Dear Editor,

Your comments and those of the three reviewers were taken into account in the revised version of the manuscript, and have contributed to its improvement: we thank you for that. Most suggestions were followed, and where this is not the case, we have provided justification. The gist of our editions targeted (1) a reorganization of the sections, so that each theoretical point is now illustrated directly by the case study, (2) additional work on an figure to illustrate practical examples as well as theoretical ones, and (3) editions on the data curation process, to document choices related to naming and data structure that were not detailed previously. We added a track-change file related to the manuscript.

In the following response to reviewers, all the changes made to the manuscript are also reported. For clarity, the editor's and reviewers' comments are in black, whereas our responses are in blue.

Best regards,

Answer to Eric Tannier, Editor

When revising your article, we remind you that your article must contain the following sections (see our Guide for Authors in the Help section of the PCI Math Comp Biol website).

Minor revision

The first draft of the manuscript was reviewed by three experts. All three praised the originality and usefulness of the approach, and are in favor of the recommendation by PCI mathematical and computational biology.

Two reviewers have produced significant work to suggest improvements to the text, and I think their reports deserve to be processed in depth by the submission authors in order to consider a revision.

Most of the suggestions require only minor rewrites, in order to facilitate the reader's understanding, the completeness of the bibliography, and the clarity and completeness of the results. Some suggestions may require major changes or additional calculations. These should be considered by the authors as optional suggestions, even if some of them could be useful. For example, I also found that some of the theoretical aspects presented in the first part only made sense when their application to an example was read later. But again it will be up to the authors to appreciate the ratio of improvement over amount of needed work. We have reorganized the manuscript so that each theoretical point is now followed and illustrated by the case study.

I am also sensitive to the question raised by a reviewer about the possible ethical reasons (confidentiality, environmental costs..) for sometimes departing from the FAIR principles. This may not be relevant in this case, it will be up to the authors to judge, but this could be useful to the global dataset/meta-analysis discussion.

PCI Mathematical and Computational Biology would be pleased to receive a corrected version, together with replies to the reviewers. And we apologize to the authors for the lengthy editing

process. The timeframe was nevertheless shorter than in standard journals, and the reviews were constructive and enriching.

Answer to reviewer #1

The manuscript by Mahmoud et al. describes a workflow for processing global datasets. The authors have collected datasets from different laboratories to study the effect of different variables related to intercropping (e.g. species composition, environmental conditions and cultural practices). They describe how to reconstruct the most complete experimental design possible from partial datasets combined into a single one. Four papers have been published based on subsets of this global dataset, focusing on different scientific questions. Overall, the manuscript is well written and I have no specific requests. I just wonder how close it is to the scope of computational biology. One improvement might be to include the actual figure from the four papers mentioned in the text in addition to Figure 3.

Thanks for your comments. We have replaced this figure to illustrate both theoretical and practical use cases, focusing on the actual global dataset presented in the paper.

Answer to Christine Dillman, reviewer

The paper proposes the concept of global datasets, as opposition to meta-analyses. Global data sets are curated aggregation of experimental data sets, far richer than summary datas that can be extracted from the litterature. The authors provide guidelines and methods to create and exploit those global datasets to answer scientific questions in agricultural research. Indeed, field experiments are costly but provide with valuable data that are often underexploited and used to answer a specific question. Aggregating the raw observations from numerous experiments into global dataset allows to study diverse phenotypic observations from varying soils and climates and may enable reliable generalizations of local findings. With the generalization of public data repositories that can handle data from field experiments, there is a real need for methodological developments like the ones that are proposed here.
Thanks for these nice comments!

The paper is organized into three main parts.

- DESIGNING GLOBAL DATASETS
  In the first part, a global workflow is presented for gathering, tidying and distributing datasets. A tidy dataset is a dataset where every column is a variable, every row an observation, and every single cell is a single value. The authors nicely review the general recommendations to end-up with FAIR open data. To me, this part lacks a section about the progresses that have been achieved during the last decade on ontologies and, in particular, plant phenotyping ontologies (see e.g. Krajewski et al, 2015, doi:10.1093/jxb/erv271 or the https://www.miappe.org).

  The paper you suggested was particularly interesting (we thank you for that). In the Introduction, we have now emphasized the recent efforts and international initiatives dealing with opening and homogenizing data, finally underlying that phenotyping data are still underrepresented in published datasets (as highlighted by the authors Krajewski et al.) in the manuscript: "*Recent efforts and international initiatives dealing with opening and standardizing data are emerging, underlying that data standardization is crucial to improve experimental results interpretation and the generalization of knowledge ascent, as well as to facilitate statistical (meta)analysis and data publication (Krajewski et al., 2015). However, phenotyping data are still underrepresented in published datasets.*"

  We also added a sentence in the Discussion ("*This step itself is very likely to necessitate more time than meta-analysis data processing step, and would greatly benefit from improved upstream data standardization practices [@Krajewski2015]*").

  However, we choose not to edit our dataset contribution by renaming variables as function of an available ontology (such as Plant or Crop ontology) for two reasons. Firstly, because we wanted to account for a range of species, and the actual ontologies are rather species dependent, and this would involve more development to fit our use in this frame. We were also quite convinced by an opinion paper from Wilson *et al.* (2017) (https://doi.org/10.1371/journal.pcbi.1005510), stating that aiming for "good enough" practices might be more efficient than aiming for "best" practices to speed up the acceptance of reproducible science workflows. They argue that adopting a consistent naming scheme, along with associated metadata, is a first but essential step to enable other research groups to appropriate and reuse datasets. But we are also aware that, if this dataset grows in the future, we should really consider consolidating it by adopting an ontology system.

- CASE STUDY
  The second part relates a case-study and describes the creation of a global dataset gathering 37 field experiments involving cereal-legume intercrops and their corresponding sole crops. The global

dataset is publicly available on Zenodo. The creation of this global dataset is a remarquable result that is insufficiently described, even in the data_report.pdf file on Zenodo.

First of all, we truly thank you for your suggestions and comments: that motivated us to quickly create a version 2 of the dataset, with a more detailed "data_report" (particularly including the links between tables).

Before answering precisely to your comments/questions, it is important to note that the majority of the work to curate the data was done manually, considering the state of the retrieved files: some were relatively clean, while others caused our computers to crash numerous times due to their heaviness, the inclusion of macros (software incompatibilities) and of various tables and graphics within the original spreadsheet files. Ultimately, this longest part is not automatable (unfortunately!). The choice of common scales shared between experiments (such as BBCH) was decided following several discussions, involving a literature search and soliciting the expertise of fellow agronomists.

In particular, on page 8 of the manuscript, the method used to redistribute the variables into four categories : trials, management, traits and climate should be better described.

We have added a figure describing the links between the four tables in the new version of the manuscript.

For example,

- trait BBCH uses the decimal code proposed for cereals in 1974. I guess this was not the code used in all experiments. When another coding system was used, was it translated or noted as NA ?

  In the experiments gathered, the phenological stages were not encoded, but rather (when present) indicated by a sentence in the language of the source country (e.g. "épi 1 cm" or "flowering"): we translated that in BBCH codes, which are much more international.

- Which method was used to end-up with the consensual trait names that figure in the global dataset ?

  The names were chosen to be as explicit as possible and are the result of numerous discussions and updates as we retrieved new files, sometimes including new variables. We settle for names containing: as few abbreviations as possible, a reference to the organizational levels (organs - leaf, shoot - individuals - plants -, population - crop), and a reference to the trait itself (biomass, number, length).

- How many traits were left-aside from the original data sets ?

  We included all traits that were measured in strictly more than 5 experiments, leaving aside 27 traits among which 14 were measured in only one experiment.

Plant variable	No. Experiments
biomass greenleaftendril	5
biomass senescleaf	5
carbon seed	5
carbon veg	5
biomass weeds	4
fruit layer	4
sla greenleaf	4
biomass greenleaf	3
biomass tendril	3
greenleaf number	2
leaf number	2
main shoot length	2

```
|senescleaf number       |                2|
|aphids                  |                1|
|biomass senescleaftendril|               1|
|carbon abs shoot        |                1|
|greenbranching          |                1|
|ndfa                    |                1|
|nitrogen abs fixation   |                1|
|nitrogen abs seed fixation |             1|
|phosphorus abs seed     |                1|
|phosphorus abs shoot    |                1|
|phosphorus abs veg      |                1|
|phosphorus seed         |                1|
|phosphorus shoot        |                1|
|phosphorus veg          |                1|
|senescbranching         |                1|
```

- Which traits are reliably informed in the original datasets (see figure below)?
  For the agricultural practices (corresponding to the figure you provide), we kept everything as this information is crucial to understand what happened and for modeling studies: that's the reason why there is so much missing data. For instance, "code_row" concerns only intercrops in alternate rows, and brings precision regarding the number of cereal's rows *vs* the number of legume's rows. In the same spirit, if there is no fertilization, then all the information related to fertilizer application (dates, doses) is meaningless.
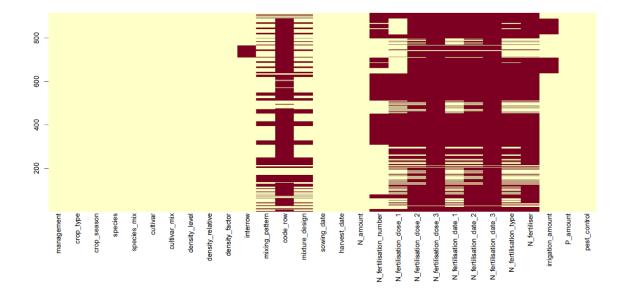

Figure 1: Description of the data concerning the management system. The 37 experiments were expanded into 960 experimental trials. The management.csv data file contains informations about the management system. In red, missing datas.

- What is the "management" trait with 65 levels ? It seems that M1 and M2 are Sole-Cropping only (M1 cereals and M2 legumes ?).
  A "management" (in the *data_management table* describing the agricultural practices) is a code given to each management within a given experiment, from M1 to Mn, with one management M corresponding to a unique combination of practices, including the crop type (sole crop *vs* intercrop), the species and genotype grown, the fertilization amount, etc. There

is a unique experiment where there are 64 different managements (experimenters have tested a lot of things!), *i.e.* Auzeville_TO_2013 (experiment_id, see below) with M1 to M10 corresponding to wheat sole crops (including 3 genotypes, several sowing densities and N-fertilization levels), M11 to M27 for sole crops of different genotypes of pea, M28 to M35 for sole crops of different genotypes of fababean, and the rest for all cereal/legume intercrops tested.

	experiment_id	manageme	crop_type	crop_seaso	species	species_mi	cultivar	cultivar_mi	density_lev	density_rel	density_fac	interrow	mixing	code	mix	sowing_dat	harvest_da	N_amount	N_fertilisat
218	Auzeville_TO_2013	M1	SC	winter	wheat_turgid	wheat_turgid	L1823	L1823	280	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
219	Auzeville_TO_2013	M2	SC	winter	wheat_turgid	wheat_turgid	L1823	L1823	140	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
220	Auzeville_TO_2013	M3	SC	winter	wheat_turgid	wheat_turgid	nefer	nefer	280	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
221	Auzeville_TO_2013	M4	SC	winter	wheat_turgid	wheat_turgid	nefer	nefer	280	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
222	Auzeville_TO_2013	M5	SC	winter	wheat_turgid	wheat_turgid	nefer	nefer	140	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
223	Auzeville_TO_2013	M6	SC	winter	wheat_turgid	wheat_turgid	nefer	nefer	140	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
224	Auzeville_TO_2013	M7	SC	winter	wheat_turgid	wheat_turgid	sculptur	sculptur	280	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
225	Auzeville_TO_2013	M8	SC	winter	wheat_turgid	wheat_turgid	sculptur	sculptur	280	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
226	Auzeville_TO_2013	M9	SC	winter	wheat_turgid	wheat_turgid	sculptur	sculptur	140	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
227	Auzeville_TO_2013	M10	SC	winter	wheat_turgid	wheat_turgid	sculptur	sculptur	140	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
228	Auzeville_TO_2013	M11	SC	winter	pea	pea	aoph10	aoph10	70	1	1	0.162	NA	NA	NA	20/11/2012	11/07/2013	0	0
229	Auzeville_TO_2013	M12	SC	winter	pea	pea	aoph10	aoph10	70	1	1	0.162	NA	NA	NA	20/11/2012	11/07/2013	140	2
230	Auzeville_TO_2013	M13	SC	winter	pea	pea	aoph10	aoph10	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	11/07/2013	0	0
231	Auzeville_TO_2013	M14	SC	winter	pea	pea	geronimo	geronimo	70	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
232	Auzeville_TO_2013	M15	SC	winter	pea	pea	geronimo	geronimo	70	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
233	Auzeville_TO_2013	M16	SC	winter	pea	pea	geronimo	geronimo	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
234	Auzeville_TO_2013	M17	SC	winter	pea	pea	geronimo	geronimo	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	140	2
235	Auzeville_TO_2013	M18	SC	winter	pea	pea	isard	isard	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
236	Auzeville_TO_2013	M19	SC	winter	pea	pea	isard	isard	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	140	2
237	Auzeville_TO_2013	M20	SC	winter	pea	pea	isard	isard	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
238	Auzeville_TO_2013	M21	SC	winter	pea	pea	kazar	kazar	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
239	Auzeville_TO_2013	M22	SC	winter	pea	pea	kazar	kazar	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	140	2
240	Auzeville_TO_2013	M23	SC	winter	pea	pea	kazar	kazar	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
241	Auzeville_TO_2013	M24	SC	winter	pea	pea	lucy	lucy	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
242	Auzeville_TO_2013	M25	SC	winter	pea	pea	lucy	lucy	70	1	1	0.162	NA	NA	NA	20/11/2012	28/06/2013	140	2
243	Auzeville_TO_2013	M26	SC	winter	pea	pea	lucy	lucy	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	28/06/2013	0	0
244	Auzeville_TO_2013	M27	SC	winter	pea	pea	lucy	lucy	35	0.5	0.5	0.162	NA	NA	NA	20/11/2012	28/06/2013	140	2
245	Auzeville_TO_2013	M28	SC	winter	fababean	fababean	castel	castel	24	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
246	Auzeville_TO_2013	M29	SC	winter	fababean	fababean	castel	castel	12	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
247	Auzeville_TO_2013	M30	SC	winter	fababean	fababean	diver	diver	24	1	1	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
248	Auzeville_TO_2013	M31	SC	winter	fababean	fababean	diver	diver	12	0.5	0.5	0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0
249	Auzeville_TO_2013	M32	SC	winter	fababean	fababean	irena	irena	24	1		0.162	NA	NA	NA	20/11/2012	25/07/2013	0	0

data_management

- Figure 2 is not very informative (8 trial sites in Europe with the three french sites overrepresented) and could be advantageously replaced by a figure showing the organization of the global data-set into four csv files and a metadata file.
We replaced this figure by the one you suggested, illustrating the relations between the 4 tables (as inspired by the design of relational databases, in a much simpler case here).

- Similarly, Table 1 would gain in being better commented. What do the colour codes stand for?
We tried to make the figure caption more informative -> "Overview of the diversity of the treatments in the global dataset by factors (columns) and experiments (rows). Within each column, each colored rectangle is a level of the factor considered. For instance, the two colors for the mixing pattern indicate that the two species intercropped were sown in alternate rows or within the row; the two colors for the nitrogen (N) fertilization indicate that the experiment included at least two N-treatments (no fertilization and N-fertilization, the latter of which may include several amounts of N); regarding Species mixture, the number of colors indicates the number of different species mixtures included in a given experiment. A rectangle in a given row and column indicates that the corresponding experiment contains at least one statistical individual with the corresponding factor level."

I guess that spatial arrangement stands for the mixing pattern (within_row or alternate_row). *Yes, we modified the name of the column in the table.*

Could the species mixtures be described? *We added a sentence listing the intercrops represented in the database: "8 resulting intercrops, i.e. i) barley intercropped with faba bean, lupin and pea, ii) durum wheat associated with chickpea, faba bean and pea, and iii) soft wheat associated with lentil and pea".*

Why were there only two Nitrogen fertilization status instead of 3 as in Mahmoud et al, 2022?

In this table, we just wanted to broadly illustrate the heterogeneity of the dataset, and thus for two main reasons: to highlight its richness and to issue warnings in case someone would like to use it to address various research questions (and make the statistical analyses associated). However, this global overview can be detailed when focusing on a given research question and that was the case dealing with N-fertilization:
- in this paper, it is broadly described as no fertilization vs N-fertilization (with N-fertilization including different amounts of nitrogen, depending on the experimentations). We used this broad categorization in the paper we wrote in 2021 on the reproductive allometry (https://doi.org/10.1111/1365-2664.13979),
- in the paper written in 2022, we split the N treatments in three levels, i.e. N0, [30-80 kg N.ha-1], > 80 kg N.ha-1.

● SUPPORTING ANALYSIS
The last part proposes a method from Graph Theory to identify subsets associated with complete factorial designs in the global dataset, allowing for statistical analyses. It is associated with the production of a very nice R package available on github that allows to visualize the structure of the global dataset and to enumerate the maximal kcliques present in the graph, each k-clique representing a factorial design. If I understand correctly, the global dataset graph has a special structure, with each experiment being a k-partite graph, uniquely described by the levels of the k factors taken into account. This could have been stated explicitly in the manuscript.
A didactic application of the method is proposed with a fictive global dataset that helps to understand the concepts. Could the edges in Figure 3 have different width, depending on the number of replicate experiments with the same k-plets? The authors claim that the method was applied to the intercropping global dataset to identify 2-factors factorial designs (field location and nitrogen fertilization) that contain two levels of N-fertilization. But no results were provided. They refer to a published analysis (Mahmoud et al 2022) but the paper cited does not refer to the method used to select the experiments. An additional figure with the selection results would be nice.
We have made an effort to make this part more instructional and we have replaced the Figure 3 by an actual case study. However (and it is now more explained in the new version), this tool is aimed at helping the potential collaborators to explore and get an overview of the dataset as a function of their factor of interest. In no case does this method replace the need for caution on the part of users: it provides an initial view of the structure of a dataset (which is not easily described by words), and by highlighting factorial design subsets.
We removed the reference to our published analysis, as the method we then used was a mixture of direct filtering on factors of interest and an early development of the graph method (and we could not present the graph method in this cited paper).

Finally, I really enjoyed this paper and would recommend it to colleagues. It raises highly relevant issues concerning the processes of data production and data analysis in agricultural sciences along with the question of opening research. Thank you!
I missed the step further in the application concerning the intercropping global dataset. It would have been very nice to compare the global analysis to single location/single mixture analyses for a basic agronomical trait like overyielding for example. Is the global-dataset really more powerful? Because several confounding factors are necessarily aggregated in the global-dataset analysis, how much larger is the residual variance?
I guess this step is the next one, i.e. the comparison of the conclusions on crop diversification brought by one or two datasets and the ones that we can have with larger datasets. Some research questions are only addressable with substantial datasets, such as the effect of the environment on intra- and interspecific variability. These issues are now among our research priorities. Furthermore, having formatted all these datasets, we have a very detailed understanding that allows us to have all

the necessary warnings regarding confounding factors. However, the methodological aspects (factorial or non-factorial designs and associated analyses/method groups) could be further explored in the future.

Answer to reviewer #2

The manuscript « A workflow for processing global datasets: application to intercropping » tackles the problem of collecting raw crop data from heterogeneous sources, in order to analyze them jointly. To this aim, the authors present a 3 step workflow that is illustrated on a concrete example (already published in a previous article). They also provide some methodological contribution by suggesting the representation of the overall experimental design as a connected graph. This representation allows one to reformulate the problem of extracting a complete factorial design subset from the global dataset design into a problem of finding cliques in the corresponding graph.

The main focus of the article (gathering and tidying datasets in order to make them amenable to a joint analysis by data analysts) is of real interest for the community, as it offers opportunities to i) increase the power when testing hypotheses compared to considering each dataset separately and ii) query the initial datasets in new ways that were not initially accessible. It is also important to emphasize that the community should support initiatives that make data reusable as they require a significant amount of time and work to deliver ready-to-use augmented datasets to the largest audience.

Thanks for these nice comments!

A first concern it the exact status of the manuscript: in its current state, it mixes different aspects, going from feedbacks from a previous experiment (the one of Gaudio, 2021, and related references), to general guidelines for global dataset construction, through methodological contribution. This combination makes it hard to get the real significant contribution of the paper. Some rewriting of the Introduction section could help here.

Yes, we initially wanted to write a data paper "augmented" with the process of data curation. The methodological addition shifted the paper tone to a more classical result paper. We still aim for the kind of paper we'd like to read and we have reorganized the manuscript with the help of your remarks. Each theoretical aspect in the dataset building approach is now associated with a concrete case study example. It clearly makes the paper easier to read and understand. We also clarified our main goal with this paper at the end of the *Introduction*, which is to promote these types of practices and to encourage the scientific community to move towards a more open approach to conducting science (making it more reproducible and shared).

Here are some additional elements of discussion of the article.

Identification of complete factorial subsets

The idea of proposing a way to automatically extract a subset of the complete global dataset that corresponds to a complete factorial design on a restricted number of levels is really interesting, for further analysis obviously, but also as a way to describe the global dataset. But the authors must provide a more detailed discussion about the use and limitations of the proposed procedure. Here are some points that should be discussed / explained more thoroughly:

- Usefulness of a complete design: while a complete design prevents the complete confounding between factors there is a huge literature on balanced incomplete design and strategies to organize the confounding to keep small order interaction distinguishable. On the other hand, note that completeness does not prevent partial confounding as the number of samples in each cell of the selected factorial crossing may be quite imbalanced, a feature

that is not accounted for in the proposed approach. I would like to see some more discussion on this aspect in the manuscript.

- Is the procedure amenable to extensions? For instance, can one investigate the graph representation to look for incomplete but connected subparts of the global dataset? How can the proposed procedure account for additional constraints (i.e. a minimum number of species/varieties? Require some combinations of levels to be present)? It seems that if the procedure explicitly enumerates all possible maximum cliques then a posteriori filtering is always possible, but can the constraints reduce the computational burden, making the procedure amenable to larger global datasets?

- It is mentioned that the procedure has an NP hard complexity, then that in the example the solutions can be found « quickly ». There is no clear quantification, so one has no idea about the size of designs that can be handled in practice. I recommend the authors to provide e.g. a table, displaying for different combinations of numbers of factors and number of levels per factor the computational time.

- The illustration of the method on a synthetic example is very clear, but the application to the real dataset is quite vague (l224-230). More precisely it is unclear whether the application is trivial or not : if one has a 2 partite graph where the second set of vertices (N fertilization) has only 2 nodes and one looks for the maximal 2-clique with the constraint that the 2 levels of N fertilization must be present, it seems that the problem boils down to looking at Table 1 and select Experiments (i.e. rows) for which the Nitrogen fertilization column is full (13 experiments satisfy this criterion). Is this what was done or am I missing something? If this is the result I guess this is not the best way to illustrate the usefulness of the procedure.

As the points you mention in the bullet points are, in a certain way, interconnected, we respond with a global answer (while addressing specific points).

As you stated in the beginning of your comment, our aim was more to provide a way to describe and explore the experimental design in a global dataset (actually giving a broad picture of the dataset) than to develop a tool to identify different types of designs suited for a range of statistical analysis. Rather than to do further research on design theory, we clarified our aims in *Introduction*, underlying that the method is useful to identify the presence/absence and combinations of plant variables and different factors in the dataset.

As an answer to point 4, we updated this figure to illustrate both a theoretical and practical example. The practical example targets the global dataset we are presenting in the paper, focusing on the main factors of interest for users (environments, type of crops, and management options). We think this application is clearer to show that this method could help to explore and glance at the structure of an experimental design, in this case showing the high level of incompleteness (few crops are common to various environments). On these low complexity examples (the design length was 66 combinations), we could probably solve the problem by reorganizing and filtering the design table.

Concerning the computational optimization (point 2) and performance (point 3), we did not search to optimize the procedure. We currently proceed as you describe: listing all the complete factorial designs with a selection of factors and filtering in this table to account for constraints on levels or their combination. In standard cases the execution time was around 100ms (new Figure 3). On a much larger theoretical design, that would be comparable to the size of a large federation of experiments, this time stayed within reasonable limits. We tested the method on 100 iterations of a random subset of 50% a large factorial design (3 factors: a first with 50 modalities, a second with 10, a last with 5, n = 2500). As iterations had an effect on the structure, the computation time (R version

4.3.1, ARM64, M2 CPU) was between 223 and 1693 ms (median was 509 ms) with the number of distinct factorial designs between 32 and 768 (median was 192).

About raw data

The authors aim at providing a sounded way to collect and tidy different datasets in view of their joint analysis, which is a useful initiative. However the procedure advocated by the authors is to provide the raw data, without any normalization. This point is roughly not commented in the manuscript, except in l265-269 where it is mentioned that researchers may be willing to access the data at different levels (e.g. plant or crop level). While I understand that such accesses requires the data to be « as raw as possible », it is important to mention that i) many data scientists experienced the frustrating case where one is unable to reproduce the results of a publication due to the impossibility to rebuild the normalized dataset from the raw data, and ii) the initial data producers are the ones aware of the experimental specificities, and consequently the ones that can suggest a sounded way to normalize the data (for e.g. spatial field effects, experimenter effects, etc). So it would be really nice to have both the raw data and the codes to rebuild the dataset as preprocessed in the inital publication on side, as an option to be used. This should be feasible as the authors mention that collecting the data requires a strong interaction with the data providers anyway. One could also think about future authors contributing to the global dataset by adding additional data but also alternative normalization codes, corresponding e.g. to new ways to analyze the data.

We added a mention to the normalization process in the revised manuscript: "*The data linked to crop measurements were not normalized (for e.g. spatial field effects, experimenter effects) as the information on experimental design type and structure was only accessible in very few trials*".

There are two main reasons for our choice.

1) The first one is very pragmatic. Actually, even if strong interactions were created with data providers, information on the experimental design type or structure, along with associated factors, were not published, or tracked and could basically be considered lost. Nevertheless, scientists who would like to add their raw data and the additional design information to the global dataset are welcome and we are going to create a new version of the dataset online precising that (and additional things pointed out by another reviewer).

2) This practice of data normalization is rarely used in agronomy (whereas it is mandatory in plant breeding trials or quantitative genetics). To our knowledge, only a few attempts were done in variety mixtures (e.g. Montazeau et al. 2020, https://doi.org/10.1111/1365-2664.13735). Thus, we can assume that if scientists want to add new datasets in the global one, it will be easier to focus on the raw data.

Scope of the paper

In many places the use of a global dataset is compared to the meta-analysis approach. Different aspects are discussed, going from the size of the dataset one can expect to collect in the two cases, to the working time these two types of data require for being processed. While the comparison makes sense in the crop science context, one can notice that similar initiatives (i.e. development of methods/ressources for meta-analysis or global datasets) are developed in other fields (e.g. quantitative genetics) where a same discussion would possibly lead to different conclusions. As an example, meta-analysis has now become a popular practice for genome-wide association studies in

human genetics, where (among others convenient features) it provides a way to share results without sharing individual data that may be protected for ethics considerations. Meta-analysis can also be a way to avoid the modeling/fitting of complex correlation patterns between traits/panels/environments. It is consequently quite important for the author to give the precise scope of their study in terms of field application where their recommendations apply. Also referring to the previous point note that in between global datasets and meta-analysis data there is the case of collected pre-normalized datasets that should be discussed.

We understand your last comment as the fact that the use of meta-analysis was domain-dependent. We stated more clearly that meta-analysis was not a mainstream analysis method in crop science as compared to a set of trials composed from yearly replication of one or two locations. We added this sentence in the *Introduction* section: "*While the use of meta-analysis to report results is growing in crop science, it is not a mainstream analysis method compared to reports based on a repeated set of field trials*."

Considering an application to crop science, meta-analysis is not the standard analysis level to report proofs.
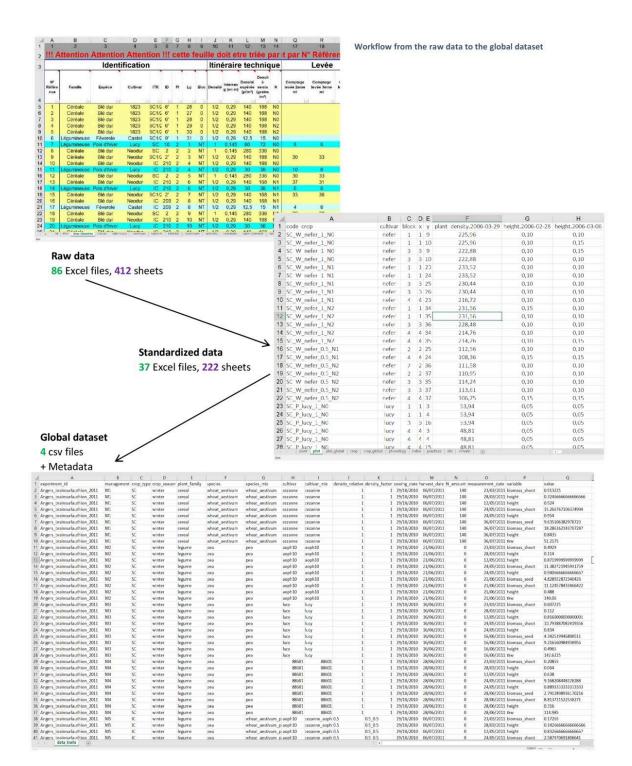
Minors

1/ It is a little bit awkward to read a paper about reusability of datasets that does not provide any link for the code associated to the procedure they present. Maybe the code can be found in the Gaudio article, but I would prefer to have this mentioned and the weblink available if any.

There are two main underlying reasons for the absence of code.
First, it is important to note that the majority of the work to curate the data was done manually, considering the state of the retrieved files: some were relatively clean, while others caused our computers to crash numerous times due to their heaviness, the inclusion of macros (software incompatibilities) and of various tables and graphics. Ultimately, this longest part is not automatable (unfortunately!). Thus, after a manual step of cleaning and homogenizing, we just gathered all the data (experiments) together using an R script which highly depends on the structure of our standardized files (see figure below).
Second, there is no original analysis linked to the data workflow justifying sharing a given code.

Workflow from the raw data to the global dataset

Raw data
86 Excel files, 412 sheets

Standardized data
37 Excel files, 222 sheets

Global dataset
4 csv files
+ Metadata

2/ I found the following sentence to be a little bit misleading: « the resulting overall design did not allow an intermediate statistical analysis… » (l185). What does that mean? One can perform an ANOVA on this dataset, including main effects and maybe some low order interactions, just as we can with any imbalanced or non-complete dataset. One just needs to be aware about the consequence of the partial confounding when interpreting the results. I emphasize here again that completeness does not amount to balance, so the complete subdesigns that are extracted will also require some caution when it comes to their interpretation.

We added the sentence you proposed in the manuscript (*"While the experimental designs had many similarities (e.g. species cultivated, agricultural practices), the resulting overall design is unbalanced. Thus, before any statistical analysis, one just needs to be aware about the consequence of the partial confounding when interpreting the results"*).

3/ The authors chose to distinguish between theory and practice by first having a section introducing the main concepts of global dataset constitution, then illustrating these concepts through the case study. When reading the conceptual part, one may not understand the implied consequences of the different guidelines (the fact e.g. that one will possibly have to deal with different programing languages to process the different datasets), so I was wondering if an alternative presentation where each concept is directly illustrated through the case study would be more sensible. This is not a strong recommendation as I'm aware it would require some significant rewriting of the paper, and both organizations (the one chosen by the authors, the alternative one I'm suggesting) make sense, just a suggestion for consideration.

We have reorganized the manuscript in order that each theoretical point is now illustrated directly by the case study.