School of Basic & Medical Biosciences
King's College London
United Kingdom

7BBG1006 Extended Research Project in Applied Bioinformatics

PlaqueMS: An Integrative Web Platform for Atherosclerosis Omics Analysis

---

Name: Nikolaos Samperis

Student Number: 24105804

Course: Applied Bioinformatics

Supervisor: Dr Konstantinos Theofilatos

Word count: 9,766

Acknowledgements

I would like to express my sincere gratitude to all individuals and institutions who contributed to the completion of this project.

First and foremost, I am deeply grateful to Dr. Konstantinos Theofilatos for his outstanding supervision, continuous guidance, and invaluable support throughout the course of this work.

I gratefully acknowledge the following collaborators for providing essential proteomics datasets and their expertise:

- Dr. Xiaoke Yin and Prof. Manuel Mayr (Imperial College London)
- Dr. Stefan Stojkovic and Prof. Johann Wojta (Medical University of Vienna)
- Dr. Sander van der Laan and Prof. Gerard Pasterkamp (University Medical Center Utrecht, Athero-Express Biobank Study)
- Dr. Clint Miller and Prof. Mete Civelek (University of Virginia)

Special thanks are due to Mr. Thakorn Pruktanakul for his technical support and assistance within the Cardiovascular Bioinformatics Lab.

The successful completion of this research would not have been possible without the collaboration, expertise, and data sharing from all the above colleagues and institutions.

Nomenclature

AUC	Area Under the Curve
CAD	Coronary Artery Disease
CI	Confidence Interval
CTA	Computed Tomography Angiography
CT	Computed Tomography
CVD	Cardiovascular Disease
CSRF	Cross-Site Request Forgery
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
ECM	Extracellular Matrix
EA	Evolutionary Algorithm
GDPR	General Data Protection Regulation
GWAS	Genome-Wide Association Studies
KNN	K-Nearest Neighbors
MCL	Markov Clustering
ML	Machine Learning
ORM	Object-Relational Mapping
PCI	Percutaneous Coronary Intervention
PPI	Protein–Protein Interaction
RBF SVM	Radial Basis Function Support Vector Machine
ROC	Receiver Operating Characteristic
scRNA-seq	Single-cell RNA sequencing
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
SYNTAX	Synergy between PCI with Taxus and Cardiac Surgery
TMT	Tandem Mass Tag
UUID	Universally Unique Identifier

Abstract

Background: Atherosclerosis remains a leading cause of cardiovascular morbidity and mortality, yet its molecular underpinnings are only partially understood due to the complexity of vascular lesions, the lack of specialised analytical tools, and the fragmented, static nature of available omics datasets.

Methodology: To address these gaps, this project introduces PlaqueMS, a Django-based web platform integrating MySQL and Neo4j databases for robust, phenotype-driven retrieval of multi-cohort proteomics and clinical metadata from human atherosclerotic plaques. The platform features interactive protein–protein network visualisation (Cytoscape.js) and embedded machine learning models (random forest, SVM, elastic net regression) trained via a multi-objective evolutionary algorithm to predict clinical endpoints such as plaque calcification status and SYNTAX score.

Results: Across 1,959 proteins from 219 carotid endarterectomy plaque samples (120 patients), the best-performing model achieved a mean AUC of 0.81 in internal cross-validation for discriminating calcification status, and demonstrated significant correlation with quantitative CT metrics (Spearman's $\rho \approx 0.55$, $p < 0.01$) and CTA categories (Spearman's $\rho = 0.48$, $p = 0.0085$), thereby uniquely capturing gradation of calcification burden. In cross-cohort validation, the second-best performing model achieved modest discrimination (AUC = 0.63), with asymptomatic cases exhibiting significantly higher median predicted probabilities of calcification than symptomatic ones (0.76 vs. 0.58; $p = 0.041$). Feature selection via the multi-objective evolutionary algorithm consistently prioritised canonical regulators of vascular calcification, including fetuin-A (AHSG) and osteopontin (OSTP), supporting the biological relevance of the predictive models.

Conclusion: Despite current limitations such as cohort heterogeneity, exclusive reliance on proteomics, and restricted external validation of the embedded predictive models, PlaqueMS represents a significant advance towards more dynamic, accessible, and translational research tools for atherosclerosis. Its modular design invites future expansion to additional omics layers and larger patient cohorts, positioning it as a valuable resource for advancing mechanistic understanding and risk assessment in cardiovascular disease research.

Contents

1. Introduction

1.1. Background on Atherosclerosis

Cardiovascular disease (CVD) is the foremost cause of global mortality, claiming nearly 18 million lives annually [1]. Atherosclerosis—a chronic, multifactorial condition—underpins most forms of CVD, including coronary artery disease, stroke, and peripheral artery disease. It involves the accumulation of lipids, immune cells, and fibrous tissue in the arterial walls, forming plaques that narrow the vascular lumen and can rupture, leading to acute clinical events like myocardial infarction and ischemic stroke [2].

While the clinical consequences of plaque rupture are well-documented, the molecular mechanisms underlying plaque instability remain only partially understood. The response-to-retention hypothesis suggests that atherogenesis begins with the binding of cholesterol-rich lipoproteins to intimal proteoglycans, triggering oxidation, foam cell formation, and inflammation [3]. Beyond lipoprotein retention, degradation of the extracellular matrix (ECM), particularly within the fibrous cap, plays a critical role in destabilising plaques [4]. Characterising ECM composition across plaque phenotypes is therefore crucial. The matrisome, encompassing ECM and associated proteins [5], can be profiled through proteomics, providing insights that go beyond histology and imaging assessments. Yet, current mapping efforts still lack comprehensive data from large blood vessels and atherosclerotic plaques [6].

1.2. Current Research on Atherosclerosis

The pathogenesis of atherosclerosis has been increasingly elucidated through the lens of omics technologies, transforming its conceptualisation from a cholesterol-driven disease to a complex interplay of genetic, transcriptomic, proteomic, and metabolic processes. Each omics domain has contributed unique insights into disease onset and progression, but substantial gaps remain in fully characterising the molecular mechanisms within vascular lesions—particularly in human tissue.

Genomic and transcriptomic investigations have made significant strides in identifying loci and regulatory elements implicated in coronary artery disease (CAD). Large-scale genome-wide association studies (GWAS) have uncovered over 160 risk loci associated with CAD, including PCSK9, SORT1, and GUCY1A3, which influence lipid metabolism, inflammation, and vascular remodelling [7]. Parallel transcriptomic analyses have revealed key cellular pathways, such as NF-κB-mediated inflammation and smooth muscle cell transdifferentiation, that underlie plaque formation and instability [8], [9]. Spatial transcriptomic methods have enabled the mapping of gene expression within specific plaque regions, but many of these studies suffer from scalability issues and limited tissue representation. For example, Nguyen et al. [10] highlighted how spatial transcriptomics in cardiovascular tissue is still constrained by high costs and small datasets, limiting its use in larger vascular biobanks.

Metabolomics and lipidomics have further enriched our understanding of atherosclerosis by profiling circulating biomarkers reflective of vascular dysfunction. Metabolite panels

including phosphatidylcholine, sphingomyelins, and ceramides have been associated with increased cardiovascular risk and have demonstrated added predictive value for major adverse events in patients with stable CAD [11]. However, while these approaches offer non-invasive diagnostic potential, they often lack specificity in linking metabolic changes to plaque-resident molecular processes [10]. This disconnection makes it difficult to infer causality or define therapeutic targets without concurrent proteomic or transcriptomic support.

In contrast, while proteomics provides direct insight into the protein machinery underpinning cellular function, its application in studies of human atherosclerotic lesions remains relatively underrepresented compared to transcriptomic and genomic approaches—particularly in spatial and systems-level investigations [12]. While several studies have employed mass-spectrometry to examine protein signatures in plasma or lipoprotein fractions, few have focused on the diseased vascular wall itself [13]. This scarcity of tissue-based proteomics is largely attributable to technical barriers, as obtaining high-quality plaque samples from human subjects is difficult, and the insoluble, complex nature of ECM proteins poses further hurdles for extraction and analysis [13].

Nevertheless, notable progress has been made in characterising the vascular wall proteome. For instance, Lorentzen et al. [14] conducted a proteomic analysis of 21 human carotid plaques and identified over 4,000 proteins, including 354 related to the ECM. Although valuable, the small cohort limits its statistical power and generalizability. Similarly, Kalló et al. [15] applied data-independent acquisition (DIA) and data-dependent acquisition (DDA) mass-spectrometry to profile protein networks in human carotid plaques but relied on fewer than 25 samples. These studies highlight the logistical and technical challenges of working with human atherosclerotic tissue, especially when attempting to profile matrix-bound proteins or stratify results by clinical outcomes.

Further complicating the landscape, many proteomic and transcriptomic studies continue to rely on non-human models, such as ApoE−/− or LDLR−/− mice [16]. While these models recapitulate some features of atherosclerosis, they do not exhibit critical human complications such as plaque rupture, thrombosis, or clinically significant coronary artery disease, which are either absent or extremely rare in murine systems [16]. For example, Martinez-Campanario et al. [17] used ApoE−/− mice with macrophage-specific ZEB1 deficiency to explore mechanisms of plaque instability, reporting increased lipid accumulation and necrotic core expansion—hallmarks of advanced lesions. Yet, their inability to model the full spectrum of human pathology underscores the translational limitations of murine systems in atherosclerosis research.

In summary, omics-driven research has uncovered important aspects of atherosclerosis biology, but its full potential remains constrained by methodological and sample-related limitations. Small cohort sizes, limited spatial resolution, and an overreliance on circulating or non-human data have impeded efforts to construct a comprehensive molecular atlas of human vascular lesions. Among the various omics, proteomics is particularly underexplored, especially in terms of large-scale, spatially resolved, and ECM-focused studies.

1.3. Existing Tools for Proteomic Analysis in Atherosclerosis

Several tools currently support the analysis and visualisation of proteomic data, each offering distinct strengths depending on the user's expertise and objectives. General-purpose tools such as Cytoscape [18] and STRING [19] are widely used for visualising protein–protein interaction networks and conducting functional enrichment analysis. These platforms support broad biological inquiries but are not tailored to specific diseases or tissue contexts such as atherosclerosis.

Applications like TraianProt [20] offer interactive modules for visualising differential expression and conducting enrichment analyses of proteomic datasets, supporting workflows from common search engines and quantification methods. However, such tools are typically generic in scope and do not provide phenotype-linked exploration or integration with clinical outcomes.

PlaqView 2.0 [21], although primarily developed for cardiovascular single-cell transcriptomic datasets, provides users with an intuitive interface to explore gene expression across cell types and conditions in vascular tissues. Even though it offers a valuable template for user interaction and disease-specific data visualisation, it is limited to transcriptomics and does not currently support proteomic data layers.

While these tools provide valuable capabilities, interactive platforms specifically designed for exploring proteomic data in atherosclerosis remain scarce. Existing tools either lack cardiovascular specialisation or do not include plaque phenotype–specific filtering, clinical linkage, or support for extracellular matrix analysis—features critical for translating proteomic findings into mechanistic and clinical insight. This scarcity of disease-focused, user-friendly, and sustainable platforms is also highlighted in recent reviews [22], which emphasise the fragmented and often short-lived nature of cardiovascular bioinformatics resources, largely due to challenges in funding and long-term maintenance. These persistent gaps underscore the need for robust, specialised tools that can support advanced and clinically relevant analysis in atherosclerosis research.

1.4. Network-Based Approaches in Proteomic Analysis

Protein–protein interaction (PPI) networks are essential for understanding the systems-level biology of atherosclerosis. Rather than analysing proteins in isolation, PPI network analysis reveals how groups of proteins function cooperatively within molecular pathways, cellular processes, or signalling cascades. In the context of atherosclerosis, this network-

based approach is particularly valuable for identifying hubs and bottlenecks—proteins that may regulate plaque development, inflammation, or stability.

As noted earlier, widely used platforms facilitate the construction and display of these networks by integrating proteomic data with known interaction databases and pathway annotations [18], [19]. These interactions are derived from a combination of experimental studies, computational predictions, and text mining of scientific literature, enabling a comprehensive view of both validated and predicted protein relationships [19]. Expanding upon such methodologies, network-oriented proteomic profiling of human atherosclerotic plaques has delineated region-specific protein signatures—such as increased levels of LUM, BGN, and VCAN in fibrous cap regions—associated with smooth muscle cell phenotypic modulation and extracellular matrix organisation [23], highlighting the complex heterogeneity that underpins plaque development and stability. Network pharmacology analyses have further demonstrated the utility of these approaches for therapeutic discovery, identifying, for instance, 31 protein targets of wogonoside within the TLR4/NF-κB signalling pathway, thereby elucidating mechanistic links between compound activity and inflammatory signalling [24].

In a complementary study, plasma proteomic profiles from individuals with CAD were subjected to network-based investigation, pinpointing core regulatory proteins—such as FN1, IL6R, and C5a—linked to inflammation and extracellular matrix remodelling, thereby reinforcing their implication in disease progression [25]. Collectively, these applications underscore the utility of network-based analyses in elucidating biologically relevant interactions and advancing the identification of candidate biomarkers and therapeutic targets in the context of atherosclerosis.

1.5. Machine Learning in Cardiovascular Research

Machine learning (ML) has become an increasingly prominent tool in cardiovascular research, offering the potential to model complex, high-dimensional datasets and uncover novel predictive insights. In the context of atherosclerosis, ML methods have been applied to a range of tasks, including risk stratification, image-based plaque assessment, and integrative multi-omics analysis. For instance, Chen et al. [26] utilised support vector machines (SVMs) and random forest algorithms to predict atherosclerotic risk from clinical and biochemical parameters, outperforming traditional scoring models and demonstrating the potential of ML in diagnostic enhancement. In parallel, ML models trained on coronary computed tomography angiography (CCTA) images have shown high accuracy in segmenting coronary plaques and quantifying stenosis severity, with strong agreement to expert interpretation and intravascular ultrasound, and predictive utility for future myocardial infarction [27].

Beyond clinical datasets, the integration of multi-omics data has opened new avenues for mechanistic insight. Usova et al. [28] combined genomic, transcriptomic, and proteomic features within ML frameworks to refine cardiovascular risk prediction and highlight biologically meaningful markers. Such integrative approaches represent a shift towards

more comprehensive modelling of disease complexity, capturing molecular heterogeneity that might otherwise be overlooked.

Despite these advances, significant challenges remain. Many ML applications suffer from limited sample sizes and cohort imbalances, reducing generalizability and increasing the risk of overfitting [29]. Furthermore, a lack of external validation using independent cohorts is common, limiting confidence in model robustness and hindering broader clinical applicability [30]. Integration across heterogeneous omics layers often lacks standardised methodologies, complicating reproducibility and downstream interpretation [31]. A further limitation lies in the interpretability of many ML models, which despite strong predictive performance, frequently operate as black boxes, offering little transparency into the biological mechanisms they exploit [32]. Feature redundancy is another prevalent issue, particularly in high-dimensional datasets—characteristic of omics technologies—where correlated or non-informative variables can obscure interpretation and inflate computational burden [33]. Most importantly, the absence of accessible, exploratory platforms means that many ML tools are not yet positioned to support real-time clinical or experimental decision-making [29].

1.6. Scientific Relevance

The significance of studying atherosclerosis lies not only in its prevalence but also in its silent progression and devastating outcomes. Most individuals remain asymptomatic until the disease reaches an advanced stage or manifests suddenly through life-threatening complications [34]. Given the multifactorial nature of atherosclerosis—driven by genetic, metabolic, immunological, and environmental factors—it is an ideal candidate for systems-level investigation using modern omics technologies [35].

To date, most efforts to characterise atherosclerotic plaques at the molecular level have focused predominantly on transcriptomics [36],[37],[38]. Nevertheless, transcriptomic analysis offers only an indirect approximation of cellular function, whereas proteomic data more directly reflect the biochemical and structural state of the tissue. At the same time, the application of machine learning to high-dimensional omics data for clinically meaningful outcome prediction in atherosclerosis remains underdeveloped [39]. Despite growing interest, few studies have successfully validated such models across independent cohorts or translated them into clinical practice [30]. Fragmented integration of multi-omics data, small retrospective datasets, feature redundancy, and limited model interpretability continue to restrict their broader utility, even as these tools hold strong potential for advancing personalised risk assessment and improving outcome prediction [39].

One of the few notable contributions addressing some of these gaps is the study by Theofilatos et al. [40], who employed tandem mass tag (TMT)-based quantitative proteomics to systematically profile the extracellular matrix composition of human atherosclerotic plaques. To enhance the biological interpretation of their proteomic data, the authors incorporated spatial transcriptomics and leveraged publicly available single-cell RNA-sequencing (scRNA-seq) datasets, enabling cell-type-specific mapping of ECM

protein expression. Through this integrative approach, they identified distinct molecular plaque phenotypes associated with calcification, inflammation, and sex-specific differences, and demonstrated that matrisome-level proteomic signatures—when combined with machine learning—can surpass conventional imaging and histological assessments in predicting clinical outcomes. Additionally, they leveraged protein interaction network analysis to contextualise ECM proteins within functional modules, identifying interaction clusters associated with specific plaque phenotypes and biological processes. Overall, these findings underscore the utility of proteomics—particularly when complemented by transcriptomic data and predictive modelling—in capturing the functional and cellular complexity underlying plaque stability and patient-specific cardiovascular risk.

Despite these advances, the translational utility of such proteomic findings remains constrained by the absence of accessible, interactive tools that allow clinicians and researchers to explore and interrogate the data in a meaningful and integrative manner. This challenge is further amplified in the context of multi-omics studies, which generate large-scale, heterogeneous datasets requiring sophisticated integration, harmonisation, and interpretation across molecular layers [41]. Yet, most existing resources present results in static formats, offering limited support for dynamic exploration, clinical metadata integration, or predictive modelling [42]. As a result, the full potential of multi-omics data to elucidate disease mechanisms, enable patient stratification, and inform therapeutic development in atherosclerosis remains insufficiently leveraged.

These limitations underscore the urgent need for integrative platforms that bridge the gap between high-dimensional omics research and practical clinical utility. Tools capable of enabling dynamic data interrogation, phenotype-specific exploration, and mechanistic interpretation are essential to unlock the translational potential of multi-omics approaches in atherosclerosis research.

1.7. Project Aims and Objectives

In response to these challenges, this project proposes the development of PlaqueMS, a unified web-based platform that integrates multiple omics layers—primarily proteomic data—derived from human atherosclerotic plaques. The platform addresses key limitations identified in current research: fragmented multi-modal omics integration, limited application of predictive modelling with demonstrated cross-cohort performance, and lack of accessible, interactive tools for data exploration. It enables users to perform phenotype-driven queries, visualise protein expression across tissue types and experimental conditions, explore PPI networks, and apply machine learning models to predict clinically relevant outcomes such as plaque calcification and SYNTAX (Synergy between PCI with Taxus and Cardiac Surgery) score. By operationalizing previously static datasets, the platform builds upon the foundational work of Theofilatos et al. [40], while also drawing on the architectural and functional concepts introduced in two prior student-led projects, one focused on protein network visualisation and the other on phenotype-based omics data exploration. Together, these elements transform complex molecular

profiles into a user-oriented analytical framework that facilitates hypothesis generation and supports more informed, translational cardiovascular research.

To achieve these aims, the project pursued the following specific objectives:

1. Populate and extend a relational database with curated proteomic datasets derived from three patient cohorts across multiple experimental conditions, building upon the schema and scripts developed by Liu [43].
2. Construct a complementary graph database to capture relationships among patients, clinical metadata, experimental contexts, tissue regions, and protein abundance measurements, adopting the schema from Tsogtbaatar's prior work [44].
3. Incorporate network visualisation functionality to enable interactive exploration of PPI networks.
4. Train and embed machine learning models for the prediction of qualitative and quantitative endpoints, such as plaque calcification status and SYNTAX score.
5. Ensure secure access to the platform by implementing user authentication and access control mechanisms to protect sensitive data and support role-specific functionalities.
6. Develop a responsive and modular web interface that unifies exploratory and predictive components.

2. Methodology

2.1. Data Acquisition and Preparation

Proteomics data for this project were assembled from three independent patient cohorts: the Medical University of Vienna [40], the Athero-Express Biobank Study [45], and Stanford University in collaboration with the University of Virginia [46]. Both the Vienna and Athero-Express cohorts comprise fresh-frozen carotid plaque specimens obtained by endarterectomy, with the Vienna cohort including 219 samples from 120 patients, and the Athero-Express cohort contributing a subset of 200 samples (200 patients) drawn from the larger biobank. Notably, the Vienna cohort provides protein expression data from distinct plaque regions, supporting spatial analysis within lesions. These cohorts have been employed as both discovery and validation sets in the landmark studies by Theofilatos et al. [40] and Palm et al. [47] (with cohort assignments reversed), forming the basis for the most extensive proteomic analyses of fresh-frozen human carotid endarterectomy tissue to date (Supplementary Table 1).

The Virginia cohort consists of 150 fresh-frozen coronary artery plaque specimens (150 patients), collected from heart transplant recipients or donor hearts at Stanford University and subsequently analysed in collaboration with the University of Virginia [46]. While distinct from the carotid cohorts in vascular origin, these samples expand the dataset to encompass advanced coronary atherosclerotic lesions, enabling comparative proteomic analyses across arterial territories. All datasets were systematically organised in a nested directory structure according to cohort, experimental protocol, and sample attributes, facilitating efficient integration into a MySQL database and ensuring traceable data management for downstream analysis.

2.2. Database Design and Implementation

A robust relational database was implemented using MySQL Server 8.0 [48] to facilitate the structured storage and efficient retrieval of protein annotation data (UniProt and gene identifiers for all proteins extracted from the three cohorts), precomputed statistical analysis results, protein network files, and comprehensive user records. The stored protein networks were generated using the Direct-AP methodology, which infers directed co-expression networks by quantifying direct associations between proteins via conditional mutual information [49].

The schema comprises core entities such as users, proteins, experiments, and datasets (each corresponding to a specific cohort), alongside specialised tables for analytical outputs (boxplots, volcano plots, heatmaps, and differential expression results) and protein network data. Unique string-based identifiers (UUIDs) serve as primary keys throughout, ensuring unambiguous referencing and maintaining data integrity, with foreign key constraints explicitly linking related entities. Intermediary tables are incorporated to efficiently represent many-to-many relationships within the database (Figure 1). All files are explicitly linked to their associated experiments and cohorts and referenced by relative paths that reflect the hierarchical directory structure of the data repository, ensuring

consistent organisation and traceability. This design enables advanced querying, flexible filtering, and seamless integration with the web application, providing a scalable foundation for reliable data management and access.

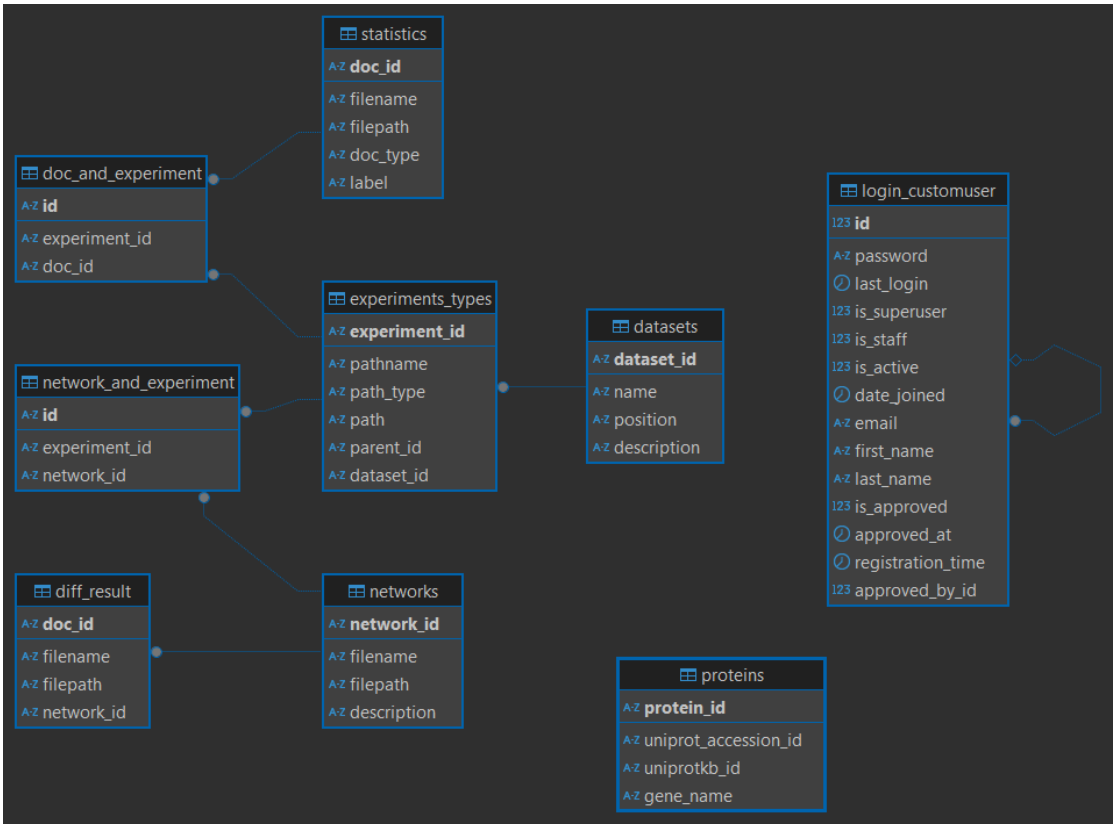


Figure 1.
Entity-relationship diagram of the core MySQL database schema. The diagram illustrates the main tables, their primary keys (indicated in bold), and the relationships between entities such as users, proteins, experiments, datasets, analytical results (statistics and diff_result tables), and protein networks. Intermediary tables are included to represent many-to-many relationships and to ensure efficient linkage between experimental data, analytical outputs, and network resources.

To complement the relational database, a graph database was developed using Neo4j (Desktop v1.6.1, database v5.24.2) [50] to capture complex biological relationships and metadata that are not efficiently represented in tabular form. The Neo4j schema currently models data from the Vienna cohort but is designed to accommodate additional cohorts with similar structure as the project evolves. It defines nodes for patients, experiments, samples, and proteins, with essential attributes—such as clinical metadata, tissue area, and protein identifiers—captured directly as node properties to streamline query performance (Figure 2). Over one million relationships capture the biological context, with "ABUNDANCE" edges linking sample to protein nodes and recording quantitative protein abundance values. Interactions between protein nodes are represented by "INTERACTS_WITH" edges, which encode network parameters including mutual information, directionality, p-values, and relevant experimental metadata.

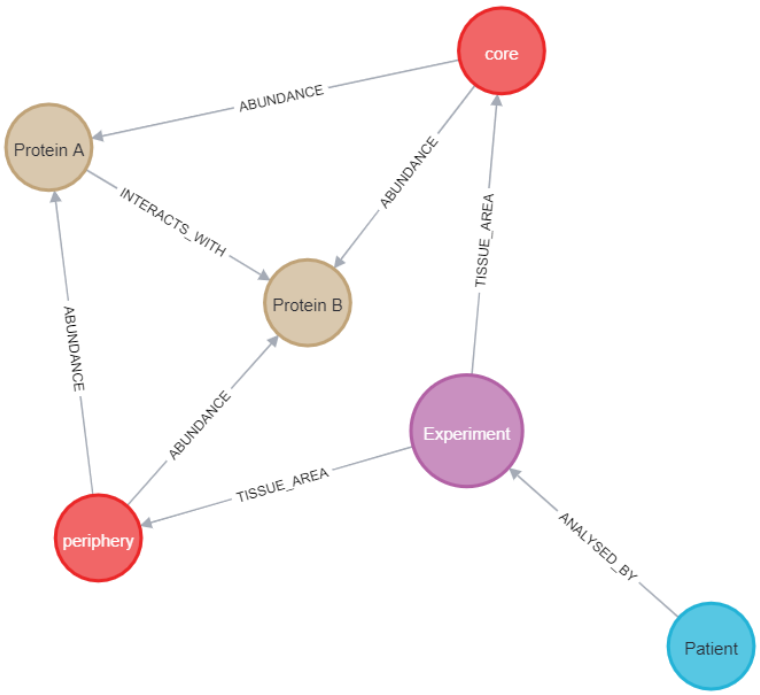


Figure 2.
Schematic representation of the Neo4j graph database structure used in this project. Nodes represent biological and experimental entities, including proteins, experiments (proteomics extraction protocol and cohort), patients, and tissue regions (e.g., core, periphery). Directed edges capture relationships such as protein abundance within specific tissue areas and protein–protein interactions, reflecting the complex connectivity of the underlying proteomics data.

The Neo4j database leverages the Cypher query language [50] for expressive and intuitive graph queries, supporting rapid traversal, filtering, and network-based exploration of proteomic data. This approach is particularly advantageous for integrating heterogeneous data types and visualising high-dimensional biological networks, enabling insights into complex relationships that extend beyond the capabilities of traditional relational models.

2.3. Web Application Development

The web application underpinning PlaqueMS was developed and tested locally (Windows 10 operating system) using Django (v5.1.7) [51], a high-level, open-source web framework based on Python, and widely adopted for constructing scalable, secure, and maintainable platforms. Django's architecture supports rapid development and modular design, while its object-relational mapping (ORM) enables seamless integration with relational databases such as MySQL, simplifying both data management and migration tasks. Additionally, Django provides a robust built-in navigation system for routing user requests, which facilitates the organisation of complex workflows and user interfaces. Its strong security features and extensive ecosystem of reusable components further enhance maintainability and extensibility, making it well-suited for research applications.

To support advanced phenotype-guided analysis, PlaqueMS incorporates the Neo4j graph database through the official Neo4j Python driver (v5.28.1) [52], allowing seamless execution of complex graph queries and dynamic data visualisation within the platform. This approach enables the efficient reconstruction and display of relationships among

patients, experimental protocols, sampled anatomical regions, and corresponding protein abundances, thereby facilitating comprehensive and interactive exploration of multi-dimensional biological data.

2.3.1. System Architecture

The overall system architecture of PlaqueMS adopts a modular and layered structure, as illustrated in Figure 3. At its core, the Django backend serves as the central controller, mediating all communication between user-facing interface modules, persistent data storage solutions, and external analytical engines. User requests are handled via HTTP/HTTPS protocols and routed through Django views and RESTful API endpoints (via Django REST Framework v3.15.2 [53]), which coordinate interactions with the underlying databases.

To support advanced protein network visualisation, PlaqueMS adopts a dual approach. Cytoscape.js (v.3.32.1) [54] is implemented on the front-end to promote fast, interactive exploration of protein interaction networks directly within the browser. For more complex tasks, such as automated network construction and node colouring based on differential expression results, the backend utilises Cytoscape Desktop (v3.10.3) [55] via the py4cytoscape Python library (v1.11.0) [56], which communicates with Cytoscape's CyREST API. This arrangement enables seamless integration of advanced analyses with interactive visualisations, ensuring both responsiveness and analytical depth.

To further enhance interactivity, Ajax calls are utilised in select frontend modules to allow for asynchronous communication with the server, supporting dynamic user actions and real-time updates without full page reloads. Furthermore, the backend incorporates pre-trained machine learning models, enabling PlaqueMS to deliver automated predictive analytics alongside interactive data exploration.

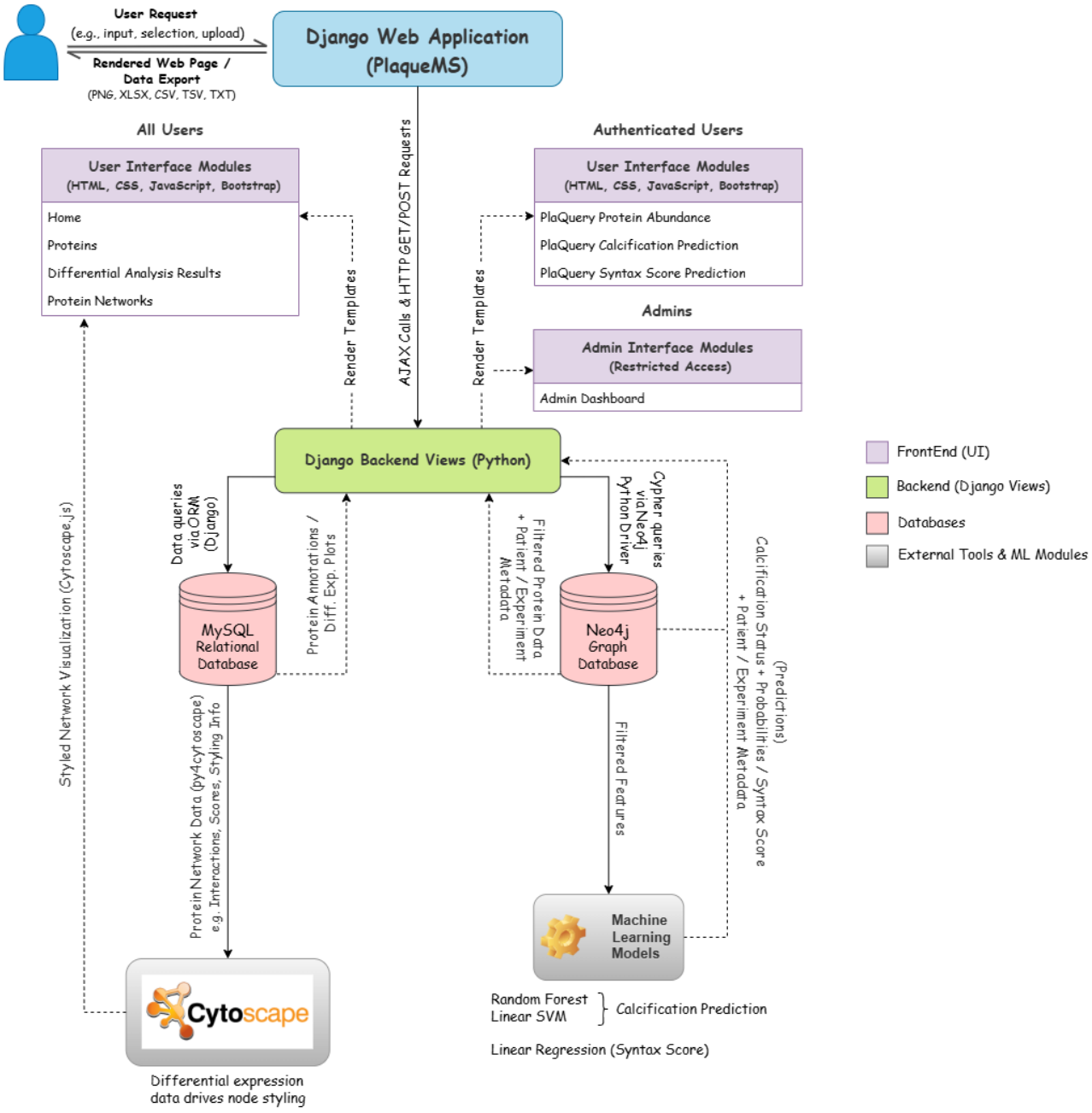


Figure 3.
High-level overview of the system architecture of PlaqueMS. Solid arrows indicate data flow from the frontend to the backend (user actions, HTTP/AJAX requests, and queries to the database), while dashed arrows represent data or results returned from the backend to the frontend, including rendered web pages, interactive visualisations, and downloadable files. Coloured boxes distinguish the platform's major components: purple denotes frontend user interface modules, green highlights backend logic and Django views, red indicates databases (MySQL and Neo4j), and grey represents external tools and machine learning modules. Certain modules and functionalities are accessible only to authenticated users or administrators. This diagram illustrates the coordinated interactions that enable seamless data analysis and visualisation within the platform.

2.3.2. Front-End Design

The front-end of the platform was developed using a suite of standard web technologies—HTML, CSS 3, and vanilla JavaScript—with Bootstrap [57] serving as the principal framework to ensure responsive design and stylistic uniformity across all modules. Both versions 3.3.7 and 5.1.3 of Bootstrap were employed in different components to maintain compatibility with legacy templates while leveraging newer features where appropriate, with custom styling used to harmonise the overall appearance. Interactivity and dynamic content rendering were achieved through JavaScript and jQuery (v3.1.0/3.6.0) [58], enabling a seamless and intuitive user experience for data querying, filtering, and visualisation.

Specialised libraries such as Bootstrap Tags Input (v0.8.0) [59] and Typeahead.js (v0.11.1) [60] support tokenised input and autocomplete, while icon sets from Font Awesome (v6.0.0-beta3) [61] and Bootstrap Icons (v1.10.5) [62] were incorporated to reinforce the platform's visual hierarchy and enhance navigational clarity. Browser-based search, pagination and table generation are implemented with bespoke JavaScript functions, removing any dependency on external table-handling plugins. Consistent typography is provided by Google Fonts Montserrat and Inter families [63], embedded across templates. Finally, Django's templating engine was employed to render views securely, integrating server-side data with client-side logic and enforcing user-specific access where required.

2.3.3. Back-End Implementation

The server layer is written in Python 3.11.9 [64], a language selected for its proven reliability, mature ecosystem, extensive adoption in scientific computing and web development, and full compatibility with the Django framework. Python's relatively gentle learning curve also facilitates efficient development and enhances the long-term sustainability of the platform.

All programmatic endpoints return data in JSON format via Django's JsonResponse [51], making responses directly consumable by the front-end's Ajax calls. Each reply encapsulates both the primary payload and metadata; for instance, prediction views bundle result dictionaries with warning flags and preprocessing notes before serialisation. This uniform contract simplifies error handling on the browser side and keeps the network footprint minimal.

A key feature of the back-end is its seamless integration of pre-trained machine learning models, all of which are developed and stored as serialised PKL files using the scikit-learn library (v1.1.3) [65]. To optimise performance, models and their associated data transformers are cached in memory after their initial load, reducing response times for repeated predictions.

Relevant modules also implement advanced file parsing and validation routines, supporting flexible data uploads in both wide and long formats (CSV, TSV, XLSX). Uploaded files are automatically harmonised and validated for downstream analyses, with users receiving informative feedback in the event of formatting or data quality issues. This

modular backend structure provides a scalable foundation for the future integration of new analytical features, ensuring that all computational logic remains centralised and reproducible across user sessions.

2.3.4. Data Protection and Access Management

To ensure compliance with the General Data Protection Regulation (GDPR) [66], rigorous access control mechanisms are enforced for all operations involving sensitive patient data. The authentication workflow utilises Django's secure session framework [67], requiring users to complete registration and subsequently obtain administrative approval prior to account activation. Unapproved accounts remain disabled, preventing unauthorised access to any protected endpoints or data resources.

User management operations are restricted to authorised administrative staff via a secured dashboard interface. The dashboard exposes granular controls for user approval, deactivation, and account deletion, with all approval actions automatically logged to an audit trail for traceability.

Account security is further reinforced by enforcing robust password complexity policies and employing cryptographically secure password hashing algorithms. All sensitive data transmissions are secured via HTTPS, and session integrity is protected through Django's middleware and Cross-Site Request Forgery (CSRF) safeguards [67].

2.4. Machine Learning Approach

To address the challenges of high-dimensionality and feature redundancy inherent to proteomics data, a multi-objective evolutionary algorithm (EA) was implemented—using code developed by Singh [68]—for feature selection prior to model training. Evolutionary algorithms are stochastic optimisation methods inspired by natural selection. In this project, the EA iteratively evolved a population of candidate feature subsets using crossover, mutation, and selection operators to efficiently explore the solution space while simultaneously choosing the classifier (random forest, linear SVM, or RBF SVM) that maximised performance for each subset. Fitness was assessed according to multiple objectives—namely, precision, recall (sensitivity), area under the receiver operating characteristic curve (ROC AUC), and minimal feature set size—with Pareto-optimal solutions retained to balance trade-offs between predictive performance and model simplicity. The specific EA configuration and training parameters are summarised in Table 1. By optimising for multiple outcomes simultaneously, this approach is particularly well suited for omics datasets, where many variables are highly correlated or uninformative.

For calcification prediction, three distinct training datasets were generated from the Vienna cohort, corresponding to soluble matrisome, core matrisome, and cellular proteome fractions (Supplementary Table 2). Each model was trained on features specific to its respective extraction protocol, capturing complementary aspects of plaque composition, with some proteins present in multiple extracts reflecting differences in solubility and subcellular localisation. Calcification status labels, determined by clinical and imaging assessment for each patient, were used as target outcomes for model training.

Missing values in each training set were imputed using the K-nearest neighbors (KNN) algorithm, and all protein intensity values were $\log_2$-transformed prior to analysis. For the calcification models, min-max scaling was applied to each feature to accommodate the requirements of SVM classifiers.

In contrast, the SYNTAX score prediction model was developed using a distinct dataset composed of label-free proteomics data (rather than TMT-labelled extracts) derived from coronary plaques in the Virginia cohort (25 patients), including proteins from both the core matrisome and the cellular proteome. Feature selection was conducted with the same multi-objective evolutionary algorithm approach, and the final predictive model was fitted using an elastic net regularised linear regression formula (Table 1).

	Calcification Prediction	SYNTAX Score Prediction
Input Data	TMT-labelled proteomics (soluble matrisome, core matrisome, cellular proteome) Carotid plaques (Vienna cohort)	Label-free proteomics (core matrisome + cellular proteome, combined) Coronary plaques (Virginia cohort)
Preprocessing	KNN imputation for missing values Log_2 transformation, Minmax scaling	KNN imputation for missing values Log2 transformation, Z-score scaling
Feature Selection	Multi-objective evolutionary algorithm (precision, sensitivity, ROC AUC; Pareto front) Trained for 200 generations, population size 400	Multi-objective evolutionary algorithm (Spearman's ρ, R^2, Mean Squared Error; Pareto front)
Model(s) Trained	Three separate models (one per dataset) Random Forest & Linear SVM (5–fold cross-validation, balanced classes, no oversampling)	One model Elastic Net Regularised Linear Regression (5–fold cross-validation)
Outcome	Predicts calcified vs non-calcified plaques All models embedded in web app	Predicts continuous syntax score Model integrated in web app

Table 1.
Summary of the machine learning pipelines implemented for calcification and SYNTAX score prediction. The table details the specific input proteomics datasets, data preprocessing workflows (including imputation and scaling methods), multi-objective evolutionary algorithm-based feature selection, model types and internal validation strategies, and target prediction outputs for each analysis.

To evaluate the clinical relevance of calcification prediction models, predicted probabilities were quantitatively assessed against independent imaging-derived measures. Specifically, correlations were computed with CT-based Agatston scores [69], calcium mass, and plaque volume measurements in 29 patients from the Vienna cohort. Additionally, model outputs were compared across CTA-derived calcification categories (non-calcified, mixed, calcified) in 35 patients from the same cohort using non-parametric statistical tests. These analyses were conducted using the scipy (v1.13.1) [70] and pandas (v2.2.2) [71] Python libraries.

Generalizability was further assessed by applying the two best-performing calcification models to the Athero-Express cohort, where direct calcification data were unavailable. In this cohort, symptom status (symptomatic vs. asymptomatic) from 200 patients served as an indirect indicator of plaque calcification, and group differences in predicted probabilities were evaluated using non-parametric statistical methods.

Cross-cohort validation was also performed for the SYNTAX score model using 109 patients from the Vienna cohort. Predicted SYNTAX scores were calculated for each patient and analysed for their association with CVD risk, symptom status, and plaque calcification. ROC (AUC) analysis was used to assess discriminatory performance for these clinical endpoints.

2.5. Code Availability

The source code and deployment scripts for the PlaqueMS platform, along with pre-trained machine learning model files and scripts for statistical analysis/plot generation, are available in a private GitHub repository. Code for model training is not included. Access is provided via the following fine-grained personal access token (read-only, valid until October 13, 2025):

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

Instructions for accessing the repository:

1. Go to: **https://github.com/NikolaosSamperis/PlaqueMS_project.git**
2. When prompted, use the above token for authentication.

The repository contains all necessary files and documentation for local deployment.

3. Results

3.1. PlaqueMS Interface & Core Features

The main dashboard of PlaqueMS organises its core exploratory and analytical functionalities into a set of modular interface components, several of which are accessible without registration. The following tools are available to all users:

3.1.1. Home Page

The Home page tab (Figure 4) functions as the primary entry point to the PlaqueMS platform, presenting an overview of the available datasets and integrated analytical capabilities. A persistent navigation bar enables users to seamlessly access the platform's main modules according to their role and authentication status. The layout is designed to optimise user orientation, directing both general and authorised users towards the resources and tools most pertinent to their access level.

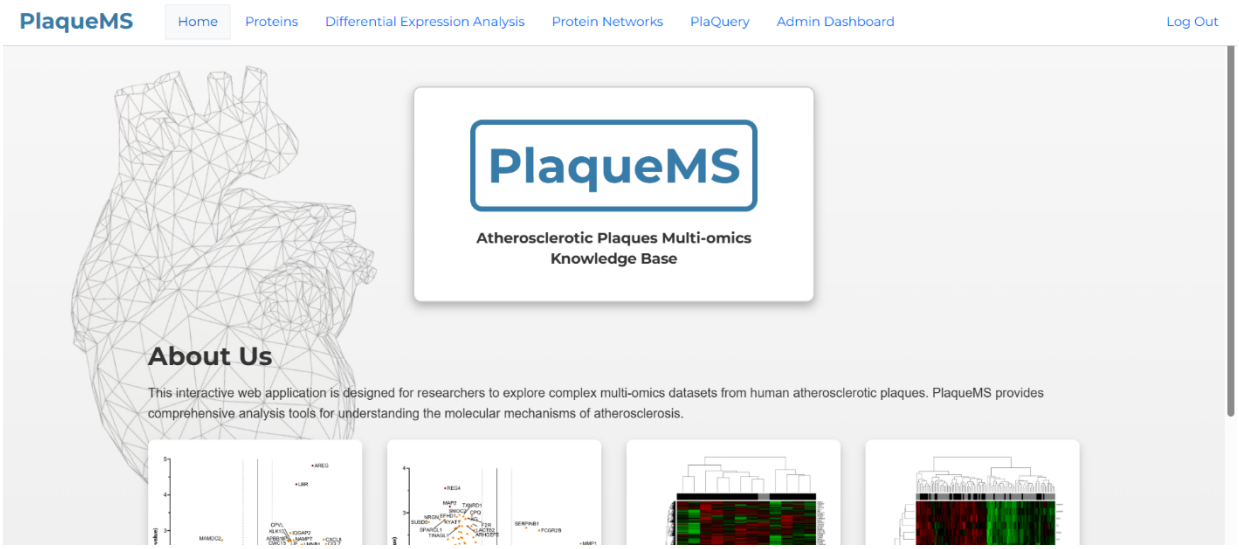


Figure 4.
Home page of the PlaqueMS platform as displayed to an administrator. All major modules—including Proteins, Differential Expression Analysis, Protein Networks, PlaQuery, and the Admin Dashboard—are accessible via the persistent navigation bar.

3.1.2. Proteins Page

Within the Proteins tab (Figure 5), users can explore a unified catalogue of unique proteins identified in human carotid and coronary atherosclerotic plaques across three different cohorts (Vienna, Athero-Express, and Virginia). Protein entries consolidate results from diverse proteomics methodologies, including tandem mass tag (TMT) labelling, label-free quantification, and sequential extraction, covering both cellular and extracellular

components. The multi-input search field allows for single or batch queries, with each entry dynamically converted into a searchable tag. Search results present standardised protein records with harmonised identifiers, enabling efficient, large-scale exploration and flexible mapping across annotation schemes.

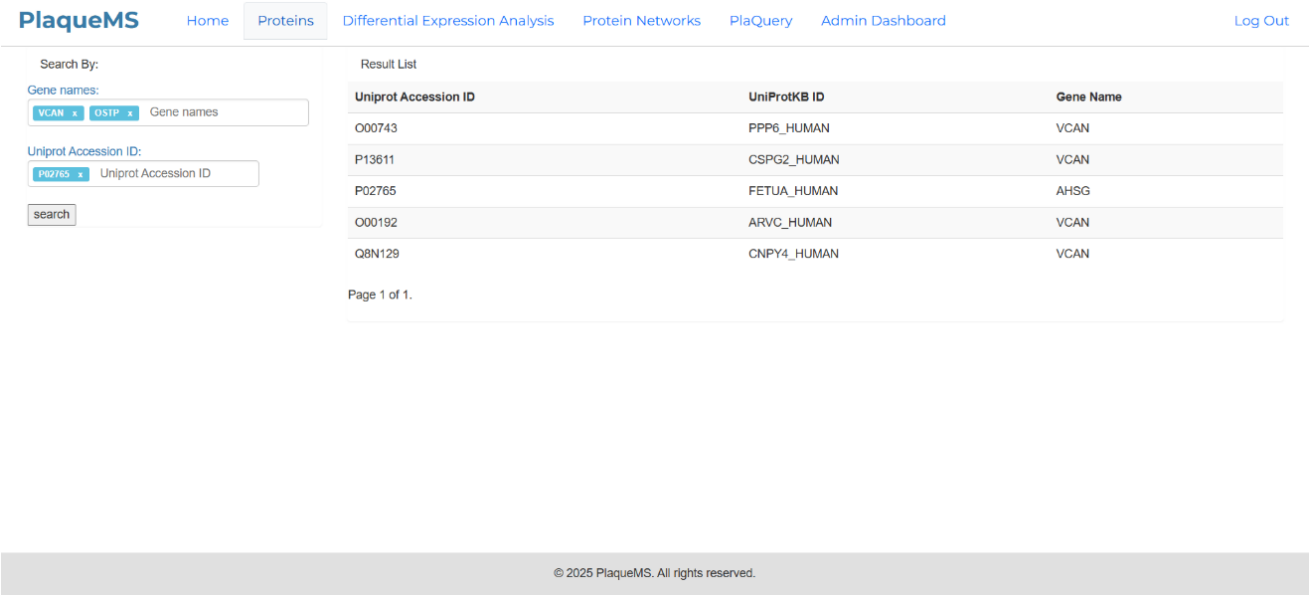


Figure 5.
Proteins page search results interface. This example shows the outcome of a combined query using both gene names and UniProt accession IDs, with each search term displayed as a tag.

3.1.3. Differential Expression Analysis Page

The Differential Expression Analysis module provides access to proteomic data from all three cohorts, enabling users to tailor their analysis by cohort, experimental protocol, and sample area (Figure 6). Statistical comparisons can be defined based on clinical phenotypes or treatment groups, generating quantitative protein expression profiles for the selected conditions. The resulting data can be visualised using precomputed boxplots, volcano plots, or heatmaps, offering multiple perspectives for examining differential expression patterns and facilitating the identification of proteins with significant (log) fold changes.

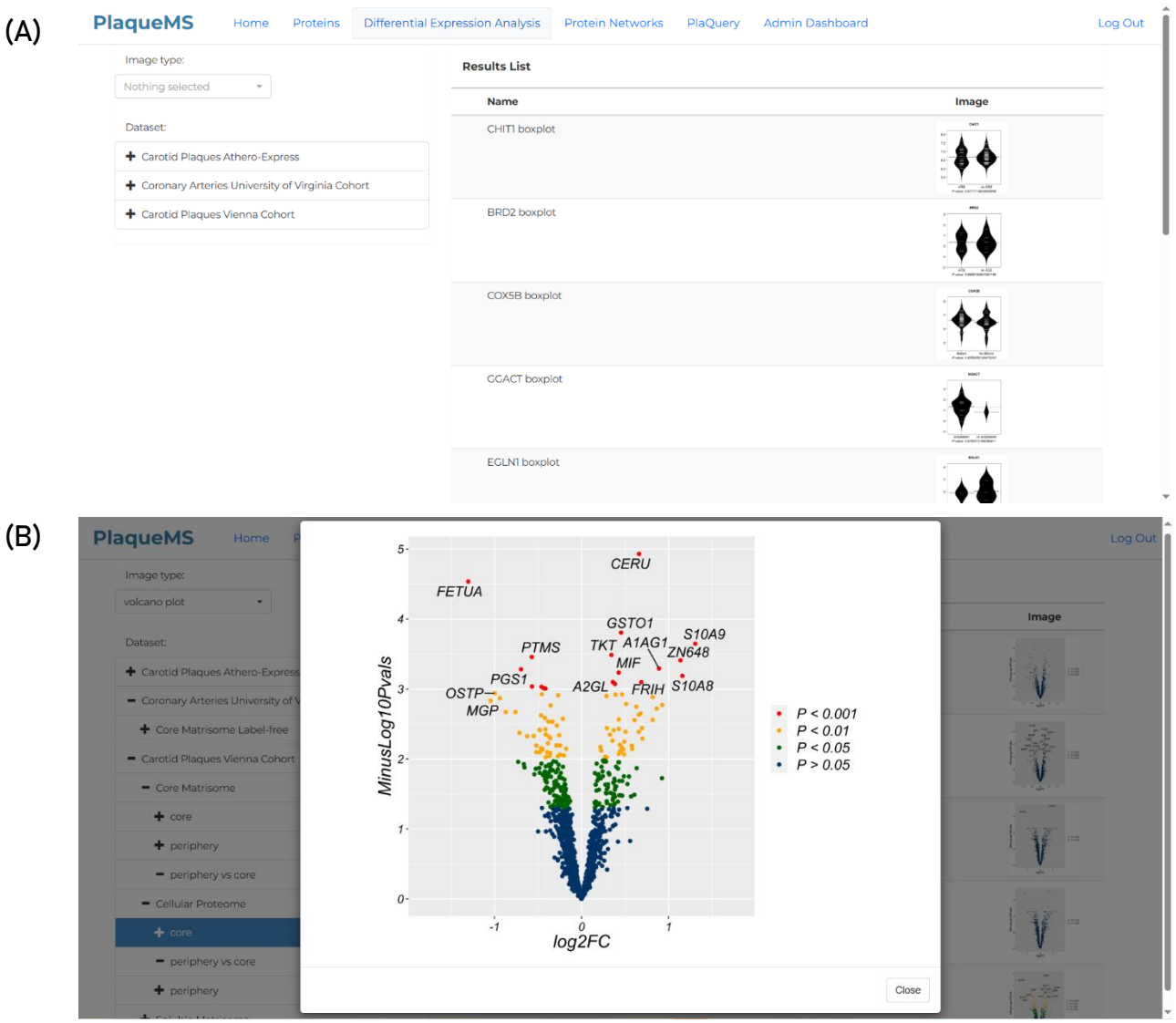
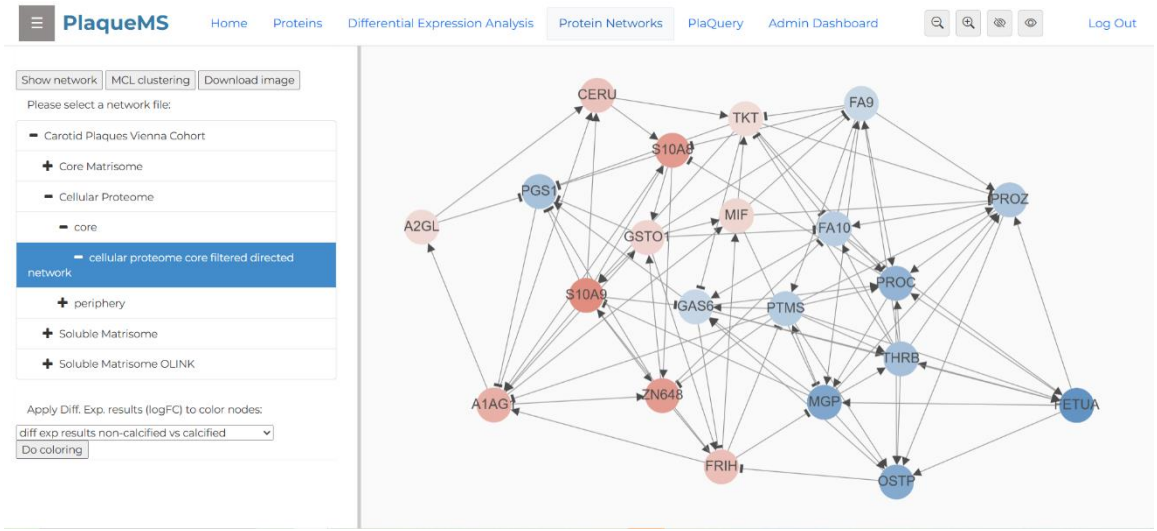


Figure 6.
Differential Expression Analysis module outputs. **(A)** Interface for selecting cohort, experimental protocol, sample area, and statistical comparison using a nested tree-like panel that organises available datasets hierarchically. Dynamic listing of protein expression plots is provided for each selected comparison. **(B)** Example volcano plot visualisation displaying differential protein expression between non-calcified and calcified plaque samples of the Vienna cohort, based on TMT proteomics analysis of the cellular proteome (core plaque region), with significance thresholds indicated by colour.

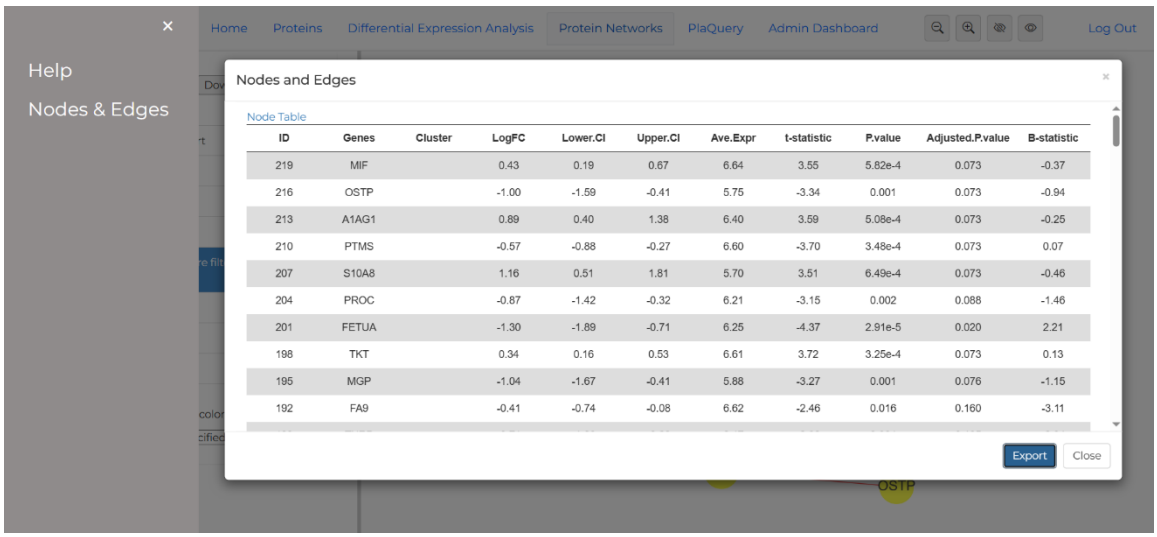
3.1.4. Protein Networks Page

The Protein Networks module (Figure 7) enables users to explore directed co-expression networks for each tissue, phenotype, and dataset. It currently supports only the Vienna cohort. Users can select specific experimental protocols and plaque regions, and choose to display either all available proteins, select from a scrollable gene list, or upload a custom file with gene names in standard formats (TXT, CSV, TSV, XLSX). The platform supports Markov Cluster (MCL) analysis [72] for community detection, and nodes can be coloured according to log fold change values from differential expression data—deeper red indicating upregulation and deeper blue indicating downregulation. Network edges show directionality: arrowheads indicate activating or directed interactions, while T-shaped heads represent inhibitory or suppressive effects. Users can view and export detailed statistics for selected nodes (expression profiles) and edges (mutual information, significance, and directionality metrics), as well as export a PNG image of the entire network or specific node subsets by hiding or unhiding selected elements, supporting comprehensive network analysis and data dissemination.

(A)



(B)



ID	Genes	Cluster	LogFC	Lower.CI	Upper.CI	Ave.Expr	t-statistic	P.value	Adjusted.P.value	B-statistic
219	MIF		0.43	0.19	0.67	6.64	3.55	5.82e-4	0.073	-0.37
216	OSTP		-1.00	-1.59	-0.41	5.75	-3.34	0.001	0.073	-0.94
213	A1AG1		0.89	0.40	1.38	6.40	3.59	5.08e-4	0.073	-0.25
210	PTMS		-0.57	-0.88	-0.27	6.60	-3.70	3.48e-4	0.073	0.07
207	S10A8		1.16	0.51	1.81	5.70	3.51	6.49e-4	0.073	-0.46
204	PROC		-0.87	-1.42	-0.32	6.21	-3.15	0.002	0.088	-1.46
201	FETUA		-1.30	-1.89	-0.71	6.25	-4.37	2.91e-5	0.020	2.21
198	TKT		0.34	0.16	0.53	6.61	3.72	3.25e-4	0.073	0.13
195	MGP		-1.04	-1.67	-0.41	5.88	-3.27	0.001	0.076	-1.15
192	FA9		-0.41	-0.74	-0.08	6.62	-2.46	0.016	0.160	-3.11

(C)

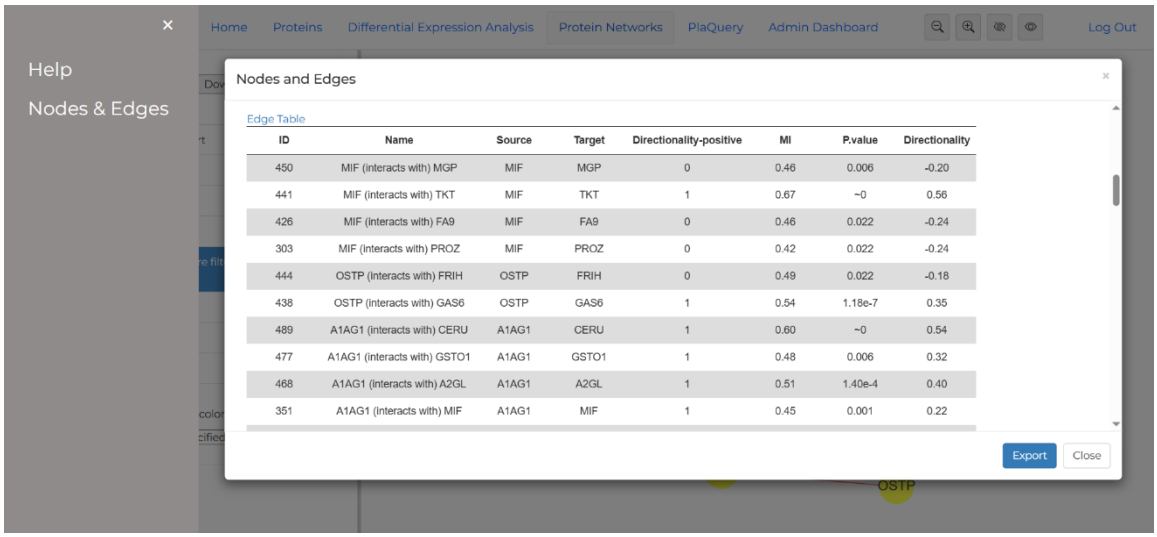
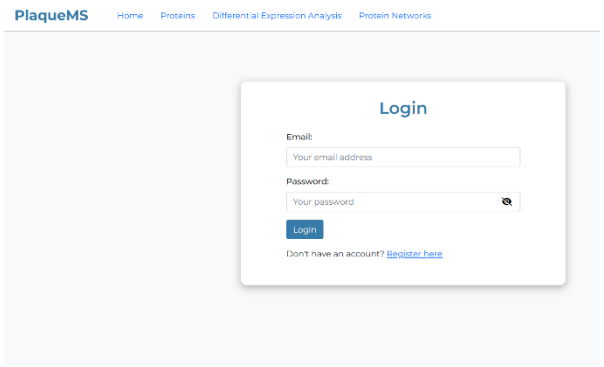


Figure 7.
Protein Networks module outputs. **(A)** Example directed co-expression network for the cellular proteome of the core plaque region in the Vienna cohort, displaying significantly differentially expressed calcification markers ($p < 0.01$), as depicted in the volcano plot of Figure 6. Node colour reflects log fold change in non-calcified versus calcified samples (red: upregulated, blue: downregulated). Edge arrows indicate activation/directed interaction; T-shaped edges indicate inhibition. **(B)** Node table displaying quantitative expression, log fold change, confidence intervals, and other statistical metrics for each protein in the network. **(C)** Edge table reporting mutual information, *p*-value, and directionality metrics for selected protein–protein interactions.

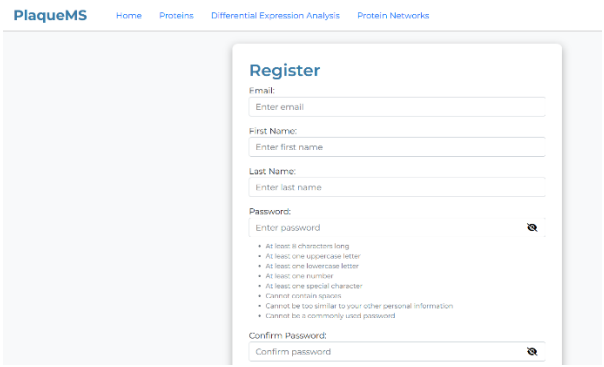
3.1.5. Authentication System Interface

The authentication system of PlaqueMS features login and registration modules (Figure 8A, 8B) that require unique email addresses per user and enforce strong password rules. Administrators manage users through a dedicated dashboard (Figure 8C), which enables approval, deactivation, and deletion of accounts, as well as providing general user information such as registration details and last login status, while ensuring all status changes are logged for auditability.

(A)



(B)



(C)

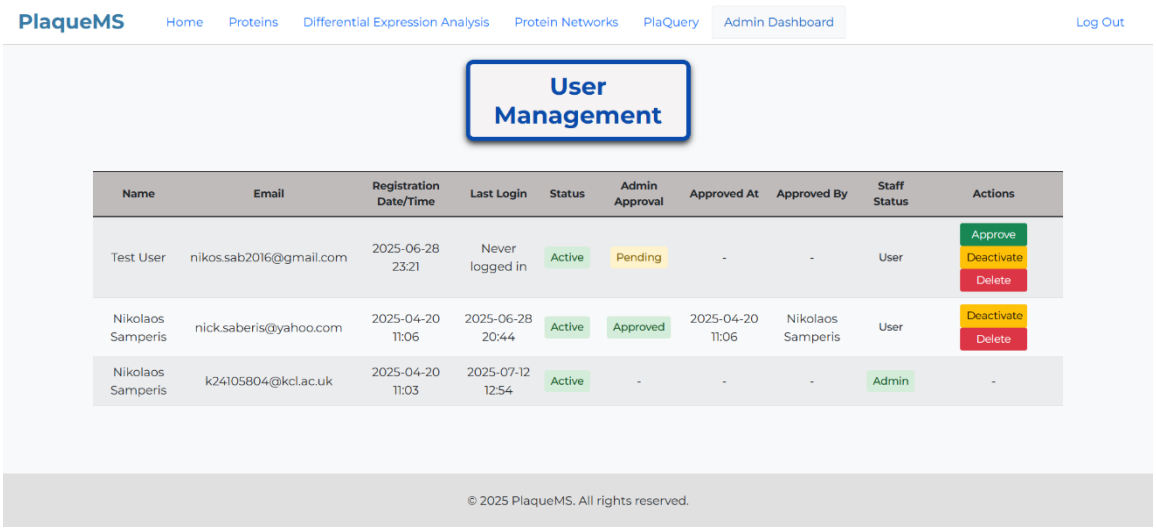


Figure 8.
Authentication system modules. **(A)** Login interface for user authentication. **(B)** Registration page, built using Django's default security forms, where new users provide contact details and create a password that meets robust security criteria. **(C)** Admin dashboard, allowing administrators to view and manage user accounts, including approval status, registration details, and last login information.

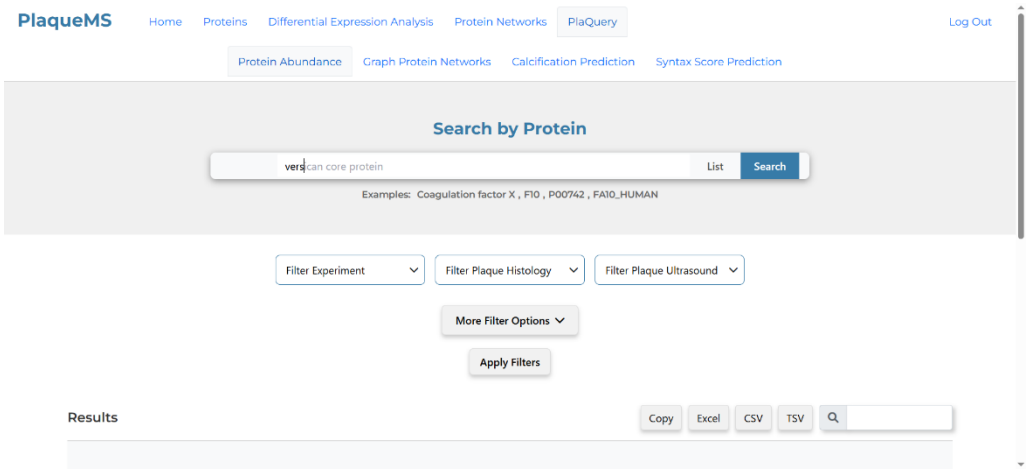
3.1.6. PlaQuery: Restricted Access Modules

In addition to the publicly accessible modules described above, PlaqueMS offers a dedicated section—PlaQuery—exclusively available to authenticated users. This restricted-access tab integrates sensitive patient metadata and comprises three additional tools that support phenotype-based queries for visualising protein abundances, alongside advanced predictive functionalities inaccessible to public users.

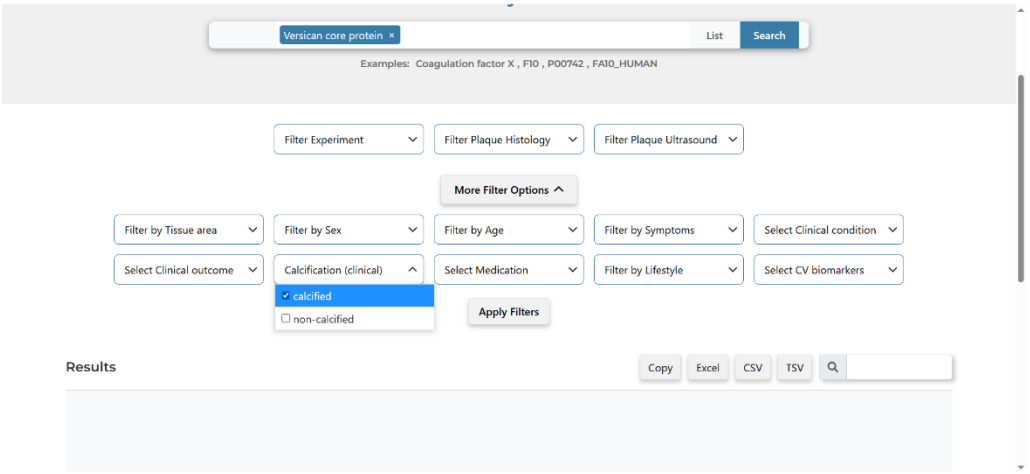
3.1.7. Protein Abundance Page

This tab provides a search engine for querying protein expression data within the Vienna cohort (Figure 9). Users can search by UniProt Accession ID, UniProtKB ID, gene name, or protein name, with auto-suggestions to streamline input. For each query, the interface returns individual protein abundances per patient, accompanied by aggregated statistics across all patients sharing the same experimental protocol and tissue region. When no filters are applied, the output includes full clinical metadata linked to each patient. Informative tool tips appear when hovering over table headers, offering additional context for each column. Batch queries of up to 50 proteins are supported via the "List" option, and all displayed data can be exported in various formats for downstream analysis.

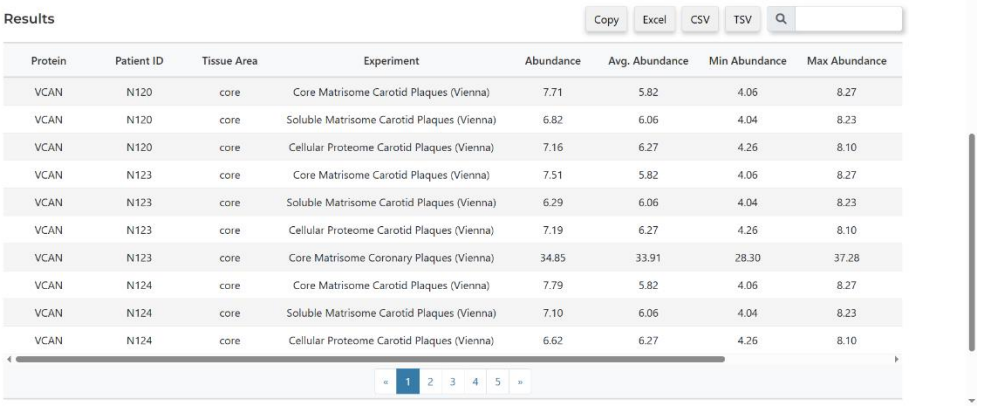
(A)



(B)



(C)



Results

Copy Excel CSV TSV

Protein	Patient ID	Tissue Area	Experiment	Abundance	Avg. Abundance	Min Abundance	Max Abundance
VCAN	N120	core	Core Matrisome Carotid Plaques (Vienna)	7.71	5.82	4.06	8.27
VCAN	N120	core	Soluble Matrisome Carotid Plaques (Vienna)	6.82	6.06	4.04	8.23
VCAN	N120	core	Cellular Proteome Carotid Plaques (Vienna)	7.16	6.27	4.26	8.10
VCAN	N123	core	Core Matrisome Carotid Plaques (Vienna)	7.51	5.82	4.06	8.27
VCAN	N123	core	Soluble Matrisome Carotid Plaques (Vienna)	6.29	6.06	4.04	8.23
VCAN	N123	core	Cellular Proteome Carotid Plaques (Vienna)	7.19	6.27	4.26	8.10
VCAN	N123	core	Core Matrisome Coronary Plaques (Vienna)	34.85	33.91	28.30	37.28
VCAN	N124	core	Core Matrisome Carotid Plaques (Vienna)	7.79	5.82	4.06	8.27
VCAN	N124	core	Soluble Matrisome Carotid Plaques (Vienna)	7.10	6.06	4.04	8.23
VCAN	N124	core	Cellular Proteome Carotid Plaques (Vienna)	6.62	6.27	4.26	8.10

« 1 2 3 4 5 »

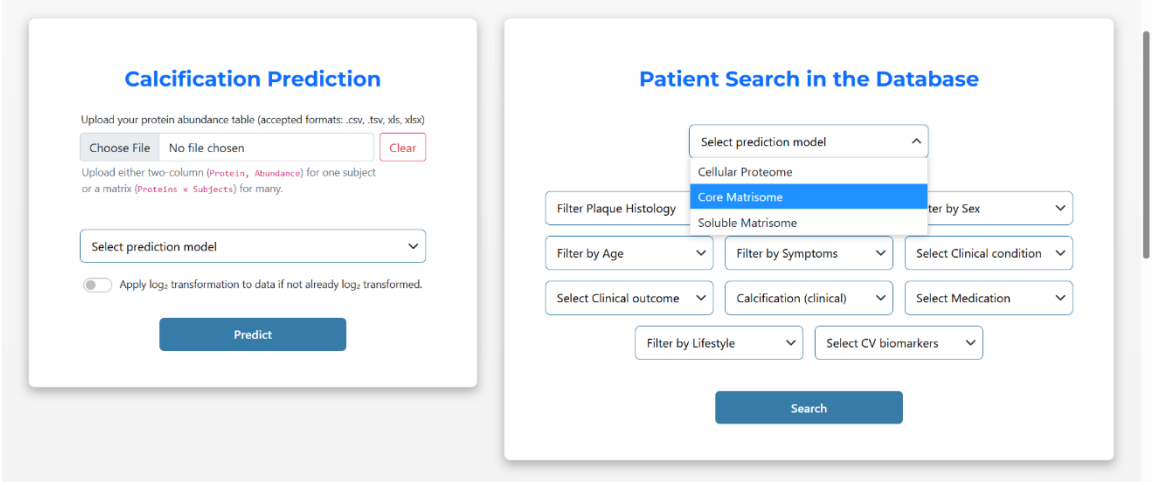
Figure 9.
PlaQuery: Protein Abundance Interface. **(A)** The main search engine for protein abundances, where users can enter UniProt IDs, Gene names, or Protein names, with support for auto-suggestions and batch queries via the "List" option. **(B)** The filtering interface, displaying advanced options for refining results based on experiment type, plaque histology, imaging, tissue area, and a range of clinical variables such as calcification status, sex, and medication. **(C)** Example results table showing individual protein abundances and summary statistics across patients, annotated with experiment protocol and tissue area. Data export options are provided in multiple formats.

3.1.8. Calcification status and SYNTAX score Prediction Pages

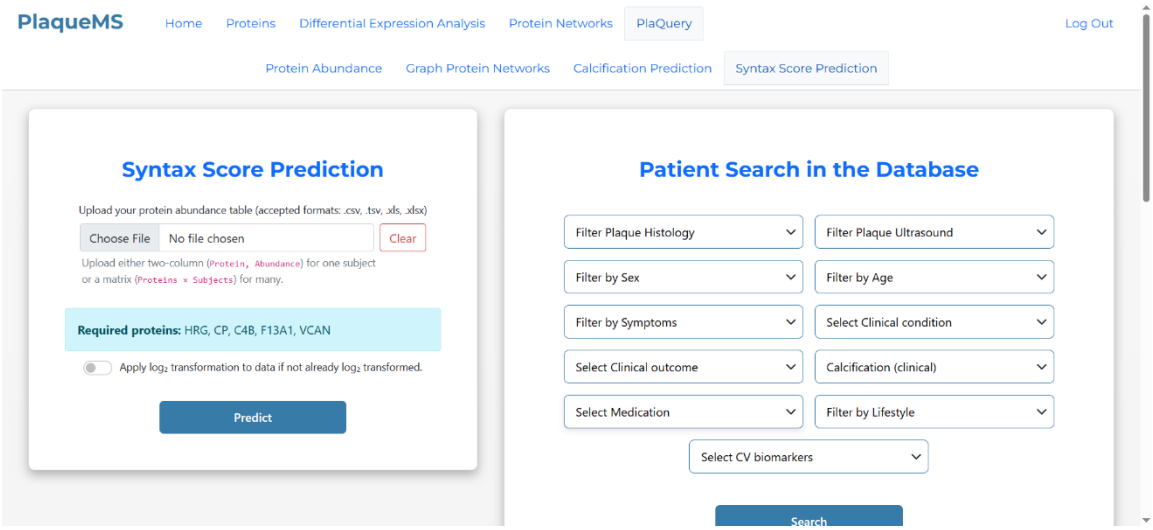
The Calcification Prediction and SYNTAX Score Prediction modules (Figure 10) provide a unified interface for applying trained models to proteomic data, either via user-uploaded protein abundance tables or direct patient queries from the Vienna cohort. For calcification prediction, users may select from three models corresponding to distinct extraction protocols, each specifying the required input features; SYNTAX score prediction utilises a single model with fixed input requirements. Log_2 transformation can be applied to input data if needed.

KNN-based imputation is automatically performed for missing protein abundances, supporting up to 50% missingness; if over 25% of required features are missing, a warning is issued with the prediction, while predictions are withheld entirely if missingness exceeds 50% to mitigate bias. Both modules support phenotype-based filtering and return results with relevant clinical metadata, similar to the PlaQuery Protein Abundance tool, and provide export options for downstream analysis.

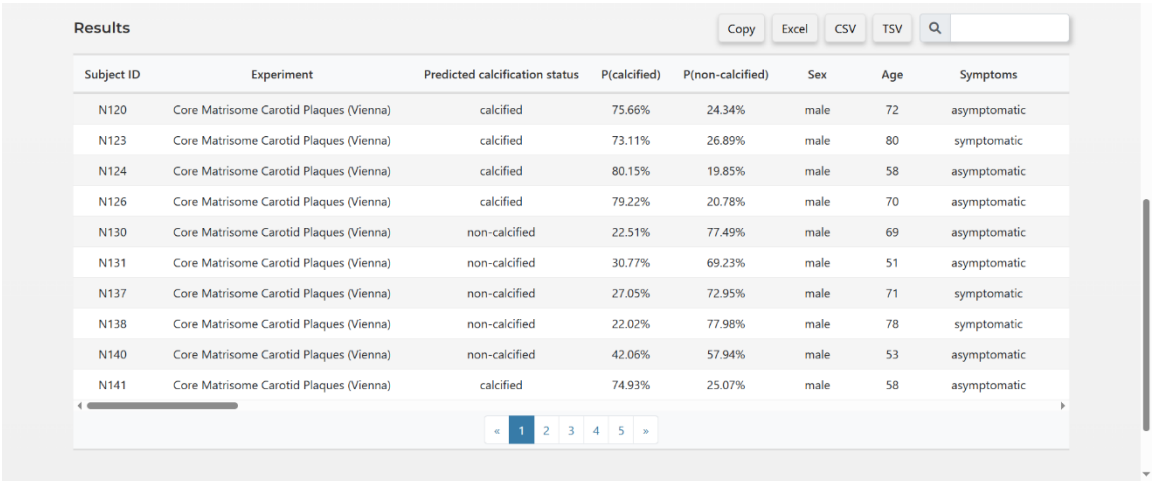
(A)



(B)



(C)



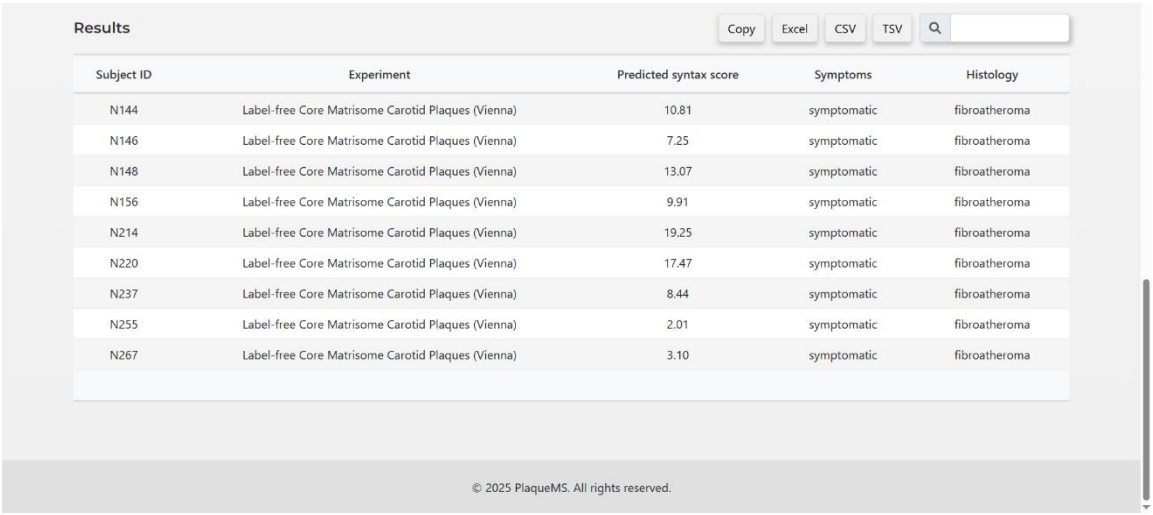
Results

Copy Excel CSV TSV

Subject ID	Experiment	Predicted calcification status	P(calcified)	P(non-calcified)	Sex	Age	Symptoms
N120	Core Matrisome Carotid Plaques (Vienna)	calcified	75.66%	24.34%	male	72	asymptomatic
N123	Core Matrisome Carotid Plaques (Vienna)	calcified	73.11%	26.89%	male	80	symptomatic
N124	Core Matrisome Carotid Plaques (Vienna)	calcified	80.15%	19.85%	male	58	asymptomatic
N126	Core Matrisome Carotid Plaques (Vienna)	calcified	79.22%	20.78%	male	70	asymptomatic
N130	Core Matrisome Carotid Plaques (Vienna)	non-calcified	22.51%	77.49%	male	69	asymptomatic
N131	Core Matrisome Carotid Plaques (Vienna)	non-calcified	30.77%	69.23%	male	51	asymptomatic
N137	Core Matrisome Carotid Plaques (Vienna)	non-calcified	27.05%	72.95%	male	71	symptomatic
N138	Core Matrisome Carotid Plaques (Vienna)	non-calcified	22.02%	77.98%	male	78	symptomatic
N140	Core Matrisome Carotid Plaques (Vienna)	non-calcified	42.06%	57.94%	male	53	asymptomatic
N141	Core Matrisome Carotid Plaques (Vienna)	calcified	74.93%	25.07%	male	58	asymptomatic

« 1 2 3 4 5 »

(D)



Results

Copy Excel CSV TSV

Subject ID	Experiment	Predicted syntax score	Symptoms	Histology
N144	Label-free Core Matrisome Carotid Plaques (Vienna)	10.81	symptomatic	fibroatheroma
N146	Label-free Core Matrisome Carotid Plaques (Vienna)	7.25	symptomatic	fibroatheroma
N148	Label-free Core Matrisome Carotid Plaques (Vienna)	13.07	symptomatic	fibroatheroma
N156	Label-free Core Matrisome Carotid Plaques (Vienna)	9.91	symptomatic	fibroatheroma
N214	Label-free Core Matrisome Carotid Plaques (Vienna)	19.25	symptomatic	fibroatheroma
N220	Label-free Core Matrisome Carotid Plaques (Vienna)	17.47	symptomatic	fibroatheroma
N237	Label-free Core Matrisome Carotid Plaques (Vienna)	8.44	symptomatic	fibroatheroma
N255	Label-free Core Matrisome Carotid Plaques (Vienna)	2.01	symptomatic	fibroatheroma
N267	Label-free Core Matrisome Carotid Plaques (Vienna)	3.10	symptomatic	fibroatheroma

Figure 10.
Overview of the Calcification Status and SYNTAX Score Prediction modules in PlaqueMS. **(A)** Calcification Prediction interface, where users can upload protein abundance data, select the extraction protocol-specific model, and optionally apply $\log_2$ transformation. **(B)** SYNTAX Score Prediction interface, which similarly allows users to upload required protein data for SYNTAX score estimation, with the model and required features fixed for this endpoint. **(C)** Example results table from the Calcification Prediction module, showing predicted calcification status, associated probabilities, and clinical metadata for each patient from the Vienna cohort; the core matrisome model has been selected and no filters have been applied in this search. **(D)** Example results table from the SYNTAX Score Prediction module, presenting predicted SYNTAX scores together with patient phenotype and histology information; filters have been applied to display only symptomatic patients with fibroatheroma.

3.2. Model Performance and Validation

The three models exhibited strong discriminatory power for plaque calcification status across all tested proteome fractions, as assessed by 5-fold cross-validation. The core matrisome model, based on a linear SVM, achieved a mean AUC of 80.55% (± 0.99%), with a mean accuracy of 79.66% (± 0.62%) and the highest F1-score of 84.33% (± 0.83%). The Random Forest classifiers trained on the cellular proteome and soluble matrisome fractions yielded mean AUCs of 82.26% (± 1.14%) and 77.26% (± 1.61%), respectively, with corresponding mean accuracies of 80.20% (± 1.22%) and 74.44% (± 1.55%). All models demonstrated high recall and precision, as well as favourable F1-scores (see Table 2 for a comprehensive breakdown of performance metrics). These results demonstrate that targeted proteomic signatures derived from distinct extraction protocols, when paired with optimised classifiers, can robustly stratify atherosclerotic plaques by calcification phenotypes.

Model	Accuracy (mean ± SD)	F1 Score (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)	ROC AUC (mean ± SD)
Core Matrisome	79.66 % ± 0.62 %	84.33 % ± 0.83 %	79.19 % ± 0.74 %	90.24 % ± 2.79%	80.55 % ± 0.99%
Soluble Matrisome	74.44 % ± 1.55 %	80.68 % ± 0.94 %	74.29 % ± 1.65 %	88.31 % ± 1.29 %	77.26 % ± 1.61 %
Cellular Proteome	80.20 % ± 1.22 %	84.27 % ± 0.85 %	79.29 % ± 1.31 %	89.94 % ± 0.68 %	82.26 % ± 1.14 %

Table 2.
Performance metrics (accuracy, F1 score, precision, recall, and ROC AUC) for the core matrisome, soluble matrisome, and cellular proteome models in predicting plaque calcification status. For each model, performance was assessed using 5-fold cross-validation, and values represent the mean ± standard deviation across 5 independent model replicates.

The observed differences in classification performance between the three proteome extracts are also reflected in the correlation of model predictions with CT-based measures of plaque calcification. As shown in the summary barplots (Figure 11), the core matrisome model demonstrated the highest and most consistent Spearman's correlation coefficients with Agatston score, calcium mass, and plaque volume, outperforming both the cellular proteome and soluble matrisome models. Specifically, the core matrisome model achieved moderate-to-strong and statistically significant correlations with these CT-derived metrics (Spearman's $\rho \approx 0.55$, $p < 0.01$), such that the predicted probability of calcification increases steadily with rising Agatston score, calcium mass, and plaque volume (Figures 12A–C). This positive monotonic relationship demonstrates that the model is not only discriminative but also well-calibrated across the spectrum of calcification burden. While all models showed significant correlations with all CT-based metrics ($p < 0.05$), predictions from the cellular and soluble matrisome models tended to separate patients into two distinct groups—high and low probability—showing little gradation across intermediate CT scores (Supplementary Figures 1A-F).

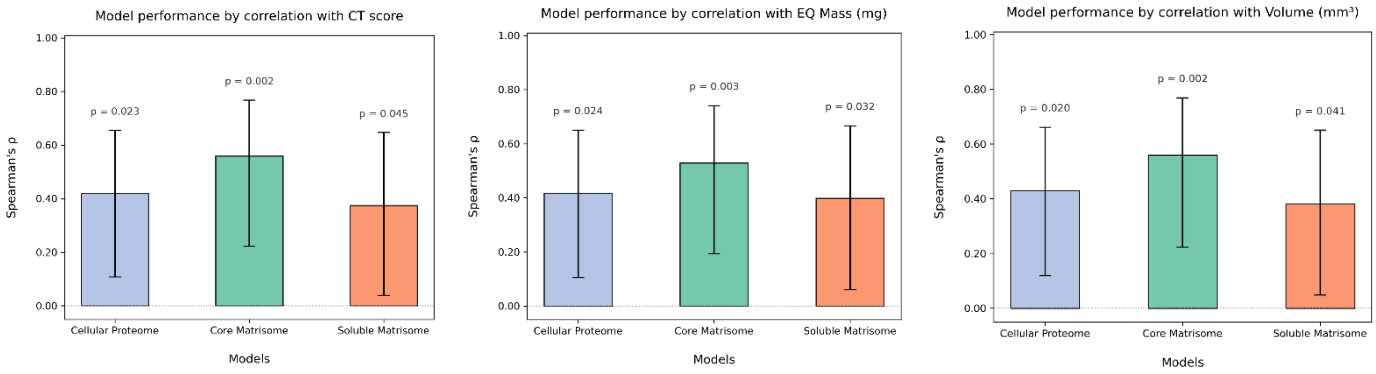


Figure 11.
Barplots display Spearman's correlation coefficients (ρ) between the predicted probability of being calcified and CT-derived measures of calcification burden for each model. While all models showed significant correlations with CT-based endpoints ($p < 0.05$), the core matrisome model consistently demonstrated the strongest and most significant associations (ρ ≈ 0.55, $p < 0.01$). Error bars indicate 95% confidence intervals, calculated by bootstrapping.

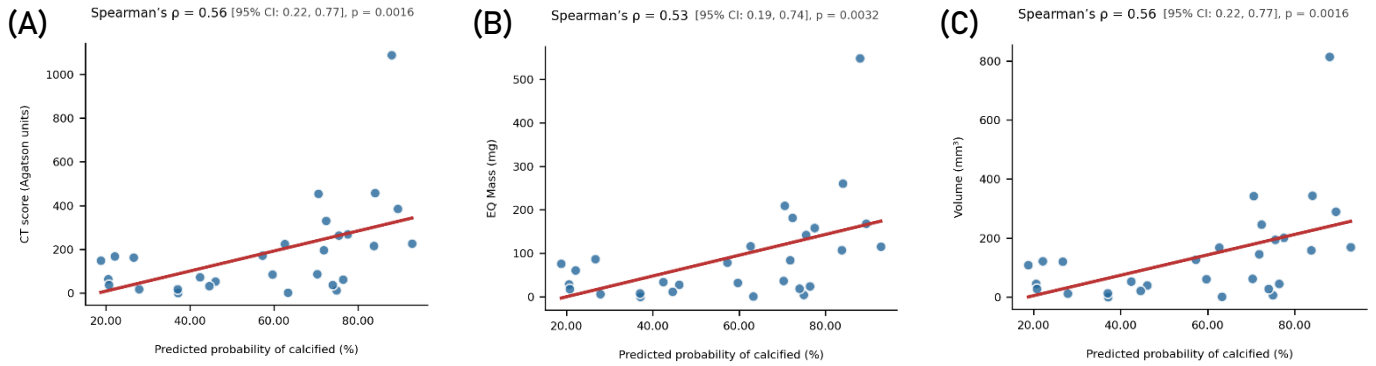


Figure 12.
Scatter plots show the relationship between the predicted probability of calcification from the core matrisome model and CT-derived metrics: **(A)** CT Agatston score, **(B)** calcium mass (mg), and **(C)** plaque volume (mm^3). For each metric, a significant ($p < 0.01$) monotonic increase is observed, with Spearman's correlation coefficients of **(A)** ρ = 0.56, 95% CI: 0.22-0.77, **(B)** ρ = 0.53, 95% CI: 0.19-0.74, and **(C)** ρ = 0.56, 95% CI: 0.22-0.77.

In addition to continuous CT-based metrics, model performance was also evaluated against CTA-derived calcification status categories (non-calcified, mixed, calcified). As shown in the summary and violin plots (Figure 13), only the core matrisome model demonstrated a moderate and statistically significant correlation with CTA status (Spearman's ρ = 0.48, 95% CI: 0.13–0.73, p = 0.0085), while the other models showed non-significant associations ($p > 0.05$). Non-parametric group comparison using the Kruskal–Wallis test (H = 9.4560, p = 0.008844) followed by Dunn's post-hoc test with Bonferroni correction revealed that only the core matrisome model significantly distinguished between mixed and calcified plaques (p = 0.0063), while no model could reliably differentiate all three groups.

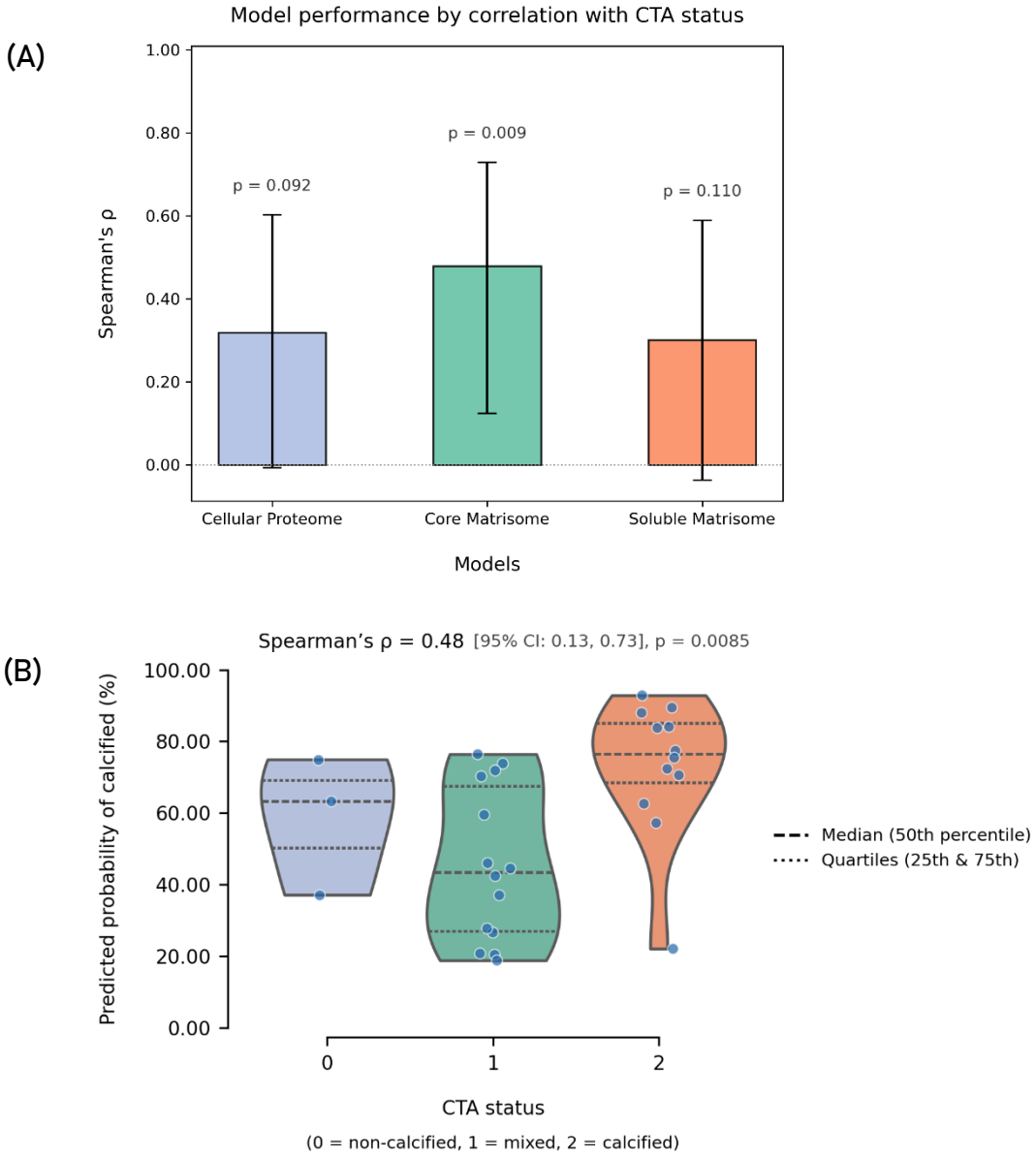


Figure 13.

(A) Barplot summarising Spearman's correlation coefficients (ρ) between the predicted probability of being calcified and CTA status categories (non-calcified, mixed, calcified) for each model. Only the core matrisome model showed a statistically significant correlation ($p = 0.0085$). Error bars indicate 95% confidence intervals, calculated by bootstrapping. **(B)** Violin plot depicting the distribution of predicted probabilities from the core matrisome model across CTA status categories. A moderate, statistically significant positive correlation is observed (Spearman's ρ = 0.48, 95% CI: 0.13–0.73, $p = 0.0085$), with the highest probabilities seen in the calcified group.

For cross-cohort validation using the Athero-Express dataset, interestingly, only the cellular proteome model achieved statistically significant discrimination between asymptomatic and symptomatic patients (Figure 14). The predicted probability of calcification was modestly but significantly higher in asymptomatic compared to symptomatic plaques, as confirmed by Mann–Whitney U test (U = 3612, p = 0.0412) and reflected by an AUC of 0.63 (95% CI: 0.52–0.74). Calcification data were not available for this cohort, so symptom status was used as a clinical proxy; nevertheless, these results demonstrate the model's ability to generalise and stratify patients by clinical phenotype across independent datasets.

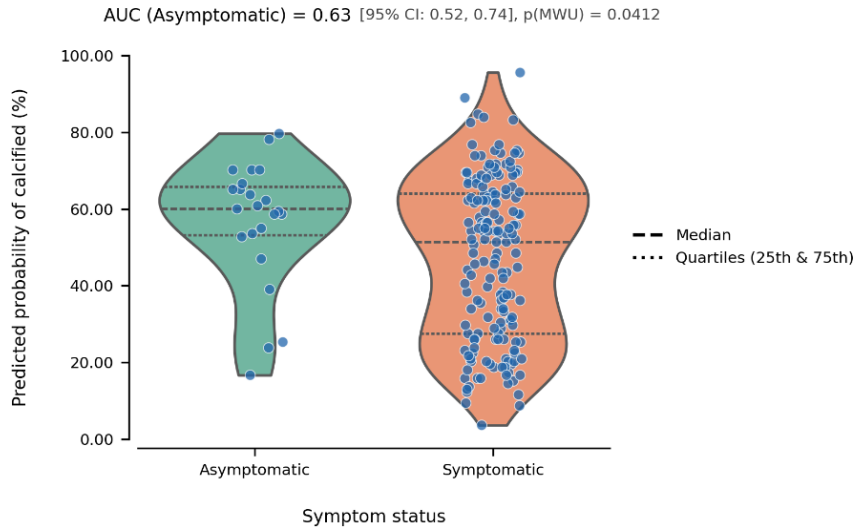


Figure 14.
Violin plot depicting the distribution of predicted probabilities of calcification from the cellular proteome model in the Athero-Express cohort, stratified by symptom status (asymptomatic vs. symptomatic). The model demonstrated modest but significant (AUC = 0.63, Mann-Whitney U test, p = 0.0412) separation between asymptomatic and symptomatic groups, with asymptomatic patients exhibiting a higher median predicted probability of calcification (76.00%) compared to symptomatic patients (58.00%).

Predicted SYNTAX scores showed a strong and statistically significant correlation with observed values (Spearman's ρ = 0.77, $p < 0.001$), demonstrating high concordance between proteomics-based estimates and angiographic assessment. Furthermore, in cross-cohort validation, the model achieved robust discrimination of CVD risk (AUC = 0.70), plaque calcification (AUC = 0.67), and symptom status (AUC = 0.65), underscoring its utility for clinically relevant risk stratification (Figure 15).

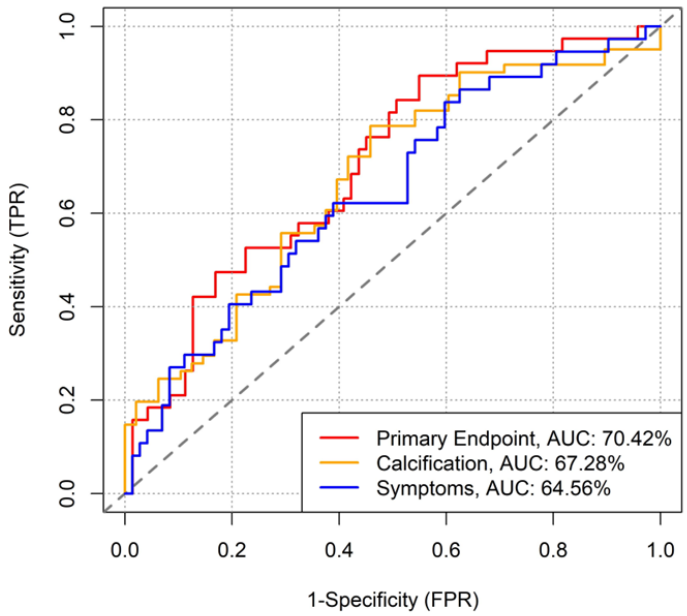


Figure 15.
Receiver operating characteristic (ROC) curves illustrating the discriminatory performance of the SYNTAX score prediction model for cardiovascular disease risk (red), plaque calcification (orange), and symptom status (blue). The diagonal dashed line indicates the reference line for random classification (AUC = 50.00%).

3.3. Feature Selection Results

Across all three proteome extracts, the application of the multi-objective evolutionary algorithm for feature selection revealed distinct patterns in how individual proteins were prioritised for the calcification prediction models (Figure 16A–C). In the core matrisome and cellular proteome datasets, fetuin-A (AHSG) and osteopontin (OSTP) were among the proteins most consistently included in the final Pareto fronts across five independent runs, with other proteins such as COL2A1, PROZ, PRDX2, and VAPB also exhibiting high selection counts within their respective fractions. For the soluble matrisome model, proteins including KNG1 and NOV were most frequently represented among the selected features. These results highlight the specific proteins repeatedly identified as important for each model.

Feature selection for the SYNTAX score prediction model identified a distinct protein signature, comprising HRG, CP, C4B, F13A1, and VCAN, reflecting the unique proteomic determinants underlying the prediction of angiographic complexity.

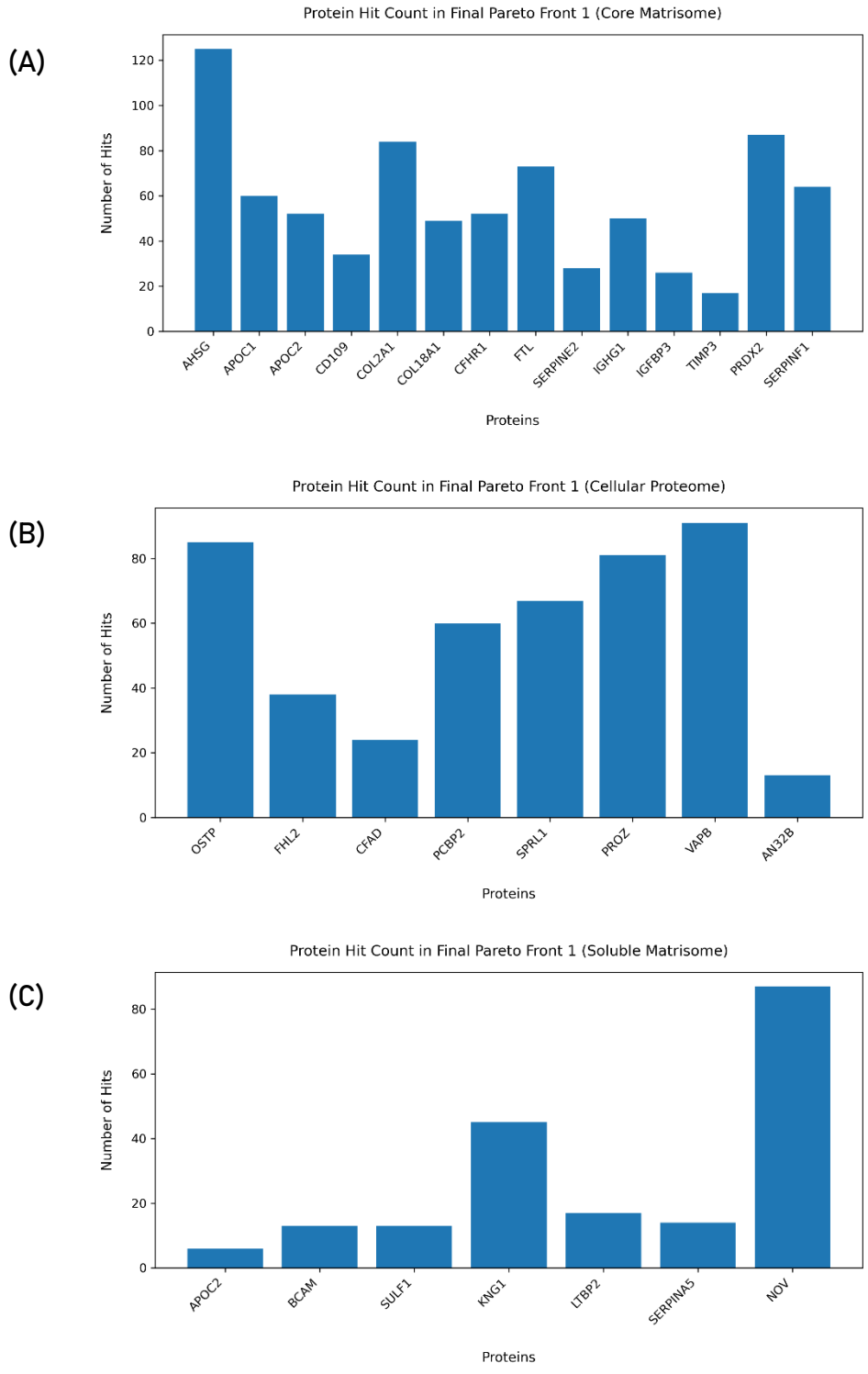


Figure 16.
Protein hit counts in the final Pareto front 1 for each model: **(A)** core matrisome, **(B)** cellular proteome, and **(C)** soluble matrisome. Bars indicate the number of times each protein was selected across five independent runs. The final solution for the soluble matrisome model consisted of only seven proteins, constituting the smallest feature set among the three models.

4. Discussion

4.1. Interpretation of Findings

Vascular calcification, once regarded as a passive byproduct of advanced atherosclerosis and traditionally considered as closely associated with inflammation, is now recognised as a highly regulated, cell-mediated process with significant implications for plaque stability and cardiovascular risk [73]. A recent large-scale proteomics study has demonstrated that calcification signatures within human atherosclerotic plaques are inversely correlated with inflammatory protein profiles, supporting the view that calcified plaques generally reflect a more quiescent, less rupture-prone phenotype [40]. However, the clinical significance of calcification could extend beyond this dichotomy. The present study builds upon these insights by demonstrating, through cross-cohort validation, that plaques with higher predicted calcification probability were more frequently observed in asymptomatic patients (Figure 14). This finding aligns with the notion that extensive calcification may serve as a stabilising, and potentially protective, mechanism in advanced atherosclerosis [73].

Recent evidence also suggests that proteomic signatures can surpass traditional imaging and histological techniques in distinguishing plaque phenotypes and predicting cardiovascular risk [40]. Leveraging this potential, machine learning algorithms were applied to high-dimensional proteomic data, enabling probabilistic prediction of calcification status at the individual plaque level. Rather than relying solely on binary classification, the use of probability scores captures the continuum of calcification and allows for a more nuanced assessment of disease state. Notably, the core matrisome model exhibited strong calibration across the spectrum of calcification burden, with predicted probabilities closely mirroring the gradation observed in clinical imaging. This was evidenced by significant positive correlations between model-derived probabilities and CTA-defined plaque phenotypes, where higher probabilities were associated with plaques classified as mixed or calcified. Additionally, while all models showed positive associations with quantitative CT metrics, the core matrisome model demonstrated the strongest correlations (Figure 11), further validating its capacity to reflect the underlying calcification burden. These results highlight that probability-based predictions not only capture the heterogeneity of atherosclerotic disease but also provide a meaningful bridge between molecular profiles and imaging-defined pathology.

Beyond multi-objective optimisation, the evolutionary algorithm proved to be a powerful tool for elucidating the molecular determinants of plaque calcification. Analysis of the most frequently selected features revealed a consistent emphasis on established calcification markers, such as osteopontin (OSTP) and fetuin-A (AHSG), which both play key inhibitory roles in vascular mineralisation [74]. Osteopontin appears to be persistently over-expressed in calcified human plaques [40] and surges during acute coronary syndromes as a compensatory response to ongoing mineral deposition, but under certain conditions it can also promote calcification [74]. Likewise, fetuin-A acts systemically to neutralise circulating calcium-phosphate crystals and is often reduced in patients with heavy

calcification, reflecting its consumption [75]. The predictive utility of both markers closely aligns with findings from previous studies, where both proteins have emerged as key predictors of vascular calcification and adverse outcomes across independent cohorts [75],[76].

Alongside these canonical calcification regulators, the algorithm identified further proteins with emerging relevance to vascular calcification, including protein Z (PROZ) and type II collagen (COL2A1). Both were found to be upregulated in calcified plaques [40], supporting their potential involvement in mineral deposition. Protein Z, a vitamin K-dependent coagulation factor, may bridge pathways between thrombosis and vascular calcification [77], while type II collagen indicates cartilage-like matrix remodelling associated with advanced plaque mineralisation [78].

Interestingly, VAPB—a vesicle-associated membrane protein with no established role in atherosclerosis or vascular calcification [79]—was selected even more frequently than osteopontin in the cellular proteome model. Although recent evidence links VAPB to osteogenic differentiation and calcification in human aortic valve tissue [80], its involvement in arterial plaque biology remains unexplored. Such findings illustrate how evolutionary algorithms can point towards novel biomarkers and biological pathways that extend beyond those currently established in literature.

The predictive performance of each model closely mirrored the biological relevance of its underlying proteome fraction. Models based on the core matrisome and cellular proteome consistently outperformed those built from the soluble matrisome, reflecting a greater enrichment of established calcification markers within the more structural and cell-associated extracts. In contrast, the soluble matrisome model, which yielded weaker predictive capacity, included fewer canonical markers, apart from NOV (CCN3), a matricellular protein recently shown to inhibit vascular calcification via modulation of the Notch signalling pathway [81].

4.2. Clinical and Research Utility

Atherosclerosis research is frequently hampered by the fragmentation of omics, imaging and clinical read-outs into static, discipline-specific silos. PlaqueMS alleviates this bottleneck by transforming 419 carotid and 150 coronary plaque proteomes and their heterogeneous annotations into a coherent, query-driven workspace that minimises manual data wrangling and broadens analytical scope.

Unlike existing cardiovascular research portals such as PlaqView 2.0 [21] or HeartBioPortal [82], which concentrate on transcriptomic and genomic information, PlaqueMS directly integrates mass-spectrometry–based proteomics with patient phenotypes. Protein abundances can be examined alongside histology, ultrasound morphology and routine clinical variables, letting investigators trace molecular signals through to observable disease features.

A central novelty is the availability of embedded prediction tools for plaque calcification status and SYNTAX score—an established angiographic metric that quantifies the

anatomical complexity of coronary artery disease and informs interventional decision-making. These models enable researchers to infer clinically pertinent endpoints from proteomic profiles alone, annotating legacy cohorts that lack imaging or angiographic data, facilitating rigorous cross-cohort comparisons, and informing the design of prospective studies.

The platform further supports integrative exploration through phenotype-informed queries and pre-computed visualisations that compare protein expression across user-defined sub-groups. Interactive PPI networks—filterable by cohort, tissue compartment or proteomics method—reveal previously overlooked molecular relationships and candidate biomarkers, stimulating new hypotheses for experimental validation.

By uniting exploratory analytics, network context and real-time risk prediction within a single web interface, PlaqueMS converts static plaque datasets into a living resource that accelerates discovery and supports translational research. Its modular design invites future expansion to additional omics layers and larger multicentre cohorts, ensuring that insights gained today can be rapidly revisited as new data emerge.

4.3. Limitations

While PlaqueMS introduces several innovations and contributes to advancing atherosclerosis research, its current implementation is subject to important limitations that define the boundaries of its present utility and guide directions for future development.

A major barrier to seamless interoperability is presented by metadata inconsistencies across cohorts. The Vienna, Athero-Express, and Virginia datasets differ in the depth and structure of their clinical and experimental annotations, limiting the generalizability of several platform modules. At present, only the Protein and the Differential Expression Analysis modules support data integration across all cohorts, whereas more advanced tools that rely heavily on detailed phenotype-related queries are currently restricted to the Vienna dataset due to missing or incompatible variables in the other cohorts. By comparison, the Protein Networks module is less constrained, as expansion to other cohorts primarily requires generating appropriate network files rather than comprehensive metadata harmonisation.

Moreover, predictive model validation remains an ongoing challenge. Although the embedded machine learning algorithms for plaque calcification status demonstrate strong performance within the discovery cohort, and the predicted probabilities show promising concordance with available CT metrics, incomplete imaging data across all patients restricts the extent of these correlations. Cross-cohort validation efforts, such as those conducted in the Athero-Express dataset, support some generalizability; however, the absence of direct calcification assessments or comparable clinical endpoints limits rigorous external validation. Future studies incorporating comprehensive and clinically relevant calcification data—ideally obtained via standardised imaging or histological approaches—

will be essential to fully establish the accuracy, translational utility, and eventual clinical implementation of PlaqueMS predictions.

Another factor limiting immediate clinical implementation is the inherent invasiveness of proteomic profiling of plaque tissue, which is rarely performed outside research settings. Consequently, PlaqueMS should be primarily regarded as a research-oriented and exploratory platform intended to facilitate hypothesis generation, biomarker discovery, and integrative proteomics analysis within atherosclerosis research. Nonetheless, the use of surrogate proteins in accessible biofluids may, in the future, allow estimation of the required protein abundances for model-based predictions, although such approaches remain to be validated.

In the same context, a further methodological constraint relates to the SYNTAX score prediction model, which was trained exclusively on normalised label-free proteomics data. Due to intrinsic scale differences between label-free and TMT-based proteomic quantification, even after standardisation of the data, the regression formula does not yield reliable results, limiting the model's generalizability across proteomics platforms.

Lastly, the current analytical framework is restricted solely to proteomics data, lacking integration with transcriptomics, metabolomics, or other omics layers. The absence of these complementary datasets reduces the platform's potential for holistic, multi-omics analysis, thereby limiting comprehensive biological interpretation.

4.4. Future Directions

Several targeted developments are planned to further enhance the analytical capabilities and applicability of PlaqueMS, while simultaneously addressing, to the greatest possible extent, the limitations identified in the preceding section. A primary objective is the establishment of standardised metadata frameworks across all participating cohorts, facilitating seamless data harmonisation and robust interoperability. Efforts are also underway to finalise and deploy the Graph Protein Networks module, the final component of PlaQuery, which leverages Neo4j's graph database architecture to efficiently represent and query complex protein-protein relationships as interconnected nodes and edges. This network-based approach will underpin the construction of advanced analytical tools, including the implementation of the PageRank algorithm to systematically identify hub proteins based on their connectivity within the network.

In parallel, future platform evolution will entail the incorporation of additional omics modalities, including transcriptomics, spatial, and single-cell data, thereby broadening the platform's scope for integrative multi-layer analyses. Improvements in interpretability are also planned through the integration of SHAP (Shapley Additive Explanations) analysis, which will provide greater transparency regarding the contribution of individual molecular markers within predictive modelling workflows. Concurrently, enhancements in user experience and platform security are anticipated, encompassing further interface refinements and advanced authentication mechanisms such as two-factor verification.

Nevertheless, certain limitations warrant ongoing attention, notably the restricted availability of comprehensive imaging data necessary for external model validation and reliance upon surrogate protein markers when direct measurements are lacking. Addressing these constraints remains imperative for improving the translational potential of PlaqueMS for clinical applications.

Collectively, these strategic enhancements will contribute significantly towards positioning PlaqueMS as an integrated, secure, and versatile analytical resource for atherosclerosis research, thereby bridging existing gaps between fundamental discovery and clinical implementation.

5. Conclusion

This work establishes PlaqueMS as an innovative and integrative platform that transforms previously static proteomic and clinical datasets into an interactive resource for atherosclerosis research. By combining detailed molecular data with comprehensive clinical and phenotypic metadata, PlaqueMS enables researchers to interrogate disease mechanisms across diverse patient cohorts in ways not previously feasible. Its unique strengths lie in facilitating phenotype-oriented queries, supporting multicohort analysis, and incorporating predictive modelling tools, thereby extending the utility of proteomic data even in the absence of complete clinical information. While challenges related to metadata standardisation, cross-cohort validation, and integration across proteomics platforms persist, PlaqueMS represents a significant advancement towards more accessible, robust, and translational research tools in cardiovascular science.

6. References

[1] S. Tan, B. Zheng, M. Tang, H. Chu, Y.-T. Zhao, and C. Weng, “Global Burden of Cardiovascular Diseases and its Risk Factors, 1990-2021: A Systematic Analysis for the Global Burden of Disease Study 2021”, *QJM: An International Journal of Medicine*, Jan. 2025.

[2] S. Jebari-Benslaiman *et al.*, “Pathophysiology of Atherosclerosis”, *International Journal of Molecular Sciences*, vol. 23, no. 6, p. 3346, Mar. 2022.

[3] K. J. Williams and I. Tabas, "The response-to-retention hypothesis of early atherogenesis", *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 15, no. 5, pp. 551–561, 1995.

[4] L. Jonasson, K. Kalogeropoulos, M. A. Karsdal, A. L. Reese-Petersen, U. Auf dem Keller, F. Genovese, J. Nilsson, and I. Goncalves, "Exploring the role of extracellular matrix proteins to develop biomarkers of plaque vulnerability and outcome", *Journal of Internal Medicine*, vol. 287, no. 5, pp. 493–513, 2020.

[5] A. Naba, K. R. Clauser, H. Ding, C. A. Whittaker, S. A. Carr, and R. O. Hynes, "The extracellular matrix: Tools and insights for the 'omics' era", *Matrix Biology*, vol. 49, pp. 10–24, 2016.

[6] M. Chandran, S. S, A. Chandran, A. Jaleel, and J. P. Ayyappan, “Defining atherosclerotic plaque biology by mass spectrometry-based omics approaches.”, *Molecular omics*, vol. 19, no. 1, pp. 6–26, Nov. 2022.

[7] H. Liu *et al.*, “Research Progress and Clinical Translation Potential of Coronary Atherosclerosis Diagnostic Markers from a Genomic Perspective”, *Genes*, vol. 16, no. 1, p. 98, Jan. 2025.

[8] A. Ajoolabady, D. Pratico, L. Lin, *et al.*, “Inflammation in atherosclerosis: pathophysiology and mechanisms”, *Cell Death and Disease*, vol. 15, no. 1, p. 817, 2024.

[9] F. Zhang, X. Guo, Y. Xia, *et al.*, “An update on the phenotypic switching of vascular smooth muscle cells in the pathogenesis of atherosclerosis”, *Cell. Mol. Life Sci.*, vol. 79, p. 6, 2022.

[10] Q. Nguyen *et al.*, “Spatial Transcriptomics in Human Cardiac Tissue”, *International Journal of Molecular Sciences*, vol. 26, no. 3, p. 995, Jan. 2025.

[11] P. Laaksonen, M. Ekroos, M. Sysi-Aho, *et al.*, “Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond LDL-cholesterol”, *European Heart Journal*, vol. 37, no. 25, pp. 1967–1976, 2016.

[12] P. Kiessling and C. Kuppe, “Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases”, *Genome Medicine*, vol. 16, no. 1, p. 14, Jan. 2024.

[13] G. Nieddu, M. Formato, and A. J. Lepedda, “Searching for atherosclerosis biomarkers by proteomics: A focus on lesion pathogenesis and vulnerability,” *International Journal of Molecular Sciences*, vol. 24, no. 20, p. 15175, 2023.

[14] M. Lorentzen, A. R. Holm, A. Mollnes, *et al.*, “Proteomic analysis of human carotid atherosclerotic plaques: identification of potential markers of plaque destabilization”, *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 42, no. 3, pp. 272–284, 2022.

[15] G. Kalló, K. Zaman, L. Potor, Z. Hendrik, G. Méhes, C. Tóth, P. Gergely, J. Tőzsér, G. Balla, J. Balla, L. Prokai, and É. Csősz, "Identification of protein networks and biological pathways driving the progression of atherosclerosis in human carotid arteries through mass spectrometry-based proteomics", *International Journal of Molecular Sciences*, vol. 25, no. 24, p. 13665, 2024.

[16] J. I. van der Vaart, R. van Eenige, P. C. Rensen, and S. Kooijman, “Atherosclerosis: an overview of mouse models and a detailed methodology to quantify lesions in the aortic root”, *Vascular Biology*, vol. 6, no. 1, p. e230017, 2024.

[17] M. C. Martinez-Campanario, M. Cortés, A. Moreno-Lanceta, *et al.*, “Atherosclerotic plaque development in mice is enhanced by myeloid ZEB1 downregulation”, *Nature Communications*, vol. 14, art. no. 8316, 2023.

[18] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks”, *Genome Research*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[19] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. von Mering, “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”, *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[20] S. de la Cámara-Fuentes, D. Gutiérrez-Blázquez, M. L. Hernáez, and C. Gil, “TraianProt: a user-friendly R Shiny application for wide-format proteomics data downstream analysis”, *Bioinformatics*, preprint, Dec. 2024.

[21] W. F. Ma, A. W. Turner, C. Gancayco, D. Wong, Y. Song, J. V. Mosquera, G. Auguste, C. J. Hodonsky, A. Prabhakar, H. A. Ekiz, S. W. van der Laan, and C. L. Miller, “PlaqView 2.0: A comprehensive web portal for cardiovascular single-cell genomics”, *Frontiers in Cardiovascular Medicine*, vol. 9, p. 969421, Aug. 2022.

[22] B. B. Khomtchouk, D. T. Tran, K. A. Vand, M. Might, O. Gozani, and T. L. Assimes, “Cardioinformatics: the nexus of bioinformatics and precision cardiology”, *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2031–2051, 2020.

[23] S. K. Banik, S. Baishya, A. Das Talukdar, and M. D. Choudhury, “Network analysis of atherosclerotic genes elucidates druggable targets,” BMC Medical Genomics, vol. 15, no. 1, p. 42, 2022.

[24] Z. Gong, H. Yang, L. Gao, Y. Liu, Q. Chu, C. Luo, L. Kang, H. Zhai, Q. Xu, W. Wu, and N. Li, “Mechanisms of wogonoside in the treatment of atherosclerosis based on network pharmacology, molecular docking, and experimental validation,” *BMC Complementary Medicine and Therapies*, vol. 25, no. 1, p. 28, 2025.

[25] Q. M. Zhu, Y. H. H. Hsu, F. H. Lassen *et al.*, "Protein interaction networks in the vasculature prioritize genes and pathways underlying coronary artery disease", *Communications Biology*, vol. 7, p. 87, 2024.

[26] Z. Chen, M. Yang, Y. Wen, S. Jiang, W. Lu, and H. Huang, “Prediction of atherosclerosis using machine learning based on operations research”, *Mathematical Biosciences and Engineering*, vol. 19, no. 5, pp. 4892–4910, Mar. 2022.

[27] A. Lin, N. Manral, P. McElhinney *et al.*, "Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study", *The Lancet Digital Health*, vol. 4, no. 4, pp. e256–e265, 2022.

[28] Ei. I. Usova et al., “Integrative Analysis of Multi-Omics and Genetic Approaches-A New Level in Atherosclerotic Cardiovascular Risk Prediction”, vol. 11, no. 11, p. 1597, Oct. 2021.

[29] G. Quer, R. Arnaout, M. Henne, and R. Arnaout, "Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review", *Journal of the American College of Cardiology*, vol. 77, no. 3, pp. 300–313, 2021.

[30] S. Allan, R. Olaiya, and R. Burhan, "Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease", *Postgraduate Medical Journal*, vol. 98, no. 1161, pp. 551–558, Jul. 2022.

[31] A. Morabito, G. De Simone, R. Pastorelli *et al.*, "Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review", *J. Transl. Med.*, vol. 23, p. 425, 2025.

[32] J. Petch, S. Di, and W. Nelson, "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology", *Can. J. Cardiol.*, vol. 38, no. 2, pp. 204–213, 2022.

[33] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction", *Front. Bioinform.*, vol. 2, p. 927312, 2022.

[34] B. Ibanez *et al.*, "Progression of Early Subclinical Atherosclerosis (PESA) Study: JACC Focus Seminar 7/8", *Journal of the American College of Cardiology*, vol. 78, no. 2, pp. 156–179, Jul. 2021.

[35] S. A. Ramsey, E. S. Gold, and A. Aderem, "A systems biology approach to understanding atherosclerosis.", *Embo Molecular Medicine*, vol. 2, no. 3, pp. 79–89, Mar. 2010.

[36] M. Mokry, A. Boltjes, L. Slenders et al., "Transcriptomic-based clustering of human atherosclerotic plaques identifies subgroups with different underlying biology and clinical presentation", *Nature Cardiovascular Research*, vol. 1, pp. 1140–1155, 2022.

[37] M. Sopić *et al.*, "Transcriptomic research in atherosclerosis: Unravelling plaque phenotype and overcoming methodological challenges", *Journal of molecular and cellular cardiology plus*, Dec. 2023.

[38] X. Wu and H. Zhang, "Omics Approaches Unveiling the Biology of Human Atherosclerotic Plaques", *American Journal of Pathology*, Jan. 2024.

[39] M. Lin, J. Guo, Z. Gu, et al., "Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice", *Journal of Translational Medicine*, vol. 23, p. 388, 2025.

[40] K. Theofilatos, S. Stojkovic, M. Hasman, S. W. van der Laan, F. Baig, J. Barallobre-Barreiro, L. E. Schmidt, S. Yin, X. Yin, S. Burnap, B. Singh, J. Popham, O. Harkot, S. Kampf, M. C. Nackenhorst, A. Strassl, C. Loewe, S. Demyanets, C. Neumayer, M. Bilban *et al.*, "Proteomic atlas of atherosclerosis: The contribution of proteoglycans to sex differences, plaque phenotypes, and outcomes", *Circulation Research*, vol. 133, no. 7, pp. 542–558, Sep. 2023.

[41] M. Sopić *et al.*, "Multiomics tools for improved atherosclerotic cardiovascular disease management.", *Trends in Molecular Medicine*, Oct. 2023.

[42] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, and J. A. Vizcaíno, "The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences", *Nucleic Acids Research*, vol. 50, no. D1, pp. D543–D552, Jan. 2022.

[43] S. Liu, *Interactive Web Tool for Multi-Omics Atherosclerotic Plaque Data Analysis*, M.Sc. final report, Dept. of Informatics, King's College London, London, U.K., 2023.

[44] Z. Tsogtbaatar, *Graph-Based Exploration of Proteomics Data in Atherosclerotic Plaques*, M.Sc. final report, Dept. of Informatics, King's College London, London, U.K., 2023.

[45] W. E. Hellings, F. L. Moll, D. P. de Kleijn, and G. Pasterkamp, "10-years experience with the Athero-Express study", *Cardiovascular Diagnosis and Therapy*, vol. 2, no. 1, p. 63, 2012.

[46] C. J. Hodonsky, A. W. Turner, M. D. Khan, N. B. Barrientos, R. Methorst, L. Ma, N. G. Lopez, J. V. Mosquera, G. Auguste, E. Farber, W. F. Ma, et al., "Multi-ancestry genetic analysis of gene regulation in coronary arteries prioritizes disease risk loci", *Cell Genomics*, vol. 4, no. 1, 2024.

[47] K. C. Palm, X. Yin, F. Baig, K. Theofilatos, S. W. van der Laan, G. J. de Borst, D. P. de Kleijn, J. Wojta, S. Stojkovic, M. Mayr, and H. M. den Ruijter, "Proteomic profiling reveals a higher presence of glycolytic enzymes in human atherosclerotic lesions with unfavourable histological characteristics", *Cardiovascular Research*, p. cvaf077, 2025.

[48] MySQL 8.0 Reference Manual, Oracle Corporation, 2024. [Online].
Available: https://downloads.mysql.com/docs/refman-8.0-en.pdf

[49] M. Hasman, “Resolving Inflammatory Networks in Atherosclerosis Using Proteomics and Bioinformatics,” Ph.D. dissertation, King’s College London, 2023.

[50] C. I. Johnpaul and T. Mathew, “A Cypher query based NoSQL data mining on protein datasets using Neo4j graph database”, in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Jan. 2017, pp. 1–6.

[51] Django Software Foundation, “Django (Version 5.1.7) [Software]”, 2024.
Available: https://www.djangoproject.com

[52] Neo4j, Inc., “Neo4j Python Driver (Version 5.28.1) [Software]”, 2024.
Available: https://neo4j.com/docs/api/python-driver/current

[53] Tom Christie et al., "Django REST Framework (Version 3.15.2) [Software]", 2024. [Online]. Available: https://www.django-rest-framework.org. [Accessed: July 2025].

[54] Cytoscape.js, “Cytoscape.js (Version 3.32.1) [Software]”, 2024.
Available: https://js.cytoscape.org

[55] The Cytoscape Consortium, “Cytoscape Desktop (Version 3.10.3) [Software]”, 2024.
Available: https://cytoscape.org

[56] B. Demchak, “py4cytoscape (Version 1.11.0) [Software],” 2024.
Available: https://py4cytoscape.readthedocs.io/en/1.11.0

[57] "Bootstrap," [Online]. Available: https://getbootstrap.com. [Accessed: July 2025].

[58] "jQuery," [Online]. Available: https://jquery.com. [Accessed: July 2025].

[59] "Bootstrap Tags Input (v0.8.0)," [Online]. Available: https://bootstrap-tagsinput.github.io/bootstrap-tagsinput/examples. [Accessed: July 2025].

[60] "Typeahead.js (v0.11.1)" [Online]. Available: https://twitter.github.io/typeahead.js.
[Accessed: July 2025].

[61] "Font Awesome (v6.0.0-beta3)" [Online]. Available: https://fontawesome.com.
[Accessed: July 2025].

[62] "Bootstrap Icons (v1.10.5)" [Online]. Available: https://icons.getbootstrap.com.
[Accessed: July 2025].

[63] Google, “Google Fonts: Montserrat and Inter”, Google Fonts. [Online].
Available: https://fonts.google.com. [Accessed: July 2025].

[64] Python Software Foundation, “Python 3.11.9 Release”, Python.org. [Online].
Available: https://www.python.org/downloads/release/python-3119. [Accessed: July 2025].

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[66] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Official Journal of the European Union, 2016. [Online]. Available: https://gdpr.eu. [Accessed: July 2025].

[67] Django Software Foundation, "Security in Django", Django Documentation, Version 5.1, 2024. [Online].
Available: https://docs.djangoproject.com/en/5.1/topics/security. [Accessed: July 2025].

[68] B. Singh, *Application of Machine Learning Regression Techniques for CVD Risk Prediction*, M.Res. project report, School of Cardiovascular Medicine & Sciences, King's College London, London, U.K., 2020.

[69] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte Jr., and R. Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *J. Am. Coll. Cardiol.*, vol. 15, no. 4, pp. 827–832, 1990.

[70] P. Virtanen *et al*., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", *Nature Methods*, vol. 17, no. 3, pp. 261-272, 2020. [Software: Version 1.13.1].
Available: https://scipy.org. [Accessed: July 2025].

[71] The pandas development team, pandas (Version 2.2.2) [Software], Zenodo, 2024.
Available: https://doi.org/10.5281/zenodo.10957263. [Accessed: July 2025].

[72] S. van Dongen, "Graph clustering via a discrete uncoupling process," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, pp. 121–141, 2008.

[73] X. Shi, J. Gao, Q. Lv, H. Cai, F. Wang, R. Ye, and X. Liu, "Calcification in atherosclerotic plaque vulnerability: Friend or foe?", *Frontiers in Physiology*, vol. 11, p. 56, 2020.

[74] N. P. Kadoglou, E. Khattab, N. Velidakis, and E. Gkougkoudi, "The role of osteopontin in atherosclerosis and its clinical manifestations (atherosclerotic cardiovascular diseases)—a narrative review", *Biomedicines*, vol. 11, no. 12, p. 3178, 2023.

[75] P. Sommer, M. Schreinlechner, M. Noflatscher, D. Lener, F. Mair, M. Theurl, R. Kirchmair, and P. Marschang, "High baseline fetuin-A levels are associated with lower atherosclerotic plaque progression as measured by 3D ultrasound," *Atherosclerosis Plus*, vol. 45, pp. 10–17, 2021.

[76] F. Carbone, F. Rigamonti, F. Burger, A. Roth, M. Bertolotto, G. Spinella, B. Pane, D. Palombo, A. Pende, A. Bonaventura, L. Liberale, A. Vecchié, F. Dallegri, F. Mach, and F. Montecucco, "Serum levels of osteopontin predict major adverse cardiovascular events in patients with severe carotid artery stenosis", *Int. J. Cardiol.*, vol. 255, pp. 195–199, 2018.

[77] L. Zhang, A. Z. Segal, D. Leifer, R. L. Silverstein, L. M. Gerber, R. B. Devereux, and J. R. Kizer, "Circulating protein Z concentration, PROZ variants, and unexplained cerebral infarction in young and middle-aged adults", *Thromb. Haemost.*, vol. 117, no. 1, pp. 149–157, 2017.

[78] A. Kuzan, A. Chwiłkowska, C. Pezowicz, W. Witkiewicz, A. Gamian, K. Maksymowicz, and M. Kobielarz, "The content of collagen type II in human arteries is correlated with the stage of atherosclerosis and calcification foci", *Cardiovasc. Pathol.*, vol. 28, pp. 21–27, 2017.

[79] Y. Ding, N. Liu, D. Zhang, L. Guo, Q. Shang, Y. Liu, G. Ren, and X. Ma, "Mitochondria-associated endoplasmic reticulum membranes as a therapeutic target for cardiovascular diseases", *Frontiers in Pharmacology*, vol. 15, p. 1398381, 2024.

[80] X. Liu, K. Xing, Q. Zheng, T. Li, X. Ma, T. Zhang, J. Sun, J. Song, and Z. Wang, "VAPB promotes osteogenic differentiation in aortic valve interstitial cells via activation of the SMAD signaling pathway," *Current Medicinal Chemistry*, vol. 32, 2025, Art. no. e09298673355591.

[81] W. Wang, Y. Li, M. Zhu, Q. Xu, J. Cui, Y. Liu, and Y. Liu, "Danlian-Tongmai formula improves diabetic vascular calcification by regulating CCN3/NOTCH signal axis to inhibit inflammatory reaction", *Frontiers in Pharmacology*, vol. 15, p. 1510030, 2025.

[82] B. B. Khomtchouk, K. A. Vand, W. C. Koehler, D. T. Tran, K. Middlebrook, S. Sudhakaran, C. S. Nelson, O. Gozani, and T. L. Assimes, "HeartBioPortal", *Circulation: Genomic and Precision Medicine*, vol. 12, no. 4, p. e002426, 2019.

7. Appendices

7.1. Supplementary Tables

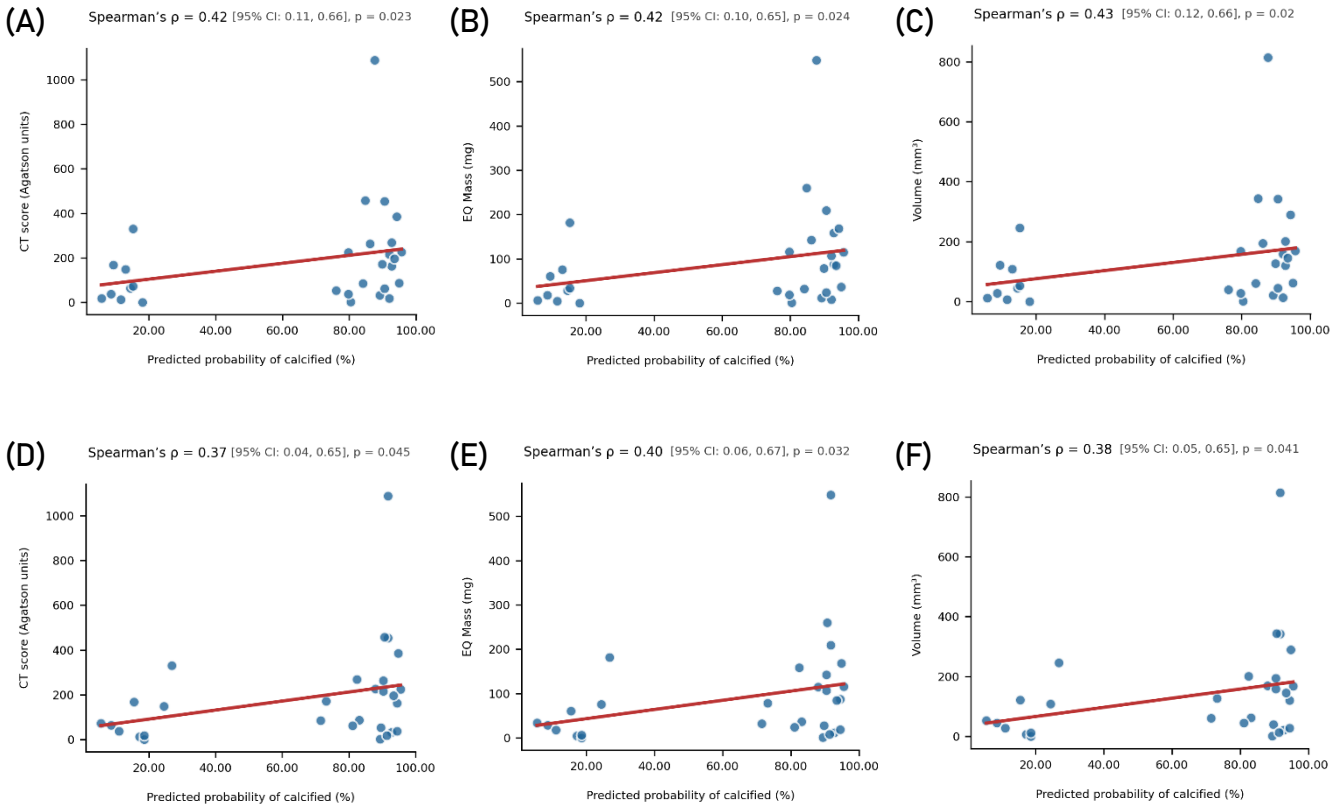
Reference	Sample Size (patients) / Discovery cohort	Sample Type	Proteomics Method
Aragones et al., *J Proteome Res.*, 2016	N=12	Human carotid endarterectomy tissue (secretome)	Untargeted MS with iTRAQ-8 labeling
Langley et al., *J Clin Invest.*, 2017	N=12	Human carotid endarterectomy tissue (fresh-frozen)	Untargeted label-free MS
Ucciferri et al., *Talanta.*, 2017	N=13	Human carotid endarterectomy tissue (fresh but not frozen)	Untargeted label-free MS
Hansmeier et al., *J. Proteome Res.*, 2018	N=12	Human carotid endarterectomy tissue (fresh-frozen)	Data independent acquisition (DIA) MS
Ward et al., *Biol Sex Differ.*, 2018	N=20	Human carotid endarterectomy tissue (fresh-frozen)	Untargeted label-free MS
Matic et al., *JACC Basic Transl Sci.*, 2018	N=18	Human carotid endarterectomy tissue (fresh-frozen) + plasma	Untargeted MS with TMT-labeling
Nehme et al., *Hypertens Res.*, 2019	N=12	Human carotid endarterectomy tissue (fresh but not frozen)	Untargeted label-free MS & targeted MS for 53 proteins
Theofilatos *et al.*, *Circ. Res.*, 2023	N=120	Human carotid endarterectomy tissue (fresh-frozen)	TMT-based LC-MS/MS (discovery) + PRM (targeted val.)
Kalló et al., *IJMS*, 2024	N=15	Human carotid tissue (endarterectomy & autopsy, fresh-frozen)	Label-free LC-MS/MS (Orbitrap, DDA & DIA)
Lai et al., *J. Transl. Med.*, 2024	N=88	Human carotid endarterectomy tissue	DIA-MS (label-free)
Lorentzen et al., *Matrix Biol. Plus*, 2024	N=21	Human carotid endarterectomy tissue (fresh-frozen in buffer)	Label-free LC-MS/MS (DIA-PASEF), ECM-focused
Palm et al., *Cardiovasc. Res.*, 2025	N=200	Human carotid endarterectomy tissue (fresh-frozen)	Label-free untargeted LC-MS/MS + TMT-LC-MS/MS (val.)
Wang et al., *BMC Med.*, 2025	N=87	Human carotid endarterectomy tissue (FFPE)	Label-free DIA LC-MS/MS (Orbitrap)

Supplementary Table 1.
Summary of the most recent and comparable proteomics studies on carotid endarterectomy samples from human subjects, most of which were limited by small cohort sizes, with the exception of the larger Theofilatos et al. [40] and Palm et al. [47] studies.

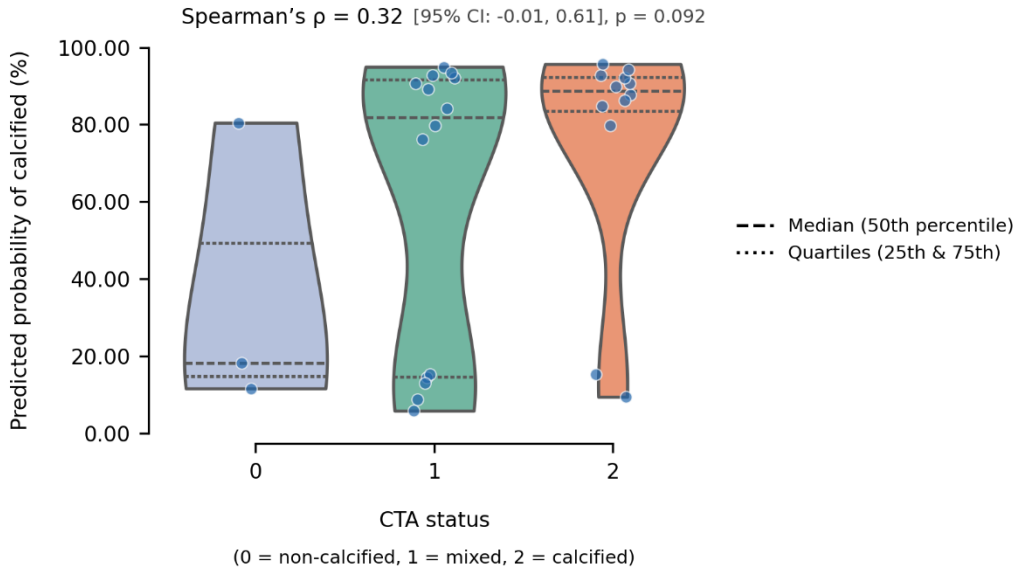
Training set	Features	Core samples	Periphery samples	Total samples	Patients
Soluble Matrisome	261	105	102	207	119
Core Matrisome	293	105	108	213	120
Cellular Proteome	1405	101	105	207	120

Supplementary Table 2.
Summary of the training datasets used for calcification prediction models. The table lists the number of quantified protein features, core and periphery plaque samples, total samples, and patients included for each extraction protocol (soluble matrisome, core matrisome, and cellular proteome) derived from the Vienna cohort.

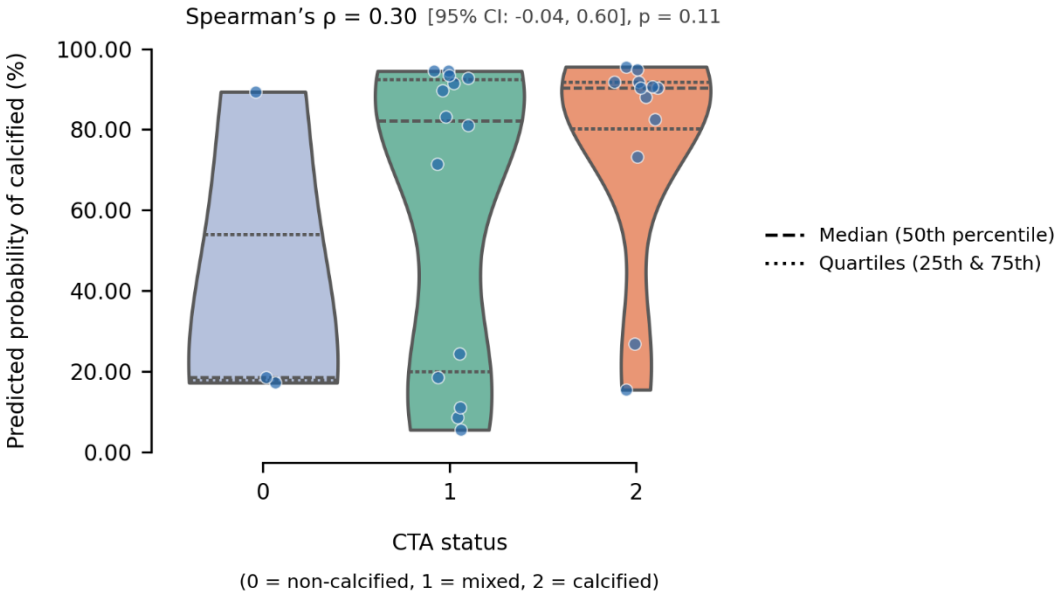
7.2. Supplementary Figures



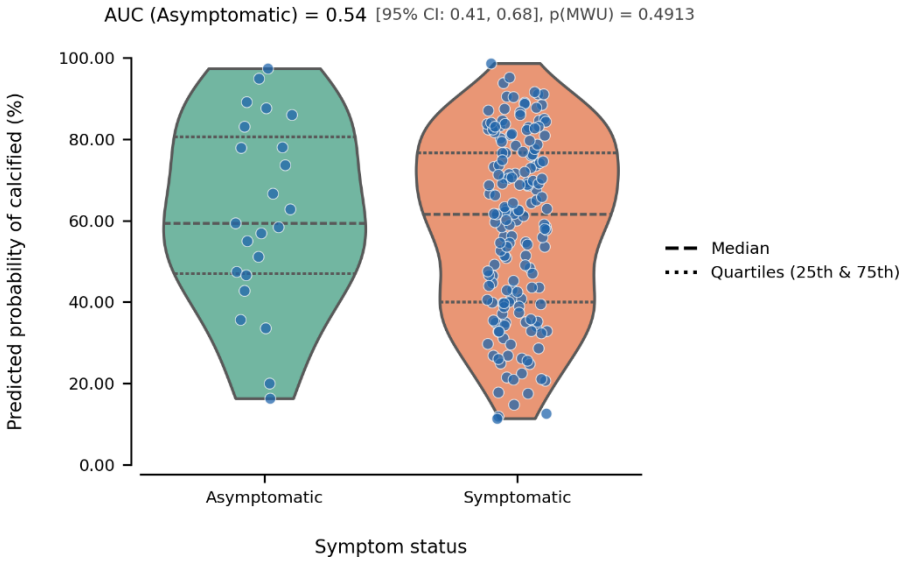
Supplementary Figure 1.
Scatter plots show the relationship between predicted probability of calcification and CT-derived metrics for the cellular proteome (**A–C**) and soluble matrisome (**D–F**) models. Both models display limited gradation across intermediate CT scores, with predictions forming two distinct clusters corresponding to high and low probabilities.



Supplementary Figure 2.
Violin plot depicting the distribution of predicted probabilities of calcification from the cellular proteome model across CTA status categories (0 = non-calcified, 1 = mixed, 2 = calcified). A weak, non-significant positive correlation was observed (Spearman's ρ = 0.32, 95% CI: –0.01 to 0.61, *p* = 0.092), with higher median probabilities in the calcified group.



Supplementary Figure 3.
Violin plot depicting the distribution of predicted probabilities of calcification from the soluble matrisome model across CTA status categories (0 = non-calcified, 1 = mixed, 2 = calcified). A weak, non-significant positive correlation was observed (Spearman's ρ = 0.30, 95% CI: –0.04 to 0.60, *p* = 0.11), with higher median probabilities in the calcified group.



Supplementary Figure 4.
Violin plot depicting the distribution of predicted probabilities of calcification from the core matrisome model in the Athero-Express cohort, stratified by symptom status (asymptomatic vs. symptomatic). The model did not achieve significant separation between groups (AUC = 0.54, 95% CI: 0.41–0.68, Mann–Whitney U test, p = 0.4913).