DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Development of a KNIME Workflow for the retrieval of molecules associated with solute carrier proteins linked to rare diseases“

verfasst von / submitted by

Marlene Kofler

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Pharmazie (Mag.pharm.)

Wien, 2020/ Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 449
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker
Mitbetreut von / Co-Supervisor:	Dipl.-Ing. (FH) Dr. Daniela Digles

ACKNOWLEDGEMENTS

Foremost, I would like to thank my supervisor Gerhard Ecker for the opportunity to write my thesis in his group, and my co-supervisor Daniela Digles, who always encouraged me and helped me with great advises and never lost her patience with replying to my questions.

Besides, I would like to thank the other members of the pharmacoinformatics research group at the University of Vienna, who provided me with additional information.

Also, I would like to take this opportunity to thank my family, who always supported me during the time of my study, financially as well as emotionally with love, patience and guidance.

Last but not least, I would like to thank my boyfriend Andreas for all his love and support, especially during the time of writing my thesis, for believing in me and motivating me when it was so well-needed.

ABSTRACT

SLCs, short for solute carrier, are a relatively unexplored group of transport proteins that control essential physiological functions. Despite being associated with several diseases, they represent a rather untapped source of new potential drug targets.

Also, around 300 million people worldwide are suffering from rare diseases which are defined as diseases that affect only a small number of people. However, despite being so rare, rare diseases are numerous, and often include a lack of basic knowledge and treatment possibilities which makes them one of the key global health priorities.

The aim of this work was to create a workflow on KNIME that shows the role of SLCs in rare diseases and the availability of possible modulators through the integration of data from, altogether, six databases, starting from a list of SLCs, provided by the RESOLUTE project.

As the data include false-positive findings and often lack essential information, like the type of association between an SLC and a rare disease or a molecule, respectively, a second workflow was created. This workflow can be accessed through the KNIME WebPortal and can be used for filtering as well as for manual curation of associations.

The collected data highly suggest that SLCs play an essential role in rare diseases. However, manual curation and research are needed to use the information further.

ZUSAMMENFASSUNG

SLCs, kurz für Solute Carrier, sind eine relativ unerforschte Gruppe von Transportproteinen, die wesentliche physiologische Funktionen steuern. Obwohl sie mit mehreren Krankheiten verbunden sind, stellen sie eine eher unerschlossene Quelle für neue potenzielle Ziele für Arzneistoffe dar.

Darüber hinaus leiden weltweit rund 300 Millionen Menschen an seltenen Krankheiten, die als Krankheiten definiert werden, von denen nur eine geringe Anzahl von Menschen betroffen sind. Trotz ihrer Seltenheit sind seltene Krankheiten zahlreich und beinhalten häufig einen Mangel an Grundkenntnissen und Behandlungsmöglichkeiten, was sie zu einer der wichtigsten globalen Gesundheitsprioritäten macht.

Ziel dieser Arbeit war es, einen Workflow auf KNIME zu erstellen, der die Rolle von SLCs in seltenen Krankheiten und die Verfügbarkeit möglicher Modulatoren durch die Integration von Daten aus insgesamt sechs Datenbanken zeigt, ausgehend von einer Liste von SLCs, die vom RESOLUTE Projekt bereitgestellt wurde.

Da die Daten falsch positive Ergebnisse enthalten und häufig wesentliche Informationen, wie die Art der Assoziation zwischen einem SLC und einer seltenen Krankheit bzw. einem Molekül, fehlen, wurde ein zweiter Workflow erstellt. Auf diesen Workflow kann über das KNIME WebPortal zugegriffen werden und er kann sowohl zum Filtern als auch zum manuellen Kuratieren von Assoziationen verwendet werden.

Die gesammelten Daten legen nahe, dass SLCs bei seltenen Krankheiten eine wesentliche Rolle spielen. Manuelle Kuratierung und Recherche sind jedoch erforderlich, um die Informationen weiter zu nutzen.

TABLE OF CONTENT

List of Figures X
List of Tables XII
1 INTRODUCTION 2
1.1 Rare diseases 2
1.1 Solute Carriers 3
1.2 Aim of the thesis 5
2 METHODS 6
2.1 KNIME 6
2.1.1 The KNIME Analytics Platform 6
2.1.2 KNIME Metanodes and Components 8
2.1.3 The KNIME Server and WebPortal 8
2.1.4 Data format: XML 9
2.1.5 Web APIs 10
2.2 Sources of data 11
2.2.1 RESOLUTE_SLCs file 12
2.2.2 UniProt KB. 12
2.2.3 Orphanet 13
2.2.4 DisGeNET 15
2.2.5 DrugBank 17
2.2.6 ChEMBL 18
2.2.7 PubChem 19
2.2.8 Datasets and content *of* file 'SLCs and rare diseases' 20
2.2.9 Datasets and content of file 'SLCs and molecules' 23
2.3 Workflow for data retrieval 26
2.3.1 Extraction of additional information about SLCs 26
2.3.2 Extraction of rare diseases 29
2.3.3 Extraction of drugs/ molecules 31
2.3.3.1 Drugbank 31
2.3.3.2 ChEMBL 33
2.3.3.3 PubChem 36
2.4 Workflow for accessing data (end-user) 38
3 RESULTS 46

3.1 Results of the workflow for data retrieval 46
3.2 Workflow for accessing data as seen by users 53
4 DISCUSSION 58
4.1 Limitations of the workflow 58
4.2 Possibilities for adaptation 59
5 REFERENCES 61
6 APPENDIX 65
6.1 Data tables 65
6.1.1 Rare diseases with number of associated molecules and SLCs 65
6.1.2 SLCs with number of associated diseases and disease classes 100

List of Figures

Figure 1: SLC transporters, based on Hediger et. al, 2013[1] 3
Figure 2: Triangulation; based on Gurinova, 2018[2] 6
Figure 3: KNIME Analytics Platform example workflow 7
Figure 4: Widgets on the KNIME Analytics Platform vs visualization via the WebPortal 8
Figure 5: Example XML 9
Figure 6: XPath in KNIME 10
Figure 7: Result of an API call performed with a GET Request node 11
Figure 8: Example of a UniProt entry (screenshot from [3], accessed 06/15/2020) 13
Figure 9: Example of an Orphanet entry (screenshot from [4], accessed 06/15/2020) 14
Figure 10: Example of gene-disease associations on DisGeNET for SLC1A1 (screenshot from [5]/, accessed 06/15/2020) 16
Figure 11: Example of DrugBank entry for SLC1A1 (screenshot from [6], accessed 06/15/20). 17
Figure 12: Example of compounds associated with SLC1A1 on ChEMBL (screenshot from [7], accessed 06/15/20) 18
Figure 13: Example of compounds associated with SLC1A1 on PubChem (screenshot from [8], accessed 06/15/2020) 20
Figure 14: Overview of the workflow for the retrieval of SLC-rare disease-molecule associations 26
Figure 15: Nodes for the Extraction of Protein & Gene Aliases from UniProt 27
Figure 16: Example XML file as retrieved via UniProt API 28
Figure 17: Table before and after joining Gene and Protein names 29
Figure 18: Example part of DisGeNET's file "ALL gene-disease-pmid associations" 29
Figure 19: Nodes used for the integration of <MeSH disease class names> 30
Figure 20: Configuration of first XPath node for Drugbank 32
Figure 21: Configuration of second XPath node for Drugbank 32
Figure 22: Nodes for the retrieval of 'protein type' node via XPath 32
Figure 23: Table creator node for the retrieval of proteins associated with drugs on Drugbank 33
Figure 24: Table after first ChEMBL API call 34
Figure 25: Screenshot of the page_meta section of the retrieved XML for NPC1 34
Figure 26: Nodes used for the retrieval of bioactivity data from ChEMBL 35
Figure 27: Nodes for the download of active CIDs associated with SLCs from PubChem 37
Figure 28: Workflow for interested users as seen in the KNIME Analytics Platform 38
Figure 29: Inside Component 'Filter disease classes/sources' 38
Figure 30: Dialog window for Interactive Filter Widget 39

Figure 31: Inside of Component ‘SLCs or rare diseases’ 39
Figure 32: Dialog window for Java IF node (screenshot) 40
Figure 33: Inside of Component ‘Bar Chart SLCs’ (screenshot)... 41
Figure 34: Combination of nodes/metanodes for manual curation 42
Figure 35: Inside of metanode 'curation_preparation' 42
Figure 36: Dialog window for Column Filter 1 43
Figure 37: Inside of component 'curation_step'.. 43
Figure 39: Combination of nodes/ components for downloading.. 44
Figure 40: Bar Chart showing SLCs with number of associated rare diseases 47
Figure 41: Bar Chart showing SLCs with number of associated molecules/drugs.................... 48
Figure 42: Number of molecules received from each database 49
Figure 43: Starting window of 'Workflow for interested users' at the KNIME WebPortal............ 53
Figure 44: First filtering options at WebPortal 53
Figure 45: Filtering option between SLCs or rare diseases... 54
Figure 46: Bar Chart and corresponding table 54
Figure 47: Curation or download 55
Figure 48: Curation example ‚Tarsal-carpal coalition syndrome’... 55
Figure 49: Curation example ‘CACT deficiency’.. 56
Figure 50: Name file and download SLC_rarediseases 57
Figure 51: Filter for molecule sources... 57

List of Tables

Table 1: attributes retrieved from the file RESOLUTE_SLCs[9] ... 21

Table 2: attributes retrieved from the UniProt KB REST API[10] ... 21

Table 3: attributes retrieved from ‘Orphanet rare diseases with their associated genes’, version 1.2.11/4.1.6 [2018-04-12] (orientdb version)[11] ... 21

Table 4: attributes retrieved from ‘Rare diseases and cross-referencing’, version 1.2.11/4.1.6 [2018-04-12] (orientdb version)[11] ... 22

Table 5: attributes retrieved from DisGeNET, 'ALL gene-disease-pmid associations', renewed every execution, analysis version: version 7.0 [2020-05-04][12] ... 22

Table 6: attributes retrieved from DisGeNET ‘BeFree gene-disease-pmid associations’, renewed every execution, analysis version: version 7.0 [2020-05-04][12] ... 23

Table 7: columns added within KNIME ... 23

Table 8: attributes retrieved from DrugBank, ‘All drugs’, version 5.1.2 [2018-12-20][13] ... 23

Table 9: attributes retrieved from ChEMBL via RESTful API, renewed every run, analysis version: version 27, [2020-05-18][14]. ... 24

Table 10: attributes retrieved from PubChem by download of .CSV files, renewed every run, analysis: last updated 05/18/2020[15] ... 24

Table 11: attributes retrieved from PubChem through API calls[16] ... 25

Table 12: Columns added within KNIME ... 25

Table 13: Counts at different positions of the workflow ... 51

1 INTRODUCTION

1.1 Rare diseases

Rare diseases are diseases that affect only a small percentage of people. However, there is no single, worldwide accepted definition. In most countries, a rare disease is defined by a maximum total number of affected patients. An example would be the US definition defining a rare disease as a disease with less than 200.000 cases in the US. [17] In the EU, however, a rare disease is characterised by a prevalence of less than 5 in 10,000 (1 in 2,000) citizens. Genetic mutations cause a big part of rare diseases with a significant part starting at childhood. An example would be the 'Fragile X syndrome'. However, rare diseases also include rare infectious diseases, caused by bacteria or viruses, autoimmune diseases and cancers.[18] Despite being individually uncommon, rare diseases are numerous. Up to 8,000 unique diseases have been described and it is estimated that around 300 million people worldwide are affected by rare diseases.[18], [19] Some diseases are generally more well-known, like Cystic Fibrosis or Huntington's disease, others have a patient population below 100.

However, rare diseases are also sometimes referred to as 'orphan diseases', as they have been neglected by researchers as well as doctors for a long time. Under normal marketing conditions, developing medicine for rare diseases would not be profitable for pharmaceutical companies, as the process of discovering a molecule and reaching marketing authorisation takes a long time and is very expensive. Therefore, developing drugs for rare diseases would operate a financial deficit, as the expected sales would not even recover the money spent on development because of the small number of treatable patients. To support the research on so-called orphan drugs, countries introduced numerous incentives such as regulatory assistance or marketing exclusivity. The United States were the first ones with introducing the Orphan Drug Act in 1983 [20] while it took the European Union until 1999 to find a harmonised regulation, the Orphan Drug Regulation.[21], [22] Despite the increasing interest in research since, still only a small percentage of diseases is well-studied when it comes to basic knowledge and treatment possibilities. This makes rare diseases to one of the key global health priorities.[23]

1.2 Solute Carriers

Transport of solutes, such as sugars, amino acids, nucleotides, neurotransmitters and ions across biological membranes, is an essential process for cellular homeostasis and is, apart from passive diffusion, controlled by transport proteins that serve as gatekeepers.[24]

These proteins can be divided according to passive and active mechanisms. Passive, also referred to as facilitated, transporters transport solutes in the direction of their electrochemical gradient while active transporters utilise energy-coupled mechanisms to move substances against their gradient. Furthermore, active transporters can be classified into primary and secondary-active ones. In primary-active transporters, the transport is directly coupled to the hydrolysis of the energy provider (e.g. ATP). However, in secondary-active transporters, the transport of one solute is directly dependent on the transport of a second, either as symporter, transferring a second solute in the same direction or as antiporter, transporting the second solute in the opposite direction. [24],[25]

Reflecting the high importance of transporters, it is, according to Hediger et al., 2013[1], estimated that around 10% of all human genes are related to transporters. Solute carrier proteins (SLCs) make the largest gene group of membrane transporters with more than 400 members. They constitute a heterogeneous group of transporters located mostly in the cell membrane, but also in intracellular organelles, like the mitochondrial SLC family 25, or vesicles as shown in Figure 1. SLCs are defined as either facilitated or secondary-active transporters. In contrast, primary-active transporters, like the ABC-transporters, aquaporins and ion channels and pumps are not members of the SLC series.[1]

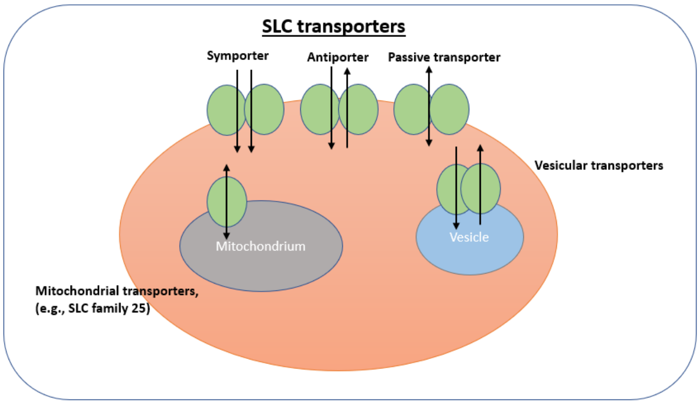


Figure 1: SLC transporters, based on Hediger et. al, 2013

Thus, the inclusion of a protein within the SLC series is based on function. This can lead to homology between different SLC families being very low to non-existent. However, members of a specific SLC family have at least 20-25% amino acid sequence identity to another member of the family.[1],[26]

The genes encoding the transporters are generally named after the HGNC (HUGO Gene Nomenclature Committee) system, starting with the root symbol ‘SLC’, followed by a number that specifies the family. The following letter defines the subfamily and is in most cases ‘A’, as most families are not further subdivided. The final number denotes the individual family member.[1], [25] However, there are some exceptions like the SLC family 21 that has its root symbol changed to SLCO.[27]

Also, some genes are referred to as ‘putative SLCs’ as they share an ancestral background with SLCs and are plausible facilitative or secondary active transporters but have not yet been classified into any of the existing SLC families and do not have a name according to the SLC root system.[28]

As solute carriers control essential biological functions, like nutrient uptake, waste removal and ion transport, genetic polymorphisms are associated with several diseases. According to Rives et al., 2017, human genetic data suggests that around 50% of SLCs are associated with human diseases compared to 20% of the broader genome, that illustrates their high importance in diseases.[29]

Some SLCs are already well studied and in use as drug targets. These drug targets include inhibitors of SGLT2, the renal sodium-glucose cotransporter 2, that is encoded by the gene SLC5A2. SGLT2 inhibitors are used in the treatment of diabetes type 2 as they lower blood sugar levels.[30] Another example would be the human monoamine transporters, mostly from SLC family 6, that are used as effective targets in the treatment of depression.[31]

Besides, SLCs can also cause rare diseases, especially monogenic (also referred to as Mendelian) diseases.[32] An example would be the association between Amish Lethal Microcephaly and SLC25A19.[33] Amish Lethal Microcephaly is a disease that has only been found in Amish families and leads to extreme microcephaly with an underdeveloped brain and early death, which suggested a defect in 2-ketoglutarate metabolism. The gene SLC25A19 that encodes the mitochondrial deoxynucleotide carrier was found responsible for this disease. A method of treatment could include drugs that enhance the transporter’s activity.[33]

Yet, the majority of SLCs have been getting only little research attention. More than 30% are even ‘orphans’ when it comes to the knowledge of their substrate specificity and function. It is assumable that many more than the SLCs are associated with diseases, especially rare diseases, and that they would represent a largely untapped source for drug targets.[32] Recently, however, the relevance of systematic research of SLCs for drug discovery is increasingly getting more attention. [32], [34]
RESOLUTE (Research Empowerment on SOLUTE carriers) is a project with 13 partners from academia and industry with the goal of intensifying worldwide research on solute carriers within a 5-year research project. The project aims to provide tools and reagents as well as assays and data- and knowledgebases. [9] For this thesis, a file with a list of SLCs originating from the RESOLUTE project, formed the starting point.

1.3 Aim of the thesis

The aim of this thesis is to collect data about the role of SLCs in rare diseases from databases through database integration. Also, possible modulators of these SLCs were aimed to be aggregated as they could form potential modulators of these diseases.
This approach is based on the diploma thesis ‘*Development of a KNIME workflow for the retrieval of associations between orphan diseases and their possible drug repurposing candidates’* by Jana Gurinova, 2018.[2] In the cited thesis, Gurinova tried to retrieve possible connections between rare diseases and drugs through their shared association with targets to propose repositioning candidates for rare diseases.
For the present work, this approach was made more specific as it was limited to SLCs as targets only. Also, new databases were included: UniProt, ChEMBL and PubChem. Besides, this workflow is not aimed at proposing drugs as repositioning candidates, but more at giving an overall overview of the role of SLCs in rare diseases and the availability of possible modulators, which includes approved drugs as well as molecules showing activity.

2 METHODS

The used method for this thesis was the aggregation of data from databases through a workflow created on KNIME in the form of a triangulation. This approach was based on the diploma thesis '*Development of a KNIME workflow for the retrieval of associations between orphan diseases and their possible drug repurposing candidates'* (Gurinova, 2018)[2].

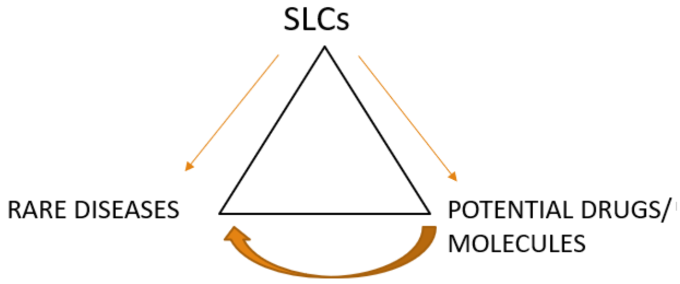


Figure 2: Triangulation; based on Gurinova, 2018

As the aggregated data needs manual curation, and this would, due to the amount of data, exceed the time constraints of a diploma thesis, a second workflow was created, that interested users can access through the KNIME WebPortal. This workflow offers the possibility to filter, curate and download the aggregated data.

The following chapter provides information about KNIME, the datasets and databases and the created workflows.

2.1 KNIME

KNIME, which derives from 'Konstanz Information Miner', is a data-analytics, reporting and integration platform that was created by a group of software engineers under the lead of Michael Berthold at the University of Konstanz. They released their first tool, the first version of the KNIME Analytics Platform, in 2006.[35]

The next subsections are going to provide an overview about the two offered, complementary tools - the freely available KNIME Analytics Platform and the KNIME Server, a commercial product, both of which were in use for this work. Besides, the data format XML, API calls, as well as meta nodes and components, are going to be explained in more detail as they were especially crucial for the creation of the workflows.

2.1.1 The KNIME Analytics Platform

The KNIME Analytics Platform is an open-source, freely available workflow management tool that provides a graphical user interface for interactive execution of a data pipeline that allows automated data analysis without extensive knowledge of programming.[35]

Workflows in the KNIME Analytics Platform are made of central, visualised units: so-called nodes. These hundreds of different available nodes can be combined by simple drag and drop and perform various tasks in processing the data. A simple workflow on KNIME starts with a node that reads in the data as the data is stored in an internal table-based format- the KNIME table. The KNIME table consists of a table with columns of a specific data type (e.g. integer, string, molecule) and an optional number of rows corresponding to the specification. Each node needs to be executed before handing the data to the following node. One of the most significant advantages of the KNIME Analytics platform is that the nodes store the data permanently. So, the workflow execution can be stopped and resumed at any time. The user can inspect intermediate results and also insert new nodes without losing previous results. [36]

Nodes on KNIME can be roughly divided into five categories:

1) Nodes, that read in the data, either directly from a file or via API call (see 2.1.5, p.10)

2) Nodes for data transformation, e.g. filters

3) Nodes for data analysis/mining

4) Nodes for visualisation that allow interactive exploration of the data

5) Nodes for data deployment

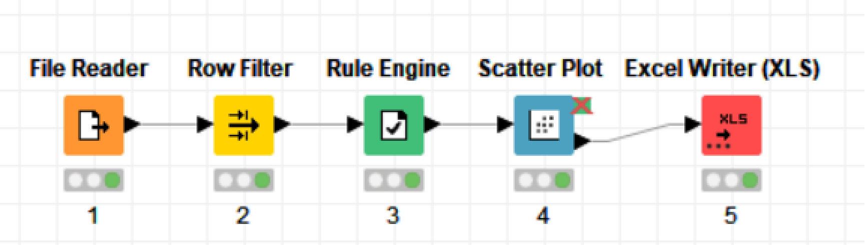


Figure 3: KNIME Analytics Platform example workflow

Figure 3 shows a screenshot of an example workflow, using one node corresponding to each of the before mentioned categories. In addition to the nodes that are included in the core KNIME Analytics implementation, it is possible to download KNIME extensions or even implement self-programmed nodes.

Besides, KNIME provides several tutorials and example workflows that can be used to get easily familiar with KNIME.

The version of the KNIME Analytics Platform used for this work was version 4.1, which was the latest update at the time of the practical part of this work. In addition, the KNIME extension 'KNIME XML-processing' was downloaded.

2.1.2 KNIME Metanodes and Components

Meta nodes look like a single node. However, they can contain several nodes. They can be used for making the workflow look 'tidied up' and make it easier for other people to understand the functionality of a workflow. [37]

Components, formerly called Wrapped Metanodes, however, are even more 'real KNIME nodes' as they bundle functionality and can have their own dialog and interactive view.[38]

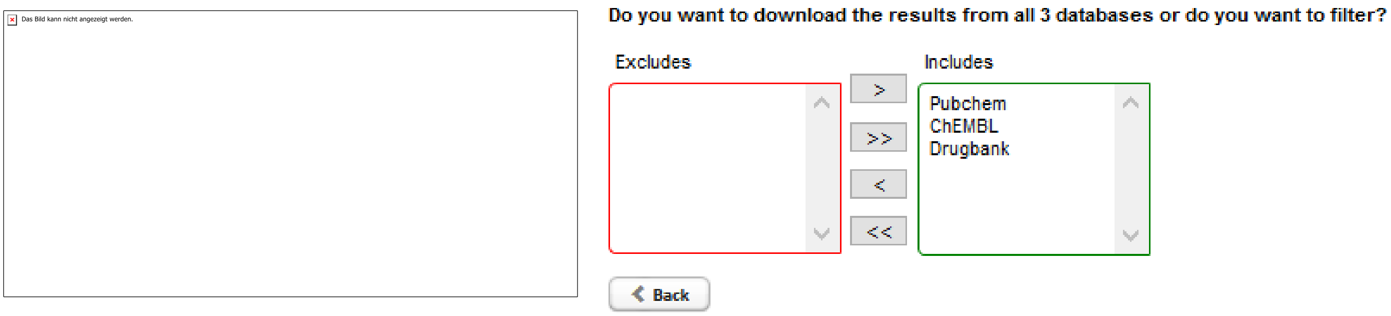
In combination with widgets and view nodes, it is possible to create interactive web pages on the WebPortal, which is a feature part of the KNIME Server.

2.1.3 The KNIME Server and WebPortal

The KNIME Server is a complementary, commercial product that offers the possibility to share workflows within a team. Workflows can be uploaded and stored on the server as well as downloaded to one's local KNIME Analytics Platform. Also, it is possible to schedule executions of workflows either for delayed or recurring jobs.[39]

This function was used for the first workflow of this work, the workflow for data retrieval (see 2.3), which is scheduled to be run and update its data at the KNIME Server every 15 days.

On the other hand, workflows can be executed through the web browser using the interactive 'KNIME WebPortal'. The KNIME WebPortal is an extension to the KNIME Server and automatically turns KNIME workflows containing components with widgets or visualisation nodes into browser-based applications.[40]



Das Bild kann nicht angezeigt werden.

Do you want to download the results from all 3 databases or do you want to filter?

Includes: Pubchem, ChEMBL, Drugbank

Figure 4: Widgets on the KNIME Analytics Platform vs visualization via the WebPortal

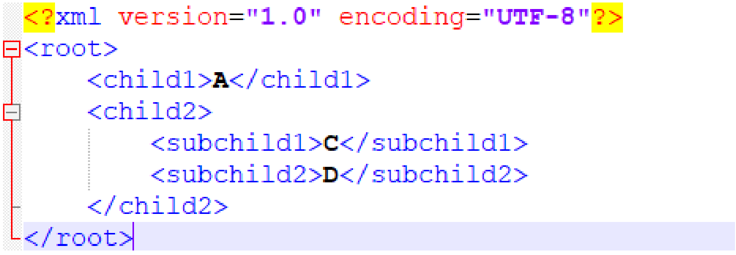
The second workflow, the workflow for interested users, is primarily dedicated to being run at the KNIME WebPortal. Figure 4 shows the inside of an example Component on the KNIME Analytics Platform versus the way it is being displayed through the WebPortal.

2.1.4 Data format: XML

The majority of data used for the workflow for data retrieval was in XML format.

XML stands for Extensive Markup Language. It is designed for the transport and storage of data and widely used in web development as it is both machine- and human-readable.[41]

Usually, an XML file starts with a prologue that contains the XML version and the character encoding. The rest of the XML format structure can roughly be compared to a tree. Like in a tree, a single ‘root’ contains all the other data elements, and it is structured in a specific way. The terms ‘parent’, ‘children’ and ‘siblings’ are used to describe the relationships between the data elements. While parents are one level above children, siblings are at the same level.[42]



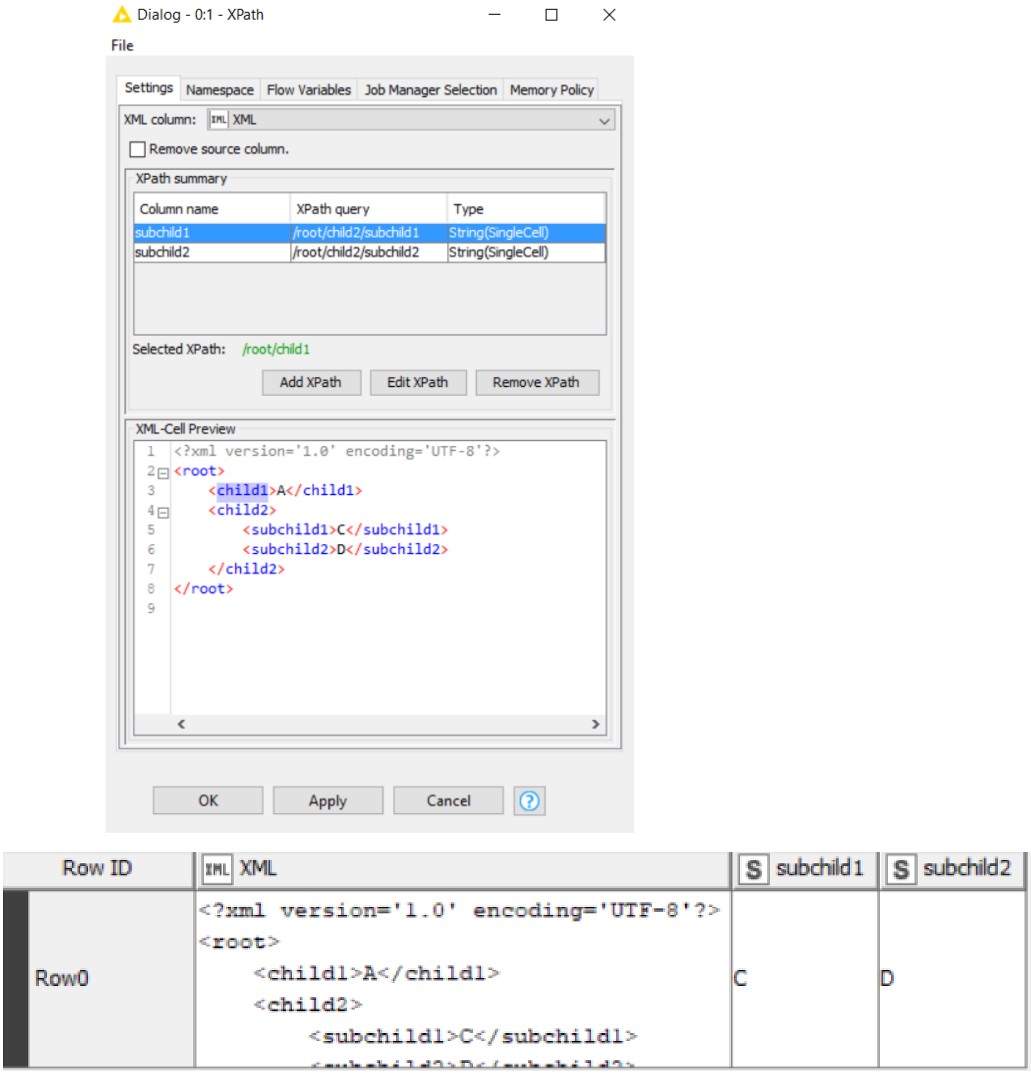
```xml
<?xml version="1.0" encoding="UTF-8"?>
<root>
    <child1>A</child1>
    <child2>
        <subchild1>C</subchild1>
        <subchild2>D</subchild2>
    </child2>
</root>
```

Figure 5: Example XML

Figure 5 shows an example XML file. A data element is always introduced and closed with a particular syntax, e.g. <child1> is used as an introduction while </child1> closes the element. All entries between the introduction and the closure define the value of the data element. In Figure 5, <child1> and <child2> are sibling elements, while <subchild1> and <subchild2> are children elements with <child2> as a parent.

The syntax makes it possible to query an XML file for specific elements. This can be done with XPath, the XML Path Language. Instruction on how to create XPaths can be found at https://www.w3schools.com/xml/xpath_intro.asp.

KNIME offers nodes to process XML files. XML files need to be imported into KNIME with the *XML reader* node. Afterwards, the XPath can be configured even easier as KNIME’s *XPath* node proposes an XPath expression when clicking on the dedicated attribute.



XML column: XML

☐ Remove source column.

XPath summary

Column name	XPath query	Type
subchild1	/root/child2/subchild1	String(SingleCell)
subchild2	/root/child2/subchild2	String(SingleCell)

Selected XPath: /root/child1

XML-Cell Preview

```
<?xml version='1.0' encoding='UTF-8'?>
<root>
    <child1>A</child1>
    <child2>
        <subchild1>C</subchild1>
        <subchild2>D</subchild2>
    </child2>
</root>
```

Row ID	XML	subchild1	subchild2
Row0	<?xml version='1.0' encoding='UTF-8'?> <root> <child1>A</child1> <child2> <subchild1>C</subchild1>	C	D

Figure 6: XPath in KNIME

Figure 6 shows the *XPath* node configuration as well as the results deriving from the Example XML file (see Figure 5).

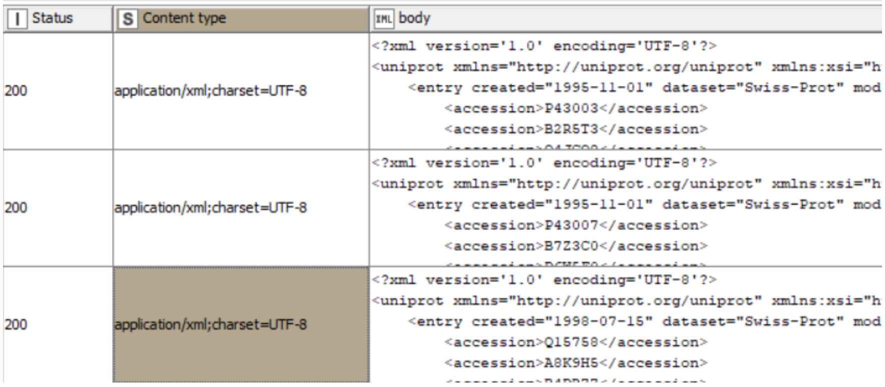
The native data sets used for the workflow are, of course, much more complex. The used files and the extraction are described in more detail in section 2.3, p.26f.

2.1.5 Web APIs

Some databases provide the possibility to download the data via Web API. Web API is short for Web Application Programming Interfaces and offers users the opportunity to get access to specific data from a database without downloading the whole one.

The API is usually a set of standard commands encoded into a URL syntax that is similar to the URL of the database. The databases having API access usually provide instruction on how their API URLs are built. Most of the times, the user can choose between different formats like JSON and XML. [14],[43],[44] For this work, API calls were used to retrieve data from three databases: UniProt, ChEMBL and PubChem. All of these API calls were performed via REST (representational state transfer) API, which refers to the type of software architectural style.

KNIME offers specific nodes for performing REST API calls that can be found within the category 'REST Web Services'. Before accessing the web API, it is necessary to create the URL using, for example, a *String Manipulation* node. Afterwards, a *GET Request* node performs the retrieval of data from the dedicated resource.



I Status	S Content type	XML body

200	application/xml;charset=UTF-8	<?xml version='1.0' encoding='UTF-8'?>
		<uniprot xmlns="http://uniprot.org/uniprot" xmlns:xsi="h
		    <entry created="1995-11-01" dataset="Swiss-Prot" mod
		        <accession>P43003</accession>
		        <accession>B2R5T3</accession>

200	application/xml;charset=UTF-8	<?xml version='1.0' encoding='UTF-8'?>
		<uniprot xmlns="http://uniprot.org/uniprot" xmlns:xsi="h
		    <entry created="1995-11-01" dataset="Swiss-Prot" mod
		        <accession>P43007</accession>
		        <accession>B7Z3C0</accession>

200	application/xml;charset=UTF-8	<?xml version='1.0' encoding='UTF-8'?>
		<uniprot xmlns="http://uniprot.org/uniprot" xmlns:xsi="h
		    <entry created="1998-07-15" dataset="Swiss-Prot" mod
		        <accession>Q15758</accession>
		        <accession>A8K9H5</accession>

Figure 7: Result of an API call performed with a GET Request node

As shown in Figure 7, the *GET Request* node results in three columns: the 'status', the 'content type' and the 'body'. The 'status', referring to the HTTP status, shows the success of a query: A status starting with 2xx indicates a successful request while a status beginning with 4xx or 5xx flags failure. The 'content type' shows the type of data accessed through the API call, and the 'body' contains the accessed data. In the case of Figure 7, an *XPath* node could follow to further process the data.

The specific URLs and XPath queries used for this work are described in more detail in section 2.3, p.26ff.

2.2 Sources of data

The data was generated via the integration of databases that include information about target-disease and target-molecule relationships starting from a list of SLCs. The workflow for data-retrieval (see 2.3, p.26ff.) results into two tables that are saved on the KNIME Server and can be accessed through the second workflow (see 2.4, p.38ff.). The first one, *SLCs and rare diseases*, contains information about SLCs and their associated rare diseases. The second table, *SLCs and molecules*, includes information on related drugs and molecules.

The information about SLCs was derived from two sources: the file RESOLUTE_SLCs that provided a list of all SLCs that were included into this work and the UniProt KB API that was used for the retrieval of additional information.

Orphanet and DisGeNET were the two databases used for the extraction of SLC-rare disease associations. DisGeNET provides an extensive set of gene-disease associations but does not include the option to filter for rare diseases. At the same time, Orphanet is dedicated to rare diseases and contains several disease identifiers that make it possible also to filter the results from DisGeNET.

Three databases were used to provide potential molecules for many of the SLC-rare disease associations and to provide interested users of the second workflow the choice between three different, widely used databases. DrugBank is fully curated, and roughly specialised to approved and experimental drugs. ChEMBL is also manually curated and specialised to bioactivity data with a significant part originating from medicinal papers. PubChem is an open database which means that everyone can upload their scientific data.

The following datasets, databases and their content are listed in the sequence of their occurrence in the workflow.

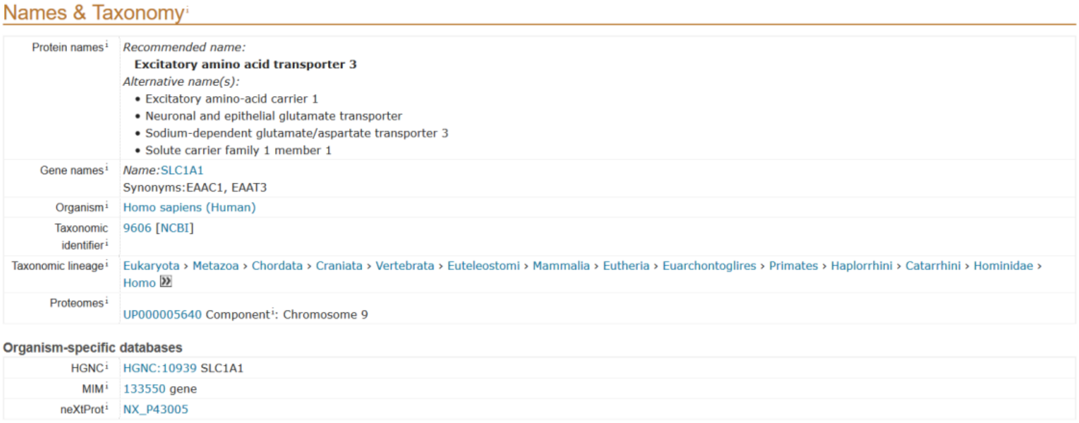
2.2.1 RESOLUTE_SLCs file

The file RESOLUTE_SLCs was provided by the RESOLUTE project and was adapted for its integration into the workflow for data retrieval as only four attributes were kept: SLC name, family, EntrezGeneID and UniProtID..

It was last updated in June 2019 and contains a list of 446 SLCs from 65 SLC families, including 16 SLCs that can be classified as putative SLCs as they are not named according to the SLC nomenclature and are not organised into any of the existing SLC families (see section 1.2, p.3f.).

2.2.2 UniProt KB.

The UniProt Knowledgebase is a database that contains annotated information about over 120 million proteins. It includes two types of entries: the curated SwissProt-Entries and the unreviewed TrEMBL entries that are annotated automatically. [45]



Names & Taxonomy

Protein names	*Recommended name:* **Excitatory amino acid transporter 3** *Alternative name(s):* • Excitatory amino-acid carrier 1 • Neuronal and epithelial glutamate transporter • Sodium-dependent glutamate/aspartate transporter 3 • Solute carrier family 1 member 1
Gene names	*Name:*SLC1A1 Synonyms:EAAC1, EAAT3
Organism	Homo sapiens (Human)
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo
Proteomes	UP000005640 Component: Chromosome 9

Organism-specific databases

HGNC	HGNC:10939 SLC1A1
MIM	133550 gene
neXtProt	NX_P43005

Figure 8: Example of a UniProt entry (screenshot from https://www.uniprot.org/uniprot/P43005, accessed 06/15/2020)

The provided information can either be accessed directly at the website at https://www.uniprot.org/ [10], downloaded as complete datasets in XML on the page 'Downloads' or accessed via several Web APIs.

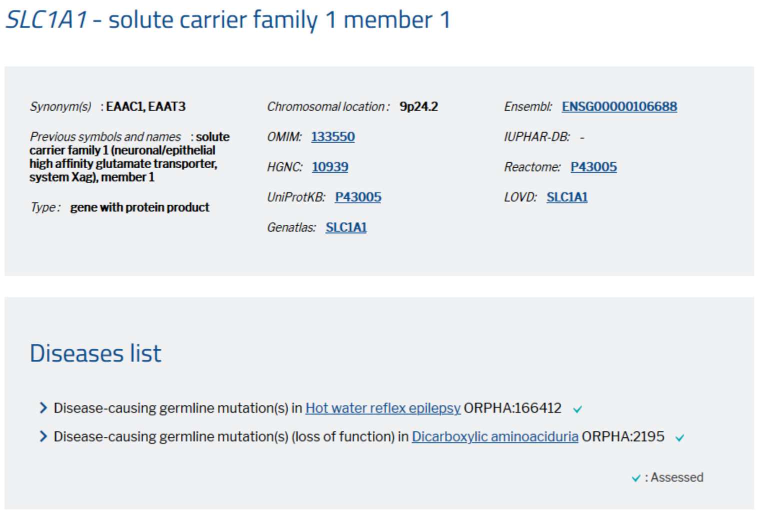
For this workflow, the UniProt website REST API was used to retrieve additional information about solute carrier proteins. The relevant part for this work of the UniProt entry is shown in Figure 8 for SLC1A1.

2.2.3 Orphanet

Orphanet was established to 'provide high-quality information on rare diseases and ensure equal access to knowledge for all stakeholders' which means that the database is adapted to the needs of patients and their families as well as of health care professionals and researchers. [46]

The curated information about rare diseases and orphan drugs provided can be accessed directly via the webpage (http://www.orpha.net) in nine languages.

An Orphanet entry for a rare disease contains information such as synonyms, several identifiers like Orphanet's specific terminology for rare diseases - the ORPHAnumber as well as cross-references to other databases (e.g. UMLS, MeSH, OMIM) and data about prevalence, age of onset and epidemiology.



SLC1A1 - solute carrier family 1 member 1

Synonym(s) : **EAAC1, EAAT3**

Previous symbols and names : **solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1**

Type : **gene with protein product**

Chromosomal location : **9p24.2**

OMIM: 133550

HGNC: 10939

UniProtKB: P43005

Genatlas: SLC1A1

Ensembl: ENSG00000106688

IUPHAR-DB: -

Reactome: P43005

LOVD: SLC1A1

Diseases list

- Disease-causing germline mutation(s) in Hot water reflex epilepsy ORPHA:166412 ✓
- Disease-causing germline mutation(s) (loss of function) in Dicarboxylic aminoaciduria ORPHA:2195 ✓

✓ : Assessed

Figure 9: Example of an Orphanet entry (screenshot from https://www.orpha.net/consor/cgi-bin/Disease_Genes.php?lng=EN&data_id=22150&Disease_Disease_Genes_diseaseGroup=SLC1A1&Disease_Disease_Genes_diseaseType=Gen&MISSING%20CONTENT=solute-carrier-family-1-member-1---SLC1A1&search=Disease_Genes_Simple&title=solute%20carrier%20family%201%20member%201%20-%20SLC1A1, accessed 06/15/2020)

Apart from searching for the name or identifiers of a rare disease, Orphanet can be directly queried for genes, leading to a list of rare diseases associated with them. Figure 9 shows the Orphanet entry for SLC1A1, presenting two associated rare diseases.

Apart from the mentioned sections, Orphanet also offers information about orphan drugs, patient organisations, expert centres, ongoing clinical trials and more. All in all, Orphanet contains information about more than 9000 rare diseases. The high number of rare diseases on Orphanet is caused by the fact that it sometimes differentiates between manifestations of diseases that are elsewhere classified as a single disease.[46]

The information provided on the website can also be downloaded from Orphadata at http://www.orphadata.org.[11] Orphadata is a platform powered by Orphanet on which it is possible to download thematically specialised data sets as XML files. It includes free datasets as well as on request data, for which a data transfer agreement needs to be signed.

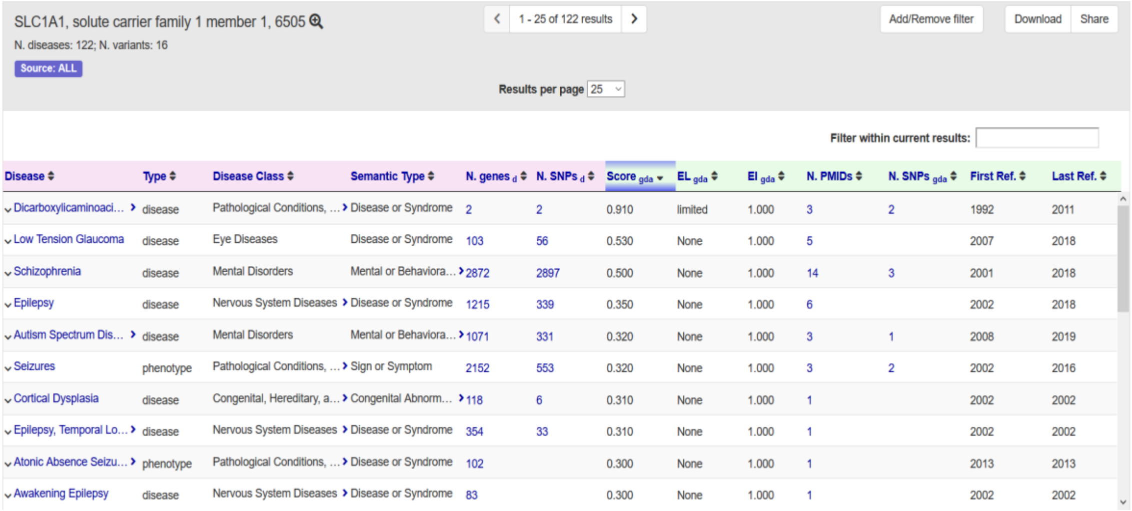
For the present work, two of these files were downloaded and integrated into the workflow for data retrieval: *'Orphanet rare diseases with their associated genes',* version 1.2.11/4.1.6 [2018/04/12] (orientdb version) [47] was used for retrieving associations between SLCs and rare diseases. The file '*Rare diseases and cross-referencing',* version 1.2.11/4.1.6 [2018/04/12] (orientdb version) [48] was used for adding external identifiers to the emerging dataset due to two reasons: the first reason is the join with the dataset from DisGeNET, as DisGeNET uses the UMLS ID as the identifier for diseases. The second reason is to provide an extensive set of identifiers to users accessing the second workflow via the KNIME WebPortal, so that it would be easily possible to join the results with datasets from other databases.

2.2.4 DisGeNET

DisGeNET [12] is a platform that offers one of the most extensive collections of gene-disease and variant-disease associations. The latest release available at the time of the analysis part of this work (version 6.0) contains more than 600.000 gene-disease-associations.

The information about gene-disease-associations in the DisGeNET platform derives from sixteen different sources. These sources are classified into the categories 'curated', 'literature-derived', 'animal models' and 'inferred'.[49] Particular mention should be made of the "literature-derived" sources LHGDN and BeFree as the data is extracted by text mining. [50], [51] 60% of the gene-disease associations listed in DisGeNET derive from text mining and are not described in any of the curated sources.[49] Text mining can be a great advantage, as there is always an unmanageable amount of newly published literature which can only be efficiently accessed by automatic tools. [49] Because of this, DisGeNET contains many new possible associations, which can be especially interesting for rare diseases, as it significantly increases the number of retrieved genes associated with these diseases. [49] On the other hand, it also poses the threat of false-positive findings as further described in section 3.2,p.53f.

There are several ways to access the data provided on DisGeNET. It can be, for example, directly queried at https://www.disgenet.org/ where it is possible to search for specific diseases, genes or variants.



SLC1A1, solute carrier family 1 member 1, 6505

N. diseases: 122; N. variants: 16

Source: ALL

1 - 25 of 122 results

Results per page 25

Filter within current results:

Disease	Type	Disease Class	Semantic Type	N. genes $_d$	N. SNPs $_d$	Score $_{gda}$	EL $_{gda}$	EI $_{gda}$	N. PMIDs	N. SNPs $_{gda}$	First Ref.	Last Ref.
Dicarboxylicaminoaci…	disease	Pathological Conditions, …	Disease or Syndrome	2	2	0.910	limited	1.000	3	2	1992	2011
Low Tension Glaucoma	disease	Eye Diseases	Disease or Syndrome	103	56	0.530	None	1.000	5		2007	2018
Schizophrenia	disease	Mental Disorders	Mental or Behaviora…	2872	2897	0.500	None	1.000	14	3	2001	2018
Epilepsy	disease	Nervous System Diseases	Disease or Syndrome	1215	339	0.350	None	1.000	6		2002	2018
Autism Spectrum Dis…	disease	Mental Disorders	Mental or Behaviora…	1071	331	0.320	None	1.000	3	1	2008	2019
Seizures	phenotype	Pathological Conditions, …	Sign or Symptom	2152	553	0.320	None	1.000	3	2	2002	2016
Cortical Dysplasia	disease	Congenital, Hereditary, a…	Congenital Abnorm…	118	6	0.310	None	1.000	1		2002	2002
Epilepsy, Temporal Lo…	disease	Nervous System Diseases	Disease or Syndrome	354	33	0.310	None	1.000	1		2002	2002
Atonic Absence Seizu…	phenotype	Pathological Conditions, …	Disease or Syndrome	102		0.300	None	1.000	1		2013	2013
Awakening Epilepsy	disease	Nervous System Diseases	Disease or Syndrome	83		0.300	None	1.000	1		2002	2002

Figure 10: Example of gene-disease associations on DisGeNET for SLC1A1 (screenshot from https://www.disgenet.org/browser/1/1/0/6505/, accessed 06/15/2020)

Figure 10 shows an example part of the query of diseases associated with the gene SLC1A1. Apart from information about the condition like its full name, the type, the MeSH disease class, and the number of associated genes, it offers metrics that can be used for ranking and filtering the gene-disease associations.

An example would be DisGeNET's in-house developed metric system: the DisGeNET score. It is calculated based on the number and the type of sources (curated, animal model, inferred and literature-derived) supporting a gene-disease association. Further details on how it is calculated are provided at https://www.disgenet.org/dbinfo#score.

Apart from the information shown in Figure 10, DisGeNET also includes details about the listed diseases such as the UMLS ID, the disease name and the associated MeSH disease class.

Besides the direct query on the website, DisGeNET offers different data sets for the download as tabulated files at http://www.disgenet.org/downloads. For this work, two of these files were implemented into the workflow. The link to the downloadable files is integrated inside the workflow for data retrieval to update the data automatically.

*'ALL gene-disease-pmid associations' (*https://www.disgenet.org/static/disgenet_ap1/files/downloads/all_gene_disease_pmid_associations.tsv.gz*)* [52] was used for retrieving SLC-rare disease associations.

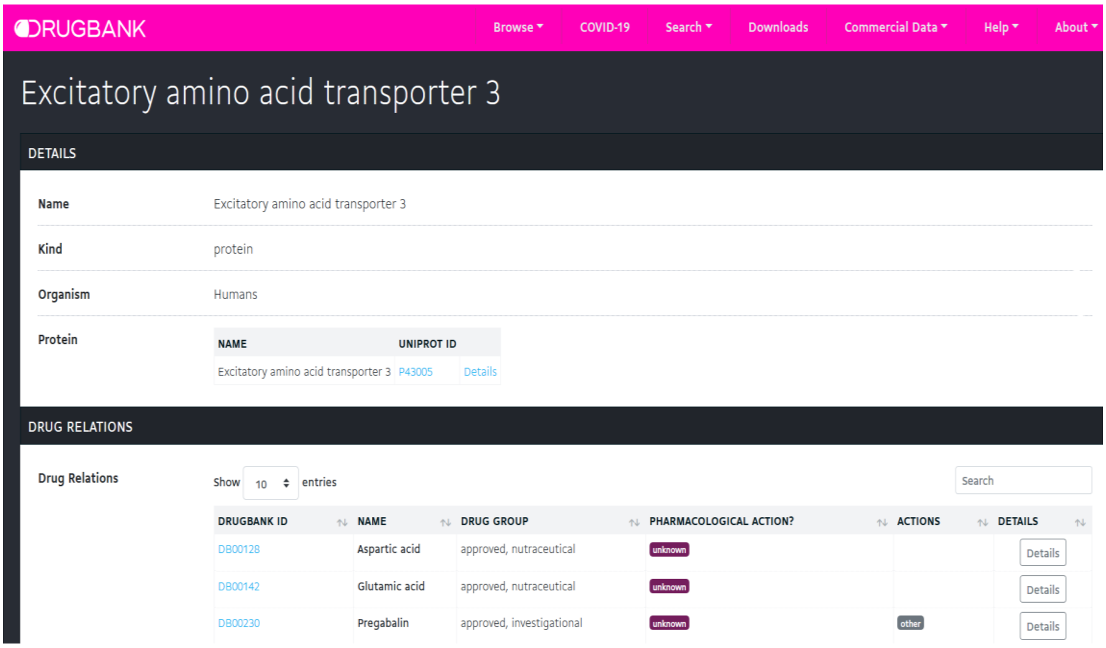
The second file was the '*BeFree gene-disease-pmid associations for Pubannotations*[53] dataset *(*https://www.disgenet.org/static/disgenet_ap1/files/downloads/pubannotator.tsv.gz) that contains the sentence on MEDLINE causal for the association retrieved through BEFREE text mining. This file can make it easier to curate the associations later on.

As the links to the files are integrated into the workflow, the dataset is designated to be updated to its newest available version. The version used for the analysis part of this work is version 7.0, released on 05/04/2020.

2.2.5 DrugBank

DrugBank is a wide-ranging, entirely curated web database that provides knowledge about drugs and drug-target-associations for drugs, that are already FDA-approved as well as experimental drugs and nutraceuticals. It was first released in 2006 and the latest big update at the time of the creation of the workflow, version 5.0, was published in 2018.[13], [54], [55]

Each drug entry on DrugBank contains more than 200 data fields, including information about the drug as well as about the associated drug targets. Compounds are annotated with detailed information about the chemical, pharmacological and pharmaceutical characteristics while the target information includes sequences, structures and pathways.



Excitatory amino acid transporter 3

DETAILS

Name: Excitatory amino acid transporter 3

Kind: protein

Organism: Humans

Protein:

NAME	UNIPROT ID	
Excitatory amino acid transporter 3	P43005	Details

DRUG RELATIONS

Drug Relations

Show 10 entries

Search

DRUGBANK ID	NAME	DRUG GROUP	PHARMACOLOGICAL ACTION?	ACTIONS	DETAILS
DB00128	Aspartic acid	approved, nutraceutical	unknown		Details
DB00142	Glutamic acid	approved, nutraceutical	unknown		Details
DB00230	Pregabalin	approved, investigational	unknown	other	Details

Figure 11: Example of DrugBank entry for SLC1A1 (screenshot from https://www.drugbank.ca/bio_entities/BE0001054, accessed 06/15/20)

Figure 11 shows the DrugBank entry for the target SLC1A1. Under 'DRUG RELATIONS', a list of drugs interacting with this protein is provided including their 'DRUGBANK ID', Drugbank's unique accession number, their stage of approval and, when known, the pharmacological drug actions.

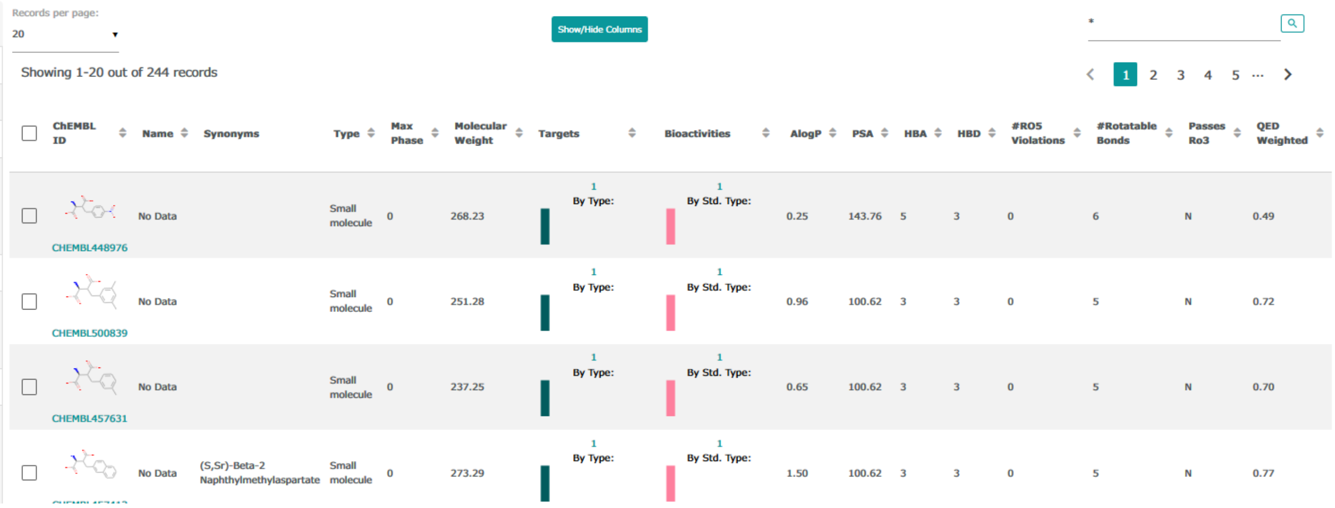
For the present work, the data from DrugBank was downloaded as an XML file from the website containing the ‘complete database’. It is necessary to create a free account (for non-commercial use) to get access to this dataset [56]. The version used for the workflow and the analysis part of this work was version 5.1.2, released on 12/20/2018 [57].

2.2.6 ChEMBL

ChEMBL is an extensive, manually curated database for bioactivity of drug-like molecules. The database contains bioactivity, molecule and target data from altogether 48 sources with a significant part of the data deriving from manual extraction of published medicinal chemistry literature. [58]–[61]

ChEMBL standardises the published activity values to make them better comparable, which means that they are, when possible, converted to a preferred standard type or unit (e.g. IC50_mean, mean IC50 and IC50 are all standardised to the standard type IC50). In addition, the pChEMBL value has been added. This value makes several measures, like the molar IC50, EC50, Ki, Kd or Potency, better comparable as they are converted to a linear scale by taking their negative logarithmic values (e.g. the pChEMBL for an IC50 measurement of 1 nM has a value of 9). [58]

The ChEMBL database can be directly accessed using the website where it can be queried for entities such as compounds, targets, assays, documents and more.



Showing 1-20 out of 244 records

□	ChEMBL ID	Name	Synonyms	Type	Max Phase	Molecular Weight	Targets	Bioactivities	AlogP	PSA	HBA	HBD	#RO5 Violations	#Rotatable Bonds	Passes Ro3	QED Weighted
□	CHEMBL448976	No Data		Small molecule	0	268.23	1 By Type:	1 By Std. Type:	0.25	143.76	5	3	0	6	N	0.49
□	CHEMBL500839	No Data		Small molecule	0	251.28	1 By Type:	1 By Std. Type:	0.96	100.62	3	3	0	5	N	0.72
□	CHEMBL457631	No Data		Small molecule	0	237.25	1 By Type:	1 By Std. Type:	0.65	100.62	3	3	0	5	N	0.70
□		No Data	(S,Sr)-Beta-2 Naphthylmethylaspartate	Small molecule	0	273.29	1 By Type:	1 By Std. Type:	1.50	100.62	3	3	0	5	N	0.77

Figure 12: Example of compounds associated with SLC1A1 on ChEMBL (screenshot from https://www.ebi.ac.uk/chembl/g/#browse/compounds/filter/_metadata.related_targets.all_chembl_ids%3ACHEMBL2721, accessed 06/15/20)

Figure 12 shows the beginning of the list of 244 compounds associated with SLC1A1. It includes the 2D structure, its ChEMBL ID, the type, the stage of approval and chemical properties like the molecular weight, the AlogP or the number of rotatable bonds.

Other possibilities to access ChEMBL are using the downloadable files, the semantic website or through the provided web services.

For this work, ChEMBL was accessed via ChEMBL web services. ChEMBL offers a RESTful API that can be accessed via the REST nodes in KNIME, which is further described at the chapter 'Workflows', starting from page 26ff. The default format is XML, but it can also be downloaded in JSON format [14]. As the API request is renewed, every time the workflow runs, it is designated to be updated automatically when a new version is released. The version used for the analysis part of this work, was ChEMBL 27, last updated on 05/18/2020.

2.2.7 PubChem

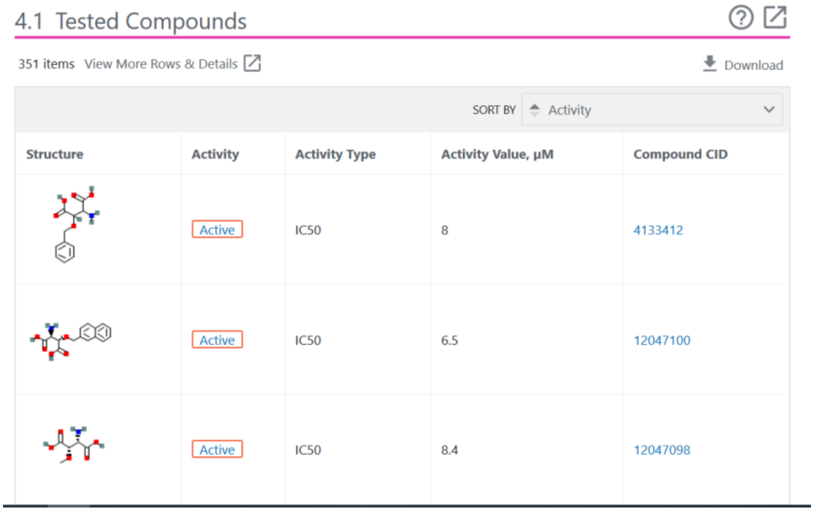
PubChem is a database containing one of the most extensive sets of publicly available information about molecules and bioactivities. It is an open database, which means that everyone can upload their scientific data to PubChem [15], [16]. In May 2020, it contained more than 100 million unique structures and almost 270 million bioactivity data points from more than 700 sources [62]. The data on PubChem is organised into three interconnected databases: PubChem Bioassays, PubChem Substances and PubChem Compounds.

'SIDs' (Substance IDs) refer to the IDs given to a substance when uploaded by a contributor, which is why one structure can have several SIDs. In contrast, 'CIDs' (Compound IDs) denote to unique structures after a standardisation process that aggregates all of the substance records for the same molecule. PubChem's assay identifier is called 'AID' [16].

Apart from the direct query via the website, PubChem offers two ways to access its data programmatically - the PUG-REST and the PUG-SOAP. More information can be found at https://pubchemdocs.ncbi.nlm.nih.gov/programmatic-access.

Unfortunately, using the programmatic access would currently lead into many inconvenient, intermediate steps as it is, starting from targets (EntrezGeneIDs), only possible to search for AIDs. Receiving active CIDs associated with SLCs and their bioactivity data would result in five API calls with several intermediate steps.

At the same time, it is possible to download CSV files with all tested compounds per target directly from the PubChem web interface. As these CSV files already contain all of the desired information, the data was automatically downloaded within the workflow. The data is updated to the newest version, every time the workflow runs on the server. The data used for the analysis part of this work has last been updated on 05/18/2020.



4.1 Tested Compounds

351 items View More Rows & Details

Download

SORT BY Activity

Structure	Activity	Activity Type	Activity Value, µM	Compound CID
	Active	IC50	8	4133412
	Active	IC50	6.5	12047100
	Active	IC50	8.4	12047098

Figure 13: Example of compounds associated with SLC1A1 on PubChem (screenshot from https://pubchem.ncbi.nlm.nih.gov/gene/6505#section=Chemicals-and-Bioactivities, accessed 06/15/2020)

Figure 13 shows the beginning of the section that includes all tested compounds and that is downloaded per target for SLC1A1. It contains the structure, the activity type and value and the PubChem CID.

The PUG-REST API was additionally used for retrieving additional information: associated Molecule/Drug names and Canonical SMILES.

2.2.8 Datasets and content of file ‘SLCs and rare diseases’

The information extracted about SLCs and rare diseases can be roughly divided into three categories: attributes describing the genes/proteins, attributes describing the rare diseases and attributes describing the gene-disease associations.

The following tables, Table 1 -

Table 7, list the datasets together with the attributes used for the emerging table. The left column shows the name used in the emerging KNIME table while the right column offers a short description when necessary.

Table 1: attributes retrieved from the file RESOLUTE_SLCs

attributes describing the gene/protein	description of attributes
SLC name	HGNC gene symbol (see SLCs, p.3f.)
SLC family	
EntrezGene ID	identifier by NCBI
UniProt ID	identifier by UniProt KB

Table 2: attributes retrieved from the UniProt KB REST API

attributes describing the gene/protein	description of attributes
Protein name and aliases	protein name & aliases from UniProt KB
Gene aliases	gene aliases from UniProt KB

Table 3: attributes retrieved from *'Orphanet rare diseases with their associated genes'*, version 1.2.11/4.1.6 [2018-04-12] (orientdb version)

attributes describing the gene/protein	description of attributes
UniProt ID	corresponding to SwissProt ID, used for joining with RESOLUTE_SLC
attributes describing the disease	**description of attributes**
OrphaNUMBER	Orphanet's specific terminology for rare diseases
disease name Orphanet	disease name
attributes describing the disease-gene association	**description of attributes**
PubMed ID	links to articles on PubMed as source of validation
DisorderGeneAssociation	e.g. 'disease-causing germline mutation(s) in', only available for few associations

Table 4: attributes retrieved from ‘Rare diseases and cross-referencing’, version 1.2.11/4.1.6 [2018-04-12] (orientdb version)

attributes describing the disease	**description of attributes**
OrphaNUMBER	Orphanet’s specific terminology for rare diseases
disease name Orphanet	disease name
synonyms	when available
ICD10	disease classification system
UMLS ID	external identifier, used for joining with DisGeNET
OMIM ID, MesH ID	external identifiers

Table 5: attributes retrieved from DisGeNET, 'ALL gene-disease-pmid associations', renewed every execution, analysis version: version 7.0 [2020-05-04]

attributes describing the gene	**description of attributes**
EntrezGene ID	corresponding to ‘geneID’, used for joining the results with RESOLUTE_SLCs
attributes describing the disease	**description of attributes**
UMLS ID	corresponding to ‘diseaseID’, used for joining the results with Orphanet
disease name DisGeNET	disease name
MeSH class code	disease classification system
attributes describing the disease-gene association	**description of attributes**
PubMed ID	links to articles on PubMed as source of validation
DisGeNET score, DSI, DPI, EI.	metrics provided by DisGeNET, see p.16
source	Sixteen different sources, see p. 15f.

Table 6: attributes retrieved from DisGeNET 'BeFree gene-disease-pmid associations', renewed every execution, analysis version: version 7.0 [2020-05-04]

attributes describing the disease-gene association	**description of attributes**
sentence	sentence on MEDLINE

Table 7: columns added within KNIME

MeSH class name	based on the MesH class code
Database	Orphanet/ DisGeNET as filtering option
Reliability of source	categories described at p. 15f., filtering option

2.2.9 Datasets and content of file 'SLCs and molecules'

The information extracted about compounds can be, again, roughly divided into three categories: attributes describing the genes/proteins, attributes describing the compounds, and attributes describing the gene-compound associations. The following tables, **Fehler! Ungültiger Eigenverweis auf Textmarke.** - Table 12, list the datasets together with the attributes used for the emerging table.

Table 8: attributes retrieved from DrugBank, *'All drugs'*, version 5.1.2 [2018-12-20]

attributes describing the gene/protein	**description of attributes**
UniProt ID	used for joining with RESOLUTE_SLCs
attributes describing the compound	**description of attributes**
Molecule/Drug Name	name of drug
DrugBank ID	specific identifier from DrugBank
attributes describing the compound-gene association	**description of attributes**
Activity Comment	corresponding to 'action', e.g. inhibitor, inducer
PubMed IDs	links to articles on PubMed as source of validation

Table 9: attributes retrieved from ChEMBL via RESTful API, renewed every run, analysis version: version 27, [2020-05-18].

attributes describing the gene/protein	**description of attributes**
UniProt ID	used for retrieving results via API
attributes describing the compound	**description of attributes**
ChEMBL ID	specific identifier from ChEMBL
Molecule/Drug name	Molecule name
Canonical SMILES	specification describing the structure, https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system
attributes describing the compound-gene association	**description of attributes**
Activity name	e.g. IC_{50}, EC_{50}, inhibition
Activity value	activity value in nm or µm
pChEMBL value	standardised value, described at subsection ‘ChEMBL’, p. 18ff.
Assay ID	ChEMBL identifier for assay
Assay description	description of assay
ChEMBL data validity comment	flags potential error, e.g. ‘outside typical range’
ChEMBL potential duplicate flag	flags potential duplicate

Table 10: attributes retrieved from PubChem by download of .CSV files, renewed every run, analysis: last updated 05/18/2020

attributes describing the gene/protein	**description of attributes**
EntrezGene ID	used for retrieving results via API
attributes describing the compound	**description of attributes**
PubChem CID	compound ID
PubChem SID	substance ID

attributes describing the compound-gene association	**description of attributes**
Activity Name	e.g. Km, EC50
Activity Value	activity value in µm
Assay ID	AID identifier
Assay Description	description of assay
PubMed IDs	links to articles on PubMed as source of validation

Table 11: attributes retrieved from PubChem through API calls

attributes describing the compound	**description of attributes**
Canonical SMILES	specification describing the structure
Molecule/Drug Name	title

Table 12: Columns added within KNIME

Database	DrugBank, ChEMBL, PubChem as a filtering option
possible inducer/inhibitor	This column is added based on the type of action (DrugBank) or the assay description (ChEMBL, PubChem), further described at chapter ‘Workflow for data retrieval’, p.31ff.

2.3 Workflow for data retrieval

The first workflow is developed to be run on the KNIME Server. It is scheduled every fifteen days, mainly to update the data from DisGeNET, ChEMBL and PubChem.

The workflow consists of three major steps.

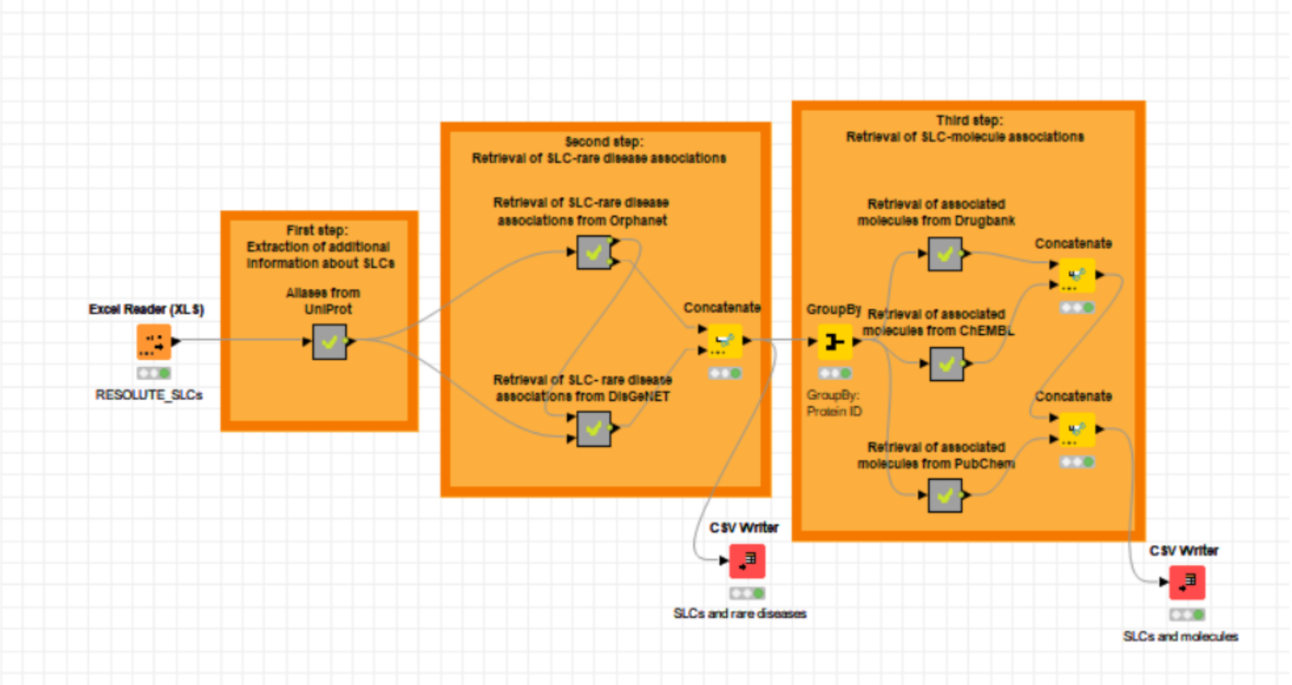

Excel Reader (XLS)
RESOLUTE_SLCs
First step:
Extraction of additional information about SLCs
Aliases from UniProt
Second step:
Retrieval of SLC-rare disease associations
Retrieval of SLC-rare disease associations from Orphanet
Retrieval of SLC- rare disease associations from DisGeNET
Concatenate
CSV Writer
SLCs and rare diseases
Third step:
Retrieval of SLC-molecule associations
Retrieval of associated molecules from Drugbank
GroupBy
Retrieval of associated molecules from ChEMBL
GroupBy: Protein ID
Concatenate
Concatenate
Retrieval of associated molecules from PubChem
CSV Writer
SLCs and molecules

Figure 14: Overview of the workflow for the retrieval of SLC-rare disease-molecule associations

Figure 14 shows an overview of the workflow, consisting of several Metanodes. Each of the Metanodes contains numerous nodes which make the workflow much more complex and are further described in the following subsections.

After the first and the second step, one tabulated file each is automatically saved directly on the KNIME Server. As the data needs to be curated, these files can be accessed, curated and downloaded through the second workflow at the WebPortal (see section 2.4, p.38ff.).

The file *RESOLUTE_SLCs* forms the starting point of the workflow. As a first step, protein aliases are extracted from the UniProt KB by using the API to provide a more extensive set of parameters. Then, rare diseases associated with SLCs are retrieved from Orphanet and DisGeNET. As a third step, associated drugs and molecules are retrieved from DrugBank, ChEMBL and PubChem.

As the workflow consists, altogether, of more than 100 nodes, and some parts were adapted from workflows created by other members of the Pharmacoinformatics Research Group, not every node and its configuration is described in detail.

2.3.1 Extraction of additional information about SLCs

The starting point of the workflow is the file RESOLUTE_SLCs. It was provided from the RESOLUTE project in June 2019 and is customised for its purpose in the workflow, as only the columns ‘SLC name’, ‘SLC family’, ‘UniProt ID.’ and ‘Entrez Gene ID’ were kept in the file used for this work.

After importing it with an *Excel Reader*, the URI for the UniProt REST API is generated with a *String Manipulation* node. The API is used to provide Protein and Gene Aliases for the curation of results in the second workflow, as shown in subsection 3.2,p.53ff.

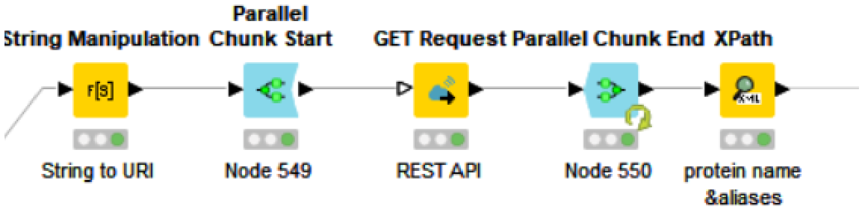


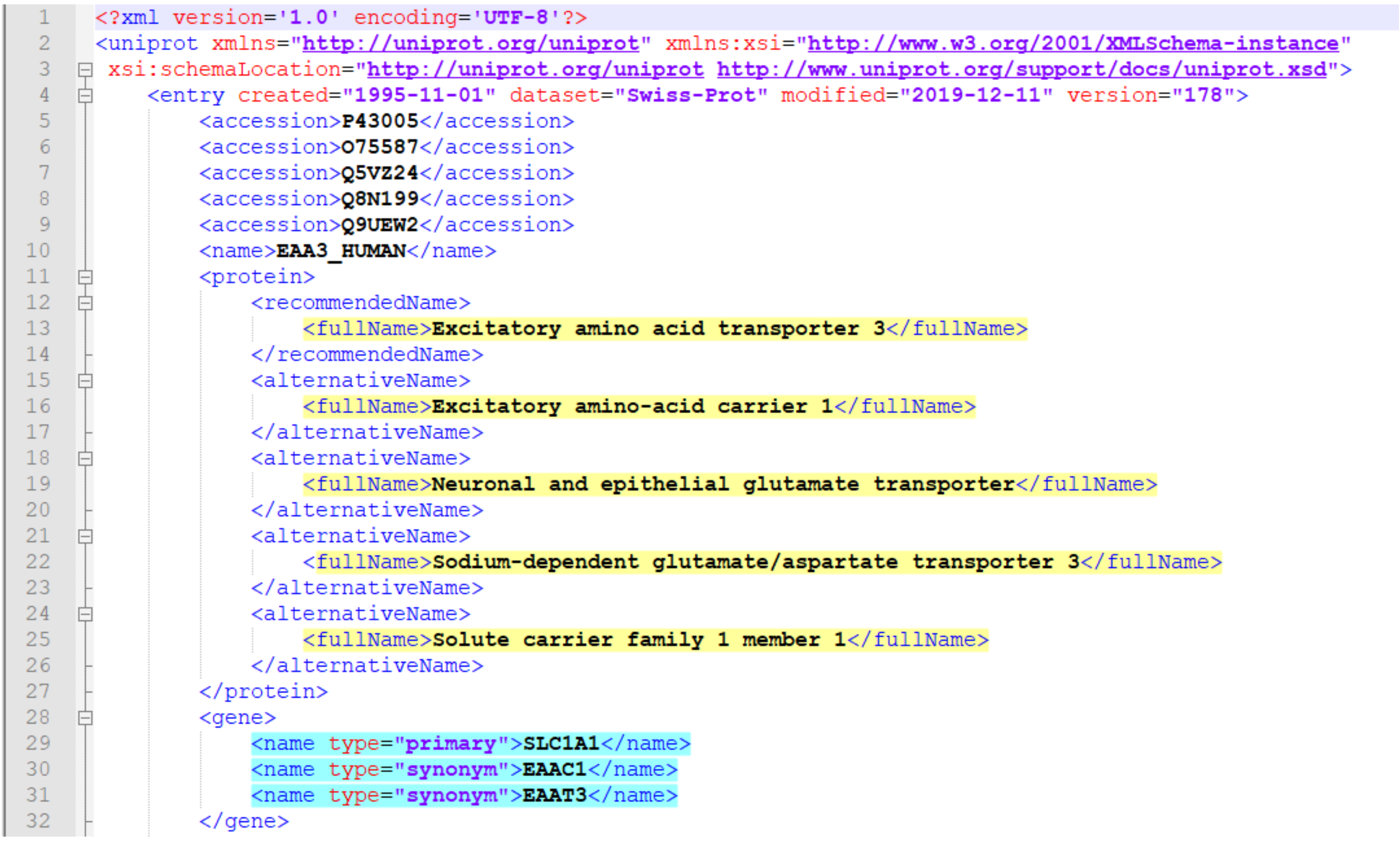
Figure 15: Nodes for the Extraction of Protein & Gene Aliases from UniProt

The URL for the REST API call is generated in a *String Manipulation* node from the expression ‘*join("https://www.uniprot.org/uniprot/",$UniProt.ID$,".xml*")’.

The URL consists of a data set, here ‘uniprot’ and the entry's unique identifier, here ‘*$UniProt.ID$*’.

The expression ‘*$UniProt.ID$*’ specifies that the value for each row is taken directly from the column *UniProt.ID* which means that the UniProt entry is downloaded for each UniProt ID listed in the SLC table. The syntax ‘*.xml*’ specifies that the entries are downloaded in XML format. Other possible formats would be for example .txt, .rdf and .fasta.

The *String Manipulation* node is followed by a *Parallel Chunk Start* node, that splits the API call rows into smaller chunks of the same size that are executed in parallel by the following *GET Request* node as this speeds up the process. The *Parallel Chunk End* node collects the results. The resulting XML files are further processed with an *XPath* node.



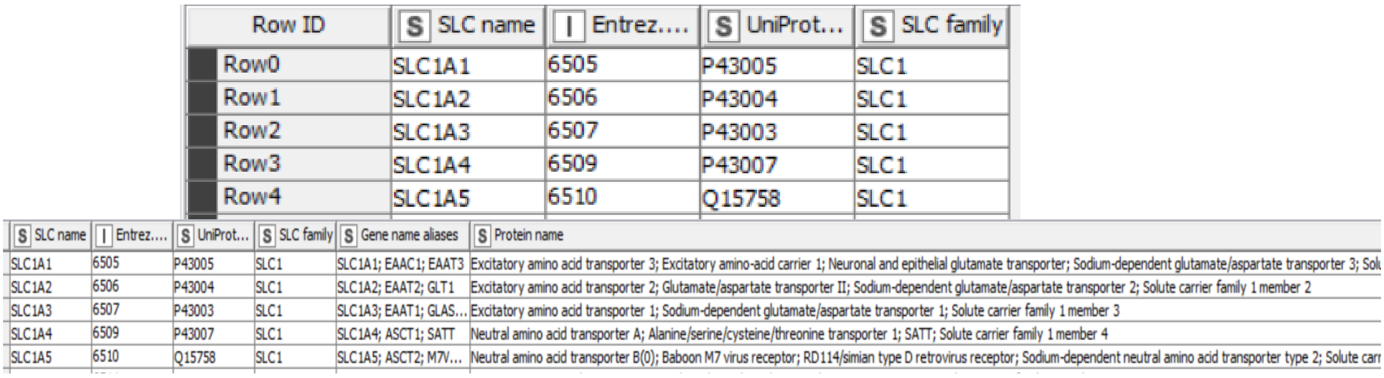
```
<?xml version='1.0' encoding='UTF-8'?>
<uniprot xmlns="http://uniprot.org/uniprot" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://uniprot.org/uniprot http://www.uniprot.org/support/docs/uniprot.xsd">
    <entry created="1995-11-01" dataset="Swiss-Prot" modified="2019-12-11" version="178">
        <accession>P43005</accession>
        <accession>O75587</accession>
        <accession>Q5VZ24</accession>
        <accession>Q8N199</accession>
        <accession>Q9UEW2</accession>
        <name>EAA3_HUMAN</name>
        <protein>
            <recommendedName>
                <fullName>Excitatory amino acid transporter 3</fullName>
            </recommendedName>
            <alternativeName>
                <fullName>Excitatory amino-acid carrier 1</fullName>
            </alternativeName>
            <alternativeName>
                <fullName>Neuronal and epithelial glutamate transporter</fullName>
            </alternativeName>
            <alternativeName>
                <fullName>Sodium-dependent glutamate/aspartate transporter 3</fullName>
            </alternativeName>
            <alternativeName>
                <fullName>Solute carrier family 1 member 1</fullName>
            </alternativeName>
        </protein>
        <gene>
            <name type="primary">SLC1A1</name>
            <name type="synonym">EAAC1</name>
            <name type="synonym">EAAT3</name>
        </gene>
```

Figure 16: Example XML file as retrieved via UniProt API

Two XPath queries are used to retrieve the recommended and alternative names of the proteins (highlighted in yellow in Figure 16) as well as the names of their encoding genes (highlighted in light blue).

The XPath query for gene name aliases can easily be created in the configuration window of the XPath node by clicking on the required attribute. It only needs to be considered that the ‘Multiple tag option’ of the XPath query setting is set to ‘Multiple Rows’ as the default setting ‘Single Cell’ would result in a single row containing only the first entry. As the protein names are not used for further automatic processing, recommended as well as alternative protein names are combined in a single column named ‘Protein Name’. The XPath expression used for this extraction is *‘//dns:fullname’.* The double slash configures that not only attribute nodes from the root element, but all nodes in the documents that match the expression are selected.

A *GroupBy* node follows that groups the rows per ‘UniProt ID’. ‘Protein Name’ and ‘Gene Aliases’ are aggregated and concatenated with commas in between. The new information is then joined with the original table. Figure 17 shows a screenshot of the table before and after joining Gene and Protein names and aliases.



Row ID	S SLC name	I Entrez....	S UniProt...	S SLC family
Row0	SLC1A1	6505	P43005	SLC1
Row1	SLC1A2	6506	P43004	SLC1
Row2	SLC1A3	6507	P43003	SLC1
Row3	SLC1A4	6509	P43007	SLC1
Row4	SLC1A5	6510	Q15758	SLC1

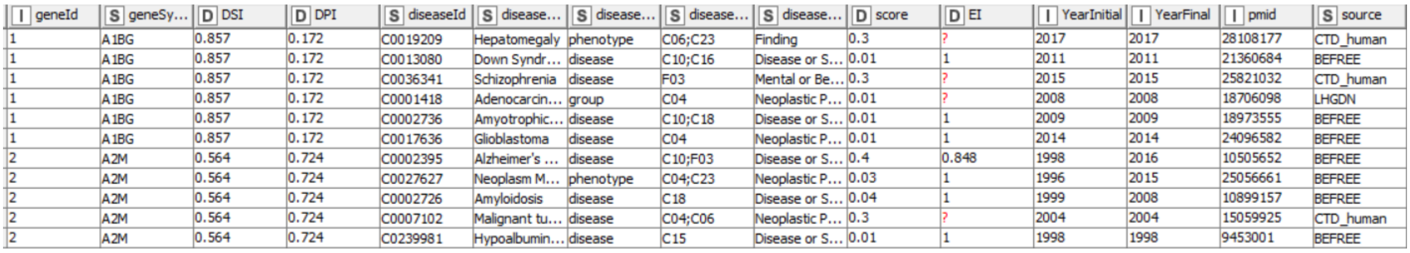
S SLC name	I Entrez....	S UniProt...	S SLC family	S Gene name aliases	S Protein name
SLC1A1	6505	P43005	SLC1	SLC1A1; EAAC1; EAAT3	Excitatory amino acid transporter 3; Excitatory amino-acid carrier 1; Neuronal and epithelial glutamate transporter; Sodium-dependent glutamate/aspartate transporter 3; Solu
SLC1A2	6506	P43004	SLC1	SLC1A2; EAAT2; GLT1	Excitatory amino acid transporter 2; Glutamate/aspartate transporter II; Sodium-dependent glutamate/aspartate transporter 2; Solute carrier family 1 member 2
SLC1A3	6507	P43003	SLC1	SLC1A3; EAAT1; GLAS...	Excitatory amino acid transporter 1; Sodium-dependent glutamate/aspartate transporter 1; Solute carrier family 1 member 3
SLC1A4	6509	P43007	SLC1	SLC1A4; ASCT1; SATT	Neutral amino acid transporter A; Alanine/serine/cysteine/threonine transporter 1; SATT; Solute carrier family 1 member 4
SLC1A5	6510	Q15758	SLC1	SLC1A5; ASCT2; M7V...	Neutral amino acid transporter B(0); Baboon M7 virus receptor; RD114/simian type D retrovirus receptor; Sodium-dependent neutral amino acid transporter type 2; Solute carr

Figure 17: Table before and after joining Gene and Protein names

2.3.2 Extraction of rare diseases

The extraction of rare diseases was adapted from a workflow created by Jana Gurinova[2].

It starts with reading in the file *ALL gene-disease-pmid associations*[52] from DisGeNET, which is automatically downloaded from the DisGeNET download page as a tab-separated file (.tsv) each time the workflow is executed. The *File Reader* node reads the data directly from the URL location.



I geneId	S geneSy...	D DSI	D DPI	S diseaseId	S disease...	S disease...	S disease...	S disease...	D score	D EI	I YearInitial	I YearFinal	I pmid	S source
1	A1BG	0.857	0.172	C0019209	Hepatomegaly	phenotype	C06;C23	Finding	0.3	?	2017	2017	28108177	CTD_human
1	A1BG	0.857	0.172	C0013080	Down Syndr...	disease	C10;C16	Disease or S...	0.01	1	2011	2011	21360684	BEFREE
1	A1BG	0.857	0.172	C0036341	Schizophrenia	disease	F03	Mental or Be...	0.3	?	2015	2015	25821032	CTD_human
1	A1BG	0.857	0.172	C0001418	Adenocarcin...	group	C04	Neoplastic P...	0.01	?	2008	2008	18706098	LHGDN
1	A1BG	0.857	0.172	C0002736	Amyotrophic...	disease	C10;C18	Disease or S...	0.01	1	2009	2009	18973555	BEFREE
1	A1BG	0.857	0.172	C0017636	Glioblastoma	disease	C04	Neoplastic P...	0.01	1	2014	2014	24096582	BEFREE
2	A2M	0.564	0.724	C0002395	Alzheimer's ...	disease	C10;F03	Disease or S...	0.4	0.848	1998	2016	10505652	BEFREE
2	A2M	0.564	0.724	C0027627	Neoplasm M...	phenotype	C04;C23	Neoplastic P...	0.03	1	1996	2015	25056661	BEFREE
2	A2M	0.564	0.724	C0002726	Amyloidosis	disease	C18	Disease or S...	0.04	1	1999	2008	10899157	BEFREE
2	A2M	0.564	0.724	C0007102	Malignant tu...	disease	C04;C06	Neoplastic P...	0.3	?	2004	2004	15059925	CTD_human
2	A2M	0.564	0.724	C0239981	Hypoalbumin...	disease	C15	Disease or S...	0.01	1	1998	1998	9453001	BEFREE

Figure 18: Example part of DisGeNET's file "ALL gene-disease-pmid associations" as seen in the KNIME table

Figure 18 shows an example part of the file from DisGeNET, as seen in KNIME. The table contains 15 columns including gene & disease ID (Entrez Gene ID and UMLS), the disease name, the MeSH disease class, several metrics, PubMed IDs and sources. It is joined with the table from the first step to filter for SLCs as targets with the setting 'Inner Join', which means that only matching rows show up in the Output Table.

Next, the file '*BeFree gene-disease-pmid associations for Pubannotations'* is also read in with a *File Reader* node via URL. The sentence from MEDLINE causal for the association retrieved via BEFREE text mining is joined into the table as this makes it easier to curate the results manually later on (see section 3.2,p.53f.).

However, DisGeNET does not include the option to filter for rare diseases. For this reason, Orphanet is used as it is a database dedicated explicitly to rare diseases.
The XML file *Orphanet: Rare diseases and cross-referencing* is read in with an *XML reader* node. Inside the *XML reader* node, it is possible to configure an XPath filter.
The XPath filter '/JDBOR/DisorderList/Disorder' results in one row per rare disease.
This file contains several external identifiers for rare diseases. The UMLS identifier is extracted to join the resulting table with the results from DisGeNET. Also, the identifiers OMIM and MeSH, the disease name, ORPHAnumber and synonyms are extracted to provide a comprehensive set of available identifiers. The extraction is achieved via an *XPath* node that follows the *XML reader* node on the workflow. The XPath queries are reused from the workflow of Jana Gurinova [2].
Some disease entries on Orphanet do not contain a known UMLS identifier. However, Orphanet offers its own dataset for target-disease-associations, *Orphanet rare diseases with their associated genes,* which is additionally used for retrieving SLC-rare disease-associations.
An *XPath* node extracts ORPHAnumber, disease name, the DisorderGeneAssociation (only provided for a few disease-target associations) as well as the UniProt ID and the GeneSymbol of the involved target and the PubMed ID as a source of validation. Also, some more adaptations to the dataset are made to provide a more extensive set of information.
The DisGeNET file contains the MeSH disease class codes. However, all disease class codes, separated with semicolons, are listed in one single cell, which would make it difficult to filter for specific MeSH disease classes via KNIME.

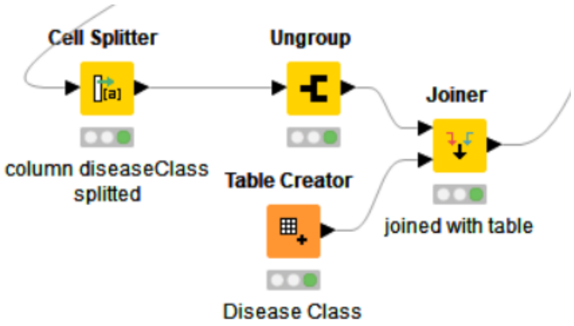


Figure 19: Nodes used for the integration of <MeSH disease class names>

For this reason, the Cell Splitter node is integrated, which splits the cells of the column ‘disease class’ into parts after each semicolon. The output setting is set to ‘list’ which results in one column that contains collection cells. The Ungroup node leads to one row for each MeSH disease class code. The codes are then joined with a table that is manually created based on the MeSH tree view and additionally contains the MeSH disease class names.

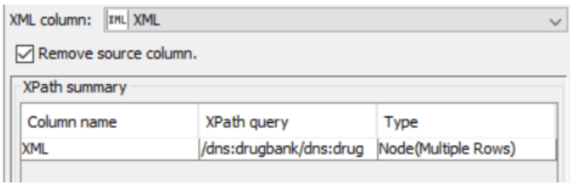
Another parameter suitable for filtering at the second workflow would be the ‘reliability of sources’. The Orphanet database as well as parts of DisGeNET is manually curated. However, some of the gene-disease-associations included in DisGeNET derive from sources that would need further curation. DisGeNET offers four categories for gene-disease associations according to the source, which are further described at subsection 2.2.4, p.15ff. As the source but not the category is included in the dataset, the categories are manually added with a *Rule Engine* node. Then, the results from both databases, Orphanet and DisGeNET, are concatenated and the resulting table is saved as a table file directly on the KNIME Server, as it can be accessed through the second workflow directly from the WebPortal (see subsection 2.4, p.38ff.)

2.3.3 Extraction of drugs/ molecules

Drugs and molecules, associated with SLCs with retrieved rare disease associations, are extracted from three databases simultaneously. As a first step, the resulting table from step two is grouped by UniProt ID (for Drugbank and ChEMBL) and EntrezGene ID (for PubChem) respectively, to extract possible drugs and molecules per unique SLC.

2.3.3.1 Drugbank

The drug extraction from Drugbank was adapted from a workflow created by Jana Gurinova[2]. The XML file is read in with an *XML reader* node. As the file is really large, the execution needs to be done through the KNIME Server, as the memory of most laptops or computers would be overloaded. The XML file contains an extensive set of information. This is why it takes four *XPath* nodes to extract all the necessary information. The first *XPath* node, as shown in Figure 20, divides the large XML file into one drug XML entry per row. The used XPath query is ‘/dns:drugbank/dns:drug’ and it is configured to create node cells, which means that the large XML file is split into smaller XML parts that can be further accessed through *XPath* nodes.



XML column: XML

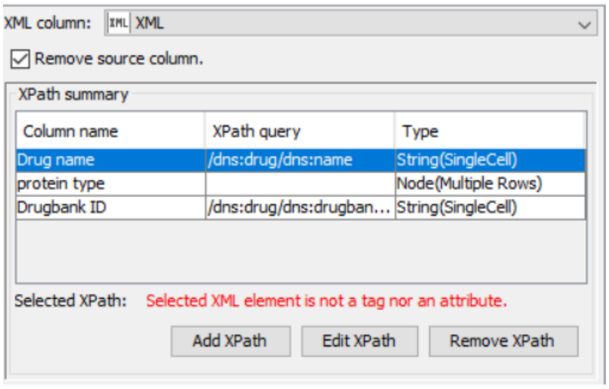
☑ Remove source column.

XPath summary

Column name	XPath query	Type
XML	/dns:drugbank/dns:drug	Node(Multiple Rows)

Figure 20: Configuration of first XPath node for Drugbank

In the second *XPath* node, the drug name and the Drugbank identifier are extracted and the protein type is, again, extracted as node cell.



XML column: XML

☑ Remove source column.

XPath summary

Column name	XPath query	Type
Drug name	/dns:drug/dns:name	String(SingleCell)
protein type		Node(Multiple Rows)
Drugbank ID	/dns:drug/dns:drugban...	String(SingleCell)

Selected XPath: Selected XML element is not a tag nor an attribute.

Figure 21: Configuration of second XPath node for Drugbank

The XPath for the Drug name was reused from Gurinova[2]. The XPath for the protein type was changed as Drugbank lists proteins associated with the listed drugs in four categories: Target, Enzyme, Transporter and Carrier. A set of nodes was used to extract all of the included information as the information provided about proteins is located in different parts of the file based on the protein type.

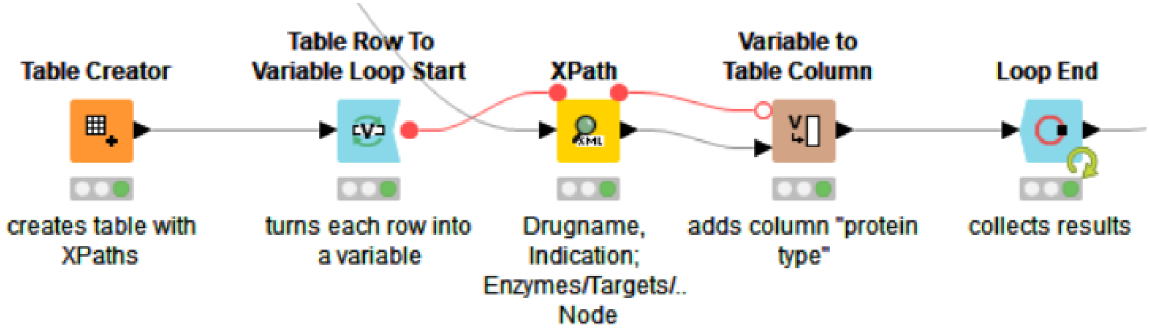
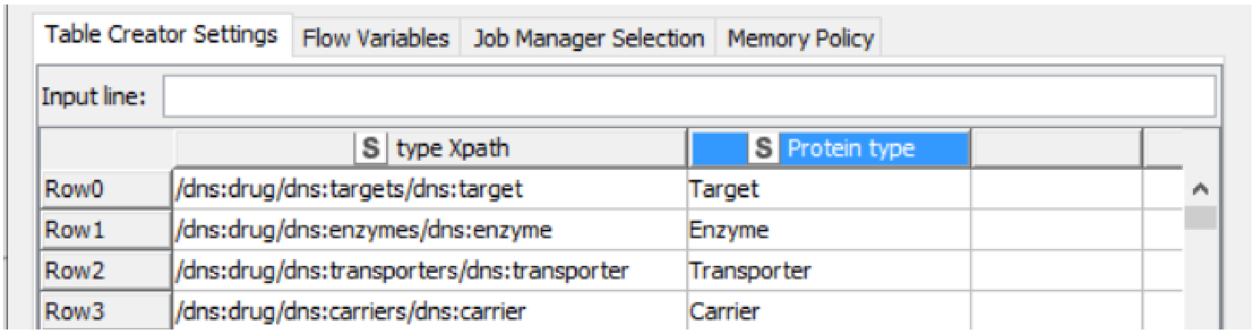


Table Creator
Table Row To Variable Loop Start
XPath
Variable to Table Column
Loop End
creates table with XPaths
turns each row into a variable
Drugname, Indication; Enzymes/Targets/.. Node
adds column "protein type"
collects results

Figure 22: Nodes for the retrieval of 'protein type' node via XPath

The first node is a *Table Creator* node, shown in Figure 23, that creates a table containing all of the four XPaths necessary for the extraction.

A *Table Row To Variable Loop* follows that turns each of the rows into variables because the following *XPath* node is configured to use the emerging variable ‘type XPath’ as XPath for the column ‘protein type’, extracted as node cell.



	S type Xpath	S Protein type	
Row0	/dns:drug/dns:targets/dns:target	Target	
Row1	/dns:drug/dns:enzymes/dns:enzyme	Enzyme	
Row2	/dns:drug/dns:transporters/dns:transporter	Transporter	
Row3	/dns:drug/dns:carriers/dns:carrier	Carrier	

Figure 23: Table creator node for the retrieval of proteins associated with drugs on Drugbank

The *Variable to Table Column* node joins the ‘protein type’ column into the resulting table. After four iterations, the results are collected with a *Loop End* node.

Two more *XPath* nodes follow, with the first one extracting the type of action (only available for few associations) and references and extracting the information about the associated polypeptides as node cell starting from the node cell ‘protein type’. The fourth *XPath* node extracts the protein’s name and UniProt ID out of the node cell ‘polypeptide’.

The extracted information is then joined with the list of UniProt IDs deriving from Step two only to keep only drugs associated with SLCs that are associated with rare diseases. In the next step, drug-protein associations listed as ‘substrate’ are excluded.

Two *Rule Engine* nodes add the two columns ‘possible inhibitor’ and ‘possible inducer’ based on the type of action (e.g., The action type ‘antagonist’ would be listed as ‘possible inhibitor’ while the action type ‘agonist’ would result in ‘possible inducer’).

In the last step, columns are renamed and an additional column named ‘Database’ is inserted that contains the value ‘Drugbank’.

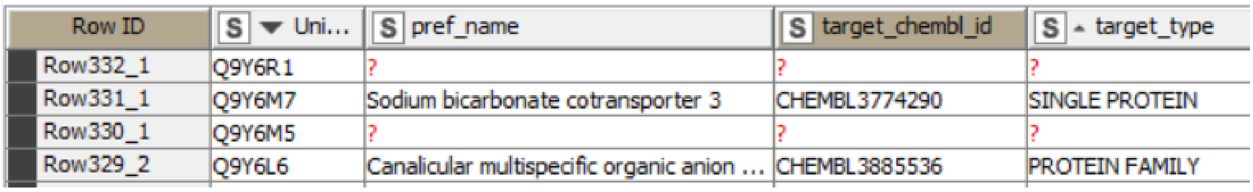
2.3.3.2 ChEMBL

The extraction of drug-like compounds from ChEMBL was adapted from a Metanode created by Daniela Digles[63]. The sequence of nodes and the used APIs were reused from the mentioned Metanode. However, the format for data retrieval was changed from JSON to XML and the following nodes from *JSON Path* to *XPath* nodes to make the workflow more consistent.

This part of the workflow consists basically of two API calls.

The first API call starts with a *String Manipulation* node with the expression: ‘join("https://www.ebi.ac.uk/chembl/api/data/target.xml?target_components__accession=",$UniProt.ID$)’. A *GET Request* node retrieves information about the targets for the provided UniProt IDs.

The following *XPath* node extracts the target’s name, its ChEMBL ID and the target type, as shown in Figure 24.



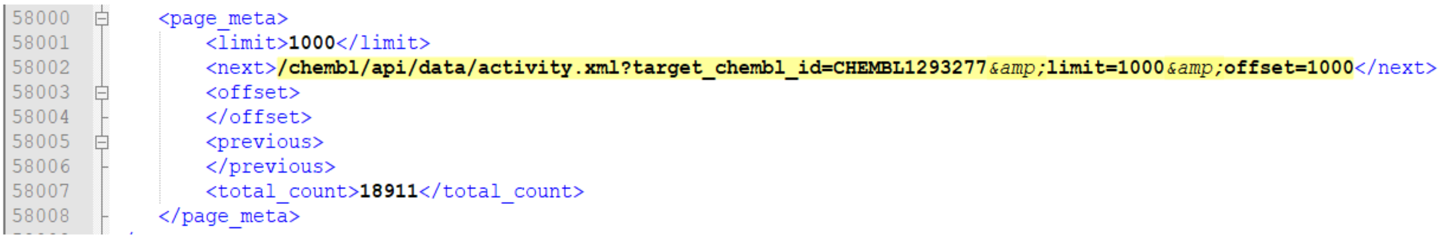
Row ID	S Uni...	S pref_name	S target_chembl_id	S target_type
Row332_1	Q9Y6R1	?	?	?
Row331_1	Q9Y6M7	Sodium bicarbonate cotransporter 3	CHEMBL3774290	SINGLE PROTEIN
Row330_1	Q9Y6M5	?	?	?
Row329_2	Q9Y6L6	Canalicular multispecific organic anion ...	CHEMBL3885536	PROTEIN FAMILY

Figure 24: Table after first ChEMBL API call

Since its update in 2014, ChEMBL distinguishes between different types of protein targets. The target type ‘SINGLE PROTEIN’ specifies that the compound is considered to interact specifically with the protein. However, the target type ‘PROTEIN FAMILY’ indicates, that either the compound interacts non-specifically with all members of a protein family or that the assay conditions make it impossible to identify the specific protein the compound is interacting with.[59]

Because of this, rows containing the target type ‘PROTEIN FAMILY’ as well as empty rows, which means that the specific UniProt ID is not associated with any targets listed on ChEMBL are excluded from the table.

The ‘target_chembl_id’ is necessary as it is part of the URI for the second API call that retrieves the bioactivity data. It starts with a *String Manipulation* node with the expression ‘join("https://www.ebi.ac.uk/chembl/api/data/activity.xml?target_chembl_id=",$target_chembl_id$,"&limit=1000")’. The syntax ‘limit=1000’ specifies that the first 1000 bioactivities are returned. The default limit for an API call on ChEMBL would be 20. The limit can be increased, but 1000 is the maximum allowed value. The ‘page_meta’ section of the resulting XML files provides information about the limit, offset and total count.



```
<page_meta>
    <limit>1000</limit>
    <next>/chembl/api/data/activity.xml?target_chembl_id=CHEMBL1293277&amp;limit=1000&amp;offset=1000</next>
    <offset>
    </offset>
    <previous>
    </previous>
    <total_count>18911</total_count>
</page_meta>
```

Figure 25: Screenshot of the page_meta section of the retrieved XML for NPC1

When a target is associated with more than 1000 bioactivities, like the NPC1 protein shown in Figure 25, the end part of the link to the next page is provided. The NPC1 protein is associated with 18911 bioactivity values, which means that, altogether, 19 API calls are needed to extract all of the data.

Because of this, a set of nodes is used, as shown in Figure 26.

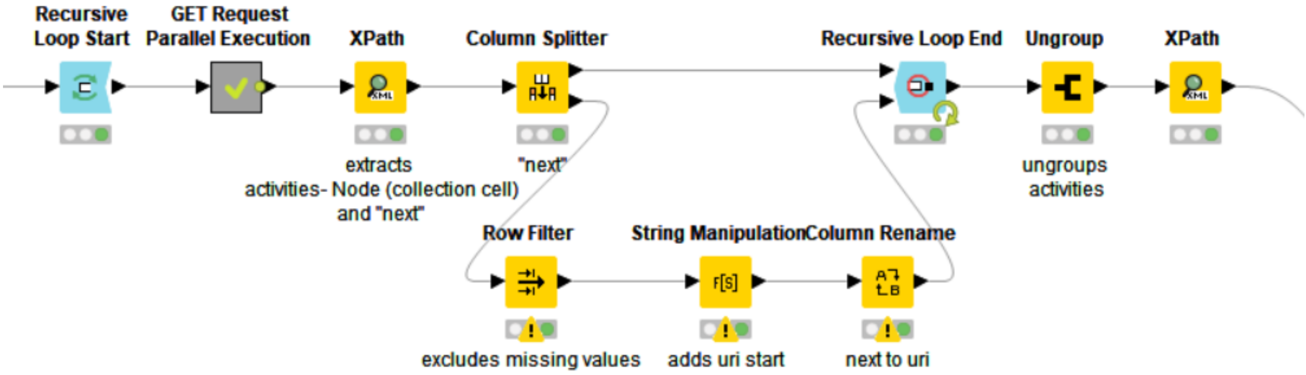


Figure 26: Nodes used for the retrieval of bioactivity data from ChEMBL

The basis of this part of the workflow is a *Recursive Loop* pair, consisting of a *Recursive Loop Start* and a *Recursive Loop End* node.

Data passed to port 0 (the top port) of the *Recursive Loop End* node is collected while data passed to port 1 (the bottom port) of the *Recursive Loop End* is returned to the *Recursive Loop Start*. The end part of the link provided in the page_meta section of the retrieved XML files is extracted with the *XPath* node that follows the API call.

In addition, this *XPath* node extracts the bioactivity data as a node collection cell, which means that every row contains a list of XML files including information about the bioactivity only. After the extraction, a *Column Splitter* node splits the columns into two tables: The bioactivity data is passed to port 0 and therefore collected. The link is completed with a *StringManipulation* node and moved to port 1 and thus to the *Recursive Loop Start* node for the next iteration. After all of the data is retrieved, an *Ungroup* node leads to one row per bioactivity value and an *XPath* node extracts the desired information.

The extracted data includes information about the bioassays, the molecules and the bioactivity data as well as the originating source.

ChEMBL contains data from altogether 48 sources, including parts of the database PubChem. As PubChem is used as another source of SLC-molecule associations (see 2.3.3.3), the bioactivity values originating from this source are excluded in a first step.

The extracted information about bioassays contains the assay ChEMBL ID, the assay description and the assay type. ChEMBL distinguishes between six types of assays including (A) for ADME data assays, (B) for Binding assays, (F) for Functional assay, (T) for Toxicity assays, (P) for Physicochemical assays and (U) for Unclassified. As only the results from binding assays are relevant for the aim of this work, the rest is filtered out.

Only a few rows include an activity comment. However, as some of the bioactivity values contain the activity comment ‘inactive’ or ‘not active’, these rows are filtered out. The ‘data validity comment’ flags activity values that are, for example, outside a typical range for that specific activity type or seem to derive from a transcription error. Furthermore, potential duplicates are flagged in an additional column. These columns are integrated into the emerging table, to let users of the second workflow decide whether or not to keep these values in their results.

In the next step, substrates are excluded as all molecule-target associations containing the activity type ’Km’ or ‘Vmax’ or the activity comment ‘substrate’ are filtered out. Two *Rule Engine* nodes add the columns ‘possible inhibitor’ and ‘possible inducer’ based on the activity comment and the assay description: Assay descriptions including the syntaxes ’inhibitor’, ’inhibition’ or ’antagonist’ are listed as ’possible inhibitor’. In contrast, assay descriptions containing ’induction’, ’inducer’, ’activator’ or ’modulator’ are listed as ‘possible inducer’.

In the last step, the columns are renamed and the additional column ‘Database’ is inserted with the value ‘ChEMBL’.

2.3.3.3 PubChem

The extraction of compounds from PubChem was adapted from a workflow created by Anna Seiler[64].

It starts with a *String Manipulation* node creating the URL links for the download of tested compounds associated with SLCs (beforehand grouped by EntrezGene ID) using the expression ‘join("https://pubchem.ncbi.nlm.nih.gov/sdq/sdqagent.cgi?infmt=json&outfmt=jsonp&query={%22download%22:%22*%22,%22collection%22:%22bioactivity%22,%22where%22:{%22ands%22:[{%22geneid%22:%22", $EntrezGene ID$,"%22},{%22cid%22:%22notnull%22},{%22activity%22:%22Active%22}]},%22order%22:[%22relevancescore,desc%22],%22start%22:1,%22limit%22:1000000}")’

As a very long list of 'inactive' compounds is listed for some of the SLCs and this would lead to a significant amount of unnecessary data as well as to a lag in time, the download is already specified to compounds listed as 'active'.

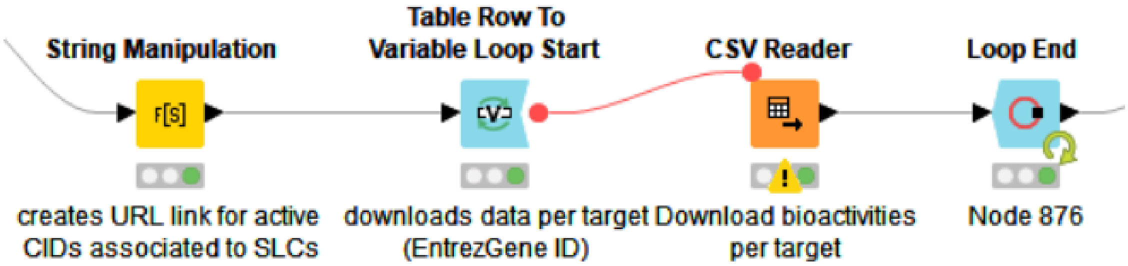

String Manipulation
creates URL link for active CIDs associated to SLCs
Table Row To Variable Loop Start
downloads data per target (EntrezGene ID)
CSV Reader
Download bioactivities per target
Loop End
Node 876

Figure 27: Nodes for the download of active CIDs associated with SLCs from PubChem

A *Table Row To Variable Loop Start* node turns row after row into variables for each loop iteration, the *CSV Reader* node reads in the CSV files deriving from the provided URL and the *Loop End* node collects all of the data.

The result is a table containing compounds listed as 'active' towards SLCs, including information about the bioassay and the bioactivity values.

As PubChem contains data from ChEMBL, but ChEMBL is used as another source for retrieving SLC-molecule-association, data deriving from the source 'ChEMBL' is excluded in a first step.

Next, two API calls follow that retrieve additional information:

The first API call deriving from the *String Manipulation* node with the expression join("https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/",string(cid),"/description/XML") retrieves all descriptions associated with the compound to extract the title which corresponds to the name of the molecule. The second API call retrieves Canonical SMILES describing the structure.

The additional columns 'possible inducer' and 'possible inhibitor' are then added based on the assay description similar to ChEMBL (see 2.3.3.2, p. 33f.). The results are joined with UniProt IDs for concatenating with the results from Drugbank and ChEMBL, and the additional column 'Database' with the value 'PubChem' is added.

The results from Drugbank, ChEMBL and PubChem are then concatenated and joined with SLC name, family and EntrezGene ID.

The resulting table is again saved as table file on the KNIME Server to be accessed through the second workflow.

2.4 Workflow for accessing data (end-user)

A second workflow was created for interested users to access, filter and curate the aggregated data at the KNIME WebPortal as the data needs manual curation, and this would, due to the amount of data, exceed the time constraints of a diploma thesis.

This part of the work describes the workflow at the KNIME Analytics Platform, as shown in Figure 28. The workflow as seen by the user at the KNIME WebPortal is further described in section 3, p.53ff.

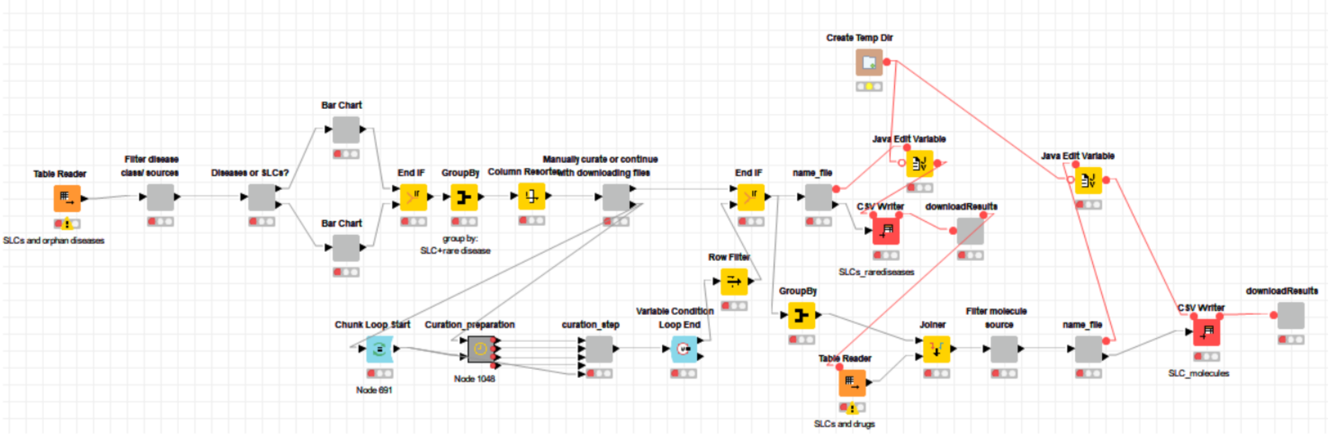


Figure 28: Workflow for interested users as seen in the KNIME Analytics Platform

It starts with importing the file ‘SLCs and rare diseases’ directly from the KNIME Server, where it has been saved during the last scheduled execution of the data retrieval workflow (see 2.3.2). Its content and originating databases are further described in section 2.2.8, p.20f.

The workflow is then made out of several components containing widgets or view nodes, offering interactive filtering options via the WebPortal.

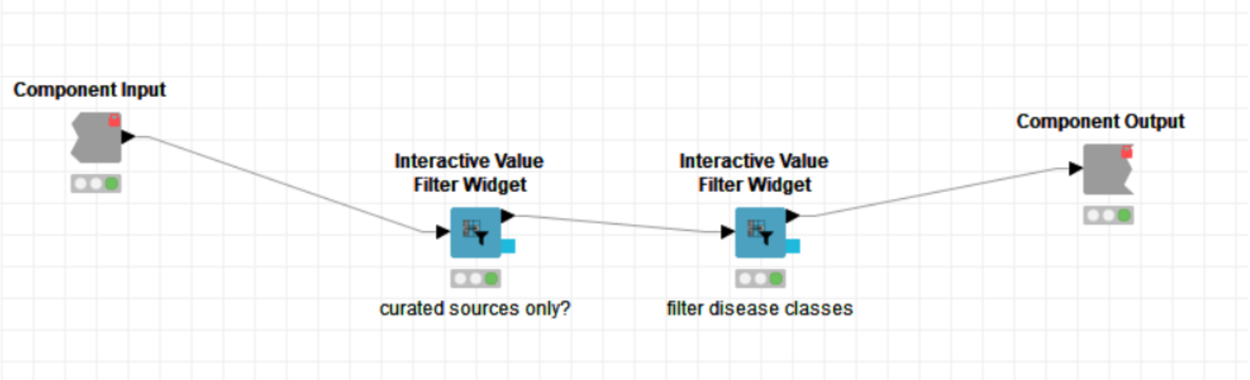
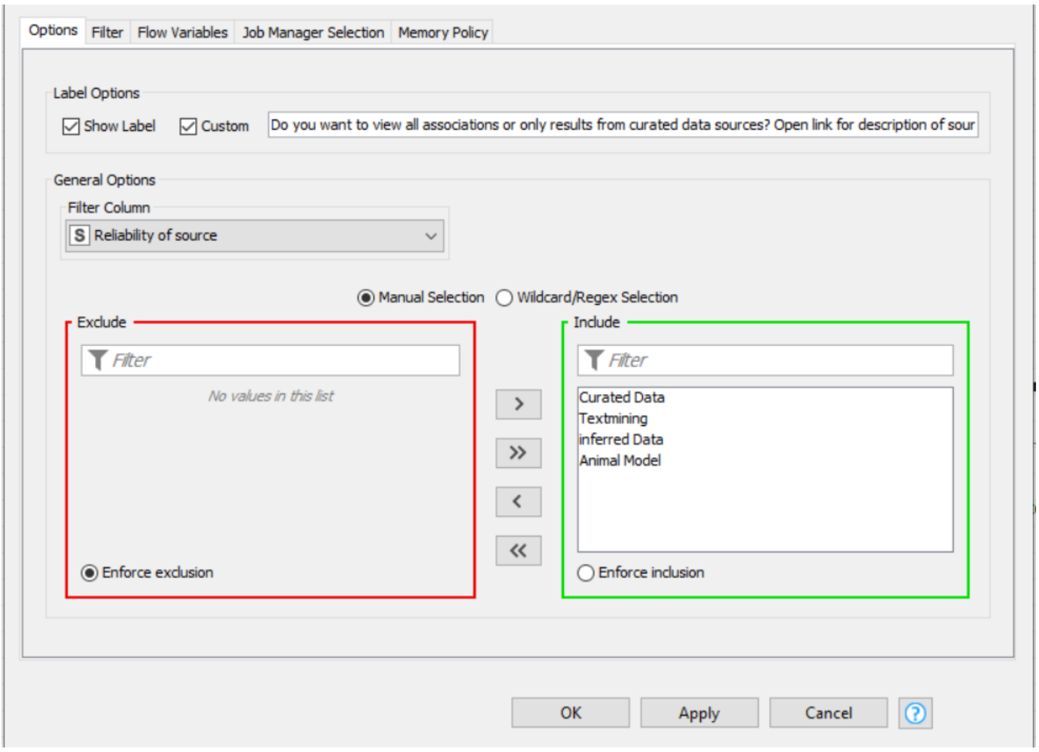


Figure 29: Inside Component ‘Filter disease classes/sources’

Figure 29 shows the inside of the first component. It consists of two *Interactive Value Filter Widget* nodes. The first node lets the user choose if results from all sources should be considered or only the ones from curated sources.



Label Options

☑ Show Label ☑ Custom Do you want to view all associations or only results from curated data sources? Open link for description of sour

General Options

Filter Column

S Reliability of source

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No values in this list

● Enforce exclusion

Include

Filter

Curated Data
Textmining
inferred Data
Animal Model

○ Enforce inclusion

Figure 30: Dialog window for Interactive Filter Widget

As shown in Figure 30, it is possible to configure a label that is shown on the WebPortal. Also, the column that is dedicated to being filtered needs to be selected.

The next component offers the user the possibility to choose between further filtering for specific SLCs or specific rare diseases.

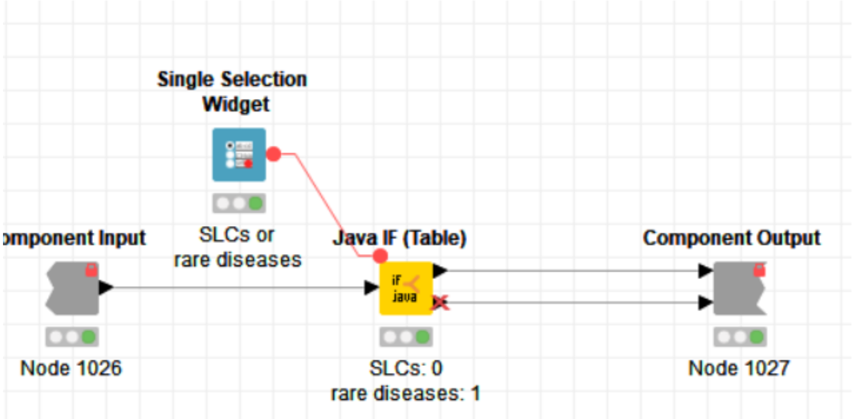
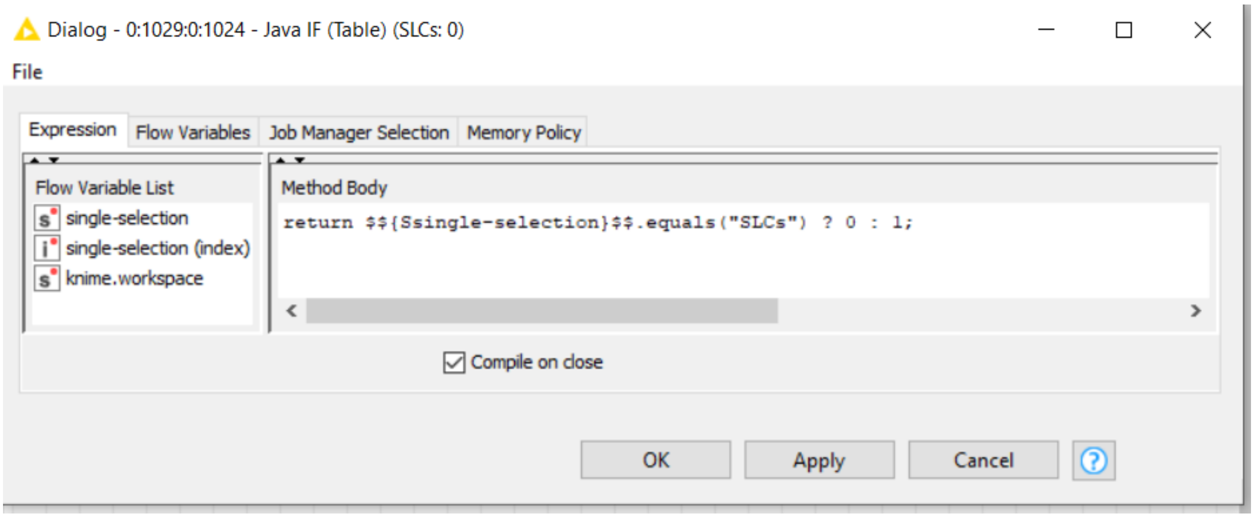


Figure 31: Inside of Component ‘SLCs or rare diseases’

As shown in Figure 31, this can be achieved with two associated nodes: The *Single Selection Widget* node allows the user to choose between the strings ‘SLCs’ or ‘rare diseases’. The selected value is handed to the *Java IF* node through a variable called ‘single-selection’.



```
return $${Ssingle-selection}$$.equals("SLCs") ? 0 : 1;
```

Figure 32: Dialog window for Java IF node (screenshot)

Figure 32 shows the configuration window for the *Java IF* node, including the Method Body with the expression ‘return $${Ssingle-selection}$$.equals("SLCs") ? 0 : 1;’. The *Java IF* node contains two output ports. When the user chooses ‘SLCs’ through the *Single Selection Widget* node, the data is handed to the first output port (port 0), making port 1 inactive. When the user chooses ‘rare diseases’, it leads to the reverse result. As this component leads to two branches with only one being active at a time, the data is then directed to either one of two nodes that are used for filtering either for SLCs or rare diseases dependent on the active port.

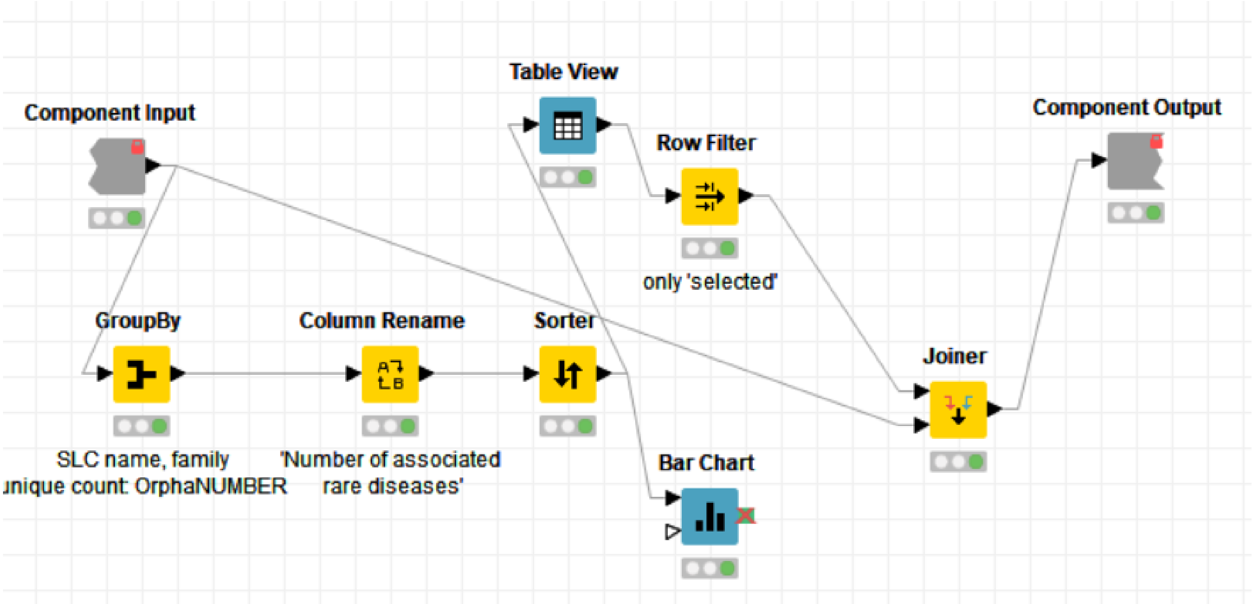


Figure 33: Inside of Component ‘Bar Chart SLCs’ (screenshot)

Figure 33 shows a screenshot of the inside of the component used for filtering for SLCs. It starts with a *GroupBy* node that groups the rows by unique SLCs (column: SLC name). At the same time, the number of associated rare diseases is aggregated through the aggregation method ‘Unique Count’. This is due to the bulk of data that would lead to a confusing complexity inside the *Bar Chart* node.

After renaming the aggregation column into ‘Number of associated rare diseases’ and sorting the rows by descending numbers, the data is handed to two *JavaScript* nodes: a *Bar Chart* node and a *Table View* node.

As both of these nodes are placed in the same component, they allow interactive filtering via the WebPortal. The user can choose the dedicated SLCs or a single SLC via the *Bar Chart* node where the associations are visually displayed in the form of a Bar Chart graph with SLC names being the category column on the x-Axis and the number of associated diseases being shown on the y-Axis.

The selected values from the *Bar Chart* node are directly handed to the *Table View* node that includes an added column called ‘Selected’. Once, SLCs are chosen via the *Bar Chart* node, the value of the corresponding row is set to ‘true’.

The following *Row Filter* node filters for only selected values and the *Joiner* node joins the selected SLCs back with the full table of SLC-rare disease associations.

When the user chooses to filter for rare diseases, the component is constructed the same way, but with the ‘disease name Orphanet’ being displayed on the x-Axis of the

Bar Chart and the number of associated SLCs on the y-Axis. An *End IF* node collects the data either from the top or bottom input depending on the active branch.

In the next step, the user can choose between manually curating the filtered data or continuing with downloading the file. This is, again, achieved with a *Java If* node in combination with a *Single Selection Widget* node as further explained at p.39.

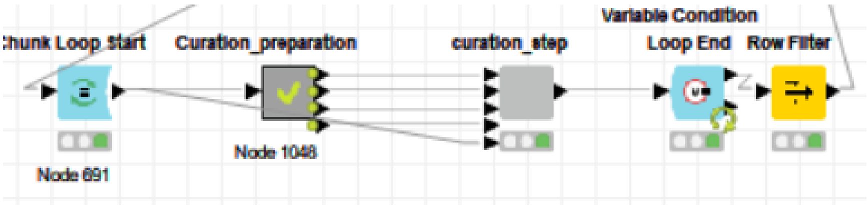


Figure 34: Combination of nodes/metanodes for manual curation

When the user decides to curate the results manually, a set of nodes follow that is based on a workflow created by Riccardo Martini.

The first node is a *Chunk Loop Start* node that splits the table into one row at a time. This node is followed by a metanode with the name ‘Curation_preparation’, shown in Figure 35.

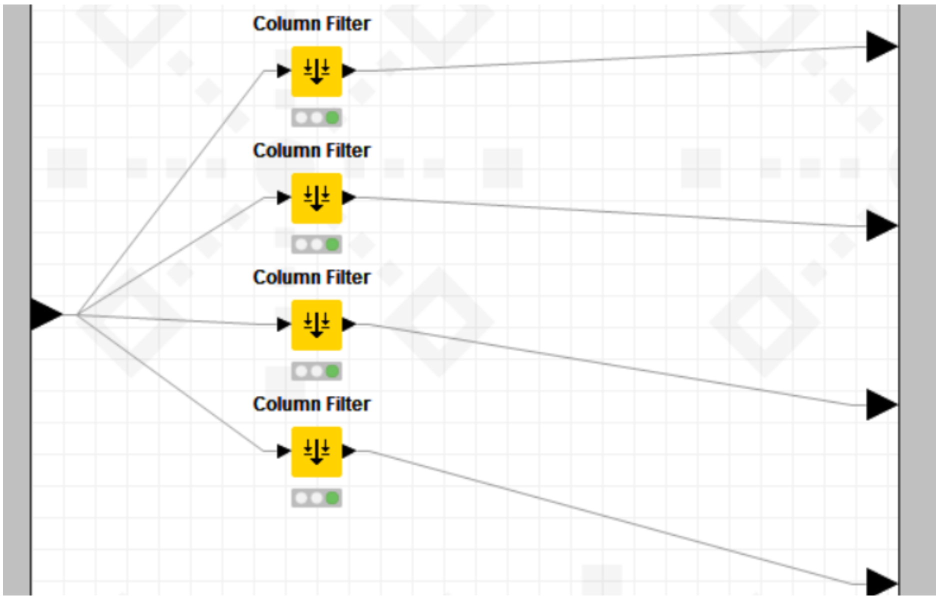
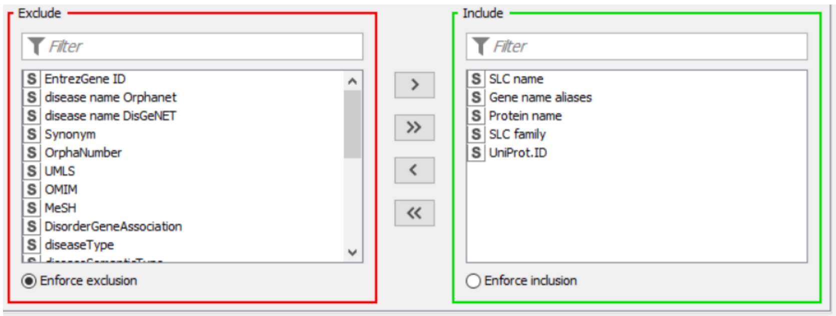


Figure 35: Inside of metanode 'curation_preparation'

This metanode contains four *Column Filter* nodes that each include four to five columns, as shown in Figure 36 and filter out the rest to build a clearly-arranged structure for the curation step.



Exclude

Filter

- S EntrezGene ID
- S disease name Orphanet
- S disease name DisGeNET
- S Synonym
- S OrphaNumber
- S UMLS
- S OMIM
- S MeSH
- S DisorderGeneAssociation
- S diseaseType

● Enforce exclusion

Include

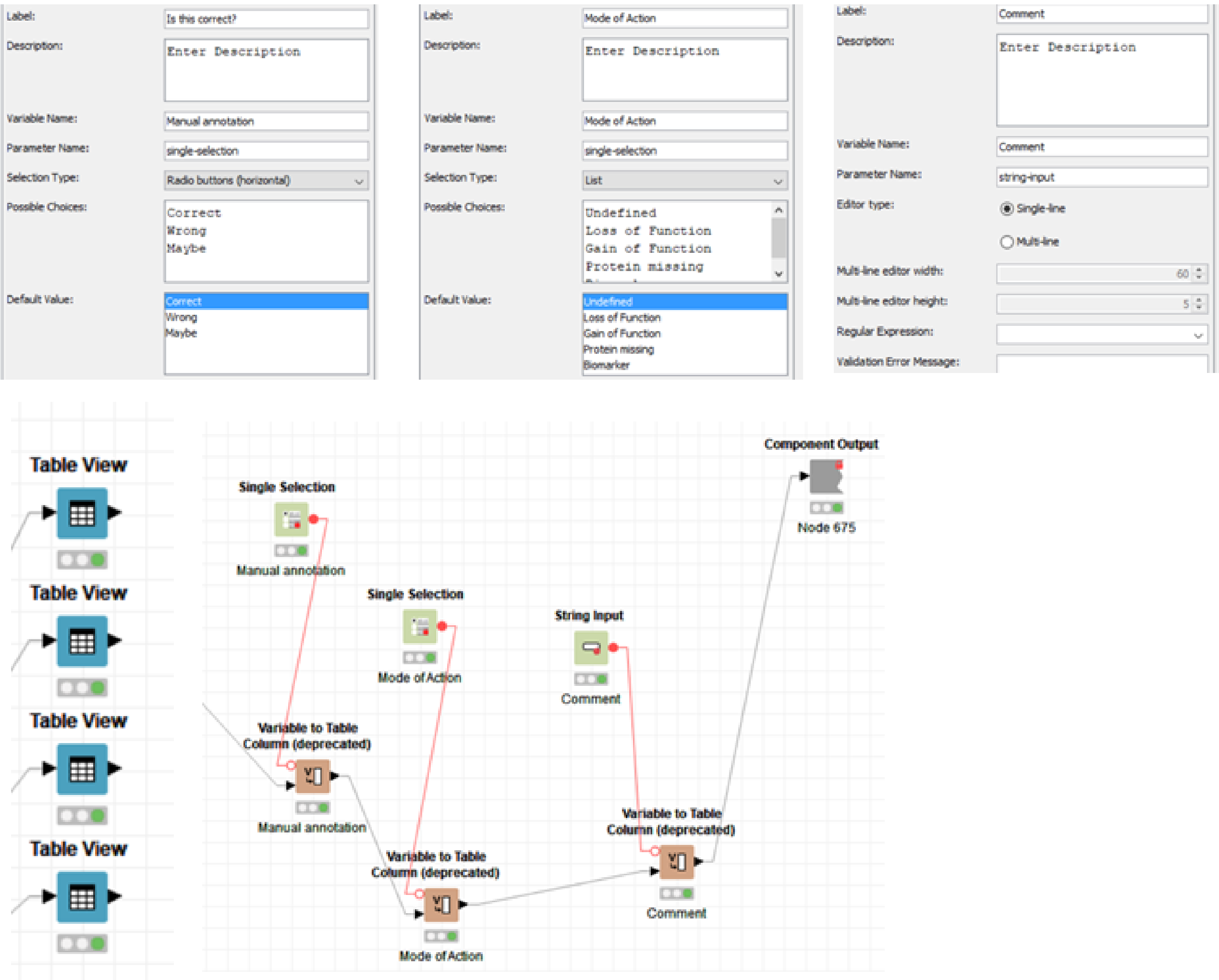
Filter

- S SLC name
- S Gene name aliases
- S Protein name
- S SLC family
- S UniProt.ID

○ Enforce inclusion

Figure 36: Dialog window for Column Filter 1

The data of these four *Column Filter* nodes is handed to the component ‘Curation_step’. The content of each of these nodes is displayed through four *Table View* nodes.



Label: Is this correct?
Description: Enter Description
Variable Name: Manual annotation
Parameter Name: single-selection
Selection Type: Radio buttons (horizontal)
Possible Choices: Correct
Wrong
Maybe
Default Value: Correct (selected), Wrong, Maybe

Label: Mode of Action
Description: Enter Description
Variable Name: Mode of Action
Parameter Name: single-selection
Selection Type: List
Possible Choices: Undefined
Loss of Function
Gain of Function
Protein missing
Default Value: Undefined (selected), Loss of Function, Gain of Function, Protein missing, Biomarker

Label: Comment
Description: Enter Description
Variable Name: Comment
Parameter Name: string-input
Editor type: ● Single-line ○ Multi-line
Multi-line editor width: 60
Multi-line editor height: 5
Regular Expression:
Validation Error Message:

Figure 37: Inside of component 'curation_step'

Also, the component contains two *Single Selection* nodes and one *String Input* node, as shown in Figure 37. The *Single Selection* nodes offer the user the possibility to decide, whether the association is correct or not and to add the Mode of Action. In contrast, the

String Input node lets the user add a comment. These three answers are added to the table as three new columns, and the results for all selected SLC-rare disease associations are collected within the following *Loop End* node. A *Row Filter* node excludes rows with the manual annotation 'Wrong' as well as rows with the Mode of action 'Protein missing' or 'Biomarker'.

Either after the manual curation or after skipping the manual curation, the user gets the chance to download the results from the filtered SLC-rare disease table.

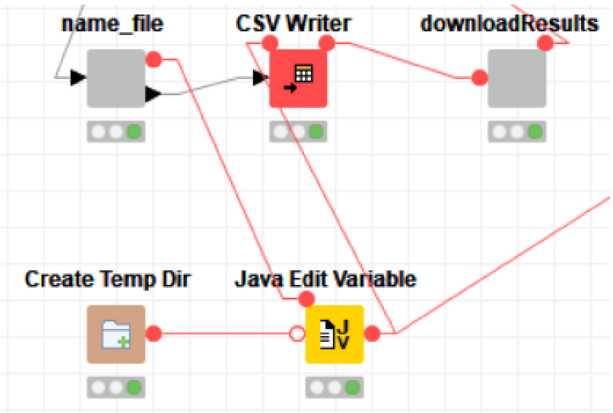


Figure 38: Combination of nodes/ components for downloading

Figure 38 shows the combination of nodes and components necessary for the download. The component *name_file* contains a *String Configuration* node that gives the user the possibility to insert the name of the file. The *Create Temp Dir* node creates a temporary directory on which the file can be saved before it is downloaded. The *Java Edit Variable* node creates the output location for the *CSV Writer* node out of the chosen file name and the name of the created temporary directory. The *CSV Writer* node saves the file at the temporary directory and can then be downloaded inside the *downloadResults* component that contains a *File Download Widget* node.

Apart from the *File Download Widget* node, the *downloadResults* component contains a *Text Output Widget* node with the Text 'By clicking "next" you can continue downloading the molecules associated with the SLCs.'. This is because after downloading the SLCs-rare diseases results, the user can directly proceed with filtering and downloading the molecules.

The file 'SLCs and molecules' is imported directly from the KNIME Server and is then joined with the results from the *Java IF* node to only pass molecules associated with the selected SLC-rare disease associations. The file's content and its originating databases are mentioned in section 2.2.9, p.23f.

The last component gives the user a choice to include molecules from all three sources, PubChem, ChEMBL and Drugbank, or to choose one or two of them only, which is again achieved through an *Interactive Value Filter Widget*. This file can also be downloaded.

3 RESULTS

This part of the thesis shows the results from the workflow, separated into two sections. While section 3.1 sums up and visualises the results from the workflow for data retrieval, 3.2 shows an application example of the second workflow as seen by users accessing it from the KNIME WebPortal.

3.1 Results of the workflow for data retrieval

The results presented and described in this chapter are derived from datasets that have last been updated at 05/19/2020. The counts at different positions of the workflow for data retrieval are presented in Table 13. Two tables containing more detailed information about concrete rare diseases and SLCs are presented in the Appendix (starting from page 65).

The start of the workflow is the adapted version of the table *RESOLUTE_SLCs,* last updated in June 2019. The file was provided by the RESOLUTE project and contains a list of 446 SLCs. The table is complemented with parameters received through the UniProt API.

The second step is the extraction of rare diseases. The information is retrieved from Orphanet and DisGeNET. The XML file 'Orphanet rare diseases and cross-references' acts as a starting point. It contains references for 9.614 diseases. As described in chapter 1.1, the number of rare diseases is usually estimated as up to 8.000. The high number of rare diseases on Orphanet is caused by the fact that it sometimes differentiates between manifestations of diseases that are elsewhere classified as a single disease.[46] 5.547 of these diseases are provided with an identifier in the Unified Medical Language System (UMLS) which is used to join the results with DisGeNET. The file '*ALL gene-disease-pmid associations'* from DisGeNET contains 3.241.576 disease-gene associations. However, the file is not specialised to rare diseases but includes all kinds of conditions. This is why it is joined with the results from 'Orphanet rare diseases and cross-references' which leads to 957.645 rows. When the results are joined with the list of SLCs from step one, the file introduces 4.295 rare disease-SLC associations with 1.021 unique rare diseases associated with 364 SLCs.

The file *'Orphanet rare diseases with their associated genes'* was accessed as a second source that additionally introduces 3.766 unique diseases. This file contains 192

SLC- rare disease associations with 178 unique rare diseases associated with 130 SLCs.

After concatenating the results from these two sources, the second part of the workflow results into a file containing 4.377 SLC- rare disease associations with 1.097 unique rare diseases with 367 SLCs with the highest number of retrieved associated diseases (143) for SLC2A1. As a considerable part of the DisGeNET data derives from text mining, the results would require manual curation as there might be false-positive results included as well. When results originating from sources based on text mining are excluded, the workflow results in 916 associations between 458 unique rare diseases with 223 SLCs, also with the highest number of retrieved associations for SLC2A1 (33).

Figure 39 shows a bar chart presenting the number of associated rare diseases per SLC when text mining is included.

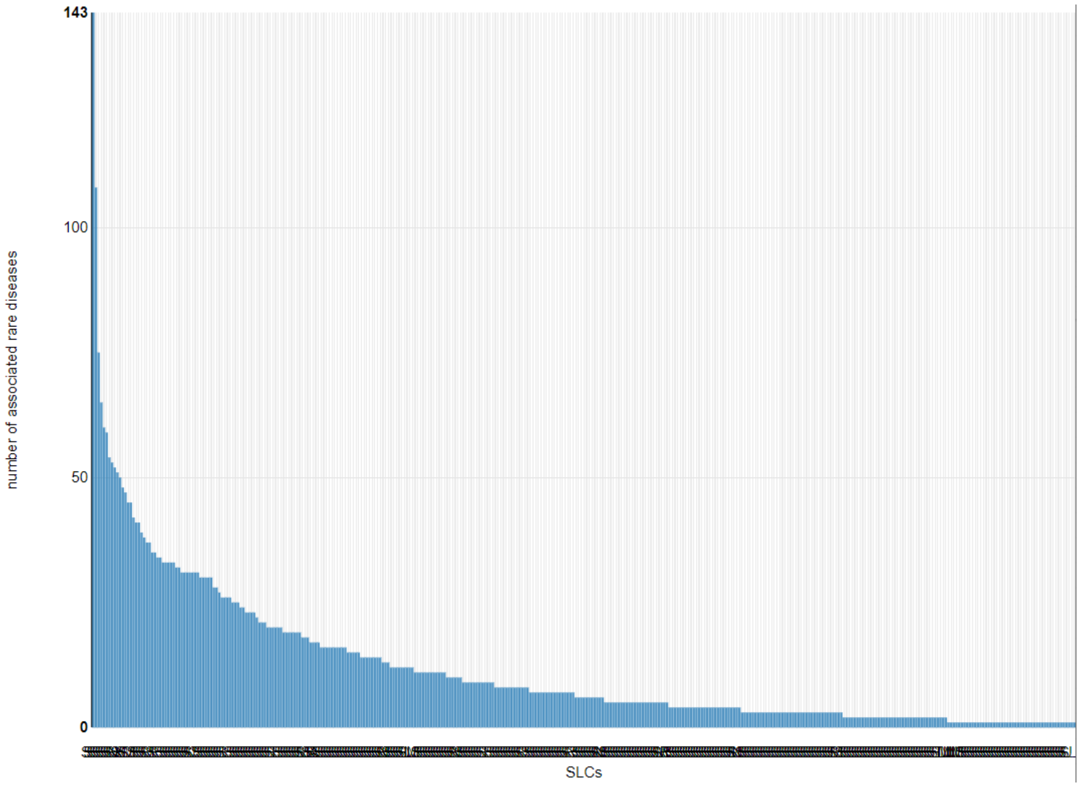


Figure 39: Bar Chart showing SLCs with number of associated rare diseases

The third part of the workflow is the extraction of associated molecules from three sources: PubChem, ChEMBL and Drugbank.

The XML file '*complete database*' from Drugbank was accessed as a source for the retrieval of SLC-molecule associations from Drugbank. It contains entries for 11.922 drugs and when joined with the results from step one and after filtering out substrates, 1.256 SLC-molecule associations with 583 unique molecules for 119 SLCs.

The extraction of molecules from ChEMBL was achieved through the use of several web APIs (see 2.2.6, p.18f.). The extraction of bioactivities leads to 21.822 SLC-molecule associations with 14.155 molecules associated with 85 SLCs.

The results from PubChem derive from a direct download of compounds associated with SLCs from the website. The download results, after the exclusion of results from ChEMBL, into the table contains 19.166 associations, including 18.375 molecules and 64 SLCs.

After concatenating the results from all three databases, 32.885 possible drugs and molecules could be retrieved for 147 out of the 367 SLCs associated with rare diseases, and accordingly when text mining is excluded, for 102 out of the 223 SLCs.

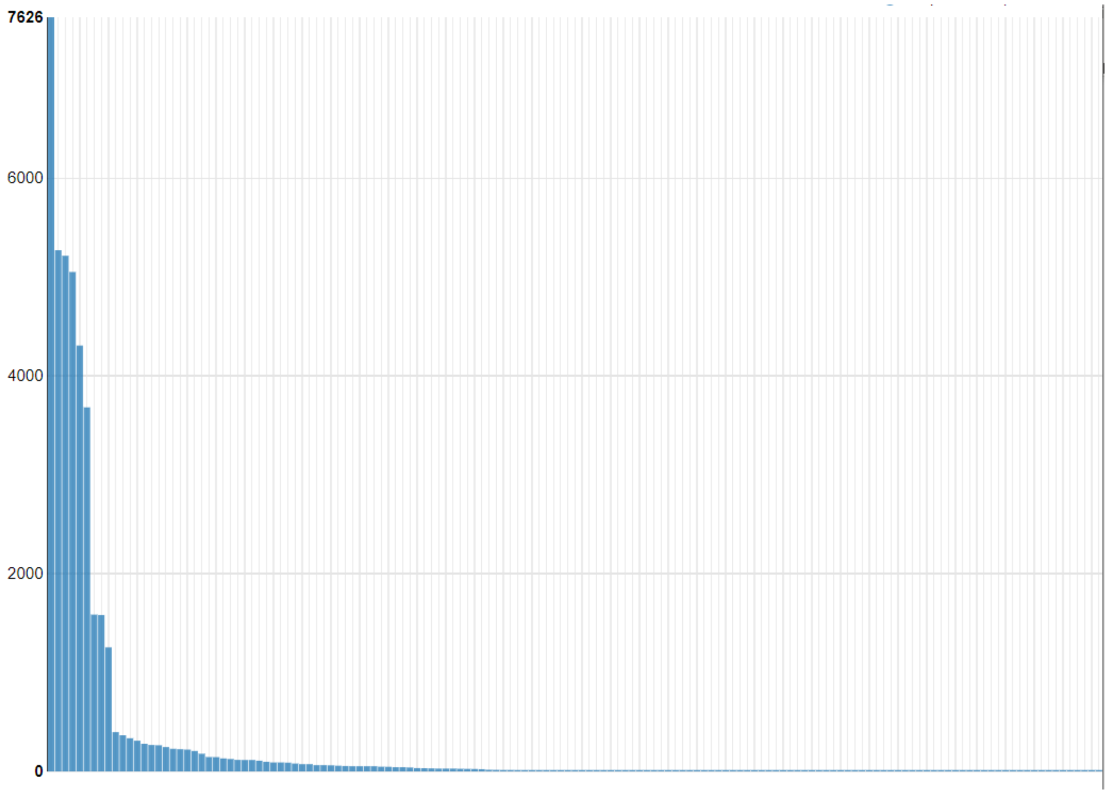


Figure 40: Bar Chart showing SLCs with number of associated molecules/drugs

As shown in Figure 40, few SLCs are associated with a high number of molecules, while the majority is associated with less than a hundred molecules. The highest number of molecules is associated with NPC1 (7626).

Figure 41 shows the number of molecules received from each database with the majority of molecules deriving from PubChem, another significant part deriving from

ChEMBL and, in relation, only a few molecules from DrugBank. This is due to results from high throughput screening assays, received from PubChem and ChEMBL, that lead to many compounds for a few solute carriers.

Although DrugBank includes the lowest number of unique molecules, it offers compounds for the broadest range of unique SLCs, as described at p. 47 and shown in Table 13.

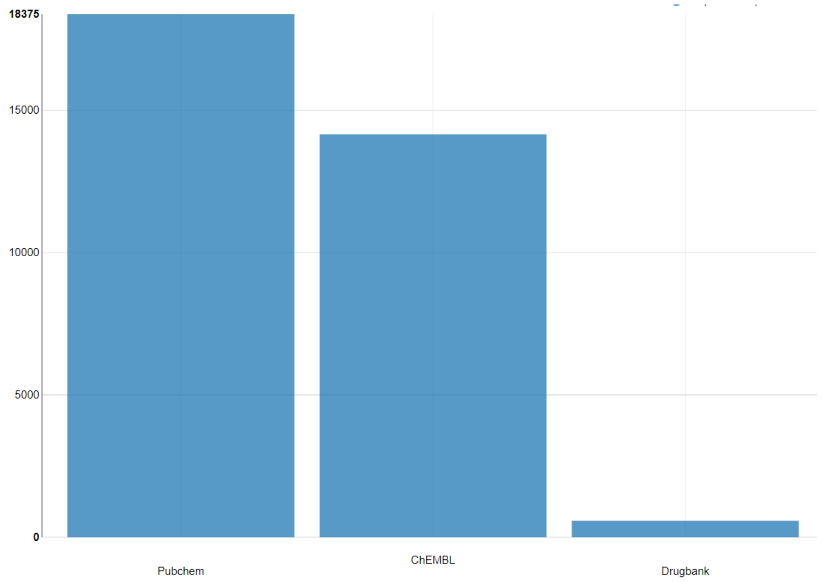


Figure 41: Number of molecules received from each database

Figure 42 shows the last stage of the triangulation, the availability of molecules for rare diseases via the intermediate step of SLCs with ‘Rare diabetes mellitus’ being the rare disease with the highest number of possible, available molecules (28.640). Altogether, the workflow proposes potential molecules for 746 rare diseases. However, the associations would need manual curation as this result includes all molecules somehow active against the SLC and does not consider the role of the SLC in the specific rare disease. Furthermore, the workflow contains false-positive results as further described in chapter 4, p.58f.

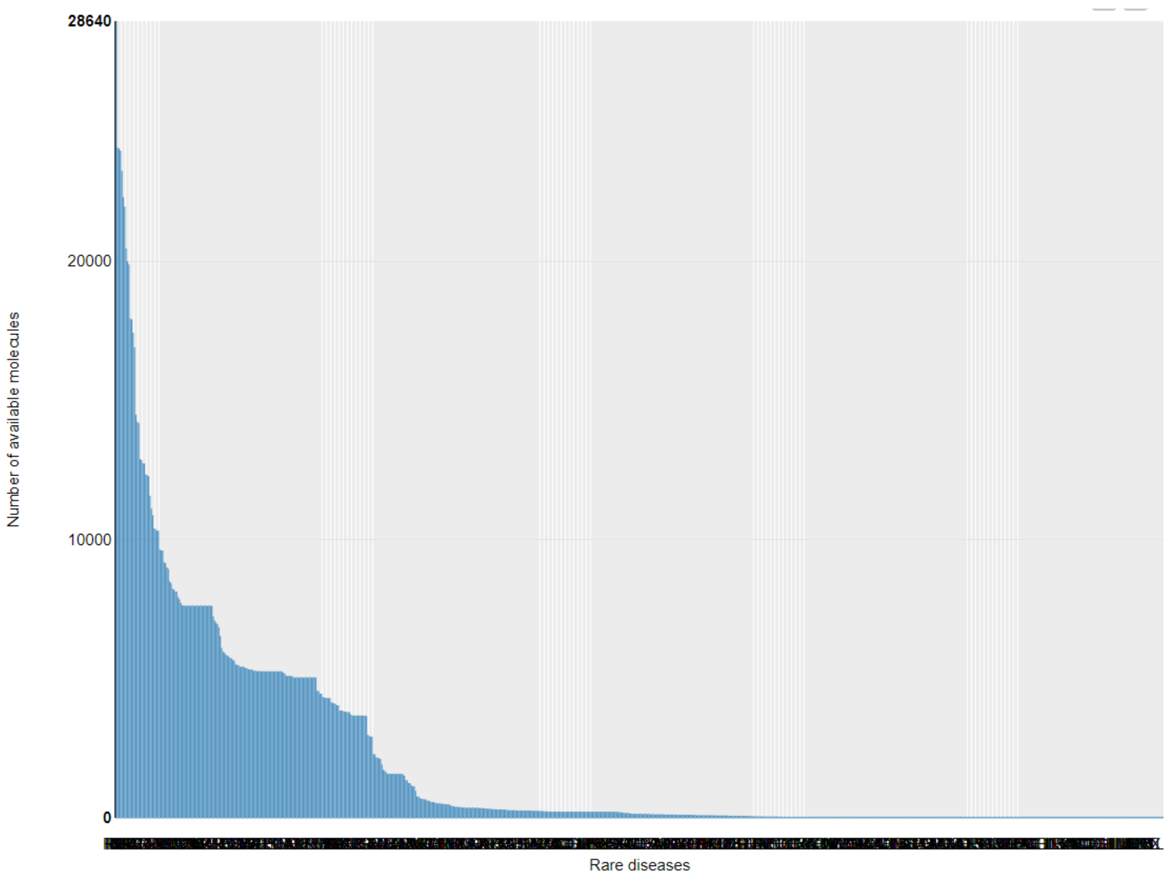


Figure 42: Bar Chart showing rare diseases with number of associated molecules

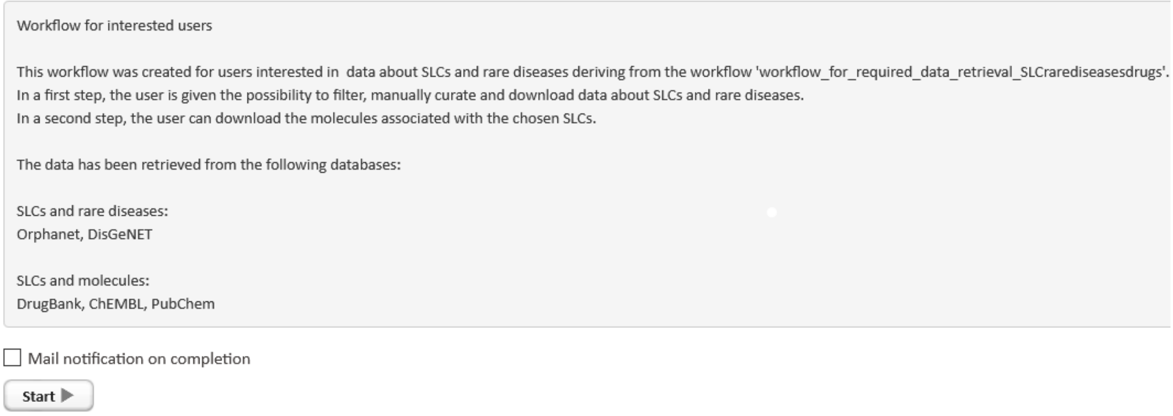
Table 13: Counts at different positions of the workflow

	DATASET	ORIGIN	COUNT	SECTION	DESCRIPTION
A	RESOLUTE_SLCs	RESOLUTE project	446	2.3.1	Dataset with SLCs + identifiers
B	UniProt Aliases	UniProt API, starting from A	446	2.3.1	API adds Aliases for proteins and genes
C	Rare diseases and cross-references	Orphadata	9.614	2.3.2	Dataset with rare diseases + identifiers from Orphanet
D	UMLS identifiers	C	5.547	2.3.2	Rare diseases from C with valid UMLS identifier
E	All gene-disease-pmid associations	DisGeNET	3.241.576	2.3.2	Dataset with all gene-disease-pmid associations from DisGeNET
F	SLC-rare disease associations from E	A+E	4.925	2.3.2	E filtered for SLCs as genes via UMLS ID and Entrez Gene ID
G	Rare diseases with their associated genes	Orphadata	3.766	2.3.2	Dataset with Rare disease-gene associations from Orphanet
H	SLC-rare disease associations from G	A+G	192	2.3.2	G filtered for SLCs as genes via UniProt ID
I	Unique SLC-rare disease associations	F+H	4.377	2.3.2	SLC- rare disease associations from both sources
J	Unique rare diseases	based on I	1.097	2.3.2	number of unique rare diseases associated with SLCs

	DATASET	ORIGIN	COUNT	SECTION	DESCRIPTION
K	Unique SLCs	based on I	367	2.3.2	number of unique SLCs associated with rare diseases
L	SLC-rare disease associations, text mining excluded	based on I	916	2.3.2	number of SLC-rare disease associations when text mining is excluded (DisGeNET)
M	full_database	Drugbank	11.922	2.3.3.1	Dataset with all drug entries from Drugbank
N	unique SLC-molecule associations Drugbank	K+M	1.817	2.3.3.1	SLC-molecule associations from Drugbank
O	Unique SLCs Drugbank	based on N	119	2.3.3.1	number of unique SLCs associated with drugs on Drugbank
P	SLC-molecule associations ChEMBL	ChEMBL API, based on K	21.822	2.3.3.2	SLC-molecule associations retrieved through API calls, PubChem excluded
Q	Unique SLCs ChEMBL	based on P	85	2.3.3.2	number of unique SLCs associated with molecules on ChEMBL
R	SLC-molecule associations PubChem	PubChem CIDs download, based on K	19.166	2.3.3.3	SLC-molecule associations retrieved through download from PubChem, ChEMBL excluded
S	Unique SLCs PubChem	based on R	64	2.3.3.3	number of unique SLCs associated with molecules on PubChem
T	Number of SLCs with retrieved molecules	O+P+R	147	2.3.3	Number of SLCs associated with rare diseases that are associated with molecules

3.2 Workflow for accessing data as seen by users

When the workflow is opened at the KNIME WebPortal, it shows a starting window containing a short description and the names of the originating databases, as shown in Figure 43.



Workflow for interested users

This workflow was created for users interested in data about SLCs and rare diseases deriving from the workflow 'workflow_for_required_data_retrieval_SLCrarediseasesdrugs'.
In a first step, the user is given the possibility to filter, manually curate and download data about SLCs and rare diseases.
In a second step, the user can download the molecules associated with the chosen SLCs.

The data has been retrieved from the following databases:

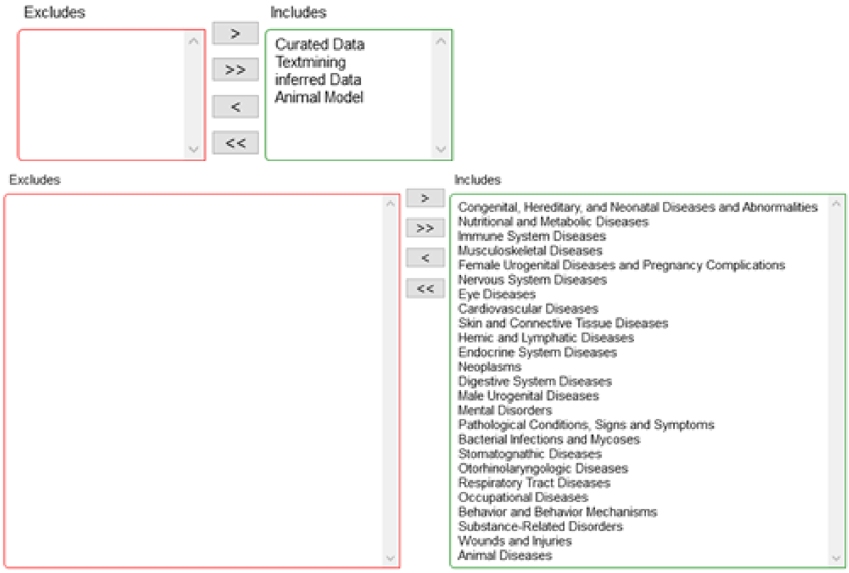
SLCs and rare diseases:
Orphanet, DisGeNET

SLCs and molecules:
DrugBank, ChEMBL, PubChem

□ Mail notification on completion

Figure 43: Starting window of 'Workflow for interested users' at the KNIME WebPortal

After clicking at ‘Start’, the user sees the first two filter possibilities. The first filter lets the user choose between curated data only or data from all sources, including text mining. Text mining offers great potential for retrieving new associations between rare diseases and SLCs. On the other hand, it also increases the risk of ‘false positive’ findings, which makes it necessary to curate the results manually afterwards. The page includes a link to the description of sources at the DisGeNET website.



Includes
Curated Data
Textmining
inferred Data
Animal Model

Includes
Congenital, Hereditary, and Neonatal Diseases and Abnormalities
Nutritional and Metabolic Diseases
Immune System Diseases
Musculoskeletal Diseases
Female Urogenital Diseases and Pregnancy Complications
Nervous System Diseases
Eye Diseases
Cardiovascular Diseases
Skin and Connective Tissue Diseases
Hemic and Lymphatic Diseases
Endocrine System Diseases
Neoplasms
Digestive System Diseases
Male Urogenital Diseases
Mental Disorders
Pathological Conditions, Signs and Symptoms
Bacterial Infections and Mycoses
Stomatognathic Diseases
Otorhinolaryngologic Diseases
Respiratory Tract Diseases
Occupational Diseases
Behavior and Behavior Mechanisms
Substance-Related Disorders
Wounds and Injuries
Animal Diseases

Figure 44: First filtering options at WebPortal

The second filtering option lets the user choose between the inclusion or exclusion of MeSH disease classes, which may be especially interesting when there is a focus on specific diseases.

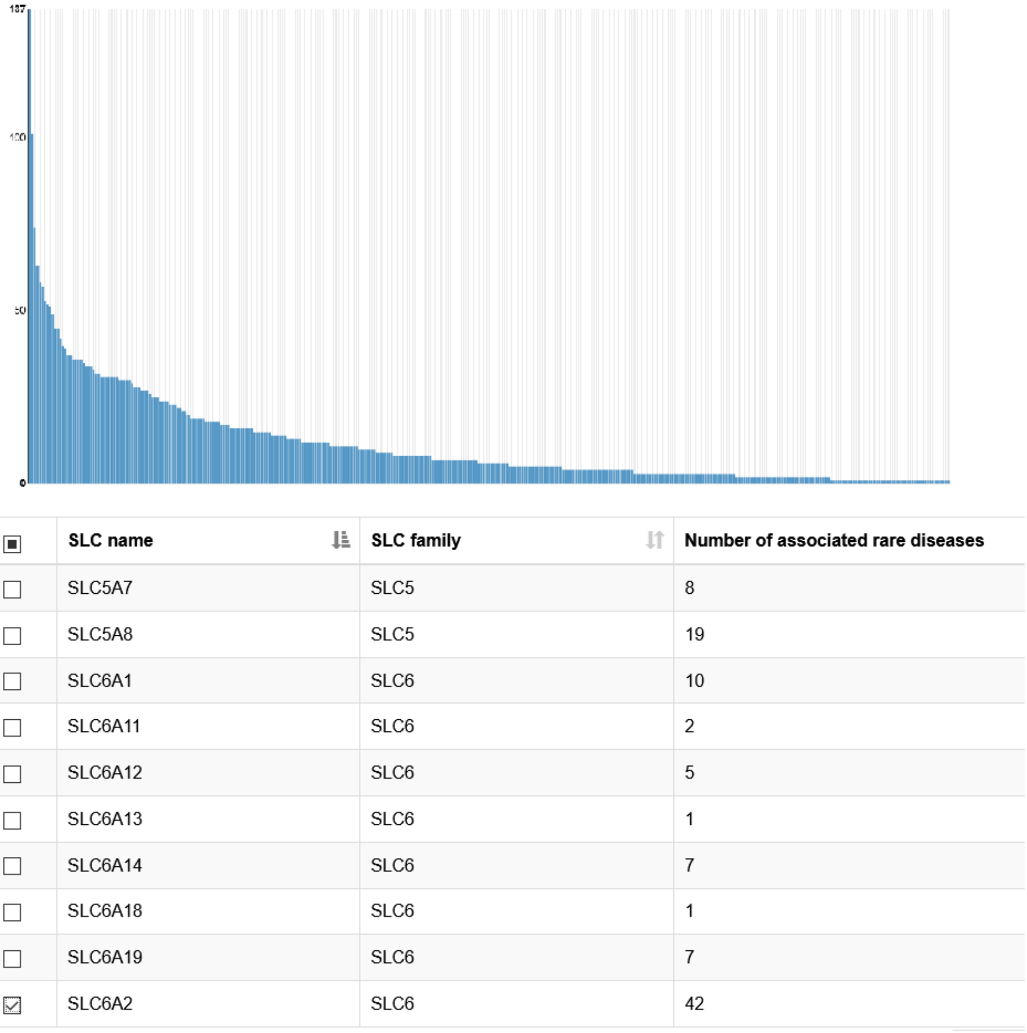
In the next step, as shown in Figure 45, the user can decide between further filtering for specific rare diseases or SLCs.

Do you want to further filter for specific diseases or SLCs?

SLCs

Figure 45: Filtering option between SLCs or rare diseases

After choosing ‘SLCs’, the user is presented with a BarChart graph and a corresponding table, as shown in Figure 46.



▣	SLC name	SLC family	Number of associated rare diseases
□	SLC5A7	SLC5	8
□	SLC5A8	SLC5	19
□	SLC6A1	SLC6	10
□	SLC6A11	SLC6	2
□	SLC6A12	SLC6	5
□	SLC6A13	SLC6	1
□	SLC6A14	SLC6	7
□	SLC6A18	SLC6	1
□	SLC6A19	SLC6	7
☑	SLC6A2	SLC6	42

Figure 46: Bar Chart and corresponding table

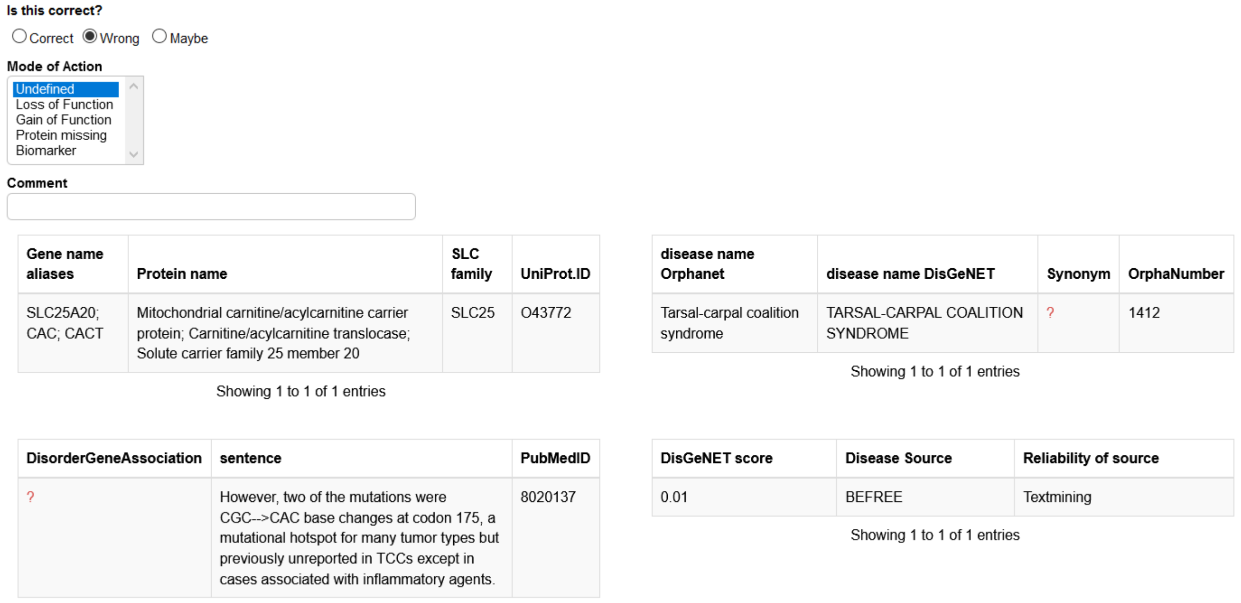
The user can now decide between choosing all SLC-rare disease associations by crossing the box at the top or selecting a specific SLC or SLCs. The table can be sorted by SLC name, SLC family and the number of associated rare diseases. Besides, the user can choose specific SLCs directly from the Bar Chart.

Do you want to curate your results manually or continue on downloading?

Curate

Figure 47: Curation or download

After selecting SLCs or rare diseases, it is possible to curate the results manually. However, this step can also be skipped, and the results can be downloaded directly.
Figure 48 shows an example screenshot of the curation step. The parameters describing SLCs, diseases and their associations are divided into four tables to arrange the page. The user can decide whether the association is correct or not and add the mode of action, which is especially essential for matching the results with possible drugs and molecules. Besides, it is possible to add a comment.
Figure 48 reveals the importance of the curation step, especially when text mining is included. DisGeNET's source BEFREE proposes an association between the rare disease 'Tarsal-carpal coalition syndrome' and SLC25A20. After taking a closer look at the sentence that supports this association, it is shown that the text mining failed in this case. The association was detected based on the abbreviation 'CAC' which can be an alias for the gene SLC25A20. However, in this case, 'CAC' marks a base sequence.



Is this correct?

○ Correct ● Wrong ○ Maybe

Mode of Action

Undefined
Loss of Function
Gain of Function
Protein missing
Biomarker

Comment

Gene name aliases	Protein name	SLC family	UniProt.ID
SLC25A20; CAC; CACT	Mitochondrial carnitine/acylcarnitine carrier protein; Carnitine/acylcarnitine translocase; Solute carrier family 25 member 20	SLC25	O43772

Showing 1 to 1 of 1 entries

disease name Orphanet	disease name DisGeNET	Synonym	OrphaNumber
Tarsal-carpal coalition syndrome	TARSAL-CARPAL COALITION SYNDROME	?	1412

Showing 1 to 1 of 1 entries

DisorderGeneAssociation	sentence	PubMedID
?	However, two of the mutations were CGC-->CAC base changes at codon 175, a mutational hotspot for many tumor types but previously unreported in TCCs except in cases associated with inflammatory agents.	8020137

DisGeNET score	Disease Source	Reliability of source
0.01	BEFREE	Textmining

Showing 1 to 1 of 1 entries

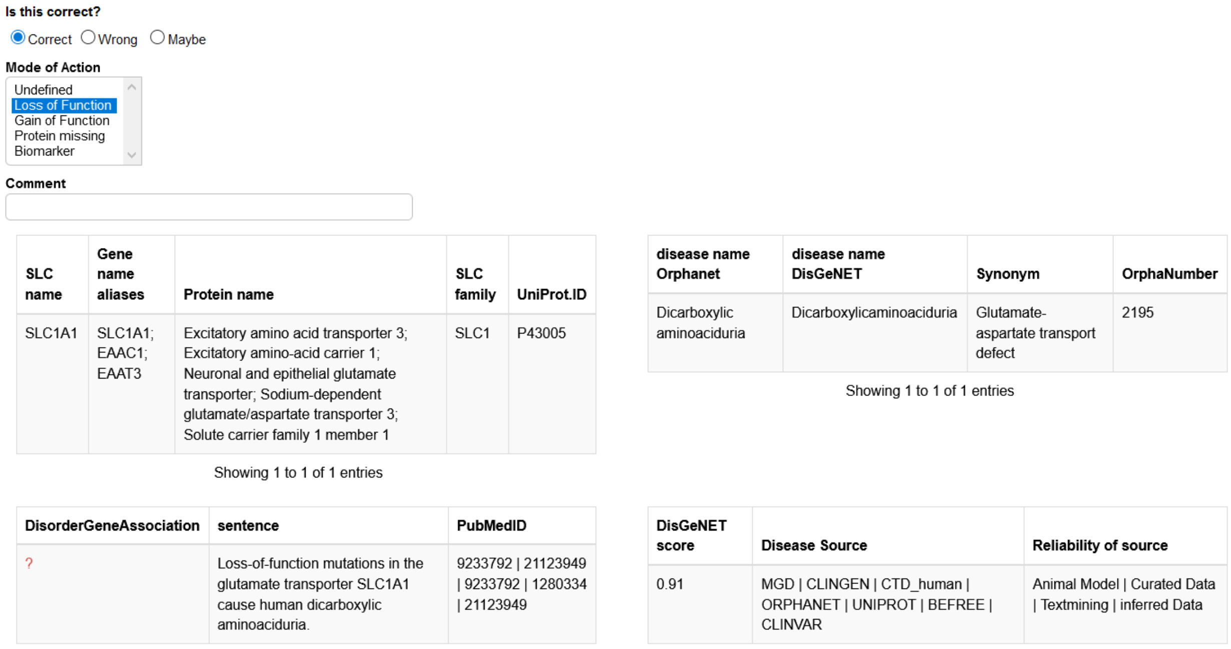
Figure 48: Curation example ‚Tarsal-carpal coalition syndrome'

Besides, BEFREE recognizes TCC as ‘Tarsal-carpal coalition syndrome’ as this abbreviation is also in use for this disease. However, when opening the entry on PubMed by using the available PubMedID, it is shown that the abbreviation TCC, in this case, stands for ‘transitional cell carcinomas’.

Thus, the user can mark the association as ‘wrong’ which leads to the exclusion of the association from the table in the next step.

Besides, especially DisGeNET includes associations with a protein being altered in disease as a Biomarker, but not as the cause. These associations are also excluded in the next step as they do not pose a potential drug target.

However, some associations are also easily detected as correct and matched with a Mode of Action, as shown in Figure 50. ‘Dicarboxylic aminoaciduria’ is associated with SLC1A1 and the sentence suggests the Mode of Action ‘loss of function’.



Is this correct?

● Correct ○ Wrong ○ Maybe

Mode of Action

Undefined
[✓ Loss of Function]
Gain of Function
Protein missing
Biomarker

Comment

SLC name	Gene name aliases	Protein name	SLC family	UniProt.ID

SLC1A1	SLC1A1;
	EAAC1;
	EAAT3
		Excitatory amino acid transporter 3; Excitatory amino-acid carrier 1; Neuronal and epithelial glutamate transporter; Sodium-dependent glutamate/aspartate transporter 3; Solute carrier family 1 member 1	SLC1	P43005

Showing 1 to 1 of 1 entries

disease name Orphanet	disease name DisGeNET	Synonym	OrphaNumber
Dicarboxylic aminoaciduria	Dicarboxylicaminoaciduria	Glutamate-aspartate transport defect	2195

Showing 1 to 1 of 1 entries

DisorderGeneAssociation	sentence	PubMedID
?	Loss-of-function mutations in the glutamate transporter SLC1A1 cause human dicarboxylic aminoaciduria.	9233792 | 21123949 | 9233792 | 1280334 | 21123949

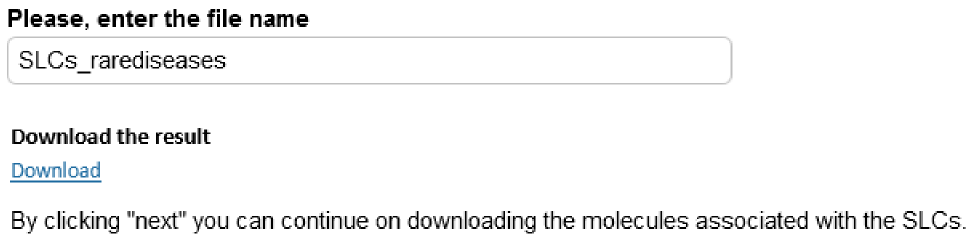
DisGeNET score	Disease Source	Reliability of source
0.91	MGD | CLINGEN | CTD_human | ORPHANET | UNIPROT | BEFREE | CLINVAR	Animal Model | Curated Data | Textmining | inferred Data

Figure 50: Curation example 'Dicarboxylic aminoaciduria'

Besides, associations from SLC1A1 to rare forms of epilepsy, Huntington disease and Amyotrophic lateral sclerosis could be verified. However, it often takes a long time to detect the corresponding Mode of Action.

After the curation step, the table is extended with three columns, one containing the curation ‘correct’ or ‘maybe’, the second one holding the mode of action and the third one presenting the ‘comment’.

The user now gets the chance to download the filtered table file by entering the file's name and then clicking on 'Download' as shown in Figure 51.



Please, enter the file name

SLCs_rarediseases

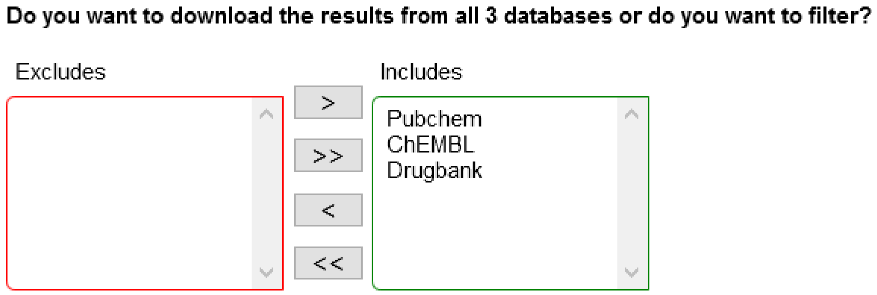
Download the result

Download

By clicking "next" you can continue on downloading the molecules associated with the SLCs.

Figure 51: Name file and download SLC_rarediseases

In the last step, the user gets the possibility to download the file with associated drugs and molecules after choosing between results from all three databases, PubChem, ChEMBL and Drugbank, or results from only one or two of these sources as shown in Figure 52.



Do you want to download the results from all 3 databases or do you want to filter?

Figure 52: Filter for molecule sources

4 DISCUSSION

The aim of the thesis was to give an overview of the role of SLCs in rare diseases via database integration and to show the availability of possible modulators. The workflow created for that purpose was capable of collecting the data from different databases. However, the retrieved data would need manual curation, which would, due to the significant amount of data, exceed the time constraints of a diploma thesis. This is why a second workflow that can be accessed at the KNIME WebPortal was created that gives interested users the option to access, filter, curate and download the aggregated data. The workflow itself, however, is not capable of showing the exact numbers of SLCs in rare diseases as it includes false-positive results.

4.1 Limitations of the workflow

As mentioned before, the data aggregated within the first workflow needs manual curation, especially before joining the SLC-rare disease associations with the SLC-drug/molecule associations. This is due to several reasons.

The first reason is that the associations retrieved from databases are not always valid, but include false-positive results. This is mostly caused by data from sources based on text mining. However, text mining, on the other hand, offers excellent potential for retrieving rather unexplored associations.

Another problem is that most SLC- rare disease associations, as well as most SLC-drug/molecule associations, do not include further information about the association. Because of this, the type of association (e.g. ‘loss of function’, ‘gain of function’ for rare diseases or ‘inhibitor’, ‘inducer’ for molecules/drugs) needs to be added manually in most cases.

While some wrong associations, as well as the types of association, are relatively easy to detect, as shown in section 3.2,p.53f., the curation of other associations takes a long time.

Besides, rare diseases are sometimes caused by the total deficiency of an SLC protein. In this case, a join with the retrieved SLC-molecule/drug associations would not bring a benefit as diseases caused by a missing protein are often treated with the specific protein itself instead of a drug that inhibits or activates the protein.

Because of this, the workflow itself is not capable of being used for drug repurposing or the proposal of active molecules as drugs straight away. However, after the manual curation of data, the results could offer potential.

4.2 Possibilities for adaptation

The workflows offer the potential for adaptations and further developments at several positions.

Possible adaptions for the workflow for data retrieval

The first workflow could be adapted especially at two positions of the workflow.

The first position would be the DisGeNET dataset. At the moment, the workflow uses the native datasets, downloaded each time the workflow runs, directly from the DisGeNET website. In 2019, DisGeNET introduced an API that gives programmatic access to its data. With this API, it would be possible to reduce the amount of data, as data could be filtered in advance, and only data associated with SLCs could be downloaded. However, as this part of the workflow was already almost finished when the API was introduced, I did not remodel it due to time constraints.

The second position would be the download of bioassays from PubChem. PubChem also offers an API, and in the beginning, it has been tried to use the API instead of the direct download. However, using the API for that purpose would, at the moment, result in a high number of consecutive REST API calls. This is why the direct download per target was preferred. Because of the sometimes high amount of data for single targets, the download is limited to associations marked as ‘active’ and does not include molecules marked as ‘unspecified’ although they could also pose potential.

Besides, the workflow could also easily be adapted for retrieving data about other proteins as targets. The first file that now includes a list of SLCs can easily be switched with a list of other proteins containing EntrezGene ID and UniProt ID.

Possible adaptions for the workflow for accessing data

The second workflow could be adapted by implementing more filtering options.

A possibility would be to filter for ‘probable’ gene-disease associations by using the scores offered by DisGeNET. The DisGeNET GDA score, for example, is based on the number and types of sources and associations with a higher rank are more likely to be valid. However, also associations with a lower rank can be correct and offer a higher potential in being rather unexplored and are therefore especially interesting.

Another possibility would be adding more filtering options for the gene-molecule file. An opportunity would include filtering for the ‘best’ molecule when a long list of molecules is offered for an SLC-disease association. The filtering option could be based on the originating database, for example with the decision of preferencing drugs and experimental drugs from Drugbank and only proposing molecules from ChEMBL and PubChem respectively, when no associations could be retrieved from Drugbank. Another possibility would include ranking the molecules from ChEMBL and PubChem based on the activity values, e.g. proposing only the molecule with the lowest IC50. However, especially results from PubChem often do not contain bioactivity values, but the association is only marked as active.

In conclusion, the collected data suggest that SLCs do play an essential role in rare diseases and could offer great potential as possible drug targets. However, the data needs manual curation to be used to repurpose drugs or find active molecules as potential drug candidates.

5 REFERENCES

[1] M. A. Hediger, B. Clémençon, R. E. Burrier, and E. A. Bruford, ‘The ABCs of membrane transporters in health and disease (SLC series): Introduction’, *Mol. Aspects Med.*, vol. 34, no. 2, pp. 95–107, Apr. 2013, doi: 10.1016/j.mam.2012.12.009.

[2] J. Gurinova, ‘Development of a KNIME workflow for the retrieval of associations between orphan diseases and their possible drug repurposing candidates’, diploma thesis, University of Vienna, Austria, 2018.

[3] ‘UniProt, entry for SLC1A1’. Accessed: Jun. 15, 2020. [Online]. Available: https://www.uniprot.org/uniprot/P43005.

[4] ‘Orphanet, entry for SLC1A1’. Accessed: Jun. 15, 2020. [Online]. Available: https://www.orpha.net/consor/cgi-bin/Disease_Genes.php?lng=EN&data_id=22150&Disease_Disease_Genes_diseaseGroup=SLC1A1&Disease_Disease_Genes_diseaseType=Gen&MISSING%20CONTENT=solute-carrier-family-1-member-1---SLC1A1&search=Disease_Genes_Simple&title=solute%20carrier%20family%201%20member%201%20-%20SLC1A1.

[5] ‘DisGeNET, entry for SLC1A1’. Accessed: Jun. 15, 2020. [Online]. Available: https://www.disgenet.org/browser/1/1/0/6505.

[6] ‘Drugbank, entry for SLC1A1’. Accessed: Jun. 15, 2020. [Online]. Available: https://www.drugbank.ca/bio_entities/BE0001054.

[7] ‘ChEMBL entry for SLC1A1’. Accessed: May 16, 2020. [Online]. Available: https://www.ebi.ac.uk/chembl/g/#browse/compounds/filter/_metadata.related_targets.all_chembl_ids%3ACHEMBL2721.

[8] ‘PubChem, entry for SLC1A1’. Accessed: Jun. 15, 2020. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/gene/6505#section=Chemicals-and-Bioactivities.

[9] G. Superti-Furga *et al.*, ‘The RESOLUTE consortium: unlocking SLC transporters for drug discovery’, *Nat. Rev. Drug Discov.*, Apr. 2020, doi: 10.1038/d41573-020-00056-6.

[10] The UniProt Consortium, ‘UniProt’. https://www.uniprot.org/ (accessed Jan. 17, 2020).

[11] Orphadata, ‘Orphadata: Free access data from Orphanet. © INSERM 1999. Data version 1.2.11/4.1.6 [2018-08-14] (orientdb version)’. http://www.orphadata.org (accessed Mar. 01, 2019).

[12] J. Piñero *et al.*, ‘DisGeNET - a database of gene-disease associations’. https://www.disgenet.org/ (accessed Jan. 05, 2020).

[13] D. S. Wishart *et al.*, ‘DrugBank: a knowledgebase for drugs, drug actions and drug targets’, *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D901–D906, Jan. 2008, doi: 10.1093/nar/gkm958.

[14] M. Davies *et al.*, ‘ChEMBL web services: streamlining access to drug discovery data and utilities’, *Nucleic Acids Res.*, vol. 43, no. W1, pp. W612–W620, Jul. 2015, doi: 10.1093/nar/gkv352.

[15] S. Kim *et al.*, ‘PubChem 2019 update: improved access to chemical data’, *Nucleic Acids Res.*, vol. 47, no. Database issue, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

[16] 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.

[17] 'Rare Disease Acts of 2002', *Rare Disease Legislative Advocates*, Nov. 07, 2002. https://rareadvocates.org/rare-disease-acts-of-2002/ (accessed Jun. 30, 2020).

[18] 'Orphanet: About rare diseases'. https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN (accessed Jul. 10, 2020).

[19] Inserm, 'Rare Diseases: Over 300 Million Patients Affected Worldwide', *Newsroom | Inserm*, Oct. 24, 2019. https://presse.inserm.fr/en/maladies-rares-plus-de-300-millions-de-patients-dans-le-monde/36980/ (accessed Jul. 10, 2020).

[20] US Food and Drug Administration., 'Orphan Drug Act of 1983'. Jan. 04, 1983.

[21] *Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products*, vol. 018. 2000.

[22] 'Orphanet: About orphan drugs'. https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanDrugs.php?lng=EN (accessed Jul. 11, 2020).

[23] W. Kaplan, V. Wirtz, A. Mante-Teeuwisse, and P. Stolk, 'Priority Medicines for Europe and the World 2013 Update'. 2013, Accessed: Jul. 12, 2020. [Online]. Available: http://www.who.int/medicines/areas/priority_medicines/ MasterDocJune28_FINAL_Web.pdf.

[24] B. Alberts *et al.*, 'Chapter 11: Membrane Transport of Small Molecules and the Electrical Properties of Membranes', in *Molecular Biology of the Cell*, Sixth Edition., New York: Garland Science, 2014, p. 597 ff.

[25] M. A. Hediger, M. F. Romero, J.-B. Peng, A. Rolfs, H. Takanaga, and E. A. Bruford, 'The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins', *Pflüg. Arch.*, vol. 447, no. 5, pp. 465–468, Feb. 2004, doi: 10.1007/s00424-003-1192-y.

[26] P. J. Höglund, K. J. V. Nordström, H. B. Schiöth, and R. Fredriksson, 'The Solute Carrier Families Have a Remarkably Long Evolutionary History with the Majority of the Human Families Present before Divergence of Bilaterian Species', *Mol. Biol. Evol.*, vol. 28, no. 4, pp. 1531–1541, Apr. 2011, doi: 10.1093/molbev/msq350.

[27] 'The SLCO (former SLC21) superfamily of transporters. - PubMed - NCBI'. https://www.ncbi.nlm.nih.gov/pubmed/23506880 (accessed Apr. 14, 2020).

[28] E. Perland and R. Fredriksson, 'Classification Systems of Secondary Active Transporters', *Trends Pharmacol. Sci.*, vol. 38, no. 3, pp. 305–315, Mar. 2017, doi: 10.1016/j.tips.2016.11.008.

[29] M.-L. Rives, J. A. Javitch, and A. D. Wickenden, 'Potentiating SLC transporter activity: Emerging drug discovery opportunities', *Biochem. Pharmacol.*, vol. 135, pp. 1–11, Jul. 2017, doi: 10.1016/j.bcp.2017.02.010.

[30] E. C. Chao, 'SGLT-2 Inhibitors: A New Mechanism for Glycemic Control', *Clin. Diabetes Publ. Am. Diabetes Assoc.*, vol. 32, no. 1, pp. 4–11, Jan. 2014, doi: 10.2337/diaclin.32.1.4.

[31] M. Tatsumi, K. Groshan, R. D. Blakely, and E. Richelson, 'Pharmacological profile of antidepressants and related compounds at human monoamine transporters', *Eur. J. Pharmacol.*, vol. 340, no. 2, pp. 249–258, Dec. 1997, doi: 10.1016/S0014-2999(97)01393-9.

[32] L. Lin, S. W. Yee, R. B. Kim, and K. M. Giacomini, 'SLC Transporters as Therapeutic Targets: Emerging Opportunities', *Nat. Rev. Drug Discov.*, vol. 14, no. 8, pp. 543–560, Aug. 2015, doi: 10.1038/nrd4626.

[33] M. J. Rosenberg *et al.*, 'Mutant deoxynucleotide carrier is associated with congenital microcephaly', *Nat. Genet.*, vol. 32, no. 1, pp. 175–179, Sep. 2002, doi: 10.1038/ng948.

[34] A. César-Razquin *et al.*, 'A Call for Systematic Research on Solute Carriers', *Cell*, vol. 162, no. 3, pp. 478–487, Jul. 2015, doi: 10.1016/j.cell.2015.07.022.

[35] 'KNIME Open Source Story | KNIME'. https://www.knime.com/knime-open-source-story (accessed Jun. 02, 2020).

[36] M. R. Berthold *et al.*, 'KNIME - the Konstanz information miner', *ACM SIGKDD Explor. Newsl.*, Nov. 2009, Accessed: Jan. 10, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/1656274.1656280.

[37] 'Metanodes for Reusability: A short story of metanodes, wrapped metanodes, and metanode templates. | KNIME'. https://www.knime.com/blog/wrapped-metanodes-and-metanode-templates-in-knime-analytics-platform (accessed Jun. 12, 2020).

[38] 'KNIME Analytics Platform 4.0: Components are for Sharing | KNIME'. https://www.knime.com/blog/knime-analytics-platform-40-components-are-for-sharing (accessed Jun. 12, 2020).

[39] 'KNIME Server | KNIME'. https://www.knime.com/knime-server (accessed Jun. 14, 2020).

[40] 'KNIME WebPortal | KNIME'. https://www.knime.com/knime-software/knime-webportal (accessed Jun. 12, 2020).

[41] 'XML Usage'. https://www.w3schools.com/xml/xml_usedfor.asp (accessed Jun. 13, 2020).

[42] 'XML Tree'. https://www.w3schools.com/xml/xml_tree.asp (accessed Jun. 13, 2020).

[43] 'PUG REST'. https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest$_Toc494865554 (accessed Jun. 13, 2020).

[44] 'Programmatic access'. https://www.uniprot.org/help/programmatic_access (accessed Jun. 13, 2020).

[45] The UniProt Consortium, 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[46] Orphanet, 'Orphanet: an online rare disease and orphan drug data base. © INSERM 1999'. http://www.orpha.net (accessed Jan. 03, 2020).

[47] Orphadata, 'Rare diseases with their associated genes; Data version: 1.2.11/4.1.6 [2018-08-14] (orientdb version)'. http://www.orphadata.org/data/xml/en_product6.xml (accessed Mar. 01, 2019).

[48] Orphadata, 'Rare diseases and cross-referencing; Data version: 1.2.11/4.1.6 [2018-08-14] (orientdb version)'. http://www.orphadata.org/data/xml/en_product1.xml (accessed Mar. 01, 2019).

[49] J. Piñero *et al.*, 'The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019)', *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkz1021 (accessed Jan. 05, 2020).

[50] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, 'Extraction of semantic biomedical relations from text using conditional random fields', *BMC Bioinformatics*, vol. 9, no. 1, p. 207, Apr. 2008, doi: 10.1186/1471-2105-9-207.

[51] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, 'Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research', *BMC Bioinformatics*, vol. 16, no. 1, p. 55, Feb. 2015, doi: 10.1186/s12859-015-0472-9.

[52] Integrative Biomedical Informatics Group GRIB/IMIM/UPF, 'ALL gene-disease-pmid association, data retrieved from DisGeNET v6.0', *ALL gene-disease-pmid associations*. https://www.disgenet.org/downloads (accessed Mar. 25, 2019).

[53] Integrative Biomedical Informatics Group GRIB/IMIM/UPF, ‘BeFree gene-disease-pmid associations for Pubannotation, data retrieved from DisGeNET v6.0’, *BeFree gene-disease-pmid associations*. https://www.disgenet.org/downloads (accessed Mar. 25, 2019).

[54] D. S. Wishart *et al.*, ‘DrugBank: a comprehensive resource for in silico drug discovery and exploration’, *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D668–D672, Jan. 2006, doi: 10.1093/nar/gkj067.

[55] 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, 04 2018, doi: 10.1093/nar/gkx1037.

[56] Wishart Research Group, ‘DrugBank’. https://www.drugbank.ca/ (accessed Dec. 20, 2018).

[57] Wishart Research Group, ‘Drugbank- full database XML; Version 5.1.2, 12-20-2018’. https://www.drugbank.ca/releases/5-1-2/downloads/all-full-database (accessed Feb. 07, 2019).

[58] A. Gaulton *et al.*, ‘ChEMBL: a large-scale bioactivity database for drug discovery’, *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1100–D1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[59] A. P. Bento *et al.*, ‘The ChEMBL bioactivity database: an update’, *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D1083–D1090, Jan. 2014, doi: 10.1093/nar/gkt1031.

[60] A. Gaulton *et al.*, ‘The ChEMBL database in 2017’, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[61] D. Mendez *et al.*, ‘ChEMBL: towards direct deposition of bioassay data’, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D930–D940, Jan. 2019, doi: 10.1093/nar/gky1075.

[62] NCBI, ‘PubChem Statistics’. https://pubchemdocs.ncbi.nlm.nih.gov/statistics (accessed Feb. 02, 2020).

[63] Daniela Digles, ‘ChEMBL Metanode, unpublished work’. .

[64] Anna Seiler, ‘diploma thesis, in preparation’.

6 APPENDIX

6.1 Data tables

6.1.1 Rare diseases with number of associated molecules and SLCs

disease name Orphanet	OrphaNumber	Number of associated molecules	Number of associated SLCs
Rare diabetes mellitus	101952	28640	129
Rare neurodegenerative disease	182070	24079	41
Rare inborn errors of metabolism	68367	24062	44
Huntington disease	399	23982	32
Rare diabetes mellitus type 2	181376	23246	96
Neuroblastoma	635	22311	46
Rare parkinsonian disorder	68402	21960	15
Rare epilepsy	101998	20459	60
Hepatocellular carcinoma	88673	20005	106
Rare malignant breast tumor	180257	19883	143
Metachromatic leukodystrophy	512	17946	3
Rare tumor of liver and intrahepatic biliary tract	306636	17921	29
Rare pulmonary disease	101944	17435	16
Rare movement disorder	102003	16909	15
Rare pervasive developmental disorder	168778	14491	12
Narcolepsy type 1	2073	14238	8
Bronchopulmonary dysplasia	70589	14191	8
Rare bacterial infectious disease	163582	12903	13
Amyloidosis	69	12854	14
Classic progressive supranuclear palsy syndrome	240071	12739	4
Progressive supranuclear palsy	683	12739	4
Rare inflammatory bowel disease	104012	12345	45
Cystic fibrosis	586	12317	27
Tuberous sclerosis complex	805	12288	7
Gastrointestinal stromal tumor	44890	11574	11
Rare sleep disorder	68354	11116	13
Rare carcinoma of pancreas	217074	10883	46
Angelman syndrome	72	10408	7

Frontotemporal dementia	282	10377	6
Alagille syndrome	52	10320	2
Testicular regression syndrome	983	10320	4
Rare dystonia	68363	9644	10
Down syndrome	870	9615	10
Juvenile myoclonic epilepsy	307	9605	4
Rare carcinoma of stomach	423771	9186	44
Rare epithelial tumor of stomach	63443	9163	42
Rare anemia	108997	8995	23
Neurofibromatosis type 1	636	8951	2
Rare viral disease	163585	8506	19
Hodgkin lymphoma	98293	8435	11
Lymphoma	223735	8235	23
Multiple myeloma	29073	8215	30
Renal cell carcinoma	217071	8142	39
Nasopharyngeal carcinoma	150	8139	15
Clear cell renal carcinoma	319276	7935	48
Precursor B-cell acute lymphoblastic leukemia	99860	7848	10
Burkitt lymphoma	543	7722	6
Tauopathy	98527	7641	4
Niemann-Pick disease type C	646	7628	3
Gaucher disease	355	7627	3
Alpha-1-antitrypsin deficiency	60	7626	2
Brucellosis	1304	7626	3
Congenital muscular dystrophy	97242	7626	1
Duchenne muscular dystrophy	98896	7626	1
Ebola hemorrhagic fever	319218	7626	2
Gangliosidosis	309144	7626	1
Lissencephaly	48471	7626	2
Muscular dystrophy	98473	7626	1
Niemann-Pick disease type A	77292	7626	1
Niemann-Pick disease type C, adult neurologic onset	216986	7626	1
Niemann-Pick disease type C, juvenile neurologic onset	216981	7626	1
Niemann-Pick disease type C, late infantile neurologic onset	216978	7626	1

Niemann-Pick disease type C, severe early infantile neurologic onset	216975	7626	1
Niemann-Pick disease type C, severe perinatal form	216972	7626	1
Niemann-Pick disease type D	79289	7626	1
Sea-blue histiocytosis	158029	7626	1
Sphingolipidosis	79225	7626	1
Tangier disease	31150	7626	1
Viral hemorrhagic fever	341	7626	2
Rare disorder with ptosis	98578	7240	19
Rare vascular disease	68362	7100	7
Arthrogryposis syndrome	109007	7019	6
Hypertrophic cardiomyopathy	217569	6959	12
Prader-Willi syndrome	739	6852	6
Glial tumor	182067	6538	52
Extrapelvic endometriosis	137820	6109	17
Tuberculosis	3389	5972	24
Rare digestive tumor	98059	5939	7
Rare intestinal disease	117569	5872	6
Diffuse large B-cell lymphoma	544	5838	10
Differentiated thyroid carcinoma	146	5802	14
Idiopathic pulmonary arterial hypertension	275766	5737	8
Pulmonary arterial hypertension	182090	5736	7
B-cell chronic lymphocytic leukemia	67038	5686	16
Cushing syndrome	553	5649	4
Cowden syndrome	201	5521	9
Systemic sclerosis	90291	5502	11
Fragile X syndrome	908	5477	6
Amyotrophic lateral sclerosis	803	5434	25
Congenital myasthenic syndrome	590	5434	3
Presynaptic congenital myasthenic syndromes	98914	5434	3
Spinocerebellar ataxia type 3	98757	5428	5
Osteochondritis dissecans	2764	5386	2
Ovarian cancer	213500	5382	39
Behçet disease	117	5347	7

Rare choreic movement disorder	306715	5342	9
Parkinsonian-pyramidal syndrome	171695	5331	3
Rare congenital non-syndromic heart malformation	88991	5325	7
Synovial sarcoma	3273	5285	5
West syndrome	3451	5285	5
Kaposi sarcoma	33276	5276	6
Rare hemolytic anemia	98363	5275	6
Early-onset generalized limb-onset dystonia	256	5274	2
Spinocerebellar ataxia type 6	98758	5274	2
Autosomal monosomy	102020	5271	3
Best vitelliform macular dystrophy	1243	5270	1
Burning mouth syndrome	353253	5270	1
Chronic thromboembolic pulmonary hypertension	70591	5270	1
Familial dysautonomia	1764	5270	1
Harlequin ichthyosis	457	5270	1
High altitude pulmonary edema	330012	5270	1
Interstitial cystitis	37202	5270	2
Postpartum psychosis	443173	5270	1
Progressive pseudorheumatoid arthropathy of childhood	1159	5270	1
Prune belly syndrome	2970	5270	1
Pulmonary venoocclusive disease	31837	5270	1
Superficial epidermolytic ichthyosis	455	5270	1
Sweet syndrome	3243	5270	2
Trigeminal neuralgia	221091	5270	1
Distal hereditary motor neuropathy type 7	139589	5214	1
Cerebrotendinous xanthomatosis	909	5171	5
Dopa-responsive dystonia	255	5107	2
Ear-patella-short stature syndrome	2554	5107	2
Elastosis perforans serpiginosa	79148	5107	2
Spinocerebellar ataxia type 2	98756	5107	3
Congenital diaphragmatic hernia	2140	5092	5

Mucolipidosis type II	576	5052	2
Autoimmune hemolytic anemia, warm type	90033	5050	1
Autoimmune lymphoproliferative syndrome	3261	5050	1
Autosomal recessive dopa-responsive dystonia	101150	5050	1
Desquamative interstitial pneumonia	98852	5050	1
Dysequilibrium syndrome	1766	5050	1
Ehlers-Danlos syndrome	98249	5050	2
Gaucher disease type 1	77259	5050	1
Idiopathic camptocormia	1320	5050	1
Infantile dystonia-parkinsonism	238455	5050	1
Kufor-Rakeb syndrome	306674	5050	1
Monosomy 5p	281	5050	1
Neurodegeneration with brain iron accumulation	385	5050	2
Pantothenate kinase-associated neurodegeneration	157850	5050	1
Partial deletion of the short arm of chromosome 5	261893	5050	1
Shwachman-Diamond syndrome	811	5050	1
Thanatophoric dysplasia type 1	1860	5050	1
Rett syndrome	778	4562	5
Dravet syndrome	33069	4558	3
Insulinoma	97279	4471	9
Rare tumor of pancreas	180824	4459	12
Fleck corneal dystrophy	98970	4348	4
Isolated focal cortical dysplasia	65683	4323	3
Congenital hypothalamic hamartoma syndrome	2113	4315	3
Pallister-Hall syndrome	672	4315	3
Malignant migrating partial seizures of infancy	293181	4307	2
Hereditary sensory and autonomic neuropathy type 2	970	4306	2
Dilated cardiomyopathy	217604	4148	9
Neuroendocrine neoplasm	877	4137	10
Rare cardiomyopathy	167848	4128	18
Small cell lung cancer	70573	4089	16
Multiple endocrine neoplasia type 1	652	4051	4
Neuroendocrine tumor of pancreas	97253	4049	4

Dentin dysplasia	1653	3859	2
Dentinogenesis imperfecta	49042	3859	2
Rare disease with dentinogenesis imperfecta	167762	3859	2
Multiple endocrine neoplasia type 2	653	3825	4
Multiple endocrine neoplasia type 2A	247698	3825	4
Retinoblastoma	790	3802	9
Von Hippel-Lindau disease	892	3800	6
Ganglioglioma	251949	3797	2
Vasculitis	52759	3708	4
Central nervous system primitive neuroectodermal tumor	251870	3682	3
Autoinflammatory syndrome	93665	3681	1
Choreoacanthocytosis	2388	3681	1
Hereditary nonpolyposis colon cancer	443909	3681	2
Lynch syndrome	144	3681	2
Neuroacanthocytosis	263440	3681	1
Osteogenesis imperfecta	666	3681	1
Postural orthostatic tachycardia syndrome due to NET deficiency	443236	3681	1
Pseudohypoaldosteronism type 1	756	3681	1
Rare autonomic nervous system disorder	423662	3681	1
Rare diabetes mellitus type 1	181371	3669	45
Glioblastoma	360	2982	59
Glycogen storage disease due to GLUT2 deficiency	2088	2940	4
Familial renal glucosuria	69076	2925	4
Glucose-galactose malabsorption	35710	2925	3
Rare disorder with lens opacification	98640	2294	18
Rare hyperlipidemia	181422	2287	14
Rare dyslipidemia	101953	2177	13
Thyroid tumor	100087	2163	19
Astrocytoma	94	2156	13
Thyroid carcinoma	100088	2109	17
Acute myeloid leukemia	519	1917	47
Syndromic diarrhea	84064	1738	9
Glomerular disease	93548	1685	3

Rare pancreatic disease	101937	1645	5
Acromegaly	963	1586	2
Atypical glycine encephalopathy	289863	1585	1
Congenital genu recurvatum	295229	1585	1
Glycine encephalopathy	407	1585	2
Infantile glycine encephalopathy	289860	1585	1
Isolated trigonocephaly	3366	1585	1
Neonatal glycine encephalopathy	289857	1585	2
Bullous pemphigoid	703	1581	1
Recurrent acute pancreatitis	64740	1581	1
Solar urticaria	97230	1581	1
Undifferentiated connective tissue syndrome	90002	1581	1
Atresia of small intestine	1201	1569	5
Cholera	173	1521	4
Bone sarcoma	223727	1350	28
Osteosarcoma	668	1350	28
Autosomal dominant optic atrophy	98672	1256	2
Listeriosis	533	1254	1
Congenital contractural arachnodactyly	115	1176	9
Rare breast tumor	180250	1134	47
Matthew-Wood syndrome	2470	1131	18
Malignant epithelial tumor of ovary	398934	975	20
Angiosarcoma	263413	774	6
Primary biliary cholangitis	186	769	15
Squamous cell carcinoma of the esophagus	99977	729	30
Ataxia-telangiectasia	100	689	10
X-linked hypophosphatemia	89936	684	5
Medulloblastoma	616	675	9
Nephroblastoma	654	672	7
Hypoxanthine-guanine phosphoribosyltransferase deficiency	206428	625	4
Lesch-Nyhan syndrome	510	625	4
Rare cancer of cervix uteri	213761	610	26
Rare hypothyroidism	181396	577	21
Oligodendroglial tumor	46484	573	4
Oligodendroglioma	251627	573	4
Rare hyperthyroidism	181399	544	12

Chronic myeloid leukemia	521	525	13
Cholangiocarcinoma	70567	524	15
Isolated congenital microcephaly	199642	523	28
Congenital sodium diarrhea	103908	512	5
Myelodysplastic syndrome	52688	511	11
Aplasia cutis congenita	1114	504	3
Hereditary spastic paraplegia	685	494	4
Immunoglobulin A vasculitis	761	494	5
Carcinoma of esophagus	70482	486	27
Leiomyosarcoma	64720	480	3
Precursor T-cell acute lymphoblastic leukemia	99861	447	7
Neonatal hypoxic and ischemic brain injury	137577	431	5
Progressive familial intrahepatic cholestasis	172	426	3
Hereditary renal hypouricemia	94088	401	2
Hereditary breast and ovarian cancer syndrome	145	397	1
Nephronophthisis	655	397	3
Myelomeningocele	93969	396	5
Pleural mesothelioma	50251	388	16
Cushing disease	96253	384	4
Rolandic epilepsy	1945	369	3
Myoclonic-astastic epilepsy	1942	368	2
Arrhythmogenic right ventricular cardiomyopathy	247	365	1
Carney complex	1359	365	1
Familial benign chronic pemphigus	2841	365	1
Kawasaki disease	2331	365	5
Loeys-Dietz syndrome	60030	365	3
Malignant tumor of penis	398043	365	1
Multiple endocrine neoplasia	276161	365	1
Multiple polyglandular tumor	100094	365	1
Sézary syndrome	3162	365	1
Familial multiple trichoepithelioma	867	357	5
Alternating hemiplegia of childhood	2131	355	2
Cytomegalic congenital adrenal hypoplasia	95702	355	2
Malaria	673	355	8

Rare hereditary hemochromatosis	220489	349	11
Collecting duct carcinoma	247203	340	4
Juvenile Huntington disease	248111	335	1
Follicular lymphoma	545	333	8
X-linked adrenoleukodystrophy	43	325	8
Asbestos intoxication	2302	321	2
Adenocarcinoma of the esophagus	99976	318	9
MODY	552	316	5
Encephalopathy due to GLUT1 deficiency	71277	314	2
Necrotizing enterocolitis	391673	311	2
Fetal and neonatal alloimmune thrombocytopenia	853	310	1
Primary sclerosing cholangitis	171	310	4
T-cell non-Hodgkin lymphoma	171918	305	4
Papillary renal cell carcinoma	319298	304	4
Channelopathy	140503	303	2
Retinopathy of prematurity	90050	303	3
Congenital bilateral absence of vas deferens	48	281	2
Soft tissue sarcoma	3394	281	6
Aceruloplasminemia	48818	279	1
Hemochromatosis type 4	139491	279	1
Porphyria cutanea tarda	101330	279	2
Rare form of salmonellosis	795	279	1
Gorlin syndrome	377	278	2
Squamous cell carcinoma of the cervix uteri	213767	270	4
Amyotrophic lateral sclerosis-parkinsonism-dementia complex	90020	269	2
Central diabetes insipidus	178029	267	2
Eosinophilic esophagitis	73247	267	4
Microvillus inclusion disease	2290	267	2
Proximal renal tubular acidosis	47159	267	3
Ewing sarcoma	319	266	3
Brooke-Spiegler syndrome	79493	264	2
MYH9-related disease	182050	264	2
Sebastian syndrome	807	264	2
Short bowel syndrome	104008	264	2

Proximal myotonic myopathy	606	257	6
Isolated spina bifida	823	256	7
Motor neuron disease	98503	256	4
Retinitis pigmentosa	791	255	12
Primary central nervous system lymphoma	46135	252	2
Familial adenomatous polyposis	733	251	11
Acute erythroid leukemia	318	247	4
Adult-onset autosomal dominant leukodystrophy	99027	246	1
Birdshot chorioretinopathy	179	246	1
Rhabdomyosarcoma	780	238	7
Rare thyroid disease	101955	237	8
Disorder of lipid metabolism	309005	235	2
Central serous chorioretinopathy	443079	230	5
Paroxysmal dyskinesia	1431	229	2
Pituitary adenoma	99408	229	2
Rare adenocarcinoma of the breast	213528	229	4
Werner syndrome	902	229	3
Progressive autosomal recessive ataxia-deafness syndrome	448251	228	1
Hirschsprung disease	388	226	3
Chromophobe renal cell carcinoma	319303	225	5
Non-Hodgkin lymphoma	547	225	8
Adenocarcinoma of the small instestine	104075	224	1
Anaplastic oligodendroglioma	251630	224	1
Aromatase deficiency	91	224	1
Arterial tortuosity syndrome	3342	224	3
Childhood absence epilepsy	64280	224	1
Classic homocystinuria	394	224	1
Coenzyme Q10 deficiency	35656	224	1
Congenital dyserythropoietic anemia type II	98873	224	2
Congenital myopathy	97245	224	2
Congenital myotonia	206973	224	1
Epilepsy with myoclonic absences	86911	224	1
Epithelioid hemangioendothelioma	157791	224	1

Follicular dendritic cell sarcoma	86902	224	1
Ganglioneuroblastoma	251877	224	1
Ganglioneuroma	251992	224	1
Hemangioblastoma	252054	224	1
Hemolytic anemia due to red cell pyruvate kinase deficiency	766	224	1
Hereditary cryohydrocytosis with reduced stomatin	168577	224	1
Idiopathic hypereosinophilic syndrome	3260	224	1
Isolated megalencephaly	268920	224	7
Megalencephaly	2477	224	7
Microlissencephaly	1083	224	2
Myeloproliferative neoplasm	98274	224	2
Paroxysmal dystonic choreathetosis with episodic ataxia and spasticity	53583	224	1
Paroxysmal exertion-induced dyskinesia	98811	224	1
Paroxysmal kinesigenic dyskinesia	98809	224	1
Paroxysmal non-kinesigenic dyskinesia	98810	224	2
Placental insufficiency	439167	224	1
Rapid-onset dystonia-parkinsonism	71517	224	1
Rare arteriovenous malformation	211266	224	2
Rare hereditary ataxia	183518	224	1
Rare lymphatic malformation	2415	224	1
Rare venous malformation	211252	224	1
Simple vascular malformation	211243	224	1
Thomsen and Becker disease	614	224	1
Uveal melanoma	39044	224	3
X-linked dystonia-parkinsonism	53351	224	1
Rare male infertility	98048	223	11
Fetal akinesia deformation sequence	994	220	1
Lambert-Eaton myasthenic syndrome	43393	220	1
Acute liver failure	90062	214	7
Dengue fever	99828	207	4

Sitosterolemia	2882	206	1
Oculocerebrorenal syndrome of Lowe	534	196	2
Mohr-Tranebjaerg syndrome	52368	192	2
Hereditary hyperekplexia	3197	178	1
Rotor syndrome	3111	178	2
Stiff person syndrome and related disorders	3198	178	2
Hepatoblastoma	449	167	8
Marfan syndrome	558	160	4
Rare primary hyperaldosteronism	181415	159	3
Rare peripheral neuropathy	98496	151	7
Acute lymphoblastic leukemia	513	147	12
Posterior urethral valve	93110	146	2
Autosomal dominant spastic paraplegia type 10	100991	145	1
Autosomal dominant spastic paraplegia type 42	171863	145	1
Congenital arteriovenous fistula	98731	145	1
Congenital cataract-hearing loss-severe developmental delay syndrome	300313	145	1
Cryptococcosis	1546	145	9
Lennox-Gastaut syndrome	2382	144	1
Phenylketonuria	716	143	5
Classic phenylketonuria	79254	142	3
Hereditary diffuse gastric cancer	26106	137	3
Undetermined early-onset epileptic encephalopathy	442835	135	2
Fanconi anemia	84	132	5
Episodic ataxia type 6	209967	131	1
Fragile X-associated tremor/ataxia syndrome	93256	131	1
Human prion disease	56970	131	2
Isolated cerebellar agenesis	1398	131	7
Spinocerebellar ataxia type 1	98755	131	1
Spinocerebellar ataxia type 7	94147	131	1
Anaplastic astrocytoma	251589	130	3
Chromosomal anomaly	68335	130	7
Acute promyelocytic leukemia	520	126	4
Alexander disease	58	125	1

Lafora disease	501	125	2
Myotonic dystrophy	206647	125	2
Neuromyelitis optica	71211	125	2
Osteoglosphonic dysplasia	2645	125	3
Periventricular leukomalacia	171676	125	1
Rasmussen subacute encephalitis	1929	125	1
Hermansky-Pudlak syndrome	79430	119	4
Glycogen storage disease due to acid maltase deficiency	365	118	4
Cerebral cortical dysplasia	268950	116	4
Dejerine-Sottas syndrome	64748	116	5
Dicarboxylic aminoaciduria	2195	116	1
Glycogen storage disease	79201	116	3
Hot water reflex epilepsy	166412	116	1
Alpha-thalassemia	846	115	1
Hemoglobin H disease	93616	115	1
Rare parasitic disease	163588	115	2
Statin toxicity	413696	115	1
Familial tumoral calcinosis	53715	112	3
Autosomal dominant hypophosphatemic rickets	89937	110	4
Rare renal tubular disease	93603	109	3
Germ cell tumor	3399	101	2
Splenic marginal zone lymphoma	86854	100	3
Lysinuric protein intolerance	470	98	7
Extranodal nasal NK/T cell lymphoma	86879	97	1
Histiocytic sarcoma	86896	97	1
Nodular lymphocyte predominant Hodgkin lymphoma	86893	97	1
Primary effusion lymphoma	48686	97	1
Primary mediastinal large B-cell lymphoma	98838	97	1
Spermatocytic seminoma	99865	97	1
T-cell/histiocyte rich large B cell lymphoma	300857	97	1
Hypocalcemic vitamin D-resistant rickets	93160	95	3
Hypophosphatemic rickets	437	95	3
Thymoma	99867	95	2
Perry syndrome	178509	94	2

Neonatal diabetes mellitus	224	92	2
Permanent neonatal diabetes mellitus	99885	92	2
Anaplastic thyroid carcinoma	142	91	4
Rare urinary tract tumor	98058	91	4
Berardinelli-Seip congenital lipodystrophy	528	90	1
Distomatosis	1685	90	1
Galactosemia	352	90	1
Hyperphenylalaninemia due to tetrahydrobiopterin deficiency	238583	90	1
Keratoderma hereditarium mutilans	494	90	2
Thymic carcinoma	99868	90	1
Carnitine-acylcarnitine translocase deficiency	159	83	2
Juvenile idiopathic arthritis	92	82	5
Mastocytosis	98292	81	2
Medullary thyroid carcinoma	1332	81	4
Familial medullary thyroid carcinoma	99361	80	2
Congenital isolated hyperinsulinism	657	79	2
Exercise-induced hyperinsulinism	165991	79	1
Heparin-induced thrombocytopenia	3325	79	1
Ketoacidosis due to monocarboxylate transporter-1 deficiency	438075	79	1
Metabolic myopathy due to lactate transporter defect	171690	79	1
Oncogenic osteomalacia	352540	79	1
Systemic primary carnitine deficiency	158	77	2
Spinocerebellar ataxia type 8	98760	73	1
Ichthyosis	79354	72	5
Sickle cell anemia	232	67	6
Chronic graft versus host disease	99921	65	2
Rare benign ovarian tumor	97293	65	2
Leptospirosis	509	63	2
Infant acute respiratory distress syndrome	70587	58	3

Brain dopamine-serotonin vesicular transport disease	352649	57	1
Hypoparathyroidism-sensorineural deafness-renal disease syndrome	2237	57	1
Bartter syndrome	112	56	4
Classic Bartter syndrome	93605	56	3
Familial hypocalciuric hypercalcemia type 1	93372	56	2
Testicular seminomatous germ cell tumor	842	56	4
Autoimmune hepatitis	2137	54	4
Systemic-onset juvenile idiopathic arthritis	85414	54	5
Endocardial fibroelastosis	2022	52	1
Mandibulofacial dysostosis	155899	52	1
Neutral lipid storage disease	165	52	1
Neutral lipid storage disease with ichthyosis	98907	52	1
Neutral lipid storage myopathy	98908	52	1
Primary hyperoxaluria	416	52	4
Primary hyperoxaluria type 1	93598	52	2
Propionic acidemia	35	52	1
Treacher-Collins syndrome	861	52	1
Embryonal carcinoma	180226	48	2
Extragonadal teratoma	883	48	3
Cleidocranial dysplasia	1452	47	2
Lymphedema-distichiasis syndrome	33001	47	1
Yolk sac tumor	876	47	1
Preeclampsia	275555	46	2
Cystinuria	214	45	7
Leishmaniasis	507	44	5
Hartnup disease	2116	43	2
Wilson disease	905	43	7
Early-onset nuclear cataract	98991	42	1
Gordon syndrome	376	42	2
IRIDA syndrome	209981	42	3
Microcytic anemia with liver iron overload	83642	42	1
Pseudohypoaldosteronism type 2	757	42	3
Glycogen storage disease type 1c	79260	37	3
Cleft palate	2014	34	9

Congenital non-bullous ichthyosiform erythroderma	79394	33	3
Familial calcium pyrophosphate deposition	1416	33	4
Ichthyosis-prematurity syndrome	88621	33	1
Rare insulin-resistance syndrome	181368	33	1
Restrictive dermopathy	1662	33	1
Severe combined immunodeficiency	183660	33	2
Acute intermittent porphyria	79276	32	1
Alopecia	79364	32	8
Porphyria	738	32	1
Inherited retinal disorder	71862	30	4
Acute graft versus host disease	99920	28	2
Arachnoid cyst	2356	28	1
Atrioventricular canal defect	98722	28	1
Cleft lip with or without cleft palate	1991	28	5
Cleft lip/palate	199306	28	1
Fetal alcohol syndrome	1915	28	1
Formiminoglutamic aciduria	51208	28	1
Gitelman syndrome	358	28	2
Hereditary folate malabsorption	90045	28	1
Idiopathic hypercalciuria	2197	28	2
Isolated cleft lip	199302	28	5
Melioidosis	31202	28	2
Methotrexate toxicity or dose selection	413690	28	1
Nance-Horan syndrome	627	28	1
Neurodegenerative syndrome due to cerebral folate transport deficiency	217382	28	1
Neurofibromatosis type 2	637	28	4
Omphalocele	660	28	1
Rare mycosis	163591	28	1
Vestibular schwannoma	252175	28	3
Citrin deficiency	247582	26	5
Citrullinemia type II	247585	26	5
Epithelioid trophoblastic tumor	254698	25	1
Placental site trophoblastic tumor	99928	25	1

Van der Woude syndrome	888	25	1
Hepatitis delta	402823	24	3
Rare renal tumor	93619	23	7
Gerstmann-Straussler-Scheinker syndrome	356	22	1
Glycogen storage disease due to glucose-6-phosphatase deficiency	364	22	1
Glycogen storage disease due to glucose-6-phosphatase deficiency type Ia	79258	22	1
Glycogen storage disease due to glucose-6-phosphatase deficiency type Ib	79259	22	1
Gorham-Stout disease	73	22	1
Hepatocellular adenoma	54272	22	1
Severe congenital neutropenia	42738	22	1
Hereditary hypophosphatemic rickets with hypercalciuria	157215	20	2
Non-acquired isolated growth hormone deficiency	631	20	4
Pendred syndrome	705	18	8
Allan-Herndon-Dudley syndrome	59	16	2
Beta-thalassemia	848	16	5
Primary myelofibrosis	824	16	3
Rare biliary tract disease	101941	16	3
Rare tumor of gallbladder and extrahepatic biliary tract	306633	16	3
Autosomal recessive infantile hypercalcemia	300547	15	1
Carcinoma of gallbladder and extrahepatic biliary tract	56044	15	5
Dominant hypophosphatemia with nephrolithiasis or osteoporosis	244305	15	1
McCune-Albright syndrome	562	15	1
Peutz-Jeghers syndrome	2869	15	3
Primary Fanconi syndrome	3337	15	1
Rare bone development disorder	139012	15	3
Acute monoblastic leukemia	514	14	4
Antenatal Bartter syndrome	93604	14	1

Autosomal dominant primary hypomagnesemia with hypocalciuria	34528	14	1
CHARGE syndrome	138	14	1
Chondrosarcoma	55880	14	7
Cysticercosis	1560	14	3
Dedifferentiated liposarcoma	99970	14	1
EAST syndrome	199343	14	1
Gardner syndrome	79665	14	1
Hemimegalencephaly	99802	14	1
Idiopathic intracranial hypertension	238624	14	1
Lateral meningocele syndrome	2789	14	1
Limb-mammary syndrome	69085	14	1
Lymphangioleiomyomatosis	538	14	1
Nephrogenic diabetes insipidus	223	14	2
Noonan syndrome with multiple lentigines	500	14	1
Rare hyperparathyroidism	181408	14	2
Relapsing fever	91547	14	2
Scleroderma	801	14	4
Subependymal giant cell astrocytoma	251618	14	1
Timothy syndrome	65283	14	1
AL amyloidosis	85443	13	1
GNE myopathy	602	13	2
Mitochondrial disease	68380	13	8
Spastic tetraplegia-thin corpus callosum-progressive postnatal microcephaly syndrome	447997	13	1
Chronic enteropathy associated with SLCO2A1 gene	468641	12	1
Cranio-osteoarthropathy	1525	12	1
Isolated congenital digital clubbing	217059	12	1
Pachydermoperiostosis	2796	12	1
Primary cutis verticis gyrata	671	12	1
Primary hypertrophic osteoarthropathy	248095	12	1
Renal dysplasia	93108	12	2
Alveolar rhabdomyosarcoma	99756	11	1
Centronuclear myopathy	595	11	1

Early-onset autosomal dominant Alzheimer disease	1020	11	3
Glycogen storage disease due to muscle glycogen phosphorylase deficiency	368	11	1
Glycogen storage disease type 1d	79261	11	1
Mitochondrial myopathy	206966	11	4
Simpson-Golabi-Behmel syndrome	373	11	2
Amelocerebrohypohidrotic syndrome	1946	10	1
Amelogenesis imperfecta	88661	10	5
Pyridoxine-dependent epilepsy	3006	10	1
Desmoplastic small round cell tumor	83469	9	1
Rare isolated myopia	98619	8	7
Panhypopituitarism	90695	7	2
Autoimmune polyendocrinopathy	282196	6	5
Autosomal dominant progressive external ophthalmoplegia	254892	6	1
Congenital cataract-hypertrophic cardiomyopathy-mitochondrial myopathy syndrome	1369	6	1
Congenital hypothyroidism	442	6	4
Corneal dystrophy	34533	6	2
Facioscapulohumeral dystrophy	269	6	2
Generalized resistance to thyroid hormone	3221	6	3
Kearns-Sayre syndrome	480	6	1
Leukodystrophy	68356	6	4
MELAS	550	6	1
MERRF	551	6	2
Mitochondrial DNA-related progressive external ophthalmoplegia	663	6	1
Rare familial disorder with hypertrophic cardiomyopathy	99739	6	1
Achondroplasia	15	5	1
Acute myelomonocytic leukemia	517	5	1

Central congenital hypothyroidism	226298	5	2
Chordoma	178	5	1
Diffuse astrocytoma	251595	5	1
Duplication/inversion 15q11	3306	5	1
Helicoid peripapillary chorioretinal degeneration	86813	5	2
Isolated follicle stimulating hormone deficiency	52901	5	1
Lipedema	77243	5	1
Neurofibromatosis type 3	93921	5	1
Partial deletion of the long arm of chromosome 6	262047	5	1
Pituitary deficiency	101957	5	1
Prolactinoma	2965	5	1
Rubinstein-Taybi syndrome	783	5	1
Septo-optic dysplasia spectrum	3157	5	1
Somatotropic adenoma	96256	5	1
Tarsal-carpal coalition syndrome	1412	5	4
ADan amyloidosis	97346	4	2
Benign familial infantile epilepsy	306	4	1
Benign familial neonatal epilepsy	1949	4	1
Benign familial neonatal-infantile seizures	140927	4	1
Bilateral striopallidodentate calcinosis	1980	4	2
Carnitine palmitoyltransferase II deficiency	157	4	1
Isolated cytochrome C oxidase deficiency	254905	4	1
Long chain 3-hydroxyacyl-CoA dehydrogenase deficiency	5	4	1
Mitochondrial trifunctional protein deficiency	746	4	1
Multiple acyl-CoA dehydrogenase deficiency	26791	4	3
Peroxisome biogenesis disorder	79189	4	1
Rare urticaria	79384	4	2
St. Louis encephalitis	83484	4	1
Very long chain acyl-CoA dehydrogenase deficiency	26793	4	1

Waldenström macroglobulinemia	33226	4	2
Wiskott-Aldrich syndrome	906	4	1
22q11.2 deletion syndrome	567	3	3
Early-onset anterior polar cataract	98988	3	4
Hyperornithinemia-hyperammonemia-homocitrullinuria syndrome	415	3	3
Immune thrombocytopenic purpura	3002	3	1
MALT lymphoma	52417	3	3
Marginal zone lymphoma	300912	3	3
Monosomy X	99226	3	3
Photosensitive epilepsy	166409	3	1
Polycythemia vera	729	3	5
Potassium-aggravated myotonia	612	3	2
Pulmonary alveolar microlithiasis	60025	3	1
Shprintzen-Goldberg syndrome	2462	3	2
Turner syndrome	881	3	3
Adrenomyeloneuropathy	139399	2	1
Autoimmune hemolytic anemia	98375	2	2
Autoimmune hemolytic anemia, cold type	228312	2	1
Barth syndrome	111	2	2
Cold agglutinin disease	56425	2	1
Congenital hydrocephalus	2185	2	2
Conotruncal heart malformations	2445	2	2
Cystinuria type B	93613	2	2
Ependymal tumor	301	2	1
Ependymoma	251636	2	1
Essential thrombocythemia	3318	2	3
Extragonadal germinoma	182127	2	1
Fish-eye disease	79292	2	1
Guanidinoacetate methyltransferase deficiency	382	2	1
Herpes simplex virus keratitis	137586	2	1
Idiopathic isolated micropenis	95707	2	5
Idiopathic pulmonary fibrosis	2032	2	8
Iminoglycinuria	42062	2	4

Interatrial communication	1478	2	3
L-Arginine:glycine amidinotransferase deficiency	35704	2	1
Leber congenital amaurosis	65	2	3
Leprosy	548	2	5
Non-functioning pituitary adenoma	91349	2	1
Penile agenesis	49	2	4
Piebaldism	2884	2	2
Rare coagulation disorder	98429	2	2
Rare hemorrhagic disorder	248308	2	2
Sideroblastic anemia	1047	2	2
Thiamine-responsive megaloblastic anemia syndrome	49827	2	1
X-linked creatine transporter deficiency	52503	2	1
2p21 microdeletion syndrome	163693	1	1
46,XX testicular disorder of sex development	393	1	2
Acquired purpura fulminans	49566	1	1
Adrenocortical carcinoma	1501	1	2
Aregenerative anemia	101096	1	2
Atypical hypotonia-cystinuria syndrome	238523	1	1
Autosomal recessive limb-girdle muscular dystrophy type 2B	268	1	1
Autosomal recessive spinocerebellar ataxia-blindness-deafness syndrome	95433	1	1
Beta-thalassemia intermedia	231222	1	2
Biotin-thiamine-responsive basal ganglia disease	65284	1	1
Cardiomyopathy-hypotonia-lactic acidosis syndrome	91130	1	1
Charcot-Marie-Tooth disease/Hereditary motor and sensory neuropathy	166	1	3
Citrullinemia	187	1	2
Congenital thrombotic thrombocytopenic purpura	93583	1	1
Corpus callosum agenesis-neuronopathy syndrome	1496	1	1
Craniosynostosis	1531	1	7
Cutis laxa	209	1	2
Cystinuria type A	93612	1	1

Disseminated superficial actinic porokeratosis	79152	1	1
Early infantile epileptic encephalopathy	1934	1	1
Early myoclonic encephalopathy	1935	1	1
Encephalitis	97275	1	2
Epileptic encephalopathy with global cerebral demyelination	353217	1	1
Familial thyroid dyshormonogenesis	95716	1	1
Giant cell tumor of bone	363976	1	1
Gray platelet syndrome	721	1	1
Gyrate atrophy of choroid and retina	414	1	1
Hereditary gingival fibromatosis	2024	1	2
Hyper-beta-alaninemia	309147	1	1
Hyperlysinemia	2203	1	1
Hypotonia-cystinuria syndrome	163690	1	1
Infantile spasms-psychomotor retardation-progressive brain atrophy-basal ganglia disease syndrome	263410	1	1
Isolated brachycephaly	35099	1	2
Isolated craniosynostosis	139390	1	7
Isolated oxycephaly	63440	1	2
Langer mesomelic dysplasia	2632	1	1
Leigh syndrome	506	1	4
Leigh syndrome with leukodystrophy	255241	1	1
Limb-girdle muscular dystrophy	263	1	1
Léri-Weill dyschondrosteosis	240	1	1
Macroglossia	156207	1	2
Malignant peripheral nerve sheath tumor	3148	1	1
Marburg hemorrhagic fever	99826	1	1
Microsporidiosis	2552	1	1
Mucopolysaccharidosis type 4	582	1	2
Myasthenia gravis	589	1	2
Neonatal intrahepatic cholestasis due to citrin deficiency	247598	1	1
Noonan syndrome	648	1	1

Ornithine transcarbamylase deficiency	664	1	1
Overhydrated hereditary stomatocytosis	3203	1	1
Porokeratosis	79358	1	1
Precocious puberty	95708	1	3
Pyruvate carboxylase deficiency	3008	1	1
Rare disorder with hypertrichosis	79365	1	4
Refractory anemia	98826	1	2
Rh deficiency syndrome	71275	1	1
Riboflavin transporter deficiency	97229	1	2
Roussy-Lévy syndrome	3115	1	2
Scrub typhus	83317	1	1
Spinocerebellar ataxia type 5	98766	1	1
Thiamine-responsive encephalopathy	199348	1	1
Thrombotic microangiopathy	93573	1	1
Thrombotic thrombocytopenic purpura	54057	1	1
Uveitis	98715	1	1
X-linked centronuclear myopathy	596	1	1
2-hydroxyglutaric aciduria	19	0	1
ALG2-CDG	79326	0	1
Achondrogenesis	932	0	1
Achondrogenesis type 1B	93298	0	4
Acquired idiopathic sideroblastic anemia	75564	0	2
Acrodermatitis enteropathica	37	0	2
Acute hepatic porphyria	95157	0	1
Acute megakaryoblastic leukemia	518	0	2
Adult T-cell leukemia/lymphoma	86875	0	2
Adult neuronal ceroid lipofuscinosis	79262	0	2
Adult-onset autosomal recessive sideroblastic anemia	255132	0	1
African trypanosomiasis	3385	0	1
Agammaglobulinemia	183669	0	1
Aicardi-Goutières syndrome	51	0	1

Allergic bronchopulmonary aspergillosis	1164	0	1
Alpha-thalassemia-X-linked intellectual disability syndrome	847	0	1
Alström syndrome	64	0	1
American trypanosomiasis	3386	0	3
Amish lethal microcephaly	99742	0	1
Anaplastic large cell lymphoma	98841	0	1
Androgen insensitivity syndrome	754	0	2
Angelman syndrome due to maternal 15q11q13 deletion	98794	0	1
Antisynthetase syndrome	81	0	1
Apparent mineralocorticoid excess	320	0	2
Atelosteogenesis type I	1190	0	1
Atelosteogenesis type II	56304	0	1
Athyreosis	95713	0	1
Audiogenic seizures	166415	0	1
Autism spectrum disorder-epilepsy-arthrogryposis syndrome	370943	0	1
Autoimmune pancreatitis	103919	0	1
Autosomal dominant Charcot-Marie-Tooth disease type 2	64746	0	1
Autosomal dominant distal renal tubular acidosis	93608	0	1
Autosomal dominant non-syndromic sensorineural deafness type DFNA	90635	0	2
Autosomal dominant spastic paraplegia type 4	100985	0	2
Autosomal dominant spastic paraplegia type 6	100988	0	1
Autosomal erythropoietic protoporphyria	79278	0	1
Autosomal recessive distal renal tubular acidosis	402041	0	1
Autosomal recessive non-syndromic intellectual disability	88616	0	2
Autosomal recessive non-syndromic sensorineural deafness type DFNB	90636	0	2

Autosomal recessive polycystic kidney disease	731	0	1
Autosomal recessive primary microcephaly	2512	0	1
Autosomal recessive proximal renal tubular acidosis	93607	0	1
Autosomal recessive sideroblastic anemia	260305	0	1
Autosomal recessive spastic paraplegia type 5A	100986	0	1
Autosomal recessive spondylocostal dysostosis	2311	0	1
Axenfeld-Rieger syndrome	782	0	1
Baraitser-Winter cerebrofrontofacial syndrome	2995	0	1
Beckwith-Wiedemann syndrome	116	0	1
Beta-thalassemia major	231214	0	2
Bile acid CoA ligase deficiency and defective amidation	276066	0	1
Bilirubin encephalopathy	415286	0	1
Blackfan-Diamond anemia	124	0	1
Bloom syndrome	125	0	2
Bowen syndrome	1271	0	1
Buerger disease	36258	0	2
CLN2 disease	228349	0	1
CLN3 disease	228346	0	1
CLN7 disease	228366	0	1
CLN8 disease	228354	0	1
Caffey disease	1310	0	1
Campomelic dysplasia	140	0	1
Camurati-Engelmann disease	1328	0	1
Central core disease	597	0	1
Chandler syndrome	98979	0	1
Charcot-Marie-Tooth disease type 1	65753	0	1
Charcot-Marie-Tooth disease type 1A	101081	0	1
Charcot-Marie-Tooth disease type 1B	101082	0	1
Chikungunya	324625	0	2
Christianson syndrome	85278	0	3
Chronic beryllium disease	133	0	1
Chronic nonbacterial osteomyelitis/Chronic recurrent multifocal osteomyelitis	324964	0	1
Chédiak-Higashi syndrome	167	0	1

Cirrhosis-dystonia-polycythemia-hypermanganesemia syndrome	309854	0	1
Classic Hodgkin lymphoma, mixed cellularity type	98844	0	1
Classic Hodgkin lymphoma, nodular sclerosis type	98843	0	1
Cleft velum	99772	0	1
Cockayne syndrome type 1	90321	0	2
Coloboma of iris	98944	0	1
Combined hyperlipidemia	79211	0	1
Congenital chloride diarrhea	53689	0	2
Congenital disorder of glycosylation	137	0	9
Congenital hereditary endothelial dystrophy type I	98975	0	1
Congenital hereditary endothelial dystrophy type II	293603	0	1
Congenital mesoblastic nephroma	2665	0	1
Congenital neuronal ceroid lipofuscinosis	168486	0	3
Congenital radioulnar synostosis	3269	0	1
Congenital stationary night blindness	215	0	1
Congenital vertical talus	178382	0	1
Constitutional sideroblastic anemia	98362	0	1
Corneal dystrophy-perceptive deafness syndrome	1490	0	1
Craniometaphyseal dysplasia	1522	0	1
Crigler-Najjar syndrome	205	0	1
Crigler-Najjar syndrome type 1	79234	0	1
Crimean-Congo hemorrhagic fever	99827	0	1
Cutaneous neuroendocrine carcinoma	79140	0	1
Cyclic neutropenia	2686	0	1
Cystic echinococcosis	400	0	1
Cystic hygroma	79486	0	1
D,L-2-hydroxyglutaric aciduria	356978	0	1
Darier disease	218	0	1
Dehydrated hereditary stomatocytosis	3202	0	1

Dentatorubral pallidoluysian atrophy	101	0	1
Dermatomyositis	221	0	2
Diastrophic dwarfism	628	0	9
Diphtheria	1679	0	5
Discoid lupus erythematosus	90281	0	5
Distal renal tubular acidosis	18	0	4
Distal renal tubular acidosis with anemia	93610	0	1
Dubin-Johnson syndrome	234	0	1
Dysosteosclerosis	1782	0	1
Dystonia-parkinsonism-hypermanganesemia syndrome	521406	0	1
EEC syndrome	1896	0	1
Ehlers-Danlos syndrome, spondylocheirodysplastic type	157965	0	1
Ehlers-Danlos syndrome, vascular type	286	0	1
Embryonal rhabdomyosarcoma	99757	0	1
Endometrial stromal sarcoma	213711	0	1
Epidermodysplasia verruciformis	302	0	4
Esophageal atresia	1199	0	1
Extramammary Paget disease	2800	0	1
Familial Mediterranean fever	342	0	1
Familial hyperaldosteronism type I	403	0	1
Familial isolated clinodactyly of fingers	295014	0	2
Familial multiple lipomatosis	199276	0	1
Familial pancreatic carcinoma	1333	0	1
Familial prostate cancer	1331	0	1
Femoral agenesis/hypoplasia	1987	0	1
Fibrochondrogenesis	2021	0	1
Fibrosarcoma	2030	0	4
Filariasis	2034	0	1
Foveal hypoplasia-optic nerve decussation defect-	397618	0	1

anterior segment dysgenesis syndrome			
Fowler syndrome	221126	0	2
Free sialic acid storage disease	834	0	1
Free sialic acid storage disease, infantile form	309324	0	1
Friedreich ataxia	95	0	1
Fuchs endothelial corneal dystrophy	98974	0	1
GM2 gangliosidosis	309152	0	1
Germ cell tumor of testis	363504	0	3
Giant cell arteritis	397	0	2
Giant cell glioblastoma	251579	0	1
Gorlin-Chaudhry-Moss syndrome	2095	0	1
Granulomatosis with polyangiitis	900	0	1
Growth and developmental delay-hypotonia-vision impairment-lactic acidosis syndrome	391348	0	1
Growth delay due to insulin-like growth factor type 1 deficiency	73272	0	1
Growth hormone insensitivity syndrome	181393	0	1
H syndrome	168569	0	1
HELLP syndrome	244242	0	1
Hemochromatosis type 2	79230	0	1
Hemoglobinopathy	68364	0	1
Hemophagocytic syndrome	158032	0	1
Hemophilia	448	0	1
Hemophilia A	98878	0	2
Hereditary breast cancer	227535	0	1
Hereditary clear cell renal cell carcinoma	422526	0	1
Hereditary cryohydrocytosis with normal stomatin	398088	0	1
Hereditary elliptocytosis	288	0	1
Hereditary motor and sensory neuropathy type 6	90120	0	1
Hereditary sensory and autonomic neuropathy	140471	0	1
Hereditary spherocytosis	822	0	4
Hereditary stomatocytosis	98365	0	1
Herpes simplex virus encephalitis	1930	0	1
Hurler syndrome	93473	0	1

Hurler-Scheie syndrome	93476	0	1
Hydranencephaly	2177	0	1
Hydrops fetalis	1041	0	2
Hyperinsulinism due to UCP2 deficiency	276556	0	1
Hyperostosis cranialis interna	443098	0	1
Hyperpigmentation of the skin	79375	0	1
Hypersensitivity pneumonitis	31740	0	1
Hypocalcified amelogenesis imperfecta	100032	0	1
Hypomaturation amelogenesis imperfecta	100033	0	1
Hypopigmentation of the skin	79376	0	5
Idiopathic achalasia	930	0	1
Idiopathic bronchiectasis	60033	0	1
Idiopathic chronic eosinophilic pneumonia	2902	0	1
Incontinentia pigmenti	464	0	1
Infantile neuronal ceroid lipofuscinosis	79263	0	1
Interdigitating dendritic cell sarcoma	86900	0	1
Intermediate severe Salla disease	309331	0	1
Interstitial lung disease	182095	0	4
Isolated Dandy-Walker malformation	217	0	2
Isolated Pierre Robin syndrome	718	0	1
Isolated agammaglobulinemia	229717	0	2
Isolated aniridia	250923	0	2
Isolated biliary atresia	30391	0	1
Isolated focal cortical dysplasia type Ia	268973	0	1
Isolated optic nerve hypoplasia/aplasia	137902	0	1
Ito hypomelanosis	435	0	1
Jeune syndrome	474	0	1
Juvenile cataract-microcornea-renal glucosuria syndrome	247794	0	1
Juvenile neuronal ceroid lipofuscinosis	79264	0	2
Kennedy disease	481	0	1

Klatskin tumor	99978	0	1
LCAT deficiency	650	0	1
Lamellar ichthyosis	313	0	1
Langerhans cell histiocytosis	389	0	1
Laron syndrome	633	0	1
Lassa fever	99824	0	1
Late infantile neuronal ceroid lipofuscinosis	168491	0	2
Leber hereditary optic neuropathy	104	0	1
Left ventricular noncompaction	54260	0	1
Legg-Calvé-Perthes disease	2380	0	1
Lemierre syndrome	137839	0	1
Leukocyte adhesion deficiency	2968	0	1
Leukocyte adhesion deficiency type I	99842	0	1
Leukocyte adhesion deficiency type II	99843	0	1
Low phospholipid-associated cholelithiasis	69663	0	1
Low-grade astrocytoma	251592	0	1
Lymphedema	79383	0	1
Lymphoproliferative syndrome	238510	0	3
Macrophage activation syndrome	158061	0	1
Macular corneal dystrophy	98969	0	1
Malignant hyperthermia of anesthesia	423	0	1
Manganese poisoning	306682	0	4
Marinesco-Sjögren syndrome	559	0	1
Maternal riboflavin deficiency	411712	0	1
Medium chain acyl-CoA dehydrogenase deficiency	42	0	1
Melkersson-Rosenthal syndrome	2483	0	1
Meningioma	2495	0	1
Meningococcal meningitis	33475	0	1
Mesomelia-synostoses syndrome	2496	0	1
Microtia	83463	0	1
Miller-Dieker syndrome	531	0	1
Mitochondrial DNA depletion syndrome	35698	0	1

Mitochondrial pyruvate carrier deficiency	447784	0	1
Moyamoya disease	2573	0	3
Mucolipidosis type IV	578	0	1
Mucopolysaccharidosis type 1	579	0	1
Mucopolysaccharidosis type 2	580	0	1
Multiple acyl-CoA dehydrogenase deficiency, mild type	394532	0	1
Multiple epiphyseal dysplasia	251	0	4
Multiple epiphyseal dysplasia type 4	93307	0	1
Multiple osteochondromas	321	0	2
Multiple symmetric lipomatosis	2398	0	1
Myofibrillar myopathy	593	0	2
Nail anomaly	79368	0	2
Nemaline myopathy	607	0	1
Neonatal severe cardiopulmonary failure due to mitochondrial methylation defect	466784	0	1
Neuroendocrine cell hyperplasia of infancy	217560	0	1
Neuromuscular disease	68381	0	1
Neuronal ceroid lipofuscinosis	216	0	3
Nevus of Ito	263432	0	1
Nijmegen breakage syndrome	647	0	1
Non-syndromic male infertility due to sperm motility disorder	276234	0	1
Non-syndromic syndactyly	90025	0	1
Occipital horn syndrome	198	0	1
Ocular albinism	284804	0	4
Oculocutaneous albinism	55	0	4
Oculocutaneous albinism type 2	79432	0	2
Oculocutaneous albinism type 4	79435	0	1
Oculocutaneous albinism type 6	370097	0	1
Ondine syndrome	661	0	1
Oral submucous fibrosis	357154	0	1

Osteopetrosis and related disorders	2781	0	2
Osteopetrosis-hypogammaglobulinemia syndrome	178389	0	1
Overlapping connective tissue disease	251312	0	1
Papillary glioneuronal tumor	251962	0	1
Papillon-Lefèvre syndrome	678	0	1
Periodic paralysis	206976	0	1
Periventricular nodular heterotopia	98892	0	1
Pitt-Rogers-Danks syndrome	98788	0	1
Pneumocystosis	723	0	2
Polycythemia	98427	0	1
Polydactyly of a biphalangeal thumb	93339	0	1
Polymicrogyria	35981	0	2
Posterior column ataxia-retinitis pigmentosa syndrome	88628	0	1
Posterior polymorphous corneal dystrophy	98973	0	1
Prader-Willi syndrome due to maternal uniparental disomy of chromosome 15	98754	0	1
Prader-Willi syndrome due to paternal deletion of 15q11q13 type 1	177901	0	1
Prader-Willi syndrome due to paternal deletion of 15q11q13 type 2	177904	0	1
Premature aging	79389	0	2
Primary cutaneous CD30+ T-cell lymphoproliferative disease	541	0	1
Primary cutaneous T-cell lymphoma	171901	0	1
Primary cutaneous anaplastic large cell lymphoma	300865	0	1
Primary immunodeficiency	101997	0	1
Primitive portal vein thrombosis	854	0	1
Progeroid syndrome, Petty type	2963	0	1
Progressive bulbar paralysis of childhood	56965	0	1

Progressive essential tremor-speech impairment-facial dysmorphism-intellectual disability-abnormal behavior syndrome	457212	0	1
Progressive multifocal leukoencephalopathy	217260	0	3
Progressive polyneuropathy with bilateral striatal necrosis	217396	0	1
Pseudoxanthoma elasticum	758	0	1
Psychomotor regression-oculomotor apraxia-movement disorder-nephropathy syndrome	505242	0	1
Pulverulent cataract	98984	0	1
Pyle disease	3005	0	2
Pyruvate dehydrogenase E1-alpha deficiency	79243	0	1
Pyruvate dehydrogenase deficiency	765	0	1
RFT1-CDG	244310	0	2
Rabies	770	0	2
Rare acquired hemolytic anemia	182047	0	1
Rare benign breast tumor	180253	0	2
Rare deafness	68361	0	4
Rare disease with Pierre Robin syndrome	138044	0	1
Rare disorder with hypogonadotropic hypogonadism	181387	0	1
Rare genetic skin disease	68346	0	1
Rare hypoaldosteronism	181419	0	3
Rare nevus	294057	0	1
Rare refraction anomaly	98618	0	2
Rare soft tissue tumor	71209	0	1
Rare tumor of intestine	104011	0	1
Rare tumor of neuroepithelial tissue	251558	0	1
Rare tumor of salivary glands	276142	0	1
Rare uterine cancer	213564	0	1
Reactive arthritis	29207	0	1
Recombinant 8 syndrome	96167	0	1
Red cell aplasia	98421	0	1
Renal agenesis, unilateral	93100	0	1
Restrictive cardiomyopathy	217632	0	1

Rhabdoid tumor	69077	0	1
Rheumatic fever	3099	0	1
Rickettsial disease	102021	0	1
SLC35A1-CDG	238459	0	1
SLC35A2-CDG	356961	0	1
SLC39A8-CDG	468699	0	1
Salla disease	309334	0	1
Sanfilippo syndrome type A	79269	0	1
Sarcoidosis	797	0	3
Schistosomiasis	1247	0	2
Schneckenbecken dysplasia	3144	0	1
Sclerosing cholangitis	447771	0	2
Shigellosis	810	0	1
Short rib-polydactyly syndrome	1505	0	1
Sialuria	3166	0	1
Situs inversus totalis	101063	0	1
Smith-Lemli-Opitz syndrome	818	0	1
Southeast Asian ovalocytosis	98868	0	1
Spondyloepimetaphyseal dysplasia-abnormal dentition syndrome	168451	0	1
Spondylometaphyseal dysplasia-cone-rod dystrophy syndrome	85167	0	1
Sporadic Creutzfeldt-Jakob disease	204	0	1
Staphylococcal toxic-shock syndrome	99919	0	1
Stargardt disease	827	0	1
Stevens-Johnson syndrome/toxic epidermal necrolysis spectrum	95455	0	1
Subcortical band heterotopia	99796	0	1
Syndromic telecanthus	98575	0	2
TMEM165-CDG	314667	0	1
Tetralogy of Fallot	3303	0	2
Thyroid hypoplasia	95720	0	1
Toxic epidermal necrolysis	537	0	1
Transient myeloproliferative syndrome	420611	0	1
Trichinellosis	863	0	1
Tropical spastic paraparesis	289326	0	1
Typhoid	99745	0	1
Vernal keratoconjunctivitis	70476	0	1

Vogt-Koyanagi-Harada disease	3437	0	1
Whooping cough	1489	0	1
Williams-Campbell syndrome	411501	0	1
Wolf-Hirschhorn syndrome	280	0	1
Wolman disease	75233	0	1
X-linked immunodeficiency with magnesium defect, Epstein-Barr virus infection and neoplasia	317476	0	1
X-linked myopathy with excessive autophagy	25980	0	1
X-linked non-syndromic intellectual disability	777	0	1
X-linked recessive ocular albinism	54	0	4

6.1.2 SLCs with number of associated diseases and disease classes

SLC name	UniProt.ID	Number of associated rare diseases	associated MeSH disease classes
SLC2A1	P11166	143	C04, C06, C10, C23, C16, C17, C05, C14, C18, C20, C15, C12, C13, C08, C19, C07, C11, F01, F03, C01, C09
SLCO6A1	Q86UG4	108	C06, C16, C17, C18, C04, C15, C20, C10, C08, C12, C13, C23, C01, C05, C19, F03, C14, C07, C09, C11
SLC12A9	Q9BXP2	75	C04, C15, C20, C06, C10, C14, C16, C18, C23, C12, C13, C05, C08, C17, C07, C19, F03, C01, C11, C09
SLC12A3	P55017	65	C04, C12, C13, C16, C18, C19, C11, C06, C01, C09, C10, C23, F01, F03, C05, C07, C15, C20, C14, C17
SLC6A3	Q01959	60	C06, C14, C16, C18, C10, C15, C20, C04, C08, C11, C23, C12, C13, C05, C07, C09, C17, F03, C01, C19
SLC16A1	P53985	59	C04, C06, C15, C20, C01, C16, C18, C11, C23, C12, C13, C05, C19, C08, C17, C10, C14, C07, C09
SLC52A2	Q9HAB3	54	C16, C04, C19, C20, C10, C06, C12, C13, C18, C15, C07, C08, C05, C23, C01, C17, C14, C11, C09
SLC6A4	P31645	53	C06, C14, C16, C23, C07, C11, C17, C08, C04, C19, C10, C13, C18, F03, C12, C05
SLC6A8	P48029	52	C04, C15, C20, C06, C14, C16, C18, C12, C13, C08, C10, C23, C05, F01, F03, C19, C07, C09, C17, C11
SLC5A5	Q92911	51	C04, C18, C15, C10, C14, C16, C20, C05, C19, C06, C07, C08, C11, C09, C12, C13, C23, C17, C01
NPC1	O15118	50	C06, C08, C16, C23, C18, C01, C04, C15, C20, C10, C11, C05, F03, C14, C07, C09, C19
SLC7A5	Q01650	48	C04, C06, C11, C10, C16, C18, C12, C13, C05, C19, C01, C09, C17, C23, C15, C20, C08, C14
XPR1	Q9UBH6	47	C04, C17, C23, C10, C18, C15, C05, C07, C11, C14, C16, C01, C12, C13, C06, C19, C08, C20
SLC17A5	Q9NRA2	45	C06, C20, C19, C16, C01, C11, C10, C18, C13, C04, C15, C23, C17, C08, F03, C14, C07, C25

SLC6A2	P23975	45	C08, C16, C04, C10, C06, C07, C19, C14, C18, F03, C05, C17, C12, C13, C01, C11
OCA2	Q04671	42	C01, C18, C10, C16, C06, C20, C04, C15, C11, C17, C05, C13, C19
MAGT1	Q9H0U3	41	C04, C15, C20, C10, C16, C06, C12, C13, C18, C01, C19, F03, C08, C14, C07, C09, C23, C17, C11
SLC4A1	P02730	41	C04, C15, C12, C13, C16, C19, C20, C18, C10, C23, C05, C17, C06, C01, C14
SLC19A1	P41440	39	C20, C04, C15, C17, C23, C14, C16, C07, C05, C10, C11, C12, C13, C06, C19, C18, C01, F03
SLC26A4	O43511	38	C05, C16, C04, C18, C12, C13, C15, C20, C19, C17, C06, C08, C07, C09, C10, C23, C11, C01
SLC1A2	P43004	37	C06, C04, C10, C16, C18, C08, F03, C14, C15, C20, C23, C05, C11, C19, C17
SLC2A3	P11169	37	C04, C06, C20, C19, C13, C05, C07, C16, C10, F03, C17, C15, C14, C12, C18, C01
SLC16A3	O15427	35	C04, C06, C01, C15, C20, C12, C13, C16, C05, C10, C08, C11, C14, C19, C17, C18
UCP2	P55851	35	C06, C10, C18, C04, C16, C15, F03, C19, C14, C17, C20, C11
SLC25A3	Q00325	34	C04, C18, C15, C12, C13, C01, C16, C06, C14, C05, C17, C19, C20, C10, C23, F03, C11
SLC2A4	P14672	34	C04, C15, C20, C05, C10, C12, C13, C14, C18, F03, C06, C16, C19, C17, C23, F01
FLVCR1	Q9Y5Y0	33	C18, C10, C19, C20, C15, C16, C04, C12, C13, C06, F03, C05, C17, C14, C11, C23, C01
SLC20A1	Q8WUM9	33	C05, C16, C10, C19, C04, C20, C13, C17, C15, C18
SLC26A3	P40879	33	C05, C16, C18, C12, C13, C19, C04, C17, C10, C23, C06, C08, C01, C15, C20, C09, C14
SLC35A2	P78381	33	C04, C06, C15, C16, C10, C18, C05, C01, C17, F03, C20, C11, C13, C19, C23, C14
SLC3A2	P08195	33	C04, C01, C15, C20, C06, C11, C12, C13, C05, C16, C10, C09, C17, C23, C18, C14, C08
SLC16A4	O15374	32	C04, C06, C01, C15, C20, C12, C13, C16, C05, C10, C08, C11, C14, C19, C18, C17
SLC27A5	Q9Y2P5	32	C06, C04, C15, C20, C01, C08, C23, C24, C26, C16, C14, C17, C18, C10, C19
SLC11A1	P49279	31	C01, C07, C11, C14, C16, C17, C04, C06, C08, C05, C20, C15, C10, C18, C19
SLC18A2	Q05940	31	C04, C08, C16, C10, C05, C07, C09, C23, C17, C18, F03, C06, C12, C13, C19, C11, C20, C14
SLC22A2	O15244	31	C04, C10, C14, C16, C18, C20, C15, C06, C12, C13, C01, C07, C09, C19, C08
SLC22A3	O75751	31	C11, C04, C12, C13, C05, C16, C06, C08, C17, C15, C09, C10, C23, F01, F03, C07, C19, C18, C20
SLC25A20	O43772	31	C15, C16, C01, C04, C20, C06, C18, C11, C17, C05, C10, C14, C13, C19
SLC26A2	P50443	31	C05, C16, C07, C04, C19, C13, C15, C20, C23, C10, C17, C18, C09, C06, C08

UCP1	P25874	31	C10, C18, C14, C04, C06, C12, C13, C16, C01, C15, C20, F03, C08, C11, C19, C17, C05, C23, F01
CLN3	Q13286	30	C10, C16, C18, C04, C06, C12, C13, C05, C19, C14, C20, C11, C17
SLC52A1	Q9NWF4	30	C19, C20, C04, C06, C05, C16, C08, C15, C13, C14, C11, C18, C07, C09, C01, C17
SLC5A2	P31639	30	C05, C10, C19, C04, C17, C20, C12, C13, C16, C18, C23, C06, C14, C11
SLC8A1	P32418	30	C14, C16, C04, C19, C10, C15, C20, C17, C06, C08, C05, C12, C11, C13, C18
SLC9A1	P19634	30	C04, C12, C13, C16, C18, C23, C06, C10, F03, C14, C19, C15, C20, C17, C01, C08
SLC25A4	P12235	28	C11, C14, C04, C10, C16, C05, C18, C23, C19, C17, C13
SLC2A2	P11168	28	C06, C16, C05, C18, C17, C12, C13, C10, C23, C04, C19, C20, C01
SLC2A10	O95528	27	C05, C10, C14, C16, C17, C04, C23, C11, C08, C20, C18, C19
SLC11A2	P49281	26	C18, C10, C08, C24, C04, C16, C23, C11, C06, F03, C15, C13, C19, F01, C17, C01
SLC12A2	P55011	26	C06, C10, C18, C04, C16, C12, C13, C19, C08, C14, C25, C11, C05, C07, F03, C09
SLC29A3	Q9BZD2	26	C17, C23, C07, C16, C05, C15, C18, C12, C10, C08, C19, C20, C04, C06, C14, C01
SLC33A1	O00400	26	C10, C18, C16, C14, C23, C13, C04, C06, C15, C20, C05, C17, C09, C11, F01, F03, C12, C19
LETM1	O95202	25	C11, C16, C04, C12, C13, C07, C05, C23, C06, C14, C10, C09, C19, C18
SLC22A5	O76082	25	C07, C11, C14, C16, C17, C04, C15, C10, C18, C06, C05, C19, C20, C12, C23, C01
SLC25A1	P53007	25	C10, C16, C18, C05, C14, C15, C19, C04, C20, C07, C11, C17, C06, F03, C23
SLC1A5	Q15758	24	C04, C10, C14, C16, C18, C20, C06, C12, C13, C08, C19, C01, C17
SLCO1B1	Q9Y6L6	24	C04, C15, C16, C10, C18, C06, C20, C13, C19, C23, C17, C14, C01
SLC1A3	P43003	23	C10, C23, C18, C04, C16, C19, C15, C06, C01, F03, C05
SLC25A37	Q9NYZ2	23	C15, C20, C06, C10, C18, C16, C17, C04, C07, C14, C05, C13, C19
SLC40A1	Q9NP59	23	C10, C18, C04, C16, C17, C08, C24, C13, C15, C20, F03, C14, C06, C19, C11, C01
SLC9A6	Q92581	23	C01, C08, C20, C10, C16, C05, C11, C23, F01, F03, C06, C07, C09, C14, C04, C18, C13, C17
SLC12A6	Q9UHW9	22	C15, C16, C10, C23, C05, C04, C17, C13, C11
SLC22A1	O15245	21	C04, C15, C06, C05, C16, C13, C19, C23, C10, C18, C17, C01
SLC25A19	Q9HC21	21	C04, C05, C10, C16, C07, C19, C01, C06, C14, C17, C20, F03, C18
SLC25A24	Q6NUK1	21	C05, C16, C04, C06, C14, C17, C12, C08, C10, F03, C23, C18

SFXN1	Q9H9B4	20	C06, C08, C16, C23, C04, C10, C15, C01, C13, C14, C20, C17, C11, C18, C19, C05
SLC29A1	Q99808	20	C04, C05, C16, C06, C08, C15, C20, C10, F03, C11, C19, C23
SLC30A10	Q6XR72	20	C10, C14, C16, C18, C20, C19, C01, C04, C06, C05, C25, C15, C11
SLC39A8	Q9C0K1	20	C16, C18, C12, C13, C19, C05, C10, F03, C14, C25, C04, C15, C06, C17, C01
SLCO2A1	Q92959	20	C15, C16, C05, C04, C06, C10, F03, C23, C11, C12, C17
TUSC3	Q13454	20	C04, C06, C10, C16, C12, C13, C18, C05, C19, C23, C17, F03, C14
SLC19A3	Q9BZV2	19	C10, C14, C05, C16, C18, C15, C04, C17, C06, C19, C20, C11, C23
SLC1A1	P43005	19	C10, C18, C04, C16, C12, C13, C23, F01, F03, C06, C05
SLC22A18	Q96BI1	19	C16, C04, C06, C15, C20, C17, C01, C13, C19, C12, C18
SLC25A21	Q9BQT8	19	C14, C16, C18, C07, C05, C04, C06, C01, C10, C15, C20, C09, C17
SLC37A4	O43826	19	C01, C10, C16, C04, C18, C05, C06, C08, C14, C19, C20, C15
SLC38A1	Q9H2H9	19	C04, C10, C14, C16, C18, C20, C05, C19, C13, C06, C15, C12, C17, C11
SLC5A8	Q8N695	19	C04, C19, C11, C16, C06, C10, C09, C23, C18, C20, C17
SLC52A3	Q9NQ40	18	C04, C06, C10, C18, C16, C17, C13, C11, C23, C09
SLC5A1	P13866	18	C06, C16, C10, C11, C18, C01, C12, C13, C04, C14, C19, C20, C17
SLCO1B3	Q9NPD5	18	C04, C15, C06, C20, C10, C16, F03, C01, C07, C09, C13, C19, C23, C17, C18
MPC1	Q9Y5U8	17	C04, C12, C13, C16, C05, C10, C17, C06, C08, C14, C15, C20, C18, C19
SLC13A5	Q86YT5	17	C07, C10, C16, F03, C04, C06, C05, C13, C19, C09, C23, C11, C18, C17
SLC22A4	Q9H015	17	C04, C07, C11, C14, C16, C17, C15, C10, C13, C06, C08, C18, C19, C20, C05, C23, C01
SLC46A1	Q96NT5	17	C06, C08, C16, C18, C04, C13, C19, C01, C10, C15, C20
ANKH	Q9HCJ1	16	C16, C18, C20, C05, C13, C10, C11, C14, C17, C15, C19
SLC12A5	Q9H2X9	16	C10, C18, C04, C16, C11, F03, C13, C19
SLC16A2	P36021	16	C10, C16, C23, C04, C19, C05, C18, C20, C11, C17, C13
SLC19A2	O60779	16	C04, C15, C11, C14, C16, C18, C19, C20, C17, C09, C10, C23
SLC27A4	Q6P1M0	16	C04, C15, C20, C12, C13, C16, C17, C08, C18, C19, C10, C05
SLC2A5	P22732	16	C04, C06, C16, C12, C13, C08, C19, C17, C18
SLC35G1	Q2M3R5	16	C04, C09, C10, C23, C05, C16, C19, C17, C18, C14, C15, C20
SLC3A1	Q07837	16	C05, C10, C16, C12, C13, C04, C18, C23, F01, F03, C08, C14, C17, C19, C11

SLC7A11	Q9UPY5	16	C10, C18, C04, C12, C13, C05, C16, C19, C06, C11, C15, C17, C01
SLC9A3	P48764	16	C12, C13, C19, C01, C16, C18, C23, C06, C08, C15, C20, C05, C10, C04
SLC25A22	Q9H936	15	C04, C06, C05, C07, C16, C10, C12, C19, C23, F03, C13
SLC31A1	O15431	15	C04, C12, C13, C07, C16, C06, C14, C19, C10, C25, C01, C18
SLC34A1	Q06495	15	C04, C05, C12, C13, C16, C18, C23, C19, C01
SLC35A1	P78382	15	C10, C18, C07, C11, C14, C16, C17, C05, C04, C06, C19, C20, C01
SLC7A7	Q9UM01	15	C04, C12, C13, C16, C17, C15, C20, C10, C23, F03, C18, C05, C19, C14
SLC12A1	Q13621	14	C12, C13, C19, C04, C05, C16, C10, F03, C23, C06
SLC14A2	Q15849	14	C17, C23, C08, C16, C15, C10, F03, C04, C14, C18, C19, C11, C01
SLC20A2	Q08357	14	C04, C10, C14, C18, C05, C16, C23, C19, C17, C20
SLC25A38	Q96DW6	14	C15, C18, C10, C11, C23, C05, C16, F03, C04, C06, C19
SLC26A5	P58743	14	C05, C16, C18, C23, C19, C01, C04, C06, C10, C11, C20, C09, C14
SLC2A9	Q9NRM0	14	C09, C10, C11, C23, F03, C04, C06, C05, C07, C16, C18, C14, C15, C20, C17, C19
SLC35B2	Q8TB61	14	C04, C15, C20, C01, C05, C16, C17, C14
SLC6A9	P48067	14	C05, C16, C04, C19, C10, C18, C11, C23, C06
SLC16A7	O60669	13	C04, C06, C15, C20, C14, C17, C18, C19, C10, C12
SLC25A10	Q9UBX3	13	C15, C23, C04, C16, C06, C10, C18, C01, C13, C17
SLC7A4	O43246	13	C05, C14, C15, C16, C19, C04, C20, C11, C23, C17
NIPA1	Q7RTP0	12	C04, C10, C18, C16, C14, C15, C20, C23
SLC15A1	P46059	12	C10, C14, C16, C18, C20, C04, C17, C12, C13, C05, C06, C09, C15, C23
SLC22A16	Q86VW1	12	C04, C06, C05, C17, C20, C13, C19, C18
SLC25A46	Q96AG3	12	C10, C16, C18
SLC29A2	Q14542	12	C04, C15, C20, C16, C06, C19, C23, C18
SLC30A8	Q8IWU4	12	C04, C10, C14, C16, C18, C20, C06, C19, C15, C05
SLC34A2	O95436	12	C04, C12, C13, C06, C05, C10, C16, C08, C18, C17
SLC39A14	Q15043	12	C04, C06, C05, C10, C25, C18, C19, C16
UCP3	P55916	12	C10, C18, C04, C05, C19, C20, F03
LETMD1	Q6P1Q0	11	C04, C15, C20, C06, C17, C13, C19
MFSD2A	Q8NA29	11	C18, C05, C10, C16, C04, C06, F03, C12, C13, C23
MFSD8	Q8NHS3	11	C10, C16, C18, F03, C23, C11
SLC18A3	Q16572	11	C05, C16, C10, C04, C20, F03, C14, C11, C23
SLC23A2	Q9UGH3	11	C04, C15, C20, C06, C08, C16, C17

SLC25A13	Q9UJS0	11	C10, C16, C18, C04, C06, C17, C23, C15, C11
SLC39A4	Q6P5W5	11	C16, C17, C23, C01, C04, C06, C10, C13, C19, C07, C09
SLC39A6	Q13433	11	C04, C06, C01, C08, C16, C17, C13, C18, C19
SLC4A11	Q8NBS3	11	C11, C16, C09, C10, C23, C04, C13, C19, C17
SLC50A1	Q9BRV3	11	C04, C12, C13, C11, C16, C01, C10, C14, C18, C19, C17
SLC6A1	P30531	11	C10, C18, C16, C04, C12, C13, C01, C19, F03
SLCO1A2	P46721	11	C04, C15, C20, C10, C11, C23, C17, C18, C19, C01, C16
SLC10A1	Q14973	10	C16, C18, C10, C01, C06, C04, C23
SLC22A12	Q96S37	10	C04, C13, C16, C17, C19, C10, C18, C06, C12, C05
SLC25A5	P05141	10	C14, C04, C06, C05, C10, C18, C15, C17, C12, C13
SLC28A3	Q9HAS3	10	C04, C15, C20, C08, C11, C16, C10, F03, C06, C14
SLC39A7	Q92504	10	C15, C20, C10, C16, C18, C04, C17, C13, C19
SLC4A4	Q9Y6R1	10	C10, C11, C12, C13, C16, C18, C23, F01, F03, C04, C14, C19, C17
NPC1L1	Q9UHC9	9	C01, C04, C06, C19, C18, C16
SLC10A2	Q12908	9	C04, C16, C17, C15, C06, C23, C18, C19
SLC14A1	Q13336	9	C15, C20, C10, C11, C16, C18, C04, C19
SLC18A1	P54219	9	C08, C16, C04, C19, C17, C10, C14
SLC22A17	Q8WUG5	9	C04, C06, C15, C12, C13, C08, C11, C16, C01
SLC28A1	O00337	9	C04, C15, C20, C06, C08, C11, C16, C17, C10, C18, C19, C14
SLC30A1	Q9Y6M5	9	C01, C17, C04, C06, C13, C19, C18
SLC5A7	Q9GZV3	9	C10, C16, C05, C04, C15, C20, F03, C11, C23, C18
SLC6A5	Q9Y345	9	C07, C16, C04, C06, C05, C10, C23, C17, F03, C20
SLC7A10	Q9NS82	9	C05, C10, C12, C13, C16, C04, C17, C20, C19
SLC7A9	P82251	9	C04, C15, C20, C12, C13, C16, C01, C18, C22, C10, C19
TMEM165	Q9HC07	9	C07, C16, C05, C18, C04, C06, C09, C10
SLC25A16	P16260	8	C04, C06, C08, C19, C17
SLC26A1	Q9H2B4	8	C05, C16, C19, C17, C18, C04, C14, C15, C20, C12, C13, C06
SLC27A2	O14975	8	C10, C16, C18, C19, C04, C13, C17
SLC2A12	Q8TD20	8	C18, C04, C14, C15, C20, C19, C17, C10
SLC2A6	Q9UGQ3	8	C09, C10, C11, C23, F03, C16, C18, C04, C14, C15, C20, C17, C19
SLC35A3	Q9Y2D2	8	C05, C16, C23, C18, C04, C10, F03
SLC39A11	Q8N1S5	8	C04, C12, C13, C19, C06
SLC44A4	Q53GD3	8	C01, C04, C06, C19, C09, C10, C23, C18, C20, C15, C08
SLC45A1	Q9Y2W3	8	C10, C16, C05, C04, C23, F03
SLC45A2	Q9UMX9	8	C17, C04, C14, C15, C20, C11, C16, C18

SLC6A19	Q695T7	8	C10, C12, C13, C16, C18, C04, C06, C01
SLC7A14	Q8TBB6	8	C16, C11, C04, C18, C19
SLCO2B1	O94956	8	C04, C06, C19, C18, C10, C17
DIRC2	Q96SL1	7	C04, C12, C13, C10, C14, C16
SLC17A6	Q9P2U8	7	C10, C16, C20, C18, C04, C09, C17
SLC25A15	Q9Y619	7	C23, C11, C16, C10, C18, C15, C06
SLC27A1	Q6PCB7	7	C04, C15, C20, C16, C17, C18, C19
SLC34A3	Q8N130	7	C05, C18, C12, C13, C16, C23
SLC36A1	Q7Z2H8	7	C10, C16, C08, C23, F01, C12, C13, C18, C06, C04, C17
SLC38A2	Q96QD8	7	C14, C04, C16, C05, C10, C12, C13, C18, C19, C17
SLC44A1	Q8WWI5	7	C04, C15, C20, C06, C23, C10
SLC4A2	P04920	7	C12, C13, C16, C04, C06, C05, C23
SLC5A11	Q8WWX8	7	C17, C23, C10, C16, C18, C19, C01
SLC5A3	P53794	7	C04, C12, C13, C10, C16, C18, C19
SLC6A14	Q9UN76	7	C06, C08, C16, C11, C04, C13, C19, C17
SLC6A6	P31641	7	C04, C13, C06, C18, C19, C20, C17, C11, C16
SLC7A1	P30825	7	C04, C06, C17, C01, C19
SLC7A2	P52569	7	C04, C06, C01, C11, C16, C18, C19, C17
SLC9A9	Q8IVB4	7	C04, C05, C10, C11, C16, C23, F01, F03, C06
SV2A	Q7L0J3	7	C18, C04, C10, C16, F03
FLVCR2	Q9UPI3	6	C10, C16, C12, C13, F03, C05, C14
RHAG	Q02094	6	C04, C15, C17, C16
SLC13A2	Q13183	6	C04, C13, C19, C09, C10, C23, C17, C12
SLC25A12	O75746	6	C10, C16, C18, C04, C06, C11, F03
SLC26A6	Q9BXS9	6	C06, C08, C16, C05, C19, C04, C15, C12, C13
SLC26A9	Q7LBE3	6	C06, C08, C16, C18, C19
SLC35D1	Q9NTN3	6	C05, C07, C16, C19, C23, C10, C15
SLC39A1	Q9NY26	6	C04, C12, C13, C01, C06, C08, C16, C18, C19
SLC39A10	Q9ULF5	6	C04, C12, C13, C17, C18, C19
SLC39A13	Q96H72	6	C14, C15, C16, C17, C05, C01, C04, C07
SLC5A6	Q9Y289	6	C05, C10, C16, C04, C13, C19, C17, C11
NIPA2	Q8N8Q9	5	C10, C16, C18, C19
NIPAL4	Q0D2K0	5	C17, C23, C16
SLC12A7	Q9Y666	5	C04, C19, C16, C17, C10
SLC17A3	O00476	5	C16, C18, C19, C04, C17
SLC17A7	Q9P2U7	5	C18, C10, C16, C04
SLC17A8	Q8NDX2	5	C06, C23, C09, C10, C01
SLC17A9	Q9BYT1	5	C16, C17, C04, C06, C18, C19
SLC1A4	P43007	5	C04, C18, C20, C05, C10, C16
SLC22A10	Q63ZE4	5	C14, C16, C04, C06, C13, C19

SLC24A3	Q9HC58	5	C07, C16, C23, C04, C17
SLC24A5	Q71RS6	5	C17, C11, C16, C18
SLC25A6	P12236	5	C04, C10, C18, C17
SLC26A7	Q8TE54	5	C05, C16, C19, C12, C13, C18, C15, C10
SLC26A8	Q96RN1	5	C04, C15, C20, C16, C11, C12
SLC2A11	Q9BYW1	5	C05, C14, C16, C17, C04, C15, C20, C18, C19
SLC2A8	Q9NY64	5	C04, C14, C15, C20, C18, C19
SLC30A3	Q99726	5	C10, C18, C05, C16, C04, C19
SLC30A6	Q6NXT4	5	C10, C18, C16, F03, C19
SLC31A2	O15432	5	C01, C04, C19, C14, C15, C20, C13, C06, C10, C16, C18
SLC47A1	Q96FL8	5	C04, C15, C18, C19, C10, C17
SLC6A12	P48065	5	C06, C04, C10, C14, C16, C18, C20, C13, C19
SLC6A20	Q9NP91	5	C06, C16, C12, C13, C18, C04, C19
SLC9C1	Q4G0N8	5	C04, C06, C18, C19, C20
UNC93B1	Q9H1C4	5	C10, C01, C04, C13, C19, C07
MPC2	O95563	4	C04, C06
RHCG	Q9UBD6	4	C04, C13, C19, C06
SLC12A4	Q9UP95	4	C10, C18, C16, C04, C15
SLC15A2	Q16348	4	C06, C16, C17, C18, C08, C04
SLC16A12	Q6ZSM3	4	C04, C12, C13, C16, C18, C11
SLC22A8	Q8TCC7	4	C04, C15, C20, C10, C12, C13, C16, C18
SLC24A4	Q8NFF2	4	C07, C16, C11, C17, C18
SLC25A11	Q02978	4	C04, C06, C11, C16, C12, C13
SLC25A27	O95847	4	C10, C18, F03, C04, C17
SLC27A6	Q9Y2P4	4	C10, C16, C18, C19, C04, C17
SLC29A4	Q7RTT9	4	C04, C06, C18, C19
SLC2A13	Q96QE2	4	C01, C10, C04, C17
SLC30A7	Q8NEW0	4	C04, C06, C18, C19
SLC32A1	Q9H598	4	C05, C07, C16, C04, C06, C19, C10
SLC35C1	Q96A29	4	C16, C18, C04, C06, C05, C10
SLC38A3	Q99624	4	C14, C16, C04, C17
SLC38A8	A6NNN8	4	C11, C16, C17, C18
SLC39A2	Q9NP94	4	C01, C06, C08, C16, C04, C17
SLC7A6	Q92536	4	C04, C12, C13, C01, C17, C15, C20, C16, C18
SLC7A8	Q9UHI5	4	C13, C04, C06, C16, C18, C19
SLC8A3	P57103	4	C10, C18, C04, C17
SLC9A5	Q14940	4	C16, C18, C23, C04, C10, F03
SLC9A7	Q96T83	4	C05, C10, C16, C04, C17
SLC9C2	Q5TAH2	4	C05, C10, C11, C16, C23, F01, F03, C06, C15, C20, C04
SLCO1C1	Q9NYB5	4	C10, C16, C23, C06, C18, C19

SLCO3A1	Q9UIG8	4	C04, C06, C19, C17
SLCO5A1	Q9H2Y9	4	C10, C16, C05, C04, C17, C08
MFSD9	Q8NBP5	3	C18, C19, C04, C17
NIPAL1	Q6NVV3	3	C11, C04, C07, C09, C16
SLC13A1	Q9BZW2	3	C16, C17, C18, C19
SLC15A4	Q8N697	3	C10, C19, C06, C14
SLC16A10	Q8TF71	3	C04, C15, C20, C11, C19
SLC16A11	Q8NCK7	3	C18, C19
SLC16A8	O95907	3	C04, C10, C16, C18
SLC1A6	P48664	3	C04, C10, C16, F03
SLC1A7	O00341	3	C01, C17, C18, C19
SLC22A23	A1A5C7	3	C13, C06, C23
SLC22A6	Q4U2R8	3	C04, C15, C20, C18, C10, C12, C13, C16
SLC23A1	Q9UHI7	3	C04, C15, C20, C06
SLC24A1	O60721	3	C11, C16, C05, C10
SLC25A18	Q9H1K4	3	C04, C15, C12, C13
SLC25A33	Q9BSK2	3	C10, C16, C18
SLC25A42	Q86VD7	3	C18, C05, C10, C14
SLC2A14	Q8TDB8	3	C05, C14, C15, C16, C19, C06
SLC30A2	Q9BRI3	3	C17, C23, C01, C04
SLC30A9	Q6PML9	3	C04, C11, C23
SLC35A4	Q96G79	3	C04, C06, C14, C15, C20, C17
SLC35F2	Q8IXU6	3	C04, C19
SLC35F6	Q8N357	3	C08, C11, C16, C04, C06, C19
SLC37A2	Q8TED4	3	C05, C16, C01, C04, C17
SLC38A5	Q8WUX1	3	C04, C18, C19
SLC38A7	Q9NVC3	3	C10, C16, C18, C04, C06, C19
SLC39A12	Q504Y0	3	C18, C19, C04, C17
SLC39A3	Q9BRY0	3	C04, C06, C19, C18
SLC39A5	Q6ZMH5	3	C04, C06, C11
SLC41A1	Q8IVJ1	3	C12, C13, C18, C19
SLC43A1	O75387	3	C04, C12, C13, C15, C20
SLC4A7	Q9Y6M7	3	C06, C08, C16, C18, C04, C17
SLC6A7	Q99884	3	C01, C05, C11, C16
SLC9A2	Q9UBY0	3	C16, C18, C23, C06, C04, C17
SLC9A8	Q9Y2E8	3	C16, C18, C23, C12, C06

SLCO4A1	Q96BD0	3	C04, C06, C19, C17, C08
SPNS1	Q9H2V7	3	C10, C16, C15, C18, C23
SV2B	Q7L1I2	3	C04, C10, C16, F03
SV2C	Q496J9	3	C10, C16, F03, C04, C17
MFSD4A	Q8N468	2	C04, C06
SFXN4	Q6P4A7	2	C18
SLC10A6	Q3KNW5	2	C04, C17
SLC10A7	Q0GE19	2	C07, C16, C05
SLC15A3	Q8IY34	2	C01, C04
SLC16A13	Q7RTY0	2	C18, C19
SLC16A6	O15403	2	C13, C04, C17
SLC22A7	Q9Y694	2	C04, C06, C18, C19
SLC25A23	Q9BV35	2	C04, C11
SLC25A31	Q9H0C2	2	C01, C12
SLC25A43	Q8WUT9	2	C04, C17
SLC25A44	Q96H78	2	C04, C12, C19
SLC25A45	Q8N413	2	C04, C13, C19, C06
SLC25A51	Q9H1U9	2	C04, C06, C17
SLC26A11	Q86WA9	2	C04, C15, C16, C17, C18
SLC27A3	Q5K4L6	2	C04
SLC30A4	O14863	2	C16, C17, C10, F03
SLC30A5	Q8TAD4	2	C18, C19
SLC35B3	Q9H1N7	2	C18, C19
SLC35B4	Q969S0	2	C04, C06
SLC35D3	Q5M8T2	2	C01, C11, C15, C16, C17, C18
SLC35F1	Q5T1Q4	2	C04, C10
SLC35G2	Q8TBE7	2	C04, C12, C13
SLC37A1	P57057	2	C18, C04, C17
SLC39A9	Q9NUM3	2	C04, C17
SLC41A2	Q96JW4	2	C18, C19
SLC4A10	Q6U841	2	C16, C10
SLC4A3	P48751	2	C10, C23
SLC4A9	Q96Q91	2	C04, C17, C08
SLC51A	Q86UW1	2	C06, C23, C05, C11, C16
SLC6A11	P48066	2	C10, C16
SLC6A13	Q9NSD5	2	C01, C10

SLC7A13	Q8TCU3	2	C12, C13, C16, C04
SLC8A2	Q9UPR5	2	C04
SLC8B1	Q6J4K2	2	C18, C10
SLC9B2	Q86UD5	2	C18, C19
SLCO4C1	Q6ZQN7	2	C18, C19
SPNS2	Q8IVW8	2	C04, C06, C19, C18
TMEM241	Q24JQ0	2	C04, C10, C16
LETM2	Q2VYF4	1	C04, C06
MFSD1	Q9H3U5	1	C18, C19
MFSD12	Q6NUT3	1	C04
MFSD13A	Q14CX5	1	C04, C06, C19
MFSD14A	Q96MC6	1	C04, C12, C13
MFSD4B	Q5TF39	1	C18, C19, C20
MTCH2	Q9Y6C9	1	C04, C06
SLC12A8	A0AV02	1	C04, C17
SLC16A5	O15375	1	C04
SLC17A1	Q14916	1	C05, C12, C13, C16, C18
SLC22A13	Q9Y226	1	C04, C12, C13
SLC22A9	Q8IVM8	1	C04, C17
SLC25A2	Q9BXI2	1	C10, C16, C18, C23
SLC25A26	Q70HW3	1	
SLC25A28	Q96A46	1	C10, C16, F03
SLC25A29	Q8N8R3	1	C10, C16, C18, C23
SLC25A32	Q9H2D1	1	
SLC25A36	Q96CQ1	1	C04, C06, C19
SLC25A41	Q8N5S1	1	C04, C17
SLC25A47	Q6Q0C1	1	C04, C06
SLC25A52	Q3SY17	1	C04, C17
SLC28A2	O43868	1	C15
SLC2A7	Q6PXP3	1	C06, C16
SLC35F3	Q8IY50	1	C04
SLC35F5	Q8WV83	1	C14, C15, C17
SLC36A2	Q495M3	1	C12, C13, C16, C18
SLC37A3	Q8NCC5	1	C11, C16

SLC38A4	Q969I6	1	C04, C06
SLC38A6	Q8IZM9	1	C04, C17
SLC38A9	Q8NBW4	1	C04, C06
SLC43A2	Q8N370	1	C04, C12, C13
SLC43A3	Q8NBI5	1	C04
SLC44A2	Q8IWA5	1	C04, C06
SLC44A3	Q8N4M1	1	C10
SLC45A3	Q96JT2	1	C04
SLC46A2	Q9BY10	1	C04, C13
SLC46A3	Q7Z3Q1	1	C04, C06
SLC48A1	Q6P1K1	1	C04, C17
SLC4A5	Q9BY07	1	C10, C11, C16
SLC51B	Q86UW2	1	C06, C23
SLC5A4	Q9NY91	1	C04, C08
SLC6A16	Q9GZN6	1	C01
SLC6A17	Q9H1V8	1	
SLC6A18	Q96N87	1	C12, C13, C16, C18
SLC7A3	Q8WY07	1	C10
SLC9B1	Q4ZJI4	1	C01
SLCO1B7	G3V0H7	1	C01
UNC93A	Q86WB7	1	C04, C13, C19
MFSD10	Q14728	0	
MFSD11	O43934	0	
MFSD14B	Q5SR56	0	
MFSD2B	A6NFX1	0	
MFSD3	Q96ES6	0	
MFSD5	Q6N075	0	
MFSD6	Q6ZSS7	0	
MFSD6L	Q8IWD5	0	
MPC1L	P0DKB6	0	
MTCH1	Q9NZJ7	0	
NIPAL2	Q9H841	0	
NIPAL3	Q6P499	0	
RHBG	Q9H310	0	
SFXN2	Q96NB2	0	
SFXN3	Q9BWM7	0	
SFXN5	Q8TD22	0	
SLC10A3	P09131	0	
SLC10A4	Q96EP9	0	
SLC10A5	Q5PT55	0	

SLC13A3	Q8WWT9	0
SLC13A4	Q9UKG4	0
SLC15A5	A6NIM6	0
SLC16A14	Q7RTX9	0
SLC16A9	Q7RTY1	0
SLC17A2	O00624	0
SLC17A4	Q9Y2C5	0
SLC18B1	Q6NT16	0
SLC22A11	Q9NSA0	0
SLC22A14	Q9Y267	0
SLC22A15	Q8IZD6	0
SLC22A24	Q8N4F4	0
SLC22A25	Q6T423	0
SLC22A31	A6NKX4	0
SLC23A3	Q6PIS1	0
SLC24A2	Q9UI40	0
SLC25A14	O95258	0
SLC25A17	O43808	0
SLC25A25	Q6KCM7	0
SLC25A30	Q5SVS4	0
SLC25A34	Q6PIV7	0
SLC25A35	Q3KQZ1	0
SLC25A39	Q9BZJ4	0
SLC25A40	Q8TBP6	0
SLC25A48	Q6ZT89	0
SLC25A53	Q5H9E4	0
SLC26A10	Q8NG04	0
SLC35A5	Q9BS91	0
SLC35B1	P78383	0

SLC35C2	Q9NQQ7	0
SLC35D2	Q76EJ3	0
SLC35E1	Q96K37	0
SLC35E2A	P0CK97	0
SLC35E2B	P0CK96	0
SLC35E3	Q7Z769	0
SLC35E4	Q6ICL7	0
SLC35F4	A4IF30	0
SLC35G3	Q8N808	0
SLC35G4	P0C7Q5	0
SLC35G5	Q96KT7	0
SLC35G6	P0C7Q6	0
SLC36A3	Q495N2	0
SLC36A4	Q6YBV0	0
SLC38A10	Q9HBR0	0
SLC38A11	Q08AI6	0
SLC41A3	Q96GZ6	0
SLC44A5	Q8NCS7	0
SLC45A4	Q5BKX6	0
SLC47A2	Q86VL8	0
SLC49A3	Q6UXD7	0
SLC4A8	Q2Y0W8	0
SLC5A10	A0PJK1	0
SLC5A12	Q1EHB4	0
SLC5A9	Q2M3M2	0
SLC6A15	Q9H2J7	0
SLC9A4	Q6AI14	0
SPNS3	Q6ZMD2	0
SVOP	Q8N4V2	0
SVOPL	Q8N434	0
TMEM104	Q8NE00	0