Comments legend:

- **Amparo:**
- **Jorge:**
- **Alex:**
- **Bertha:**

Associate Editor Comments to the Author

Dear Mr. Jorge Paz Ruza,

Your manuscript entitled "Positive unlabelled learning for identifying new candidate diet restriction-related genes among ageing-related genes" has been reviewed and needs to be revised.

If you care to revise the manuscript in accord with the comments of the reviewer(s) found at the end of this message, we will then reconsider it for publication.

pleasantly,
Editor

Reply: Thank you very much for your comments. We have revised the manuscript aiming at tackling clearly all issues raised by the reviewers: we improved the discussion of the limitations of the existing approach, added missing details of the experimental setup, and corrected two minor inaccuracies in the algorithm pseudocodes. Moreover, we have extended our experiments with the requested measurements of model efficiency, to show our PU Learning approach to DR-related gene identification is not only more effective but also more sustainable. Finally, we added two Appendixes with auxiliary information on model hyperparameters and a figure with finer detail of the F1 Scores.

Response to the reviewers

We thank the reviewers for their critical assessment of our work. We are submitting a revised version of our manuscript, with all major changes formatted in blue. In the following, we address their concerns point by point.

Reviewer 1 (#1)

Dear Authors,
The manuscript has been written well, I enjoyed reading it to learn more about applying ML to identify DR aging related genes.

Reviewer Point P 1.1 — *It would be good to add the representation graphs for your statistical analysis in the result section. Also, the accuracy scores can be represented by curve graphs to show how accurate the learning of your ML classifier you used among all other classifiers.*

Reply:

We agree that a teardown of the most important statistical results can add more detail to the manuscript. We have now added Appendix B with a box plot figure detailing the results for the F1 metric, which is the most relevant and reliable in PU Learning, as well as the statistical confidence against the best non-PU method in the metric.

Reviewer 2 (#2)

This paper introduces a similarity-based positive-unlabelled learning classifier to predict dietary-restriction genes among ageing-related genes. Overall, the manuscript's structure and methodology are satisfactory. However, the following corrections need to be carefully addressed before possible publication:

Reviewer Point P 2.1 — *[Abstract] Include the major achieved results (in numbers)*

Reply: We agree that including some precise results in the abstract can be useful for readers. Following the reviewer's suggestion, we have improved the abstract by including both the number and name of the top genes identified as DR-related (4 genes: PRKAB1, PRKAB2, IRS2, PRKAG1), the statistical strength that proves the higher reliability of our method ($p < 0.05$), and the improvement in computational cost (up to 40%). We have opted not to include any raw accuracy values in favor of statistical strength because, as discussed in Section 4.2, the use of genuine data distorts the F1 Score of all methods by a common unknown constant factor, so stating a single accuracy value in the abstract would be misleading for readers.

Reviewer Point P 2.2 — *Improve the discussion of existing related approaches and their main challenges, as it surprisingly mentions only one approach ([16]).*

Reply: We agree that the wording of the discussion about existing approaches was misleading. The target task in our paper is the identification of new DR-Related genes. Magdaleno et al.'s recent work [16] is the only existing approach to solve this task at the time of this manuscript's submission. To the best of our knowledge, no later research has pushed to overcome any of the the limitations of their work, which is the objective of our research. We have now applied rewording across the manuscript to clarify this fact, and stated it explicitly at the beginning of Section 2.1.

Reviewer Point P 2.3 — *Briefly highlight the major limitations of ML-based dietary restriction approaches that motivated the authors to propose this classifier.*

Reply: We understand that our mention of the limitations of Magdaleno et al.'s previous research in the Introduction could be more clear. We have now improved the clarity of that part of the Introduction (Section 1, par. 4-5) to help readers comprehend earlier the motivation behind our work.

Reviewer Point P 2.4 — *Merge Section 2.3 with Section 2.2 (Existing Approaches) since both discuss related works*

Reply: The reviewer is right that Sections 2.2 and 2.3 are both related-work-oriented, but we believe they cover fundamentally different aspects of our research's background, thus merging them would convolute the structure of the manuscript.

- Section 2.2 discusses the related work in the specific biological problem we address: the existing methodology for ML-based DR-related gene identification (a non-PU approach originally introduced by Magdaleno et al.) and its core limitation.
- Section 2.3 discusses PU Learning methods independently from any target problem: it is an introduction to PU Learning's core ideas and rationale, and mentions relevant use-case examples in the bioinformatics literature.

We understand that the titles of Sections 2.2 and 2.3 could be clearer and more precise to reflect the above distinction between these two sections. We have now renamed these sections to "*2.2. - Existing Methodology for DR-Related Gene Identification*" and "*2.3. - Essential Notions on Positive-Unlabelled (PU) Learning*" and improved their descriptions at the end of the Introduction (Section 1 par. 11) and the beginning of the Background (Section 2, par.1).

Reviewer Point P 2.5 — *Mention at the end of this section how your proposed method is distinguished from existing approaches*

Reply: We are confident that, in the original manuscript, this is stated with clarity towards the end of Section 1 (par. 7) and near the start of Section 3 (par. 1), but we agree with the reviewer that it can be worth refreshing readers about this in Section 2.2. We have added some closing remarks at the end of Section 2.2 to make this more clear for readers.

Reviewer Point P 2.6 — *There is no justification for using a threshold of 0.8 (even if it has been used in existing literature) and how the model's performance will be affected if this threshold is set as a learnable parameter.*

Reply: We found that the manuscript was somewhat misleading in this regard, as the selection threshold t is indeed a tunable hyperparameter not fixed to $t = 0.8$. In our original manuscript, this was explicitely said in the text of Section 3 (second-to-last par.), but not elsewhere. We have improved the caption of Fig.3 to properly reflect that $t = 0.8$ is only an example for the Figure, and corrected a typo in Algorithm 2 (l.9) that did not show t as a tunable parameter alongside k.

Reviewer Point P 2.7 — *Replace "memoize" with "memorize" in Algorithm 1 (line 10)*

Reply: We understand that "memoize", while correct (not to be confused with "memorize"), can be an unnecessarily technical term. From Wikipedia:

> *In computing, memoization or memoisation is an optimization technique used primarily to speed up computer programs by storing the results of expensive function calls to pure functions and returning the cached result when the same inputs occur again.*

We agree that a more widespread term can be more adequate; to preserve the intended meaning, we have changed "memoized" to "cached" in Algorithm 1.

Reviewer Point P 2.8 — *Define RN as an input in Algorithm 1*

Reply: We think that the reviewer means "output" rather than "input", as RN (the selected set of reliable negative examples for training) is the object returned by Algorithm 1. We agree with the reviewer and have modified the preamble of Alg. 1 to declare RN as the function output.

Reviewer Point P 2.9 — *Justify the use of two unbalanced datasets, 60% of aging-related genes in PathDIP against 13% of aging-related genes in GO!*

Reply: Based on the reviewer's comment, we noted that Table 1's column headers may have been misleading. To clarify, all genes are ageing-related (i.e. they belong to $\mathcal{G}_{\mathsf{AGE}}$), as we're looking to identify new DR-related genes among these ageing-related genes. The set of known DR-related genes (i.e. $\mathcal{G}_{\mathsf{DR}\cap\mathsf{AGE}} \subset \mathcal{G}_{\mathsf{AGE}}$) act as the known positive examples, and appear in similar proportion on both feature sets (PathDIP: $110/1096 \simeq 10\%$, GO: $114/1238 \simeq 9\%$). We have now improved the clarity of Table 1's column headers and the table's title using the formal notation from Section 2.1.

Reviewer Point P 2.10 — *Clarify the nature of dataset features, for example, show some representative examples of positive and negative AGE-related genes.*

Reply: We agree that the structure of the data (examples and features) is not trivial to understand despite the lengthy explanation. We have now extended our explanation of the feature and gene extraction in Section 4.1 (par. 3-5) by giving specific examples of a DR-related (positive) gene and a gene with absence of DR-relation (unlabelled), and specific examples of a PathDIP and a GO feature, alongside their meaning in the data.

Reviewer Point P 2.11 — *Summarize the initial values of all experimental hyperparameters in Section 4.3 (Implementation Details), including k and t introduced in Section 3, even if they were fine-tuned.*

Reply:

We agree that summarizing the hyperparameters of the classifiers is useful for the reader. We have added Table 5 describing the particular hyperparameter configurations of BRF and CAT, and the PU Learning hyperparameter search space. For the sake of flow, we have put this table in a new Appendix A, but it could be moved to the main text upon the reviewer's request.

Reviewer Point P 2.12 — *Since the PU learning achieved relatively similar results as reported in [16] in terms of 3 metrics, you need to explain the efficiency (training time, complexity, memory cost) of the proposed method to demonstrate - its significance; otherwise, there is no point in additional steps for genes ranking or prioritization.*

Reply: To clarify, the predictive performance of our proposed PU learning method was statistically significantly better ($p < 0.05$) than the predictive performance of the state-of-the-art method in [16], as shown in Section 5.1. However, we fully agree with the reviewer that it is crucial to do so in a computationally sustainable way. We had not included our experiments about this matter to maintain the compactness of the manuscript, but we have followed the reviewer's suggestion and added Figure 4 containing the computational cost results (in terms of CO_2 emissions, more precisely gCO_2e), where we show that our PU Learning method requires less computational effort to identify DR-related genes in the best-performing scenarios. This experiment has been discussed in Section 5.1 (par. 5) and referenced throughout the manuscript.

Alex: OK, the parallelisation affecting the training time is a good point, Let's report the CO_2 emissions then. I read now the paragraph about this in Section 4.2. It is a short paragraph, I wonder if it could be expanded a bit, perhaps saying a bit more about codecarbon. You cite a reference to it, which goes straight to the code itself, a few words about it here, in the main text, might help a bit.

E.g., it seems it was developed by the machine learning community, right? (The GitHub for the code includes the folder "mlco2", I assume ml stands for ml here ?). Or perhaps cite also a paper discussing the codecarbon software, its motivation, its principles etc. ALso, in Figure 4 with the CO_2 results, I understand the vertical axis is measuring green house gas emissions, but what is the exact meaning of the numerical values in that axix? For example, the first bar in the figure, for PathDIP+CAT and Original method, has a value in the vertical axis of about 0.12, what does that mean? Presumably 12%, but 12% of what? Also, you mentioned that BRF was parallelized in your experiments, was CAT parallelized in your experiments too? If so we have a good reason to explain to the reviewer why we are not reporting training time. Anyway, I agree in reporting CO_2 emissions rather than training time.

Jorge: Figure 4 has a vertical axis label (gCO2e) which is the specific measurement codecarbon does. Because maybe it wasn't clear, I have also now added it in the Figure caption, and explained with another phrase what codecarbon exactly measures in Section 4.2, i.e. what gCO2e is. Regarding the cites, I'm currently citing what the codecarbon team explicitly asks to cite (which is the codecarbon repository with a specific citation text), as said in their repository. Regarding CAT, it cannot be paralellised like BRF because it is a boosting ensmeble method.

Reviewer Point P 2.13 — *Discuss the PU results (scores) shown in Tables 3 and 4 in the context of k and threshold values.*

Reply: We understand, as said in Reviewer Points P2.6 and P2.11, that our explanation of how k and t are selected to train the model and compute the gene DR-relatedness probabilities was not clear and may have caused confusion in Tables 3 and 4. To further clarify this, we have improved the headers of Tables 3 and 4 with information about the setup for computing feature importance and gene DR Probabilities.

To clarify, recall we are performing a nested 10x5 Cross-Validation, where the inner 5-fold CV optimizes k and t for the current training set, and uses that k and t to predict on the test set. For all experiments, the complete 10x5 CV is also repeated 10 times for result robustness and to test statistical significance of results; therefore, it is not useful to discuss the results of Tables 2, 3 and 4 in terms of t and k values. We have now added the following information in the headers of Table 3 and Table 4:

- The feature importance values in Table 3 are averages of 100 training-test partition pairs (10 executions of the 10x5 CV). As the inner CV is in charge of optimizing k and t for each of them, there is not a single or unique value for t and k associated with the results in Table 3.
- Regarding Table 4, the reported DR-probability of each gene is an average of 10 predictions; for each of the 10 executions of the 10x5 CV, the predicted DR-probability of the gene is simply the predicted value for that gene in the outer test fold where the gene was. Again, as in each outer fold the k and t values are optimized using the inner CV, there is not a single or unique value k and t for the results in Table 4.

Reviewer Point P 2.14 — *No analysis provided for the experimental results in the context of previous works apart from the single referenced work [16]! Articulate the results and main findings regarding the most related works rather than comparing the performance with the original method presented in [16].*

Reply: With respect to Reviewer Points P2.14 and P2.15, we tie our response to the answers to Reviewer Points P2.2 and P2.4: we have improved our wording across the manuscript and made clarifications to

help readers to understand that Magdaleno et al. [16] is the only previous work which has proposed a solution for novel DR-related gene identification (the same problem addressed in our work). Hence, there are no other suitable related works for comparison against our work. In the revised paper, we have improved the structure of Section 2 to make more clear that Section 2.3 does not contain other related works in the specific biological problem we tackle, but rather an introduction for readers to the PU Learning ML paradigm and an illustration of use-case examples in other unrelated bioinformatics tasks which cannot be compared with ours.

Reviewer Point P 2.15 — *To compare with the literature rather than with [16] only, consider implementing other classifiers with the PU-learning approach to see if better results are achievable. It's worth trying different classifiers, such as ANN, SVM, or logistic regression.*

Reply: We agree that we the explanation of the choice of classifiers (CAT, BRF) could be more clear. There are two fundamental reasons for maintaining the use of the tree ensembles explored by Magdaleno et al.: 1) to ensure fairness of comparison with their work (as we want to prove that PU Learning in the task surpasses the existing non-PU approach, not just test new classifiers), and 2) because the State of the Art in tabular data is tree ensembles, surpassing options like SVM, Logistic Regression or Neural Networks [29, 52]. We have now properly explained this in Section 4.1 (par. 6).

Reviewer Point P 2.16 — *The (conclusions) text is long with redundant details! Instead, focus on the main findings of this study in the first part and keep the future suggestions concise.*

Reply: We gladly agree with the reviewer on making the Conclusions more clear and concise. We have improved Section 6 (pp. 23-24) by leaving finer experimental details to the inner sections of the manuscript and getting rid of redundant arguments.

Reviewer Point P 2.17 — *I'd suggest citing a few more recent studies (2023 and 2024)*

Reply: We agree that having up-to-date discussion of the State of the Art is important, and we believe our manuscript currently complies with this; our manuscript discusses recent research (2023-24) in the use of ML for other ageing genomics problems [14], the use of PU Learning in other bioinformatics tasks [38], and most recent supporting evidence of the possible DR-related nature of our promising identified genes [60, 67]. If the reviewer is aware of any other recent works that would be insightful for our work, we will gladly discuss them in the manuscript.

Reviewer Point P 2.18 — *Use the acronyms after defining them the first time they appear in the text, and avoid the use of uppercase letters afterward. e.g., PU (page 3), DR (page 4), Machine Learning, Dietary-Restricted, Positive-Unlabelled Learning.*

Reply: We agree that this is a good stylistic improvement and have revised the text to apply it.

Reviewer Point P 2.19 — *Use a consistent language style throughout the whole manuscript, e.g., prioritization (in Abstract) and prioritisation (in Introduction).*

Reply: We have now applied the British "prioritisation" spelling throughout the text.

Reviewer Point P 2.20 — *Replace the opening parenthesis with a comma on page 14 (i.e., whether it belongs...")*

Reply: We have corrected this typo by closing the parenthesis, for writing style consistency.

Reviewer 3 (#3)

The manuscript entitled "Positive-Unlabelled Learning for Identifying New Candidate Dietary Restriction-related Genes among Ageing-related Genes" has been investigated in detail. The paper presents a novel gene prioritization method for identifying dietary restriction (DR)-related genes using a two-step Positive-Unlabelled (PU) Learning paradigm. While the approach addresses the critical issue of unreliable negative examples in machine learning (ML) models for gene identification, the paper lacks clarity, methodological detail, and rigorous evaluation in several areas. Significant revisions are needed to enhance the paper's comprehensibility and impact.

The proposed gene prioritization method for identifying DR-related genes using a two-step PU Learning paradigm is promising but requires substantial revisions to enhance clarity, detail, and rigor. Addressing the detailed critiques will significantly improve the paper's quality and provide a more comprehensive understanding of the proposed method's effectiveness and potential applications.

Reviewer Point P 3.1 — *The introduction briefly outlines the importance of DR and ML in gene identification. However, it lacks depth in explaining why the current methods are inadequate and the specific gaps the proposed method aims to fill.*

Reply: We agree that this could be more clear in the Introduction. As said in Reviewer Point P2.3, we have applied wording changes and expanded that discussion to properly highlight the limitations of Magdaleno et al.'s work and our contributions in overcoming them.

Reviewer Point P 3.2 — *The review of previous works is superficial. Provide a more comprehensive discussion of existing ML methods for gene identification, their limitations, and how your method overcomes these issues.*

Reply:

We understand that the structure of Section 2 was misleading: it should be more clear that the only existing approach to our specific problem (ML-based identification of DR-related genes) is the one proposed by Magdaleno et al [16], and that Section 2.3 does not cover other methods for DR-related gene identification, but instead introduces the essentials of PU Learning citing use-case examples in other biological problems. As mentioned in our answers to Reviewer Points P2.2, P2.3, P2.4 and P2.5, we have made clearer the narrative structure and subsection naming of Section 2 to improve the manuscript in this regard.

Reviewer Point P 3.3 — *The description of the two-step PU Learning paradigm is vague. Provide a detailed explanation of the PU Learning concept, its relevance to the problem, and the specific steps involved in your method.*

Reply: We understand that Section 2.2 didn't explicitly justify the relevance of PU Learning to solve the limitations of Magdaleno et al.'s state-of-the-art approach. As said in Reviewer Point P2.3, we now conclude Section 2.2 discussing why PU Learning is relevant and adequate for this problem. We have also renamed Section 2.3 to properly reflect that this Subsection covers in detail the basic concepts of PU Learning and the two-step paradigm.

Reviewer Point P 3.4 — *Explain how the similarity-based, KNN-inspired approach selects reliable negative examples. Describe the similarity metrics used, the rationale behind their selection, and how the KNN method is adapted for this purpose.*

Reply: We understand that the justification for using the Jaccard Measure in the KNN method in Step 1 of the PU Learning algorithm was too brief. We have expanded that justification (Section 3, par.4) by discussing how, due to the sparsity of the datasets and the fact that all predictive features are binary (indicating the presence or absence of a biological property for a gene), the Jaccard measure is more adequate than the Euclidean or Cosine distances, which are not cost-effective in binary spaces and suffer more from the curse of dimensionality. Regarding the use of the nearest-neighbours paradigm to design the reliable negative selection, we are confident the current manuscript has sufficient detail, containing a step-by-step explanation (Section 3, par. 4-6), a more formal pseudocode (Alg. 1) and an example for illustrative purposes (Fig. 3).

Reviewer Point P 3.5 — *Detail the classifier used for differentiating DR-related and non-DR-related genes. Specify the algorithm, features, training process, and hyperparameter tuning. Explain why this classifier is suitable for the task.*

Reply: We agree some aspects of the classifier discussion, including justification and configuration details, was not detailed in a self-contained manner in this manuscript. Tying with our answer to Reviewer Points P2.11 and P2.16, we have added a new Appendix A that fully details the selected hyperparameters (tunable and non-tunable) for both classifiers and the PU Learning algorithm, and we have properly justified the classifiers choice for the second step of the PU Learning method.

Reviewer Point P 3.6 — *Clarify the method used to generate the ranking of promising genes for novel DR-relatedness. Provide the mathematical formulation and rationale behind this ranking process.*

Reply: We understand that Section 4.2 didn't explicitly formalize how the promising DR-related gene rankings are to be constructed. We have improved Section 4.2 (par. 5) by properly tying our explanation with the formalization of the task given in Section 2.1, which does include the mathematical formulation for the selection of promising genes.

Reviewer Point P 3.7 — *Specify the datasets used for training and testing the model, including the source, number of genes, and any preprocessing steps. Explain the criteria for selecting known DR-related and non-DR-related genes.*

Reply: We understand that some aspects of the dataset information in Table 1 were unclear. Particularly, tying to Reviewer Point P2.9, we have improved the column headers and caption of Table 1 to properly formalize that all genes are ageing-related, and the only labels are the positive labels for known DR-related genes, using the formal notation from Section 2.1. With respect to the source, preprocessing and labelling of the known DR-related genes, we are confident Section 4.1 (par. 2-5) contains all required information to understand and/or reproduce the dataset creation and/or usage.

Reviewer Point P 3.8 — *Provide detailed comparisons with existing state-of-the-art non-PU approaches. Include quantitative metrics such as accuracy, precision, recall, and F1-score to demonstrate the superiority of your method.*

Reply: With respect to the metrics, we agree that the rationale for not using accuracy, recall and precision was not explicit in the manuscript. We have now stated in Section 4.2 (par. 3) that these three metrics are avoided because, in isolation, they are highly inadequate for unbalanced datasets like those used in this work; this is why we favour the use of the F1 Measure, which combines Recall and Precision and also happens to be the most adequate for PU Learning tasks. With respect to existing state-of-the-art approaches, as explained in Reviewer Points P2.2, P2.14 and P3.2, Magdaleno et al.'s approach [16] is the only other proposed solution for the ML-based identification of DR-related genes, so our evaluation comprehensively compares with three relevant metrics our PU-based method with [16] in the four {Features, Classifiers} scenarios that showed highest performance in [16]. As explained in Section 5.1, our PU approach significantly ($p < 0.05$) outperforms the existing State of the Art.

Reviewer Point P 3.9 — *Explain the three relevant performance metrics used to evaluate your model. Justify why these metrics are appropriate for assessing the effectiveness of DR-related gene prediction.*

Reply: We agree that the information about evaluation metrics (Section 4.2) can be more self-contained. Following the reviewer's suggestion, we now added a simple mathematical definition of F1, G.Mean and AUC-ROC (Eq. 9) and clarified better in the text why they are suitable for PU Learning tasks (Section 4.2, par. 2-4).

Reviewer Point P 3.10 — *Describe the process of curating existing literature to find support for the top-ranked candidate genes. Explain how this supports the validity of your model's predictions.*

Reply:

We understand that our definition of the literature curation process may have been ambiguous. We have improved the final portion of Section 4.2 with some finer details of the literature curation process: for each gene, we used two popular research literature indexes (*PubMed* and *Google Scholar*) to search for studies that linked the gene or its encoded protein to any biological mechanisms potentially related to DR.

Reviewer Point P 3.11 — *The authors should clearly emphasize the contribution of the study. Please note that the up-to-date of references will contribute to the up-to-date of your manuscript. The studies named- "Overcoming nonlinear dynamics in diabetic retinopathy classification: a robust AI-based model with chaotic swarm intelligence optimization and recurrent long short-term memory; Classification of diabetic retinopathy by machine learning algorithm using entropy-based features; Agricultural crop classification with R-CNN and machine learning methods"- can be used to explain the methodology and machine learning technique in the study or to indicate the contribution in the "Introduction" section.*

Reply:

We understand that our discussion on the main contributions of our study was previously convoluted in some parts of the manuscript. Tying with Reviewer Points P2.3, P2.5, P2.17, P3.1, P3.2 and P3.3, we have applied several improvements throughout the text (mainly in Sections 1, 2.2 and 6) to clearly state how we overcome the limitations of the existing approach for ML-based identification of DR-related genes through the use of the PU Learning paradigm. With respect to the articles suggested by the reviewer, we have examined them in detail but have opted not to discuss them in the text as none of

them addresses PU Learning, the identification of DR-related genes, or other topics directly relevant for our work.

Reviewer Point P 3.12 — *Provide a detailed analysis of the experimental results. Discuss why your method outperforms the existing approaches and any observed limitations or failure cases.*

Reply: With respect to the justification of why our method outperforms the existing approaches, this was previously discussed explicitly in Section (pp. 3-4) and in Section 3 (par. 1) but, as said in our answer to Reviewer Point P2.6, we have also added extended Section 2.2 further justifying the need for PU Learning due to the existing approach's limitation. With respect to potential limitations or failure cases, we are confident they are correctly discussed in our mention of future work (Section 6, par. 4) and the discussion of computational experiment results (Section 5.1, par. 3), where we provided a potential explanation for the lower performance of our approach in the single {PathDIP, BRF} scenario.

Reviewer Point P 3.13 — *Discuss the biological relevance of the top-ranked genes identified by your model. Explain their potential roles in DR mechanisms and any supporting evidence from the literature.*

Reply: We think the reviewer refers to literature evidence that supports the potential DR-related role of the four new genes we propose as DR-related. We are confident Section 5.3 adequately presents this evidence for each of the four proposed genes (PRKAB1, PRKAB2, PRKAG1 and IRS1), dicussing the findings of each of the referenced studies and how each of them can influence or be influenced by the DR mechanisms.

Reviewer Point P 3.14 — *The conclusion should succinctly summarize the key contributions of the paper. Currently, it lacks a clear wrap-up of the main findings and their implications.*

Reply: We agree that the Conclusions section was previously too redundant with the discussions found in the inner sections of the article. Tying with Reviewer Point 2.17, we have improved this Section 6 by more succinctly stating the main contributions and findings of our work.

Reviewer 4 (#6)

The manuscript titled "Positive-Unlabelled Learning for Identifying New Candidate Dietary Restriction-related Genes among Ageing-related Genes" presents a novel and effective approach for identifying DR-related genes. The methodology is well-structured, and the results are significant, showing a clear improvement over existing methods. However, some areas need further clarification and improvement:

Reviewer Point P 4.1 — *While the methodology is innovative, providing more details about the implementation, especially the selection criteria for reliable negatives, would enhance the reproducibility of the study.*

Reply: We agree that some of the implementation details could be more clear, especially the tuning of k ant t for the reliable negative selection and the hyperparameters used for PU and the classifiers. Tying with our answers to Reviewer Points P2.6, P2.11 and P3.4, we have improved the clarity of Figure

3's caption and fixed a typo in Algorithm 2 to properly convey that t is a tunable hyperparameter, and added Appendix A to detail the used hyperparameters for the classifiers CAT and BRF, as well as the search of space of the PU Learning hyperparameters.

Reviewer Point P 4.2 — *The methodology is robust. However, more details on the feature extraction process and the computational complexity of the method would be useful.*

Reply:

We agree that analyzing the computational cost of the method can be interesting. Tying with Reviewer Point P2.12, we have extended our experiments with measurements of the computational complexity (in CO_2 emissions) of the model (pp. 18-19 and Figure 4). Regarding the details of the feature extraction, as said for Reviewer Point P2.10, we have now given a more detailed explanation with examples of the PathDIP and GO gene features used in the work (p.14).

Reviewer Point P 4.3 — *The results are well-presented. Including more visual aids, such as summary tables or graphs, could enhance readability.*

Reply: We think the reviewer refers to a more visual approach to the results, especially those in Table 2, which are purely numerical. We understand a more visual teardown of the more important results can be useful, as said in Reviewer Point P1.1, and have have added Appendix B with Figure 5 visualizing in detail the results for F1 (the most important and reliable metric) and the statistical test of F1 against the best non-PU method.

Reviewer Point P 4.4 — *A discussion on the limitations of the proposed method and potential areas for future research would add depth to the study.*

Reply: We understand that our discussion of the possible future work on the task at the end of the Conclusions section may have been convoluted. Tying with our response to Reviewer Points 2.17 and 3.14, we have made the Conclusions section more compact and more succinctly enumerated the possible avenues for further research on the topic and the end of the section.