Planning for data analysis

In this task, you will be constructing a plan for a project that strongly relies on data analysis (machine learning). During this course, we will be mostly using examples from non-targeted LC/HRMS. Non-targeted screening aims to detect and identify as many compounds as possible important chemicals from the samples of interest. Both detection and identification are complicated. The identification relies on predicting the LC/HRMS properties of the compound from the potential structure and comparing the predicted property and the measured property. In such a way retention time in LC, exact mass, fragmentation in MS/MS as well as collision cross-sections from ion mobility can be used to aid compound identification even if such properties have not been previously measured for these compounds.

Retention times

You are working in the field of metabolomics and you use non-targeted LC/HRMS to detect small molecules that might possess an important role in metabolomic cycles. You already use database matching of MS/MS spectra to identify the metabolites. However, you often end up with multiple matches. As you run LC/HRMS you also collect information about the retention time of the metabolites. Therefore, you would be interested in using also retention times to confirm or rule out the proposed structures.

You usually run a standard workflow that most metabolomics labs use: reversed-phase chromatography (RP) with water-acetonitrile (both contain 0.1% of formic acid) gradient. The samples are run alternatingly in positive and negative mode.

Your plan should consist of:

1. What is the purpose and how does it affect the plan?
2. Which measurements need to be conducted?
    a. Which samples/compounds/… do you choose? How? Is any data science needed here?
    b. How many samples/compounds/… do you need to choose?
    c. How to conduct the measurements? Are there any instrumental aspects that need to be considered?
    d. Is time an issue here?
3. In case you do not need to do measurements yourself, which data sources you will use?
4. How to prepare the data for the analysis?
    a. Any preprocessing required? Noise filtrations? Data quality measures?
    b. Think about, how to evaluate that the model is performing as intended (train-test-validate).
5. Which type of data science do you need?
    a. Is it a classification, clustering, regression, or any other task?
    b. Are there any limitations to the models you can use (interpretability vs best performance)? Which? Why? Think about the purpose of the methods.
6. How is the model used in real life?
    a. Is it an in-house model? Or can others use it? How can others access your model?

Planning for data analysis

In this task, you will be constructing a plan for a project that strongly relies on data analysis (machine learning). During this course, we will be mostly using examples from non-targeted LC/HRMS. Non-targeted screening aims to detect and identify as many compounds as possible important chemicals from the samples of interest. Both detection and identification are complicated. The identification relies on predicting the LC/HRMS properties of the compound from the potential structure and comparing the predicted property and the measured property. In such a way retention time in LC, exact mass, fragmentation in MS/MS as well as collision cross-sections from ion mobility can be used to aid compound identification even if such properties have not been previously measured for these compounds.

MS/MS spectra

You are working in the field of metabolomics and you use non-targeted LC/HRMS to detect small molecules that might possess an important role in metabolomic cycles. You already use database matching of MS/MS spectra to identify the metabolites. However, a significant part of the compounds that you detect does not have a good match in the database. This is likely to results from the fact that these compounds have simply not been measured by the curator of the library. You are a keen problem solver so you would like to still use the measured MS/MS spectra to unravel the structure of these unknown-unknowns in a scalable manner. So you need to build/develop an algorithm that you could use to gain insight into the structure. You have access to the database of MS/MS spectra with respective chemical structures (~50 000 compounds).

Your plan should consist of:

1. What is the purpose and how does it affect the plan?
2. Which measurements need to be conducted?
   a. Which samples/compounds/… do you choose? How? Is any data science needed here?
   b. How many samples/compounds/… do you need to choose?
   c. How to conduct the measurements? Are there any instrumental aspects that need to be considered?
   d. Is time an issue here?
3. In case you do not need to do measurements yourself, which data sources you will use?
4. How to prepare the data for the analysis?
   a. Any preprocessing required? Noise filtrations? Data quality measures?
   b. Think about, how to evaluate that the model is performing as intended (train-test-validate).
5. Which type of data science do you need?
   a. Is it a classification, clustering, regression, or any other task?
   b. Are there any limitations to the models you can use (interpretability vs best performance)? Which? Why? Think about the purpose of the methods.
6. How is the model used in real life?
   a. Is it an in-house model? Or can others use it? How can others access your model?

Planning for data analysis

In this task, you will be constructing a plan for a project that strongly relies on data analysis (machine learning). During this course, we will be mostly using examples from non-targeted LC/HRMS. Non-targeted screening aims to detect and identify as many compounds as possible important chemicals from the samples of interest. Both detection and identification are complicated. The identification relies on predicting the LC/HRMS properties of the compound from the potential structure and comparing the predicted property and the measured property. In such a way retention time in LC, exact mass, fragmentation in MS/MS as well as collision cross-sections from ion mobility can be used to aid compound identification even if such properties have not been previously measured for these compounds. The identification runs parallel with figuring out which of the detected compounds are important.

Toxicity

You are working in the field of exposomics and your task is to identify and assess the importance of the detected contaminants. You have a sample bank of blood samples from kids collected over the last 20 years and you have analyzed all these samples in one LC/HRMS sequence in a randomized manner. Now when you have all the raw data at hand you have two tasks: (1) what are the important chemicals (signals from LC/HRMS) that need to be identified; and (2) how toxic these chemicals could be? For the second task please assume that most of these chemicals do not have LC_{50} or similar values available but some of the fairly similar compounds may have.

Your plan should consist of:

1. What is the purpose and how does it affect the plan?
2. Which measurements need to be conducted?
   a. Which samples/compounds/… do you choose? How? Is any data science needed here?
   b. How many samples/compounds/… do you need to choose?
   c. How to conduct the measurements? Are there any instrumental aspects that need to be considered?
   d. Is time an issue here?
3. In case you do not need to do measurements yourself, which data sources you will use?
4. How to prepare the data for the analysis?
   a. Any preprocessing required? Noise filtrations? Data quality measures?
   b. Think about, how to evaluate that the model is performing as intended (train-test-validate).
5. Which type of data science do you need?
   a. Is it a classification, clustering, regression, or any other task?
   b. Are there any limitations to the models you can use (interpretability vs best performance)? Which? Why? Think about the purpose of the methods.
6. How is the model used in real life?
   a. Is it an in-house model? Or can others use it? How can others access your model?

Planning for data analysis

In this task, you will be constructing a plan for a project that strongly relies on data analysis (machine learning). During this course, we will be mostly using examples from non-targeted LC/HRMS. Non-targeted screening aims to detect and identify as many compounds as possible important chemicals from the samples of interest. Both detection and identification are complicated. The identification relies on predicting the LC/HRMS properties of the compound from the potential structure and comparing the predicted property and the measured property. In such a way retention time in LC, exact mass, fragmentation in MS/MS as well as collision cross-sections from ion mobility can be used to aid compound identification even if such properties have not been previously measured for these compounds. The identification runs parallel with figuring out which of the detected compounds are important.

Toxicity

You are working in the field of exposomics and your task is to identify and assess the importance of the detected contaminants. You have analyzed a set of blood samples from kids collected over the last 20 years. You have identified that a big portion of these compounds are metabolites of PCBs, namely, hydroxylated polychlorinated biphenyls (OH-PCBs). Now you need to assess the toxicological importance of these compounds and you would need to consider both the quantity as well as intrinsic toxicity of the detected compounds. Assume that most of these chemicals do not have LC_{50} or similar values available in the databases and you also do not have exactly these compounds to quantify them with calibration graphs. However, some of the previously known metabolites of PCBs with fairly similar structures may be available in the lab, and also some LC_{50} values may be available in databases.

Your plan should consist of:

1. What is the purpose and how does it affect the plan?
2. Which measurements need to be conducted?
   a. Which samples/compounds/… do you choose? How? Is any data science needed here?
   b. How many samples/compounds/… do you need to choose?
   c. How to conduct the measurements? Are there any instrumental aspects that need to be considered?
   d. Is time an issue here?
3. In case you do not need to do measurements yourself, which data sources you will use?
4. How to prepare the data for the analysis?
   a. Any preprocessing required? Noise filtrations? Data quality measures?
   b. Think about, how to evaluate that the model is performing as intended (train-test-validate).
5. Which type of data science do you need?
   a. Is it a classification, clustering, regression, or any other task?
   b. Are there any limitations to the models you can use (interpretability vs best performance)? Which? Why? Think about the purpose of the methods.
6. How is the model used in real life?
   a. Is it an in-house model? Or can others use it? How can others access your model?