

1 A plasma peptidomic signature reveals extracellular matrix remodeling and predicts prognosis in
2 alcohol-related hepatitis.

3

4 #Khaled Sayed^{1,2}, #Christine E. Dolin³, Daniel W. Wilkey³, Jiang Li⁴, Toshifumi Sato⁴, Juliane I
5 Beier^{4,5}, Josepmaria Argemi^{4,6}, Ramon Bataller⁷, Abdus S Wahed⁸, *Michael L Merchant³,
6 *Panayiotis V Benos¹, *Gavin E Arteel^{4,5}

7

8 ¹Department of Epidemiology, University of Florida, Gainesville, Florida, USA.

9 ²Department of Electrical & Computer Engineering and Computer Science, University of New
10 Haven, West Haven, Connecticut, USA.

11 ³Department of Medicine, University of Louisville, Louisville, Kentucky, USA.

12 ⁴Department of Medicine, Division of Gastroenterology, Hepatology and Nutrition, University of
13 Pittsburgh, Pittsburgh, PA, USA.

14 ⁵Pittsburgh Liver Research Center, University of Pittsburgh, Pittsburgh, PA, USA.

15 ⁶Department of Internal Medicine, Clinical University of Navarra, Navarra, Spain.

16 ⁷Liver Unit, Hospital Clinic. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS),
17 Barcelona, Spain.

18 ⁸Department of Biostatistics and Computational Biology, University of Rochester, Rochester,
19 NY, USA.

20

21 #Equally contributing first authors

22 *Equally contributing senior authors

23 Running title: Plasma peptidome of alcohol-related hepatitis.

24

25 Keywords: causal models, protein degradomics, extracellular matrix, ALD, LC-MS/MS

1 Send all correspondence to: Gavin E. Arteel, PhD, FAASLD
2 Thomas E. Starzl Biomedical Science Tower
3 West 1143
4 200 Lothrop Street
5 Pittsburgh, PA 15213
6 Phone: +1-412-648-4187
7 Email: gearteel@pitt.edu
8
9

10 **Abbreviations:** ABIC, age, bilirubin, INR and creatinine; AH, alcohol-related hepatitis; ALD,
11 alcohol-related liver disease; ALT, alanine aminotransferase; AP, alkaline phosphatase; ASH,
12 alcohol-related steatohepatitis; AST, aspartate aminotransferase; AUD, alcohol use disorder;
13 AUDIT, alcohol use disorders identification test; BMI, body mass index; BTO, Brenda tissue
14 ontology; CTP, Child-Turcotte-Pugh; DAG, directed acyclic graphs; DF, discriminant function;
15 DSM, diagnostic and statistical manual of mental disorders; ECM, extracellular matrix; FGES,
16 fast greedy equivalence search; FLIGHT, functional liver-image guided therapy; GO, gene
17 ontology; LC-MS/MS, liquid Chromatography with tandem mass spectrometry; LTDH, lifetime
18 drinking history; MELD, model for end-stage liver disease; PCA, principal component analysis;
19 PGM, probabilistic graphical models; oPLS-DA, orthogonal partial least squared-discriminant
20 analysis; TIC, total ion chromatogram; VIP, variable importance plot.

1 **ABSTRACT**

2 Alcohol-related hepatitis (AH) is plagued with high mortality and difficulty in identifying at-risk
3 patients. The extracellular matrix undergoes significant remodeling during inflammatory liver
4 injury that can be detected in biological fluids and potentially used for mortality prediction. EDTA
5 plasma samples were collected from AH patients (n= 62); Model for End-Stage Liver Disease
6 (MELD) score defined AH severity as moderate (12-20; n=28) and severe (>20; n=34). The
7 peptidome data was collected by high resolution, high mass accuracy UPLC-MS. Univariate and
8 multivariate analyses identified differentially abundant peptides, which were used for Gene
9 Ontology, parent protein matrix composition and protease involvement. Machine learning
10 methods were used on patient-specific peptidome and clinical data to develop mortality
11 predictors. Analysis of plasma peptides from AH patients and healthy controls identified over
12 1,600 significant peptide features corresponding to 130 proteins. These were enriched for ECM
13 fragments in AH samples, likely related to turnover of hepatic-derived proteins. Analysis of
14 moderate versus severe AH peptidomes showed a shift in abundance of peptides from collagen
15 1A1 and fibrinogen A proteins. The dominant proteases for the AH peptidome spectrum appear
16 to be CAPN1 and MMP12. Increase in hepatic expression of these proteases was orthogonally-
17 validated in RNA-seq data of livers from AH patients. Causal graphical modeling identified four
18 peptides directly linked to 90-day mortality in >90% of the learned graphs. These peptides
19 improved the accuracy of mortality prediction over MELD score and were used to create a
20 clinically applicable mortality prediction assay. A signature based on plasma peptidome is a
21 novel, non-invasive method for prognosis stratification in AH patients. Our results could also
22 lead to new mechanistic and/or surrogate biomarkers to identify new AH mechanisms.

- 1 **Lay summary.** We used degraded proteins found the blood of alcohol-related hepatitis patients
- 2 to identify new potential mechanisms of injury and to predict 90 day mortality.

1 Alcohol-related hepatitis (AH) is a subacute form of alcohol-related liver disease with a
2 high mortality rate of 30-50% at 3 months and 40% at 6 months [1, 2]. AH is characterized by
3 hepatic decompensation, jaundice and multiple organ failure [3]. AH occurs in patients with
4 heavy chronic alcohol consumption (80-100 g per day) and can be the first manifestation of
5 clinically silent ALD or an exacerbation of pre-existing cirrhosis [3].

6
7 Accurately predicting AH patient outcome risks is important for clinical decision-making.
8 For example, AH patients with higher negative outcome (e.g., mortality) risks are better
9 candidates for corticosteroid treatment, and patients with lower risk could be candidates for
10 long-term clinical studies [2, 3]. Currently, the best approach for predicting outcome risk is
11 combining static scores, such as the modified Maddrey's discriminant function (MDF), Model for
12 End-stage Liver Disease (MELD), prognostic algorithm score constituting Age, Bilirubin, INR
13 and Creatinine (ABIC), and/or Glasgow with the dynamic Lille scoring system [4, 5], with the
14 MELD score favored globally [6]. These clinical scores are useful for predicting outcome risks in
15 patients with severe AH, but are limited in predicting outcome risks in patients with moderate
16 disease [7]. Our group demonstrated significant extracellular matrix (ECM) remodeling during
17 inflammatory liver injury [8]. During such remodeling, altered protein turnover shifts the
18 distribution of peptide fragments including degraded ECM in biologic fluids (e.g., plasma) [9].
19 Peptidomic analysis of the degraded ECM (i.e., 'degradome') is a useful diagnostic/prognostic
20 tool in metastatic cancers and other diseases of ECM remodeling [9].

21
22 Probabilistic Graphical Models (PGMs), in general (i.e., "causal graphs") have recently
23 gained popularity, because of their simplicity and straightforward interpretability. When certain
24 assumptions are met,[10] there are theoretical guarantees that DAGs will recover the true
25 cause-effect associations [11, 12]. Current methods can handle mixed data types (continuous
26 and discrete) [13, 14], and have reduced processing time [15, 16], overcoming past obstacles.

1 Importantly, these methods can be used to build robust predictive models of an outcome [17,
2 18].

3
4 It was hypothesized that the severe inflammatory liver injury caused by AH would yield a
5 unique peptidome profile in human patient plasma, and that unique ECM peptides or peptide
6 grouping would vary between patient groups. The goals of this work were three-fold: 1) identify
7 novel surrogate candidate biomarkers for AH, 2) develop new mechanistic hypotheses by
8 predicting proteases that generated the observed peptidome, and 3) employ PGM to identify
9 unique predictors of outcome from the peptidome profile (see **Figure 1** for scheme).

10

1 **Methods**

2 **Study participants.** The University of Louisville Human Studies Committee Human
3 approved sample collection and use of de-identified samples provided in this study. All study
4 subjects provided informed consent prior to sample collection. All studies were conducted in
5 compliance with the Declaration of Helsinki. A total of 70 adult male and female individuals
6 participated in this NIH-funded study (**Figure 1A**). This investigation constitutes a single time
7 point assessment of patients between the study subgroups. Data were collected from
8 biobanked samples from a large national multisite clinical trial (clinicaltrials.gov: NCT01922895
9 and NCT01809132). Inclusion and exclusion criteria are listed in those studies. Informed
10 consent was obtained from all study participants before collection of data and bodily samples.
11 All AH patients were enrolled at the University of Louisville, the University of Massachusetts
12 Medical School, the University of Texas-Southwestern and the Cleveland Clinic. AH diagnosis
13 was done using clinical and laboratory criteria described by the NIAAA consortium on AH [19].
14 Individuals with liver injury met the criteria for AUD based on DSM 4 XR or DSM 5 manual. All
15 healthy participants were recruited at the University of Louisville free of any clinically diagnosed
16 disease (liver or organ systems) that might contribute to altered laboratory values in comparison
17 analyses.

18
19 The subgroups included healthy participants (n=7) and AH patients (n=63; **Figure 1A**).
20 Clinical variables did not exist for healthy controls and for one AH patient and were therefore
21 excluded from categorical analyses (n=62). For categorical comparisons, AH patients were
22 stratified as “moderate” (MELD=12-20; n=28) and “severe” AH (>20; n=34).[20] Out of the 63
23 AH patients, survival information was lacking for 5 patients, so they were excluded from the
24 causal graphical modeling (n=58; see **Supplemental Material**). A variety of clinical data was
25 gathered for these patients, including transaminases, alkaline phosphatase, and total bilirubin.

1 **Table 1** shows a list of demographics and clinical data for the moderate AH, and severe AH
2 participants.

3
4 **Analytical approaches, data analysis and causal graphical modeling.** Peptidomic
5 analysis of patient samples were conducted as recently described with some modifications
6 **(Figure 1B)** [21]. That and all other detailed methods are provided in Supplemental Material.

1 **Results**

2 For initial analysis, we compared the degradome in moderate (MELD 12-20) and severe
3 (MELD >20) AH versus healthy controls. Moderate and severe AH patients did not differ
4 significantly by age, sex, or race (**Table 1**). The median age was 51 years, 66% of patients were
5 male, and 90% were white. In addition to MELD, severe AH patients had higher MDF (median
6 58 vs. 18), CTP/child-Pugh score (median 11 vs. 9), bilirubin (median 18.6 vs. 4.9 mg/dl), and
7 INR (median 2.0 vs 1.4) and Ascites score (88.2% vs 51.9% 1-2) than the moderate AH
8 patients (all $p < 0.001$). AST and ALT did not differ significantly by MELD score severity. The
9 peptidomic dataset consisted of 1,693 primary peptidome features identified by PEAKS X Pro,
10 corresponding to degradation products of 134 unique proteins. There was significant qualitative
11 overlap between peptides changed in Moderate and Severe AH (vs healthy control), as
12 visualized by Venn Diagram (**Figure 1C**). Differences in relative peptide abundance between
13 moderate and severe AH were determined using t-test (on preprocessed TIC-normalized data);
14 volcano plots visualize these results (**Figure 1D**). These data demonstrate a shift toward
15 increased relative abundances of collagen 1A1 (COL1A1) and collagen 1A2 (COL1A2)
16 fragments and relative decreases in some fibrinogen A (FGA) peptide fragments in severe AH.

17
18 **Feature analysis of the peptidome.** PCA showed that the two largest principal
19 components account for 8.0% and 5.8% variability between the three participant categories
20 (healthy, and moderate and severe AH; **Figure 2A**). Repeated analysis of healthy versus
21 moderate and healthy versus severe categories demonstrated a slight increase in PC1 to
22 explaining 10% of the data separation (**Figure 2B**). Comparison of the moderate versus severe
23 AH samples decreased PC1 to 7% thus suggesting most of the variability in the data, could be
24 attributed to the differences between healthy versus moderate and healthy versus severe.

25

1 Impact of AH on plasma peptidome profile: dominance of matrisome-derived peptides. The
2 matrisome is an expanded definition of the ECM that also incorporates ECM-affiliated proteins
3 and ECM-modifying proteins.[22] The gene names associated with all identified peptides
4 (**Supplemental Table S1**) were submitted for annotation by matrisome category and division
5 using the matrix-annotator tool *MatrisomeAnalyzer*
6 (<https://matrinet.shinyapps.io/MatrisomeAnalyzer/>). The plasma peptidome comparison of the
7 healthy versus AH samples were enriched in AH samples for peptides belonging to components
8 of the core matrisome (collagen, ECM glycoprotein, proteoglycan) or the matrisome associated
9 compartment (ECM regulators, ECM-affiliated proteins, secreted factors; **Supplemental Table**
10 **S2**). The pattern of proteins from significantly differential abundant peptides that belong to
11 matrisome did not differ with AH severity, comprising 60% of the total peptide signal (vs. healthy
12 controls). The majority were defined as core matrisome proteins (collagens, ECM glycoproteins
13 and proteoglycans) in all comparisons.

14
15 To further assess the molecular differences between the moderate and severe AH, we
16 performed a supervised oPLS-DA score plot analysis of the plasma peptidome profiles of
17 individual patients (**Figure 2C**) grouped by moderate and severe AH. The variation both within
18 (7.3%) and between (4.4%) was small suggesting the differences between AH groups are
19 driven by a small set of peptides. Variable importance plot (VIP) identified the top 25 peptides
20 separating moderate and severe AH groups. In total, 23 of the top 25 peptides were collagen
21 fragments. Interestingly, orthogonal partial least squared-discriminant analysis (oPLS-DA)
22 indicated several overlapping ECM-derived peptides (e.g., degraded fibrinogen and collagen
23 proteins) were dominant in the top-scored peptides (**Figure 2D**). These differences in plasma
24 peptidomes suggest that biomarkers could be developed.

25

1 Biological pathway analyses indicate remodeling of hepatic tissue. The significantly different
2 peptides in AH (moderate AH vs controls, severe AH vs controls) were analyzed using StringDB
3 [23], which contains physical and functional protein-protein associations from both
4 experimentally validated and homology-based associations. Molecular pathways related to
5 altered metabolism and remodeling were also enriched in Gene Ontology (GO) terms for
6 Biological Process and Cellular Component (**Supplemental Tables S3 and S4**). Alcohol-
7 related hepatitis is a systemic disorder and extrahepatic dysfunction (e.g., skeletal muscle and
8 kidneys) is a key driver of mortality [24]. Despite this factor, the liver was the most common solid
9 organ enriched in the data set as determined by Brenda Tissue Ontology (BTO:0000759;
10 **Supplemental Table S5**) in the AH-moderate and -severe versus healthy controls ($p=5.82\times 10^{-11}$
11 and 4.90×10^{-16} , respectively), and second only to plasma proteins (e.g., BTO:0000131; $p=$
12 2.22×10^{-14} and 4.39×10^{-23} , respectively). Interestingly, another tissue that was highly enriched
13 in the peptidome in AH-moderate and -severe versus healthy controls was determined to be of
14 fetal origin (BTO:0000449; $p=1.56\times 10^{-9}$ and $p=7.68\times 10^{-8}$, respectively), which is in-line with
15 previous studies (e.g., [25]).

16

17 Calpains and MMP-14 proteases are predicted to regulate the observed peptidome. Many
18 proteases cleave substrates with high specificity only at certain sequence sites. Thus,
19 information on the fragment sequence of degraded proteins can inform on proteases that may
20 have generated this pattern. Proteasix (proteasix.org) is an open-source peptide-centric tool to
21 predict *in silico* the proteases involved in generating these peptides [26]. **Figure 2E** shows the
22 relative frequency (node size) with which the top 16 proteases were predicted to generate the
23 resultant peptidome peptides by this analysis. The two top predicted upregulated proteases,
24 Calpain -1/-2 (CAPN1/CAPN2) and MMP-14 (MMP14) were also robustly induced in publicly-
25 available RNAseq expression data from human AH (**Figure 2E**, node color) [25]. Indeed, there

1 was generally good concordance between the Proteasix prediction and the hepatic gene
2 expression data from that study.

3
4 Peptide features that are directly linked to 90-day survival in AH. The initial analysis was to
5 categorically describe the gestalt changes in the peptidome caused by moderate and severe AH
6 (as determined by MELD). These results may yield useful insight for future mechanistic or
7 interventional studies. However, as mentioned in the Introduction, a key limitation in the clinical
8 management of AH is an accurate tool to predict outcome after clinical presentation, namely
9 patient 90-day survival. We hypothesized that representative peptides from the peptidome
10 could serve as surrogate biomarkers to predict this outcome.

11
12 Identifying differentially abundant peptides and pathways related to AH severity is
13 undoubtedly useful for future studies. However, for clinical purposes, we were also interested in
14 identifying a few peptides that could inform AH patient survival in conjunction or independently
15 of the MELD score. To identify such potential effector peptides, we used (causal) probabilistic
16 graphical models. Given the relatively small size of our unique dataset, we followed a leave-
17 one-out approach, where we learned 58 graphs (one for each sample that was left out) at 10
18 different scarcities (see Methods). Then we counted the times, out of a total of 580 models, a
19 particular variable was independently linked (i.e., belongs to its Markov blanket) to the 90-day
20 mortality (binarized) variable. The graph models that did not include clinical features had four
21 peptide fragments that consistently appeared in >90% of the graphs: X83A, X54A, X79C, and
22 X142A. These fragments correspond to parent genes VIM, APOC1, TUBB, CALD1, respectively
23 (**Supplemental Figure S2**). A fifth fragment (X231B, gene BIN2) appears in 76% of the models
24 (**Supplemental Table S6**). Three of these genes (CALD1, TUBB, VIM) belong to the
25 cytoskeleton pathway.

26

1 When we included the clinical features and the MELD score in our models, the same four
2 fragments still appeared in >90% of the graphs, but fragment X231B appeared in 24.5% of the
3 models (**Supplemental Table S7**). The MELD score was also found in the top informative
4 features, as it was included in 58% of the models. We noted that MELD and X231B (BIN2) appear
5 together in only 10 of the models (1.7%), indicating they may contain redundant information
6 regarding the 90-day survival. **Supplemental Figure S8** shows the distribution of the identified
7 variables within each predicted outcome group (i.e., alive vs deceased).

8
9 Additionally, we performed stepwise regression analysis on the five informative peptides to
10 further reduce the number of relevant peptides that can be measured in a clinical setting when
11 predicting the mortality of new patients. We found that only X54A (APOC1), X142A (CALD1), and
12 X231B (BIN2) are the most relevant features, and these are the features we used for our
13 predictive model.

14
15 Development of a three-peptide signature for clinical application. In clinical practice it is not
16 feasible to use whole peptidome measurements as a prognostic test. Therefore, we wanted to
17 test whether the model we learned from TIC-normalized data can be used to make predictions
18 about survival when a limited number of peptides are measured. Thus, we simulated a clinical
19 application of the predictive ability of the 3-peptidome signature as follows. We divided the
20 dataset into five folds. In each fold, we used 80% of the samples (training set) to select the
21 optimum parameters of the three peptides (“signature peptides”) through 10X cross-validation.
22 For the training phase, we used the TIC-normalized data. The performance of the model of this
23 fold was assessed on the raw measurements of the 20% of the left-out samples (validation set).

24
25 To avoid sample-to-sample variation in these three peptides, we used a small number of
26 peptides to act as normalization standards. We chose the four most invariant peptides (p -value

1 >0.95) with mean raw concentration >14 across all samples (“invariant peptides”) as our
2 normalization standard. To overcome the problem of spontaneous missing values, normalization
3 was done based on the median of the 4 invariant peptides. The selected most invariant peptides
4 were: X38A, X157A, X61A, and X154C, corresponding to parent proteins CO3, LASP1, FETUA,
5 and ITIH4 respectively. The median of the four invariant peptides was calculated for all samples
6 with 3 or more non-zero values to normalize the peptides associated with mortality as shown in
7 Equation 1:

$$8 \quad X_i|_{normalized} = \frac{\log_2[X_i]}{\text{median}(\sum_{k=1}^4 \log_2[X_k])} \quad Eq. 1$$

9 where X_i represents a model variable and X_k represents an invariant variable. Both X_i and
10 X_k can be non-TIC or TIC-normalized peptides. It is worth mentioning that only one sample had
11 two non-zero values. In this way, in a potential clinical application, one has to measure only
12 these 7 peptides (3 signature and 4 invariant peptides).

13
14 Using the above procedure, we learned and compared three logistic regression models.
15 *Model 1* consisted of the MELD score only and served as our baseline model. *Model 2*
16 consisted of the three peptidomic features only. *Model 3* included the three peptidomic features
17 and the MELD score. After learning the optimum weights using 10X cross-validation on the 80%
18 of the data in each fold, we selected the optimum classification threshold as the point of
19 intersection of sensitivity and specificity (**Figure 3A**). The threshold value where the sensitivity
20 and specificity curves intersect was selected as the optimal threshold in each fold
21 (**Supplemental Table S8**). The density function for the two categories (alive and deceased)
22 across thresholds for each model is plotted in **Figure 3B**. This shows that the model only with
23 the MELD score did not separate the two distributions well, while the peptidome models (with or
24 without the MELD score) performed better. The validation results of each cross-validation fold
25 are shown in **Table 2**. Like the initial training analysis (see **Supplemental Table S7**), we see

1 that the 3-peptide model and the composite model (MELD+3-peptides) improve the MELD-only
2 model in terms of average sensitivity, specificity, and balanced accuracy (**Figure 3C**).

3
4 For the final model to be used in future clinical settings, we compared the three models,
5 trained on the full dataset (parameter setting through 10× cross-validation on the TIC-
6 normalized data), and tested them using the raw data (normalized by the median of the invariant
7 peptides). Confusion matrices, sensitivity, specificity, balanced accuracy, and survival curves,
8 obtained for each model when the models were tested using non-TIC normalized data, are
9 presented in **Figure 3C**. Consistently with the results above, the highest performance measures
10 were obtained for Model 3 whereas the lowest performance measures were obtained for the
11 MELD-only model (Model 1). Additionally, Models 2 and 3 split the data into two clusters (i.e.,
12 Predicted Alive and Predicted Deceased) with 12 out of 13 Deceased samples grouped in the
13 Predicted Deceased cluster. Model 3 outperformed Model 2 in detecting the Alive samples with
14 a sensitivity of 80%. Finally, the survival curves in **Figure 3C** show that Models 2 and 3 can
15 predict the survival of AH patients better than Model 1 which is based on the MELD score only.
16 The distribution of the final model variables (i.e., X54A, X142A, and X231B) over the predicted
17 classes obtained by each model is presented in **Supplemental Figure S3**. Additionally, the
18 demographics and clinical characteristics of each predicted class in each model are shown in
19 **Supplemental Tables S9-S11**.

20

1 **Discussion**

2 Both the AH diagnosis and prognosis could be impacted by the development of more
3 sensitive and specific surrogate candidate biomarkers. Our group previously demonstrated that
4 inflammatory stress causes the hepatic ECM to undergo dynamic transitional remodeling [8].
5 Others have shown that ECM remodeling causes degradation products to be secreted into the
6 blood and that analysis is a useful prognostic tool in diseases [9]. Therefore, here we aim to
7 study the plasma ECM peptidome changes with AH severity and to develop a new, clinically
8 useful method, to predict 90-day survival in AH patients.

9
10 Informatics analysis of these peptides demonstrated an enrichment of ECM fragments in the
11 AH patient samples that is likely related to the turnover of hepatic proteins. Six fibrinopeptide A
12 peptides were more abundant in plasma from moderate AH patients (compared to severe AH),
13 whereas 27 peptides (24 COL1A1 fragments, one COL1A2, one COL1A3, and one BIN2
14 peptide fragment) were more abundant in plasma from severe AH patients. The findings of
15 increased collagen in more severe AH is in line with previous studies indicating that underlying
16 fibrosis drives AH prognosis [27]. Using a discriminant analysis (oPLS-DA; **Figure 2C**) the
17 moderate and severe AH samples were well resolved into two sample groups and the VIP
18 scores with a similar pattern (**Figure 2D**).

19
20 Analysis of this spectrum also identified 2 proteases that appear to be dominant in
21 generating the pattern associated with AH (CAPN1 and MMP12); the increase in hepatic
22 expression of these proteases was orthogonally validated in a separate analysis of publicly
23 available bulk RNA-seq data of livers from AH patients (**Figure 2E**). Little is known regarding
24 the role of MMP14 in liver disease. Recent work by this group showed that CAPN2 is
25 progressively induced in post-transplant NASH fibrosis severity [28]. These proteases have not

1 been identified to be involved in AH prior to this work. Future studies should investigate this
2 finding further.

3

4 Besides investigating the molecular mechanisms of AH progression, we also wanted to
5 develop a predictor for 90-day survival (outcome) using PGM. Significantly, the variables
6 selected through this process provide independent information about 90-day survival, and
7 collectively are the most informative variable set, even though some of them, individually, may
8 not appear to be significantly different in the two categories (alive/deceased; **Supplemental**
9 **Figure S2**). The final models showed that the 3-peptide model was equally accurate in
10 predicting 90-day mortality as the composite one (MELD+3-peptides; **Figure 3C**). One peptide
11 (APOC1) has been shown to be a marker of AH severity in a recent plasma proteomic study
12 [29]. The other two peptides (CALD1 and BIN2) are both cytoskeletal proteins that have not
13 been previously associated with AH and warrant further investigation.

14

15 While our goal was to differentiate moderate and severe AH and AH outcome (i.e.,
16 mortality), we acknowledge the ratio of HC to AH in our cohort was imbalanced in “n” values and
17 this may be insufficient to adequately power the study. Additionally, the potential for a “clinical
18 site effect” may be present, although the AH samples were collected from multiple institutions.
19 Despite these study limitations, the statistical modeling and informatics filtering of the
20 peptidomics data supports the hypothesis that the ECM plasma peptidome is associated with
21 the AH spectrum. These patterns of select peptidome “features” can be investigated further in
22 future studies as biomarkers for AH severity and outcome.

23

24 Another important consideration for surrogate biomarker discovery is the impact of the
25 sample preparation method on the peptidome. Here, we used K3EDTA plasma, which inhibits
26 coagulation-driven proteases that are dependent on divalent cations. K3EDTA is an

1 anticoagulant of choice for plasma proteomic studies, in general, and peptidomics in particular
2 [30]. Although direct comparisons between plasma preparations were not performed, several of
3 the peptides identified as key biomarkers for AH outcome in this study (e.g., vimentin) have an
4 affinity for heparin [31], and thus may not be present in studies involving heparinized plasma.
5 Therefore, these study results are applicable at this point with use in studies of EDTA-based
6 plasma samples.

7
8 Taken together, the results of this study indicate that analysis of the plasma peptidome can
9 yield useful new information on mechanism and outcome prediction in AH. This study validates
10 previous mechanistic findings (e.g., ECM remodeling and fetal-like reprogramming), as well as
11 identifies new potential “players” at the level of degraded proteins and proteases that may
12 generate these signals. Moreover, our curated algorithm identified by CGM proved to be
13 superior to MELD both in sensitivity and specificity to predict mortality in AH. Future studies will
14 investigate these prospects further.

15

1 Reference List

- 2
- 3 [1] Torok NJ. Update on Alcoholic Hepatitis. *Biomolecules* 2015;5:2978-2986.
- 4 [2] Lucey MR, Mathurin P, Morgan TR. Alcoholic hepatitis. *N Engl J Med* 2009;360:2758-
5 2769.
- 6 [3] Seitz HK, Bataller R, Cortez-Pinto H, Gao B, Gual A, Lackner C, et al. Alcoholic liver
7 disease. *Nat Rev Dis Primers* 2018;4:16.
- 8 [4] Louvet A, Labreuche J, Artru F, Boursier J, Kim DJ, O'Grady J, et al. Combining Data
9 From Liver Disease Scoring Systems Better Predicts Outcomes of Patients With
10 Alcoholic Hepatitis. *Gastroenterology* 2015;149:398-406 e398; quiz e316-397.
- 11 [5] Degré D, Wandji LCN, Moreno C, Louvet A. Alcoholic hepatitis: Towards an era of
12 personalised management. *United European Gastroenterol J* 2020;8:995-1002.
- 13 [6] Morales-Arráez D, Ventura-Cots M, Altamirano J, Abrales JG, Cruz-Lemini M, Thursz
14 MR, et al. The MELD Score Is Superior to the Maddrey Discriminant Function Score to
15 Predict Short-Term Mortality in Alcohol-Associated Hepatitis: A Global Study. *Am J*
16 *Gastroenterol* 2022;117:301-310.
- 17 [7] Dunn W, Jamil LH, Brown LS, Wiesner RH, Kim WR, Menon KV, et al. MELD accurately
18 predicts mortality in patients with alcoholic hepatitis. *Hepatology* 2005;41:353-358.
- 19 [8] Massey VL, Dolin CE, Poole LG, Hudson SV, Siow DL, Brock GN, et al. The hepatic
20 "matrisome" responds dynamically to injury: Characterization of transitional changes to
21 the extracellular matrix in mice. *Hepatology* 2017;65:969-982.

- 1 [9] Sand JM, Leeming DJ, Byrjalsen I, Bihlet AR, Lange P, Tal-Singer R, et al. High levels of
2 biomarkers of collagen remodeling are associated with increased mortality in COPD -
3 results from the ECLIPSE study. *Respir Res* 2016;17:125.
- 4 [10] Spirtes P, Glymour CN, Scheines R. *Causation, Prediction, and Search*. Cambridge, MA:
5 MIT Press; 2000.
- 6 [11] Glymour C, Cooper GF, editors. *Computation, Causation, and Discovery*. Cambridge,
7 MA: MIT Press; 1999.
- 8 [12] Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge
9 University Press; 2009.
- 10 [13] Sedgewick AJ, Buschur K, Shi I, Ramsey JD, Raghu VK, Manatakis DV, et al. Mixed
11 graphical models for integrative causal analysis with application to chronic lung disease
12 diagnosis and prognosis. *Bioinformatics* 2019;35:1204-1212.
- 13 [14] Andrews B, Ramsey J, Cooper GF. Learning High-dimensional Directed Acyclic Graphs
14 with Mixed Data-types. *Proc Mach Learn Res* 2019;104:4-21.
- 15 [15] Ramsey JD. A scalable conditional independence test for nonlinear, non-Gaussian data`.
16 arXiv:14015031 [csAI]; 2014.
- 17 [16] Fucello AN, Yuan DY, Benos PV, Raghu VK. Improving Constraint-Based Causal
18 Discovery from Moralized Graphs. *NeurIPS workshop on Causal Discovery and*
19 *Causality-inspired Machine Learning*. on-line; 2020.
- 20 [17] Raghu VK, Zhao W, Pu J, Leader JK, Wang R, Herman J, et al. Feasibility of lung cancer
21 prediction from low-dose CT scan and smoking factors using causal models. *Thorax*
22 2019;74:643-649.

- 1 [18] Raghu VK, Beckwitt CH, Warita K, Wells A, Benos PV, Oltvai ZN. Biomarker identification
2 for statin sensitivity of cancer cell lines. *Biochem Biophys Res Commun* 2018;495:659-
3 665.
- 4 [19] Crabb DW, Bataller R, Chalasani NP, Kamath PS, Lucey M, Mathurin P, et al. Standard
5 Definitions and Common Data Elements for Clinical Trials in Patients With Alcoholic
6 Hepatitis: Recommendation From the NIAAA Alcoholic Hepatitis Consortia.
7 *Gastroenterology* 2016;150:785-790.
- 8 [20] Mitchell MC, Friedman LS, McClain CJ. Medical Management of Severe Alcoholic
9 Hepatitis: Expert Review from the Clinical Practice Updates Committee of the AGA
10 Institute. *Clin Gastroenterol Hepatol* 2017;15:5-12.
- 11 [21] Li J, Sato T, Hernández-Tejero M, Beier JI, Sayed K, Benos PV, et al. The plasma
12 degradome reflects later development of NASH fibrosis after liver transplant. *Sci Rep*
13 2023;13:9965.
- 14 [22] Naba A, Clauser KR, Hoersch S, Liu H, Carr SA, Hynes RO. The matrisome: in silico
15 definition and in vivo characterization by proteomics of normal and tumor extracellular
16 matrices. *Molecular & cellular proteomics : MCP* 2012;11:M111 014647.
- 17 [23] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING
18 database in 2021: customizable protein-protein networks, and functional characterization
19 of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;49:D605-d612.
- 20 [24] Bataller R, Arab JP, Shah VH. Alcohol-Associated Hepatitis. *N Engl J Med*
21 2022;387:2436-2448.
- 22 [25] Argemi J, Latasa MU, Atkinson SR, Blokhin IO, Massey V, Gue JP, et al. Defective
23 HNF4alpha-dependent gene expression as a driver of hepatocellular failure in alcoholic
24 hepatitis. *Nature communications* 2019;10:3126.

- 1 [26] Klein J, Eales J, Zurbig P, Vlahou A, Mischak H, Stevens R. Proteasix: a tool for
2 automated and large-scale prediction of proteases involved in naturally occurring peptide
3 generation. *Proteomics* 2013;13:1077-1082.
- 4 [27] Israelsen M, Misas MG, Koutsoumourakis A, Hall A, Covelli C, Buzzetti E, et al. Collagen
5 proportionate area predicts long-term mortality in patients with alcoholic hepatitis. *Dig
6 Liver Dis* 2022;54:663-668.
- 7 [28] Sato T, Head KZ, Li J, Dolin CE, Wilkey D, Skirtich N, et al. Fibrosis resolution in the
8 mouse liver: Role of Mmp12 and potential role of calpain 1/2. *Matrix Biology Plus*
9 2023;17:100127.
- 10 [29] Argemi J, Kedia K, Gritsenko MA, Clemente-Sanchez A, Asghar A, Herranz JM, et al.
11 Integrated Transcriptomic and Proteomic Analysis Identifies Plasma Biomarkers of
12 Hepatocellular Failure in Alcohol-Associated Hepatitis. *Am J Pathol* 2022;192:1658-1669.
- 13 [30] Banfi G, Salvagno GL, Lippi G. The role of ethylenediamine tetraacetic acid (EDTA) as in
14 vitro anticoagulant for diagnostic purposes. *Clin Chem Lab Med* 2007;45:565-576.
- 15 [31] Pan Y, Lei T, Teng B, Liu J, Zhang J, An Y, et al. Role of vimentin in the inhibitory effects
16 of low-molecular-weight heparin on PC-3M cell adhesion to, and migration through,
17 endothelium. *J Pharmacol Exp Ther* 2011;339:82-92.

18

1 **Acknowledgements:** Access to healthy control and AH consortium samples provided by Craig
2 McClain, MD and Vatsalya Vatsalya, MD through the University of Louisville Alcohol Research
3 Center (P50 AA024337). Supported, in part, by grants from NIH (R01 DK130294, R01
4 AA021978, R01 HL157879, R01 HL127349, P20 GM113226, P30 DK120531).

5
6 **Author Contributions:** KS: Visualization, Investigation, Validation, Formal Analysis, Writing-
7 Original Draft. CED: Investigation, Validation, Formal Analysis, Visualization, Writing-Original
8 Draft. DWW: Investigation, Validation. JL: Visualization, Investigation, Validation, Formal
9 Analysis. TS: Investigation, Validation, Formal Analysis. JIB: Investigation, Validation, Formal
10 Analysis. JA: Visualization, Investigation, Resources. RB: Visualization, Investigation,
11 Resources. ASW: Visualization, Formal Analysis, Writing-Review and Editing. . MLM: Project
12 Administration, Visualization, Conceptualization, Investigation, Supervision, Writing-Review and
13 Editing, Funding acquisition, Resources. PVB: Project Administration, Visualization,
14 Conceptualization, Investigation, Supervision, Writing-Review and Editing, Funding acquisition,
15 Resources. GEA: Project Administration, Visualization, Conceptualization, Investigation,
16 Supervision, Writing-Review and Editing, Funding acquisition, Resources.

17
18 **Data availability statement:** Proteomic files were deposited in MassIVE
19 (<http://massive.ucsd.edu/>) as study (MassIVE MSV000093513) entitled “Alcoholic hepatitis
20 plasma degradome”. Data include (A) the primary data files (.RAW), (B) peak list files (.mzML),
21 (C) sample key, (D) the sequence databases (human UniprotKB reviewed reference
22 proteomes), and (E) excel files containing Peaksdb results for de novo peptide sequence
23 assignment. The shared data will be released from private embargo for public access upon the
24 manuscript's acceptance for publication. All other data will be made available on request.

25

- 1 **Additional Information:** The authors declare that they have no known competing financial
- 2 interests or personal relationships that could have appeared to influence the work reported in
- 3 this paper.

1 **Table 1: Baseline Demographic and Characteristics by Disease Severity**

Variable	All N=63*	Moderate*** N=28	Severe*** N=34	p-value
Age (in years)	N=62	N=28	N=34	0.64
Median (IQR)	50.5 (44.0 : 57.0)	51.0 (44.5 : 58.0)	50.0 (41.0 : 56.0)	
Sex	N=62	N=28	N=34	0.43
Male	41 (66.1%)	17 (60.7%)	24 (70.6%)	
Female	21 (33.9%)	11 (39.3%)	10 (29.4%)	
Race	N=62	N=28	N=34	0.12
White	56 (90.3%)	28 (100.0%)	28 (82.4%)	
Black/African-American	4 (6.5%)	0 (0.0%)	4 (11.8%)	
Asian/Asian-American	1 (1.6%)	0 (0.0%)	1 (2.9%)	
Other	1 (1.6%)	0 (0.0%)	1 (2.9%)	
LTDH	N=51	N=23	N=28	0.18
Median (IQR)	26.0 (10.0 : 34.0)	30.0 (15.0 : 36.0)	16.5 (10.0 : 33.5)	
MELD Score	N=62	N=28	N=34	<0.001
Median (IQR)	22.0 (17.0 : 26.0)	16.0 (13.0 : 19.0)	26.0 (24.0 : 28.0)	
Maddrey's Discriminant Function	N=61	N=27	N=34	<0.001
Median (IQR)	42.3 (18.5 : 59.7)	18.0 (9.5 : 26.5)	58.0 (47.9 : 66.3)	
CTP/Child-Pugh Score	N=61	N=27	N=34	<0.001
Median (IQR)	10.0 (9.0 : 11.0)	9.0 (8.0 : 10.0)	11.0 (10.0 : 12.0)	
AST (SGOT) (IU/L)	N=62	N=28	N=34	0.14
Median (IQR)	113.5 (85.0 : 178.0)	109.0 (72.0 : 160.0)	126.0 (97.0 : 186.0)	
ALT (SGPT) (IU/L)	N=62	N=28	N=34	0.64
Median (IQR)	41.0 (29.0 : 66.0)	41.5 (27.5 : 72.5)	39.0 (32.0 : 64.0)	
Alkaline phosphatase (IU/L)	N=62	N=28	N=34	0.32
Median (IQR)	150.5 (119.0 : 207.0)	144.0 (98.0 : 201.5)	156.5 (133.0 : 209.0)	
Bilirubin (mg/dL)	N=62	N=28	N=34	<0.001
Median (IQR)	13.6 (5.2 : 18.8)	4.9 (2.9 : 7.1)	18.6 (14.1 : 23.4)	
Creatinine (mg/dL)	N=62	N=28	N=34	0.07
Median (IQR)	0.7 (0.6 : 1.1)	0.7 (0.5 : 0.9)	0.8 (0.6 : 1.3)	
Albumin (g/L)	N=62	N=28	N=34	0.09
Median (IQR)	2.6 (2.3 : 2.9)	2.6 (2.4 : 3.0)	2.4 (2.3 : 2.7)	
Ascites	N=61	N=27	N=34	0.001
0	17 (27.9%)	13 (48.1%)	4 (11.8%)	
1	37 (60.7%)	14 (51.9%)	23 (67.6%)	
2	7 (11.5%)	0 (0.0%)	7 (20.6%)	
Encephalopathy	N=61	N=27	N=34	0.008
0	46 (75.4%)	25 (92.6%)	21 (61.8%)	
1	14 (23.0%)	2 (7.4%)	12 (35.3%)	

Variable	All N=63*	Moderate*** N=28	Severe*** N=34	p-value
2	1 (1.6%)	0 (0.0%)	1 (2.9%)	
Patients PT/INR	N=61	N=27	N=34	<0.001
Median (IQR)	1.7 (1.4 : 2.0)	1.4 (1.2 : 1.6)	2.0 (1.7 : 2.2)	
Total Protein (g/dL)	N=62	N=28	N=34	0.14
Median (IQR)	6.0 (5.5 : 6.6)	6.3 (5.6 : 7.4)	5.9 (5.5 : 6.4)	
AUDIT Total	N=55	N=26	N=29	0.54
Median (IQR)	24.0 (18.0 : 29.0)	23.0 (16.0 : 29.0)	25.0 (19.0 : 29.0)	
Vital status at 90 days**	N=58	N=26	N=31	0.003
Alive	45 (77.6%)	25 (96.2%)	19 (61.3%)	
Dead	13 (22.4%)	1 (3.8%)	12 (38.7%)	

1

2 *Clinical information including MELD score used for Severe and moderate classification was
3 unavailable for one participant

4 ** 90-day vital status was missing for five participants.

5 *** Moderate: MELD score 12-20; Severe: MELD score > 20.

6

7 Patient baseline and clinical characteristics were summarized by disease severity using
8 frequencies and percentages for categorical variables and using median and 25th and 75th
9 percentiles (referred to as IQR) for continuous variables. The distribution of categorical variables
10 across moderate and severe AH patients was compared using chi-square or Fisher's exact tests
11 whereas that of the continuous variables was compared using Wilcoxon's test.

12

13 Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUDIT,
14 alcohol use disorders identification test; CTP, Child-Turcotte-Pugh; NR, international normalized
15 ratio; LTDH, lifetime drinking history; MELD, model for end-stage liver disease; PT, prothrombin
16 time.

1 **Table 2: Causal graphical modeling validation results.** Sensitivity, specificity, and accuracy
2 values are provided for each fold of each model, while average values are also given. Red/bold
3 font designates the maximum average value of each metric.

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Model 1 (MELD only)	Sensitivity	0.8889	0.8889	0.8889	0.4444	0.6667	0.7556
	Specificity	1.0000	0.3333	0.3333	1.0000	1.0000	0.7333
	Accuracy	0.9444	0.6111	0.6111	0.7222	0.8333	0.7444
Model 2 (Peptides only)	Sensitivity	0.7778	0.8889	0.6667	0.6667	1.0000	0.8000
	Specificity	1.0000	0.6667	0.6667	0.5000	1.0000	0.7666
	Accuracy	0.8889	0.7778	0.6667	0.5833	1.0000	0.7833
Model 3 (Peptides + MELD)	Sensitivity	0.8889	0.8889	0.8889	0.7778	0.7778	0.8444
	Specificity	1.000	1.0000	0.3333	0.5000	1.0000	0.7666
	Accuracy	0.9444	0.9444	0.6111	0.6389	0.8889	0.8055

4
5
6
7

1 **Figure Legends**

2 **Figure 1. Study design and peptidome.**

3 Panel A: Consort Diagram.

4 Panel B: Analytic workflow. Plasma proteins were precipitated with trichloroacetic acid (TCA).

5 The peptidome was concentrated and desalted using solid phase extraction prior to data

6 collection by high resolution, high mass accuracy UPLC-MS. Database and de novo MS

7 spectral assignments were made using Peaks Xpro. Raw peptide abundances were normalized

8 based on total extracted ion chromatograms (XIC) and then preprocessed within Metaboanalyst.

9 Data were mined by univariate and multivariate statistical methods for differentially abundant

10 peptides and peptide groups, for Gene Ontology (Panther), parent protein matrixomal

11 composition (MatrisomeAnnotator) and for protease involvement (Proteasix). Machine learning

12 methods were initiated with patient-specific TIC normalized peptidome and clinical scoring data

13 (e.g., MELD, 90day mortality). Data were preprocessed to address missing values and leave-

14 one out causal graphs to building a selected variables data set. The performance of the

15 selected variables with or without MELD scores was compared to MELD alone using a 5-fold

16 validation and logistic regression to establish model parameters and prediction of 90-day

17 mortality in AH.

18 Panel C: Plasma peptidome analysis by Venn diagram for prevalence (AH moderate vs. AH
19 severe).

20 Panel D: Volcano plot for significant differences ($FC > \pm 1.5$; $p < 0.05$). Significant peptide data

21 points were labeled using the gene name. The analysis defines shifts of increased

22 fibrinopeptide A (FBA) in moderate and increased collagen (e.g., CO1A1) peptides in severe

23 AH.

24

25 **Figure 2. Plasma peptidome features analysis.**

1 Panel A: PCA analysis showing principal components PC1 and PC2 for self-sorting of healthy
2 control (green), moderate AH (blue) and severe AH (red) samples as defined by 95%
3 confidence intervals. Healthy control samples are resolved from AH samples.

4 Panel B: Two-group analysis of moderate AH versus severe AH samples demonstrates
5 emerging self-sorting properties of the peptidome.

6 Panel C: oPLS-DA analysis comparing AH severity. Complete separation of the moderate and
7 severe AH peptidomes is achieved using discriminate analysis.

8 Panel D: Major peptide features sorted by oPLS-DA of AH samples are prolyl-hydroxylated
9 CO1A1 fragments (severe AH) and FBA fragments (moderate). Peptide fragments defined by
10 parent protein Gene Name, amino acid (start, stop) location, and site-specific modifications: *,
11 prolyl hydroxylation; a, acetylation; d, dehydration

12 Panel E: Cluster analysis of the peptidome/degradome in AH. The peptides significantly
13 increased in AH were analyzed by the Proteasix (<http://www.proteasix.org>) algorithm using a
14 positive predictive value (PPV) cut-off to 80%. Protein-protein interaction network analysis of
15 regulated proteomic data sets (q-value <0.05) was performed using Search Tool for the
16 Retrieval of Interacting Genes/Proteins, STRING v11,[23] with the highest confidence score
17 (0.900). The resultant matrix of both Proteasix and STRING analyses were visualized using
18 Cytoscape v3.9.1. Node sizes of the predicted proteases represented the relative frequency
19 with which the top 16 proteases were predicted to mediate the observed cleavage (0.2-25%).
20 Node shape for the proteases represents protease family subtype: serine (diamond), cysteine
21 (square), aspartyl (parallelogram), and metalloproteases (octagon). Node color for protease
22 corresponds to the Log2FC (vs healthy control) of hepatic mRNA expression from previously
23 published work.[25] Raw data and metadata are publicly available in the Database of
24 Genotypes and Phenotypes of the National Library of Medicine under the accession study code
25 phs001807. Node sizes of the peptides represented the relative number of unique peptides (1-
26 61) identified from each parent protein. Node colors of the peptides represented the median

1 Log2FC vs healthy controls for all peptides derived from that parent protein. Solid lines depict
2 connections between the parent proteins identified by STRING; broken lines depict predicted
3 protease events identified by Proteasix.

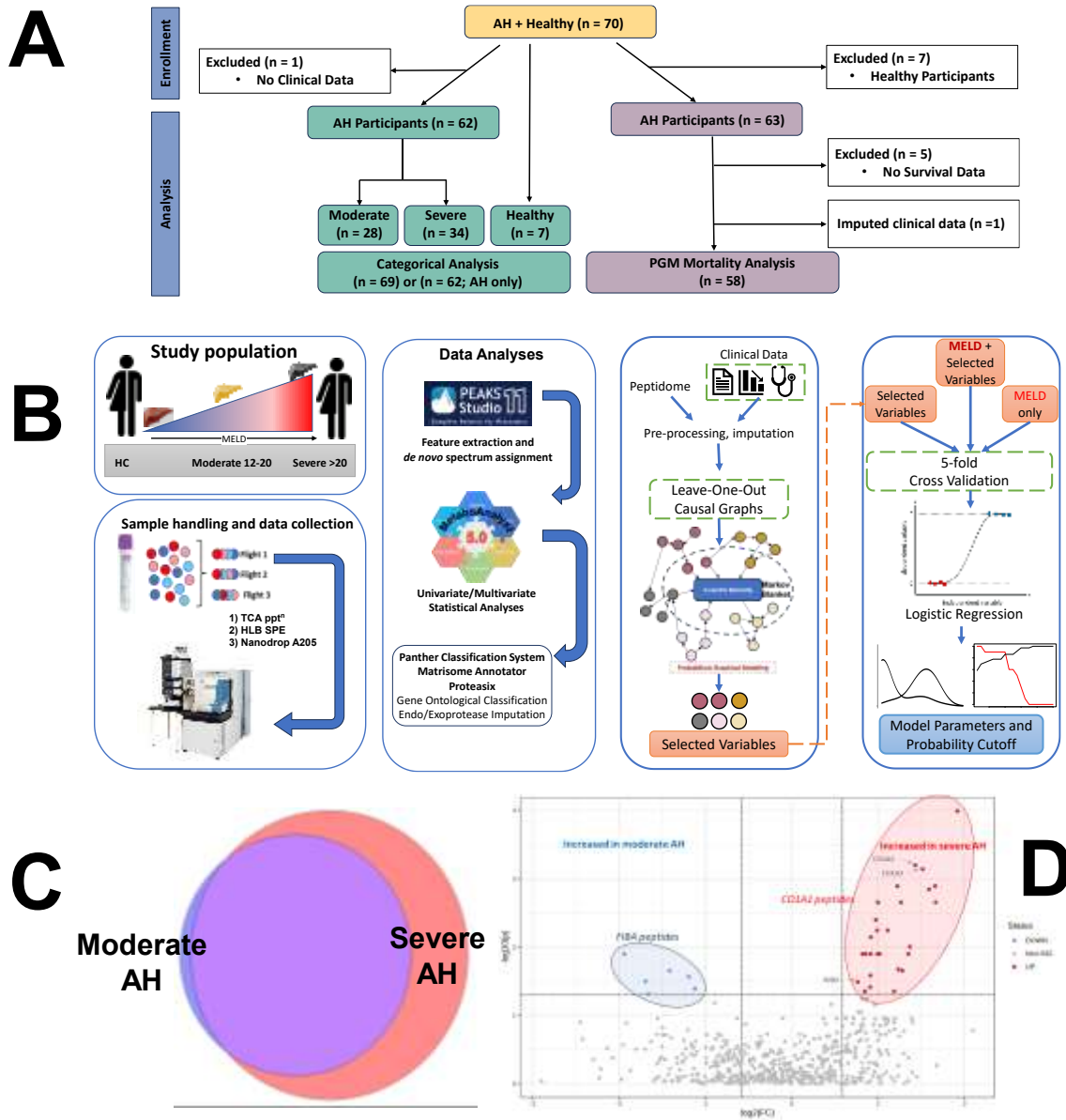
4

5 **Figure 3. CGM modeling of the peptidome and clinical features to predict AH outcome.**

6 Panel A: Sensitivity and Specificity of the 5-fold cross-validation during the prediction phase of
7 model development. X-axis: the threshold used in the parameter sweep (range 0.1-1.0). The
8 intersection of sensitivity and specificity was used to determine the optimal threshold for each
9 fold in each model.

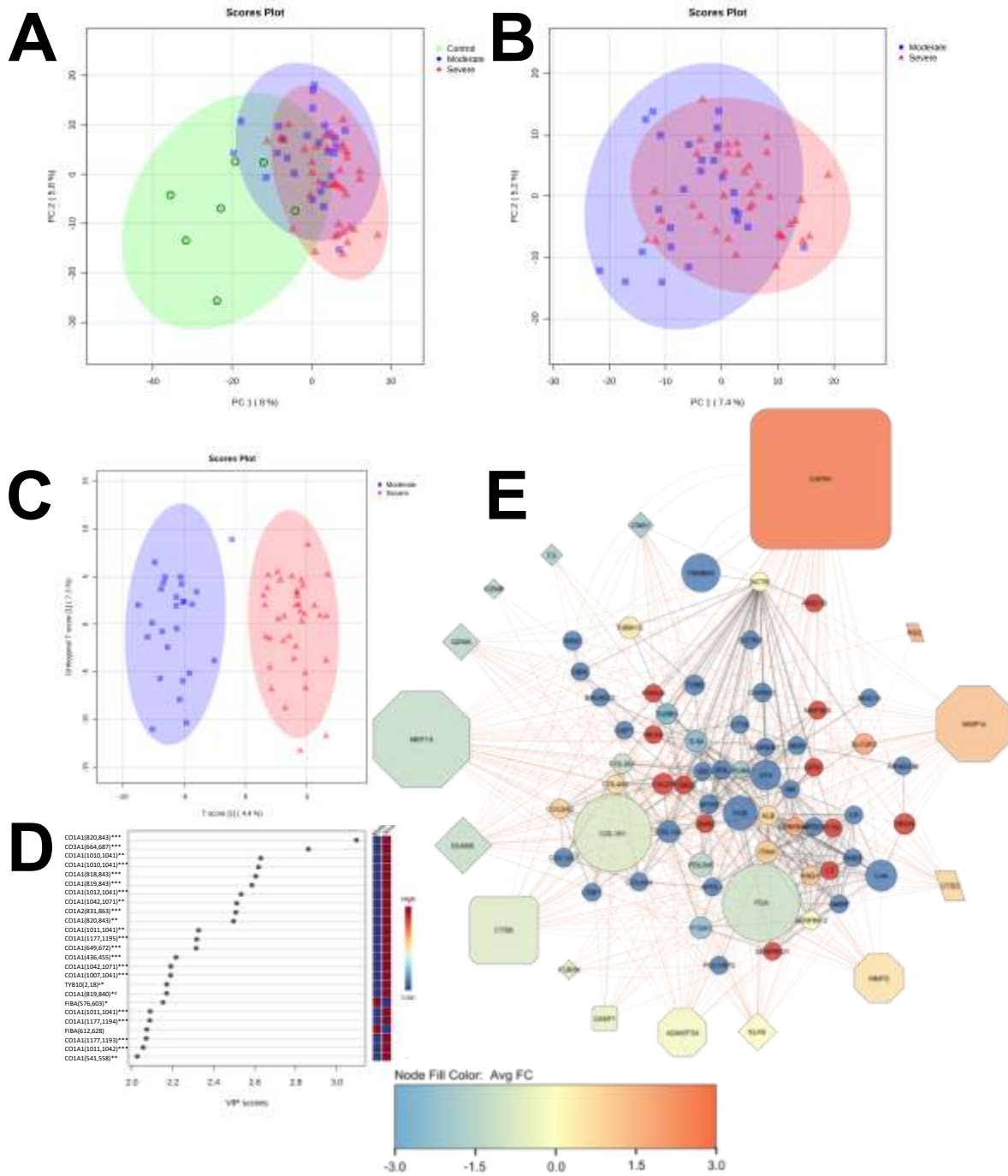
10 Panel B: Density distribution of 90-day survival classification over different cutoff probability
11 thresholds. Model 2 and Model 3 offer better separation of the two categories than the MELD
12 score alone.

13 Panel C: Comparison of model performance using the complete dataset. The tables show the
14 number of correctly and incorrectly classified samples. Sensitivity, specificity, and balanced
15 accuracy summarize these results. Kaplan-Maier survival plots depict the discrimination ability
16 of the three models.



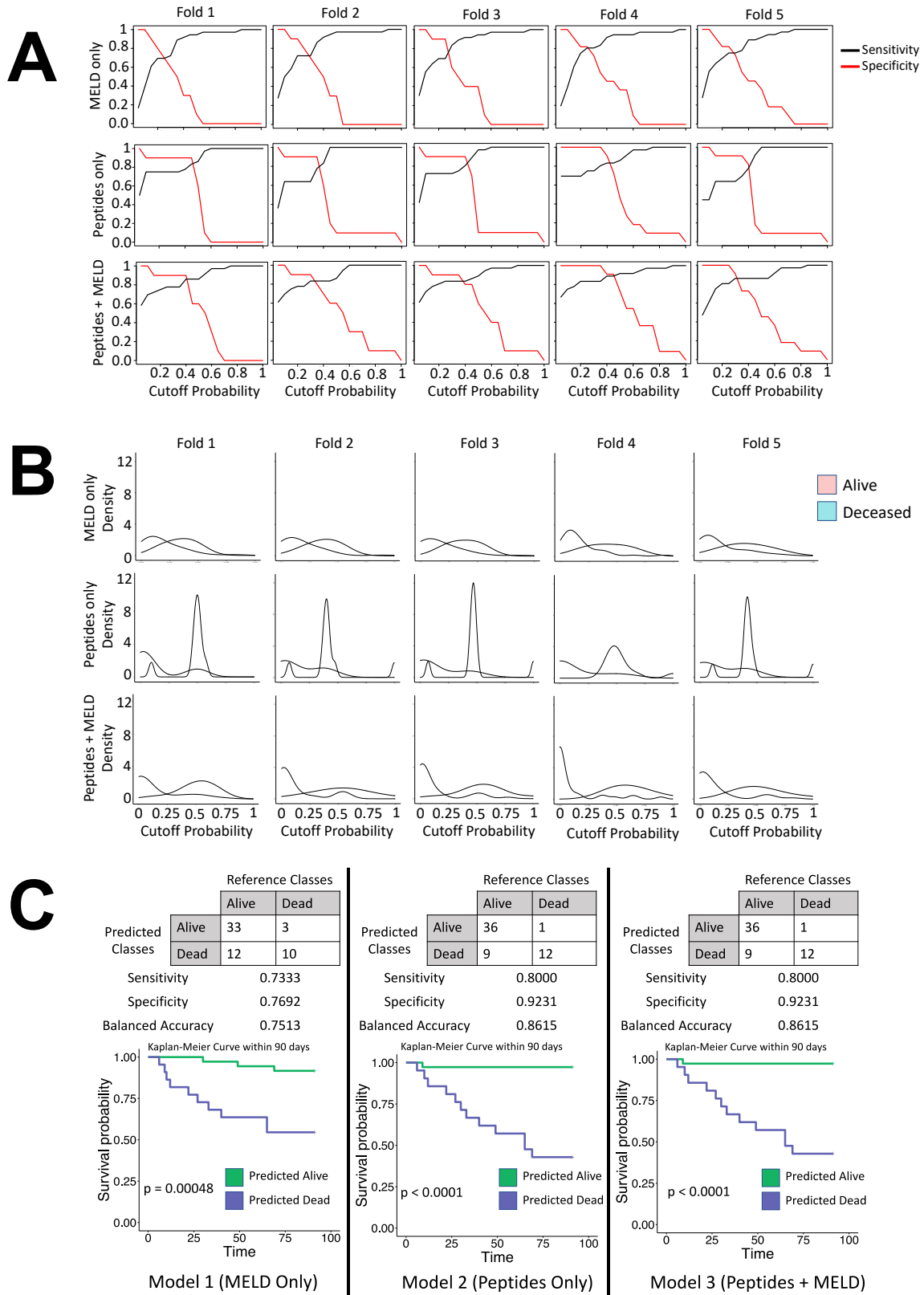
1
2
3
4
5
6
7
8

Figure 1. Sayed et al.



1
2
3
4
5
6

Figure 2. Sayed et al.



1

2

Figure 3. Sayed et al.