#1

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 3:04:59 PM
Last Modified: Monday, December 04, 2017 3:19:22 PM
Time Spent: 00:14:22
IP Address: 69.204.40.71

Page 1

Q1 What is your team name?

hyu

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Statistics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

University at Buffalo

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: standard scaling
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): Pathways: KEGG; Interaction Networks
Q15 Describe your method. (Choose all that apply)	**Tree based,** **Penalized regression ,** **Ensemble**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Predictor variables: KNN
Response variables: not imputed

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#2

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 3:47:16 PM
Last Modified: Monday, December 04, 2017 3:55:38 PM
Time Spent: 00:08:21
IP Address: 12.42.251.74

Page 1

Q1 What is your team name?

Bitsplease

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**BA/BS**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**BA/BS**	**Biology**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

New York

Q7 Describe your method. (Choose all that apply)

Prediction model --- Linear regression (i.e. LASSO regression)

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**Yes**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#3

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 4:06:29 PM
Last Modified: Monday, December 04, 2017 4:11:39 PM
Time Spent: 00:05:10
IP Address: 47.154.228.40

Page 1

Q1 What is your team name?

Roland Luethy

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Los Angeles, CA

Q7 Describe your method. (Choose all that apply)

Prediction model --- non linear regression (i.e. Tree model)

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Tree based,** **Ensemble**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNA, protein
Q20 Did you normalize or standardize the data before carrying any analysis?	**No**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Tree based,** **Ensemble**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

independently

#4

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 4:35:56 PM
Last Modified: Monday, December 04, 2017 5:35:45 PM
Time Spent: 00:59:49
IP Address: 67.194.232.73

Page 1

Q1 What is your team name?

Hongyang Li and Yuanfang Guan

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

University of Michigan - Ann Arbor, MI

Q7 Describe your method. (Choose all that apply)

Constant imputation

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

very simple - a few seconds on one CPU

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: For breast, we use both RNAseq and CNV.
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Quantile normalization
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): For ovary, we use CPTAC colorectal cancer data (but the improvement is very minimal).
Q15 Describe your method. (Choose all that apply)	**Non-Linear,** **Tree based,** **Ensemble,** **Use input feature as proxy (e.g use mRNA as protein abundance)**

Q16 What is the computational complexity of your method (time and nodes) ?

It takes about 5 mins on 14 CPU nodes if models are pre-trained.

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

predictor variable: we exclude genes with missing values
response variable: we exclude samples with missing values to train our models

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
Proteomics and RNAseq

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Quantile normalization. Proteomics data are normalized by sample mean and sd.

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Non-Linear,

Tree based,

Ensemble,

Use input feature as proxy (e.g use mRNA as protein abundance)

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

predictor variable: we exclude genes with missing values
response variable: we exclude samples with missing values to train our models

Q24 What is the computational complexity of your method (time and nodes) ?

It takes about 5 mins on 14 CPU nodes if models are pre-trained.

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Train each site separately and add weighted average predictions of all sites from the same protein.

#5

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 5:35:07 PM
Last Modified: Monday, December 04, 2017 5:45:10 PM
Time Spent: 00:10:02
IP Address: 137.82.157.80

Page 1

Q1 What is your team name?

Manuel Belmadani

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Vancouver, Canada, University of British Columbia

Q7 Describe your method. (Choose all that apply)

Other (please specify):

I tried to use the imp4p R package but it did not terminate fast enough. I did not put much time here.

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**Yes**

Q11 What is the computational complexity of your method (time and nodes) ?

I would have to read the imp4p documentation, though it doesn't seem published yet.

Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Per sample normalization to centre distribution mean to 0
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): Sequence specific features (Amino acid counts, CDS length), GTEX data, COSMIC genes
Q15 Describe your method. (Choose all that apply)	**Non-Linear,** **Tree based,** **Penalized regression**

Q16 What is the computational complexity of your method (time and nodes) ?

Assuming pearson test is O(n), and a constant population size, O(n*m + n*n) for one generation.

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

I simply filled in missing values with mean of other gene values.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**1**
CNA	**3**

Q19 Did you use multiple data types to make your predictions?

No

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Same method as in 2

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
Same method as in 2

Q22 Describe your method. (Choose all that apply)

Non-Linear,

Tree based,

Penalized regression

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Same method as in 2

Q24 What is the computational complexity of your method (time and nodes) ?

Same method as in 2

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Did not get this far.

#6

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 7:33:07 PM
Last Modified: Monday, December 04, 2017 7:35:54 PM
Time Spent: 00:02:47
IP Address: 76.181.209.239

Page 1

Q1 What is your team name?

PersianGulf

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly. **Sc1**

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Ohio University

Q7 Describe your method. (Choose all that apply)	**Constant imputation** , Other (please specify): Stacking model
Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

N^2

Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#7

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 8:37:14 PM
Last Modified: Monday, December 04, 2017 8:42:51 PM
Time Spent: 00:05:36
IP Address: 190.195.13.231

Page 1

Q1 What is your team name?

martinguerrero89

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**BA/BS**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Biology**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

IMBECU-CONICET Mendoza, Argentina

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear,** **Penalized regression** , **Multivariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

24hs 8 threads, core i7 6th generation 4GHz, Personal computer

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

a missing value was considered as not expressed in both RNA data and Proteomic data. Since data is already scaled a missing value was replaced with the minor value for the gene across all samples minus an empirically chosen factor of 0.1.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#8

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 11:53:30 PM
Last Modified: Tuesday, December 05, 2017 12:02:00 AM
Time Spent: 00:08:29
IP Address: 103.25.231.102

Page 1

Q1 What is your team name?

sherry Bhalla

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**Others**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

IIIT Delhi

Q7 Describe your method. (Choose all that apply)

Prediction model --- non linear regression (i.e. Tree model)

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

one node and 30 minutes

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: gene expression and CNA
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Non-Linear,** **Multivariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

1 node and 20 minutes

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We imputed mean values in case of missing values

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
protein expression,gene expression (rnaseq) and CNA

Q20 Did you normalize or standardize the data before carrying any analysis?

No

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Non-Linear,

kernel method ,

Multitask,

Multivariate regression

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We took mean value of the protein across the given samples

Q24 What is the computational complexity of your method (time and nodes) ?

aprox 1 hour and 1 node

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

we took protein, Rnaseq and cNA whereever it was available aling with other most correlated 97 features. Where protein expression was not availabe we took 100 most correlated features

#9

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, December 04, 2017 11:38:34 PM
Last Modified: Tuesday, December 05, 2017 12:08:44 AM
Time Spent: 00:30:10
IP Address: 103.25.231.101

Page 1

Q1 What is your team name?

Raghava India

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**Others**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

IIIT Delhi, India

Q7 Describe your method. (Choose all that apply)

Prediction model --- non linear regression (i.e. Tree model)

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

Approximately 20 minutes

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: mRNA (RNA seq), CNA
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Multivariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

1 node, 20-30 minutes

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Mean based imputation

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: mRNA (RNA seq), CNA, Protein
Q20 Did you normalize or standardize the data before carrying any analysis?	**No**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Multivariate regression**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Protein based mean imputation

Q24 What is the computational complexity of your method (time and nodes) ?

1 Node, Approximately 60 minutes (1Hour)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

We have predicted each phosphosite of each protein using its corresponding protein expression, mRNA and CNA and 97 other highly correlated features.

#10

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 3:40:08 AM
Last Modified: Tuesday, December 05, 2017 4:12:56 AM
Time Spent: 00:32:47
IP Address: 157.25.120.74

Page 1

Q1 What is your team name?

Ardigen

Q2 How many people made up your team?

4

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Others**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Kraków, Poland

Q7 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q8 Did you use iterative algorithm to update imputation in your model? **Respondent skipped this question**

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: mRNA (RNA seq), CNA
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): List of transcription factors from: http://www.tfcheckpoint.org/index.php/introduction
Q15 Describe your method. (Choose all that apply)	**Linear,** **Ensemble,** **Multivariate regression** , **Univariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

12-24 hours on 4 cores (on a laptop with i7-6820HQ)

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We used mean imputation for predictor variables.
We did not use imputation of response variables (we discarded observations with a missing response variable).

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
mRNA (RNA seq), CNA, protein

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Each predictor variable was normalized before regression by subtracting the mean and dividing by the L2-norm

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
List of transcription factors from:
http://www.tfcheckpoint.org/index.php/introduction

Q22 Describe your method. (Choose all that apply)

Linear,

Penalized regression ,

Ensemble,

Multivariate regression ,

Univariate regression

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We used mean imputation for predictor variables.
We did not use imputation of response variables (we discarded observations with a missing response variable).

Q24 What is the computational complexity of your method (time and nodes) ?

around 12 hours on 4 nodes (on a laptop with i7-6820HQ)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

independently

#11

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 5:15:33 AM
Last Modified: Tuesday, December 05, 2017 5:26:27 AM
Time Spent: 00:10:54
IP Address: 130.232.69.182

Page 1

Q1 What is your team name?

Aboensis II

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Mathematics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Turku, Finland

Q7 Describe your method. (Choose all that apply)	**Constant imputation** , Other (please specify): LLS
Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

~10h, 10 CPUs

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNA, CNA
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Median normalization as in doi: 10.1016/j.cell.2016.05.069
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear**

Q16 What is the computational complexity of your method (time and nodes) ?

~3h, single CPU

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

predictor: LLS imputation
response: filtering

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
CNA, RNA, Protein

Q20 Did you normalize or standardize the data before carrying any analysis?

No

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Linear,

Tree based,

Penalized regression ,

Ensemble

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Respondent skipped this question

Q24 What is the computational complexity of your method (time and nodes) ?

~1h, single CPU

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Same value for all

#12

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 8:13:42 AM
Last Modified: Tuesday, December 05, 2017 8:17:41 AM
Time Spent: 00:03:59
IP Address: 171.61.76.90

Page 1

Q1 What is your team name?

Alfajiri

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Statistics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Bangalore, India

Q7 Describe your method. (Choose all that apply)

Other (please specify):
MICE

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

-

Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear,** **Penalized regression** , **Multivariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

-

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Constant imputation

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: Copy number and proteomics datasets
Q20 Did you normalize or standardize the data before carrying any analysis?	**No**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Linear,** **Penalized regression** , **Multivariate regression**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Constant imputation

Q24 What is the computational complexity of your method (time and nodes) ?

-

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**3**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

independently

#13

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 10:52:28 AM
Last Modified: Tuesday, December 05, 2017 10:58:23 AM
Time Spent: 00:05:55
IP Address: 124.110.62.173

Page 1

Q1 What is your team name?

NGT

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**BA/BS**	**Biology**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Niigata, Japan

Q7 Describe your method. (Choose all that apply)

Constant imputation

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

N

Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: 2 ** x
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Non-Linear**

Q16 What is the computational complexity of your method (time and nodes) ?

N?

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

2 ** x => 0

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**1**
CNA	**3**

Q19 Did you use multiple data types to make your predictions?	**No**
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: 2 ** x
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Non-Linear**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

2 ** x => 0

Q24 What is the computational complexity of your method (time and nodes) ?

N?

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**3**
mRNA (microarray, moresamples)	**2**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Neural Nets

#14

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 12:15:15 PM
Last Modified: Tuesday, December 05, 2017 12:23:42 PM
Time Spent: 00:08:27
IP Address: 153.1.76.192

Page 1

Q1 What is your team name?

TamBio

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Tampere, Finland

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear,** **Penalized regression**

Q16 What is the computational complexity of your method (time and nodes) ?

<2 hours, 5 nodes for training. <5 minutes for prediction

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#15

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 12:23:00 PM
Last Modified: Tuesday, December 05, 2017 12:36:31 PM
Time Spent: 00:13:30
IP Address: 108.216.109.223

Page 1

Q1 What is your team name?

GACT-NCI

Q2 How many people made up your team?

Four

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

United State

Q7 Describe your method. (Choose all that apply)

Other (please specify):
Chained Random Forest

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

Node=1, ~30minutes

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNAseq and CNV
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear**

Q16 What is the computational complexity of your method (time and nodes) ?

node=1, ~1 hour

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Imputation using 'impute' or 'missRanger' packages

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNAseq, CNV and protein
Q20 Did you normalize or standardize the data before carrying any analysis?	**No**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Linear**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Imputation using 'impute' and 'missRanger' packages.

Q24 What is the computational complexity of your method (time and nodes) ?

node=1, time ~2 hours

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Each phosphosite belonging to a given protein was predicted separately by taking the parent protein/gene as reference.

#16

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, December 05, 2017 2:13:18 PM
Last Modified: Tuesday, December 05, 2017 2:20:26 PM
Time Spent: 00:07:08
IP Address: 88.232.169.252

Page 1

Q1 What is your team name?

ABU_Proteogenomics

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**BA/BS**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**MS**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Antalya Turkey

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNA-Seq and CNV
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: We standardized the data to have zero mean and unit variance
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): We downloaded the CPTAC proteome data from Zhang et al paper supplimentary
Q15 Describe your method. (Choose all that apply)	**Linear,** **Use input feature as proxy (e.g use mRNA as protein abundance)**

Q16 What is the computational complexity of your method (time and nodes) ?

n2

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We filled the missing values in the protein expression data by replacing them with the column-wise minimum value – 0.01. We assumed that missing values in protein expression data correspond to cases where the corresponding protein is not expressed in detectable levels.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions? **Respondent skipped this question**

Q20 Did you normalize or standardize the data before carrying any analysis? **Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases) **Respondent skipped this question**

Q22 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively. **Respondent skipped this question**

Q24 What is the computational complexity of your method (time and nodes) ? **Respondent skipped this question**

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them) **Respondent skipped this question**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide? **Respondent skipped this question**

#17

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Wednesday, December 06, 2017 2:51:03 AM
Last Modified: Wednesday, December 06, 2017 5:22:06 AM
Time Spent: 02:31:03
IP Address: 108.173.249.52

Page 1

Q1 What is your team name?

UoA

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

University of Alberta, Edmonton, Canada

Q7 Describe your method. (Choose all that apply)

Constant imputation ,

Prediction model --- Linear regression (i.e. LASSO regression)

,

Other (please specify):

Middle value imputation by weighting the 'true' mean with percentage of non-zero missing values

Q8 Did you use iterative algorithm to update imputation in your model?

No

Q9 Does your method provide confidence scores for the imputed values?

No

Q10 Does your method treat biological missing and technical missing differently?

Yes

Q11 What is the computational complexity of your method (time and nodes) ?

Simple model. 5 -10 mins for each datamatrix.

Q12 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
RNA,
CNA

Q13 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Standardization (z-score)

Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
Gene-Ontology (GO)

Q15 Describe your method. (Choose all that apply)

Linear,

Penalized regression ,

Ensemble,

Multivariate regression

Q16 What is the computational complexity of your method (time and nodes) ?

2-4 days to train on multicore (8) CPU. 20 to 40 min to run the model.

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

- Missing values in protein (response variable) were just ignored (Those instances were discarded during training for that particular protein.)

- For predictor variables (RNA and CNA) - complete(filtered) dataset was used for training. None were missing. If any feature values were found to be missing during performance of model, they were imputed by simple mean.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**1**
CNA	**3**

Q19 Did you use multiple data types to make your predictions? **Respondent skipped this question**

Q20 Did you normalize or standardize the data before carrying any analysis? **Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases) **Respondent skipped this question**

Q22 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively. **Respondent skipped this question**

Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#18

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Wednesday, December 06, 2017 12:04:41 PM
Last Modified: Wednesday, December 06, 2017 4:33:27 PM
Time Spent: 04:28:46
IP Address: 176.63.29.212

Page 1

Q1 What is your team name?

NPG-MGL

Q2 How many people made up your team?

4

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Budapest, Hungary

Q7 Describe your method. (Choose all that apply)	**Prediction model --- clustering method (i.e. , KNN)** **Prediction model --- Linear regression (i.e. LASSO regression)**
Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**Yes**

Q11 What is the computational complexity of your method (time and nodes) ?

O(P*N)

Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: CNA, RNA-seq, microarray
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: for each feature (genes) as an option 0 mean and 1 std
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Non-Linear,** **kernel method ,** **Neural network,** **Penalized regression**

Q16 What is the computational complexity of your method (time and nodes) ?

O((P+M)*N) P:number of samples, N:number of inputs, M:number of outputs

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

predictors (RNA) were imputed with constant values (-16), response variables were imputed by KNN

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
CNA, RNA-seq, microarray, protein

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
for each feature (genes) as an option, 0 mean and 1 std

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Non-Linear,

kernel method ,

Neural network,

Penalized regression

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

predictors and response variables were imputed together by KNN

Q24 What is the computational complexity of your method (time and nodes) ?

O((P+M)*N) P:number of samples, N:number of inputs, M:number of outputs

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

separately

#19

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Thursday, December 07, 2017 2:43:48 AM
Last Modified: Thursday, December 07, 2017 3:19:29 AM
Time Spent: 00:35:41
IP Address: 163.152.20.205

Page 1

Q1 What is your team name?

DMIS-PTG

Q2 How many people made up your team?

8

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Korea University

Q7 Describe your method. (Choose all that apply)

Low rank matrix completion

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	If Yes, please briefly describe the procedure used to generate the confidence score: : We selected random 3 validation samples per proteins and calculated a correlation coefficient between prediction and true value of validation points. we calculated correlation for every proteins and samples, and averaged corresponding sample's correlation and protein's correlation.
Q10 Does your method treat biological missing and technical missing differently?	**No**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: CNA, Gene Expression
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: The organizer provided the gene expression data with two types(microarray, rna-seq). We selected the microarray data because our model showed the better correlation when using it. The problem was the test data only gave rna-seq values, so it was necessary to rescale the microarray values to rna-seq values. Our solution was applying Z-normalization to both gene expression values.
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): BioGRID(Protein-Interaction network), BTNET(Gene regulatory network), Protein-complex information(CORUM)
Q15 Describe your method. (Choose all that apply)	**Linear,** **Penalized regression ,** **Multivariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

We made a model for each protein. When we trained the all models at the same time with 10 cpu cores, it took around 10 minutes to train.

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We imputed the missing values with averaging all the existing values of the corresponding feature.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**1**
CNA	**2**

Q19 Did you use multiple data types to make your predictions? **No**

Q20 Did you normalize or standardize the data before carrying any analysis? **No**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases) **No**

Q22 Describe your method. (Choose all that apply) **Linear, kernel method, Multivariate regression**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We imputed the missing values with the mean values of corresponding protein.

Q24 What is the computational complexity of your method (time and nodes) ?

We build a model for each phosphoprotein.

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

We ignored the protein/peptide information of each phosphosites. We consider each phosphosite as an independent target.

#20

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, December 17, 2017 3:03:18 AM
Last Modified: Sunday, December 17, 2017 3:15:07 AM
Time Spent: 00:11:48
IP Address: 108.172.104.198

Page 1

Q1 What is your team name?

Jubilee

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Public health agency of Canada, Winnipeg, Canada

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: Amino acid sequence properties, codon properties, RNA-seq and CNV
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): Amino acid sequence properties, codon properties
Q15 Describe your method. (Choose all that apply)	**Tree based,** **Multitask,** **Ensemble,** **Use input feature as proxy (e.g use mRNA as protein abundance)**

Q16 What is the computational complexity of your method (time and nodes) ?

For building the models, if with one node, ~6hours

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

As I construct a prediction model per sample and had thousands of proteins as instances, I screened out those proteins with missing data in either input features or response variable.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions? **Respondent skipped this question**

Q20 Did you normalize or standardize the data before carrying any analysis? **Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases) **Respondent skipped this question**

Q22 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively. **Respondent skipped this question**

Q24 What is the computational complexity of your method (time and nodes) ? **Respondent skipped this question**

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them) **Respondent skipped this question**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide? **Respondent skipped this question**

#21

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, December 17, 2017 1:36:00 PM
Last Modified: Sunday, December 17, 2017 1:42:03 PM
Time Spent: 00:06:02
IP Address: 192.88.140.108

Page 1

Q1 What is your team name?

Stargazer

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

New Mexico State University

Q7 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q8 Did you use iterative algorithm to update imputation in your model? **Respondent skipped this question**

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear,** **Univariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

polynomial (product of the number of proteins and the number of possible predictors)

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

missing values are imputed by the mean values of the same gene across all other samples

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**2**
CNA	**3**

Q19 Did you use multiple data types to make your predictions?	**No**

Q20 Did you normalize or standardize the data before carrying any analysis?	**No**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q22 Describe your method. (Choose all that apply)	**Linear,** **Univariate regression**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

the missing values of proteins are imputed by applying the prediction model in SC2

Q24 What is the computational complexity of your method (time and nodes) ?

polynomial ((product of the number of predicted variables and the number of possible predictors)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

N/A

#22

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, December 17, 2017 3:43:58 PM
Last Modified: Sunday, December 17, 2017 3:58:04 PM
Time Spent: 00:14:05
IP Address: 88.10.213.7

Page 1

Q1 What is your team name?

CGATeam

Q2 How many people made up your team?

6

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Valencia, Spain

Q7 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q8 Did you use iterative algorithm to update imputation in your model? **Respondent skipped this question**

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Gene expression data was normalized using quantiles
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): Pathways: KEGG
Q15 Describe your method. (Choose all that apply)	**Use input feature as proxy (e.g use mRNA as protein abundance)**

Q16 What is the computational complexity of your method (time and nodes) ?

It is quite fast and do not require too much memory

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Missing values in the input files were assigned a 0.5 expression level.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**2**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**No**

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Gene expression data was normalized using quantiles

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
Pathways: KEGG

Q22 Describe your method. (Choose all that apply)

Use input feature as proxy (e.g use mRNA as protein abundance)

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Missing values in the input files were assigned a 0.5 expression level.

Q24 What is the computational complexity of your method (time and nodes) ?

It is quite fast and do not require too much memory

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**2**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Our method infers protein activity status in a global manner, it can not difference between different phosphosites in the same protein

#23

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Wednesday, December 20, 2017 5:49:04 AM
Last Modified: Wednesday, December 20, 2017 5:56:23 AM
Time Spent: 00:07:18
IP Address: 61.216.1.86

Page 1

Q1 What is your team name?

BiSheng

Q2 How many people made up your team?

8

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Taipei, Taiwan

Q7 Describe your method. (Choose all that apply)

Prediction model --- Comprehensive model(i.e. Neural Network)

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

time

Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#24

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Friday, December 22, 2017 2:03:52 AM
Last Modified: Friday, December 22, 2017 7:01:11 AM
Time Spent: 04:57:18
IP Address: 61.50.104.82

Page 1

Q1 What is your team name?

Stanford BMIR

Q2 How many people made up your team?

5

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Stanford University

Q7 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q8 Did you use iterative algorithm to update imputation in your model? **Respondent skipped this question**

Q9 Does your method provide confidence scores for the imputed values?

Respondent skipped this question

Q10 Does your method treat biological missing and technical missing differently?

Respondent skipped this question

Q11 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q12 Did you use multiple data types to make your predictions?

No

Q13 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
z-score
normalization

Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q15 Describe your method. (Choose all that apply)

Non-Linear,

Neural network

Q16 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Apply singular-value decomposition to fill missing values

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?

No

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
z-score
normalization

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Non-Linear,

Neural network

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Apply singular-value decomposition to fill missing values

Q24 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Respondent skipped this question

#25

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Friday, December 22, 2017 4:10:20 PM
Last Modified: Friday, December 22, 2017 4:24:40 PM
Time Spent: 00:14:20
IP Address: 128.42.39.198

Page 1

Q1 What is your team name?

Big-S2-2017-proteogenomic

Q2 How many people made up your team?

6

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Statistics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Statistics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Houston and Chapel Hill

Q7 Describe your method. (Choose all that apply)

Low rank matrix completion

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**Yes**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: Genotype and RNA-seq data
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**
Q15 Describe your method. (Choose all that apply)	**Linear,** **Non parametric ,** **Univariate regression**

Q16 What is the computational complexity of your method (time and nodes) ?

about15-30 minutes for each model, using one node

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

The missing values were imputed using our method in sub-challenge 1.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
Genotype, RNA-seq, and proteomics data

Q20 Did you normalize or standardize the data before carrying any analysis?

No

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Non parametric ,
Univariate regression ,
Linear

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

The missing values were imputed using our method in sub-challenge 1.

Q24 What is the computational complexity of your method (time and nodes) ?

about 30 minutes for each model, using one node

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**N/A**
mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

By using the residuals after regressing on its corresponding protein.

#26

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, December 31, 2017 11:49:02 AM
Last Modified: Sunday, December 31, 2017 11:57:42 AM
Time Spent: 00:08:40
IP Address: 75.166.118.250

Page 1

Q1 What is your team name?

jjacob_cub

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Others**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Boulder, CO

Q7 Describe your method. (Choose all that apply)

Other (please specify):

RIDGE regression

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	If Yes, please briefly describe the procedure used to generate the confidence score: : Imputed values are confidence ranked by inverses coefficient of variation for each row. Any estimate lower than 0.5 is given a confidence score of 0.5, equal to chance.
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

~O(C^2*N) where C is number of columns, and N is number of rows. In addition to regression complexity there are other factors so this is very much approximate.

Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**

Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#27

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Tuesday, January 02, 2018 9:59:46 AM
Last Modified: Tuesday, January 02, 2018 10:04:55 AM
Time Spent: 00:05:08
IP Address: 134.174.140.32

Page 1

Q1 What is your team name?

LSP

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Harvard Medical School

Q7 Describe your method. (Choose all that apply) **Respondent skipped this question**

Q8 Did you use iterative algorithm to update imputation in your model? **Respondent skipped this question**

Q9 Does your method provide confidence scores for the imputed values?

Respondent skipped this question

Q10 Does your method treat biological missing and technical missing differently?

Respondent skipped this question

Q11 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q12 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
CNV,
mRNA

Q13 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Quantile normalization to a reference (exponential) distribution

Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
PathwayCommons

Q15 Describe your method. (Choose all that apply)

Non-Linear,

Tree based,

Ensemble

Q16 What is the computational complexity of your method (time and nodes) ?

Low to medium

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Imputed as median value

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
CNV, mRNA, Protein abundance

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
Quantile normalization to a reference (exponential) distribution

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
PathwayCommons,
PhosphoSitePlus

Q22 Describe your method. (Choose all that apply)

Non-Linear,

Tree based,

Ensemble

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Imputed as median value

Q24 What is the computational complexity of your method (time and nodes) ?

Low to medium

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Trained a separate model for each

#28

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Friday, January 19, 2018 4:14:25 PM
Last Modified: Friday, January 19, 2018 4:21:00 PM
Time Spent: 00:06:34
IP Address: 171.66.213.131

Page 1

Q1 What is your team name?

MolProLab

Q2 How many people made up your team?

2

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Others**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Biology**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Stanford University

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	**No**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): GO
Q15 Describe your method. (Choose all that apply)	**Penalized regression**

Q16 What is the computational complexity of your method (time and nodes) ?

Did not investigate - used glmnet off the shelf

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

k nearest neighbor imputation

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**

Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#29

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Friday, January 19, 2018 4:33:37 PM
Last Modified: Friday, January 19, 2018 4:40:40 PM
Time Spent: 00:07:03
IP Address: 170.54.61.190

Page 1

Q1 What is your team name?

pins

Q2 How many people made up your team?

4

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Hyderabad, India

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: RNA, CNA and protein
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Quantile normalization

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Multivariate regression

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

knn based imputation.

Q24 What is the computational complexity of your method (time and nodes) ?

5 minutes on 10 nodes with 32Gb RAM

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**2**
mRNA (RNAseq, less samples)	**1**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Multivariate adaptive regression splines

#30

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Friday, January 19, 2018 8:18:48 PM
Last Modified: Friday, January 19, 2018 8:32:08 PM
Time Spent: 00:13:19
IP Address: 115.64.200.224

Page 1

Q1 What is your team name?

Qut students

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**MS**	**Computer Science**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Queensland University of Technology, Brisbane, Australia

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?

Respondent skipped this question

Q10 Does your method treat biological missing and technical missing differently?

Respondent skipped this question

Q11 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q12 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
We augmented the data with 3-mer and 5-mer counts.

Q13 Did you normalize or standardize the data before carrying any analysis?

No

Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q15 Describe your method. (Choose all that apply)

Non-Linear,

Multivariate regression ,

Use input feature as proxy (e.g use mRNA as protein abundance)

Q16 What is the computational complexity of your method (time and nodes) ?

Variable

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

In the final result we used a matrix mean matching imputation method. Our plan was to train another algorithm to impute missing values for us but we ran out of time.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**N/A**
mRNA (microarray, moresamples)	**N/A**
CNA	**N/A**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#31

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, January 21, 2018 5:46:04 AM
Last Modified: Sunday, January 21, 2018 5:56:44 AM
Time Spent: 00:10:40
IP Address: 129.22.124.176

Page 1

Q1 What is your team name?

Cavaliers

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computer Science**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

case western reserve university

Q7 Describe your method. (Choose all that apply)

Low rank matrix completion

Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	If Yes, please briefly describe the procedure used to generate the confidence score: : use matrix completion
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

n^2

Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#32

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, January 21, 2018 7:53:25 AM
Last Modified: Sunday, January 21, 2018 8:07:10 AM
Time Spent: 00:13:44
IP Address: 116.33.11.116

Page 1

Q1 What is your team name?

DEARGENpg

Q2 How many people made up your team?

8

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1,
Sc2,
Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Daejeon, Suouth Korea

Q7 Describe your method. (Choose all that apply)

Low rank matrix completion

Q8 Did you use iterative algorithm to update imputation in your model?	**No**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**

Q11 What is the computational complexity of your method (time and nodes) ?

30 minute

Q12 Did you use multiple data types to make your predictions?	**No**
Q13 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: feature oriented standardization
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.): CODON count, GC percentage, Protein folding energy
Q15 Describe your method. (Choose all that apply)	**Ensemble**

Q16 What is the computational complexity of your method (time and nodes) ?

Train : about 1 minute, Test : about 10 second for each protein

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We filled the missing values as mean value of across the samples in RNA / DNA data.
We didn’t use missing protein.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions? **No**

Q20 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
feature oriented
standardization

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

If Yes, please specify, (for instance, Pathways: KEGG; GO; Reactome. Interaction Networks: Biogrid; String.):
CODON count, GC percentage, Protein folding energy, KEGG pathway

Q22 Describe your method. (Choose all that apply) **Ensemble**

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

We filled the missing values as mean value of across the samples in RNA / DNA data.
We didn't use missing protein.

Q24 What is the computational complexity of your method (time and nodes) ?

Train : about 1 minute, Test : about 10 second for each phospho site

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**N/A**
CNA	**3**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

predicted independently, and gave a protein name as a feature.

#33

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Sunday, January 21, 2018 11:44:56 AM
Last Modified: Sunday, January 21, 2018 11:50:05 AM
Time Spent: 00:05:09
IP Address: 183.172.97.180

Page 1

Q1 What is your team name?

BruinGo

Q2 How many people made up your team?

3

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc1

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**Others**	**Others**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**BA/BS**	**Mathematics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

UCLA, LA, CA, USA/Tsinghua University, Beijing, China

Q7 Describe your method. (Choose all that apply)	**Low rank matrix completion** , Other (please specify): With Bootstrap Aggregating
Q8 Did you use iterative algorithm to update imputation in your model?	**Yes**
Q9 Does your method provide confidence scores for the imputed values?	**No**
Q10 Does your method treat biological missing and technical missing differently?	**No**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**

Q19 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**

#34

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, January 22, 2018 1:36:00 AM
Last Modified: Monday, January 22, 2018 1:58:39 AM
Time Spent: 00:22:38
IP Address: 158.64.14.170

Page 1

Q1 What is your team name?

BIOMOD-LIH

Q2 How many people made up your team?

7

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computer Science**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Luxembourg Institute of Health (LIH), Strassen, Luxembourg

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: all
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: feature re-scaling.
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**

Q22 Describe your method. (Choose all that apply)

Linear,

Non-Linear,

Tree based

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

It was data-dependent, e.g., proteomics missing data estimated with LCMD.

Q24 What is the computational complexity of your method (time and nodes) ?

training (days) and testing (few hours)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

predicted independently

#35

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, January 22, 2018 4:02:45 AM
Last Modified: Monday, January 22, 2018 4:14:43 AM
Time Spent: 00:11:57
IP Address: 158.64.14.170

Page 1

Q1 What is your team name?

BIOMOD-LIH

Q2 How many people made up your team?

6

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Luxembourg, Luxembourg, Luxembourg Institute of Health (LIH)

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: All
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Feature rescaling
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**

Q22 Describe your method. (Choose all that apply)

Linear,

Non-Linear,

Tree based

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Data dependent (for instance, for protein data, we used the LCMD method - Left-Censored Missing Data).

Q24 What is the computational complexity of your method (time and nodes) ?

Training (days), testing (hours)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Independently

#36

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, January 22, 2018 2:33:33 AM
Last Modified: Monday, January 22, 2018 5:43:42 AM
Time Spent: 03:10:09
IP Address: 158.64.14.170

Page 1

Q1 What is your team name?

BIOMOD-LIH

Q2 How many people made up your team?

6

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**Others**	**Computer Science**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Luxembourg Institute of Health

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: proteomics + transcriptomics + genomics
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: Feature rescaling

Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q22 Describe your method. (Choose all that apply)

Linear,

Non-Linear,

Tree based

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

NA values has been Input in a data dependent manner (LCMD).

Q24 What is the computational complexity of your method (time and nodes) ?

Training: several days, testing: several hours.

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Independently.

#37

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, January 22, 2018 5:22:56 AM
Last Modified: Monday, January 22, 2018 6:46:56 AM
Time Spent: 01:23:59
IP Address: 158.64.14.170

Page 1

Q1 What is your team name?

BIOMOD-LIH

Q2 How many people made up your team?

6

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**PhD**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Others**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Luxembourg Institute of Health, Strassen, Luxembourg

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: all
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: feature rescaling
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**

Q22 Describe your method. (Choose all that apply)

Linear,

Non-Linear,

Tree based

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Data dependent (LCMD for protein data). Microarrays were rescaled to RNA-seq

Q24 What is the computational complexity of your method (time and nodes) ?

Training (days) and testing (hours)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Predicted independently

#38

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Monday, January 22, 2018 6:40:08 AM
Last Modified: Monday, January 22, 2018 6:48:23 AM
Time Spent: 00:08:14
IP Address: 158.64.14.170

Page 1

Q1 What is your team name?

BIOMOD-LIH

Q2 How many people made up your team?

7

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc3

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Computational Biology/Bioinformatics**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**MS**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Luxembourg

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?	**Respondent skipped this question**
Q10 Does your method treat biological missing and technical missing differently?	**Respondent skipped this question**
Q11 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q12 Did you use multiple data types to make your predictions?	**Respondent skipped this question**
Q13 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q15 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q16 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q19 Did you use multiple data types to make your predictions?	If Yes, please list the data types:: All
Q20 Did you normalize or standardize the data before carrying any analysis?	If Yes, please explain how you normalized each data type.: feature rescaling
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**No**

Q22 Describe your method. (Choose all that apply)

Linear,

Non-Linear,

Tree based

Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

Data dependent (LCMD for protein data)

Q24 What is the computational complexity of your method (time and nodes) ?

Training (days) and testing (hours)

Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

Protein	**1**
mRNA (RNAseq, less samples)	**2**
mRNA (microarray, moresamples)	**3**
CNA	**4**

Q26 How did you predict multiple phosphosites observed for the same protein/peptide?

Predicted independently

#39

COMPLETE

Collector: Web Link 1 (Web Link)
Started: Wednesday, January 24, 2018 3:34:07 AM
Last Modified: Wednesday, January 24, 2018 3:38:48 AM
Time Spent: 00:04:41
IP Address: 213.243.163.170

Page 1

Q1 What is your team name?

Ari Siitonen

Q2 How many people made up your team?

1

Q3 In which subchallenges did you participate ? (Choose all that apply) Please complete the questions related to subchallenge you participated in accordingly.

Sc2

Q4 Please select your highest degree conferred and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please selcect:	**MS**	**Biology**

Q5 Please select your ongoing (if any) degree and focus ? (PhD, MS, BS/BA) (e.g. PhD computational biology)

	Degree	Focus
Please select:	**PhD**	**Computational Biology/Bioinformatics**

Q6 What is the key location of your team (City, State/Country, or Institution) ?

Helsinki

Q7 Describe your method. (Choose all that apply)

Respondent skipped this question

Q8 Did you use iterative algorithm to update imputation in your model?

Respondent skipped this question

Q9 Does your method provide confidence scores for the imputed values?

Respondent skipped this question

Q10 Does your method treat biological missing and technical missing differently?

Respondent skipped this question

Q11 What is the computational complexity of your method (time and nodes) ?

Respondent skipped this question

Q12 Did you use multiple data types to make your predictions?

If Yes, please list the data types::
CNA, RNA
seq

Q13 Did you normalize or standardize the data before carrying any analysis?

If Yes, please explain how you normalized each data type.:
sklearn standard scaler

Q14 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)

No

Q15 Describe your method. (Choose all that apply)

Linear

Q16 What is the computational complexity of your method (time and nodes) ?

10h? 1 node?

Q17 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.

filled missing values with 0 and used PCA to reduce dimensions.

Q18 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)

mRNA (RNAseq, less samples)	**1**
mRNA (microarray, moresamples)	**N/A**
CNA	**2**

Q19 Did you use multiple data types to make your predictions?

Respondent skipped this question

Q20 Did you normalize or standardize the data before carrying any analysis?	**Respondent skipped this question**
Q21 Did you use external data / prior knowledge to build your model? (e.g. pathway databases)	**Respondent skipped this question**
Q22 Describe your method. (Choose all that apply)	**Respondent skipped this question**
Q23 How did you treat the missing values in the dataset? Please elaborate your effort on predictor variables and response variable respectively.	**Respondent skipped this question**
Q24 What is the computational complexity of your method (time and nodes) ?	**Respondent skipped this question**
Q25 Which data types provide most power to your model for prediction ? (select most important to least (1 being most important). If you do not know, please click N/A for all of them)	**Respondent skipped this question**
Q26 How did you predict multiple phosphosites observed for the same protein/peptide?	**Respondent skipped this question**