Healthcraft Frontiers
https://www.acadlore.com/journals/HF

Leakage-Safe Machine Learning and Explainable Artificial Intelligence for Baseline Proteomic Signal Prioritization in Preclinical Rheumatoid Arthritis

Turker Berk Donmez[1], Mohammed Mansour[2*]

[1] Department of Biomedical Engineering, Sakarya University of Applied Sciences, 54050 Sakarya, Turkey
[2] Department of Mechatronics Engineering, Sakarya University of Applied Sciences, 54050 Sakarya, Turkey

[*] Correspondence: Mohammed Mansour (mohammedmansour@subu.edu.tr)

Received: 01-10-2026 **Revised:** 02-20-2026 **Accepted:** 03-04-2026

Citation: Donmez, T. B. & Mansour, M. (2026). Leakage-safe machine learning and explainable artificial intelligence for baseline proteomic signal prioritization in preclinical rheumatoid arthritis. *Healthcraft. Front.*, *4*(1), 1–20. https://doi.org/10.56578/hf040101.

Abstract: Reliable baseline biomarkers for progression toward rheumatoid arthritis in anti-citrullinated protein antibody-positive at-risk individuals remain insufficiently characterized. An exploratory leakage-safe machine learning framework combined with explainable artificial intelligence was developed to prioritize circulating proteomic signals associated with progression status in a small at-risk cohort. Baseline clinical and Olink proteomic data from 47 individuals (16 progressors and 31 non-progressors) were analyzed, although limited follow-up among non-progressors rendered the endpoint exploratory rather than prognostic. Of 1,472 quantified proteins, 1,449 were retained after application of a ≤20% missingness threshold. Fold-internal feature selection, including Cohen's *d*-based ranking, correlation filtering ($|r| < 0.85$), and top-30 protein selection, was embedded within repeated stratified five-fold cross-validation. Predictive performance remained modest, with the primary support vector classifier achieving a mean Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.675 and Precision-Recall Area Under the Curve (PR-AUC) of 0.447, while calibration remained weak (Brier score = 0.231; calibration slope = 0.455). A 500-iteration permutation audit was not statistically significant (p = 0.164). Regularized logistic regression failed to improve discrimination, whereas incorporation of routine clinical covariates did not yield a reproducible advantage over proteomic features alone. Extreme gradient boosting demonstrated lower discriminative performance and was retained only for secondary interpretability analyses. Across Tree-based SHapley Additive exPlanations (TreeSHAP), Kernel SHapley Additive exPlanations (KernelSHAP) for the support vector classifier, and bootstrap perturbation analyses, trefoil factor 2 (TFF2), KIT proto-oncogene receptor tyrosine kinase (KIT), cadherin 3 (CDH3), angiopoietin-like 2 (ANGPTL2), interleukin-5 (IL5), and glypican-1 (GPC1) emerged as recurrent candidate proteins. Given the limited cohort size, weak calibration, and non-significant permutation testing, all findings should be regarded as exploratory. The primary contribution therefore lies in the establishment of a transparent, leakage-aware workflow for proteomic signal prioritization in severely underpowered $p \gg n$ settings, thereby supporting future longitudinal validation studies in preclinical rheumatoid arthritis.

Keywords: Rheumatoid arthritis; Proteomics; Machine learning; SHapley Additive exPlanations; Explainable artificial intelligence

1. Introduction

Rheumatoid arthritis is a chronic immune-mediated inflammatory disease characterized by substantial heterogeneity in onset, progression, and therapeutic response. Contemporary rheumatoid arthritis research increasingly views the disease as a continuum extending from genetic and environmental susceptibility through systemic autoimmunity and symptomatic preclinical phases to clinically apparent inflammatory arthritis (Deane, 2024; Deane & Holers, 2021; Frazzei et al., 2023; O'Neil et al., 2024; Toyoda & Mankia, 2024). This conceptual shift is clinically important because biological perturbations that precede overt synovitis may offer an opportunity for earlier risk stratification, closer monitoring, and eventually preventive intervention (Deane, 2024; Deane &

Holers, 2021; O'Neil et al., 2024). Rather than treating rheumatoid arthritis only after persistent inflammatory arthritis becomes clinically evident, recent literature argues for identifying informative molecular and clinical signals during earlier phases of disease development (Frazzei et al., 2023; Toyoda & Mankia, 2024).

A major implication of this continuum model is that progression risk is not uniformly distributed among individuals with arthralgia, autoantibody positivity, or subclinical inflammatory abnormalities. Reviews focusing on prediction and prevention have shown that combinations of anticitrullinated protein antibodies, rheumatoid factor, musculoskeletal symptoms, and imaging findings outperform isolated markers when estimating the likelihood of future rheumatoid arthritis (Deane, 2024; O'Neil et al., 2024; Toyoda & Mankia, 2024). At the same time, prevention-oriented literature emphasizes that at-risk populations are biologically diverse: some individuals progress rapidly to clinically apparent disease, whereas others remain stable for prolonged periods or never convert at all (Deane & Holers, 2021; Frazzei et al., 2023). These observations suggest that progression-oriented studies should move beyond one-dimensional biomarkers and adopt multimodal frameworks capable of capturing heterogeneous disease trajectories. Within this context, biomarker research in rheumatoid arthritis has broadened substantially beyond traditional serological testing. Recent reviews describe an expanding landscape of diagnostic, prognostic, and management-oriented biomarkers spanning clinical variables, autoantibodies, imaging, transcriptomics, and proteomics (Sahin et al., 2025). Among these modalities, circulating proteomics is especially attractive because blood-based protein profiles can reflect immune activation, stromal remodeling, metabolic stress, and systemic inflammatory burden in a clinically accessible form (Cuesta-López et al., 2024; Jin et al., 2023; Sahin et al., 2025). In addition, targeted high-throughput assays and modern mass spectrometry workflows now make it feasible to characterize large protein panels with growing analytical reproducibility and translational relevance (Jin et al., 2023; Zhao et al., 2023).

Recent related work further indicates that proteomic alterations may be temporally informative rather than merely cross-sectional correlates of established disease. In individuals at risk of rheumatoid arthritis, serum proteomic networks have been linked to rheumatoid arthritis-related autoantibodies and longitudinal outcomes before the onset of inflammatory arthritis (O'Neil et al., 2022). More recent cohort studies have identified plasma or serum protein signatures that predate clinical rheumatoid arthritis, illuminate evolving biological pathways during disease development, and associate with activity or treatment-response phenotypes (He et al., 2025b; Zaim et al., 2025). Other proteomic studies have reported candidate biomarkers connected to cardiometabolic and inflammatory pathways in rheumatoid arthritis, as well as biologically distinct patient clusters derived from circulating protein signatures (Cuesta-López et al., 2024; Ferreira et al., 2023). Together, these findings support the idea that baseline proteomic measurements may contain informative signals relevant to subsequent disease progression. Proteomics has also become increasingly relevant to precision medicine in rheumatoid arthritis. Beyond risk prediction, recent studies have identified blood-based protein candidates associated with methotrexate resistance, disease activity, and mechanistically distinct patient subgroups (Escal et al., 2024; Lewis, 2024; Rivellese et al., 2022). More broadly, inflammatory protein studies using targeted proteomic platforms have demonstrated that circulating proteins can reveal causal pathways and potential therapeutic targets across immune-mediated diseases (Zhao et al., 2023). This broader literature is important for progression modeling because it establishes two principles: first, clinically meaningful heterogeneity in rheumatoid arthritis is biologically structured; second, molecular measurements can complement patient-level clinical information in ways that may improve prediction and biological interpretability (Lewis, 2024; Rivellese et al., 2022).

Machine learning has emerged as a natural analytical framework for modeling this complexity. In rheumatoid arthritis, recent studies have applied machine learning to identify inadequate responders to methotrexate, predict biologic inefficacy, estimate remission probability in registry cohorts, and screen for difficult-to-treat disease phenotypes (Alsaber et al., 2024; Baloun et al., 2025; Duquesne et al., 2023; Sonomoto et al., 2024; Ukalovic et al., 2024). These studies show that nonlinear models can integrate interacting clinical and laboratory variables more flexibly than conventional approaches and can support individualized risk estimation in rheumatology settings (Alsaber et al., 2024; Duquesne et al., 2023; Ukalovic et al., 2024). However, most of this literature has focused on treatment response, remission, or drug-related outcomes in established rheumatoid arthritis rather than on progression prediction from baseline clinical and molecular profiles in earlier disease states.

For predictive modeling to be clinically meaningful in rheumatology, performance alone is not sufficient. Black-box systems are difficult to trust when the goal is not merely forecasting but also understanding which signals are driving predictions and whether those signals are biologically plausible. Systematic reviews across healthcare consistently report that explainable artificial intelligence is increasingly important for transparency, accountability, and adoption in medical decision support (Ali et al., 2023; Alkhanbouli et al., 2025; Allgaier et al., 2023; Loh et al., 2022). Feature-attribution methods such as SHapley Additive exPlanations (SHAP) are particularly attractive because they allow strong nonlinear learners to be interpreted at both global and local levels, helping researchers quantify how individual features contribute to model output (Allgaier et al., 2023; Loh et al., 2022). In rheumatoid arthritis specifically, explainable machine-learning approaches have already been used for remission prediction, biologic ineffectiveness modeling, and adverse event prediction, demonstrating that interpretable machine-learning pipelines are feasible and clinically relevant in this domain (Alsaber et al., 2024; Jang et al., 2025;

Ukalovic et al., 2024).

Despite these advances, an important gap remains at the intersection of rheumatoid arthritis progression modeling, circulating proteomics, and explainable machine learning. The recent literature strongly supports early prediction within a disease-continuum framework (Deane, 2024; Deane & Holers, 2021; O'Neil et al., 2024), highlights the promise of blood-based proteomics for identifying preclinical and progression-related biology (He et al., 2025b; O'Neil et al., 2022; Zaim et al., 2025), and shows that machine learning and explainable artificial intelligence can generate clinically interpretable predictions in rheumatoid arthritis (Alsaber et al., 2024; Duquesne et al., 2023; Jang et al., 2025; Ukalovic et al., 2024). Yet comparatively few studies have integrated baseline clinical variables with targeted proteomic measurements in a way that prioritizes candidate baseline proteins while explicitly auditing leakage risk and small-sample instability.

Despite these advances, three lines of work, preclinical rheumatoid arthritis progression modeling, high-dimensional circulating proteomics, and explainable machine learning, have rarely been integrated within a single methodological framework that explicitly controls for information leakage in small-sample, high-dimensional settings. Most existing rheumatoid arthritis prediction studies operate on established disease, on treatment-response endpoints, or on lower-dimensional feature sets where naive global feature selection introduces only modest bias. Few studies have combined baseline clinical variables with targeted proteomic measurements while simultaneously embedding feature selection inside the cross-validation loop, reporting permutation-based significance, auditing calibration, and using explainable artificial intelligence to map candidate proteins back to biology. The specific contribution of the present study is to demonstrate such an end-to-end leakage-aware prioritization workflow on a 47-subject anti-citrullinated protein antibody-positive at-risk cohort with 1,449 baseline proteins, and to document, transparently, both what this workflow can and cannot deliver in the $p \gg n$ regime. The aim is therefore not to propose an externally validated clinical predictor, but to provide a reusable methodological template for narrowing a high-dimensional baseline proteomic feature space under endpoint uncertainty.

2. Methodology

2.1 Analytical Workflow Overview

Before describing each component in detail, this study summarizes the overall analytical pipeline so that readers unfamiliar with leakage-safe machine-learning workflows can follow the subsequent technical sections. Starting from the publicly available Allen Institute clinical and Olink proteomic tables, the pipeline (i) restricted the cohort to 47 anti-citrullinated protein antibody-positive at-risk subjects with progressor or non-progressor status, (ii) derived a single baseline visit per subject and applied an assay-level missingness filter of $\leq 20\%$ to retain 1,449 of 1,472 proteins, (iii) embedded fold-internal protein ranking (Cohen's d), correlation filtering ($|r| < 0.85$), and top-30 selection inside a repeated stratified five-fold cross-validation loop (10 repeats; 50 held-out folds), (iv) compared three predictor views (clinical-only, proteomic-only, and combined) across regularized logistic regression, support vector classification with radial-basis kernel, and a secondary extreme gradient boosting model, (v) summarized internal performance with subject-level out-of-fold discrimination, calibration, and a 500-iteration permutation audit, and (vi) applied a dual-model explainability strategy combining Tree-based SHapley Additive exPlanations (TreeSHAP) on extreme gradient boosting with Kernel SHapley Additive exPlanations (KernelSHAP) and permutation importance on the better-performing support vector classifier. Figure 1 depicts these stages and how feature selection is kept fully inside each training fold to avoid information leakage. As shown in the figure, feature ranking, correlation filtering, and top-k selection are performed strictly within each training fold of the repeated stratified five-fold cross-validation before any estimator is fitted. Discrimination, calibration, the 500-iteration permutation audit, and dual-model explainability (TreeSHAP on extreme gradient boosting, KernelSHAP and permutation importance on the support vector classifier) operate on these fold-internal selections.

2.2 Study Design and Analytical Objectives

This investigation was conducted as a retrospective exploratory multimodal biomarker-prioritization study in an anti-citrullinated protein antibody-positive at-risk rheumatoid arthritis cohort. The primary objective was to test whether leakage-safe baseline modeling could prioritize candidate proteins for future validation while providing transparent internal discrimination estimates. Because follow-up among non-progressors was limited, the binary progressor/non-progressor label was treated as an immature exploratory status label rather than as a mature prognostic endpoint. The secondary objectives were (i) to compare focused clinical-only, proteomic-only, and combined baseline representations under a leakage-safe internal validation design; (ii) to assess how sensitive the extreme gradient boosting workflow was to different top-N protein subsets; and (iii) to characterize model behavior using SHAP-based global and model-comparison explanations. A final descriptive objective was to document longitudinal proteomic change patterns in the subset with repeated Olink sampling.

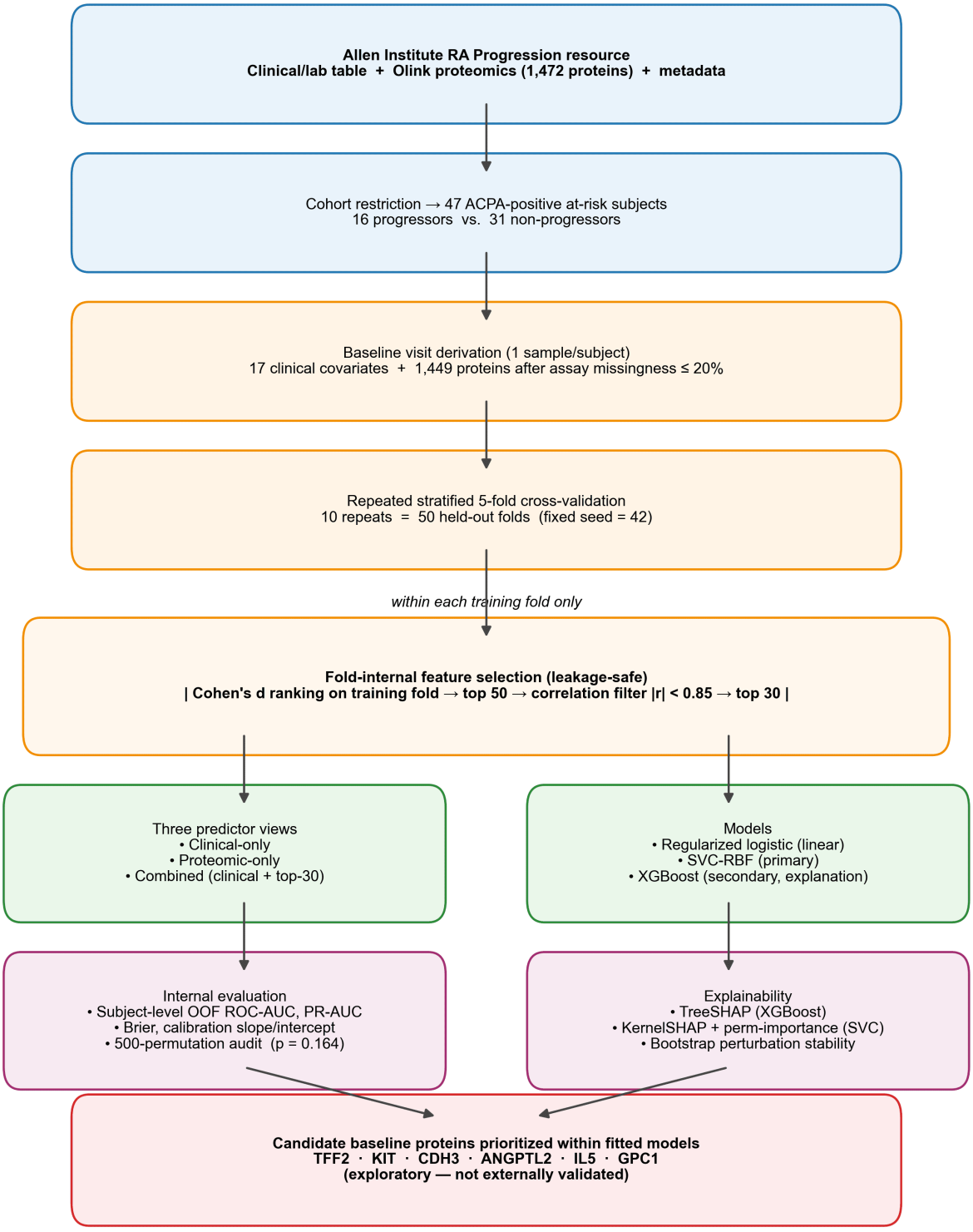

Figure 1. Schematic of the leakage-safe analytical workflow

Note: RA = rheumatoid arthritis; ACPA = anti-citrullinated protein antibody; OOF ROC-AUC = out-of-fold Receiver Operating Characteristic Area Under the Curve; PR-AUC = Precision-Recall Area Under the Curve; SVC-RBF = support vector classifier with radial basis function; XGBoost = extreme gradient boosting; TreeSHAP = Tree-based SHapley Additive exPlanations; KernelSHAP = Kernel SHapley Additive exPlanations; SVC = support vector classifier.

2.3 Source Data and Cohort Definition

Three curated source tables were integrated for analysis: a clinical/laboratory results table, a variable-descriptor table, and an Olink proteomics table. These files were downloaded from the Allen Institute for Immunology

Human Immune System Explorer "Systemic Inflammation in At-Risk Individuals Advancing to Clinical Rheumatoid Arthritis" project and its companion cohort report (Allen Institute for Immunology, 2025; He et al., 2025a), specifically from the downloadable Clinical Labs & Metadata and Plasma Proteomics resources. The analytic population was restricted to subjects classified as anti-citrullinated protein antibody-positive at-risk individuals and to the two longitudinal outcome strata, progressors and non-progressors. The primary study endpoint was encoded as a binary label, with 1 for progressors and 0 for non-progressors. For progressors, time from baseline sampling to diagnosis was available for descriptive summaries; for non-progressors, observed follow-up was summarized from the available visit timeline. Because many non-progressors had limited observed follow-up and only 16 progression events were available, the endpoint was retained as an exploratory baseline discrimination target rather than a full survival-model target. After cohort restriction and baseline derivation, the primary analytic cohort comprised 47 unique subjects, including 16 progressors and 31 non-progressors. The principal analytic subsets and feature-selection stages are summarized in Table 1. The top-50 ranking was recomputed within each training fold during internal validation, and the full-cohort export used for the final SHAP model retained 49 nonredundant proteins after correlation filtering.

Table 1. Summary of cohort derivation and multimodal feature-selection workflow

Analytic Quantity	Value
Baseline at-risk subjects	47
Progressors	16
Non-progressors	31
Baseline time window (days since first visit)	0–175
Total proteins in baseline Olink matrix	1,472
Proteins retained at missingness ≤ 20%	1,449
Top-ranked proteins requested by absolute Cohen's *d*	50
Proteins retained after correlation filtering (full-cohort export)	49
Proteins used in the final explanation dataset	30
Subjects with longitudinal follow-up (≥2 visits)	19
Longitudinal progressors	16
Longitudinal non-progressors	3

2.4 Baseline Visit Derivation and Outcome Encoding

To derive a single baseline observation per subject, the clinical table was first restricted to at-risk subjects belonging to one of the two progression groups of interest and then ordered by subject identifier, days since first visit, age at blood draw, and sample identifier. The earliest available visit for each subject was retained as the baseline sample. This procedure yielded a patient-level baseline cohort with one observation per individual and a baseline sampling interval spanning 0 to 175 days from the first recorded visit. Metadata retained for downstream integration included the subject identifier, sample identifier, disease group, progression group, days since first visit, and days to diagnosis when available.

2.5 Baseline Clinical Variables and Missing-Data Handling

The candidate baseline predictor set consisted of 17 numeric clinical/laboratory variables together with biological sex as a categorical covariate. These variables were chosen to capture demographic, anthropometric, inflammatory, hematologic, and routine chemistry domains (Table 2). Overall baseline missingness was limited, with the highest variable-level missingness equal to 4.26% before imputation. Missing numeric values were handled by median imputation. Median imputation was preferred over alternative approaches such as *k*-nearest neighbors imputation because overall baseline missingness was low (≤4.26% at the variable level), reducing the practical advantage of multivariate imputation methods, and because the median is inherently robust to the skewed distributions common in biomarker and serological variables, such as third-generation anti-cyclic citrullinated peptide antibody and rheumatoid factor immunoglobulin A, unlike mean imputation which would be distorted by extreme values. Given the small cohort size ($n = 47$), the simplicity and stability of median imputation also reduce the risk of introducing imputation-driven noise that could inflate downstream model performance, although multivariable imputation approaches such as multiple imputation by chained equations might better preserve correlation structure in larger cohorts. Biological sex was imputed using the most frequent category. Standardization was applied only for algorithms sensitive to scale, specifically logistic regression, radial-basis support vector classification, and *k*-nearest neighbors.

2.6 Olink Proteomic Preprocessing

Baseline proteomic profiling was derived from the Olink table after restriction to the same at-risk progression

cohort and alignment of proteomic records to baseline clinical samples using subject and sample identifiers. Protein measurements were represented as normalized protein expression values generated by a proximity extension assay platform (Assarsson et al., 2014). When duplicate entries for the same subject, sample, and assay were present, assay-specific normalized protein expression values were averaged before matrix construction. The filtered long-format proteomic data were subsequently reshaped into a subject-by-protein-wide matrix and merged with the baseline clinical dataset. The resulting baseline proteomic matrix contained 1,472 proteins, of which 1,449 satisfied the prespecified assay-level missingness threshold of ≤20% and were retained for downstream screening (Table 3).

Table 2. Baseline clinical covariates included in multimodal modeling

Domain	Variables
Demographic and anthropometric	Age at draw, biological sex, and body mass index
Inflammatory and serological markers	Third-generation anti-cyclic citrullinated peptide antibody, rheumatoid factor immunoglobulin A, rheumatoid factor immunoglobulin M, and erythrocyte sedimentation rate
Hematology	White blood cell count, absolute neutrophil count, absolute lymphocyte count, absolute monocyte count, platelet count, and hematocrit
Blood chemistry	Albumin, creatinine, alanine transaminase, glucose, and total protein

Table 3. Compact summary of Olink preprocessing and final baseline feature-space sizes

Olink/Feature-Processing Item	Value
Baseline samples aligned to Olink	47
Total proteins in baseline Olink matrix	1,472
Proteins retained after missingness filter (≤20%)	1,449
Top-ranked proteins requested by absolute Cohen's *d*	50
Proteins retained after full-cohort correlation filtering	49
Proteins used in final explanation dataset	30

2.7 Leakage-Safe Internal Validation Framework

Given the high dimensionality of the proteomic feature space relative to sample size, baseline protein prioritization was embedded directly within the internal validation workflow. At each resampling iteration, the data were split using repeated stratified five-fold cross-validation with shuffling and a fixed seed sequence (10 repeats; 50 held-out folds in total). Repeated rather than nested cross-validation was used because hyperparameters were prespecified rather than tuned; the goal was to characterize internal variability under fixed settings while preserving as much training data as possible in a 47-subject cohort. Within each training fold only, the retained proteins were ranked by absolute Cohen's *d* between progressors and non-progressors, as shown in Eq. (1):

$$d = \frac{\bar{x}_{CONV} - \bar{x}_{NONC}}{s_p}, s_p = \sqrt{\frac{(n_{CONV} - 1)s_{CONV}^2 + (n_{NONC} - 1)s_{NONC}^2}{n_{CONV} + n_{NONC} - 2}} \quad (1)$$

where, for a given protein within a training fold, $\bar{x}_{CONV}$ and $\bar{x}_{NONC}$ denote the mean Olink NPX values in progressors and non-progressors, respectively; s_{CONV} and s_{NONC} denote the corresponding group standard deviations; n_{CONV} and n_{NONC} denote the corresponding group sample sizes; and s_p denotes the pooled standard deviation.

The top 50 proteins from this fold-specific ranking were taken forward as candidate markers. To reduce redundancy, proteins were processed sequentially in ranked order and removed if their absolute pairwise correlation with any already retained higher-ranked protein was ≥0.85. This yielded a leakage-safe fold-specific nonredundant protein pool from which the top 30 proteins were used to define the primary combined model. This univariate screen was used for computational tractability within the repeated cross-validation loop, while acknowledging that it can miss proteins whose signal is primarily interaction-driven rather than individually strong.

Three predictor views were evaluated in each held-out fold: clinical-only, proteomic-only, and combined clinical + proteomic (top 30). The main text focuses on two primary model families: regularized logistic regression, representing a linear baseline, and support vector classification with a radial-basis kernel, representing a nonlinear interaction-sensitive classifier. Extreme gradient boosting was retained only as a secondary explanation-compatible comparator. Additional classifiers were examined in supplementary benchmarking but are not emphasized in the main narrative because the dataset is too small for broad algorithm shopping to be especially informative. Numeric predictors were processed through median-imputation pipelines; standardization was added

only for scale-sensitive models. Predicted probabilities were thresholded at 0.5 only for descriptive threshold-based metrics; because no threshold optimization or clinical utility analysis was performed, discrimination and calibration metrics were emphasized over threshold-derived summaries. Performance was summarized both across held-out folds and via subject-level mean out-of-fold predictions aggregated across repeats. In addition, the combined extreme gradient boosting workflow was re-evaluated across top-20, top-30, and top-49 protein settings to assess sensitivity to feature-set size. Stability analyses included fold-level selection frequencies, bootstrap perturbation summaries, and a leakage-safe logistic-regression coefficient check.

Model settings were prespecified rather than tuned. Conservative fixed settings were used throughout so that the analysis would remain focused on leakage-safe signal prioritization rather than on further optimization in a 47-subject cohort. No model family was optimized within this cohort beyond that single prespecified parameterization.

2.8 Final Explainability Model and Shapley Additive Explanations Workflow

After internal benchmarking, extreme gradient boosting was retained as one explanation model because SHAP TreeExplainer provides exact, efficient, and compositionally consistent local and global attribution for nonlinear tree ensembles (Chen & Guestrin, 2016; Lundberg et al., 2020). Recognizing that explaining a weaker classifier risks attributing noise rather than signal, the extreme gradient boosting TreeSHAP analysis was complemented with two model-agnostic explainability methods applied to the better-performing support vector classifier model: (i) permutation importance, which quantifies the decrease in Receiver Operating Characteristic Area Under the Curve (ROC-AUC) when each feature is randomly shuffled (Pedregosa et al., 2011), and (ii) KernelSHAP, a model-agnostic SHAP estimator that provides feature-level attribution without assuming a tree-based model structure (Lundberg et al., 2020). This dual-model explanation strategy allows the reader to assess whether the feature-importance signals identified by the extreme gradient boosting TreeSHAP workflow are corroborated by the better-performing support vector classifier, thereby mitigating the concern that extreme gradient boosting-based explanations are dominated by noise. For the full-cohort explanation fit, the same baseline ranking and correlation-filtering logic was applied to the complete baseline dataset, yielding 49 correlation-filtered proteins and a final top-30 multimodal matrix used for both extreme gradient boosting and support vector classifier explanation analyses.

The final extreme gradient boosting model was fit to the full top-30 multimodal dataset using median imputation for numeric variables and one-hot encoding for biological sex, without numeric standardization. SHAP values were then computed on the transformed feature matrix and mapped back to readable clinical and proteomic labels. In the main text, the explanation results were organized as a global SHAP summary, illustrative local exemplar plots, and a compact set of dependence plots, followed by cross-model comparison with the better-performing support vector classifier. In parallel, permutation importance and KernelSHAP were computed on the support vector classifier model fitted to the same top-30 feature matrix to provide a model-agnostic comparison of feature rankings.

2.9 Permutation Testing for Statistical Significance

To assess whether the observed cross-validated ROC-AUC of the best-performing support vector classifier model exceeded what would be expected by chance given the high-dimensional feature space and small sample size, a permutation test was conducted. The full leakage-safe repeated stratified five-fold cross-validation pipeline (including fold-internal protein ranking, correlation filtering, and top-30 selection) was executed 500 times with randomly permuted target labels. The empirical p-value was computed as $(k + 1)/(n_{\text{perm}} + 1)$, where k is the number of permutations runs in which the null ROC-AUC equaled or exceeded the observed ROC-AUC.

2.10 Internal Calibration and Batch-Sensitivity Analyses

Because calibration is a core dimension of prediction-model performance, the primary support vector classifier and regularized logistic regression models were additionally evaluated using subject-level mean out-of-fold predicted probabilities aggregated across the 10 repeated five-fold resamples. Internal calibration was summarized with the Brier score, calibration intercept, calibration slope, quintile-based calibration bins, and subject-level bootstrap 95% confidence intervals. Subject-level bootstrap 95% confidence intervals were also computed for ROC-AUC and Precision-Recall Area Under the Curve (PR-AUC) to provide a more transparent uncertainty summary than repeated-fold means alone. As a simple assay-wise batch sensitivity analysis, Olink normalized protein expression values were centered within each assay-by-batch combination using training-fold means only, after which the same leakage-safe support vector classifier pipeline was rerun. This analysis was intended as a sensitivity check for obvious batch dependence rather than as definitive batch correction.

2.11 Exploratory Longitudinal Delta Proteomic Analysis

The following exploratory analysis is presented only in the Appendix and should be interpreted strictly as hypothesis-generating rather than as a statistically powered longitudinal investigation. The longitudinal subset comprises only 19 subjects with repeated Olink measurements, of which only 3 are non-progressors. With a control group of $N = 3$, variance estimates and effect sizes (Cohen's d) are inherently unstable and statistically unreliable. This study retains this section solely to document observed patterns and motivate future studies with adequately sized longitudinal cohorts; no inferential conclusions should be drawn from these results.

Subjects with at least two Olink visits were identified, yielding a longitudinal subset of 19 subjects (16 progressors and 3 non-progressors). For each subject and assay, the earliest available Olink observation was designated as baseline and the latest available observation as follow-up after ordering by days since the first visit and sample identifier. Protein-specific longitudinal change was defined in Eq. (2):

$$\Delta NPX_{i,a} = NPX_{i,a}^{last} - NPX_{i,a}^{baseline} \tag{2}$$

where, i denotes subjects and a denotes Olink assays; $NPX_{i,a}^{last}$ denotes the normalized protein expression value for assay a at subject i's latest available Olink visit; and $NPX_{i,a}^{baseline}$ denotes the corresponding value at the baseline visit. Cohen's d was then recomputed for each assay using these delta values to descriptively rank proteins by differential temporal change between progressors and non-progressors. Given the extreme group imbalance ($n = 3$ non-progressors), these rankings should be regarded as anecdotal pattern summaries rather than as statistically meaningful effect-size estimates. The highest-ranked delta proteins were exported for qualitative comparison against baseline SHAP-prioritized proteins, and overlap and theme-summary tables were generated to characterize concordance and divergence between baseline and longitudinal molecular signatures.

2.12 Software and Statistical Reporting

All analyses were implemented in Python using standard scientific computing, machine-learning, and explainability libraries (Chen & Guestrin, 2016; Lundberg et al., 2020; Pedregosa et al., 2011; Prokhorenkova et al., 2018). Permutation importance was computed with 100 repeats, and KernelSHAP used k-means summarization of the background dataset ($k = 10$) with $n_{\text{samples}} = 200$. A fixed random seed of 42 was used throughout the reproducible workflow. No external validation cohort, nested hyperparameter optimization, or multiplicity-adjusted inferential testing was performed, and calibration was assessed only internally from repeated out-of-fold predictions. Accordingly, all biomarker signals and performance summaries should be interpreted as exploratory rather than confirmatory.

3. Results

3.1 Cohort Description and Baseline Characteristics

The baseline analytic cohort comprised 47 anti-citrullinated protein antibody-positive at-risk subjects, including 16 eventual progressors and 31 non-progressors. Baseline clinical differences were generally modest (Table 4). The largest standardized differences were observed for the third-generation anti-cyclic citrullinated peptide antibody (0.76), age at sampling (−0.58; younger in progressors), rheumatoid factor immunoglobulin A (0.50), and total protein (−0.37), whereas most routine hematologic and chemistry variables showed only small-to-moderate separation between groups. No baseline clinical variable reached conventional statistical significance in descriptive two-group testing; the smallest p-values were observed for age at sampling ($p = 0.086$) and the third-generation anti-cyclic citrullinated peptide antibody ($p = 0.106$). This limited baseline clinical separation foreshadowed the weak performance of clinical-only models. In the table below, values are reported as mean ± deviation for continuous variables and count (%) for biological sex. Standardized differences summarize imbalance between progressors and non-progressors. Descriptive p-values are from Mann–Whitney U tests for continuous variables and Fisher's exact test for biological sex.

3.2 Endpoint Timing, Internal Calibration, and Batch Sensitivity

Although the endpoint was encoded as a binary progressor/non-progressor label, the available timing information was asymmetric across groups. Among progressors, baseline samples preceded rheumatoid arthritis diagnosis by a median of 614.5 days (interquartile range 257.5–714.8; range 106–1144). In contrast, observed follow-up among non-progressors was limited (median 0 days, interquartile range 0–0; range 0–347), indicating that many non-progressors were observed only at baseline. This pattern reinforces that the present endpoint is better interpreted as an exploratory progression-status label than as a mature time-to-event outcome.

Internal calibration of the best-performing support vector classifier was also weak. Subject-level mean out-of-fold predictions yielded a Brier score of 0.231 (95% confidence interval 0.167–0.298), a calibration intercept of −0.393 (95% confidence interval −0.770 to 0.012), and a calibration slope of 0.455 (95% confidence interval 0.116–1.106) (Figure 2 and Table 5). These values suggest overconfident probability separation and argue against any clinically actionable interpretation of the current risk estimates. In a leakage-safe assay-wise batch-centering sensitivity analysis, the mean ROC-AUC of the support vector classifier remained similar to the main benchmark (0.664 ± 0.157 versus 0.642 ± 0.168), and 7 of the original top 10 ranked proteins were preserved after batch centering. The original top-10 candidate proteins also spanned four Olink panels (Cardiometabolic, Inflammation, Neurology, and Oncology), reducing the likelihood that the ranking was driven by a single-panel artifact. In Figure 2, the dashed diagonal indicates ideal calibration; the observed quintile-bin curve deviates materially from this line, consistent with the weak calibration intercept and slope reported in Table 5.

Table 4. Baseline clinical characteristics of the 47-subject at-risk cohort

Variable	Overall	Progressors	None-progressors	Standardized Difference	*p*
Age of subject at time of sample collection	56.79 ± 16.26	50.75 ± 16.22	59.90 ± 15.64	−0.58	0.086
Biological sex	Female: 37 (78.7%); male: 10 (21.3%)	Female: 13 (81.2%); male: 3 (18.8%)	Female: 24 (77.4%); male: 7 (22.6%)	0.04	1.000
Body mass index	27.13 ± 5.11	27.84 ± 5.36	26.76 ± 5.02	0.21	0.494
Third generation anti-cyclic citrullinated peptide antibody	427.57 ± 851.85	833.81 ± 1255.61	217.90 ± 436.96	0.76	0.106
Rheumatoid factor immunoglobulin A result	5.40 ± 13.74	9.82 ± 22.00	3.11 ± 5.63	0.50	0.417
Rheumatoid factor immunoglobulin M result	18.69 ± 32.57	23.92 ± 35.13	15.98 ± 31.42	0.24	0.661
Westergren erythrocyte sedimentation rate	17.02 ± 14.15	17.53 ± 12.88	16.77 ± 14.95	0.05	0.563
White blood cell count	6.64 ± 2.16	6.29 ± 2.31	6.81 ± 2.10	−0.24	0.306
Absolute neutrophil count	4060.00 ± 1575.58	3806.67 ± 1614.43	4186.67 ± 1567.94	−0.24	0.354
Absolute lymphocyte count	1880.00 ± 630.51	1826.67 ± 616.98	1906.67 ± 645.91	−0.13	0.682
Absolute monocyte count	486.67 ± 137.51	460.00 ± 91.03	500.00 ± 155.36	−0.29	0.362
Platelet count	280.73 ± 65.06	290.47 ± 73.38	275.87 ± 61.23	0.22	0.647
Hematocrit	44.31 ± 8.15	46.21 ± 13.44	43.36 ± 3.30	0.35	0.895
Albumin	4.31 ± 0.29	4.26 ± 0.35	4.33 ± 0.26	−0.23	0.255
Creatinine	0.80 ± 0.18	0.82 ± 0.19	0.79 ± 0.17	0.19	0.695
Alanine transaminase	16.96 ± 6.41	18.25 ± 8.10	16.27 ± 5.33	0.31	0.652
Glucose	95.59 ± 18.33	99.19 ± 27.56	93.67 ± 10.84	0.30	0.872
Protein, total	7.24 ± 0.48	7.12 ± 0.50	7.30 ± 0.47	−0.37	0.276
Days between first sample collection and collection of this sample	9.36 ± 36.80	0.00 ± 0.00	14.19 ± 44.79	−0.39	0.213

Table 5. Endpoint-timing and batch-sensitivity summaries supporting the endpoint reframing

Summary Item	Value
Progressor baseline-to-diagnosis, median (interquartile range) (days)	614.5 [257.5, 714.8]
Non-progressor observed follow-up, median (interquartile range) (days)	0.0 [0.0, 0.0]
Non-progressor observed follow-up range (days)	0–347
Batch-centered Receiver Operating Characteristic Area Under the Curve (ROC-AUC) for the support vector classifier	0.664 ± 0.157
Top-10 overlap after batch centering	7/10

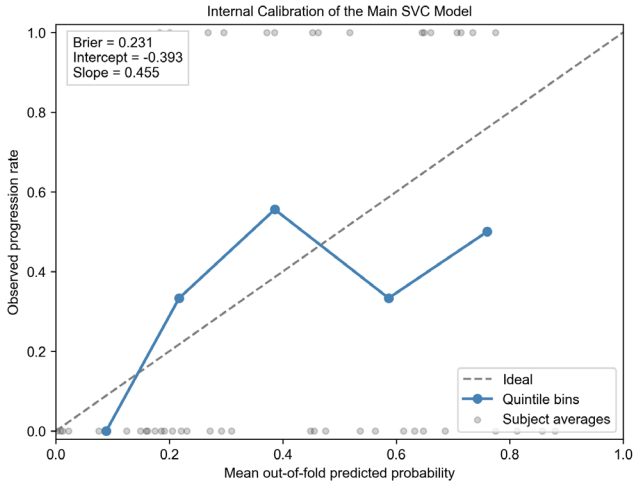

Figure 2. Internal calibration of the best-performing support vector classifier model using subject-level mean out-of-fold predicted probabilities

Note: SVC = support vector classifier.

3.3 Leakage-Safe Internal Benchmark Performance

Once protein ranking and filtering were moved inside the training folds, internal discrimination remained modest. Because repeated five-fold cross-validation creates non-independent fold means across repeats, the main results are reported from subject-level mean out-of-fold predictions aggregated across repeats, with fold means shown only as secondary context. The primary support vector classifier achieved a subject-level ROC-AUC of 0.675 (95% confidence interval 0.516–0.825) and PR-AUC of 0.447 (95% confidence interval 0.353–0.666), whereas regularized logistic regression reached 0.643 (95% confidence interval 0.484–0.796) and 0.433 (95% confidence interval 0.340–0.657), respectively (Table 6). Both models showed weak calibration, but the support vector classifier remained the strongest internally benchmarked model. Extreme gradient boosting, retained only as a secondary explanation-compatible model, achieved a mean fold ROC-AUC of 0.539 and was therefore not treated as a primary predictive analysis.

Table 6. Primary-model performance based on subject-level mean out-of-fold predictions aggregated across the 10 repeated five-fold resamples, with fold-wise Receiver Operating Characteristic Area Under the Curve (ROC-AUC) means only as secondary context

Model	ROC-AUC (95% Confidence Interval)	Precision–Recall Area Under the Curve (PR-AUC) (95% Confidence Interval)	Brier (95% Confidence Interval)	Calibration Slope (95% Confidence Interval)	Calibration Intercept (95% Confidence Interval)	Fold ROC-AUC Mean ± Standard Deviation
Support vector classifier (radial basis function)	0.675 (0.516–0.825)	0.447 (0.353–0.666)	0.231 (0.167–0.298)	0.455 (0.116–1.106)	−0.393 (−0.770–0.012)	0.642 ± 0.168
Regularized logistic regression	0.643 (0.484–0.796)	0.433 (0.340–0.657)	0.250 (0.185–0.317)	0.317 (0.018–0.776)	0.308 (−0.154–0.810)	0.606 ± 0.166

3.4 Modality Comparison and Top *N* Sensitivity

The modality comparison demonstrated that the proteomic signal was substantially more informative than the routine baseline clinical covariates, but that adding routine clinical covariates did not produce a stable gain over proteomics alone in the two primary model families (Table 7). For the support vector classifier, the combined clinical+proteomic representation achieved a slightly higher mean ROC-AUC than proteomics alone (0.641 vs. 0.622) but not a higher PR-AUC (0.558 vs. 0.564). For regularized logistic regression, the proteomic-only representation slightly outperformed the combined representation on both ROC-AUC (0.611 vs. 0.603) and PR-

AUC (0.542 vs. 0.535). In both families, the clinical-only representation was clearly weakest. Therefore, stable evidence, which the 17 routine clinical and laboratory covariates added incremental discriminative information beyond the proteomic baseline in this dataset, was not observed.

Changing the number of proteins in the combined extreme gradient boosting workflow did not materially alter performance (Table 8). Across top-20, top-30, and top-49 settings, mean ROC-AUC ranged only from 0.519 to 0.544 and mean Matthews correlation coefficient from −0.001 to 0.031. This relative flatness suggests that the present dataset supports feature prioritization more readily than sharp model optimization around a single top-*N* cutoff. In the table, values are means across 50 held-out folds. For the top-49 setting, the mean number of proteins actually retained per fold was 48.48 because fold-local correlation filtering occasionally produced fewer than 49 proteins.

Table 7. Comparison of baseline data modalities for the two primary model families, with means across 50 held-out folds as the values

Dataset	Model	Receiver Operating Characteristic Area Under the Curve (ROC-AUC)	Precision–Recall Area Under the Curve (PR-AUC)	Balanced Accuracy	Recall	Specificity	Matthews Correlation Coefficient
Proteomic only	Support vector classifier (radial basis function)	0.622	0.564	0.528	0.362	0.693	0.053
Clinical + proteomic (top 30)	Support vector classifier (radial basis function)	0.641	0.558	0.597	0.520	0.673	0.195
Clinical only	Support vector classifier (radial basis function)	0.380	0.423	0.480	0.013	0.946	−0.049
Proteomic only	Regularized logistic regression	0.611	0.542	0.564	0.390	0.739	0.125
Clinical + proteomic (top 30)	Regularized logistic regression	0.603	0.535	0.543	0.312	0.774	0.084
Clinical only	Regularized logistic regression	0.484	0.455	0.489	0.263	0.715	−0.032

Table 8. Sensitivity of the combined extreme gradient boosting workflow to the number of fold-selected proteins

Requested Top-N	Mean Selected per Fold	Receiver Operating Characteristic Area Under the Curve (ROC-AUC)	Precision–Recall Area Under the Curve (PR-AUC)	Matthews Correlation Coefficient
20	20.00 ± 0.00	0.519	0.479	−0.001
30	30.00 ± 0.00	0.539	0.487	−0.033
49	48.48 ± 0.76	0.544	0.493	0.031

Table 9. Top 10 features in the final extreme gradient boosting SHapley Additive exPlanations (SHAP) analysis together with fold-level and perturbation-based stability summaries

Feature	Mean	SHAP	Fold Selection Freq.	Higher Values Tended Toward	Fold Selection Frequency
Trefoil factor 2 (TFF2)	0.648	Non-progression	0.72	0.342	0.156
KIT proto-oncogene receptor tyrosine kinase (KIT)	0.528	Progression	0.82	0.402	0.212
Cadherin 3 (CDH3)	0.486	Progression	1.00	0.770	0.558
Angiopoietin-like 2 (ANGPTL2)	0.396	Progression	0.70	0.248	0.088
Interleukin-5 (IL5)	0.351	Progression	0.98	0.490	0.228
Glypican-1 (GPC1)	0.342	Progression	0.48	0.224	0.094
N-terminal pro–B-type natriuretic peptide (NTproBNP)	0.316	Non-progression	0.86	0.342	0.152
Oxytocin (OXT)	0.305	Progression	0.94	0.508	0.310
Phospholipase A2 group X (PLA2G10)	0.229	Non-progression	0.64	0.260	0.114
Fc gamma receptor IIIb (FCGR3B)	0.196	Progression	0.64	0.302	0.126

3.5 Global Shapley Additive Explanations Ranking, Directionality, and Feature Stability

These outputs are in-sample, model-dependent prioritization maps from a weak secondary model rather than association estimates. Although extreme gradient boosting was not a strong discriminator in the benchmark analysis (mean fold ROC-AUC 0.539), it provided the most tractable TreeSHAP workflow and was therefore retained only for secondary explanation. The global SHAP ranking was dominated by proteins rather than routine clinical variables (Table 9 and Figure 3). The top six features were trefoil factor 2 (TFF2), KIT proto-oncogene receptor tyrosine kinase (KIT), cadherin 3 (CDH3), angiopoietin-like 2 (ANGPTL2), interleukin-5 (IL5), and glypican-1 (GPC1). Higher TFF2 values tended to shift the fitted model toward non-progression, whereas higher KIT, CDH3, ANGPTL2, IL5, and GPC1 values tended to shift it toward progression within the fitted model. Perturbation-based stability was mixed rather than uniform: CDH3 remained selected in 77.0% of bootstrap perturbations and remained in the bootstrap top 10 in 55.8%; oxytocin (OXT) showed 50.8% and 31.0%, IL5 49.0% and 22.8%, KIT 40.2% and 21.2%, TFF2 34.2% and 15.6%, and GPC1 22.4% and 9.4%, respectively. This pattern supports candidate prioritization but not robust biomarker validation.

In the table, bootstrap columns summarize how often each protein remained selected, and how often it remained in the bootstrap top 10, under repeated perturbation of the baseline cohort. The highest-ranking clinical variable in the final SHAP importance list was rheumatoid factor immunoglobulin M (mean absolute SHAP 0.035), far below the leading proteins. This disparity is consistent with the modality-comparison results and indicates that, within the present workflow, baseline proteomic structure contributed more strongly than routine clinical covariates to the fitted extreme gradient boosting decision function.

Figure 3. Global SHapley Additive exPlanations (SHAP) summary plot for the final extreme gradient boost

Note: RFIgM = Rheumatoid factor immunoglobulin M.

3.6 Illustrative Local Shapley Additive Explanations Exemplars

To make the extreme gradient boosting explanation outputs concrete without over-interpreting them, one reproducibly selected non-progressor exemplar and one progressor exemplar were retained from the fitted secondary model (Figure 4). The non-progressor exemplar was at-risk subject 28 (sample KT00463), for whom the fitted extreme gradient boosting model assigned a progression probability of 0.027. Its local explanation was dominated by contributions toward non-progression from TFF2 (−0.692), KIT (−0.601), N-terminal pro–B-type natriuretic peptide (NTproBNP) (−0.421), IL5 (−0.382), OXT (−0.382), and insulin-like growth factor-binding protein 2 (IGFBP2) ((−0.300), partially offset by smaller progression-oriented contributions from CDH3 (+0.252), ANGPTL2 (+0.214), and programmed cell death protein 1 (PDCD1) (+0.161). The selected progressor exemplar was at-risk subject 32 (sample KT00114), with a predicted progression probability of 0.903. In that case, the dominant progression-oriented contributions were KIT (+0.721), CDH3 (+0.483), OXT (+0.475), IL5 (+0.416), ANGPTL2 (+0.346), NTproBNP (+0.225), and PDCD1 (+0.210), whereas TFF2 again served as the strongest opposing feature. These examples remain illustrative model-specific decompositions from a weak secondary model rather than patient-level biological proof.

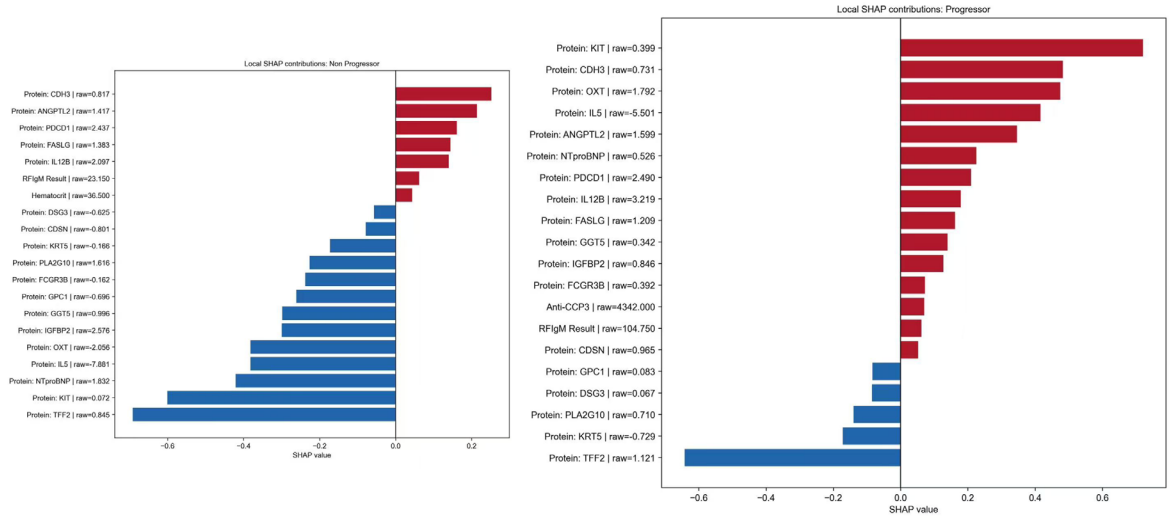

Figure 4. Illustrative local SHapley Additive exPlanations (SHAP) contribution plots for one reproducibly selected non-progressor (top) and one progressor (bottom)

In the figure, each bar shows a single feature's SHAP contribution: bars extending rightward (positive) push the model toward predicting progression; bars extending leftward (negative) push toward non-progression. These are illustrative model-specific decompositions, not patient-level biological proof.

3.7 Illustrative Dependence-Pattern Plots

The first six dependence plots broadly mirrored the global SHAP summaries (Figure 5). Higher TFF2 values were generally associated with more negative SHAP values, whereas higher KIT, CDH3, ANGPTL2, IL5, and GPC1 values tended to shift SHAP values upward. The order of the six exported dependence plots (TFF2, KIT, CDH3, ANGPTL2, IL5, and GPC1) matched the leading protein-importance ranking, supporting a directional attribution pattern within the fitted extreme gradient boosting model while remaining strictly model-dependent and exploratory.

3.8 Permutation Testing and Model-Agnostic Feature Validation

To assess whether the best cross-validated performance exceeded chance, a standalone permutation audit was conducted by re-running the leakage-safe support vector classifier pipeline 500 times with randomly shuffled target labels (Figure 6). The empirical value was 0.164. Accordingly, the observed internal discrimination was not statistically distinguishable from chance at conventional thresholds, reinforcing that these results should be treated as exploratory rather than confirmatory.

In Figure 6, the histogram shows the null distribution of mean ROC-AUC obtained from 500 random label permutations under the leakage-safe pipeline. The red dashed line indicates the observed audit statistic. The empirical *p*-value quantifies the probability of observing this performance by chance.

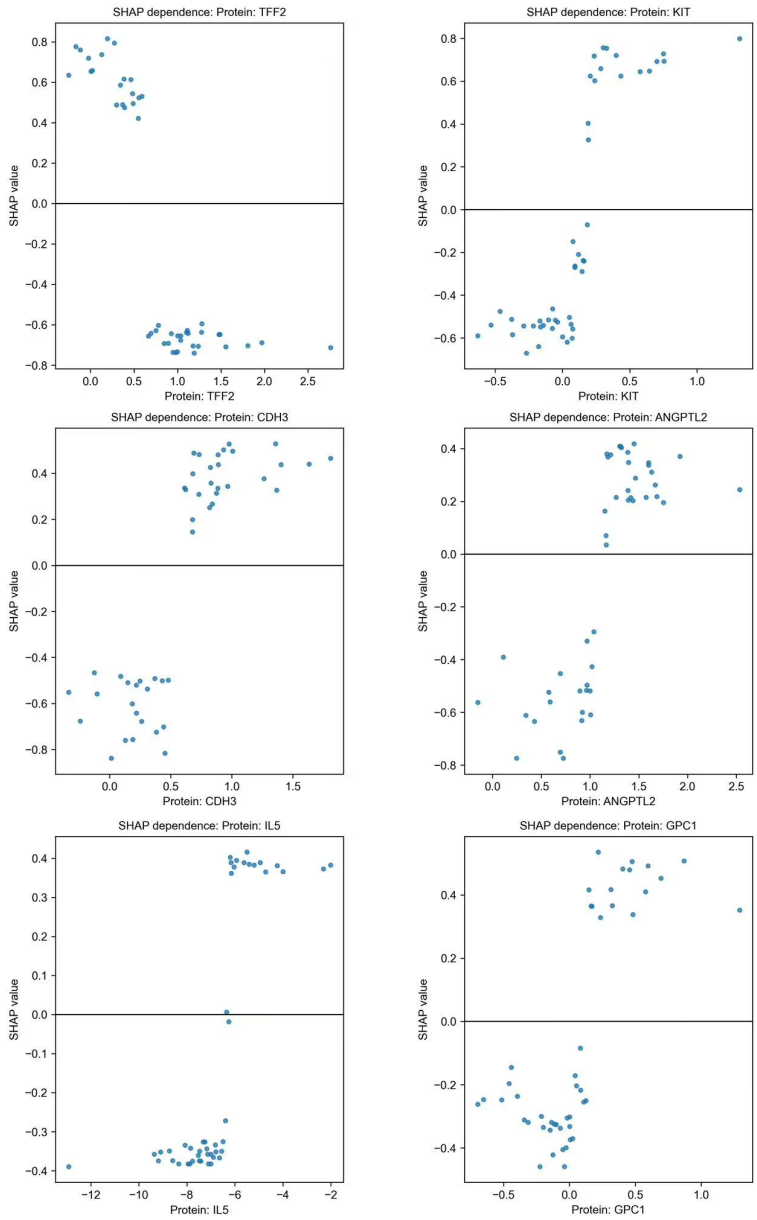

Figure 5. SHapley Additive exPlanations (SHAP) dependence plots for six proteins: TFF2, KIT, CDH3, ANGPTL2, IL5, and GPC1

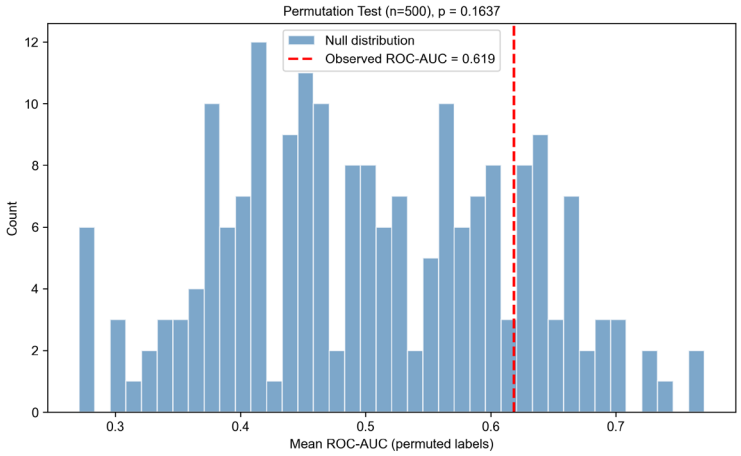

Figure 6. Permutation audit for the support vector classifier model

To complement the extreme gradient boosting TreeSHAP analysis and address the concern that explaining a weaker model may attribute noise, KernelSHAP was applied to the better-performing support vector classifier model fitted to the same top-30 feature matrix. Standard permutation importance was also computed but yielded near-zero values for all features except Hematocrit, consistent with the known limitation that the (radial basis function kernel) support vector classifier relies on holistic, high-dimensional interactions rather than individual feature contributions, making single-feature permutation an insensitive probe for this model family. In contrast, the KernelSHAP ranking for the support vector classifier (Figure 7) showed partial concordance with the extreme gradient boosting TreeSHAP results: among the top 10 KernelSHAP features for the support vector classifier, 7 overlapped with the extreme gradient boosting top 10 such as CDH3, OXT, Fc gamma receptor IIIb (FCGR3B), NTproBNP, IL5, TFF2, and KIT. This cross-model convergence is more consistent with candidate-signal convergence within the present dataset than with robust biomarker validation, and the KernelSHAP outputs for the support vector classifier should also be read as in-sample, model-dependent prioritization maps.

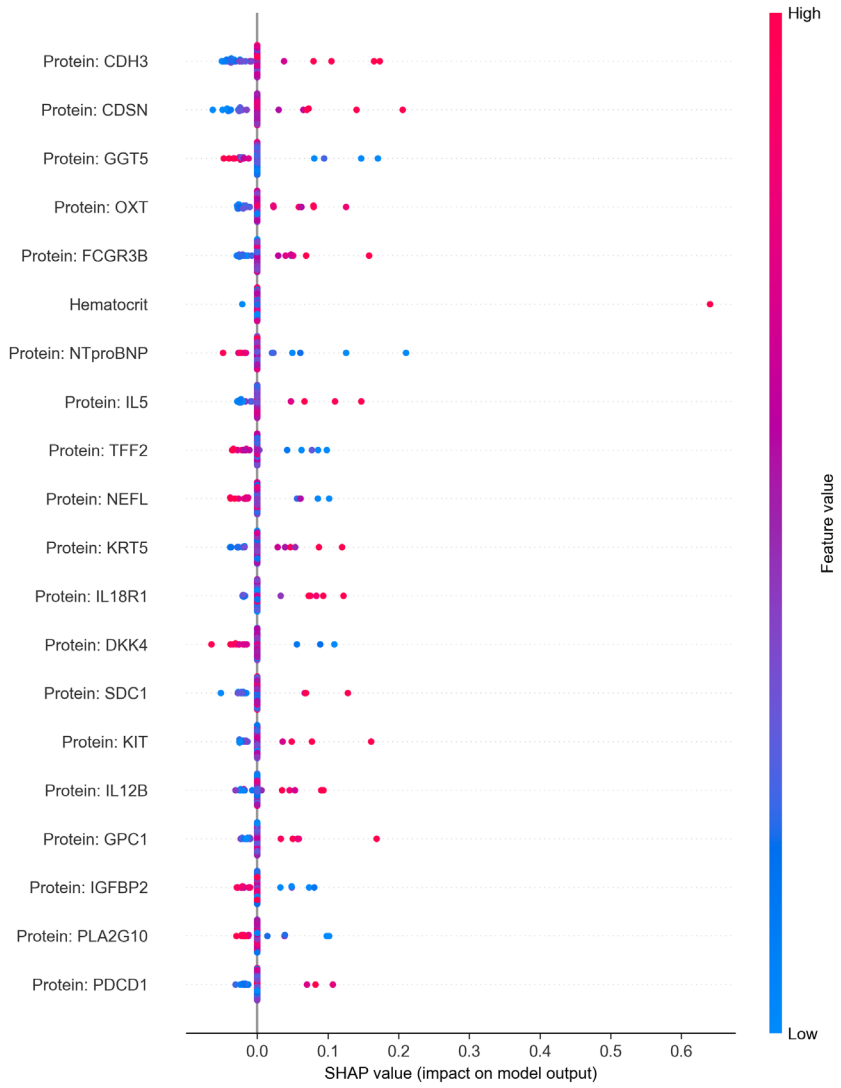

Figure 7. Model-agnostic Kernel SHapley Additive exPlanations (KernelSHAP) summary plot for the support vector classifier model fitted to the same top-30 feature matrix

Note: SHAP = SHapley Additive exPlanations.

In the figure, among the top 10 features, 7 overlap with the extreme gradient boosting TreeSHAP top 10 (CDH3, OXT, FCGR3B, NTproBNP, IL5, TFF2, and KIT), providing cross-model convergence of candidate signals within this exploratory dataset.

4. Discussion

This study should be interpreted as an exploratory multimodal biomarker-prioritization and explainability analysis rather than as a high-confidence clinical prediction study. Once feature ranking and correlation filtering were moved fully inside the training folds, baseline classification performance remained modest, with the primary support vector classifier reaching a subject-level ROC-AUC of 0.675 (95% confidence interval 0.516–0.825) and the prespecified extreme gradient boosting explanation model reaching only a mean fold ROC-AUC of 0.539. The standalone permutation audit was not statistically significant (p = 0.164), and internal calibration of the main support vector classifier was weak (Brier score 0.231; calibration slope 0.455). In addition, observed follow-up among non-progressors was limited (median 0 days; range 0–347), so the binary endpoint should not be interpreted as a mature time-to-event label. Taken together, these findings indicate that the present dataset is better suited to internal signal prioritization than to strong predictive claims. In that sense, the current results are more consistent with the difficulty of forecasting progression across heterogeneous at-risk rheumatoid arthritis states described in the contemporary prevention literature (Deane, 2024; Deane & Holers, 2021; O'Neil et al., 2024; Toyoda & Mankia, 2024).

To place this performance in context, an internal subject-level ROC-AUC of 0.675 with a 95% confidence interval of 0.516–0.825 lies in the lower range typically reported for preclinical or early-rheumatoid arthritis prediction studies. Recent multimodal rheumatoid arthritis prediction and treatment-response models built on substantially larger and more mature cohorts have achieved ROC-AUC values in the 0.70–0.85 range, such as machine-learning models predicting biologic ineffectiveness, methotrexate inadequate response, or remission in registry data (Alsaber et al., 2024; Duquesne et al., 2023; Sonomoto et al., 2024; Ukalovic et al., 2024). Pre-clinical proteomic studies of progression to rheumatoid arthritis have likewise reported areas under the curve around 0.70–0.80 once cohorts are sufficiently large and follow-up is mature (He et al., 2025b; Zaim et al., 2025). Against this backdrop, a value near 0.67 with a 95% confidence interval that spans 0.52–0.83 and a non-significant permutation audit (p = 0.164) is consistent with weak-to-moderate internal discrimination and is not, in its current state, sufficient to support individualized risk prediction or clinical triage. Therefore, these numbers can be interpreted as evidence that the leakage-safe workflow can extract a coherent but modest baseline proteomic signal from this cohort, while explicitly reinforcing that such performance is below the level usually considered adequate for clinical decision support and that any application beyond hypothesis generation would require larger, longitudinally complete cohorts and external validation.

A second key finding is that the baseline proteomic representation consistently outperformed the routine baseline clinical covariates, while adding clinical covariates did not produce a stable gain over proteomics alone. In the support vector classifier, the combined representation slightly improved mean ROC-AUC but not PR-AUC; in regularized logistic regression, the combined representation was marginally worse than proteomics alone on both metrics. Therefore, this study did not observe stable evidence of incremental predictive value from the 17 routine clinical and laboratory variables beyond what was already captured by the top-ranked circulating proteins. This pattern suggests that the dominant baseline signal in the present cohort was primarily molecular rather than clinical, although the absence of a clear combined-model gain may still reflect sample-size constraints, redundant information content between clinical and proteomic features, or both. Even so, the result is informative: it aligns with recent rheumatoid arthritis biomarker studies, indicating that circulating proteomic measurements may capture aspects of disease biology not well summarized by conventional serologic and laboratory markers alone (Cuesta-López et al., 2024; He et al., 2025b; O'Neil et al., 2022; Sahin et al., 2025; Zaim et al., 2025).

The explanation analyses narrowed this proteomic signal to a compact set of candidate baseline proteins within the fitted models. TFF2, KIT, CDH3, ANGPTL2, IL5, and GPC1 dominated the global TreeSHAP ranking, while KernelSHAP for the support vector classifier and bootstrap perturbation summaries suggested partial but uneven stability rather than uniformly robust recurrence. CDH3 showed the strongest perturbation robustness, whereas TFF2, ANGPTL2, and GPC1 were noticeably less stable. In addition, a coefficient-based sanity check from the leakage-safe logistic-regression workflow recovered several overlapping signals, including OXT, TFF2, CDH3, FCGR3B, phospholipase A2 group X (PLA2G10), and IL5, suggesting that part of the prioritized feature structure was not unique to a single nonlinear estimator. These convergences do not establish mechanism, causality, or clinical utility, but they do support carrying the leading proteins forward as candidate baseline proteins for downstream validation rather than dismissing them as obvious one-model artifacts.

A legitimate concern with using extreme gradient boosting for explanation, given its near-chance ROC-AUC of 0.539, is that the SHAP attributions may largely reflect noise rather than genuine biological signal. To mitigate this risk, a dual-model explanation strategy was adopted. In addition to extreme gradient boosting TreeSHAP, model-agnostic permutation importance and KernelSHAP were applied to the better-performing support vector classifier model (mean fold ROC-AUC 0.641). The resulting cross-model overlap is better interpreted as candidate-signal convergence within the present dataset than as reassurance that the extreme gradient boosting attributions are robust. Extreme gradient boosting TreeSHAP was retained because it provides exact, compositionally consistent, and computationally efficient Shapley values with both direction and magnitude at the

individual-subject level capabilities that permutation importance and KernelSHAP (which required approximate sampling with $n_{samples}$ = 200) do not match in interpretive granularity (Ali et al., 2023; Alkhanbouli et al., 2025; Allgaier et al., 2023; Chen & Guestrin, 2016; Loh et al., 2022; Lundberg et al., 2020). Even so, the extreme gradient boosting-based explanations should be viewed as illustrative in-sample prioritization outputs from a weak secondary model rather than as stable biomarker evidence.

A self-appraisal based on the prediction model risk of bias assessment tool would place the greatest risks of bias in the participants/follow-up and analysis domains. Participant applicability is constrained by the limited follow-up among many non-progressors; predictor handling is challenged by the extreme $p \gg n$ setting; the analysis remains vulnerable because validation is entirely internal; hyperparameters were not tuned in a nested fashion; and calibration remained weak. These are not minor caveats; they define the scope of the study.

Several limitations are central to the interpretation of this work. First, the baseline cohort was small (n = 47), class-imbalanced (16 progressors versus 31 non-progressors), and evaluated only with internal resampling; no independent external validation set was available. The high ratio of candidate proteins (p = 1,449) to subjects (n = 47) creates a challenging curse-of-dimensionality setting in which spurious correlations can arise despite leakage-safe fold-internal feature selection. Second, the outcome formulation is limited by incomplete follow-up in the non-progressor group, and the present binary label should not be treated as equivalent to a fully observed time-to-event endpoint. Third, although fold-local feature ranking and filtering reduced information leakage, the univariate Cohen's d followed by correlation filtering approach may still capitalize on sampling variability and may discard proteins whose value is primarily interaction-based; embedded sparse methods such as elastic net or the least absolute shrinkage and selection operator would be worth exploring in larger cohorts. Fourth, median imputation was stable for this small dataset but does not preserve multivariable covariance structure as well as more complex methods such as the multiple imputation by chained equations. Fifth, the assay-wise batch-centering sensitivity analysis was reassuring but does not replace a more comprehensive treatment of batch, plate, and panel effects. Sixth, the extreme gradient boosting explanation model achieved a ROC-AUC of only 0.539, so its TreeSHAP attributions require particular caution even after cross-model comparison. Finally, SHAP explains contribution within a fitted model; it does not establish biological causality, clinical actionability, or mechanistic priority on its own. Taken together, these constraints mean that the present manuscript should be read as an exploratory report aimed at candidate prioritization and methodological transparency rather than as a definitive prognostic study.

5. Conclusions

In conclusion, this 47-subject study does not support a clinically usable progression predictor. Under leakage-safe internal validation, subject-level support vector classifier discrimination remained modest (ROC-AUC 0.675, 95% confidence interval 0.516–0.825); the standalone permutation audit was not statistically significant (p = 0.164); internal calibration was weak; and the non-progressor endpoint remained immature because follow-up was often absent or minimal. What the data do support is exploratory baseline biomarker prioritization. Across secondary extreme gradient boosting TreeSHAP, KernelSHAP for the support vector classifier, and bootstrap perturbation summaries, a recurrent set of candidate baseline proteins–including TFF2, KIT, CDH3, ANGPTL2, IL5, and GPC1–emerged within the fitted models. The main contribution is, therefore, a transparent leakage-aware workflow for narrowing a high-dimensional baseline proteomic feature space under severe $p \gg n$ imbalance and endpoint uncertainty. Independent cohorts with longer non-progressor follow-up, explicit time-to-event modeling, stronger batch adjustment, and external validation are required before any prognostic claim can be justified.

Beyond this specific cohort, the broader contribution of this study is methodological. The proposed leakage-safe pipeline, fold-internal Cohen's d ranking, correlation filtering, top-k selection, dual-model SHAP explainability, permutation auditing, and explicit calibration reporting are fully general and can be applied to other small-sample, high-dimensional biomarker-prioritization problems in which the candidate feature space (e.g., proteomic, transcriptomic, or metabolomic panels) is far larger than the available number of subjects. By clearly separating signal prioritization from clinical prediction, and by transparently reporting both what the data support and what they do not, this workflow offers a reusable template for exploratory biomarker studies in early or rare-disease cohorts where naive pipelines tend to overstate predictive performance. Therefore, the present analysis is considered a worked example of how to extract the maximum amount of honest information from a small $p \gg n$ cohort while remaining methodologically defensible.

Author Contributions

Conceptualization, T.B.D. and M.M.; methodology, T.B.D. and M.M.; software, T.B.D. and M.M.; validation, T.B.D. and M.M.; formal analysis, T.B.D. and M.M.; investigation, T.B.D. and M.M.; resources, T.B.D. and M.M.; data curation, T.B.D. and M.M.; writing—original draft preparation, T.B.D. and M.M.; writing—review and editing, T.B.D. and M.M.; visualization, T.B.D. and M.M.; supervision, T.B.D. and M.M.; project administration, T.B.D. and M.M.; funding acquisition, T.B.D. and M.M. All authors have read and agreed to the published version

of the manuscript.

Data Availability

The clinical/laboratory and Olink proteomic data used in this study were obtained from the Allen Institute for Immunology Human Immune System Explorer "Systemic Inflammation in At-Risk Individuals Advancing to Clinical Rheumatoid Arthritis" project (Allen Institute for Immunology, 2025; He et al., 2025a). The specific downloadable sources used for this manuscript were the Clinical Labs & Metadata and Plasma Proteomics resources from that RA Progression project, available at https://apps.allenimmunology.org/aifi/insights/ra-progression/.

Acknowledgements

The authors acknowledge Sakarya University of Applied Sciences (https://subu.edu.tr/) for the technical support provided to publish the present manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Declaration on the Use of Generative AI and AI-assisted Technologies

The authors declare that no generative artificial intelligence (AI) or AI-assisted technologies were used in the preparation of this manuscript.

References

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Comput. Biol. Med.*, *166*, 107555. https://doi.org/10.1016/j.compbiomed.2023.107555.

Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Med. Inform. Decis. Mak.*, *25*(1), 110. https://doi.org/10.1186/s12911-025-02944-6.

Allen Institute for Immunology. (2025). *Systemic inflammation in at-risk individuals advancing to clinical rheumatoid arthritis*. Allen Institute for Immunology. https://apps.allenimmunology.org/aifi/insights/ra-progression/

Allgaier, J., Mulansky, L., Draelos, R. L., & Pryss, R. (2023). How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artif. Intell. Med.*, *143*, 102616. https://doi.org/10.1016/j.artmed.2023.102616.

Alsaber, A. R., Al-Herz, A., Alawadhi, B., Doush, I. A., Setiya, P., Al-Sultan, A. T., Saleh, K., Al-Awadhi, A., Hasan, E., & Al-Kandari, W. et al. (2024). Machine learning-based remission prediction in rheumatoid arthritis patients treated with biologic disease-modifying anti-rheumatic drugs: Findings from the Kuwait rheumatic disease registry. *Front. Big Data*, *7*, 1406365. https://doi.org/10.3389/fdata.2024.1406365.

Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Bucht Thorsen, S., Ekman, D., Erikson, A., Dickens, E. R., Ohlsson, S., & Edfeldt, G. et al. (2014). Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One*, *9*(4), e95192. https://doi.org/10.1371/journal.pone.0095192.

Baloun, J., Cerezo, L. A., Kropáčková, T., Prokopcová, A., Marešová, K. B., Mann, H., Vencovský, J., Pavelka, K., & Šenolt, L. (2025). Machine learning-assisted screening of clinical features for predicting difficult-to-treat rheumatoid arthritis. *Sci. Rep.*, *15*(1), 34747. https://doi.org/10.1038/s41598-025-18298-y.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Cuesta-López, L., Escudero-Contreras, A., Hanaee, Y., Perez-Sanchez, C., Ruiz-Ponce, M., Martínez-Moreno, J. M., Pérez-Pampin, E., González, A., Plasencia-Rodriguez, C., & Martínez-Feito, A. et al. (2024). Exploring candidate biomarkers for rheumatoid arthritis through cardiovascular and cardiometabolic serum proteome profiling. *Front. Immunol.*, *15*, 1333995. https://doi.org/10.3389/fimmu.2024.1333995.

Deane, K. D. (2024). Rheumatoid arthritis: Prediction of future clinically-apparent disease, and prevention. *Curr. Opin. Rheumatol.*, *36*(3), 225–234. https://doi.org/10.1097/BOR.0000000000001013.

Deane, K. D. & Holers, V. M. (2021). Rheumatoid arthritis pathogenesis, prediction, and prevention: An emerging paradigm shift. *Arthritis Rheumatol.*, *73*(2), 181–193. https://doi.org/10.1002/art.41417.

Duquesne, J., Bouget, V., Cournede, P. H., Fautrel, B., Guillemin, F., de Jong, P. H., Heutz, J. W., Verstappen, M., van der Helm-van Mil, A. H. M., & Mariette, X. et al. (2023). Machine learning identifies a profile of inadequate responder to methotrexate in rheumatoid arthritis. *Rheumatology*, *62*(7), 2402–2409. https://doi.org/10.1093/rheumatology/keac645.

Escal, J., Neel, T., Hodin, S., Boussoualim, K., Amouzougan, A., Coassy, A., Locrelle, H., Thomas, T., Delavenne, X., & Marotte, H. (2024). Proteomics analyses of human plasma reveal triosephosphate isomerase as a potential blood marker of methotrexate resistance in rheumatoid arthritis. *Rheumatology*, *63*(5), 1368–1376. https://doi.org/10.1093/rheumatology/kead390.

Ferreira, M.B., Kobayashi, M., Costa, R.Q., Fonseca, T., Brandão, M., Oliveira, J.C., Marinho, A., Cyrne Carvalho, H., Rodrigues, P., & Zannad, F. et al. (2023). Unsupervised clustering to differentiate rheumatoid arthritis patients based on proteomic signatures. *Scand. J. Rheumatol. 52*(6), 619–626. https://doi.org/10.1080/03009742.2023.2196781.

Frazzei, G., Musters, A., de Vries, N., Tas, S. W., & van Vollenhoven, R. F. (2023). Prevention of rheumatoid arthritis: A systematic literature review of preventive strategies in at-risk individuals. *Autoimmun. Rev.*, *22*(1), 103217. https://doi.org/10.1016/j.autrev.2022.103217.

He, S., Zhu, C., Liu, Y., Xu, Z., Sun, R., Yang, B., Guo, X., Herrmann, M., Muñoz, L. E., & Gjertsson, I. et al. (2025a). A longitudinal cohort study uncovers plasma protein biomarkers predating clinical onset and treatment response of rheumatoid arthritis. *Nat. Commun.*, *16*(1), 6692. https://doi.org/10.1038/s41467-025-62032-1.

He, Z., Glass, M. C., Venkatesan, P., Feser, M. L., Lazaro, L., Okada, L. Y., Tran, N. T. T., He, Y. D., Zaim, S. R., & Bennett, C. et al. (2025b). Progression to rheumatoid arthritis in at-risk individuals is defined by systemic inflammation and by T and B cell dysregulation. *Sci. Transl. Med.*, *17*(817), eadt7214. https://doi.org/10.1126/scitranslmed.adt7214.

Jang, J., Kim, W. J., Park, S. W., & Moon, K. W. W. (2025). Development of explainable machine learning models to predict side effects in patients with rheumatoid arthritis taking methotrexate treatment: A nationwide multicentre cohort study. *BMJ Open*, *15*(11), e108527. https://doi.org/10.1136/bmjopen-2025-108527.

Jin, L., Wang, F., Wang, X., Harvey, B. P., Bi, Y., Hu, C., Cui, B., Darcy, A. T., Maull, J. W., & Phillips, B. R. et al. (2023). Identification of plasma biomarkers from rheumatoid arthritis patients using an optimized sequential window acquisition of all theoretical mass spectra (SWATH) proteomics workflow. *Proteomes*, *11*(4), 32. https://doi.org/10.3390/proteomes11040032.

Lewis, M. J. (2024). Predicting best treatment in rheumatoid arthritis. In *Seminars in Arthritis and Rheumatism* (Vol. 64, p. 152329). WB Saunders. https://doi.org/10.1016/j.semarthrit.2023.152329.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.*, *226*, 107161. https://doi.org/10.1016/j.cmpb.2022.107161.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

O'Neil, L. J., Alpízar-Rodríguez, D., & Deane, K. D. (2024). Rheumatoid arthritis: The continuum of disease and strategies for prediction, early intervention, and prevention. *J. Rheumatol.*, *51*(4), 337–349. https://doi.org/10.3899/jrheum.2023-0334.

O'Neil, L. J., Meng, X., Mcfadyen, C., Fritzler, M. J., & El-Gabalawy, H. S. (2022). Serum proteomic networks associate with pre-clinical rheumatoid arthritis autoantibodies and longitudinal outcomes. *Front. Immunol.*, *13*, 958145. https://doi.org/10.3389/fimmu.2022.958145.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, *12*, 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.*, *31*.

Rivellese, F., Surace, A. E. A., Goldmann, K., Sciacca, E., Çubuk, C., Giorli, G., John, C. R., Nerviani, A., Fossati-Jimack, L., & Thorborn, G. et al. (2022). Rituximab versus tocilizumab in rheumatoid arthritis: Synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat. Med.*, *28*(6), 1256–1268. https://doi.org/10.1038/s41591-022-01789-0.

Sahin, D., Di Matteo, A., & Emery, P. (2025). Biomarkers in the diagnosis, prognosis and management of rheumatoid arthritis: A comprehensive review. *Ann. Clin. Biochem.*, *62*(1), 3–21. https://doi.org/10.1177/00045632241285843.

Sonomoto, K., Fujino, Y., Tanaka, H., Nagayasu, A., Nakayamada, S., & Tanaka, Y. (2024). A machine learning approach for prediction of CDAI remission with TNF inhibitors: A concept of precision medicine from the FIRST registry. *Rheumatol. Ther.*, *11*(3), 709–736. https://doi.org/10.1007/s40744-024-00668-z.

Toyoda, T. & Mankia, K. (2024). Prevention of rheumatoid arthritis in at-risk individuals: Current status and future prospects. *Drugs*, *84*(8), 895–907. https://doi.org/10.1007/s40265-024-02061-0.

Ukalovic, D., Leeb, B. F., Rintelen, B., Eichbauer-Sturm, G., Spellitz, P., Puchner, R., Herold, M., Stetter, M., Ferincz, V., & Resch-Passini, J. et al. (2024). Prediction of ineffectiveness of biological drugs using machine learning and explainable AI methods: Data from the Austrian Biological Registry BioReg. *Arthritis Res. Ther.*, *26*(1), 44. https://doi.org/10.1186/s13075-024-03277-x.

Zaim, S. R., Savage, A. K., Gillespie, M. A., Castillo, J. D., Bennett, C., Torgerson, T. R., Becker, L. A., Mahler, M., Moss, L., & Feser, M. L. et al. (2025). Serum proteomic signatures before the diagnosis of rheumatoid arthritis: Evolving biologic pathways and specific periods of disease development. *Arthritis Rheumatol.*, *77*(9), 1166–1178. https://doi.org/10.1002/art.43175.

Zhao, J. H., Stacey, D., Eriksson, N., Macdonald-Dunlop, E., Hedman, Å. K., Kalnapenkis, A., Enroth, S., Cozzetto, D., Digby-Bell, J., & Marten, J. et al. (2023). Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat. Immunol.*, *24*(9), 1540–1551. https://doi.org/10.1038/s41590-023-01588-w.

Appendix: Exploratory Longitudinal Proteomic Analysis

The longitudinal subset contains only 19 subjects with 3 non-progressors. Consequently, all effect-size estimates and rankings reported below are statistically unreliable and should be treated as anecdotal observations, not as formal statistical findings.

The exploratory longitudinal proteomic analysis was restricted to 19 subjects with repeated Olink measurements, making it smaller and more progression-enriched than the baseline cohort. Comparison of the top 20 baseline proteins with the top 20 delta-ranked proteins showed no direct overlap (intersection count = 0), but broad thematic continuity remained at a descriptive level (Table A1). The baseline top-20 set contained more immune/inflammatory proteins than the delta-ranked set, whereas the delta-ranked set was more heavily enriched in the other/mixed category. These observations are reported as descriptive patterns only; the extreme class imbalance (N = 3 controls) renders the underlying Cohen's d estimates unstable and precludes any statistical inference. The absence of protein-level overlap may reflect genuine biological differences between baseline state and temporal change, but it could equally reflect the instability of rankings derived from such a small and imbalanced subset. This study retains this appendix solely to document a possible future direction for larger longitudinal cohorts. In the table below, the protein-level overlap is absent, but broad biological themes remain represented in both rankings.

Table A1. Compact comparison of baseline and longitudinal delta proteomic prioritization

Comparison Item	Baseline	Longitudinal Delta
Top-ranked proteins compared	20	20
Direct overlap with the other ranking	0 proteins (none)	
Immune/inflammatory	6	3
Metabolic/stress/systemic	4	3
Other/mixed	6	11
Structural/adhesion/trafficking	4	3