

1 Latent Interacting Variable-Effects Modeling of Gut Microbiome Multi-Omics in Inflammatory Bowel

2 Disease

3 Javier. E. Munoz^{1,2}, Douglas. K. Brubaker^{1,3*}

4 ¹Weldon School of Biomedical Engineering, Purdue University, West Lafayette IN USA.

5 ²Purdue Interdisciplinary Life Science Program, Purdue University, West Lafayette IN USA.

6 ³Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette IN USA

7 * To whom correspondence should be addressed, Email: dkbrubak@purdue.edu

8

9 ABSTRACT

10 Latent Interacting Variable Effects (LIVE) modeling is a framework to integrate different types of
11 microbiome multi-omics data by combining latent variables from single-omic models into a structured meta-
12 model to determine discriminative, interacting multi-omics features driving disease status. We implemented
13 and tested LIVE modeling in publicly available metagenomics and metabolomics datasets from Crohn's
14 Disease and Ulcerative Colitis patients. Here, LIVE modeling reduced the number of feature correlations from
15 the original data set for CD and UC to tractable numbers and facilitated prioritization of biological associations
16 between microbes, metabolites, enzymes and IBD status through the application of stringent thresholds on
17 generated inferential statistics. We determined LIVE modeling confirmed previously reported IBD biomarkers
18 and uncovered potentially novel disease mechanisms in IBD. LIVE modeling makes a distinct and
19 complementary contribution to the current methods to integrate microbiome data to predict IBD status because
20 of its flexibility to adapt to different types of microbiome multi-omics data, scalability for large and small cohort
21 studies via reliance on latent variables and dimensionality reduction, and the intuitive interpretability of the
22 linear meta-model integrating -omic data types. The results of LIVE modeling and the biological relationships
23 can be represented in networks that connect local correlation structure of single omic data types with global
24 community and omic structure in the latent variable VIP scores. This model arises as novel tool that allows
25 researchers to be more selective about omic feature interaction without disrupting the structural correlation
26 framework provided by sPLS-DA interaction effects modeling. It will lead to form testable hypothesis by
27 identifying potential and unique interactions between metabolome and microbiome that must be considered for
28 future studies.

29

30 **AUTHOR SUMMARY**

31 Latent Interacting Variable Effects (LIVE) modeling integrates microbiome multiomics features by encoding
32 them in a set of latent variables (LVs) from single-omic sparse Partial Least Squares models, and then
33 combine these LVs into structured metamodel to determine the most discriminative features driving IBD. We
34 used publicly available metagenomic and metabolomics data from Crohn's Disease and Ulcerative Colitis
35 patients to develop LIVE modeling. LIVE modeling reduced data dimensionality efficiently and identified
36 statistical interactions among microbiome multi-omics data, which can be visualized as a mineable network
37 data structure. LIVE modeling confirmed features previously reported and revealed novel microbiome
38 interactions in IBD. LIVE offers a flexible framework for multi-omic modeling that may aid in interpretation of
39 complex microbiome datasets.

40

41

12 INTRODUCTION

13 Alterations in the composition and function of intestinal microbiota are a hallmark of Crohn's Disease
14 (CD) and Ulcerative Colitis (UC), two subtypes of Inflammatory Bowel Diseases (IBD) [1-4]. These alterations
15 trigger local and systemic host inflammatory responses that progressively damage the intestines over time [5-
16 6]. A hypothesis of how alterations in gut microbiota drive chronic inflammation in Crohn's Disease and
17 Ulcerative Colitis is through the gut metabolome which may act as an interface between the host and
18 microbiome [7]. However, these mechanisms of metabolome-mediated host-microbiome interactions are poorly
19 understood. A better understanding of how gut microbiome alterations drive IBDs could enable less-invasive
20 diagnostic methods for CD and UC, facilitate identification of therapeutic biomarkers, or reveal strategies to
21 prolong disease remission and reduce the likelihood of relapse [8]. As validated pathways of host-microbiome
22 interactions in IBD continue to be established, interpretable computational models of microbiome multi-omics
23 data could potentially reveal translatable insights into IBD progression and drug-resistance, data which is
24 increasingly available from large cohorts of CD and UC patients.

25 Much of the current state of the art to determine microbe-metabolite associations is focused on pairwise
26 correlation approaches, supervised machine learning, unsupervised clustering methods, and multivariate linear
27 models [9-11]. These techniques are often challenged by the dimensionality of the multi-omics data, often do
28 not account for a correlation structure within microbiomes or metabolomes and are challenged by high multiple
29 hypothesis testing burdens. Some examples of how computational workflows try to address high data
30 dimensionality include applying feature selection methods to random forest (RF) classifiers or reducing the
31 feature space with latent variable methods like principal component analysis (PCA), and partial least squares –
32 discriminant analysis (PLS-DA). Latent variable approaches provide additional advantages because of the
33 global covariance-based frameworks that reduce data dimensionality by encoding the data on latent variables
34 (LV) that weight the relative importance of the original features [13]. PLS-DA uses latent variables to establish
35 a system of linear relationships between blocks of independent and dependent variables which estimate causal
36 relationships between observed indicators and their latent variables.

37 A variant of PLS-DA, sparse PLS-DA (sPLS-DA), uses Lasso Penalization to select the most
38 discriminant features to create the LVs to classify the data [16-18]. An example of these methods was
39 presented by Priya *et al.*, 2022 in which host transcriptomic and gut microbiome data were integrated in a

70 machine learning framework that uses sparse canonical correlation analysis and Lasso penalize regression to
71 characterize the most relevant associations between gut microbiota and host genes and pathways across
72 three intestinal diseases [19]. This regularization penalty on PLS latent variables helps reduce the curse of
73 dimensionality in high-dimensional data. Critically however, if multiple -omics data types are modeled on the
74 same LVs, the Lasso penalty may be overly-restrictive and exclude consequential biological features.
75 Therefore, there is a need to balance dimensionality reduction with preserving potentially mechanistic
76 biological information in different molecular data types.

77 Here, we introduce a computational framework for causal modeling of microbiome multi-omics data
78 based upon a Partial Least Square Path Modeling (PLS-PM) framework, and which will be denoted as Latent
79 Interacting Variable-Effects (LIVE) Modeling. PLS-PM approaches incorporate the advantages of LV modeling
80 for data dimensionality reduction while also including information about the modeled system through structural
81 connections between model components, allowing the model to encode hypotheses of causality [14-15]. Here,
82 we construct PLS-PM models of microbiome composition, metabolomics, and bacterial proteomics data from a
83 cohort of 155 IBD patients and controls in which each molecular data stream is encoded as a set of latent
84 variables. We then construct a structural model in which we quantify interactions between latent variables with
85 linear model interaction effects. We show that LIVE Modeling accounts for global data covariance structure in
86 the single-omic LV's, efficiently reduces data dimensionality, and allows us to identify statistical interactions
87 between microbiome multi-omics data associated with disease onset in UC and CD.

88

39 RESULTS

40 Model Description

41 LIVE modeling integrates multi-omics data while preserving covariance structure within data modalities.
42 SPLS-DA models are built to model the relationship between individual types of molecular features and an
43 output variable. A variety of microbiome multi-omics data types, such as relative abundances of genes, taxa,
44 and metabolites, can be analyzed individually with respect to the dependent variable in their own
45 dimensionless latent variable space. By building a *bottom-up* multi-omic model from individual grassroots
46 models of single data types, this framework can incorporate several types of microbiome multi-omics data as
47 inputs, determining the most discriminative latent variables per data type. Second, the latent variables are

8 extracted from the single-omic sPLS-DA models and integrated together as main and interaction effects terms
9 in a generalized linear model, formalizing a structured relationship between the response variable and multiple
0 predictive -omic features. This allows us to evaluate data type-specific effects and determine how the
1 predictive power of data types depend on each other. Using the LV's as regression terms preserves internal
2 covariance within data type LVs and simplifies the interpretability of the multi-omic effects in the model through
3 analysis of the regression coefficients. Finally, the relationships between multi-omics data types and specific
4 features implicated by the meta-model can be studied by modeling the relationships between microbes,
5 metabolites, and enzymes via Spearman correlation analysis and PLS Variable Importance of Projection (VIP)
6 scores (Figure 1).

7 We applied our LIVE modeling framework to predict CD or UC status versus control from publicly
8 available metabolomics, and metagenomics data obtained from the PRISM cohort (the Prospective Registry in
9 IBD Study at MGH) and reported by Franzosa *et al.*, in 2019. In this study, gut metabolic profile and
10 microbiome composition in IBD were characterized by analyzing stool samples of 155 patients diagnosed with
11 Crohn's Disease (68 patients), Ulcerative Colitis (53 patients), and non-IBD control (34 patients). The collected
12 stool samples were subjected to shotgun metagenomic sequencing to determine taxonomic compositional and
13 functional potential, as well as four liquid chromatography tandem mass spectrometry (LC-MS) methods to
14 quantify composition of polar metabolites, lipids, free fatty acids, and bile acids. The datasets contained the
15 relative abundances of 201 microbiota features, 3829 metabolic features, and 2113 microbial enzyme features
16 [9]. To validate our model trained on PRISM data and determine performance metrics, we used a public
17 independent validation cohort also reported by Franzosa *et al.*, consisting of 22 control subjects from LifeLines-
18 DEEP (LLD) general population study, and 43 IBD subjects from a study of the Department of
19 Gastroenterology and Hepatology at University Medical Center Groningen (UMCG).

20

21 **Single-omic sPLS-DA are as predictive of Crohn's Disease status as naive multi-omics models**

22 Before combining single-omic latent variables (LV) from sparse Partial Least Square Discriminant Analysis
23 (sPLS-DA) via the LIVE multi-omic meta-model, we assessed the disease-status predictive power of these
24 data streams individually training on the PRISM cohort and testing on the LLD cohort for Crohn's Disease and
25 Ulcerative Colitis. The individual microbiome, metabolome and microbial enzyme sPLS-DA models were

16 significantly predictive of CD status. (Metabolome AUC = 0.959: Figure 2a and 2b, Microbiome AUC = 0.972:
17 Figure 2e and 2f, Enzymatic composition AUC = 0.939: Figure 2i and 2j). The most discriminative features per
18 each single sPLS-DA models were sorted by the PLS Variable Importance of Projection (Figure 2c, 2g, 2k) and
19 the top features are listed here: the most discriminative metabolites for CD are the flavonoids, steroidal
20 glucuronide conjugates, fatty acids and conjugates, dipeptides, terpene glycosides, tetrapyrroles and derivates,
21 benzenesulfonic acids and derivatives, triterpenoids, benzopyrans and alpha-acyloxy carbonyl compounds
22 (Figure 2d). The most discriminant microbial species are *Coprococcus catus*, *Subdoligranulum unclassified*,
23 *Alistipes shahii*, *Roseburia hominis*, *Bacteriodales bacterium ph8*, *Eubacterium ventriosum*, *Ruminococcus*
24 *obeum*, *Gordonibacter pamelaeae* (Figure 2h). Finally, the most discriminative enzymatic features are
25 1.14.13.81: Magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase, 1.1.1.35: 3-hydroxyacyl-CoA
26 dehydrogenase, 3.5.3.9: Allantoate deiminase, 3.1.4.1: Phosphodiesterase I, Saccharopine dehydrogenase
27 (NAD(+), L-lysine-forming) (Figure 2i).

28 As a computational control, we trained sPLS-DA models on a concatenated matrix that integrated all
29 data types (microbiome, metabolome, enzymes) in one dataset. Doing this for the CD patient subsets allowed
30 us to assess whether predicting disease status using multi-omics outperformed single-omic sPLS-DA models.
31 The control sPLS-DA model also was similarly predictive of CD status as the individual -omic sPLS-DA models
32 (AUC = 0.959) (Figures 2m and 2n). However, even though the control model has comparable AUC with
33 single-omic models, the concatenated control has higher balanced error rate (CD = 0.45) than single omic
34 models (Table 1). This indicated that unstructured combining of multi-omics alone does not enhance predictive
35 power.

36 sPLS-DA employs a regularization penalty when constructing LV's, incorporating different features into
37 different LV's based on the ability to discriminate between Crohn's Disease, and control status. The severity of
38 this penalty is proportional to the number of features in the original dataset, making it much higher for the multi-
39 omics sPLS-DA model compared to the single-omic sPLS-DA models. In our models for CD, 2 microbiome
40 LV's, explaining 5% and 2% variance encoded 100 and 2 features respectively, 3 metabolome LV's explaining
41 17%, 3%, and 6% variance encoded 70, 40, and 90 features respectively, and 2 microbial enzyme LV's,
42 explaining 12% and 2% variance, encoded 100, and 2 features respectively.

43 By comparison, the concatenated data control sPLS-DA model resulted in lower feature coverage in the

54 IBD microbiome multi-omics data compared to the single-omic sPLS-DA models. The control model that
55 combines all omics for CD was comprised of 2 LV's that explained 13% and 1% variance and encoded 70
56 features (0 microbes, 70 metabolites, 0 enzymes) and 1 feature (0 microbe, 1 metabolite, 0 enzymes)
57 respectively. Given the number of features that encoded the covariance, individual microbiome, metabolome,
58 and microbial enzyme sPLS-DA models show a higher combined data coverage than the control model (Table
59 2). It could be attributed to not restrictive Lasso penalties on individual models than the control model because
60 of the small number of features in individual omics data sets than omics concatenation matrix data set.

61

62 **Structured integration of multi-omics latent variables Predicting Crohn's Disease Status**

63 We sought to test whether structured integration of multi-omics data better predicted disease status than
64 single-omic sPLS-DA models or the concatenated control sPLS-DA models. Given AUC values and balanced
65 error rates, we determine the single omic sPLS-DA models can classify control patients versus CD, and
66 predictions are more accurate than concatenated control model. The most discriminative latent variables from
67 these single sPLS-DA models were integrated in a structured multi-omics model. We first examined a main-
68 effects only structured multi-omics model for predicting CD status. The form these models took was as a linear
69 model with a slope intercept coefficient and main-effects coefficients for patient scores on each latent variable
70 as features predicting disease status. A main-effects only regression model of microbiome, metabolite and
71 microbial enzyme LVs was overall predictive of Crohn's Disease status ($p = 2.2 \times 10^{-16}$). By evaluating the p-
72 values from each data-type latent variables, we determine that the 2 Metabolomics LV's emerge as the
73 strongest data type to predict CD status from control patients in comparison with microbiota and enzymatic
74 composition (Table 3).

75 We incorporated interaction effects between single-omic LV's into the structured microbiome multi-
76 omics PLS-PM framework to extract candidate microbe-metabolite-enzyme interactions driving CD.
77 Incorporating microbiome, metabolome and microbial enzyme LV interaction effects into the Crohn's Disease
78 model identified four significant main effects from metabolome LV1, LV2 and LV3, as well as microbiota LV1.
79 The model also identified significant interaction effects between microbiome LV1 - metabolome LV1, and
80 metabolite LV2 - microbial enzyme LV1. The strongest microbe-metabolite-microbial enzyme LV interaction
81 was present between metabolite LV1 - microbiota LV1 - enzymes LV1 which together separated CD patients

32 from controls (Figure 3). This indicated that omic features allocated in these microbe-metabolite-microbial
33 enzyme interacting Latent variables depends on the value of other omic variables to precisely define Crohn
34 Disease status. In other words, a metabolite in conjunction with specific microbial species and microbial
35 enzymes can determine CD status.

36

37 **LIVE Modeling Prioritizes Multi-omic Features Associated with Crohn's Disease**

38 Having demonstrated that a structured PLS-PM of microbiome multi-omics data outperforms single-
39 omic and concatenated multi-omics sPLS-DA models, we sought to extract biological insights from the
40 predictive, interacting LV's in our CD model. The most discriminant bacteria, metabolites and microbial enzyme
41 features that determine disease status were prioritized from interacting LV's by using as criteria Lasso
42 Penalization and the variable importance of projection (VIP) scores of feature loadings on LVs (VIP > 1). After
43 combining these two selection steps, correlation analysis was applied to identify microbe-metabolite-enzymes
44 triads that were synergistically predictive of Crohn's Disease status. The correlational analysis at this stage,
45 restricted to features prioritized in our PLS-PM model, was prioritized to 23,188 microbe-metabolites-microbial
46 enzymes pairs in the CD data based on the Lasso and VIP criteria. This represents a substantial decrease in
47 the multiple testing burden of the original datasets of over 1.62×10^9 microbe-metabolites-microbial enzymes
48 pairs.

49 For Crohn's Disease, analysis of these 23,188 microbe-metabolites-microbial enzymes pairs, identified
50 7,562 significant correlations (FDR q < 0.01) predictive of CD status on interacting multi-omic LV's (FDR q
51 <0.001 = 6,471 significant correlations; and FDR q < 0.0001 = 5305 significant correlations). LIVE modeling
52 framework reduces the total number of features to a tractable number (7,562 CD) and the number of pairwise
53 comparisons between features to that equivalent to a standard RNA-seq experiment. The prioritization of
54 pairwise relationships between microbes, metabolites, and enzyme can be further refined by implementing
55 more stringent thresholds on correlational coefficients, log fold change values, VIP scores, and correlation p-
56 values or by framing interpretation in terms of a particular data type. We can visualize the impact of such
57 thresholding by first constructing a microbe-metabolite-enzyme correlation network in Cytoscape [20] and then
58 creating subnetworks that meet different thresholding criteria.

9 Here, we examine the effect of refining the significant pairwise correlations (i.e. pairwise local-structure)

10 by increasing the threshold on the VIP score (i.e. global model importance) a given microbe, metabolite, or
11 enzyme needs to achieve. For the CD model, we gradually increased the VIP threshold by 1, creating
12 subnetworks of features with minimum global importance in our PLS-PM model until we reached a threshold
13 that eliminated all bacterial nodes from the correlation network. At VIP score greater than 5 for Crohn's
14 Disease features, the network contained 72 nodes that include 44 metabolites, 27 microbial enzymes, and 1
15 bacterium, with a VIP threshold greater than 6 removing all bacteria. At our identified VIP threshold of 5, our
16 model predicted *Coprococcus cactus* as one of the most discriminating bacterial features between CD and
17 control samples, and Sphingolipids and Leucyl aminopeptidase as upregulated determinant metabolite and
18 microbial enzyme features in Crohn's Disease (Figure 4).

19

20 **Single-omic sPLS-DA are as predictive of Ulcerative Colitis status as naive multi-omics models**

21 For UC, the individual microbiome, metabolome, and microbial enzyme sPLS-DA models were
22 significantly predictive of UC status (metabolome AUC = 0.932 Figure 5a and 5b, microbiome AUC = 0.969:
23 Figure 5e and 5f, enzymatic composition AUC = 0.922: Figure 5i and 5j). Following LIVE modeling workflow,
24 the most discriminative features of UC status were sorted by the PLS Variable Importance of Projection (Figure
25 5c, 5g, 5k) and the top features per each single sPLS-DA model were identified. The most discriminative
26 metabolite features for UC are the diphenylmethanes, diterpenoids, pyridoxamines, tetrapyrroles and
27 derivatives, cardiolipins, pyridinecarboxilic acids and monoalkyglycerophosphates (Figure 5d). The most
28 relevant microbial species to classify UC patients are: *Gordonibacter pamelaeae*, *Roseburia hominis*,
29 *Eubacterium rectale*, *Eubacterium hallii*, *Eubacterium ramulus*, *Alistipes shahii*, *Subdoligranulum unclassified*.
30 (Figure 5h). The most discriminative enzymatic features of UC status are 2.7.1.66: Undecaprenol kinase,
31 1.5.3.1: Sarcosine oxidase, 2.7.1.66: Undecaprenol kinase, 1.5.3.1: Sarcosine oxidase, 2.4.1.4: Amylosucrase,
32 3.1.2.14: Oleoyl-[acyl-carrier-protein] hydrolase, 4.2.1.52: Transferred entry: 4.3.3.7, and 1.1.1.31: 3-
33 hydroxyisobutyrate dehydrogenase (Figure 5l).

34 By comparing single-omics sPLS-DA models with the concatenated matrix as a computational control,
35 we determined for UC status, similarly to the CD study case, that the control model has comparable AUC with
36 single-omic models (AUC = 0.959) (Figures 5m and 5n), but, has higher balanced error rate (UC = 0.5) than

37 single omic models (Table 4). This supports that unstructured combining of multi-omics alone does not
38 necessarily enhance predictive power.

39 The regularization penalty from sPLS-DA is proportional to the number of features in the original data
40 set. Similar our CD study case, the penalty applied to the number of features for the multiomics sPLS-DA
41 model is higher than single-omic sPLS-DA models, causing more error during predictions. For our UC sPLS-
42 DA models, 2 microbiome LV's, explaining 5% and 2% variance, encoded 80 and 1 features respectively, 2
43 metabolome LV's, explaining 10%, and 3% variance, encoded 100 and 100 features respectively, and 2
44 microbial enzyme LV's, explaining 13% and 6% variance, encoded 100, and 90 features respectively.

45 By comparison, the concatenated data control sPLS-DA model resulted in lower feature coverage in the
46 UC microbiome multi-omics data compared to the single-omic sPLS-DA models. The control sPLS-DA model
47 for predicting UC status was comprised of 2 LV's, explaining 8% and 4% variance, and encoded 100 (0
48 microbes, 100 metabolites, 0 enzymes) and 1 feature (0 microbe, 1 metabolite, 0 enzymes) respectively (Table
49 5). Given the number of features that encoded the covariance, individual microbiome, metabolome, and
50 microbial enzyme sPLS-DA models for UC show a higher coverage than the control model. Like the CD
51 analysis, it could be attributed to not restrictive Lasso penalties on individual models than the control model
52 because of the small number of features in individual omics data sets than omics concatenation matrix data
53 set.

54

55 **Structured integration of multi-omics latent variables Predicting Ulcerative Colitis Status**

56 Given that single omic sPLS-DA models can classify control patients versus UC, and predictions are more
57 accurate than concatenated control model, the structured integration of multi-omics data was implemented
58 similar CD case study. The most discriminative latent variables of UC status from these single sPLS-DA
59 models were integrated in a structured multi-omics model. Following the steps in LIVE modeling workflow, we
60 first examined a main-effects only multi-omics model for predicting UC status, and we determined that a main-
61 effects only regression model of microbiome, metabolite and microbial enzyme LVs was overall predictive of
62 Ulcerative Colitis status ($p = 2.2 \text{ e-16}$). By evaluating the p-values from each data-type latent variables, we
63 determine that the 2 Metabolomics LV's emerge as the strongest data type to predict CD status from control
64 patients in comparison with microbiota and enzymatic composition (Table 6).

55 We incorporated interaction effects between single-omic LV's into the structured multi-omics model to
56 extract candidate microbe-metabolite-enzyme interactions driving UC. Our LV's interaction effects model for
57 UC identified four significant main effects from metabolome LV1 and LV2, microbiota LV1, and microbial
58 enzyme LV1. The model also identified significant interaction effects between metabolites LV2 - microbiota
59 LV1, microbiome LV1 -metabolome LV2, and microbiome LV1 - microbial enzyme LV1. The strongest microbe-
60 metabolite-microbial enzyme LV interaction was present between metabolite LV1 - microbiota LV1 - enzymes
61 LV1 which together separated UC patients from controls (Figure 6). This indicated that the omic features
62 projected in the strongest microbe-metabolite-microbial enzyme interacting Latent Variables depends on the
63 value of other omic variables to precisely define Ulcerative Colitis status.

74

75 **LIVE Modeling Prioritizes Multi-omic Features Associated with Ulcerative Colitis**

76 The most discriminant bacteria, metabolites and enzyme features that determine Ulcerative Colitis
77 disease status were prioritized from the predictive and interacting LV's to extract biological insights, by using
78 as criteria Lasso Penalization and the variable importance of projection (VIP) scores of feature loadings on LVs
79 (VIP > 1). Like CD study case, by combining these two selection steps, and applying correlation analysis
80 restricted to features prioritized in our PLS-PM model, the number of microbe-metabolites-microbial enzymes
81 pairs for UC model was 21,514, a decrease in the feature space by 5 orders of magnitude from 1.62×10^9
82 microbe-metabolites-microbial enzymes pairs in the original data. Analysis of these 21,514 microbe-
83 metabolites-microbial enzymes pairs identified 2,271 significant correlations (FDR q < 0.01) (S5b Table: FDR q
84 <0.001 = 935 significant pairs; FDR q < 0.0001 = 445 significant pairs) predictive of UC status on interacting
85 LV's.

86 In the microbe-metabolite-enzyme network for Ulcerative Colitis, VIP score was constrained greater
87 than 4 and the model identified 69 nodes that include 39 metabolites, 27 microbial enzymes, and 3 bacteria,
88 with a VIP score greater than 5 removing all bacterial features. Our UC model predicted *Eubacterium rectale*,
89 *Gordonibacter parmeliaeae* and *Roseburia hominis* as the most relevant bacterial features and showed several
90 metabolites were specifically associated with each bacterium. For instance, the Tetrapyrroles and derivatives
91 category is described as a downregulated metabolite associated with *Eubacterium rectale*. Fatty acid esters,
92 Diphenylmethanes and Pyridoxamines as downregulated metabolites associated with *Gordonibacter*

3 *pamelaeae*, and Tetrapyrroles and derivatives, Fatty acid esters, and Pyridoxamines as downregulated
4 metabolites associated with *Roseburia hominis* and *Gordonibacter pamelaeae* in Ulcerative colitis (Figure 7).

5

6 DISCUSSION

7 Our structural LIVE (Latent Interacting Variable Effects) modeling framework not only predicts features
8 that are in concordance with previous studies, but also estimates novel omic associations that must be studied
9 *in vitro* and *in vivo* models. This framework uncovers more feature interactions that increase the likelihood to
10 establish testable hypothesis and studying biological mechanisms in IBD. LIVE modeling is a microbiome
11 multi-omics PLS-PM framework reduces the total number of features to tractable numbers (7,562 CD and
12 2,271 UC FDR $q < 0.01$) and the number of pairwise comparisons between features to that equivalent to a
13 standard RNA-seq experiment. This represents a substantial decrease in the multiple testing burdens of the
14 original datasets. Additionally, given the greater number of significant correlations between CD microbiome
15 multi-omic features than UC features reinforces the stratification of IBD into these two molecularly subtypes in
16 addition to their differences in clinical presentation.

7 The prioritization of pairwise relationships between microbes, metabolites, and enzyme can be further
8 refined in LIVE modeling by implementing more stringent thresholds on correlational coefficients, log fold
9 change values, VIP scores, and correlation p-values or by framing interpretation in terms of a particular data
10 type. We can visualize the impact of such thresholding by first constructing a microbe-metabolite-enzyme
11 correlation network in Cytoscape and then creating subnetworks that meet different thresholding criteria. LIVE
12 modeling workflow examines the effect of refining the significant pairwise correlations (i.e. pairwise local-
13 structure) by increasing the threshold on the VIP score (i.e. global model importance) a given microbe,
14 metabolite, or enzyme needs to achieve. It allows us to identify the most discriminative and interacting
15 metabolites, microbial species and enzymes correlations for CD and UC.

6 This framework identified features associated with Inflammatory Bowel Diseases that have been
7 reported previously by different studies [9-10]. Some of these features were reported by the original publication
8 we drew data from, Franzosa *et al.* Here, the authors analyzed this data using other methods such as pairwise
9 correlational analysis and random forest classifiers. In total, forty-four metabolites that were downregulated in
10 Crohn's Disease were predicted by our model at one of the highest levels of restriction (VIP > 5). Among the

11 most important metabolite features are flavonoids, steroidal glucuronide conjugates, fatty acids and
12 conjugates, dipeptides, terpene glycosides, tetrapyrroles and derivates, benzenesulfonic acids and derivatives,
13 triterpenoids, benzopyrans and alpha-acyloxy carbonyl compounds [9, 21-23].

14 The Sphingolipids category was also predicted by our model as a determinant metabolite that was up
15 regulated in Crohn's Disease and reported by Franzosa *et al* and other studies [24]. From thirty-seven
16 determinant microbiota species predicted by our model at the level of restriction of VIP >1 for Crohn's Disease,
17 thirteen were also reported by Franzosa *et al.* as representative microbial species. If the level of restriction is
18 more stringent VIP>5, *Coprococcus catus* was predicted as the most determinant microbiota species and it
19 also has been reported by Franzosa *et al.*, but not prioritized as canonical feature, as well as in other studies
20 [25-27]. Fold changes calculated from the relative abundance values of 37 predicted microbial species are
21 about zero which is associated with the loss of species diversity, a characteristic phenomenon in IBD. From
22 twenty-seven microbial enzymes predicted by our model at VIP > 5 for Crohn's Disease, propionyl-CoA
23 carboxylase, precorrin-2 dehydrogenase, leucyl aminopeptidase, NAD (+) synthase (glutamine-hydrolyzing)
24 enzymes were also reported by Franzosa *et al.* as representative microbial enzymes.

25 Though bile acids and dicarboxylic acids were reported by Franzosa *et al.*, these features were not
26 predicted by our model at three distinctive levels of restrictions: VIP >5 (72 features), VIP > 1 (185 features),
27 and q < 0.01 (7,562 significant features). We determined that these features were excluded in the single omic
28 SPLS-DA models when individual omics classify between CD and control patients, it means that bile acids and
29 dicarboxylic acids were not the most discriminative metabolites features of CD status. Several studies
30 associated Bile acids with Inflammation chronic diarrhea which is connected to bile acid malabsorption in CD
31 patients.

32 Our workflow generates several inferential statistics that facilitate prioritization of biological associations
33 between microbes, metabolites, enzymes, and IBD status. In the preset IBD dataset, by using VIP, metabolite,
34 microbiota, and enzyme interactions synergistically predict disease status can be visualized at different levels
35 of restriction. These statistics provide flexibility in the to be more stringent with feature selection. For instance,
36 in the Crohn's Disease model, one of the highest levels of restriction was VIP > 5, and it estimated
37 *Coprococcus catus* as the most determinant microbial species correlated with 44 metabolites and 27 microbial
38 enzymes. Here, *Coprococcus catus* becomes the central focus of a network of metabolite and microbial

19 enzyme interactions in Crohn's Disease. Nevertheless, this framework can be even more selective if it is
50 needed. If a higher resolution of feature interactions between metabolites and microbial enzymes with
51 *Coprococcus catus* is required, the level of restriction can be incremented in the VIP regarding metabolites and
52 microbial enzyme features. At VIP >11, our model estimate *Coprococcus catus* interact with 1.1.1.35: 3-
53 hydroxyacyl-CoA dehydrogenase, 1.14.13.81: Magnesium-protoporphyrin IX monomethyl ester (oxidative)
54 cyclase, and 1.2.99.3: aldehyde dehydrogenase; and the following metabolites: triterpenoids, benzenesulfonic
55 acids and derivatives, tetrapyrroles and derivatives, terpene glycosides, dipeptides, fatty acids and conjugates,
56 steroid glucuronide conjugates, and flavonoids (Figure 4c). This structural omic interaction framework that
57 keeps the correlation structure within metabolites, microbiota, metabolome, and enzymes provides high-
58 resolution property to evaluate omic interactions based on research needs.

59 A potential application of the LIVE Modeling framework's identification of disease-driving molecules and
60 taxa is to leverage features that discriminate disease status from control to identify biomarkers for drug
61 response, diagnosis, and therapeutics in IBD. In 2021, Lee *et al.* reported features associated with anti-
62 cytokine and anti-integrin therapies in this same cohort we analyzed in this study [11].. We compared the
63 reported drug response features in Lee *et al.* with disease associated features selected by our structural
64 interaction model. *Gordonibacter parmelaeae* was the main microbial species they identified as a determinant
65 in drug response for Crohn's Disease and Ulcerative Colitis, and it also was predicted by our structural
66 correlational model for CD (VIP>1) and UC (VIP>4). Particularly in Crohn's Disease, common microbial
67 enzyme features predicted by our model and reported as drug response associated features were precorrin-4
68 C (11)-methyltransferase, precorrin-6B C (5,15) -methyltransferase (decarboxylating), glucosamine-1-
69 phosphate N-acetyltransferase, and histidinol-phosphatase. For Ulcerative Colitis, the metabolite acyl
70 carnitines, and the microbial enzymes arabinogalactan endo-beta-1,4-galactanase, and phosphoglucomutase
71 (alpha-D-glucose-1,6-bisphosphate-dependent) were features associated with therapy response and disease
72 status classification (Table 7). In sum, there are several commonalities between omics features associated
73 with disease status in LIVE modeling and secondary association with drug response in IBD [9,11].

74 LIVE modeling used the sample projections of the sPLS-DA model to build an interaction effects linear
75 regression and predict omic features that are associated with disease and drug response in Inflammatory
76 Bowel Diseases. These predictions were in good agreement with previous studies that reported disease and

17 drug response features, which applying other statistical integrative methods. For instance, a combination of
18 multivariable linear models and Random Forest classifiers with feature selection were applied to find
19 associations between metabolome and microbiome features with clinical outcomes before and after exposure
30 and integrate multi-omics and predict disease status and treatment response, respectively. Likewise, Metwaly
31 *et al.* described linkages between IBD and sulfur metabolism using a Generalized Canonical Correlation
32 Analysis (DIABLO-MixOmics) to integrated microbiota and metabolite profiles [24]. This last approach uses
33 another correlation structure established by the weighted linear composites which vary to maximize the
34 correlation between variables by overlapping their distributions [28]. Unlike LIVE Modeling however, it does not
35 consider a regression modeling between sample projections of independent variables, and dependent
36 variables.

37 Our LIVE Modeling framework makes a distinct and complementary contribution to these integrative
38 methods that aim to predict metabolite and microbial species features that can predict IBD status. What
39 distinguishes our approach is the flexibility to adapt to different types of microbiome multi-omics data,
40 scalability for large and small cohort studies via reliance on latent variables and dimensionality reduction, and
41 the intuitive interpretability of the linear meta-model integrating -omic data types. The results of LIVE modeling
42 and the biological relationships can be represented in networks that connect local correlation structure of single
43 omic data types with global community and omic structure in the latent variable VIP scores. This model arises
44 as novel tool that allows researchers to be more selective about omic feature interaction without disrupting the
45 structural correlation framework provided by sPLS-DA interaction effects modeling. It will lead to form testable
46 hypothesis by identifying potential and unique interactions between metabolome and microbiome that must be
47 considered for future studies.

48

49

50

51

52

53

54

15

16

17

18

19

20

21

22

23

MATERIALS AND METHODS

Preprocessing the Data

Publicly available metabolomics, and taxonomic and functional metagenomic data was obtained from 155 patients diagnosed with Crohn Disease, Ulcerative Colitis and non IBD patients (control) [9].. Relative Abundance profiles were log-transformed to variance stabilize the data with pseudo count 1 for zero values. This data set includes 201 microbiota features, 3829 metabolic features, and 2113 microbial enzyme features.

29

Data Integration and Feature selection by sPLS-DA

We trained a sparse Partial Least Square – Discriminant Analysis (sPLS-DA) model on each single-omic data to predict the disease status using MixOmics R Package(17)(18). We used the function *tune.splsda* to select the optimal number of variables and components, and AUC plots were obtained with *auroc* function. Then, loadings, variance importance of projection (VIPs) and coefficients were exported from MixOmics.

25

Multi-omics modeling with sample projection extracted from latent variables

Sample projections on the single-omic latent variables inferred from sPLS-DA models were used to train a generalized linear model with interaction effects terms. The alr4 R Package with the function *lm* were implemented to training the linear model. The main effects were patient projections on microbiome or metabolome, or microbial enzyme composition LV's. Interactions effects were coded for each pair of microbiome, metabolome, and enzymatic LV's. Significant interaction effects indicated that the predictive

32 power of the bacterial loadings on a particular microbiome LV was conditioned upon the loadings of
33 metabolites and enzymes on a particular metabolome and enzymatic LV.

34

35 ***Feature Selection from Lasso Penalizations and Variance Importance Projection.***

36 From sPLS-DA models which identify the most predictive feature that classify disease status, we identified
37 microbe, metabolite and enzymes features on significant interacting Latent variables with significant variable
38 importance of projection (VIP) scores. The Lasso Penalization Formula is calculated by the following equation:

$$\text{minimize} \left(SSE + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

39

40 Variance Importance in Projection (VIP) scores for features are calculated using the regression coefficient b ,
41 weight vector w_j , and score vector t_j as given in the following equation:

$$VIP = \sqrt{n \frac{\sum_{j=1}^a b_j^2 t_j^T t_j \left(\frac{w_{kj}}{\|w_j\|} \right)^2}{\sum_{j=1}^a b_j^2 t_j^T t_j}}$$

42

43 By considering Lasso Penalization and VIP greater than 1, we selected the most predictive and significant
44 metabolic, microbiota and enzymes features that synergistically determine IBD status, for prioritized microbe-
45 metabolite-enzyme correlation analysis. Correlational values were determined by Spearman correlational
46 analysis. Duplicate correlational values were removed and same type omic data type interactions were
47 removed. Then, p-value and false discovery rate were calculated from these omic matrices, and most relevant
48 features were determined by using thresholds such as q-value < 0.01, and q-value < 0.001.

49

50 ***Building Network Plots***

51 Cytoscape software was used to visualize metabolites, microbiota, and enzyme interaction networks.
52 To build these network plots, a node and edge files were used which containing the name and type of the
53 features, log fold change and the VIP score, and correlation values, p value and false discovery rate,

54 respectively. Different levels of restriction can be applied to visualize interaction network by filtering the data by
55 VIP score, fold change, correlation, or other metadata.
56 **Acknowledgments:** This work is based upon efforts supported by the EMBRIO Institute, contract #2120200, a
57 National Science Foundation (NSF) Biology Integration Institute. The views and conclusions contained herein
58 are those of the authors and should not be interpreted as representing the official policies, either expressed or
59 implied, of NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute
60 reprints for governmental purposes notwithstanding any copyright annotation therein.

61

52 **References:**

- 53 1. Ramos GP, Papadakis KA. Mechanisms of Disease: Inflammatory Bowel Diseases. Mayo Clin Proc. 2019;94(1):155–65.
- 54 2. de Mattos BRR, Garcia MPG, Nogueira JB, Paiatto LN, Albuquerque CG, Souza CL, et al. Inflammatory Bowel Disease: An Overview of Immune Mechanisms and Biological Treatments. *Mediators Inflamm.* 2015; 2015:1–11.
- 55 3. Schirmer M, Garner A, Vlamicis H, Xavier RJ. Microbial genes and pathways in inflammatory bowel disease. *Nat Rev Microbiol.* 2019;17(8):497–511.
- 56 4. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature.* 2020; 27;578(7796):527–39.
- 57 5. Schirmer M, Smeekens SP, Vlamicis H, Jaeger M, Oosting M, Franzosa EA, et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell.* 2016 Nov;167(4):1125-1136.e8.
- 58 6. Glassner KL, Abraham BP, Quigley EMM. The microbiome and inflammatory bowel disease. *J Allergy Clin Immunol.* 2020;145(1):16–27.
- 59 7. Karpinska-Leydier, K., Amirthalingam, J., Alshoaiikh, K., Iroshani Jayarathna, A., Salibindla, D., Paidi, G., and Ergin, H. E. Correlation Between the Gut Microbiome and Immunotherapy Response in Inflammatory Bowel Disease: A Systematic Review of the Literature. *Cureus* (2021), 13(8):e16808.
- 60 8. Plichta DR, Graham DB, Subramanian S, Xavier RJ. Therapeutic Opportunities in Inflammatory Bowel Disease: Mechanistic Dissection of Host-Microbiome Relationships. *Cell.* 2019;178(5):1041–56.
- 61 9. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol.* 2019 Feb;4(2):293–305.
- 62 10. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature.* 2019. 569(7758):655-62.
- 63 11. Lee JWJ, Plichta D, Hogstrom L, Borren NZ, Lau H, Gregory SM, et al. Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease. *Cell Host Microbe.* 2021, 29(8), 1294–1304.e4.
- 64 12. Lin E, Lane H-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark Res.* 2017;5(1):2.

13. Lee LC, Liou C-Y, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *The Analyst*. 2018;143(15):3526–39.

14. Tenenhaus M, Vinzi VE, Chatelin Y-M, Lauro C. PLS path modeling. *Comput Stat Data Anal*. 2005; 1;48(1):159–205.

15. Sanchez G. Sanchez, G. (2013) PLS Path Modeling with R, Trowchez Editions. Berkeley, 2013. Trowchez Editions. 2013. Available from: <http://www.gastonsanchez.com/PLS Path Modeling with R.pdf>

16. Lê Cao K-A, Martin PG, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*. 2009; 26;10(1):34.

17. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*. 2011; 22;12(1):253.

18. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;13(11): e1005752.

19. Priya, S., Burns, M.B., Ward, T. *et al.* Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat Microbiol* 7.2022; 780–795.

20. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003; 1;13(11):2498–504.

21. Vezza T, Rodríguez-Nogales A, Algieri F, Utrilla MP, Rodriguez-Cabezas ME, Galvez J. Flavonoids in Inflammatory Bowel Disease: A Review. *Nutrients*. 2016; 9;8(4):211.

22. Neurath MF. Host–microbiota interactions in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol*. 2020;17(2):76–7.

23. Huang X, Chen Z, Li M, Zhang Y, Xu S, Huang H, *et al.* Herbal pair Huangqin-Baishao: mechanisms underlying inflammatory bowel disease by combined system pharmacology and cell experiment approach. *BMC Complement Med Ther*. 2020;20(1):292.

24. Brown EM, Ke X, Hitchcock D, Jeanfavre S, Avila-Pacheco J, Nakata T, *et al.* Bacteroides-Derived Sphingolipids Are Critical for Maintaining Intestinal Homeostasis and Symbiosis. *Cell Host Microbe*. 2019;25(5):668-680.e7.

17 25. Metwaly A, Dunkel A, Waldschmitt N, Raj ACD, Lagkouvardos I, Corraliza AM, *et al.* Integrated microbiota
18 and metabolite profiles link Crohn's disease to sulfur metabolism. *Nat Commun.* 2020;11(1):4322.

19 26. Sankarasubramanian J, Ahmad R, Avuthu N, Singh AB, Guda C. Gut Microbiota and Metabolic Specificity
20 in Ulcerative Colitis and Crohn's Disease. *Front Med.* 2020; 7:606298.

21 27. van der Giessen J, Binyamin D, Belogolovski A, Frishman S, Tenenbaum-Gavish K, Hadar E, *et al.*
22 Modulation of cytokine patterns and microbiome during pregnancy in IBD. *Gut.* 2020;69(3):473–86.

23 28. Sørensen M, Kanatsoulis CI, Sidiropoulos ND. Generalized Canonical Correlation Analysis: A Subspace
24 Intersection Approach. *IEEE Trans Signal Process.* 2021; 69:2452–67.

25

26

27

28

29

30

31

32

33

34

35

36

37

38 **FIGURE CAPTIONS**

39 **Figure 1.- Latent Interacting Variable Effects (LIVE) Modeling Workflow** Multi-Omics Microbiome Data is
40 encoded in a set of Sparse Partial Least Square Models. Then, structured model combines discriminative
41 latent variable per each single-omics sPLS-DA models. Next, significant discriminative and predictive omic
42 features are thresholding by VIP, correlational analysis, and q-values to build meta model networks.
43

44 **Figure 2.- sPLS-DA model performance and variance importance of projections (VIP) for CD** (a) sPLS-
45 DA sample scores plot for metabolomics data. (b) ROC curve for metabolomics sPLS-DA model. (c) VIP score
46 histogram for metabolites. (d) Top 14 metabolites by VIP score. (e) sPLS-DA sample scores plot for microbiota
47 composition data. (f) ROC curve for microbiota sPLS-DA model. (g) VIP scores histogram for microbiota. (h)
48 Top 10 microbial taxa VIP scores. (i) sPLS-DA sample scores plot for enzymes data. (j) ROC curve for
49 enzymes sPLS-DA model. (k) VIP score histogram for enzymes. (l) Top 8 enzymes by VIP score. (m) sPLS-DA
50 sample scores plot for concatenation control multi-omic model. (n) ROC curve for concatenation control model.
51 (o) VIP score histogram for concatenation control model. (p) Top 12 feature VIP scores for the control model.
52

53 **Figure 3.- Main and Interaction Effects in a structured model regression for CD** (a) LV main effects model
54 p-values. (b) LV Interaction effects p-values. (c) Scores plot of samples on enzymes LV1 and metabolites LV1.
55 (d) Scores plot of samples on enzymes LV1 and microbiota LV1. (e) Scores plot of samples on microbiota LV1
56 and metabolites LV1.
57

58 **Figure 4.- Meta Model Networks and Subnetworks for CD in Cytoscape** (a) Meta-Model networks VIP > 1.
59 (b) Top 20% highest VIP per each omic data type. (c) VIP > 11 and *Coprococcus catus*. (d) Top 2 highest VIP.
60 Edges reflect Spearman correlation coefficient.
61

62 **Figure 5.- sPLS-DA model performance and variance importance of projections (VIP) for UC** (a) sPLS-
63 DA sample scores plot for metabolomics data. (b) ROC curve for metabolomics sPLS-DA model. (c) VIP score
64 histogram for metabolites. (d) Top 14 metabolites by VIP score. (e) sPLS-DA sample scores plot for microbiota
65 composition data. (f) ROC curve for microbiota sPLS-DA model. (g) VIP scores histogram for microbiota. (h)
66 Top 10 microbial taxa VIP scores. (i) sPLS-DA sample scores plot for enzymes data. (j) ROC curve for
67 enzymes sPLS-DA model. (k) VIP score histogram for enzymes. (l) Top 8 enzymes by VIP score. (m) sPLS-DA
68 sample scores plot for concatenation control multi-omic model. (n) ROC curve for concatenation control model.
69 (o) VIP score histogram for concatenation control model. (p) Top 11 feature VIP scores for the control model.
70

71 **Figure 6.- Main and Interaction Effects in a structured model regression for UC**. (a) LV main effects
72 model p-values. (b) LV Interaction effects p-values. (c) Scores plot of samples on microbiota LV1 and
73 metabolites LV1. (d) Scores plot of samples on enzymes LV1 and microbiota LV1. (e) Scores plot of samples
74 on enzyme LV1 and metabolites LV1.
75

76 **Figure 7.- Meta Model Networks and Subnetworks for UC in Cytoscape** (a) Networks VIP > 1. (b) Top 20%
77 highest VIP per each omic data type. (c) VIP > 7.5 and Top 3 microbial species. (d) Top 2 highest VIP. Edges
78 reflect Spearman correlation coefficient.
79

80

81

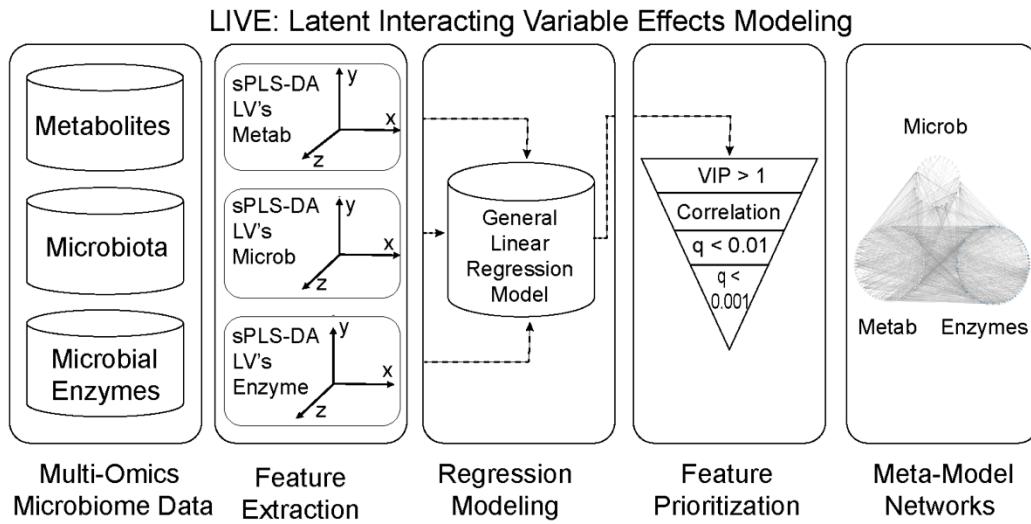
82

33 **FIGURES**

34

35

36



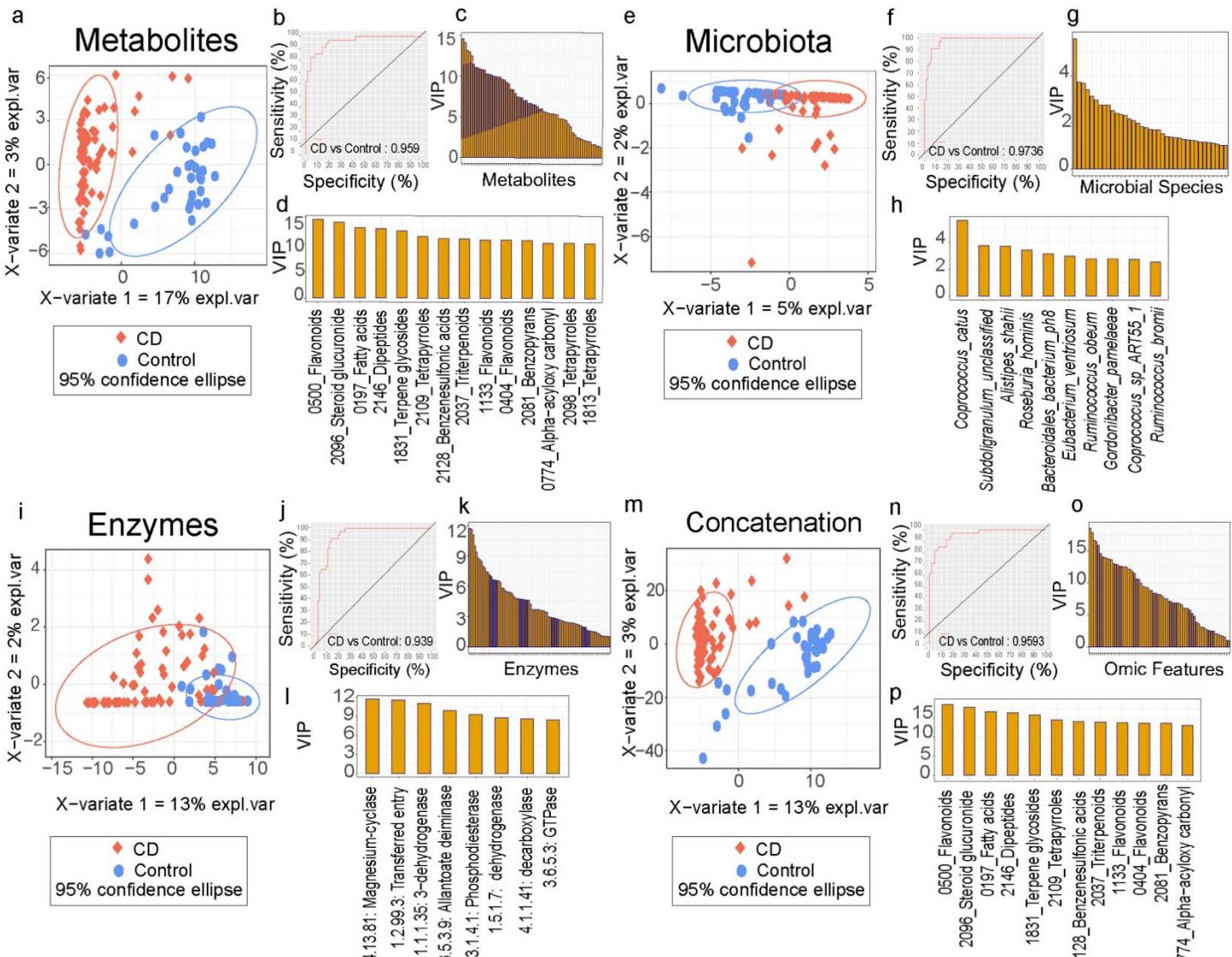
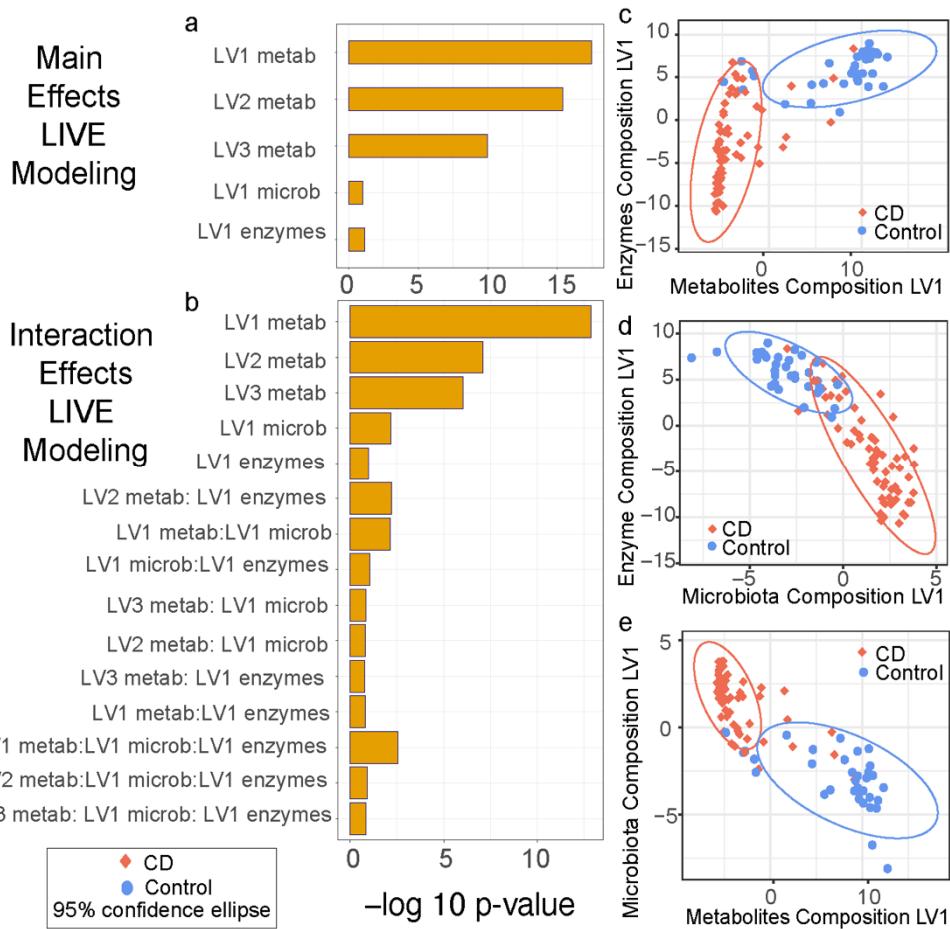
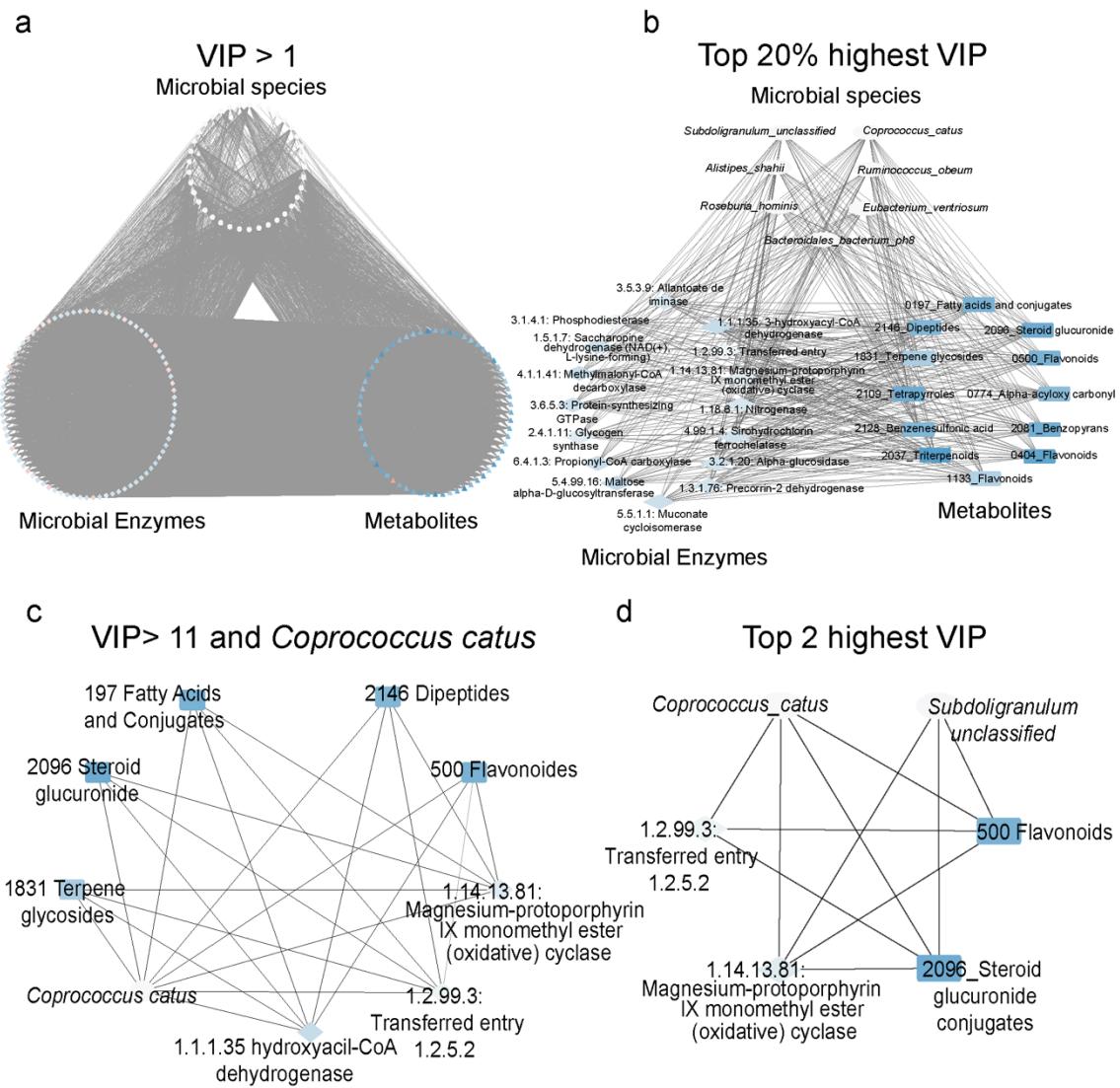


Figure 2.- sPLS-DA model performance and variance importance of projections (VIP) for CD (a) sPLS-DA sample scores plot for metabolomics data. (b) ROC curve for metabolomics sPLS-DA model. (c) VIP score histogram for metabolites. (d) Top 14 metabolites by VIP score. (e) sPLS-DA sample scores plot for microbiota composition data. (f) ROC curve for microbiota sPLS-DA model. (g) VIP scores histogram for microbiota. (h) Top 10 microbial taxa VIP scores. (i) sPLS-DA sample scores plot for enzymes data. (j) ROC curve for enzymes sPLS-DA model. (k) VIP score histogram for enzymes. (l) Top 8 enzymes by VIP score. (m) sPLS-DA sample scores plot for concatenation control multi-omic model. (n) ROC curve for concatenation control model. (o) VIP score histogram for concatenation control model. (p) Top 12 feature VIP scores for the control model.



)2

)3 **Figure 3.- Main and Interaction Effects in a structured model regression for CD** (a) LV main effects model
)4 p-values. (b) LV Interaction effects p-values. (c) Scores plot of samples on enzymes LV1 and metabolites LV1.
)5 (d) Scores plot of samples on enzymes LV1 and microbiota LV1. (e) Scores plot of samples on microbiota LV1
)6 and metabolites LV1.
)7



8

9 **Figure 4.- Meta Model Networks and Subnetworks for CD in Cytoscape** (a) Meta-Model networks VIP > 1.
10 (b) Top 20% highest VIP per each omic data type. (c) VIP > 11 and *Coprococcus catus*. (d) Top 2 highest VIP.
11 Edges reflect Spearman correlation coefficient.
12

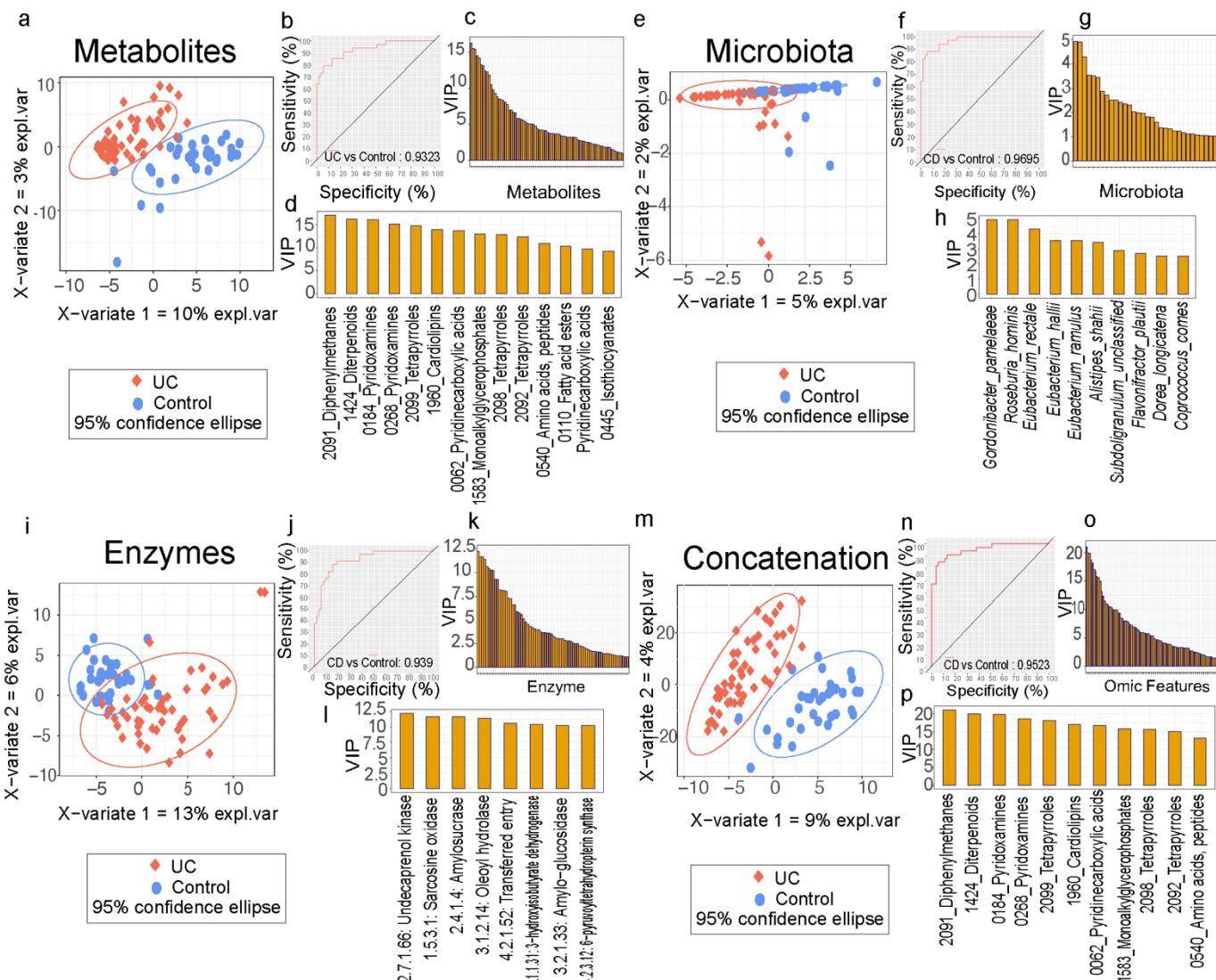
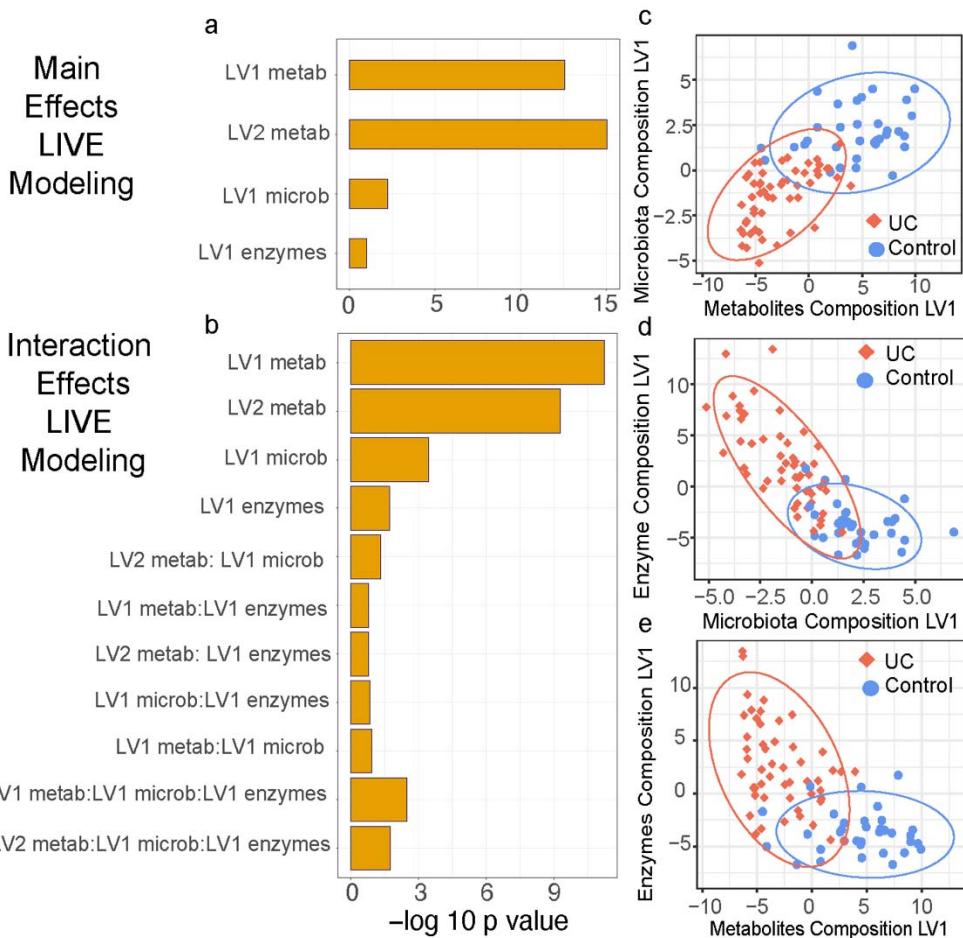
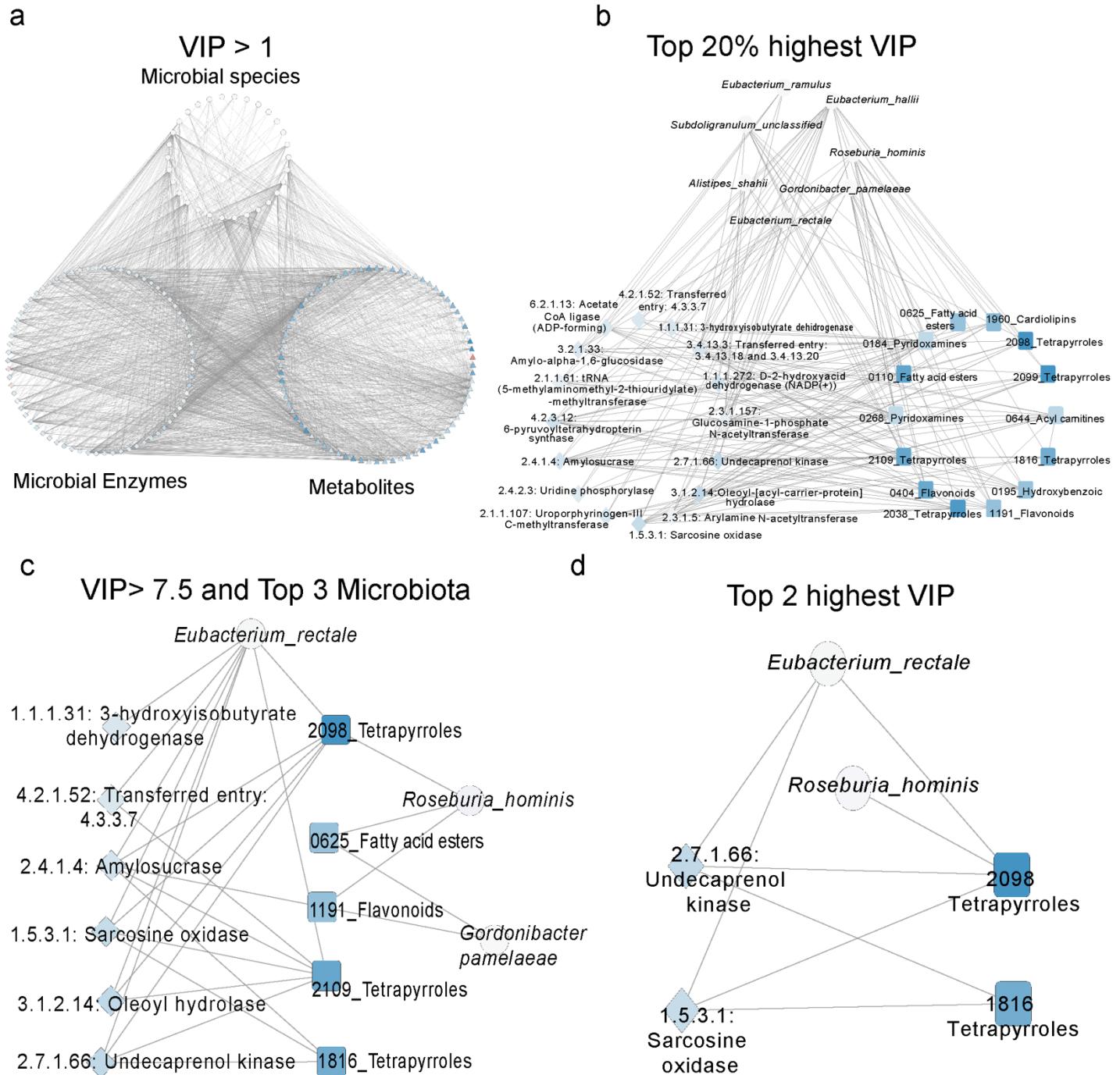


Figure 5.- sPLS-DA model performance and variance importance of projections (VIP) for UC (a) sPLS-DA sample scores plot for metabolomics data. (b) ROC curve for metabolomics sPLS-DA model. (c) VIP score histogram for metabolites. (d) Top 14 metabolites by VIP score. (e) sPLS-DA sample scores plot for microbiota composition data. (f) ROC curve for microbiota sPLS-DA model. (g) VIP scores histogram for microbiota. (h) Top 10 microbial taxa VIP scores. (i) sPLS-DA sample scores plot for enzymes data. (j) ROC curve for enzymes sPLS-DA model. (k) VIP score histogram for enzymes. (l) Top 8 enzymes by VIP score. (m) sPLS-DA sample scores plot for concatenation control multi-omic model. (n) ROC curve for concatenation control model. (o) VIP score histogram for concatenation control model. (p) Top 11 feature VIP scores for the control model.





34

35 **Figure 7.- Meta Model Networks and Subnetworks for UC in Cytoscape** (a) Networks VIP > 1. (b) Top 20%
36 highest VIP per each omic data type. (c) VIP > 7.5 and Top 3 microbial species. (d) Top 2 highest VIP.
37 Edges reflect Spearman correlation coefficient.

38

39

10 **TABLES**

11 **Table 1.- AUC and Balanced Error Rate comparison from single-omics sPLS-DA models for CD**

| Performance Metric | Concatenation Control | Metabolome | Microbiome | Enzymatic Composition |
|---------------------|-----------------------|------------|------------|-----------------------|
| AUC | 0.959 | 0.959 | 0.9736 | 0.939 |
| Balanced Error Rate | 0.45 | 0.43 | 0.24 | 0.20 |

12 **Table 2.- Single-omic sPLS-DA models show higher coverage than concatenation control model in CD.**

| Latent Variable | Concatenation Control | | Metabolome | | Microbiome | | Enzymatic Composition | |
|-----------------|-----------------------|--------------------|------------------|--------------------|------------------|--------------------|-----------------------|--------------------|
| | Encoded Features | Explained Variance | Encoded Features | Explained Variance | Encoded Features | Explained Variance | Encoded Features | Explained Variance |
| LV1 | 70 | 13% | 70 | 17% | 100 | 5% | 100 | 12% |
| LV2 | 1 | 3% | 40 | 3% | 2 | 2% | 2 | 2% |
| LV3 | 0 | 0% | 90 | 6% | 0 | 0% | 0 | 0% |
| Total | 71 | 16% | 200 | 20% | 102 | 7% | 102 | 14% |

15 **Table 3.- LIVE Modeling Outcome Table with Coefficients and p-values for CD.**

| Crohn Disease Coefficient and p-values | Feature | Main Effects | | Interactions | |
|---|---------|--------------|----------|--------------|------------|
| | | coefficients | p-value | coefficients | p-value |
| Metabolite LV1 | | -0.0579 | < 2 e-16 | -0.0541 | 1.36 e-13 |
| Metabolite LV2 | | 0.0624 | 3.98e-16 | 0.0572 | 8.10 e -08 |
| Metabolite LV3 | | 0.0394 | 1.07e-10 | 0.0359 | 8.86 e-07 |
| Microbiota LV1 | | 0.0256 | 0.141 | 0.0495 | 0.006 |
| Enzymes LV1 | | 0.0115 | 0.083 | -0.0002 | 0.976 |
| Metabolite LV1 x Microbiota LV1 | | - | - | 0.0079 | 0.007 |
| Metabolite LV2 x Microbiota LV1 | | - | - | 0.0069 | 0.300 |
| Metabolite LV3 x Microbiota LV1 | | - | - | -0.0049 | 0.264 |
| Metabolite LV1 x Enzymes LV1 | | - | - | -0.0006 | 0.569 |
| Metabolite LV2 x Enzymes LV1 | | - | - | 0.0078 | 0.006 |
| Metabolites LV3 x Enzymes LV1 | | - | - | 0.0013 | 0.446 |
| Microbiota LV1 x Enzymes LV1 | | - | - | -0.0048 | 0.118 |
| Metabolite LV1 x Microbiota LV1 x Enzymes LV1 | | - | - | -0.0010 | 0.002 |
| Metabolite LV2 x Microbiota LV1 x Enzymes LV1 | | - | - | 0.0007 | 0.176 |
| Metabolite LV3 x Microbiota LV1 x Enzymes LV1 | | - | - | 0.00005 | 0.245 |
| Intercept | | 0.6666 | < 2 e-16 | 0.7274 | < 2 e-16 |
| | | F-stat | 140.9 | F-stat | 77.84 |
| | | p-value | 2.2 e-16 | p-value | 2.2 e-16 |

19 **Table 4.- AUC and Balanced Error Rate comparison from single-omics sPLS-DA models for UC**

| Performance Metric | Concatenation Control | Metabolome | Microbiome | Enzymatic Composition |
|---------------------|-----------------------|------------|------------|-----------------------|
| AUC | 0.952 | 0.922 | 0.932 | 0.969 |
| Balanced Error Rate | 0.5 | 0.343 | 0.5 | 0.125 |

59

Table 5.- Single-omic sPLS-DA models show higher coverage than concatenation control model in UC.

| Latent Variable | Concatenation Control | | Metabolome | | Microbiome | | Enzymatic Composition | |
|-----------------|-----------------------|--------------------|------------------|--------------------|------------------|--------------------|-----------------------|--------------------|
| | Encoded Features | Explained Variance | Encoded Features | Explained Variance | Encoded Features | Explained Variance | Encoded Features | Explained Variance |
| LV1 | 100 | 8% | 100 | 10% | 80 | 5% | 100 | 13% |
| LV2 | 1 | 4% | 100 | 3% | 1 | 2% | 90 | 6% |
| Total | 101 | 16% | 200 | 13% | 81 | 7% | 190 | 19% |

60

61

62

Table 6.- LIVE Modeling Outcome Table with Coefficients and p-values for UC.

| Ulcerative Colitis Coefficient and p-values | Main Effects | | Interaction Effects | | |
|---|--------------|--------------|---------------------|--------------|----------|
| | Feature | coefficients | p-value | coefficients | p-value |
| Metabolite LV1 | | -0.0590 | 1.05e-12 | -0.059 | 6.01e-12 |
| Metabolite LV2 | | 0.0595 | 1.38e-13 | 0.061 | 5.23e-10 |
| Microbiota LV1 | | -0.0486 | 0.002 | -0.066 | 0.0003 |
| Enzymes LV1 | | -0.0003 | 0.45 | 0.023 | 0.019 |
| Metabolite LV1 x Microbiota LV1 | | - | - | -0.001 | 0.675 |
| Metabolite LV2 x Microbiota LV1 | | - | - | 0.012 | 0.050 |
| Metabolite LV1 x Enzymes LV1 | | - | - | 0.001 | 0.477 |
| Metabolite LV2 x Enzymes LV1 | | - | - | 0.001 | 0.487 |
| Microbiota LV1 x Enzymes LV1 | | - | - | -0.0005 | 0.900 |
| Metabolite LV1 x Microbiota LV1 x Enzymes LV1 | | - | - | -0.002 | 0.003 |
| Metabolite LV2 x Microbiota LV1 x Enzymes LV1 | | - | - | 0.002 | 0.018 |
| Intercept | | 0.609 | 2e-16 | 0.642 | 2e-16 |
| | | F-stat | 87.07 | F-stat | 47.94 |
| | | p-value | 2.2 e-16 | p-value | 2.2 e-16 |

63

Table 7.- CD and UC features associated with drug response that were predicted by LIVE modeling.

| Features | Crohn Disease | Ulcerative Colitis |
|--------------------|--|---|
| Metabolites | - | C18-neg_Cluster_1016_Acyl carnitines |
| Microbiota Species | Gordonibacter_pamelaeae | Gordonibacter_pamelaeae |
| Enzymes | 2.1.1.133 Precorrin-4 C (11)-methyltransferase | 2.3.1.157 Glucosamine-1-phosphate N-acetyltransferase |
| | 2.1.1.132 Precorrin-6B C (5,15) -methyltransferase (decarboxylating) | 2.7.7.72 CCA tRNA nucleotidyltransferase |
| | 2.3.1.157 Glucosamine-1-phosphate N-acetyltransferase | 3.2.1.89 Arabinogalactan endo-beta-1,4-galactanase |
| | 2.7.7.72 CCA tRNA nucleotidyltransferase | 5.4.2.2 Phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent) |
| | 3.1.3.15 Histidinol-phosphatase | - |

64

65

58 **SUPPORTING INFORMATION AND CAPTIONS**

59 **List of Supplemental Materials**

- 70 1. Supplemental Table 1.- Linn VIP from Metabolomic single-omic sPLS-DA model for CD
- 71 2. Supplemental Table 2.- Linn VIP from Microbiota composition single-omic sPLS-DA model for CD
- 72 3. Supplemental Table 3.- Linn VIP from Enzymatic composition single-omic sPLS-DA model for CD
- 73 4. Supplemental Table 4.- Model Performance single-omic sPLS-DA models and concatenation control.
- 74 5. Supplemental Table 5.- Significant metabolic, microbiota and enzyme pairs that determine CD status
- 75 6. Supplemental Table 6.- Node File for CD in Cytoscape
- 76 7. Supplemental Table 7.- Edge File for CD in Cytoscape
- 77 8. Supplemental Table 8.- Node File for CD VIP > 5 to build Meta-Model Networks
- 78 9. Supplemental Table 9.- Linn VIP from Metabolomic single-omic sPLS-DA model for UC
- 79 10. Supplemental Table 10.- Linn VIP from Microbiota composition single-omic sPLS-DA model for UC
- 80 11. Supplemental Table 11- Linn VIP from Enzymatic composition single-omic sPLS-DA model for UC
- 81 12. Supplemental Table 12.- Model Performance single-omic sPLS-DA models and concatenation control.
- 82 13. Supplemental Table 13.- Significant metabolic, microbiota and enzyme pairs that determine UC status
- 83 14. Supplemental Table 14.- Node File for UC in Cytoscape
- 84 15. Supplemental Table 15.- Edge File for CD in Cytoscape
- 85 16. Supplemental Table 16.- Node File for UC VIP > 4 to build Meta-Model Networks