NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

DAVID JOÃO COSTA CASTRO

BSc in Computer Science and Engineering

PREDICTION OF PROTEIN PURIFICATION PROTOCOLS

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

NOVA University Lisbon
February, 2026

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

PREDICTION OF PROTEIN PURIFICATION PROTOCOLS

DAVID JOÃO COSTA CASTRO

BSc in Computer Science and Engineering

Adviser: André Lamúrias
Assistant Professor, NOVA University Lisbon

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

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

NOVA University Lisbon
February, 2026

Abstract

Protein purification is vital for protein biology studies and biopharmaceuticals production. 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.

This project proposes an approach to predict purification protocols for new proteins, based on information extracted from peer-reviewed literature. By combining physico-chemical properties of proteins with information extracted from papers describing purification protocols of those same proteins, we aim to train a model that can efficiently predict a sequence of purification steps, reducing the effort necessary to purify new proteins.

The first step of this project is to create a data extraction tool that allows us to compile a database of proteins, their physico-chemical properties and their purification processes. An initial prototype has been developed in this phase to attempt to extract purification protocols from scientific literature. The second step is using the data we extracted to train a model to predict the purification process of a given protein. We discuss both of these steps in this work, with their success measured by the efficiency and accuracy of the predictions. By achieving our goal, we will be able to lower costs, chemical waste and time consumed in the discovery of added value proteins.

Keywords: Information Extraction · Natural Language Processing · Cheminformatics · Protein Purification

Resumo

A purificação de proteínas é vital para estudos de biologia proteica e produção de biofármacos. No entanto, otimizar esses métodos de purificação pode ser demorado devido às variações nas técnicas e etapas do protocolo necessárias para cada nova proteína, que precisam ser determinadas experimentalmente por meio de tentativa e erro.

Este projeto propõe uma abordagem para prever um protocolo de purificação para novas proteínas, com base em informações extraídas de literatura revista por pares. Ao combinar as propriedades físico-químicas das proteínas com informações extraídas de artigos que descrevem protocolos de purificação dessas mesmas proteínas, pretendemos treinar um modelo que possa prever com eficiência uma sequência de etapas de purificação, reduzindo o esforço necessário para purificar novas proteínas.

A primeira etapa deste projeto é criar uma ferramenta de extração de dados que nos permita compilar uma base de dados de proteínas, das suas propriedades físico-químicas e dos seus processos de purificação. Um protótipo inicial foi desenvolvido nesta fase para tentar extrair protocolos de purificação da literatura científica. A segunda etapa é usar os dados que extraímos para treinar um modelo para prever o processo de purificação de uma determinada proteína. Discutimos ambas as etapas neste trabalho, com o seu sucesso medido pela eficiência e exatidão das previsões. Ao atingir o nosso objetivo, seríamos capazes de reduzir custos, resíduos químicos e tempo consumido na descoberta de proteínas de valor agregado.

Palavras-chave: Extração de informação · Processamento de Linguagem Natural · Quimioinformática · Purificação de Proteínas

Contents

List of Figures

1

Introduction

1.1 Protein Purification

Proteins are essential molecules that perform a vast range of critical functions in all living organisms. The ability to isolate these molecules is a fundamental requirement for progress in many scientific and industrial fields, including medical research and the development of new biopharmaceuticals [1]. Because proteins naturally exist within complex biological environments, they must be separated from other cellular components before they can be studied or used effectively, therefore, obtaining a pure sample is a vital step in biotechnology. The speed and success of many scientific advancements depend on the efficiency of this isolation process.

Protein purification is the process of isolating a specific protein of interest from a complex biological mixture, such as a cell lysate or a tissue sample. The objective is to remove all non-protein contaminants and other undesirable proteins while maintaining the biological activity and structural integrity of the target molecule.

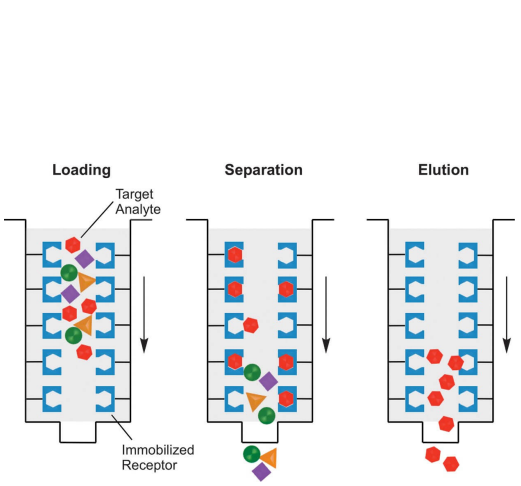
The primary challenge in this field lies in the immense diversity of proteins. Every protein possesses a unique combination of physico-chemical properties, including molecular weight, net charge, surface hydrophobicity, and specific binding affinities. Because these characteristics vary significantly even between similar proteins, there is no universal one-size-fits-all protocol. For proteins that do not yet have an established protocol, researchers must rely on a labor-intensive trial and error approach. This involves testing various experimental conditions and chemical buffers, which is both time-consuming and resource-heavy, often becoming a bottleneck in biochemical research.

A common strategy used to simplify this process is the use of affinity tags. Affinity tags are short amino acid sequences or proteins genetically fused to the target molecule that act as standardized "handles", providing a predictable way to bind the protein to a purification resin, allowing for efficient separation based on the tag's known chemical affinity. This does however come at a price, since tags can lead to interference with a protein's function [2].

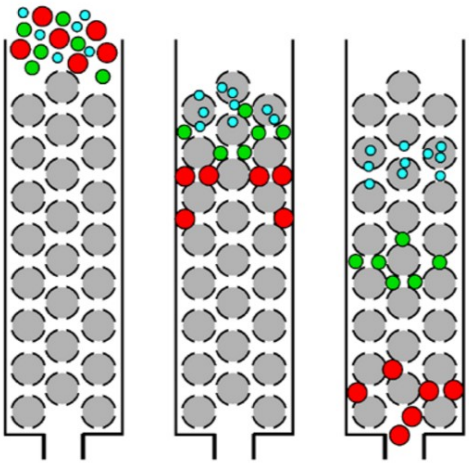
To achieve high levels of purity, laboratory workflows rely almost exclusively on

liquid chromatography. Chromatography techniques function by passing a mobile phase containing the protein mixture through a stationary phase. The components of the mixture are separated based on how they interact with the stationary phase (see Figure 1.1):

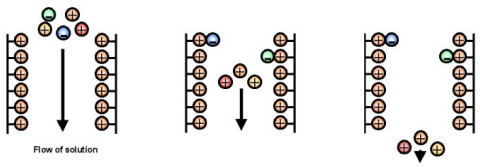
- **Affinity:** Based on specific biological interactions between the protein and a ligand.
- **Size Exclusion:** Based on the physical dimensions and shape of the molecule.
- **Ion Exchange:** Based on the net surface charge of the protein.
- **Hydrophobic Interaction:** Based on the distribution of non-polar groups on the protein surface.



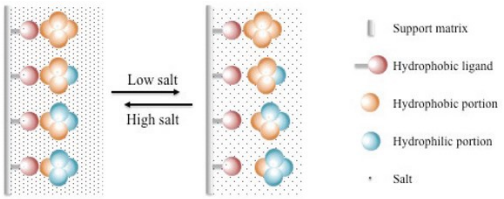
(a) Affinity [3]



(b) Size Exclusion [4]



(c) Ion Exchange [5]



(d) Hydrophobic Interaction [6]

Figure 1.1: Visualizations for the Different Types of Chromatography

In practice, a single step is rarely sufficient. A complete purification protocol is a chronological sequence of these techniques, organized to progressively refine the sample. These multi-step sequences form the purification protocols that this project aims to be able to predict.

1.2 Motivation

Despite more than a century of methodological advances, protein purification remains a major bottleneck in both basic research and industrial biotechnology [1]. While expression

systems and analytical techniques have become increasingly standardized, the design of purification protocols (particularly for non-tagged proteins) continues to rely heavily on empirical optimization. In practice, this often involves iterative testing of chromatography media, buffer compositions, and elution conditions, guided primarily by expert intuition and prior experience rather than formalized predictive principles.

This trial-and-error paradigm has several limitations. It is time-consuming, costly in terms of reagents and labor, and poorly scalable when applied to large numbers of proteins, such as those emerging from modern genomics and structural biology initiatives, like *AlphaFold* [7], a deep learning model that predicts protein structures.

Recent advances in machine learning, particularly sequence-based modeling and natural language processing, provide an opportunity to address this gap. Large public repositories such as the Protein Data Bank [8] implicitly encode decades of successful purification efforts, while biomedical literature contains detailed experimental protocols that are now accessible in machine-readable form. We assume that by leveraging these resources we can accelerate the discovery of protein purification protocols using deep learning tools and techniques.

1.3 Goal

The primary goal of this dissertation is to develop a computational system capable of predicting protein purification protocols directly from protein sequence data and derived physico-chemical properties into an ordered, laboratory-ready purification recipe composed of chromatography steps and associated techniques, reflecting strategies that have been validated in prior experimental work.

To achieve this, the project aims to integrate large-scale data mining from structural databases and the biomedical literature with a Transformer-based model architecture designed for sequential purification protocol prediction. The expected output is not a single optimized protocol, but a plausible and informative starting strategy that can guide experimental design and reduce the search space explored during laboratory optimization.

2

Related Work

2.1 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 [9]. 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 (Figure 2.1) 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 [10], 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

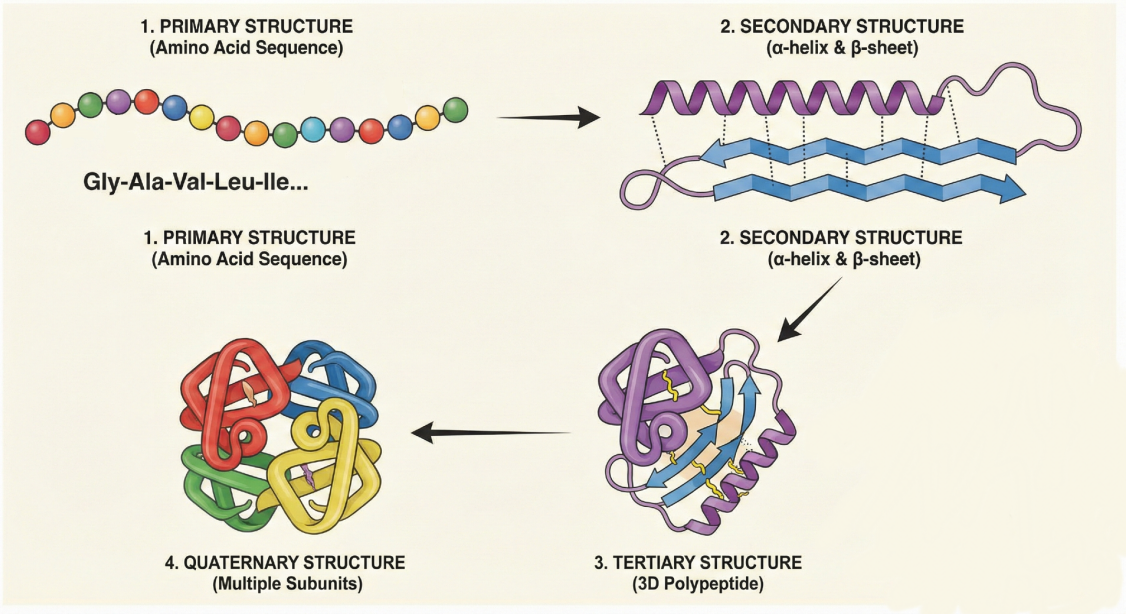


Figure 2.1: Representations of the Different Levels of Protein Structures. A linear amino acid chain (primary structure) folds into helices and sheets (secondary structure), forms a three-dimensional shape (tertiary structure), and may assemble with other chains into a multi-subunit protein (quaternary structure).Image generated by AI (Nano Banana), manually adapted and verified.

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 [11]. 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 biopharmaceuticals [12].

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.

2.2 Sourcing relevant data

Because the objective of this work is to train a predictive model, the quality and source of the training data is of upmost importance. Reliable predictions require datasets that accurately reflect successful protein purification outcomes. Considering that, this section outlines the public resources used to assemble the training corpus, including databases providing physico-chemical protein properties and literature sources from which experimentally validated purification protocols can be extracted.

2.2.1 Protein Data Bank

The Protein Data Bank (PDB) is the global repository for three-dimensional structural data of biological macromolecules [8]. Established in 1971, it serves as a central resource for structural biology by providing open access to validated models of proteins and nucleic acids through their website and an API, which we will be using. Each entry in the PDB represents a successful experiment where a protein was expressed, purified, and its structure determined.

In addition to atomic coordinates, PDB entries contain metadata such as crystallization conditions—including pH and temperature—and references to the primary literature. While the structural data is highly standardized and machine-readable, the specific purification protocols used to obtain these samples are not stored in a structured format within the database. Instead, these procedural details are typically contained within the "Materials and Methods" sections of the cited research papers. As a result, the PDB acts as a link between structured protein data and the unstructured purification processes found in scientific literature.

2.2.2 Europe PMC

Europe PMC serves as a primary open-access repository for life science literature, managed by the European Bioinformatics Institute (EMBL-EBI)[13]. At the time of writing, the platform indexes more than 47.5 million abstracts and 11.5 million full-text articles, aggregating data from major sources such as PubMed and PubMed Central. A critical feature relevant to this project is the availability of machine-readable full-text content in XML format, which is specifically designed to support large-scale text and data mining.

Programmatic access is facilitated through RESTful APIs and FTP bulk download services, allowing for the systematic retrieval of research data using standard identifiers like PMIDs or DOIs. Beyond simple document hosting, Europe PMC enriches its corpus with over 2 billion text-mined annotations for biological entities, including proteins, chemicals, and experimental methods, while maintaining reciprocal links to over 60 external life science databases [14].

2.2.3 UniProt

The Universal Protein Resource (UniProt) serves as the primary central repository for protein sequence data and functional annotation. Its core component, the UniProt Knowledgebase (UniProtKB), is structured into two main sections: UniProtKB/Swiss-Prot, which contains high-quality, manually curated entries, and UniProtKB/TrEMBL, which provides computationally annotated sequences [15].

Beyond simple sequence storage, UniProt facilitates data interoperability by providing a unified identification system with cross-references to other biological databases, such as the Protein Data Bank (PDB) for structural data and PubMed/Europe PMC for primary literature. This integration is essential for my data extraction pipeline, as it allows for the correlation of amino acid sequences with biochemical properties and experimental evidence.

2.2.4 Literature mining approaches

The challenge of extracting structured information from unstructured biomedical literature has been extensively documented in meta-research contexts [16–18]. The fundamental scalability problem inherent to manual literature curation is: with over 1 million papers indexed by PubMed annually, traditional manual extraction approaches become prohibitively time-consuming and difficult to reproduce [18]. This limitation is particularly evident when dealing with methodological details buried within specific, unstandardized sections of natural language text, like what we find in the scientific literature.

To address these challenges in their domain, Dockès et al. [18] developed `pubget`, a command-line tool for bulk downloading and processing articles from PubMed Central, and `labelbuddy` (Figure 2.2), a lightweight annotation application for creating ground-truth datasets. While these tools target neuroimaging literature specifically, their underlying approach to automated content extraction and manual validation may prove valuable for protein purification protocol mining, particularly when validation datasets become necessary for training or evaluating other extraction methods.

In Figure 2.3 we see a table with a breakdown of the pros and cons of the different approaches for their case, which is similar to ours. In our case, after researching the possibility of re-using an existing corpus, the ones closest in similarity to what we need were not readily available nor ideal. Manually collecting papers, given its low scalability, was also not suited for our needs, considering we needed a very large amount of data for our purpose. This left us no choice but to automatically collect papers, which is what our data mining tool attempts to do.

The authors' comparison of extraction methodologies provides important context for rule-based approaches like the keyword hierarchy employed in this work. When extracting participant demographics from neuroimaging studies, their heuristic method achieved exact matches in only 36% of cases, compared to 50% for zero-shot GPT-3.5 prompting [18]. However, the heuristic approach demonstrated comparable accuracy

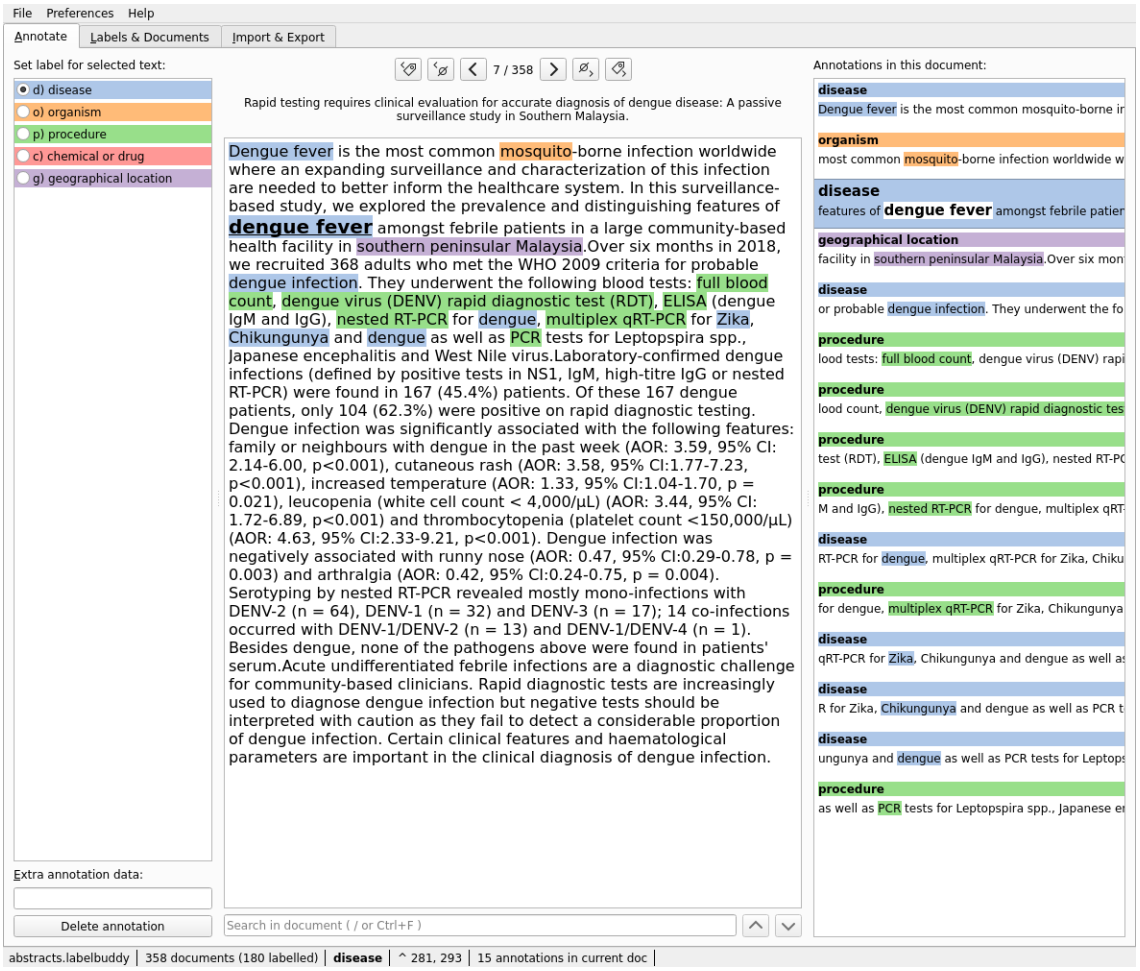


Figure 2.2: Screenshot of the Lightweight Data Annotation Tool `labelbuddy` [18]

	Existing workflows for collecting literature-mining data			Using our *litmining* ecosystem
	1) Re-using an existing corpus	**2) Manually** collecting papers	**3) Automatically** collecting papers	
Accessible (low technical expertise needed)	medium	high	low	medium
Scalable (not time consuming)	low	low	high	high
Reproducible	low (dataset) – high (analysis)	low	medium	high

Figure 2.3: Table Breakdown of Literature Mining Approaches by J. Dockès et al. [18]

when it did make predictions (0% median absolute percent error for both methods), with the primary difference being in recall—GPT-3.5 made predictions for 100% of papers versus 54% for the rule-based system. These results suggest that dictionary-based extraction, while potentially limited in coverage, can achieve acceptable precision for domain-specific information retrieval when patterns are sufficiently regular.

For the current phase of this project, a keyword-based approach offers the advantages of transparency, reproducibility, and computational efficiency. However, should initial results indicate insufficient recall or accuracy in capturing the full diversity of purification protocols, the precedent set by Dockès et al. demonstrates that large language models represent a viable alternative extraction strategy worth investigating.

2.3 Transformer-based models for text mining

The extraction of structured information from biomedical literature has traditionally relied on rule-based or dictionary-based approaches. While these methods can achieve high precision in identifying specific chromatography techniques or experimental parameters within defined contexts, they often struggle with the inherent linguistic variability and complex terminology found in scientific text [19]. Rule-based systems are frequently limited by the requirement for exhaustive keyword hierarchies and their inability to capture the broader semantic context of a sentence. To address these limitations, recent advancements in Natural Language Processing (NLP) have shifted toward deep learning architectures, most notably the Transformer.

2.3.1 BioBERT

The Bidirectional Encoder Representations from Transformers (BERT) model introduced contextualized word representations, allowing for a more nuanced understanding of text [20]. However, general-domain models often perform poorly on specialized scientific literature due to the significant shift in word distribution between general corpora and biomedical text [19].

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) addresses this domain gap by initializing with general-purpose BERT weights and undergoing further pre-training on large-scale biomedical corpora, specifically PubMed abstracts and PubMed Central full-text articles [19]. This domain-specific pre-training enables the model to effectively recognize complex biomedical entities and relationships that rule-based systems might miss. By leveraging BioBERT's capabilities, it is possible to automate the extraction of relevant purification keywords and chromatography steps with significantly higher accuracy than traditional methods. This would be achieved by identifying each purification step through Named Entity Recognition (NER) and correctly structuring them through Relation Extraction (RE).

2.3.2 GLiNER-BioMed

While models like BioBERT have significantly improved biomedical NER, they remain constrained by a fixed taxonomy, requiring the fine-tuning of a classification head for a pre-defined set of entities [19, 21]. This limitation makes them less adaptable to more specific domains, where certain entity types have no training data, as is our case of chromatography techniques. To overcome these challenges, GLiNER-BioMed introduces an "open NER" framework that treats entity recognition as a matching problem between text spans and natural language labels [21].

The primary advantage of GLiNER-BioMed for automated data mining lies in its ability to perform zero-shot recognition, enabling the extraction of arbitrary entity types without model retraining. This flexibility is achieved through a domain-specific adaptation of the Generalist and Lightweight Model for NER (GLiNER) [22]. The development of GLiNER-BioMed involved several key techniques, most notably synthetic data distillation. In this process, a large-scale teacher model (OpenBioLLM-70B) was used to generate high-quality NER annotations, which were then used to train a smaller student model to efficiently annotate a 105,000-sample pre-training corpus [21]. We can see a diagram of this workflow in Figure 2.4. This particular technique might prove useful if we come to train our own purification protocol extractor.

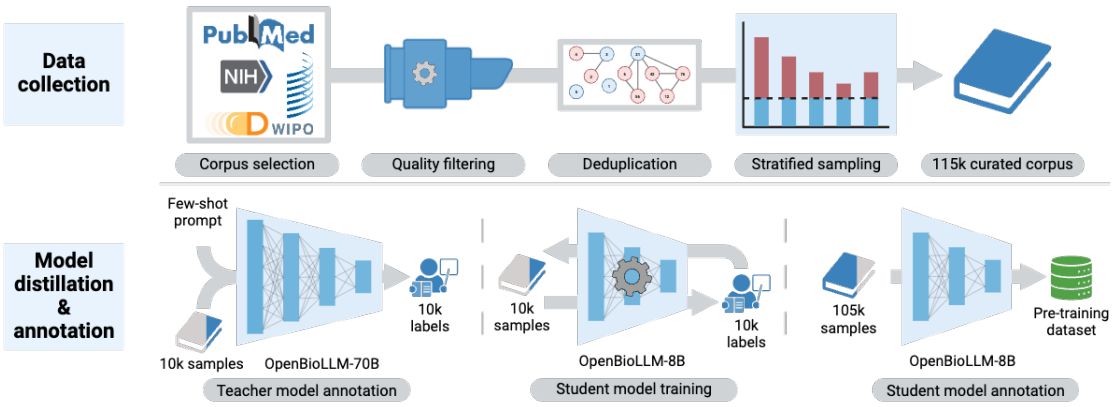


Figure 2.4: Overview of GLiNER-BioMed's Synthetic Pre-Training Data Generation Pipeline [21]

2.3.3 Large-scale relation extraction and knowledge integration

Building upon entity recognition, the final stage of automated data mining involves identifying the complex relationships between extracted entities to reconstruct sequential protocols. While BioBERT and GLiNER-BioMed excel at isolating specific technical terms, the work of Zhang et al. (2023) demonstrates how transformer-based models can be optimized for large-scale biomedical relation extraction (RE) across diverse categories [23]. Their research highlights several techniques that are particularly relevant for transforming raw extraction outputs into structured "laboratory-ready" recipes.

A key finding from their study is that the performance of relation extraction is significantly enhanced when entities are enriched with detailed semantic information. By incorporating semantic type names into the model's input representation, the architecture can better understand the functional role of an entity, thereby improving the accuracy of predicted relationships [23]. In our work this technique could be replicated using our chromatography techniques dictionary, where we would instead enrich our entities with the hierarchical information of the purification technique, for example.

For the objective of generating purification protocols, the integration of these extracted relations into a Knowledge Graph (KG) framework, as implemented by Zhang et al. (2023), offers a robust method for storing and querying chronological sequences [23]. By representing purification steps as nodes and their sequential connections as edges, it becomes possible to treat protocol generation as a structured sequence discovery problem.

2.4 Protein representation

In the context of predicting laboratory protocols, the representation of a protein must encapsulate more than its sequence to account for the diverse physico-chemical behaviors encountered during chromatography. Recent developments in protein representation learning have focused on addressing the inherent "knowledge gap" in standard protein language models (LMs), which often fail to capture factual biological context. A notable contribution is the Knowledge-exploited Auto-encoder for Protein (KeAP), which introduces a more granular, token-level exploration of knowledge graphs to enrich primary structure modeling [24].

Unlike earlier models that integrated knowledge at a coarse sequence-wide level, KeAP utilizes a cross-attention mechanism where individual amino acids iteratively query associated knowledge tokens, specifically relation and attribute terms derived from Gene Ontology (GO) [25, 26]. This interaction allows the model to integrate functional and structural descriptors, such as molecular functions and cellular localizations, directly into the protein representation. This granular integration enables the production of a better-contextualized representation that has shown superior performance in predicting downstream properties like protein stability and binding affinity. For a system designed to generate purification recipes, this paradigm is highly applicable, as it allows the input representation to explicitly leverage biological metadata that dictate chromatography behavior, providing a richer foundation for purification protocol prediction.

In Figure 2.5 we have a diagram of KeAP's cross attention mechanism from the original paper. Given a knowledge-graph triplet (Protein, Relation, Attribute), the protein sequence is encoded into amino-acid embeddings, while the associated relation and attribute texts are encoded into word-level representations. Knowledge integration is performed in the protein decoder through stacked Protein–Knowledge exploration (PiK) blocks.

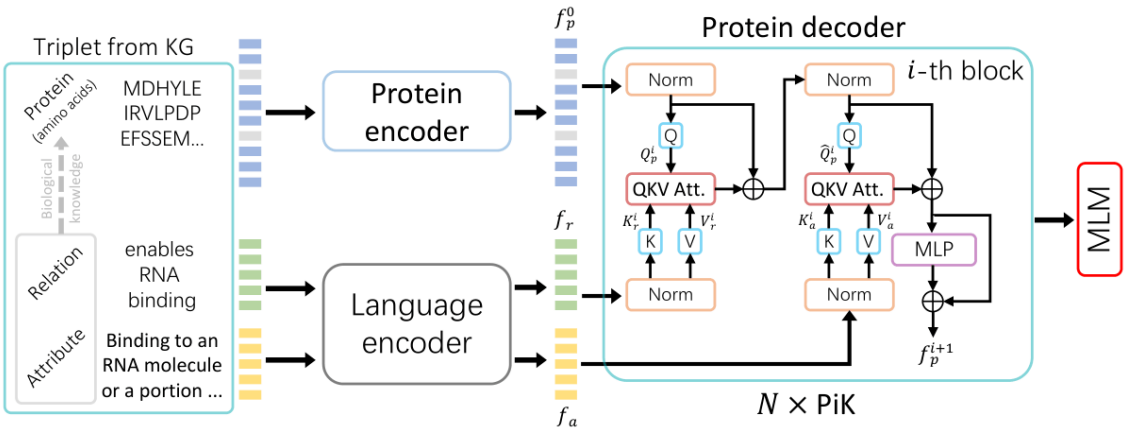


Figure 2.5: KeAP's Cross-Attention Mechanism [24]

2.5 Previous Works on Protocol Information Extraction

Previous work has addressed problems similar to those considered here. Training a model to predict protein purification protocols requires a large amount of high-quality data, which motivates the development of our data mining tool. Although such work would ideally rely on an existing database linking proteins to their purification steps, we were unable to find any accessible database that provides this information.

2.5.1 PurificationDB

PurificationDB is a curated knowledge base specifically designed to aggregate experimentally validated protein purification conditions from the literature [27]. The database was constructed by first identifying protein structures deposited in the PDB, under the assumption that successful structure determination implies prior successful purification. Associated publications were retrieved via PDB-linked DOIs, and full texts were collected primarily from crystallography reports. From these documents, purification-related information was extracted using a rule-based named-entity recognition (NER) framework informed by expert-defined vocabularies covering chromatography techniques, buffer components, and concentration units.

From the perspective of our work, PurificationDB represents a highly relevant prior effort, as it targets the systematic extraction and structuring of purification knowledge from unstructured literature, closely aligning with our initial objective of compiling a protein purification protocol database. However, despite the authors of the paper providing a URL[1] where we should be able to access the "open-access and user-friendly knowledge base", it was not accessible at the time of our investigation. This limitation further motivates the development of reproducible, fully automated data mining pipeline that operates on publicly available repositories and does not depend on the long-term availability of third-party curated databases.

[1] https://purificationdatabase.herokuapp.com/

2.5.2 Extraction of purification protocol information using a LLM

A closely related study by Chen and Sivaraman [28] demonstrates the feasibility of leveraging large language models (LLMs) to systematically extract protein expression and purification strategies from scientific papers indexed in the PDB. Similar to our work, their central premise is that proteins with solved three-dimensional structures represent successful prior purification efforts, and thus constitute a high-quality empirical foundation for learning purification strategies.

Methodologically, their pipeline parallels our data mining framework in several ways. First, both approaches use the PDB as an entry point to anchor experimental protocols to validated protein sequences via UniProt identifiers. Second, the authors perform large-scale full-text processing of structural biology articles, with a strong focus on sections where purification details are reported. Third, both studies aim to convert unstructured experimental descriptions into structured, machine-interpretable representations of purification procedures.

A notable contribution from this paper is their hybrid information extraction strategy, which combines dense text embeddings with a multi-step LLM prompting scheme to localize and extract protocol-relevant passages. Specifically, they employ embedding-based section ranking to restrict LLM attention to purification-related text segments, followed by a two-step LLM extraction process and structured prompts to reduce hallucination and misclassification errors. These design choices are directly relevant to our work, as they provide validated techniques for improving the precision of protocol reconstruction from long and unstructured biomedical articles. In Figure 2.6 we see a diagram representing this strategy.

Although the goals of this paper are closely aligned with those of our data extraction tool, its contributions are purely theoretical, as the authors do not make any of their developed resources publicly available. While they state that the data supporting their findings can be obtained from the corresponding author upon reasonable request, access would be restricted to the specific dataset extracted at the time of their study, thereby limiting the scope of data we would have to work with.

Furthermore, our project uses the extracted data for a fundamentally different purpose—namely, the training of a predictive model—, whereas the aforementioned study employs it for a statistical analysis of purification strategies. In addition, we aim to construct a larger database that incorporates other protein-related information to be used as input features for prediction and model training. We emphasize this distinction because different objectives can lead to subtle but significant differences in the approaches used for data extraction and preprocessing.

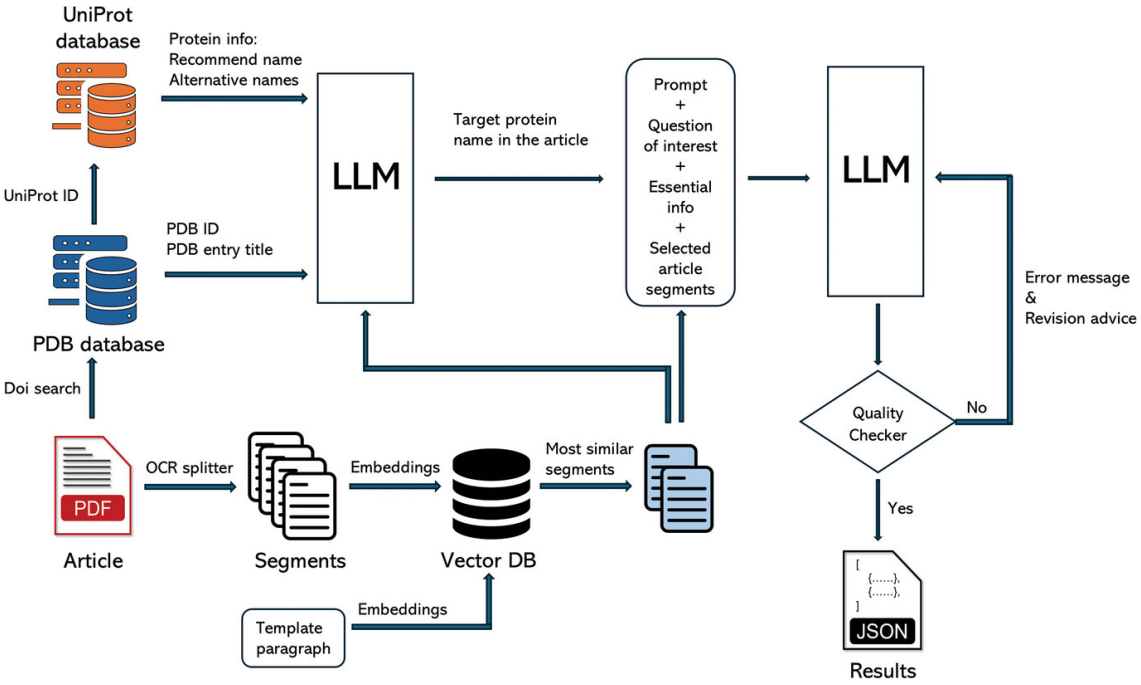


Figure 2.6: Workflow of the Paper's Efficient Article Information Extraction Tool [28]. Articles are processed from PDF files using OCR and segmented into section-based text chunks. All segments are embedded and stored in a vector database, enabling retrieval of the most relevant passages for a given question using semantic similarity. Protein names are identified and standardized using information from UniProt and the PDB, with assistance from an LLM. The selected article segments, target protein names, and question of interest are combined into a prompt for LLM-based information extraction. A quality-checking step validates the output before structured results are generated.

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. This model will need to be part of a larger pipeline 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 plan 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. *Document 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 protocols

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 Figure 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 Figure 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 Figure 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/

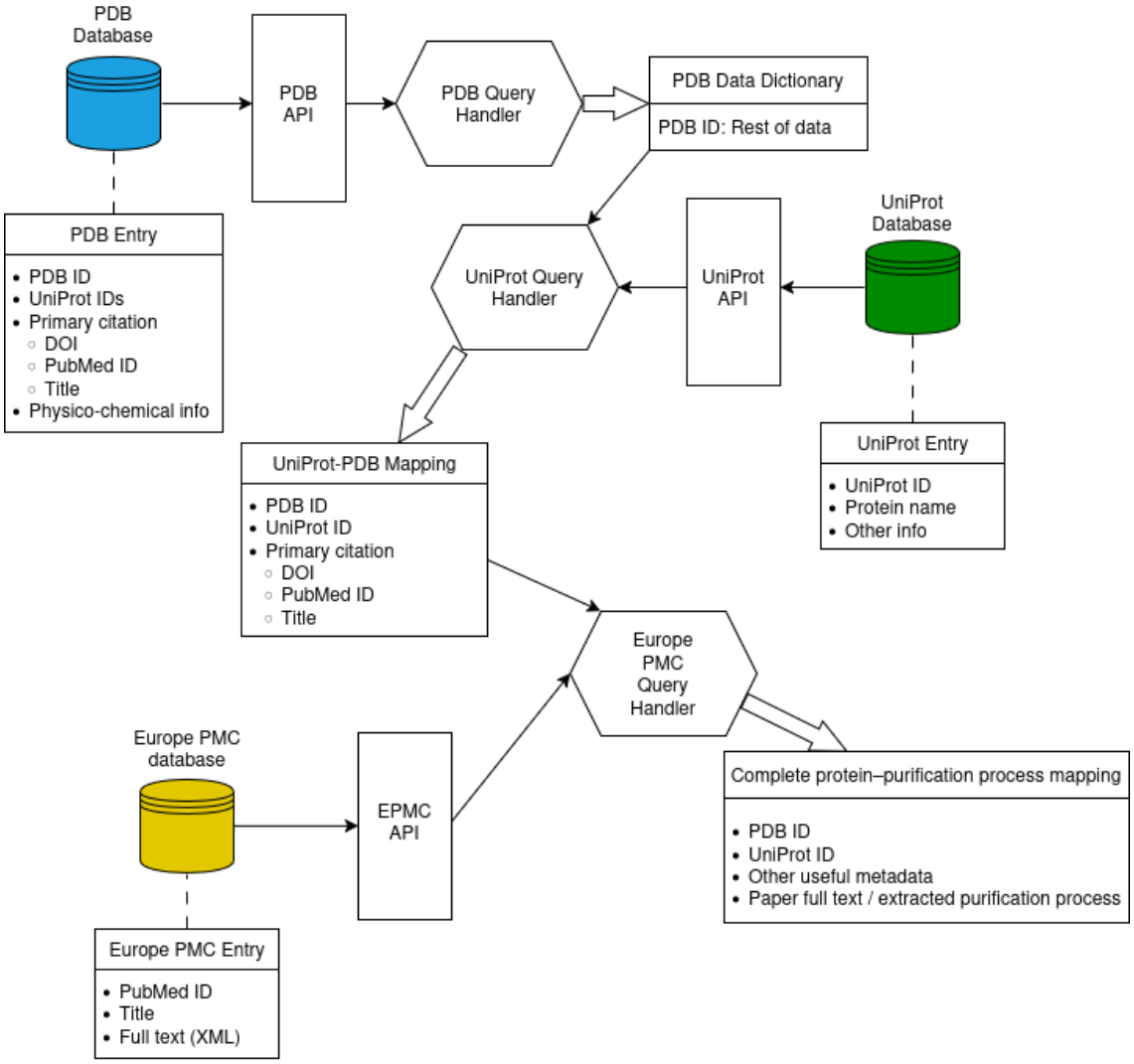


Figure 3.1: Diagram of Old Method of Information Extraction. The PDB query handler queries the PDB API for its data, resulting in a dictionary of data of the structures. The PDB data contains UniProt IDs, which we use to map structures to their corresponding proteins. Finally, we use the PDB's primary citation information to query Europe PMC's API for the full-text of the paper, which we use to extract the protein's purification protocol.

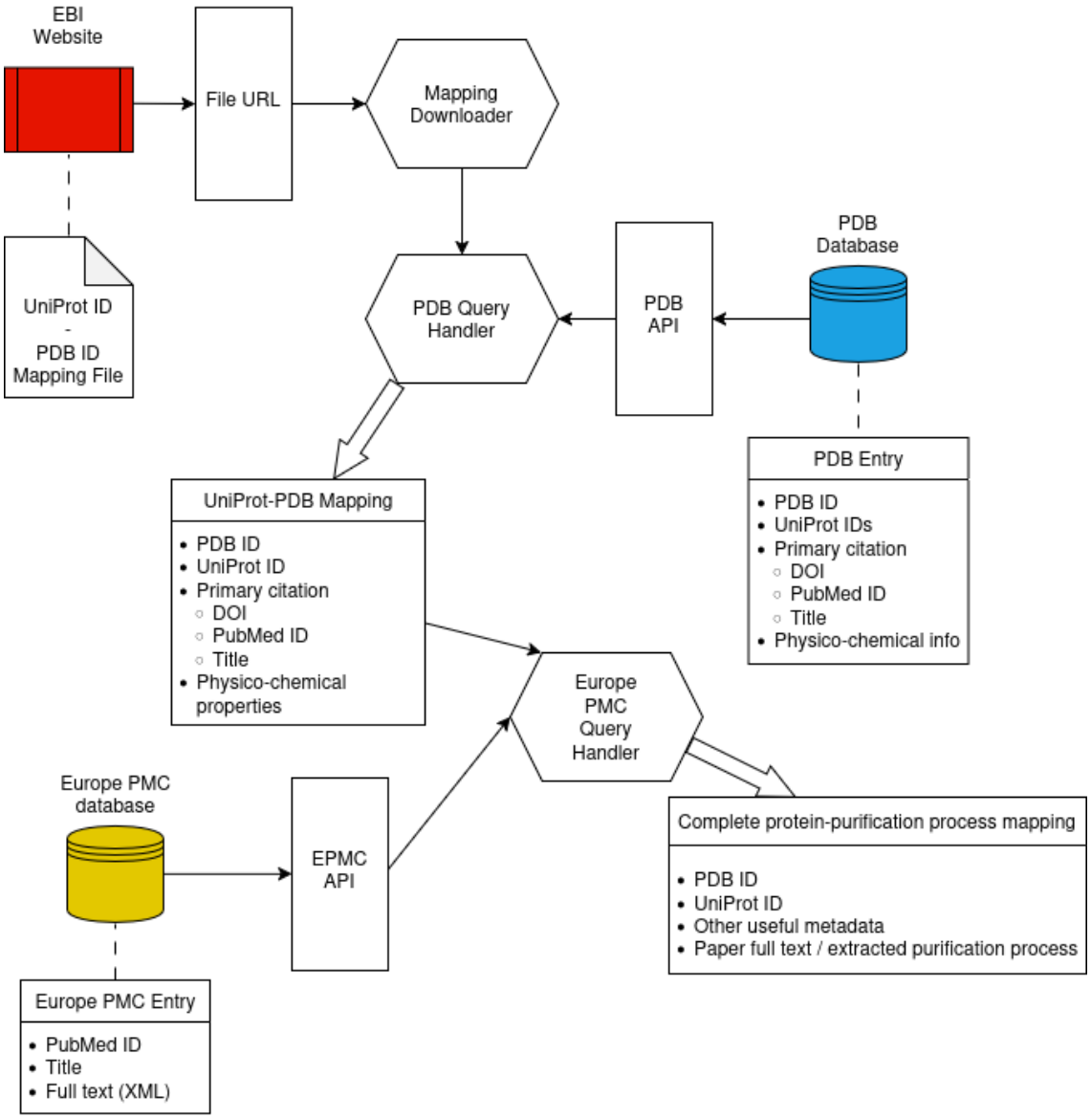
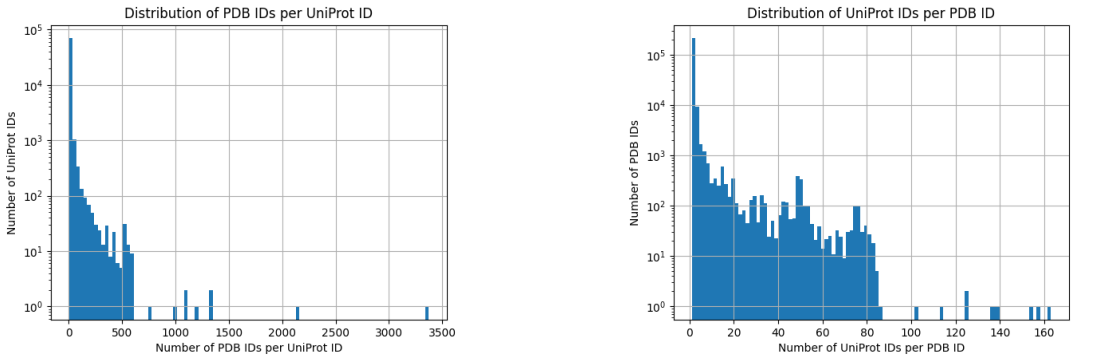


Figure 3.2: Diagram of New Method of Information Extraction. In the new method of information extraction the mapping between a UniProt ID and a PDB ID is downloaded from EBI's website. Thanks to this, all we have to do is fetch the primary citation information for each PDB entry and fetch the paper full-text from EuropePMC.



(a) Number of PDB IDs per UniProt ID

(b) Number of UniProt IDs per PDB ID

Figure 3.3: PDB/UniProt ID Distribution

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.

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:

- **Repeated PMIDs:** 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.

- **Not open access:** 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.
- **No citation available:** 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.
- **Failed to get full text:** 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.

Given the size of the full mapping and the databases we query, using every entry for testing purposes would take a prohibitive amount of time and compute, so a small subset of the first 200 entries was used instead. In Figure 3.4 we see the results for this phase.

Each arm of the sankey diagram represents a different error received by the data extraction tool, but upon closer inspection the cases where we received an empty result was actually due to lack of open access. Considering this, the main problem was the lack of open access to papers (20%), followed by the lack of citations in PDB (19.5%).

Many PDB entries cite the same primary publication because a single study often reports the experimental determination of multiple protein structures. As a result, the same PMID frequently appears across multiple entries, leading to repeated PMIDs in our dataset (18.5% of examples). Extracting a single protein purification protocol from unstructured text is already challenging; extracting multiple distinct protocols from the same source and correctly associating each with its corresponding protein will be substantially more difficult. Nevertheless, we anticipate that with the employment of more advanced NLP methods, particularly the use of LLMs, we will be able to achieve a high success rate.

Although repeated PMIDs are challenging to handle, they correspond to successful full-text extractions and, together with the other successful cases, yield an overall success rate of 52.5%. If this rate were to hold across the more than 249,000 PDB entries, we could, in principle, compile up to approximately 130,000 distinct protein purification protocols, assuming a 100% success rate in protocol extraction and ignoring potential duplicates.

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

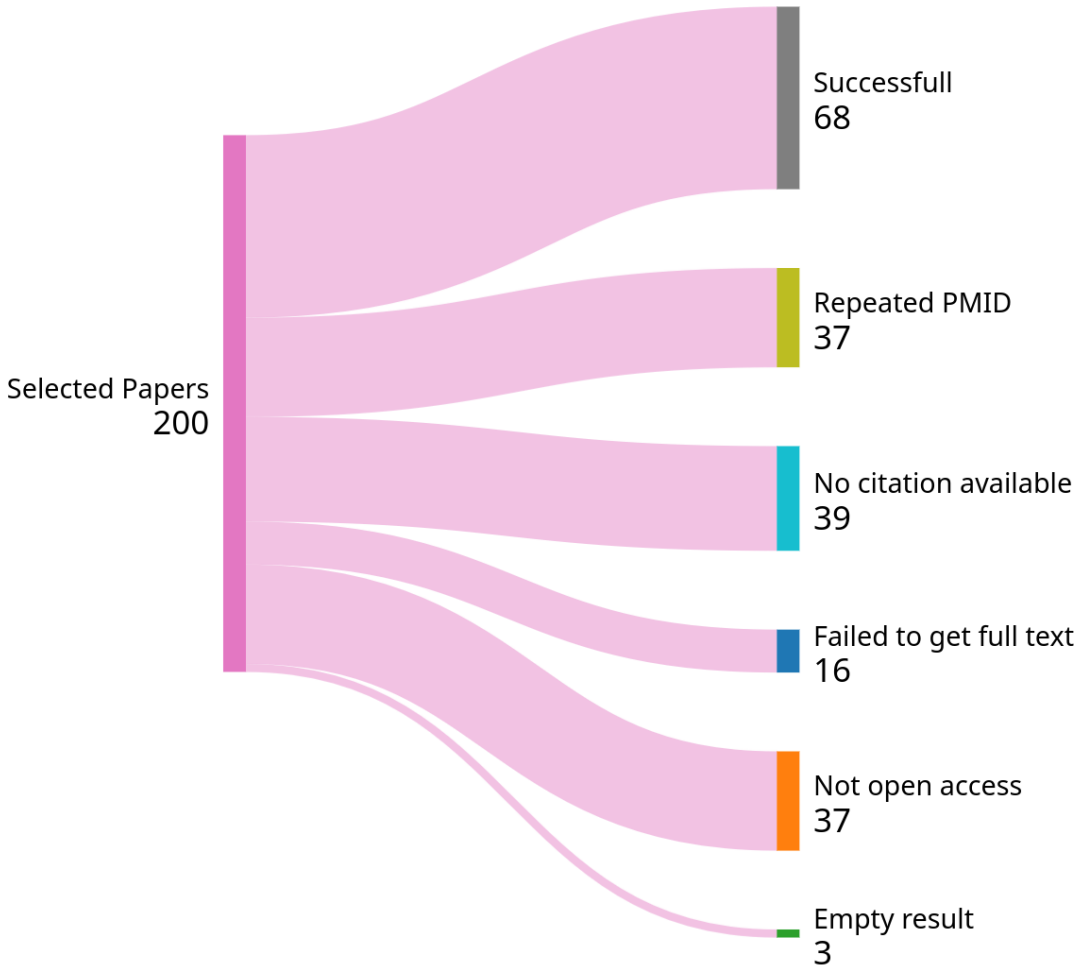


Figure 3.4: Sankey Diagram of the Results of the Full-Text Extraction Algorithm. These results correspond to the papers associated to the 200 proteins we selected and their outcomes are explained in Section 3.2.3.

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 (Figure 1.1).

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 Figure 3.5 we can see an example of the keywords' hierarchy for the "Size Exclusion" technique category.

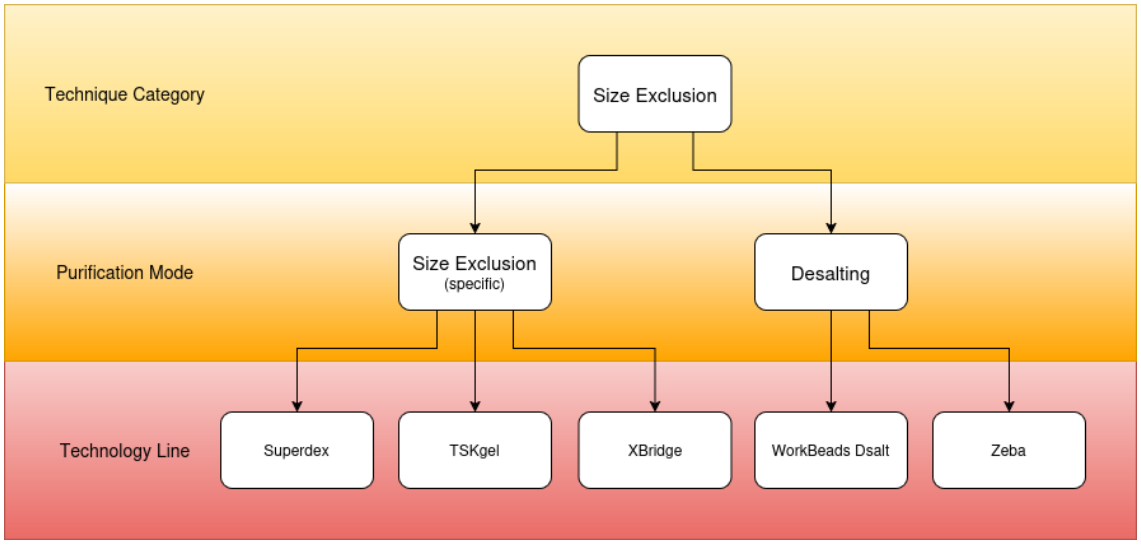


Figure 3.5: Example of Chromatography Techniques Hierarchy

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

While substantial progress has been made on the data mining pipeline, significant work remains to ensure data quality, develop the predictive model, and validate its performance. The remaining tasks include finalizing the extraction tool, executing it on the complete dataset, defining and training the Transformer architecture, and documenting the findings. Figure 3.6 presents the proposed timeline for the upcoming semester.

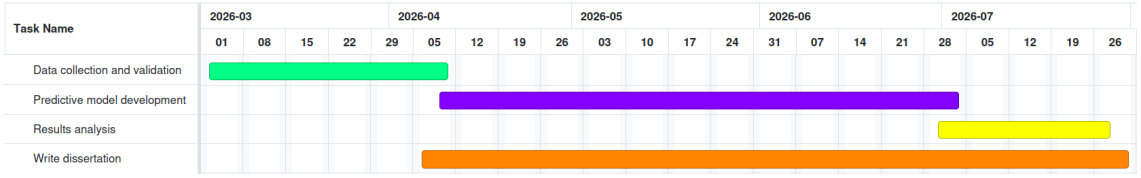


Figure 3.6: Work Plan Chart

3.3.1 Data Collection and Validation

Three interdependent subtasks will complete the data mining phase: finalizing the extraction script, validating its output, and executing it on the full dataset. Currently, the pipeline processes only 200 PDB–UniProt entries for testing purposes, as processing the complete dataset during development would be computationally prohibitive.

The overlapping nature of these subtasks reflects the iterative development process inherent to data mining tools. Script refinement proceeds through cycles of execution, error identification, and correction. Validation will involve systematic verification of extracted protocols against source publications and identifying gaps in the extraction logic that require adjustment. Once validation confirms acceptable accuracy, the pipeline will be scaled to the complete dataset, with ongoing validation to address any previously undetected edge cases.

3.3.2 Predictive Model Development

Model development comprises several interconnected subtasks. Initial work will focus on architecture design, involving the evaluation of alternative Transformer configurations based on sequence modeling requirements, attention mechanisms, and the vocabulary of chromatography techniques, as seen in Appendix A.

The protein entries from the PDB that have purification protocols described in the literature will be curated and several structure-based physico-chemical properties will be calculated using the software MOE [29]. First, the PDB structures will be fixed for missing atoms/amino acids using the QuickPrep Tool. Then, physico-chemical properties will be calculated to be used as protein descriptors, Table 3.1. These descriptors will be what the model will base itself on to make predictions. This task will be done in collaboration with the Biomolecular Engineering Lab, Chemistry Department, NOVA FCT, UNL.

Table 3.1: Table of Protein Descriptors

Name	Description	Class
patch_hyd	Area of hydrophobic protein patch(es)	Protein Patch
patch_pos	Area of positive protein patch(es)	Protein Patch
patch_neg	Area of negative protein patch(es)	Protein Patch
patch_ion	Area of ionic protein patch(es)	Protein Patch
ens_charge	Ensemble Net Charge	Titration
mass	Protein Mass in kDa	Molecular
pI_3D	Structure-based pI Prediction	Molecular
r_gyr	Radius of Gyration	Molecular
r_solv	Hydrodynamic Radius	Molecular
asa_vdw	Accessible Surface Area (Water Probe)	Molecular
asa_hyd	Hydrophobic Surface Area	Molecular
asa_hph	Hydrophilic Surface Area	Molecular
volume	Protein Volume	Molecular
mobility	Protein Mobility	Molecular
dipole_moment	Protein Dipole Moment	Molecular
hyd_moment	Hydrophobicity Moment	Molecular
res_exp	Residue Percent Exposure	Residue
pI_seq	Sequence-based pI Prediction	Sequence

The training phase represents the most resource-intensive component of the project. Beyond the computational demands of model training itself, this period encompasses the development of data preprocessing pipelines, implementation of the wrapper algorithm, and iterative refinement based on preliminary results. The wrapper algorithm serves as a gap filling layer that addresses practical limitations of the model, handling tasks such as input normalization, output formatting and constraint enforcement to maximize the generated protocols' usefulness, resulting in a complete end-to-end tool.

3.3.3 Results Analysis

Once the development of the predictive model is mostly finished, we will enter a phase of result analysis and validation. During this phase we will apply our tool to specific case studies and perform a formal qualitative evaluation. Depending on the outcome, some changes may still be applied, but we expect to have a mostly finished tool by this point, so they should be minor. Finally, we will proceed with the publication of the code and model.

Bibliography

[1] M. Du et al. "Progress, applications, challenges and prospects of protein purification technology". In: *Frontiers in Bioengineering and Biotechnology* 10 (2022), p. 1028691. DOI: 10.3389/fbioe.2022.1028691 (cit. on pp. 1, 2).

[2] W. T. Booth et al. "Impact of an N-terminal Polyhistidine Tag on Protein Thermal Stability". In: *ACS Omega* 3.1 (2018), pp. 760–768. DOI: 10.1021/acsomega.7b01598 (cit. on p. 1).

[3] C. Biostructure. *Custom Affinity Cromatography Service*. URL: https://www.creative-biostructure.com/custom-affinity-chromatography-service-257.htm (cit. on p. 2).

[4] C. Biostructure. *Custom Size Exclusion Cromatography Service*. URL: https://www.creative-biostructure.com/custom-size-exclusion-chromatography-service-259.htm (cit. on p. 2).

[5] C. Biostructure. *Custom Ion Exchange Cromatography Service*. URL: https://www.creative-biostructure.com/custom-ion-exchange-chromatography-service-258.htm (cit. on p. 2).

[6] C. Biostructure. *Custom Hydriphobic interaction Cromatography Service*. URL: https://www.creative-biostructure.com/custom-hydrophobic-interaction-chromatography-service-260.htm (cit. on p. 2).

[7] J. Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596 (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 3).

[8] H. M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (2000-01), pp. 235–242. DOI: 10.1093/nar/28.1.235 (cit. on pp. 3, 6).

[9] E. Martz. *Timeline of Protein Chemistry*. 2002. URL: https://www.umass.edu/microbio/chime/pe_beta/pe/protexpl/histprot.htm (cit. on p. 4).

[10] J. Pietzsch. *Speed read: Preparing pure proteins*. Nobel Prize Outreach. URL: https://www.nobelprize.org/prizes/chemistry/1946/speedread (cit. on p. 4).

[11] C. H. Taron, J. C. Samuelson, and L. Morrison. *Over 40 Years in Protein Expression and Purification*. New England Biolabs, Inc. URL: https://www.neb.com/en/tools-and-resources/feature-articles/over-40-years-in-protein-expression-and-purification (cit. on p. 5).

[12] M. E. Kimple, A. L. Brill, and R. L. Pasker. "Overview of affinity tags for protein purification". In: *Current Protocols in Protein Science* 73 (2013-09), pp. 9.9.1–9.9.23. DOI: 10.1002/0471140864.ps0909s73 (cit. on p. 5).

[13] S. Rosonovski, M. Levchenko, R. Bhatnagar, et al. "Europe PMC in 2023". In: *Nucleic Acids Research* 52.D1 (2024), pp. D1668–D1676. DOI: 10.1093/nar/gkad1085 (cit. on p. 6).

[14] A. Venkatesan et al. "SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data [version 2; peer review: 2 approved, 1 approved with reservations]". In: *Wellcome Open Research* 1.25 (2017). DOI: 10.12688/wellcomeopenres.10210.2 (cit. on p. 6).

[15] The UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2025". In: *Nucleic Acids Research* 53.D1 (2025-11), pp. D609–D617. DOI: 10.1093/nar/gkae1010 (cit. on p. 7).

[16] I. Beltagy, K. Lo, and A. Cohan. *SciBERT: A Pretrained Language Model for Scientific Text*. 2019. arXiv: 1903.10676 [cs.CL] (cit. on p. 7).

[17] F. Liu et al. "Learning for Biomedical Information Extraction: Methodological Review of Recent Advances". In: (2016-06). DOI: 10.48550/arXiv.1606.07993 (cit. on p. 7).

[18] J. Dockès et al. "Mining the neuroimaging literature". In: *eLife* 13 (2024), RP94909. DOI: 10.7554/eLife.94909 (cit. on pp. 7, 8).

[19] J. Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240 (cit. on pp. 9, 10).

[20] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019-06, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cit. on p. 9).

[21] A. Yazdani, I. Stepanov, and D. Teodoro. "GLiNER-BioMed: A Suite of Efficient Models for Open Biomedical Named Entity Recognition". In: *arXiv preprint arXiv:2504.00676* (2025) (cit. on p. 10).

[22] U. Zaratiana et al. "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics, 2024-06, pp. 5364–5376. DOI: 10.18653/v1/2024.naacl-long.300 (cit. on p. 10).

[23] Z. Zhang et al. "Large-Scale Biomedical Relation Extraction Across Diverse Relation Types: Model Development and Usability Study on COVID-19". In: *Journal of Medical Internet Research* 25 (2023), e48115 (cit. on pp. 10, 11).

[24] C. Bian et al. "Protein Representation Learning via Knowledge Enhanced Primary Structure Modeling". In: *International Conference on Learning Representations (ICLR)*. 2023 (cit. on pp. 11, 12).

[25] M. Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), pp. 25–29. DOI: 10.1038/75556 (cit. on p. 11).

[26] T. G. O. Consortium. "The Gene Ontology knowledgebase in 2026". In: *Nucleic Acids Research* 54.D1 (2025-12), pp. D1779–D1792. ISSN: 1362-4962. DOI: 10.1093/nar/gkaf1292 (cit. on p. 11).

[27] O. Garland et al. "PurificationDB: database of purification conditions for proteins". In: *Database* 2023 (2023), baad016 (cit. on p. 12).

[28] Z. Chen and J. Sivaraman. "Using Large Language Model to Optimize Protein Purification: Insights from Protein Structure Literature Associated with Protein Data Bank". In: *Advanced Science* 12 (2025), p. 2413689. DOI: 10.1002/advs.202413689 (cit. on pp. 13, 14).

[29] Chemical Computing Group ULC. *Molecular Operating Environment (MOE)*. Version 2024.0601. 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, 2026 (cit. on p. 25).

A

Chromatography techniques dictionary

Brand	Technology Line (Resin)	Primary Column Formats	Purification Mode	Specific Method
Cytiva	**Superdex / Superose / Sephacryl**	HiLoad, Increase (10/300), HiPrep	Size Exclusion (SEC)	Size Exclusion
	MabSelect (PrismA, SuRe)	HiTrap, HiScreen, ReadyToProcess	Affinity (Protein A/Antibody)	Antibody Affinity
	Capto (Q, S, DEAE, MMC, Adhere)	HiTrap, HiScreen, HiPrep	Ion Exchange / Mixed-Mode	Ion Exchange, Mixed mode chromatography
	HisTrap / GSTrap / StrepTrap	HiTrap (1 mL, 5 mL)	Affinity (Tagged Proteins)	Affinity chromatography, Immobilized metal (ion) affinity chromatography
	Sepharose (Fast Flow, High Performance)	HiTrap, HiPrep, XK (Empty)	General IEX, HIC, Affinity	IEX, HIC, Affinity chromatography
	Resource (Q, S, PHE)	Resource 1 mL, 6 mL, 15 mL	High-Res Polishing (Fast)	IEX
	Mono Q / Mono S	Tricorn 5/50, 10/100	Ultra-High-Res Polishing	IEX
	HiTrap Fibro	PrismA, Tellus (Fiber units)	Rapid Capture (Seconds)	Affinity chromatography
Bio-Rad	**ENrich**	SEC 70, SEC 650, Q, S	High-Res SEC & IEX	Size Exclusion, Ion Exchange
	Nuvia (Q, S, HR-S, IMAC)	EconoFit, Foresight, Bio-Scale	High-Capacity IEX / Affinity	Ion Exchange, Affinity chromatography
	CHT (Ceramic Hydroxyapatite)	EconoFit, Foresight, XT	Mixed-Mode (Unique Resolution)	Hydroxyapatite
	Profinity (IMAC, GST, eXact)	EconoFit, Bio-Scale Mini	Tagged Protein Affinity	Immobilized metal (ion) affinity chromatography, Affinity chromatography
	UNSphere (Q, S)	EconoFit, Macro-Prep	Rapid Ion Exchange	Ion Exchange, Hydrophobic interaction chromatography
	Macro-Prep	High Q, High S, Methyl HIC	Preparative Scale IEX/HIC	Ion Exchange, Affinity chromatography
Thermo Scientific	**POROS (A, G, HQ, HS, XS)**	GoPure (Prepacked), PEEK/SS	Perfusion IEX / Affinity	IEX, Affinity chromatography
	CaptureSelect	POROS-based or Resin-only	Specialized Affinity (VHH)	Affinity chromatography
	MAbPac	HPLC (4x50mm, 4x250mm)	Analytical Antibody IEX/SEC	IEX, SEC, Affinity chromatography
	ProPac (Elite, WCX, SCX)	HPLC / UHPLC Formats	Analytical Charge Variant IEX	Ion Exchange
	Zeba	Spin Columns / 96-well	Desalting / Buffer Exchange	Desalting / Buffer Exchange
Tosoh Bioscience	**TSKgel (SW, SWXL, SuperSW)**	Stainless Steel (Analytical)	High-Res SEC (Silica-based)	Size Exclusion
	TSKgel (PW, PWXL)	Stainless Steel / PEEK	SEC for Large Polymers/Proteins	Size Exclusion
	Toyopearl (Q, S, Hexyl, Butyl)	ToyoScreen, MiniChrom	Preparative IEX, HIC, Affinity	IEX, HIC, Affinity chromatography
	TSKgel BioAssist (Q, S)	Stainless Steel / PEEK	High-Throughput IEX	Ion Exchange
Merck / MilliporeSigma	**Eshmuno (A, Q, S, CPX, CMX)**	MiniChrom, RoboColumn	High-Productivity IEX/Affinity	Ion Exchange, Affinity chromatography
	Fractogel (EMD Q, S, SO3)	MiniChrom, Scout Columns	Tentacle-based IEX/HIC	Ion Exchange, Hydrophobic interaction chromatography
	Hibar	Stainless Steel	Analytical Prep	
Waters	**BioResolve mAb**	4.6 x 50/150/300 mm (HPLC)	SEC, RP, SCX (mAb specific)	Affinity chromatography
	XBridge / ACQUITY	Premier (Bio-inert hardware)	Protein SEC (125, 250, 450 Å)	Size Exclusion
	Protein-Pak	Hi Res Q, Hi Res S, Hi Res CM	Ion Exchange (IEX)	Ion Exchange
	BioSuite	SEC, Phenyl (PA)	SEC & Specialized Affinity	Size exclusion chromatography
Agilent	**AdvanceBio SEC**	130, 300, 500, 1000 Å	Size Exclusion (Peptides to VLP)	Size exclusion chromatography
	AdvanceBio IEX	Bio MAb, Bio Q, Bio S	High-Res Charge Variant IEX	Ion Exchange
	AdvanceBio HIC	Phenyl, Butyl, Ether	Hydrophobic Interaction	Hydrophobic interaction chromatography
	Bio-Monolith	5.2 x 4.9 mm (Monolith Discs)	Ultra-fast Protein A, G, IEX	Ion Exchange, Affinity chromatography
	PLRP-S	Analytical & Prep (Polymer)	Reversed-Phase (RP)	Reversed-Phase Chromatography
YMC	**BioPro IEX**	Smart Sep, QA, SP, QF, SF	Porous & Non-porous IEX	Ion Exchange
	BioPro HIC	HIC BF (Butyl), HIC HT	High-res HIC for ADCs	Hydrophobic interaction chromatography
	YMC-SEC MAB	250 Å (Silica)	Specialized mAb SEC	Size exclusion chromatography
	YMC-Triart Bio	C4, C18 (Hybrid Silica)	RP for Peptides & Proteins	Reversed-Phase Chromatography
Bio-Works	**WorkBeads affimAb**	GoBio Mini (1, 5mL), Prep	Alkali-stable Protein A	Affinity chromatography
	WorkBeads 40/100	GoBio Mini, Screen, Prep	Q & S Ion Exchange	Ion Exchange
	WorkBeads IMAC	Ni-NTA, IDA, TREN	His-tag Affinity	Immobilized Metal Affinity
	WorkBeads Dsalt	GoBio Mini Dsalt	Desalting / Buffer Exchange	Desalting / Buffer Exchange
Sartorius	**Sartobind Q / S**	Nano, Mini, Capsules	Rapid Membrane IEX	Ion Exchange
	Sartobind Rapid A	Capsules, Cassettes	Rapid Membrane Protein A	Affinity chromatography
	Sartobind Phenyl	Nano, Capsules	Rapid Membrane HIC	Hydrophobic interaction chromatography
	Sartobind STIC PA	Capsules, Jumbo	Salt-tolerant Polishing	
Shodex	**PROTEIN KW-800**	8.0 x 300 mm (KW-802.5, 803)	High-recovery SEC	Size Exclusion
	PROTEIN LW-803	8.0 x 300 mm	High-load mAb SEC	Size Exclusion
Phenomenex	**Biozen dSEC**	dSEC-2, dSEC-7	Analytical SEC (1k - 450kDa)	Size Exclusion
	Biozen WCX / SCX	6 µm (BioTi Hardware)	Charge Variant IEX	Ion Exchange
	Biozen Intact	C8, C4 (Core-Shell)	RP for Intact Mass Spec	Reversed-Phase Chromatography