

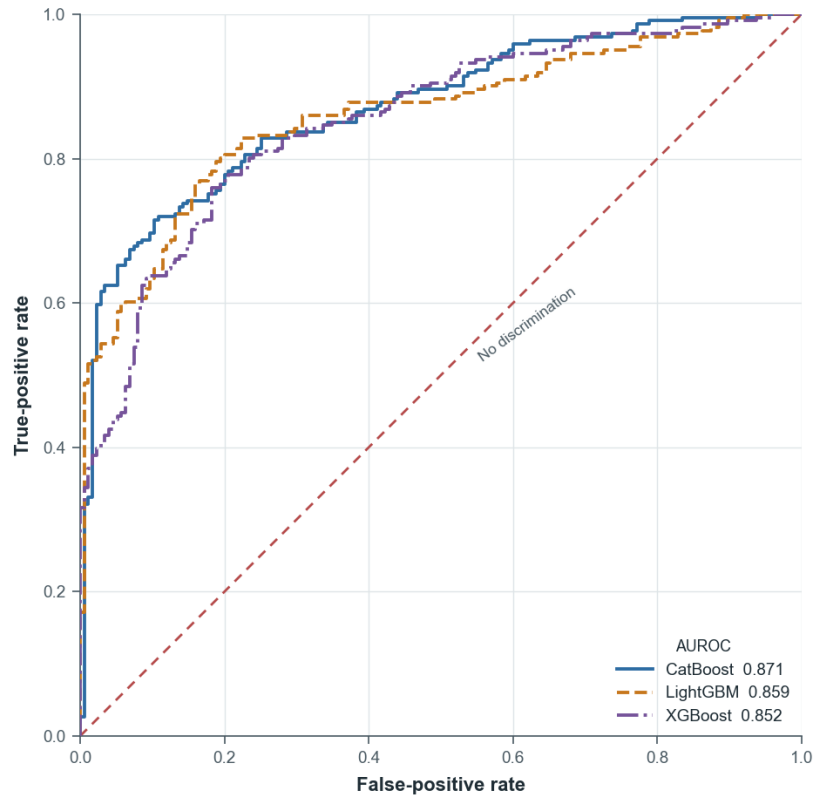
Supplementary Figures

Interpretable machine learning for classifying time-defined post-injury recovery stage from gait data after spinal cord injury
Canonical Workflow 00-15 outputs. Unless otherwise stated, figures use 396 records from 115 animals and strictly out-of-fold predictions.

Supplementary Figure S1A. Pooled out-of-fold ROC curve: Early boundary

A Pooled out-of-fold ROC · Early boundary

Event: Middle or Late · 396 records / 115 rats

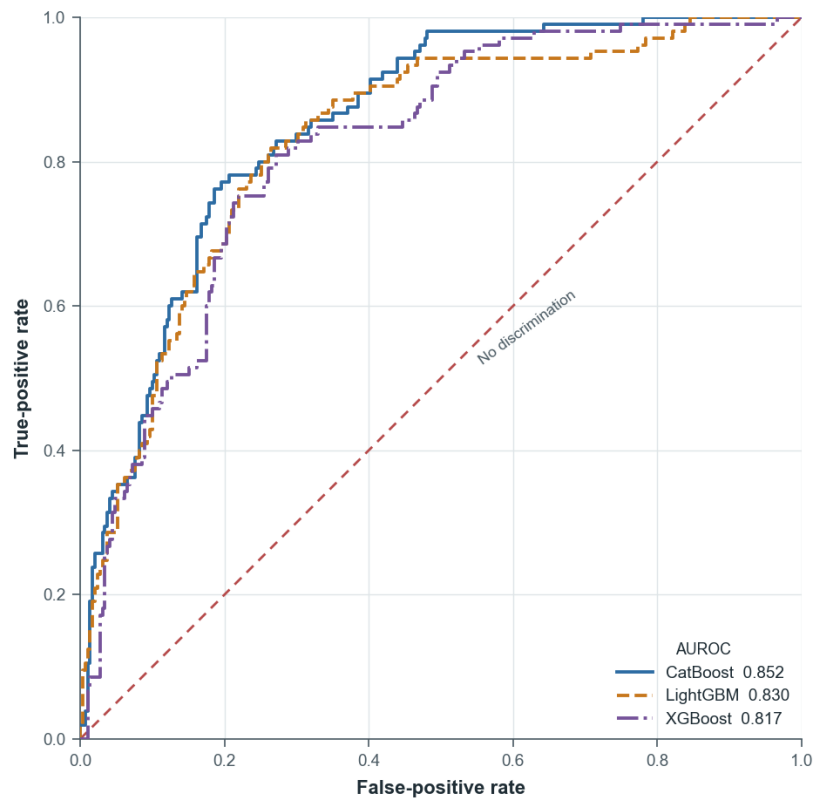


The cumulative Early boundary treated Middle/Late as the event and Early as the reference. Every record was predicted once by a model that had not been trained on that animal. AUROC was 0.871 for CatBoost, 0.859 for LightGBM, and 0.852 for XGBoost. The red dashed line denotes no discrimination.

Supplementary Figure S1B. Pooled out-of-fold ROC curve: Late boundary

B Pooled out-of-fold ROC · Late boundary

Event: Late · 396 records / 115 rats

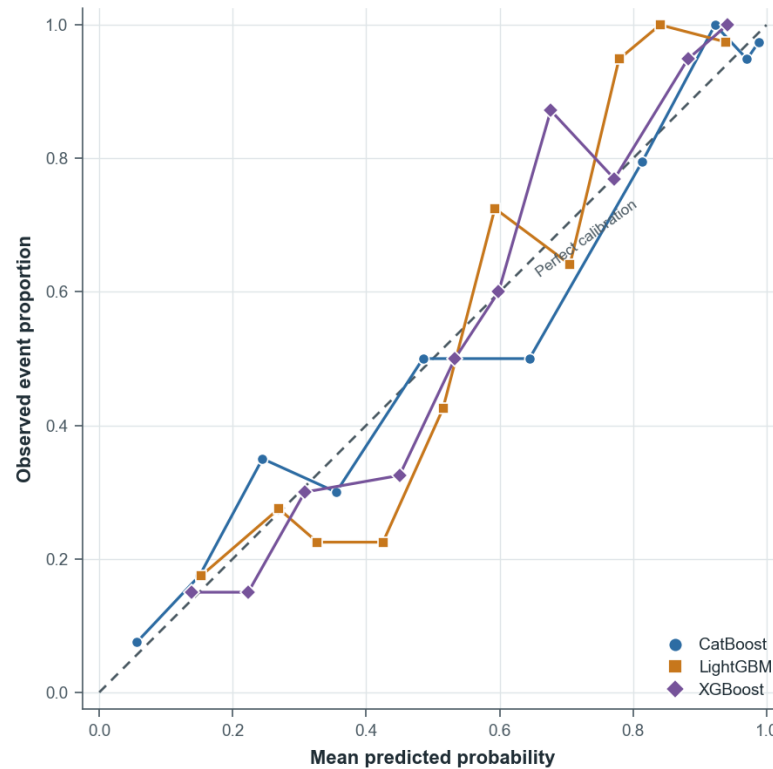


The cumulative Late boundary treated Late as the event and Early/Middle as the reference. AUROC was 0.852 for CatBoost, 0.830 for LightGBM, and 0.817 for XGBoost. Curves concatenate the five mutually exclusive animal-grouped outer test folds.

Supplementary Figure S2A. Pooled out-of-fold calibration: Early boundary

A Pooled out-of-fold calibration · Early boundary

Uncalibrated 26-feature models · decile bins · P(Middle or Late)

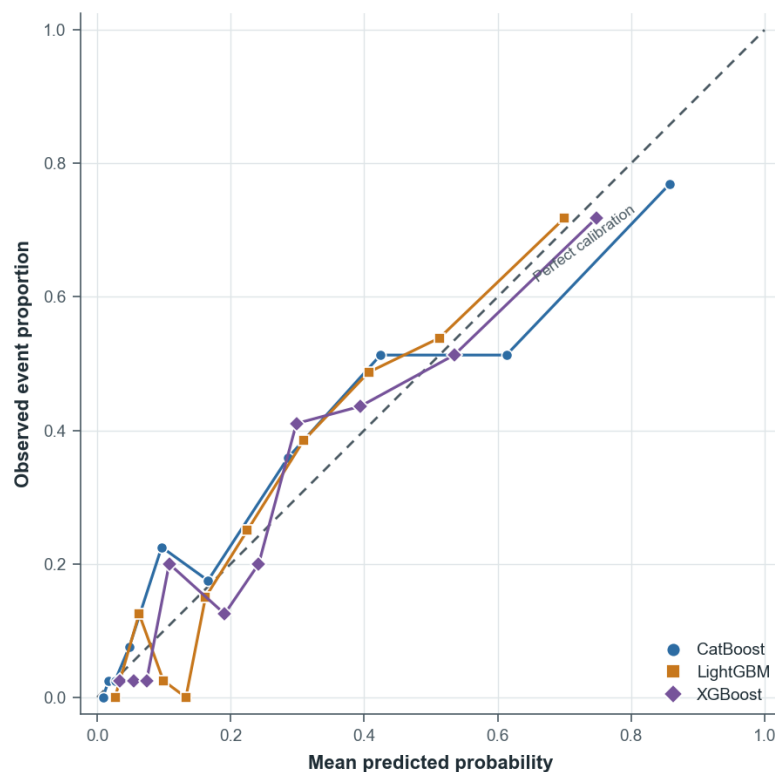


Observed Middle/Late event proportions are plotted against mean uncalibrated predicted probabilities within probability bins. The dashed line denotes perfect calibration. These are descriptive pooled OOF curves for all three 26-feature classifiers; calibration methods were evaluated separately within each framework using cross-fitted calibrators.

Supplementary Figure S2B. Pooled out-of-fold calibration: Late boundary

B Pooled out-of-fold calibration · Late boundary

Uncalibrated 26-feature models · decile bins · $P(\text{Late})$

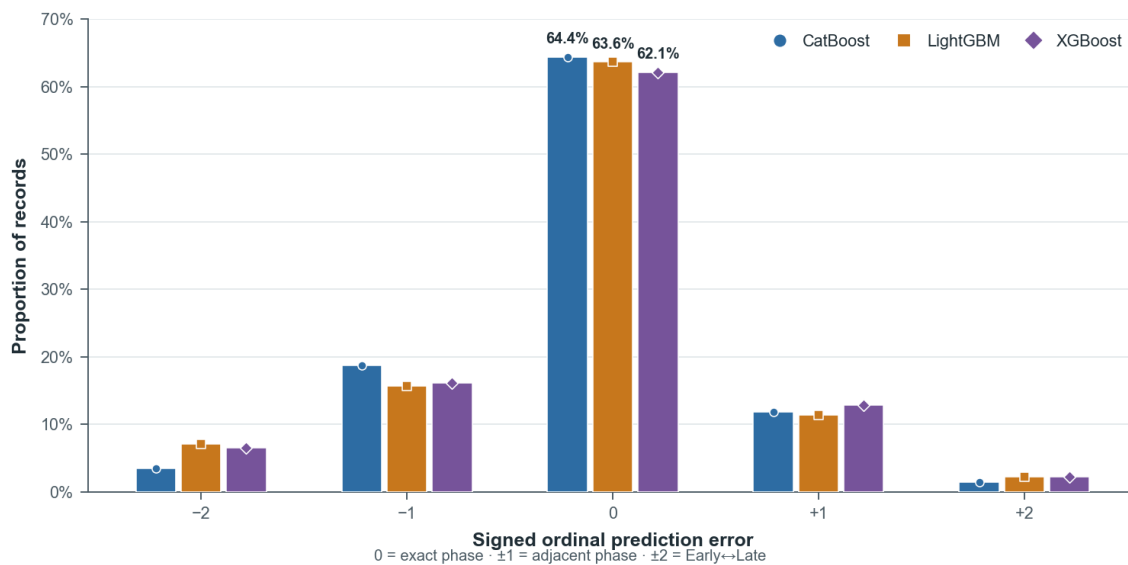


Observed Late event proportions are plotted against mean uncalibrated predicted probabilities within probability bins. The dashed line denotes perfect calibration. The lower event frequency at this boundary yields fewer observations in the higher-probability range.

Supplementary Figure S3. Ordinal prediction-error distribution

Ordinal prediction-error distribution

Predicted phase index – observed phase index · pooled out-of-fold predictions

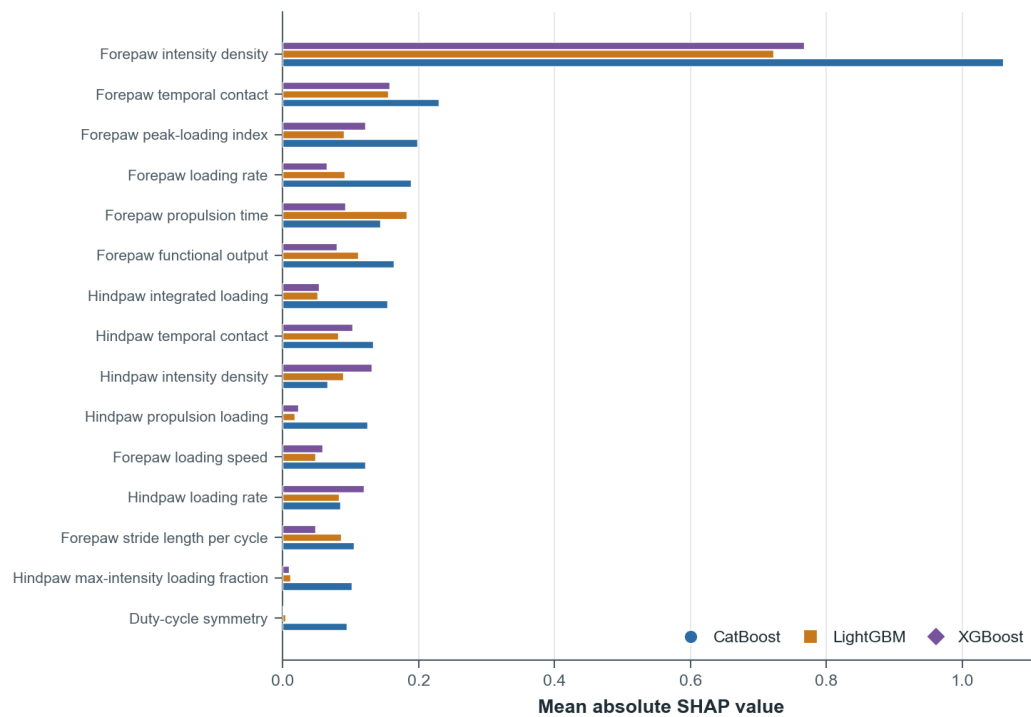


Prediction error is predicted ordinal index minus observed ordinal index: 0 is exact, ±1 is an adjacent-phase error, and ±2 is an Early-to-Late or Late-to-Early error. Exact-match rates were 64.4%, 63.6%, and 62.1% for CatBoost, LightGBM, and XGBoost, respectively.

Supplementary Figure S4. Cross-framework native TreeSHAP importance: Late boundary

Cross-framework SHAP importance · Late boundary

Held-out outer-fold records · top 15 features · raw-margin scale

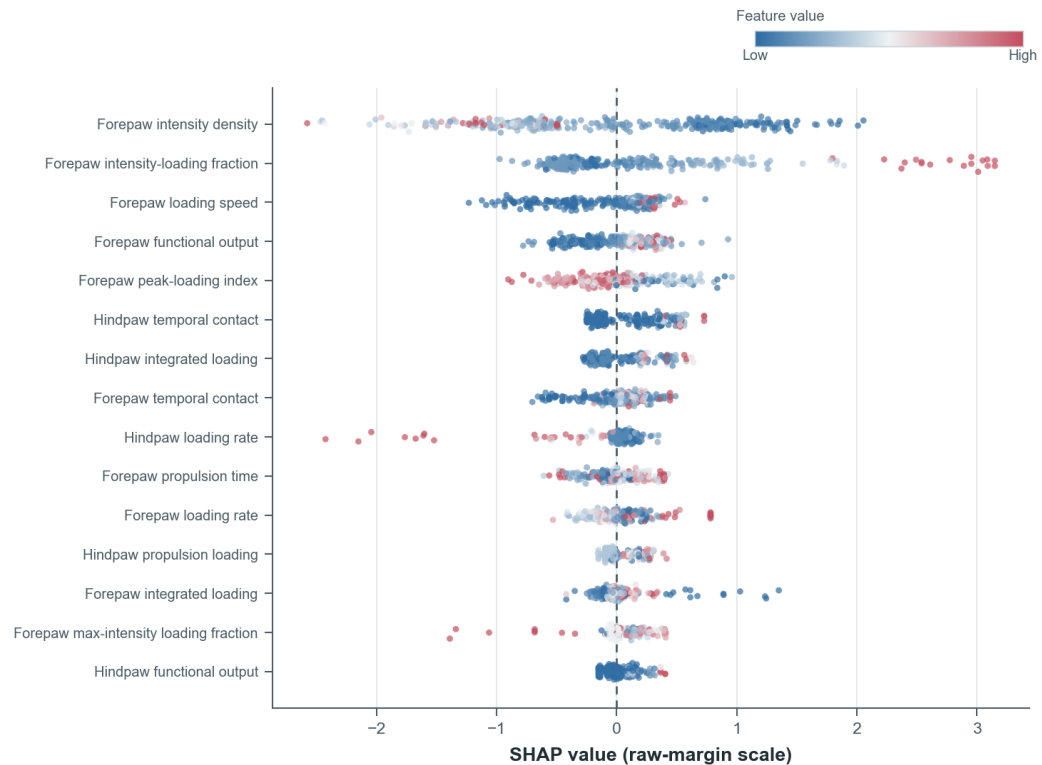


Bars show mean absolute native TreeSHAP values for $q_2=P(\text{Late})$, aggregated only over held-out outer-fold predictions from the 15 fitted framework-fold models. Values are on each estimator's raw-margin scale and are therefore most appropriate for ranking within a framework. Additivity passed all 30 checks.

Supplementary Figure S5A. CatBoost held-out TreeSHAP summary: Early boundary

A CatBoost SHAP summary · Early boundary

Held-out outer-fold records · color indicates observed feature value

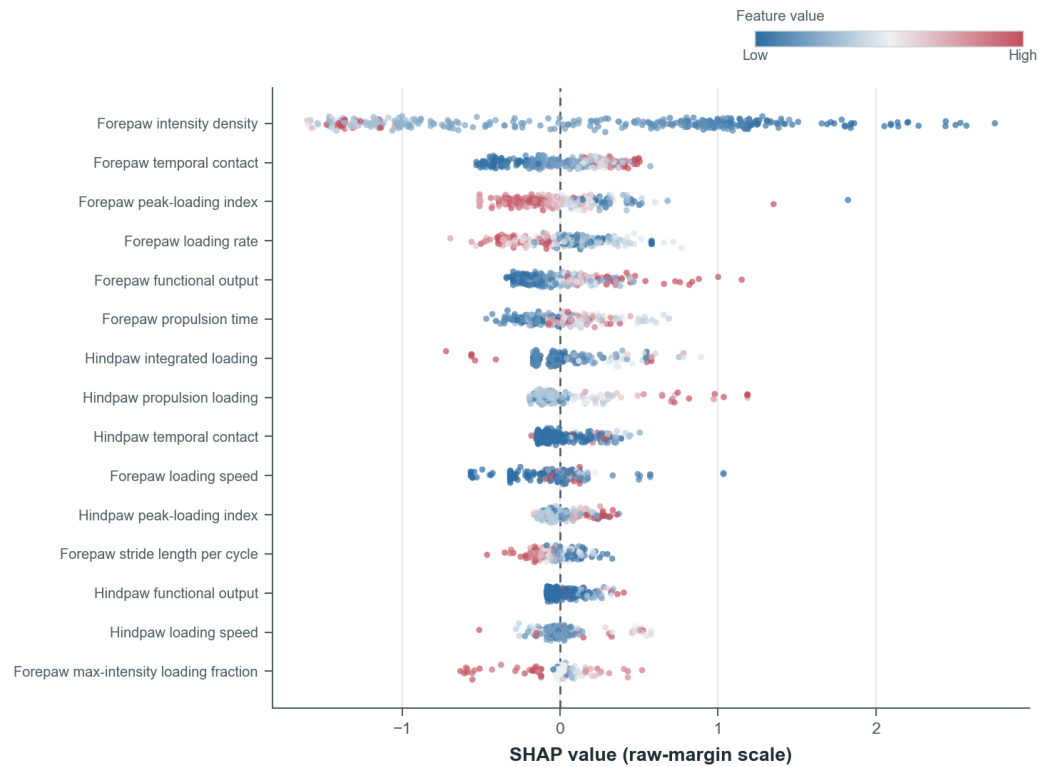


CatBoost $q1=P(\text{Middle or Late})$. Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S5B. CatBoost held-out TreeSHAP summary: Late boundary

B CatBoost SHAP summary · Late boundary

Held-out outer-fold records · color indicates observed feature value

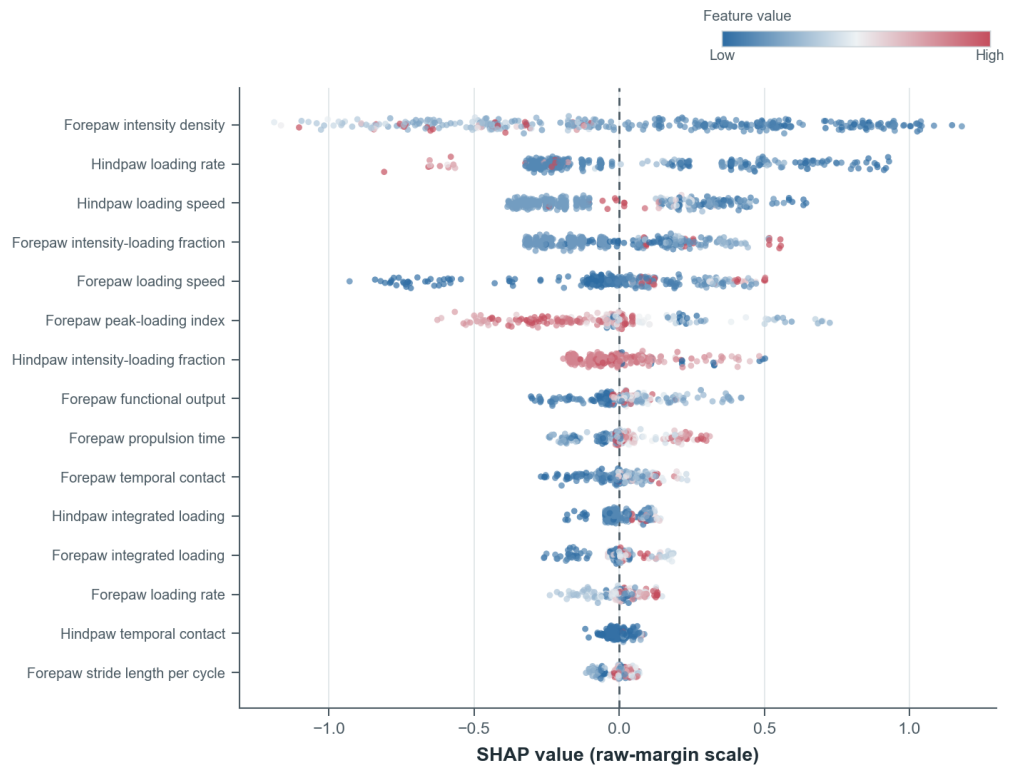


CatBoost $q_2=P(\text{Late})$. Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S6A. LightGBM held-out TreeSHAP summary: Early boundary

A LightGBM SHAP summary · Early boundary

Held-out outer-fold records · color indicates observed feature value

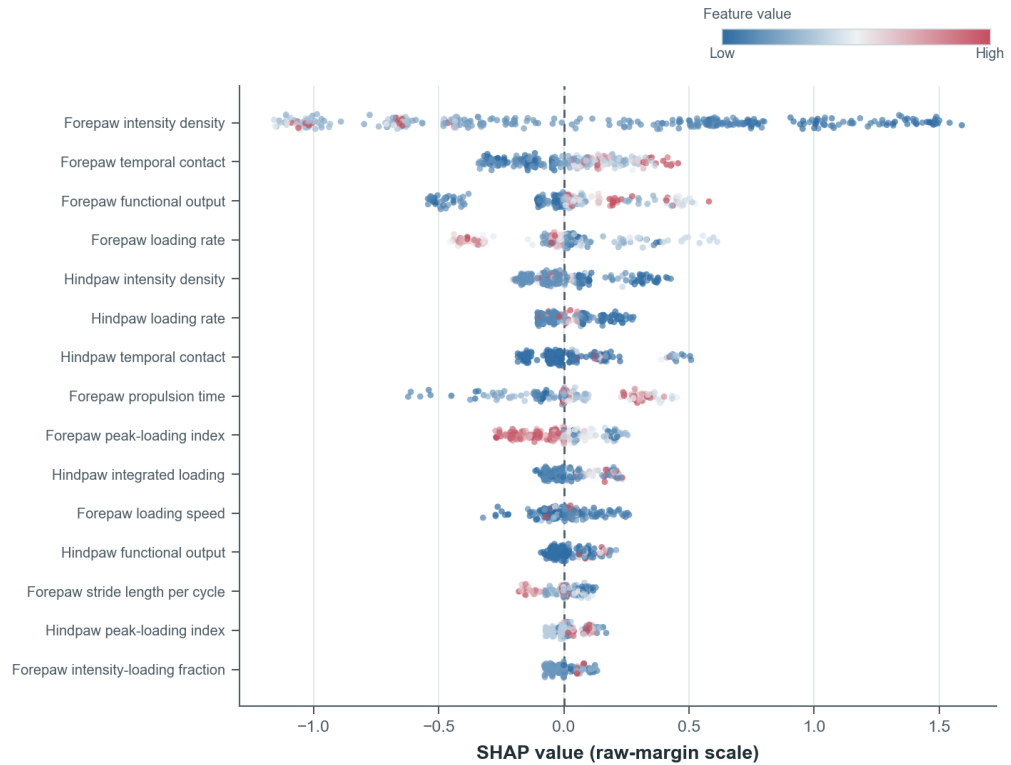


LightGBM q1=P(Middle or Late). Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S6B. LightGBM held-out TreeSHAP summary: Late boundary

B LightGBM SHAP summary · Late boundary

Held-out outer-fold records · color indicates observed feature value

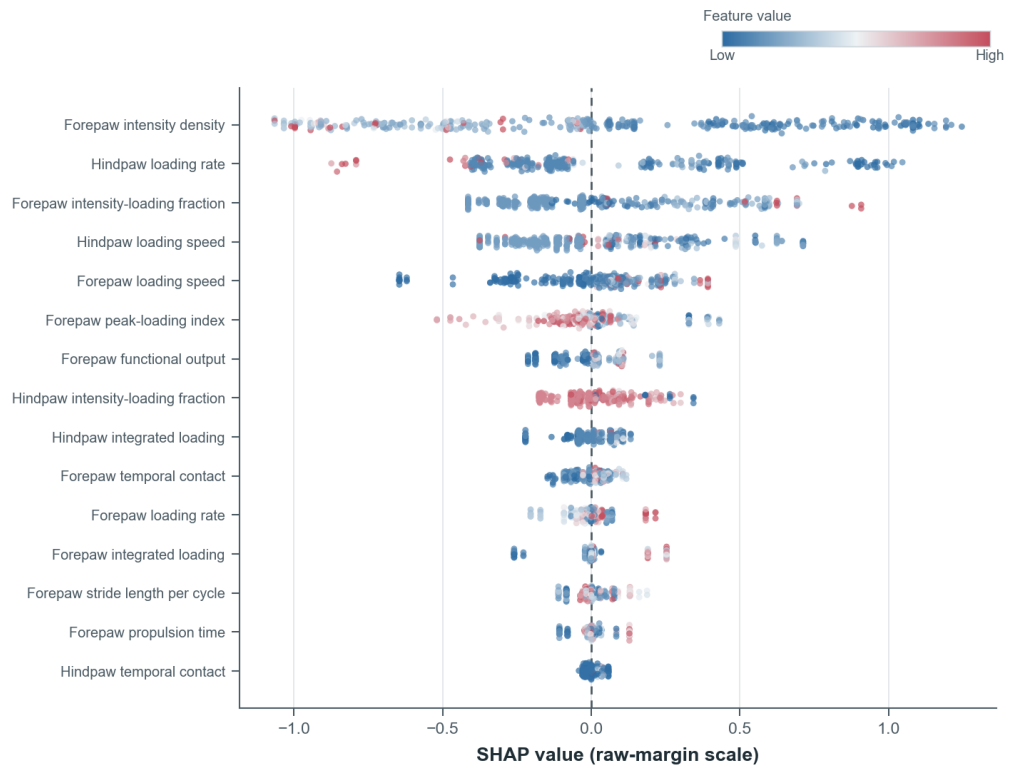


LightGBM $q_2=P(\text{Late})$. Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S7A. XGBoost held-out TreeSHAP summary: Early boundary

A XGBoost SHAP summary · Early boundary

Held-out outer-fold records · color indicates observed feature value

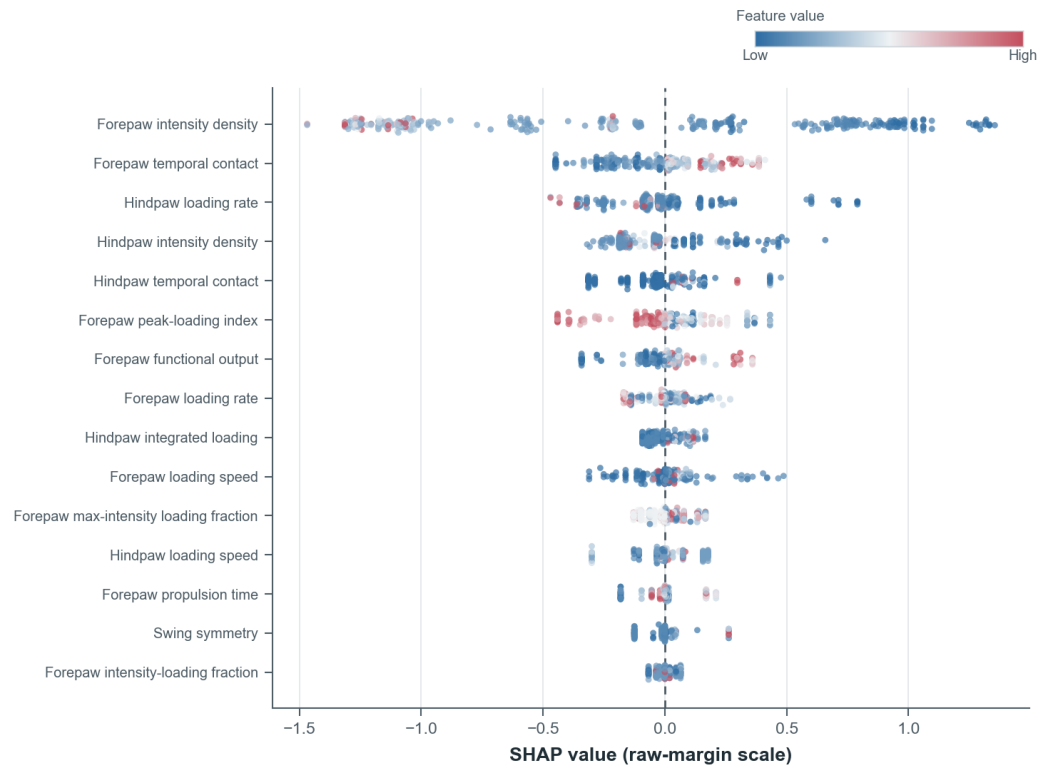


XGBoost $q1=P(\text{Middle or Late})$. Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S7B. XGBoost held-out TreeSHAP summary: Late boundary

B XGBoost SHAP summary · Late boundary

Held-out outer-fold records · color indicates observed feature value



XGBoost $q_2=P(\text{Late})$. Each point is a held-out record. Horizontal position is the native TreeSHAP contribution on the raw-margin scale; color is the normalized observed feature value from low (blue) to high (red). Features are ordered by mean absolute contribution. Explanations are descriptive and do not imply causal effects.

Supplementary Figure S8. Full 26-feature ablation heatmap

Full 26-feature ablation

Δ mean cumulative AUROC (baseline – ablated) · positive values indicate loss

	CatBoost	LightGBM	XGBoost
Forepaw intensity density	+0.021	+0.022	+0.025
Forepaw intensity-loading fraction	-0.001	+0.002	+0.003
Hindpaw loading rate	-0.003	+0.006	+0.007
Forepaw loading speed	-0.002	+0.002	+0.003
Hindpaw loading speed	-0.005	+0.004	+0.004
Hindpaw temporal contact	+0.002	+0.003	+0.004
Hindpaw integrated loading	-0.001	+0.003	+0.002
Forepaw temporal contact	-0.001	+0.004	+0.003
Forepaw peak-loading index	-0.008	+0.003	+0.002
Forepaw loading rate	-0.002	+0.001	+0.002
Forepaw functional output	-0.009	+0.002	+0.003
Hindpaw intensity density	-0.002	+0.004	+0.004
Forepaw propulsion time	-0.004	+0.005	+0.005
Hindpaw propulsion loading	-0.001	+0.003	+0.006
Hindpaw intensity-loading fraction	-0.003	+0.002	+0.003
Forepaw stride length per cycle	-0.004	+0.003	+0.005
Forepaw integrated loading	-0.004	+0.003	+0.003
Hindpaw functional output	-0.009	+0.003	+0.002
Forepaw max-intensity loading fraction	-0.001	+0.002	+0.002
Stand symmetry	-0.003	+0.003	+0.004
Duty-cycle symmetry	-0.002	+0.003	+0.005
Swing symmetry	-0.003	+0.003	+0.002
Hindpaw stride length per cycle	-0.001	+0.003	+0.006
Hindpaw max-intensity loading fraction	-0.007	+0.003	+0.002
Hindpaw peak-loading index	-0.006	+0.002	+0.003
Hindpaw propulsion time	-0.003	+0.003	+0.005

Improved -0.025 No change Loss +0.025

Each cell is the pooled mean-AUROC loss after removing one engineered feature from the indicated framework and refitting on the same outer-training folds with saved model settings. Positive values indicate worse discrimination after removal; negative values indicate no loss or a small improvement. fp_intensity_density produced the largest consistent loss (0.021–0.025). Hyperparameters were not reoptimized per ablation.

Supplementary Figure S9. Feature-ranking correlation matrix

Feature-ranking agreement

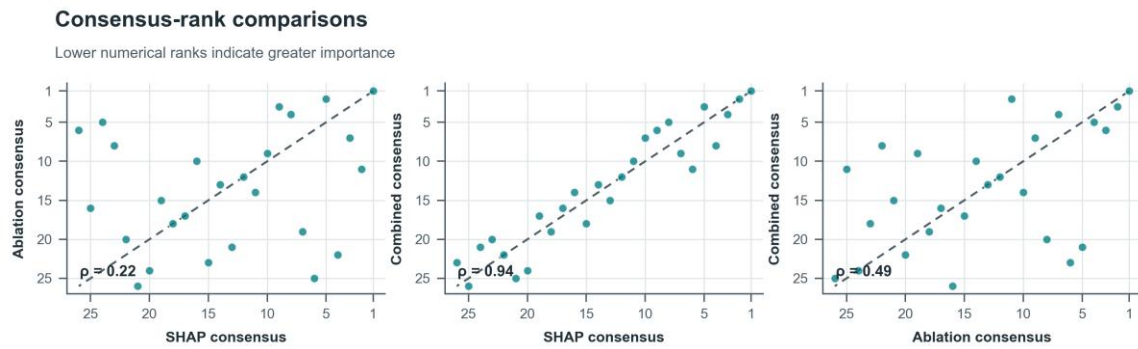
Pairwise Spearman ρ across all 26 engineered features



Pairwise Spearman correlations compare framework-by-outer-fold SHAP ranks, matched ablation ranks, their method consensus, and the equal-weight combined consensus over all 26 engineered features. Concordance was stronger among the 15 SHAP judges (Kendall $W=0.686$) than among the 15 ablation judges ($W=0.168$); SHAP- versus ablation-consensus ranks had $\rho=0.217$.

Supplementary Figure S10. Consensus feature-ranking diagnostics

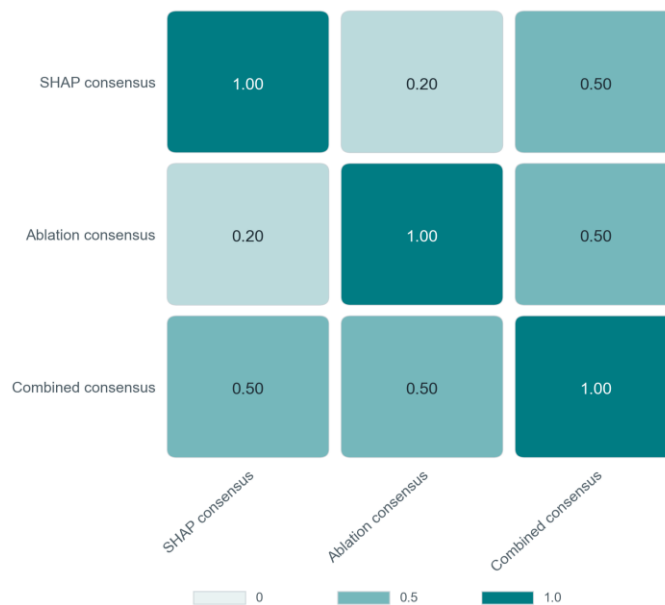
A



B

Top-3 consensus-ranking overlap

SHAP, ablation, and combined consensus; cells are Jaccard similarity

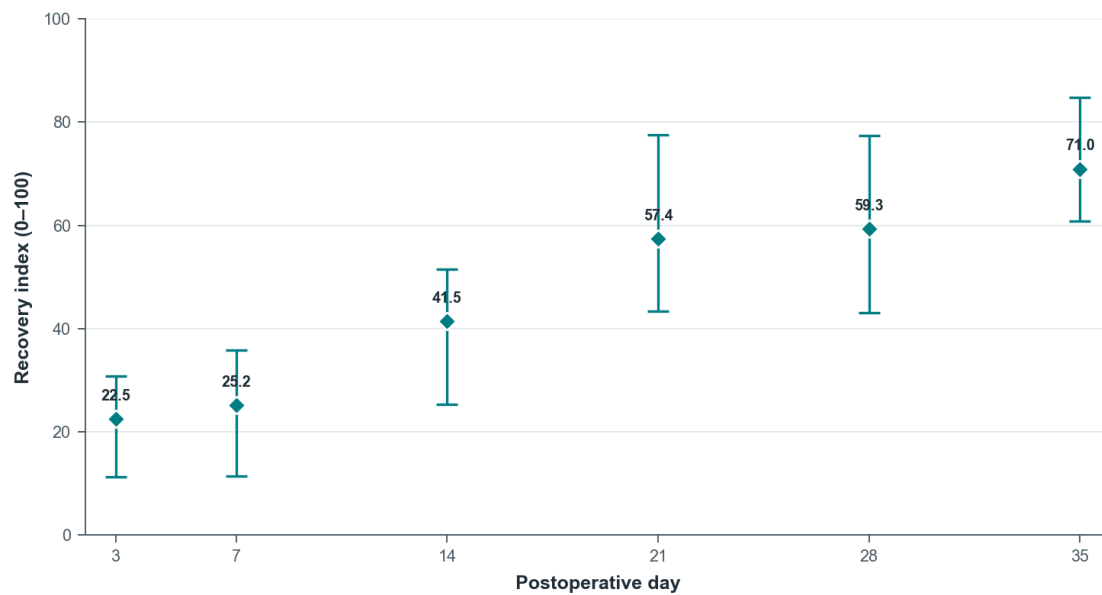


A, pairwise consensus-rank comparisons; lower numerical ranks denote greater importance. Spearman rho was 0.945 for SHAP versus combined consensus, 0.491 for ablation versus combined consensus, and 0.217 for SHAP versus ablation consensus. B, overlap of the top-three lists; cell values are Jaccard similarity. The combined consensus shared two of three features with each method consensus, whereas the SHAP and ablation consensus shared one.

Supplementary Figure S11. Ensemble recovery index over observed timepoints

Ensemble recovery index over observed timepoints

Mean with interquartile range · Platt-calibrated 3-feature ensemble



Timepoint summaries are independent; no longitudinal trajectory is implied.

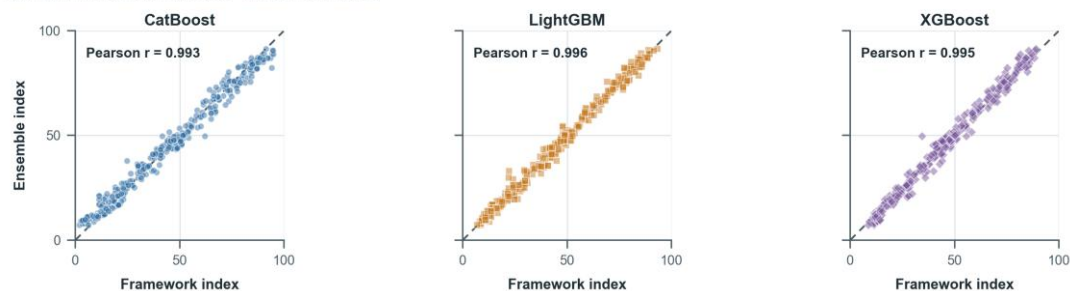
Diamonds show the mean equal-weight ensemble recovery index at each observed timepoint, and vertical whiskers span the first to third quartiles. The index is $50 \times P(\text{Middle}) + 100 \times P(\text{Late})$, using Platt-calibrated out-of-fold probabilities from the three post-selection three-feature frameworks. Repeated measurements are retained, so this descriptive trajectory is not an independent longitudinal effect estimate.

Supplementary Figure S12. Recovery-index framework agreement and disagreement

A

Framework-ensemble agreement

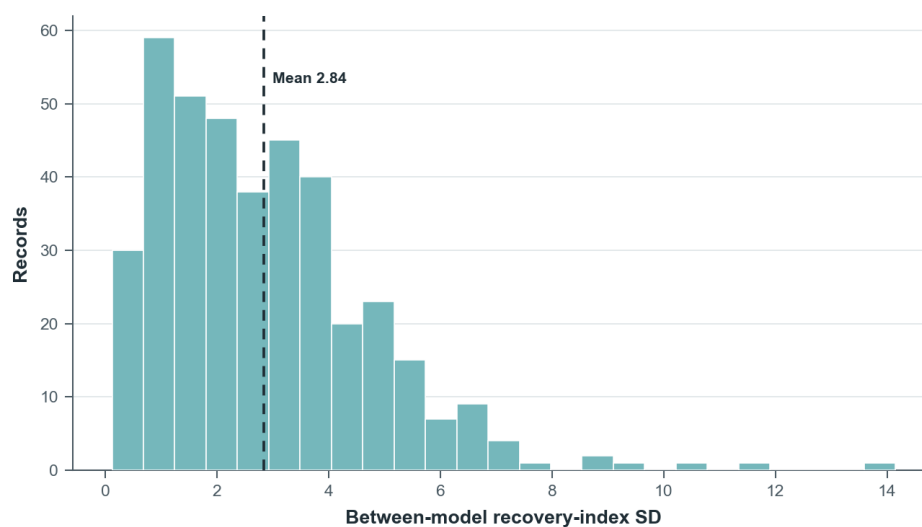
Each point is one out-of-fold record · dashed line is identity



B

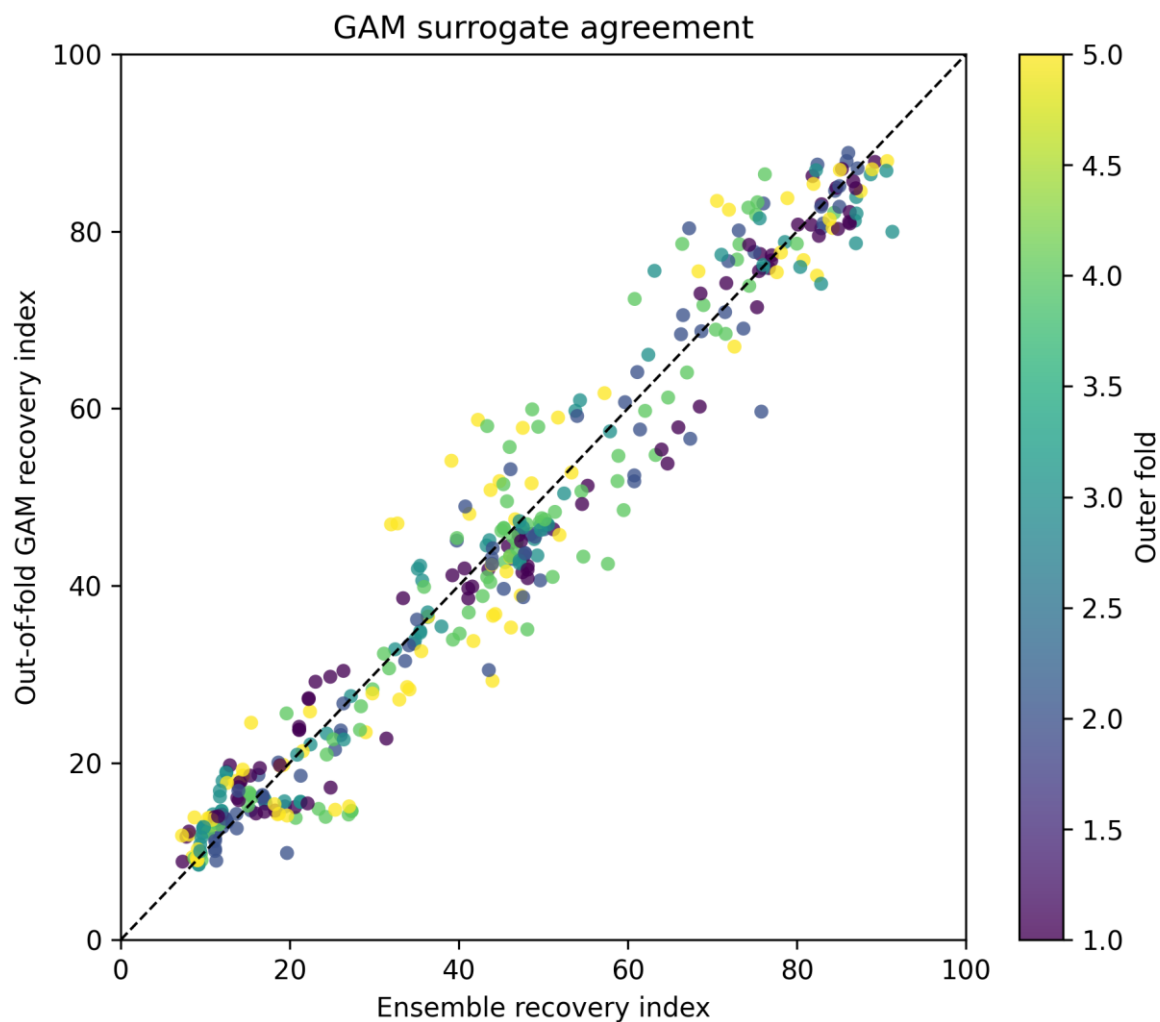
Between-model disagreement

Record-level SD across CatBoost, LightGBM, and XGBoost recovery indices



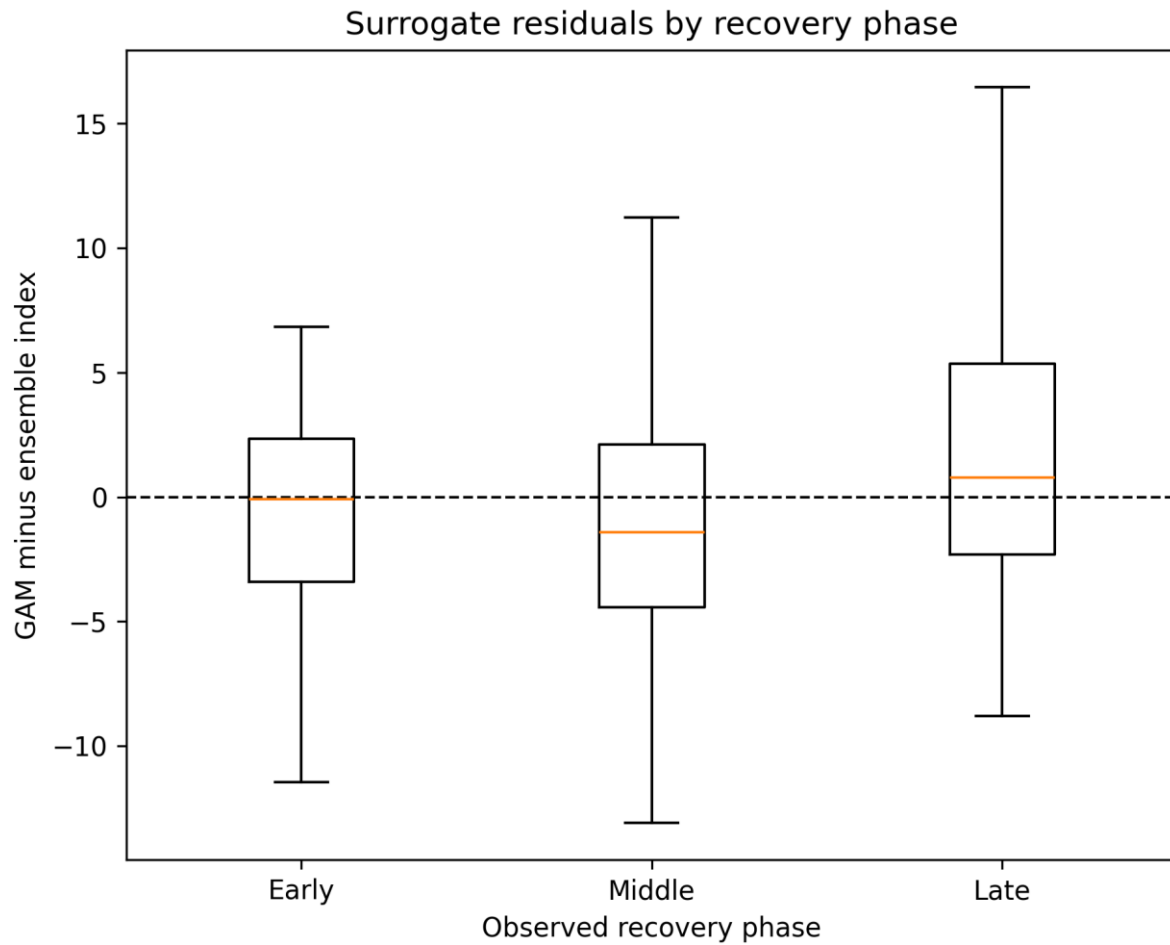
A, framework-specific recovery indices versus the equal-weight ensemble, with the dashed identity line; pairwise Pearson correlations among framework indices ranged from 0.980 to 0.992. B, distribution of the record-level standard deviation across the three framework indices. Mean between-model SD was 2.84 index points.

Supplementary Figure S13. Out-of-fold GAM surrogate agreement



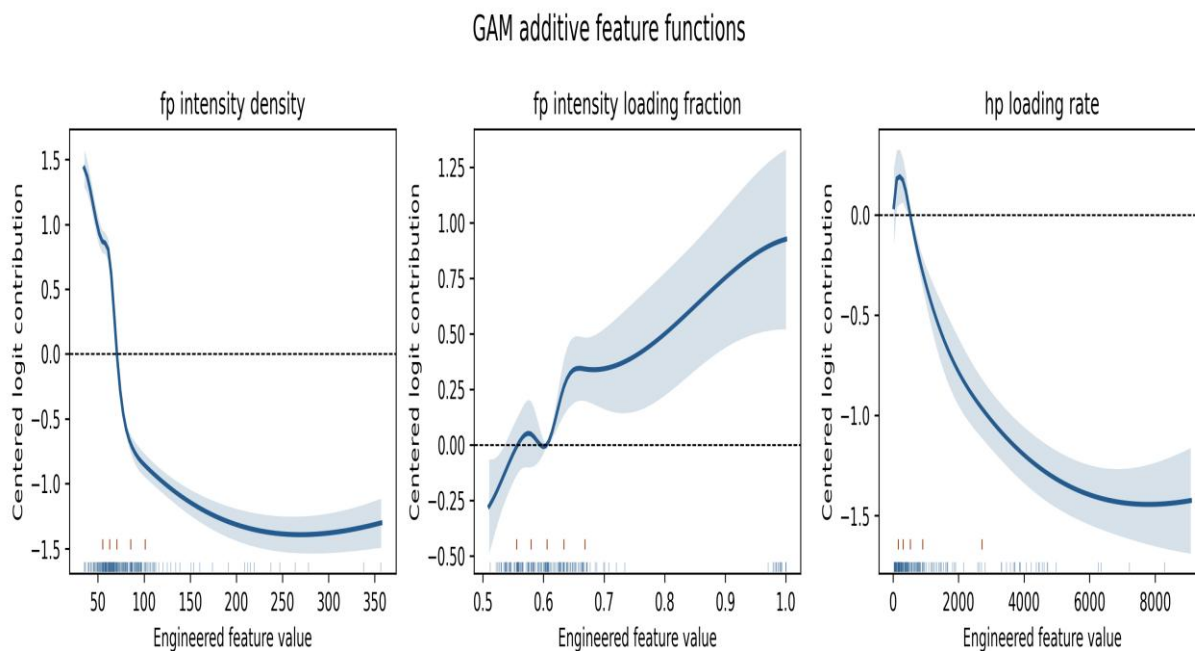
Each point compares the ensemble recovery index with its out-of-fold generalized additive model approximation; color identifies the held-out outer fold and the dashed line is identity. Fidelity estimates were $R^2=0.957$, $RMSE=5.26$, $MAE=4.01$, and Spearman $\rho=0.969$. The GAM approximates the ensemble index rather than observed phase labels.

Supplementary Figure S14. GAM surrogate residuals by recovery phase



Residual is out-of-fold GAM index minus ensemble recovery index; the dashed line denotes zero. Boxes summarize the phase-specific residual distributions with outliers suppressed for display. Phase-specific RMSE was 4.80 for Early, 5.60 for Middle, and 5.60 for Late.

Supplementary Figure S15. GAM additive feature functions



Functions from the final all-data GAM surrogate for the three consensus features used to approximate the ensemble recovery index: `fp_intensity_density`, `fp_intensity_loading_fraction`, and `hp_loading_rate`. Lines are centered additive contributions on the fitted logit scale; shaded regions are 95% conditional intervals. Curves describe the surrogate and should not be interpreted as causal or independently validated biological dose-response relationships.