# Full epistatic maps retrieve part of missing heritability and improve phenotypic predictions

**Jean-Baptiste Carluer**[1,2], **Clément Carre**[3], Christian Chaux[3], Nicolas Roche[3], Gabriel Krouk[1,3], André Mas[2,3]

[1] *Institut des Sciences des Plantes de Montpellier - Systems Team.*
[2] *IMAG, Univ. Montpellier, CNRS*
[3] *Bionomeex*

## Context

Phenotype prediction has potentially important implications from agronomic selection to medical applications such as personalized medicine for instance. Being able to predict a particular phenotype from genetic data is however dependent on the identification of explanatory genetic variants. One of the successful and popular techniques to identify such variants is Genome Wide Association Study (GWAS). However GWAS signals quite often explain a relatively small portion of the phenotypic variance (concept of *missing heritability [h2]*).
Here we developed a pipeline of analysis to predict phenotypic values from genetic data. This work is built upon a new kind of GWAS (called Next Gen GWAS -NGG or 2DGWAS) that provides full epistatic 2D maps [1].
Here we evaluate i) if 2D GWAS signal indeed contains some of the missing heritability, and ii) if 2D-GWAS help to predict phenotypic values of plant (Arabidopsis) nutritional content. To do so we combined, seven machine learning methods : Deep Neural Network (DNN), Gaussian process, Gradient boosting, Random Forest (RF), Support Vector Machine (SVM), Linear regression, Lasso and Elastic-net. We show that 2D signals indeed contain part of the *missing heritability* and that they can be used to predict classified phenotype with a maximal f1-score set at 73%. We thus demonstrate that indeed important information lies in epistatic signals that bring us a bit further towards good predictions of very diverse phenotypes.

## Workflow

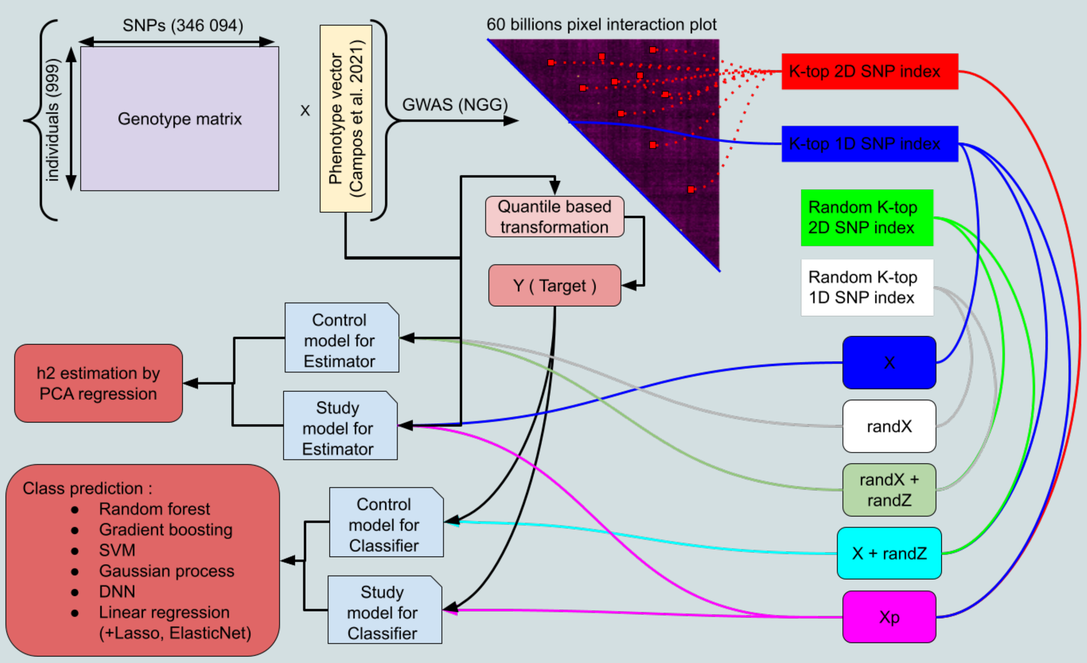


**Figure 1 : Estimation of retrieved missing heritability and of predictability :** Next-Gen GWAS allow us to make SNP's selection. The selected SNP, which come from 1D and 2D, are compared to SNP randomly picked. The comparision is based on narrow sense heritability estimation and class prediction to estimate heritability gain and prediction power gain.

## h2 estimation

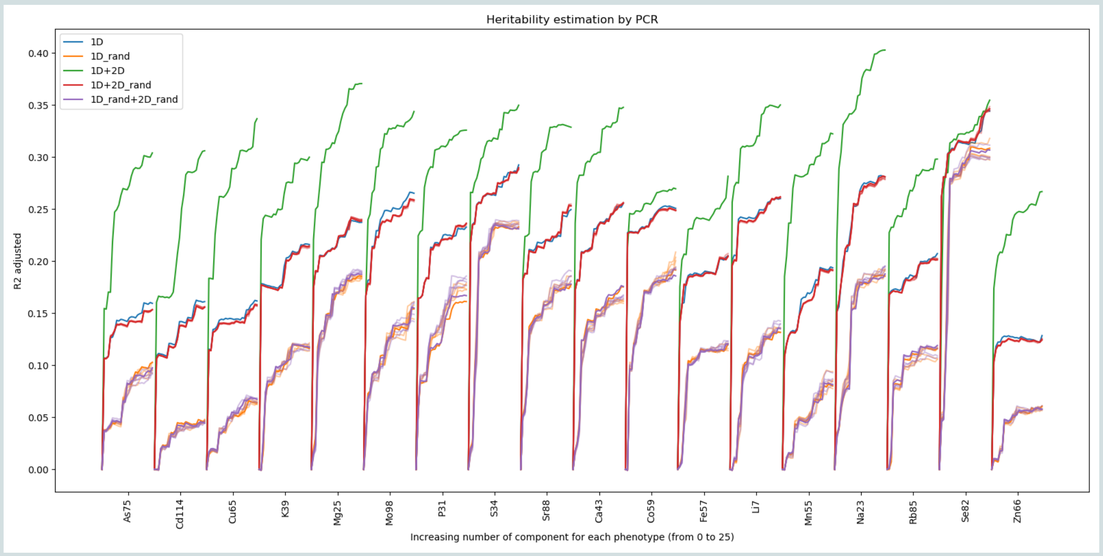


**Figure 3 : Estimation of heritability gain offered by SNP interaction :** Heritability (h2 seen as adjusted R2) is measured for an increasing number of PCA components, and for signal retrieved only from 1D-GWAS or 1D-GWAS + 2D-NGG.

## Acknowledgements and References

[1] Carre Clement and Carluer Jean Baptiste et al. "Full epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction". In: *bioRxiv* (2022). eprint: `https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500572.full.pdf`.

## Class Prediction

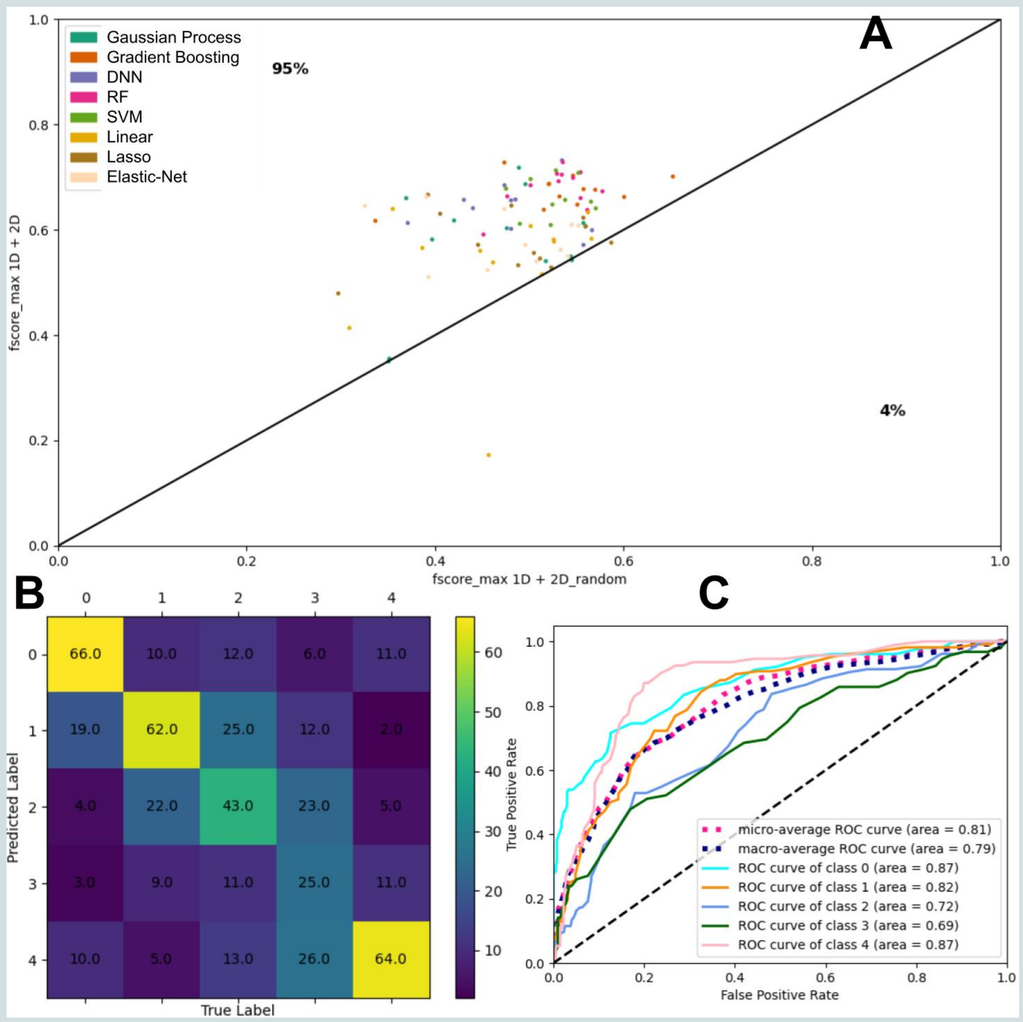


**Figure 2 : Estimation of prediction power gain offered by SNP interaction :** A) In this dot plot each color corresponds to a given machine learning model (among: SVM, RF, DNN, Gaussian processes, LASSO, Elastic-Net Classifier) trying to predict Molybdenum combined with different learning data inputs including a different number of classes (3 or 5) and different number of SNPs (50, 100, 500, 1000, 5000, 10 000). The x axis reports max F1 score for the mode provided with SNPs simple 1D signals and randomly picked 2D epistatic SNP combinations (our control). The y axis reports 1D signals and 2D signals retrieved by NGG for the sample model and parameter combinations, respectively. We observe a clear improvement (above the y=x line) of >95 % of the models. B,C) Example of the good prediction of classified molybdenum concentrations (Mo98 phenotype) for 10,000 SNPs for the RF method. B) confusion matrix, C) panel ROC curves for each class of Mo98 phenotype.

## Conclusions

We gave here some evidence of the interest that can be found in epistatic interaction. This epistatic interaction signal gain has been quantified for 18 phenotypes using Principal Component Regression. We have been able to target most of our phenotype as having interaction signal (Fig.3).
The epistatic interaction and causal SNP have been used in a class prediction study and has allow to reach F1-score of 73% for extreme class of some phenotype (Fig.2).
We aim to use these method in the future on different biological models with more complex genetic architectures.