Feature Impact Assessment: A New Score to Identify Relevant Metabolomics Features in Artificial Neural Networks

Danhui Wang
University of Massachusetts Amherst
Peyton Greenwood
The Ohio State University
Matthias S Klein (klein.663@osu.edu)
The Ohio State University https://orcid.org/0000-0001-7455-5381

Short Report

Keywords: Artificial neural networks, Deep learning, Artificial intelligence, Feature selection, Metabolomics, R

Posted Date: February 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1366354/v1

Abstract

Introduction Artificial Neural Networks (ANN) are increasingly used in metabolomics. Objectives Given the multitude of implementations of ANN, there is no straightforward way to identify important features (metabolites). We developed a simple numeric score, the FIA score, to identify features of high importance. Methods FIA analysis was implemented in R and tested on microbial and human datasets. Results FIA scores correlated significantly to p -values and can provide information on the stability of ANN models. Conclusion FIA scores are a novel, simple score to assess the impact of features that will help interpreting ANN outcomes in the metabolomics area.

1. Introduction

Artificial neural networks and deep learning methods are emerging as powerful new tools in the area of metabolomics (Pomyen *et al.*, 2020; Sen *et al.*, 2021; Wang *et al.*, 2021a). However, while these methods are oftentimes excellent in generating predictive models, the interpretation of the biological backgrounds that underpin these predictions may be complicated. This complication is caused by the fact that the results of such algorithms are produced through the actions of complex networks of mathematical operations, the parameters of which being hard or impossible to visualize, effectively acting as “black boxes”. ANN can be set up in many different configurations, varying the number of layers, changing the connectivity patterns, adding residual connections, and so on. These varying set-ups make interpreting such results even more complicated, as a solution that works on one ANN architecture might not work on other set-ups. In the area of metabolomics, it is usually of high interest to identify the signals or features that allow for predicting an outcome. Metabolomics usually measures a plethora of metabolite intensities, only a few of which might be of importance for the respective study. Also, translating basic research results into clinical practice usually requires the use of a small number of metabolites of interest in a targeted way rather than relying on expensive metabolomics platforms. This practice also needs a tool to identify the most impactful metabolites in a dataset. Unfortunately, identifying these metabolites remains a complicated task when using ANN methods.

Many data analysis methods return their main finding as a simple, numeric result, such as a VIP score (PLS-da, OPLS-da) or a *p*-value (*t*-tests, ANOVA). Such easy-to-interpret numeric values enable researchers to identify features of importance based on agreed-upon metrics and to compare outcomes from different data sets. Such a numeric value has yet to be defined for ANN models.

Here we present a novel score to analyze results from ANN classifiers, named the feature impact assessment (FIA, pronounced as *fee-uh*) score.

2. Materials And Methods

For calculating FIA scores, we need a dataset containing samples from different groups to be able to change selected features while staying within the range that would be observed in real samples. We also

need a well-trained ANN model with categorical predictive outcomes.

To calculate the FIA scores, we start by selecting from the dataset a sample that was correctly predicted by the ANN model. Then, a single feature is chosen and replaced first by the 1% percentile and then by the 99% percentile of that feature (dataset-wide percentiles). If prediction outcomes change in either or both of these cases, we conclude that this feature is of high importance. This feature gets a “raw” FIA score of 1. This process is repeated for all features.

Next, combinations of two features (excluding the features that were positive in the previous step) are chosen and both of these features are consecutively replaced by 1% and 99% percentile values. If at least one of the two resulting predictions differs from the original prediction, both features are assigned a raw FIA value of 2. This process is repeated for all possible pairs. Then, this procedure is repeated for all possible combinations of three features, four features, and so on, assigning raw FIA scores of 3, 4, etc. A faster algorithm for doing this assignment is described in the Supplemental Materials.

This procedure is repeated for all correctly predicted samples of one categorical group. In the end, each feature is assigned the lowest raw FIA score observed in any sample of that group.

The final FIA score for a specific feature is then calculated as the sum of the raw FIA value and a percentage value that is always less than 1 but larger than, or equal to, 0. This percentage value indicates how often this feature was **not** observed at this raw FIA score in the dataset. If the feature has the same raw FIA score in all samples, the fractional part is 0%, making the final FIA score 1.0 (if raw FIA = 1). If it is observed in only 5% of samples, the fractional part is 1-0.05=.95, making the final FIA score 1.95. It is obvious that a feature found in many samples will be given a stronger impact score, and thus FIA = 1.0 has higher impact than FIA = 1.95 in this schematic.

It is noteworthy that the integer part of the final FIA score, that is the digits to the left of the decimal point, are always equal to the raw FIA score. The raw score indicates whether changing the feature can change the prediction outcome on its own or only in a combination of multiple features. Features that can change the prediction outcome on their own will be assigned a stronger impact value than features that can change this outcome only when combined with other features, hence FIA = 1.0 is more impactful than, for example, FIA = 3.0.

Not all features in a dataset will be assigned a FIA score. Features without measurable impact on the prediction accuracy will not be assigned a score.

We tested the FIA algorithm on data from an experiment on microbial metabolism (Wang *et al.*, 2021a). Ten microbial strains were cultivated in a defined medium at 37 °C (Wang *et al.*, 2021b). After 4 hours, medium samples were collected and measured by 1D ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy on an 850 MHz Avance III HD Ascend spectrometer (Bruker BioSpin, Billerica, MA, USA). Spectra were binned with a bin width of 0.005 pm using mrbin (Klein, 2021). From each spectrum, a spectrum of pure medium was subtracted, resulting in data that reflect the changes of feature intensities

during microbial growth; positive values indicate increasing signals, while negative values indicate decreasing signals. A fully connected ANN with two equally sized hidden layers of 800 neurons each was trained on the dataset, with 10 output neurons corresponding to the 10 bacterial strains in the dataset (Wang *et al.*, 2021a). Calculations were performed in R (3.5.1) using the packages keras (2.6.0) and tensorflow (2.6.0). FIA scores were calculated using the R package mrbin (1.6.3).

For comparison, two-tailed unpaired one-sample *t*-tests were used to test each feature for significant changes during microbial growth of each bacterial strain. *p*-values were corrected for multiple testing using false discovery rate controlling (FDR) at the 20% level.

Volcano plots were generated by plotting the negative decadic logarithm of the FIA scores versus the intensity change of this feature during microbial growth.

3. Results And Discussion

The FIA score was defined as a simple numeric value that allows us to assess the importance of a given feature for a specific classification. A FIA score can take values between 1 and the number of features of the respective data; one example would be FIA = 1.0 or FIA = 4.33. Lower FIA scores indicate more impactful features. A feature that was not assigned an FIA score has no measurable impact on the classification results in the analyzed dataset.

The FIA method was tested on a dataset of microbial metabolites measure by means of 1D ^{1}H NMR (Wang *et al.*, 2021a). The data set consisted of NMR spectra from various bacterial strains grown in a culture medium. This dataset has the advantages of having well-defined, distinctive groups and controlled experimental conditions, thus minimizing variability. Also, the metabolomics data have been characterized in an earlier publication. The data set contained 80 spectra with 1384 bins (features), some of which have been previously annotated with metabolite identities.

3.1 Visualization of FIA scores

We visualized the obtained FIA scores using adapted volcano plots, shown in Fig. 1. Features with FIA scores of less than 10 are highlighted in red (decreasing during microbial growth) or blue (increasing during growth). Identified signals were labeled in the plots. These plots bear great resemblance to the volcano plots based on *p*-values obtained for this data (Wang *et al.*, 2021a). It should be noted that not all features in a dataset will be assigned a FIA score, therefore not all features are shown in this kind of volcano plot, in contrast to “regular” volcano plots. Next, we further investigated the relationship between FIA scores and *p*-values.

3.2 Relationship of FIA scores to p-values

We hypothesized that the FIA algorithm should be able to identify features that were identified as significant by statistical tests, assuming that successful ANN models will make (at least partial) use of

such features when predicting outcomes. For this, we compared *p*-values of features with FIA scores <4 to all other features of this group. This comparison is shown in Fig. 2A for all groups in the dataset. It is obvious that low FIA scores were connected to lower *p*-vales in all groups. In all but one group, this difference was significant ($p \leq 0.05$). The group that was not significant ($p = 0.147$) is *Pseudomonas*. It is noteworthy that in the original publication on this dataset, this group had by far the lowest number of significant features, combined with low overall change in feature intensities, potentially linked to relatively slow growth rates of *Pseudomonas* (Wang *et al.*, 2021a). Groups with only a few significant features might force ANN to use more and less significant features to predict the outcome, as compared to groups with strongly correlated features. From the volcano plots we concluded that FIA scores of less than 4 seemed to be the most impactful in this dataset.

To further investigate the relationship between FIA scores and *p*-values, we modeled the number of FIA scores less than 4 versus the number of significant features after FDR correction (Fig. 2B). A strong and significant inverse linear correlation was observed ($p = 0.0211$). This inverse correlation means that if many FIA scores were <4, there were only few significant features. Groups with few significant *p*-values are poorly defined and it makes sense that many different features can disrupt correct prediction of group membership in such cases. On the other hand, there will be only a few features capable of changing the prediction at FIA < 4 for well-defined groups (high number of significant *p*-values).

While the cutoff of 4 worked well to characterize features of high importance in this dataset, there may be other datasets and/or ANN models in which this cutoff might be too high or too low to deliver sensible outcomes. In these cases, it might be better to analyze, for example, the lowest 100 FIA scores (“top 100”). Fig. 2C shows the correlation between the mean of the top 100 FIA scores and the number of significant *p*-values. Even stronger positive correlations were observed in this case ($p = 0.00084$). This result means that lower averages of the FIA top 100 were found in less well-defined groups and vice versa. In this way, analysis of the characteristics of observed FIA scores may provide additional information about how well-defined a group is regarding metabolite signatures. In case less than 100 FIA scores are available, similar trends are seen when using the top 10 FIA scores (Figure S1 in the Supplemental Materials).

In conclusion, the FIA score method identified features that also had significant *p*-values in a separate statistical test. It is important to notice that in contrast to *t*-tests, FIA score calculation has no prerequisites such as data normality, making FIA scores more generally applicable than hypothesis tests.

FIA algorithm performance was validated on an additional dataset of human samples (Shearer *et al.*, 2021), employing a different ANN architecture. Results were comparable and are shown in the Supplemental Materials. Interestingly, FIA analysis was able to identify features of interest that were missed by other data analysis approaches in this validation dataset.

3.3 Recommendations for interpreting FIA scores

Based on our analyses, we recommend the following rules for calculating and interpreting FIA scores:

For calculations, a real-life dataset containing a variety of samples from the different observed groups is required to allow for selecting meaningful 1% and 99% percentile feature values.

FIA scores of 1.0 indicate features of maximum impact on the prediction in all samples of the dataset. Scores between 1.01 and 1.99 indicate very strong impact, but only in part of the samples. Scores between 2.0 and 3.99 indicate strong impact of this feature. Scores equal to or larger than 4 indicate signals of medium to low impact. These recommendations are summarized in Table S1 in the Supplemental Materials.

FIA < 4 seemed to be sensible cutoff to find important features in the analyzed data. As other datasets and/or predictive models might require more complex combinations of features, no FIA scores of less than 4 might be observed. In these cases, analyzing the top 10 or top 100 FIA scores might be a more meaningful analysis in these cases. One should notice that all features that share the minimum FIA score are of equal importance. For example, if the lowest observed FIA score is 20.5, and there are 20 features with FIA = 20.5, all of these 20 features should be considered in the analysis, not just the top 10, as the feature order will be arbitrarily chosen in the case of ties.

4. Conclusion

We here present a new, numeric metric to assess the impact of features on the prediction outcome of ANN. FIA analysis works independently of the ANN network architecture by varying the original data and comparing classification results. It also works for non-ANN classifiers, although its use might be of limited interest in these cases, as such methods usually require advanced feature selection ahead of training and thus provide information on which features are of importance.

FIA scores will be of most value in datasets where each feature bears a meaning in itself, such as metabolomics datasets or similar data. In other common applications of ANN, such as speech or picture recognition, features might only gain meaning in combination with other features, in these cases the FIA analysis might not be as useful.

The FIA algorithm was implemented in the function FIA in the R package mrbin, version 1.6.2 or higher, available at https://CRAN.R-project.org/package=mrbin (Klein, 2021). Sample R code to perform FIA analysis can be found in the Supplemental Materials.

Statements And Declarations

Funding

This research was funded by the USDA National Institute of Food and Agriculture, Hatch project 1018603, and The Ohio State University's Foods for Health Discovery Theme.

Data Availability Statement

The metabolomics and metadata reported in this paper are available via the Zenodo repository at https://doi.org/10.5281/zenodo.6037351.

Software or Database Availability Statement

The software developed in this study is available via https://CRAN.R-project.org/package=mrbin

Acknowledgments

The authors would like to thank Professor Ahmed Yousef for providing bacterial strains, and Molly Davis for helpful discussions.

Conflict of Interest

D.W. declares they have no conflict of interest. P.G. declares they have no conflict of interest. M.S.K. declares they have no conflict of interest.

Author Contributions

Conceptualization, M.S.K.; methodology, M.S.K. and D.W.; software, M.S.K.; validation, M.S.K.; formal analysis, M.S.K., D.W., and P.G.; investigation, M.S.K., D.W., and P.G.; resources, M.S.K.; data curation, M.S.K.; writing—original draft preparation, M.S.K.; writing—review and editing, M.S.K., D.W., and P.G.; visualization, M.S.K.; supervision, M.S.K.; project administration, M.S.K.; funding acquisition, M.S.K. All authors have read and agreed to the published version of the manuscript.

References

1. Klein, M. S. (2021). Affine Transformation of Negative Values for NMR Metabolomics Using the mrbin R Package. *Journal of Proteome Research*, 20, 1397–1404
2. Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal*, 18, 2818–2825
3. Sen, P., Lamichhane, S., Mathema, V. B., McGlinchey, A., Dickens, A. M., Khoomrung, S., & Orešič, M. (2021). Deep learning meets metabolomics: a methodological perspective. *Briefings in Bioinformatics*, 22, 1531–1542
4. Shearer, J., Klein, M. S., Vogel, H. J., Mohammad, S., Bainbridge, S., & Adamo, K. B. (2021). Maternal and Cord Blood Metabolite Associations with Gestational Weight Gain and Pregnancy Health Outcomes. *Journal of Proteome Research*, 20, 1630–1638
5. Wang, D., Greenwood, P., & Klein, M. S. (2021a). Deep Learning for Rapid Identification of Microbes Using Metabolomics Profiles. *Metabolites*, 11, 863
6. Wang, D., Greenwood, P., & Klein, M. S. (2021b). A Protein-free Chemically Defined Medium for the Cultivation of Various Microorganisms with Food Safety Significance. *Journal of Applied*

Microbiology, 131, 844–854

Figures

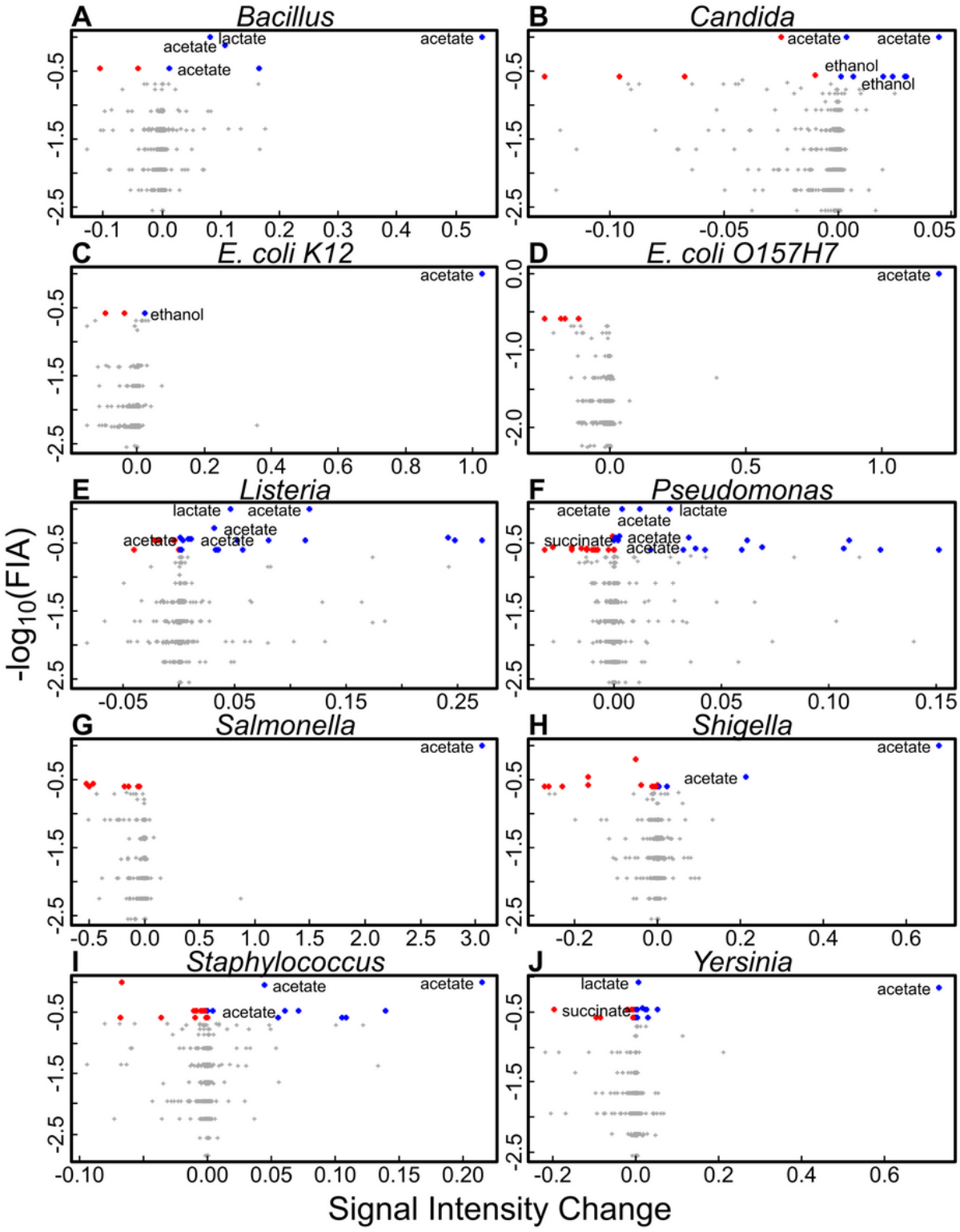

Figure 1

Volcano plots showing -$\log_{10}$ of FIA scores versus signal intensity change for all groups in the data set (**A**-**J**). FIA scores less than 4 are highlighted in blue (increasing/to the right) or red (decreasing/to the left).

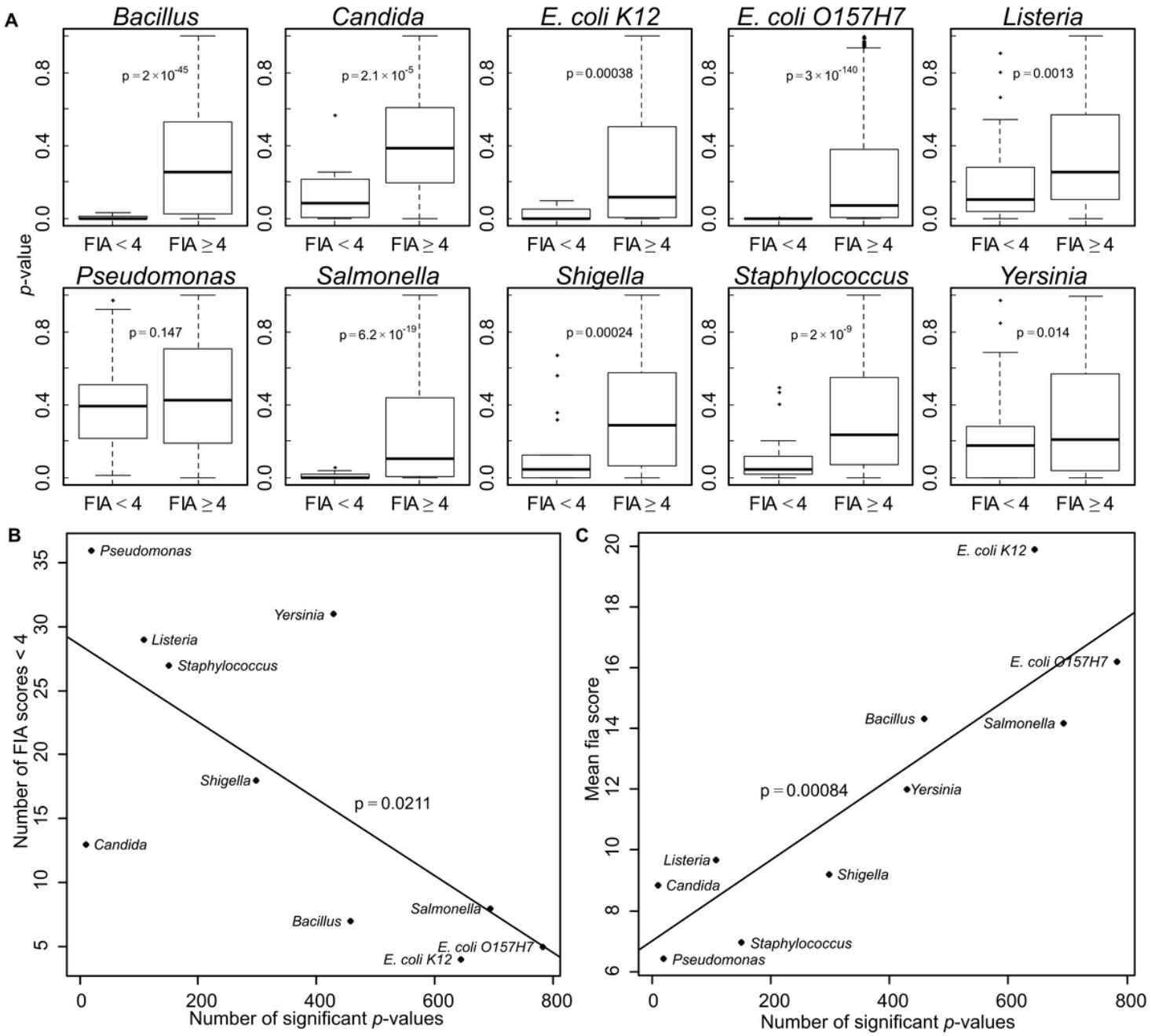

Figure 2

(**A**) Boxplots of *p*-values of features with FIA score <4 compared to all other features. (**B**) Number of FIA scores <4 versus number of significant *p*-values for this group. (**C**) Mean value of the top 100 FIA scores versus the number of significant *p*-values for this group.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FIASupplementalMaterials.pdf