scientific reports

OPEN

Analyzing the relationship between gene expression and phenotype in space-flown mice using a causal inference machine learning ensemble

James A. Casaletto[1]✉, Ryan T. Scott[2], Makenna Myrick[3], Graham Mackintosh[4], Hamed Chok[1], Amanda Saravia-Butler[2], Adrienne Hoarfrost[5], Jonathan M. Galazka[6], Lauren M. Sanders[6] & Sylvain V. Costes[6]

Spaceflight has several detrimental effects on human and rodent health. For example, liver dysfunction is a common phenotype observed in space-flown rodents, and this dysfunction is partially reflected in transcriptomic changes. Studies linking transcriptomics with liver dysfunction rely on tools which exploit correlation, but these tools make no attempt to disambiguate true correlations from spurious ones. In this work, we use a machine learning ensemble of causal inference methods called the Causal Research and Inference Search Platform (CRISP) which was developed to predict causal features of a binary response variable from high-dimensional input. We used CRISP to identify genes robustly correlated with a lipid density phenotype using transcriptomic and histological data from the NASA Open Science Data Repository (OSDR). Our approach identified genes and molecular targets not predicted by previous traditional differential gene expression analyses. These genes are likely to play a pivotal role in the liver dysfunction observed in space-flown rodents, and this work opens the door to identifying novel countermeasures for space travel.

Rodent studies demonstrate that spaceflight negatively impacts liver function[1–3]. Astronaut studies including the seminal NASA Twins Study also reveal a theme of lipid dysregulation[4,5]. Despite these findings, there has been relatively little research studying the impact of microgravity or space radiation on the liver, with more research emphasis on central nervous system effects and carcinogenesis. This is a key knowledge gap considering the disruption of such a critical organ could impact astronaut health and jeopardize the success of future long-term space missions. Identifying the genetic and molecular mechanisms implicated in spaceflight-induced liver dysfunction is required as a first step in precisely mitigating those deleterious effects. Traditional statistical methods identify correlations which may or may not be spurious, especially in high-dimensional, high-throughput data analysis[6]. While randomized controlled trials are considered the gold standard for identifying non-spurious, causal relationships between dependent and independent variables[7], such experiments can be very expensive and time consuming, or logistically infeasible, especially in a spaceflight environment where sample sizes are limited. Instead, we turn to new machine learning (ML) approaches to identify genes in transcriptomic data predictive of a lipid metabolic response from spaceflight and ground control rodent liver samples.

Tools which are commonly used to analyze high-dimensional data, and ML algorithms in general, share an intrinsic flaw. They discover those patterns in data which minimize training error, but training data are often flawed by selection bias, label bias, capture bias, and negative set bias[8]. Algorithms which train on biased datasets inherit these data biases. Minimizing training error encourages algorithms to indiscriminately absorb all the correlations found in training data, real or spurious. Spurious correlations resulting from data biases are unrelated to the true underlying signal[9]. Recently, disambiguating true correlations from spurious ones has been studied in the context of causal inference. For this reason, we leverage tools from the causal inference domain to identify genes which are robustly correlated with a phenotype. While such genes are putatively causal,

[1]Blue Marble Space Institute of Science, NASA Ames, Mountain View, USA. [2]KBR, NASA Ames, Mountain View, USA. [3]Department of Chemistry, University of Florida, Gainesville, USA. [4]Bay Area Environmental Research, NASA Ames, Mountain View, USA. [5]Department of Marine Science, University of Georgia, Athens, USA. [6]NASA Ames Research Center, Moffett Field, Mountain View, USA. ✉email: james.casaletto@gmail.com

validating true causality is beyond the scope of this research. In this research, we use CRISP—an ensemble machine learning platform developed by the Frontier Development Laboratory (FDL) 2020 Astronaut Health team[10] to enhance biological and medical research with heterogeneous and high-dimensional observational data[11]. The FDL team used CRISP to identify genetic drivers that differentiate two subtypes of colorectal cancer and to implicate operational taxonomic units of the associated microbiome.

The algorithms in the CRISP platform are based on the concept of invariance as a proxy for causal inference. Invariance is a property of a feature which reflects how well a classification algorithm performs using that feature to predict a response invariantly; that is, on data which were generated in different environments, under different circumstances, different conditions, or using different interventions[12,13]. An algorithm based on invariance can identify those features that predict the target label regardless of the background data generating processes that gave rise to the dataset. The classic example is a machine learning classifier built to distinguish images of cows from images of camels[14]. A machine learning classifier that is overfit to a particular environment may learn that a cow is an animal that lives in green pastures while a camel is an animal that lives in beige deserts. Given a cow on a sandy beach, this classifier would likely call it a camel. By contrast, a classifier based on invariance would be optimized to ignore the background environment and learn the salient features which truly distinguish a cow from a camel, such as the dimensions of the neck and legs and the shape of the face. Classification algorithms which exploit invariance promote learning correlations that are stable across training environments, as these are expected to persist on out-of-distribution data (i.e. data generated in environments not seen by the algorithm during training) and therefore be robustly correlated to and more likely causal of the response variable[15].

In this research, we leverage several data transformations for augmenting the dataset that additionally provide an environment in which to leverage the CRISP invariance strategy. We use CRISP to identify genes potentially causal of a high lipid density phenotype in space-flown mice liver tissue. We establish a binary threshold of lipid density using scalar values associated with oil red O (ORO) stained tissue. To compare our method with traditional differential gene expression tools, we use EdgeR and DESeq2. We also compare our results with those derived from generic machine learning classifiers including random forest and empirical risk minimization. Overall, we find that CRISP identifies a biologically relevant set of genes which are uniquely predictive of a high lipid density response in space-flown mice. Gene set enrichment and pathway analyses reveal that the dysfunctional regulation of the genes identified by CRISP is implicated in the spectrum of diseases caused by non-alcoholic fatty liver disease (NAFLD). The mice in the experiment flight group were only in space for a maximum of 54 days, yet their gene expression profiles were altered significantly enough to manifest markers of NAFLD. NASA has gathered a significant amount of biomedical data on the effect of short-term spaceflight (< 6 months), from the astronauts of the Apollo missions to those who fly on the International Space Station (ISS). What matters now is not merely the impact of being in space, but rather the impact of living there. Surviving long-term spaceflight (> 6 months) is necessary for the success of the planned Artemis and Mars missions[16]. Our study provides the first machine learning analysis of gene expression predictive of a disease-related response to spaceflight in the liver.

Methods and data

The data we used for our experiment include transcriptomic and histology data from the liver tissue of space-flown and ground-control mice. The overall workflow for our data and methods is depicted in Fig. 1.

NASA open science data repository

The NASA Open Science Data Repository (OSDR) provides AI-ready datasets allowing rapid deployment of machine learning algorithms for data mining. This is possible because OSDR is a FAIR database (Findable, Accessible, Interoperable, Reusable)[16,17] with rich metadata providing full context for the data and experiments. The full set of metadata for the samples in our experiment is shown in Supplementary Table 1.

RNA-seq data

This study uses transcriptomic data from four OSDR datasets: OSD-47[17] (version 11), OSD-48[17] (version 10), OSD-137[18] (version 6), and OSD-168[17] (version 10). These datasets were generated from three rodent research (RR) missions: RR-1 CASIS, RR-1 NASA, and RR-3. Two different strains of mice were used: C57 and Balb/C. The RR-1 CASIS experiment was designed to study the effects of microgravity of C57 mice on muscle degeneration due to spaceflight (OSD-47). The RR-1 NASA mission was designed to validate the experimental hardware and scientific capabilities on the International Space Station (OSD-48). The RR-3 mission was designed to study countermeasures in Balb/C mice for loss of mass in muscle and bone that have been observed in spaceflight (OSD-137 and OSD-168). The OSD-168 dataset was not based on a separate mission but rather to test the utility of External RNA Control Consortium (ERCC) RNA sequencing controls and therefore constitute technical replicates in our experiments. These rodent research missions were originally designed as randomized controlled experiments, with mice randomly assigned to the groups described in Table 1.

In our analyses, we consider the mice in the basal, vivarium, and ground experimental groups as "non-flight" and compare them with the flight group. We are re-using these data to explore the relationship between the transcriptomes and a phenotype, constituting the data as observational in our research. Indeed, the causal inference algorithms in the CRISP platform ensemble were designed to run on observational data.

Liver histology phenotype data

Liver tissues used for gene expression were quantified for lipid density using the oil red O (ORO) staining protocol. ORO is a fat soluble, hydrophobic dye that stains lipid molecules red[19]. ORO percent positivity was calculated for each sample from the stained images, providing a scalar value that measures the lipid density—higher ORO positivity values directly correspond to higher lipid densities. ORO positivity is the de facto

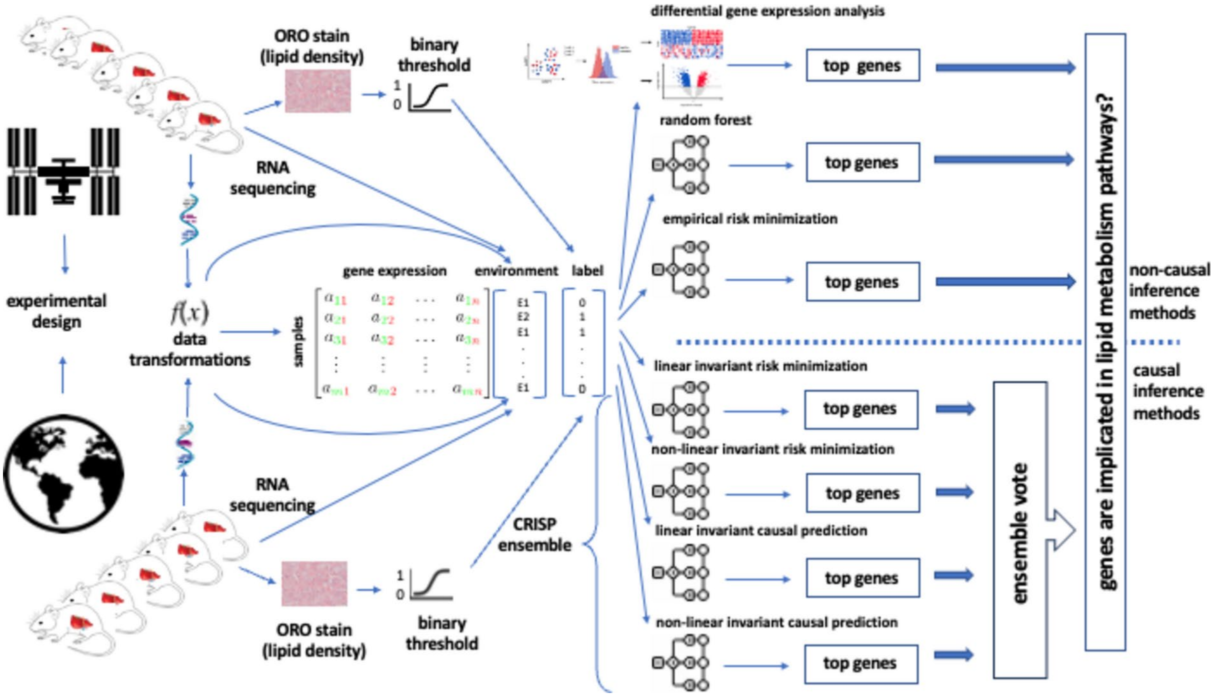


Fig. 1. Experiment setup to compare the most predictive genes of space-flown and ground-control murine lipid density from liver tissue RNA-seq data using both traditional (non-causal inference) methods and the CRISP ensemble of causal inference methods. Gene expression data were transformed to augment the data set and then combined into a single matrix. Environment strings were assigned to each sample based on the name of the transformation performed and the RNA-seq library preparation method. Binary labels (0 and 1) were assigned to each sample based on a threshold value of lipid density. We compare which methods (causal or non-causal inference) perform better at identifying the top genes implicated in the lipid density phenotype.

Group	Description
Basal	Housed in standard vivarium cages on Earth, euthanized 1 day after launch
Vivarium	Housed in standard vivarium cages on Earth, euthanized n days after launch
Ground	Housed in ISS habitat cages on Earth, euthanized n days after launch
Flight	Housed in ISS habitat cages on ISS, euthanized n days after launch

Table 1. The four experimental groups of mice from the OSDR datasets. The basal, vivarium, and ground groups were combined as "non-flight" samples.

histological biomarker for diagnosing the spectrum of disorders in non-alcoholic fatty liver disease (NAFLD) *post mortem*. Because OSD-168 is comprised of technical replicates, the ORO positivity data are associated with the biological replicates in OSD-47, OSD-48, and OSD-137 and not in OSD-168 itself.

Data preparation

A typical machine learning pipeline includes a data preprocessing step. At the very least, the data must be prepared to satisfy the assumptions and requirements of the algorithms which use the data. CRISP requires that the features be real-valued, that the target be binary, and that the environment string be ASCII text, as described in the following sections.

Binarized target

The ORO positivity scalar values in our dataset range from 0.91 to 26.94, but the CRISP platform only permits binary targets (low and high) for classification. We converted the scalar value to a per-mission binary value using the mean value between flight and non-flight medians. The thresholds are depicted in the box-and-whisker plot of Fig. 2 as a horizontal dashed blue line.

Using the thresholds per mission indicated in Fig. 2, a sample was assigned a binary target of 0 if its ORO positivity is less than the threshold value and 1 if its ORO positivity is greater than or equal to the threshold value.

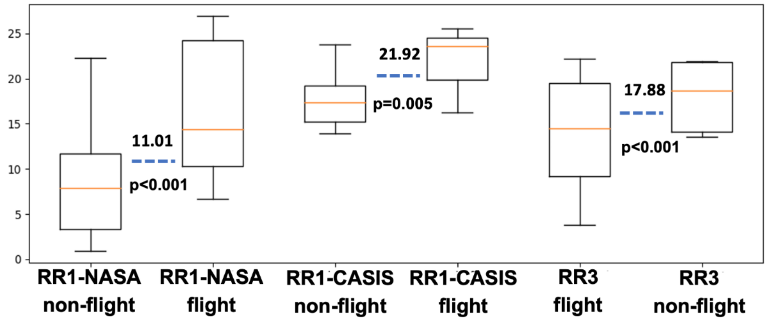


Fig. 2. Box-and-whisker plots for ORO values (y-axis) based on mission (x-axis). The dashed blue line is the rodent research mission threshold that was calculated as the mean of the two group medians (yellow lines). The 2-sided student t-test *p*-values show that the differences between the medians of flight and non-flight samples are all significant. The top of each box represents the 75th percentile, the bottom of each box represents the 25th percentile, and the solid black lines on the very top and bottom represent the maxima and minima, respectively.

Transformation	Description	Reference
Log scale	Scales values to their log base 2	Quinn et al.[25]
Square root	Scales values to their square roots	Zhang et al.[26]
Median of ratios	Scales values to account for sequencing depth, gene length, and outliers	Robinson et al.[27]
Centered log ratio	Transforms data to eliminate over-dispersion	Anders et al.[28]
Box-Cox	Transforms non-normal data into a normal shape	Sun et al.[29]
Z-score	Scales values to their number of standard deviations from the mean	Zwiener et al.[30]

Table 2. Description of and reference for each gene expression data transformation used in pre-processing of data. These 5 transformations (log, square root, median of ratios, centered log ratio, and Box-Cox) provide 6 times more data for building the models and create environments in which to leverage invariance for CRISP. The z-scores of each transformation were individually calculated prior to merging the datasets.

Feature transformations and data augmentation

There are many types of transformations, including power-scaling and normalization, that are commonly performed on data as pre-processing steps in a machine learning pipeline. Gonzalez et al.[20] and others have shown that while data preprocessing is a necessary step in a machine learning pipeline, there isn't much agreement as to which is the best. Some data transformations are more volatile than others. A data transformation may change the data so drastically that it destroys some of the underlying signals of interest. This lack of data preprocessing standard exists in transcriptomic data analysis as well[21]. In our research, instead of choosing one pre-processing method, we used several methods as shown in Table 2. This technique of adding differently transformed samples to a dataset is referred to as data augmentation and is a common practice in machine learning[22]. Standardization (converting values to z-scores) is a proven method of data harmonization when combining multiple datasets into one dataset[23]. Each transformation was considered a separate environment across which CRISP must find genes invariantly correlated to the target. We exploit CRISP's built-in search for invariance across environments and consider each transformation as a perturbation of the data akin to a causal intervention[24].

Figure 3 shows the original data distribution (named "identity") and the data after having been transformed and plotted as (variance vs mean) coordinates in log scale.

The mean–variance plots of the differently transformed data in Fig. 3 reveals that certain transformations change the data significantly while other transformations are relatively mild in effect, compared to the original raw data. In addition to these transformations, we applied some basic filtering of the input to remove transcripts which don't have ENSEMBL identifiers, don't code for a protein, have counts of 0 in 80% or more of the samples, have a coefficient of variation less than 0.1, or have counts less than 50 in 80% or more of the samples, as shown in Supplementary Table 6. This filtering produced the final set of 8,092 genes which we subsequently used in all downstream analysis.

To prevent data leakage, we first separated the data into training, validation, and test datasets and standardized the training and validation sets separately[31]. We held out 90% of the raw data for testing and did not perform any transformations on it[32]. We held out 10% of the raw data for validation, and we used all the augmented data during the training phase.

Technical batch effects

We examined the dataset for batch effects and found that only the type of library preparation of the RNA-seq experiments generating the transcriptomic data clearly separate the samples, as shown in Fig. 4a.

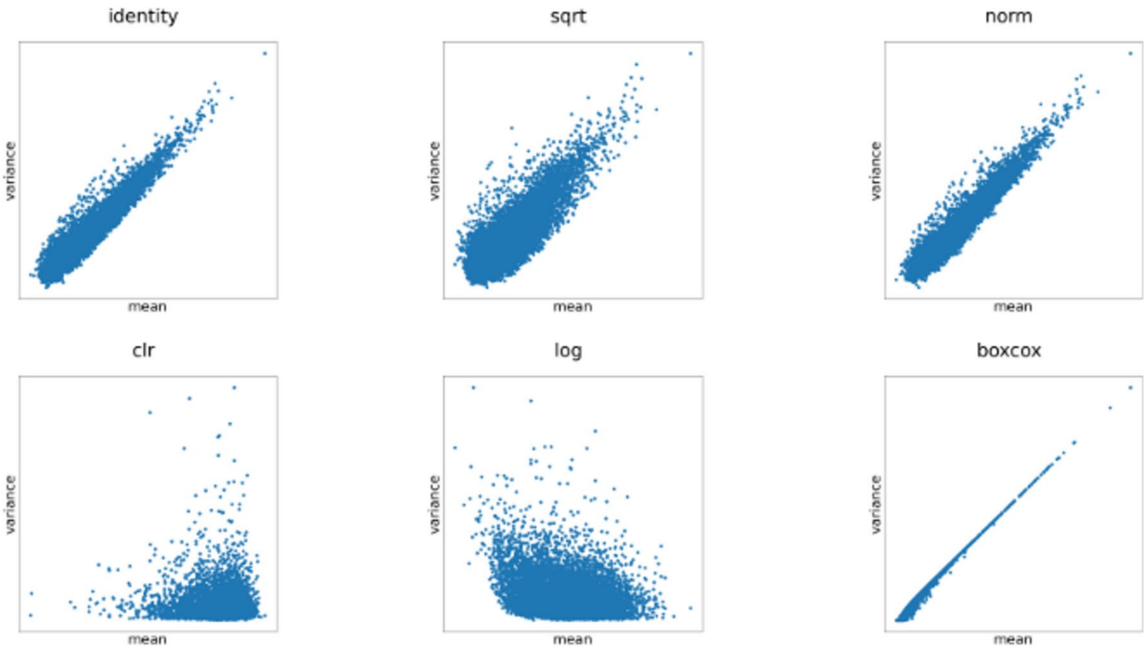


Fig. 3. Scatter plots of variance versus mean for different transformations used in preprocessing. The vertical and horizontal axes are shown in log2 scale after having standardized the data. The identity transformation represents the original, untransformed data. The square root (sqrt) transformation computes the square root of each expression value. The normalization (norm) transformation uses the median of ratios method from the R DESeq2 package. The centered log ratio (clr) method divides each row of gene expression data by the mean of that row and returns the logarithms of those ratios. The log transformation computes the logarithm (base 2) of each expression value. The boxcox transformation is a type of power transformation that makes the gene expression data more normally distributed.

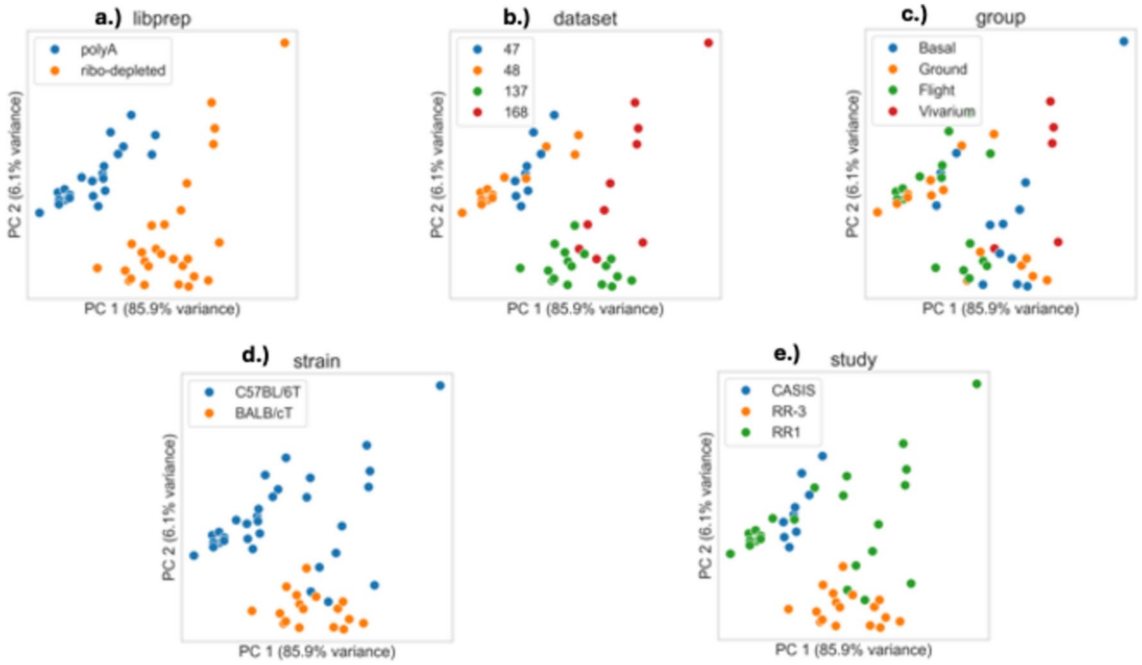


Fig. 4. PCA plots of OSD datasets colorized by the different covariates, including (**a**) library preparation, (**b**) dataset, (**c**) experimental group, (**d**) mouse strain, and (**e**) Rodent Research study. The first two principal components capture 92% of the total variance of the gene expression. Colorizing by library preparation in Fig. 4a shows a distinct separation of all samples.

Unlike Fig. 4a (library preparation), Fig. 4b–e do not show a clean separation of samples colored by covariate. To account for this library preparation batch effect, we include the library preparation in the environment string, as discussed in the following section.

Constructing the environment string
CRISP defines the environments of each sample by associating with it an environment string value. Different values of this string represent known perturbations in the data generating process for that sample. In our experiment, the two known perturbations of the data include the library preparation and the data transformation. Because the other covariates do not cluster in the PCA plot, we conclude they do not create a batch effect and therefore are not included in the environment string. Causal inference algorithms which exploit environment invariance theoretically perform better on out-of-distribution data when the number of environments used to train the model is high[12]. However, if the number of samples in the dataset is small (which is intrinsic to transcriptomic experiments), then the environment string must be selected such that there are enough samples in each partition to test for significance. Therefore, we restricted our choice of environment string to include only known perturbations of the data—i.e. transformation and library preparation—and excluded perturbations of unknown effect such as study and group. We show results using strain in the environment string in Supplemental Figs. 4 and 5.

Table 3 shows a snippet of one fictitious sample's input data after having performed the 5 data transformations and the standardization. The identity transformation is the original data, standardized. The environment string (here, called "env") is a concatenation of the sample library preparation (here, "polyA") and the transformation name (e.g. "boxcox"). Only the binary ORO threshold (called "oro_thresh") for the sample remains unchanged across the same sample as well as for its respective technical replicates of OSD-168, if they exist. As a result of performing these 5 data transformations, our combined dataset size of 51 samples was increased six-fold to 306 samples.

Running CRISP experiments

With the data in place, we turn now to how we run the CRISP in silico experiments.

CRISP ensemble methods
CRISP is an ensemble of machine learning methods which have been designed to identify features causal of a target[11]. We leveraged CRISP to identify genes potentially causal of a phenotype. There are 2 types of algorithms in CRISP: one based on invariant causal prediction (ICP)[13]; and one based on invariant risk minimization (IRM)[12]. There is a linear and non-linear version of each of these algorithms. The linear versions use linear combinations of the feature space to construct a binary output, and the non-linear versions use non-linear combinations (multi-layer perceptrons with non-linear activation functions) of the feature space to construct a binary output. Both algorithms rely on invariance as a proxy for causality, and invariance in turn relies on having data partitioned by known interventions or perturbations. Each of these methods trains a model using a training set, selects model hyper-parameters using a validation set, and provides a final model accuracy on a held-out test set. CRISP partitions the data into different environments such that the model is trained on data from one environment and then validated and tested on the data from the other environments. CRISP uses two methods of feature reduction to reduce the number of variables to search. The first method removes features whose correlation to the target is below a certain threshold, and the second method uses a multi-layer perceptron to identify the most salient features based on model feature importance.

CRISP configuration
CRISP experiments are configured with several parameters in a JSON configuration file. By default, the test_val_split parameter defines how much data to dedicate for training, testing, and validation. We overrode this default data splitting configuration and explicitly defined the training, testing, and validation data subsets. The training data is used to train the model parameters, the validation data is used to select the optimal set of model hyperparameters, and the test data is held out after the model is trained and validated to report accuracy. It is this held-out test accuracy that we subsequently report on. The max_features parameter defines the number of features each model in the ensemble should find as most predictive of the target. We used the default value of 20 in our CRISP experiment. The data_options parameter defines the file location containing the dataset,

Sample	Gnai3	Apoh	…	env	oro_thresh
Mmus_FLT_I_boxcox	−0.012137	−0.011788	…	polyA:boxcox	1
Mmus_FLT_I_clr	0.969207	3.344341	…	polyA:clr	1
Mmus_FLT_I_log	−0.011233	2.079650	…	polyA:log	1
Mmus_FLT_I_norm	0.969207	3.344341	…	polyA:norm	1
Mmus_FLT_I_sqrt	0.388949	5.967345	…	polyA:sqrt	1
Mmus_FLT_I_identity	−0.011230	2.079653	…	polyA:identity	1

Table 3. One fictitious mouse sample's gene expression data (columns truncated to 2 genes) after 5 data transformations and standardization. The environment string ("env") includes the name of the library preparation ("polyA" or "ribo-depleted") and the name of the transformation. The binary ORO threshold ("oro_thresh") is set to either 0 (low ORO positivity) or 1 (high ORO positivity).

which columns in that dataset are to be used as predictors, which column is the environment variable, and which column is the target variable. There are several other parameters that may be configured in this JSON configuration file which get consumed by the machine learning algorithms configured in the ensemble. The JSON configuration we used in our experiment is shown in Supplementary Fig. 1.

CRISP ensemble voting

CRISP is an ensemble of machine learning algorithms. Ensembles are used to combine a set of multiple "weak" learners into a single "strong" learner to minimize training errors[33]. Each model in the ensemble is trained on the dataset to identify the features most predictive of the target. After training, each model selects the features (20 by default) which it found as most predictive of the target. For the linear models—linear invariant risk minimization (linear IRM) and linear invariant causal prediction (linear ICP)—the features which are most predictive are those whose coefficients of the linear model have the highest absolute values. For non-linear models—non-linear invariant risk minimization (non-linear IRM) and non-linear invariant causal prediction (non-linear ICP)—the most predictive features are selected through sensitivity analysis (the features which have the highest impact on the response variable). After each model has selected its top-most predictive features, the ensemble votes to elect a single set of (by default, 20) features to present as the final result of the experiment. Non-Causal Empirical risk minimization (ERM) and random forest (RF) are not causal predictors and do not participate in CRISP's selection of the features. Non-causal ERM relies on statistical minimization of risk or error, while RF is another non-causal method which combines decision trees for a classification result. They are included in the experiment only as a basis of comparison to the causal predictors (non-linear IRM, linear ICP, non-linear ICP, and linear IRM). CRISP attributes the highest weight to the feature that the highest accuracy model gives the largest coefficient (or highest sensitivity). Conversely, the lowest ranking feature from the worst performing model will be attributed the lowest weight. Furthermore, the higher degree of concordance across the ensemble (i.e. how many models found the feature in their top 20 list), the higher the weighted coefficient of that feature. This attribution weight *w(k)* of feature *k* is calculated as shown in the following equation.

$$w\left(k\right)=p_k\cdot\frac{1}{M}\sum_{m=1}^{M}a_m\cdot s_m\left(k\right)$$

The values p_k represent the fraction of models which found this feature *k* in its list of top (by default, 20) features. The number *M* is the number of models CRISP is configured to use. The values $s_m(k)$ are the absolute values of the linear coefficients (for linear models) or sensitivity factors (for non-linear models) scaled to the unit interval. The values a_m are the predictive accuracies of the models. In this way, CRISP identifies those features that are most predictive of the target from the highest number of best performing models.

CRISP results on synthetic data

To help validate whether CRISP is truly identifying features robustly correlated to the target, we ran an experiment in which we synthetically generated a feature space with a known causal relationship to a target using structural equation modeling. Of the 6 known causal features, the CRISP ensemble found 5 of them in its list of top 20 features, as shown in Supplementary Fig. 6. The probability of this happening at random is 0.00124 as shown in Supplementary Eq. 1. By contrast, the RF and ERM classifiers didn't find any of the 6 causal features in their top 20 list. This is further evidence that CRISP can distinguish robustly correlated features from uncorrelated ones.

CRISP updates

Accompanying this paper we publish a CRISPv1.1 code release, with the following updates. First, we updated the linear IRM code to output continuous feature weights such that the coefficients are now on the unit interval [0, 1]. We create confusion matrices for the test results and append them to the results so that we can calculate sensitivity, specificity, and F-1 scores. We changed the default batch size from 128 to 8 to improve accuracy. We updated the configuration to provide a means to override the default method used to split the data into training, testing, and validation sets. Last, we updated the post-processing step by putting their feature coefficients on the scale of z-scores.

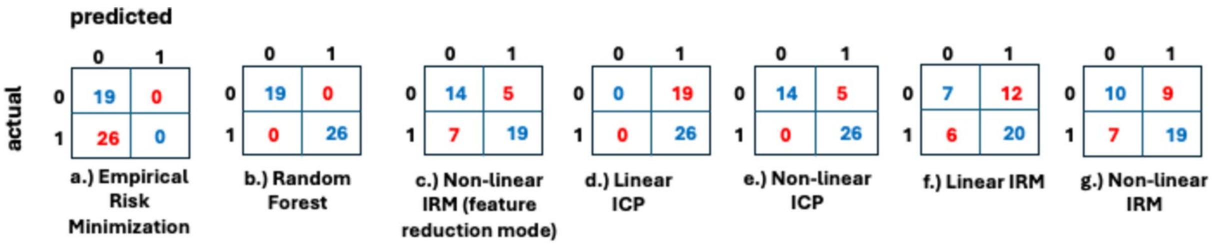
Results

In this section, we show the results of the CRISP experiment. We validate our findings using pathway analysis, gene set enrichment analysis, the Scalable Precision Medicine Open Knowledge Engine (SPOKE), and a search of relevant literature. We show the results of the gene expression analysis using DESeq2 and EdgeR to compare the CRISP results with commonly used bioinformatic tools.

CRISP results

Each of the models in the ensemble trains on the same data set to identify the features which best predict the target across all environments. Figure 5 depicts the confusion matrices and performance metrics for each model in the ensemble.

The best and worst accuracies in the experiment do not participate in the ensemble vote. The RF test accuracy was the best out of all the models, scoring a 100% accuracy rate (and therefore perfect precision, recall, and F1-score). Conversely, ERM accuracy was the lowest. Non-linear ICP has 89% test accuracy and therefore contributes the most to the ensemble results. Conversely, linear IRM performed the worst across the ensemble (58% test accuracy) and therefore contributes the least to the ensemble results. The linear ICP algorithm did not predict any samples to have a low lipid density, so its precision and F1-scores are undefined.



h.) performance metrics

Algorithm	Accuracy	Precision	Recall	F1-score
Empirical Risk Minimization	0.42	0.42	1.0	0.59
Random Forest	1.0	1.0	1.0	1.0
Non-linear IRM feature reduction mode	0.73	0.67	0.74	0.89
Linear ICP	0.58	undefined	0.0	undefined
Non-linear ICP	0.89	1.0	0.74	0.85
Linear IRM	0.56	0.54	0.37	0.44
Non-linear IRM	0.64	0.59	0.53	0.56

Fig. 5. (**a**–**g**) Show confusion matrices and (**h**) shows performance metrics for each model in the CRISP ensemble. The horizontal rows are the actual classes, and the vertical columns are the predicted classes. We assigned the low lipid density class (0) to "positive" and the high lipid density class (1) to "negative" for the purpose of calculating precision, recall, and F1-score. Precision and F1-score values are undefined for the linear ICP model because it did not predict any positive classes, as shown in (**d**). The higher the accuracy of the model, the higher the weight associated with the results of that model in the ensemble.

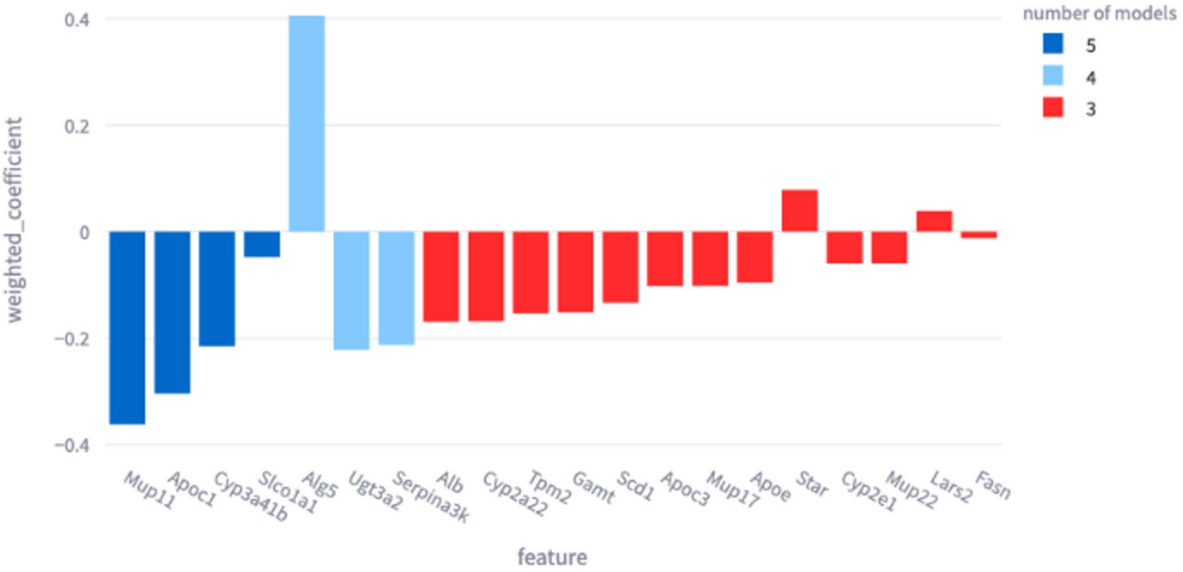


Fig. 6. The weight of the top 20 genes and their degree of concordance across the ensemble. The direction of the bar indicates whether the gene expression is predictive of high or low lipid density. Each bar is colored by the number of models which found that gene in their list of the top 20 most predictive genes.

The bar chart in Fig. 6 shows the degree to which each gene across the ensemble is predictive of the lipid density response variable. Based on Fig. 6, we see that the *Mup11* gene was found in the top 20 genes most predictive of lipid density in all 5 models. By contrast, the *Fasn* gene was found in the top 20 genes most predictive of lipid density in only 3 of the models.

The biological functions of the 20 CRISP genes are shown in Table 4. The functions were derived from the information provided in the NCBI Gene tool at https://www.ncbi.nlm.nih.gov/gene/.

Gene	Function of protein
Alb	Carrier protein for various molecules such as steroids and fatty acids in blood
Alg5	Enables dolichyl-phosphate beta-glucosyltransferase activity
Apoc1	Transports lipids from the intestines to other locations in the body
Apoc3	Regulates metabolism of lipoproteins and play a role in lipid storage
Apoe	Transports lipoproteins in the blood
Cyp2a22	Enables enzyme binding activity; heme binding activity; and monooxygenase activity
Cyp2e1	Enables monooxygenase activity
Cyp3a41b	Enables several functions, including caffeine oxidase activity; iron ion binding activity; and monooxygenase activity
Fasn	Enables fatty acid synthase activity
Gamt	Enables guanidinoacetate N-methyltransferase activity and identical protein binding activity
Lars2	Involved in leucyl-tRNA aminoacylation and mitochondrial translation
Mup11	The MUP family proteins bind to, concentrate, and stabilize many volatile scent substances (e.g. pheromones), thereby controlling both pheromone transport in circulation and pheromone release into the air
Mup17	
Mup22	
Scd1	Enables metal ion binding activity; palmitoyl-CoA 9-desaturase activity; and stearoyl-CoA 9-desaturase activity
Serpina3k	Enable serine-type endopeptidase inhibitor activity
Slco1a1	Enables anion transmembrane transporter activity and identical protein binding activity
Star	Enables cholesterol binding activity
Tpm2	Binds to actin filaments and stabilize them by regulating access to actin modifying proteins
Ugt3a2	Enables UDP-glycosyltransferase activity

Table 4. The top 20 mouse genes identified in the CRISP experiment as robustly predictive of the threshold lipid density phenotype (listed alphabetically) and a short description of the protein they encode.

Gene set name	Description	FDR q-value
GOBP MONOCARBOXYLIC ACID BIOSYNTHETIC and METABOLIC PROCESS	Chemical reactions and pathways resulting in the formation and metabolism of monocarboxylic acids	< 0.001
GOBP LIPID and CELLULAR LIPID METABOLIC PROCESS	Chemical reactions and pathways involving lipids	< 0.001
GOBP SMALL MOLECULE BIOSYNTHETIC PROCESS	Chemical reactions and pathways resulting in the formation of small molecules	< 0.001
GOBP ORGANIC ACID BIOSYNTHETIC and METABOLIC PROCESS	Chemical reactions and pathways resulting in the formation and metabolism of organic acids	< 0.001
GOBP TRIGLYCERIDE RICH and VERY LOW DENSITY LIPOPROTEIN PARTICLE CLEARANCE	Process in which a triglyceride-rich and very-low-density lipoprotein particles are removed from the blood via receptor-mediated endocytosis and its constituent parts degraded	< 0.001

Table 5. Gene set enrichment analysis using gene ontology gene sets, showing the false discovery rate (FDR) adjusted significance q-value. All of these gene sets statistically significantly overlap the CRISP gene result set and their associated biological processes are directly involved in the metabolism of lipid proteins.

We will discuss later in this section that the genes CRISP found are not only involved in lipid metabolism but many have also been implicated in NAFLD.

Validation using gene set enrichment analysis

We used the 20 genes from the CRISP experiment to perform Gene Set Enrichment Analysis (GSEA, https://www.gsea-msigdb.org/gsea/msigdb/annotate.jsp) using the Human Molecular Signatures Database (MSigDB) ontology gene sets (C5) collection version v2023.2.Mm. This tool finds those gene ontologies from all ontology gene sets (GO: Gene Ontology and HPO: Human Phenotype Ontology) that have a significant overlap with the query gene set in the mouse genome. Table 5 shows the top gene sets in the C5 collection that significantly overlap with these 20 mouse genes.

The genes which enrich these gene sets include *Apoc3*, *Apoc1*, *Cyp2e1*, *Scd1*, *Fasn*, *Star*, *Gamt*, *Apoe*, *Cyp3a41b*, and *Alg5*. These gene sets are consistent with our lipid density phenotype as they are involved in the metabolism of fats, and we will explore the scientific literature to determine how these genes are implicated in NAFLD in the "Validation from scientific literature" section.

Validation using pathway analysis

We submitted the 20 genes resulting from our CRISP experiment, and as background all 8,092 genes which were used in our CRISP experiment, to the KEGG pathway[34] of the ShinyGO 0.80 pathway analysis tool (https://bioinformatics.sdstate.edu/go/). This tool finds the Gene Ontology pathways which overlap with the query gene set as compared to the background gene set. Table 6 shows the enriched pathway that ShinyGO identified.

Pathways	Number of CRISP genes	Number of KEGG pathway genes	Fold enrichment	Enrichment FDR
Cholesterol metabolism	4	49	45.2	< 0.001
Fatty acid metabolism	2	62	19.8	0.049
Alcoholic liver disease	3	139	15.8	0.01

Table 6. Enriched gene ontology pathways involving the genes from the CRISP experiment. The fold enrichment column uses the number of CRISP genes and KEGG pathway genes to measure how over-represented the CRISP genes are as compared to the background. This fold enrichment is then used to provide an enrichment false discovery rate (FDR), all of which are below our 0.05 threshold of significance.

Disease ontology	Adjusted *p*-value
Metabolic dysfunction-associated steatotic liver disease (DOID: 0080208)	< 0.001
Lipid storage disease (DOID:9455)	< 0.001
Steatotic liver disease (DOID:9452)	0.001
Disease of metabolism (DOID: 0014667)	0.001

Table 7. Disease ontologies and associated Bonferroni-corrected p-values. All the disease ontology identifiers (DOID) that SPOKE found as statistically significantly associated with the CRISP result gene set are related to dysfunctional lipid metabolism.

Using the enrichment false discovery rate metric of significance, the significant pathways relating to our gene set include cholesterol metabolism, fatty acid metabolism, and alcoholic liver disease. Cholesterol and fatty acid metabolism are consistent with both our gene set enrichment analysis as well as our lipid density phenotype. Lipid dysregulation is a key factor of fatty liver disease: the accumulation of lipids creates a lipotoxic environment that can lead to inflammation, oxidative stress, fibrosis, and liver damage. The genes enriching these 3 pathways include *Apoe*, *Star*, *Apoc3*, *Apoc1*, *Fasn*, *Scd1*, and *Cyp2e1*. We will explore the scientific literature to determine how these genes are implicated in NAFLD in the "Validation from scientific literature" section.

Validation from SPOKE

The Scalable Precision Medicine Open Knowledge Engine (SPOKE) is a massive knowledge graph of biomedical information which may be leveraged for a variety of purposes from drug discovery to disease diagnosis[35]. We used SPOKEv5 to identify disease conditions (derived from the Disease Ontology Knowledgebase[36]) from the list of 20 genes that CRISP found. Table 7 identifies those disease conditions and their associated adjusted *p*-value.

All these statistically significant disease ontologies are consistent with a high lipid density phenotype and NAFLD. Metabolic dysfunction-associated steatotic liver disease (MASLD) is the more recent term used for NAFLD.

Validation from scientific literature

Table 8 shows one relatively recent and relevant research article that discusses the relationship of NAFLD to each of the CRISP genes which enrich the pathways and gene sets previously discussed.

Except for *Alg5*, all the genes from the gene set and pathway enrichment analyses have previously been shown to be dysregulated in liver inflammation, liver disease, non-alcoholic steatohepatitis (NASH), and NAFLD. Fuior et al. discuss how the overexpression of *Apoc1* in mice may lead to hyperlipidemia. Lu et al. demonstrated that overexpression of *Apoc3* manifests in features similar to NAFLD, including increased hepatic lipid content. Harjumaki et al. discuss the role of *Cyp2e1* in the development of NASH which is a precursor condition to NAFLD, suggesting that the over-expression of the *Cyp2e1* enzyme and subsequent oxidative stress is sufficient to induce hepatic lipid accumulation. Woolsey et al. found that hepatic mRNA expression of *Cyp3a41b* in NASH was 69% lower than control livers. Testing its pharmacological inhibition in human cell culture, O'Farrell et al. consider *Fasn* to be a promising therapeutic target for NASH as they observed that its reduced expression reduced overall triglyceride content, markers of fibrosis and inflammation, and development of hepatocellular carcinoma. The *Gamt* gene encodes an enzyme which helps synthesize creatine, and Marinello et al. found that its supplementation has been demonstrated to prevent NAFLD progression. Qin et al. describe how aberrant transcriptional activation of *Scd1* causes excessive lipid accumulation and promotes progression of NAFLD (among other metabolic complications), making it a promising therapeutic target. Finally, Qiu et al. found that *Star* overexpression, among other protective roles, can reduce hepatic lipid accumulation—suggesting it may be a potential therapeutic target for NAFLD.

Comparing CRISP results to other analyses

In this section, we compare the results of our CRISP experiments with other tools which associate features with targets including DESeq2, random forest, and empirical risk minimization.

CRISP genes	Research implicating gene expression in lipid dysregulation
Alg5	N/A
Apoc1	Fuior, E. V. & Gafencu, A. V. Apolipoprotein C1: Its Pleiotropic Effects in Lipid Metabolism and Beyond. *IJMS***20**, 5939 (2019)
Apoc3	Paiva, A. A., Raposo, H. F., Wanschel, A. C. B. A., Nardelli, T. R. & Oliveira, H. C. F. Apolipoprotein CIII Overexpression-Induced Hypertriglyceridemia Increases Nonalcoholic Fatty Liver Disease in Association with Inflammation and Cell Death. *Oxidative Medicine and Cellular Longevity***2017**, 1838679 (2017)
Apoe	Lu, W. *et al.* ApoE deficiency promotes non-alcoholic fatty liver disease in mice via impeding AMPK/mTOR mediated autophagy. *Life Sciences***252**, 117601 (2020)
Cyp2e1	Harjumäki, R., Pridgeon, C. S. & Ingelman-Sundberg, M. CYP2E1 in Alcoholic and Non-Alcoholic Liver Injury. Roles of ROS, Reactive Intermediates and Lipid Overload. *IJMS***22**, 8221 (2021)
Cyp3a41b	Woolsey, S. J., Mansell, S. E., Kim, R. B., Tirona, R. G. & Beaton, M. D. CYP3A Activity and Expression in Nonalcoholic Fatty Liver Disease. *Drug Metab Dispos***43**, 1484–1490 (2015)
Fasn	O'Farrell, M. *et al.* FASN inhibition targets multiple drivers of NASH by reducing steatosis, inflammation and fibrosis in preclinical models. *Sci Rep***12**, 15661 (2022)
Gamt	Marinello, P. C. *et al.* Creatine supplementation protects against diet-induced non-alcoholic fatty liver but exacerbates alcoholic fatty liver. *Life Sciences***310**, 121064 (2022)
Scd1	Sun, Q. *et al.* SCD1 is the critical signaling hub to mediate metabolic diseases: Mechanism and the development of its inhibitors. *Biomedicine & Pharmacotherapy***170**, 115586 (2024)
Star	Qiu, Y. *et al.* Steroidogenic acute regulatory protein (StAR) overexpression attenuates HFD-induced hepatic steatosis and insulin resistance. *Biochimica et Biophysica Acta (BBA)—Molecular Basis of Disease***1863**, 978–990 (2017)

Table 8. Research articles implicating the changes in expression of the CRISP-identified genes in the spectrum of lipid dysregulation and NAFLD. Selection criteria for these articles include journal impact factor (greater than 3), date of publication (within the last 10 years), and relevance to a causal relationship between expression of the gene and the progression of NAFLD.

Gene	Function
Atp1a2	Integral membrane protein responsible for establishing and maintaining the electrochemical gradients of Na and K ions across the plasma membrane
Des	This gene encodes a muscle-specific class III intermediate filament. Homopolymers of this protein form a stable intracytoplasmic filamentous network connecting myofibrils to each other and to the plasma membrane and are essential for maintaining the strength and integrity of skeletal, cardiac and smooth muscle fibers
Hsd3b1	Predicted to be involved in several processes, including C21-steroid hormone metabolic process; hippocampus development; and response to corticosterone
Tpm2	This gene encodes beta-tropomyosin, a member of the actin filament binding protein family, and mainly expressed in slow, type 1 muscle fibers
Star	Predicted to enable cholesterol binding activity. Acts upstream of or within cellular lipid metabolic process; glucocorticoid metabolic process; and regulation of steroid biosynthetic process
Akr1b7	Predicted to act upstream of or within cellular lipid metabolic process

Table 9. Gene result set using EdgeR to find differentially expressed genes between low and high ORO groups. Half of these genes are involved in lipid metabolism (*Akr1b7*, *Tpm2*, and *Star*).

Comparing results from CRISP and DESeq2

We used DESeq2 version 1.34.0 to perform differential gene expression analysis (DGEA) on the data on the same set of genes as we used in CRISP to identify which genes are significantly differentially expressed between the high and low ORO groups. DESeq2 performs its own filtering and normalization steps, so we did not transform the data. Because of the distinct batch effect due to library preparation, we added the library preparation covariate to the column data and included it in the DESeq2 design formula along with the ORO threshold value. DESeq2 did not find any genes that were significantly differentially expressed when setting the Benjamini–Hochberg adjusted *p*-value cutoff to 0.05.

Comparing results from CRISP and EdgeR

We used EdgeR version 3.36.0 to perform DGEA on the data on the same set of genes as we used in CRISP to identify which genes are significantly differentially expressed between the high and low ORO groups. EdgeR performs its own filtering and normalization steps, so we did not transform the data. Because of the distinct batch effect due to library preparation, we added the library preparation covariate to the column data and included it in the EdgeR design formula along with the ORO threshold value. Table 9 shows the 6 differentially expressed genes from the EdgeR experiment using 0.05 as the Bonferroni adjusted p-value threshold and not using a log2 fold change threshold.

Submitting these genes to the GSEA tool, we see that these genes are implicated in hyperplasia, abnormal myocardium morphology, reduced systolic function, weakness of facial musculature, and areflexia (shown in Supplementary Tables 2 and 3). The ShinyGO tool finds several pathways overlapping these 6 genes, from cardiomyopathy to ovarian steroidogenesis, folate biosynthesis, cortisol synthesis and secretion, thyroid hormone synthesis, pancreatic secretion, cardiomyopathy, Cushing Syndrome, and various metabolic pathways including galactose, cholesterol, fructose, and lipid metabolism. Given such a wide range of pathways and gene sets that affect so many different organs and pathologies, it is not at all clear what, if any, conclusions could be drawn from the EdgeR gene result set.

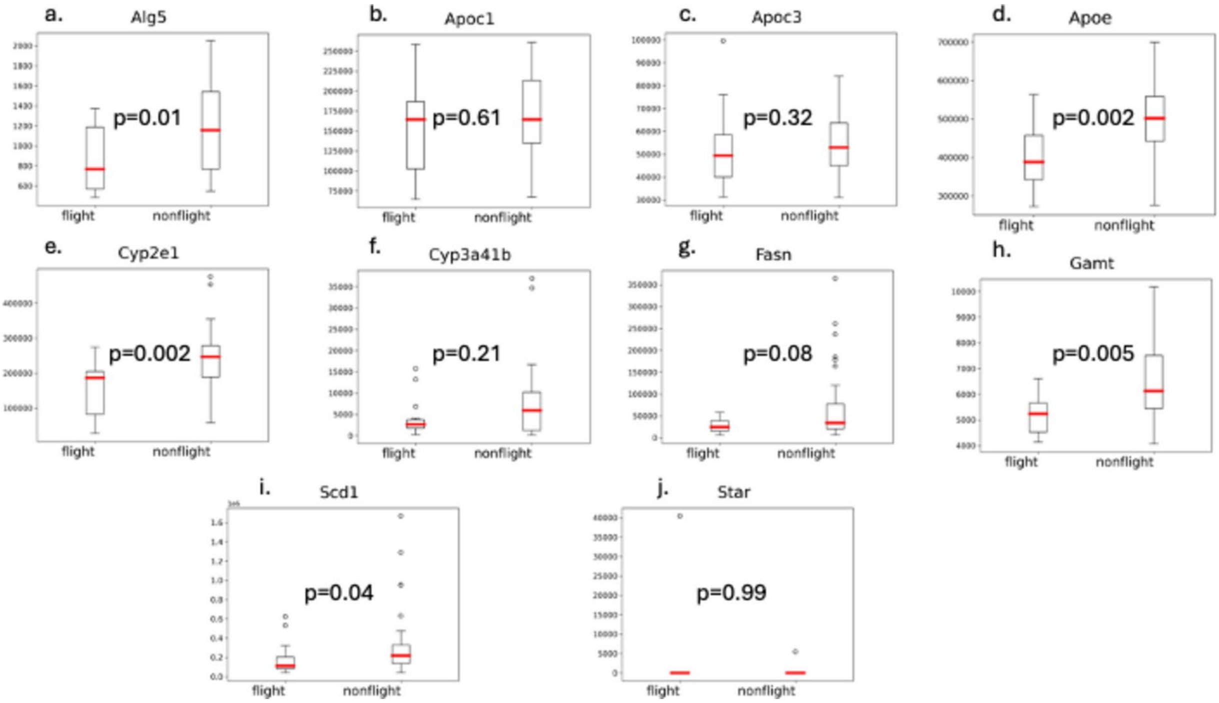


Fig. 7. Box-and-whisker plots of the distributions of gene expression for each of the genes with associated Wilcoxon *p*-values comparing flight and non-flight groups. Several *p*-values are not statistically significant, suggesting that only using statistics to analyze differences in gene expression between groups may not be sufficient.

Statistical analysis of gene distributions

Taken together, the genes enriched in the pathways and gene sets shown in Tables 5 and 6 include *Alg5*, *Apoc1*, *Apoc3*, *Apoe*, *Cyp2e1*, *Cyp3a41b*, *Fasn*, *Gamt*, *Scd1*, and *Star*. Figure 7 shows the box-and-whisker plots of the distributions of the raw counts of gene expression for each of these 10 genes between the flight samples and non-flight samples.

As indicated in Fig. 7, not all the differences between flight and non-flight lipid densities are statistically significant. This suggests that using statistics alone to identify genes predictive of a response may be insufficient in some circumstances. Indeed, the CRISP platform is intended to complement current approaches which rely heavily, if not entirely, on statistical methods.

Analyzing results from CRISP when using non-augmented data

Running CRISP without data augmentation (i.e. using the original gene expression of the n = 51 samples and using only library preparation in the environment string), CRISP identified the genes in Supplementary Fig. 7 as most predictive of the phenotype. However, pathway enrichment analysis on these 20 genes reveals a long list of pathways, not all of which are related to the phenotype (e.g. hormone synthesis, nitrogen metabolism, and coronavirus disease). We attribute this result to the lack of training data, poor model performance, and insufficient exploitation of invariance in the CRISP experiment.

Analyzing results from CRISP when using other environment strings

We include Supplementary Figs. 2 and 3 to show the CRISP results and model performance when using group in the environment variable rather than library preparation. Because the mean lipid density is significantly different between flight and non-flight groups, we expected the accuracy to be higher in the CRISP models predicting the lipid density response when using the group (flight vs non-flight) in the environment variable rather than library preparation. However, the performance of the ensemble was not as good, and the downstream analyses were inconsistent with respect to liver metabolism. We attribute this to the fact that while the metabolic response to spaceflight in the liver was statistically significantly different than ground control, the gene expression data between the 2 groups does not separate in the PCA plot, as shown in Fig. 4c.

We include Supplementary Figs. 4 and 5 to show the CRISP results and model performance when using strain in the environment variable rather than library preparation. We observe that while the CRISP experiments have 50% concordance in their gene sets, using library preparation as the environment variable yields more biologically specific and plausible results. The gene ontologies and pathways enriched by the genes that CRISP found when using strain in the environment variable are not uniformly related to lipid metabolism and include pathways such as cardiomyopathy and thyroid hormone synthesis.

Analyzing results from the random forest classifier

CRISP includes non-causal algorithms in its ensemble to compare results with causal algorithms. While random forest lays no claim to identify features causal of a target, the algorithm is one of the best-performing and highly used ML classification algorithms[37]. The CRISP ensemble uses the RandomForestClassifier module from version 0.24.2 of the scikit-learn package. Like its causal counterparts, the random forest algorithm calculates a feature importance metric which can be used to define confidence in the results. Among the top 20 genes that random forest identified as predictive of the lipid density response, none of them are involved in the lipid metabolic process according to the NCBI Gene tool. Neither the ShinyGO nor the GSEA enrichment tools found any overlapping pathways or gene ontologies using these 20 genes as input. We conclude that the random forest classifier found spurious correlations between the expression of those 20 genes and the lipid density response. The 20 genes which random forest found as most predictive of the lipid density phenotype are shown in Supplementary Table 4.

Analyzing results from empirical risk minimization

Following the same analysis as with random forest, here we analyze the results from empirical risk minimization (ERM). ERM uses the canonical minimization of the sum of the squares of the residuals as its unconditional objective function. It does not partition the data into environments like IRM. As such, it is perhaps the closest comparison to the linear IRM and non-linear IRM results from CRISP. Among the 20 genes that ERM identified as predictive of the lipid density response, none of them are involved in the lipid metabolic process according to NCBI. Neither the ShinyGO nor the GSEA enrichment tools found any overlapping pathways or gene ontologies using these 20 genes as input. We conclude that the ERM classifier found spurious correlations between the expression of those 20 genes and the lipid density response. The 20 genes which ERM found as most predictive of the lipid density phenotype are shown in Supplementary Table 5.

Discussion

The CRISP platform trains an ensemble of binary classifiers on data in one environment and validates and tests them using the data from the other environments. Separating the environments in this way ensures that the ensemble algorithms are rewarded for finding correlations that are independent of biases in the data generating processes. Classifying data independently of the environment is the definition of invariance in this context and leads to more robustly correlated results. Using the CRISP ensemble, we've identified a set of genes which are predictive of high and low lipid density. These genes are significantly associated with gene sets and genetic pathways that are consistent with NAFLD or similar liver dysfunction. On the other hand, machine learning methods such as empirical risk minimization and random forest do not distinguish between spurious and non-spurious correlations. In our experiment, their top 20 genes are not biologically related to lipid metabolism, despite the random forest classifier having the highest held-out test accuracy (100%) of all the models in predicting lipid density.

The DESeq2 package did not find any genes significantly correlated to our phenotype. The EdgeR package found 6 genes whose expression was significantly correlated to the lipid density response variable—one of which (*Star*) was also found by CRISP. However, gene set enrichment and pathway analysis tools found a wide variety of conditions and processes associated with these 6 genes, leaving the results difficult to interpret. We attribute these quantitative and qualitative differences in results to the environment invariance modeling approach that the CRISP ML algorithms use in conjunction with the data augmentation methods we leveraged.

As opposed to the non-specific pathways enriched by the EdgeR genes, the CRISP genes enrich lipid metabolism pathways including cholesterol, fatty acid, and NAFLD pathways—all of which are consistent with using lipid density as a response in our machine learning models. The NAFLD pathway is significantly enriched by the CRISP gene result set and includes *Cyp2e1*, *Fasn*, and *Scd1* genes with a 0.014 strength of association, as shown in Table 6. As noted in Table 8, the literature is consistent in finding these genes implicated in experiments on NAFLD. Moreover, the SPOKE knowledge graph found MASLD (the new name for NAFLD) to be the most significantly associated disease ontology with the CRISP gene result set.

The use of data augmentation in this CRISP experiment is crucial. Not only does augmentation provide known perturbations of the data as required by CRISP, data augmentation also provides more samples to train, test, and validate the ensemble of machine learning classifiers. Even so, having more samples than 51 would enrich the underlying gene expression signals which would lead to even more robust correlations. We therefore recommend future spaceflight experiments to increase the cohort size and make the liver a standard tissue to collect, process, and analyze. We acknowledge a caveat that the OSD-137 study used the Balb/c mouse strain, while the other three studies used the C57BL6 strain. These two mouse strains are known to have differing responses to spaceflight[5]. Thus, a future study would benefit from further investigating the responses of the different mouse strains in larger cohorts.

To further validate our findings, one could explore other data associated with these same rodents including proteomic and methylation data, as well as leverage existing molecular tests of liver function. A future analysis of the gene expression of ground models of NAFLD may help validate the genes identified here by CRISP. Future work could focus on randomized controlled experiments specifically designed to manipulate the function of each putative gene to verify they do, in fact, cause high or low lipid density in liver tissue. For example, reverse transcriptase PCR or quantitative immunohistochemistry studies on liver samples from controlled experiments could provide key validation results. Adverse outcome pathways (AOPs) related to fatty liver, steatosis, and NAFLD could be explored to identify the molecular pathways responsible for the lipid density phenotype and to gain a more complete purview of the putative dysfunction. Additionally, to further elucidate a link to human manifestation of NAFLD or similar liver dysfunction, it will be important to address the species differences between mouse and human.

In this research, we provide a novel approach to identify potentially causal genes associated with liver dysfunction during spaceflight. These genes constitute potential biomarkers of NAFLD for targeted monitoring or therapeutics development in the future that would otherwise be more time consuming or impossible to identify with traditional statistical or experimental approaches. Given the expense of randomized controlled experiments, having a targeted set of genes putatively causal of the response is invaluable. While the results of our research link the expression of certain genes to a lipid dysregulation phenotype in liver tissue, our approach can be generalized to other tissues, phenotypes, and even other -omics data. For NASA to embark on longer and more frequent space missions, understanding the impact of spaceflight on biological function is paramount. To this end, there are many data sets in OSDR left to explore, and more are continuing to be published.

Data availability

All the mouse liver transcriptomic data and histological lipid data are available at https://osdr.nasa.gov. Direct links to the individual datasets are as follows: https://osdr.nasa.gov/bio/repo/data/studies/OSD-47, https://osdr.nasa.gov/bio/repo/data/studies/OSD-48, https://osdr.nasa.gov/bio/repo/data/studies/OSD-137, https://osdr.nasa.gov/bio/repo/data/studies/OSD-168.

Code availability

All software for this experiment is at https://github.com/nasa/AI4LS/tree/main/crispv1.1.

Received: 12 April 2024; Accepted: 26 November 2024
Published online: 18 January 2025

References

1. Jonscher, K. R. et al. Spaceflight activates lipotoxic pathways in mouse liver. *PLoS ONE* **11**, e0152877 (2016).
2. Beheshti, A. et al. Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver. *Sci. Rep.* **9**, 19195 (2019).
3. Vinken, M. Hepatology in space: Effects of spaceflight and simulated microgravity on the liver. *Liver Int. Liv.* https://doi.org/10.1111/liv.15444 (2022).
4. Garrett-Bakelman, F. E. et al. The NASA twins study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**, eaau8650 (2019).
5. da Silveira, W. A. et al. Comprehensive multi-omics analysis reveals mitochondrial stress as a central biological hub for spaceflight impact. *Cell* **183**, 1185-1201.e20 (2020).
6. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
7. Hariton, E. & Locascio, J. J. Randomised controlled trials—The gold standard for effectiveness research: Study design: Randomised controlled trials. *BJOG Int. J. Obstet. Gynaecol.* **125**, 1716–1716 (2018).
8. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. in *CVPR 2011* 1521–1528 (IEEE, Colorado Springs, CO, USA, 2011). https://doi.org/10.1109/CVPR.2011.5995347.
9. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
10. Ganju, S. *et al.* Learnings from Frontier Development Lab and SpaceML—AI Accelerators for NASA and ESA. (2020) https://doi.org/10.48550/ARXIV.2011.04776.
11. Budd, S. *et al.* Prototyping CRISP: A Causal Relation and Inference Search Platform applied to Colorectal Cancer Data. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* 517–521 (IEEE, Nara, Japan, 2021). https://doi.org/10.1109/LifeTech52111.2021.9391819.
12. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at http://arxiv.org/abs/1907.02893 (2020).
13. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference using invariant prediction: Identification and confidence intervals. Preprint at http://arxiv.org/abs/1501.01332 (2015).
14. Beery, S., van Horn, G. & Perona, P. Recognition in terra incognita. Preprint at https://doi.org/10.48550/ARXIV.1807.04975. (2018)
15. Bühlmann, P. Invariance, Causality and Robustness. Preprint at https://doi.org/10.48550/ARXIV.1812.08233 (2018).
16. Vernice, N. A., Meydan, C., Afshinnekoo, E. & Mason, C. E. Long-term spaceflight and the cardiovascular system. *Preci. Clin. Med.* **3**, 284–291 (2020).
17. Globus, R., Cadena, S. & Galazka, J. Rodent Research-1 (RR1) National Lab Validation Flight: Mouse liver transcriptomic, proteomic, and epigenomic data. NASA GeneLab https://doi.org/10.26030/K5C1-JD05 (2015).
18. Globus, R., Galazka, J., Smith, R. & Cramer, M. Rodent Research-3-CASIS: Mouse liver transcriptomic, proteomic, and epigenomic data. NASA GeneLab https://doi.org/10.26030/9K6W-4C28 (2017).
19. Levene, A. P., Kudo, H., Thursz, M. R., Anstee, Q. M. & Goldin, R. D. Is oil red-O staining and digital image analysis the gold standard for quantifying steatosis in the liver?. *Hepatology* **51**, 1859–1859 (2010).
20. Gonzalez Zelaya, C. V. Towards Explaining the Effects of Data Preprocessing on Machine Learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* 2086–2090 (IEEE, Macao, Macao, 2019). https://doi.org/10.1109/ICDE.2019.00245.
21. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol.* **22**, 258 (2021).
22. Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 1–6 (IEEE, Gold Coast, Australia, 2016). https://doi.org/10.1109/DICTA.2016.7797091.
23. Ilangovan, H. et al. Harmonizing heterogeneous transcriptomics datasets for machine learning-based analysis to identify spaceflown murine liver-specific changes. *npj Micrograv.* **10**, 61 (2024).
24. Ilse, M., Tomczak, J. M. & Forré, P. Selecting data augmentation for simulating interventions. Preprint at https://doi.org/10.48550/ARXIV.2005.01856 (2020).
25. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. *BMC Bioinf.* **19**, 274 (2018).
26. Zhang, Z. et al. Novel data transformations for RNA-seq differential expression analysis. *Sci. Rep.* **9**, 4820 (2019).
27. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

29. Sun, Z. & Zhu, Y. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics* **28**, 2584–2591 (2012).
30. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9**, e85150 (2014).
31. Bernett, J. et al. Guiding questions to avoid data leakage in biological machine learning applications. *Nat Methods* **21**, 1444–1453 (2024).
32. Lones, M. A. Avoiding common machine learning pitfalls. *Patterns* https://doi.org/10.1016/j.patter.2024.101046 (2024).
33. Schapire, R. E. The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
34. Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30 (2000).
35. Morris, J. H. et al. The scalable precision medicine open knowledge engine (SPOKE): A massive knowledge graph of biomedical information. *Bioinformatics* **39**, btad080 (2023).
36. Schriml, L. M. et al. Human disease ontology 2018 update: Classification, content and workflow expansion. *Nucl. Acids Res.* **47**, D955–D962 (2019).
37. Parmar, A., Katariya, R. & Patel, V. A Review on Random Forest: An Ensemble Classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (eds. Hemanth, J., Fernando, X., Lafata, P. & Baig, Z.) vol. 26 758–763 (Springer International Publishing, Cham, 2019).

Acknowledgements

This research was supported by the NASA Human Research Program (HRP), Science Mission Directorate, and Biological and Physical Sciences (BPS) Division. We thank the HRP Radiation Element Translational Radiation Research and Countermeasures team and the GeneLab team for collaboration. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. We also wish to thank Charlotte Nelson from Mate Bioservices, Inc. for providing us the results from the SPOKE tool.

Author contributions

JC designed and executed the CRISP experiments, made the updates to the CRISP software, and wrote most of the manuscript. RS and JG identified the mouse liver transcriptomic and histological lipid data. MM interpreted the pathway analyses. HC analyzed some of the underlying algorithms of the CRISP platform. ASB, AH, LS, GM, and SC reviewed the manuscript and provided integral feedback.

Competing interests

The authors declare no competing financial or non-financial interests with respect to this research.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-81394-y.

Correspondence and requests for materials should be addressed to J.A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.