NirsDB Application Manual

NirsDB is a Shiny application designed to allows users to make predictions about phenotypic traits of A. *thaliana* by submitting related NIRS spectrum which will be treated by deep-learning algorithms. It also contains a database built from 5325 unique spectrum and 81 trait measurements across *Arabidopsis thaliana* plants grown in various condition from published and unpublished datasets. It is available at the following URL : https://shiny.cefe.cnrs.fr/NirsDB/

This application has three functionalities described below.

1. Obtain NIRS-based predictions of phenotypic traits

The first functionality of NirsDB ("Submit your spectrum" tab) is to submit your data set to get trait of interest's predictions. To do so it's possible to choose among three operating modes according to your needs that we'll discuss further. Depending on the chosen mode, one or more file must be upload. Then, phenotypic traits of interest can be chosen among two lists, one for the functional traits and one for the metabolites. Finally, to launch the job a valid email address must be provided and the Run button can be hit. Once the job is complete an email will be sent to this address with all the results.

- The first mode ("Predictions using our models") allows to use trait's specific models already trained on the NirsDB data set, then only spectrum data is needed to be submitted to the algorithm. A csv file must be prepared containing all spectrum data corresponding to a wavelength from 350 to 2500 with headers. Each individuals should be represented by 2151 value of absorbance in line corresponding to a wavelength in column with a header. If this wavelength range can't be matched the dataset will be re-sample to match the fited model, only a minimum of 400 absorbance values will be accepted. A file with missing values will not be accepted, no data completion or gap-filling method is performed in the application. The csv file can then be uploaded by clicking the "Browse" button in "upload spectrum CSV file" section.

- The second mode ("Build your own model + Predictions") allows to create and use new models trained on the provided data set. As for the first mode a Spectrum csv file must be supplied but moreover a second file containing traits value should be prepared. It is mandatory to have as many values in the spectrum file as in the trait file, and the row have to correspond to each other. it's possible to make predictions for only one trait or multiple ones, the traits file can contain from one to several columns. The file should be organized in columns whose headings correspond to the trait names spelt out exactly as shown in the selection list for Functional traits / Metabolites. Trait values should be expressed in the units shown in Table 1. Sugars have to be expressed in µmol/gFW , hormones in ng/gFW, for glucosinolates and secondary metabolites the unit used is foliar relative concentration. Like the first file no missing values will be accepted. The csv file can then be uploaded by clicking the "Browse" button in "upload traits CSV file" section. Traits submitted in this file absolutely must be also selected in the functional traits / metabolites lists before launching the job.

- The third mode ("Complete, Test dataset needed") also allows to create and use new models trained on the provided data set but with more precision because, in addition to the training data set, a test data set must be provide. The spectrum test file must be prepared following the same rules as the spectrum train file and the trait test file must be prepared following the same rules as the trait train file. Concerning the traits files, the test file must contain exactly

the same traits in the same order as the train one. Like the second mode traits submitted in those files have to be selected in the functional traits/ metabolites lists.

Table 1 Database's functional traits.

Traits	Definition	Abbreviation	Unit
Leaf dry matter content	ratio of leaf dry mass to fresh mass	LDMC	mg g-1
Specific leaf area	ratio of leaf area to leaf dry mass	SLA	mm^2 mg-1
Leaf nitrogen content	ratio of nitrogen in leaf	LNC	%
Leaf thickness	thickness of leaf	Thickness	µ m
Relative water content	amount of water in leaf	RWC	%
Leaf carbon content	ratio of carbon in leaf	LCC	%
Fraction of 13C isotope	ratio 13C /12C	Delta13C	
Fraction of 15N isotope	ratio 15N/14N	Delta15N	
Plant lifespan	plant lifetime	Plant_lifespan	days
Plant growth rate	relative increase in leaf area over time	Plant_growth_rate	mg d-1
C score	C score	Csr_c	%
R score	R score	Csr_s	%
S score	S score	Csr_r	%

2. Consult and get data from the database

The second functionality of NirsDB ("Consult Database" tab) allows to consult the database used by the deep-learning algorithms to make predictions, the search can be specify through several filters. Moreover graphic analysis of the filtered sample will be printed with a mean comparison and a PCA and finally the result of the search can be download among multiple formats.
The results of the search will be downloadable as a Csv file but the content of it depends on the output format chosen. This database is made from A. *thaliana* data used to build and calibrate deep learning algorithms (Vasseur et al. 2022), each of the 5325 individuals are associated to spectra values from a wavelength of 350 to 2500 as well as several sample's information and conditions and finally 81 traits values.

First choice ("All Data") allows to get spectrum values and traits values of the individuals corresponding to the chosen filters. Second one ("Spectrum only") allows to get only spectrum values while third one ("Phenotypic traits only") allows to get only traits values. Finally last choice ("Custom") allows to choose which traits should appear in the downloadable result file.

Once the filter are established and the output format is chosen the submit button should be hit to visualize the graphics analysis of the corresponding sample after a short loading time, then the download button will appear at the bottom of the page to get the expected data set.

The chosen sample is visualized as (a) as a curve of absorption of sample spectrum's mean compared to the curve of absorption of all data set spectrum's mean; and (b) a principal component analysis of sample's spectrum values compared to another PCA of all database spectrum values.

3. Contribute to the database

The third functionality of NirsDB (“Become Contributor” tab) allows the user to submit his own data set by uploading it to the administrators of NirsDB to maybe integrate it in the database. Same as in the first mode, a csv file must be prepared containing all spectrum data corresponding to a wavelength from 350 to 2150 with headers. Each individuals should be represented by 2151 value of absorbance in line corresponding to a wavelength in column with a header. There is no size constraint in term of spectrum numbers.

Once the file is ready, it can be uploaded by clicking the “Browse” button in “upload CSV file” section. An email address have to be provided so the owner of the submitted data set can be contacted for further information by the NirsDB administrators. The process is finalized when the “Send” button is clicked an a confirmation modal appear on screen.

Table 2 Database’s calibration coverage per traits.

Traits	Number of calibration data (associated spectra)	Ratio	Coverage percentage
csr_c	2902	0,54	54,5
csr_s	2472	0,46	46,42
csr_r	2737	0,51	51,4
plant_lifespan	1398	0,26	26,25
sla	3399	0,64	63,83
ldmc	2836	0,53	53,26
delta13c	1218	0,23	22,87
delta15n	1170	0,22	21,97
lcc	1905	0,36	35,77
thickness	2513	0,47	47,19
plant_growth_rate	700	0,13	13,15
rwc	1285	0,24	24,13
lnc	1958	0,37	36,77
Hormones	$\mu = 144,8$	$\mu = 0,027$	$\mu = 2,72$
Sugars	$\mu = 117,3$	$\mu = 0,022$	$\mu = 2,2$
Glucosinolates	$\mu = 151,5$	$\mu = 0,028$	$\mu = 2,85$
Secondary metabolites	$\mu = 155,47$	$\mu = 0,029$	$\mu = 2,92$