DreamAI: algorithm for the imputation of proteomics data

Weiping Ma[1*], Sunkyu Kim[2*]; Shrabanti Chowdhury[1], Zhi Li[3], Mi Yang[4,5,6], Seungyeul Yoo[1], Francesca Petralia[1], Jeremy Jacobsen[7], Jingyi Jessica Li[8], Xinzhou Ge[8], Kexin Li[9], Thomas Yu[10], Nathan Edwards[11], Samuel Payne[12], Paul C. Boutros[13], Henry Rodriguez[14], Gustavo Stolovitzky[15], Jun Zhu[1], Jaewoo Kang[2], David Fenyo[3], Julio Saez-Rodriguez[5,16], Pei Wang[1#]

[1]Icahn School of Medicine at Mount Sinai (USA),

[2]Department of Computer Science and Engineering, Korea University (South Korea),

[3]New York University (USA),

[4]Faculty of Biosciences, Heidelberg University (Germany) ,

[5]JRC for Computational Biomedicine, RWTH Aachen University, Faculty of Medicine, Aachen, (Germany),

[6]Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University(USA),

[7]University of Colorado (USA),

[8]Department of Statistics, University of California Los Angeles (USA),

[9]Department of Mathematics, Tsinghua University (China),

[10]Sage Bionetworks (USA),

[11]Georgetown University (USA),

[12]Pacific Northwest National Laboratory (USA),

[13]University of California, Los Angeles (USA),

[14]National Cancer Institute (USA),

[15]IBM Research & Mount Sinai (USA),

[16]Institute for Computational Biomedicine, Heidelberg University, Faculty of Medicine, & Heidelberg University Hospital (Germany),

* Co-first authors

[#] Corresponding author: Pei Wang (pei.wang@mssm.edu)

Abstract

Deep proteomics profiling using labelled LC-MS/MS experiments has been proven to be powerful to study complex diseases. However, due to the dynamic nature of the discovery mass spectrometry, the generated data contain a substantial fraction of missing values. This poses great challenges for data analyses, as many tools, especially those for high dimensional data, cannot deal with missing values directly. To address this problem, the NCI-CPTAC Proteogenomics DREAM Challenge was carried out to develop effective imputation algorithms for labelled LC-MS/MS proteomics data through crowd learning. The final resulting algorithm, DreamAI, is based on an ensemble of six different imputation methods. The imputation accuracy of DreamAI, as measured by correlation, is about 15%-50% greater than existing tools among less abundant proteins, which are more vulnerable to be missed in proteomics data sets. This new tool nicely enhances data analysis capabilities in proteomics research.

Introduction

Proteins are responsible for nearly every task of cellular life and are important molecules for disease diagnosis, prevention and treatment. The technique of Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) using isobaric labeling methods, including isobaric tags for absolute and relative quantification (iTRAQ) and tandem mass tags (TMT), allows detection and quantification of thousands of proteins and tens of thousands of their post-translational modifications (PTM) in a given biological sample [1,2]. Isobaric labeling not only greatly enhance the precision of quantification, but also improve the throughput [3,4], as multiple samples can be combined into one multiplex and profiled simultaneously. These technology developments greatly accelerate the application of proteomics to study various diseases [1,2,5-8].

Due to the proteome complexity of many biological samples, in combination with the stochastic sampling procedure and limited duty cycle of mass spectrometry based discovery proteomics, only a subset of peptides and PTMs in a sample can be detected and quantified in each LC-MS/MS experiment, and the members of this subset vary from experiment to experiment. Thus, when proteomics profiles from a collection of LC-MS/MS experiments are analyzed together, a substantial number of missing values are present [9]. In addition, in isobaric labeling experiments, the missingness is correlated with the multiplex structure since the detection of a peptide is done together for all samples in MS1 within the multiplex. Consequently, a peptide is either observed or missing simultaneously for all samples analyzed together. This type of experimental induced multiplex-level missing constitutes the majority of missing events when using isobaric labeling. For example, in proteomics data sets generated in CPTAC ovarian cancer study with iTRAQ platform[2], among all detected proteins and phosphosites, 31.1% proteins and 98.3% phosphosites had missing values in at least one sample (**Fig. 1a-b, Supplemental Fig. 1a-b**). And more than 95% or 99% of total missing events in the whole global or phospho-proteomics data sets are multiplex-level missing (**Fig. 1c**). This multiplex-level missing is also prevalent in data from TMT platforms, as illustrated in **Fig. 1a-b** based on data examples from the CPTAC ovarian cancer confirmatory study [7] (**Supplemental Fig. 1a-b**)

Moreover, as indicated in previous works [10-12], missing in mass spectrometry (MS) based proteomics data is non-random: probabilities of a peptide being missing depend on their abundances in the sample, such that peptides with higher abundance tend to have lower missing rates. Furthermore, the degree of this dependence often varies across different experiments and studies (**Fig. 1d-e, Supplemental Fig. 1c-d**). This dependence between the propensity of a value to be missing and its values is referred to as MNAR --- *missing not at random* [13]. It has been well established in the statistical literature that analysis based on the observed data only in the presence of MNAR shall lead to biased estimates and incorrect inference[13].

The substantial missing rates combined with multiplex dependent MNAR bring great challenges to the downstream data analysis. The common strategy of focusing only on proteins observed in all samples [1,2] makes the downstream data analysis convenient, but abandons a large amount of information from hundreds or thousands of proteins in each proteomics data set. These abandoned proteins could, unfortunately, be very interesting for understanding disease mechanisms, as disease-relevant proteins are often low abundant or subtypes specific and therefore less likely to be measured in all samples.

Thus, there is a pressing need to have strategies other than simply ignoring proteins and PTMs with missing values in proteomics data analysis. Two commonly used methods for handling data with missing values are: 1. to substitute missing values with some constants (e.g., a small number or an estimated mean/median value)[14]; and 2. to perform analysis using observed data only [1,2]. The constant imputation, as well as its enhanced variation (Perseus [15]) which fills in missing values with random variables independently drawn from a pre-specified Gaussian distribution, obviously, will not work for labelled proteomics data, due to the experimentally induced multiplex-level missing patterns. On the other hand, for mass spectrometry data with MNAR, it is dangerous to perform analyses based on observed data points only, which could lead to biased estimates and incorrect inferences [10,13]. In addition, for multivariate and high-dimensional analysis, a subset of samples with completely observed data in multiple features could be small or non-existent.

A more sensible solution is to perform stage-wise learning: firstly use information from observed data points to “learn” the unobserved data points, i.e. impute the missing values; and then conduct statistical analysis based on the imputed matrices. Since proteins and PTMs that interact with each other usually

have correlated abundances, the measured abundances in a given sample contain substantial information of other unobserved proteins and PTMs. Information of other samples with shared properties can also be useful in this learnings step. A few imputation strategies have been proposed to handle missing values in high dimension omics data sets in the past decades. Some of the strategies take advantage of local similarity of the data set. For example, the commonly used *KNN* imputation predicts missing values based on information from K nearest neighbors (proteins or samples) [16,17]. This strategy has been applied to a few proteogenomics studies [5]. To better accommodate the MNAR in proteomics data, in another work [6], the authors proposed a modified KNN algorithm, *ADMIN*, which employs weighted average incorporating abundance dependent missing mechanisms in proteomics data [6]. In addition, *MissForest*, which builds Random Forest models to predict missing values of one feature based on observed values of all other features [18], is another effective local similarity based imputation strategy and has been adopted in multiple genomic studies [19,20].

Besides methods relying on local similarity in the data, there is a collection of imputation algorithms utilizing global structure of the data based on low rank matrix completion. Those methods stemmed from the field of image de-noising [16,21-23], has flourished in a broad range of applications to solve various imputation problems, such as completion of single cell RNA-seq data [24] and GWAS data [25], as well as prediction of miRNA-Disease association [26]. Low rank matrix completion techniques have been recently applied to proteomic data imputation too. For example, *pcaMethods*, a PCA-based method for matrix completion [27], has been applied to impute missing values in TMT proteomics data sets in a recent publication.[28]

Good efforts have been made to evaluate performances of different imputation strategies on label free proteomics data [12,29]. Consensus conclusions from these studies suggest that local similarity based methods and global structure based methods perform better than the constant imputation methods in the presence of MNAR [12,29]. In addition, one study [29] reported superior performance of methods based on global structure, such as low-rank matrix completion [17] and linear model based maximum likelihood estimate [30] [31] to those of local similarity based methods (KNN) for label free proteomics data. Moreover, as expected, it is more challenging to impute missing values for features with missing rate higher than 50% than those with lower missing rates [29].

Despite these various efforts, there has not been any systematic evaluation on whether and how various imputation tools work on labelled LC-MS/MS data sets. The pioneer investigation by Palstrøm et. al.[28] is informative and confirms the advantage of KNN and low rank matrix completion over constant imputation for labelled proteomics data. But this investigation is incomprehensive due to the limited number of imputation methods considered and the inadequate numerical examples with rather simplified missing mechanism assumptions. Therefore it is of great interest to perform more systematically assessment on which tools may best solve the missing value imputation problem for proteomics data from labelled LC-MS/MS experiments.

Towards this goal, we carried out a NCI-CPTAC DREAM Proteogenomics Imputation Challenge, aiming to leverage techniques from multiple research field such as statistical computation and machine learning, and to achieve a superior solution for the data imputation problem for labelled LC-MS/MS proteomics data sets through crowd learning (https://sagebionetworks.org/research-projects/nci-cptac-dream-proteogenomics-challenge/).

The Challenge included a competition phase and a collaborative phase. In the competition phase, participants were invited to submit imputation algorithms trained on labelled LC-MS/MS proteomics data sets, and the performances of these algorithms were evaluated on a collection of test datasets generated from the CPTAC breast data [1]. In the collaborative phase, together with the three winning teams from the competition phase, we further enhanced and integrated different imputation techniques and developed the final *Aggregation based Imputation algorithm* --- DreamAI, which is based on ensemble of six different imputation methods including two low-rank matrix completion methods, two prediction based imputation methods, and two KNN type methods. The performance of DreamAI and other imputation tools were then systematically evaluated and compared using the CPTAC ovarian proteomics data sets, which contains profiles of duplicate tumor samples from the same patients [2]. The imputation accuracy of DreamAI, as measured by correlation, is about 15%-50% greater than the few leading popular tools, including ADMIN [6], KNN[16,17], missForest[18] and *pcaMethods*[27].

To illustrate the usage of imputation in proteomics data analysis, we performed proteogenomic integrative analysis using a newly published data of deep TMT proteomic profiling of 103 clear cell renal cell carcinoma (CCRCC) samples and 80 adjacent normal tissue samples[32]. We observed better RNA-protein

concordances between transcriptomic data and proteomic data with imputation than that without imputation. When evaluating the power to detect proteins having significantly different abundances in tumor and adjacent normal tissues, we further observed an advantage of using data with DreamAI imputation over that with KNN imputation or no imputation.

In summary, this work represents a landmark crowdsourced community effort to address the problem of imputation for labelled LC-MS/MS proteomics data sets. The R package of DreamAI is provided through github. This tool can benefit data analysis practice in a broad range of proteomics research.

Result

Challenge overview

The NCI-CPTAC DREAM Proteogenomics Imputation Challenge was carried out to develop a benchmark imputation strategy for labelled LC-MS/MS proteomics data sets through crowd learning. The challenge consists of two phases: a challenging and a community phase. In the challenging phase, participants were invited to build their own imputation algorithms and winners were identified based on performances of submitted imputation algorithms on test data sets. In the community phase, top-performing participants worked jointly to develop a benchmark imputation strategy for labelled LC-MS/MS proteomics data. In both phases, imputation performances were assessed based on two metrics: protein-wise correlation and normalized root mean squared error (NRMSD) between imputed and true values.

The challenging phase

Since imputation is an unsupervised learning, to objectively evaluate different imputation algorithms, in the challenge phase, we implemented a simulation framework to generate decoy data sets with missing patterns mimicking that of the real data sets, based on protein profiles from labelled LC-MS/MS experiments in CPTAC breast cancer studies.[1,8] Specifically, we started with subsets of protein intensity matrices with complete measurements and superimposed pseudo missing data points generated from a probability model, which incorporates both biological and instrumental missing events, with the

probability of the latter depending on protein abundance measurements (see **Online Methods**).

In total 10 training data sets and 100 testing data sets were generated. The large number of test data sets is to allow a thorough evaluation of performances of submitted imputation algorithms (**Fig 2a**, see **Online Methods**). Specifically, training data sets were generated based on global proteome data from CPTAC retrospective breast cancer study [1] and were shared with participants, while testing datasets were based on global proteomics from CPTAC breast cancer confirmatory study[8] and were not shared with participants. Each participant team needed to firstly develop an imputation algorithm based on training data sets, and then submit their final algorithm to Synapse to be evaluated on the testing data sets. The final ranking of participating teams during the challenge phase was determined by a tie breaking strategy (see **Online Methods** and **Supplementary Table 1-2**).

Among 21 teams participating in this challenge, 17 got valid scores on the final leaderboard. Names and affiliations of all participants were listed in **Supplementary Table 3**. The corresponding 17 imputation methods include 6 methods based on prediction models, 5 using matrix completion techniques, 2 relying on constant imputation, 2 employing multiple strategies and 2 other method without algorithm strategies reports in the survey. The performances of these 17 algorithms were illustrated in **Fig. 2b**, **2c**. Interestingly, diverse performances were observed for teams employing the same category of methods. For example, among the five low-rank matrix completion based imputation methods by five different teams, two showed superior performance, but the other three got much worse results than KNNimpute [16,17], a baseline imputation method (Fig. 2b). This observation suggests that customized treatment for labelled proteomics data in employing these imputation techniques is important to assure good performance. Also, as expected, the two methods based on constant imputation showed poor performances, suggesting this simple treatment does not work well for proteomics data with complicated missing mechanisms.

Three methods --- SpectroFM, RegImpute, and Birnn --- demonstrate better performance than the baseline algorithm KNNimpute [16,17]. Both SpectroFM and Birnn use matrix completion techniques, while RegImpute employs prediction models. Please see next section and **Online Methods** for more details. The corresponding teams of the three winning algorithms --- SpectroFM,

RegImpute, and Birnn --- were then invited to participate in the community phase.

The community phase

In the community phase, the goal is to construct a consensus imputation algorithm by integrating multiple methods with diverse strategies. We not only utilized the winning algorithms from the challenging phase, but also leveraged existing tools that provide complementary strengths. We extensively evaluated different integration strategies, and developed a bagging based aggregation framework that enhances the robustness of the final algorithm ---DreamAI: Aggregated Imputation algorithms based on bagging procedure. Please see next Section for methodology and performance details of DreamAI.

We utilized protein profiles of 32 pairs of duplicate tumor samples quantified by two independent proteomics labs in the CPTAC ovarian study [2] to evaluate imputation performances. Specifically, one set of the 32 tumor samples were processed by the Pacific Northwest National Lab; and the duplicate set of the 32 tumors were processed by a proteomics lab from John Hopkins University. We thus referred to these two data sets of 32 samples as PNNL-data and JHU-data respectively.

All imputation methods were firstly applied to the PNNL-data of 3027 genes (n=32) and the results were then evaluated against corresponding data points in JHU-data, which is regarded as good approximation for the true values that was missing in PNNL-data. There are 3700 missing values in the PNNL-data, and most (>99%) of them were not missing in the JHU-data. In addition, to account for technical and biological factors contributing to different protein abundance measurements in PNNL- and JHU data sets, we employed scaled correlation and NRMSD-**δ** as performance evaluation. Specifically, for each protein, background correlation and NRMSD were obtained using paired data points observed in both PNNL- and JHU-data. Scaled correlation was then calculated by dividing the correlation between imputed values and ground truths with the background correlation of each protein. NRMSD-**δ** was calculated as the NRMSD performance of the imputed values minus the background NRMSD. In addition, to ensure robust evaluation, we select a subset of 289 proteins which have at least 5 missing data points and background correlation between PNNL and JHU-data greater than 0.3 for imputation performance evaluation.

DreamAI: Methodology and Performance

DreamAI utilizes an aggregated imputation framework [33] including three steps (**Fig. 3a**): generates 100 bagging sets with pseudo missing values based on the original data; imputes each bagging set with a consensus imputation strategy; and averages imputated values of each missing spot across different bagging sets.

The consensus imputation strategy

The central piece of DreamAI --- the consensus imputation strategy, is based on results from six imputation algorithms: the three winning algorithms in the challenging phase (spectroFM: Team DMIS_PTG; RegImpute: Team Jeremy Jacobsen; Birnn: Team BruinGo) and 3 baseline algorithms (ADMIN[6], KNN[16,17], missForest[18]) (**Fig. 3b**).

Both spectroFM and Birnn are based on low rank matrix completion methods. Specifically, spectroFM employs LibFM, a factorization machine library [34] to approximate the normalized protein abundance matrix (with missing values) with the product of two dense latent low rank matrices corresponding to proteins and samples respectively. In addition, a regularized MCMC algorithm is implemented in spectroFM to solve the optimization problem. Birnn, while employs a similar low rank matrix decomposition framework, uses a different regularization technique --- the smoothly clipped absolute deviation (SCAD) penalty [35] --- to constrain the ranks of the decomposed matrices, and implements an iteratively reweighted nuclear norm (IRNN) [36] algorithm to solve the optimization problem (see Online Methods).

Similar as missForest [18], RegImpute tackles the problem of imputation through prediction. The idea is to use observed abundances of other proteins (samples) to estimate the missing abundance of a given protein (sample). While random forest models are used by missForest, ridge regressions [37] are utilized by RegImpute (see Online Methods). Specifically, RegImpute incorporates an iterative procedure to refit the prediction models leveraging the imputed values from the last iteration. This iterative procedure helps to improve the prediction accuracy, and usually converges after 10 iterations.

KNN based imputation, the most commonly used imputation strategy in omics studies, can also be viewed as a prediction approach: a small set of features (samples) in the neighborhood of the feature (sample) to be imputed are used to

fit a prediction model, which often takes the form of a linear combination (weighted average). ADMIN [6] is an enhanced version of KNN. It specifically models the abundance-dependent missing mechanism in proteomics data set, and uses the joint likelihood of protein abundances and missing mechanisms to calculate the optimal weight for predicting the missing values (see **Online Methods**).

In addition, when selecting baseline methods to be included in DreamAI aggregation, we also considered *pcaMethods* [27], a low-rank matrix completion method that has been applied to missing value imputation of labelled proteomics data [28]. However, the performance of *pcaMethods* is substantially worse than that of KNN, MissForest, and ADMIN on the CPTAC2 ovarian cancer data set (Fig S3). Thus we did not include this algorithm in the final consensus of DreamAI.

All selected methods provide complementary strengths. While the low rank matrix completion based methods take good advantage of the strong global covariance structure among proteins, the prediction-based methods provide more flexible imputation solution to small neighbors (individual features) in the data. In addition, missFroest helps to capture non-linear relationship among proteins, and ADMIN utilized the abundance-dependent missing trend in proteomics data. Thus, by aggregating all these strategies in an effective way, we expect to achieve more optimal and robust imputation performance. Specifically, we propose to average the imputation results of all the 6 methods on one data set as the consensus imputation strategy. The bagging procedure, described below, makes this simple average rather robust and effective.

Model aggregation through bagging

A modified bagging strategy is adopted in DreamAI to improve the robustness and accuracy of imputation algorithms. Instead of sub-sampling subjects or proteins, DreamAI generates “bagging” (perturbed) data matrices by setting a small subset of observed data points in the original data matrix as pseudo NAs. Specifically, these data points were selected according to a probability model reflecting the abundance-dependent missing mechanism with parameters estimated based on the original data matrix (see **Online Methods**). Then DreamAI applies imputation algorithms on a collection of bagging matrices with both true and pseudo missing values, and reports the average of the imputed values of each missing spot across all bagging matrices as the final imputed

values. For the application on the PNNL-data, we utilized 100 bagging matrices, and set the missing rates in the bagging matrices to double that of the original data set.

Performance evaluation

We first illustrated the benefit of bagging aggregation on imputation. We applied individual imputation method with or without bagging aggregation on the PNNL data. For each method, correlation between imputed values and the observed "true" values from the JHU data set of the corresponding data points for protein groups based on different stratification criterions were used for evaluation. Specifically, proteins were divided into multiple groups with different (a) protein closeness in observed data, (b) NRMSD of pseudo missing data from all bagging sets and (c) average protein abundances in observed data. Note, protein closeness measures correlation strength between each protein and its neighboring proteins (see Methods). As shown in **Fig. 3C**, the results based on bagging aggregation showed overall improved correlations compared to those without using bagging aggregation. And the improvement is more dramatic for baseline methods than the winning algorithms from the challenging phase.

We then compared the performance of DreamAI to that of the individual imputation algorithm (with bagging). The average scaled correlation and NRMSD based on all proteins are shown in **Fig. 3d**. DreamAI achieves higher correlation and lower NRMSD than all the six individual imputation methods. Specifically, the imputation accuracy of DreamAI, as measured by scaled-correlation, is about 20% greater than KNN and ADMIN, and 15% greater than missForest. In addition, the performance of DreamAI was also compared to that of *pcaMethods,* and a 50% improvement on performance in term of correlation was observed (Fig S3*)*. In addition, the dashed line in the NRMSD plot represents the reference NRMSD based on all paired data points observed in both the PNNL and the JHU data sets. Interestingly, NRMSD of DreamAI is smaller than the reference NRMSD, implying superior performance of DreamAI.

As illustrate in **Fig. 3d**, the three winning algorithms from the Challenge all outperformed the three baseline methods, which is consistent with what we observed in the challenge phase. An immediate question, then, is whether it helps, in the aggregation exercise, to include any or all of the baseline methods, which have suboptimal performances. We thus also evaluated strategies of aggregating none or a subset of the baseline methods in DreamAI. As illustrated

in **Supplementary Fig. 2a**, without any of the baseline methods, the scaled correlation of imputation result is about 13% lower than the result from aggregating all 6 methods. This clearly demonstrates the benefit of aggregating methods with complementary strengths. Moreover, ADMIN appears to be a more important player than KNN and missForest, such that the scaled correlation drops more if ADMIN was left out from the aggregation than when missForest or KNN was left out. This illustrates the benefit of incorporating the abundance dependent missing mechanism, a common feature of proteomics data, in the imputation framework. Between KNN and missForest, KNN is less helpful in the aggregation, such that the method by leaving KNN out achieves even slightly better performance in terms of scaled correlation. More detailed investigation further suggests that KNN helps only for proteins with close neighbors and high abundances (**supplementary Fig. 2b-c**).

In practice, DreamAI R-package provides the flexibility for users to specify any combination of the 6 individual methods to perform DreamAI imputation. When the data dimension or computational cost is not a concern, one may choose to include ADMIN and missForest, in addition to the three winning algorithms, to achieve the optimal performance. When the data matric has a large dimension, computational time required by missForest could be substantial, and the users may choose to include ADMIN and KNN instead of missForest to balance the tradeoff between performance and computational burden.

To further understand the impact of various protein characteristics on the imputation performances, we compared imputation results of different protein groups stratified by three criterions: (a) protein closeness based on observed data; (b) NRMSD of pseudo missing across all bagging sets; and (c) average protein abundances based on observed data. Please see Methods for details. Average scaled-correlation and NRMSD-**δ** are calculated for each protein group. The results are shown in **Fig. 4**.

Imputation performance of DreamAI, in term of (scaled-)correlation, shows an increasing trend with protein closeness. Moreover, the improvement of DreamAI over KNN is the most dramatic, more than 65%, for the protein cluster with the lowest closeness, suggesting the advantage to leverage the information in the whole data set for data points with uninformative neighbors when performing imputation (**Fig. 4a**). Similar pattern is observed based on NRMSD-**δ** as well.

Across the four protein clusters with different pseudo missing performance evaluations, both DreamAI and KNN showed better imputation accuracy in term of correlation for the cluster with the best pseudo missing performance than the others. The improvements of DreamAI over KNN, however, are quite comparable across the four clusters (**Fig. 4b**).

Protein abundance, a metric correlates with imputation performance of KNN, however does not show obvious association with performance of DreamAI (**Fig. 4c**). And DreamAI showed the biggest improvement over KNN for the protein group with the lowest abundances. NRMSD-**δ** of both DreamAI and KNN appeared to be negatively associated with the protein abundance, which seems to imply that NRMSD depends on the scale of the value to be imputed, and thus its interpretation needs to be taken with cautious.

Imputation helps to gain biological insights

To illustrate the improvement of data analysis power based on proteomics data with proper imputation, we applied DreamAI to a large TMT proteomics data set from a newly published proteogenomic study of clear cell renal cell carcinoma (CCRCC) [32]. In this study, 103 treatment naïve renal cell carcinoma and 80 paired normal adjacent tumor (NAT) tissue samples were profiled using a proteogenomic approach wherein each tissue was homogenized via cryopulverization and aliquoted to facilitate genomic, transcriptomic, and proteomic analyses on the same tissue sample. In the global proteomics TMT experiments, protein abundance measurements of 9209 genes were obtained in at least 50% of the samples, with 2059 genes having missing abundance measurements in at least one sample. The overall missing rate of the protein abundance matrix of these 2059 genes was 20.4%, and sample wise missing rate ranges from 2.5% to 7%. The abundance dependent missing (MNAR) trends in proteomics data of tumor and NAT samples are illustrated in **Fig. 5a**, **S4a** respectively.

We first evaluated gene-wise correlations between RNAseq and global proteomics data with or without DreamAI imputation among tumors samples. For 2012 proteins with at least one missing value in tumor samples, we observed improved protein-RNA concordance in proteomic data with DreamAI imputation than that without imputation, including significantly higher gene-wise protein-RNA correlations (wilcox test pvalue<10e-16) (**Fig 5b**), as well as greater numbers of genes with significantly non-zero protein-RNA correlation at

various p-value cutoffs (**Figs 5c, 5d**). Parallel analysis applied to proteogenomic data of NAT samples reveals similar improvement of protein-RNA concordance based on proteomic data with DreamAI imputation over that without imputation (**Fig. S4**).

We then evaluate whether different treatment of missing values may impact statistical powers to detect proteins associated with normal-tumor status. Specifically, we focused on a subset of 49 genes in the CCRCC proteomic data, whose imputed protein abundances by KNN and that by DreamAI are rather different (the NRMSD between the imputed abundance by KNN and that by DreamAI is greater than 0.5). As illustrated in **Fig. S5a**, the distribution of p-values from Wilcox two-sample t-tests comparing tumor and NAT samples based on proteomic data with imputation by DreamAI is more significant than that by KNN as well as that based on data without imputation. Similar benefit of power gain by DreamAI imputation over KNN as well as no-imputation is also observed in **Fig. S5b** when screening for proteins associated with four different immune subtypes of CCRCC samples[32] using Kruskal–Wallis tests. These examples illustrate the advantage of using proteomic data with DreamAI imputation in downstream statistical analysis over other alternative strategies.

Discussion

How to handle missing values in MS based proteomics data has been a long-standing challenge in proteomics research. The larger the study size is, the worse the issue of missing will be, as data from more mass spectrometry experiments need to be merged together. The isobaric labelling technique, which on one hand greatly enhances the quantitation precision and experiment throughput, on the other hand, further exacerbates the missing data problem. With experimental induced multiplex-level missing pattern as well as the abundance dependent missing trend, proteomics data from labelled MS experiments cannot be properly or effectively analyzed by using observed data only (either ignoring all features with missing values or ignoring subsets of samples with missing data points in feature-wise modeling).

Another strategy to handle missing data is through imputation, which has been widely adopted in many research fields, such as image processing, single-cell RNAseq studies, as well as label free proteomics data analysis. Its usage in proteomics data from labelled MS experiments is still limited, largely due to a lack of a benchmark imputation method suitable for this type of data. Because of

the complicated missing structure in labelled proteomics data, imputation tools developed for other data types do not apply or does not perform well.

The goal of this study is to develop a benchmark imputation algorithm for labelled proteomics data sets. Specifically, we conducted the NCI-CPTAC DREAM Proteogenomics Imputation Challenge to achieve this goal through crowd learning. 21 teams from a broad range of research fields participated in the Challenge and contributed diverse expertise. As expected, many general imputation algorithms used in other disciplines/applications do not perform well on labelled proteomics data sets. Indeed, only a subset of teams achieved better performance than the KNN imputation on Challenge data sets, suggesting customized treatment of the imputation algorithm for labelled proteomics data is important in order to effectively tackle this problem.

The three winning teams from the Challenge further participated in a collaborative phase, and we jointly developed the final algorithm --- DreamAI --- an ensemble based imputation method. DreamAI employs a bagging framework to aggregate results from 6 diverse imputation methods: three winning algorithms from the Challenge (two based on low-rank matrix completion and one based on prediction model fitting), as well as three baseline imputation methods --- KNN, ADMIN, and missForest, which have been used in previous proteogenomics data analysis [5,6,19,20]. This ensemble strategy of DreamAI leads to greatly improved performance compared to that of individual algorithm: the imputation accuracy of DreamAI in terms of correlation is 15-50% better than that of individual baseline tool, or 9-15% better than that of the individual winning algorithm on an ovarian cancer proteomics data set.

The bagging framework in DreamAI not only enhances the imputation performance, but also helps one gain insights on imputation quality of each feature. Specifically, for a given feature, DreamAI estimates its imputation quality using the correlation between the true and imputed values of pseudo missing data points of this feature across different bagging iterations. In the CPTAC ovarian data application, the correlation assessment for the protein group with the best pseudo missing performance is 0.75, at least 26% higher than the rest protein groups. Therefore, the pseudo missing performance score of each feature is informative to shed light on feature-specific imputation quality.

Since imputation is an unsupervised learning problem, it has been a challenging task to objectively assess the performance of imputation methods. Thus, one of

the major efforts during the Dream Challenge was to create high-quality benchmark simulation data sets to objectively evaluate imputation performances. Specifically, simulations were set up to mimic missing patterns in real proteomics data sets as closely as possible. Multiple testing data sets with varying proportions of biological and experimental missing rates, as well as different degrees of abundance dependent missing trend were generated based on two CPTAC breast cancer proteomics data sets.[1,8] Moreover, to complement the usage of simulated data sets during the Challenge phase, in the community phase, we utilized the CPTAC ovarian cancer proteomic data set [2], which contains proteomics profiles of two replicate biological samples of 32 ovarian tumors. This provides a unique opportunity to directly assess imputation performances on real missing data points in cancer proteomics studies.

The benefit of using imputed data in downstream analyses stems from the improvement of sample size and thus the analysis power. As illustrated in the CCRCC application, imputation helps to capture more molecular features in proteomics data and improves the RNA-protein concordance overall. In the real data analysis, we removed features with missing rates higher than 50% in imputation and downstream analysis. The choice of 50% cutoff is a tradeoff between imputation accuracy and information (data feature) loss in the downstream analysis. For features with high missing rate, the tasks to accurately identify close neighbors or to fit prediction model based on observed data points become very challenging due to the sample size limitation. It has been suggested that, in general, imputation methods perform better on features with less missing values (<50%) than on features with more missing values (>50%)[29]. Also, in downstream analyses, it's preferred that the observed data points out weight the imputed data points to ensure robustness. Thus, we settled with a cutoff of 50%.

Although we provided NRMSD values on all examples, we used Spearman correlation as the main metric for evaluating imputation performance. NRMSD measures the distance between the imputed values and the true values of missing data points normalized by the varying range of abundances of each protein. Despite being a normalized distance measurement, NRMSD still depends on the scale and distribution of the protein abundances. On the other hand, Spearman correlation is a scale free measurement which is robust to any outliers and the absolute scale of the data distribution. As illustrated in **Fig. 4c**, among protein groups with different mean abundance levels, performance

based on correlation is very stable, but NRMSD has an obvious trend to be positively associated with protein mean abundances.

For data analysis of label free proteomics data, it has been suggested that directly model peptide abundance could be more efficient than performing imputation at the protein abundance level [12]. This is because the summary (or average) based peptide-protein intensity roll-up used for label free proteomics data is vulnerable to many confounding factors, and then modeling the peptide level abundances directly could effectively get around the variabilities induced in the roll-up step. However, in isobaric labeled proteomics experiments, rolled-up from peptides to proteins can be performed at the log-ratio intensity level (i.e. log-ratio between intensity of a target sample and that of the reference sample in the same TMT multiplex for one peptide). This strategy greatly improves the robustness and precision of protein quantification, while at the same time, effectively reduces the missing data percentage in protein level data compared to the peptide-level data. Thus, for isobaric labeled global proteomics experiments, we recommend working with protein/gene level data. For phosphorproteomics experiment, since phosphosite-site is the meaningful biological unit for downstream analysis, we actually work with the quantification at phosphor-site level and perform imputation on phosphor-site level data directly.

Although DreamAI has a general framework and can be applied to other proteomics data from label free experiments, its performance on those applications warrants future study. In addition, for proteomics data from targeted mass spectrometry experiments, such as MRM (multiple reaction monitoring), imputation could be less of a concern due to the relatively low missing rate. However, MRM experiments right now can handle at most a few hundred proteins/peptides in one run, and thus are not suitable for deep profiling in discovery studies.

An R package of DreamAI has been implemented and is available to public at Github (https://github.com/WangLab-MSSM/DreamAI). Performing DreamAI imputation with this R package on the CCRCC data matrix with 9209 genes and 183 samples took 4.3 hours on a PC with Intel Core i7-7700HQ CPU (2.80GHz).

ONLINE METHODS

Design and Data Sets of Challenging Phase

Multiple stages were set up in the challenging phase: two leaderboard rounds, and one final ranking round, to allow self-correction on the algorithm of each participant and also to achieve fair competition for the final ranks.

The process of generating data matrices with missing value is the same in both training and testing. We collect protein with complete observation as the basis matrix of underlying truth (7927 proteins of 80 samples from CPTAC2 breast cancer retrospective study for training data and around 8203 proteins of 83 samples from CPTAC2 breast cancer confirmatory study for testing data).

Biological missing spots were assigned to basis matrix with missing spot correlated among proteins with protein intensity correlation of the basis matrix. Basis matrix with biological missing was considered as underlying truth. Since biological missing are difficult to identify from the missing data, to raise the challenge of imputation in the synthetic data set we set the biological missing rate to be much higher than the non batch level missing rate in real data set.

Next, we simulate instrumental missing with abundance dependent missing mechanism, learned from the real data set. Both instrumental missing and biological missing were indicated as ‘NA’ in the observed data sets.

Imputation algorithm will be applied on the observed data sets and evaluated on the missing spot with underlying truth. We setup multiple replicates of training and testing data sets to assess robust evaluation on the imputation algorithms. In total, we generated 10 training data sets with same missing mechanism and 200 data sets of testing with same instrumental missing mechanism but diverse level of biological missing rate (**Fig. 2b**).

After opening of the challenge competition, we released the 10 training data set to public, participants were allowed to build and train their algorithms in the training data. Leader board were presented and updated during the period of Round 1 and 2 by evaluating algorithms of participants using 100 testing data sets. Final Score ranking were generated in the final round by evaluation on the other 100 testing data sets.

Evaluation of Imputation performance and Tie Breaking for Final Round Leaderboard

Performance of imputation algorithms are evaluated through normalized root-mean-square errors (NRMSD) and correlation coefficients between imputed data and underlying truth. NRMSD is calculated on all missing spots of each protein, and correlation is calculated on instrumental missing spots of each protein.

Given X to be imputed value and Y to be underlying true value,

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n_{missing}} (y_i - x_i)^2 / n_{missing}}}{y_{max} - y_{min}}, \quad r = \frac{1}{n_{missing} - 1} \sum_{i=1}^{n_{missing}} \frac{(x_i - \overline{X})(y_i - \overline{Y})}{S_x S_y}.$$

Evaluation metrics of 100 different observed data sets in the final round were compared to identify the winning team. Specifically, we compared NRMSD first, and if there are ties on NRMSD, we will compare the correlation to break the tie. Significance of score differences is tested using two criteria:

1. *Confidence Intervals* For each team, we computed 95% Confidence Intervals (CI) across different data sets. Since difference of biological missing rate will lead to different levels of scores, to make the variance estimation more meaningful we calculate CI for 4 groups with different biological missing rate separately. We declared two teams statistically different, when one team has (all) CI non-overlapped with (and higher than) the corresponding interval of the other team.

2. *Bayes Factor* Given two teams, we estimated the Bayes Factor (BF) via a 100 paired imputed matrix. Each pair came from the results of the same observed data set. We declared two teams statistically different if the Bayes Factor of their scores is larger than 10 or smaller than 0.1.

We consider the four teams having the lowest average NRMSD scores across 100 data sets, since the baseline method KNN will beat the 5th team with our tie breaking criterion. Those teams are *Hongyang Li and Yuanfang Guan, DMIS_PTG, BruinGo, Jeremy*.

Comparison of CI was showing in the **Supplementary Table 1**. If the number equals 4, scores of the team at row will be significantly higher than the scores of the team at column. From **Supplementary Table 1A,** we found out none of those team can beat any other team by NRMSD. Therefore we look at the correlation of them in **1B**, and infer that the team *DMIS_PTG* has the best correlation scores based on the confidence intervals. We also compared BF. For each team pairs (**Supplementary Table 2**) If the number is larger than 10, scores of the team at row will be significantly higher than the scores of the team at column. If the number is smaller than 0.1, scores of the team at row will be significantly lower than the scores of the team at column. We found out only team *DMIS_PTG* can beat some of the other teams by NRMSD (**Supplementary Table 2A**), but none of the team is dominant in this criterion. Therefore we look at the comparison of correlation (**Supplementary Table 2B**) and infer that the team *DMIS_PTG* has the best correlation scores based on the BFs. In conclusion, this sub-challenge was won by team *DMIS_PTG.*

Evaluation of Imputation performance in Community Phase

To fully understand the improving of DreamAI from the baseline method KNN, and in the mean time to study the impact on the imputation performance by the protein behavior, we summarized the performance at cluster level. We defined cluster by three different criteria: **protein closeness**, **pseudo missing performance**, and **protein abundance**. Those clusters were constructed with following procedure

1. **Protein Closeness:** We calculate the pairwise correlation of all proteins having at least one missing datapoint in the PNNL data, and protein closeness is calculated using average of largest 50 correlations of each

protein(those 50 proteins were considered as its neighbor proteins, and the average of correlation is regarded as closeness of that protein among all neighbors). We split 289 proteins that are eligible to evaluation into 4 clusters based on the 4 quantiles, with the average closeness from lowest (first cluster) to highest (4^{th} cluster).

2. **Pseudo missing performance:** NRMSD was calculated between pseudo missing values of bagging datasets from the PNNL data and corresponding observed value in the same data set. We used NRMSD to form 4 clusters of the 289 proteins. These clusters are ordered from low performance to high performance by the average pseudo NRMSD values, meaning that meaning that the 1^{st} cluster has the highest average pseudo NRMSD and the 4^{th} cluster has the lowest average pseudo NRMSD.

3. **Protein abundance:** Finally, we also defined gene cluster by the range of observed mean protein abundance and ordered the clusters from lowest (first cluster) to highest (4^{th} cluster) mean protein abundances. Genes within each cluster have similar protein abundance.

Methods of 3 baseline algorithms

ADMIN: Abundance Dependent Missing Data Imputation

The method is designed for imputation of isotopic labeling proteomics data in which batch effects exist and missing data is dependent on protein abundances.[6] Observed abundance data is assumed to follow a linear mixed-effect model. Random intercept is accounted for batch effect at protein level. Each protein is fitted by the linear regression of its close neighbors regardless of the random intercepts in the model. Close neighbors are determined by the pairwise correlation. A fixed number of neighbors are included in the linear regression for each protein. On the other hand, a non-random missing mechanism is assumed: missing rate is exponentially linear correlated with the ‘true’ abundance. Based on these assumptions, an EM(expectation-maximization) based algorithm is employed to iteratively solve the linear

prediction of missing values and estimation of the abundance dependent missing parameters in one model: given a current estimation of imputation values, in next M step random effects and parameters of missing mechanism are estimated with both observed and imputed values; in the following E step, for a given protein, the missing elements are predicted from the close neighbors with linear model on both observed and imputed value after removing the bias from missing mechanism and random effect values. To avoid huge computation consumption, the default number of neighbors in algorithm is set to be 10.

knn.impute

impute.knn is a function designed to impute missing values of gene expression data, using K-nearest neighbor averaging.[16,17] For each gene with missing values, k nearest neighbors were found using a Euclidean distance metric, confined to the columns for which that gene is NOT missing. After the k nearest neighbors are identified for a gene, imputed value of a missing element is the average of those (non-missing) elements of its neighbors. For categorical variables the mode of the neighbors is used, and for continuous variables the median value is used instead. To increase computation efficiency, gene sets over certain threshold (set as 1500 in the package) were broken into blocks using two-mean clustering. This is done recursively till all blocks have less than the max number of genes. For each block, k-nearest neighbor imputation is done separately.

missForest

missForest is developed to impute missing values particularly in mixed-type data: continuous and/or categorical data including complex interactions and nonlinear relations.[18] The missing data problem is addressed using an iterative imputation scheme by training a Random forest model on observed values, followed by predicting the missing values. Imputation problem is solved by iteratively fitting and predicting procedure, since the imputed value on predictors can help to obtain better prediction. Random forest is chosen to model the missing value because it can handle mixed-type data and is known to perform very well under conditions like high dimensions, complex interactions and non-linear data structures. In case of high-dimensional data some

parameters in the algorithm are suggested with a relatively small value, for example: number of trees to grow in each forest and number of variables randomly sampled at each split to obtain an appropriate imputation result within a feasible amount of time. Moreover, it can be run parallel to save computation time using an appropriate backend.

Methods of top 3 participants

SpectroFM: Matrix factorization-based imputation

In the computer science domain, the imputation of missing values, which has been the focus of many studies, can be considered as a recommendation task since a user's unobserved preferences are represented as missing values in a user-item matrix. Given a user-item matrix, a recommendation system predicts a user's preferences for an item based on other users' existing preferences for the item and the user's preferences for other items. This is analogous to the task in this challenge. If we consider proteins as items and patients as users, it is possible to exploit collaborative filtering algorithms. We first apply Z-normalization to a protein abundance data matrix to make the data fit a normal distribution. We save the mean and variance to revert the data to its original scale when we perform imputation. We train a low-rank matrix factorization model on existing values in the normalized abundance matrix. For the implementation of the matrix factorization model, we use LibFM, a factorization machine library [34]. Using the calculated latent parameter matrix of proteins and the latent parameter matrix of patients in the model, we reconstruct the best approximation of the original input matrix by multiplying the two latent matrices. Since the latent matrices are dense, the missing values in the original matrix are imputed in the reconstructed approximated matrix. We set the dimensions of the latent protein and patient matrices to 40. Consequently, the rank of the reconstructed approximated matrix is 40. We use a Markov chain Monte Carlo (MCMC)[38] algorithm to optimize parameters. One of the advantages of MCMC is that it integrates regularization parameters into the model, which allows us to skip hyper parameter optimization. After the imputation of missing values by the multiplication of the latent matrices, we revert the normalized values to their original scale using the saved mean and variance.

Algorithm 1: SpectroFM

Input: binary indicator feature vector x and observed protein abundance values y^{obs}

Output: imputed protein abundance values y^{miss}

Initialize model parameter θ

Z-Normalize y values to $\tilde{y}$ using

$$\mu^{obs} = \frac{1}{N} \sum_{(u,i) \in obs}^{N} y_{u,i}$$

$$\sigma^{obs} = \sqrt{\frac{1}{N} \sum_{(u,i) \in obs}^{N} (y_{u,i} - \mu^{obs})^2}$$

$$\tilde{y}_{u,i} = \frac{y_{u,i} - \mu^{obs}}{\sigma^{obs}}, (u,i) \in obs$$

Obtaini optimal parameter:

For t in 1,...,7

1: θ^{*} = MCMCOptimizer(x, $\tilde{y}$, θ)

2: $\theta = \theta^{*}$

Impute the missing values $\tilde{y}_{u,i}^{miss}$ using

$$\tilde{y}_{u,i}^{miss} = h_{\theta}(x_{u,i}^{miss}) = w_0 + w_u + w_i + \sum_{f=1}^{k} v_{u,f} v_{i,f}, (u,i) \in miss$$

Return $\tilde{y}^{miss}$ to original scale values y^{miss} using

$$y_{u,i}^{miss} = \mu_{obs} + \sigma_{obs} \times \tilde{y}_{u,i}^{miss}$$

RegImpute: Regression-based imputation

A conventional, post-processed proteomics dataset usually takes the form of a two-dimensional array. From the perspective of training a regression model, the columns of an array can be interpreted as features (dimensions), and the rows can be considered as training instances (or vice-versa). The features and instances can be used to train a predictive model to impute unobserved contains missing values. One solution is to divide data sets into subsets, on which

models can be trained. However, this approach can be very time consuming. A second approach is to train a model on only complete dataset without missing values. The drawback of this approach is that samples with missing values may be characteristically different from samples without missing values (e.g., not missing at random (NMAR) versus missing at random (MAR)). RegImpute is a combination of the two approaches above and uses a simple imputation method such as mean imputation on the existing values to generate a complete training set. In addition, users can impute missing values using the values (e.g., zeros) selected by the users. Then, we use ridge regression, which is a fast and robust linear regression technique. Ridge regression is an extension of linear regression, and its regularization prevents it from overfitting. Ridge regression performs regularization by adjusting weights to avoid focusing on only a few features [37]. Using single regression on the dataset may be sufficient if the initial guesses are nearly correct or if there are few missing values. However, in some cases, the initial regression values are heavily influenced by a prior assumption(s). For this reason, performing regression several times may reduce estimation errors. At each iteration, we use the imputed missing values from the previous imputation to improve regression for the current imputation. At some point, usually after ~10 iterations, convergence is reached.

Algorithm 2: RegImpute

Input: *data matrix Y of protein abundance*

Output: Y^*

For each iteration n:

1. For each column Y_i in data Y (i = 1,...,n) Split the data into two subsets:

   $Y_{miss,i}$: rows with Y_i missing

   $Y_{obs,i}$: rows with Y_i observed

2 Fill NAs in $Y_{obs,i}$ with the imputed values of $Y_{obs,i}$ in iteration n-1 (fill in 0s if n=1)

3 Train ridge regression model on data $Y_{obs,i}$, to associate ith column with all the other columns, obtain β to solve the minimization:

$$\min_{\beta}(Y^i_{obs,i} - Y^i_{obs,i}\beta)^T(Y^i_{obs,i} - Y^i_{obs,i}\beta) + \lambda(\beta^T\beta - c)$$

4 Fill in NAs in $Y_{miss,i}$ with the imputed values of $Y_{obs,i}$ in iteration n-1 (fill in 0s if

n=1)

for all but not ith column

5 Use trained model on $Y_{obs,i}$ to predict ith column in $Y_{miss,i}$ with the other columns as predictor

$$Y_{miss,i}^{i*} = Y_{miss,i}^{-i}\beta$$

6 repeat 1 to 5 until convergence of imputed value.

Birnn: Matrix completion and Bagging-based imputation

We consider the imputation of missing protein abundances in a protein-sample matrix as a matrix completion problem. We assume that all the protein abundances have the same data distribution because they are from the same type of cancer, and thus the matrix is assumed to have a low rank structure. Based on this assumption, we used the iteratively reweighted nuclear norm (IRNN)[36] algorithm with the smoothly clipped absolute deviation (SCAD)[35] penalty, which is a non-convex penalty function on singular values, to better approximate the rank function and enhance low rank matrix approximation. Moreover, we use the bootstrap aggregating algorithm to train multiple models on sampled sub-datasets of the original dataset. The final prediction is given by aggregating the outputs of the multiple models. The bootstrap aggregating algorithm can help prevent models from over fitting by reducing model variance, which contributes to performance improvement.

Algorithm 3: Birnn

Initialize: $k = 0$, X^k, w_i^k, i = 1,2,...,min(m,n)

Output: X^*

1. while not converge do
2. Update X^k by solving

$$\min_x \sum_{i=1}^{\min(m,n)} w_i^k \sigma_i + \frac{1}{2}\left\| X - (X^k - s\nabla f(X^k)) \right\|_F^2$$

with **Weighted Singular Value Thresholding (WSVT).**

Update the weights w_i^k, i = 1,2,...,min(m,n) by

$$w_i^{k+1} = g_\lambda(\sigma_i(X^{k+1}))$$

where

$$g_\lambda(\theta) = \begin{cases} \lambda\theta, \theta \le \lambda \\ \dfrac{-\theta^2 + 2\gamma\lambda\theta - \lambda^2}{2(\gamma-1)}, \lambda < \theta \le \gamma\lambda \\ \dfrac{\lambda^2(\gamma+1)}{2}, \theta > \gamma\lambda \end{cases}$$

3. end while

ACKNOWLEDGEMENT

We would like to thank the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of proteogenomics, on providing the data used in this challenge and making it freely available to the public. We also like to thank Dream Challenges organization for providing the good opportunity to encourage researchers all around the world to take parts in this cutting-edge research topic and all the participants in this challenge for building the algorithms and submitting the results. This work was partly supported by grant (U24 CA210993), from the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC).

AUTHOR CONTRIBUTIONS

WM, ZL, MY, FP, TY, NE, SP, PB, HR, GS, JZ, DF, JSR and PW organized the challenge. WM and PW designed the challenge problem. WM, ZL, FP, NE processed and prepared the data for challenge. TY organized challenge data on sage and implemented the challenge infrastructure. WM, MY and TY evaluated performances of participants in final round of challenge. SK and JK developed the best-performing algorithm in the challenge. WM and PW designed and organized the community phase. SK, JK, JJ, JL, XG, and KL participated in the community phase. WM, SY, and SC carried out the performance evaluation in the community phase. WM, SC, SY and PW wrote the manuscript. SK, JK, MY, KL, XG, JJ, FP, SP, PC, HR, GS, JZ, DF and JSR helped with the manuscript writing. SC built the DreamAI R package. NCI-CPTAC-DREAM Consortium participated in the challenge and submitted their predictions. HR initiated the challenge. PW supervised the project. DF, and JSR assisted in supervising the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing interests.

Reference

1. Mertins, Philipp, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang et al. "Proteogenomics connects somatic mutations to signalling in breast cancer." *Nature* 534, no. 7605 (2016): 55.

2. Zhang, Hui, Tao Liu, Zhen Zhang, Samuel H. Payne, Bai Zhang, Jason E. McDermott, Jian-Ying Zhou et al. "Integrated proteogenomic characterization of human high-grade serous ovarian cancer." *Cell* 166, no. 3 (2016): 755-765.

3. Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS." *Analytical chemistry* 75, no. 8 (2003): 1895-1904.

4. Ross, Philip L., Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski et al. "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents." *Molecular & cellular proteomics* 3, no. 12 (2004): 1154-1169.

5. Gao, Qiang, Hongwen Zhu, Liangqing Dong, Weiwei Shi, Ran Chen, Zhijian Song, Chen Huang et al. "Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma." Cell 179, no. 2 (2019): 561-577.

6. Wang, Minghui, Noam D. Beckmann, Panos Roussos, Erming Wang, Xianxiao Zhou, Qian Wang, Chen Ming et al. "The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease." *Scientific data* 5 (2018): 180185.

7. Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). “CPTAC Ovarian Cancer Confirmatory Study.” Distributed by NCI Proteomic Data Commons. https://cptac-data-portal.georgetown.edu/cptac/s/S038

8. Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). “CPTAC Breast Cancer Confirmatory Study.” Distributed by NCI Proteomic Data Commons. https://cptac-data-portal.georgetown.edu/cptac/s/S039

9. Brenes, Alejandro, Jens L. Hukelmann, Dalila Bensaddek, and Angus I. Lamond. "Multi-batch TMT reveals false positives, batch effects and missing values." *Molecular & Cellular Proteomics* (2019): mcp-RA119.

10. Chen, Lin S., Jiebiao Wang, Xianlong Wang, and Pei Wang. "A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments." *The annals of applied statistics* 11, no. 1 (2017): 114.

11. Chen, Lin S., Ross L. Prentice, and Pei Wang. "A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation." *Biometrics* 70, no. 2 (2014): 312-322.

12. Webb-Robertson, Bobbie-Jo M., Holli K. Wiberg, Melissa M. Matzke, Joseph N. Brown, Jing Wang, Jason E. McDermott, Richard D. Smith et al. "Review, evaluation, and discussion of the challenges of missing value

imputation for mass spectrometry-based label-free global proteomics." *Journal of proteome research* 14, no. 5 (2015): 1993-2001.

13. Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

14. Clough, Timothy, Safia Thaminy, Susanne Ragg, Ruedi Aebersold, and Olga Vitek. "Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs." *BMC bioinformatics* 13, no. 16 (2012): 1-17.

15. Tyanova, Stefka, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y. Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. "The Perseus computational platform for comprehensive analysis of (prote) omics data." *Nature methods* 13, no. 9 (2016): 731.

16. Hastie, Trevor, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. "Imputing missing data for gene expression arrays." (1999).

17. Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17, no. 6 (2001): 520-525.

18. Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28, no. 1 (2011): 112-118.

19. Causey, Dwight R., Jin-Hyoung Kim, David A. Stead, Samuel AM Martin, Robert H. Devlin, and Daniel J. Macqueen. "Proteomic comparison of selective breeding and growth hormone transgenesis in fish: Unique pathways to enhanced growth." *Journal of proteomics* 192 (2019): 114-124.

20. Trilla-Fuertes, Lucía, Angelo Gámez-Pozo, Jorge M. Arevalillo, Mariana Díaz-Almirón, Guillermo Prado-Vázquez, Andrea Zapater-Moros, Hilario Navarro et al. "Molecular characterization of breast cancer cell response to metabolic drugs." *Oncotarget* 9, no. 11 (2018): 9645.

21. Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. "Spectral regularization algorithms for learning large incomplete matrices." Journal of machine learning research 11, no. Aug (2010): 2287-2322.

22. Candès, Emmanuel J., and Benjamin Recht. "Exact matrix completion via convex optimization." *Foundations of Computational mathematics* 9, no. 6 (2009): 717.

23. Sun, Ruoyu, and Zhi-Quan Luo. "Guaranteed matrix completion via non-convex factorization." *IEEE Transactions on Information Theory* 62, no. 11 (2016): 6535-6579.

24. Mongia, Aanchal, Debarka Sengupta, and Angshul Majumdar. "McImpute: Matrix completion based imputation for single cell RNA-seq data." *Frontiers in genetics* 10 (2019): 9.

25. Jiang, Bo, Shiqian Ma, Jason Causey, Linbo Qiao, Matthew Price Hardin, Ian Bitts, Daniel Johnson, Shuzhong Zhang, and Xiuzhen Huang.

"SparRec: An effective matrix completion framework of missing data imputation for GWAS." *Scientific reports* 6 (2016): 35534.

26. Tang, Chang, Hua Zhou, Xiao Zheng, Yanming Zhang, and Xiaofeng Sha. "Dual Laplacian regularized matrix completion for microRNA-disease associations prediction." *RNA biology*16, no. 5 (2019): 601-611.

27. Stacklies, Wolfram, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. "pcaMethods—a bioconductor package providing PCA methods for incomplete data." *Bioinformatics* 23, no. 9 (2007): 1164-1167.

28. Palstrøm, Nicolai Bjødstrup, Rune Matthiesen, and Hans Christian Beck. "Data imputation in merged isobaric labeling-based relative quantification datasets." In *Mass Spectrometry Data Analysis in Proteomics*, pp. 297-308. Humana, New York, NY, 2020.

29. Lazar, Cosmin, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies." *Journal of proteome research* 15, no. 4 (2016): 1116-1125.

30. Schafer, J. L. "NORM: Analysis of incomplete multivariate data under a normal model." *University Park, PA: The Methodology Center, The Pennsylvania State University, version* 3 (2016).

31. Karpievitch, Yuliya, Jeff Stanley, Thomas Taverner, Jianhua Huang, Joshua N. Adkins, Charles Ansong, Fred Heffron et al. "A statistical framework for protein quantitation in bottom-up MS-based proteomics." *Bioinformatics* 25, no. 16 (2009): 2028-2034.

32. Clark, David J., Saravana M. Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost et al. "Integrated proteogenomic characterization of clear cell renal cell carcinoma." *Cell* 179, no. 4 (2019): 964-983.

33. Breiman, Leo. "Bagging predictors." *Machine learning* 24, no. 2 (1996): 123-140.

34. Rendle, Steffen. "Factorization machines with libfm." *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, no. 3 (2012): 57.

35. Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96, no. 456 (2001): 1348-1360.

36. Lu, Canyi, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. "Generalized nonconvex nonsmooth low-rank minimization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4130-4137. 2014.

37. Tikhonov, Andrei Nikolaevich, A. V. Goncharsky, V. V. Stepanov, and Anatoly G. Yagola. *Numerical methods for the solution of ill-posed problems*. Vol. 328. Springer Science & Business Media, 2013.

38. Salakhutdinov, Ruslan, and Andriy Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo." In *Proceedings of the 25th international conference on Machine learning*, pp. 880-887. ACM, 2008.

Main Figure Captions

Figure 1. Missing rates and missing patterns in various proteomics data sets from CPTAC ovarian studies [2,7]. **(a)** Distribution of protein level missing rates in a 4plex-iTRAQ global-proteomics data set of 112 tumor samples and a 10plex-TMT global proteomics data set of 120 tumor samples. The iTRAQ and TMT data sets consist of 7126 proteins and 8290 proteins respectively. **(b)** Distribution of phosphosite-level missing rates in a 4plex-iTRAQ phospho-proteomics data set of 92 samples and a 10plex-TMT phospho-proteomics data set of 120 samples. The iTRAQ and TMT data sets consist of 20746 and 45625 phosphosites respectively. **(c)** Percentage of multiplex-level and non multiplex-level missing data in iTRAQ global- and phospho-proteomics data sets. **(d)** Scatter plot of protein-level missing rates v.s. mean protein abundances based on observed data in the iTRAQ global-proteomics data set. **(e)** Scatter plot of phosphosite level missing rates v.s. mean phosphosite abundances based on observed data in the iTRAQ phospho-proteomics data set.

Figure 2. Proteomics Data Imputation Challenge competition design and performance results of participants (a): Design of data simulation in challenge phase. **(b)**: Correlation and NRMSD evaluations of 17 submitted imputation algorithms. Different colors and shapes represent different imputation strategy categories. The dotted lines illustrate the performance level of KNN imputation. Three leading algorithms with better performance than KNN imputation have their names labeled. **(c)** Performance rank of all algorithms summarized for each strategy category (*algorithms using multiple strategies were listed multiple times in all relevant categories).

Figure 3. **DreamAI algorithm and its performance**. **(a)** Bagging procedure in DreamAI. Firstly, different set of pseudo missing are introduced to original observed data to generate a collection of bagging data sets. Then imputation is performed for each bagging set using the consensus imputation method. The final imputed matrix is the average of all bagging sets at the missing spots of the original data. **(b)** Consensus imputation method in DreamAI: average of 6 algorithms including 3 baseline methods and 3 winning algorithms from the Challenge. **(c)** Imputation performance (correlation) of all individual imputation

method with and without bagging strategy. Average correlations are reported for different protein strata based on different protein closeness, average abundance, or pseudo missing performance evaluations. **(d)** Performance comparison between DreamAI and all individual methods. Scaled correlation was computed by dividing the performance correlation (imputed values v.s. "ground truth" values) by the correlation between the observed data points of this feature from PNNL- and JHU-data (please see the text). The dashed line in the NRMSD panel represents the background NRMSD between PNNL- and JHU-data based on data points observed in both data sets (please see Online Method). The numbers on the bars represent the ranks of the performance.

Figure 4. **Performance of DreamAI and KNN across different protein strata**: **(a)** protein closeness in observed data; **(b)** NRMSD performance of pseudo missing data of all bagging sets; and **(c)** average protein abundances based on observed data. Scaled correlation was computed by dividing the performance correlation (imputed values v.s. "ground truth" values) by the correlation between the observed data points of this feature from PNNL- and JHU-data (please see the text). NRMSD-δ was the difference of performance NRMSD and background NRMSD (based on data points observed in both data sets).

Figure 5. **For a set of CCRCC tumors, proteomic data with DreamAI imputation shows improved concordance with their corresponding transcriptome data. (a)** Scatter plot of protein-level missing rates vs. mean protein abundances based on observed values in the global proteomics data of 103 CCRCC tumor samples [32]. **(b)** Scatter plot of protein-RNA correlation based on the proteomics data with DreamAI imputation (y-axis) vs. that without imputation (x-axis). **(c)** Scatter plot of significance levels (- log 10 p-value) for testing protein-RNA association based on proteomics data with DreamAI imputation (y-axis) vs. that without imputation (x-axis). **(d)** Number of genes showing significant protein-RNA correlation based on proteomics data with DreamAI imputation (pink) or without imputation (blue) at different p-value cutoffs.

Supplementary Figure Captions

Supplementary Figure 1. **Missing rates and Missing patterns of ovarian cancer global- and phospho-proteomics data [2,7]**. **(a)** Proportion of proteins with different level of missing multiplexes in global- and phospho-proteomics iTRAQ data. **(b)** Proportion of proteins with different level of missing Multiplexes in global- and phospho-proteomics TMT data. **(c)** Scatter plot of protein-level missing rates v.s. mean protein abundances based on observed data in the TMT global-proteomics data set. **(d)** Scatter plot of phosphor-site level missing rates v.s. mean phosphosite abundances based on observed data in the TMT phospho-proteomics data set.

Supplementary Figure 2. Imputation performance of DreamAI with absence of one or all baseline methods on CPTAC2 ovarian cancer data set. (a) Average imputation performance (scaled correlation and NRMSD) of all proteins. **(b) and (c)** Average imputation performance of different protein groups stratified by protein closeness and abundance.

Supplementary Figure 3. Comparing imputation performance (scaled correlation and NRMSD) of baseline methods on CPTAC2 ovarian cancer data set. Scaled correlation was computed by dividing the performance correlation (imputed values v.s. “ground truth” values) by the correlation between the observed data points of this feature from PNNL- and JHU-data (please see the text). The dashed line in the bottom panel represents the background level of NRMSD between PNNL- and JHU-data based on data points observed in both data sets.

Supplementary Figure 4. For the CCRCC NAT (normal adjacent normal) tissue samples, proteomic data with DreamAI imputation shows improved concordance with their corresponding transcriptomic data. (a) Scatter plot of protein-level missing rates vs. mean protein abundances based on observed values in the global proteomics data of 80 CCRCC NAT samples [32]. **(b)** Scatter plot of protein-RNA correlation based on the proteomics data with imputation (y-axis) vs. that without imputation (x-axis). **(c)** Scatter plot of

significance levels (- log 10 p-value) for testing protein-RNA association based on proteomics data with imputation (y-axis) vs. that without imputation (x-axis). **(d)** Number of genes showing significant protein-RNA correlation based on proteomics data with imputation (pink) or without imputation (blue) at different p-value cutoffs.

Supplementary Figure 5. Improved power to detect proteins associated with tumor/normal status or immune subtypes based on the CPTAC-CCRCC proteomic data with imputation by DreamAI than that by KNN. Focusing on 49 proteins with substantially different imputed values by DreamAI and KNN (NRMSD>0.5), the violin plots in **(a)** illustrate the distributions of p-values from two-sample t-tests searching for differential expressed proteins between tumor and NAT samples based on the proteomic data matrix without imputation (grey), with imputation by KNN (light blue) and with imputation by DreamAI (red) respectively. **(b)** is the same as **(a)** except that the p-values were from Kruskai-Wallis tests searching for proteins associated with immune subtypes.

Supplementary Tables

Supplementary Table 1. Comparing results of CI. For each team in the row, number of intervals none-overlapped with (and higher than) the reference team at each column was showing in the table. Table 1A is showing comparison of NRMSD and Table 1B is showing comparision of correlation.

a

NRMSD Confidence Interval	**R1**	**R2**	**R3**	**R4**	**Ref**
Hongyang Li and Yuanfang Guan (R1)		2	2	2	2
DMIS_PTG(R2)	2		0	0	0
BruinGo(R3)	2	0		0	0
Jeremy(R4)	2	0	0		0
KNN(Ref)	2	0	0	0	

b

Correlation Confidence Interval	**R1**	**R2**	**R3**	**R4**	**Ref**
Hongyang Li and Yuanfang Guan (R1)		0	0	0	0
DMIS_PTG(R2)	4		4	4	4
BruinGo(R3)	4	0		0	0
Jeremy(R4)	4	0	4		4
KNN(Ref)	4	0	0	0	

Supplementary Table 2 Comparing results of BF. For each team in the row, BF against the reference team at each column was showing in the table.Table 2A is showing comparision of NRMSD and Table 2B is showing comparision of correlation

a

NRMSD BayesFactor	**R1**	**R2**	**R3**	**R4**	**Ref**
Hongyang Li and Yuanfang Guan (R1)		1	1	1	1
DMIS_PTG(R2)	1		Inf	Inf	Inf
BruinGo(R3)	1	0		1	1.38
Jeremy(R4)	1	0	1		1
KNN(Ref)	1	0	0.72	1	

b

Correlation BayesFactor	**R1**	**R2**	**R3**	**R4**	**Ref**
Hongyang Li and Yuanfang Guan (R1)		0	0	0	0
DMIS_PTG(R2)	Inf		Inf	Inf	Inf
BruinGo(R3)	Inf	0		0	1.04
Jeremy(R4)	Inf	0	Inf		Inf
KNN(Ref)	Inf	0	0.96	0	

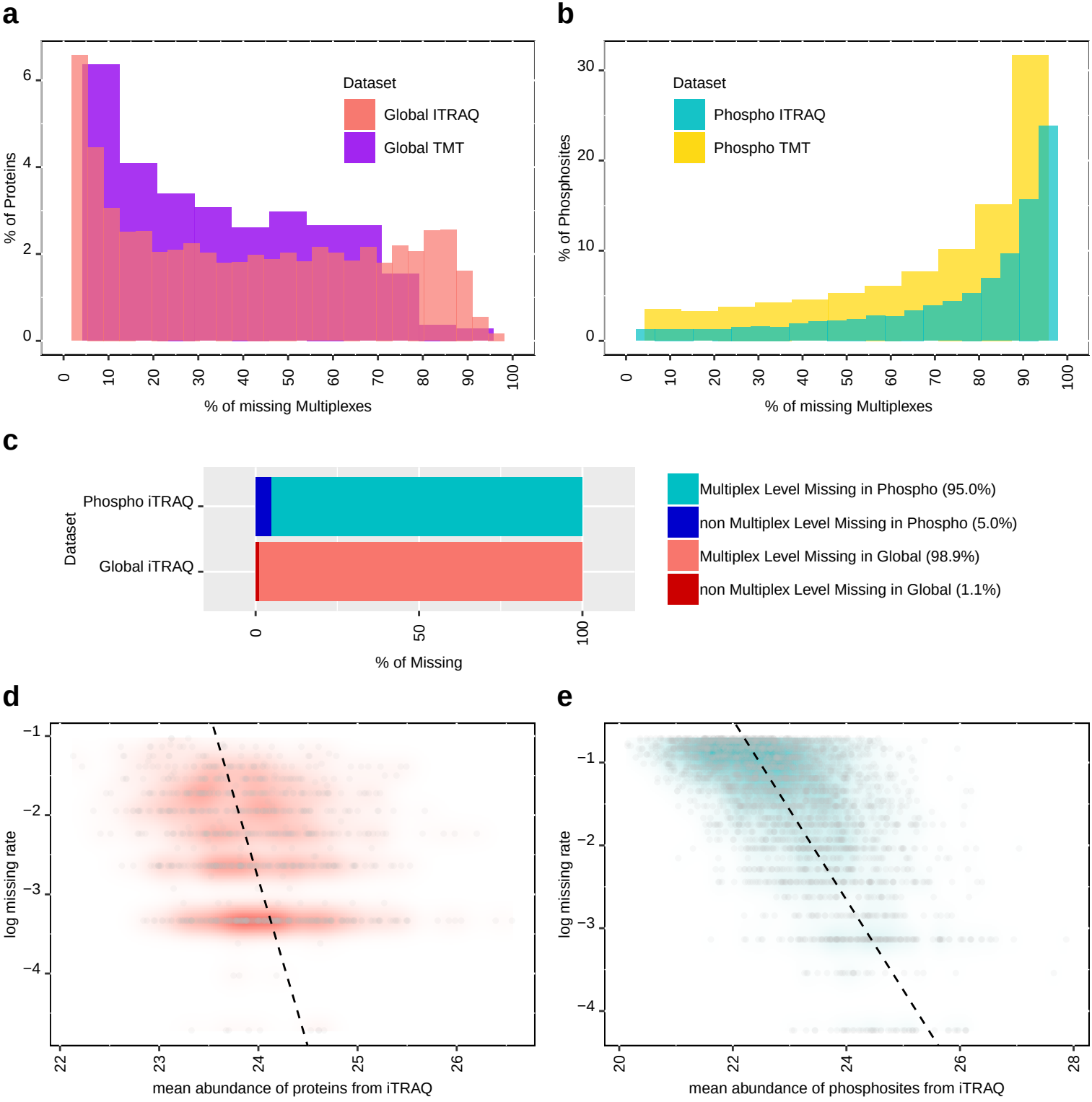
Supplementary Table 3. The NCI-CPTAC Proteogenomics DREAM imputation Challenge Participants

Full Name	Affiliation
Gajendra Pal Singh Raghava	Indraprastha Institute of Information Technology, Delhi, INDIA
Sunil V Kalmady	University of Alberta,Edmonton, Alberta, Canada
Harpreet Kaur	CSIR-Institute of Microbial Technology, Chandigarh, INDIA
Piyush Agrawal	CSIR-Institute of Microbial Technology, Chandigarh, INDIA
Salman Sadullah Usmani	CSIR-Institute of Microbial Technology, Chandigarh, INDIA
Eunji Heo	Deargen Inc. & School of Computing, KAIST, Daejeon, South Korea.
Bora Lee	Deargen Inc., Daejeon, South Korea
Yunpeng Liu	Department of Biology, Massachusetts Institute of Technology, Cambridge MA, USA
Wei Chen,	Department of Biology, Southern University of Science and Technology, Shenzhen, China.
Yue Shan	Department of Biostatistics, The University of North Carolina at Chapel Hill, USA
Hongtu Zhu	Department of Biostatistics, the University of Texas MD Anderson Cancer Center, USA
Kaixian Yu	Department of Biostatistics, the University of Texas MD Anderson Cancer Center, USA
Hongyang Li	Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

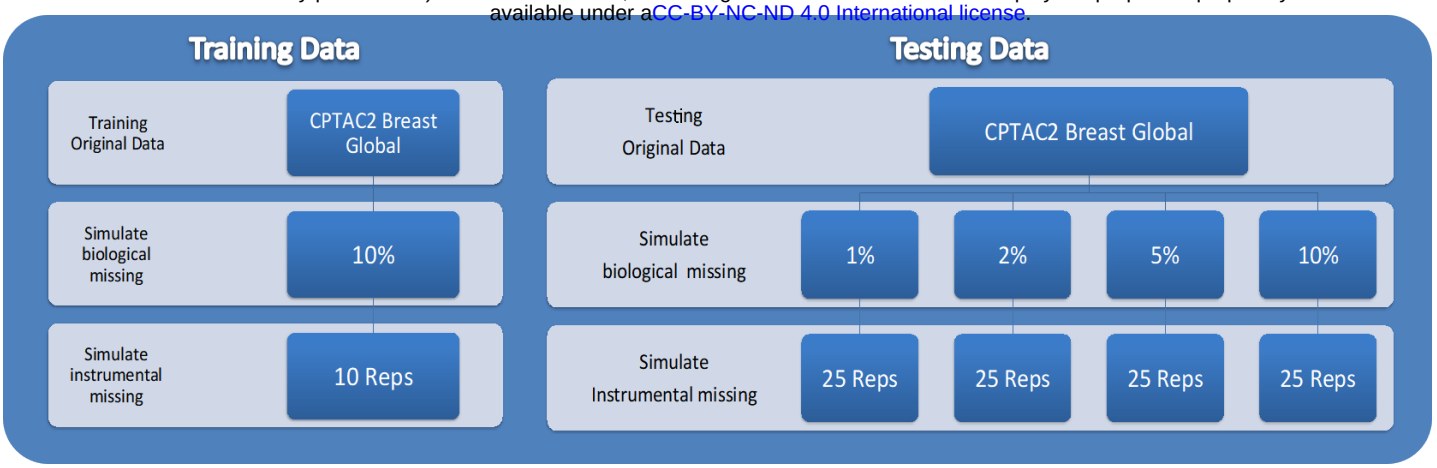
Yuanfang Guan	Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
Jaewoo Kang	Department of Computer Science and Engineering & Interdisciplinary Graduate Program in Bioinformatics, College of Informatics, Korea University
Daehan Kim	Department of Computer Science and Engineering, College of Informatics, Korea University
Keonwoo Kim	Department of Computer Science and Engineering, College of Informatics, Korea University
Minji Jeon	Department of Computer Science and Engineering, College of Informatics, Korea University
Sunkyu Kim	Department of Computer Science and Engineering, College of Informatics, Korea University
Yonghwa Choi	Department of Computer Science and Engineering, College of Informatics, Korea University
Tengfei Li	Department of Radiology, The University of North Carolina at Chapel Hill, USA
Liuqing Yang	Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, USA
Maomao Ding	Department of Statistics, Rice University, USA
Jingyi Jessica Li	Department of Statistics, University of California, Los Angeles, CA, USA
Kexin Li	Department of Statistics, University of California, Los Angeles, CA, USA
Xinzhou Ge	Department of Statistics, University of California, Los Angeles, CA, USA
Huiyuan Chen,	Electrical Engineering and Computer Science, Case Western Reserve University, USA

Ke Hu,	Electrical Engineering and Computer Science, Case Western Reserve University, USA
Kumardeep Chaudhary	Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, 96813, USA
Nai-Wen Chang	Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
Jia Xin Yu,	Icahn School of Medicine at Mount Sinai, New York, New York
Devishi Kesar	Indraprastha Institute of Information Technology, Delhi, INDIA
Sherry Bhalla	Indraprastha Institute of Information Technology, Delhi, INDIA
Mehreen Ali	Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Finland
Ábel Fóthi	Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
Ching-Tai Chen	Institute of Information Science, Academia Sinica, Taipei, Taiwan
Ting-Yi Sung	Institute of Information Science, Academia Sinica, Taipei, Taiwan
Heewon Lee	Interdisciplinary Graduate Program in Bioinformatics, College of Informatics, Korea University
Hwisang Jeon	Interdisciplinary Graduate Program in Bioinformatics, College of Informatics, Korea University
Sandeep Kumar Dhanda	La Jolla Institute for Immunology, La Jolla, CA, USA
Swapnil Mahajan	La Jolla Institute for Immunology, La Jolla, CA, USA

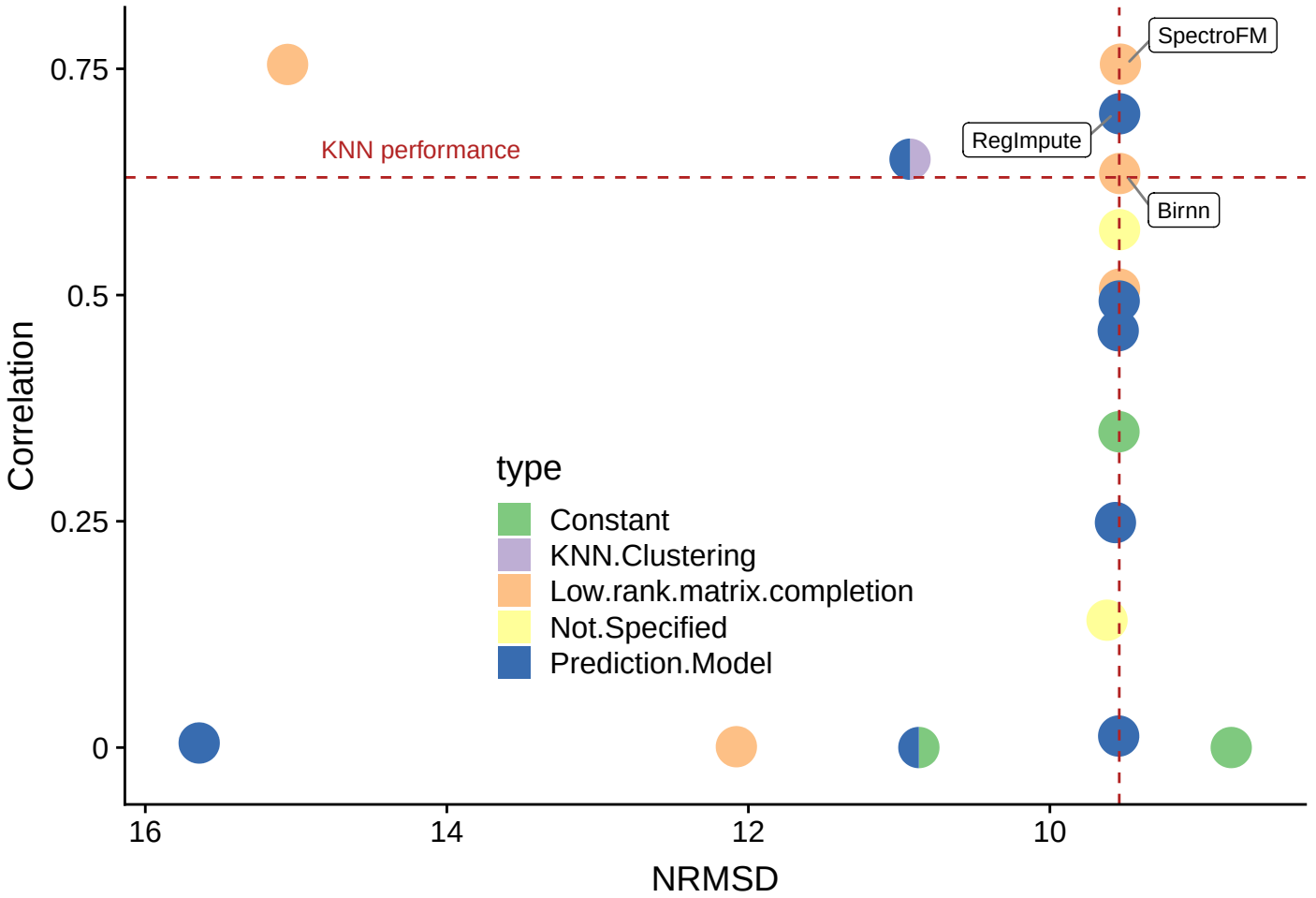
San-Yuan Wang	Master Program in Clinical Pharmacogenomics and Pharmacoproteomics, College of Pharmacy, Taipei Medical University, Taipei, Taiwan
Shujiro Okuda	Niigata University, Niigata, Japan
Yasuhiro Kambara	Niigata University, Niigata, Japan
Laura L. Elo	Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
Mehrad Mahmoudian	Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
Sohrab Saraei	Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
Tomi Suomi	Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
Tommi Välikangas	Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
Russell Greiner	University of Alberta, Edmonton, Alberta, Canada
Roberto Vega	University of Alberta,Edmonton, Alberta, Canada
Jeremy R. Jacobsen	University of Colorado Boulder, Boulder, Colorado, USA


a
% of Proteins
Dataset
Global ITRAQ
Global TMT
% of missing Multiplexes
b
% of Phosphosites
Dataset
Phospho ITRAQ
Phospho TMT
% of missing Multiplexes
c
Dataset
Phospho iTRAQ
Global iTRAQ
% of Missing
Multiplex Level Missing in Phospho (95.0%)
non Multiplex Level Missing in Phospho (5.0%)
Multiplex Level Missing in Global (98.9%)
non Multiplex Level Missing in Global (1.1%)
d
log missing rate
mean abundance of proteins from iTRAQ
e
log missing rate
mean abundance of phosphosites from iTRAQ

a

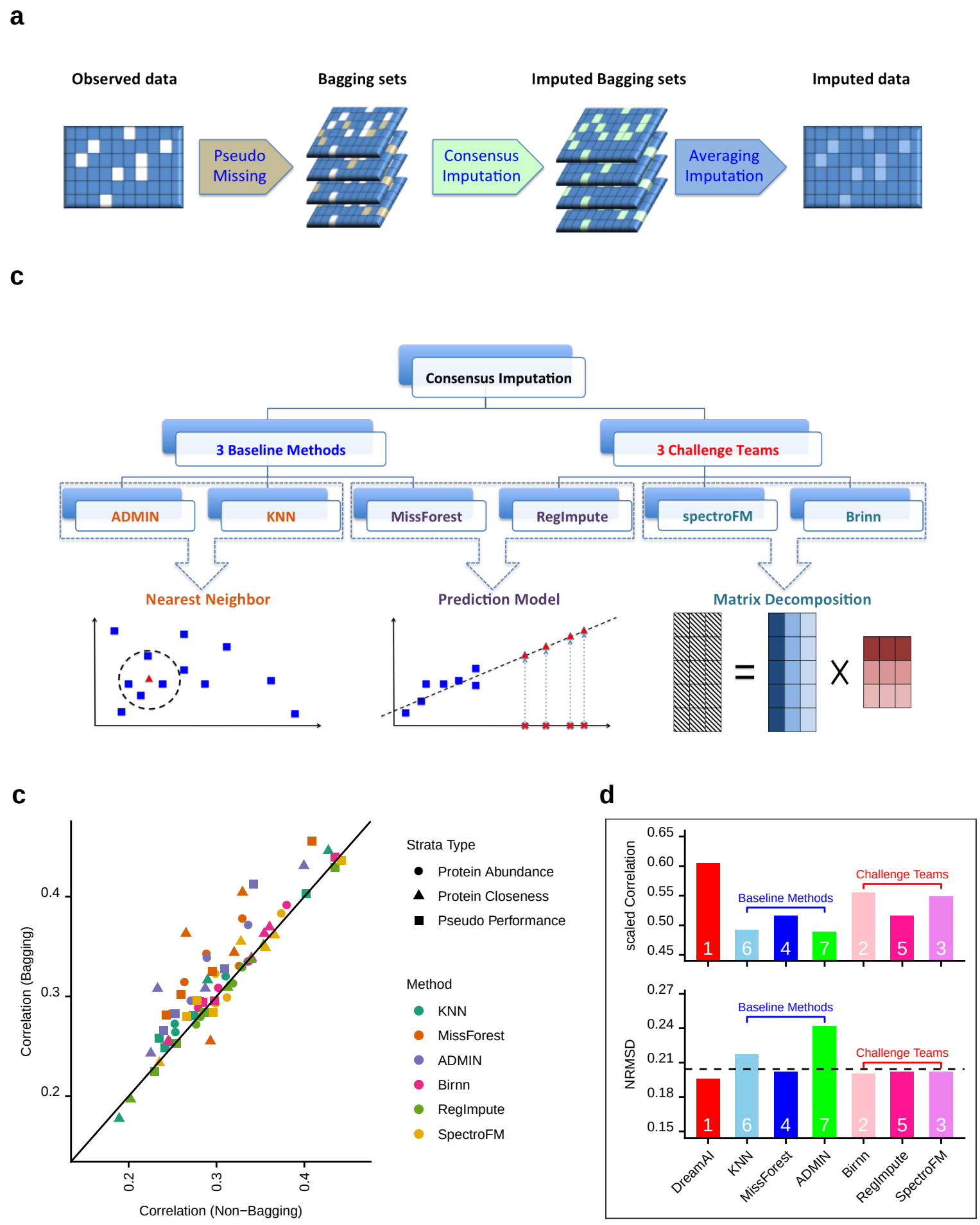


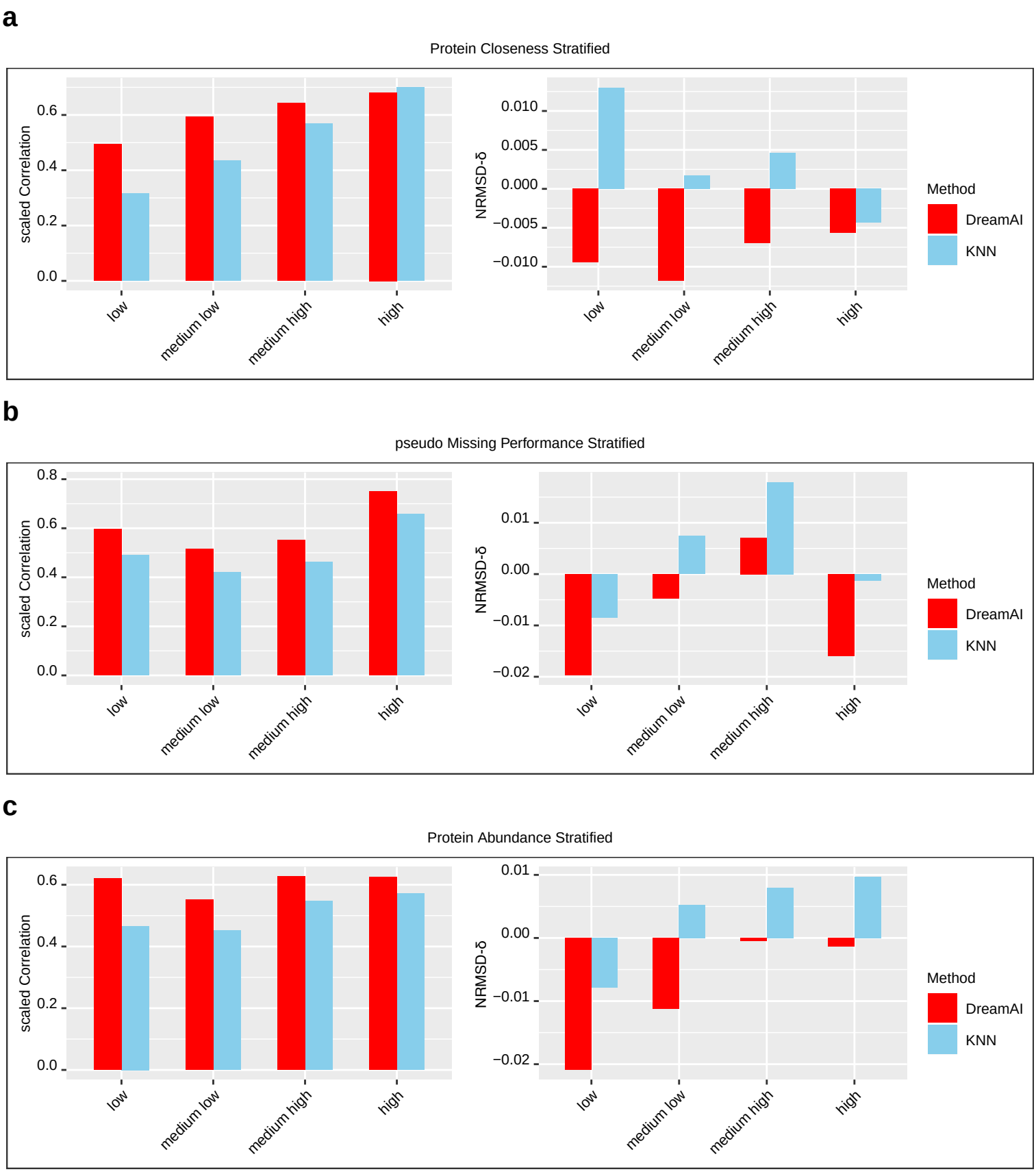
b

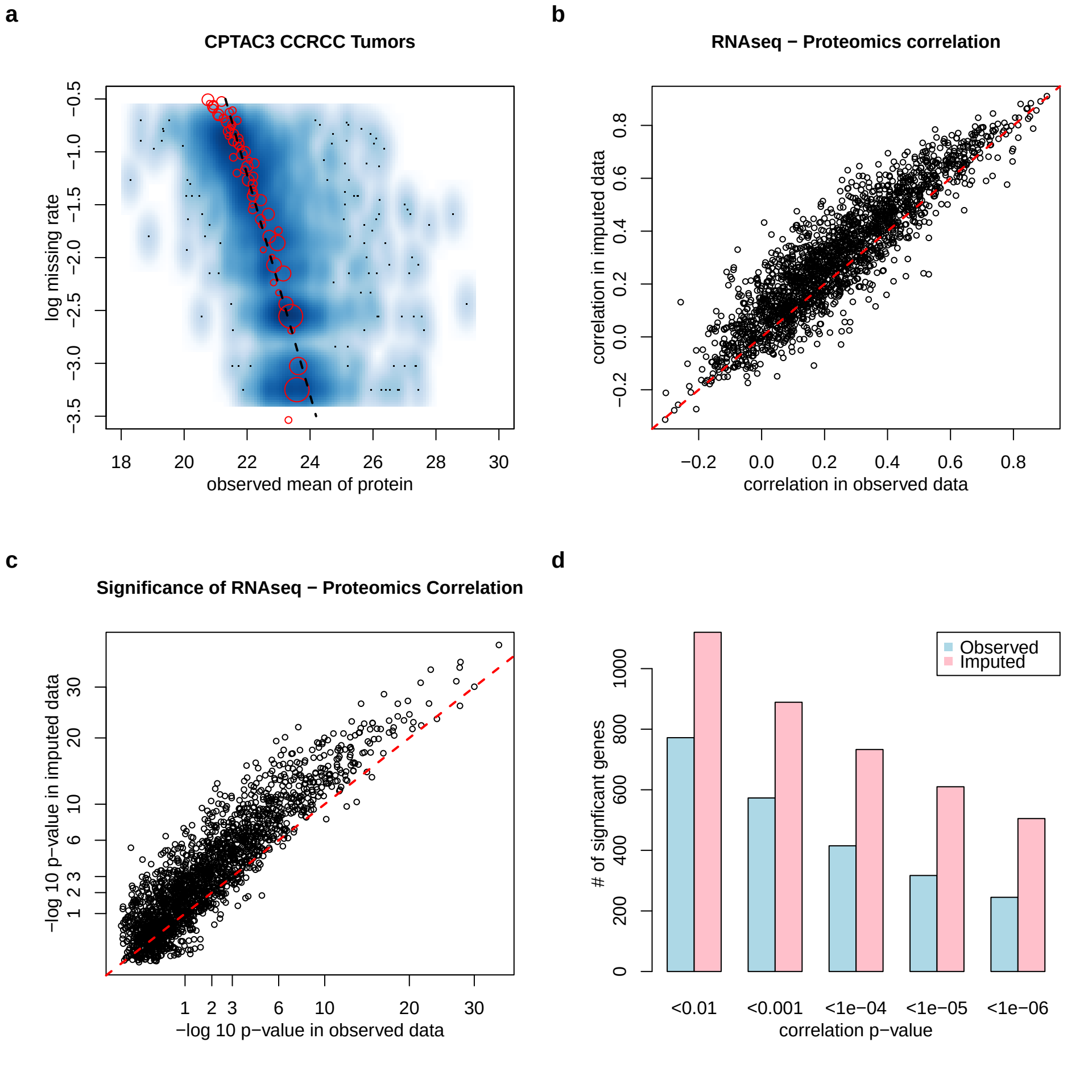


c

Strategies	Type	# of Teams	Corr Ranks	NRMSD Ranks
Constant Imputation	Single Point	3	16.5*/16.5/10	1/8/13*
KNN Clustering	Local Similarity	1	4*	14*
Prediction Model	Global Structure	8	3/4*/8/9/11/13/14/16.5*	4/7/9/10/11/13*/14*/17
Low Rank Matrix Completion	Global Structure	5	1/2/5/7/15	2/3/6/15/16
Not Specified		2	6/12	5/12


a
Observed data
Bagging sets
Imputed Bagging sets
Imputed data
Pseudo Missing
Consensus Imputation
Averaging Imputation
c
Consensus Imputation
3 Baseline Methods
3 Challenge Teams
ADMIN
KNN
MissForest
RegImpute
spectroFM
Brinn
Nearest Neighbor
Prediction Model
Matrix Decomposition
c
Correlation (Bagging)
Correlation (Non−Bagging)
0.2
0.3
0.4
Strata Type
Protein Abundance
Protein Closeness
Pseudo Performance
Method
KNN
MissForest
ADMIN
Birnn
RegImpute
SpectroFM
d
scaled Correlation
NRMSD
Baseline Methods
Challenge Teams
DreamAI
KNN
MissForest
ADMIN
Birnn
RegImpute
SpectroFM


a
Protein Closeness Stratified
scaled Correlation
0.0
0.2
0.4
0.6
low
medium low
medium high
high
NRMSD-δ
0.010
0.005
0.000
−0.005
−0.010
Method
DreamAI
KNN
b
pseudo Missing Performance Stratified
scaled Correlation
0.0
0.2
0.4
0.6
0.8
NRMSD-δ
0.01
0.00
−0.01
−0.02
Method
DreamAI
KNN
c
Protein Abundance Stratified
scaled Correlation
0.0
0.2
0.4
0.6
NRMSD-δ
0.01
0.00
−0.01
−0.02
Method
DreamAI
KNN


a
CPTAC3 CCRCC Tumors
log missing rate
observed mean of protein
b
RNAseq − Proteomics correlation
correlation in imputed data
correlation in observed data
c
Significance of RNAseq − Proteomics Correlation
−log 10 p−value in imputed data
−log 10 p−value in observed data
d
Observed
Imputed
# of signficant genes
<0.01
<0.001
<1e−04
<1e−05
<1e−06
correlation p−value