**Documentation to Supplementary Information of Dissertation
"Improving compound synthesis efficiency through laboratory automation and artificial intelligence"**

**David F. E. Nippa**

**2024**

**(1) Data:**

Download all files from the repository and build a folder structure with the following directory:

**\data\_package\literature\_analysis** - all files covering the literature analysis.

*literature\_analysis.yxmd* (Alteryx Designer Workflow)

*literature\_analysis.xml* (Alteryx Designer Workflow in .xml format)

**\data\_package\literature\_analysis\database\_download** - all downloaded files from the databases.

*literature\_analysis\_database\_download\_scifinder.xlsx*, *literature\_analysis\_database\_download\_scopus.xlsx*, *literature\_analysis\_database\_download\_webofscience.xlsx*,

*literature\_analysis\_database\_download\_scopus\_full.csv*,
*literature\_analysis\_database\_download\_webofscience\_full.xlsx***\data\_package\literature\_analysis\output\_data** - file output folder from the workflow.

*literature\_analysis\_output\_data\_cleaned\_literature\_data.csv* was added to the repository to compare the results of the output

**\data\_package\literature\_analysis\surf\_file** - location for the SURF file.

As this file was manually curated, no workflow is related to this file. It is solely an example SURF data set. For further SURF data sets, see: <https://github.com/alexarnimueller/surf/tree/main/data>.

**\data\_package\data\_structure** - all files to build the database structure (see Figure 3.8)

**\data\_package\data\_structure\compound\_data** - all files containing the compound data.

*data\_structure\_compound\_data\_analytics.csv*

*data\_structure\_compound\_data\_chemicals.csv*

**\data\_package\data\_structure\experiments** - all files containing the experiment information.

*data\_structure\_requests.csv
data\_structure\_substrate.csv*

*data\_structure\_experiments.csv*

*data\_structure\_process\_controls.csv*

**\data\_package\data\_structure\experiments** - all files covering the plate definition.

*data\_structure\_plates.csv
data\_structure\_products.csv*

**\data\_package\lcms\_analysis** - all files covering the lcms reaction data analysis to generate visualization and SURF files.

*lcms\_analysis.yxmd (Alteryx Designer Workflow)*

*lcms\_analysis.xml (Alteryx Designer Workflow in .xml format)*

**\data\_package\lcms\_analysis\01\_rpt\_file** - location of an exemplary .rpt file (lcms output).

*lcms\_analysis\_rpt\_file\_MSA-22-00984.rpt*

No workflow is related to this file. The output files of the Roche confidential parsing script of this .rpt file, which is needed for further analysis can be found in \data\_package\lcms\_analysis\02\_structured\_lcms\_data.

**\data\_package\lcms\_analysis\02\_structured\_lcms\_data** - location of structured lcms data files.

*lcms\_analysis\_structured\_lcms\_data\_channels.csv*

*lcms\_analysis\_structured\_lcms\_data\_masses.csv*

*lcms\_analysis\_structured\_lcms\_data\_sample\_id.csv*

**\data\_package\lcms\_analysis\03\_cleaned\_lcms\_data** - location of file with cleaned lcms data (one line represents one peak in the lcms spectrum and includes all corresponding masses).

*lcms\_analysis\_cleaned\_lcms\_data.csv*

No workflow is related to this file. The file serves as an example of how the LCMS data can be cleaned to a tabular format containing the information of one peak in one row.

**\data\_package\lcms\_analysis\04\_output\_data** - file output folder from the workflow.

*lcms\_analysis\_output\_data\_surf.tsv
lcms\_analysis\_output\_data\_visualization\_1.tsv*

*lcms\_analysis\_output\_data\_visualization\_2.tsv*

*lcms\_analysis\_output\_data\_visualization\_3.tsv*

*lcms\_analysis\_output\_data\_visualization\_4.tsv*

*lcms\_analysis\_output\_data\_visualization\_5.tsv*

The six files were added to the repository to compare the results of the workflow output.

**\data\_package\potential\_products** - workflow to generate potential products and file output.

*pot\_products.yxmd (Alteryx Designer Workflow)*

*pot\_products.xml (Alteryx Designer Workflow in .xml format)*

The output files was added to the repository to compare the results of the workflow output.

potential\_products\_output.csv

**(2) General notes:**

Within the Alteryx workflows, each data transformation step is meticulously annotated. For operations with high complexity, a concise description is embedded within the annotation field of the corresponding tool to elucidate its function. Prior to execution, it is imperative to establish a coherent folder structure (see **1**), which necessitates the adjustment of the input file paths within the Alteryx workflows to align with the established directory hierarchy.

The adjustment of input locations is streamlined by the predefined folder structure; the file paths necessitate only the prefixing of the primary directory of the 'data\_package' folder (*e.g.*, C:\Users\Desktop\data\_package...). Consistency is maintained across all Alteryx inputs, which are uniformly labelled in the format "input: file\_type", where "file\_type" is indicative of the specific data source, such as "input: scifinder" for the SciFinder literature download.

Employing this standardized approach ensures that the workflows are compatible with the exemplary data sets provided. After the workflow execution, the resultant output files can be systematically compared against the benchmark files housed within the repository for validation purposes.

**(3) Literature analysis workflow:**

The methodology described in the dissertation (Chapter 3.2.1, Figure 3.4) outlines a literature analysis workflow (*literature\_analysis.yxmd*) that integrates three primary data sources: SciFinder, Reaxys, and Web of Science. These sources provide initial data sets comprising solely the Digital Object Identifiers (DOIs) of the retrieved records. Augmentation of these datasets is achieved through the incorporation of two comprehensive files from Scopus and Reaxys, which contain exhaustive details of the records (denoted as scopus\_full and webofscience\_full, respectively).

The workflow is designed to refine the literature data set by executing a de-duplication process and subsequently enriching the dataset with additional bibliometric information obtained from Scopus and Web of Science, such as citation counts and publication years. The culmination of this process yields an output file that encapsulates the curated literature data. This file serves as a foundation for conducting an in-depth and systematic review of the literature.

The provided files cover a literature search for Minisci-type alkylation reactions.

**(4) Database structure:**

The files contained within the database structure folder are provided as illustrative exemplars, corresponding to the platform architecture delineated in the dissertation (refer to Chapter 3.2.4 and Section 3.8). These eight representative files encapsulate data of a single Minisci-type alkylation 24-well library screening, which was conducted as part of the case study presented in Chapter 6.2 of the thesis. These files are requisite for the execution of the potential product (**5**) and the liquid chromatography-mass spectrometry (LCMS) reaction data analysis (**6**) workflows.

The files proffered serve as a template, enabling users to tailor and augment the platform to meet their specific research requirements. The platform is designed for collaborative use; when files are uploaded and merged within a single Google Sheet, multiple contributors can operate the system in parallel.

**(5) Potential products workflow:**

The workflow (*pot\_products.yxmd*) as described in Chapter 3.2.5 (Figure 3.10) facilitates the automated derivation of expected chemical formulae and corresponding masses for potential products based on the reactants and the specified chemical transformation. This process requires defined data sets from the data structure (**4**), encompassing details from the experiments, substrate, chemicals, and product tabs. Utilizing this dataset, the workflow modifies the chemical formula of the starting material to enumerate the potential products, which are subsequently cataloged in the output file.

The workflow will generate potential products for the Minisci-type alkylation 24-well library screening, as detailed in the case study of Chapter 6.2 of the dissertation. The output file containing these potential products is needed for the operation of the LCMS reaction data analysis workflow (**6**), which necessitates the potential product data for accurate analysis. Modifications to the input files (**4**) can be implemented to facilitate the generation of potential products for a diverse array of chemical reactions.

**(6) LCMS reaction data analysis workflow:**

The LCMS analysis workflow (lcms\_analysis.yxmd) is designed to process structured LCMS data files (data\_channels.csv, data\_masses.csv, sample\_id.csv) into five visualization files for Spotfire analysis, as well as into a Simple User-Friendly Reaction Format (SURF) file. This process is elaborated in Chapter 3.5.2 and depicted in Figure 3.11 of the dissertation. Execution of the workflow necessitates not only the structured LCMS data but also the potential product file generated by (**5**) and the associated files from (**4**). These files must be interconnected as per the instructions detailed within the workflow, analogous (**3**) and (**5**). Workflow execution will then produce the aforementioned output files covering the outcome of the Minisci-type alkylation reaction example provided through the exemplary data set.

Disclaimer: The conversion of the .rpt file, which is the output format from Waters (Milford, US) LCMS systems as referenced in Figure 3.11, into structured LCMS data files is conducted on an internal server infrastructure. The specifics of this infrastructure are proprietary and cannot be disclosed. Additionally, users may encounter varying LCMS output formats when using other machines. Consequently, the transformation of raw LCMS data into the requisite structured format is a prerequisite step that must be completed before initiating the workflow.