

~Supplementary notes~

NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses

Shisheng Wang¹, Wenzhe Li², Liqiang Hu¹, Jingqiu Cheng¹, Hao Yang^{1,*} and Yansheng Liu^{2,3,*}

¹ West China-Washington Mitochondria and Metabolism Research Center; Key Lab of Transplant Engineering and Immunology, MOH, West China Hospital, Sichuan University, Chengdu, 610041, China

² Yale Cancer Biology Institute, Yale University, West Haven, CT, 06516, USA

³ Department of Pharmacology, Yale University School of Medicine, New Haven, CT, 06520, USA

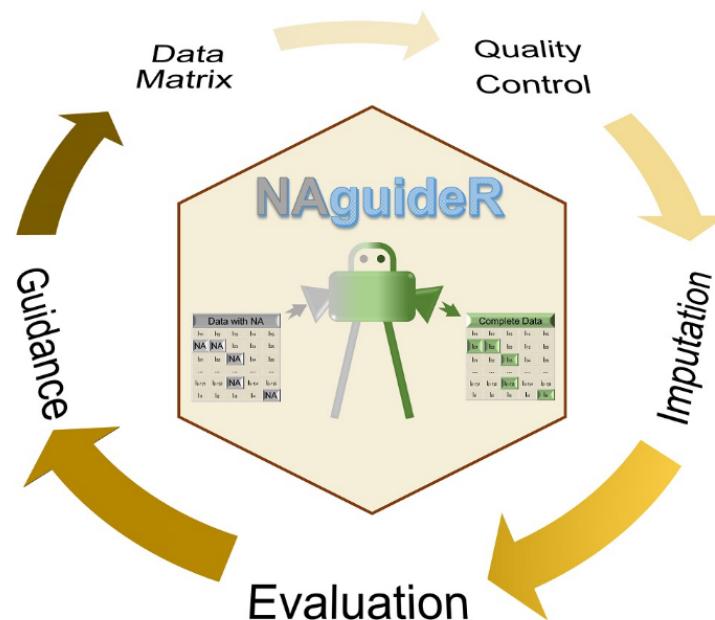
Corresponding Author

*Email address: yanghao@scu.edu.cn; yansheng.liu@yale.edu.

Supplementary notes

NAguideR integrates up to 23 commonly used missing value imputation methods (described in Table S1) and provides two categories of evaluation criteria (four classic computational criteria and four empirical proteomics criteria) to assess the imputation performance of various methods. Here we present the detailed introduction and operation of NAguideR, by which users can follow to analyze their own data freely and conveniently.

Users can visit this site: <http://www.omicsolution.org/wukong/NAguideR>. Then the website homepage can be shown like this:



Basically, there are four main steps in NAguideR:

1. Uploading proteomics expression data and sample information data;
2. Data quality control;
3. Missing value imputation;
4. Performance evaluation;

After this, NAguideR can provide valuable guidance for users to select one proper method for their own data based on the evaluation results. Detailed introduction can be found in the **Help** part.

Finally, NAguideR is developed by R shiny (Version 1.3.2), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at wssdandan2009@outlook.com.

Optional: For large-scale analysis, enter your email here and come back any time (Note: Please also check junk mail if possible.):

wssdandan2009@outlook.com

^_^ Enjoy yourself in NAguideR ^_^\n

1. Data Preparation

NAguideR supports four basic file formats (.csv, .txt, .xlsx, .xls). Before analysis, users should prepare two required data: (i) Proteomics expression table for quantification and (ii) Sample information. The data required here could be readily generated based on results of several popular tools such as MaxQuant (1), PEAKS (2), Spectronaut (3), DIA-NN (4), OpenSWATH (5), and so on. The users then can upload the two data into NAguideR with right formats respectively and start subsequent analysis.

1.1 Expression data

There are currently four types of proteomics expression data supported in NAguideR (i.e., 'Peptides+Charges+Proteins', 'Peptides+Charges', 'Peptides+Proteins', 'Proteins'), among which the main differences are the first few columns. In addition, users may upload other kinds of omics data (i.e. Genomics, Metabolomics), they can just choose the fifth type ('Others'). Please note, the fifth type cannot generate the results based on those proteomic criteria.

The screenshot shows the 'Step 1: Upload Original Data' interface. On the left, there's a sidebar with 'Load experimental data' (selected) and 'Load example data'. The main area has a section titled '1. Expression data:' with '1.1 File format:' and options for '.csv/txt' (selected), '.xls', and '.xlsx'. Below that is '1.2 Import your data:' with a 'Browse...' button and a 'No file selected' message. On the right, a panel titled '1. Expression data:' shows a dropdown menu for 'The first few column types' with five options: 'Peptides+Charges+Proteins' (selected), 'Peptides+Charges+Proteins', 'Peptides+Charges', 'Peptides+Proteins', and 'Proteins'. A red box highlights the dropdown menu. At the bottom of the panel, it says 'Showing 1 to 1 of 1 entries'.

1.1.1 Expression data with peptide sequences, peptide charge states, and protein ids

In this situation, peptide sequences, peptide charge states, and protein ids are sequentially provided in the first three columns of input file. Peptide sequences in the first column can be peptides with any post-translational modification (PTM, written in any routine format) or stripped peptides (without PTM). The second column is peptide charge status. The protein ids in the third column should be UniProt ids. From the fourth column, peptides/proteins expression intensity or signal abundance in every sample should be listed. The data structure is shown as below:

Sample names

Peptides	Charges	Uniprot IDs	Phos_Cyc					
			1	2	3	4	5	6
MLISAVS[Phospho (2	A0AVK6	157593.6	203318.3	125540.8	131965.7	195625.7	180664.5
KINS[Phospho (STY	2	A0AVK6	712596.2	744410.4	998674.3	1139399	956570.4	1120046
EPT[Phospho (STY)	3	A0FGR8	511129.7	639703 NA	562894.8	829802.6	645625.4	
SSSSLLAS[Phospho	3	A0FGR8	74890.52	80801.82	80222.84	88827.91	115544.8	80334.69
SSSSLLAS[Phospho	2	A0FGR8	34336.86	28903.73	NA	30830.68	33390.47	NA
TQDPVPETPSDS[Pho	3	A0JLT2	221698.9	NA	270359.6	312614.6	345215.3	284286.5
SMS[Phospho (STY)	3	A0JNW5	248274	427877.3	358461	316457.7	352716.8	285275.5
M[Acetyl (Protein	2	A1KXE4	79679.09	NA	110380.5	130927.4	82461.96	155724.4
QNSLGC[Carbamidom	3	A1L020	558781.1	676339.8	594215.1	692863.3	587093.6	756873
ASS[Phospho (STY)	2	A1L170	344653.8	413764.8	287084	286627.3	417670.3	295301.9
SHS[Phospho (STY)	2	A1L390	3293265	2527386	2685655	NA	2318149	4120553
GPLS[Phospho (STY)	2	A1L390	1551857	1596314	1253587	1406729	1723560	1502006
IWEGMESSGGS[Phosp	2	A1L390	686212.1	703314.1	697566	580441.7	808891.8	745552.6
SHS[Phospho (STY)	3	A1L390	NA	604569.9	NA	784035	554084.8	NA
SPLS[Phospho (STY)	2	A1L390	833264.3	1034303	867998.1	714042.4	990010.2	1039246
S[Phospho (STY)]P	2	A1L390	801754.1	825141.1	840367.5	709674.4	895440	913974.4
SSVLS[Phospho (S	2	A1L390	729638	795307.5	637437.1	714943.1	806124.7	818071.6
RES[Phospho (STY)	2	A1L390	386283.1	NA	371454.2	364010.8	415586.1	373595.6

Peptide sequences Protein ids Peptide Charge status Intensity matrix

1.1.2 Expression data with peptide sequences and peptide charge states

Similar to the above situation, peptide sequences and peptide charge status are sequentially provided in the first two columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM, written in any routine format) or stripped peptides (without PTM). The second column is peptide charge states. From the third column, peptides/proteins expression intensity or signal abundance in every sample should be listed. The data structure is shown as below:

Sample names

Peptides	Charges	Phos_Cyc					
		1	2	3	4	5	6
MLISAVS[F	2	157593.6	203318.3	125540.8	131965.7	195625.7	180664.5
KINS[Phos	2	712596.2	744410.4	998674.3	1139399	956570.4	1120046
EPT[Phosp	3	511129.7	639703	NA	562894.8	829802.6	645625.4
SSSSLLAS[3	74890.52	80801.82	80222.84	88827.91	115544.8	80334.69
SSSSLLAS[2	34336.86	28903.73	NA	30830.68	33390.47	NA
TQDPVPETI	3	221698.9	NA	270359.6	312614.6	345215.3	284286.5
SMS[Phosp	3	248274	427877.3	358461	316457.7	352716.8	285275.5
M[Acetyl	2	79679.09	NA	110380.5	130927.4	82461.96	155724.4
QNSLGC[Ca	3	558781.1	676339.8	594215.1	692863.3	587093.6	756873
ASS[Phosp	2	344653.8	413764.8	287084	286627.3	417670.3	295301.9
SHS[Phosp	2	3293265	2527386	2685655	NA	2318149	4120553
GPLS[Phos	2	1551857	1596314	1253587	1406729	1723560	1502006
IWEGMESSC	2	686212.1	703314.1	697566	580441.7	808891.8	745552.6
SHS[Phosp	3	NA	604569.9	NA	784035	554084.8	NA
SPLS[Phos	2	833264.3	1034303	867998.1	714042.4	990010.2	1039246
S[Phospho	2	801754.1	825141.1	840367.5	709674.4	895440	913974.4
SSVLS[PR	2	729638	795307.5	637437.1	714943.1	806124.7	818071.6
RES[Phosp	2	386283.1	NA	371454.2	364010.8	415586.1	373595.6

Peptide sequences Peptide Charge status Intensity matrix

1.1.3 Expression data with peptide sequences, and protein ids

Under this circumstance, peptide sequences, and protein ids are sequentially provided in the first two columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM, written in any routine format) or stripped peptides (without PTM). The protein ids in the second column should be UniProt ids. From the third column, peptides/proteins expression intensity or signal abundance in every sample should be listed. The data structure is shown as below:

Peptides	Uniprot IDs	Sample names					
		Phos_Cyc 1	Phos_Cyc 2	Phos_Cyc 3	Phos_Cyc 4	Phos_Cyc 5	Phos_Cyc 6
MLISAVS[Phospho (STY)]	A0AVK6	157593.6	203318.3	125540.8	131965.7	195625.7	180664.5
KINS[Phospho (STY)]AP	A0AVK6	712596.2	744410.4	998674.3	1139399	956570.4	1120046
EPT[Phospho (STY)]PSI	A0FGR8	511129.7	639703	NA	562894.8	829802.6	645625.4
SSSSLLAS[Phospho (STY]	A0FGR8	74890.52	80801.82	80222.84	88827.91	115544.8	80334.69
SSSSLAS[Phospho (STY]	A0FGR8	34336.86	28903.73	NA	30830.68	33390.47	NA
TQDPVPPETPSDS[Phospho	A0JLT2	221698.9	NA	270359.6	312614.6	345215.3	284286.5
SMS[Phospho (STY)]VDL	A0JNW5	248274	427877.3	358461	316457.7	352716.8	285275.5
M[Acetyl (Protein N-t	A1KXE4	79679.09	NA	110380.5	130927.4	82461.96	155724.4
QNSLGC[Carbamidomethy	A1L020	558781.1	676339.8	594215.1	692863.3	587093.6	756873
ASS[Phospho (STY)]PSL	A1L170	344653.8	413764.8	287084	286627.3	417670.3	295301.9
SHS[Phospho (STY)]VPE	A1L390	3293265	2527386	2685655	NA	2318149	4120553
GPLS[Phospho (STY)]PF	A1L390	1551857	1596314	1253587	1406729	1723560	1502006
IWEGMESSGGS[Phospho (A1L390	686212.1	703314.1	697566	580441.7	808891.8	745552.6
SHS[Phospho (STY)]VPE	A1L390	NA	604569.9	NA	784035	554084.8	NA
SPLS[Phospho (STY)]PT	A1L390	833264.3	1034303	867998.1	714042.4	990010.2	1039246
S[Phospho (STY)]PLSPT	A1L390	801754.1	825141.1	840367.5	709674.4	895440	913974.4
SSVLS[Phospho (STY)]	A1L390	729638	795307.5	637437.1	714943.1	806124.7	818071.6
RES[Phospho (STY)]LSY	A1L390	386283.1	NA	371454.2	364010.8	415586.1	373595.6

The diagram illustrates the data structure. At the top, a yellow arrow labeled "Sample names" points down to the header row of the table. Below the table, three blue arrows point up from the bottom left to the second column ("Uniprot IDs"), the third column ("Peptides"), and the fourth column ("Intensity matrix").

1.1.4 Expression data with protein ids

In this situation, protein ids are provided in the first two columns of input file. The protein ids here should be UniProt ids. From the second column, peptides/proteins expression intensity or signal abundance in every sample should be listed. The data structure is shown as below:

Sample names

Uniprot_ IDs	Phos_Cyc					
	_1	_2	_3	_4	_5	_6
AOAVK6	157593.6	203318.3	125540.8	131965.7	195625.7	180664.5
AOAVK6	712596.2	744410.4	998674.3	1139399	956570.4	1120046
AOFGR8	511129.7	639703	NA	562894.8	829802.6	645625.4
AOFGR8	74890.52	80801.82	80222.84	88827.91	115544.8	80334.69
AOFGR8	34336.86	28903.73	NA	30830.68	33390.47	NA
AQJLT2	221698.9	NA	270359.6	312614.6	345215.3	284286.5
AQJNW5	248274	427877.3	358461	316457.7	352716.8	285275.5
A1KXE4	79679.09	NA	110380.5	130927.4	82461.96	155724.4
A1L020	558781.1	676339.8	594215.1	692863.3	587093.6	756873
A1L170	344653.8	413764.8	287084	286627.3	417670.3	295301.9
A1L390	3293265	2527386	2685655	NA	2318149	4120553
A1L390	1551857	1596314	1253587	1406729	1723560	1502006
A1L390	686212.1	703314.1	697566	580441.7	808891.8	745552.6
A1L390	NA	604569.9	NA	784035	554084.8	NA
A1L390	833264.3	1034303	867998.1	714042.4	990010.2	1039246
A1L390	801754.1	825141.1	840367.5	709674.4	895440	913974.4
A1L390	729638	795307.5	637437.1	714943.1	806124.7	818071.6
A1L390	386283.1	NA	371454.2	364010.8	415586.1	373595.6

1.1.5 Other kinds of omics data

If users want to use NAguideR for other omics data (i.e. genomics, metabolomics), gene/metabolite ids/names should be provided in the first columns of input file. From the second column, genes/metabolites expression intensity or signal abundance in every sample should be listed. The data structure may be shown as below:

Sample names

names	QC1	QC2	QC3	A1	A2	A3	B1	B2	B3
7.61_520.3395	29071.81	38643.44	28234.63	30406.31	71802.95	63405.6	27240.34	34531.1	86799.11
9.32_561.3968	873.295	915.3884	881.619	740.8663	602.5532	726.5324	671.0133	740.4401	892.6893
2.81_393.2095	912.2716	921.4828	912.8439	729.7479	639.6943	738.2883	527.454	636.2253	895.943
6.22_431.2773	NA	NA	7.788453	1050.224	925.818	1072.712	14.42876	NA	NA
2.54_333.1165	795.9917	844.7863	745.8313	1772.859	684.9352	1137.385	496.0765	187.7253	461.9466
9.34_517.3708	1280.567	1262.694	1167.615	NA	NA	NA	925.803	1077.253	1251.438
4.62_259.0950	518.5587	517.8369	357.9304	31.09069	30.04137	11.52148	4459.264	14.84036	21.40445
8.97_551.3551	2038.019	2334.945	2083.307	2214.768	1951.39	2441.952	2392.231	2949.036	3632.063
7.94_496.3398	94826.86	133532.9	92989.35	118231.6	205200.5	158300.1	143394.8	140966.2	216601.7
8.88_727.4591	1051.915	1154.645	1091.713	1110.864	857.2376	1206.524	1286.299	1591.188	1952.897
8.91_683.4332	1383.729	1507.906	1411.877	1389.049	1140.618	1482.737	1551.337	1994.083	2400.456
1.73_366.0592	5932.028	6010.801	4609.131	111.3436	86.55886	46.36566	1207.938	59.87836	30.88989
2.95_437.2356	1463.975	1524.707	1381.035	1241.034	1259.203	1216.24	1155.202	1138.482	1471.988
3.76_509.2928	1568.452	1210.315	2169.963	1925.416	991.1353	1924.902	1644.312	NA	2337.039
8.93_639.4072	1659.309	1836.396	1750.505	1679.314	1371.43	1845.365	1867.644	2435.767	2920.997
8.95_595.3807	NA	2190.768	2191.84	2172.971	1688.99	2321.392	2247.959	2886.744	3449.909
8.39_449.2872	1682.981	1830.449	957.0811	604.8445	1224.764	691.8988	710.3057	705.4097	885.0374
3.31_548.2924	137.8952	148.7875	152.2905	657.6478	55.35951	443.2037	627.0334	84.37179	38.64603
3.08_476.3064	1154.09	1296.454	1097.852	NA	930.2538	951.6176	814.497	927.2989	1172.239
3.19_509.2188	327.6037	375.0294	419.5559	1801.861	242.4814	1489.241	928.8548	203.6372	320.4381
2.64_349.1836	1369.609	1482.483	1431.126	1292.658	1102.289	1333.388	1257.245	1160.143	1530.776

1.2 Samples information data

Sample information here means that users should provide sample group identity information. This information could e.g., enable filtration strategy for different group respectively in a later step (see below). The sample names are in the first column and their orders are same as those in the expression data. Group information is in the second column. The data structure is shown as below:

Sample names

↓

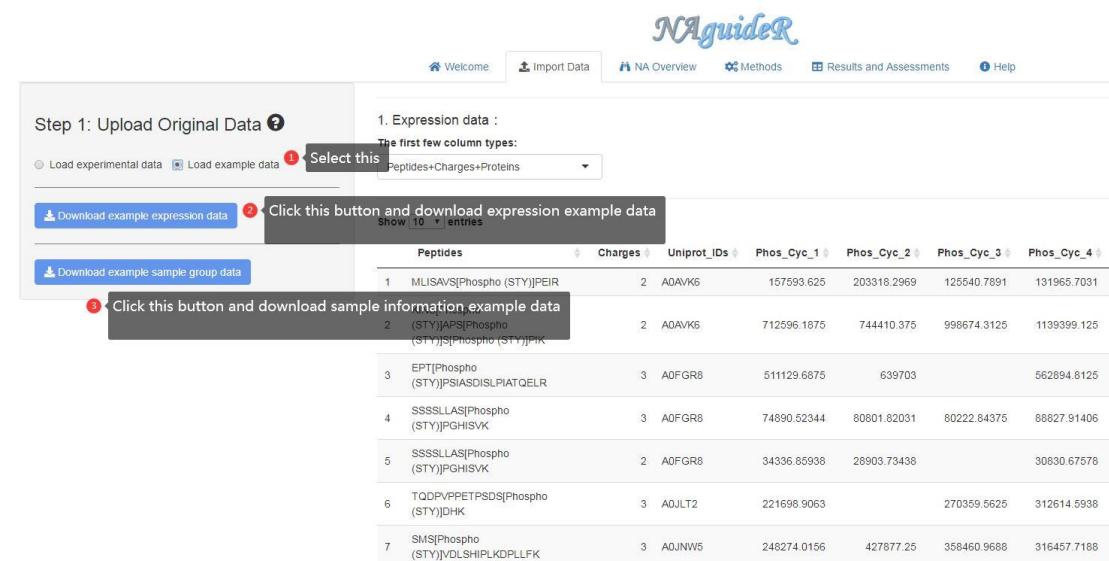
Samples	Groups
Phos_Cyc_1	Cyc
Phos_Cyc_2	Cyc
Phos_Cyc_3	Cyc
Phos_Cyc_4	Cyc
Phos_Cyc_5	Cyc
Phos_Cyc_6	Cyc
Phos_Cyc_7	Cyc
Phos_Cyc_8	Cyc
Phos_Cyc_9	Cyc
Phos_Cyc_10	Cyc
Phos_Noco_1	Noco
Phos_Noco_2	Noco
Phos_Noco_3	Noco
Phos_Noco_4	Noco
Phos_Noco_5	Noco
Phos_Noco_6	Noco
Phos_Noco_7	Noco
Phos_Noco_8	Noco
Phos_Noco_9	Noco
Phos_Noco_10	Noco

↑

Sample groups

1.3 Download example datasets

If users want to download the example datasets to their own computer and check the data format locally, they can download them from here:



Peptides	Charges	Uniprot_IDs	Phos_Cyc_1	Phos_Cyc_2	Phos_Cyc_3	Phos_Cyc_4
MLISAVS[Phospho (STY)]PEIR	2	A0AVK6	157593.625	203318.2969	125540.7891	131965.7031
(STY)APS[Phospho (STY)]PSIPIK	2	A0AVK6	712596.1875	744410.375	998674.3125	1139399.125
EPT[Phospho (STY)]PSIASDILPATQELR	3	A0FGR8	511129.6875	639703	562894.8125	
SSSSLLAS[Phospho (STY)]PGHISVK	3	A0FGR8	74890.52344	80801.82031	80222.84375	88827.91406
SSSSLLAS[Phospho (STY)]PGHISVK	2	A0FGR8	34336.85938	28903.73438		30830.67578
TQDPVPPETPSDS[Phospho (STY)]DHK	3	A0JLT2	221698.9063		270359.5625	312614.5938
SMS[Phospho (STY)]VDSLHPLKDPLLFK	3	A0JNW5	248274.0156	427877.25	358460.9668	316457.7188

First, select “Load example data” and the example data will be shown on the right panel interactively. Users can visually observe what the data looks like.

Second, users can download the example data (expression data and sample information data) by clicking the corresponding button. The data are saved as .csv format and users can open them in other software, such as Excel.

2. Import data.

This is the first step, in which users should upload data here or load the example data to learn the data formats. By default, we use the example data to show each result of every step.

2.1 Uploading data. When users prepare their data (expression and sample information data set), they can upload these data from here:

There are two main panels: first, *parameters panel*, users can adjust parameters here; second, *results panel*, many results after users set the parameters will be shown here and users can also download these results.

In the *parameters panel* of “Import Data”, there are two choices for users:

a. *Load experimental data*. When users choose this option, they can upload their own data here. Users should select the right format based on their data and then click “Browse” button to import the data;

First row as column names: this means whether the first row is column names. If true, you should choose this parameter.

First column as row names: this means whether the first column is row names. If true, you should choose this parameter.

b. *Load example data*. As described in part 1.3, users can choose this option and download the example data to check them locally.

In the *results panel* of “Import Data”, if users don’t upload their data, here will show “NAguideR detects that you did not upload your data. Please upload the expression data (or sample information data), or load the example data to check first” to warn users.

Before uploading expression data, users should also recognize which type their data belongs to and choose the right parameter by adjusting the “*The first few column types*”. The instruction of the

column types can be found above (*Data Preparation* part).

Step 1: Upload Original Data

Load experimental data Load example data

1. Expression data:

1.1 File format:

.csv/txt .xls .xlsx

1.2 Import your data :

1. Expression data :

The first few column types:

Peptides+Charges+Proteins

Peptides+Charges+Proteins

Peptides+Charges

Peptides+Proteins

Proteins

Others

Showing 1 to 1 of 1 entries

3. NA Overview

Users can check the missing value situation of their own data and filter those data with a high proportion of missing value in this step. Note, “NA” is short for Not Available, which means missing value here (see below).

The screenshot shows the NAguideR web application. At the top, there is a navigation bar with links: Welcome, Import Data, NA Overview (which is the active tab), Methods, Results and Assessments, and Help. Below the navigation bar, there are three tabs: NA Distribution, NA Filter (which is the active tab), and Input data check. Under the NA Filter tab, there are several input fields and checkboxes:

- 1. Missing value type: A text input field containing "NA".
- 2. Count NA by each group or not?: A checkbox that is checked.
- 3. NA ratio: A text input field containing "0.5".
- 4. Median normalization or not?: A checkbox that is checked.
- 5. Log or not?: A checkbox that is checked.
- 6. CV threshold (raw scale): A text input field containing "0.3".
- Height for figure: A text input field containing "900".

At the bottom of the form is a large orange "Calculate" button.

3.1 Parameters

This is a configuration panel for the Step 2: NA Overview step. It contains the same set of parameters as the main interface:

- 1. Missing value type: A text input field containing "NA".
- 2. Count NA by each group or not?: A checkbox that is checked.
- 3. NA ratio: A text input field containing "0.5".
- 4. Median normalization or not?: A checkbox that is checked.
- 5. Log or not?: A checkbox that is checked.
- 6. CV threshold (raw scale): A text input field containing "0.3".
- Height for figure: A text input field containing "900".

- Missing value type:* what the missing values look like in the expression data, for example, Spectronaut (6,7) software usually export “Filtered” as missing values, so users should change this parameter to “Filtered” if their data contain “Filtered”. NAguideR will recognize these characters and replace them with NAs. Any other characters indicating a missing value can be similarly defined.
- Count NA by each group or not:* if true, NAguideR will count the number of missing values in each group and calculate the NA ratio. Otherwise, it calculates the NA ratio across all groups, for example, as below:

Peptides	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	Charges	Uniprot_ids	Phos_Cyc	Phos_Noc																			
Malacetyl (Protein N-ter)	2	AIK384	79679.09	NA	110380.5	130927.4	82461.96	159724.4	113495.3	136404.3	56171.31	98299.7	NA	NA	151027.6	NA	210179.9	182829.7	151426.3	NA	NA	181321.2	

There are 2 groups (10 biological replicates in each group) here, if users select this parameter, NAguideR will calculate 2 NA ratios for this peptide (first group: 1/10=0.1, second group: 5/10=0.5), otherwise, only one NA ratio: 6/20=0.3.

3. *NA ratio*: the threshold of NA ratio. Those peptides/proteins with NA ratio above this threshold will be removed.

4. *Median normalization or not*: if true, NAguideR will process median normalization for original data. (Note, NAguideR was not designed to perform sophisticated normalization analysis. Any normalized datasets with NA can be accepted for analysis).

5. *Log or not*: if true, the data will be transformed to the logarithmic scale with base 2.

6. *CV threshold (raw scale)*: the threshold of coefficient of variation. Those peptides/proteins with NA ratio above this threshold will be removed. “raw scale” here means the CV of each peptide/protein is calculate using the data before logarithm transformation.

7. *Height for figure*: users can adjust the height of figures by changing this parameter.

If users set these parameters well, then click “calculate” button, the results will appear on the right panel.

	Phos_Cyc_1	Phos_Cyc_2	Phos_Cyc_3	Phos_Cyc_4	Phos_Cyc_5	Phos_Cyc_6	Phos_Cyc_7	Phos_Cyc_8	Phos_Cyc_9
MLISAV[S]Phospho (STY)[PEIR_2_ADAVKG]	157593.625	203318.2969	125540.7891	131965.7031	195625.6563	180664.5469	148941.4688	143790.9375	91102.99219
KIN[S]Phospho (STY)[APS]Phospho (STY)[P]Phospho (STY)[PHK_2_ADAVKG]	712596.1875	744410.375	998674.3125	1139399.125	956570.375	1120045.625	860231.875	823408.5625	
EPT[Phospho (STY)[SIAASDILPATQELR_3_ADFGR8]	511129.6875	639703		562894.8125	829802.625	645625.4375			608952.875
SSSSLLAS[Phospho (STY)[PQHISVK_3_ADFGR8]	74890.52344	80801.82031	80222.84375	88827.91406	115544.7813	80334.6875	80562.07031	61538.41406	53648.84766
SSSSLLAS[Phospho (STY)[PQHISVK_2_ADFGR8]	34336.85938	28903.73438		30830.67578	33390.47266			31978.69141	29228.26758
TQDPVPPETPSD[Phospho (STY)[DHK_3_ADLT2]	221698.9063		270359.5625	312614.5938	345215.25	284286.4688	203317.4063	218004.125	185125.5156
SMS[Phospho (STY)[VLSHPLKDPLFLK_3_AQJNWS]	248274.0156	427877.25	358460.9688	316457.7188	352716.75	285275.5	331924.5625	174794.2344	241767.2344
M[Acetyl] (Protein N-term)[NPV*YSPGSSGV[P]Phospho (STY)[ANAK_2_A1KOE4]	79679.09375			110380.5	130927.3672	82461.96094	155724.3594	113495.2891	136404.2969
									56171.30859

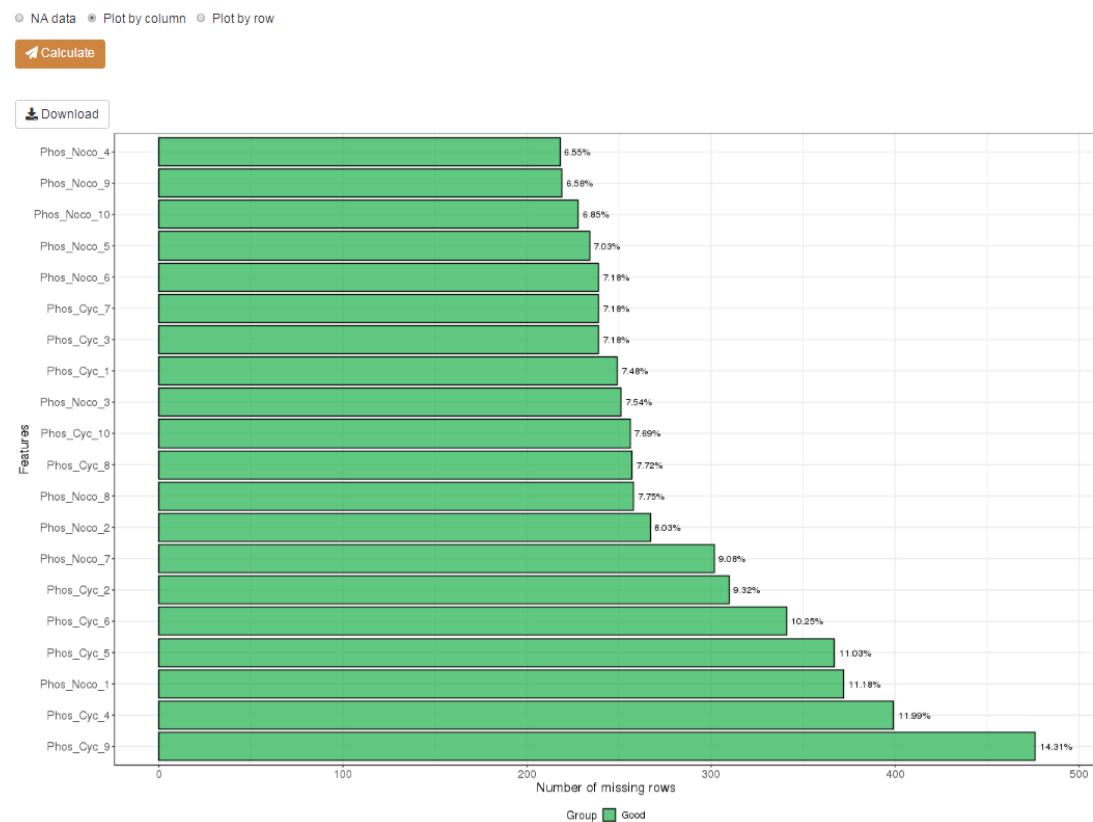
3.2 results of NA overview

a. *NA Distribution*. This part contains three sub-parts:

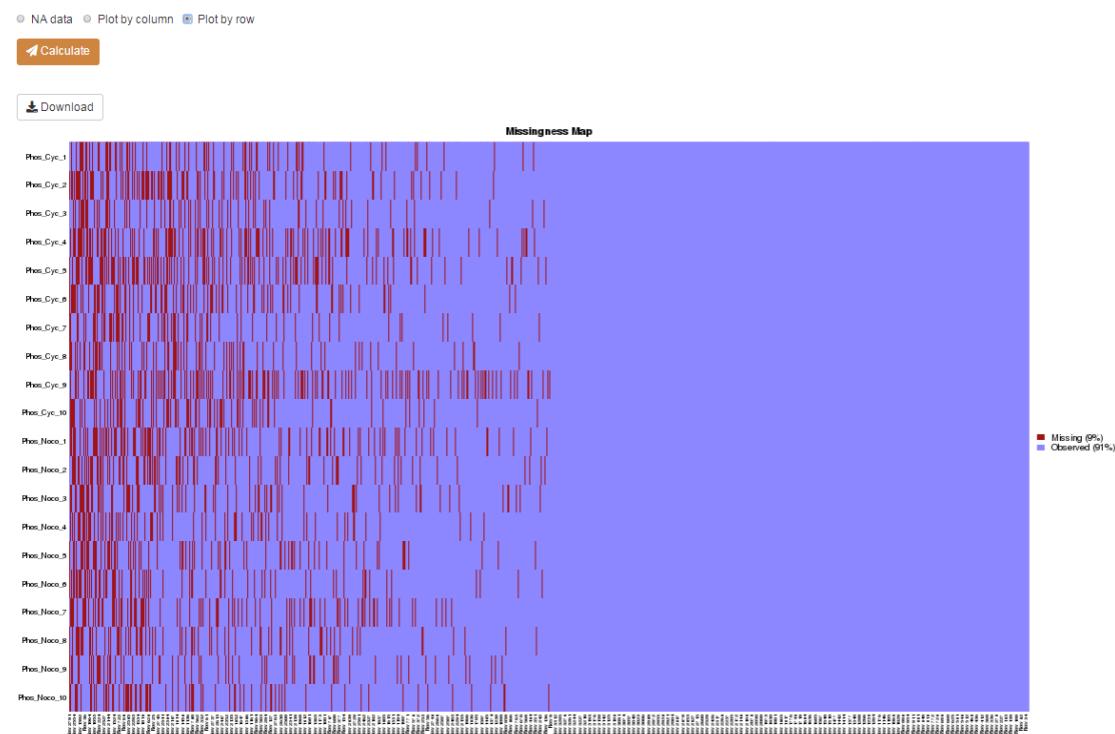
a.1 *NA data*. Here shows the result where the “Missing value type” defined by “NA” will be shown with a blank cell and users can click “Download” button to download this result to their own computer:

	Phos_Cyc_1	Phos_Cyc_2	Phos_Cyc_3	Phos_Cyc_4	Phos_Cyc_5	Phos_Cyc_6	Phos_Cyc_7	Phos_Cyc_8	Phos_Cyc_9	Phos_Cyc_10	Phos_Nocco_1	Phos_Nocco_2
MLISAV[S]Phospho (STY)[PEIR_2_ADAVKG]	157593.625	203318.2969	125540.7891	131965.7031	195625.6563	180664.5469	148941.4688	143790.9375	91102.99219	140345.125	59488.09766	92400.46094
KIN[S]Phospho (STY)[APS]Phospho (STY)[P]Phospho (STY)[PHK_2_ADAVKG]	712596.1875	744410.375	998674.3125	1139399.125	956570.375	1120045.625	860231.875	823408.5625		888177.75	509135.625	595305.5
EPT[Phospho (STY)[SIAASDILPATQELR_3_ADFGR8]	511129.6875	639703		562894.8125	829802.625	645625.4375		608952.875		620510.4375	323346.5313	334969.9063
SSSSLLAS[Phospho (STY)[PQHISVK_3_ADFGR8]	74890.52344	80801.82031	80222.84375	88827.91406	115544.7813	80334.6875	80562.07031	61538.41406	53648.84766	65030.57031	516738.8125	782993.875
SSSSLLAS[Phospho (STY)[PQHISVK_2_ADFGR8]	34336.85938	28903.73438		30830.67578	33390.47266		31978.69141	29228.26758		26532.99219	333476.9688	297875.2188
TQDPVPPETPSD[Phospho (STY)[DHK_3_ADLT2]	221698.9063		270359.5625	312614.5938	345215.25	284286.4688	203317.4063	218004.125	185125.5156	245305	81982.05469	81776.67188
SMS[Phospho (STY)[VLSHPLKDPLFLK_3_AQJNWS]	248274.0156	427877.25	358460.9688	316457.7188	352716.75	285275.5	331924.5625	174794.2344	241767.2344	284089.4375	170259.6094	207056.5
M[Acetyl] (Protein N-term)[NPV*YSPGSSGV[P]Phospho (STY)[ANAK_2_A1KOE4]	79679.09375			110380.5	130927.3672	82461.96094	155724.3594	113495.2891	136404.2969	156171.30859	98299.69531	
QNSLGC[Carbamidomethyl (C)]IGEC[Carbamidomethyl (C)]GVDS[Phospho (STY)[GFEAPR_3_AL020]	558781.125	676339.75	594215.0625	692863.25	587093.5625	756873	569292.25	648059.3125		626625.4375	379149.5625	361978.75

a.2 Plot by column. Here shows the result of the NA distribution of every sample.



a.2 Plot by row. Here shows the result of the NA distribution of every peptide/protein.



b. NA filter. This part will show the filtered result. That means, on the basis of the preset parameters

(i.e. NA ratio, CV threshold), those objects (peptides/proteins/genes/metabolites) without meeting these requirements would be removed.

c. *Input data check*. This part will show the checking information as a summary note for input data. By default, if there still remain more than half (>50%) objects in the filtered data, NAguideR will think it is acceptable, and will give users a message like below:

Otherwise, NAguideR will give some warnings to users, which means users should pay more attention to their own data and those preset parameters. It is recommended that the users should then make sure that there are no problems before they can proceed to the next step:

NAguideR

Welcome Import Data NA Overview Methods Results and Assessments Help

Step 2: NA Overview

1. Missing value type:

2. Count NA by each group or not?

3. NA ratio:

4. Median normalization or not?

5. Log or not?

6. CV threshold (raw scale):

Height for figure:

~~ Check information for input data ~~

1. There are 54075 rows and 20 columns in the input expression data;
2. After removing those rows with high proportion of missing values and coefficient of variation (the threshold can be set on the left parameter panel), there are 13946 rows left in the filtered data;

Warning: 74 % of the input data are removed, we suggest you check or adjust your input data and the parameters again. If you can be sure there are no problems on the input data and parameters, you can proceed to the next step.

4. Methods

In this step, users can select any of 23 missing value imputation methods that are currently supported. All methods have been classified into three categories based on their algorithm (Single value approaches, global structure approaches and local similarity approaches). In order to control the running time, we set these fast methods (17 methods) chosen by default. If users choose those slow methods (6 methods), that means the running time will be longer. If users want to try these slow methods, they just need to select the corresponding methods. The detailed information about each method can be found in Table S1. In addition, we also provide the reference for every method just blow each option on the web:

NAguideR

Welcome Import Data NA Overview Methods Results and Assessments Help

Step 3: Missing value imputation. All methods have been classified based on their algorithm, please select the imputation methods you want (by default, fast methods are chosen in each category), then click the 'Calculate' button.

A. Single value approaches

- Method 1: Zero

Using zero method or not?

DOI: 10.1021/acs.jproteome.5b00981
- Method 2: Minimum

Using minimum method or not?

DOI: 10.1038/s41586-019-0967-8
- Method 3: Column median (colmedian)

Using colmedian method or not?

Package: e1071
- Method 4: Row median (rowmedian)

Using row median method or not?

Package: e1071
- Method 5: Deterministic minimal value (mindet)

Using mindet method or not?

Package: imputeCMD
- Method 7: Perseus imputation (P)

Using perseus imputation method or not?

DOI: 10.1038/math.2001

B. Global structure approaches

- Method 8: Singular value decomposition (svd)

Using svd method or not?

DOI: 10.1093/bioinformatics/17.8.520
- Method 10: Sequential imputation (impseq)

Using impseq method or not?

DOI: 10.19165/compbiochem.2007.07.001
- Method 12: Bayesian principal component analysis (bpca)

Using bpca method or not?

DOI: 10.1093/bioinformatics/btg287
- Method 9: Maximum likelihood estimation (mle)

Using mle method or not?

Package: norm
- Method 11: Robust sequential imputation (impseqrob)

Using impseqrob method or not?

DOI: 10.1016/j.compbiochem.2008.07.019

C. Local similarity approaches

- Method 13: K-nearest neighbor (knn)

Using knn method or not?

DOI: 10.1093/bioinformatics/17.6.520
- Method 15: Quantile regression (qr)

Using qr method or not?

Package: imputeCMD
- Method 17: Gimmel Ridge Regression (GRR)

Using GRR method or not?

Package: DreamAI
- Method 19: Truncation knn (trknn)

Using trknn method or not?

DOI: 10.1186/12859-017-1547-6
- Method 21: Generalized Mass Spectrum (GMS)

Using GMS method or not?

DOI: 10.1093/bioinformatics/bt2488
- Method 23: Random forest model (rf)

Using rf method or not?

Number of trees:
20

DOI: 10.1093/bioinformatics/bt5597
- Method 14: Sequential knn (seq-knn)

Using seq-knn method or not?

DOI: 10.1186/1471-2105-5-160
- Method 16: Local least squares (lls)

Using lls method or not?

DOI: 10.1093/bioinformatics/btm499
- Method 18: Multiple imputation bayesian linear regression (mice-norm)

Using mice-norm method or not?

DOI: 10.1837/jea.v045.i03
- Method 20: Iterative robust model (irm)

Using irm method or not?

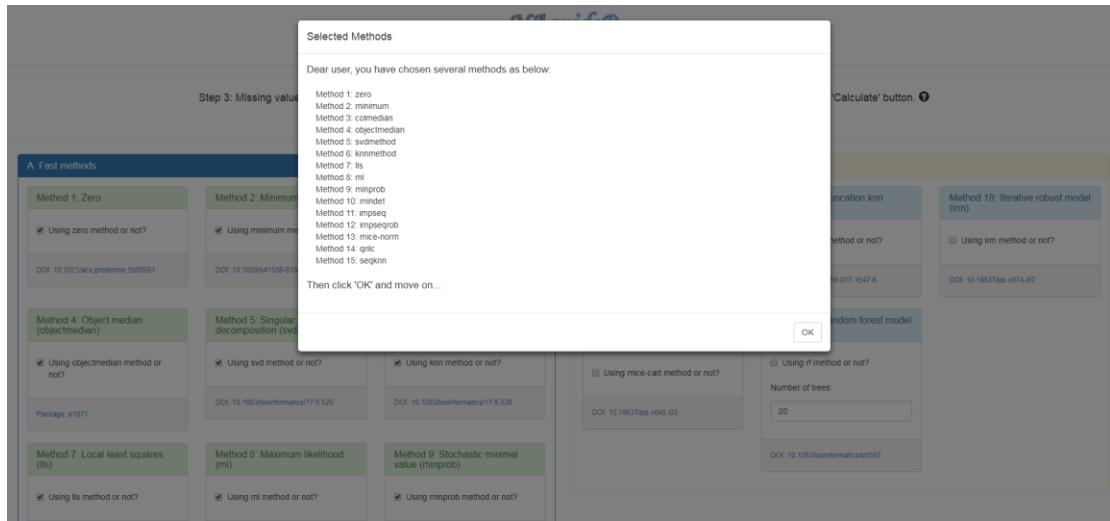
DOI: 10.1837/jea.v074.i07
- Method 22: Multiple imputation classification and regression trees (mice-cart)

Using mice-cart method or not?

DOI: 10.1837/jea.v045.i03

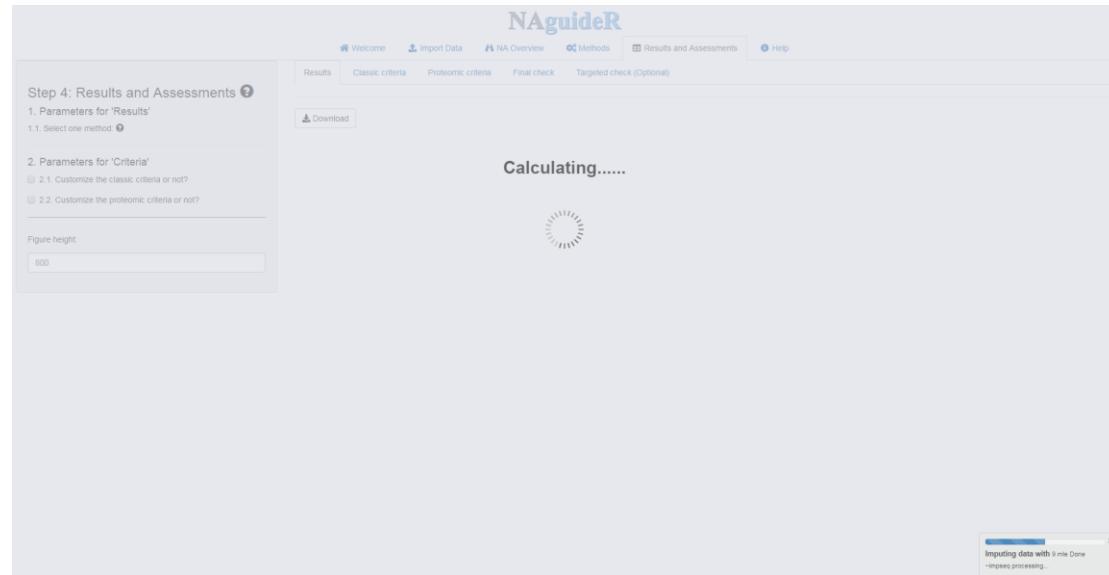
Calculate

After selecting suitable methods, users need to click 'Calculate' button, and a popup window will be jumped out to show the selected methods, then click 'OK' button and continue:



5. Results and Assessments

This step will process missing value imputation and performance evaluation of every method that users select in “Methods” step. Click “Results and Assessments”, NAguideR will start to impute these missing value items, a process bar will appear in the bottom right corner to tell users where it goes:



The result from every imputation method will be shown on the “Results” panel:

The screenshot shows the NAguideR interface with the 'Results' tab selected. On the left, there are three sections: 'Parameters for 'Results'' (dropdown set to 'zero'), 'Parameters for 'Criteria'' (checkboxes for classic and proteomic criteria), and 'Figure height' (input field set to '600'). In the center, there is a table titled 'Phos_Cyc_1' with columns for Phos_Cyc_1 through Phos_Cyc_9. The table contains 14 rows of data, each with a protein name and its corresponding imputed values. The table includes a search bar at the top right and a 'Show' dropdown set to '20 entries'.

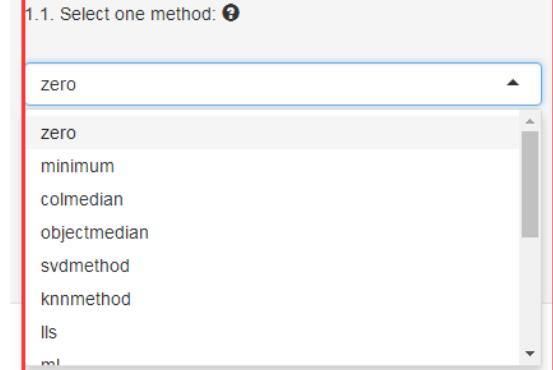
	Phos_Cyc_1	Phos_Cyc_2	Phos_Cyc_3	Phos_Cyc_4	Phos_Cyc_5	Phos_Cyc_6	Phos_Cyc_7	Phos_Cyc_8	Phos_Cyc_9
MILISVIPSPhospho (STY)PPIER_2_A0AVK6	-0.8907	-0.60469	-1.19361	-1.29635	-0.71933	-0.86619	-0.951	-0.98901	-1.50231
KNSIPPhospho (STY)APSPhospho (STY)SPPhospho (STY)PIK_2_A0AVK6	1.28617	1.26767	1.79805	1.81369	1.57044	1.74598	1.57898	1.52863	0
EPTIPPhospho (STY)PPIASDIBLPIATQELR_3_A0FGR8	0.80678	1.04897	0	0.79635	1.36534	0.95119	0	1.0933	0
SSSSLASIPPhospho (STY)PHHSIVK_3_A0FGR8	-1.96406	-1.93597	-1.83988	-1.86743	-1.47898	-2.05541	-1.83757	-2.21342	-2.26626
SSSSLASIPPhospho (STY)PHHSIVK_2_A0FGR8	-3.08908	-3.4191	0	-3.39407	-3.26992	0	-3.17056	-3.28755	0
TGDPVPETPSQSIPPhospho (STY)DHDH_3_A0L1T2	-0.39631	0	-0.06709	-0.05213	0.10007	-0.23216	-0.50201	-0.38863	-0.47937
SMSJIPPhospho (STY)VLSHIPLKPLFLFK_3_A0JNW8	-0.23498	0.46877	0.31985	-0.0345	0.13108	-0.22715	0.20511	-0.70733	-0.09426
MFACK1 (Protein N- terminal YSPPCQSSGVIPYIPPhospho (STY)JANAK_2_A1KXE4	-1.87464	0	-1.37948	-1.30774	-1.95653	-1.10051	-1.34311	-1.06509	-2.19997
QNSLGGCCarbamidomethyl (C)GCGCCarbamidomethyl (C)GCGCCarbamidomethyl (STY)GFPEAPP_3_A1L020	0.93537	1.12932	1.04902	1.09606	0.86616	1.16054	0.98343	1.18314	0
ASSIPPhospho (STY)PSLIER_2_A1L170	0.23824	0.42038	-0.00049	-0.17734	0.37494	-0.17732	-0.183	-0.19322	-0.52493
SHSIPPhospho (STY)NPENMVPEPLSGR_2_A1L390	3.49454	3.03115	3.22524	0	2.84747	3.62526	3.51308	3.67061	3.37297

a. **Parameters for ‘Results’.** Herein users can change the parameter “Select one method” on the left panel to check relative result, for example, if users select “zero”, it will show the result derived from zero method:

Step 4: Results and Assessments 

1. Parameters for 'Results'

1.1. Select one method: 



zero

minimum

colmedian

objectmedian

svdmethod

knnmethod

IIs

ml

b. *Parameters for 'Criteria'*. Users can customize the criteria and relative weighting for specific experimental designs and aims. By default, these parameters are not selected and all criteria weights are equal.

2. Parameters for 'Criteria'

2.1. Customize the classic criteria or not?

2.2. Customize the proteomic criteria or not?

b.1 *Customize the classic criteria or not?* If true, users can set the classic criteria and relative weight they want, by default, four classic criteria (NRMSE, SOR, ACC_OI, PSS) are chosen and their weights are equal. Please note, the number of criteria and weights should be equal, for example, if users select 'NRMS', 'SOR', and 'PSS', the weights parameter should be type in '1;1;1', which are separated by semicolons, and in this situation, the three criteria weights are all 0.333 (1/3). If users think 'NRMS' should have a higher weight and type in '3;1;1', this means the weight of 'NRMS' is 0.6 (3/5), 'SOR' and 'PSS' is 0.2 (1/5), respectively:

2. Parameters for 'Criteria'

2.1. Customize the classic criteria or not?

2.1.1. Please select the criterion/criteria you want:

NRMSE SOR ACC_OI PSS

2.1.2. Please set the weighting for each criterion you select:

1;1;1;1

b.2 *Customize the proteomic criteria or not?* If true, users can set the proteomic criteria and relative weight they want, by default, four proteomic criteria (Charge, PepProt, CORUM and PPI) are chosen and their weights are equal. Please also note, the number of criteria and weights should be equal and other descriptions are similar to those for classic criteria as above. Note, the b.1 and b.2 options enable users to customize the criteria and set relative weightings for those specific experimental designs (e.g., a mixture of protein standards being measured in which no in-vivo protein complex formation or interactions expected).

2.2. Customize the proteomic criteria or not?

2.2.1. Please select the criterion/criteria you want:

Charge PepProt CORUM PPI

2.2.2. Please set the weighting for each criterion you select:

1;1;1;1

Especially for type ‘Proteins’ dataset (see part 1 above), Charge and PepProt criteria cannot be used (As there are no information about charges and peptides in the data), so users should change the parameters like this if they decide to customize the proteomic criteria:

2.2. Customize the proteomic criteria or not?

2.2.1. Please select the criterion/criteria you want:

CORUM PPI

2.2.2. Please set the weighting for each criterion you select:

1;1

Next, click “Classic criteria” and “Calculate” button. NAguideR will assess every method under the four classic criteria:

The tables and figures are provided here under the four classic criteria.

1. This table shows the comprehensive ranks of every imputation method. By default, all criteria weights are equal, if users change their weights, and the comprehensive ranks would also change correspondingly based on the new criteria and weights;
- 2-5, the tables show the scores of every imputation method based on 'Normalized root mean squared Error (NRMSE)', 'NRMSE-based sum of ranks (SOR)', 'Procrustes sum of squared errors (PSS)', and 'Average correlation coefficient between original value and imputed value (ACC_OI)', respectively;
6. Figures here show the normalized scores of every imputation method under the four classic criteria. 'Normalized Values' here means that every score is divided by the corresponding max value.

1. Comprehensive ranks under classic criteria:

Methods	NRMSE_Rank	SOR_Rank	ACC_OI_Rank	PSS_Rank	Rank_Mean
Method 2	imposeq	1	1	1	1
Method 3	impseqrob	2	2	2	2
Method 13	segknn	4	3	3	4
Method 10	ml	3	6	5	4.25
Method 4	knnmethod	5	4	4	4.5
Method 5	ls	6	5	6	5.25
Method 6	mice-norm	7	7	7	7
Method 11	objectmedian	8	8	8	8
Method 12	qrilc	9	10	9	9.75
Method 14	svdmethod	10	9	10	10.25
Method 1	colmedian	11	12	11	11
Method 15	zero	12	11	12	11
Method 7	mindet	13	13	13	13
Method 9	minprob	14	14	14	14
Method 8	minimum	15	15	15	15

Showing 1 to 15 of 15 entries

Previous 1 Next

2. Normalized root mean squared Error (NRMSE):

Methods	NRMSE	
Method 11	imposeq	0.07796
Method 12	impseqrob	0.07814
Method 8	ml	0.10625
Method 15	segknn	0.11049
Method 6	knnmethod	0.11513
Method 7	ls	0.1237
Method 13	mice-norm	0.16857
Method 4	objectmedian	0.5063
Method 14	qrilc	0.8632
Method 5	svdmethod	0.93162
Method 3	colmedian	1.00393
Method 1	zero	1.06355
Method 10	mindet	2.2209
Method 9	minprob	2.25375
Method 2	minimum	3.28021

Showing 1 to 15 of 15 entries

3. NRMSE-based sum of ranks (SOR):

Methods	SOR	
Method 11	imposeq	2122
Method 12	impseqrob	2142
Method 15	segknn	3536
Method 6	knnmethod	3625
Method 7	ls	3676
Method 8	ml	3696
Method 13	mice-norm	4296
Method 4	objectmedian	6526
Method 5	svdmethod	8030
Method 14	qrilc	8110
Method 1	zero	8406
Method 3	colmedian	8418
Method 10	mindet	10135
Method 9	minprob	10313
Method 2	minimum	11769

Showing 1 to 15 of 15 entries

4. Procrustes sum of squared errors (PSS):

Methods	PSS	
Method 11	imposeq	0.00048
Method 12	impseqrob	0.00051
Method 8	ml	0.00064
Method 7	ls	0.00094
Method 6	knnmethod	0.00109
Method 15	segknn	0.00129
Method 13	mice-norm	0.00556
Method 4	objectmedian	0.02591
Method 1	zero	0.05313
Method 3	colmedian	0.05389
Method 14	qrilc	0.05468
Method 5	svdmethod	0.06779
Method 10	mindet	0.10707
Method 9	minprob	0.10904
Method 2	minimum	0.13141

Showing 1 to 15 of 15 entries

Previous 1 Next

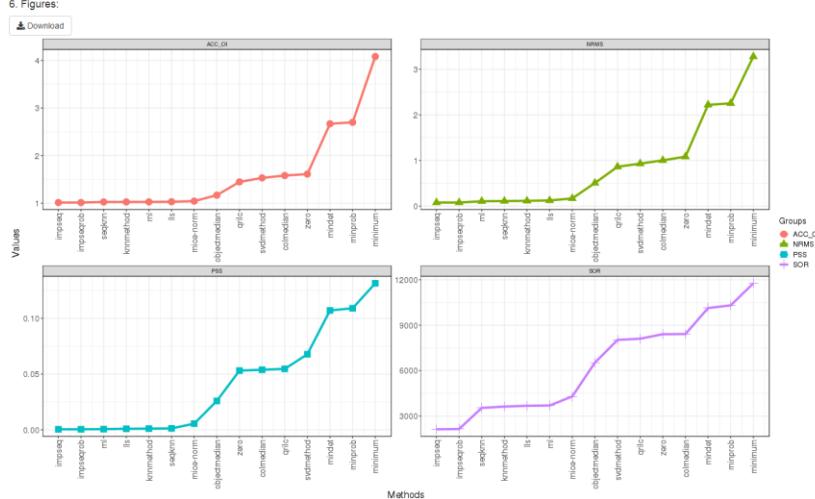
5. Average correlation coefficient between original value and imputed value (ACC_OI):

Methods	Cor_mean	
Method 11	imposeq	0.98755
Method 12	impseqrob	0.98748
Method 15	segknn	0.9797
Method 6	knnmethod	0.975
Method 8	ml	0.97447
Method 7	ls	0.97116
Method 13	mice-norm	0.96947
Method 4	objectmedian	0.8567
Method 14	qrilc	0.69105
Method 5	svdmethod	0.653
Method 3	colmedian	0.63258
Method 1	zero	0.62062
Method 10	mindet	0.37464
Method 9	minprob	0.37038
Method 2	minimum	0.24487

Showing 1 to 15 of 15 entries

Previous 1 Next

6. Figures:



Then click “Proteomic criteria” and “Calculate” button. NAguideR will assess every imputation method under the four proteomic criteria:

The screenshot shows the NAguideR web application. At the top, there's a navigation bar with links like Welcome, Import Data, NA Overview, Methods, Results and Assessments, and Help. Below the navigation, a sub-menu for 'Results and Assessments' is open, showing 'Step 4: Results and Assessments'. This step has several sections: '1. Parameters for 'Results'' (with a dropdown menu set to 'zero'), '2. Parameters for 'Criteria'' (with two checkboxes: '2.1. Customize the classic criteria or not?' and '2.2. Customize the proteomic criteria or not?'), and 'Figure height' (set to 800). In the center, a large orange button labeled 'Calculate' is prominent. To its right, a progress bar with the text 'Calculating.....' is shown. Below the progress bar, five items are listed with download buttons: '1. Comprehensive ranks under proteomic criteria' (Download), '2. Average correlation coefficient between peptides with different charges (ACC_Charge)' (Download), '3. Average correlation coefficient between peptides in a same protein (ACC_PepProt)' (Download), '4. Average correlation coefficient between protein complexes (ACC_CORUM)' (Download), and '5. Average correlation coefficient between protein complexes (ACC_PPI)' (Download). At the bottom right, a small tooltip window displays the text 'Methods for ACC_PepProt mle processing' and 'Calculating each object processing'.

The tables and figures are provided here under the four proteomic criteria.

1. This table shows the comprehensive ranks of every imputation method. By default, all criteria weights are equal, if users change their weights, and the comprehensive ranks would also change correspondingly based on the new criteria and weights;
- 2-5, the tables show the scores of every imputation method based on 'Average correlation coefficient between peptides with different charges (ACC_Charge)', 'Average correlation coefficient between peptides in a same protein (ACC_PepProt)', 'Average correlation coefficient between protein complexes (ACC_CORUM)', 'Average correlation coefficient between protein complexes (ACC_PPI)', respectively;
6. Figures here show the correlation coefficient distribution of the original values and the imputed values from every imputation method under the four proteomic criteria. Figures will be instantly updated for a particular NA method that can be specified in “1.1 Select one method” parameter under Step 4 (left panel). The figure example below shows the results of method “zero”.

1. Comprehensive ranks under proteomic criteria:

Methods	Charge_Rank	PepProt_Rank	CORUM_Rank	PPI_Rank	Rank_Mean
Method 4	knnmethod	2	1	1	2
Method 13	seqknn	1	2	4	1
Method 2	imseq	3	3	2	3
Method 3	imsegrob	4	4	3	4
Method 5	lts	5	5	6	5.25
Method 10	ml	6	6	5	5.75
Method 6	mice-norm	7	7	7	7
Method 11	objectmedian	8	8	8	8
Method 12	qrilc	9	9	9	9.5
Method 14	svdmethod	10	10	10	9.75
Method 1	colmedian	11	11	12	11
Method 15	zero	12	12	11	11.75
Method 7	mindet	13	13	13	13
Method 9	minprob	14	14	14	14
Method 8	minimum	15	15	15	15

Showing 1 to 15 of 15 entries

Previous 1 Next

2. Average correlation coefficient between peptides with different charges (ACC_Charge):

Methods	ACC_Charge
Method 15	0.84803
Method 6	0.84666
Method 11	0.84525
Method 12	0.84508
Method 7	0.84018
Method 8	0.83723
Method 13	0.82996
Method 4	0.73897
Method 14	0.62586
Method 5	0.60933
Method 3	0.59157
Method 1	0.58832
Method 10	0.43458
Method 9	0.42645
Method 2	0.35983

Showing 1 to 15 of 15 entries

Previous 1 Next

3. Average correlation coefficient between peptides in a same protein (ACC_PepProt):

Methods	ACC_peppro
Method 6	0.54688
Method 15	0.54877
Method 11	0.54602
Method 12	0.54588
Method 7	0.54151
Method 8	0.54064
Method 13	0.53333
Method 4	0.47951
Method 14	0.40258
Method 5	0.38689
Method 3	0.37715
Method 1	0.37693
Method 10	0.27806
Method 9	0.27274
Method 2	0.22728

Showing 1 to 15 of 15 entries

Previous 1 Next

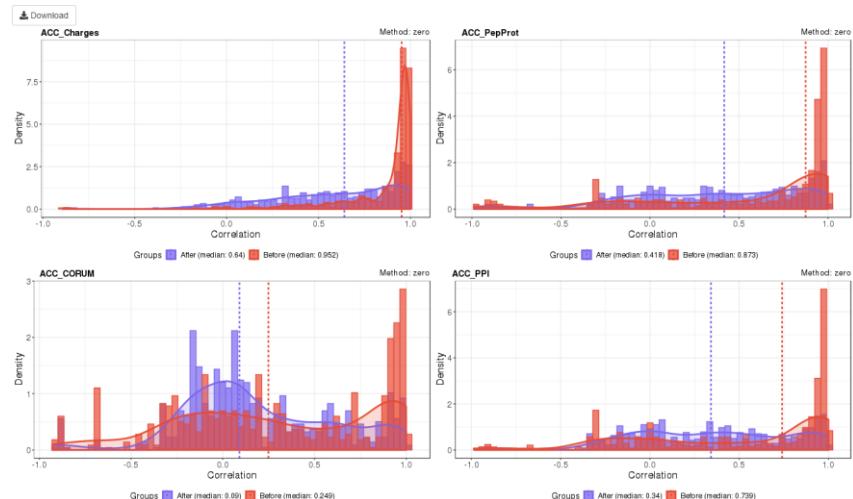
4. Average correlation coefficient between protein complexes (ACC_CORUM):

Methods	ACC_CORUM
Method 6	0.30498
Method 11	0.30475
Method 12	0.30471
Method 15	0.30459
Method 8	0.29933
Method 7	0.29666
Method 13	0.29583
Method 4	0.2485
Method 14	0.21802
Method 5	0.19725
Method 1	0.19269
Method 3	0.18941
Method 10	0.15264
Method 9	0.15054
Method 2	0.127

Showing 1 to 15 of 15 entries

Previous 1 Next

6. Figures:



Next, click ‘Final check’ for checking final imputation results as a summary note. NAguideR will re-check those scores based on every criterion. If everything is acceptable (see below), NAguideR will show a message like:

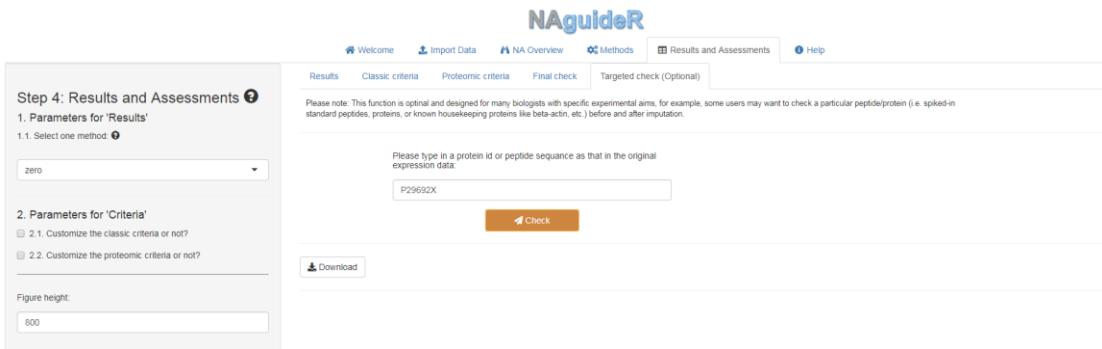
Here NAguideR performs a simple check to report if there is any big difference among these imputation methods under more than half of the criteria (by default, NAguideR check the fold change between the maximum score and the minimum score for each criterion, if the fold change is below 2, a fact suggesting that no big difference under the corresponding criterion, i.e., that NAguideR cannot provide a significantly discriminant guidance on NA method selection), NAguideR will give some warnings and possible solutions for users to review/re-calculate these imputation results:

Last but not least, NaguideR implements one optional function, ‘Targeted check’, which is designed for many biologists with specific experimental aims. For example, this feature conveniently allows users to directly visualize the results of a particular peptide or protein item (i.e., spiked-in standard peptides, proteins, or known housekeeping proteins like beta-actin, etc.). Therefore, by following their experimental design, they can type in the peptide sequence or protein id in the text area and click the ‘Check’ button.

Then, NAguideR will locate this peptide or protein id in the input and resultant matrix (if the id was not listed in the user’s input, it will give a message, “Target protein/peptide not found. Please make sure the item is included in the input table”, example 1 as below). Then, NAguideR will show the results before and after imputation by using bar plots and provide a note “Target protein/peptide was missed in N=X samples among all N=Y samples” (example 2 as below). This plot should help the users to inspect results following their particular experimental design. If the target protein/peptide is quantified without the need of NA imputation, NAguideR will still display the bar plots and provide a

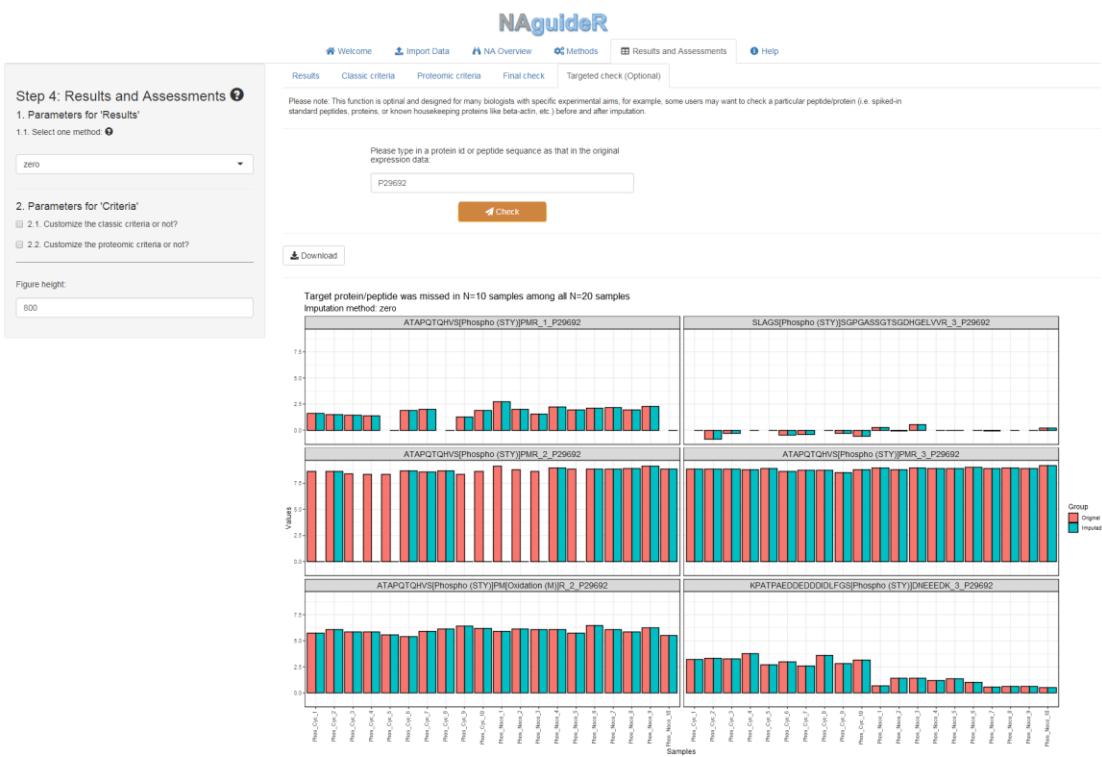
note, "Target protein/peptide was not missed in any sample" (example 3 as below).

Example 1 (Target protein/peptide not found. Please make sure the item is included in the input table):

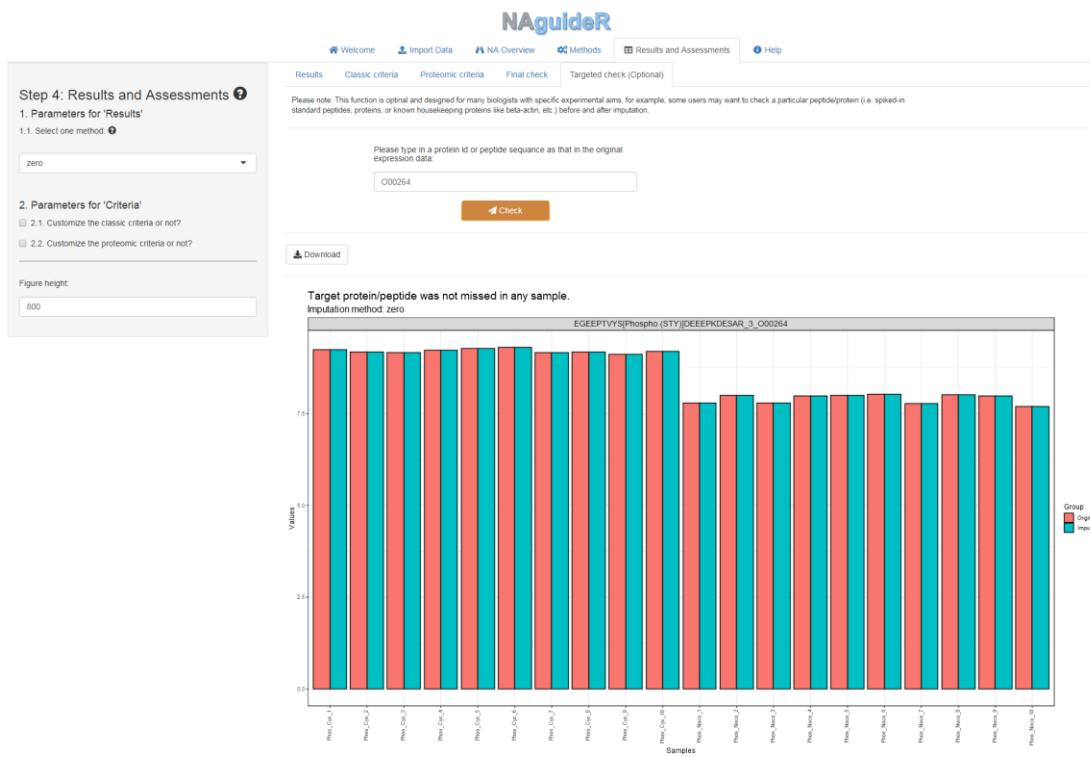


Target protein/peptide not found. Please make sure the item is included in the input table!

Example 2 (Target protein/peptide was missed in N=10 samples among all N=20 samples):



Example 3 (Target protein/peptide was not missed in any sample):



6. Help

This part provides brief introductions and operation manual about NAguideR for users to quickly learn this tool and start to use this tool.

Detailed description

[1. Overview of NAguideR](#)

[2. User manual](#)

1.1 Abstract

Mass-spectrometry (MS) based quantitative proteomics experiments frequently generate data with missing values, which may profoundly affect downstream analyses. A wide variety of missing value imputation methods have been established to deal with the incomplete data. To date, however, there is a scarcity of effective, systematic, and user-friendly tools that are tailored for proteomics community. Herein, we develop a user-friendly and powerful web tool, NAguideR, for the implementation and evaluation of different missing value methods offered by twenty popular missing-value imputation algorithms. Evaluation of data imputation results can be performed through classic computational criteria and, unprecedentedly, proteomic empirical criteria such as quantitative consistency between different charge-states of the same peptide, different peptides belonging to the same protein, and individual proteins participating functional protein complexes. We applied NAguideR into three label-free proteomic datasets featuring peptide-level, protein-level, and phosphoproteomic variables respectively, all generated by data independent mass spectrometry (DIA-MS) with substantial biological replicates. The results indicate that NAguideR is able to discriminate the optimal imputation methods that are facilitating DIA-MS experiments over these sub-optimal and low-performance algorithms. NAguideR web-tool further provides downloadable tables and figures supporting flexible data analysis and interpretation. The flowchart below summarizes the process of data analysis in NREVA.

A: Original intensity data with missing values (NAs)

	A1	A2	A3	A4	A5	...	B1	B2	B3	B4	B5	...
Feature 1	Int	NA	Int	NA	Int	...	Int	Int	NA	NA	Int	...
Feature 2	Int	Int	NA	Int	Int	...	Int	NA	NA	Int	Int	...
...
Feature n-1	Int _{n-1}	Int _{n-1}	NA	Int _{n-1}	Int _{n-1}	...	Int _{n-1}	Int _{n-1}	Int _{n-1}	NA	Int _{n-1}	...
Feature n	NA	Int	Int	Int	Int	...	Int	Int	NA	Int	Int	...

B: Data quality control

The diagram illustrates the relationship between the number of missing values (NA count) and the coefficient of variation (CV). A large blue oval represents the total dataset. Inside it, a red circle represents the NA count, and an orange circle represents the CV. The overlapping area between the two circles indicates where both quality metrics are problematic.

Detailed description

[1. Overview of NAguideR](#)

[2. User manual](#)

2.1 Input data preparation

NAguideR supports four basic file formats (.csv, .txt, .xlsx, .xls). Before analysis, users should prepare two required data: (1) Proteomics expression data and (2) Sample information data. The data required here could be readily generated based on results of several popular tools such as MaxQuant, PEAKS, Spectronaut, and so on. Then can upload the two data into NAguideR with right formats respectively and start subsequent analysis.

2.1.1 Proteomics expression data

There are four types of proteomics expression data supported in NAguideR, among which the main differences are the first few columns.

2.1.1.1 Expression data with peptide sequences, peptide charge status, and protein IDs

In this situation, peptide sequences, peptide charge status, and protein ids are sequentially provided in the first three columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM) or stripped peptides (without PTM). The second column is peptide charge status. The protein ids in the third column should be UniProt ids. From the fourth column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:

Sample names

↓

Peptides	Charges	Uniprot IDs	1	2	3	4	5	6
MLISAVS[Phospho (-2)]	2	A0A9K6	1579936	2053183	3125540	8131965	7195625	7180664
LTPL[Phospho (-2)]	2	A0A9K6	712596	2744010	4998674	1119321	1119320	41120046
EPT[Phospho (-2)]	3	A0P28	52089	12899	45899	86268	82298	5645625
SSSSLAS[Phospho (-2)]	2	A0PGR8	74890	928081	828022	948827	9111544	580044
SSSSLAS[Phospho (-2)]	2	A0PGR8	34338	8628903	73NA	30830	6833394	47NA
TQGPWPPFTPSDSD[Phospho (-2)]	3	A0JLT2	221698	9NA	270359	6312614	6345215	3284286
SMS[Phospho (-2)]	3	A0JNW5	248274	4278773	358461	316457	7352716	8285275
SMS[Phospho (-2)]	3	A0JNW5	79674	09NA	110380	513927	482461	96155724
QEDL[Protein]	2	A1K84	50000	50000	50000	50000	50000	50000
AAI[Phospho (-2)]	2	A1L170	344653	841764	5287094	1287627	31417670	3229501
SRS[Phospho (-2)]	2	A1L390	3293265	2527386	2685655	NA	2318149	4120553
GFLS[Phospho (-2)]	2	A1L390	1551857	1596314	1253887	1406729	1723560	1502006
IWCMESSCGC[Phospho (-2)]	2	A1L390	686212.1	703314.1	697561	580441.7	808891.8	745552.6
SRS[Phospho (-2)]	3	A1L390	1A	604569.9	NA	784035	554084.5	NA
SFLS[Phospho (-2)]	2	A1L390	833284	31934303	797986.1	7146361	99901.5	1039246
S1P[Phospho (-2)]	2	A1L390	807754.1	825141.1	840387.5	845764.4	89854	913974
SSSVL5[Phospho (-2)]	2	A1L390	729638	795307.5	637437.1	714943.1	806124.7	818071.4
EEE[Phospho (-2)]	2	A1L390	386283.1	NA	371454.2	364010.8	415588.1	373595.6

7. How to run this tool locally?

NAguideR is an open source software for non-commercial use and all codes can be obtained on our GitHub: <https://github.com/wangshisheng/NAguideR>. If users want to run NAguideR on their own computer, they should operate as below:

7.1 As this tool was developed with R, you may:

- a) Install R. You can download R from here: <https://www.r-project.org/>.
- b) Install RStudio. (Recommenatory but not necessary). You can download RStudio from here: <https://www.rstudio.com/>.
- c) Check packages. After installing R and RStudio, you should check whether you have installed these packages (shiny, shinyBS, shinyjs, shinyWidgets, DT, gdata, ggplot2, ggsci, openxlsx, data.table, DT, raster, Metrics, vegan, tidyverse, ggExtra, cowplot, Amelia, e1071, impute, SeqKnn, pcaMethods, norm, imputeLCMD, VIM, rrcovNA, mice, missForest, GMSimpute, DreamAI). You may run the codes below to check them:

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(shiny, shinyBS, shinyjs, shinyWidgets, DT, gdata, ggplot2, ggsci, openxlsx,
data.table, DT, raster, Metrics, vegan, tidyverse, ggExtra, cowplot, Amelia, e1071, impute,
SeqKnn, pcaMethods, norm, imputeLCMD, VIM, rrcovNA, mice, missForest, GMSimpute,
DreamAI)
```

Please note, you may find the SeqKnn package (<https://github.com/cran/SeqKnn>) cannot be installed rightly as it has not been updated for a long time. If so, please download this package from here:

https://github.com/wangshisheng/NAguideR/blob/master/SeqKnn_1.0.1.tar.gz. Then you can install this separate package locally:

```
setwd('path') #path is where the two packages are.
install.packages("SeqKnn_1.0.1.tar.gz", repos = NULL, type = "source")
```

d) Run this tool locally

```
if(!require(NAguideR)) devtools::install_github("wangshisheng/NAguideR")
library(NAguideR)
NAguideR_app()
```

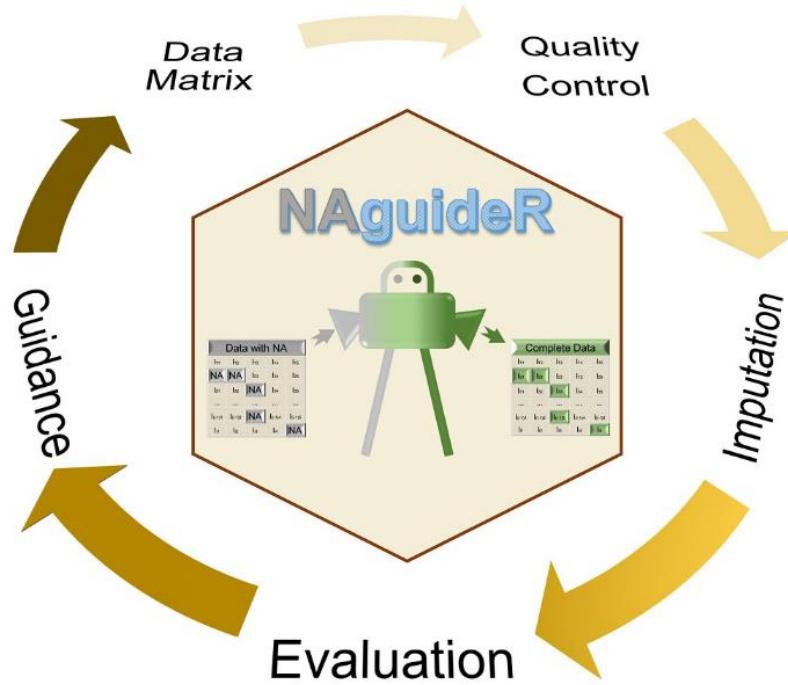
Then NAguideR will be started as below, and the detailed operation about NAguideR can be found in the Supplementary Notes part 1-6:

NAguideR

Welcome Import Data NA Overview Methods Results and Assessments Help

~~ Dear Users, Welcome to NAguideR ~~

NAguideR is a web-based tool, which integrates 23 commonly used missing value imputation methods and provides two categories of evaluation criteria (4 classic criteria and 4 proteomic criteria) to assess the imputation performance of various methods. We hope this tool could help scientists impute the missing values systematically and present valuable guidance to select one proper method for their own data. In addition, this tool supports both online access and local installation.



Basically, there are four main steps in NAguideR:

1. Uploading proteomics expression data and sample information data;
2. Data quality control;
3. Missing value imputation;
4. Performance evaluation;

After this, NAguideR can provide valuable guidance for users to select one proper method for their own data based on the evaluation results. Detailed introduction can be found in the **Help** part.

Finally, NAguideR is developed by R shiny (Version 1.3.2), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at wsslearning@omicsolution.com.

^_^ Enjoy yourself in NAguideR ^_^

III. Reference

1. Tyanova, S., Temu, T. and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*, **11**, 2301-2319.
2. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM*, **17**, 2337-2342.
3. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinović, S.M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y. and Escher, C. (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics*, **14**, 1400-1410.
4. Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S. and Ralser, M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods*, **17**, 41-44.
5. Rost, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmstrom, J., Malmstrom, L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology*, **32**, 219-223.
6. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinovic, S.M., Cheng, L.Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C. *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & cellular proteomics : MCP*, **14**, 1400-1410.
7. Bruderer, R., Bernhardt, O.M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D. and Reiter, L. (2017) Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & cellular proteomics : MCP*, **16**, 2296-2309.