



**HAL**  
open science

# Machine Learning classifier built with heavy metal signature biomarker genes as features to distinguish between heavy metal exposure from non-heavy metal exposure gene expression samples

Shradha Mukherjee

► **To cite this version:**

Shradha Mukherjee. Machine Learning classifier built with heavy metal signature biomarker genes as features to distinguish between heavy metal exposure from non-heavy metal exposure gene expression samples. 2023. hal-04084188

**HAL Id: hal-04084188**

**<https://hal.science/hal-04084188v1>**

Preprint submitted on 27 Apr 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **Title:** Machine Learning classifier built with heavy metal signature biomarker genes as  
2 features to distinguish between heavy metal exposure from non-heavy metal exposure  
3 gene expression samples.

4  
5 **NAMES AND AFFILIATIONS OF EACH AUTHOR**

6 Shradha Mukherjee, PhD

7 Independent Researcher.

8 This work is self-owned and not affiliated with any organization/university.

9  
10 **CORRESPONDING AUTHOR'S NAME AND EMAIL ADDRESS**

11 Address correspondence to: Shradha Mukherjee; Independent Researcher, Calcutta,  
12 West Bengal, India

13 Email [smukher2@yahoo.com](mailto:smukher2@yahoo.com)

14  
15 **FINANCIAL INTERESTS OR CONFLICT OF INTEREST (IF APPLICABLE)**

16 NA

17  
18 **Abstract:**

19 There are over 350,000 registered chemicals and chemical combinations in use today  
20 globally. This is a public health concern and an active area of research, as for majority  
21 of these chemicals no scientific data is available on potential adverse effects on human  
22 health. Both U.S. Environmental Protection Agency (EPA) and World Health  
23 Organization (WHO) have listed heavy metals, lead (Pb), mercury (Hg), cadmium (Cd)  
24 and arsenic (As) among the top chemicals of public health concern. Adverse health  
25 effects, induced by heavy metals and other non-heavy metal chemicals include  
26 neurodegenerative diseases (Alzheimer's disease, Parkinson's disease), cognitive  
27 decline, behavioral problems, kidney diseases, cancer and cardiovascular diseases.  
28 Thus, it is important to detect not only active chemical exposure but also past chemical  
29 exposures. In this paper differential gene expression (DEG) analysis and machine  
30 learning (ML) were combined to identify differentially expressed genes (DEGs) or heavy  
31 metal toxicity signature genes that were used as features in ML to classify test samples  
32 into heavy metal and non-heavy metal control groups. From NIH-GEO, RNA-seq gene  
33 expression data from a total of 827 human neuronal cell culture samples treated with 87  
34 different chemicals were downloaded and normalized. Two groups of DEGs consisting  
35 of 80 genes (consensus of limma, edgeR and simple DEG analysis) and 879 genes  
36 (consensus of at least 2 of the three DEG methods limma, edgeR and simple) were  
37 identified and designated as heavy metal biomarkers. The heavy metal biomarker gene  
38 sets were enriched with metal metabolism gene ontology, kidney disease and cancer  
39 diseases genes. Comparison of different ML models built with 80 DEGs and 879 DEGs  
40 showed that Logistic Regression and Support Vector Machine (SVM) were accurate  
41 (>90% success in classifying test samples into heavy metal and non-heavy metal  
42 groups) for both 80 DEG and 879 DEG features. In this paper, a combined DEG  
43 analysis and ML pipeline has been developed that can successfully detect heavy metal  
44 exposure from gene expression data. This pipeline can be applied for identification of  
45 chemical exposure, which is the first step for developing a treatment plan for patients  
46 exposed to toxic chemicals.

47 **Introduction:**

48 There are over 350,000 registered chemicals and chemical combinations in use today  
49 globally[1]. This is a public health concern and an active area of research, as for  
50 majority of these chemicals no scientific data is available on potential adverse effects on  
51 human health[2]. Children are especially susceptible to chemical exposure induced  
52 diseases such as autism, cerebral palsy, mental retardation, obesity and respiratory  
53 diseases[3]. Both U.S. Environmental Protection Agency (EPA) and World Health  
54 Organization (WHO) have listed heavy metals, lead (Pb), mercury (Hg), cadmium (Cd)  
55 and arsenic (As) among the top chemicals of public health concern[4; 5]. Heavy metal  
56 refers to chemical elements with high molecular weight that cause toxicity in humans.  
57 As heavy metals accumulate in the human body overtime (bioaccumulate), exposure to  
58 even small amounts of heavy metals can cause toxicity and adverse health effects.  
59 Adverse health effects, induced by heavy metals and other non-heavy metal chemicals  
60 include neurodegenerative diseases (Alzheimer's disease, Parkinson's disease),  
61 cognitive decline, behavioral problems, kidney diseases, cancer and cardiovascular  
62 diseases[6].

63  
64 Long-term adverse effects induced by chemical exposure in humans, remain even when  
65 the chemical exposure itself has stopped. Animal studies showed that chronic lead (Pb)  
66 exposure results in lead accumulation in choroid plexus and correlates with reduced  
67 production of transthyretin by choroid plexus, required for regulation of thyroid hormone  
68 critical for prenatal and early postnatal development [7; 8]. Chronic exposure to  
69 chromium (Cr) and arsenic (As) through drinking water increased incidence and size of  
70 tumors in mice [9]. Exposure to low levels of mercury (Hg) in form of methylmercury  
71 (MeHg) prenatally was associated with learning and memory deficits in children [10].  
72 Long-term effects may occur because of persistent gene expression changes that were  
73 induced by the chemical during exposure, gene expression changes induced by release  
74 of bioaccumulated chemicals later in life or due to other hereto unknown reasons.  
75 Though the mechanism of how chemical induced gene effects persist after chemical  
76 exposure has stopped is not clearly understood, it is certain that as genes regulate all  
77 biological phenotypes, detection of these gene expression changes during the chemical  
78 exposure and after chemical exposure has stopped, is the first step towards  
79 understanding and treating the adverse disease phenotypes. Identification of chemicals  
80 by analyzing its induced gene expression profile, must also account for presence of  
81 other chemicals that work in combination with it, to produce a net chemical induced  
82 gene expression profile in humans.

83  
84 Accurate toxicity predictions have the potential can reduce uncertainty and expense of  
85 clinical trials of drugs. To decipher chemical effects, artificial intelligence (AI) and  
86 machine learning (ML) techniques are now being widely used. AI/ML algorithms can  
87 learn from patterns in chemical assay results (activity) and chemical characteristics  
88 (structural) data from known chemicals, and build AI/ML models that can make  
89 predictions about unknown chemicals. Stress assay results from chemical-protein  
90 binding assays, known toxicity endpoints (hepatotoxicity, oral toxicity, cardiotoxicity,  
91 mutagenicity) and types of chemical functional groups in chemical structure are some of  
92 the training data features employed to build toxicity predictor AI/ML models. These

93 chemical structure and activity parameters called Quantitative Structure-Activity  
94 Relationships (QSAR) parameters are commonly used for AI/ML methods, while  
95 chemical induced gene expression data is not always used as training data features to  
96 build toxicity predictor AI/ML models. eToxPred is a ML model trained on chemical  
97 structural features and toxicity classification data for known chemicals, and can predict  
98 toxicity of new unknown chemicals [11]. DeepTox, a deep learning model normalizes  
99 and computes features to build a machine learning model that can predict toxicity of  
100 new unknown chemical compounds [12]. DTox (Deep Learning for Toxicology) is a  
101 deep learning model that takes as input chemical structure to predict its probable  
102 toxicity assay result and part of its gene expression profile underlying toxicity effects  
103 [13].

104  
105 Neurotoxicity of a chemical is an underutilized toxicity endpoint relative to other toxicity  
106 endpoints hepatotoxicity, oral toxicity, cardiotoxicity and mutagenicity in AI/ML models  
107 because its difficult to maintain neural samples in culture that represent in vivo  
108 neurotoxicity. Now, its known that AI/ML models perform better when they are trained  
109 with more volume of data and more data types. However, comparison of ability of ML  
110 models to reliably predict chemical toxicity was shown to be more robust for 2D neural  
111 cell culture, than 3D neural cell culture (organoid) [14; 15]. As more drug induced gene  
112 expression data gets generated from chemical treatment of 2D and 3D neural cell  
113 culture assays, it would be a missed opportunity to not incorporate all data sources, 2D  
114 and 3D, to increase size and type of data utilized for building AI/ML models and  
115 potentially increase performance of AI/ML models. Variability in gene expression data  
116 originating from different sources, is a known bottleneck for gene expression based  
117 meta-analysis. Thus in the present study it was hypothesized that by reducing batch or  
118 source variability in gene expression data using normalization, it will be possible to use  
119 data from different sources for training a reliable and robust AI/ML model for prediction  
120 of chemical toxicity. The goal of the present study was to develop a computational  
121 pipeline that would normalize and enable utilization of different data types, 2D and 3D,  
122 and from different sources, for building a reliable AI/ML model for toxicity prediction,  
123 specifically heavy metal toxicity. Here, a pipeline was developed to normalize gene  
124 expression data from different sources and identify heavy metal signature genes  
125 (biomarkers). The biomarker genes were used as features train an AI/ML model for  
126 prediction of heavy metal toxicity. The ML algorithms, logistic regression and Support  
127 Vector Machine (SVM) were able to distinguish between heavy metal and non-heavy  
128 metal samples for biomarker gene sets of both 80 genes and 879 genes.

129  
130

## 131 **Methods**

### 132 Code Availability

133 Computational code with html or pdf rendering showing input and output of code chunks  
134 is available as a git local repository at  
135 <https://icedrive.net/s/h3P65RbNvf5Dh8yT1DabXxyNgWq6> with all files and as a git  
136 remote repository at [https://gitlab.com/smukher2/pbothers\\_rnaseq\\_ml\\_feb2023](https://gitlab.com/smukher2/pbothers_rnaseq_ml_feb2023) with  
137 large files ignored or removed.

138

139  
140 Source of RNA-seq gene expression data  
141 NCBI GEO repository <https://www.ncbi.nlm.nih.gov/geo/>, was searched for RNA-seq  
142 gene expression data from human pluripotent or embryonic stem cell derived neural  
143 tissue culture samples exposed to different chemicals. The results were filtered to retain  
144 those datasets, which contained untreated and/or DMSO solvent treated samples. The  
145 series numbers GSE166297, GSE128431, GSE63935 and GSE126786 were selected  
146 as they met the filtering criteria.

147  
148 Normalization of RNA-seq gene expression data  
149 RNA-seq raw counts are number of reads overlapping annotated genes in human gene  
150 annotation file Homo\_sapiens.GRCh38.p13.gtf. RNA-seq raw counts generated by  
151 NCBI were obtained from NCBI GEO repository,  
152 <https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html#locate>. A bash script with  
153 commandline tool wget was used to programmatically download the NCBI generated  
154 raw RNA-seq counts available at url of format  
155 <https://www.ncbi.nlm.nih.gov/geo/download/?acc=<GSE Series ID> example>,  
156 <https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE166297>, under 'Series RNA-seq  
157 raw counts matrix' at url of format  
158 [https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq\\_counts&acc=<GSE Series ID>&format=file&file=<GSE Series ID>\\_raw\\_counts\\_GRCh38.p13\\_NCBI.tsv.gz](https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq_counts&acc=<GSE Series ID>&format=file&file=<GSE Series ID>_raw_counts_GRCh38.p13_NCBI.tsv.gz)  
159 example,  
160 [https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq\\_counts&acc=GSE166297&format=file&file=GSE166297\\_raw\\_counts\\_GRCh38.p13\\_NCBI.tsv.gz](https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq_counts&acc=GSE166297&format=file&file=GSE166297_raw_counts_GRCh38.p13_NCBI.tsv.gz). In R/Rstudio using  
161 package limma\_3.38.3, the downloaded NCBI generated raw counts were quantile  
162 normalized and scaled to log<sub>2</sub>+1. Effect of limma quantile model normalization on batch  
163 variation was visualized by comparing pre-limma quantile normalized  
164 (before) and limma quantile normalized (after) expression data plots. Both raw counts  
165 (not scaled to log<sub>2</sub>+1) and normalized counts (scaled to log<sub>2</sub>+1), were visualized with  
166 density plots, Box-Whiskers plots, correlation plots and PCA analyses to estimate effect  
167 of normalization on variability between samples from different GSE series. Density plots  
168 and Box-Whiskers plots were made using ggplot2\_3.1.1, while prcomp stats4\_3.5.0 and  
169 corrplot\_0.84, were used to perform PCA analyses and correlation analysis,  
170 respectively [16; 17].

171  
172  
173  
174 Source of metadata for RNA-seq gene expression data  
175 GEOquery R package was used to obtain metadata for all samples in the GSE series  
176 numbers GSE166297, GSE128431, GSE63935 and GSE126786 [18]. The metadata  
177 fields, sample id, source tissue (embryonic stem cells or pluripotent stem cells derived  
178 neural tissue), organism (homo sapiens) and chemical name (87 unique chemicals) was  
179 fetched with GEOquery for each GSE series. In the list of chemicals, lead, arsenic,  
180 cadmium and mercury were heavy metals and others were non-heavy metal chemicals.

181  
182 Identification of heavy metal biomarkers  
183 To identify heavy metal biomarker genes, normalized log<sub>2</sub>+1 expression values were  
184 used as input for estimation of DEGs (Differentially Expressed Genes). Gene

185 expression normalization and differential gene expression (DEG) analysis was done  
186 using previously published methods [19; 20]. Differentially expressed genes (DEGs)  
187 were identified by comparing gene expression of heavy metal and non-heavy metal  
188 exposed samples using limma\_3.38.3, edgeR\_3.24.3 and simple comparison  
189 expression means [21; 22; 23]. In limma and edgeR model design step variable  
190 (metadata fields) requires variables have more than one value, thus organism was not  
191 used as it contained only one value 'homo sapiens'. DEGs between heavy metal and  
192 non-heavy metal was calculated while correction for false discovery and multiple testing  
193 was done using Benjamini and Hochberf (BH) to correct for the presence of multiple  
194 genes or features in the analysis [24]. In limma and edgeR, significantly upregulated  
195 genes in heavy metal category with BH corrected adjP-values <0.05 and a fold changes  
196 >5, were designated as heavy metal specific genes or biomarker genes. In simple  
197 comparison of means method of DEG identification, gene expression from heavy metal  
198 and non-heavy metal samples was compared, and genes with P-values <0.05 and a  
199 fold changes >5, were designated as heavy metal specific genes or biomarker genes.  
200 Gene expression of DEGs for all three methods of DEG analysis, was visualized with  
201 volcano plots, Box-Whiskers plots, barplots and density plots using ggplot2\_3.1.1 R  
202 package[17]. Consensus strict DEGs were those DEGs that were present in DEG lists  
203 of all three methods, limma, edgeR and simple comparison of means. Consensus  
204 relaxed DEGs were those DEGs that were present in DEG lists of atleast two of the  
205 three methods, limma, edgeR and simple comparison of means. Significance of overlap  
206 between DEG lists were calculated using GeneOverlap\_1.18.0 R package and number  
207 of genes overlapping were visualized as Venn-diagrams using VennDiagram\_1.6.20 R  
208 package[25; 26].

209

#### 210 Gene Ontology (GO) analysis of heavy metal biomarker genes

211 To determine biological significance of the heavy metal biomarkers, EnrichR\_1.0 a R  
212 package was performed to determine biological process gene ontology (GO) enriched in  
213 the heavy metal strict DEGs and relaxed DEGs [27]. GOs with adjP-values <0.05 were  
214 considered significant GOs.

215

#### 216 Machine Learning Models

217 Two sets of differentially expressed genes were used as features (set of 80 DEGs  
218 consensus of all 3 methods and set of 880 DEGs consensus of atleast 2 methods) to  
219 train the machine learning algorithm with gene expression data for classifying or  
220 distinguishing between heavy metal exposure and non-heavy metal exposure human  
221 samples. To avoid class sample size bias, the non-heavy metal chemical exposure  
222 samples were not included in the ML training, as there were more non-heavy metal  
223 chemical exposure samples than heavy metal chemical exposure samples. Only 33  
224 heavy metal chemical exposure samples, 25 untreated control samples and 15 solvent  
225 DMSO treated control samples, were included in the ML training. In  
226 Python/Spyder/Jupyter Notebook, samples were split into training (70%) and test  
227 datasets (30%) using scikit-learn train\_test\_split function. The models were built with  
228 scikit-learn using five ML methods Logistic Regression, Support Vector Machine (SVM),  
229 Naïve Bayes, K-means, Random Forest, and XGB.

230

231 For ML model building, supervised learning methods learn about patterns in dataset  
232 from labeled datasets, while unsupervised learning methods learn about patterns in the  
233 dataset without being given labeled datasets. Logistic Regression is a supervised  
234 learning method, which relies on odds ratio calculation to predict the probability of an  
235 event (here heavy metal and non-heavy metal) occurrence depending on value of input  
236 features (here biomarker genes). SVM is a supervised learning method, which uses  
237 labeled training dataset to create a partition or hyperplane that separates the groups  
238 (here heavy metal and non-heavy metal) using input features (here biomarker genes).  
239 Naïve Bayes, is a supervised learning method, which classifies based on probability of a  
240 feature (here biomarker genes) occurring in a group (here heavy metal and non-heavy  
241 metal). K-means, is an unsupervised leaning method, in which the datasets are split into  
242 clusters or groups (here heavy metal and non-heavy metal) based on their input  
243 features (here biomarker genes), and by an iterative process the centroid of the cluster  
244 is matched with the mean of the samples assigned into the cluster until changing  
245 assignment of samples in the cluster does not change the mean or centroid does not  
246 move anymore. Random Forest is a supervised learning method, which consists of  
247 many decision trees built using features (here biomarker genes) and prediction of class  
248 (here heavy metal and non-heavy metal) is done by calculating average prediction of  
249 each tree. XGB or XGBoost, is another decision tree based method that is fast because  
250 of parallel processing of decision trees.

#### 251 252 Evaluation of Machine Learning (ML) heavy metal and non-heavy metal classifier 253 models

254 To compare performance of these ML models, accuracy, f1-score, ROC curve and AUC  
255 were calculated and confusion matrix was plotted. To visually display quality of  
256 classification models a confusion matrix was plotted for each using scikit-learn\_1.2.1  
257 python package[28]. In confusion matrix, the columns represent model predicted labels  
258 and rows represent true labels of samples in the test dataset. Receiver Operating  
259 Characteristic (ROC) curve was plotted using scikit-learn\_1.2.1 python package, where  
260 Area Under the Curve (AUC) in the ROC plot indicates resolving power of classifier  
261 models[28]. Accuracy, precision, recall and f1-score were calculated using  
262 classification\_report function of the scikit-learn\_1.2.1 python package[28].

#### 263 264 265 **Results:**

266 Quantile normalization reduces batch effects in RNA-seq gene expression count data  
267 As publicly available chemical exposure induced neural tissue gene expression count  
268 data was obtained from different research studies or batches, the counts were  
269 normalized to reduce cross-study variability and effect of normalization was visualized  
270 with different plots. Density and Box-Whiskers plots showed a greater overlap of  
271 samples from different batches and greater overlap of gene expression means,  
272 respectively, in limma quantile normalized count data relative to pre-normalized count  
273 data (Figure 1A, Figure 1B). Correlation plots showed greater correlation of batches  
274 after quantile normalization relative to that before quantile normalization (Figure 1C).  
275 PCA plots showed that variability or wide dispersion of samples was less in quantile  
276 normalized count data relative to pre-normalized count data (Figure 1D). Thus, limma

277 quantile normalization reduced variability unrelated to chemical exposure gene  
278 expression counts, making the data suitable for further analysis.

279  
280 Limma identified 85 DEGs upregulated in heavy metals relative to other samples

281 To identify heavy metal signature genes, heavy metal samples were contrasted with  
282 samples that were treated with non-heavy metal samples (other chemical treatments  
283 and untreated controls), using limma DEG analysis. DEGs significantly upregulated or  
284 downregulated ( $p$ -value  $< 0.05$ ) in heavy metals relative to other samples were  
285 visualized in volcano plot (Figure 2A). From the DEGs, 85 genes were significantly ( $p$ -  
286 value  $< 0.05$ ) upregulated (fold-change  $> 5$ ) in heavy metals relative to other samples  
287 were putative heavy metal signature genes. Density, Box-Whiskers and bar plots  
288 showed that the average normalized gene expression count of upregulated heavy metal  
289 DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure  
290 2B, Figure 2C, Figure 2D).

291  
292 EdgeR identified 1072 DEGs upregulated in heavy metals relative to other samples

293 To identify heavy metal signature genes, heavy metal samples were contrasted with  
294 samples that were treated with non-heavy metal samples (other chemical treatments  
295 and untreated controls), using edgeR DEG analysis. DEGs significantly upregulated or  
296 downregulated ( $p$ -value  $< 0.05$ ) in heavy metals relative to other samples were  
297 visualized in volcano plot (Figure 3A). From the DEGs, 1072 genes were significantly  
298 ( $p$ -value  $< 0.05$ ) upregulated (fold-change  $> 5$ ) in heavy metals relative to other samples  
299 were putative heavy metal signature genes. Density, Box-Whiskers and bar plots  
300 showed that the average normalized gene expression count of upregulated heavy metal  
301 DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure  
302 3B, Figure 3C, Figure 3D).

303  
304 Simple method identified 2237 DEGs upregulated in heavy metals relative to other  
305 samples

306 To identify heavy metal signature genes, heavy metal samples were contrasted with  
307 samples that were treated with non-heavy metal samples (other chemical treatments  
308 and untreated controls), using edgeR DEG analysis. DEGs significantly upregulated or  
309 downregulated ( $p$ -value  $< 0.05$ ) in heavy metals relative to other samples were  
310 visualized in volcano plot (Figure 4A). From the DEGs, 2237 genes were significantly  
311 ( $p$ -value  $< 0.05$ ) upregulated (fold-change  $> 5$ ) in heavy metals relative to other samples  
312 were putative heavy metal signature genes. Density, Box-Whiskers and bar plots  
313 showed that the average normalized gene expression count of upregulated heavy metal  
314 DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure  
315 4B, Figure 4C, Figure 4D).

316  
317 Heavy metal biomarker genes identified by overlapping of limma, edgeR and simple  
318 DEGs

319 Overlapping DEGs identified by limma, edgeR and simple comparison of means  
320 methods, identified 80 DEGs common to all 3 methods (strict overlap), and 879 DEGs  
321 common to at least 2 methods (relaxed overlap) (Figure 5A). Jaccard index calculations



322 showed that edgeR and limma has significant overlap, while simple method did not  
323 significantly overlap with edgeR or limma (Figure 5B).

324

### 325 Gene Ontology (GO) of heavy metal biomarkers

326 Metal related cellular response and homeostasis were the most enriched biological  
327 processes in heavy metal GO Biological Process (Figure 5C, Figure 5D). GOs related to  
328 cellular response to metals, "cellular response to zinc ion" (GO:0071294 p-values  
329 1.71E-10 and 1.87E-05), "response to copper ion" (GO:0046688 p-values 1.71E-10 and  
330 1.87E-05) and "cellular response to cadmium ion" (GO:0071276 p-values 1.47E-09 and  
331 5.11E-05), were the most significant GOs in both heavy metal biomarkers in strict group  
332 of 80 genes and relaxed group of 879 genes (Figure 5C, Figure 5D). These results are  
333 consistent with the group of genes being heavy metal biomarkers. GOs "cellular  
334 response to unfolded protein" (GO:0034620 p-value 5.29E-08) and "negative regulation  
335 of growth" (GO:0045926 p-value 4.88E-09) were significant in heavy metal biomarkers  
336 in strict group of 80 genes (Figure 5C).

337

### 338 Performance of ML models with heavy metal 80 and 879 biomarker as features

339 To build and test ML models, randomly picked 70% of samples (25 heavy metal and 26  
340 control DMSO or untreated samples) were used for training and 30% of samples (8  
341 heavy metal and 14 control DMSO or untreated samples) were used as testing datasets  
342 (Figure 6A, Figure 7A). A total of seven models, Logistic Regression, K-means, Naïve-  
343 Bayes, SVM, Random Forest, XGB with grid and XGB without grid were compared to  
344 determine their ability to distinguish or classify heavy metal and control non-heavy metal  
345 samples. The performance of the models was compared using confusion matrix,  
346 accuracy, precision, f1-score, recall, ROC and AUC (Figure 6, Figure 7). For 80 heavy  
347 metal biomarker genes as features, both Logistic Regression and SVM outperformed  
348 other models, with 95.4% accuracy, 100% precision, 0.94 AUC, 0.875 recall and 0.933  
349 f1-score (Figure 6I). Random Forest also showed good performance, with 90.9%  
350 accuracy, 100% precision, 0.88 AUC, 0.75 recall and 0.857 f1-score (Figure 6I). Other  
351 models, K-Means, Naïve-Bayes, XGB with grid and XGB without grid showed an  
352 accuracy between 27.27% to 81.81% (Figure 6I). For 879 heavy metal biomarker genes  
353 as features, both Logistic Regression and SVM outperformed other models, with 95.4%  
354 accuracy, 100% precision, 0.94 AUC, 0.875 recall and 0.933 f1-score (Figure 7I). XGB  
355 without grid also showed good performance, with 90.9% accuracy, 100% precision, 0.88  
356 AUC, 0.75 recall and 0.857 f1-score (Figure 7I). Other models, K-Means, Naïve-Bayes,  
357 Random Forest and XGB with grid showed an accuracy between 45.45% to 86.36%  
358 (Figure 7I). For both 80 genes and 879 gene features (heavy metal biomarkers),  
359 confusion matrix constructed from the different ML models showed visually the true and  
360 predicted, heavy metal and control non-heavy metal labels consistent with the precision,  
361 AUC, recall and f1-score calculations (Figure 6B-H, Figure 7B-H). Taken together, these  
362 results show that Logistic Regression and SVM outperform other models with 879 gene  
363 features, as well as 80 gene features.

364

365

366

367

368 **Discussion:**

369 *Neurotoxicity cell culture models*

370 The nervous system is protected by the blood-brain-barrier from many infectious agents  
371 and harmful chemicals. However, this protection weakens making the nervous system  
372 susceptible to chemicals such as neurotoxic heavy metals, when the blood-brain-barrier  
373 shows dysfunction in adults with aging and disease [29; 30]. Thus, it is important to  
374 study neurotoxicity to enable detection of neurotoxicity, understand its molecular  
375 mechanisms and discover therapeutic targets for intervention. However, this has been  
376 challenging as there are no standard neural cell culture protocols, so gene expression  
377 data originating from different 2D or 3D neural culture conditions is highly variable. In  
378 iPSC derived organoids developed to study Alzheimer's disease, large variability was  
379 found in morphology and electrophysiological activity of neurons inside organoids even  
380 when they were prepared from the same cell line [31]. This variability could result from  
381 inherent stochastic nature of in vitro self-organization which makes neural differentiation  
382 process and neuronal cell type characteristics inside organoids variable. In the present  
383 study, a pipeline was developed to reduce variability in gene expression with  
384 normalization to make it usable for detection of heavy metal neurotoxicity with ML using  
385 heavy metal molecular biomarkers as features.

386

387 *Meta-analysis and normalization of RNA-seq*

388 Most gene expression RNA-seq raw reads from published studies are stored in National  
389 Centre for Biotechnology Information Sequence Read Archive (NCBI SRA) or National  
390 Centre for Biotechnology Information Gene Expression Omnibus (NCBI GEO) public  
391 repository that can be used for meta-analysis. Meta-analysis involves integration of  
392 samples from different studies related to a research topic, to increase sample size which  
393 improves robustness of results. Different studies have processing variability as they use  
394 different methods to convert raw RNA-seq reads to annotated gene expression [32].  
395 Thus, for uniformity in meta-analysis combining gene expression data from different  
396 studies requires extensive pre-processing where all the raw reads from different studies  
397 are processed with same pipeline to annotated gene expression. To overcome this  
398 caveat of having to re-run RNA-seq data from different studies available in the public  
399 NCBI GEO repository, NCBI GEO made annotated gene expression available for RNA-  
400 seq studies in their repository. Annotated gene expression has variability due to different  
401 source of tissue origin that can be normalized using several methods such as TPM,  
402 FPKM and quantile normalization. Comparison of these normalization methods in a  
403 study of human tumour xenograft showed normalization of counts had lower median  
404 coefficient of variation than FPKM and TPM normalization [33]. Thus, to reduce variation  
405 in annotated gene expression downloaded from NCBI GEO, quantile normalization  
406 method was applied. Quantile normalization of annotated gene expression for meta-  
407 analysis of combined GSE166297, GSE128431, GSE63935 and GSE126786 batches,  
408 significantly reduced batch variability (Figure 1).

409

410 *Feature or biomarker gene selection for ML*

411 Feature selection is process of selecting most informative features, here genes, which  
412 are most likely to be relevant distinguishing characteristics of the sample labels, here  
413 heavy metal toxicity and non-heavy metal toxicity. More features a ML model has, more

414 time it takes to run and more challenging it becomes to understand effects of features  
415 on classification prediction made by ML model. Feature selection algorithms can be  
416 broadly classified into filter methods, wrapper methods, embedded methods and hybrid  
417 methods [34]. However, these feature selection methods are based on 'trial and error'  
418 where features or genes are put through several rounds of iteration, features are pruned  
419 after each iteration and model is re-build with only those features that are most relevant  
420 to ML model's prediction accuracy. In the present study to make feature selection for  
421 more explainable, instead of prevalent computational feature selection methods,  
422 bioinformatics DEG selection methods were used. It was hypothesized that DEG genes  
423 that are significantly upregulated in heavy metals toxicity, designated heavy metal  
424 toxicity biomarkers, by virtue of their biological relevance will help create better ML  
425 models. Heavy metal DEGs or biomarkers that were upregulated significantly by at least  
426 5 fold, were identified with three commonly used DEG methods, limma, edgeR and  
427 simple comparison of means (Figure 2, Figure 3, Figure 4). Overlap of all 3 DEG  
428 methods resulted in 80 genes (strict overlap) and overlap of at least 2 DEG methods  
429 resulted in 879 (relaxed overlap) genes (Figure 5A, Figure 5B). ML models were built  
430 with both 80 genes and 879 genes, so that robustness of the ML models could be  
431 tested.

432

#### 433 *Insights into selected biomarker features from GO analysis*

434 GO analysis of 80 genes and 879 genes was done to understand molecular mechanism  
435 of heavy metal neurotoxicity and make the ML model more explainable, by  
436 determination of major biological groups to which the selected features belonged. All the  
437 GOs related to cellular response to metals, contained metallothionein (MT) genes  
438 MT2A, MT1M, MT1F, MT1G, MT1X, MT1H and MT1E (Figure 5C, Figure 5D).  
439 Metallothionein proteins are rich in amino acid cysteine (formula  $\text{HOOC-CH}(-\text{NH}_2)\text{-CH}_2\text{-SH}$ ) and the sulphur (S) gives them ability to bind with metal ions. MT-1 and  
440 MT2 family of MTs are induced by presence of metals such as zinc (Zn), cadmium (Cd),  
441 copper (Cu) and lead (Pb) [35; 36; 37]. On the DNA, MT genes contain metal response  
442 elements (MREs) on their promoters, that get activated by binding of metal regulatory  
443 transcription factor (MTFs). The binding of MTFs to MREs and consequent activation of  
444 MT genes is regulated by present of heavy metals [36; 38]. MTs are required for zinc  
445 (Zn) and copper (Cu) homeostasis, oxidative stress response and detoxification of  
446 heavy metals from the body [39; 40].

448

#### 449 *Robustness of ML models for heavy metal toxicity classification*

450 AI and ML models are being widely investigated, and applied in biomedical field.  
451 Artificial intelligence (AI) used in monitoring cancer disease state by analysis of  
452 radiographic images from patients can provide quantitative assessment, while  
453 physicians can only provide a qualitative assessment [41]. Machine Learning (ML)  
454 analysis revealed cancer biomarkers and therapeutic targets in soft tissue sarcoma  
455 datasets [42]. Research into ML models for classification is focused on development of  
456 new ML models (algorithms), reducing errors in ML models and finding avenues where  
457 ML models can be applied. Thus, here existing ML models were applied and tested in  
458 the context of heavy metal exposure detection. As the dataset was imbalanced, with  
459 more samples for other types of chemicals than heavy metals, to avoid overfitting issue,

460 for training and testing of ML models, only heavy metals and controls were used as their  
461 sample size was comparable (Figure 6A, Figure 7A). This comparable number of  
462 samples in each group is conducive to Logistic Regression modelling and makes it well  
463 suited for comparison with other models. Comparison of ML models to determine their  
464 ability to distinguish or classify heavy metal and control non-heavy metal samples,  
465 showed that Logistic Regression and Support Vector Machine (SVM) outperformed  
466 other models (Figure 6, Figure 7). Logistic Regression and SVM performances were  
467 consistent for both 80 feature genes, and 879 feature genes, with 95.4% accuracy,  
468 100% precision, 0.94 AUC, 0.875 recall and 0.933 f1-score (Figure 6B, Figure 6E,  
469 Figure 6I, Figure 7B, Figure 7E, Figure 7I). Random Forest does not work well with  
470 missing values, while XGB can automatically fill missing values. Now as there were no  
471 missing values of gene expression for features of 80 genes and 879 genes, expectedly  
472 performance of Random Forest and XGB with/without gird were comparable. The  
473 performance of Random Forest and XGB with/without gird, were moderate for both 80  
474 feature genes, and 879 feature genes, with range 72 to 90% accuracy, range 66 to  
475 100% precision, range 0.68 to 0.81 AUC, range 0.5 to 0.75 recall and range 0.57 to 0.85  
476 f1-score (Figure 6F, Figure 6G, Figure 6H, Figure 6I, Figure 7F, Figure 7G, Figure 7H,  
477 Figure 7I). The performance of K-Means got worse with increase in number of features  
478 (accuracy 27.2% for 80 genes and 68.1% for 879 genes), while the performance of  
479 Naïve Bayes got better with increase in number of features (accuracy 63.6% for 80  
480 genes and 45.4% for 879 genes) (Figure 6C, Figure 6D, Figure 6I, Figure 7C, Figure  
481 7D, Figure 7I). This could be because Naïve Bayes model works better with high  
482 dimension data, so with more features its performance got better.

#### 483 484 *Summary*

485 This pipeline combines normalization, DEG analysis, GO analysis, and ML modelling  
486 that is a reusable in silico method that can be adapted for assay of various potentially  
487 toxic chemicals. This pipeline can be re-used for other datasets to study effects of  
488 chemical exposure by detection of neurotoxicity, understand molecular mechanism and  
489 discover therapeutic targets for toxic chemicals. Neurodegenerative diseases, bone  
490 diseases and cancers are some of the adverse effects of heavy metal toxicity. Detection  
491 of active and past transient heavy-metal chemical exposure is critical to device a  
492 treatment plan and plan lifestyle changes to safeguard the patient from adverse short-  
493 term and long-term effects of heavy metal toxicity. Robust classification of patients into  
494 putative heavy metal and non-heavy metal exposure classes, based patient's gene  
495 expression profile, will help detect cases of heavy metal toxicity. These results can then  
496 guide a healthcare provider to take necessary actions to treat the patient for heavy  
497 metal toxicity. The methods developed in this paper can also be applied and extended  
498 to distinguish between any other toxic chemicals or chemical combinations and  
499 untreated controls. Chemicals for which toxicity scientific data is not available but  
500 chemical induced gene expression profile is available in patients, ML methods  
501 developed in this paper can be used to determine if the chemical induced gene  
502 expression profile is more like toxic chemicals or untreated controls. For example, gene  
503 expression profile from a patient could be run through the ML model and scored for  
504 similarity with available toxic chemical and non-toxic chemical gene expression profiles.

505

506 **Bibliography:**

- 507 [1] Z. Wang, G.W. Walker, D.C.G. Muir, and K. Nagatani-Yoshida, Toward a Global Understanding of  
508 Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical  
509 Inventories. *Environmental science & technology* 54 (2020) 2575-2584.
- 510 [2] E.D. Pellizzari, T.J. Woodruff, R.R. Boyles, K. Kannan, P.I. Beamer, J.P. Buckley, A. Wang, Y. Zhu, and  
511 D.H. Bennett, Identifying and Prioritizing Chemicals with Uncertain Burden of Exposure:  
512 Opportunities for Biomonitoring and Health-Related Research. *Environ Health Perspect* 127  
513 (2019) 126001.
- 514 [3] L.R. Goldman, and S. Koduru, Chemicals in the environment and developmental toxicity to children: a  
515 public health and policy perspective. *Environ Health Perspect* 108 Suppl 3 (2000) 443-8.
- 516 [4] WHO, 10 chemicals of public health concern. [https://www.who.int/news-room/photo-story/photo-](https://www.who.int/news-room/photo-story/photo-story-detail/10-chemicals-of-public-health-concern)  
517 [story-detail/10-chemicals-of-public-health-concern](https://www.who.int/news-room/photo-story/photo-story-detail/10-chemicals-of-public-health-concern) (2020).
- 518 [5] U.S.E.P.A. (EPA), Chemicals and Toxics Topics \_ US EPA. [https://www.epa.gov/environmental-](https://www.epa.gov/environmental-topics/chemicals-and-toxics-topics)  
519 [topics/chemicals-and-toxics-topics](https://www.epa.gov/environmental-topics/chemicals-and-toxics-topics) (2022).
- 520 [6] M. Balali-Mood, K. Naseri, Z. Tahergorabi, M.R. Khazdair, and M. Sadeghi, Toxic Mechanisms of Five  
521 Heavy Metals: Mercury, Lead, Chromium, Cadmium, and Arsenic. *Front Pharmacol* 12 (2021)  
522 643972.
- 523 [7] W. Zheng, W.S. Blaner, and Q. Zhao, Inhibition by lead of production and secretion of transthyretin in  
524 the choroid plexus: its relation to thyroxine transport at blood-CSF barrier. *Toxicol Appl*  
525 *Pharmacol* 155 (1999) 24-31.
- 526 [8] W. Zheng, H. Shen, W.S. Blaner, Q. Zhao, X. Ren, and J.H. Graziano, Chronic lead exposure alters  
527 transthyretin concentration in rat cerebrospinal fluid: the role of the choroid plexus. *Toxicol*  
528 *Appl Pharmacol* 139 (1996) 445-50.
- 529 [9] X. Wang, A.K. Mandal, H. Saito, J.F. Pulliam, E.Y. Lee, Z.J. Ke, J. Lu, S. Ding, L. Li, B.J. Shelton, T. Tucker,  
530 B.M. Evers, Z. Zhang, and X. Shi, Arsenic and chromium in drinking water promote tumorigenesis  
531 in a mouse colitis-associated colorectal cancer model and the potential mechanism is ROS-  
532 mediated Wnt/beta-catenin signaling pathway. *Toxicol Appl Pharmacol* 262 (2012) 11-21.
- 533 [10] S.T. Orenstein, S.W. Thurston, D.C. Bellinger, J.D. Schwartz, C.J. Amarasiriwardena, L.M. Altshul, and  
534 S.A. Korrick, Prenatal organochlorine and methylmercury exposure and memory and learning in  
535 school-age children in communities near the New Bedford Harbor Superfund site,  
536 Massachusetts. *Environ Health Perspect* 122 (2014) 1253-9.
- 537 [11] L. Pu, M. Naderi, T. Liu, H.C. Wu, S. Mukhopadhyay, and M. Brylinski, eToxPred: a machine learning-  
538 based approach to estimate the toxicity of drug candidates. *BMC Pharmacol Toxicol* 20 (2019) 2.
- 539 [12] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, DeepTox: Toxicity Prediction using Deep  
540 Learning. *Frontiers in Environmental Science* 3 (2016).
- 541 [13] Y. Hao, J.D. Romano, and J.H. Moore, Knowledge-guided deep learning models of drug toxicity  
542 improve interpretation. *Patterns (N Y)* 3 (2022) 100565.
- 543 [14] F. Kuusisto, V.S. Costa, Z. Hou, J. Thomson, D. Page, and R. Stewart, Machine learning to predict  
544 developmental neurotoxicity with high-throughput data from 2D bio-engineered tissues. *Proc*  
545 *Int Conf Mach Learn Appl* 2019 (2019) 293-298.
- 546 [15] M.P. Schwartz, Z. Hou, N.E. Propson, J. Zhang, C.J. Engstrom, V. Santos Costa, P. Jiang, B.K. Nguyen,  
547 J.M. Bolin, W. Daly, Y. Wang, R. Stewart, C.D. Page, W.L. Murphy, and J.A. Thomson, Human  
548 pluripotent stem cell-derived neural constructs for predicting neural toxicity. *Proc Natl Acad Sci*  
549 *U S A* 112 (2015) 12516-21.
- 550 [16] T. Wei, and V. Simko, R package "corrplot": Visualization of a Correlation Matrix (Version 0.84).  
551 (2017).
- 552 [17] H. Wickham, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (2016).

- 553 [18] S. Davis, and P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and  
554 BioConductor. *Bioinformatics* 23 (2007) 1846-7.
- 555 [19] S. Mukherjee, Immune gene network of neurological diseases: Multiple sclerosis (MS), Alzheimer's  
556 disease (AD), Parkinson's disease (PD) and Huntington's disease (HD). *Heliyon* 7 (2021) e08518.
- 557 [20] S. Mukherjee, Quiescent stem cell marker genes in glioma gene networks are sufficient to  
558 distinguish between normal and glioblastoma (GBM) samples. *Scientific reports* 10 (2020)  
559 10937.
- 560 [21] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, and G.K. Smyth, limma powers differential  
561 expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43  
562 (2015) e47.
- 563 [22] M.D. Robinson, D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential  
564 expression analysis of digital gene expression data. *Bioinformatics* 26 (2010) 139-40.
- 565 [23] Z. Zhao, F. Meng, W. Wang, Z. Wang, C. Zhang, and T. Jiang, Comprehensive RNA-seq transcriptomic  
566 profiling in the malignant progression of gliomas. *Scientific data* 4 (2017) 170024.
- 567 [24] A. Reiner, D. Yekutieli, and Y. Benjamini, Identifying differentially expressed genes using false  
568 discovery rate controlling procedures. *Bioinformatics* 19 (2003) 368-75.
- 569 [25] L. Shen, and M. Sinai, GeneOverlap: Test and visualize gene overlaps. R package version 1.22.0  
570 (2019).
- 571 [26] H. Chen, and P.C. Boutros, VennDiagram: a package for the generation of highly-customizable Venn  
572 and Euler diagrams in R. *BMC bioinformatics* 12 (2011) 35.
- 573 [27] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, and A. Ma'ayan, Enrichr:  
574 interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* 14  
575 (2013) 128.
- 576 [28] Pedregosa F, G.e. Varoquaux, Gramfort A, Michel V, Thirion B, Grisel O, and e. al., Scikit-learn:  
577 Machine learning in Python. *Journal of Machine Learning Research* 12 (2011) 2825-2830.
- 578 [29] E.G. Knox, M.R. Aburto, G. Clarke, J.F. Cryan, and C.M. O'Driscoll, The blood-brain barrier in aging  
579 and neurodegeneration. *Mol Psychiatry* 27 (2022) 2659-2673.
- 580 [30] W. Zheng, M. Aschner, and J.F. Gherzi-Egea, Brain barrier systems: a new frontier in metal  
581 neurotoxicological research. *Toxicol Appl Pharmacol* 192 (2003) 1-11.
- 582 [31] J.H. Lee, G. Yoo, J. Choi, S.H. Park, H. Shin, R. Prasad, Y. Lee, M.R. Ahn, I.J. Cho, and W. Sun, Cell-line  
583 dependency in cerebral organoid induction: cautionary observations in Alzheimer's disease  
584 patient-derived induced pluripotent stem cells. *Mol Brain* 15 (2022) 46.
- 585 [32] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szczesniak,  
586 D.J. Gaffney, L.L. Elo, X. Zhang, and A. Mortazavi, A survey of best practices for RNA-seq data  
587 analysis. *Genome Biol* 17 (2016) 13.
- 588 [33] Y. Zhao, M.C. Li, M.M. Konate, L. Chen, B. Das, C. Karlovich, P.M. Williams, Y.A. Evrard, J.H.  
589 Doroshov, and L.M. McShane, TPM, FPKM, or Normalized Counts? A Comparative Study of  
590 Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models  
591 Repository. *J Transl Med* 19 (2021) 269.
- 592 [34] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, and J.M. O'Sullivan, A Review of Feature Selection  
593 Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform* 2 (2022) 927312.
- 594 [35] E. Cobb, J. Hall, and D.L. Palazzolo, Induction of Metallothionein Expression After Exposure to  
595 Conventional Cigarette Smoke but Not Electronic Cigarette (ECIG)-Generated Aerosol in  
596 *Caenorhabditis elegans*. *Front Physiol* 9 (2018) 426.
- 597 [36] R.D. Palmiter, Regulation of metallothionein genes by heavy metals appears to be mediated by a  
598 zinc-sensitive inhibitor that interacts with a constitutively active transcription factor, MTF-1.  
599 *Proc Natl Acad Sci U S A* 91 (1994) 1219-23.

- 600 [37] H. Ikebuchi, R. Teshima, K. Suzuki, T. Terao, and Y. Yamane, Simultaneous induction of Pb-  
601 metallothionein-like protein and Zn-thionein in the liver of rats given lead acetate. *Biochem J*  
602 233 (1986) 541-6.
- 603 [38] M.K. Yagle, and R.D. Palmiter, Coordinate regulation of mouse metallothionein I and II genes by  
604 heavy metals and glucocorticoids. *Mol Cell Biol* 5 (1985) 291-4.
- 605 [39] R. Dallinger, B. Berger, P. Hunziker, and J.H. Kagi, Metallothionein in snail Cd and Cu metabolism.  
606 *Nature* 388 (1997) 237-8.
- 607 [40] D.H. Hamer, Metallothionein. *Annu Rev Biochem* 55 (1986) 913-51.
- 608 [41] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, and H. Aerts, Artificial intelligence in radiology.  
609 *Nat Rev Cancer* 18 (2018) 500-510.
- 610 [42] I.D.G.P. van, K. Szuhai, I.H. Briaire-de Bruijn, M. Kostine, M.L. Kuijjer, and J. Bovee, Machine learning  
611 analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and  
612 identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput Biol* 15 (2019) e1006826.

613  
614

### 615 **Figure Legend**

616 **Figure 1.** Effect of quantile normalization on chemical exposure and control RNA-seq  
617 datasets (GSE166297, GSE128431, GSE63935 and GSE126786). **A:** Density plot  
618 representation of gene expression per study before and after quantile normalization. **B:**  
619 Box and whisker plot representation of gene expression per study before and after  
620 quantile normalization. **C:** Pearson correlation plot of studies before and after quantile  
621 normalization. **D:** Principal Component Analysis (PCA) of datasets before and after  
622 quantile normalization. For this figure the following R packages and built-in R functions  
623 were used: `prcomp` from `stats4_3.5.0`, `corrplot_0.84`, `ggplot2_3.1.1` and `limma_3.38.3`.

624  
625 **Figure 2.** Heavy metal biomarkers or DEGs identified with `limma`. **A:** Volcano plot of  
626 differentially expressed genes showing  $\log_{FC}$  (log fold change) and its p-value in -  
627  $\log_{10}(\text{adjusted.p-value})$ . Red dots have significant p-value  $<0.05$  and black boxed  
628 genes have fold-change  $> 5$ . **B:** Box and whisker plot representation of gene expression  
629 per group, here heavy metal and non-heavy metal group. **C:** Density plot representation  
630 of gene expression per group, here heavy metal and non-heavy metal group. **D:** Bar  
631 plot of  $\log_{FC}$  (log fold change) of top 100 genes (here 85 genes) upregulated in heavy  
632 metal relative to non-heavy metal group. For this figure the following R packages and  
633 built-in R functions were used: `ggplot2_3.1.1` and `limma_3.38.3`.

634  
635 **Figure 3.** Heavy metal biomarkers or DEGs identified with `edgeR`. **A:** Volcano plot of  
636 differentially expressed genes showing  $\log_{FC}$  (log fold change) and its p-value in -  
637  $\log_{10}(\text{adjusted.p-value})$ . Red dots have significant p-value  $<0.05$  and black boxed  
638 genes have fold-change  $> 5$ . **B:** Box and whisker plot representation of gene expression  
639 per group, here heavy metal and non-heavy metal group. **C:** Density plot representation  
640 of gene expression per group, here heavy metal and non-heavy metal group. **D:** Bar  
641 plot of  $\log_{FC}$  (log fold change) of top 100 genes upregulated in heavy metal relative to  
642 non-heavy metal group. For this figure the following R packages and built-in R functions  
643 were used: `ggplot2_3.1.1` and `edgeR_3.24.3`.

644

645 **Figure 4.** Heavy metal biomarkers or DEGs identified with simple comparison of means.  
646 **A:** Volcano plot of differentially expressed genes showing logFC (log fold change) and  
647 its p-value in  $-\log_{10}(\text{adjusted.p-value})$ . Red dots have significant p-value  $< 0.05$  and  
648 black boxed genes have fold-change  $> 5$ . **B:** Box and whisker plot representation of  
649 gene expression per group, here heavy metal and non-heavy metal group. **C:** Density  
650 plot representation of gene expression per group, here heavy metal and non-heavy  
651 metal group. **D:** Bar plot of logFC (log fold change) of top 100 genes upregulated in  
652 heavy metal relative to non-heavy metal group. For this figure the following R packages  
653 and built-in R functions were used: ggplot2\_3.1.1.

654  
655 **Figure 5.** Heavy metal biomarker 80 genes from overlap of all 3 DEG methods and 879  
656 biomarker genes from overlap of at least 2 methods, and their GO analysis. **A:** Venn  
657 Diagram for overlapping DEGs significantly upregulated by  $> 5$  FC (fold change). **B:**  
658 Heat-map for overlap of significance for genes significantly upregulated by  $> 5$  FC (fold  
659 change) in limma, edgeR and simple comparison of means. **C:** GO Analysis (Gene  
660 Ontology Analysis) Biological Process of heavy metal biomarker 80 genes from overlap  
661 of all 3 DEG methods. **D:** GO Analysis (Gene Ontology Analysis) Biological Process of  
662 heavy metal biomarker 879 genes from overlap of at least 2 methods. For this figure the  
663 following R packages and built-in R functions were used: GeneOverlap\_1.18.0,  
664 VennDiagram\_1.6.20, EnrichR\_1.0 and ggplot2\_3.1.1.

665  
666 **Figure 6.** Evaluation of machine learning (ML) models built with heavy metal biomarker  
667 80 genes (strict overlap of DEG methods) as features. **A:** Number of datasets from  
668 heavy metal and non-heavy metal groups used to train and test ML models. Confusion  
669 matrix and ROC curve for **B:** Logistic Regression AUC=0.94 **C:** K-Means AUC=0.38 **D:**  
670 Naïve Bayes AUC=0.58 **E:** Support Vector Machine (SVM) AUC=0.94 **F:** Random  
671 Forest AUC=0.88 **G:** XGB Grid AUC=0.68 **H:** XGB No Grid AUC=0.75 **I:** Summary of  
672 accuracy, precision, recall and f1-score for all models. For this figure the scikit-  
673 learn\_1.2.1 python package was used.

674  
675 **Figure 7.** Evaluation of machine learning (ML) models built with heavy metal biomarker  
676 879 genes (relaxed overlap of DEG methods) as features. **A:** Number of datasets from  
677 heavy metal and non-heavy metal groups used to train and test ML models. Confusion  
678 matrix and ROC curve for **B:** Logistic Regression AUC=0.94 **C:** K-Means AUC=0.56 **D:**  
679 Naïve Bayes AUC=0.57 **E:** Support Vector Machine (SVM) AUC=0.94 **F:** Random  
680 Forest AUC=0.81 **G:** XGB Grid AUC=0.68 **H:** XGB No Grid AUC=0.88 **I:** Summary of  
681 accuracy, precision, recall and f1-score for all models. For this figure the scikit-  
682 learn\_1.2.1 python package was used.

683  
684  
685 **Access to Codes:**

686 <https://icedrive.net/s/h3P65RbNvf5Dh8yT1DabXxyNqWq6> and  
687 [https://gitlab.com/smukher2/pbothers\\_rnaseq\\_ml\\_feb2023](https://gitlab.com/smukher2/pbothers_rnaseq_ml_feb2023).

688 Please cite this paper if you use these codes. Thank you.



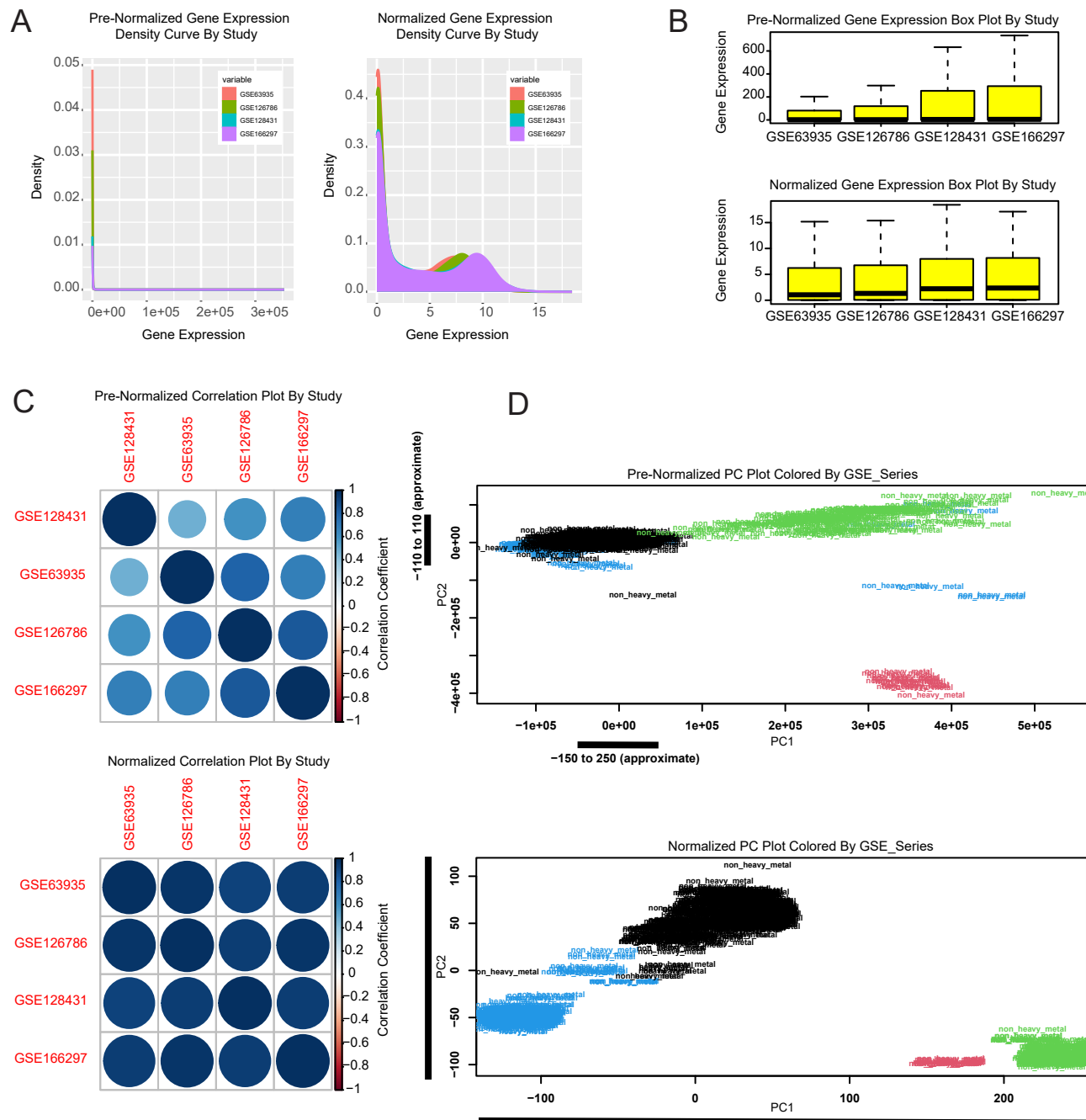


Figure 1 Mukherjee

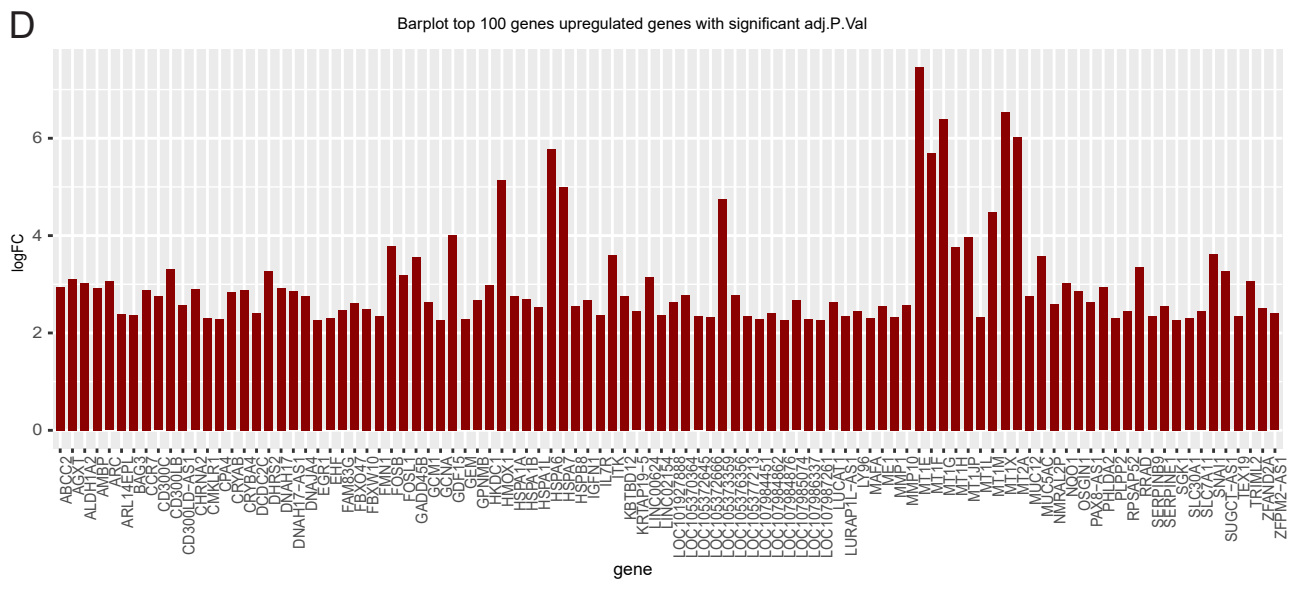
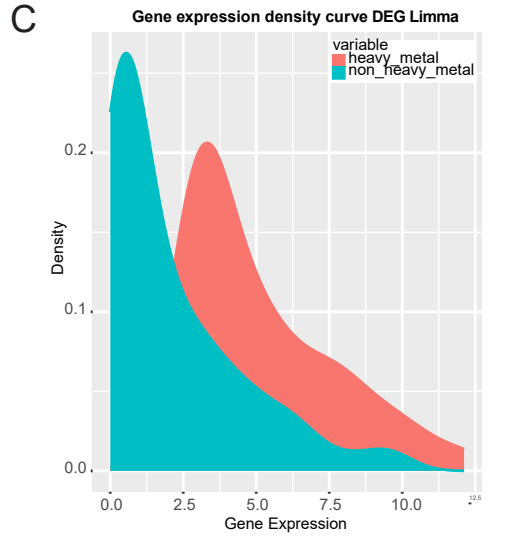
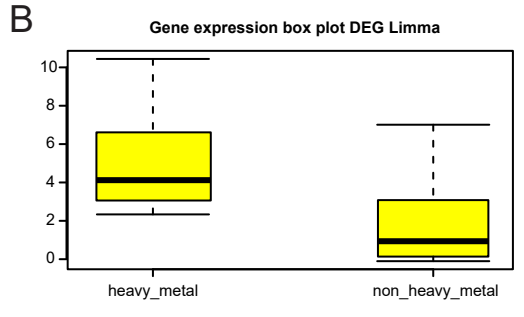
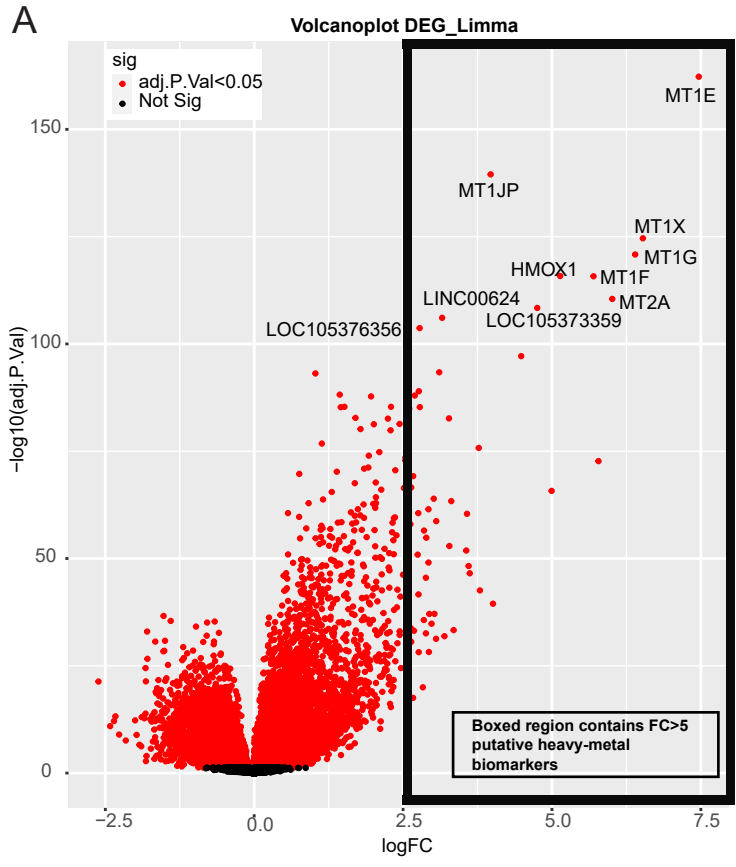


Figure 2 Mukherjee

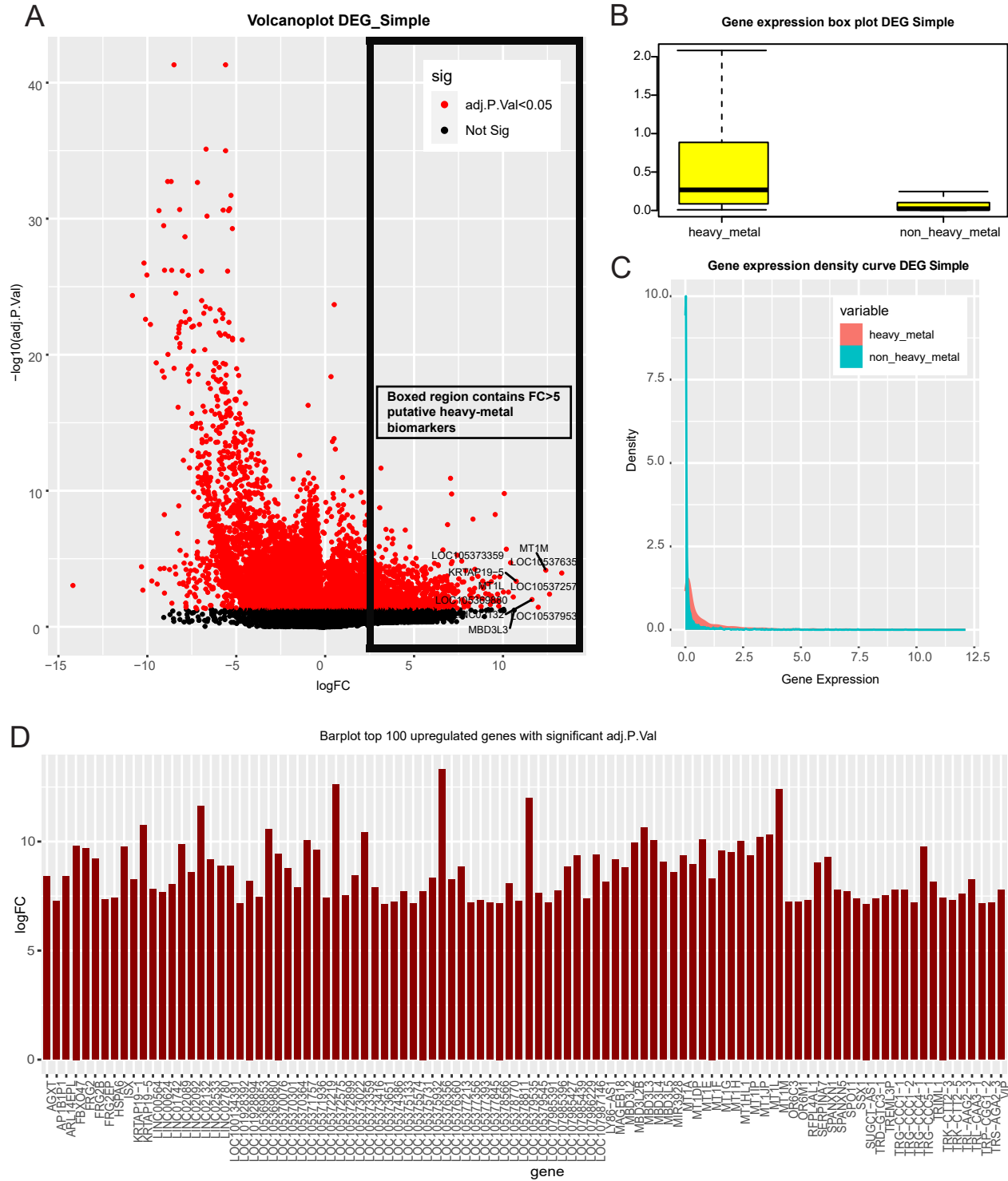


Figure 3 Mukherjee

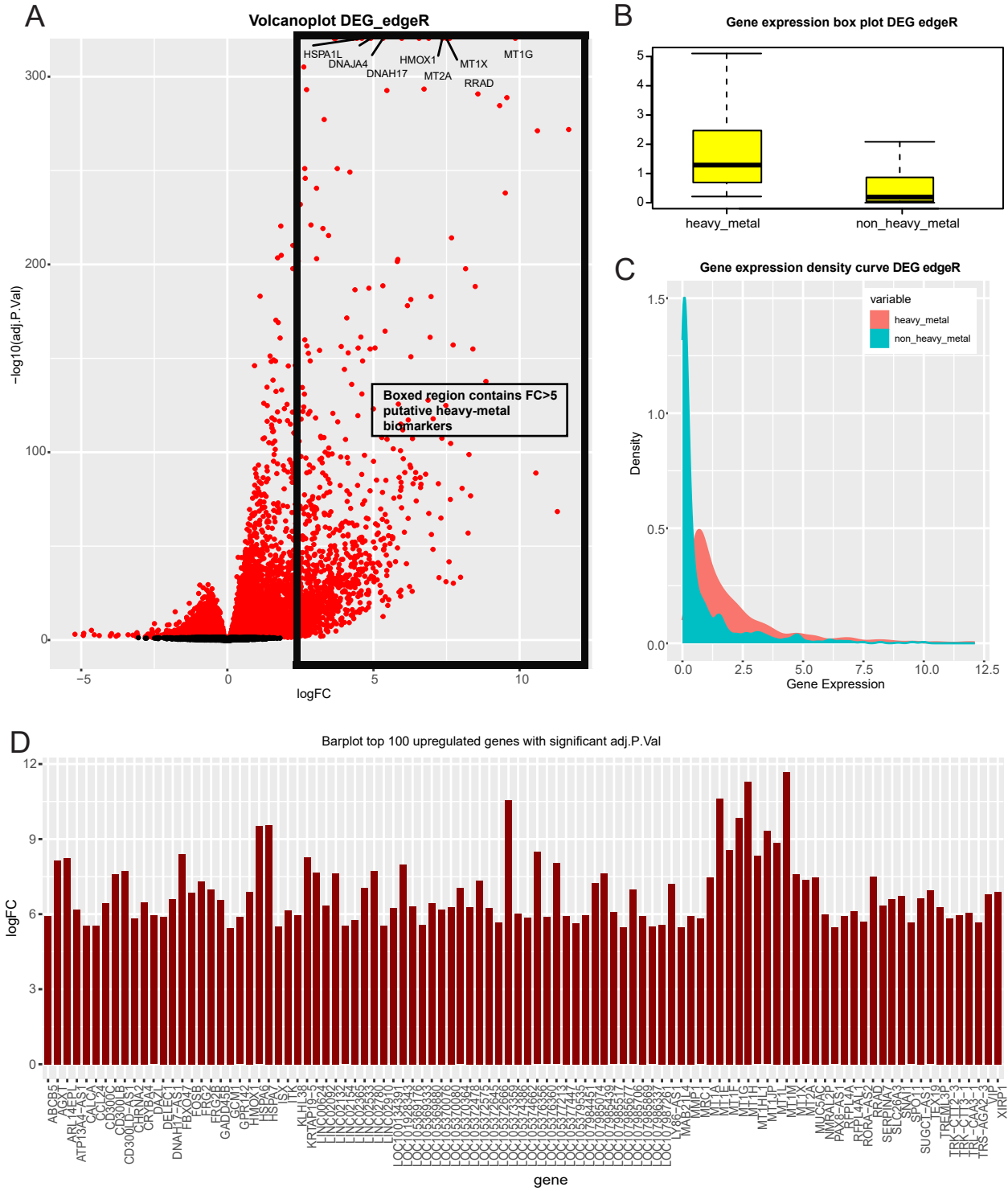
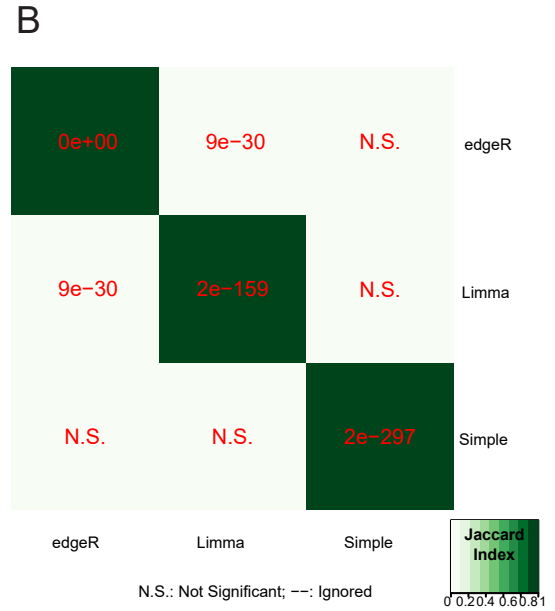
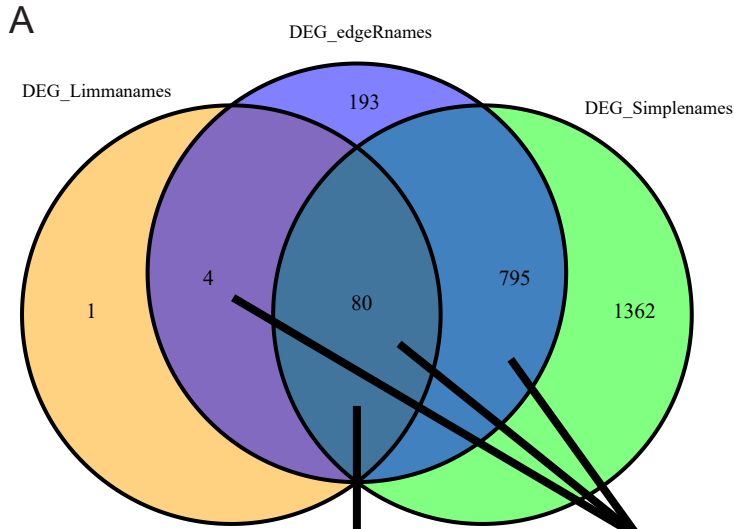


Figure 4 Mukherjee



**C** Overlap of atleast 3 DEG methods (80 heavy-metal biomarker genes):  
GO Biological Process

	Term	Adjusted.P. value	Genes
1	cellular response to zinc ion (GO:0071294)	1.71E-10	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
2	cellular response to copper ion (GO:0071280)	1.71E-10	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
3	response to copper ion (GO:0046688)	5.49E-10	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
4	cellular response to cadmium ion (GO:0071276)	1.47E-09	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
5	cellular zinc ion homeostasis (GO:0006882)	1.47E-09	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
6	response to cadmium ion (GO:0046686)	1.47E-09	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
7	response to zinc ion (GO:0010043)	1.47E-09	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
8	zinc ion homeostasis (GO:0055069)	1.67E-09	MT2A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E
9	negative regulation of growth (GO:0045926)	4.88E-09	MT2A;MT1M;MT1F;OSGIN1;MT1G;MT1X;MT1H;HSPA1B;MT1E;HSPA1A
10	cellular response to unfolded protein (GO:0034620)	5.29E-08	HSPA1L;BAG3;HSPB8;HSPA6;HSPA1B;HSPA1A

**D** Overlap of atleast 2 DEG methods (879 heavy-metal biomarker genes):  
GO Biological Process

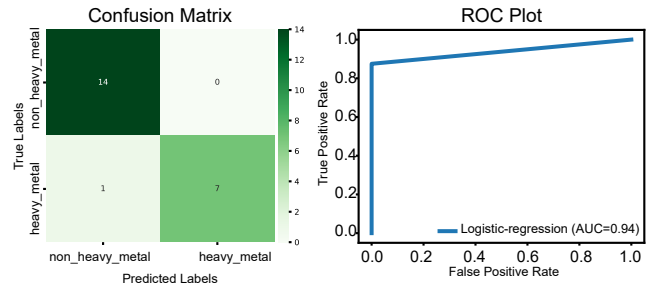
	Term	Adjusted.P. value	Genes
1	response to copper ion (GO:0046688)	1.87E-05	IL1A;MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E;MT1HL1
2	cellular response to zinc ion (GO:0071294)	1.87E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E;MT1HL1
3	cellular response to copper ion (GO:0071280)	2.28E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1E;MT1HL1
4	cellular response to cadmium ion (GO:0071276)	5.11E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;FOS;MT1HL1;MT1E
5	cellular zinc ion homeostasis (GO:0006882)	6.10E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;SLC30A1;MT1H;MT1HL1;MT1E
6	response to cadmium ion (GO:0046686)	7.46E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;FOS;MT1HL1;MT1E
7	zinc ion homeostasis (GO:0055069)	9.21E-05	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;SLC30A1;MT1H;MT1HL1;MT1E
8	response to zinc ion (GO:0010043)	0.000622445	MT2A;MT1A;MT1M;MT1F;MT1G;MT1X;MT1H;MT1HL1;MT1E
9	negative regulation of inclusion body assembly (GO:0090084)	0.130986911	DNAJB1;DNAJA4;HSPA1B;HSPA1A
10	cellular response to unfolded protein (GO:0034620)	0.162569664	HSPA1L;BAG3;HSPB8;HSPA6;HSPA1B;HSPA1A

Figure 5 Mukherjee

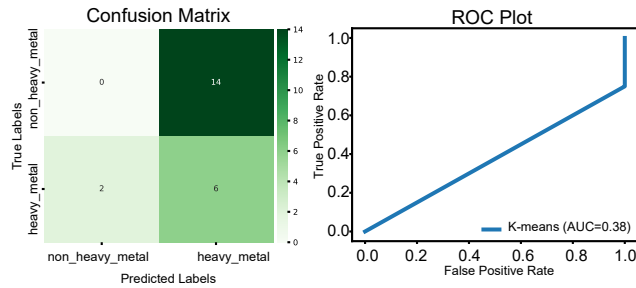
### A Number of test and train datasets

Train Datasets	
26	non_heavy_metal
25	heavy_metal
Test Datasets	
14	non_heavy_metal
8	heavy_metal

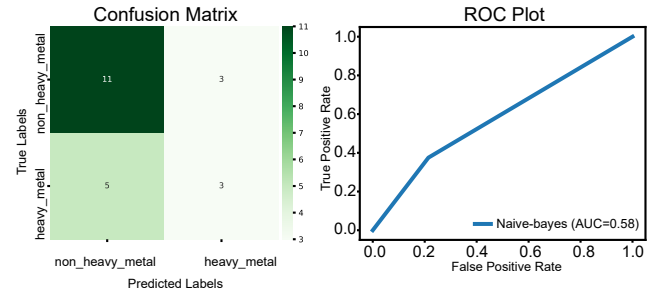
### B Logistic Regression 80 genes



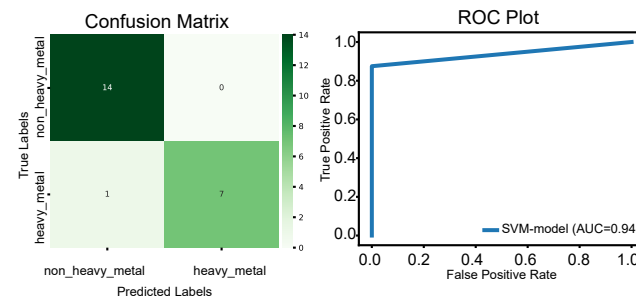
### C K-Means 80 genes



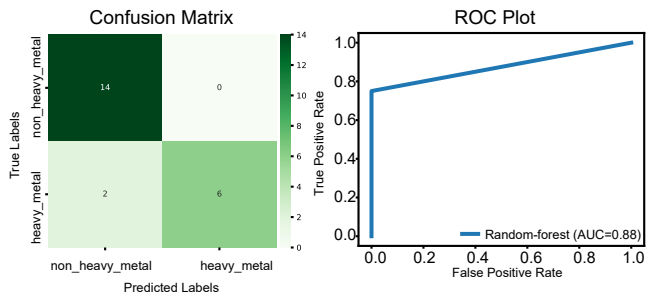
### D Naive Bayes 80 genes



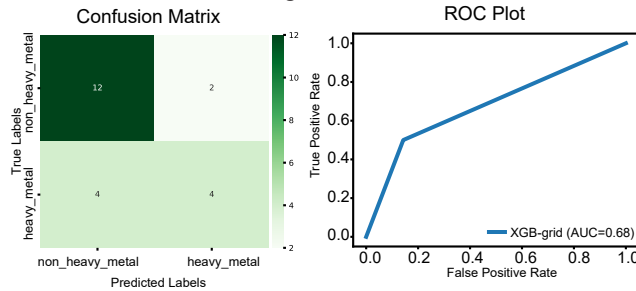
### E SVM Model 80 genes



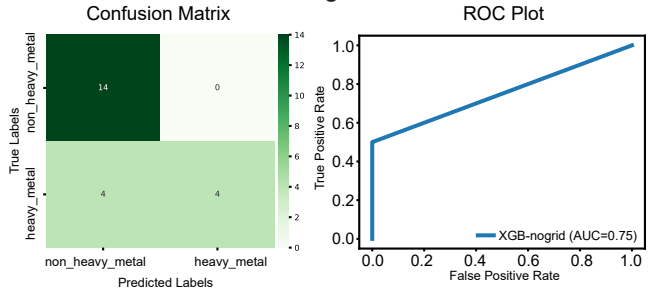
### F Random Forest 80 genes



### G XGB Grid Model 80 genes



### H XGB No Grid Model 80 genes



### I Summary of Models

Logistic Regression	precision	recall	f1-score
non heavy metal	0.9333333	1	0.9655172
heavy metal	1	0.875	0.9333333
accuracy	0.9545455		

K-Means	precision	recall	f1-score
non heavy metal	0	0	0
heavy metal	0.3	0.75	0.4285714
accuracy	0.2727273		

Naïve-Bayes	precision	recall	f1-score
non heavy metal	0.6875	0.7857143	0.7333333
heavy metal	0.5	0.375	0.4285714
accuracy	0.6363636		

SVM Model	precision	recall	f1-score
non heavy metal	0.9333333	1	0.9655172
heavy metal	1	0.875	0.9333333
accuracy	0.9545455		

Random-Forest	precision	recall	f1-score
non heavy metal	0.875	1	0.9333333
heavy metal	1	0.75	0.8571429
accuracy	0.9090909		

XGB-Grid	precision	recall	f1-score
non heavy metal	0.75	0.8571429	0.8
heavy metal	0.6666667	0.5	0.5714286
accuracy	0.7272727		

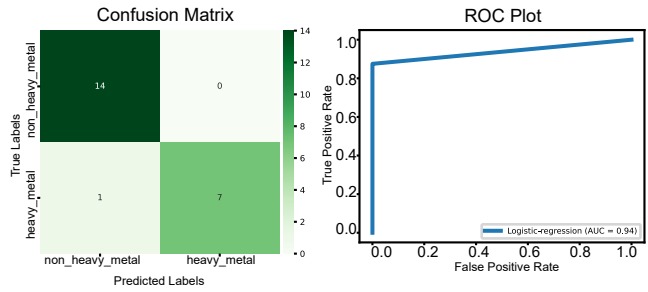
XGB-NoGrid	precision	recall	f1-score
non heavy metal	0.7777778	1	0.875
heavy metal	1	0.5	0.6666667
accuracy	0.8181818		

Figure 6 Mukherjee

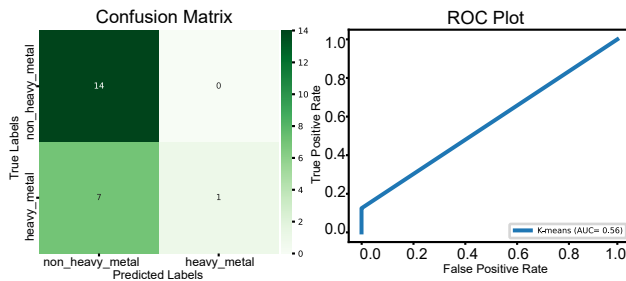
### A Number of test and train datasets

Train Datasets	
26	non_heavy_metal
25	heavy_metal
Test Datasets	
14	non_heavy_metal
8	heavy_metal

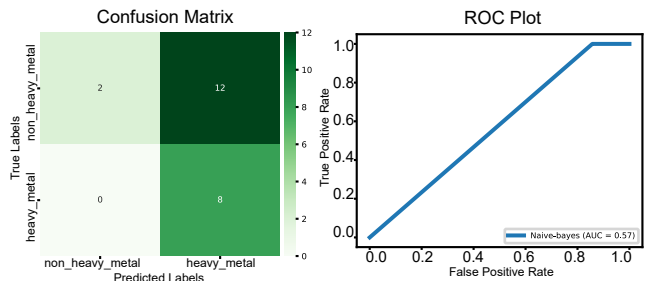
### B Logistic Regression 879 genes



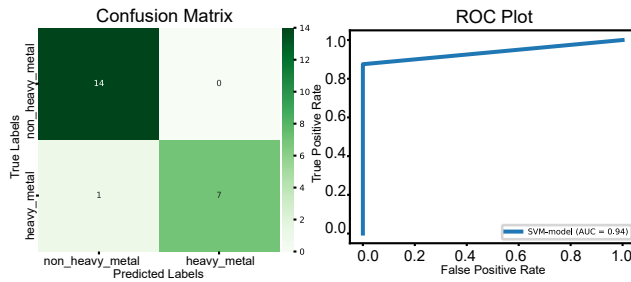
### C K-Means 879 genes



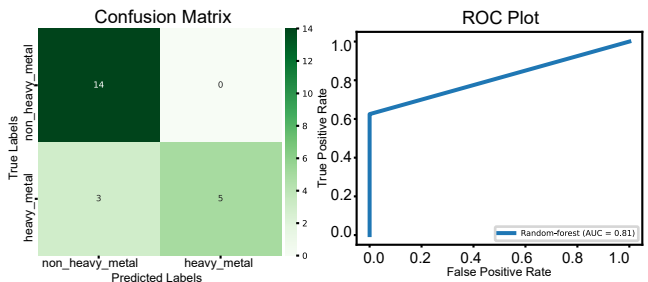
### D Naive Bayes 879 genes



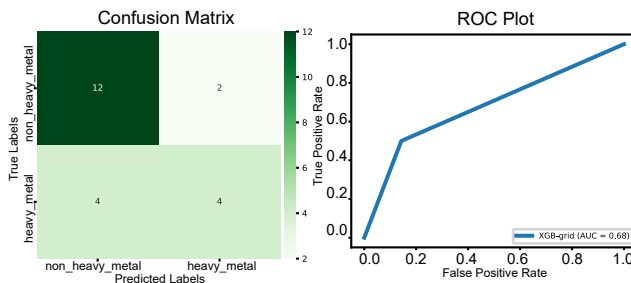
### E SVM Model 879 genes



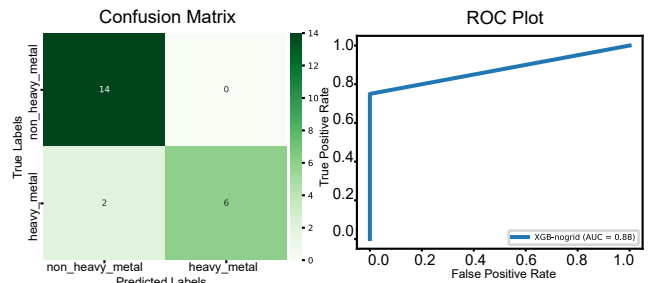
### F Random Forest 879 genes



### G XGB Grid Model 879 genes



### H XGB No Grid Model 879 genes



### I Summary of Models

Logistic Regression	precision	recall	f1-score	K-means	precision	recall	f1-score	Naive-bayes	precision	recall	f1-score
0	0.93333333	1	0.96551724	0	0.66666667	1	0.8	0	1	0.142857143	0.25
1	1	0.875	0.93333333	1	1	0.125	0.22222222	1	0.4	1	0.571428571
accuracy	0.95454545			accuracy	0.68181818			accuracy	0.45454545		
SVM-model	precision	recall	f1-score	Random-forest	precision	recall	f1-score				
0	0.93333333	1	0.96551724	0	0.82352941	1	0.90322581				
1	1	0.875	0.93333333	1	1	0.625	0.76923077				
accuracy	0.95454545			accuracy	0.86363636						
XGB-grid	precision	recall	f1-score	XGB-NoGrid	precision	recall	f1-score				
0	0.75	0.85714	0.8	0	0.875	1	0.93333333				
1	0.66666667	0.5	0.57142857	1	1	0.75	0.85714286				
accuracy	0.72727273			accuracy	0.90909091						

Figure 7 Mukherjee