

# Development of an R Toolbox for Near-Infrared Spectroscopy Data Processing and Analysis of Plant Metabolic Phenotypes

DEGGENDORF INSTITUTE OF TECHNOLOGY

MSC. LIFE SCIENCE INFORMATICS

*Methun George*

Supervised by  
PD Dr. habil. rer. nat. Steffen Neumann  
Prof. Dr. Melanie Kappelmann-Fenzl

October 28, 2024

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Related Work . . . . .	3
<b>2</b>	<b>Background</b>	<b>4</b>
2.1	Near Infrared Spectroscopy (NIRS) . . . . .	4
2.1.1	Introduction . . . . .	4
2.2	Metabolomics . . . . .	4
2.2.1	Introduction . . . . .	4
2.2.2	Mass Spectrometry . . . . .	4
2.3	Machine Learning . . . . .	4
2.3.1	Introduction . . . . .	4
2.3.2	Partial Least Square Regression (PLS) . . . . .	4
2.3.3	Random Forest (RF) . . . . .	4
2.3.4	Convolutional Neural Network (CNN) . . . . .	4
<b>3</b>	<b>Methods and Implementation</b>	<b>5</b>
3.1	MSNovelist . . . . .	6
3.1.1	Basic Idea of the Software . . . . .	6
3.1.2	Deep Learning Architecture . . . . .	6
3.1.3	Preparation of Data . . . . .	6
3.2	Data . . . . .	6
3.2.1	Data Records . . . . .	6
3.2.2	File Formats . . . . .	6
3.3	Evaluation of the Training . . . . .	6
3.4	Implementations in the TensorFlow Framework . . . . .	6
3.4.1	Preprocessing of Data . . . . .	6
3.4.2	Training . . . . .	6
3.4.3	Evaluation . . . . .	6
3.4.4	Termination Criterion . . . . .	6
3.4.5	Training on the Pubchem Dataset . . . . .	6
3.5	Implementations in the PyTorch Framework . . . . .	6
3.5.1	Reimplementation of the LSTM Architecture . . . . .	6
3.5.2	Training with DeepSMILES and SELFIES . . . . .	6
3.5.3	Transformer Implementation . . . . .	6
3.6	Software . . . . .	6
3.7	Hardware . . . . .	6
<b>4</b>	<b>Results and Discussion</b>	<b>7</b>
4.1	Baseline Model . . . . .	7
4.1.1	Termination Criterion . . . . .	7
4.1.2	Impact of the Amount of Data . . . . .	7

4.1.3	Duplicates in the Fingerprints . . . . .	7
4.2	Alternative Tokenization of SMILES Sequences . . . . .	7
4.2.1	QBMG Tokenization . . . . .	7

# Chapter 1

## Introduction

Near infrared spectroscopy (NIRS) has become a valuable tool in plant science due to its non destructive nature and ability to provide detailed information on plant metabolomics.

### 1.1 Related Work

The related work include, the fillowings

# Chapter 2

## Background

The background of this study include

### 2.1 Near Infrared Spectroscopy (NIRS)

.....

.....

#### 2.1.1 Introduction

### 2.2 Metabolomics

#### 2.2.1 Introduction

#### 2.2.2 Mass Spectrometry

### 2.3 Machine Learning

#### 2.3.1 Introduction

#### 2.3.2 Partial Least Square Regression (PLS)

#### 2.3.3 Random Forest (RF)

#### 2.3.4 Convolutional Neural Network (CNN)



# Chapter 3

## Methods and Implementation

### 3.1 MSNovelist

#### 3.1.1 Basic Idea of the Software

#### 3.1.2 Deep Learning Architecture

#### 3.1.3 Preparation of Data

### 3.2 Data

#### 3.2.1 Data Records

#### 3.2.2 File Formats

### 3.3 Evaluation of the Training

### 3.4 Implementations in the TensorFlow Framework

#### 3.4.1 Preprocessing of Data

#### 3.4.2 Training

#### 3.4.3 Evaluation

#### 3.4.4 Termination Criterion

#### 3.4.5 Training on the Pubchem Dataset

### 3.5 Implementations in the PyTorch Framework

#### 3.5.1 Reimplementation of the LSTM Architecture

#### 3.5.2 Training with DeepSMILES and SELFIES

#### 3.5.3 Transformer Implementation

### 3.6 Software

### 3.7 Hardware

# Chapter 4

## Results and Discussion

### 4.1 Baseline Model

#### 4.1.1 Termination Criterion

#### 4.1.2 Impact of the Amount of Data

#### 4.1.3 Duplicates in the Fingerprints

### 4.2 Alternative Tokenization of SMILES Sequences

#### 4.2.1 QBMG Tokenization