SUPPORTING INFORMATION

Table S1 Expert pathologist annotations collected for model training by substance. Annotation numbers refer to pixel level bounding boxes (regions) drawn by pathologists using the PathAI digital slide platform. Bolded annotations reflect those used as classes during model training.

Table S2 Manual scoring data for NAS components and fibrosis staging supporting intra-reader and inter- reader variability. Slide IDs indicate unique patient biopsies. Rows with repeated IDs were used to assess intra-pathologist reproducibility. Reading intervals specify time between reads of H&E and Trichrome slides.

Table S3 Intra-observer reproducibility of pathologist manual scores of NAS components and fibrosis staging. Values are weighted Cohen's kappa (1 is perfect agreement) for each pathologist's repeated measurements. Cohen's Kappa measures the rate of agreement while accounting for the degree of disagreement based on the ordinal gradation.

Table S4 Inter-observer reproducibility of fibrosis staging. Values are weighted Cohen's Kappa for each pathologist's and the model's score against the pathologist consensus. For each pathologist their score was computed against the consensus of the other

pathologists.

Table S5 Prognostic value of each ML feature for predicting progression to cirrhosis and adjudicated clinical events in STELLAR-3 and STELLAR-4, respectively. Each row represents results from a univariate Cox regression model. Confidence intervals (ci-low, and ci-high) represent the 95% confidence interval for the hazard ratio.

Table S6 Prognostic significance of feature clusters for prediction of adjudicated clinical events and progression to cirrhosis. Feature column lists all features assigned to each cluster via agglomerative clustering. Q column lists the Benjamini-Hochberg corrected p-values. P column lists nominal cluster-wise p-values derived from EBM and univariate Cox regression.

Figure S1 Confusion matrix indicating model performance for labeling held-out pathologist annotations on H&E slides. Each column in the x-axis represents the predicted label and each column in the y- axis represents the true label. Box colors indicate the accuracy (0.0, white to purple, 1.0). Normal liver represents an amalgam of normal liver histologic features including normal hepatocytes (**Table S1**).

Figure S2 Associations between ML-based model measurements and grading of NAS features by the CP according to trial. Model values describe the percentage of tissue predicted to be the substance in question (steatosis, lobular inflammation, or hepatocellular ballooning). Boxes show the IQR and whiskers show 1.5 x the limit of the IQR. Points show values beyond this range. Values shown are Spearman correlations (rho) and corresponding p-values. No models were trained using the ATLAS dataset.

Figure S3 Confusion matrix indicating model performance for labeling held-out pathologist annotations on trichrome-stained slides. Each column in the x-axis represents the predicted label and each column in the y-axis represents the true label. Box colors indicate the accuracy (0.0, white to purple, 1.0). Confusion between starred classes (Blood Vessels and Lumen) are expected since these annotations are overlapping.

Figure S4 Associations between ML-based model measurements and staging of fibrosis by the CP according to A) the Ishak classification in all studies, and B) the Ishak and NASH CRN classifications by trial. Model values describe the weighted average score for each patient by pathologist-derived fibrosis stage. The weighted average score is computed by multiplying each fraction of tissue area by its corresponding stage and summing across the slide (**Methods**). Boxes show the IQR

and whiskers show 1.5 x the limit of the IQR. Points show values beyond this range. Values shown are Spearman correlations (rho) and corresponding p-values. No models were trained using the ATLAS dataset.

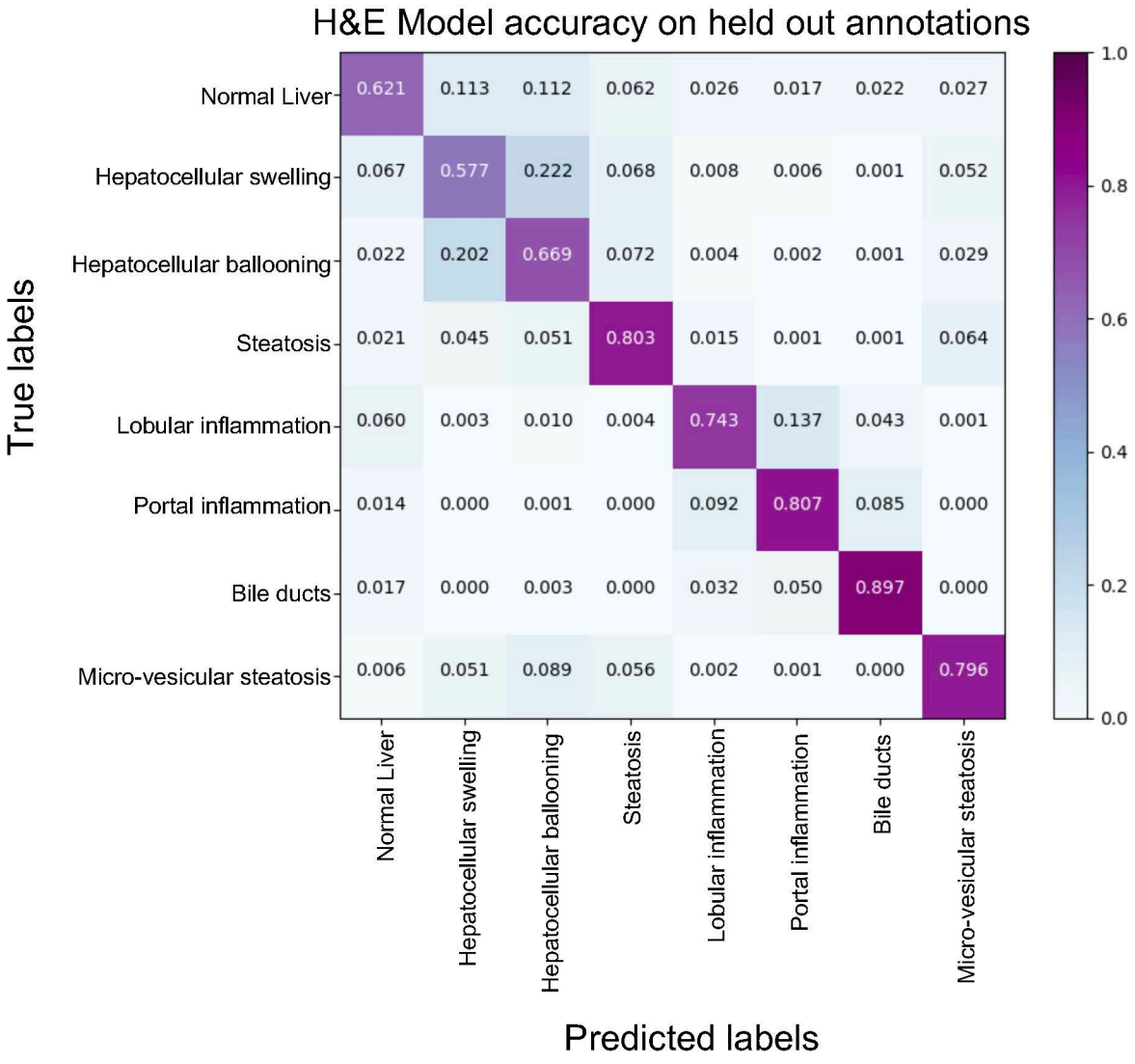
Figure S5 Reproducibility and variability of Ishak fibrosis staging by pathologists in 166 slides. A) Bar chart showing intra-rater agreement (weighted Cohen's kappa) for Ishak fibrosis stage. B) Bar chart showing inter-rater agreement (weighted Cohen’s kappa) for Ishak fibrosis stage. Each pathologist’s agreement is measured with all other pathologists in the study.

Figure S6 Feature clustering and significance testing for patient prognosis in STELLAR-3 and STELLAR- 4. A) Heatmap showing the inter-feature absolute Spearman correlation (0=black, white=1). Features are sorted using agglomerative clustering. B) “Elbow” plot showing the number of clusters by height of the dendrogram in (A). Blue line indicates selected cut point corresponding to 22 clusters. C and D) Quantile-quantile plots of cluster-wise nominal p-values (blue dots) for association with progression to cirrhosis in patients with bridging (F3) fibrosis (STELLAR-3) or liver-related clinical events in patients with cirrhosis (F4; STELLAR-4). Significant clusters (FDR corrected q-value) are denoted by red circles. P-values are sorted and plotted against the corresponding uniform quantile value.

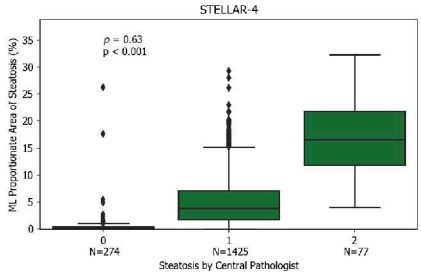
Figure S7 Correspondence of DELTA Liver Fibrosis score with markers of treatment response. A-C) Box and whisker plot showing the difference in DELTA Liver Fibrosis score for patients in the ATLAS trial by treatment group (x-axis) and achievement of a ≥1-stage improvement in NASH CRN fibrosis stage as evaluated by the CP (A), reduction of at least 0.5 in ELF score (B) and 25% reduction in FibroScan (C). Boxes show the IQR and whiskers show 1.5 x the limit of the IQR. P-values are computed using the Mann-Whitney test.

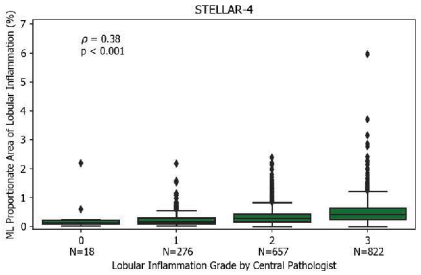
Figure S8 Bar charts comparing the proportion of patients with a reduction in fibrosis as assessed by the DELTA Liver Fibrosis score and the CP in experimental arms (red) compared with placebo-treated patients (grey) in ATLAS. P-values are computed using Fisher's exact test.

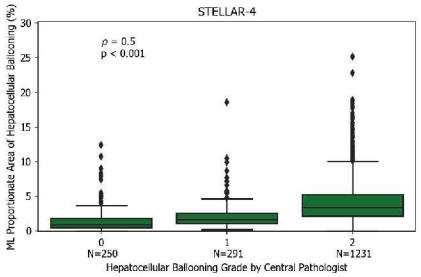
S. Figure 1

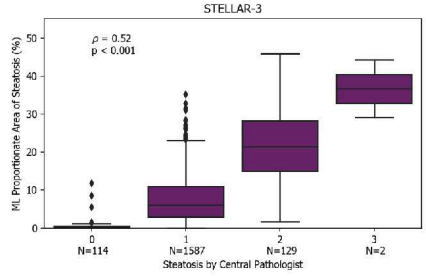

H&E Model accuracy on held out annotations
True labels
Predicted labels
Normal Liver
Hepatocellular swelling
Hepatocellular ballooning
Steatosis
Lobular inflammation
Portal inflammation
Bile ducts
Micro-vesicular steatosis
0.621 0.113 0.112 0.062 0.026 0.017 0.022 0.027
0.067 0.577 0.222 0.068 0.008 0.006 0.001 0.052
0.022 0.202 0.669 0.072 0.004 0.002 0.001 0.029
0.021 0.045 0.051 0.803 0.015 0.001 0.001 0.064
0.060 0.003 0.010 0.004 0.743 0.137 0.043 0.001
0.014 0.000 0.001 0.000 0.092 0.807 0.085 0.000
0.017 0.000 0.003 0.000 0.032 0.050 0.897 0.000
0.006 0.051 0.089 0.056 0.002 0.001 0.000 0.796
1.0
0.8
0.6
0.4
0.2
0.0

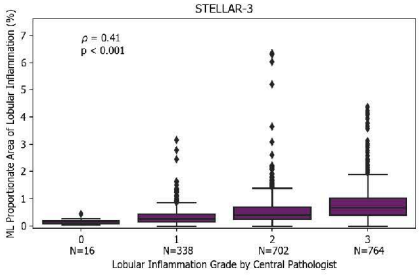
S. Figure 2

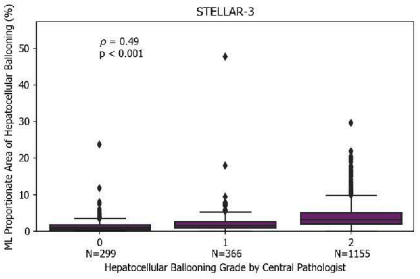

STELLAR-4
ρ = 0.63
p < 0.001
ML Proportionate Area of Steatosis (%)
0 N=274
1 N=1425
2 N=77
Steatosis by Central Pathologist

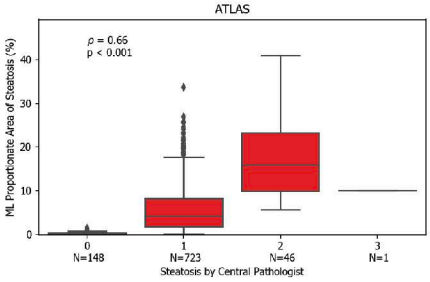

STELLAR-4
ρ = 0.38
p < 0.001
ML Proportionate Area of Lobular Inflammation (%)
0 N=18
1 N=276
2 N=657
3 N=822
Lobular Inflammation Grade by Central Pathologist

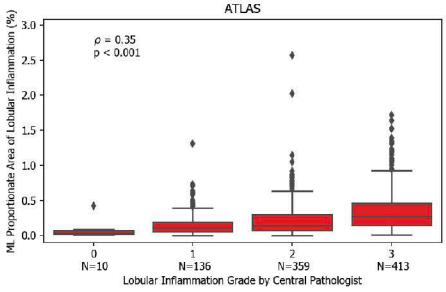

STELLAR-4
ρ = 0.5
p < 0.001
ML Proportionate Area of Hepatocellular Ballooning (%)
0 N=250
1 N=291
2 N=1231
Hepatocellular Ballooning Grade by Central Pathologist

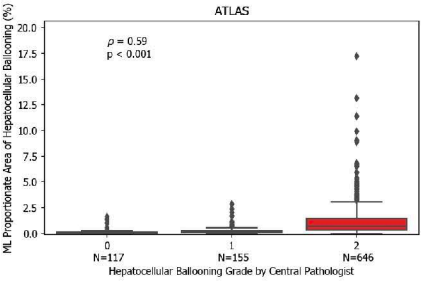

STELLAR-3
ρ = 0.52
p < 0.001
ML Proportionate Area of Steatosis (%)
0 N=114
1 N=1587
2 N=129
3 N=2
Steatosis by Central Pathologist


STELLAR-3
ρ = 0.41
p < 0.001
ML Proportionate Area of Lobular Inflammation (%)
0 N=16
1 N=338
2 N=702
3 N=764
Lobular Inflammation Grade by Central Pathologist


STELLAR-3
ρ = 0.49
p < 0.001
ML Proportionate Area of Hepatocellular Ballooning (%)
0 N=299
1 N=366
2 N=1155
Hepatocellular Ballooning Grade by Central Pathologist


ATLAS
ρ = 0.66
p < 0.001
ML Proportionate Area of Steatosis (%)
0 N=148
1 N=723
2 N=46
3 N=1
Steatosis by Central Pathologist


ATLAS
ρ = 0.35
p < 0.001
ML Proportionate Area of Lobular Inflammation (%)
0 N=10
1 N=136
2 N=359
3 N=413
Lobular Inflammation Grade by Central Pathologist

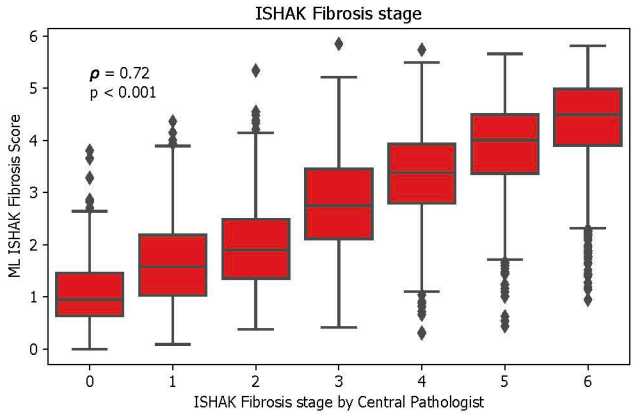

ATLAS
ρ = 0.59
p < 0.001
ML Proportionate Area of Hepatocellular Ballooning (%)
0 N=117
1 N=155
2 N=646
Hepatocellular Ballooning Grade by Central Pathologist

S. Figure 3

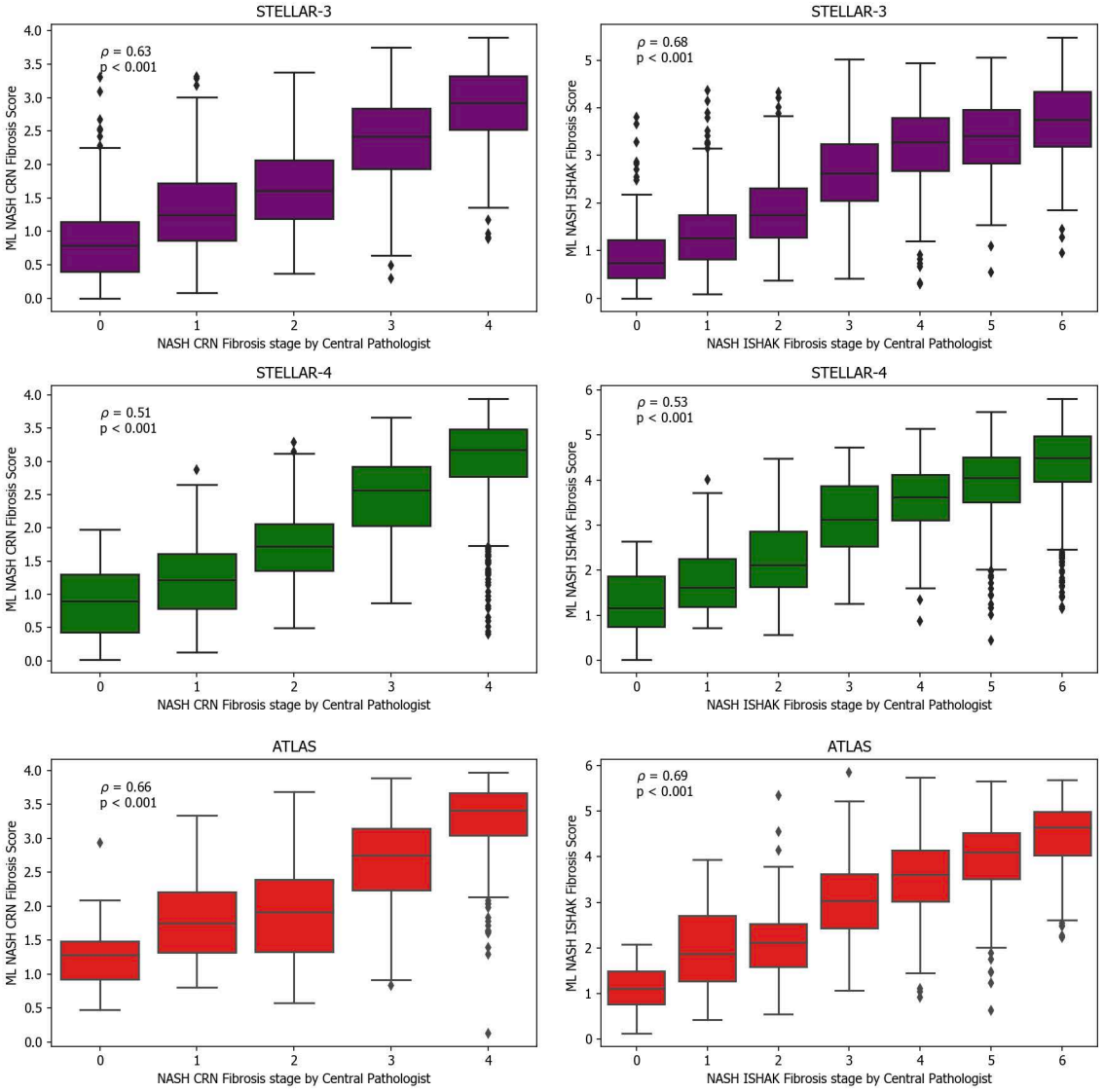

Trichrome Model accuracy on held out annotations
True labels
Predicted labels
Normal Liver
Fibrosis
Bile ducts
*Lumen
*Blood vessel
0.844 0.021 0.019 0.064 0.052
0.037 0.893 0.012 0.011 0.047
0.073 0.025 0.873 0.009 0.020
0.050 0.003 0.008 0.879 0.060
0.101 0.025 0.033 0.419 0.422
Normal Liver
Fibrosis
Bile ducts
*Lumen
*Blood vessels
1.0
0.8
0.6
0.4
0.2
0.0

S. Figure 4

a

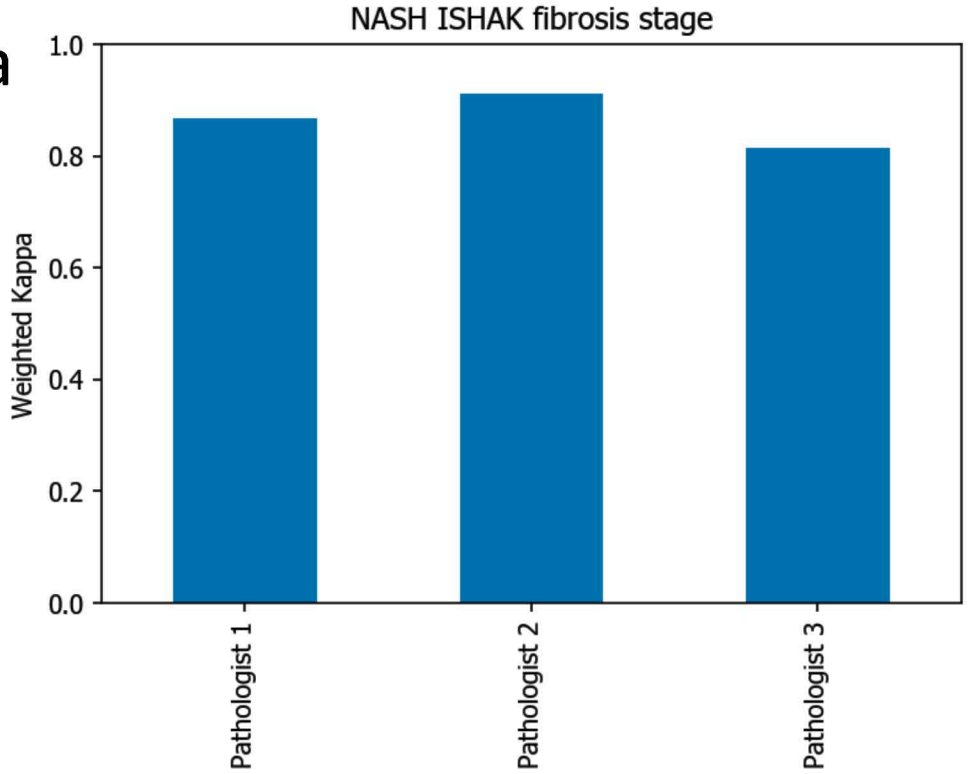

ISHAK Fibrosis stage
ρ = 0.72
p < 0.001
ML ISHAK Fibrosis Score
ISHAK Fibrosis stage by Central Pathologist

b

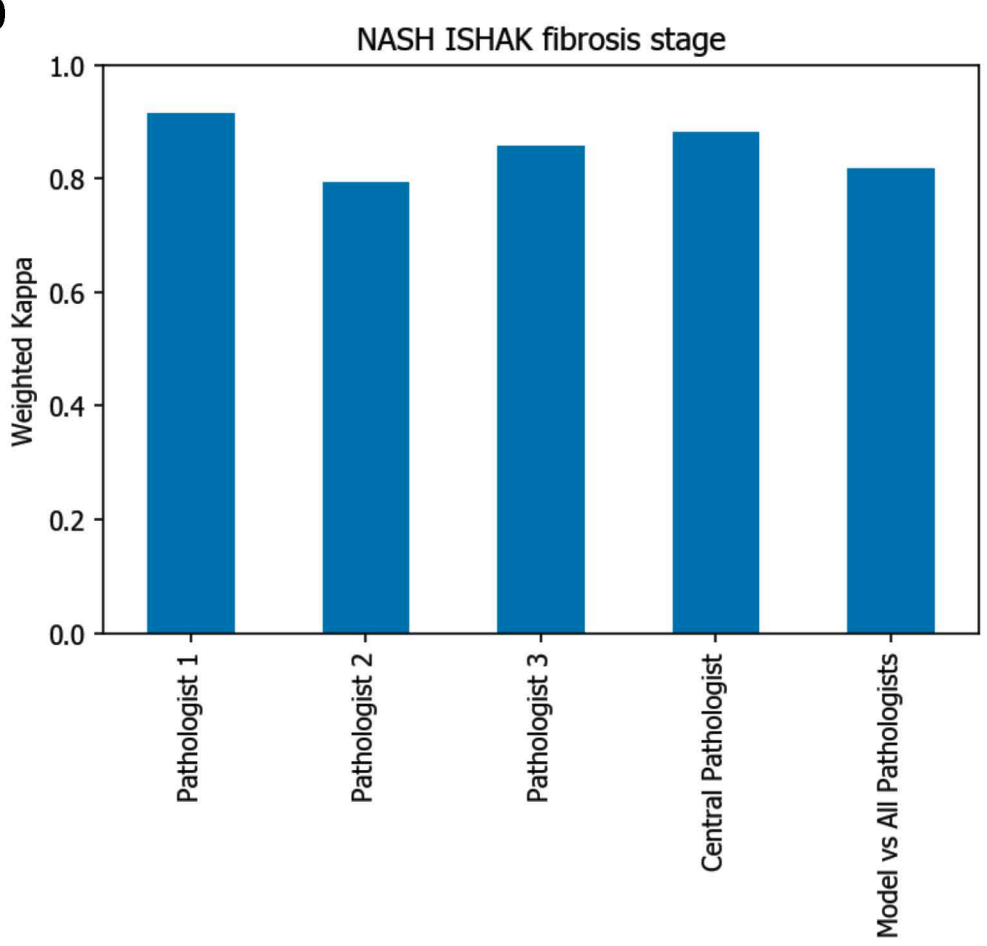

STELLAR-3
ρ = 0.63
p < 0.001
ML NASH CRN Fibrosis Score
NASH CRN Fibrosis stage by Central Pathologist
STELLAR-3
ρ = 0.68
p < 0.001
ML NASH ISHAK Fibrosis Score
NASH ISHAK Fibrosis stage by Central Pathologist
STELLAR-4
ρ = 0.51
p < 0.001
ML NASH CRN Fibrosis Score
NASH CRN Fibrosis stage by Central Pathologist
STELLAR-4
ρ = 0.53
p < 0.001
ML NASH ISHAK Fibrosis Score
NASH ISHAK Fibrosis stage by Central Pathologist
ATLAS
ρ = 0.66
p < 0.001
ML NASH CRN Fibrosis Score
NASH CRN Fibrosis stage by Central Pathologist
ATLAS
ρ = 0.69
p < 0.001
ML NASH ISHAK Fibrosis Score
NASH ISHAK Fibrosis stage by Central Pathologist

S. Figure 5

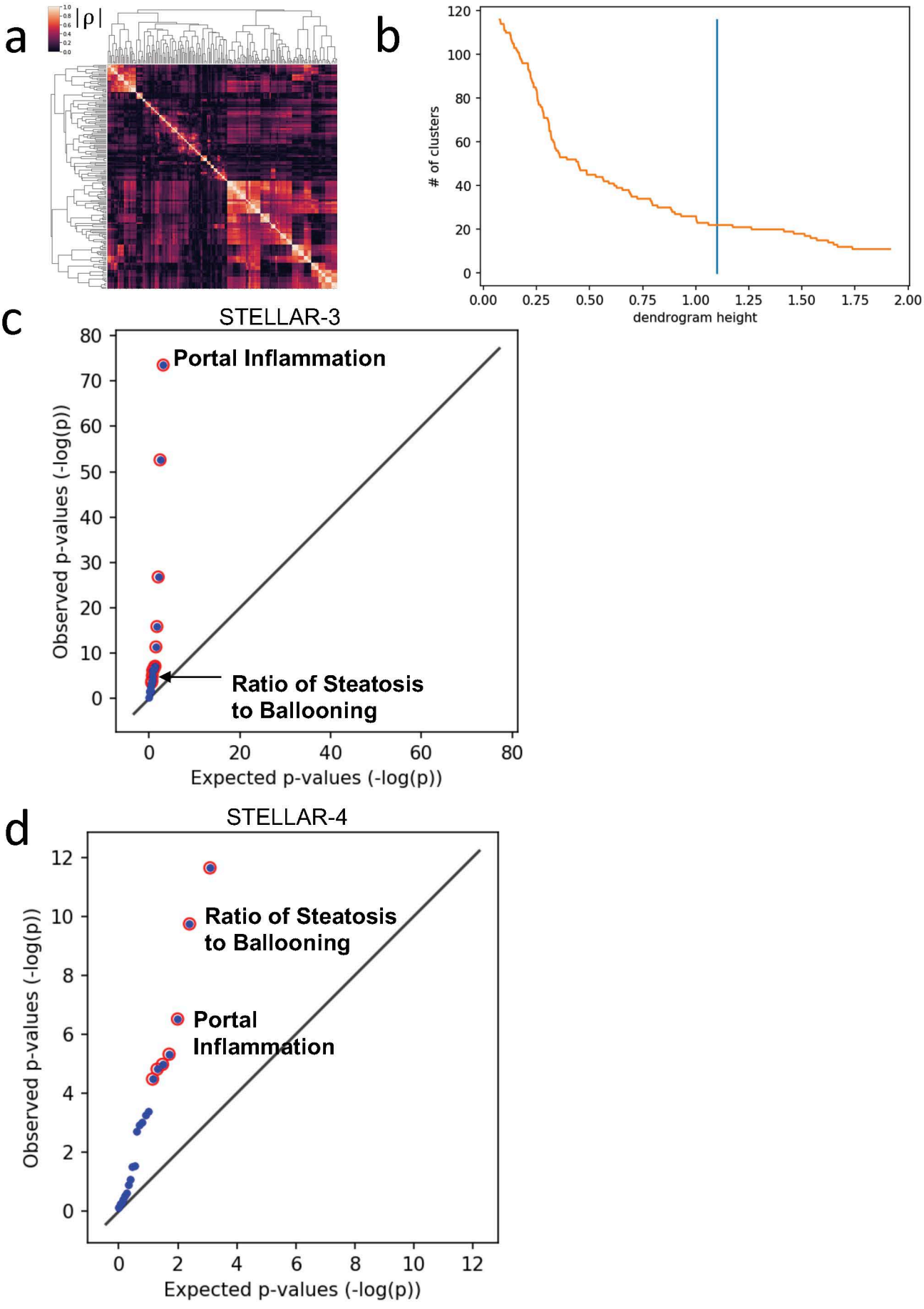
a


NASH ISHAK fibrosis stage
Weighted Kappa
1.0
0.8
0.6
0.4
0.2
0.0
Pathologist 1
Pathologist 2
Pathologist 3

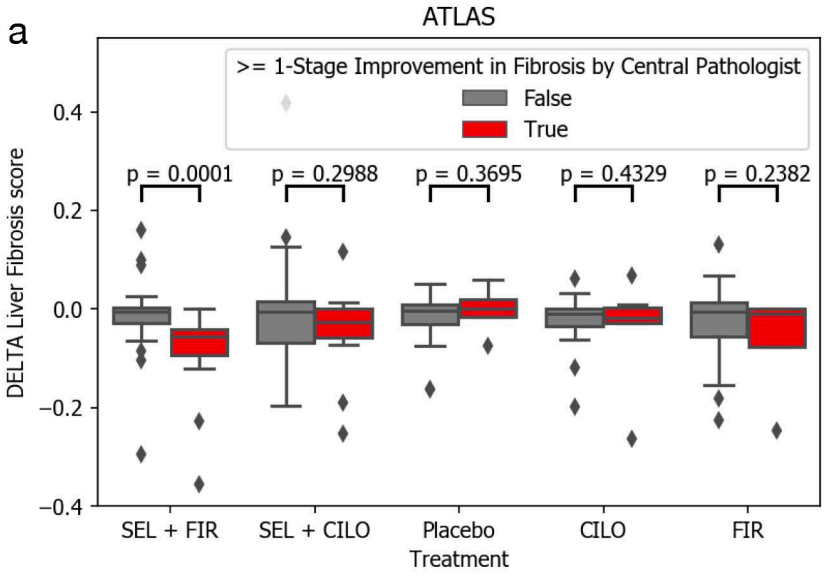
b

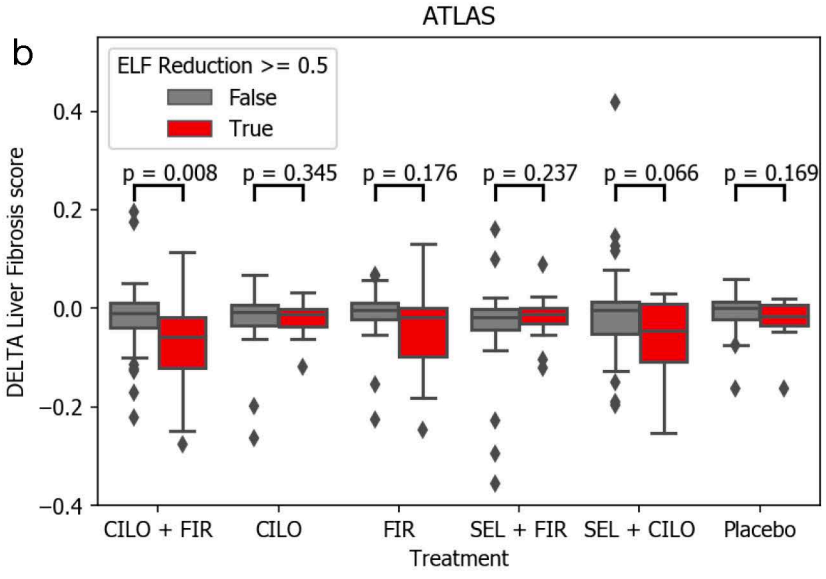

NASH ISHAK fibrosis stage
Weighted Kappa
1.0
0.8
0.6
0.4
0.2
0.0
Pathologist 1
Pathologist 2
Pathologist 3
Central Pathologist
Model vs All Pathologists

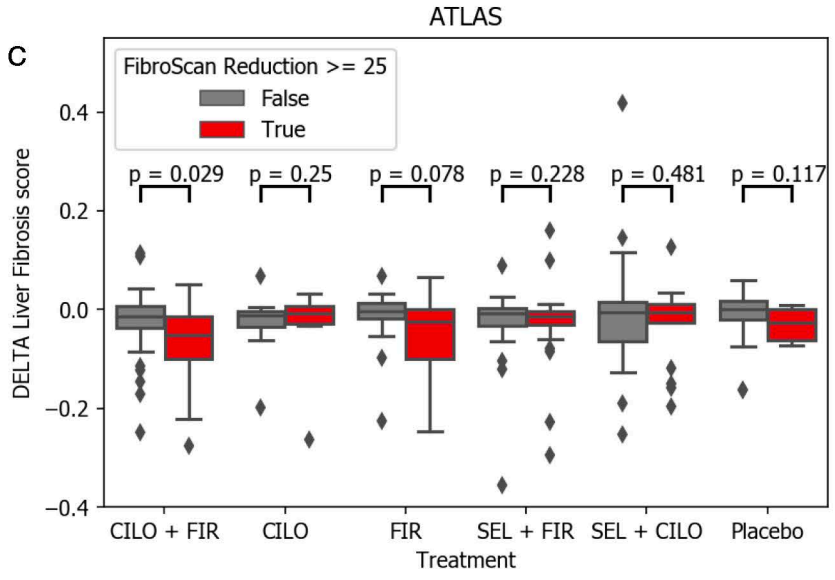
S Figure 6


a
|ρ|
b
# of clusters
dendrogram height
c
STELLAR-3
Portal Inflammation
Ratio of Steatosis to Ballooning
Observed p-values (-log(p))
Expected p-values (-log(p))
d
STELLAR-4
Ratio of Steatosis to Ballooning
Portal Inflammation
Observed p-values (-log(p))
Expected p-values (-log(p))

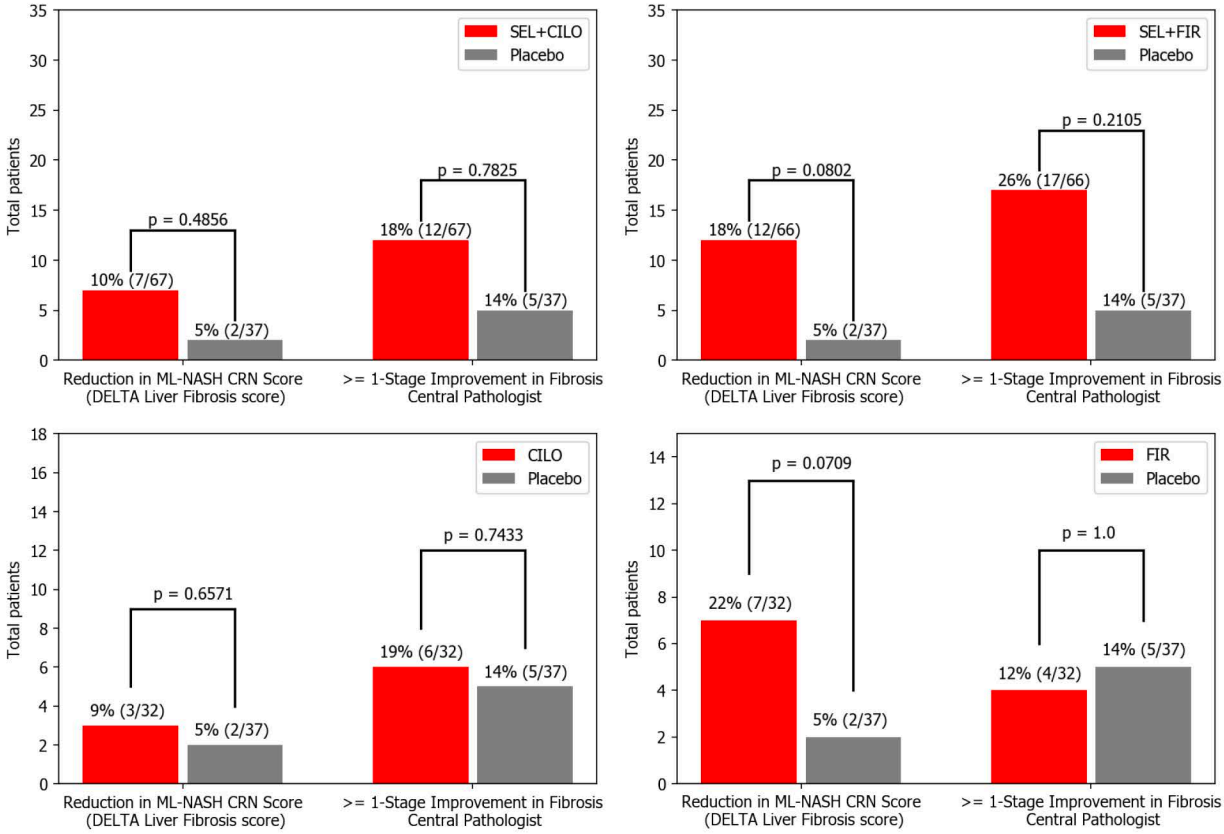
S Figure 7


a
ATLAS
>= 1-Stage Improvement in Fibrosis by Central Pathologist
False
True
p = 0.0001
p = 0.2988
p = 0.3695
p = 0.4329
p = 0.2382
DELTA Liver Fibrosis score
0.4
0.2
0.0
−0.2
−0.4
SEL + FIR
SEL + CILO
Placebo
CILO
FIR
Treatment


b
ATLAS
ELF Reduction >= 0.5
False
True
p = 0.008
p = 0.345
p = 0.176
p = 0.237
p = 0.066
p = 0.169
DELTA Liver Fibrosis score
0.4
0.2
0.0
−0.2
−0.4
CILO + FIR
CILO
FIR
SEL + FIR
SEL + CILO
Placebo
Treatment


c
ATLAS
FibroScan Reduction >= 25
False
True
p = 0.029
p = 0.25
p = 0.078
p = 0.228
p = 0.481
p = 0.117
DELTA Liver Fibrosis score
0.4
0.2
0.0
−0.2
−0.4
CILO + FIR
CILO
FIR
SEL + FIR
SEL + CILO
Placebo
Treatment

S Figure 8


SEL+CILO
Placebo
Total patients
p = 0.4856
10% (7/67)
5% (2/37)
p = 0.7825
18% (12/67)
14% (5/37)
Reduction in ML-NASH CRN Score (DELTA Liver Fibrosis score)
>= 1-Stage Improvement in Fibrosis Central Pathologist
SEL+FIR
Placebo
Total patients
p = 0.0802
18% (12/66)
5% (2/37)
p = 0.2105
26% (17/66)
14% (5/37)
Reduction in ML-NASH CRN Score (DELTA Liver Fibrosis score)
>= 1-Stage Improvement in Fibrosis Central Pathologist
CILO
Placebo
Total patients
p = 0.6571
9% (3/32)
5% (2/37)
p = 0.7433
19% (6/32)
14% (5/37)
Reduction in ML-NASH CRN Score (DELTA Liver Fibrosis score)
>= 1-Stage Improvement in Fibrosis Central Pathologist
FIR
Placebo
Total patients
p = 0.0709
22% (7/32)
5% (2/37)
p = 1.0
12% (4/32)
14% (5/37)
Reduction in ML-NASH CRN Score (DELTA Liver Fibrosis score)
>= 1-Stage Improvement in Fibrosis Central Pathologist