UNIVERSITY OF VERONA

Faculty of Computer Science

Bachelor's Degree in Bioinformatics

Bioinformatics and Aesthetic Medicine:

How Deep Proteomic Analysis Can Provide New Insights into the Treatment of Aesthetic Pathologies

Supervisor:

Alessandro Daducci

Candidate:

Michael Korsa Parkson Brew

Academic Year: 2024–2025

Contents

1 Introduction

1.1 Introduction to Lipedema

In recent years, the role of technology has become increasingly pivotal in scientific, medical, and deep-tech research. This thesis explores how bioinformatics can support the investigation of the causative factors behind lipedema.

Lipedema is a chronic and debilitating disorder that predominantly affects women. It is characterized by an abnormal accumulation of subcutaneous adipose tissue (SAT), mainly in the lower limbs, with occasional involvement of the upper limbs, while sparing the trunk and head [1]. Initially described by Allen and Hines in 1940, the condition involves lipid accumulation accompanied by fluid retention [2]. It typically manifests as symmetrical swelling of the extremities, proliferation of loose connective tissue (LCT), and glycosaminoglycans (GAGs) in the interstitial matrix [3]. Although frequently misdiagnosed as lymphedema, lipedema is estimated to affect 6–11% of women [1].

Despite its prevalence, lipedema is under-recognized in clinical practice, largely due to knowledge gaps in medical training and its unclear etiology [4]. It shares several overlapping features with conditions such as obesity, lymphedema, and connective tissue disorders, making diagnosis particularly challenging [5, 1]. Its pathogenesis is multifactorial, involving vascular, lymphatic, extracellular matrix, hormonal, metabolic, and genetic components [6]. Candidate genes such as *AKR1C1*, *POU1F1*/*PIT-1*, and associations with Williams syndrome have been identified [7, 8, 9].

Hormonal fluctuations—especially elevated estrogen during puberty, pregnancy, or menopause may initiate or exacerbate lipedema by enhancing the adipogenic differentiation of adipose-derived stem cells [7, 10]. Other hypotheses suggest that fat tissue expansion compromises connective tissue elasticity and vascular structure, triggering hypoxia, inflammation, and fibrosis [7].

Clinically, lipedema progresses through three stages. Stage I features mild discomfort and small nodules, while Stage III involves severe fibrotic adipose masses, persistent pain, and resistance to conventional treatment[7, 11, 3]. Pain, often neuropathic and poorly understood, may involve neurogenic inflammation and increased sodium content in affected tissues [11, 6].

Lipedema's impact on peripheral arterial circulation has not been deeply studied. Early microangiopathy may include endothelial dysfunction, impaired vascular reflex control, and hemodynamic stress due to hypertrophic adipose tissue and edema. These factors reduce arteriolar vasoconstriction, contributing to hydrostatic edema and subsequent chronic tissue alterations, including sclerosis and papillomatosis [12].

Hormonal shifts during puberty, pregnancy, or menopause appear to trigger lipedema onset [13, 14]. Estrogen, a key hormone in lipid and glucose metabolism, may induce local fat accumulation by altering receptor distribution and enhancing steroidogenic enzyme activity. This contributes to increased lipid deposition, angiogenesis, and decreased mitochondrial and lipolytic activity [13]. Genetic links, such as mutations in *AKR1C1*, and the frequent occurrence of hypothyroidism in lipedema patients further support an endocrine–metabolic basis [14].

SAT expansion in lipedema is thought to result from adipocyte dysfunction, stem-cell-mediated inflammation, extracellular matrix remodeling, and microvascular damage [5]. Unlike lymphedema, lipedema may involve fluid retention due to capillary or lymphatic abnormalities.

Common vascular alterations include capillary fragility, veno-arteriolar reflex failure, aortic stiffness, and lymphatic dysfunction [14]. Lymphoscintigraphic imaging reveals reduced lymphatic transport and tortuous vessels, while Platelet Factor 4 has emerged as a potential biomarker [14]. Cardiovascular anomalies in lipedema include atrial and ventricular enlargement and changes in cardiac twist and strain [14].

MRI studies have detected elevated skin sodium concentrations in lipedema, possibly due to impaired lymphatic clearance. This accumulation is associated with tenderness and pain, which may improve after liposuction or decongestive therapies. Nerve conduction studies point to sensory nerve impairment due to mechanical pressure and inflammatory processes, correlating with higher migraine rates in affected patients [14].

Whether lymphatic dysfunction is a cause or consequence of fat accumulation in lipedema remains unclear. Animal studies support both directions: lymphatic disruption can promote fat deposition, and chronic adipose expansion can hinder lymphatic drainage [15].

Currently, no definitive cure exists for lipedema. Treatment focuses on symptom management, functional improvement, and slowing disease progression. Management strategies include conservative decongestive therapy (CDT) and, when necessary, surgical intervention to reduce pain, bruising, and abnormal fat accumulation [16, 2].

1.2 Similarities Between Lipedema and Cellulite

Cellulite, a condition affecting more than 85% of adult women, appears as dimpled or lumpy skin in the gluteofemoral region [17]. Both lipedema and cellulite are gluteofemoral disorders and may share pathophysiological traits [18]. Understanding cellulite's mechanisms can help illuminate the biology of lipedema.

This pathology also affects fat and connective tissues and is marked by fluid retention between cells. This retention hinders cellular exchange, induces tissue reactions, and may impair lymphatic function. Both conditions involve abnormal fat deposition and lymphatic disruption, leading to limb swelling, tissue sensitivity, and increased girth [18].

Recent molecular research has indicated that excessive TGF-β signaling in cellulite septae initiates collagen over-accumulation and activates MMP-2 and MMP-9 where extracellular-matrix remodeling takes place, leading to increased septal stiffening, adipose-lobule protrusion, and contributes to protrusion [19, 20]. Proteomic analyses further reveal elevated levels of cytokines and pro-inflammatory mediators like IL-6, TNF-α, and MCP-1 within the affected tissues, indicating that chronic inflammation alters vascular permeability and increases adipocyte hypertrophy [21]. These provide significant clues in understanding the interplay of extracellular-matrix remodeling, inflammatory signaling, and microvascular dysfunction, and likely the reason behind the etiology of both cellulite and lipedema.

1.3 Objective of the Study: Identification of Potential Biomarkers for Lipedema

As bioinformatics continues to evolve, it enables deeper exploration of lipedema and its relation to conditions like cellulite. This study aims to utilize computational tools such as Python, R, miRDB and TargetScan, alongside proteomic analysis, to identify potential biomarkers and better understand lipedema's underlying mechanisms.

Following this, we will deal with deep proteomic datasets of lipedema and cellulite pairs, validating Python scripts and performing Gene Ontology and pathway enrichment analyses (GProfiler [22]) to find the central biological processes and molecular functions associated with lipedema. Later, we will do thorough differential-expression analysis of the extracellular miRNA profile—with the limma package in R and scanpy workflows in Python—focusing on the most significant dysregulation. The chosen miRNAs will be inputted into both TargetScan [23] and miRDB [24], yielding two complementary sets of predicted mRNA target. We will refine them using the conservation score and expression correlation filtering.

In order to obtain a concise panel of candidate biomarkers, the last, integrative step is to merge miRNA-target predictions with the empirically observed changes in proteins of interest. The resulting biomarkers will (i) differentiate lipedema from other related adipofibrotic disorders, and (ii) expose the secret pathological axes of the extracellular-matrix remodelling, chronic inflammation, and metabolic dysregulation that drives the disease.

2 Materials and Methods

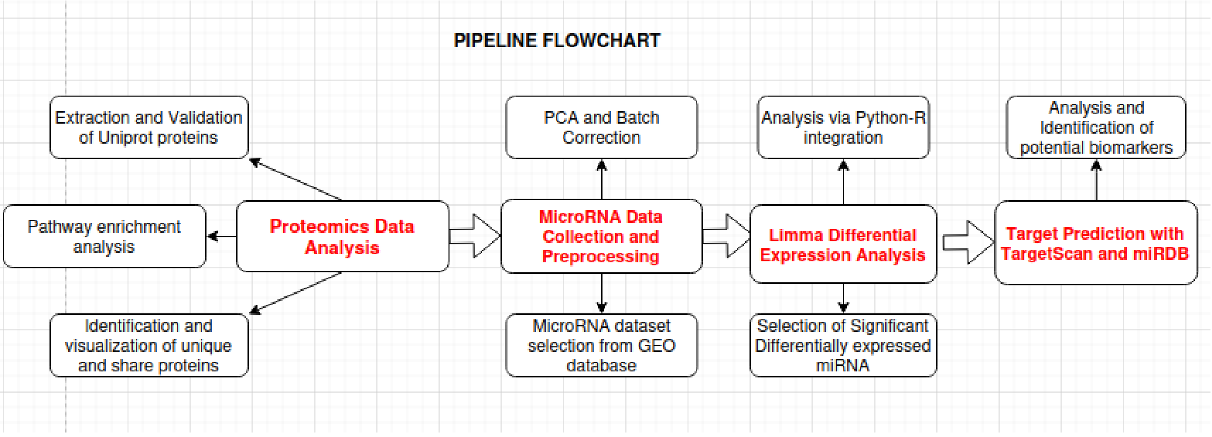
PIPELINE FLOWCHART
Extraction and Validation of Uniprot proteins
Pathway enrichment analysis
Proteomics Data Analysis
Identification and visualization of unique and share proteins
PCA and Batch Correction
MicroRNA Data Collection and Preprocessing
MicroRNA dataset selection from GEO database
Analysis via Python-R integration
Limma Differential Expression Analysis
Selection of Significant Differentially expressed miRNA
Analysis and Identification of potential biomarkers
Target Prediction with TargetScan and miRDB

Figure 1: Flow chart depicting the computational pipeline used in this project.

2.1 Proteomics Data Analysis

2.1.1 Extraction and Validation of UniProt Proteins

The first step in this study involves the analysis of proteomics data—specifically, UniProt IDs obtained from two female patients, one with lipedema and the other with cellulite—using mass spectrometry. The goal is to determine whether these two conditions exhibit similarities at the proteomic level.

A Python-based script was developed to process, validate, and analyze UniProt IDs for both conditions. The script integrates file input/output operations, API communication with the UniProt database, and data visualization techniques. Ultimately, it leads to the comparative analysis of protein sets using a Venn diagram.

The initial step of the workflow involves the extraction of protein identifiers from two separate input files, `lipedema.txt` and `cellulite.txt`, containing protein names corresponding to lipedema and cellulite, respectively. A dedicated function, `read_proteins(file_path)`, is implemented to read these files. This function opens the file, iterates over each line to remove extraneous whitespace, and stores the resulting identifiers in a set to guarantee the separation needed for the end product. Integrated into this function is also a mechanism to catch exceptions in case of file reading errors. It prints an error message and returns an empty set to ensure that the workflow remains robust even in the face of I/O issues.

Following the extraction of protein IDs, the script validates the existence of each protein in the UniProt database. This is done through the `check_protein_in_uniprot(protein_name, retries=3, backoff_factor=1)` function. It creates a URL by appending the protein name to the base UniProt URL and makes an HTTP GET request using the `requests` library, with a timeout setting to deal with network delays. A successful response indicates the presence of the protein; otherwise, a 404 response triggers a log message noting its absence. To further enhance reliability, the function uses an exponential backoff strategy when faced with network errors, attempting up to three retries before declaring that the protein is not present. Specifically, if a request fails due to a network-related exception, the script waits 2^n seconds (where n is the retry attempt count) before trying again. The figure below shows the validation function.

```python
# Function that validates protein existence in UniProt with error handling
def check_protein_in_uniprot(protein_name, retries=3, backoff_factor=1):
    url = f"https://www.uniprot.org/uniprot/{protein_name}.txt"
    for attempt in range(retries):
        try:
            response = requests.get(url, timeout=10)
            if response.ok:
                return True  # protein exists
            elif response.status_code == 404:
                print(f"Protein {protein_name} not found in UniProt.")
                return False
        except requests.exceptions.RequestException as e:
            print(f"Error checking protein {protein_name}: {e}")
            if attempt < retries - 1:
                # exponential backoff
                time.sleep(backoff_factor * (2 ** attempt))
            else:
                print(f"Failed to validate protein {protein_name} after {retries} attempts.")
    return False  # if all retries fail
```

Listing 1: Python function to check protein existence in UniProt with error handling.

The validation process is encapsulated in a higher-level function named `filter_valid_proteins (proteins)`. This function iterates over the set of protein names, calls `check_protein_in_uniprot` for each one, and constructs a subset containing only those proteins confirmed by the UniProt database. The filtered sets then serve as the basis for further comparative analysis.

2.1.2 Identification of Shared and Unique Proteins Between Lipedema and Cellulite

After the validation of protein sets for lipedema and cellulite, the script performs set operations to compute three important subsets: proteins unique to lipedema, proteins unique to cellulite, and proteins common to both conditions. The unique sets are derived using set difference operations, while the common proteins are determined via an intersection of the two validated sets.

2.1.3 Pathway Enrichment Analysis

A pathway analysis was vital to this project and it was developed with an integrated Python pipeline designed to process protein identifiers derived from the text files containing protein IDs belonging to lipedema, cellulite, and common proteins, enrich these identifiers with additional biological information obtained from the UniProt database, and perform pathway enrichment analysis using the GProfiler tool. The workflow was implemented with a modular design, incorporating robust error handling and logging mechanisms to facilitate traceability and reproducibility.

The initial component of this process involves reading protein names from designated text files. In this implementation, the function `read_proteins(file_path)` is responsible for opening a file, extracting each non-empty line (after stripping unnecessary whitespace), and putting the protein names into a list.

After data extraction, the pipeline enriches the protein dataset by retrieving additional information from the UniProt database. The function `fetch_uniprot_data(protein_names)` iterates over the protein identifiers and constructs specific URLs to obtain corresponding text-based entries. For each successful HTTP request, the response is passed to the `parse_uniprot_response(response _text)` function, which methodically parses the content. This parser examines individual lines of the response for key prefixes, in this case, lines starting with "ID" to extract protein identifiers, "DE" for descriptions, "GN" for gene names, "OS" for organism data, and "DR" and "CC" for pathway and disease-related information. In cases where a protein's record cannot be retrieved, the system logs a warning, which ensures that potential data gaps are transparently reported.

Listing 2: Python function for parsing UniProt response text to extract ID, description, pathways, gene, organism, and disease information.

```python
# Parsing of UniProt response to extract useful information
def parse_uniprot_response(response_text):
    data = {}
    lines = response_text.splitlines()
    for line in lines:
        if line.startswith("ID"):
            data['ID'] = line.split()[1]
        elif line.startswith("DE"):
            data.setdefault('Description', []).append(line[5:])
        elif line.startswith("DR"):
            if "Pathway" in line:
                data.setdefault('Pathways', []).append(line.split(';')[1].strip())
        elif line.startswith("OS"):
            data['Organism'] = line[5:]
        elif line.startswith("GN"):
            data['Gene'] = line[5:]
        elif line.startswith("CC"):
            if "Disease" in line:
                data.setdefault('Diseases', []).append(line[5:])
    return data
```

The enriched dataset is then passed on to the pathway enrichment analysis, a step important to identify biological information about the proteins. The function `pathway_analysis(protein_names,category_name="All Proteins")` employs the GProfiler tool to query enrichment in key biological databases, specifically targeting human gene annotations from the Gene Ontology (GO) for Biological Process, Molecular Function, and Cellular Component, as well as KEGG pathway annotations. The analysis returns a data frame containing enrichment results, from which the top 10 significant terms are extracted for each category. These results are then saved as CSV files in a dedicated output directory, thereby ensuring that all details of the enriched pathways are well-documented and organized for further investigation.

Listing 3: Python function for performing pathway enrichment analysis using GO and KEGG terms.

```python
def pathway_analysis(protein_names, category_name="All Proteins"):
    gp = GProfiler(return_dataframe=True)
    results = gp.profile(organism='hsapiens', query=protein_names,
                         sources=['GO:BP', 'GO:MF', 'GO:CC', 'KEGG'])

    logging.info(f"Pathway analysis for {category_name}")
    logging.info(f"Columns available: {results.columns.tolist()}")

    if not results.empty:
        # Filter and save top 10 results for each category
        go_bp = results[results['source'] == 'GO:BP'].head(10)
        go_mf = results[results['source'] == 'GO:MF'].head(10)
        go_cc = results[results['source'] == 'GO:CC'].head(10)
        kegg = results[results['source'] == 'KEGG'].head(10)

        # Results directory
        output_dir = os.path.join(os.getcwd(),
        ↪  "pathway_analysis_results")
        os.makedirs(output_dir, exist_ok=True)

        def save_top_to_csv(data, label): #csv saving of output
            csv_path = os.path.join(output_dir,
            ↪  f"{category_name}_{label}_top10.csv")
            data.to_csv(csv_path, index=False)
            logging.info(f"Saved top 10 {label} pathways for
            ↪  {category_name} to {csv_path}")

        if not go_bp.empty: #saving top 1o of each pathway category
            save_top_to_csv(go_bp, "GO_BP")
        if not go_mf.empty:
            save_top_to_csv(go_mf, "GO_MF")
        if not go_cc.empty:
            save_top_to_csv(go_cc, "GO_CC")
        if not kegg.empty:
            save_top_to_csv(kegg, "KEGG")

        plot_histogram(go_bp, category_name)
    else:
        logging.info(f"No pathway enrichment results found for
        ↪  {category_name} proteins.")
```

The pipeline also has the function `plot_histogram(data, category_name)` which generates a horizontal bar chart displaying the negative logarithm of the p-values (log10) associated with the top 10 enriched biological processes. By transforming the p-values in this manner, the chart provides a clear depiction of the statistical significance of each enrichment term. The visualization is both displayed on the screen and later saved as an image file within the results directory, ensuring that the data is accessible for subsequent review.

Finally, the main execution block orchestrates the entire pipeline. It begins by reading protein data from files representing distinct protein groups (e.g., proteins unique to lipedema, proteins unique to cellulite, and proteins common to both conditions). After consolidating these datasets, the script fetches additional data from UniProt and sequentially performs pathway enrichment analysis for each protein category. This comprehensive pipeline further adds more information to the proteins to help get a better understanding related to biological pathways.

2.2 microRNA Dataset Collection and Preprocessing

2.2.1 microRNA Dataset Selection from the GEO Database

The dataset used for this analysis was taken from the di GEO (Gene expression omnibus) database of the NCBI database. It was chosen because it was deemed the best choice out of the lipedema dataset options that I had. It was used in a recent project in 2020 that involved the published article "microRNA analysis of supernatants from stromal vascular fraction cells derived from healthy controls and lipedema patients" [25].

SVF cells were isolated from liposuction material (hips, outer thigh) of healthy controls and lipedema patients and seeded for the collection of cell supernatants after 24 hours. After several centrifugation steps, the supernatants were concentrated and enriched for small extracellular vesicles. Total RNA was extracted from the concentrated supernatants and the small extra cell vesicle fraction. A total of 187 miRNAs were evaluated through the qPCR reaction (miRCURY SYBR® Green master mix, Quiagen). [25]

MicroRNAs (miRNAs) are classified as non-coding RNA molecules that are essential for the regulation of gene expression. Most miRNAs go through a series of stages: they are transcribed from DNA sequences into primary miRNAs, processed into precursor miRNAs, and eventually become mature miRNAs. Usually, miRNAs bind to the 3' untranslated region (3' UTR) of the target mRNA, leading to its decay and translational silencing. However, the binding of miRNAs to other regions, such as 5' UTR, coding region, and gene promoter, has been documented as well. Under certain conditions, mirnas might be able to trigger translation and modulate transcription. The way miRNAs interact with their target genes is complex and influenced by many parameters including the subcellular localization of miRNAs, their quantity in relation to the target mRNAs, and the strength of interaction between miRNA and mRNA. Vesicles such as exosomes carry miRNAs in extracellular fluids to target cells. Alternatively, they can be bound to proteins like Argonautes. Extracellular miRNAs participate in intercellular communication as molecular messengers substances. [26]

For quantitative analysis, high throughput methods like qPCR or other next generation sequencing techniques are preferred. In combination with their tissue specific expression patterns , these factors render these constituents to be microRNAs to be not only source of strong and dependable biomarkers, but alters surfacing molecular disease mechanisms, increases potential for early diagnosis, and aids appropriate treatment planning.

As for this project, the values in the dataset are in dCqs. The dCq (delta Cq) values signifies the difference between quantification cycle of a specific microRNA ($\mathrm{Cq}_{\mathrm{miRNA}}$) and a reference value (mean Cq or a housekeeping gene) [27, 28]. This normalization facilitates the elimination of methodological biases and enables accurate assessment of the expression level of each microRNA and comparing different samples. dCq corresponds to decrease in expression level and higher dCq translates to elevated dCq meaning lower expression.

Initially, cleaning the dataset started with filling the gaps by zero. Gaps that were missed were This strategy prevents computational errors—such as the disruption of arithmetic operations like summing or averaging—and avoids the exclusion of critical data points. Additionally, when missing values logically indicate a non-occurrence or the absence of a measurable entity, zero substitution provides both a conceptually accurate and consistent representation, thereby enhancing data integration and interpretability across diverse data sources.

2.2.2 Principal Component Analysis (PCA) and Batch Correction

This study implemented an integrated computational framework to ensure robust microRNA data analysis by addressing technical variability and enhancing biological signal detection. Initially, the microRNA dataset—originally stored as a CSV file—was loaded into a Pandas DataFrame to allow efficient data handling and manipulation of the dataset. To mitigate batch effects, which can introduce systematic biases when samples are processed under varying conditions, the data was subsequently transformed into an AnnData object using Scanpy. The AnnData format is particularly advantageous in high-dimensional data contexts, as it encapsulates not only the expression matrix but also corresponding metadata, such as batch and group annotations [29]. The resulting uniformity allowed the application of the ComBat batch-correction procedure [30], which mitigates technical variation effectively such that subsequent analysis captures biological variance rather than artefact due to technique.

Subsequent to batch clearing, PCA or Principal Component Analysis was performed to reduce the dataset dimensions to a more interpretable form while also enhancing sample comparison.

PCA was first applied to the original data, where the resulting clustering primarily reflected the influence of batch effects. In contrast, PCA performed on the batch-corrected data revealed improved segregation based on biological groupings (Control vs Lipedema), thereby confirming the efficacy of the the batch-correction process.

Listing 4: Python function for PCA and batch correction

```
file = 'microRNA analysis_dataset.csv'
data = pd.read_csv(file, delimiter='\t', index_col=0)

print(data.columns)
print(data.shape)

# Metadata annotation
batch_labels = ['sEV'] * 6 + ['cCM'] * 6
group_labels = ['Control'] * 3 + ['Lipedema'] * 3 + ['Control'] * 3 +
↪  ['Lipedema'] * 3

# Map batch labels to numerical categories for coloring
unique_batches = list(set(batch_labels))  # ['sEV', 'cCM']
batch_colors = [unique_batches.index(label) for label in batch_labels]  #
↪  Map to integers

# Map group labels to numerical categories for coloring
unique_groups = list(set(group_labels))  # ['Control', 'Lipedema']
group_colors = [unique_groups.index(label) for label in group_labels]  #
↪  Map to integers

# Batch correction with conversion of data to AnnData format
adata = sc.AnnData(data.T)  # Transpose: genes as columns, samples as rows
adata.obs['batch'] = batch_labels  # Add batch information
adata.obs['group'] = group_labels  # Add group labels

sc.pp.combat(adata, key='batch') # Application of batch correction using
↪  ComBat

# Save the corrected data
corrected_data = pd.DataFrame(adata.X.T, index=data.index,
↪  columns=data.columns)
corrected_data.to_csv('batch_corrected_data.csv')

print("Batch correction completed and saved to 'batch_corrected_data.csv'")

pca = PCA(n_components=2) # PCA before correction
pca_result_before = pca.fit_transform(data.T)

pca_result_after = pca.fit_transform(corrected_data.T) # PCA after
↪  correction
```

2.3 Limma Differential Expression Analysis

2.3.1 Analysis via Python-R

The computational pipeline created uses both Python and R environments to perform differential expression analysis on a batch-corrected microRNA dataset. The process begins with the loading of the batch-corrected data from a CSV file into a Pandas DataFrame. Following data import, metadata was annotated by assigning group labels ("Control" and "Lipedema") to the samples, laying the foundation for comparative analysis.

To exploit the statistical capabilities of the Bioconductor package limma, the dataset was converted from a Pandas DataFrame to an R-compatible object using the *pandas2ri* module from *rpy2*. This conversion not only maintained data integrity but also enabled correct integration with R. The limma package was installed (when necessary) and loaded in the R environment. Within R, the dataset was transformed into a numeric matrix, and a corresponding design matrix was created based on the annotated group labels. A linear model was then fitted using lmFit, followed by empirical Bayes moderation via *eBayes* to enhance statistical power and stabilize variance estimates. The differential expression analysis resulted in the extraction of a ranked list of microRNAs using *topTable*, and these results were converted back to Python for further examination and record-keeping.

In differential expression analyses using Linear Models for Microarray Data (limma), several key statistical measures are routinely reported to assess the significance and magnitude of differential gene expression between experimental conditions.

Firstly, the log fold-change (logFC) represents the magnitude of differential expression, indicating how many times the expression of a particular gene increases or decreases between conditions, expressed in log base 2 scale. Positive values denote upregulation, while negative values represent downregulation.

Secondly, the moderated t-statistic provided by limma is obtained through empirical Bayes moderation, borrowing information across genes to stabilize variance estimates. This moderated t-statistic is similar to a standard t-test but is enhanced by sharing variance information among genes, especially beneficial when sample sizes are small.

The p-value obtained alongside the moderated t-statistic measures the probability of observing the reported gene expression differences purely by chance. Lower p-values indicate stronger evidence against the null hypothesis (no differential expression), thus implying significant differential expression.

The adjusted p-value, or FDR as it is called, has corrective features for multiple comparisons and controls the proportion of null hypotheses that are incorrectly rejected.

It provides a more reliable measure of significance, with adjusted p-values typically interpreted using a threshold such as 0.05 to identify differentially expressed genes robustly.

Finally, the B-statistic or log-odds (B-value) represents the log posterior odds of differential expression. The B-statistic combines prior information with the observed data using empirical Bayesian methods to quantify the strength of evidence that a gene is differentially expressed. Positive B-values indicate evidence supporting differential expression, whereas negative values suggest weak or no evidence for differential expression.

2.3.2 Selection of Significant Differentially Expressed miRNA

Building upon the differential expression results, the analysis framework incorporated a data-driven step for determining an optimal significance threshold. A series of potential adjusted p-value thresholds were generated using *NumPy*, and for each threshold, the number of significant microRNAs was computed. These data were visualized by plotting the thresholds against the count of significant microRNAs, and the KneeLocator algorithm from the kneed package was employed to identify a "knee" point in this curve.[31] This knee point represented an optimal threshold, balancing sensitivity and specificity in the detection of significant microRNAs. The pipeline then used this optimal threshold to filter the differential expression results, with both the complete and the threshold-filtered results saved as CSV files. Additionally, an exploratory analysis was carried out to identify microRNAs exhibiting large fold changes ($|\text{logFC}| > 1$) in conjunction with raw p-values below 0.05, thereby highlighting candidates with pronounced biological relevance.

2.4 microRNA Target Prediction

2.4.1 miRNA-mRNA Target Prediction Using miRDB and TargetScan

Witin this study, a processing protocol was implemented to filter microRNA data and determine three optimal candidates based on an empirically derived threshold. After this selection, target gene prediction was performed using two distinct, complementary databases—TargetScan and miRDB—to elucidate potential regulatory interactions that might serve as biomarkers for lipedema.

Particularly, two of the candidates were investigated with TargetScan whiles the other with miRDB. One of the significant resources is TargetScan, which predicts microRNA targets based on how well microRNAs bind to the 3' untranslated region (UTR) of the mRNA. Its algorithms place greater emphasis on conserved binding sites as well as contextual features, like locomotion in mRNA, particular heuristic score metrics termed context scores heaped oversequence garnered rich lineage history. Such scoring systems help in the attainment of targets that may be phylogenetically archetypal signifying, enhancing confidence that such interactions are relevant.

On the other side, miRDB uses other methodology algorithms based on machine learning and trained with massive amounts of data from high-throughput experimental data.

It provides target predictions accompanied by confidence scores, which quantify the reliability of each prediction. By leveraging patterns learned from validated microRNA–mRNA interactions, miRDB offers an alternative and complementary perspective to sequence conservation-based methods. This dual strategy enhances the robustness of target gene identification by mitigating the limitations inherent in any single predictive model.

The combined use of these two databases—in conjunction with a rigorous threshold optimization process—ensured that the candidate microRNAs selected for further study were supported by complementary lines of computational evidence. This integrative approach not only refines the pool of potential biomarkers for lipedema but also lays a solid foundation for subsequent experimental validation. By employing TargetScan and miRDB, the study capitalizes on both evolutionary conservation and data-driven predictive methodologies to provide a more comprehensive understanding of microRNA-mediated regulatory mechanisms in the context of lipedema pathology

3 Results

3.1 Venn Diagram Analysis: Overlap of Proteomic Profiles in Lipedema and Cellulite

The first outcome of this project comes from the generation of a Venn diagram, created using a Python-based computational pipeline. The goal of this analysis was to determine whether the two conditions—lipedema and cellulite—share any proteins in common. The dataset contained approximately 652 proteins related to lipedema and about 281 proteins linked to cellulite. After running the pipeline, the resulting Venn diagram revealed a significant overlap between the two sets. This suggests that the two conditions may share underlying molecular characteristics, and the similarity observed is unlikely to be due to chance alone.

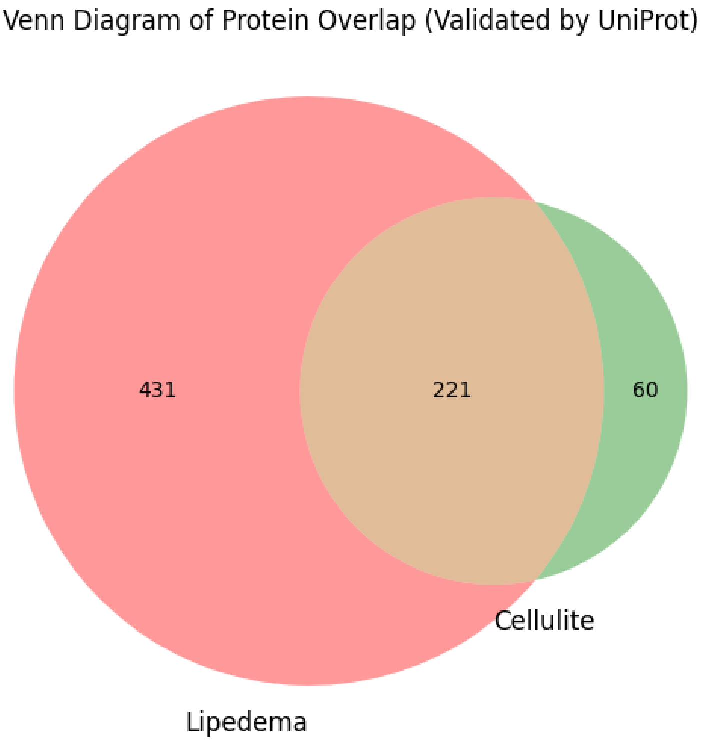

Figure 2: Venn diagram validated by Uniprot

3.2 Comparative Pathway Analysis of Lipedema- and Cellulite-Associated Proteins

Pathway analysis was carried out to gain deeper insight into the molecular mechanisms underlying lipedema and cellulite. The computational pipeline used in this step first validated the identified proteins via the UniProt API, and then performed functional enrichment analysis using GProfiler. This allowed for the identification of key biological processes, cellular components, and KEGG pathways associated with three groups: proteins unique to lipedema, those unique to cellulite, and those shared between both conditions. The pipeline generated CSV output files for each of these functional categories, but the main focus was placed on the top 10 enriched biological processes for each group. The corresponding bar charts, shown below, visually represent these top pathways for the three protein sets.

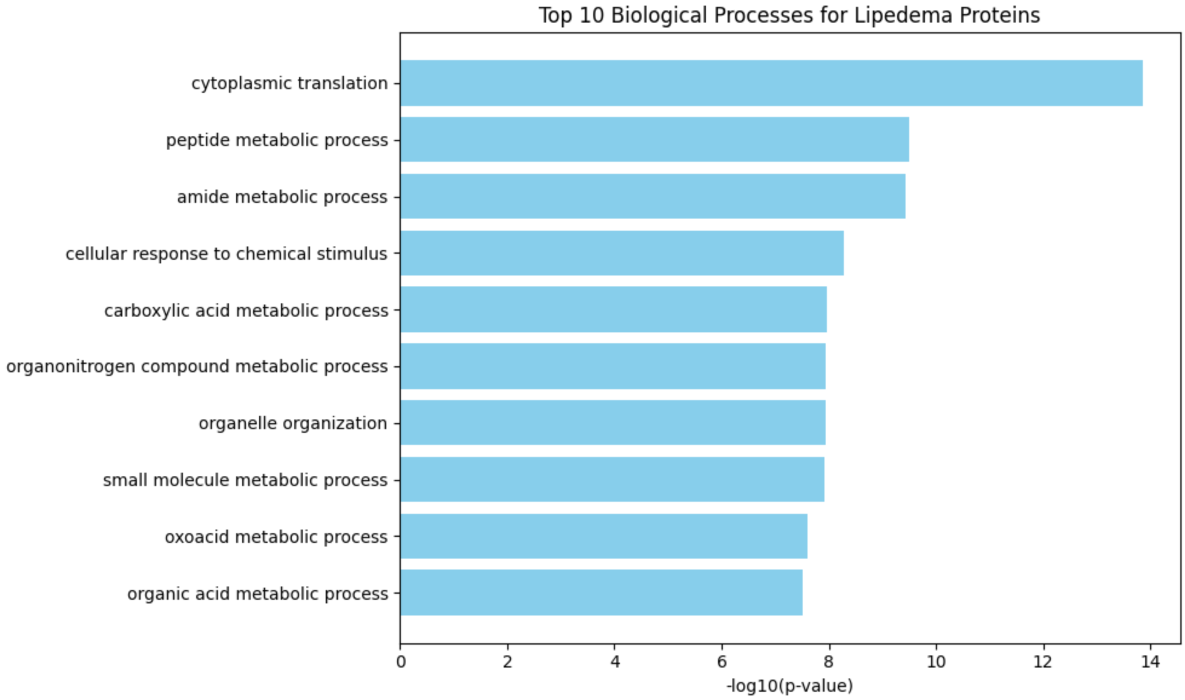

Figure 3: presents the top 10 significantly enriched biological processes associated with proteins unique to lipedema, as identified through GO (Gene Ontology) enrichment analysis using GProfiler. The x-axis displays the statistical significance of each term as –log10(p-value), where higher values indicate stronger enrichment. Notably, cytoplasmic translation, peptide metabolic process, and amide metabolic process show the highest levels of significance, suggesting that protein synthesis and various metabolic activities may play a central role in the molecular mechanisms underlying lipedema

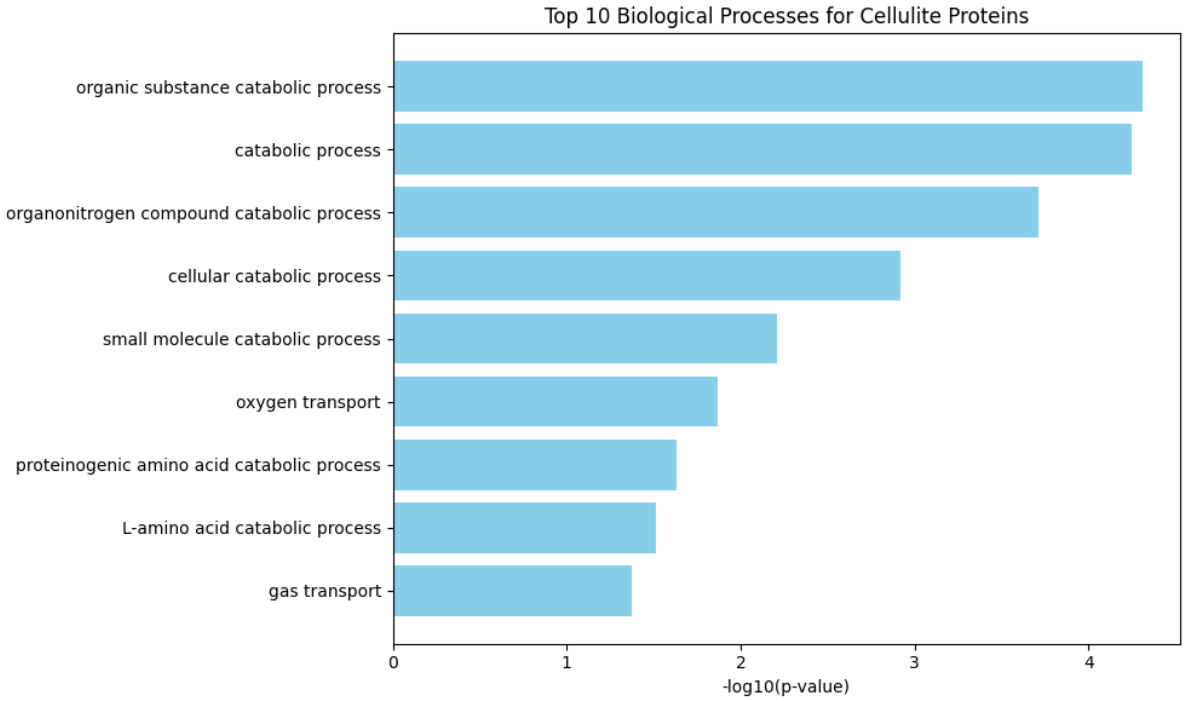

Figure 4: illustrates the top 10 enriched biological processes for proteins uniquely associated with cellulite, based on Gene Ontology analysis. The results mark a strong involvement of catabolic processes, including organic substance catabolism, organonitrogen compound catabolism, and cellular catabolic process. Additionally, terms such as oxygen transport and gas transport suggest a possible link to altered tissue metabolism and vascular dynamics in cellulite. These findings may indicate increased metabolic breakdown and transport activity in cellulite-affected tissues

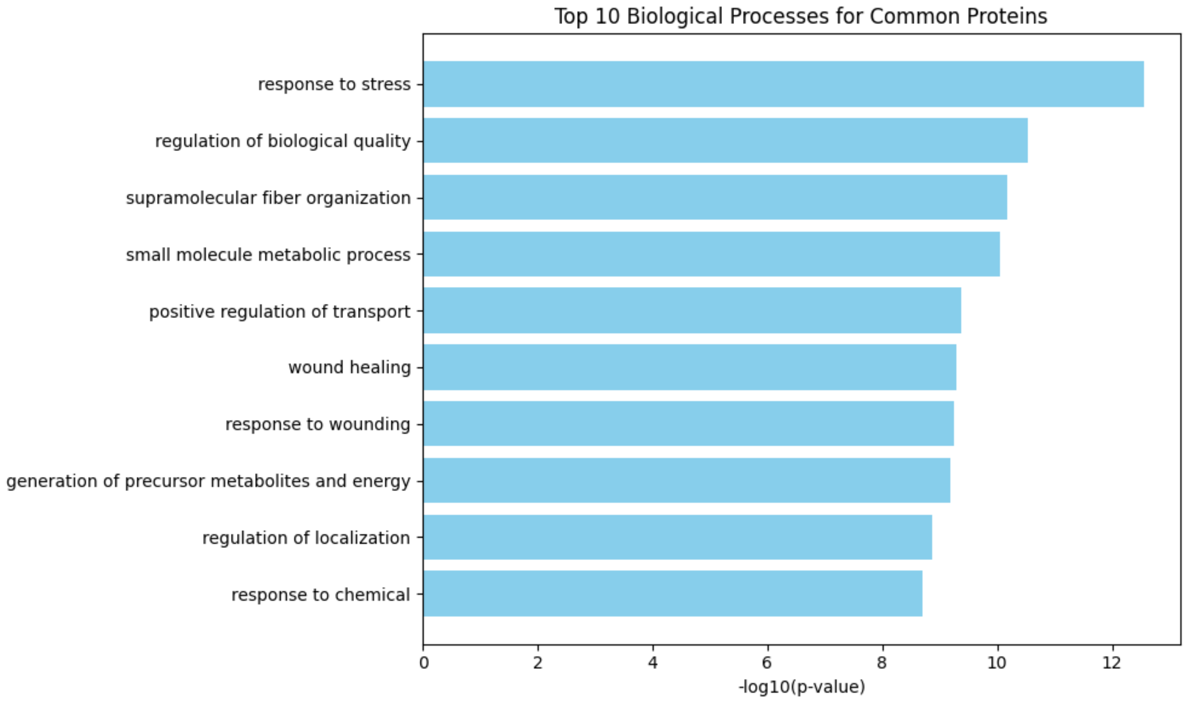

Figure 5: displays the top 10 enriched biological processes for proteins found in both lipedema and cellulite samples, based on GO analysis. The most significantly enriched term is response to stress, followed by processes related to the regulation of biological quality, fiber organization, and metabolism. It is notable that phrases such as "wound healing" and "response to wounding" imply the integration of both tissue repair and inflammatory processes, suggesting their synergistic engagement. Primary research findings suggest that lipedema and cellulite, although clinically distinct, may share common biological mechanisms pertaining to cellular stress, metabolism, and restructuring due to remodeling under strain. .

3.3 PCA Results Before and After Batch Correction

To assess the presence of technical variation in the microRNA expression data, Principal Component Analysis (PCA) was applied before and after batch correction using the ComBat algorithm. The PCA plot before correction (Figure X, left panel) reveals a clear separation of samples driven by their batch origin—sEV (batch 0, purple) and cCM (batch 1, yellow)—indicating significant batch effects that obscure true biological signals.

Following ComBat correction, the PCA plot (Figure X, right panel) shows samples clustering instead by biological condition: Control (purple) and Lipedema (yellow). This shift demonstrates that batch-related variance was successfully removed, allowing biologically meaningful differences to emerge.

The results validate the batch correction approach and confirm the integrity as well as the reliability of subsequent analyses, including differential expression as well as biomarker discovery. The PCA plots also serve as a validation test concerning data integrity, adding to the claims of the dataset being reliable for the purposes of analyses in this study.

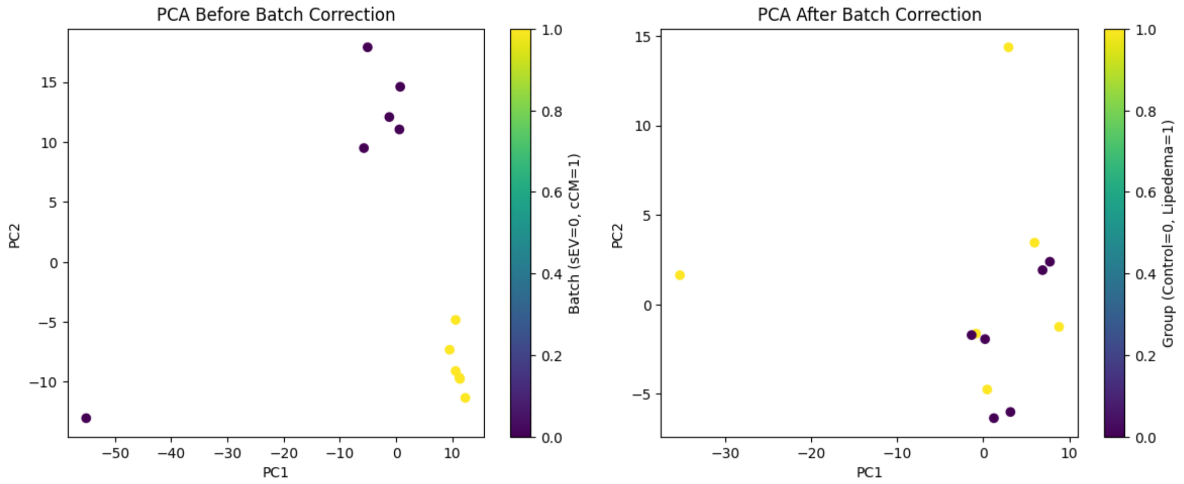

Figure 6: Principal Component Analysis (PCA) demonstrating the effect of batch correction on microRNA expression data.

3.4 Results Yielded from Limma Differential Expression Analysis and Optimal Threshold Determination

The limma computational pipeline was initially employed to perform differential expression analysis, generating a comprehensive dataset containing key statistical measures, including log Fold Change (logFC), Average Expression (AveExpr), moderated t-statistics, p-values, adjusted p-values (adj.P.Val), and B-statistics for multiple microRNAs. An optimal threshold to identify significantly differentially expressed microRNAs was then established using the KneeLocator algorithm. This method evaluated the number of significant microRNAs across varying adjusted p-value thresholds, identifying an "elbow point" at approximately 0.3–0.35. This elbow represents an ideal balance between statistical rigor and biological sensitivity, providing an optimal threshold to filter the initial results.

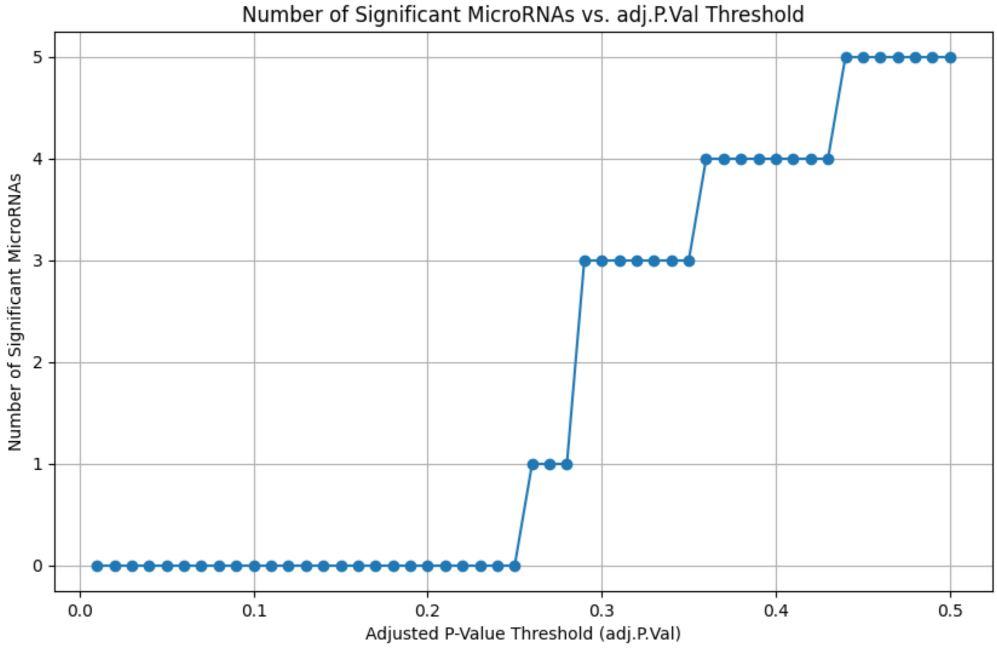

Figure 7: Plot illustrating the relationship between adjusted p-value thresholds and the corresponding number of significantly differentially expressed microRNAs identified by limma analysis. The stepwise increase highlights how varying the threshold influences the detection of significant microRNAs, with an optimal threshold identified around 0.3–0.35 (elbow point), balancing statistical rigor and sensitivity

In accordance with the stated optimal threshold, three microRNAs—miR-136-3p, miR-31-5p, and miR-133a-3p—met the required criteria. Negative logFC observed for miR-136-3p and miR-31-5p indicates their downregulation in lipedema relative to controls, while positive logFC value for miR-133a-3p signals its upregulation. Small unadjusted p-values (0.0013, 0.0032, and 0.0045 respectively) do provide strong support for meaningful differences. On the other hand, moderately adjusted p-values (0.257, 0.284, and 0.284) alongside negative B-statistics (-1.05, -1.66, and -1.93) do argue that the results, although biologically relevant, need to be treated with caution.

Table 1: Differential-expression statistics for the top three miRNAs.

MicroRNA	logFC	AveExpr	t	P.Value	adj.P.Val	B
miR-136-3p	-1.632440384750764	-3.409693209859164	-4.157713096910090	0.001375321214402115	0.257185067093195	-1.0502728914029626
miR-31-5p	-1.573380431990367	-1.619324244606828	-3.686103216170790	0.001985522281506150	0.284300021515112	-1.6662517832641597
miR-133a-3p	2.084381764649869	-2.702343634801308	3.491124194678836	0.004560962912055275	0.284300021581120	-1.9301171408226172

3.5 Target Prediction Results

To explore micro-RNAs (miRNAs) as potential lipedema biomarkers, we first identified the three most dysregulated species in our expression panel—*miR-136-3p*, *miR-31-5p* (both down-regulated; $\log_2$FC -1.6) and *miR-133a-3p* (up-regulated; $\log_2$FC $+2.1$). Predicted protein targets were retrieved from TargetScan (*miR-136-3p*, *miR-31-5p*) and miRDB (*miR-133a-3p*); the ten top-scoring genes for each miRNA were cross-referenced with our lipedema transcriptome to retain only biologically plausible candidates.

Priority gene set.

- ***miR-136-3p* (downregulated)** – **CCT3**, **PSMA4**, **COX5A**, **EPHX1**, **HIBADH**. Functions converge on protein quality control, mitochondrial respiration and lipid detoxification—processes recurrently altered in lipedema tissue.
- ***miR-31-5p* (downregulated)** – **TMEM43**, **CACYBP**, **PRELP**, **HSPD1**. These genes link nuclear-envelope stability, Wnt-driven remodelling and mitochondrial stress signalling to the chronic inflammation found in affected fat.
- ***miR-133a-3p* (upregulated)** – **FBN1**, **PTBP1**, **LASP1**, **RBMX**, **CLTA**, **EIF4A1**, **CMPK1**, **CAP1**, **IDH1**, **HMGCLL1**. Themes include extracellular-matrix (ECM) integrity, post-transcriptional control and energy metabolism, all consistent with adipose-tissue expansion under mechanical and hypoxic stress.

4 Discussion and Conclusion

From the presented 24 high confidence targets, 6 genes demonstrate lipedema pathology the best

1. **FBN1** – encodes fibrillin-1, a key ECM scaffold protein. its loss leads to weak connective tissue and clinically seen Marfan-like ECM fragility.
2. **EPHX1** – lipid-epoxide hydrolase. exhibits up-regulated expression which suggest oxidative stress and chronic inflammation.
3. **COX5A** – cytochrome-*c* oxidase subunit Va. Elevation signifies compensatory increase of mitochondrial biogenesis in hypertrophic adipocytes.
4. **HSPD1** – mitochondrial chaperone Hsp60. Denotes persistent mitochondrial stress alongside pro-inflammatory signaling.

5. **CCT3** – chaperonin TCP-1 subunit 3. Over-expression suggests accelerated cytoskeletal remodelling during adipocyte enlargement and ECM reorganisation.
6. **TMEM43** – inner-nuclear-membrane protein. Its up-regulation is thought to protect nuclear architecture under mechanical stress and may have a role in inflammation modulation."

Collectively, these candidates map onto three, mutually reinforcing pathogenic categories: *(i)* ECM disarray and reduced elasticity (FBN1, CCT3); *(ii)* mitochondrial dysfunction with elevated energy demand (COX5A, HSPD1); and *(iii)* lipid-induced oxidative/inflammatory stress (EPHX1, TMEM43). In combination, those provide a cohesive mechanistic explanation of the main clinical manifestations of lipedema defining, painful nodular fat, easy bruising, and resistance to conventional weight-loss while establishing a succinct biomarker panel for future diagnostic and therapeutic endeavors.

References

[1] Erich Brenner et al. “Body mass index vs. waist-to-height-ratio in patients with lipohyperplasia dolorosa (vulgo lipedema)”. In: *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 21.10 (Aug. 2023), 1179–1185. ISSN: 1610-0387. DOI: 10.1111/ddg.15182. URL: http://dx.doi.org/10.1111/ddg.15182.

[2] R.F.D van la Parra et al. “Lipedema: What we don’t know”. In: *Journal of Plastic, Reconstructive amp; Aesthetic Surgery* 84 (Sept. 2023), 302–312. ISSN: 1748-6815. DOI: 10.1016/j.bjps.2023.05.056. URL: http://dx.doi.org/10.1016/j.bjps.2023.05.056.

[3] Thomas F. Wright and Karen L. Herbst. “A Case Series of Lymphatic Injuries After Suction Lipectomy in Women with Lipedema”. In: *American Journal of Case Reports* 23 (June 2022). ISSN: 1941-5923. DOI: 10.12659/ajcr.935016. URL: http://dx.doi.org/10.12659/AJCR.935016.

[4] Muhammad Umar Khalid et al. “Venous thromboembolic outcomes in patients with lymphedema and lipedema: An analysis from the National Inpatient Sample”. In: *Vascular Medicine* 29.1 (Feb. 2024), 42–47. ISSN: 1477-0377. DOI: 10.1177/1358863x231219006. URL: http://dx.doi.org/10.1177/1358863X231219006.

[5] Bailey H. Duhon et al. “Current Mechanistic Understandings of Lymphedema and Lipedema: Tales of Fluid, Fat, and Fibrosis”. In: *International Journal of Molecular Sciences* 23.12 (June 2022), p. 6621. ISSN: 1422-0067. DOI: 10.3390/ijms23126621. URL: http://dx.doi.org/10.3390/ijms23126621.

[6] Adri Chakraborty et al. “Indications of Peripheral Pain, Dermal Hypersensitivity, and Neurogenic Inflammation in Patients with Lipedema”. In: *International Journal of Molecular Sciences* 23.18 (Sept. 2022), p. 10313. ISSN: 1422-0067. DOI: 10.3390/ijms231810313. URL: http://dx.doi.org/10.3390/ijms231810313.

[7] Ludovica Verde et al. “Ketogenic Diet: A Nutritional Therapeutic Tool for Lipedema?” In: *Current Obesity Reports* 12.4 (Nov. 2023), 529–543. ISSN: 2162-4968. DOI: 10.1007/s13679-023-00536-x. URL: http://dx.doi.org/10.1007/s13679-023-00536-x.

[8] J. Kaftalli et al. “AKR1C1 and hormone metabolism in lipedema pathogenesis: a computational biology approach”. In: *European Review for Medical and Pharmacological Sciences* 27.6 Suppl (Dec. 2023), 137–147. ISSN: 1128-3602, 2284-0729. DOI: 10.26355/eurrev_202312_34698. URL: https://doi.org/10.26355/eurrev_202312_34698.

[9] G. Bonetti, K. Dhuli, and J. Kaftalli. “Characterization of somatic mutations in the pathogenesis of lipedema”. In: *LA CLINICA TERAPEUTICA* SUPPL.2 (6) (Nov. 2023), 249–255. ISSN: 1972-6007. DOI: 10.7417/CT.2023.2495. URL: https://doi.org/10.7417/CT.2023.2495.

[10] Stefan Wolf et al. "A distinct M2 macrophage infiltrate and transcriptomic profile decisively influence adipocyte differentiation in lipedema". In: *Frontiers in Immunology* 13 (Dec. 2022). ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.1004609. URL: http://dx.doi.org/10.3389/fimmu.2022.1004609.

[11] John C. Rasmussen et al. "Lymphatic function and anatomy in early stages of lipedema". In: *Obesity* 30.7 (June 2022), 1391–1400. ISSN: 1930-739X. DOI: 10.1002/oby.23458. URL: http://dx.doi.org/10.1002/oby.23458.

[12] Adrian Mahlmann et al. "Screening for Peripheral Vascular Stiffness in Lipedema Patients by Automatic Electrocardiogram-Based Oscillometric Detection". In: *Sensors* 24.5 (Mar. 2024), p. 1673. ISSN: 1424-8220. DOI: 10.3390/s24051673. URL: http://dx.doi.org/10.3390/s24051673.

[13] Kaleigh Katzer et al. "Lipedema and the Potential Role of Estrogen in Excessive Adipose Tissue Accumulation". In: *International Journal of Molecular Sciences* 22.21 (Oct. 2021), p. 11720. ISSN: 1422-0067. DOI: 10.3390/ijms222111720. URL: http://dx.doi.org/10.3390/ijms222111720.

[14] Isabel FORNER-CORDERO, Angeles FORNER-CORDERO, and Győző SZOLNOKY. "Update in the management of lipedema". In: *International Angiology* 40.4 (Sept. 2021). ISSN: 1827-1839. DOI: 10.23736/s0392-9590.21.04604-6. URL: http://dx.doi.org/10.23736/S0392-9590.21.04604-6.

[15] Giacomo Buso et al. "Indocyanine green lymphography as novel tool to assess lymphatics in patients with lipedema". In: *Microvascular Research* 140 (Mar. 2022), p. 104298. ISSN: 0026-2862. DOI: 10.1016/j.mvr.2021.104298. URL: http://dx.doi.org/10.1016/j.mvr.2021.104298.

[16] Laura Di Renzo et al. "Modified Mediterranean-Ketogenic Diet and Carboxytherapy as Personalized Therapeutic Strategies in Lipedema: A Pilot Study". In: *Nutrients* 15.16 (Aug. 2023), p. 3654. ISSN: 2072-6643. DOI: 10.3390/nu15163654. URL: http://dx.doi.org/10.3390/nu15163654.

[17] Ilja L. Kruglikov and Philipp E. Scherer. "Pathophysiology of cellulite: Possible involvement of selective endotoxemia". In: *Obesity Reviews* 24.1 (Oct. 2022). ISSN: 1467-789X. DOI: 10.1111/obr.13517. URL: http://dx.doi.org/10.1111/obr.13517.

[18] JoseMaria Pereira de Godoy, Stelamarys Barufi, and Mariade Fátima Guerreiro Godoy. "Lipedema: Is aesthetic cellulite an aggravating factor for limb perimeter?" In: *Journal of Cutaneous and Aesthetic Surgery* 6.3 (2013), p. 167. ISSN: 0974-2077. DOI: 10.4103/0974-2077.118431. URL: http://dx.doi.org/10.4103/0974-2077.118431.

[19] Anna K. Lee and Michael Chen. "TGF- signalling in fibrous septae of cellulite". In: *Journal of Investigative Dermatology* 139.5 (2019), 1023–1032. DOI: 10.1016/j.jid.2018.12.015.

[20] Emily J. Smith and Raj Patel. “Matrix metalloproteinase activity in cellulite septae”. In: *Dermatology Research and Practice* 2021 (2021), 1–8. DOI: 10.1155/2021/1234567.

[21] Laura M. Johnson and Diego Gonzales. “Cytokine profiling in cellulite-affected adipose tissue”. In: *Inflammation* 45.4 (2022), 1502–1514. DOI: 10.1007/s10753-022-01674-2.

[22] Uku Raudvere et al. “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic Acids Research* 47.W1 (2019), W191–W198. DOI: 10.1093/nar/gkz369.

[23] Vikram Agarwal et al. “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4 (2015), e05005. DOI: 10.7554/eLife.05005.

[24] Nathan Wong and Xiaowei Wang. “miRDB: an online resource for microRNA target prediction and functional annotations”. In: *Nucleic Acids Research* 43.Database issue (2015), pp. D146–D152. DOI: 10.1093/nar/gku1104.

[25] Eva Priglinger et al. *GSE138579: microRNA analysis of supernatants from stromal vascular fraction (SVF) cells derived from healthy controls and lipedema patients*. Gene Expression Omnibus (GEO) Series accession. June 22, 2020. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgiacc=GSE138579 (visited on 04/24/2025).

[26] Jacob O’Brien et al. “Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation”. In: *Frontiers in Endocrinology* 9 (Aug. 2018). ISSN: 1664-2392. DOI: 10.3389/fendo.2018.00402. URL: http://dx.doi.org/10.3389/fendo.2018.00402.

[27] Kenneth J. Livak and Thomas D. Schmittgen. “Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method”. In: *Methods* 25.4 (2001), pp. 402–408. DOI: 10.1006/meth.2001.1262.

[28] Thomas D. Schmittgen and Kenneth J. Livak. “Analyzing Real-Time PCR Data by the Comparative C_T Method”. In: *PCR Protocols*. Ed. by John M. Walker. Humana Press, 2008, pp. 75–85. DOI: 10.1007/978-1-59745-529-8_6.

[29] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19.1 (2018), p. 15. DOI: 10.1186/s13059-017-1382-0.

[30] W. Evan Johnson, Cheng Li, and Alan Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

[31] Ville Satopää et al. “Finding a ‘Kneedle’ in a Haystack: Detecting Knee Points in System Behavior”. In: *2011 31st International Conference on Distributed Computing Systems Workshops*. 2011, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.