AGR6932 MULTI-OMIC INTEGRATION FOR APPLIED PREDICTION BREEDING

Graduate Level – 3 credit hours
Spring 2023

Instructor:

Dr. Diego Jarquin
2089 McCarty Hall B
(352) 392-1811
jhernandezjarqui@ufl.edu
Office Hours: TBD. Please feel free to talk to me about any issue related to the course. It is recommended to start this process early in the semester.

Lecture time: Tuesdays period 6 (12:50-14:45); and Thursdays period 6 (12:50-13:40)

Location: McCarty Hall B – TBD, Gainesville, FL 32611

Course Description

This course is intended for students in Plant and Animal Sciences with emphasis on breeding applications. Students will learn the basis for modeling trait performance of genotypes assisted by the integration of multiple data types ‘omics’ considering different approaches (parametric and non-parametric Artificial Intelligence AI). In addition, some of the most relevant and novel topics of interest for the private industry and research institutions will be covered in the course such as artificial intelligence methods implemented for genomic prediction GP, GP aided by high-throughput phenotyping platforms, multi-omics integration, prediction and estimation of Genotype-by-Environment G×E Interactions, multi-trait prediction, sparse testing designs, prediction of time-related traits, etc.

Learning Objectives.

- Students will understand the underlying conditions and foundations of the different prediction paradigms (parametric [frequentist, Bayesian], and non-parametric Artificial Intelligence AI).
- Students will be able to write R scripts replicating the results of elaborated functions implemented under parametric and non-parametric frameworks.
- Students will learn to identify the best strategy to adopt according to the needs of the trait(s) under study.
- Students will identify the presence of genotype-by-environment G×E interaction and leverage it in prediction models.
- Students will integrate different ‘Omics’ of information in prediction models to outperform conventional models.
- Students will develop sophisticated prediction pipelines for plant and animal breeding.
- Students will learn when and how to implement the different cross-validation scenarios in breeding pipelines.
- Students will deploy complex prediction pipelines using parallel computing on High-Throughput Computing (HTC) and High-Performance Computing (HPC) facilities such as HiPerGator.

Intended Audience

The course is designed for MS and PhD graduate students in the following disciplines: plant breeding, agronomy, horticultural sciences, environmental horticulture, forestry, and animal breeding.

Course Objectives

This course will demonstrate the development and utilization of prediction models in plant breeding programs and how to implement these at different stages of the breeding pipeline. The focus of the course is to facilitate students the foundations of the different paradigms (parametric, and non-parametric AI) in which these implementations are based. This course is designed to complement other plant breeding, quantitative genetics, and statistical genetic courses. At the end of this course, students will be able to distinguish the differences between the prediction paradigms assisting in the selection of superior cultivars, integrate multiple layers or 'Omics' (genomics, enviromics/weather, soil, high-throughput phenotyping, etc.) of information in the prediction models, apply quality control on the data, data collection, and data alignment. In addition, students will also be able to estimate the genotype-by-environment G×E interaction and leverage it in prediction models, exploit genetic correlation by considering multi-trait analysis, and achieve hybrid prediction based on the General Combining Ability GCA and Specific Combining Ability SCA terms using the marker information of the parents in interaction with environments. The students will also be able to write their own modules to conduct customized analyses adapted to the needs of specific breeding programs (annual crops, perennial crops including tree species, repeated measures, time-related traits). Finally, the students will be able to develop complex prediction pipelines using parallel computing in High-Throughput Computing (HTC) and High-Performance Computing (HPT) facilities such as HiPerGator.

Evaluation

Quizzes

Three quizzes will happen randomly during lectures. They will consist of questions pertinent to the topics being discussed in class. Students are required to email the answers to the instructor during the allotted time period when the quiz is assigned. <u>There is NO make up for quizzes</u>. Answers will be reviewed and discussed in class after completion to provide comprehensive feedback.

Breeding pipeline development.

The students will develop a genomic prediction pipeline including at least two different omics or layers of information (genomics plus another [soil, weather, images, high-throughput phenotyping, satellite, etc.]) as part of the final project.

The pipeline development will be divided into three parts:

Part 1. Due on Thursday, January 26, by 5 PM (ET) in Canvas. Identification of a suitable multi-environments data set containing phenotypic data and at least two data 'omics' to be used as covariates (marker SNPs and any other layer [weather, soil, high-throughput phenotypic, etc.]) Describe the data set and covariates, including your name, the source, trait(s) of interest, and the objective of the study (limited to 200 words). Each student will receive feedback from the instructor and TA.

Part 2. Due on Tuesday, March 28, by 5 PM (ET) in Canvas. It should include all the sections presented in Part 1, plus the following components of a prediction pipeline: quality control on the data, compute the covariance structures to model main effects of the covariates and the interactions between these and environments and/or environmental covariates, and the simulation of training testing prediction scenarios (prediction of *i*) tested genotypes in unobserved environments, *ii*) untested genotypes in unobserved environments, *iii*) tested genotypes in unobserved environments, *iv*) untested genotypes in unobserved environments).

Part 3. Due on Tuesday, April 18, by 5 PM (ET) in Canvas. It should include all the sections presented in Part 1 and Part 2, plus the prediction stage using different paradigms (parametric [frequentist, and Bayesian], and non-parametric [Artificial Intelligence, and kernels]). Each student will receive feedback from the instructor and TA.

Oral presentation. Each student will present their project during a 15-minute presentation in class on Tuesday (April 20) or Thursday (April 25) (students will be randomly assigned to these dates). The presentation format is open, and the students are encouraged to use their preferred delivery method. Each student will receive feedback from the instructor and TA.

Exam
There will be a final take-home exam on Tuesday May 2 during the regular class time (12:50 to 2:45).

Activity	Number	Points/activity	Total
Quizzes	3	4	15
Pipeline development	3 stages	10	30
Oral presentation	1	20	20
Final Take-home Exam	1	35	35

A > 91 B+ 85 to 90 B 80 to 84 C+ 75 to 79 C 70 to 74 D+ 65 to 69 D 60 to 64 E < 60

UF grading policies https://catalog.ufl.edu/UGRD/academic-regulations/grades-grading-policies/

Software and computer
You will need to use your own laptop for data analyses demos and homework. Different software will be used during the semester. Specific instructions about where and how to obtain them will be given in class. All analysis will be conducted using R-CRAN statistical software (standalone version only).
To access the High-Throughput Computing facilities (e.g., HiperGator), OSX and Windows users can use the terminal (command line); however, I strongly recommend Windows users to install Putty. To exchange files between personal computers and the HTC facilities, Windows users are encouraged to install WinSCP while OSX users can install Cyberduck. There are free versions for all the mentioned software; no need to pay for these.

Recommended Literature – specific scientific papers will be assigned as the semester progresses.

Montesinos López OA, Montesinos López A, Crossa J. Multivariate statistical machine learning methods for genomic prediction. Cham: Springer; 2022. https://link.springer.com/book/10.1007/978-3-030-89010-0
Ahmadi N, Bartholome J. Genomic Prediction of Complex 2022. Humana, New York, NY. https://link.springer.com/book/10.1007/978-1-0716-2205-6

Course Schedule and Topics (Tentative)

Week	Description
1. Jan 9-Jan 13	**Theme: Software overview and connectivity** Day 1: Tuesday, January 10 a. Introductions, reviewing the syllabus, and assessment of expectations b. Review of the software to use in the class (R, WinSCP, Putty, Cyberduck, etc.) c. Quick review of useful commands in R Day 2: Thursday, January 12

		a. Accessing the high-throughput computing facilities HTC using the command line (terminal, Putty). b. Review of basic commands useful to navigate in the cluster.
2.	Jan 17- Jan 20	**Theme: Genomic Selection and Statistical learning - Basic concepts** Day 1: Tuesday, January 18 a. Introduction to Genomic Selection and Genomic Prediction. b. Current paradigms, approaches, and models. c. Cross-validation schemes and Assessment of predictive ability. Day 2. Thursday, January 19 a. Prediction versus Inference b. Statistical Machine Learning Models – Basic ideas c. Data types and types of learning (unsupervised, supervised, semi-supervised).
3.	Jan 23 - Jan 27	**Theme: Data preparation for implementing prediction models.** Day 1: Tuesday, January 24 a. Fixed vs Random effects b. BLUEs and BLUPs Day 2: Thursday, January 26 (Due date for part 1 of the breeding pipeline development project) c. Quality Control on the genomic data. d. Methods for computing genomic covariance structures.
4.	Jan 30 - Feb 3	**Theme: Basic elements for building supervised statistical Machine learning models** Day 1: Tuesday, February 1 a. Linear Multiple Regression Model LMRM b. Fitting LMRM using Ordinary Least Squares (OLS) c. Fitting LMRM using Maximum Likelihood (ML) d. Fitting LMRM using Gradient Descent Method (GD) Day 2: Thursday, February 2 a. Pros and cons of standard LMRM b. Regularization parameters to overcome the curse of dimensionality. c. Ridge regression, LASSO regression d. Logistic regression, Logistic Ridge regression, and LASSO logistic regression.
5.	Feb 6 - Feb 10	**Theme: Overfitting, Model tunning, and evaluation of Prediction** Day 1. Tuesday, February 7 a. Overfitting and Underfitting b. Prediction accuracy vs. Model interpretability c. Cross-validation Day 2. Thursday, February 9 a. Model tunning b. Evaluating model performance for prediction
6.	Feb 13 - Feb 17	**Theme: Linear Mixed Models**

	Day 1. Tuesday, February 14 a. Intro to Linear Mixed Models b. Estimation Methods Day 2. Thursday, February 16 a. Linear Mixed Models in Genomic Prediction b. Multi-trait Genomic Linear Mixed-Effects Models
7. Feb 20 - Feb 24	**Theme: Bayesian Genomic Linear Regression** Day 1. Tuesday, February 21 a. Bayes Theorem and Bayesian Linear Regression b. Bayesian Genome-Based Ridge Regression c. Bayesian GBLUP Genomic Model Day 2. Thursday, February 23 a. Genomic-Enabled Prediction Bayes A Model b. Genomic-Enabled Prediction Bayes B and C Models c. Genomic-Enabled Bayesian LASSO Model
8. Feb 27 - March 3	**Theme: Estimation of Covariance Structures and Multi-Trait Prediction Models** Day 1. Tuesday, February 28 a. Bayesian Factor Analytic Model and other Co-variance structures Day 2. Thursday, March 2 a. Genomic Prediction Models using structured Co-variances b. Multi-trait methods for Genomic Prediction
9. March 6 - Mar 10	**Theme: Estimation of Genotype-by-Environment Interaction** Day 1. Tuesday, March 6. a. Additive Main-effects and Multiplicative Interaction Model AMMI b. Singular Value Decomposition SVD c. Estimation of AMMI model parameters. OLS Day 2. Thursday, March 9 a. Bayesian AMMI model b. General Bayesian AMMI model c. Hierarchical Bayesian AMMI model
10.	**NO CLASS – SPRING BREAK**
11. Mar 20 - Mar 24	**Theme: GxE Genomic Prediction models** Day 1. Tuesday, March 21 a. Genomic Prediction including interactions between markers and environments b. Genomic Prediction including interactions between markers and weather covariates c. Practical GxE d. Practical GxW Day 2. Thursday, March 23 a. Multi-omics integration using co-variance structures b. General and Specific Combining Ability (GCA and SCA) in interaction with environments c. Sparse Testing Designs d. Prediction of time-related traits e. Pipeline development for Genomic Selection

12. Mar 27 - Mar 31	**Theme: Reproducing Kernel Hilbert Spaces Regression and other Kernels** Day 1: Tuesday, March 28 (Due date for part 2 of the breeding pipeline development project) a. The RKHS b. Generalized Kernel model Day 2. Thursday, March 30 a. Linear Mixed Model with Kernels b. Hyperparameter Tunning c. Bayesian Kernel Methods
13. Apr 3 - Apr 7	**Theme: Artificial Intelligence Methods for Genomic Prediction** Day 1. Tuesday, April 4 a. Support Vector Machines and Support Vector Regression Day 2. Thursday, April 6 a. Fundamentals of Artificial Neural Networks and Deep Learning
14. Apr 10 - Apr 14	**Theme: Artificial Intelligence Methods for Genomic Prediction** Day 1. Tuesday, April 11 a. Artificial Neural Networks and Deep Learning for Genomic Prediction Day 2. Thursday, April 13 a. Convolutional Neural Networks
15. Apr 17 - Apr 21	**Theme: Artificial Intelligence Methods for Genomic Prediction** Day 1. Tuesday, April 18 (Due date for part 3 of the breeding pipeline development project) a. Introduction to Convolutional Neural Networks b. Random Forest for Genomic Prediction **Theme: Presentations** Day 2. Thursday, April 20. a. Oral Presentation
16. Apr 24 - Apr 28	**Theme: Presentations** Day 1. Tuesday, April 25 a. Oral Presentation
17. May 1 - May 5	EXAM

ADDITIONAL REFERENCES

Acosta-Pech, R., Crossa, J., de los Campos, G., Teyssedre, S., Claustres, B., Pérez-Elizalde, S., et al. (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor. Appl. Genet. 130, 1431–1440. DOI: 10.1007/s00122-017-2898-0

Crossa J., Perez-Elizalde S., Jarquin D., Miguel Cotes, J., Viele, K., Liu, G., and Cornelius, P. (2011) Bayesian Estimation of the Additive Main Effects and Multiplicative Interaction Model. Crop Science, 51:1468-1469. doi.org/10.2135/cropsci2010.06.0343

Jarquin D., Pérez-Elizalde S., Burgueno J., Crossa J*. (2016) A Hierarchical Bayesian Estimation Model for Multienvironment Plant Breeding Trials in Successive Years. Crop Science, 56(5): 2260-2276. doi:10.2135/cropsci2015.08.0475.

Jarquin D., Crossa J., Lacaze X., Pérez P., Cheyron P.D., Daucourt J., Lorgeou J., Piraux F., Guerreiro L., Burgueno J., de los Campos G*. (2014) A Reaction Norm Model for Genomic Selection Using High-Dimensional Genomic and Environmental Data. Theoretical and Applied Genetics, 127(3):595-607. doi:10.1007/s00122-013-2243-1.

Jarquin D., Howard R., Xavier A., Choudhury S.D. (2018). Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans. Agronomy, 8(4), 51. doi:10.3390/agronomy8040051.

Jarquin D., Kajiya-Kanegae H. Taishen C., Persa R.,Yabe S., Iwata H. Coupling Day Length Data and Genomic Prediction tools for Predicting Time-Related Traits under Complex Scenarios. Scientific Reports. Sci Rep 10, 13382 (2020). doi.org/10.1038/s41598-020-70267-9.

Jarquin D., de Leon N., Romay C., Bohn M., Buckler E.S, Ciampitti I., Edwards J., Ertl D., Flint-Garcia S., Gore M.A, Graham C., Hirsch C.N, Holland J.B., Hooker D., Kaeppler S.M., Knoll J., Lee E.C., Lawrence-Dill C.J, Lynch J.P., Moose S.P, Murray S.C, Nelson R., Rocheford T., Schnable J.C., Schnable P.S., Smith M., Springer N., Thomison P., Tuinstra M., Wisser R.J., Xu W., Yu J., Lorenz A. Utility of Climatic Information via Combining Ability Models to Improve Genomic Prediction for Yield within the Maize Genomes to Fields Project. Front. Genet., 08 March 2021 | https://doi.org/10.3389/fgene.2020.592769.

Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–1829. DOI: 10.1093/genetics/157.4.1819

Perez-Elizalde S., Jarquin D., Crossa J. (2011) A General Bayesian Estimation Method of Linear-Bilinear Models Applied to Plant Breeding Trials with Genotype by Environment Interaction. Journal of Agricultural, Biology and Environmental Statistics, 17 (1): 15-37. doi:10.1007/s13253-011-0063-9.

Persa R., Grondona M., Jarquin D*. (2021). Development of a Genomic Prediction Pipeline for Maintaining Comparable Sample Sizes in Training and Testing Sets across Prediction Schemes Accounting for the Genotype-by-Environment Interaction. Agriculture 2021, 11(10), 932; doi.org/10.3390/agriculture11100932

Ray S., Jarquin D., Howard R (2022). Comparing Artificial Intelligence Techniques with State-of-the-Art Parametric Prediction Models for Predicting Soybean Traits. The Plant Genome.

VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–4423. doi: 10.3168/jds.2007-0980.

Viera C.C., Persa R., Chen P., Jarquin D. (2022) Incorporation of Soil-derived Covariates in Progeny Testing and Line Selection to Enhance Genomic Prediction Accuracy in Soybean Breeding. Front. Genet., 08 September 2022 Sec. Plant Genomics. https://doi.org/10.3389/fgene.2022.905824

Attendance and Make-Up Work

"Requirements for class attendance and make-up exams, assignments, and other work in this course are consistent with university policies. Those can be found at: https://catalog.ufl.edu/UGRD/academicregulations/attendance-policies/

Online Course Evaluation Process
Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at https://gatorevals.aa.ufl.edu/students/. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, located in the Canvas course menu under GatorEvals, or via https://ufl.bluera.com/ufl/. Summaries of course evaluation results are available to students at https://gatorevals.aa.ufl.edu/public-results/.

Software Use
All faculty, staff, and students of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate.

Academic Honesty
As a student at the University of Florida, you have committed yourself to uphold the Honor Code, which includes the following pledge: "We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honesty and integrity." You are expected to exhibit behavior consistent with this commitment to the UF academic community, and on all work submitted for credit at the University of Florida, the following pledge is either required or implied: "On my honor, I have neither given nor received unauthorized aid in doing this assignment."

It is assumed that you will complete all work independently in each course unless the instructor provides explicit permission for you to collaborate on course tasks (e.g. assignments, papers, quizzes, exams). Furthermore, as part of your obligation to uphold the Honor Code, you should report any condition that facilitates academic misconduct to appropriate personnel. It is your individual responsibility to know and comply with all university policies and procedures regarding academic integrity and the Student Honor Code. Violations of the Honor Code at the University of Florida will not be tolerated. Violations will be reported to the Dean of Students Office for consideration of disciplinary action. For more information regarding the Student Honor Code, please see:
http://www.dso.ufl.edu/sccr/process/student-conduct-honor-code.
If you have any questions or concerns, please consult with the instructor or TAs in this class.

.
Services for Students with Disabilities
"Students with disabilities requesting accommodations should first register with the Disability Resource Center (352-392-8565; https://disability.ufl.edu/get-started/) by providing appropriate documentation. Once registered, students will receive an accommodation letter which must be presented to the instructor when requesting accommodation. Students with disabilities should follow this procedure as early as possible in the semester."

Campus Resources
Students experiencing crises or personal problems that interfere with their general well-being are encouraged to utilize the university's counseling resources. The Counseling& Wellness Center provides confidential counseling services at no cost for currently enrolled students. Resources are available on campus for students having personal problems or lacking clear career or academic goals, which interfere with their academic performance.
University Counseling & Wellness Center, 3190 Radio Road, 352-392-1575, https://counseling.ufl.edu/

Counseling Services
Groups and Workshops
Outreach and Consultation
Self-Help Library
Wellness Coaching

U Matter, We Care: If you or someone you know is in distress, please contact umatter@ufl.edu, 352-392-1575, or visit U Matter, We Care website to refer or report a concern and a team member will reach out to the student in distress.

Counseling and Wellness Center: Visit the Counseling and Wellness Center website or call 352-392-1575 for information on crisis services as well as non-crisis services.

Student Health Care Center: Call 352-392-1161 for 24/7 information to help you find the care you need, or visit the Student Health Care Center website.

University Police Department: Visit UF Police Department website or call 352-392-1111 (or 9-1-1 for emergencies).

UF Health Shands Emergency Room / Trauma Center: For immediate medical care call 352-733-0111 or go to the emergency room at 1515 SW Archer Road, Gainesville, FL 32608; Visit the UF Health Emergency Room and Trauma Center website.

GatorWell Health Promotion Services: For prevention services focused on optimal wellbeing, including Wellness Coaching for Academic Success, visit the GatorWell website or call 352-273-4450.

Academic Resources
E-learning technical support: Contact the UF Computing Help Desk at 352-392-4357 or via e-mail at helpdesk@ufl.edu

Career Connections Center: Reitz Union Suite 1300, 352-392-1601. Career assistance and counseling services.

Library Support: Various ways to receive assistance with respect to using the libraries or finding resources.

Teaching Center: Broward Hall, 352-392-2010 or to make an appointment 352- 392-6420. General study skills and tutoring.

Writing Studio: 2215 Turlington Hall, 352-846-1138. Help brainstorming, formatting, and writing papers.
Student Complaints On-Campus: Visit the Student Honor Code and Student Conduct Code webpage for more information.

On-Line Students Complaints: Visit the Distance Learning Student Complaint Process.

In-Class Recording
Students are allowed to record video or audio of class lectures. However, the purposes for which these recordings may be used are strictly controlled. The only allowable purposes are (1) for personal educational use, (2) in connection with a complaint to the university, or (3) as evidence in, or in preparation for, a criminal or civil proceeding. All other purposes are prohibited. Specifically, students may not publish recorded lectures without the written consent of the instructor.

A “class lecture” is an educational presentation intended to inform or teach enrolled students about a particular subject, including any instructor-led discussions that form part of the presentation, and delivered by any instructor hired or appointed by the University, or by a guest instructor, as part of a University of Florida course. A class lecture does not include lab sessions, student presentations, clinical presentations such as patient history, academic exercises involving solely student participation, assessments (quizzes, tests, exams), field trips, private conversations between students in the class or between a student and the faculty or lecturer during a class session.

Publication without permission of the instructor is prohibited. To “publish” means to share, transmit, circulate, distribute, or provide access to a recording, regardless of format or medium, to another person (or persons), including but not limited to another student within the same class section. Additionally, a recording, or transcript of a recording, is considered published if it is posted on or uploaded to, in whole or in part, any media platform, including but not limited to social media, book, magazine, newspaper, leaflet, or third-party note/tutoring services. A student who publishes a recording without written consent may be subject to a civil cause of action instituted by a person injured by the publication and/or discipline under UF Regulation 4.040 Student Honor Code and Student Conduct Code.

Privacy statement

Our class sessions may be audio visually recorded for students in the class to refer back and for enrolled students who are unable to attend live. Students who participate with their camera engaged or utilize a profile image are agreeing to have their video or image recorded. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live. The chat will not be recorded or shared. As in all courses, unauthorized recording and unauthorized sharing of recorded materials is prohibited.

Services for Students with Disabilities

The Disability Resource Center coordinates the needed accommodations of students with disabilities. This includes registering disabilities, recommending academic accommodations within the classroom, accessing special adaptive computer equipment, providing interpretation services and mediating faculty-student disability related issues. Students requesting classroom accommodation must first register with the Dean of Students Office. The Dean of Students Office will provide documentation to the student who must then provide this documentation to the instructor when requesting accommodation
0001 Reid Hall, 352-392-8565, https://disability.ufl.edu/

The instructor reserves the right to make changes in the assignments and syllabus as needed. Notification will be made via E-Learning, e-mail or class announcements