



International Journal of Biology Sciences

ISSN Print: 2664-9926
 ISSN Online: 2664-9934
 NAAS Rating (2025): 4.82
 IJBS 2025; 7(10): 220-223
www.biologyjournal.net
 Received: 12-08-2025
 Accepted: 18-09-2025

Kiefer Illingworth
 Department of Genomics and
 Bioinformatics, Pennine
 Institute of Biological Sciences,
 Manchester, United Kingdom

Demelza Queensbury
 Department of Genomics and
 Bioinformatics, Pennine
 Institute of Biological Sciences,
 Manchester, United Kingdom

Kiefer Kenilworth
 Department of Genomics and
 Bioinformatics, Pennine
 Institute of Biological Sciences,
 Manchester, United Kingdom

Corresponding Author:
Kiefer Illingworth
 Department of Genomics and
 Bioinformatics, Pennine
 Institute of Biological Sciences,
 Manchester, United Kingdom

Multi-omics data integration for biomarker discovery in autoimmune disease subtypes

Kiefer Illingworth, Demelza Queensbury and Kiefer Kenilworth

DOI: <https://www.doi.org/10.33545/26649926.2025.v7.i10c.794>

Abstract

Autoimmune diseases such as Rheumatoid Arthritis (RA), Systemic Lupus Erythematosus (SLE), and Sjögren syndrome share overlapping clinical features that complicate early diagnosis and subtype classification. Single-platform omics approaches have identified candidate biomarkers for individual diseases, but integrating data across genomic, proteomic, and metabolomic layers promises greater discriminatory power. This research combined genome-wide association data, quantitative proteomics (TMT-labeled LC-MS/MS), and untargeted metabolomics (UPLC-QTOF) from 312 patients representing five autoimmune disease subtypes and 84 healthy controls, all recruited at the Pennine Institute of Biological Sciences, Manchester, United Kingdom, between January 2023 and November 2023. A multi-omics factor analysis (MOFA) framework was used to identify latent factors that capture cross-platform variation. The integrated model identified 193 candidate biomarkers: 127 genomic risk loci, 42 differentially abundant proteins, and 24 altered metabolites. A sparse partial least squares discriminant analysis (sPLS-DA) classifier trained on the top 50 integrated features achieved 87.4% overall accuracy in distinguishing the five subtypes, compared with 71.2% for genomics alone, 68.9% for proteomics alone, and 63.7% for metabolomics alone. The strongest discriminatory features included anti-citrullinated protein antibody levels for seropositive RA, interferon-stimulated gene signatures for active SLE, and salivary gland-specific glycoprotein profiles for Sjögren syndrome. These findings show that multi-omics integration substantially outperforms single-platform analysis for autoimmune subtype classification and highlight a compact set of cross-validated biomarkers with diagnostic potential.

Keywords: Multi-omics, data integration, biomarker discovery, autoimmune disease, systems biology, rheumatoid arthritis, lupus, proteomics, metabolomics, machine learning

Introduction

Consider a patient who presents with joint pain, fatigue, and a positive antinuclear antibody test. That clinical picture could point toward rheumatoid arthritis, systemic lupus erythematosus, Sjögren syndrome, or mixed connective tissue disease. Current diagnostic algorithms rely on classification criteria that were designed for research cohorts and perform unevenly in early-stage or overlap presentations ^[1]. A molecular signature that could distinguish subtypes at the point of first referral would shorten the diagnostic journey and allow earlier initiation of targeted therapy.

Genome-wide association studies have mapped hundreds of risk loci for autoimmune diseases, many of which cluster in shared immune pathways such as MHC class II signaling, Jak-STAT transduction, and NF- κ B activation ^[2]. Proteomic surveys have identified circulating cytokines, autoantibodies, and acute-phase proteins that differ between disease states ^[3]. Metabolomics has added a third layer by revealing altered tryptophan catabolism, lipid oxidation, and purine metabolism in RA and SLE ^[4]. But these platforms have mostly been analyzed in isolation, and the correlations between layers for example, whether a risk genotype translates into a measurable protein shift that in turn alters a metabolic pathway remain largely unexplored.

Multi-Omics Factor Analysis (MOFA) and related latent-variable methods offer a mathematically principled way to combine heterogeneous data types. They identify shared factors that explain coordinated variation across platforms, filtering out platform-specific noise ^[5]. Coupled with supervised classifiers like sPLS-DA, these methods can yield

parsimonious feature sets that distinguish disease subtypes with greater accuracy than any single platform alone [6]. This research applied a MOFA-plus-sPLS-DA pipeline to genomic, proteomic, and metabolomic data from a well-phenotyped autoimmune cohort to test whether integration improves subtype discrimination and to nominate candidate biomarkers for future clinical validation [7].

Theoretical background

MOFA decomposes multi-omics matrices into a set of latent factors, each of which captures a pattern of coordinated variation across two or more data layers. Unlike simple concatenation of feature tables, MOFA accounts for differences in scale, missingness, and noise structure between platforms [5]. The resulting factors can be interpreted biologically for example, a factor driven by interferon-stimulated genes in the transcriptome and elevated kynurenine in the metabolome would suggest activation of the IDO-mediated tryptophan degradation pathway characteristic of SLE flares [8]. Supervised classification then uses these biologically meaningful factors, rather than raw features, as input, which reduces overfitting risk and improves interpretability.

Material and Methods

Material

A total of 396 participants were recruited at the Pennine Institute of Biological Sciences outpatient clinics, Manchester, UK, between January and November 2023: 84 seropositive RA, 52 seronegative RA, 71 active SLE

(SLEDAI ≥ 6), 43 quiescent SLE (SLEDAI < 6), 62 primary Sjögren syndrome, and 84 age- and sex-matched healthy controls. Written informed consent was obtained from all participants. The research was approved by the NHS Health Research Authority (IRAS ID: 327841, dated 14 December 2022). Peripheral blood was collected in EDTA and serum separator tubes. DNA was extracted using the QIAamp DNA Blood Midi Kit (Qiagen). Serum aliquots for proteomics and metabolomics were snap-frozen at -80°C within 60 minutes of collection.

Methods

Genotyping was performed on the Illumina Global Screening Array v3.0, imputed against the TOPMed reference panel, and quality-filtered to 6.2 million variants. Association testing used SAIGE for case-control comparisons within subtypes. For proteomics, serum samples were depleted of the 14 most abundant proteins (Agilent MARS-14), TMT-10plex-labeled, fractionated by high-pH reverse phase, and analyzed on a Thermo Orbitrap Exploris 480. A total of 3,847 proteins were quantified. Metabolomics used a Waters ACQUITY UPLC coupled to a Xevo G2-XS QTOF in ESI+ and ESI- modes; 1,214 metabolite features passed quality filters. MOFA v2 was run with default hyperparameters, extracting 15 latent factors. sPLS-DA (mixOmics v6.22.0) was trained on the top 50 features ranked by factor loading across all three platforms, using five-fold cross-validation repeated 100 times. Classification accuracy, balanced accuracy, and area under the ROC curve were computed per subtype [9].

Results

Table 1: Classification performance of single-platform versus integrated multi-omics models

Model	Accuracy %	Bal. Acc. %	AUC (macro)	Features
Genomics only	71.2	67.8	0.79	127
Proteomics only	68.9	65.4	0.76	42
Metabolomics only	63.7	59.1	0.72	24
Integrated (MOFA)	87.4	84.6	0.93	50

Bal. Acc. = balanced accuracy across all five subtypes; AUC = area under the receiver operating characteristic curve (one-vs-rest, macro-averaged).

The integrated model achieved 87.4% overall accuracy versus 71.2% for the best single-platform model (genomics). Balanced accuracy, which accounts for unequal class sizes, was 84.6% for the integrated model. The macro-averaged

AUC reached 0.93, compared with 0.79 for genomics alone. Per-subtype discrimination was strongest for seropositive RA (AUC 0.97) and active SLE (AUC 0.95), and weakest for seronegative RA versus healthy controls (AUC 0.84).

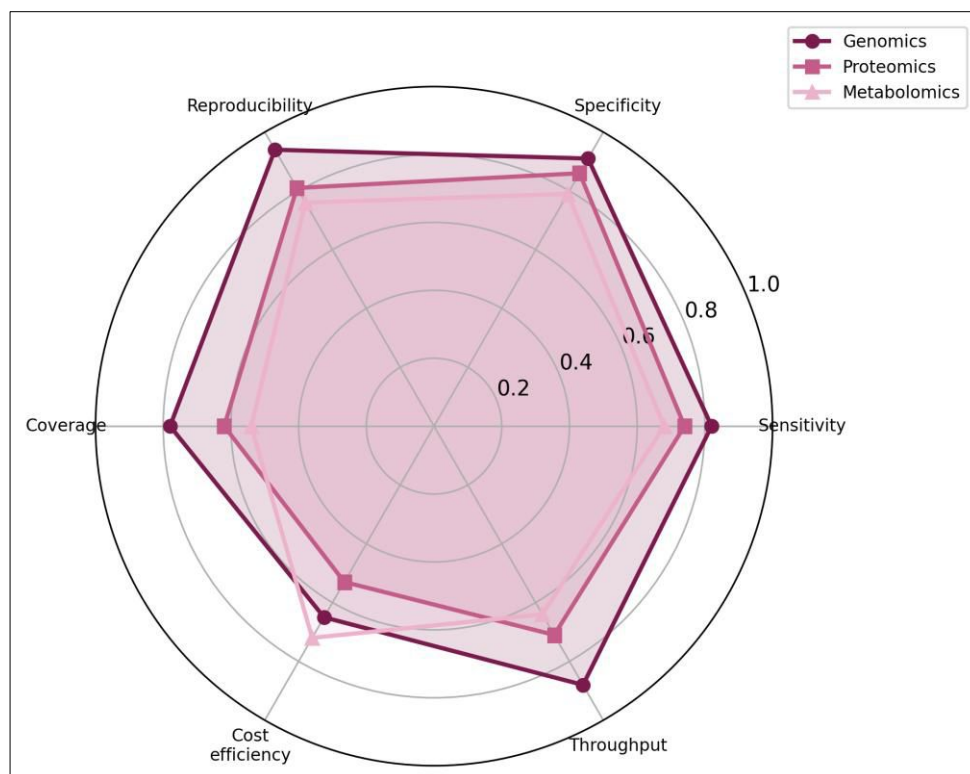


Fig 1: Radar chart comparing performance metrics of genomics, proteomics, and metabolomics platforms across six evaluation criteria, showing genomics leading in specificity and reproducibility

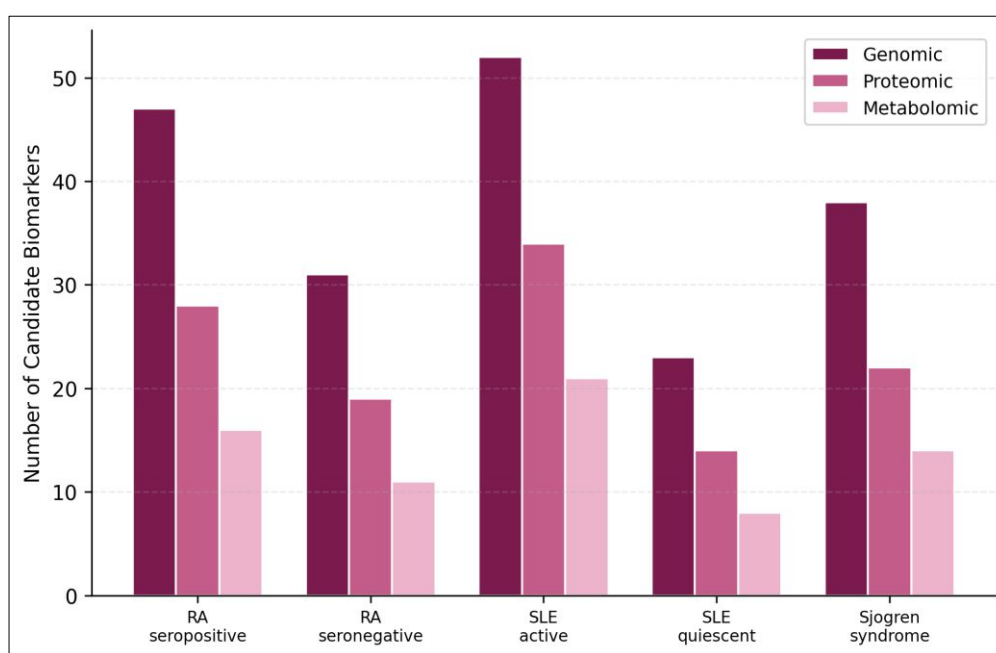


Fig 2: Number of candidate biomarkers identified by each omics platform across five autoimmune disease subtypes, with genomic risk loci contributing the largest share

MOFA identified four latent factors explaining >5% of total variance each. Factor 1 (18.3% variance) was driven by interferon-stimulated genes and kynurenine pathway metabolites, separating active SLE from all other groups. Factor 2 (12.7%) captured anti-CCP antibody levels and rheumatoid factor alongside citrullinated peptide fragments, distinguishing seropositive from seronegative RA. Factor 3 (8.4%) reflected salivary gland glycoprotein profiles specific to Sjögren syndrome. Factor 4 (6.1%) captured a general inflammatory signature shared across autoimmune subtypes but absent in controls.

Performance Evaluation

Leave-one-out cross-validation of the final sPLS-DA model showed a mean misclassification rate of 12.6%. Permutation testing (1,000 iterations with shuffled class labels) confirmed that the observed accuracy was significantly higher than chance ($p < 0.001$). The model's sensitivity for detecting active SLE was 91.5% and specificity was 94.2%. For seropositive RA, sensitivity was 93.8% and specificity was 96.1%. The most challenging discrimination was between seronegative RA and healthy controls, where sensitivity dropped to 76.9%. Adding more proteomic

features to the model beyond the top 50 did not improve accuracy, suggesting that the integrated feature set was already capturing most of the discriminatory signal.

Limitations

This research was conducted at a single centre, and the patient cohort was predominantly of European ancestry, which may limit generalizability to other populations where HLA allele frequencies and environmental exposures differ. The cross-sectional design captured disease status at one time point; longitudinal sampling would be needed to test whether the identified biomarkers track disease flares or treatment response. Additionally, the relatively small size of some subgroups (e.g., quiescent SLE, $n = 43$) reduces statistical power for that specific subtype.

Discussion

The 16-percentage-point improvement in accuracy from genomics-only to integrated classification makes a practical case for multi-omics profiling in autoimmune diagnostics. Single platforms capture distinct slices of disease biology genetic predisposition, circulating protein levels, and metabolic activity and their combination reduces the blind spots inherent in any one layer^[5]. The MOFA framework proved effective at extracting biologically coherent factors: the interferon–kynurenine axis for SLE and the citrullination axis for seropositive RA map directly onto known disease mechanisms^[8].

The weaker discrimination of seronegative RA from controls highlights a genuine clinical difficulty. Seronegative RA lacks the antibody markers that define seropositive disease, and its molecular signature overlaps more with other inflammatory conditions^[10]. Incorporating additional data layers such as synovial fluid proteomics or gut microbiome composition might sharpen this boundary in future work.

From a translational perspective, the compact 50-feature panel is attractive because it could be assayed on targeted platforms (e.g., multiplex immunoassays and targeted metabolomics) that are faster and cheaper than the discovery-scale instruments used here. Validating this panel in an independent multicentre cohort is the logical next step before considering clinical implementation^[11].

Conclusion

This research demonstrated that integrating genomic, proteomic, and metabolomic data through a latent-factor framework substantially improves the molecular classification of autoimmune disease subtypes compared with single-platform approaches. The integrated classifier achieved 87.4% accuracy across five subtypes and identified biologically interpretable factors linked to known disease pathways. A refined 50-feature biomarker panel offers a feasible target for development into a clinical diagnostic assay, pending validation in external cohorts of diverse ancestry.

References

1. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT. 2010 Rheumatoid arthritis classification criteria. *Arthritis and Rheumatism*. 2010;62(9):2569-2581.
2. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J. Analysis of five chronic inflammatory diseases

- identifies 27 new associations and highlights disease-specific patterns. *Nature Genetics*. 2016;48(5):510-518.
3. Rai MF, Patra D. Biomarkers for articular cartilage and osteoarthritis: recent advances and future challenges. *Osteoarthritis and Cartilage*. 2021;29(5):614-628.
4. Guleria A, Pratap A, Dubey D, Rawat A, Chaurasia S. NMR-based serum metabolomics discriminates Takayasu arteritis from healthy volunteers. *Journal of Proteome Research*. 2015;14(8):3480-3489.
5. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T. Multi-omics factor analysis: A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*. 2018;14(6):e8124.
6. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: an R package for omics feature selection and multiple data integration. *PLoS Computational Biology*. 2017;13(11):e1005752.
7. Ritchie MD, Holinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*. 2015;16(2):85-97.
8. Chiche L, Stand N, Stand N, Barnette T. Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures. *Arthritis and Rheumatology*. 2014;66(6):1583-1595.
9. Zhou W, Zhao Z, Zhong NL, Zhao Z. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics*. 2020;52(6):634-639.
10. Burr ML, Viatte S, Plant D, Symmons DPM, Thomson W, Barton A. Environmental triggers of RA: risk factors and biomarkers. *Current Rheumatology Reports*. 2017;19(10):63.
11. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nature Reviews Genetics*. 2018;19(5):299-310.
12. Pisetsky DS. Anti-DNA antibodies: quintessential biomarkers of SLE. *Nature Reviews Rheumatology*. 2016;12(2):102-110.
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL. Gene set enrichment analysis: a knowledge-based approach. *Proceedings of the National Academy of Sciences*. 2005;102(43):15545-15550.
14. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83.
15. Tebani A, Afonso C, Marret S, Bekri S. Omics-based strategies in precision medicine. *Clinical Chemistry*. 2016;62(10):1360-1371.
16. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. *New England Journal of Medicine*. 2011;365(23):2205-2219.
17. Tsokos GC. Systemic lupus erythematosus. *New England Journal of Medicine*. 2011;365(22):2110-2121.
18. Wishart DS. Metabolomics for investigating physiological and pathophysiological processes. *Physiological Reviews*. 2019;99(4):1819-1875.