DEPARTMENT OF
COMPUTER SCIENCE

DAVID JOÃO COSTA CASTRO

BSc in Computer Science and Engineering

AUTOMATED NON-TAGGED PROTEIN PURIFICATION PROTOCOLS

Dissertation Plan
MASTER IN COMPUTER SCIENCE AND ENGINEERING
SPECIALIZATION IN ARTIFICIAL INTELLIGENCE

NOVA University Lisbon
Draft: *January 23, 2026*

DEPARTMENT OF
COMPUTER SCIENCE

AUTOMATED NON-TAGGED PROTEIN PURIFICATION PROTOCOLS

DAVID JOÃO COSTA CASTRO

BSc in Computer Science and Engineering

Adviser: André Lamúrias
Associate Professor, NOVA University Lisbon

Co-adviser: Arménio Barbosa
Associate Professor, NOVA University Lisbon

Dissertation Plan
MASTER IN COMPUTER SCIENCE AND ENGINEERING
SPECIALIZATION IN ARTIFICIAL INTELLIGENCE

NOVA University Lisbon
Draft: *January 23, 2026*

Abstract

Protein purification is vital for protein biology studies, yet optimizing these purification methods can be time-consuming because of variations in the techniques and protocol steps necessary for each new protein that have to be determined experimentally through trial and error. Previous works have focused on creating a database of protein purification conditions and using LLMs to extract relevant information from articles. The development of Protein Language Models. In recent years, several Protein Language Models have been proposed which could also be used to learn useful representations of proteins.

This project proposes an approach to predict a purification protocol for new proteins, based on information extracted from the literature. By combining chemical properties of proteins with information extracted from papers describing purification protocols of those same proteins, we aim to train a model that can do this task efficiently in order to reduce the effort necessary to purify new proteins.

Keywords: One keyword · Another keyword · Yet another keyword · One keyword more · The last keyword

Resumo

Palavras-chave: Primeira palavra-chave · Outra palavra-chave · Mais uma palavra-chave · A última palavra-chave

Contents

Appendices

Annexes

List of Figures

List of Tables

1

Introduction

1.1 Historical context

The history of protein purification is intrinsically linked to our understanding of life at the molecular level. This journey began in 1789 when Antoine Fourcroy first distinguished several types of complex organic substances, which he categorized as "albumins," including fibrin, gelatin, and gluten. Although these substances were not yet recognized as proteins, their consistent presence in biological processes made them a primary focus for early chemists. The identification of the building blocks of these substances was a slow process; while asparagine was the first amino acid isolated in 1809, its role as a fundamental constituent of proteins was not fully established until 1873. A critical link was formed earlier, in 1819, with the isolation of leucine, which helped researchers begin to understand the chemical nature of these "albuminous" materials.

By 1837, Gerrit J. Mulder determined the elemental composition of several proteins and proposed that they shared a common core substance. In response to these findings, Jacob Berzelius suggested the name "protein" in 1838, derived from the Greek word *proteios*, meaning "primary" or "of the first rank." Despite this naming, the chemical diversity of proteins remained largely unknown; at the time, only glycine and leucine had been identified. It would take nearly another century, until the discovery of threonine in 1936, for the complete set of 20 standard amino acids to be recognized.

A defining moment in the field occurred in 1926, amidst a heated debate over whether enzymes were distinct chemical entities or simply "catalytic forces" associated with proteins. James Sumner settled this by isolating and crystallizing the enzyme urease from jack beans. This achievement provided the first definitive proof that enzymes were proteins with specific, defined chemical structures that could be purified to homogeneity. Sumner's work, which earned him the Nobel Prize in 1946, effectively birthed the field of structural biochemistry and established purification as a prerequisite for understanding protein function.

In the decades following Sumner's breakthrough, the field saw the development of diverse biophysical techniques designed to separate proteins based on their intrinsic properties, such as electrical charge, molecular size, and polarity. These methods—including various forms of chromatography and electrophoresis—became the standard toolkit for biochemists. The landscape of protein science changed again in 1973, when Stanley Cohen and Herbert Boyer developed recombinant DNA technology. This allowed scientists to insert specific DNA sequences into host organisms like *E. coli*, turning bacteria into "factories" for the mass production of specific proteins.

While recombinant technology solved the problem of protein "sourcing," it introduced new challenges for purification. In the 1980s, the development of affinity tags (such as the polyhistidine tag or GST-tag) revolutionized the field by allowing researchers to add a universal "handle" to any recombinant protein. This made purification significantly easier and more predictable. However, these tags can often interfere with the protein's native folding, biological activity, or its suitability for therapeutic use in humans.

Consequently, the purification of "non-tagged" proteins remains the gold standard for many high-precision applications. Because every non-tagged protein has a unique combination of surface charges and hydrophobic patches, designing an effective purification protocol remains a labor-intensive process of trial and error. This historical difficulty is the primary driver for the current research, as we seek to automate the design of these complex protocols through computational modeling. We have not yet discarded the possiblity of increasing the scope of our project to include tagged proteins, however it is currently hung up on analysis of our preliminary results.

1.2 Motivation

Expand on why we are doing this/its important.

Designing a protein purification protocol is currently a time and resource consuming process defined by significant trial and error. When working with a protein that lacks an established protocol, scientist must often commit substantial time and financial resources to discover an effective sequence of techniques. Our work aims to bridge this gap by creating a system that predicts these protocols using only basic biochemical information.

1.3 Goal

Expand and give some details on what the goal of the final "product" will be and the success metrics.

The goal is to produce a tool that translates raw protein sequence data into a laboratory-ready recipe. By providing this automated starting point, we hope to significantly reduce the manual effort, time and cost currently required to purify new proteins.

2

Related Work

3

Work Plan

This chapter aims at defining the scope of our work, what our goal is, the plan to execute it and what has already been done, along with its preliminary results.

3.1 Overview

At this stage, we have developed a data extraction tool to build the necessary foundation for such a system. This tool leverages the Protein Data Bank (PDB) to identify proteins with known 3D structures and then targets their associated scientific literature. We operate on the logic that a protein must be successfully purified before its structure can be experimentally determined, so these papers represent a reliable source of proven purification methodologies.

With this data, we plan to train a predictive model, using a Transformer-based architecture due to its strength in generating sequential instructions. We assume that this model will need to be part of a larger algorithm that manages data inputs and ensures the final output is technically consistent. These later stages remain theoretical and will be discussed further in the future work plane section.

3.2 Work Done

The development of a robust predictive model is fundamentally dependent on the quality and volume of the training data. For this project, the primary objective is to correlate the physico-chemical properties of a protein with its optimal purification strategy. Because no centralized database currently exists that maps these biochemical attributes to specific experimental protocols, a significant portion of the initial work focused on the design and implementation of an automated data mining pipeline.

The pipeline was engineered to identify, retrieve, and process scientific literature to build a structured dataset. This process was divided into four distinct phases:

1. *Source Identification:* Determining from where we could source protein purification protocols.

2. *Entity Linking:* Mapping specific protein sequences and structures to the papers that describe their purification.

3. *Full-text Retrieval:* Automating the retrieval of the complete text of identified papers.

4. *Information Extraction:* Processing the unstructured text to isolate chromatography steps and related biochemical metadata.

3.2.1 Determining sources of purification protocol

In order to get a high volume of quality training data, we would need a source of correctly documented protein purification protocols and the proteins they are being applied to. Given that there is no centralized database with such information, and selecting by hand would prove unfeasible, the first step was to find a solution to this problem. In the context of biochemistry, the most comprehensive descriptions of protein purification processes are found within peer-reviewed scientific papers. However, the vast volume of published literature demands a systematic method for identifying relevant documents, as manual selection is not scalable for a dataset of the size required for deep learning.

To solve this, I utilized the PDB[1] as the primary gateway for data discovery. Each entry in the PDB is typically associated with a primary citation, which is the scientific paper detailing the methods used to determine that specific structure.

The logical basis for this approach is that the determination of a protein's structure requires a highly purified sample. Consequently, the primary citation for a PDB entry almost universally includes a methodology section describing the purification protocol used to reach the required level of purity. At the time of this research, the PDB contained 247,417 entries. While a subset of these entries may lack an associated paper or detailed methodology, the sheer scale of the database provides a sufficient foundation for training a predictive model.

To extract this information programmatically, I interfaced with the PDB API. This allowed for the automated querying of specific metadata fields for every entry in the database. For the current phase, the relevant fields were PDB ID, UniProt ID and the associated paper's bibliographic metadata, such as the DOI, PubMed ID and title.

The UniProt ID is particularly important because it allows us to link the 3D structure in the PDB to the protein sequence it corresponds to. Not only that, since UniProt[2] is cross-referenced with 185 other databases, from genomics to biochemistry, biology and chemistry, it allows us to build a complete protein profile, which will be very useful later on.

[1] https://www.rcsb.org/

[2] https://www.uniprot.org/

3.2.2 Mapping specific protein sequences and structures to the papers that describe their purification

To gather the most comprehensive metadata possible, my first priority was to establish a reliable link between PDB entries (3D structures) and their corresponding protein's UniProt entry (proteins). While both databases provide APIs that allow for programmatic queries, developing a custom mapping tool from scratch proved to be more complex than initially anticipated.

The primary challenge was the sheer volume of data. With nearly 250,000 entries in the PDB, each potentially mapping to multiple protein chains and UniProt IDs, the number of required requests was massive. Both the PDB and UniProt APIs enforce strict rate limits to ensure server stability. Considering this, we selected the 200 first entries we get from PDB's API to run our tests on.

In fig. 3.1 we have a diagram illustrating our initial setup for the data extraction. We start by fetching PDB entries by querying its API. Most PDB entries link to one or more UniProt IDs, which identify the specific proteins contained in that structure, so we use those IDs to query UniProt's API. With UniProt's data we are able to complete our ID mapping, leaving just the full-text of the paper corresponding to that protein missing. This is primarily obtained through the PubMed Central's ID present in the PDB, which allows us to query Europe PMC's[3] API for the full-text of the respective paper. Finally, we would need to apply text mining logic to the resulting XMLs to extract the purification processes. This last part was not yet flushed out by this point.

My initial tests showed that attempting to complete sync both databases would take an impractical amount of time—potentially days of continuous running—just to establish a baseline mapping. Furthermore, handling the edge cases where one PDB structure corresponds to multiple distinct proteins added significant logic overhead to the scripts.

Recognizing that this manual mapping was becoming a bottleneck, I sought a more efficient alternative. This led me to the European Bioinformatics Institute (EBI) public file server[4]. EBI provides precomputed mapping files that are updated weekly, specifically designed to correlate PDB and UniProt entries, along other cross-references. In fig. 3.2 we have the update diagram representing this change.

Having this simplified mapping allowed us to create better visualizations of the data, such as the distribution of PDB and UniProt IDs that we see in fig. 3.3.

3.2.3 Automating the retrieval of the complete text of identified papers

Once the relevant scientific citations were identified and linked to their respective proteins, the next objective was to acquire the full text of these papers. This is a critical step because the specific details of a purification protocol, such as the chromatography steps

[3] https://europepmc.org/

[4] https://ftp.ebi.ac.uk/

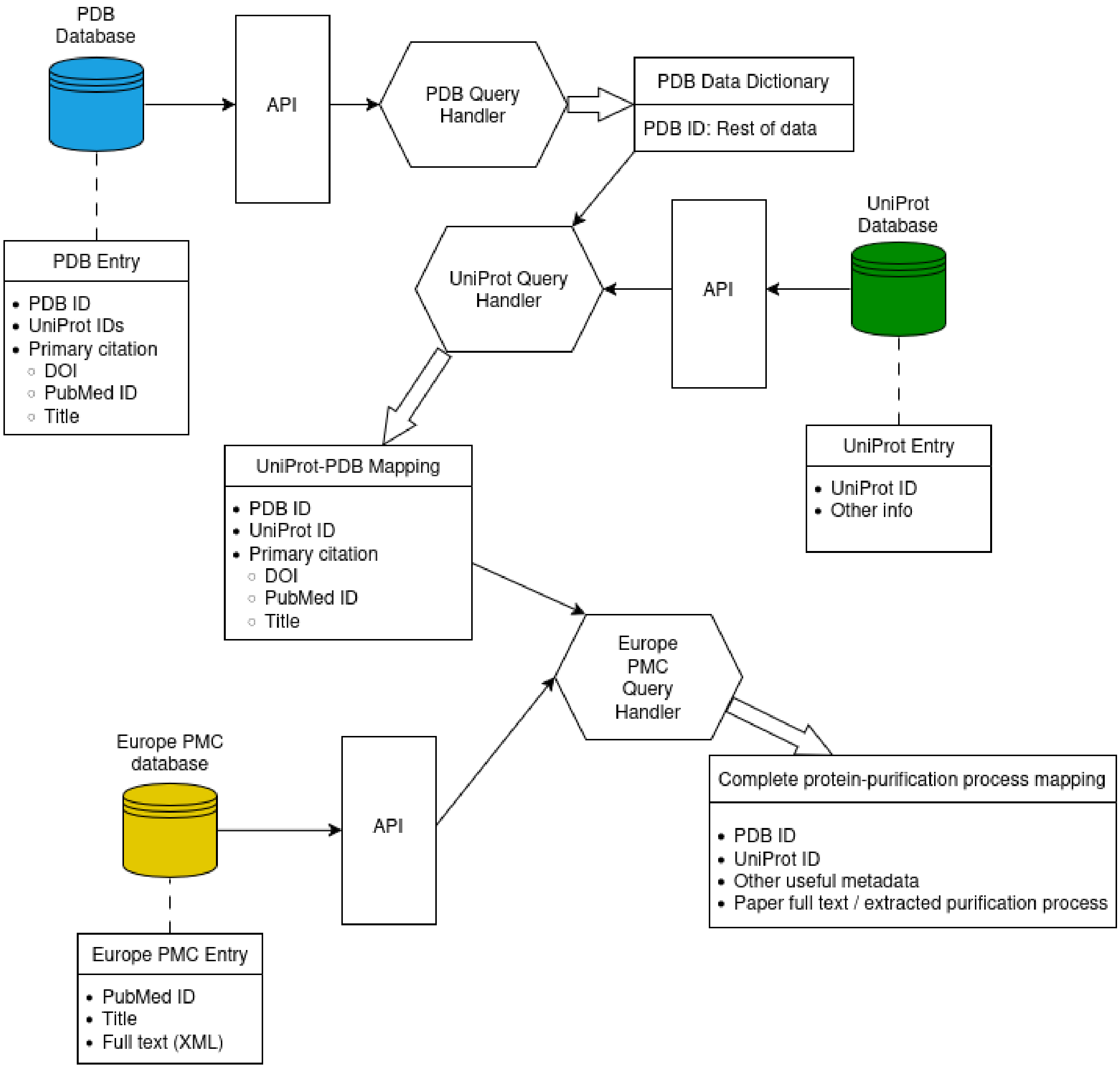


Figure 3.1: Diagram of old method

and their order, are almost exclusively contained within the "Materials and Methods" or "Experimental Procedures" sections of a full manuscript, rather than in the abstract.

After evaluating several biological literature repositories, I concluded that the Europe PMC REST API offered the most robust solution for automated full-text acquisition. Europe PMC is particularly advantageous because it provides a centralized access point for a vast collection of scientific literature and offers a dedicated endpoint for retrieving papers in a machine-readable XML format.

While the PDB provides both DOIs and PMIDs, the Europe PMC API is most efficient when queried using the PMID. Consequently, I implemented a preprocessing step to ensure every entry had a valid PMID. In cases where only a DOI was available, I utilized the NCBI ID Converter API to programmatically resolve the DOI into its corresponding PubMed identifier. With a clean list of PMIDs, the pipeline was then able to systematically request the full-text XML for each record.

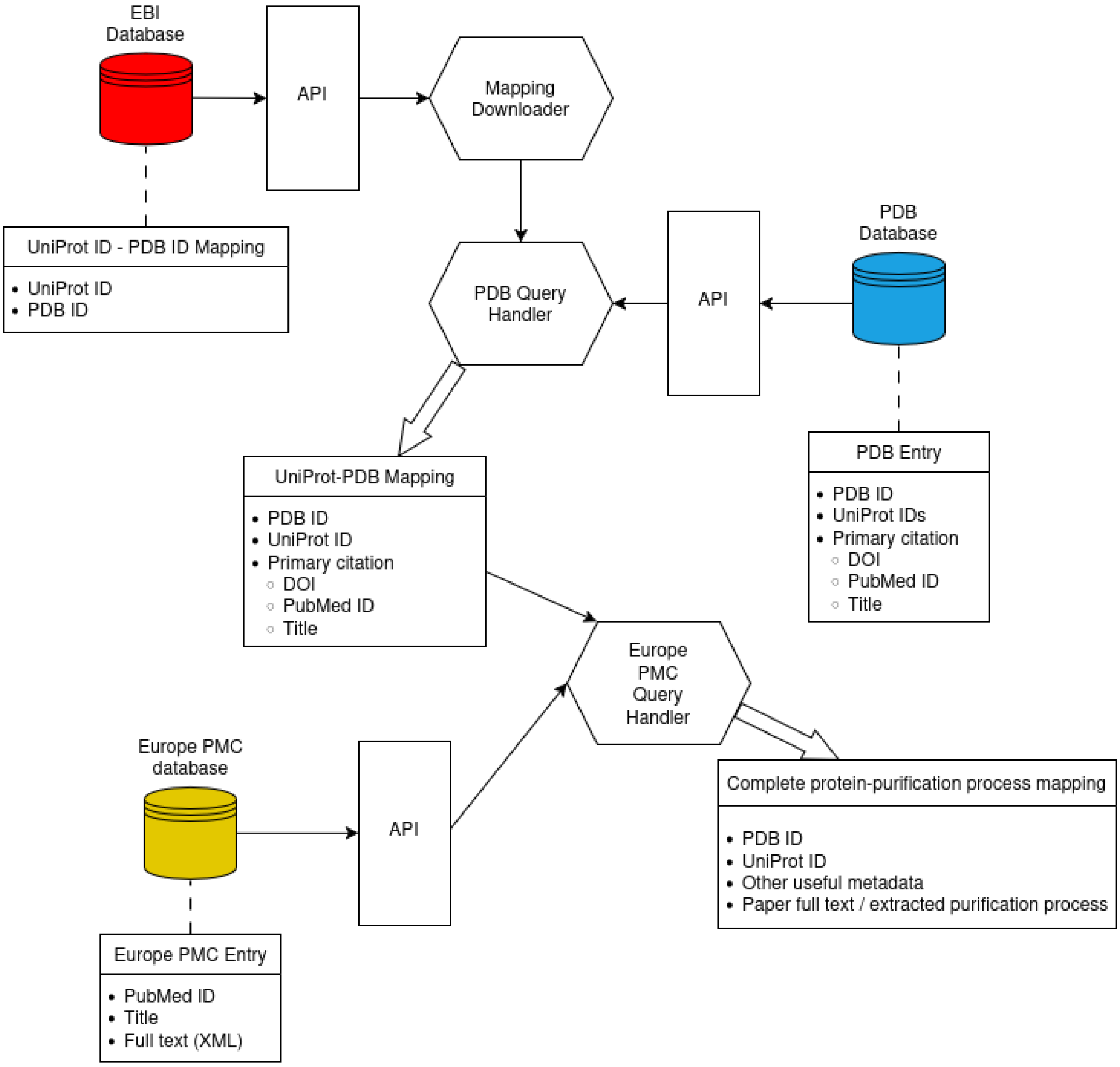
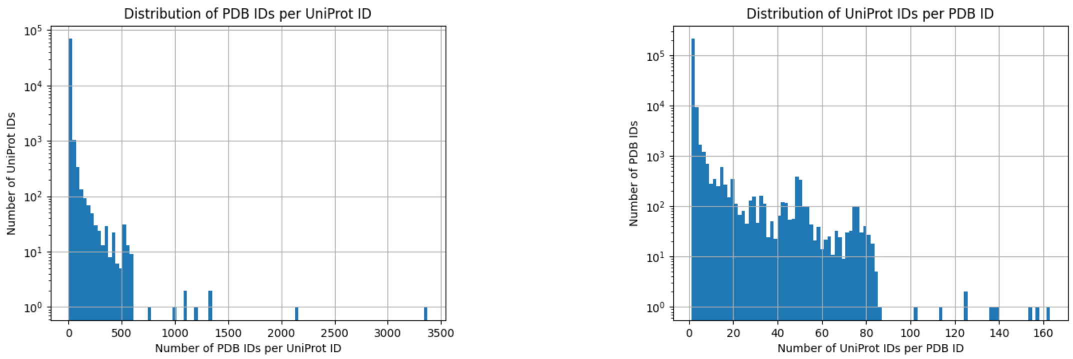


Figure 3.2: Diagram of new method



(a) Number of PDB IDs per UniProt ID

(b) Number of UniProt IDs per PDB ID

Figure 3.3: PDB/UniProt ID distribution

After evaluating several biological literature repositories, I concluded that the Europe PMC REST API offered the most robust solution for automated full-text retrieval. Europe PMC is particularly advantageous because it provides a centralized access point for a vast collection of scientific literature and offers a dedicated endpoint for retrieving papers in an easily parsable XML format.

One technical challenge encountered during this phase was the inconsistency of the available metadata. While most entries are complete, a significant number of records lack a PubMed ID or DOI, providing only a publication title. This demanded the development of a flexible retrieval strategy that could fall back on title-based searches when unique digital identifiers were unavailable, ensuring that the maximum amount of relevant literature could be captured for the next stage of the pipeline. The following issues were identified:

- **Data Redundancy:** Frequently, multiple PDB entries (representing different structural configurations or mutants of the same protein) reference the exact same primary citation. While not an error, the pipeline had to be optimized to recognize these duplicates to avoid redundant API calls and unnecessary storage use.

- **The Open Access Barrier:** The most significant hurdle is that not all papers are available in the Europe PMC Open Access subset. Many papers remain behind paywalls, meaning the API can only return the abstract or metadata rather than the complete paper.

- **Missing Identifiers:** For some older or more obscure PDB entries, neither a DOI nor a PMID is recorded. Without these unique identifiers, automated retrieval becomes significantly more difficult, often requiring title-based fuzzy matching which is less reliable.

- **Inconsistent Availability:** In some instances, a record might exist in the database, but the full-text version has not been deposited or processed into the XML format required by our extraction tools.

3.2.4 Processing the unstructured text to isolate chromatography steps and related biochemical metadata

The final and most complex phase of the pipeline involves transforming the retrieved full-text papers into a structured sequence of purification steps. Having the papers in XML format, as provided by the Europe PMC API, proved to be a significant advantage. Unlike PDFs, which are notoriously difficult to parse due to inconsistent layouts, XML documents use standardized tags to identify specific sections, titles, and paragraphs. This structure allowed me to programmatically navigate the document and isolate the most relevant portions of the text.

In the early stages of development, I needed a simple and reliable way to identify where the purification process was described within a paper. After analyzing the structure of around 20 scientific papers, I found that two primary indicators were highly effective:

1. Searching for the keyword "Purification" within subsection titles (e.g., *<title>Protein expression and purification</title>*).
2. Identifying paragraphs in the main body that contained the term "chromatography."

While this approach was successful for locating the general area of interest during preliminary tests, it was not sufficient for our ultimate goal. The specific steps, the tools being used, their sizes, concentrations and other parameters and their order are extremely important to systemically define each purification protocol, so we had to refine our approach.

For this, I leveraged the fact that protein purification is a specialized field with a relatively finite set of techniques. Most protocols follow a logical progression using a limited number of standard methods, such as Affinity Chromatography, Ion Exchange, or Size Exclusion.

By recognizing this pattern, I developed a comprehensive dictionary of chromatography terms, tools, and specific biochemical markers (such as "His-tag," "IMAC," "gradient elution," or "superdex"). The underlying logic is that these technical terms are highly specific; they are rarely mentioned in biological literature outside the context of an actual purification protocol. In fig. 3.4 we can see an example of the keywords' hierarchy for the "Size Exclusion" technique category.

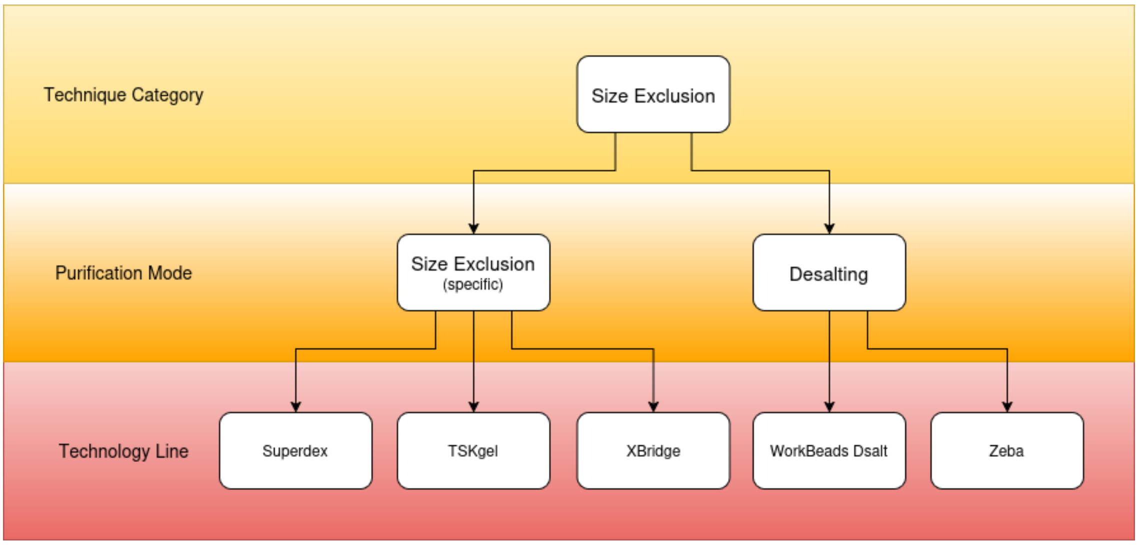


Figure 3.4: Chromatography techniques hierarchy

The current version of the tool scans the identified "Purification" sections and extracts these terms in the order they appear. By capturing this sequence, the pipeline theoretically reconstructs the "recipe" used in the laboratory. For example, if the tool detects "Affinity

Chromatography" followed by "Dialysis" and then "Gel Filtration," it records these as three distinct chronological steps.

This methodology is currently a work in progress. While the dictionary-based approach provides a structured way to handle unstructured text, it is not yet perfect. Scientific writing can be nuanced, and the tool must be able to distinguish between a technique that was actually performed and one that is merely being discussed or referenced.

At this stage, I have not yet produced definitive preliminary results from this extraction phase. It remains an iterative process, and I expect to refine the dictionary and the extraction logic as I begin to validate the output against known manual protocols.

3.3 Future Work Plan

4

Adding Support to a New School (work in progress)

Bibliography

[1] R. J. Dias et al. “Verification of Snapshot Isolation in Transactional Memory Java Programs”. In: *Proceedings of the 26th European conference on Object-oriented programming (ECOOP’12)*. Springer-Verlag, 2012-06 (cit. on p. 15).

A

NOVAthesis covers showcase

B

Appendix 2 Lorem Ipsum

This is a test with citing something [1] in the appendix.

I

Annex 1 Lorem Ipsum