Project Number: 952914

Project Acronym: FindingPheno

Project Title: Unified computational solutions to disentangle biological interactions in multi-omics data

D3.4 List of Genotype-Phenotype Associations in Chicken, Salmon and Maize

WP3 BEYOND GWAS: FROM PAIRWISE ASSOCIATIONS TO MULTIOMICS

Original due date of deliverable:	29/02/2024
Revised submission date:	16/09/2024
Submission date:	16/09/2024
Dissemination level:	Confidential

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952914

DATA ANALYSIS & AUTHORS	
FChampalimaud	Dylan FELDNER-BUSZTIN Panos FIRBAS Gonzalo DE POLAVIEJA
CER	Marton CSILLAG
CONTRIBUTORS & PROVISION OF DATA	
UCPH	Shyam Gopalakrishnan
UCPH	Antton Alberdi Estibaritz
UCPH	Sofia Marcos
UCPH	Jazmín Ramos Madrigal
UCPH	Xin Sun
UCPH	Morten LIMBORG
UCPH	Jaelle BREALEY

DOCUMENT HISTORY

Version	Date	Description	Changes
0.1	12/09/2024	First draft	Typing up of introduction, methods, results, discussion and conclusion.
0.2	13/09/2024	Second draft	Finalising details for submission
1.0	16/09/2024	Submission	Submission to EU portal

List of Genotype-Phenotype Associations in Chicken, Salmon, Maize and Wheat

Introduction

There are many different ways to generate genotype-phenotype relationships. Among the simplest is to treat each genotype-phenotype pair as a pair of isolated vectors and calculate their statistical relationship. Even then, there are options such as which type of statistical association to calculate and how to adapt the technique to different experimental conditions. We used this approach, but we also looked wider and used an approach which considers the dataset more holistically (attribution method) and the research field more broadly (KGE method).

Another approach to generating a list of genotype-relationships is to use the feature importance (attribution) of machine learning methods trained to predict phenotype from genotype. This considers the dataset more holistically than the correlation methods measured above, as the ML methods try to find combinations of features that best predict the phenotype. However, not all predictions are good. Therefore, we look at the attributions of the best predicted validation fish from each CV split. While there are many methods to calculating feature importance, axiomatic attribution (Sundararajan et al., 2017) stands out as being a theoretically sound technique.

The axiomatic attribution method, (also known as integrated gradients method), allows us to calculate feature importance values with respect to a given input datapoint and a neural network. It satisfies two fundamental axioms that are essential for any feature importance method.

1) Sensitivity - if any two input data points differ in only one feature, but have different predictions, then the differing feature must have a non-zero attribution.
2) Implementation invariance - if the outputs of two neural networks are equal for all inputs, then their attributions must be identical, regardless of the implementation details of the network.

Another advantage of the method is that it requires no modification to the original neural network, and is applicable to a variety of networks (e.g. text, vision) and also to deep networks trained for regression tasks, as in our case.

Using the method is relatively simple. It requires defining a baseline datapoint which has to be a "neutral" datapoint. For example in an image classification network, an all-black image is an ideal baseline. In our case the baseline is represented by the fish with all input feature values set to zero. After defining the baseline datapoint, we take a straight line path between this baseline datapoint and the input we would like to compute the attributions for. After passing all the points along this straight line path through the network, the gradients are computed for each of them. The attribution values of each input feature are then obtained by taking the path integral of the gradients along the straight line path.

Working solely with individual datasets ignores all prior work. For multi-omics there are already thousands of papers published, so it seems like a worthwhile opportunity to use genotype-phenotype relationships that

were previously found. In general, literature is stored in PDFs, which cannot be immediately parsed for genotype-phenotype relationships. However, text-mining projects such as Knetminer (Hassani-Pak et al., 2021), have made strides towards addressing this. Knowledge graph embedding (KGE) is a field in its own right. In KGE, the aim is to re-represent KGs in continuous space. An interesting aspect of this is that in a KG of genes, proteins and other biological entities, there are also natural embeddings arrived at using LLMs. Here we use AgroNT (Mendoza-Revilla et al., 2024) to embed the sequence of genes in Knetminer and then use that embedding to do KGE. One of the ways to evaluate KGE is link-prediction. Given a head and relation, the model tries to predict the tail. It does this by assigning a score to each possible tail. People in the field use a metric called hits@10. So if the correct tail is in the top 10, it is a hit.

There are many KGE methods as well as applications for biology as reviewed in Mohamed et al., (2021), one that has stood the test of time is transE (Bordes et al., 2013). transE is a relatively simple method, which readily allows for customization needed to incorporate the AgroNT embeddings. Knowledge graphs consist of heat (h), relation (r), and tail (t) triplets. transE treats the KGE process as a translation problem: $h + \square \approx \square$. In other words, the head embedding, plus the relation embedding, should equal something close to the tail embedding. To get to this,

Methods

Datasets

HoloFood

For the investigation in chicken (*Gallus domesticus*) we used data produced from the HoloFood project (Tous et al., 2022; Marcos et al., 2023a; Marcos et al., 2023b). The dataset our collaborators provided includes metadata (per animal and pen), metabolites, metagenomics, metatranscriptomics and host-transcriptomics. Host-transcriptomics stood out as the most valuable -omic in previous reports and we were able to utilise it to a certain degree here but unfortunately host-genomic data, on which some of our most promising previously developed techniques relied on, did not become available in time to be used in this report.

HappyFish

The HappyFish dataset originally described in Brealey et al., 2024 is the largest salmon multi-omics dataset to date. It covers over 400 salmon individuals with different subsets of them being profiles for genotype, transcriptome, metagenome, metabolome and all of whom have phenotypic information.

Maize

We are using the genomes2fields dataset (https://www.genomes2fields.org/), this is a large-scale project in the US that measures grain yields in different locations each year. It generally provides the genotypic information of the plants as well as environmental data.

Wheat

For wheat (Triticum Aestivum) we are using Knetminer (Hassani-Pak et al., 2021) knowledge graphs. Here we query the graph database for 50 different agronomically relevant keywords (including heat tolerance, disease resistance and so on). For each of these, we get a list of genes with a relevance score with respect to each keyword. We then create knowledge graphs for the top 100 genes in each keyword search. We process this further into head-relation-tail triples (e.g., gene HSP70, participates in, heat acclimation). We then split all the triples into training and validation sets, resulting in 8115 training triples and 4695 validation triples.

Causality

We applied the LGR (LASSO and generalization robustness) method as described in our previous deliverable on the new HoloFood data of chicken transcriptomics. LGR is a multi-step method that aims to find either causes or effects. Briefly, it selects an alpha (penalization parameter) for LASSO, runs LASSO regression to eliminate features, predicts labels using the remaining features, and finally performs attribute validation to individual features.

Attribution

The ML method as the basis for the attribution calculation was the multilayer perceptron from D3.2, which we describe again briefly below:

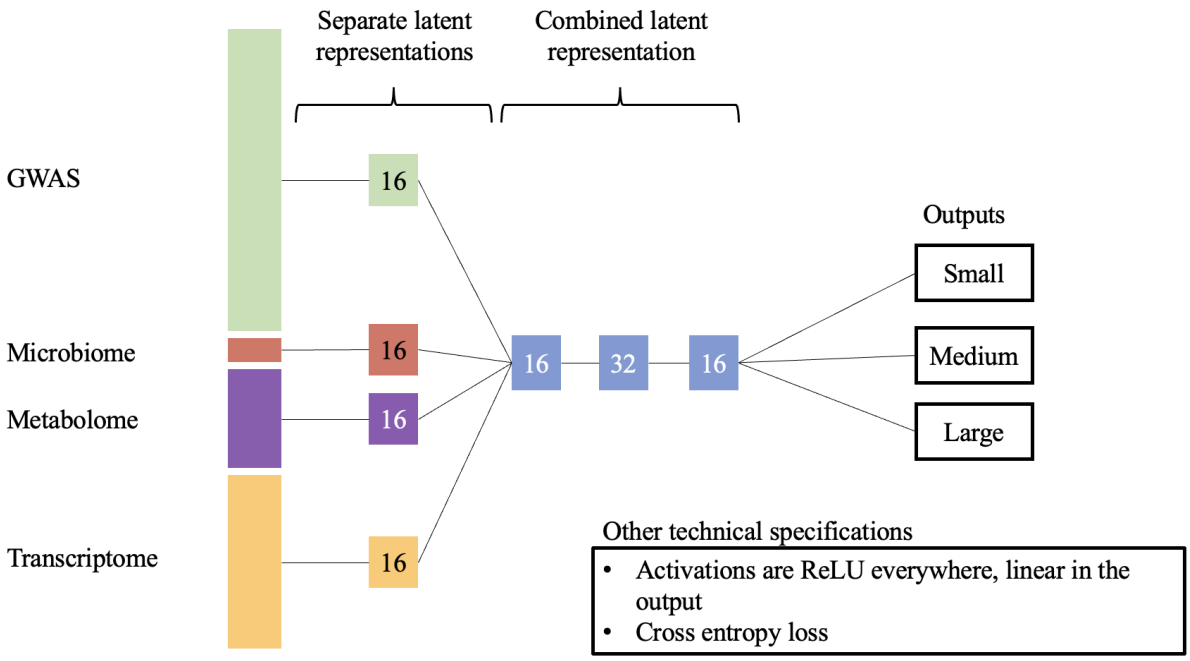

Figure 1. A neural network that takes in the GWAS and multi-omic data, projects them into a comparable lower dimensional space of 16 neurons each and then passes them through successive layers of 16, 32 and 16 units, respectively.

Knowledge graph embedding

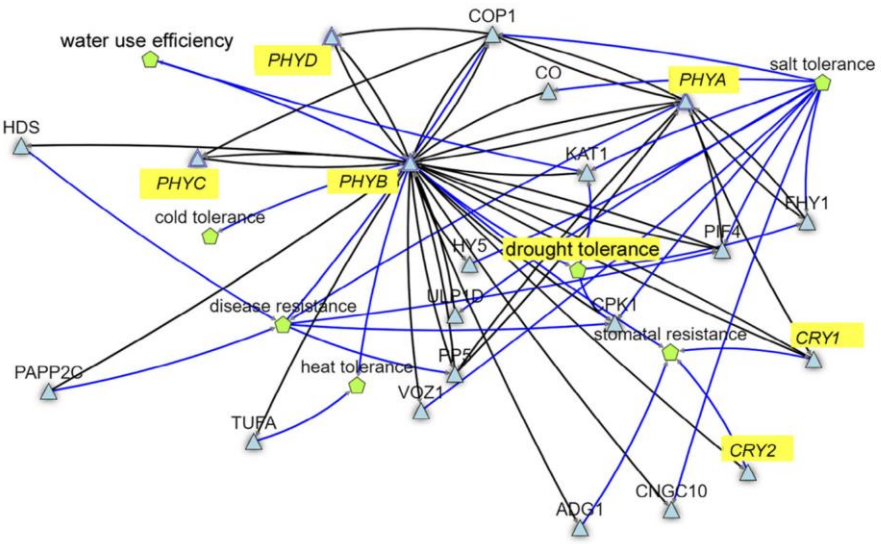

Figure 2. A Knetminer knowledge graph (image from Mawkieh et al., 2021).

In this experimental approach, we take the following steps:

- Enter the Knetminer website, navigate to the wheat (Triticum Aestivum) resource section.
- Download Cytoscape JSON. This is a textual representation of a subset of the knowledge graph.
- Process the JSON into head, relation, tail triplets.
- Feed the genes in the triplets dataframe through a nucleotide language model (e.g., AgroNT (Mendoza-Revilla et al., 2024)) to get their embeddings.
- Perform knowledge graph embedding (e.g., using transE), but with the embeddings of the genes fixed.
- In future work, we will generate new sequences with mutations on the LM embedded genes, see if they move closer to the trait of interest. In this draft however, we select a particular tail - 'heat acclimation' which will be important for crops in the next few years of climate change. We then feed in all genes to the relation 'participates in' and return the top 10 ones for the best scores to 'heat acclimation'.

Results

Maize - Correlation

Table 1.

Mutation Chromosome and Location	**Mean absolute correlation across years (2014 - 2020)**	**Lowest correlation across years**	**Highest correlation across years**
S5_76500404	0.129769	-0.397960	0.133281

S3_176083970	0.127518	-0.365411	0.144378
S3_130345140	0.126174	-0.118796	0.341226
S5_78476319	0.124617	-0.341194	0.070489
S3_114689619	0.123514	-0.127446	0.339909
S1_300195867	0.123488	-0.048645	0.409273
S7_1107165	0.123344	-0.344465	0.121607
S10_5364506	0.123204	-0.327803	0.131917
S4_181586232	0.122874	-0.374295	0.100363
S3_59240512	0.122550	-0.138215	0.276621

Chicken - Causality

While the method identifies genes potentially linked to the weight phenotype (see supplementary Excel file), the numerous conditions involved (Caecum/Ileum, tissue/mucosa, and three different animal ages) require splitting the dataset into at least 12 distinct trials, each with around 50 samples. This fragmentation means that the statistical strength of the suggested relationships might not be significant.

With the salmon data that was used in the previous work, the method produces more reliable suggestions, and identifies the haemoglobin complex as a clearly interesting target for further investigation. Furthermore in that case, thanks to the availability of host genome data, we can implement further techniques which when combined yield even more reliable results.

Table 2. Top 10 genes identified by the LGR along with their full names looked up on Salmobase (https://salmobase.org/).

Gene symbol	Gene name/ product
LOC106566372	glutathione S-transferase Mu 3 [product]
hba	hemoglobin subunit alpha
LOC106607380	hemoglobin subunit beta-1 [product]
duox	dual oxidase [product]
LOC106583594	heme transporter hrg1-A-like [product]
LOC106601072	hemoglobin subunit alpha [product]
pgfrl	Platelet-derived growth factor receptor-like protein
LOC106601071	hemoglobin subunit alpha-4 [product]
LOC106578021	purine nucleoside phosphorylase-like [product]
LOC123724067	hemoglobin subunit beta-like [product]

Salmon - Attribution

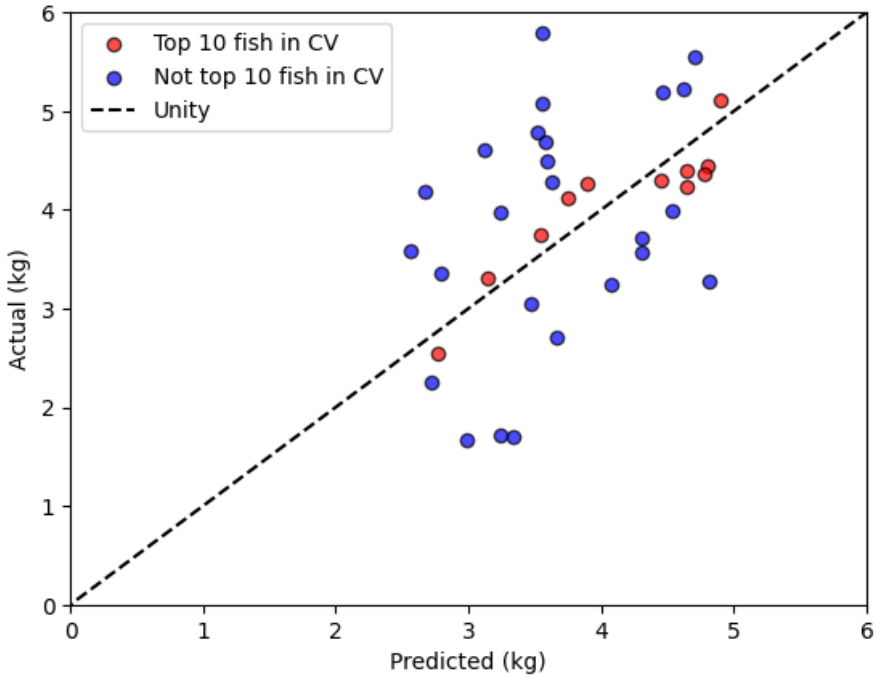

Figure 3. Scatter plot of prediction vs actual salmon fish weights in kilograms.

Table 3. Top 10 genes by mean attribution.

Feature	**Mean attribution**
LOC106588400	0.017704
hba	0.016162
NC_059454.1_78018296	0.016010
NC_059454.1_77772094	0.015071
LOC106601072	0.015001
LOC106601071	0.014249
NC_059451.1_111586271	0.014208
NC_059442.1_89613241	0.014052
LOC106607380	0.013993
NC_059442.1_89613243	0.013867

Wheat - Knowledge graph embedding

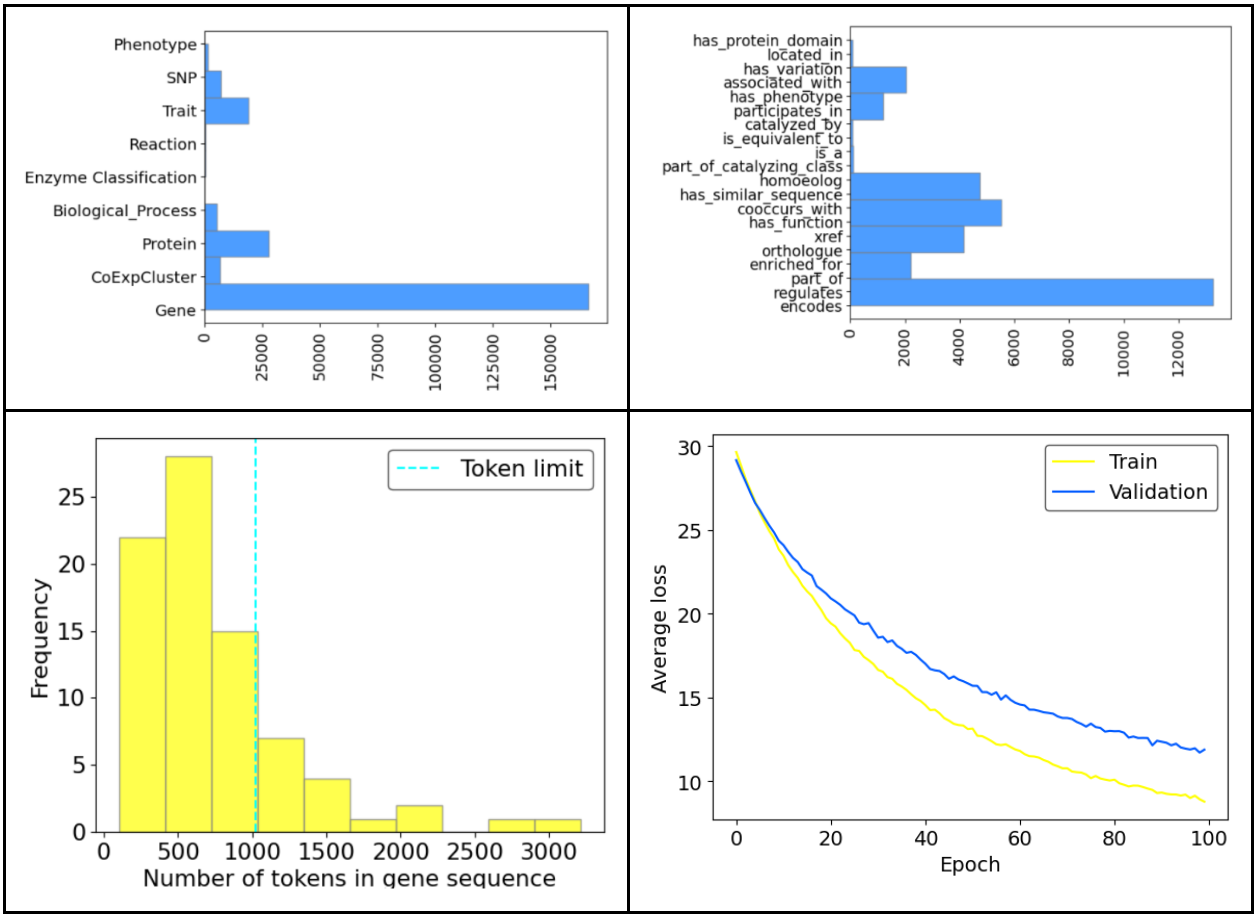

Figure 4. Results of the KGE pipeline. Top left: distribution of edge types in the graph, top right: distribution of edges in the graph, bottom left: distribution of tokens arising from the nucleotide sequences, bottom right: learning curves.

Table 4. Top 10 genes from the KGE method.

head	score
MADS-box transcription factor 6	29.839184
Auxin efflux carrier component 1	32.830154
SWI3C	34.15934
NAC domain-containing protein 2	35.497692
Dehydration-responsive element	35.72682
B3 domain-containing transcript	35.94995
Regulatory protein NPR1	36.314632
Floral homeotic protein APETALA 2	37.153297
WRKY transcription factor WRKY24	37.216694
Zinc finger protein HD1	37.457985

Discussion

Correlation

Apart from ranking genes by mean absolute correlations as in Table 1, it could be interesting to investigate how consistent correlations are. Do we then rank them according to their highest average/ median correlation? Alternatively, do we take the spread of correlations into account. For example, using the Sharpe ratio (Sharpe 1994) we divide the mean correlation by the standard deviation of correlations. This penalises correlations that are more highly variable. But again, this opens questions. Should we penalise correlations with a high upside variability i.e. if the mean correlation is X but it can go up to 3X. So instead we can look at penalising for downside variability (i.e., down Sharpe). Future work could consider in more depth the pros and cons of these other options.

Attribution

The best predictions were spread across the weights, which is fortunate because if the model only predicted fish where the average weight, then fish that really were the average weight would be selected in the top 10. This points to a challenge of this attribution technique, which is model-dependent. It should only be used when the prediction model works better than random. Then if there are two models, giving different weights importance to features, the model with the higher validation accuracy should be used.

Causality

The heterogeneous nature of multi-omic data proves to be a formidable challenge. In our previous work we described a combination of methods (LGR and GFGWA) that when used together provide more reliable suggestions. Unfortunately the necessary data for GFGWA is not yet available in chicken or maize, while the available number of samples for the data used in LGR is significantly lower in chicken than in salmon and unavailable in maize.

Knowledge Graph Embedding (KGE)

This is a "discovery" approach where there is no ground truth until an experiment is conducted. We are still working on it. One thing we can do is test the robustness of the technique by taking subsets of the data and see if they come up with the same answer. However this does not test accuracy or provide groundtruth. Taking the approach to the next level, creating new sequences that are fed into a language model to see where they rank in the knowledge graph embedding prediction is even more uncertain. It is possible that there are mutations that produce embeddings that have good predictions in terms of the model, but would be harmful for the organism. In other words, it would be difficult to know if we are interpolating or extrapolating. However, in the limit of having a very comprehensive knowledge graph and good KGE model, we cannot think of a reason why this method should not work.

Conclusion

- We have presented top 10 candidate genes for agronomically relevant phenotypes using several different approaches: causal discovery, axiomatic attribution, correlation and knowledge graph embedding.
- The heterogeneity of the multi-omic data *causes* heterogeneity in methods. We have to use different approaches depending on the collection of measurements in each organism.
- These could be taken as starting points to look for targets for intervention in a genetically modified organism initiative.

References

- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, *26*.
- Brealey, J.C., Kodama, M., Rasmussen, J.A., Hansen, S.B., Santos-Bay, L., Lecaudey, L.A., Hansen, M., Fjære, E., Myrmel, L.S., Madsen, L. and Bernhard, A., 2024. Host–gut microbiota interactions shape parasite infections in farmed Atlantic salmon. *Msystems*, *9*(2), pp.e01043-23.
- Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J.D., Amberkar, S., Phillips, A.L., Doonan, J.H. and Rawlings, C., 2021. KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnology Journal*, *19*(8), pp.1670-1678.
- Marcos, S., Odriozola, I., Eisenhofer, R., Aizpurua, O., Tarradas, J., Martin, G., Estonba, A., Gilbert, M.T.P., Kale, V., Baldi, G. and Finn, R., 2023a. Reduced metabolic capacity of the gut microbiota associates with host growth in broiler chickens.
- Marcos, S., Odriozola, I., Langa, J., Baldi, G., Sahin, E., Mak, S.S.T., Pless, L., Tarradas, J., Estonba, A. and Alberdi, A., 2023b. Priority effects and microbial cross-feeding shape zoonotic agent spread in broiler chickens.
- Mawkhiew, H.B., Sahoo, L. and Kharshiing, E.V., 2021. Gene-to-trait knowledge graphs show association of plant photoreceptors with physiological and developmental processes that can confer agronomic benefits. *Genetic Resources and Crop Evolution*, *68*(7), pp.2727-2735.
- Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B.P., Richard, G., Caton, J., Lopez Carranza, N., Skwark, M. and Laterre, A., 2024. A foundational large language model for edible plant genomes. *Communications Biology*, *7*(1), p.835.
- Mohamed, S.K., Nounu, A. and Nováček, V., 2021. Biological applications of knowledge graph embedding models. *Briefings in bioinformatics*, *22*(2), pp.1679-1693.
- Sharpe, W.F., 1994. The sharpe ratio. *Journal of portfolio management*, *21*(1), pp.49-58.
- Sundararajan, M., Taly, A. and Yan, Q., 2017, July. Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.
- Tous, N., Marcos, S., Goodarzi Boroojeni, F., Pérez de Rozas, A., Zentek, J., Estonba, A., Sandvang, D., Gilbert, M.T.P., Esteve-Garcia, E., Finn, R. and Alberdi, A., 2022. Novel strategies to improve chicken performance and welfare by unveiling host-microbiota interactions through hologenomics. *Frontiers in physiology*, *13*, p.884925.