Annabel Margalef Burgos

Integrative multi-omics profiling and pathway analysis for human biofluids

Final degree project

Directed by Dr. Raquel Cumeras Olmeda

Bachelor's Degree in Biomedical Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
(2024)

Index

Abstract

Blood and cerebrospinal fluid are two biofluids with very important functions in the human body. To study these two biofluids, multi-omics analysis and pathway analysis are necessary. Therefore, a source of information that includes all the omics data of these biofluids is needed in order to conduct comparative studies and and metabolic pathway analysis. With the current information sources available on the internet, this is not possible, as many of them do not contain specific information about the fluids, are incomplete, outdated, and have redundant information.

Therefore, the LifeFluids project emerged with the objective of creating multi-omics databases of the composition of the body's biofluids. In this final degree project, two unified databases were created for blood and cerebrospinal fluid, compiling information on their composition of metabolites, proteins, transcripts, and chemical elements. Afterwards, multi-omics profiling and pathway analysis were performed with the obtained data to evaluate the metabolic pathways involving both biofluids.

The methodology involved identifying necessary and available sources for collecting information on the mentioned omics, creating the databases, studying correlations between the two biofluids to identify common elements, and analyzing metabolic pathways.

For the creation of the databases, existing information sources were consulted and integrated, such as recognized online databases like the Human Metabolome Database (HMDB), Vesiclepedia (VSCP), and Marker DB (MDB), among others. Additionally, literature and repositories were consulted. These sources use different nomenclatures, so standardized identifiers were needed to normalize the data included in our database: InChIKey for metabolites and chemical elements, Uniprot ID for proteins, and HGNC ID for transcripts.

As a result, the Life Fluids database for blood contains 90.482 metabolites, 15.763 proteins, 1.463 transcripts, and 98 chemical elements. The Life Fluids database for cerebrospinal fluid contains 2.090 metabolites, 5.447 proteins, 1.431 transcripts, and 47 chemical elements.

Afterwards, all this data was used to investigate associated diseases by studying the metabolic pathways involved in each biofluid and the consequences that their perturbations could have.

The newly created databases provide a valuable resource for researchers studying these biofluids, enabling a more comprehensive analysis of the previously existing data.

Key words: blood, CSF, databases, metabolites, proteins, transcripts, chemical elements, multi-omics profiling, pathway analysis.

1. Introduction

Blood and cerebrospinal fluid are two human body fluids or biofluids. Their study is important in health research, as they are involved in the development of tools for the diagnosis, monitoring, and treatment of different types of diseases.

Blood is a specialized biofluid with four main components: plasma, red blood cells, white blood cells, and platelets. It has many diverse functions within the human body, including transporting oxygen and nutrients from the lungs to the rest of the body's tissues, transporting cells and antibodies that fight infections, bringing waste products to the kidneys and liver, which filter and clean the blood, and many more [1].

On the other hand, we have cerebrospinal fluid (CSF), a clear, colorless, watery fluid that is situated inside and around the brain and spinal cord, that is, the central nervous system. Its main function is to act as a cushion that protects the brain and spinal cord from impacts and other types of injuries. It also serves to remove waste products, and nourishment of the brain [2].

The relationship between blood and CSF is close, as CSF is considered an ultrafiltrate of blood plasma, the liquid portion of blood. Most of the CSF is produced through a network of cuboidal cells known as the choroid plexus, a type of epithelial cells with microvilli that cover the ventricles of the brain and allow the filtration of plasma through them [3]. This forms what is known as blood-CSF barrier, shown in Figure 1.

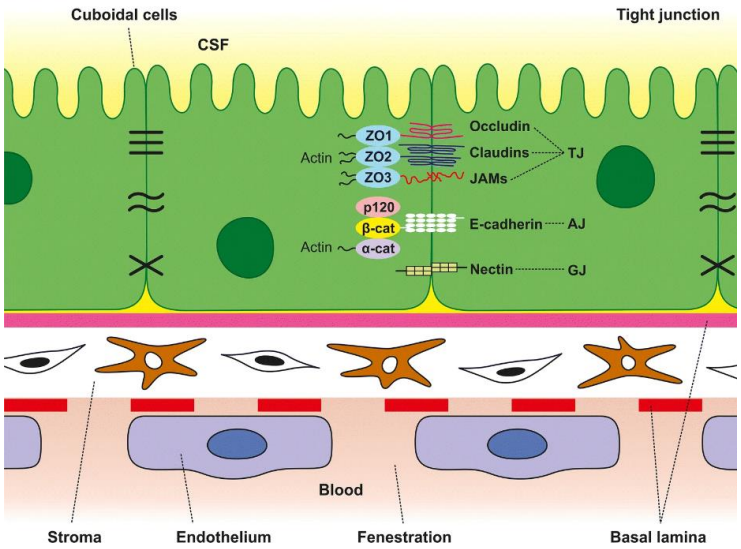


Figure 1. Schematic image of the blood-CSF barrier[4].

Some time ago, came about the need to develop new technologies to manage and store data from all the components that biological systems have, such as biofluids. From here, “omic” sciences emerged, technologies that can be applied to biological systems to obtain a measurement of the biological molecules they contain, to obtain relevant information about the underlying biology at a resolution that has never before been possible [5].

The "omic" sciences include all sciences ending with the suffix -omics, such as genomics, transcriptomics, proteomics and metabolomics. The objective of omic sciences is to identify, characterize, and quantify all biological molecules that are involved in the structure, function, and dynamics of a cell, tissue, or organism [6].

Genomics is the study of the structure, function, evolution, mapping, and editing of genomes. A genome is the complete set of DNA (including all of its genes) in an organism. Close to genome, we have transcriptomics which is the science dedicated to the identification and quantification of the transcriptome. The transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue [5].

Proteomics is the study of the proteins present in a cell, tissue, or organism. Therefore, proteomics is the application of technologies for the identification and quantification of proteins present in a cell, tissue, or organism [7]. Proteomics is one of the most complex omic sciences to study because the proteome is highly dynamic due to complex regulatory systems that control protein expression levels [7].

Metabolomics involves the application of technologies for the identification and quantification of metabolome, the complete set of small molecule metabolites found within a biological sample, including metabolic intermediates in carbohydrate, lipid, amino acid, nucleic acid, and other biochemical pathways, along with hormones and other signaling molecules, as well as exogenous substances such as drugs and their metabolites. It is subject to many factors such as diet, stress, physical activity, diseases, etc., of the studied system[5].

As a subset of metabolomics, we could find lipidomics (the study of lipids), volatilomics (the study of volatile metabolites), and ionomics (the study of chemical elements ions).

We can say that metabolism is the set of chemical reactions that take place within living organisms and that aim at the synthesis or degradation of substances[8]. These reactions generally occur in chain, forming what we call "metabolic pathways," a series of reactions that lead to the conversion of a substance to a final product [8].

In Figure 2, we can see a general map where an overview of the metabolic pathways of biological systems is represented schematically, extracted from the iPath3.0 website.[9]. In this map, it is represented how metabolites, proteins and different metabolic pathways are interconnected, and the color classification of each pathway according to the type of metabolism to which they belong.

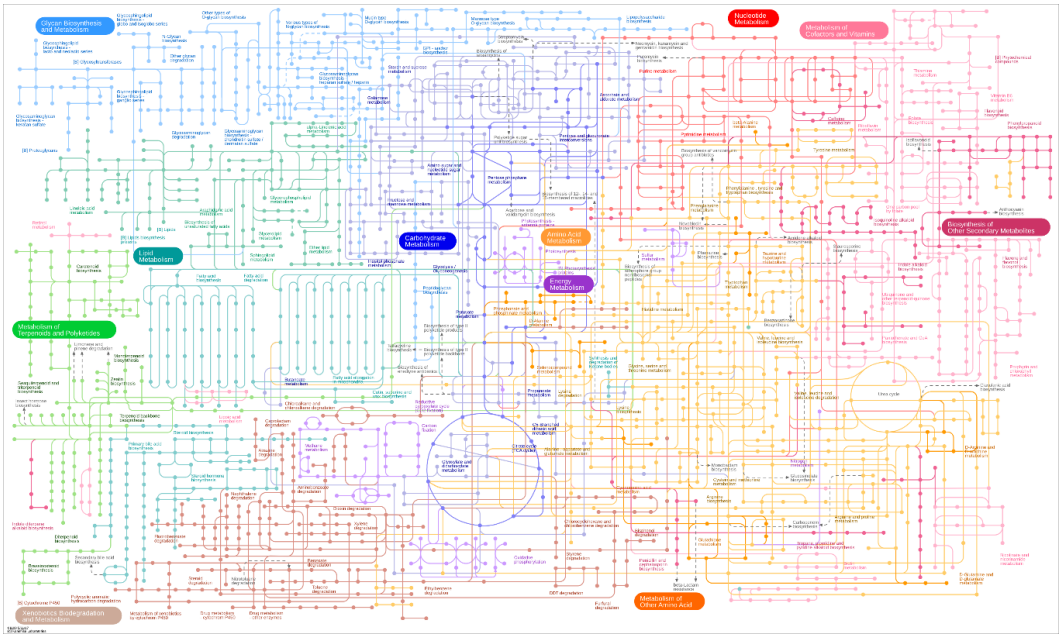


Figure 2. Diagram with the metabolic pathways. Image from iPath 3.0 [9]

.

Thanks to the study of metabolic pathways in organisms, we can get an insight of the biochemical processes that take place. As a result, we can identify the main metabolites, proteins genes and transcripts involved in these pathways with the aim of studying associated diseases, identifying the mechanisms involved, or understanding the effect of drugs on specific pathways and, therefore, verifying their effectiveness [10].

Biomedical engineers in omics sciences play a key role in analyzing complex biological data by cleaning and organizing it, applying statistical methods, and using machine learning to identify patterns and predict outcomes. They design and manage large databases to store and integrate various types of omics data, facilitating comprehensive insights into biological processes.

Given the large amount of omics data available spread across various databases performing multi-omics analysis present a significant challenge due to its disorganized, decentralized and outdated nature. This problem arises from the existence of numerous databases that collect this different information, often with incomplete records, duplicate entries and is not up to date with the new publications. Furthermore, most sources are general and fail to provide detailed omics information specific to biofluids, or the available information is limited. This is why the Life Fluids (www.lifefluids.com) project has been developed. It has the aim to create a set of databases for human biofluids, where information on metabolites, proteins and associated genes, chemical elements, and transcripts of these biofluids can be found, solving the previously mentioned problem.

2. Objectives

The main objective is to obtain an integrative multi-omics profiling and pathway analysis for human biofluids, specifically blood and cerebrospinal fluid (CSF).

The specific objectives are:

Objective 1: **Obtain merged and curated databases of proteins, metabolites, transcripts and chemical elements for blood.**

To achieve this objective, I needed to identify the available sources of information on the composition of blood for each of the omics included (metabolites, proteins and associated genes, transcripts, and chemical elements), extract the data and translate it to specific identifiers for each omics. This step will imply large amounts of data to be merged. Then, I needed to identify the attributes that we will need to extract from these sources in order to create a comprehensive database of the composition of the two biofluids. Finally, the created databases for the metabolites, proteins and associated genes, and chemical elements for blood and cerebrospinal fluid, are manually checked.

Objective 2: **Obtain merged and curated databases of proteins, metabolites, transcripts and chemical elements for CSF.**

To achieve this objective, I will follow the same procedure as Objective 1 but applied to CSF biofluid.

Objective 3: Pathway analysis of LifeFluids Blood databases created. Analyze the created databases for blood to perform a multi-omics profiling.

To achieve this objective, I used the created databases, along with the previous work, to analyze the different metabolic pathways and the correlations between them. Identify the metabolic pathways present in the blood biofluid.

Objective 4: Pathway analysis of LifeFluids CSF database created. Analyze the created databases for blood to perform a multi-omics profiling.

To achieve this objective, I will follow the same procedure as Objective 3 but applied to CSF biofluid.

3. Methodology

During this thesis, two omics databases of blood and cerebrospinal fluid composition were developed. To do so, first I divided the composition of these fluids into metabolites, proteins and transcripts, and chemical elements.

Python was used for the creation and manipulation of all the database's tables. I used the functions of the Pandas Python library, that allows the manipulation of data in a *dataframe* format and has enabled the combination of the necessary data from the different information sources, to finally obtain each of the tables that constitute the database.

For all the tables, a 3-step methodology was used:

- Omics data translation (to a specific identifier for each omics used) and merging on the metabolites data.
- Attributes retrieval.
- Manual curation to eliminate duplicates and retrieve missing information.

This is a data harmonization process, which has been performed because each source of information has its own compound identifiers, and in some cases, different nomenclature.

To perform the multi-omics profiling and pathway analysis, the web application iPath 3.0 [9] was used. This tool allows for the observation of metabolic pathways involved in the biofluids by entering the data obtained during the creation of the metabolite and protein databases for both blood and cerebrospinal fluid.

3.1. Sources of information for databases

The content of this section is confidential.

3.2. New comprehensive blood and cerebrospinal fluid databases

The first step in developing each one of the databases is to choose the attributes that each entry or compound will have in the databases of proteins, metabolites, transcripts, and chemical elements. For each omics, specific attributes are needed to describe each compound accurately.

The attributes will represent the characteristics or information we will have for each one of the entries in the database. Once the attributes are chosen, it will be necessary to search for the sources of information from which they will be extracted, ensuring that the sources are reliable and up to date in order to guarantee the quality of the data included in our database.

3.2.1. Proteins

The key attribute that defines the proteins in the database is the Uniprot identifier. The Uniprot ID, is a unique and universal identifier assigned to proteins in the UniProt database. This identifier provides a minimal description of the protein sequence and enables the search of information about it in other sources[29].

Starting with the UniProt, the rest of the attributes, as shown in table 4, of the protein database have been extracted.

Table 1. Proteins attributes description

Attribute	Description
Life Fluids ID	Identifier within the Life Fluids database
Protein Name	Protein name
Gene symbol	It's the gene associated with the protein short form abbreviation approved by the HGNC [30]
EV	If it has been associated with extracellular vesicles
Uniprot ID	The Uniprot code that identifies this protein
Diseases	The diseases associated with the protein
Database	The databases from which the protein has been extracted
Literature	The articles from which the protein has been extracted or identified

The attributes shown in Table 4 were extracted from the following sources:

First, human proteins were downloaded from Uniprot (https://www.uniprot.org). Next, proteins were downloaded from Marker DB (https://markerdb.ca/downloads) along with the diseases or conditions they are associated with. This provides a sort of dictionary of proteins and associated diseases, containing the attributes shown in the previous table.

Then, files with blood proteins were downloaded from the sources listed in section 3.1. a), obtaining the Uniprot identifiers of the blood proteins.

The next step was to match each file containing Uniprot IDs from the different sources with the previously created "dictionary". If a match was found, the data for: Uniprot ID, Name, Gene Symbol, and Diseases from the dictionary were copied, adding for each entry the database from which it was extracted in the field 'Database'.

Finally, the columns 'EV' and 'Literature' were filled in at the end, adding new proteins found in the literature and indicating in the 'EV' field if the study is related to extracellular vesicles.

In Figure 3, we can see the sources used to extract the information for the proteins, starting from the Uniprot ID of each one.

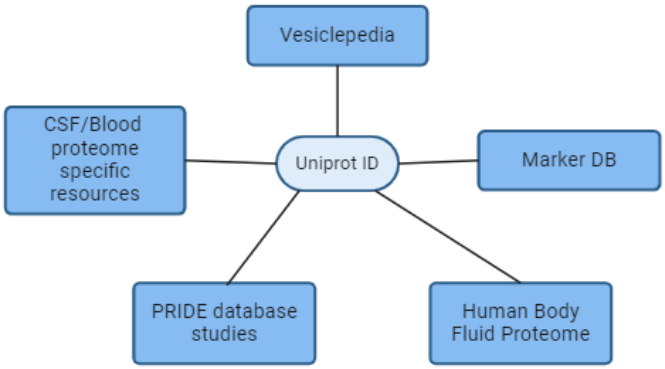


Figure 3. Diagram with the sources from which the proteins were extracted

In Figure 4, we can see the first rows of the blood proteins Life Fluids database, with its corresponding attributes.

	NumberID	Name	Gene_Symbol	EV	Uniprot	Diseases	Database	Literature
2	LFPR00001	HLA class I hi	HLA-A		P04439	DISEASE: A	HPPA | HBFP | VSCP	LTPR0080 | LTPR0078 | LTPR0075 | LTPR0072 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 | LTPR0055 |
3	LFPR00002	HLA class I hi	HLA-B		P01889		HPPA | HBFP | VSCP	LTPR0079 | LTPR0078 | LTPR0075 | LTPR0072 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 | LTPR0055 |
4	LFPR00003	Alpha-1-antit	SERPINA1		P01009	DISEASE: A	HPPA | HBFP | MDB |	LTPR0083 | LTPR0080 | LTPR0079 | LTPR0077 | LTPR0075 | LTPR0071 | LTPR0069 | LTPR0066 | LTPR0065 |
5	LFPR00004	Basigin (5F7)	BSG		P35613		HPPA | HBFP | VSCP	LTPR0082 | LTPR0081 | LTPR0079 | LTPR0077 | LTPR0075 | LTPR0068 | LTPR0069 | LTPR0056 | LTPR0065 |
6	LFPR00005	Alpha-amino	ALDH7A1		P49419		HPPA | HBFP | VSCP	LTPR0077 | LTPR0075 | LTPR0056 | LTPR0065 | LTPR0053 | LTPR0047 | LTPR0046
7	LFPR00006	Neurexin-1-b	NRXN1	Yes	P58400		VSCP	
8	LFPR00007	Complement	C1R		P00736		HPPA | HBFP | VSCP	LTPR0083 | LTPR0072 | LTPR0071 | LTPR0063 | LTPR0058 | LTPR0057 | LTPR0053 | LTPR0050 | LTPR0047
9	LFPR00008	Tenascin-X (T	TNXB		P22105		HPPA | HBFP | VSCP	LTPR0077 | LTPR0059
10	LFPR00009	Transmembr	TMPRSS2		O15393	DISEASE: L	HBFP | VSCP	LTPR0053
11	LFPR00010	Cathepsin D (	CTSD		P07339		HPPA | HBFP | VSCP	LTPR0082 | LTPR0080 | LTPR0043 | LTPR0079 | LTPR0078 | LTPR0077 | LTPR0076 | LTPR0075 | LTPR0072 |
12	LFPR00011	Angiotensin-	ACE		P12821	DISEASE: V	HPPA | HBFP | MDB |	LTPR0069 | LTPR0058 | LTPR0053
13	LFPR00012	CD9 antigen (	CD9		P21926	DISEASE: E	HPPA | HBFP | VSCP	LTPR0077 | LTPR0075 | LTPR0072 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 | LTPR0057 | LTPR0053 |
14	LFPR00013	Haptoglobin	HP		P00738	DISEASE: C	HPPA | HBFP | VSCP	LTPR0083 | LTPR0080 | LTPR0079 | LTPR0077 | LTPR0075 | LTPR0071 | LTPR0069 | LTPR0056 | LTPR0065 |
15	LFPR00014	Renin recept	ATP6AP2		O75787	DISEASE: N	HPPA | HBFP | VSCP	LTPR0082 | LTPR0079 | LTPR0077 | LTPR0065 | LTPR0059 | LTPR0057 | LTPR0053 | LTPR0048 | LTPR0047 |
16	LFPR00015	HLA class I hi	HLA-C		P10321	DISEASE: R	HPPA | VSCP	LTPR0075 | LTPR0072 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 | LTPR0057 | LTPR0055 | LTPR0054 |
17	LFPR00016	Radixin	RDX		P35241	DISEASE: N	HPPA | HBFP | VSCP	LTPR0043 | LTPR0079 | LTPR0077 | LTPR0075 | LTPR0068 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 |
18	LFPR00017	Acid ceramid	ASAH1		Q13510	DISEASE: I	HPPA | HBFP | VSCP	LTPR0082 | LTPR0080 | LTPR0079 | LTPR0078 | LTPR0077 | LTPR0075 | LTPR0069 | LTPR0056 | LTPR0065 |
19	LFPR00018	Receptor-typ	PTPRS		Q13332		HPPA | HBFP | VSCP	LTPR0082 | LTPR0043 | LTPR0057 | LTPR0053 | LTPR0048 | LTPR0046
20	LFPR00019	Hypoxia up-r	HYOU1		Q9Y4L1		HPPA | HBFP | VSCP	LTPR0079 | LTPR0077 | LTPR0075 | LTPR0069 | LTPR0056 | LTPR0065 | LTPR0059 | LTPR0057 | LTPR0053 |
21	LFPR00020	Latent-transf	LTBP4		Q8N2S1		HPPA | HBFP | VSCP	LTPR0077

Figure 4. Screenshot of the first 20 rows of the blood proteins database

Finally, the attribute Life Fluids ID is the identifier assigned to the elements that are part of the Life Fluids project databases. In Annex 1, it can be seen how this identifier is calculated for each of the omics.

3.2.2. Metabolites

The most significant attribute for metabolites in this database is the InChIKey, the International Chemical Identifier Key, developed by the IUPAC (the International Union of Pure and Applied Chemistry). It is a compact version of the InChI code, which provides information about molecular connectivity, charge, stereochemistry, isotopic enrichment, hydrogen atom positions, and bonding in metal compounds, organized in a hierarchical, layered format [31].

The InChIKey is a condensed, 27-character version of this descriptive identifier, making it more practical for use in databases or internet searches [31].

Using the InChIKey as a reference, the remaining attributes of the metabolites database have been extracted from various sources of information described in later sections. The rest of the attributes are the following:

Table 2. Metabolites attributes description

Attribute	Description
Life Fluids ID	Identifier within the Life Fluids Database
Metabolite Name	The metabolite's common name
RefMet ID	RefMet (Reference Set of Metabolite Names) is a highly curated database of standardized names for metabolites [32].
Molecular_Weight	The molecular weight of the metabolite
Monoisotopic_Mass	Is the mass of the metabolite when it is composed of the most abundant isotopes of its constituent elements.
Formula	The molecular formula that describes the metabolite.
Xlogp	The partition coefficient. Is the ratio of a substance in one medium or phase to the concentration in a second phase when the two concentrations are at equilibrium[33]. A high XLogP value indicates that the molecule is more lipophilic, while a low XLogP indicates that it is more hydrophilic.
InChIKey	The InChIKey that describes the metabolite
InChI	The InChI code that describes the metabolite
SMILES	The Simplified Molecular Input Line Entry System is a written form of representing and describing the structure and composition of molecules in a simple way using ASCII table characters[34].
HMDB	The Human Metabolome Database identification number
PubChemID	The compound identification number (CID) assigned to the metabolite within the PubChem repository.
KEGG_ID	The metabolite identification number assigned to the metabolite in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[35].
Bin ID	The identifier assigned in BinBase[36], a database based on GC-MS untargeted metabolomics
LipidMAPS	The identifier assigned to the metabolite within the Lipid MAPS (Metabolites and Pathways Strategy) website and database, which is responsible for lipid curation, classification, and nomenclature[37].
CAS	The CAS (Chemical Abstract Service) registry number. CAS REGISTRY compiles substances identified in the literature from 1957 to the present. Each chemical substance it contains has a unique numeric identifier, with up to 10 digits divided into three parts separated by hyphens, and with no chemical meaning[38].
ChEBI	The identifier of the metabolite within the Chemical Entities of Biological Interest (ChEBI) dictionary, a dictionary of molecular entities focused on small chemical compounds[39].
DrugBank_ID	The metabolites identifiers in DrugBank database, an online database containing information on drugs and drug targets[40]
Kingdom / Superclass / Class / Subclass	These four columns describe the taxonomic classification of the metabolite.
Disease	The diseases associated with the metabolite
Database	The databases from which the metabolite has been extracted
Literature	The articles from which the metabolite has been extracted or identified.

The attributes shown in the table, were extracted from the following sources:

PubChem (https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi) is a public repository for information on chemical substances and their biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH)[41].

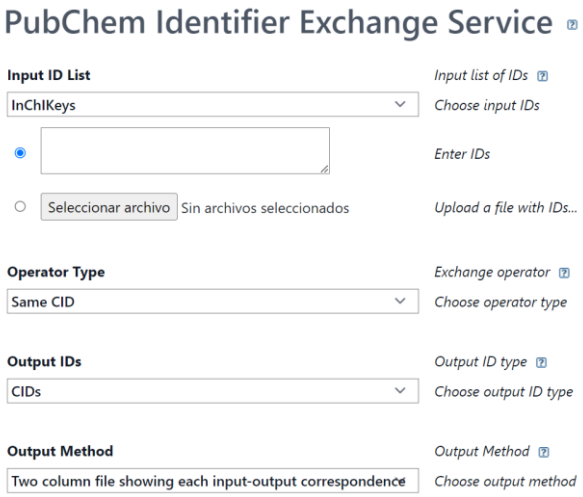


Figure 5. Screenshot of the PubChem Identifier Exchange Service interface through which some of the data for our database was obtained

This website (figure 5 shows its interface) allows you to enter a series of identifiers as input; in our case, we will choose the InChIKey. Next, you enter the identifiers in text format and select the type of output you want, which in our case will be the CID (the PubChem identifier), InChI code, SMILES, and the compound's synonyms. Finally, you click on 'submit job,' and it returns a text document in list format with the InChIKeys entered as input along with their corresponding output.

PubChem allows translation between these types of identifiers: CID, InChI, InChIKey, SMILES, synonyms and IUPAC Names.

The Chemical Translation Service (CTS) (https://cts.fiehnlab.ucdavis.edu/batch) works in a similar way to PubChem. Figure 6 shows the interface of this tool, which allows to enter a list of compound identifiers and translates them to one or several types of compounds at once. That is to say, we can enter a list of InChIKeys, and CTS will translate them, for example, to their corresponding chemical name, ChEBI, InChI code, etc. Unlike PubChem, CTS allows for batch conversion and also enables translation to more types of compounds.

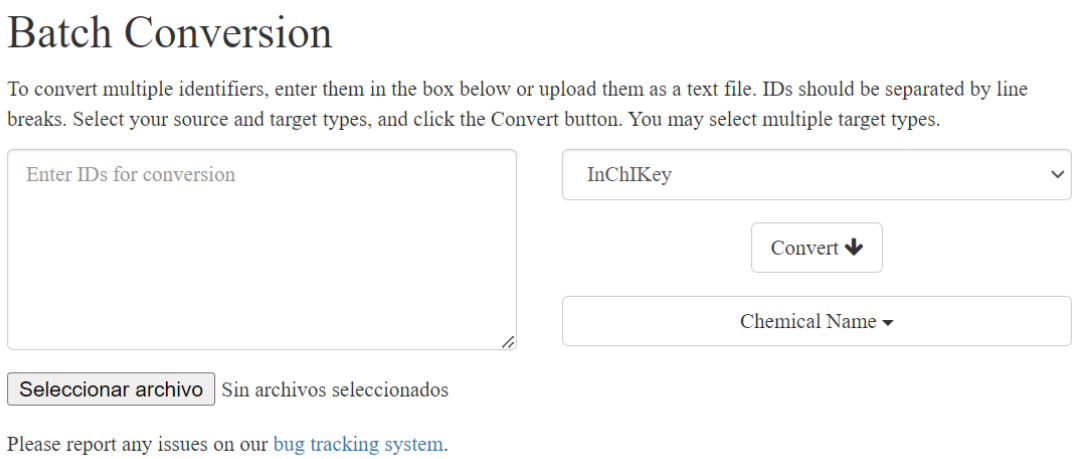


Figure 6. Screenshot of the PubChem Identifier Exchange Service interface through which some of the data for our database was obtained.

Finally, ClassyFire (http://classyfire.wishartlab.com/) is a web-based application for automated structural classification of chemical entities. It provides a hierarchical chemical classification of chemical entities, based on a chemical taxonomy [42].

Works similarly to the two previous ones. In ClassyFire, the users input the InChI of the compounds from which they want to obtain taxonomic data, and the web returns as an output a list containing the superclass, class, subclass, and kingdom of the introduced compounds.

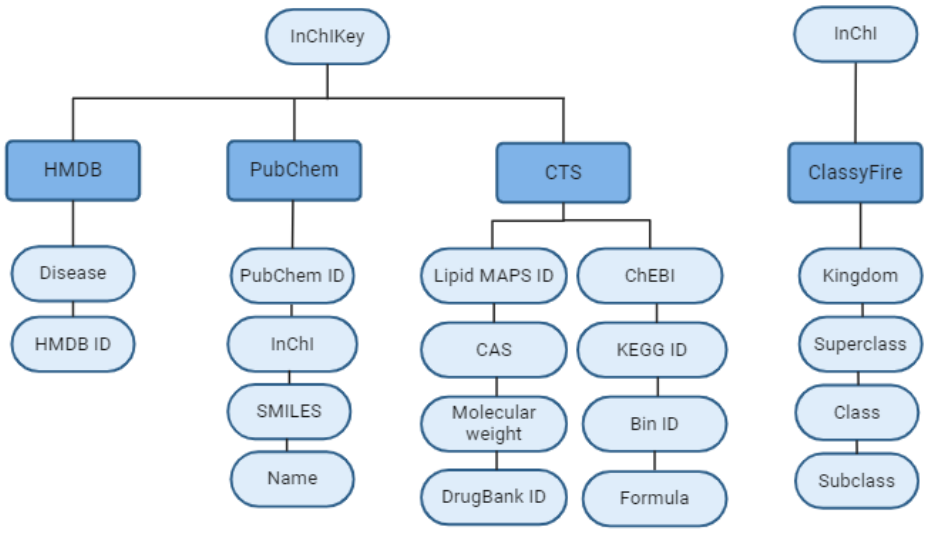


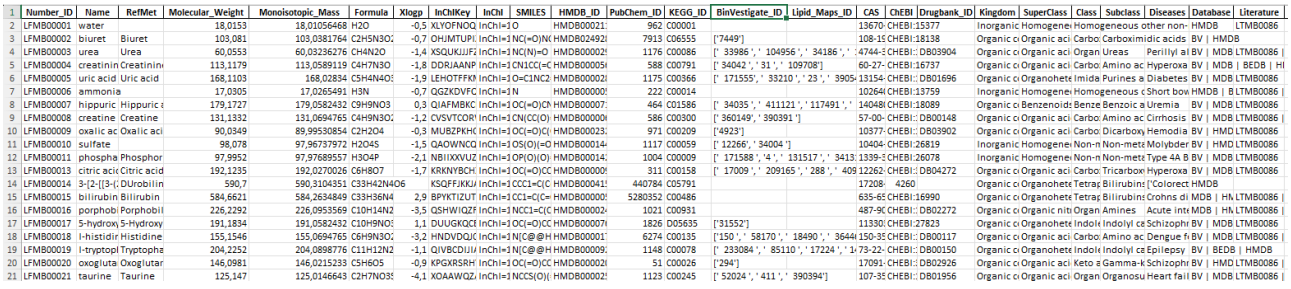
Figure 7. Diagram made with BioRender [43] *showinga summary from where the metabolites' attributes were extracted*

In Figure 7 is shown a summarized and visual representation of the sources from which the data were taken to fill the fields of the metabolites database, starting with the InChIKey for each of the metabolites. There are attributes that do not appear in the previous figure

because they are extracted from a different source. The attributes xlogP and monoisotopic mass were extracted using Python with the library *pubchempy*, which allows obtaining data such as the monoisotopic mass, molecular weight or xlogP from the PubChem ID (CID) of a compound.

The RefMet attribute was extracted from the Metabolomics Workbench website (https://www.metabolomicsworkbench.org), where entering the names of the metabolites returns a list with the standardized name. Finally, the Literature and Database attributes have been updated each time new information from a new source was added.

In Figure 8, we can see the first 20 rows of the Life Fluids database of blood metabolites.



	Number_ID	Name	RefMet	Molecular_Weight	Monoisotopic_Mass	Formula	Xlogp	InChIKey	InChI	SMILES	HMDB_ID	PubChem_ID	KEGG_ID	BinVestigate_ID	Lipid_Maps_ID	CAS	ChEBI	Drugbank_ID	Kingdom	SuperClass	Class	Subclass	Diseases	Database	Literature
2	LFMB00001	water		18,0153	18,01056468	H2O	-0,5	XLYOFNOQ	InChI=1	O	HMDB00021	962	C00001			13670-	CHEBI:15377		Inorganic	Homogene	Homogeneous other non-			HMDB	LTMB0086	
3	LFMB00002	biuret	Biuret	103,081	103,0381764	C2H5N3O2	-0,7	OHJMTUPI	InChI=1	NC(=O)N(	HMDB02492	7913	C06555	['7449']		108-19-	CHEBI:18138		Organic co	Organic aci	Carboxi	Carboximidic acids		BV | HMDB		
4	LFMB00003	urea	Urea	60,0553	60,03236276	CH4N2O	-1,4	XSQUKJJJF	InChI=1	NC(N)=O	HMDB00002	1176	C00086	[' 33986 ', ' 104956 ', ' 34186 ', '		4744-9	CHEBI::	DB03904	Organic co	Organic aci	Organ	Ureas	Perillyl al	BV |	MDB	LTMB0086	
5	LFMB00004	creatinin	Creatinine	113,1179	113,0589119	C4H7N3O	-1,8	DDRJAANP	InChI=1	CN1CC(=C	HMDB00005	588	C00791	[' 34042 ', ' 31 ', ' 109708']		60-27-	CHEBI:16737		Organic co	Organic aci	Carbo	Amino ac	Hyperoxa	BV |	MDB |	BEDB |	HI
6	LFMB00005	uric acid	Uric acid	168,1103	168,02834	C5H4N4O3	-1,9	LEHOTFFK	InChI=1	O=C1NC2	HMDB00002	1175	C00366	[' 171555', ' 33210 ', ' 23 ', ' 3905		13154-	CHEBI::	DB01696	Organic co	Organohete	Imida	Purines a	Diabetes	BV |	MDB	LTMB0086	
7	LFMB00006	ammonia		17,0305	17,0265491	H3N	-0,7	QGZKDVFQ	InChI=1	N	HMDB00005	222	C00014			10264(	CHEBI:13759		Inorganic	Homogene	Homogeneous c	Short bow	HMDB |	B	LTMB0086	
8	LFMB00007	hippuric	Hippuric a	179,1727	179,0582432	C9H9NO3	0,3	QIAFMBKC	InChI=1	OC(=O)CN	HMDB00007	464	C01586	[' 34035 ', ' 411121 ', ' 117491 ', '		14048(	CHEBI:18089		Organic co	Benzenoids	Benze	Benzoic a	Uremia	BV |	MDB	LTMB0086	
9	LFMB00008	creatine	Creatine	131,1332	131,0694765	C4H9N3O2	-1,2	CVSVTCOR	InChI=1	CN(C(O)=	HMDB00000	586	C00300	[' 360149', ' 390391 ']		57-00-	CHEBI::	DB00148	Organic co	Organic aci	Carbo	Amino ac	Cirrhosis	BV |	MDB	LTMB0086	
10	LFMB00009	oxalic ac	Oxalic aci	90,0349	89,99530854	C2H2O4	-0,3	MUBZPKHO	InChI=1	OC(=O)C(	HMDB00023	971	C00209	['4923']		10377-	CHEBI::	DB03902	Organic co	Organic aci	Carbo	Dicarboxy	Hemodia	BV |	HMD	LTMB0086	
11	LFMB00010	sulfate		98,078	97,96737972	H2O4S	-1,5	QAOWNCQ	InChI=1	OS(O)(=O	HMDB00014	1117	C00059	[' 12266', ' 34004 ']		10404-	CHEBI:26819		Inorganic	Homogene	Non-n	Non-mete	Molybder	BV |	HMD	LTMB0086	
12	LFMB00011	phospha	Phosphor	97,9952	97,97689557	H3O4P	-2,1	NBIIXXVUZ	InChI=1	OP(O)(O)=	HMDB00014	1004	C00009	[' 171588 ', '4 ', ' 131517 ', ' 3413		1339-2	CHEBI:26078		Inorganic	Homogene	Non-n	Non-mete	Type 4A B	BV |	MDB	LTMB0086	
13	LFMB00013	citric acic	Citric acid	192,1235	192,0270026	C6H8O7	-1,7	KRKNYBCH	InChI=1	OC(=O)CC	HMDB00000	311	C00158	[' 17009 ', ' 209165 ', ' 288 ', ' 409		12262-	CHEBI::	DB04272	Organic co	Organic aci	Carbo	Tricarboxi	Hyperoxa	BV |	MDB	LTMB0086	
14	LFMB00014	3-[2-[[3-(	DUrobilin	590,7	590,3104351	C33H42N4O6		KSQFFJKJ	InChI=1	CCC1=C(C	HMDB00041	440784	C05791			17208-	4260		Organic co	Organohete	Tetrap	Bilirubins	['Colorect	HMDB			
15	LFMB00015	bilirubin	Bilirubin	584,6621	584,2634849	C33H36N4	2,9	BPYKTIZUT	InChI=1	CC1=C(C=	HMDB00005	5280352	C00486			635-65	CHEBI:16990		Organic co	Organohete	Tetrap	Bilirubins	Crohns di	MDB |	HM	LTMB0086	
16	LFMB00016	porphobi	Porphobil	226,2292	226,0953569	C10H14N2	-3,5	QSHWIQZF	InChI=1	NCC1=C(C	HMDB00002	1021	C00931			487-90	CHEBI::	DB02272	Organic co	Organic niti	Organ	Amines	Acute inte	MDB |	HM	LTMB0086	
17	LFMB00017	5-hydroxy	5-Hydroxy	191,1834	191,0582432	C10H9NO3	1,1	DUUGKQCE	InChI=1	OC(=O)CC	HMDB00007	1826	D05635	['31552']		11330(	CHEBI:27823		Organic co	Organohete	Indole	Indolyl ca	Schizophr	BV |	MDB	LTMB0086	
18	LFMB00018	l-histidir	Histidine	155,1546	155,0694765	C6H9N3O2	-3,2	HNDVDQJC	InChI=1	N[C@@H	HMDB00001	6274	C00135	['150 ', ' 58170 ', ' 18490 ', ' 3644		150-35	CHEBI::	DB00117	Organic co	Organic aci	Carbo	Amino ac	Dengue fe	BV |	MDB	LTMB0086	
19	LFMB00019	l-tryptopl	Tryptopha	204,2252	204,0898776	C11H12N2	-1,1	QIVBCDIJ	InChI=1	N[C@@H	HMDB00009	1148	C00078	[' 233084 ', ' 85110 ', ' 17224 ', ' 1		73-22-	CHEBI::	DB00150	Organic co	Organohete	Indole	Indolyl ca	Epilepsy	BV |	BEDB |	HMDB		
20	LFMB00020	oxoglutar	Oxoglutar	146,0981	146,0215233	C5H6O5	-0,9	KPGXRSRH	InChI=1	OC(=O)CC	HMDB00002	51	C00026	['294']		17091-	CHEBI::	DB02926	Organic co	Organic aci	Keto a	Gamma-k	Schizophr	BV |	HMD	LTMB0086	
21	LFMB00021	taurine	Taurine	125,147	125,0146643	C2H7NO3	-4,1	XOAAWQZ	InChI=1	NCCS(O)(	HMDB00002	1123	C00245	[' 52024 ', ' 411 ', ' 390394']		107-35	CHEBI::	DB01956	Organic co	Organic aci	Organ	Organosu	Heart fail	BV |	MDB	LTMB0086	

Figure 8. Screenshot of the first 20 rows of the blood metabolites database

3.3.3. Transcripts

First, the complete set of genes from HUGO Gene Nomenclature Committee or HGNC (www.genenames.org) was downloaded, which will serve as a dictionary. Second, the set of genes related to blood or cerebrospinal fluid was downloaded from the exRNA Atlas website (www.exrna-atlas.org/). From here, with the help of a Python script, the two datasets were merged, creating a new one that contained the exRNA set genes that matched 70% with the HGNC genes.

Table 3. Transcripts attributes description

Attribute	Description
Life Fluids ID	Identifier within the Life Fluids Database
Approved symbol	The HGNC approved gene symbol.
Approved name	The gene name approved by the HGNC[30]
HGNC_ID	A unique ID created by the HGNC for every approved symbol
Location	The location or region of the gene on the chromosome
miR	Micro-non-protein coding genes that encode microRNAs (miRNA)
Alias symbols	Other symbols used to name this gene
Database	The databases from which the transcript has been extracted

In Figure 9, you can see what the first rows of the database look like.

NumberID	Approved symbol	Approved name	HGNC ID	Location	miR	Alias symbols	Database
LFTR00001	MIRLET7A1	microRNA let-7a-1	HGNC:31476	9q22.32	MI0000060	hsa-let-7a-1	exRNA | VSCP
LFTR00002	MIRLET7A2	microRNA let-7a-2	HGNC:31477	11q24.1	MI0000061	hsa-let-7a-2	exRNA | VSCP
LFTR00003	MIRLET7A3	microRNA let-7a-3	HGNC:31478	22q13.31	MI0000062	hsa-let-7a	exRNA
LFTR00003	MIRLET7A3	microRNA let-7a-3	HGNC:31478	22q13.31	MI0000062	hsa-let-7a-3	exRNA | VSCP
LFTR00004	MIRLET7B	microRNA let-7b	HGNC:31479	22q13.31	MI0000063	hsa-let-7b	exRNA | VSCP
LFTR00005	MIRLET7C	microRNA let-7c	HGNC:31480	21q21.1	MI0000064	hsa-let-7c	exRNA | VSCP
LFTR00006	MIRLET7D	microRNA let-7d	HGNC:31481	9q22.32	MI0000065	hsa-let-7d	exRNA | VSCP
LFTR00007	MIRLET7E	microRNA let-7e	HGNC:31482	19q13.41	MI0000066	hsa-let-7e	exRNA | VSCP
LFTR00008	MIRLET7F1	microRNA let-7f-1	HGNC:31483	9q22.32	MI0000067	hsa-let-7f-1	exRNA | VSCP
LFTR00009	MIRLET7F2	microRNA let-7f-2	HGNC:31484	Xp11.22	MI0000068	hsa-let-7f	exRNA
LFTR00009	MIRLET7F2	microRNA let-7f-2	HGNC:31484	Xp11.22	MI0000068	hsa-let-7f-2	exRNA | VSCP
LFTR00010	MIR21	microRNA 21	HGNC:31586	17q23.1	MI0000077	hsa-mir-21	exRNA | VSCP
LFTR00011	MIR25	microRNA 25	HGNC:31609	7q22.1	MI0000082	hsa-mir-25	exRNA | VSCP
LFTR00012	MIR26B	microRNA 26b	HGNC:31612	2q35	MI0000084	hsa-mir-26b	exRNA | VSCP
LFTR00013	MIR28	microRNA 28	HGNC:31615	3q28	MI0000086	hsa-mir-28	exRNA | VSCP
LFTR00014	MIR30A	microRNA 30a	HGNC:31624	6q13	MI0000088	hsa-mir-30a	exRNA | VSCP
LFTR00015	MIR31	microRNA 31	HGNC:31630	9p21.3	MI0000089	hsa-mir-31	exRNA | VSCP
LFTR00016	MIR32	microRNA 32	HGNC:31631	9q31.3	MI0000090	hsa-mir-32	exRNA | VSCP
LFTR00017	MIR33A	microRNA 33a	HGNC:31634	22q13.2	MI0000091	hsa-mir-33a	exRNA

Figure 9. Screenshot of the first rows of the blood transcripts Life Fluids database

3.3.4. Chemical Elements

The tables of chemical elements share most attributes with those of metabolites. They differ in that chemical elements do not have taxonomic attributes (kingdom, superclass, class and subclass), instead, they have other attributes specific to chemical elements: Usual Form, Usual Form Formula, KEGG ID for the Usual Form, Main Group and Group. These attributes are explained below.

Table 4. Chemical elements attributes description

Attribute	Description
Usual_Form	The form in which the element is commonly found in nature
Usual Form Formula	The formula of the element when it is in its usual form
KEGG_ID For Usual Form	The identification number assigned to the metabolite in the Kyoto Encyclopedia of Genes and Genomes (KEGG)
Main Group	They can be metals, metalloids, and non-metals
Group	The group to which they belong according to their chemical properties and the position they occupy in the periodic table.

These attributes have been extracted from the Human Metabolome Database (HMDB). Additionally, two studies have been consulted to complete the chemical elements databases, explained in section 3.1. d).

Figure 10 shows the first rows of the blood chemical elements Life Fluids database.

	NumberID	Name	Symbol	Monoisotopic_Mass	Molecular_Weight	InChIKey	InChI	HMDB_ID	PubChem_ID	KEGG_ID	Usual Form	Usual Form Formula	KEGG_ID For Usual Form	CAS_#	ChEBI_ID	Drugbank_ID	Main Group	Group	Diseases	Database	Literature
2	LFCE001	Hydrog	H	1,007825032	1,008	UFHFLCQG	InChI=	HMDB0001	5362549		Diatomic		C00282	12408-(	CHEBI:18276		Non-metal	Alkali Metals		BEDB	
3	LFCE002	Helium	He	4,002603254	4,0026	SWQJXJO	InChI=	HMDB0037	23987	D04420				71086-:	CHEBI:30219		Non-metal	Noble Gas		BEDB | HMDB	
4	LFCE003	Lithiur	Li	7,016003437	7	WHXSMM	InChI=	HMDB0005	3028194	C15473	Cation	Li+		13982-(	CHEBI:30145		Metal	Alkali N	["Alzheir	BEDB | HMDB	
5	LFCE004	Beryllii	Be	9,012183065	9,012183	ATBAMAF	InChI=	HMDB0002	5460467	C16460	Cation	Be+2		7440-41-7			Metal	Alkaline Earth Me		BEDB | HMDB	
6	LFCE005	Boron	B	11,00930536	10,81	ZOXJGFHI	InChI=	HMDB0002	5462311		Cation	B+3	C06266	11129-:	CHEBI:27560		Metalloids	Metalloids		BEDB | HMDB	
7	LFCE007	Nitroge	N	14,003074	14,007	UGRMHO	InChI=	HMDB0001	57370662		Diatomic		C00697	117302	CHEBI:17997		Non-metal	Non-m	["Pregnan	BEDB	
8	LFCE008	Oxygei	O	15,99491462	15,999	MYMOFIZ	InChI=	HMDB0001	159832		Diatomic		C00007	17778-(	CHEBI:15379		Non-metal	Chalcogens		BEDB	
9	LFCE009	fluorin	F	18,99840316	18,99840316	YCKRFDG	InChI=	HMDB0000	5360525		Anion	F-	C00742	7782-41-4			Non-metal	Halogens		BEDB | HMDB	
10	LFCE010	Neon	Ne	19,99244018	20,18	GKAOGPI	InChI=1S/Ne		23935					12794-67-1			Non-metal	Noble Gas		BEDB	
11	LFCE011	Sodiur	Na	22,98976928	22,9897693	KEAYESYH	InChI=	HMDB0000	5360545		Cation	Na+	C01330	17341-:	CHEBI:26708		Metal	Alkali N	["Primary	BEDB | HMDB	
12	LFCE012	Magne	Mg	23,9850417	24,305	FYYHWM	InChI=	HMDB0000	5462224		Cation	Mg+2	C00305	13446-:	CHEBI:25107		Metal	Alkalin	["Alzheir	BEDB | HMDB	
13	LFCE013	Alumir	Al	26,98153853	26,9815	XAGFODP	InChI=	HMDB0001	5359268	C06264				39302-:	CHEBI:289	DB01370	Metal	Other N	["Alzheir	MDB | HM	LTMB0086
14	LFCE014	Silicon	Si	27,97692653	28,085	XUIMIQQ	InChI=	HMDB0002	5461123		Cation	Si+4	C06263	13537-(	CHEBI:27573		Metalloids	Metalloids		BEDB | HMDB	
15	LFCE015	Phospl	P	30,973762	30,9738	OAICVXF	InChI=	HMDB0001	5462309	C06262				7723-14-0			Non-metal	Non-m	["Cerebro	MDB | BEDB | HMDB	
16	LFCE016	Sulfur	S	31,97207117	32,07	NINIDFK	InChI=	HMDB0000	5362487		Anion	S-2	C00087	13981-57-2			Non-metal	Chalcogens		BEDB | HMDB	
17	LFCE017	Chlorir	Cl	34,96885268	35453	ZAMOUSC	InChI=	HMDB0303	5360523		Anion	Cl-	C00698	22537-:	CHEBI:23116		Non-metal	Halogens		HMDB	
18	LFCE018	Argon	Ar	39,96238312	39,9	XKRFYHL	InChI=	HMDB0037	23968					129004	CHEBI:49475		Non-metal	Noble Gas		BEDB | HMDB	
19	LFCE019	Potass	K	38,96370649	39,0983	ZLMJMSJ	InChI=	HMDB0000	5462222		Cation	K+	C00238	1415-9(	CHEBI:26216		Metal	Alkali N	["21-Hydr	BEDB | HMDB	
20	LFCE020	Calciur	Ca	39,96259086	40,08	OYPRJOB	InChI=	HMDB0000	5460341		Cation	Ca+2	C00076	14092-(	CHEBI:293	DB01373	Metal	Alkalin	["Alzheir	BEDB | HMDB	

Figure 10. Screenshot of the first rows of the blood chemical elements Life Fluids database

3.3. Code development

To obtain each one of the tables of the database, a Python code was developed to combine the different sources used. The following list, shows a simple explanation of the steps taken to create each one of the databases, using the example of the CSF proteins database creation, although the same procedure was applied for each one of the databases.

1. Creation of an empty dataframe with the names of the attributes (Figure 11)

Name	RefMet	Molecular weight	Monoisotopic mass	…	Database	Literature

Figure 11. Empty metabolites dataframe with the names of the attributes

2. Filling the empty dataframe with the data from the first source (Figure 12)

Figure 12. Filling the empty metabolites dataframe with the HMDB data

3. Continue filling the table with the information of the different sources and saving the origin of the data in the table (Figure 13)

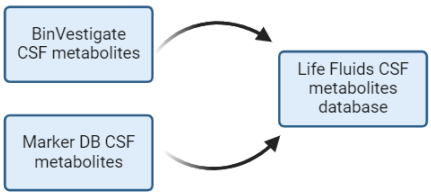


Figure 13. Filling the table with information from other sources

4. Complementing the table with literature from different repositories

5. In case of obtaining new compounds from the literature or other sources, fill the empty fields with the on-line data sources (Figure 14).

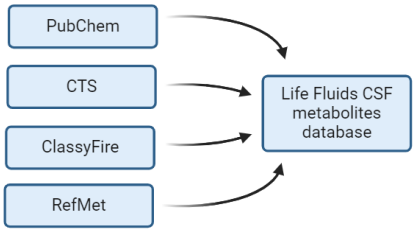


Figure 14. Filling the empty fields with the on-line data sources

6. Manual curation of the data, ensuring the data is correct, and that there are no duplicate entries, removing one of the copies if found.

7. Finally, assign a Life Fluids ID for each entry in the database.

The code shown in Figure 15, is a summarized representation of the Python code that has been used, in general, for merging a Life Fluids database (for example, blood proteins) with another information source (for example, the blood proteins extracted from Vesiclepedia). In the following code, generic variables representing different elements of the table are used: the *columns* represent the attributes of the database, the *identifier* is what will be used to compare the tables, and *lf_database* is the Life Fluids database that is being developed, and *new_source* is the online database or source we are merging it with. The rest of the elements are explained in the comments of the code in Figure 15.

Additionally, the Python scripts developed throughout this work are available in the link found in the Annex 2.

```
# Create an empty dataframe
new_dataframe = pd.DataFrame(columns = ['Column1','Column2','Column3','Column4'])

# Iterate through the rows of the new source to add
for a in new_source.index:

  # The identifier of each entry (for example, the InChIKey for metabolites
  # or the Uniprot ID for proteins)
  identifier = new_source['Column1'][a]

  # Search for the coincidences of the identifier in our database
  coincidences = lf_database['Column1'].str.contains(fr'\b{identifier}\b',na=False)
  total = coincidences.sum()

  # If there is are coincidences, add the information source to our database
  if total.item() >= 1:

    # All the positions where there are matches
    positions = [index for index, element in enumerate(coincidences) if element]
    for b in positions:

      # Let's imagine that column4 is where we store the source of origin.
      # These lines of code update this column based on whether the information was already added.
      if pd.isna(lf_database.loc[b,'Column4']):
        lf_database.loc[b,'Column4'] = new_source.loc[a,'Column4']
      elif str(new_source.loc[a,'Column4']) not in str(lf_database.loc[b,'Column4']):
        lf_database.loc[b,'Column4'] = str(lf_database.loc[b,'Column4'])+
                                       " | " +str(new_source.loc[a,'Column4'])

  # If there are no coincidences, add a new entry to the database
  elif total.item() == 0:
    new_row = pd.DataFrame([{
        'Column1':new_source['Column1'][a],
        'Column2':new_source['Column2'][a],
        'Column3': new_source['Column3'][a],
        'Column4': new_source['Column4'][a]}],
        columns = ['Column1','Column2','Column3','Column4'])
    new_dataframe = pd.concat([new_dataframe, new_row], ignore_index=True)

lf_database = pd.concat([lf_database,new_dataframe], ignore_index=True)
lf_database = lf_database.drop_duplicates().reset_index(drop=True)
```

Figure 158. General structure of the Python code developed.

3.4.1. Python libraries used.

To perform the merging between the different sources to create the databases, I used various Python libraries that provide the necessary tools to perform this task.

a) Pandas

It has been the main library used, because this Python package allows the manipulation and analysis of data in a *dataframe* format. With the tools offered by Pandas, we can read, write, and handle files of different types such as CSV and Excel. Pandas has facilitated working with databases, especially enabling merging, concatenation, and the handling of missing and duplicate entries.

b) Pubchempy

Provides a way to interact with PubChem in Python. It allows, among others, chemical searches by name, substructure and similarity of chemical substances[44]. We used it to get metabolites' properties like the Molecular weight, Molecular formula and xlogp starting with the CID (the PubChem ID) of the compound.

c) Molmass

Molmass was used to calculate the molecular mass or the monoisotopic mass starting with the molecular formula of a compound. However, it has more functionalities, such as obtaining information about the composition of the chemical formula of a compound or simply obtaining information about the physicochemical and descriptive properties of a specific chemical element [45].

3.4.2. API connection to a database

Some of the databases we have used to extract data, such as MarkerDB or Binvestigate, offer an API (Application Programming Interface) service. APIs are mechanisms that allow two software components (in our case, the database from which we want to extract information and the computer with the Python code we will use to extract the data) to communicate via a series of protocols and functions[46].

For example, the MarkerDB API has a series of 'GET' functions that allow the extraction of data on conditions, chemicals, gene biomarkers, protein biomarkers, and karyotype biomarkers[20].

The example in Figure 16 shows how to extract the 'Uniprot ID' attribute of a series of proteins, using the protein name as input, through the MarkerDB's API:

```
# Create a list to store the Uniprot ID obtained through the API
markerDB_uniprots = []

# Iterate through the previously loaded MarkerDB dataframe containing the protein names
for a in MarkerDB.index:
    # GET request to the Marker DB's API
    response = requests.get("http://markerdb.ca/api/v1/proteinapi/proteinrequest?api_key=cc69fb7c1a4809bbae7c0701d5188c17&page=1
                            + MarkerDB.Name[a])
    # Extraction of the proteins data in a .json format
    data = response.json()
    data = data['proteins']
    data = data[0]

    # Store the Uniprot ID of the protein in the previously created list
    markerDB_uniprots.append(data['uniprot_id'])

print(markerDB_uniprots)
```

Figure 16. Code used to extract Uniprot IDs starting with the protein names, via Marker DB's API

3.4. Pathways Analysis

Pathway analysis is a critical step in understanding the biological significance of large scale omics data, such as genomics, proteomics and metabolomics. The Interactive Pathways Explorer or "iPath3", which is a web-based tool for the visualization, analysis, and customization of various metabolic pathway maps [9].

This internet tool offers an interactive format that allows you to navigate and search for information about metabolic pathways in various ways. Firstly, there is a list with four options of interactive metabolic pathway maps that can be explored:

- Metabolic pathways
- Biosynthesis of secondary metabolites
- Microbial metabolism in diverse environments
- Biosynthesis of antibiotics.

In the metabolic pathways option, we can see a map of the human body's metabolic pathways, where the proteins and metabolites of each pathway are color-coded, as seen in figure 2. The map allows you to zoom in and select different metabolites and proteins of each pathway to obtain information about them.

This application allows to create personalized metabolic pathway maps. To do so, I had to enter into the search engine the proteins and metabolites of the biofluids' metabolic pathway I want to evaluate, and the application returns a metabolic pathway map with the introduced data highlighted in the map (as seen in Figure 30 and 31).

To enable iPath 3.0 to highlight specific compounds, we will input the protein data obtained from the databases created. For proteins, we will provide the UniProt IDs, and for metabolites, we will input the KEGG IDs of the biofluid compounds.

4. Results

4.1. Life Fluids' blood database

The content of this section is confidential.

4.2. Life Fluids' cerebrospinal fluid database

The content of this section is confidential.

4.3. Metabolites correlation between the two fluids

The content of this section is confidential.

4.4. Metabolic profiling and pathway analysis

The content of this section is confidential.

4.5. Diseases association

The content of this section is confidential.

4. Discussion

5.1. Creation of the databases

The content of this section is confidential.

5.2. Manual curation

The content of this section is confidential.

5.3. Proteins

The content of this section is confidential.

5.4. Metabolites

The content of this section is confidential.

5.5. Transcripts

The content of this section is confidential.

5.6. Chemical elements

The content of this section is confidential.

5. Conclusions

For this project, I have developed a set of databases of the omic composition of blood and cerebrospinal fluid. For each biofluid, I have created a table of metabolites, proteins, transcripts, and chemical elements.

The objective was to create a single source by merging various information sources such as online databases or repositories of scientific studies, where all the omic data of these two biofluids, blood and cerebrospinal fluid, can be found for subsequent use in metabolic pathway analysis and profiling.

The existing information has been unified, improving previous databases that offered duplicated, decentralized, and non-biofluid-specific information, and thereby meeting the objectives.

As a result, we have obtained the databases that will form part of the Life Fluids project: the blood databases with 90.482 metabolites, 15.763 proteins, 1.463 transcripts, and 98 chemical elements; and the CSF databases with 2.090 metabolites, 5.447 proteins, 1.431 transcripts and 47 chemical elements.

Additionally, the data obtained for proteins and metabolites has been used for a practical purpose: multi-omics profiling and pathway analysis. This allows the study of metabolic pathways involved in different processes occurring in blood, cerebrospinal fluid, or both, to investigate associated diseases or conditions by examining the affected pathways and their potential consequences.

This work highlights the importance of integrating existing omics data into a single source to facilitate subsequent multi-omics analyses. For the future, it is suggested to expand the database to include more biofluids.

6. My contribution

During this work, I contributed to the Life Fluids project by creating databases of the composition of blood and cerebrospinal fluid. The databases contain information on the metabolites, proteins, transcripts, and chemical elements of these two biofluids.

Previous to developing this work, I completed an internship within the same research group (GIOTEC at the Institut d'Investigació Sanitària Pere Virgili), during which I improved the pre-existing databases that had been previously developed by the research group.

The pre-existing databases had information on the composition of urine, breath, saliva, and blood metabolites, although the last one was incomplete.

My contribution also involved improving the previous databases by adding columns for monoisotopic mass and XlogP to the metabolite tables of all biofluids, associating compounds with diseases, and creating the transcript tables for the breath, saliva, and urine databases. Additionally, I completed the blood protein database and created the comprehensive databases for blood and cerebrospinal fluid.

7. Acknowledgements

First, I would like to thank my supervisors Dr. Raquel Cumeras Olmeda PhD and PhD student Maria Llambrich Rodríguez, for guiding me throughout the development of this project and giving me the opportunity to work with this group. I would like to make a special mention for Héctor Villalba Ordóñez, Biomedical Engineer and my project tutor, for supervising my work and helping me solve all my doubts.

Last but not least, I want to thank my family and friends for supporting me throughout the entire process and being there when I needed it.

8. Annex 1

Blood metabolites increase calculation.

Database	Number of entries
HMDB	37.228
Blood_LFMB	90.482

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{90428 * 100}{37228}$ – 100 = 142,90 % ≅ 143 %

Blood proteins increase calculation.

Database	Number of entries
HPPA	4608
Blood_LFPR	15763

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{15763 * 100}{4608}$ – 100 = 242,08 % ≅ 242 %

Blood transcripts increase calculation.

Database	Number of entries
exRNA	1361
Blood_LFTR	1463

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{1463 * 100}{1361}$ – 100 = 7,49 % ≅ 7,5 %

CSF metabolites increase calculation.

Database	Number of entries
HMDB	396
CSF_LFMB	2090

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{2090 * 100}{396}$ – 100 = 427,78 % ≅ 428 %

CSF proteins increase calculation.

Database	Number of entries
CSF-PR	2729
CSF_LFPR	5447

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{5447 * 100}{2729}$ – 100 = 99,60 % ≅ 90 %

CSF transcripts increase calculation.

Database	Number of entries
exRNA	1375
CSF_LFTR	1431

Percentage of increase = $\frac{LF * 100}{DB}$ – 100 = $\frac{1431 * 100}{1375}$ – 100 = 4,07 %

Calculation of the proportion of blood and CSF metabolites

Proportin of blood and CSF metabolites = $\frac{Blood\ metabolites}{CSF\ metabolites}$ = $\frac{90.482}{2.090}$ = 43,29 ≅ 43

Calculation of the proportion of lipid-like molecules in the metabolites database

Proportin of lipid-like molecules = $\frac{29.602 * 100}{90.4298}$ = 32,71 % ≅ 33%

Calculation of the Life Fluids ID

For each omics, a specific abbreviation is used: MB for metabolites, PR for proteins, TR for transcripts and CE for chemical elements.

LF is the abbreviation of Life Fluids.

LFID = LF + omics abbreviation + position added

For example, the first metabolite would be:

LFID = $\frac{LF}{Life\ Fluids}$ + $\frac{MB}{Metabolites}$ + $\frac{0000001}{Position\ added}$ = LFMB0000001

9. Annex 2

Here is the link to the folder containing the Python scripts and the tables that make up the Life Fluids Project Database:

https://drive.google.com/drive/folders/1WnbpPZvOJpJk9oG0n9OzlnpfGMHyGCFn?usp=sharing

10. References

[1] "Hematology Glossary - Hematology.org." Accessed: Jun. 03, 2024. [Online]. Available: https://www.hematology.org/education/patients/blood-basics

[2] "Cerebrospinal Fluid (CSF) Analysis: MedlinePlus Medical Test." Accessed: Jun. 03, 2024. [Online]. Available: https://medlineplus.gov/lab-tests/cerebrospinal-fluid-csf-analysis/

[3] L. N. Telano and S. Baker, "Physiology, Cerebral Spinal Fluid," *StatPearls*, Jul. 2023, Accessed: Jun. 27, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK519007/

[4] P. Solár, A. Zamani, L. Kubíčková, P. Dubový, and M. Joukal, "Choroid plexus and the blood–cerebrospinal fluid barrier in disease," *Fluids and Barriers of the CNS 2020 17:1*, vol. 17, no. 1, pp. 1–29, May 2020, doi: 10.1186/S12987-020-00196-2.

[5] C. M. Micheel *et al.*, "Omics-Based Clinical Discovery: Science, Technology, and Applications," Mar. 2012, Accessed: Jun. 04, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK202165/

[6] M. Vailati-Riboni, V. Palombo, and J. J. Loor, "What Are Omics Sciences?," *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*, pp. 1–7, Jan. 2017, doi: 10.1007/978-3-319-43033-1_1.

[7] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool, "Proteomics: Technologies and Their Applications," *J Chromatogr Sci*, vol. 55, no. 2, pp. 182–196, Feb. 2017, doi: 10.1093/CHROMSCI/BMW167.

[8] A. Blanco and G. Blanco, "Metabolism," *Med Biochem*, pp. 307–314, Jan. 2022, doi: 10.1016/B978-0-323-91599-1.00026-2.

[9] Y. Darzi, I. Letunic, P. Bork, and T. Yamada, "iPath3.0: interactive pathways explorer v3," *Nucleic Acids Res*, vol. 46, no. W1, pp. W510–W513, Jul. 2018, doi: 10.1093/NAR/GKY299.

[10] "Metabolic Pathways Analysis - Creative Proteomics." Accessed: Jul. 01, 2024. [Online]. Available: https://metabolomics.creative-proteomics.com/metabolic-pathways-analysis.htm

[11] S. V. Chitti *et al.*, "Vesiclepedia 2024: an extracellular vesicles and extracellular particles repository," *Nucleic Acids Res*, vol. 52, no. D1, p. D1694, Jan. 2024, doi: 10.1093/NAR/GKAD1007.

[12] D. Shao *et al.*, "HBFP: a new repository for human body fluid proteome," *Database (Oxford)*, vol. 2021, 2021, doi: 10.1093/DATABASE/BAAB065.

[13] F. Desiere *et al.*, "The PeptideAtlas project," *Nucleic Acids Res*, vol. 34, no. Database issue, 2006, doi: 10.1093/NAR/GKJ040.

[14] "What is PRIDE? | PRIDE." Accessed: Jun. 03, 2024. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/pride-quick-tour/what-is-pride/

[15] A. Guldbrandsen *et al.*, "CSF-PR 2.0: An Interactive Literature Guide to Quantitative Cerebrospinal Fluid Mass Spectrometry Data from Neurodegenerative Disorders," *Mol Cell Proteomics*, vol. 16, no. 2, pp. 300–309, Feb. 2017, doi: 10.1074/MCP.O116.064477.

[16] I. S. Khomiakov, R. F. Salamov, and N. A. Semenova, "[Age-related dynamics of the shape and size of the subpubic angle in children and adolescents].," *Arkh Anat Gistol Embriol*, vol. 93, no. 7, pp. 47–51, Jul. 1987, Accessed: Jun. 06, 2024. [Online]. Available: https://europepmc.org/article/MED/3675212

[17] S. Sirko *et al.*, "Injury-specific factors in the cerebrospinal fluid regulate astrocyte plasticity in the human brain," *Nat Med*, vol. 29, no. 12, pp. 3149–3161, Dec. 2023, doi: 10.1038/S41591-023-02644-6.

[18] P. Kapoor-Vazirani *et al.*, "SAMHD1 deacetylation by SIRT1 promotes DNA end resection by facilitating DNA binding at double-strand breaks," *Nat Commun*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/S41467-022-34578-X.

[19] D. S. Wishart *et al.*, "HMDB: the Human Metabolome Database," *Nucleic Acids Res*, vol. 35, no. suppl_1, pp. D521–D526, Jan. 2007, doi: 10.1093/NAR/GKL923.

[20] D. S. Wishart *et al.*, "MarkerDB: an online database of molecular biomarkers," *Nucleic Acids Res*, vol. 49, no. D1, pp. D1259–D1267, Jan. 2021, doi: 10.1093/NAR/GKAA1067.

[21] Z. Lai *et al.*, "Identifying epimetabolites by integrating metabolome databases with mass spectrometry cheminformatics," *Nat Methods*, vol. 15, no. 1, p. 53, Jan. 2018, doi: 10.1038/NMETH.4512.

[22] D. K. Barupal and O. Fiehn, "Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach," *Environ Health Perspect*, vol. 127, no. 9, Sep. 2019, doi: 10.1289/EHP4713.

[23] J. Goveia *et al.*, "Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities," *EMBO Mol Med*, vol. 8, no. 10, p. 1134, Oct. 2016, doi: 10.15252/EMMM.201606798.

[24] Y. Xu *et al.*, "An atlas of genetic scores to predict multi-omic traits," *Nature*, vol. 616, no. 7955, p. 123, Apr. 2023, doi: 10.1038/S41586-023-05844-9.

[25] M. Sud *et al.*, "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Res*, vol. 44, no. Database issue, p. D463, Jan. 2016, doi: 10.1093/NAR/GKV1042.

[26] O. Yurekten *et al.*, "MetaboLights: open data repository for metabolomics," *Nucleic Acids Res*, vol. 52, no. D1, pp. D640–D646, Jan. 2024, doi: 10.1093/NAR/GKAD1045.

[27] O. D. Murillo *et al.*, "exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids," *Cell*, vol. 177, no. 2, pp. 463-477.e15, Apr. 2019, doi: 10.1016/J.CELL.2019.02.018.

[28] I. M. Vincent *et al.*, "Metabolomics Identifies Multiple Candidate Biomarkers to Diagnose and Stage Human African Trypanosomiasis," *PLoS Negl Trop Dis*, vol. 10, no. 12, Dec. 2016, doi: 10.1371/JOURNAL.PNTD.0005140.

[29] A. Bateman, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Res*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/NAR/GKY1049.

[30] "About the HGNC | HUGO Gene Nomenclature Committee." Accessed: Jun. 03, 2024. [Online]. Available: https://www.genenames.org/about/

[31] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *J Cheminform*, vol. 7, no. 1, pp. 1–34, May 2015, doi: 10.1186/S13321-015-0068-4/FIGURES/11.

[32] E. Fahy and S. Subramaniam, "RefMet: a reference nomenclature for metabolomics," *Nat Methods*, vol. 17, no. 12, pp. 1173–1174, Dec. 2020, doi: 10.1038/S41592-020-01009-Y.

[33] K. A. Anderson and W. E. Hillwalker, "Bioavailability," *Encyclopedia of Ecology, Five-Volume Set*, vol. 1–5, pp. 348–357, Jan. 2008, doi: 10.1016/B978-008045405-4.00375-X.

[34] D. Weininger, "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules," *J Chem Inf Comput Sci*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/CI00057A005/ASSET/CI00057A005.FP.PNG_V03.

[35] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/NAR/28.1.27.

[36] K. Skogerson, G. Wohlgemuth, D. K. Barupal, and O. Fiehn, "The volatile compound BinBase mass spectral database," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–15, Aug. 2011, doi: 10.1186/1471-2105-12-321/FIGURES/7.

[37] M. Sud *et al.*, "LMSD: LIPID MAPS structure database," *Nucleic Acids Res*, vol. 35, no. Database issue, Jan. 2007, doi: 10.1093/NAR/GKL838.

[38] "Acerca de CAS | CAS." Accessed: Jun. 04, 2024. [Online]. Available: https://www.cas.org/es-es/about-us

[39] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res*, vol. 44, no. D1, pp. D1214–D1219, 2016, doi: 10.1093/NAR/GKV1031.

[40] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/NAR/GKX1037.

[41] S. Kim *et al.*, "PubChem Substance and Compound databases," *Nucleic Acids Res*, vol. 44, no. D1, pp. D1202–D1213, Jan. 2016, doi: 10.1093/NAR/GKV951.

[42] Y. Djoumbou Feunang *et al.*, "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy," *J Cheminform*, vol. 8, no. 1, pp. 1–20, Nov. 2016, doi: 10.1186/S13321-016-0174-Y.

[43] "Scientific Image and Illustration Software | BioRender." Accessed: Jun. 12, 2024. [Online]. Available: https://www.biorender.com/

[44] "PubChemPy documentation — PubChemPy 1.0.4 documentation." Accessed: Jul. 03, 2024. [Online]. Available: https://pubchempy.readthedocs.io/en/latest/

[45] "molmass · PyPI." Accessed: Jul. 03, 2024. [Online]. Available: https://pypi.org/project/molmass/

[46] "¿Qué es una API? - Explicación de interfaz de programación de aplicaciones - AWS." Accessed: Jul. 02, 2024. [Online]. Available: https://aws.amazon.com/es/what-is/api/

[47] E. Smedler *et al.*, "Metabolomics analysis of cerebrospinal fluid suggests citric acid cycle aberrations in bipolar disorder," *Neuroscience Applied*, vol. 1, p. 100108, Jan. 2022, doi: 10.1016/J.NSA.2022.100108.

[48] "Trastornos mitocondriales - Salud infantil - Manual MSD versión para público general." Accessed: Jul. 04, 2024. [Online]. Available: https://www.msdmanuals.com/es-es/hogar/salud-infantil/trastornos-metab%C3%B3licos-hereditarios/trastornos-mitocondriales

[49] J. Merino Pérez, M. José, and N. Borge, "FISIOLOGÍA GENERAL VÍAS METABÓLICAS DE SÍNTESIS AMINOÁCIDOS Y NUCLEÓTIDOS INTRODUCCIÓN".

[50] M. A. Sambrook, "The relationship between cerebrospinal fluid and plasma electrolytes in patients with meningitis," *J Neurol Sci*, vol. 23, no. 2, pp. 265–273, Oct. 1974, doi: 10.1016/0022-510X(74)90230-5.

[51] W. Barmore, F. Azad, and W. L. Stone, "Physiology, Urea Cycle," *StatPearls*, May 2023, Accessed: Jul. 04, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK513323/

[52] E. Croom, "Metabolism of Xenobiotics of Human Environments," *Prog Mol Biol Transl Sci*, vol. 112, pp. 31–88, Jan. 2012, doi: 10.1016/B978-0-12-415813-9.00003-9.

[53] A. Varki, "Biological roles of glycans," *Glycobiology*, vol. 27, no. 1, pp. 3–49, Jan. 2017, doi: 10.1093/GLYCOB/CWW086.

[54] N. S. Chandel, "Amino Acid Metabolism," *Cold Spring Harb Perspect Biol*, vol. 13, no. 4, 2021, doi: 10.1101/CSHPERSPECT.A040584.

[55] G. Berezhnoy, C. Laske, and C. Trautwein, "Metabolomic profiling of CSF and blood serum elucidates general and sex-specific patterns for mild cognitive impairment and Alzheimer's disease patients," *Front Aging Neurosci*, vol. 15, p. 1219718, 2023, doi: 10.3389/FNAGI.2023.1219718/FULL.

[56] G. Dante, V. Vaccaro, and F. Facchinetti, "Use of progestagens during early pregnancy," *Facts Views Vis Obgyn*, vol. 5, no. 1, p. 66, 2013, Accessed: Jul. 05, 2024. [Online]. Available: /pmc/articles/PMC3987350/

[57] J. Mathew, P. Sankar, and M. Varacallo, "Physiology, Blood Plasma," *StatPearls*, Apr. 2023, Accessed: Jun. 26, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK531504/

[58] "The human blood proteins - secreted to blood - The Human Protein Atlas." Accessed: Jun. 26, 2024. [Online]. Available: https://www.proteinatlas.org/humanproteome/blood+protein/secreted+to+blood

[59] M. Vieira and M. J. Saraiva, "Transthyretin: A multifaceted protein," *Biomol Concepts*, vol. 5, no. 1, pp. 45–54, Mar. 2014, doi: 10.1515/BMC-2013-0038/ASSET/GRAPHIC/BMC-2013-0038_CV1.JPG.

[60] A. A. J. Vaillant, Z. Jamal, P. Patel, and K. Ramphul, "Immunoglobulin," *Encyclopedia of Respiratory Medicine: Volume 1-4*, vol. 1–4, pp. V2-314-V2-320, Aug. 2023, doi: 10.1016/B0-12-370879-6/00184-8.

[61] H. H. Billett, "Hemoglobin and Hematocrit," *Anesthesiology*, vol. 28, no. 4, pp. 763–763, Jul. 1990, doi: 10.1097/00000542-196707000-00028.

[62] G. D. O. Lowe, A. Rumley, and I. J. Mackie, "Plasma fibrinogen," *http://dx.doi.org/10.1258/0004563042466884*, vol. 41, no. 6, pp. 430–440, Nov. 2004, doi: 10.1258/0004563042466884.

[63] T. Manon-Jensen, N. G. Kjeld, and M. A. Karsdal, "Collagen-mediated hemostasis," *J Thromb Haemost*, vol. 14, no. 3, pp. 438–448, Mar. 2016, doi: 10.1111/JTH.13249.

[64] X. Yuan and D. M. Desiderio, "Proteomics analysis of human cerebrospinal fluid," *Journal of Chromatography B*, vol. 815, no. 1–2, pp. 179–189, Feb. 2005, doi: 10.1016/J.JCHROMB.2004.06.044.

[65] H. Reiber, "Dynamics of brain-derived proteins in cerebrospinal fluid," *Clinica Chimica Acta*, vol. 310, no. 2, pp. 173–186, Aug. 2001, doi: 10.1016/S0009-8981(01)00573-3.

[66] C. Levecque, A. Velayos-Baeza, Z. G. Holloway, and A. P. Monaco, "The dyslexia-associated protein KIAA0319 interacts with adaptor protein 2 and follows the classical clathrin-mediated endocytosis pathway," *Am J Physiol Cell Physiol*, vol. 297, no. 1, p. C160, Jul. 2009, doi: 10.1152/AJPCELL.00630.2008.

[67] C. S. Musaeus, H. S. Gleerup, P. Høgh, G. Waldemar, S. G. Hasselbalch, and A. H. Simonsen, "Cerebrospinal Fluid/Plasma Albumin Ratio as a Biomarker for Blood-Brain Barrier Impairment Across Neurodegenerative Dementias," *J Alzheimers Dis*, vol. 75, no. 2, pp. 429–436, 2020, doi: 10.3233/JAD-200168.

[68] O. Kiseleva, I. Kurbatov, E. Ilgisonis, and E. Poverennaya, "Defining blood plasma and serum metabolome by gc-ms," *Metabolites*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/METABO12010015/S1.

[69] R. Mandal *et al.*, "Multi-platform characterization of the human cerebrospinal fluid metabolome: A comprehensive and quantitative update," *Genome Med*, vol. 4, no. 4, Apr. 2012, doi: 10.1186/GM337.

[70] J. García Pérez, L. Sastre Sisto, and L. Caridad Menéndez, "El ácido láctico en las meningoencefalitis bacterianas. Su valor diagnóstico," *Rev Cubana Med*, vol. 25, no. 8, Jan. 2021, Accessed: Jul. 04, 2024. [Online]. Available: https://revmedicina.sld.cu/index.php/med/article/view/2395

[71] E. Vigorito *et al.*, "microRNA-155 regulates the generation of immunoglobulin class-switched plasma cells," *Immunity*, vol. 27, no. 6, pp. 847–859, Dec. 2007, doi: 10.1016/J.IMMUNI.2007.10.009.

[72] C. Esau *et al.*, "MicroRNA-143 regulates adipocyte differentiation.," *J Biol Chem*, vol. 279, no. 50, pp. 52361–52365, Oct. 2004, doi: 10.1074/JBC.C400438200.

[73] S. Wang *et al.*, "The endothelial-specific microRNA miR-126 governs vascular integrity and angiogenesis," *Dev Cell*, vol. 15, no. 2, pp. 261–271, Aug. 2008, doi: 10.1016/J.DEVCEL.2008.07.002.

[74] R. A. DeGiosio, M. J. Grubisha, M. L. MacDonald, B. C. McKinney, C. J. Camacho, and R. A. Sweet, "More than a marker: potential pathogenic functions of MAP2," *Front Mol Neurosci*, vol. 15, Sep. 2022, doi: 10.3389/FNMOL.2022.974890.

[75] A. John, H. Wildner, and S. Britsch, "The homeodomain transcription factor Gbx1 identifies a subpopulation of late-born GABAergic interneurons in the developing dorsal spinal cord," *Dev Dyn*, vol. 234, no. 3, pp. 767–771, Nov. 2005, doi: 10.1002/DVDY.20568.

[76] "What Chemical Elements are Found in the Human Body?" Accessed: Jun. 28, 2024. [Online]. Available: https://www.news-medical.net/life-sciences/What-Chemical-Elements-are-Found-in-the-Human-Body.aspx

[77] R. Kandimalla, J. Vallamkondu, E. B. Corgiat, and K. D. Gill, "Understanding Aspects of Aluminum Exposure in Alzheimer's Disease Development," *Brain Pathology*, vol. 26, no. 2, p. 139, Mar. 2016, doi: 10.1111/BPA.12333.

[78] P. G. Righetti and E. Boschetti, "Biomedical Involvements of Low-Abundance Proteins," *Low-Abundance Proteome Discovery*, pp. 197–231, Jan. 2013, doi: 10.1016/B978-0-12-401734-4.00006-3.

[79] C. Borbinha, F. Serrazina, M. Salavisa, and M. Viana-Baptista, "Bismuth encephalopathy- a rare complication of long-standing use of bismuth subsalicylate," *BMC Neurol*, vol. 19, no. 1, pp. 1–5, Aug. 2019, doi: 10.1186/S12883-019-1437-9/TABLES/1.