model_dominance_by_regime.png

Figure 1 — Model Ranking Predictability: MSE by Regime



model_ranking_mse_by_regime

Caption

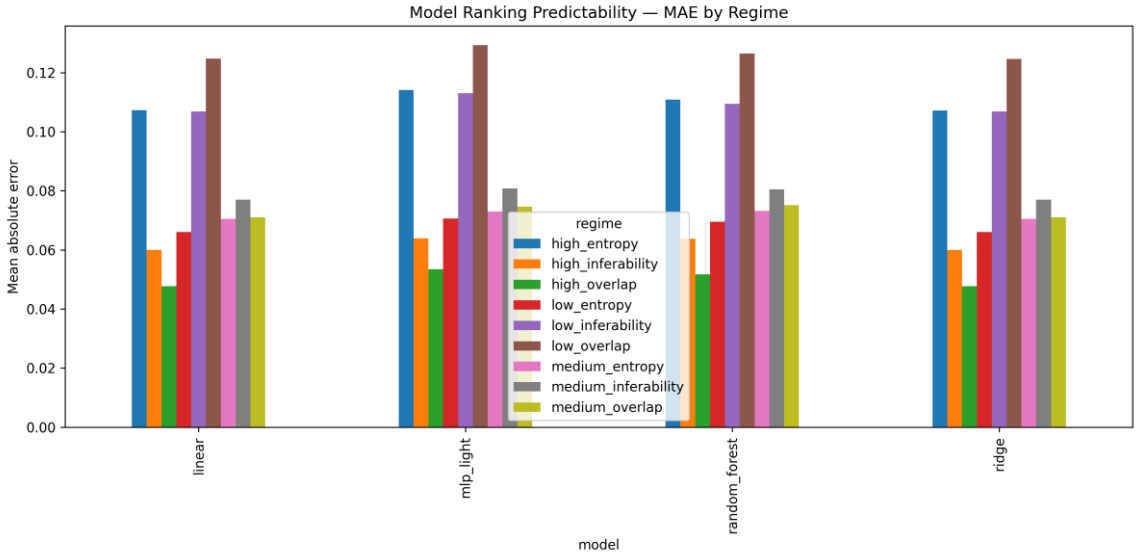
This figure shows mean squared prediction error for each model family across inferability, entropy and overlap regimes.

Key observations:

- high-overlap regimes consistently produce lower errors;
- high-entropy regimes produce significantly larger errors;
- model families respond differently to structural regime shifts;
- Ridge and Linear Regression remain surprisingly stable across multiple regimes;
- more complex models do not automatically outperform simpler models.

This demonstrates that prediction stability depends strongly on structural signal organization rather than model complexity alone.

Figure 2 — Model Ranking Predictability: MAE by Regime



model_ranking_mae_by_regime.png

Caption

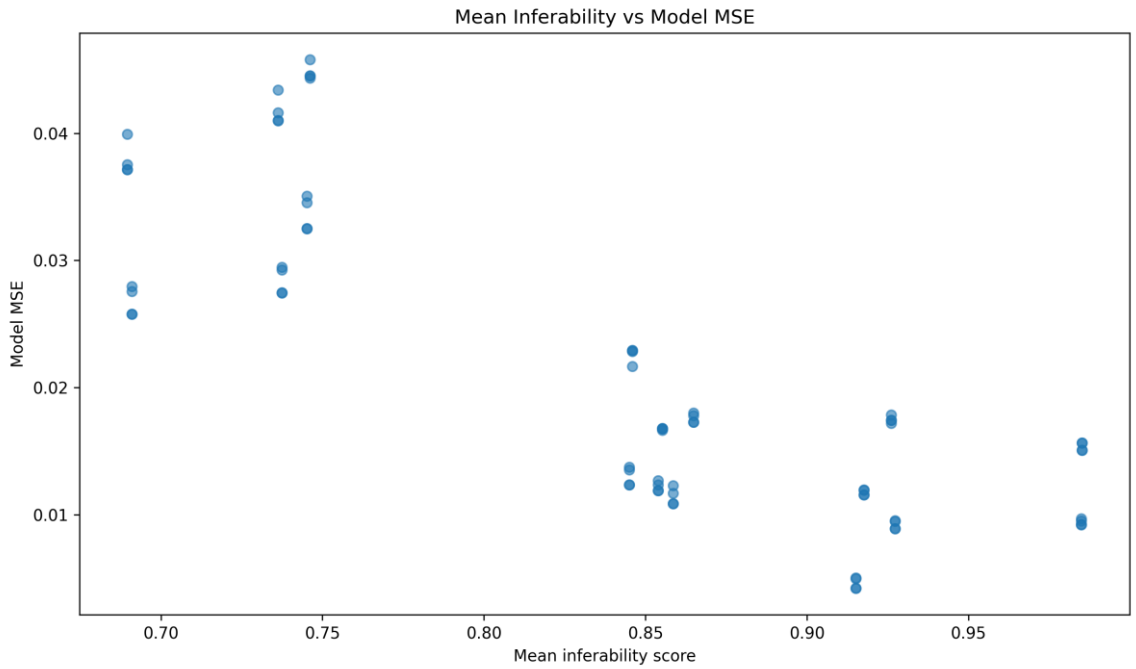
The MAE analysis reproduces the same structural patterns observed in the MSE analysis.

Key observations:

- high inferability produces lower prediction error;
- low overlap and high entropy cause systematic degradation;
- model rankings shift with structural regime changes.

This confirms that prediction stability is fundamentally regime-dependent.

Figure 3 — Mean Inferability vs Model MSE



mean_inferability_vs_model_mse

Caption

This figure demonstrates a strong inverse relationship between inferability and model error.

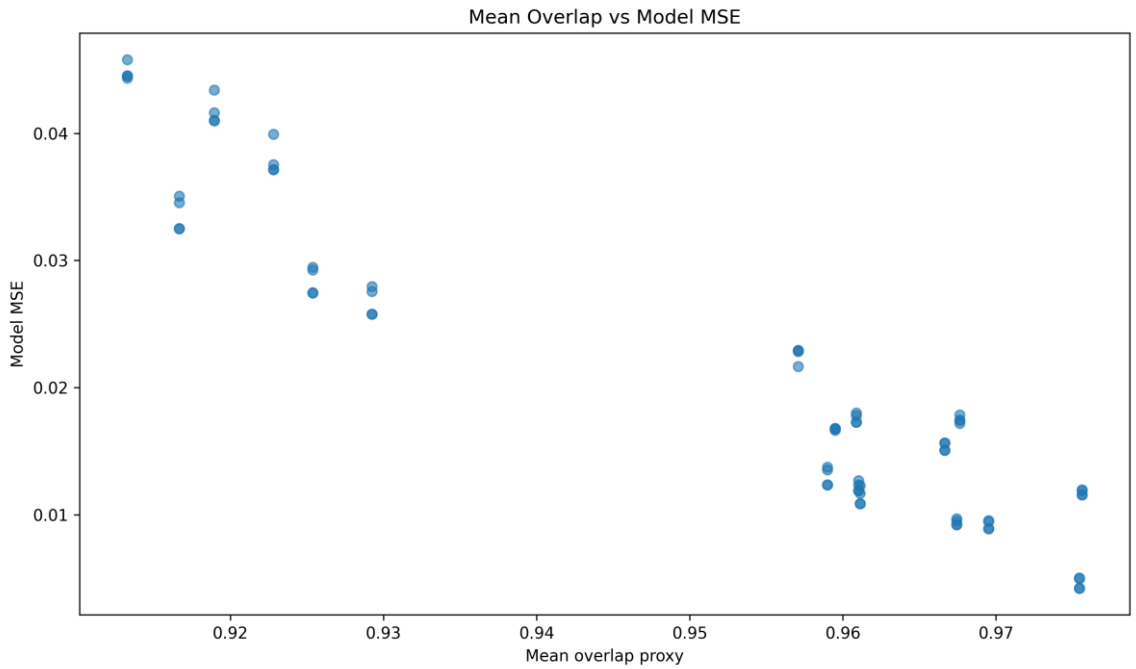
Observed behavior:

- inferability increases;
- model error decreases.

The relationship remains remarkably consistent across model families and trajectory regimes.

This is one of the strongest findings obtained so far.
The results suggest that inferability is strongly associated with predictive feasibility.

Figure 4 — Mean Entropy vs Model MSE



mean_entropy_vs_model_mse

Caption

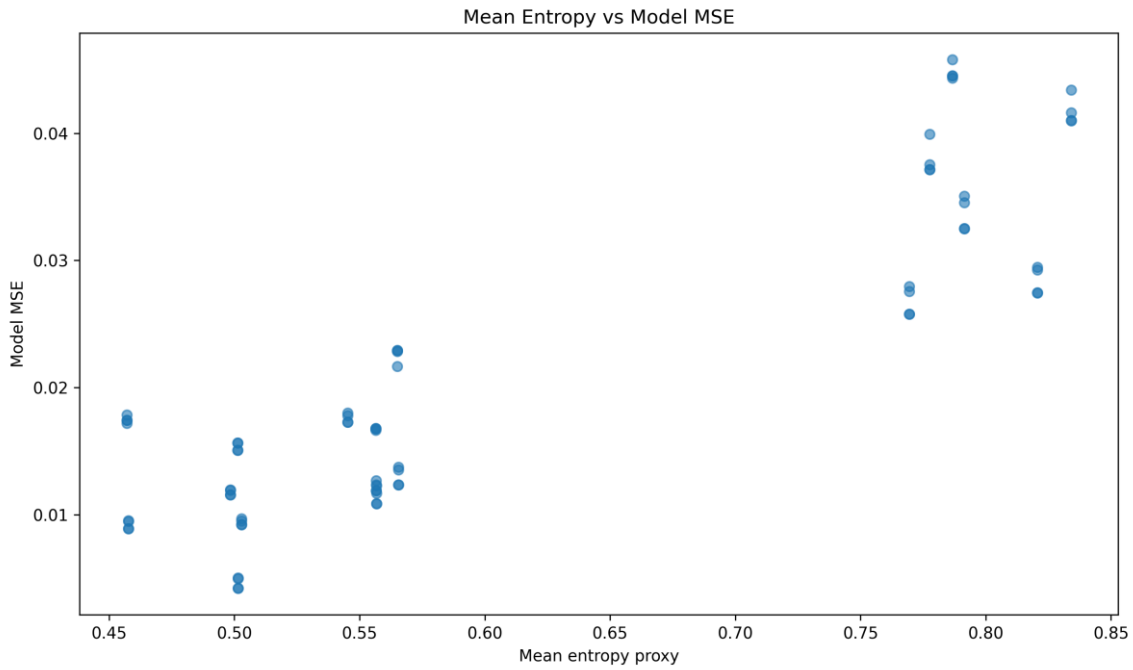
This figure shows a strong positive relationship between entropy and prediction error.

Observed behavior:

- higher entropy produces larger prediction error;
- structural chaos leads to increased instability;
- deployment robustness decreases as entropy rises.

This strongly supports the central entropy-sensitive degradation hypothesis of the framework.

Figure 5 — Mean Overlap vs Model MSE



mean_overlap_vs_model_mse.png

Caption

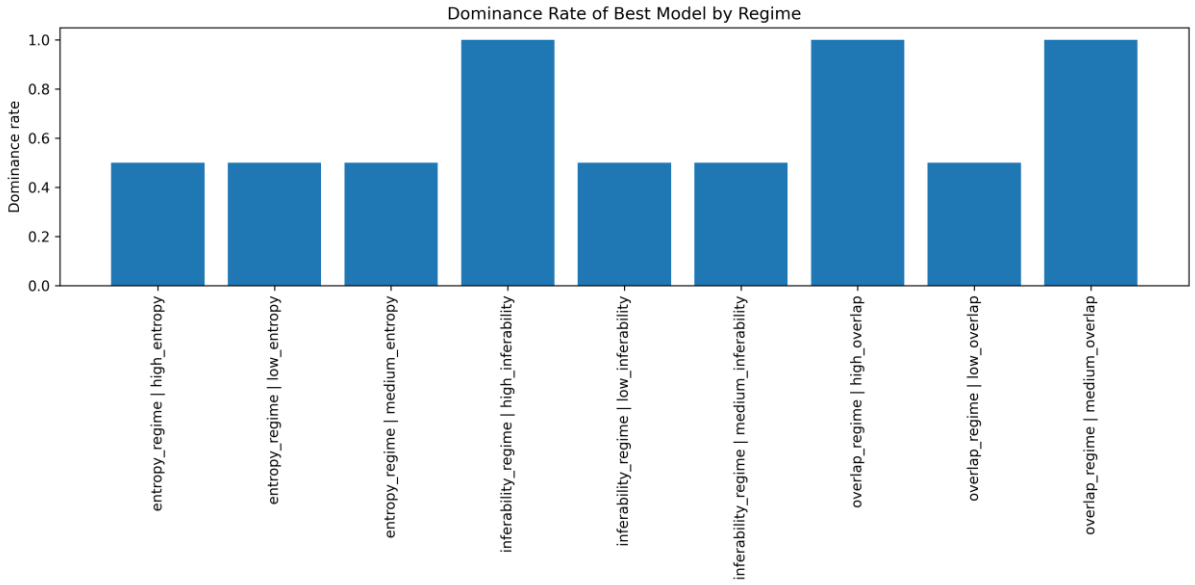
Overlap emerges as one of the strongest stability indicators in the entire benchmark.

Observed behavior:

- increasing overlap reduces prediction error;
- model consistency improves;
- generalization becomes more stable.

This suggests that overlap measures structural reproducibility rather than merely local similarity.

Figure 6 — Dominance Rate of Best Model by Regime



best_model_dominance_by_regime

Caption

This figure may be the most important visualization of the benchmark.

It reveals:

- some structural regimes produce stable model dominance;
- other regimes generate unstable model rankings;
- high-inferability regimes consistently favor a dominant best-performing model;
- high-overlap regimes produce stable model selection behavior;
- entropy-dominated regimes produce less stable rankings.

This suggests that structural signal organization can be used to predict model-family suitability before deployment.

Scientific Interpretation

This validation represents a major transition for the framework.

The framework now moves from:

Prediction feasibility detection

toward:

Pre-deployment model-family selection.

For the first time, inferability metrics appear to provide information not only about whether prediction is possible, but also about which model family is most likely to succeed.

Most Important Result

The strongest outcome of this benchmark is:

structural signal organization predicts model-family behavior.

Specifically:

- inferability predicts prediction stability;
- entropy predicts degradation;
- overlap predicts reproducibility;
- model-family performance shifts systematically with regime structure.

This substantially increases the practical value of the framework.

Industrial Relevance

This validation has direct implications for:

- predictive maintenance;
- condition monitoring;
- deployment screening;
- AI model selection;
- edge deployment;
- reliability engineering.

The framework now begins to answer a critical industrial question:

Which model family is likely to remain stable before development even starts?

This type of information is currently missing in many industrial AI workflows.

Conclusion

This validation demonstrates that inferability is associated not only with prediction success, but also with:

- model degradation;
- transfer robustness;
- model-ranking stability;
- and model-family suitability.

The results suggest that:

- inferability,
- entropy,
- and overlap

can be used as structural indicators for model selection before deployment.

This represents the strongest step so far toward a practical pre-deployment predictive-feasibility framework.

References (IEEE)

[1] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence,
Dataset Shift in Machine Learning,
MIT Press, 2009.

[2] R. Geirhos, J.-H. Jacobsen, C. Michaelis, et al.,
“Shortcut Learning in Deep Neural Networks,”
Nature Machine Intelligence, vol. 2, pp. 665–673, 2020.

[3] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar,
“Do ImageNet Classifiers Generalize to ImageNet?”
Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.

[4] S. Deng, Z. Xiao, and M. de Rijke,
“Domain Generalization in Time Series Forecasting,”
ACM Transactions on Knowledge Discovery from Data, 2024.

[5] J. Kim et al.,
“A Comprehensive Survey of Deep Learning for Time Series Forecasting: Architectural Diversity and Open Challenges,”
Artificial Intelligence Review, 2025.

[6] X. Kong et al.,
“Deep Learning for Time Series Forecasting: A Survey,”
Artificial Intelligence Review, 2025.

[7] K. Liao et al.,
“Deep Learning for Time Series Forecasting: A Survey of Recent Advances,”
Frontiers of Computer Science, 2026.

[8] A. Windmann, B. Stratmann, M. Lyashenko, and O. Niggemann,
“Industrial AI Robustness Card: Evaluating and Monitoring Time Series Models,”
arXiv, 2025.

[9] G. E. P. Box,
“Science and Statistics,”
Journal of the American Statistical Association, 1976.

[10] H. Kantz and T. Schreiber,
Nonlinear Time Series Analysis,
Cambridge University Press, 2004.

[11] E. N. Lorenz,
“Deterministic Nonperiodic Flow,”
Journal of the Atmospheric Sciences, 1963.

[12] Dryad Dataset:
“Recovering Mixtures of Fast Diffusing States from Short Single Particle Trajectories,”
DOI: 10.6078/D13H6N.

[13] NASA Ames Prognostics Center of Excellence,
Battery Aging Dataset.

[14] NASA C-MAPSS Turbofan Engine Degradation Simulation Dataset.