MESSAGES

16 of 18

Ruiquan Ge via Frontiers: Action needed: Interactive review for your manuscript has been activated - 885929

Ruiquan Ge
To me

Manuscript

18 Mar 2022 03:46 PM (GMT)

Dear Dr Indriani,

The interactive review of your manuscript "ProtTrans-Glutar: Incorporating Features from Pre-trained Transformer-Based Models for Predicting Glutarylation Sites" submitted to Frontiers in Genetics, section Computational Genomics has now been activated.

The reviewers recommended that you make substantial amendments to your manuscript. Please respond within the next 14 days to all comments raised by the reviewers and editor in the online review forum. If a reviewer has finalized the review and discussion on the Reviewer tab is closed, you should submit a reply to pending comments in a new thread in the Editor tab. You can also submit a revised version of your manuscript at that time. We encourage you to submit your documents with tracked changes to highlight the revisions.
There can be more than one iteration between authors and reviewers, but only when all comments by each reviewer have been addressed successfully can the review be finalized.

To access the review forum and respond to the reviewers, please click on the following link:
http://www.frontiersin.org/Review/EnterReviewForum.aspx?activationno=800d9a40-827e-4f4b-b0cb-db0b208cb490

Journal: Frontiers in Genetics, section Computational Genomics
Article type: Original Research
Manuscript title: ProtTrans-Glutar: Incorporating Features from Pre-trained Transformer-Based Models for Predicting Glutarylation Sites
Manuscript ID: 885929
Authors: Fatma Indriani, Kunti Robiatul Mahmudah, Bedy Purnama, Kenji Satou
Submitted on: 28 Feb 2022
Interactive review started on: 18 Mar 2022

Please do not hesitate to contact us if you have any questions. Your timely response would be much appreciated. Note that if we do not hear from you by the revision deadline, the editorial office reserves the right to withdraw your manuscript from consideration for publication, as we cannot hold manuscripts in review without any updates from the authors.

With best regards,

Ruiquan Ge
Guest Associate Editor,
www.frontiersin.org

--

Frontiers: Congratulations! Your article is published

Genetics Production Office
To me

Manuscript

31 May 2022 04:43 AM (GMT)

Dear Fatma Indriani,

Genetics Production Office has sent you a message. Please click 'Reply' to send a direct response

Congratulations on the publication of your article: ProtTrans-Glutar: Incorporating Features from Pre-trained Transformer-Based Models for Predicting Glutarylation Sites, by Fatma Indriani, Kunti Robiatul Mahmudah, Bedy Purnama, Kenji Satou, published in Frontiers in Genetics, section Computational Genomics.

To view the online publication, please click here:

http://journal.frontiersin.org/article/10.3389/fgene.2022.885929/full?&utm_source=Email_to_authors_&utm_medium=Email&utm_content=T1_11.5e1_author&utm_campaign=Email_publicat

This article is an open access publication accessible to readers anywhere in the world. Share the link with your network and track the impact of your research with our Article and Author Impact Metrics. This includes metrics on citations, views and downloads, as well as the social media attention your article receives.

If you have not done so already, please update your Loop profile to maximise your readership:

http://loop.frontiersin.org/people/me/?utm_source=WFPOFAut&utm_medium=Email&utm_campaign=WF11.5E-1

Authors with fully populated profiles receive 4X more profile views and 6X more publication views.

*** BE AWARE OF SCAM ATTEMPTS - PLEASE READ **********

There has been a recent surge in fraud attempts against Frontiers authors. Please be alert if someone contacts you from a fake email address that pretends to be a Frontiers Staff member providing you with a fake invoice or payment instructions.
- Frontiers will always contact you from a @frontiersin.org or @frontiersin.com email address.
- We never amend payment details via email.
- Scammers might try to contact the corresponding author and/or the payer to request a payment, or to request an additional payment. The accurate invoice is always available in the payer's account and in the corresponding author's account (if different) on our website. If you don't have one yet, please register with this email address. You will find the invoice in My Frontiers > My Invoices

If you have any doubt or believe to have followed the wrong instructions, please reach out to us at accounting@frontiersin.org
****

We look forward to your future submissions!

Best regards,

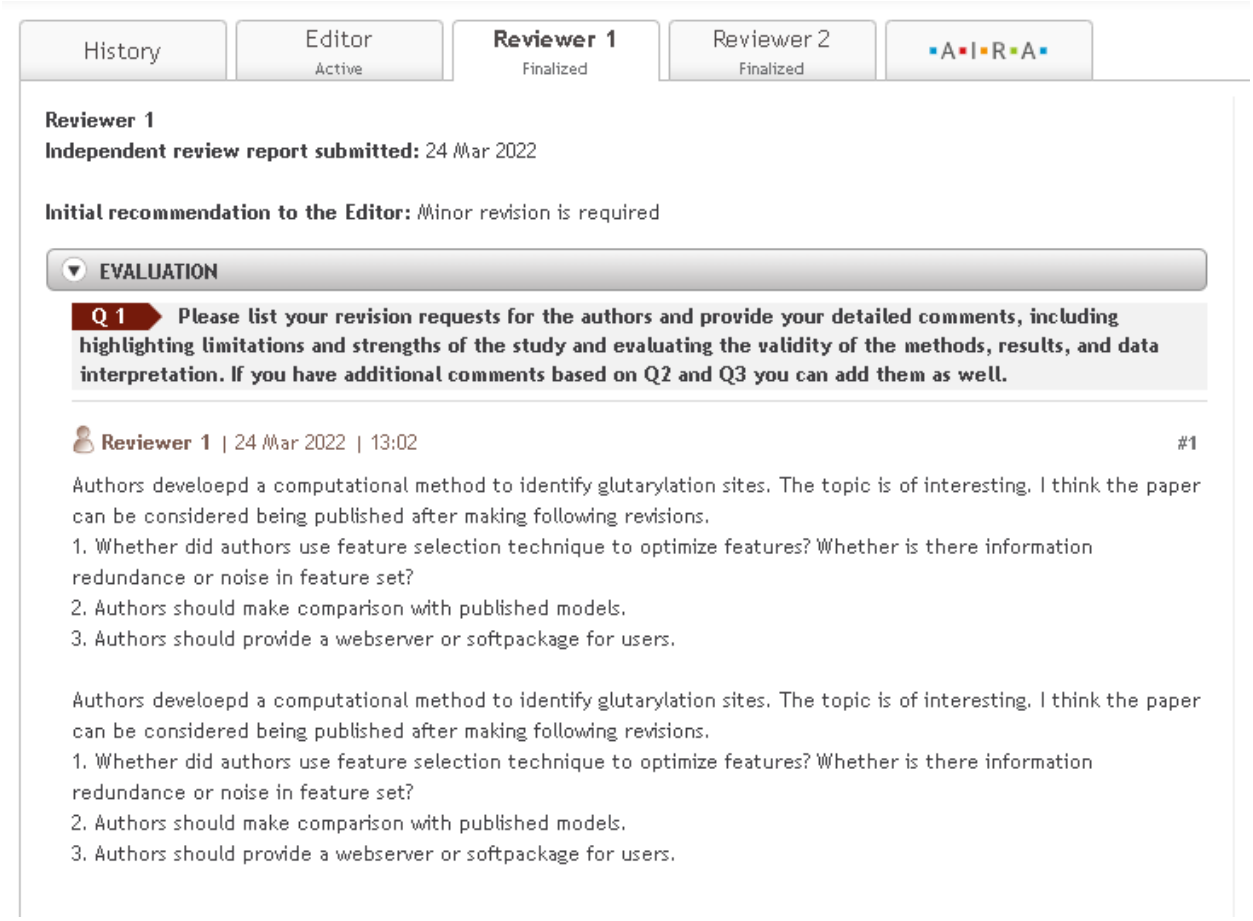
Frontiers Genetics Production Office
genetics.production.office@frontiersin.org
www.frontiersin.org

History

History | Editor (Active) | Reviewer 1 (Finalized) | Reviewer 2 (Finalized) | AIRA

Date	Updates
26 Apr 2022	Article accepted for publication.
25 Apr 2022	Corresponding Author Fatma Indriani posted new comments in the Editor tab.
21 Apr 2022	Guest Associate Editor Ruiquan Ge posted new comments in the Editor tab.
20 Apr 2022	Corresponding Author Fatma Indriani posted new comments in the Editor tab.
18 Apr 2022	Guest Associate Editor Ruiquan Ge posted new comments in the Editor tab.
	Corresponding Author Fatma Indriani re-submitted manuscript.
17 Apr 2022	Editorial Office reminded you to respond to a comment in the Editor tab.
	Editorial Office reminded you to respond to a comment in the Editor tab.
	Editorial Office reminded you to respond to a comment in the Editor tab.
	Editorial Office reminded you to respond to a comment in the Editor tab.
13 Apr 2022	Guest Associate Editor Ruiquan Ge posted new comments in the Editor tab.
	Guest Associate Editor Ruiquan Ge requested Corresponding/Submitting Author to revise the manuscript.
12 Apr 2022	Review of Review Editor 2 finalized.
08 Apr 2022	Corresponding Author Fatma Indriani re-submitted manuscript.
24 Mar 2022	Review of Reviewer 1 is finalized.
18 Mar 2022	Interactive review forum activated.
28 Feb 2022	Corresponding Author Fatma Indriani submitted manuscript.

REVIEWER 1


History
Editor
Active
Reviewer 1
Finalized
Reviewer 2
Finalized
▪A▪I▪R▪A▪
Reviewer 1
Independent review report submitted: 24 Mar 2022
Initial recommendation to the Editor: Minor revision is required
EVALUATION
Q 1
Please list your revision requests for the authors and provide your detailed comments, including highlighting limitations and strengths of the study and evaluating the validity of the methods, results, and data interpretation. If you have additional comments based on Q2 and Q3 you can add them as well.
Reviewer 1 | 24 Mar 2022 | 13:02
#1
Authors develoepd a computational method to identify glutarylation sites. The topic is of interesting. I think the paper can be considered being published after making following revisions.
1. Whether did authors use feature selection technique to optimize features? Whether is there information redundance or noise in feature set?
2. Authors should make comparison with published models.
3. Authors should provide a webserver or softpackage for users.
Authors develoepd a computational method to identify glutarylation sites. The topic is of interesting. I think the paper can be considered being published after making following revisions.
1. Whether did authors use feature selection technique to optimize features? Whether is there information redundance or noise in feature set?
2. Authors should make comparison with published models.
3. Authors should provide a webserver or softpackage for users.

Q 2 **Check List**

Reviewer 1 | 24 Mar 2022 | 13:02 #1

a. Is the quality of the figures and tables satisfactory?
- Yes

b. Does the reference list cover the relevant literature adequately and in an unbiased manner?
- Yes

c. Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)
- Yes

d. Is a statistician required to evaluate this study?
- No

e. Are the methods sufficiently documented to allow replication studies?
- Yes

QUALITY ASSESSMENT

Q 3	Rigor	
Q 4	Quality of the writing	
Q 5	Overall quality of the content	
Q 6	Interest to a general audience	

REVIEWER 2

History | Editor Active | Reviewer 1 Finalized | **Reviewer 2** Finalized | ▪A▪I▪R▪A▪

Reviewer 2
Independent review report submitted: 18 Mar 2022
Interactive review activated: 18 Mar 2022
Review finalized: 12 Apr 2022

Initial recommendation to the Editor: Major revision is required

EVALUATION

Q 1 **Please list your revision requests for the authors and provide your detailed comments, including highlighting limitations and strengths of the study and evaluating the validity of the methods, results, and data interpretation. If you have additional comments based on Q2 and Q3 you can add them as well.**

Reviewer 2 | 18 Mar 2022 | 08:24 #1

In this work, the authors developed a machine learning model named ProtTrans-Glutar to predict glutarylation sites. The first impression is that the manuscript is very poorly written. It must be sent to English editing service so that the work can be understood by audience. The second impression is that the authors did not pay much attention to present their work in a clear manner and it seems that the authors rush to submit the manuscript. A lot of first-used abbreviations without full-names. Although the authors have put a lot of effort in the experiments, the poor presentation reduces the value of the study. Furthermore, the prediction performance of ProtTrans-Glutar is not better than existing predictors. In addition, the method for feature extraction (from BERT or BERT-based models) are not new.

My comments are below:

Q1. Several places in the abstract are not clear.

For example, this sentence: “In this study, we propose a model named ProtTrans-Glutar to classify protein sequence into positive glutarylation or negative by combining traditional sequence based features with features derived from pre-trained transformer-based protein model”. It is not clear whether the authors want to perform sequence classification or residue classification.

Furthermore, what is CTD?

Q2. In the abstract, there are some language mistakes. For example, the below sentence should be separated into 2 sentences to make them grammatically correct.

“Identifying glutarylated peptides using proteomics techniques is expensive and time consuming, it is important to investigate computational models and predictors to identify glutarylation speedily”.

→ should be

“Identifying glutarylated peptides using proteomics techniques is expensive and time consuming. It is important to investigate computational models and predictors to identify glutarylation speedily”.

Furthermore, “The features for the model is constructed...” should be “The features for the model are constructed ...”
“... several feature set” should be “... several feature sets”
“... gave recall, specificity, and AUC scores 0.7864, 0.6286, and 0.7075 respectively” should be “... obtained recall, specificity, and AUC scores of 0.7864, 0.6286, and 0.7075, respectively”
“... to identify new glutarylation sited ” should be “... to identify new glutarylation sites”
“... the performance of these methods is still limited” should be “... the performance of these methods are still limited”
I believe there are still more typos and grammar mistakes and I don’t list them here. The author must send the manuscript to English editing service so that the writing is of scientific standard. After that step, it can be considered for publication.
The abstract is the first thing that tells what the authors do and it is often written very carefully. The authors should pay a lot of attention to this.

Q3. PTM identification using features derived from BERT has been studies by several groups (Please see in the paper list below)
[1] Ho, Quang-Thai, Nguyen Quoc Khanh Le, and Yu-Yen Ou. "FAD-BERT: improved prediction of FAD binding sites using pre-training of deep bidirectional transformers." Computers in Biology and Medicine 131 (2021): 104258.
BERT-based features can be used for protein family classification and for DNA sequences.
[2] Le, Nguyen Quoc Khanh, Quang-Thai Ho, Trinh-Trung-Duong Nguyen, and Yu-Yen Ou. "A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information." Briefings in bioinformatics 22, no. 5 (2021): bbab005.
[3] Shah, Syed Muazzam Ali, Semmy Wellem Taju, Quang-Thai Ho, and Yu-Yen Ou. "GT-Finder: Classify the family of glucose transporters with pre-trained BERT language models." Computers in Biology and Medicine 131 (2021): 104259.
[4] Shah, Syed Muazzam Ali, and Yu-Yen Ou. "TRP-BERT: Discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT." Computers in Biology and Medicine 137 (2021): 104821.
Q4. From this sentence, it is not clear about the dataset.
“The dataset comprised 2103 training set and 247 testing set, and is imbalanced with 444 positive sites and 1906 negative sites”
Let’s ignore the typos, how many datasets were used? 2103 and 247 are the number of protein sequences? From the sequences, the authors retrieves the binding sites?
The author should read the below papers and follow the way the authors express their data processing method to comprise the datasets.
[5] Kusuma, Rosdyana Mangir Irawan, and Yu-Yen Ou. "Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network." Journal of Molecular Graphics and Modelling 92 (2019): 86-93.
[6] Tran, The-Anh, Dinh-Minh Pham, and Yu-Yen Ou. "Incorporating a transfer learning technique with amino acid embeddings to efficiently predict N-linked glycosylation sites in ion channels." Computers in Biology and Medicine 130 (2021): 104212.
I am also wondering why 2103 training and 247 test sets yeild only 444 positive sites and 1906 sites? For binding sites prediction using machine learning, the number of samples (binding and non binding sites) are often large. Please check your manuscript again and provide an explanation or a step-by-step description of how you retrieve your data.
Q5. In the feature extraction section, please give the full-names of the 7 methods for AAC, EAAC, CTDC, CTDT, CTDD, PAAC, and APAAC as this is the first time they appear in the manuscript.

Q6. I suggestthe authors explain AAC, EAAC, CTDC, CTDT, CTDD, PAAC, and APAAC by using an example and put them in the Supplementary document. From one protein sequences, i.e. with a length of 50, how these 7 features are calculated. It will make the manuscript clearer and easy the readers.
Q7. In section 2.2.2, there should be an explanation for N(s,r).
Q8. The formulae for MCC is wrong. The numerator should be TP x TN - FP x FN instead.
Q9. What does SBF in section 3.2 mean?
Q10. In section 3.3, based on results from Table 7, it is hard to conclude that ProtTrans-Glutar is the best model as both its AUC and MCC are not higher than RF-GlutarySite.
Please note that even though RF-GlutarySite used balanced test data, to make a fair comparison, the authors can also compare the performance on balanced test data.
Q11. The data and source code are not provided. I suggest the authors deposit these things in a public repository such as github. This is the minimum thing the author should do to allow replication. Some other groups provide web server for prediction.

Corresponding Author: Fatma Indriani | 08 Apr 2022 | 23:37 #2

Q1. The classification task is sequence classification. We have modified the abstract to make it clearer.

Composition, Transition and Distribution (CTD) is a protein descriptor introduced by (Dubchak, 1995) and has been used extensively throughout the years. CTD is one of the many protein sequence feature extraction methods available in the iFeature Python package which we use. The other feature sets (AAC, EAAC, PAAC, and APAAC) are also provided by the package.

Dubchak, Inna, Ilya Muchnik, Stephen R. Holbrook, and Sung-Hou Kim. "Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence." Proceedings of the National Academy of Sciences of the United States of America 92, no. 19 (1995): 8700-8704. http://www.jstor.org/stable/2368330.

Q2. We have modified the manuscript according to your suggestions. We have also given careful attention to the English editing.

Q3. We have included some of your suggested papers as reference in our manuscript (lines 74-88).

Q4. We modified the text to make it clearer. The data was acquired from a public dataset from another group's previous research so we did not do any preprocessing. The number of positive and negative sites in training and test set is shown in Table 1 of the manuscript.

Q5. We have given the full names of the methods as the reviewer suggested.

Q6. We have prepared more explanation of the 7 methods in the supplementary document.

Q7. Section 2.2.2 is modified slightly to explain N(s,r).

Q8. The formula of MCC has been fixed.

Q9. SBF is defined for the first time in line 124.

Q10. We have conducted further experiment with balanced test data. The result is shown in Table 8 and discussed in lines 301-307.

Q11. We have uploaded the dataset and code at https://github.com/findriani/ProtTrans-Glutar

Q 2 **Check List**

Reviewer 2 | 18 Mar 2022 | 08:24 #1

a. Is the quality of the figures and tables satisfactory?
- Yes

b. Does the reference list cover the relevant literature adequately and in an unbiased manner?
- Yes

c. Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)
- No

d. Is a statistician required to evaluate this study?
- No

e. Are the methods sufficiently documented to allow replication studies?
- No

QUALITY ASSESSMENT

Q 3	Rigor	■■■□□
Q 4	Quality of the writing	■■□□□
Q 5	Overall quality of the content	■■■□□
Q 6	Interest to a general audience	■■■□□

EDITOR

History | **Editor** Active | Reviewer 1 Finalized | Reviewer 2 Finalized | ▪A▪I▪R▪A▪

Handling Editor: Ruiquan Ge
Received date: 28 Feb 2022
Editorial assignment start date: 28 Feb 2022
Independent review start date: 08 Mar 2022
Interactive review activated date: 18 Mar 2022
Review finalized date: 12 Apr 2022
Final validation date: 26 Apr 2022

Revision request

Guest Associate Editor: Ruiquan Ge | 13 Apr 2022 | 02:14 #1

Please give the point-by-point response to Reviewer 1 in this section (Editor tab) and update the manuscript.

Guest Associate Editor: Ruiquan Ge | 13 Apr 2022 | 09:00 #2

Dear authors,

However, the writing of this manuscript is of low-quality. The reviewer suggest the authors send it to any editing service. In the first round of revision, you have paid attention to it and have some editings but that is not enough. Please update the manuscript to make it easier for readers to understand.

Guest Associate Editor: Ruiquan Ge | 18 Apr 2022 | 15:03 #3

Please give the point-by-point response to Reviewer 1 in this section (Editor tab) and update the manuscript.

For example, the author may need make the following revisions:

1. Whether did authors use feature selection technique to optimize features? Whether is there information redundance or noise in feature set?
2. Authors should make comparison with published models.
3. Authors should provide a webserver or softpackage for users.

Corresponding Author: Fatma Indriani | 20 Apr 2022 | 07:33 #4

Dear Editor,

We apologize for the delay. I thought I have posted this forum reply two days ago when I resubmitted the manuscript, but it turned out it just went into "draft" (unposted).

The manuscript has been edited by a professional English editing service. We have uploaded the updated manuscript. Thank you for your understanding.

Our response to Reviewer 1 is posted below.

Q1. We believe there is noise in the original feature sets, but after feature selection the noise should be reduced. Originally we extracted 7 traditional feature sets (AAC, EAAC, CTDC, CTDT, CTDD, PAAC, APAAC) and 6 features set from protein embedding (ProtBERT, and its variations). We selected features by evaluating combinations of features sets (Figure 1, manuscript line 97-101). The best combination is the one we proposed for the model: CTDD, EAAC, and ProtT5-XL-UniRef50. We did not do more fine grained feature selection.

Q2. In Part 4 (Discussion section), we compared the result to previous models, those that used the same dataset as our model (iGlu_AdaBoost, RF-GlutarySite) as well as those that used different dataset (GlutPred, iGlu-Lys, MDDGlutar).

Q3. We have provided code and dataset https://github.com/findriani/ProtTrans-Glutar

We are also in the process of providing script to run the pretrained model.

Guest Associate Editor: Ruiquan Ge | 21 Apr 2022 | 04:41 #5

Dear Dr. Indriani,

In https://github.com/findriani/ProtTrans-Glutar, there is only data set and no code. Please check it.

Best wishes,

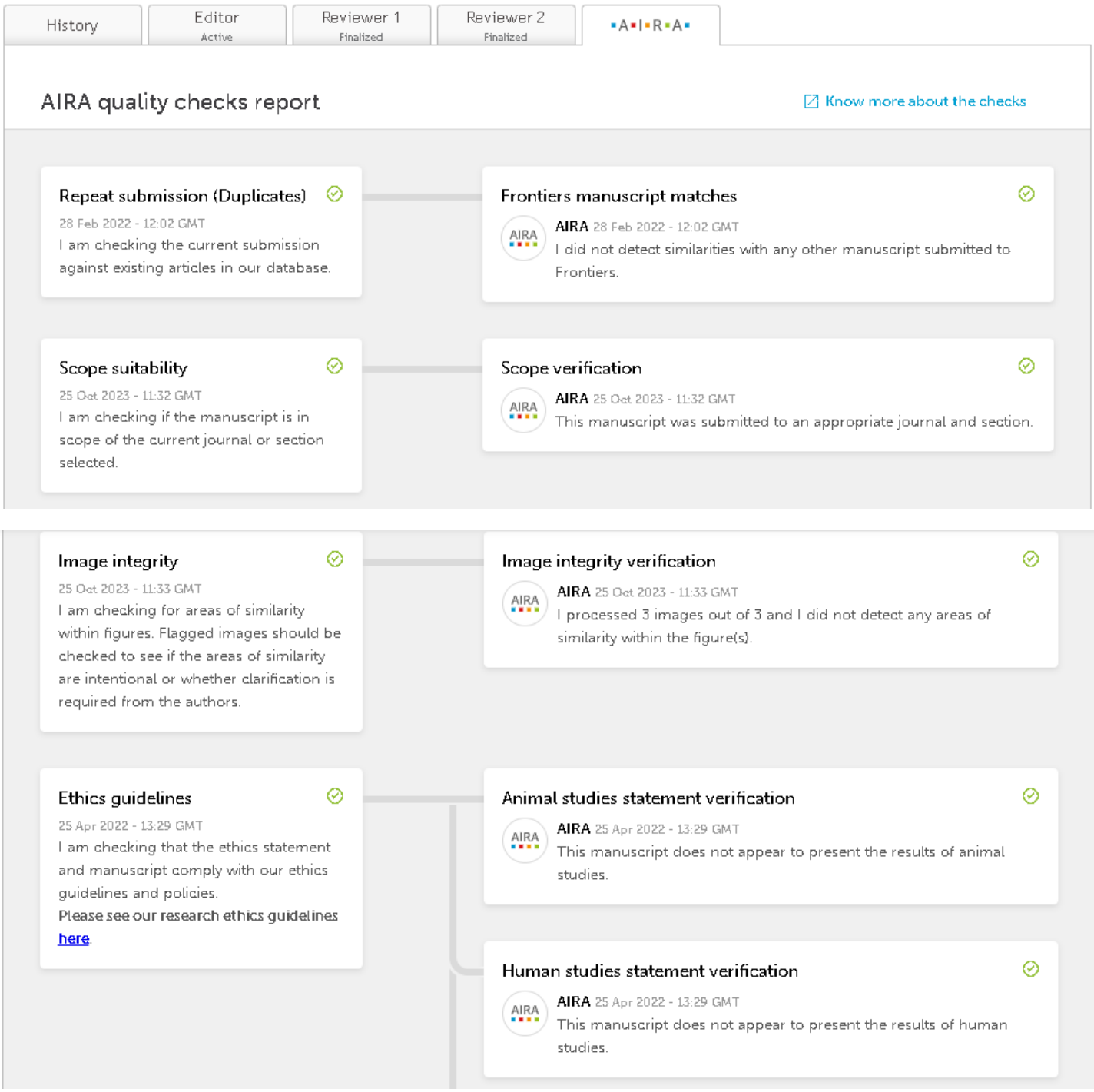
Ruiquan Ge

Corresponding Author: Fatma Indriani | 25 Apr 2022 | 04:13 #6

Dear Editor,

We have fixed the files and the links. Please check again. Thank you

QUALITY CHECKS



Identifiable images and information

AIRA 25 Apr 2022 - 13:29 GMT
This manuscript does not appear to present potentially identifiable human images or data.

Human images

28 Feb 2022 - 12:03 GMT
I am checking for human images in the figure files and supplementary files, as these require consent statements.

Face and body detection

AIRA 28 Feb 2022 - 12:03 GMT
I did not detect human images in the figure files.

Data availability

16 Mar 2022 - 10:17 GMT
I am checking that the manuscript complies with our data availability guidelines.

Data availability statement verification

Editorial Office 16 Mar 2022 - 10:17 GMT
Data availability is OK.

Commercial conflicts

25 Oct 2023 - 11:32 GMT
I am checking the submission for potential commercial conflicts.

Commercial keyword detection

AIRA 25 Oct 2023 - 11:32 GMT
I did not detect any potential commercial conflicts of interest.

Controversial topics

28 Feb 2022 - 12:02 GMT
I am checking the manuscript for controversial topics.

Controversial themes

AIRA 28 Feb 2022 - 12:02 GMT
I didn't detect any controversial theme in this manuscript.

Controversial keywords (global and journal specific)

AIRA 28 Feb 2022 - 12:02 GMT
I didn't detect any controversial elements in this manuscript.

Language quality

16 Mar 2022 - 10:17 GMT
I am checking the language quality of the manuscript and assigning it a recommended copy-editing level score.

Language evaluation

Editorial Office 16 Mar 2022 - 10:17 GMT
Manuscript L2