

1 **Deep Learning-Based Multimodal Clustering Model for Endotyping and Post-Arthroplasty**  
2 **Response Classification using Knee Osteoarthritis Subject-Matched Multi-Omic Data**

3  
4 Jason S. Rockel<sup>1,2\*</sup>, Divya Sharma<sup>1,3,4\*</sup>, Osvaldo Espin-Garcia<sup>1,2,3,5\*</sup>, Katrina Hueniken<sup>1,2,4\*</sup>, Amit  
5 Sandhu<sup>1,2\*</sup>, Chiara Pastrello<sup>1,2\*</sup>, Kala Sundararajan<sup>1,2</sup>, Pratibha Potla<sup>1,2</sup>, Noah Fine<sup>1,2</sup>, Starlee S.  
6 Lively<sup>1,2</sup>, Kimberly Perry<sup>1,2</sup>, Nizar N. Mohamed<sup>1,2,6</sup>, Khalid Syed<sup>1,2,6</sup>, Igor Jurisica<sup>1,2,#</sup>, Anthony  
7 V. Perruccio<sup>1,2,6,8,#</sup>, Y. Raja Rampersaud<sup>1,2,6,#</sup>, Rajiv Gandhi<sup>1,2,6,7,#</sup> and Mohit Kapoor<sup>1,2,6,7,#,+</sup>

8  
9 1. Division of Orthopaedics, Osteoarthritis Research Program, Schroeder Arthritis Institute,  
10 University Health Network, Toronto, Ontario, Canada

11 2. Krembil Research Institute, University Health Network, Toronto, Ontario, Canada

12 3. Department of Biostatistics, Dalla Lana School of Public Health, University of Toronto,  
13 Toronto, Ontario, Canada

14 4. Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

15 5. Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada

16 6. Department of Surgery, University of Toronto, Toronto, Ontario, Canada

17 7. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto,  
18 Ontario, Canada

19 8. Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health,  
20 University of Toronto, Toronto, Ontario, Canada

21  
22 \* equal first author contribution

23 # equal senior author contribution

24  
25 + Correspondence: [mkapoor@uhnresearch.ca](mailto:mkapoor@uhnresearch.ca); Schroeder Arthritis Institute, 60 Leonard Avenue,  
26 5<sup>th</sup> floor Krembil Discovery Tower, Toronto, Ontario, Canada.

27 **Abstract**

28

29 **Background:** Primary knee osteoarthritis (KOA) is a heterogeneous disease with clinical and  
30 molecular contributors. Biofluids contain microRNAs and metabolites that can be measured by  
31 omic technologies. Deep learning captures complex non-linear associations within multimodal  
32 data but, to date, has not been used for multi-omic-based endotyping of KOA patients. We  
33 developed a novel multimodal deep learning framework for clustering of multi-omic data from  
34 three subject-matched biofluids to identify distinct KOA endotypes and classify one-year post-  
35 total knee arthroplasty (TKA) pain/function responses.

36

37 **Materials and Methods:** In 414 KOA patients, subject-matched plasma, synovial fluid and urine  
38 were analyzed by microRNA sequencing or metabolomics. Integrating 4 high-dimensional  
39 datasets comprising metabolites from plasma (n=151 features), along with microRNAs from  
40 plasma (n=421), synovial fluid (n=930), or urine (n=1225), a multimodal deep learning variational  
41 autoencoder architecture with K-means clustering was employed. Features influencing cluster  
42 assignment were identified and pathway analyses conducted. An integrative machine learning  
43 framework combining 4 molecular domains and a clinical domain was then used to classify  
44 WOMAC pain/function responses post-TKA within each cluster.

45

46 **Findings:** Multimodal deep learning-based clustering of subjects across 4 domains yielded 3  
47 distinct patient clusters. Feature signatures comprising microRNAs and metabolites across  
48 biofluids included 30, 16, and 24 features associated with Clusters 1-3, respectively. Pathway  
49 analyses revealed distinct pathways associated with each cluster. Integration of 4 multi-omic  
50 domains along with clinical data improved response classification performance, with Cluster 3  
51 achieving AUC=0.879 for subject pain response classification and Cluster 2 reaching AUC=0.808  
52 for subject function response, surpassing individual domain classifications by 12% and 15%  
53 respectively.

54

55 **Interpretation:** We have developed a deep learning-based multimodal clustering model capable  
56 of integrating complex multi-fluid, multi-omic data to assist in KOA patient endotyping and test  
57 outcome response to TKA surgery.

58 **Funding:** Canada Research Chairs Program, Tony and Shari Fell Chair, Campaign to Cure  
59 Arthritis, University Health Network Foundation.

60

61 **Keywords:** Osteoarthritis; Variational Autoencoder; Machine-Learning; Endotyping; Multi-  
62 omics; Response Classification

63

64

## 65 **Introduction**

66  
67 Osteoarthritis (OA) is a degenerative, painful and disabling joint disease affecting over 500 million  
68 people worldwide,<sup>1</sup> with the knee most commonly afflicted<sup>2</sup>. Primary knee (K)OA patients are  
69 heterogeneous<sup>3</sup>. Risk factors of KOA include age, sex, and obesity status.<sup>2</sup> Mental health and  
70 persistent pain status have also associated with KOA clinical phenotypes.<sup>4,5</sup> Total joint  
71 arthroplasty (TKA) is the only available therapy for KOA patients who no longer respond to  
72 conservative management; however, up to 34% of KOA patients with TKA fail to achieve  
73 clinically-relevant pain reduction.<sup>6</sup> Identifying those at high risk of non-response is of significant  
74 interest. It is possible that KOA heterogeneity captured by biological features may improve our  
75 ability to classify patient responses to TKA.

76 Biofluid microRNAs (miRNAs) and metabolites can provide highly descriptive, individualized  
77 categorizations of patients beyond clinical measures. MiRNAs epigenetically modify target RNA  
78 expression. Biofluid metabolomes represents snapshots of the metabolic activity contributed by  
79 associated cells and tissues. Advanced omic technologies can measure miRNAs (miRNomics) and  
80 metabolites (metabolomics), primarily by next generation sequencing (NGS) and liquid  
81 chromatography-mass spectrometry/mass spectrometry (LC-MS/MS), respectively.<sup>7,8</sup>

82 In a case-control study using the UK Biobank cohort, 14 distinct OA risk phenotypes were  
83 identified by multi-modal machine learning (ML) using clinical factors alone.<sup>9</sup> The inclusion of  
84 individual proteomics, genomics or metabolomics data showed no prediction improvement of  
85 case-control status over clinical factor modeling alone.<sup>9</sup> In contrast, studies using individual  
86 biofluids have identified endotypes of OA patients. Three endotypes of OA patients (low tissue  
87 turnover, structural damage and systemic inflammation) were identified from a panel of 16 serum  
88 and urine proteins/peptides using unsupervised ML.<sup>10</sup> Plasma metabolomics alone uncovered  
89 multiple endotypes of KOA patients,<sup>11,12</sup> with some endotype-related metabolite ratios able to  
90 differentiate specific endotypes of KOA subjects from control participants.<sup>11</sup> KOA patient biofluid  
91 miRNA signatures were also able to differentiate between slow and fast progressors,<sup>13</sup> early and  
92 late KOA,<sup>14,15</sup> and patients requiring TKA or not.<sup>16</sup> Thus, endotype data from biofluids is  
93 important for understanding KOA heterogeneity, and, consequently, may be associated with  
94 patient outcomes. To our knowledge, KOA endotypes have yet to be evaluated across multiple  
95 biofluids using multi-omic technologies in an integrated approach.<sup>8</sup>

96 Our experience to date suggests that multi-biofluid, multi-omic endotyping requires more complex  
97 modeling systems. Deep learning aids in extracting complex patterns from data. However,  
98 integrating multimodal data from multiple sources presents challenges due to the diversity within  
99 and across data types. Variational Autoencoders (VAEs) address this challenge by embedding  
100 diverse data domains into reduced latent dimensions, facilitating improved data clustering.<sup>17-19</sup>  
101 Despite the potential of VAEs, there is a lack of unified frameworks for leveraging these methods  
102 to identify clusters from multimodal data and to classify clinical responses by integrating diverse  
103 data domains. Additionally, VAEs have not been applied to investigate OA endotypes.  
104 In this study, we developed a novel multimodal deep learning framework employing VAEs for  
105 integrative clustering using 4 high-dimensional domains of subject-matched multi-omic data from  
106 synovial fluid, urine and plasma and tested its ability to determine distinct clusters (endotypes) of  
107 a sample of KOA patients. Leveraging these endotypes, we further developed an integrative ML  
108 framework and tested the potential of this methodology to assess pain and function responses to  
109 TKA surgery.

110

## 111 **Methods**

112

### 113 ***Study Sample***

114 A sample of 414 patients with primary KOA who underwent TKA within the Longitudinal  
115 Evaluation in the Arthritis Program-OA cohort (University Health Network, Toronto, ON,  
116 Canada), as previously described,<sup>20</sup> and who had synovial fluid (collected intra-operatively),  
117 plasma and urine (collected up to 3 months prior to surgery) available, were selected for analysis.  
118 Subjects completed self-reported, multidimensional questionnaires from which baseline Western  
119 Ontario and McMaster Universities Arthritis Index (WOMAC) pain and function,<sup>21</sup> Hospital  
120 Anxiety and Depression Scale (HADS)<sup>22</sup> and painDETECT<sup>23</sup> at baseline (completed within the 3-  
121 months preceding surgery), and WOMAC pain and function 1-year post-TKA, were calculated.  
122 Improvement in WOMAC pain and function from baseline to 1 year post-TKA was calculated and  
123 individuals were categorized as responders (>33% improvement) or non-responders (≤33%  
124 improvement). HADS depression and anxiety scores were each categorized as normal (score 0-7),  
125 borderline case (score 8-11), and definite case (score 11-21)<sup>22</sup>. PainDETECT neuropathic-like pain  
126 scores were used to classify patients pain as likely nociceptive (score 0-12), unclear (score 13-18),

127 or likely neuropathic (score 19-38).<sup>23</sup> Baseline age, sex, and weight and height [from which body  
128 mass index (BMI; kg/m<sup>2</sup>) was calculated] were also collected. Biofluids and patient data were  
129 collected with written informed consent from each patient and Research Ethics Board approval of  
130 the University Health Network, Toronto, ON (REB# 07-0383-BE; 14-7592-AE).

131  
132 ***MiRNA sequencing (miRNomics) and metabolomics***  
133 MiRNA extraction was performed using 200  $\mu$ L plasma (N=414), 100  $\mu$ L synovial fluid (N=414),  
134 and 1 ml urine (N=414). cDNA libraries were prepared using a protocol we previously reported.<sup>15</sup>  
135 NGS was conducted at the Schroeder Arthritis Institute (Toronto) sequencing facility using the  
136 Illumina NextSeq550 platform. Alignment, processing and quality assessment was performed  
137 using a previously reported pipeline.<sup>24</sup> Targeted metabolomics was used to profile 188 metabolites  
138 (Biocrates AbsoluteIDQ p180 kit, Biocrates Life Sciences AG, Austria) in N=414 plasma samples  
139 at The Metabolomics Innovation Centre (Calgary, Alberta) by LC-MS/MS, as previously  
140 described.<sup>12</sup> Metabolite quantification and batch correction was conducted using the Absolute  
141 IDQ-coupled MetIDQ software (Biocrates). MiRNA count data and metabolite concentrations  
142 were normalized using sum normalization, log-transformation, and Pareto scaling.<sup>25</sup> To stabilize  
143 variance estimates within differential expression analysis, empirical Bayes moderation techniques  
144 were applied.

145  
146 ***OmicVAE: integrative variational autoencoder architecture for multimodal clustering***  
147 We generated a novel variational autoencoder (VAE) architecture named ‘*omicVAE*’ designed to  
148 cluster multimodal multi-omic data (Figure 1). *OmicVAE* consists of a single encoder network  
149 followed by 4 individual decoder networks, to perform integrative clustering combining 4  
150 modalities: metabolomics, miRNA plasma, miRNA synovial fluid, and miRNA urine.  
151 The encoder network inputs concatenated multimodal multi-omic data and maps it to a shared  
152 latent space representation using multiple fully connected neural network (FNN) layers with  
153 sigmoid activation. The encoder network's output layers parameterize the mean and variance of a  
154 Gaussian distribution representing the shared latent space. Each decoder network reconstructs its  
155 respective domain's input data from samples drawn from this latent space, using multiple fully  
156 connected neural network (FNN) layers. During training, variational inference optimizes the  
157 VAE's parameters. The objective function  $L_{total}$  includes the reconstruction loss ( $L_{rec}$ ) and the

158 Kullback-Leibler (KL) divergence ( $L_{KL}$ ) between the learned latent distribution and a predefined  
 159 prior (eqs. 1-3). Minimizing KL divergence regularizes the latent space, preventing overfitting and  
 160 ensuring it remains structured and interpretable. The reconstruction loss measures the discrepancy  
 161 between the input data and its reconstruction by the VAE decoder for each modality.

$$\begin{aligned}
 162 \quad L_{rec} = & \frac{1}{N_{metabolites}} \sum_{i=1}^N \|X_{metabolites(i)} - Decoder_{metabolites(i)}(z)\|^2 + \\
 163 & \frac{1}{N_{miRNA_{plasma}}} \sum_{i=1}^N \|X_{miRNA_{plasma}(i)} - Decoder_{miRNA_{plasma}(i)}(z)\|^2 + \\
 164 & \frac{1}{N_{miRNA_{synovial}}} \sum_{i=1}^N \|X_{miRNA_{synovial}(i)} - Decoder_{miRNA_{synovial}(i)}(z)\|^2 + \\
 165 & \frac{1}{N_{miRNA_{urine}}} \sum_{i=1}^N \|X_{miRNA_{urine}(i)} - Decoder_{miRNA_{urine}(i)}(z)\|^2 \quad (1)
 \end{aligned}$$

$$\begin{aligned}
 166 \\
 167 \quad L_{KL} = & -\frac{1}{2} \left( (1 + \log(\sigma_{metabolites}^2) - \mu_{metabolites}^2 - \sigma_{metabolites}^2) + \right. \\
 168 & \quad \left. (1 + \log(\sigma_{miRNA_{plasma}}^2) - \mu_{miRNA_{plasma}}^2 - \sigma_{miRNA_{plasma}}^2) + \right. \\
 169 & \quad \left. (1 + \log(\sigma_{miRNA_{synovial}}^2) - \mu_{miRNA_{synovial}}^2 - \sigma_{miRNA_{synovial}}^2) + \right. \\
 170 & \quad \left. (1 + \log(\sigma_{miRNA_{urine}}^2) - \mu_{miRNA_{urine}}^2 - \sigma_{miRNA_{urine}}^2) \right) \quad (2)
 \end{aligned}$$

$$171 \\
 172 \quad L_{total} = L_{rec} + L_{KL} \quad (3)$$

173 where  $i$  represents the samples in each modality,  $X$  is the input data,  $Decoder(z)$  is the  
 174 reconstructed data, and  $\mu$  and  $\sigma$  denote the mean and variance of the Gaussian distribution in the  
 175 latent space. Once omicVAE is trained, K-means clustering on the learned latent space is used to  
 176 identify distinct subpopulations within the multimodal multi-omic data.

### 177 178 **Multi-omic signature identification within each cluster**

179 We employed a comprehensive approach to identify signature features (miRNAs and/or  
 180 metabolites within three biofluids) influencing cluster assignment within each domain. We  
 181 concurrently conducted standardized mean differences (SMD) analysis and differential expression  
 182 (DE) analysis for pairwise cluster comparisons (one cluster vs others), using Benjamini-Hochberg  
 183 (BH) adjusted p-values ( $q < 0.05$ ) to identify significant features. By integrating these analyses,

184 we identified features with both large SMDs and significant DE, capturing robust signature  
185 features distinguishing clusters.

186

### 187 ***Endotype pathway analysis***

188 MiRNAs gene targets per cluster were identified using the top 1% of targets per miRNA using  
189 mirDIP v. 5.2 (<https://ophid.utoronto.ca/mirDIP>)<sup>26</sup> We performed pathway enrichment analysis  
190 for sets of gene targets in each cluster using pathDIP 5 (<https://ophid.utoronto.ca/pathDIP>).<sup>27</sup>

191 Diseases, drugs and vitamins, and genetic information processing pathway types were excluded  
192 from enrichment analysis. Only pathways with q-value (BH adjusted) <0.01 were considered.

193 Metabolite pathway enrichment analysis was not possible, so we identified pathways that included  
194 metabolites specific for each cluster for further analyses. Selected pathways specific for each  
195 cluster were visualized using NAViGaTOR 3.0.19

196 (<https://navigator.ophid.utoronto.ca/navigatorwp>).<sup>28</sup> Mapping of pathways to consolidated

197 categories in pathDIP was used to calculate the number of pathways per category. ggradar2\_1.1.0

198 in R 4.3.0 was subsequently used to plot their distribution per cluster, scaling category pathway  
199 counts from 0% to 100%.

200

### 201 ***Integrative machine learning framework for classifying response***

202 We developed a comprehensive two-step ML framework (Figure 4a) to integrate plasma  
203 metabolites, miRNA plasma, miRNA synovial fluid, and miRNA urine domain, along with clinical

204 domain (consisting of age, sex, BMI, depression and anxiety categories, and neuropathic pain  
205 category), to classify 1-year pain and function responses (i.e. responders vs non-responders). In

206 the first step, we trained separate unimodal ML models for each domain to extract features  
207 classifying 1-year response. We utilized the MICE library in R for imputing missing clinical data

208 (missingness <8%). We explored various ML algorithms, including logistic regression, lasso

209 regression, ridge regression, support vector machines and random forests, selecting models based  
210 on 10-fold cross-validation performance.

211

212 In the second step, we integrated features from all domains using a naïve-Bayes meta-classifier,  
213 trained with classifiers from the unimodal models. Cross-validation was used for performance

214 estimation and hyperparameter tuning. The final classification was generated by the meta-



215 classifier based on integrated features. Model evaluation involved assessing the overall framework  
216 performance using area under the receiver operating characteristic curve (AUC) and analyzing  
217 feature importance with mean Gini impurity metrics.

218

### 219 ***Role of the Funding Sources***

220 Funders had no role in study design, data collection and analysis, decision to publish, or  
221 manuscript writing.

222

### 223 **Findings**

224

#### 225 ***Endotype and signature identification using omicVAE and K-means clustering***

226 We first sought to identify endotypes from our sample of 414 KOA patients. The patient sample  
227 was 57% female, with a mean age ( $\pm$ sd) of  $65.7 \pm 8.7$  years, and BMI ( $\pm$ sd) of  $31 \pm 7.1$  kg/m<sup>2</sup>. The  
228 majority of subjects had anxiety or depression symptom scores in the normal range and the  
229 majority had painDETECT scores indicating likely nociceptive pain. Mean baseline WOMAC  
230 pain score for the sample was  $10.1 \pm 3.5$  points on a 20 total point scale, and baseline WOMAC  
231 function score was  $34.9 \pm 11.9$  points on a 68 total point scale (Supplementary Table 1).

232 After metabolomics and miRNomics analyses of plasma, synovial fluid and urine, our analytical  
233 dataset consisted of 2727 molecular features from 4 domains: 151 plasma metabolites, 421 plasma  
234 miRNAs, 930 synovial fluid miRNAs, and 1225 urine miRNAs. We then developed *omicVAE*  
235 with K-means clustering (Figure 1), which uncovered three clusters of patients using the 4 domains  
236 (Figure 2a). Distribution of most baseline clinical, demographic and anthropometric measures  
237 were similar across clusters, except cluster 3 that had a higher proportion of subjects with  
238 depression scores in the normal range, and cluster 1 that had a higher proportion of subjects with  
239 likely neuropathic pain (Table 1).

240 Significant features associated with each cluster were identified by the intersection of differential  
241 expression and standardized mean difference analyses. Distinct signatures consisting of 30, 16 and  
242 24 features for clusters 1-3, respectively, were identified (Figure 2b and Supplementary Table 2).  
243 Notably, each signature contained features from all 4 domains. In Cluster 1, the highest mean  
244 value difference was observed for synovial fluid hsa-miR-2053. In Cluster 2, synovial fluid hsa-

245 miR-496 exhibited the highest mean value. In Cluster 3, plasma hsa-miR-570-3p had the highest  
246 mean value. Thus, each cluster represents a group of subjects with a distinct endotype.

247

### 248 *Endotype feature signatures are enriched for unique pathways*

249 We next sought to determine if cluster endotype signatures were associated with unique  
250 physiological pathways. We first identified putative miRNA-gene targets using mirDIP,<sup>26</sup>  
251 identifying 3257, 2211, and 2319 individual genes targeted by the miRNAs in each of the endotype  
252 signatures associated with clusters 1-3, respectively. Using these gene set lists, we performed  
253 enrichment analysis using pathDIP (Supplementary Table 3-5).<sup>27</sup> For metabolites in each endotype  
254 signature, pathway annotations were also identified using pathDIP (Supplementary Table 6-8). All  
255 pathways were also annotated with categories in pathDIP. For each endotype, individual miRNA-  
256 gene targets and metabolites were linked to some common as well as unique pathways. The top  
257 unique enriched and annotated pathways linked to miRNA-targeted genes or metabolites,  
258 respectively, for each endotype are displayed in a network showing individual pathways and  
259 categories (Figure 3a, Supplementary Figures 1-3).

260 We next used pathway categories to evaluate physiologically-relevant mechanisms linked to each  
261 endotype. Each enriched miRNA-derived pathway or annotated metabolite pathway was counted  
262 based on its category annotation in pathDIP, scaled and visualized using radar plots (Figure 3b).  
263 The cluster 1 endotype signature was most linked to pathway categories associated with  
264 development and regeneration, membrane transport, metabolism of various molecules, and the  
265 nervous system. The cluster 2 endotype signature was most linked to aging, and cellular  
266 community categories. Finally, the cluster 3 endotype signature was most linked to transport and  
267 catabolism, signal transduction, sensory system, endocrine system, excretory system, immune  
268 system, catabolism, and lipid metabolism categories. Overall, these analyses suggested that  
269 features associated with each endotype were uniquely associated with distinct physiological  
270 pathways.

271

### 272 *Evaluation of classification performance for WOMAC pain and function responses*

273 To determine the classification performance of our clusters for identifying post-TKA WOMAC  
274 pain and function response status, we used an integrative ML framework using five domains—  
275 plasma metabolites, plasma miRNAs, synovial fluid miRNAs, urine miRNAs, and clinical data

276 (age, sex, BMI, anxiety and depression categories, and neuropathic pain category; Figure 4a).  
277 Subject clusters had similar mean pain and function scores 1 year post-TKA, change in scores  
278 from baseline to 1 year, as well as pain and function response rates (Table 1). We first conducted  
279 differential expression analysis in each cluster, identifying metabolites with a fold-change of 1·1  
280 and miRNAs within each biofluid with a fold change of 1·5 between responders and non-  
281 responders. Subsequently, we employed 10-fold cross-validation to estimate the AUC. Of the ML  
282 approaches compared, random forests consistently outperformed others in each individual domain  
283 (Supplementary Table 9) and was used for the unimodal ML models.

284 After differential analysis-based feature selection, cluster 1-3 retained 250, 87, and 49 features,  
285 respectively for the ML analysis to classify pain response (Supplementary Table 10). Initially,  
286 unimodal models were applied to each domain. Within cluster 1, miRNA plasma demonstrated  
287 the highest unimodal performance with an AUC of 0·735. However, the integrative performance,  
288 combining all 4 domains in the meta-classifier, notably improved AUC to 0·841 (highlighted in  
289 red in the ROC plot). For cluster 2, the clinical domain had the highest unimodal AUC of 0·740,  
290 while the integrative AUC was 0·816. Cluster 3 achieved the highest integrative AUC of 0·879,  
291 with miRNA urine showing the highest unimodal AUC of 0·786 (Figure 4b).

292 Based on differential analysis-based feature selection to classify function response, clusters 1-3  
293 retained 63, 46, and 46 features, respectively, for the ML analysis (Supplementary Table 11).  
294 Across clusters 1-3, the clinical data domain consistently exhibited the best performance among  
295 unimodal domains, with AUCs of 0·702, 0·791, and 0·722, respectively. In terms of integrative  
296 performance, clusters 1-3 achieved AUCs of 0·786, 0·808, 0·738 (Figure 4c), respectively.

297

### 298 ***Identifying key response classification features in our multimodal machine learning framework***

299 To enhance the interpretability of our model we identified the most important features (molecular  
300 and clinical) contributing to response classification in each cluster. The top 20 features  
301 contributing to WOMAC pain or function response classification are shown in Supplementary  
302 Figures 4 & 5, respectively. Each top 20 list consisted of features from all 4 molecular domains,  
303 with a notable absence of clinical features; however, all molecular and clinical features inherently  
304 played a role in response classification (Supplementary Tables 12&13). Interestingly, only 3  
305 miRNAs overlapped in the top 20 important features for WOMAC pain response between clusters  
306 2 & 3, namely synovial fluid hsa-miR-1265 and hsa-mir-642a-3p, plasma hsa-3942-5p. In

307 addition, only the metabolite glutamine overlapped in the top 20 important feature lists of clusters  
308 2 & 3 for WOMAC function response. Thus, using our integrative approach, the vast majority of  
309 the most important features for response classification were unique for each cluster and were  
310 primarily driven by molecular entities. Overall, these findings highlight the diverse molecular  
311 features associated with outcome classification in each cluster, emphasizing the importance of  
312 integrating multiple domains for classification modeling of WOMAC pain and function responses  
313 post-TKA.

314

### 315 **Interpretation**

316

317 Demographic, anthropometric and clinical characteristics of OA patients are heterogeneous,  
318 influencing outcomes to therapy, including TKA.<sup>3,29</sup> Heterogeneity has also been identified  
319 through biofluid data, however, most studies to date have used single biofluids with single  
320 molecular type measures to identify endotypes within OA cohorts.<sup>10,11</sup> We developed a novel  
321 multimodal deep learning algorithm, *omicVAE*, to cluster a sample of 414 KOA subjects who  
322 underwent TKA using preoperative miRNA and metabolite feature sets, identified by miRNomics  
323 and targeted metabolomics, from plasma, synovial fluid and urine, and uncovered three unique  
324 cluster endotypes. To our knowledge, our study is the first to use patient-matched multi-fluid,  
325 multi-omic approach to KOA patient endotyping. Not only did we uncover three unique multi-  
326 omic-based cluster endotypes, each was linked to unique biologically-relevant pathways. Despite  
327 a similar clinical phenotype, the cluster 1 endotype was primarily linked to metabolic processes  
328 and nervous system pathways, the cluster 2 endotype was primarily associated with aging  
329 pathways, and the cluster 3 endotype was primarily linked to immune, endocrine, and lipid  
330 metabolism pathways. Overall, the cluster endotypes uncovered are likely to contribute to, or be  
331 the result of, distinct mechanisms associated with KOA patients. Finally, using this novel approach  
332 to cluster endotyping, combined with integrative multimodal ML, we enhanced classification of  
333 patient-reported pain and function responses beyond that achieved using clinical measures alone.  
334 Surprisingly, it was the molecular entities that primarily drove classification of pain and function  
335 responses using our integrative modeling. Overall, our unique methodological approach reduced  
336 OA patient heterogeneity by defining patient clusters that had intra-cluster molecular differences  
337 that enhanced classifying pain and function responses to TKA. As endotypes are further refined and

338 molecular entities best associated with classification are further characterized, it will be important  
339 to understand how these response-based signatures may relate to physiological responses post-  
340 surgery.

341 The strength of our methodology lies in developing integrative deep learning and ML techniques  
342 for efficient multi-omic endotyping and response classification in KOA patients. VAEs offer  
343 advantages to integrating multiple domains for clustering compared to traditional approaches by  
344 effectively capturing the underlying structure of heterogeneous data through a joint latent  
345 representation. While traditional approaches such as dimensionality reduction or sequential  
346 clustering may provide insights, they often suffer from limitations such as difficulty in capturing  
347 non-linear relationships, and inadequate integration of domain-specific characteristics. Unlike  
348 standard VAEs,<sup>19</sup> we employed 4 separate decoders, enabling domain-specific reconstruction  
349 facilitating robust subject clustering, accounting for the inherent uncertainty in the latent space via  
350 the VAE's probabilistic nature. Overall, VAEs effectively leverage complementary information  
351 from multiple modalities for a more comprehensive characterization of KOA patient endotypes.

352 The integration of multiple data domains through our comprehensive two-step ML framework  
353 represents a significant advancement in response modeling for KOA outcomes by combining  
354 complementary information inherent in metabolomics, miRNA, and clinical data domains.  
355 Utilization of unimodal ML models in the first step allowed for extraction of domain-specific  
356 features that could classify 1-year TKA pain and function responses, while subsequent integration  
357 of these features using a naïve-Bayes meta-classifier enhanced classification accuracy. Naïve-  
358 Bayes classifiers emulate aspects of clinical decision-making by probabilistically combining  
359 evidence from multiple sources to make classifications.<sup>30</sup> Importantly, our novel framework  
360 demonstrated improvements in classification performance compared to unimodal domain-specific  
361 models, underscoring the utility of an integrative approach.

362 Although we included three biofluids to integrate miRNA or metabolite features to identify  
363 endotypes among KOA patients, additional endotypes may exist. Incorporating additional omic  
364 technologies (e.g. proteomics, genomics, methylomics) in the presented framework, as well as  
365 comprehensively evaluating omic measures across all biofluids may further refine endotypes, or  
366 uncover additional endotypes to further improve our understanding of KOA and ability to more  
367 accurately classify responses to interventions. Future studies should also focus on easily obtained  
368 patient biofluids, such as urine and blood, to determine if the presented approach can show similar

369 endotyping capability and response classification accuracy. In response classification modelling,  
370 we only evaluated a subset of clinical and demographic variables associated with KOA and patient  
371 outcomes to TKA, but incorporation of additional patient-related clinical and sociodemographic  
372 variables (e.g. comorbidities, medication use, race etc.), alongside endotype data, may also help  
373 improve modeling accuracy. Although we extensively validated our integrative ML unimodal  
374 models using a 10-time, 10-fold cross-validation, a lack of external validation remains. For  
375 external validation to be accomplished, better patient clinical and sociodemographic annotations,  
376 omic data and biosample sharing practices, and harmonization are needed.<sup>7,8</sup> Lastly, similar  
377 evaluations in additional patient cohorts, such as those with early-KOA and other afflicted joints,  
378 or evaluating other response measures, would also be of interest moving forward.  
379 Overall, using our novel modelling framework, we were able to unravel some heterogeneity of a  
380 sample of late-stage surgical KOA patients and evaluate post-TKA response classification. We  
381 anticipate this methodological approach will aid in understanding underlying molecular  
382 contributors and pathways to clusters of OA patients, and define molecular signatures contributing  
383 to intervention response. With additional studies, our methodological approach could ultimately  
384 help in shared patient-clinician decision making with regard to proceeding with selected therapies,  
385 including TKA for primary KOA.

386

### 387 **Contributors**

388 JSR, DS, OEG, KH, AS, YRR, AVP, RG and MK conceptualized the study. NNM, K. Syed, AVP,  
389 YRR, RG and MK supervised patient data and biofluid collection. KP managed biofluid and  
390 patient data storage. JSR, DS, OEG, KH, AS, CP, IJ, K. Sundararajan, YRR, AVP, RG and MK  
391 processed and curated the data. JSR, DS, OEG, KH, AS, CP, PP, NF, IJ and MK created the  
392 methodology and validated the data. DS developed the deep learning and machine learning  
393 algorithms. DS, OEG, and KH performed statistical analyses. JSR, DS, CP, IJ and MK created  
394 figures and tables. JSR, DS, KH, AS, CP and MK wrote the manuscript. MK supervised and  
395 acquired funding for multi-omic analysis. All authors had full access to the study data, were  
396 involved in manuscript editing, and were responsible for the decision to submit for publication.

397

### 398 **Data Sharing**

399 De-identified subject primary microRNA sequencing datasets are available on the Gene  
400 Expression Omnibus under accession number GSE222979. Software code and the dataset of  
401 processed miRNA counts, metabolite concentrations and demographic, anthropometric and  
402 clinical questionnaire responses used in this study is available at  
403 [https://github.com/divya031090/DeepLearning\\_KOA](https://github.com/divya031090/DeepLearning_KOA).

404

#### 405 **Declaration of Interests**

406 We declare no competing interests.

407

#### 408 **Acknowledgments**

409 Funding for this project was provided by the Canada Research Chairs Program (MK), Tony and  
410 Shari Fell Platinum Chair in Arthritis Research (MK), Campaign to Cure Arthritis, University  
411 Health Network Foundation. AVP is supported by the Arthritis Society Canada STAR Award-20-  
412 0000000012 and YRR is supported by J. Bernard Gosevitz Chair in Arthritis Research at  
413 University Health Network. Computational analysis was supported in part by funding from  
414 Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2024-04314),  
415 Canada Foundation for Innovation (CFI #225404, #30865), and Ontario Research Funds (RDI  
416 #34876, RE010-020). The funders had no role in study design, data collection and analysis,  
417 decision to publish, or preparation of the manuscript. Authors would like to thank the clinical  
418 research team within the Division of Orthopedics and members of the Buchan Arthritis Center at  
419 the Schroeder Arthritis Institute for their assistance in study recruitment. We also thank Dr. Max  
420 Kotlyar for his assistance during initial discussions related to the study.

421

## 422 **References**

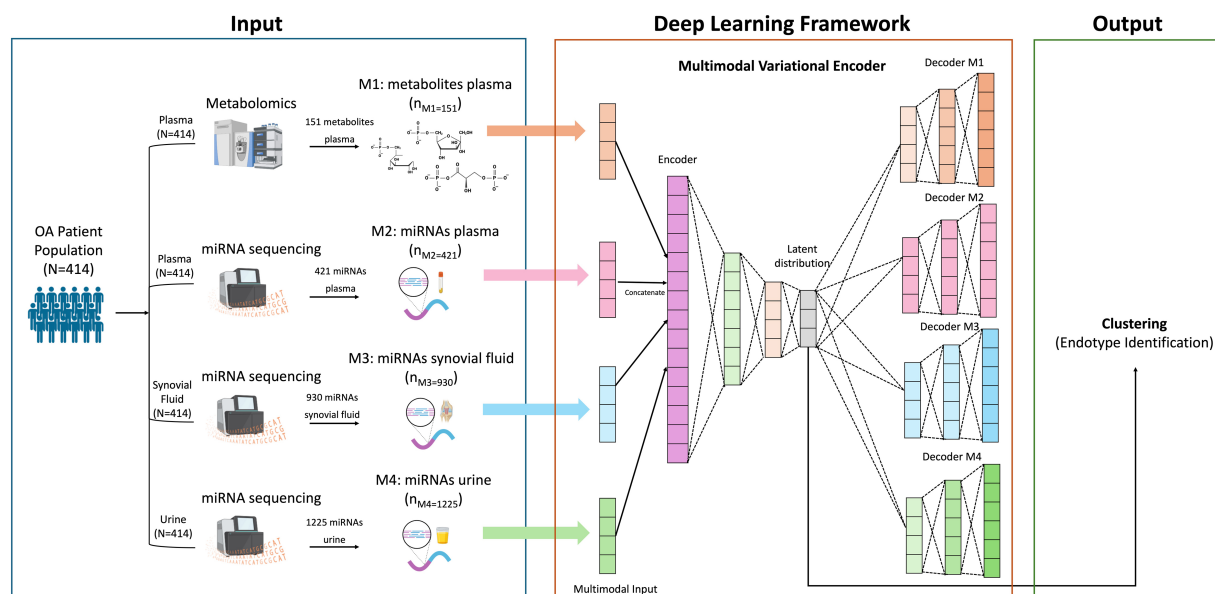
- 423 1. Long H, Liu Q, Yin H, et al. Prevalence Trends of Site-Specific Osteoarthritis From 1990  
424 to 2019: Findings From the Global Burden of Disease Study 2019. *Arthritis Rheumatol* 2022;  
425 **74**(7): 1172-83.
- 426 2. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *Lancet* 2019; **393**(10182): 1745-59.
- 427 3. Bierma-Zeinstra SM, Verhagen AP. Osteoarthritis subpopulations and implications for  
428 clinical trial design. *Arthritis Res Ther* 2011; **13**(2): 213.
- 429 4. Dell'Isola A, Allan R, Smith SL, Marreiros SS, Steultjens M. Identification of clinical  
430 phenotypes in knee osteoarthritis: a systematic review of the literature. *BMC Musculoskelet Disord*  
431 2016; **17**(1): 425.
- 432 5. Deveza LA, Melo L, Yamato TP, Mills K, Ravi V, Hunter DJ. Knee osteoarthritis  
433 phenotypes and their relevance for outcomes: a systematic review. *Osteoarthritis Cartilage* 2017;  
434 **25**(12): 1926-41.
- 435 6. Beswick AD, Wylde V, Gooberman-Hill R, Blom A, Dieppe P. What proportion of  
436 patients report long-term pain after total hip or knee replacement for osteoarthritis? A systematic  
437 review of prospective studies in unselected patients. *BMJ Open* 2012; **2**(1): e000435.
- 438 7. Ramos YFM, Rice SJ, Ali SA, et al. Evolution and advancements in genomics and  
439 epigenomics in OA research: How far we have come. *Osteoarthritis Cartilage* 2024.
- 440 8. Rai MF, Collins KH, Lang A, et al. Three decades of advancements in osteoarthritis  
441 research: insights from transcriptomic, proteomic, and metabolomic studies. *Osteoarthritis*  
442 *Cartilage* 2024; **32**(4): 385-97.
- 443 9. Nielsen RL, Monfeuga T, Kitchen RR, et al. Data-driven identification of predictive risk  
444 biomarkers for subgroups of osteoarthritis using interpretable machine learning. *Nat Commun*  
445 2024; **15**(1): 2817.
- 446 10. Angelini F, Widera P, Mobasher A, et al. Osteoarthritis endotype discovery via clustering  
447 of biochemical marker data. *Ann Rheum Dis* 2022; **81**(5): 666-75.
- 448 11. Werdyani S, Liu M, Zhang H, et al. Endotypes of primary osteoarthritis identified by  
449 plasma metabolomics analysis. *Rheumatology (Oxford)* 2021; **60**(6): 2735-44.
- 450 12. Rockel JS, Layeghifard M, Rampersaud YR, et al. Identification of a differential  
451 metabolite-based signature in patients with late-stage knee osteoarthritis. *Osteoarthr Cartil Open*  
452 2022; **4**(3): 100258.



- 453 13. Ali SA, Espin-Garcia O, Wong AK, et al. Circulating microRNAs differentiate fast-  
454 progressing from slow-progressing and non-progressing knee osteoarthritis in the Osteoarthritis  
455 Initiative cohort. *Ther Adv Musculoskelet Dis* 2022; **14**: 1759720X221082917.
- 456 14. Li YH, Tavallae G, Tokar T, et al. Identification of synovial fluid microRNA signature in  
457 knee osteoarthritis: differentiating early- and late-stage knee osteoarthritis. *Osteoarthritis*  
458 *Cartilage* 2016; **24**(9): 1577-86.
- 459 15. Ali SA, Gandhi R, Potla P, et al. Sequencing identifies a distinct signature of circulating  
460 microRNAs in early radiographic knee osteoarthritis. *Osteoarthritis Cartilage* 2020; **28**(11): 1471-  
461 81.
- 462 16. Beyer C, Zampetaki A, Lin NY, et al. Signature of circulating microRNAs in osteoarthritis.  
463 *Ann Rheum Dis* 2015; **74**(3): e18.
- 464 17. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a  
465 multimodal deep learning method. *Nat Commun* 2022; **13**(1): 7705.
- 466 18. Kopf A, Fortuin V, Somnath VR, Claassen M. Mixture-of-Experts Variational  
467 Autoencoder for clustering and generating from similarity-based representations on single cell  
468 data. *PLoS Comput Biol* 2021; **17**(6): e1009086.
- 469 19. Rong Z, Liu Z, Song J, et al. MCluster-VAEs: An end-to-end variational deep learning-  
470 based clustering method for subtype discovery using multi-omics data. *Comput Biol Med* 2022;  
471 **150**: 106085.
- 472 20. Sandhu A, Espin-Garcia O, Rockel JS, et al. Association of synovial fluid and urinary  
473 C2C-HUSA levels with surgical outcomes post-total knee arthroplasty. *Osteoarthritis Cartilage*  
474 2024; **32**(1): 98-107.
- 475 21. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of  
476 WOMAC: a health status instrument for measuring clinically important patient relevant outcomes  
477 to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;  
478 **15**(12): 1833-40.
- 479 22. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand*  
480 1983; **67**(6): 361-70.
- 481 23. Freynhagen R, Baron R, Gockel U, Tolle TR. painDETECT: a new screening questionnaire  
482 to identify neuropathic components in patients with back pain. *Curr Med Res Opin* 2006; **22**(10):  
483 1911-20.

- 484 24. Potla P, Ali SA, Kapoor M. A bioinformatics approach to microRNA-sequencing analysis.  
485 *Osteoarthr Cartil Open* 2021; **3**(1): 100131.
- 486 25. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering,  
487 scaling, and transformations: improving the biological information content of metabolomics data.  
488 *BMC Genomics* 2006; **7**: 142.
- 489 26. Hauschild AC, Pastrello C, Ekaputeri GKA, et al. MirDIP 5.2: tissue context annotation  
490 and novel microRNA curation. *Nucleic Acids Res* 2023; **51**(D1): D217-D25.
- 491 27. Pastrello C, Kotlyar M, Abovsky M, Lu R, Jurisica I. PathDIP 5: improving coverage and  
492 making enrichment analysis more biologically meaningful. *Nucleic Acids Res* 2024; **52**(D1):  
493 D663-D71.
- 494 28. Brown KR, Otasek D, Ali M, et al. NAViGaTOR: Network Analysis, Visualization and  
495 Graphing Toronto. *Bioinformatics* 2009; **25**(24): 3327-9.
- 496 29. Fernandez-de-Las-Penas C, Florencio LL, de-la-Llave-Rincon AI, et al. Prognostic Factors  
497 for Postoperative Chronic Pain after Knee or Hip Replacement in Patients with Knee or Hip  
498 Osteoarthritis: An Umbrella Review. *J Clin Med* 2023; **12**(20).
- 499 30. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using  
500 confounding patient and healthcare variables. *NPJ Digit Med* 2019; **2**: 31.
- 501

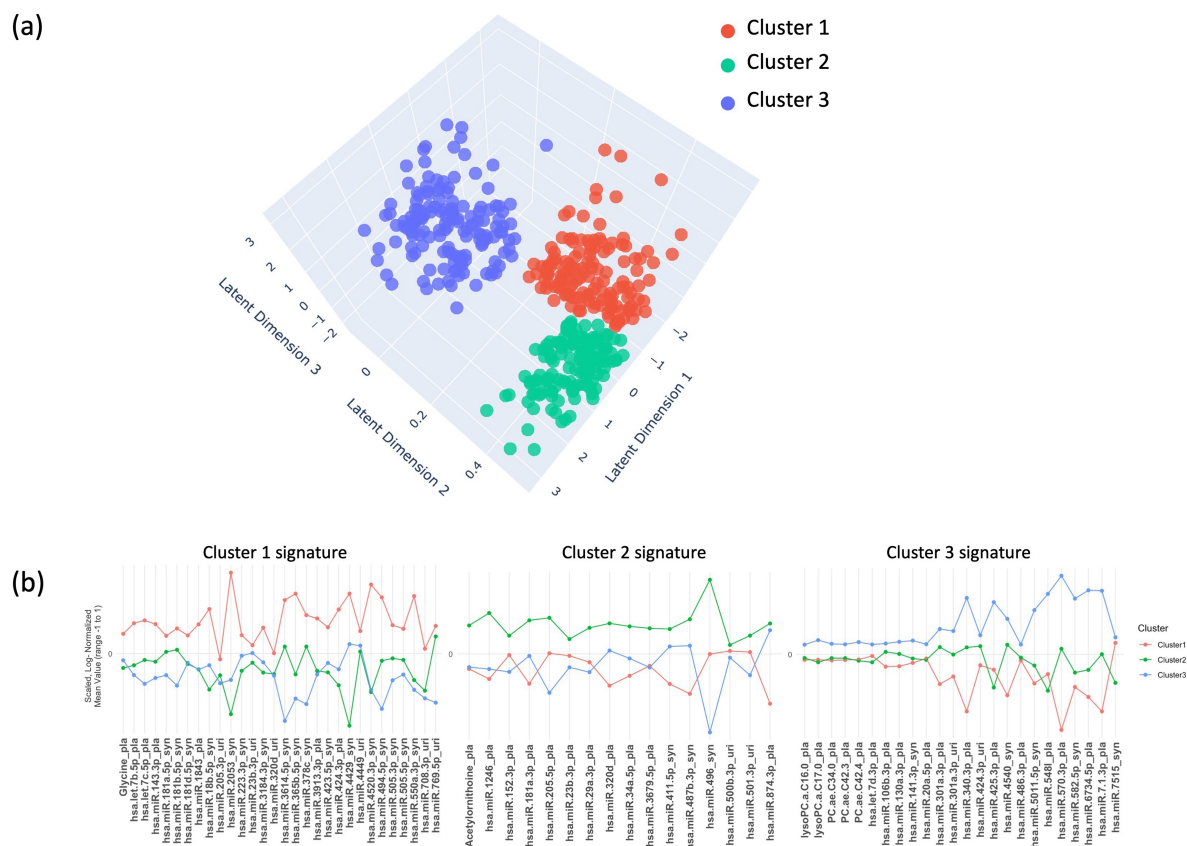
502  
503  
504



505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515

**Figure 1: Overall framework for Deep Learning-based Multimodal Clustering.** Deep-learning framework for multimodal integration and clustering using Variational Autoencoder (VAE) modeling. Four individual decoders in the proposed VAE accurately identify the latent distribution within each domain (M1-M4) capturing the complex non-linear associations within the multimodal data. K-means approach is used for identifying three clusters from the obtained latent distribution. ‘N’ represents the no. of patients and ‘nMx’ represents the number of features in each data domain.

516



517

518

519 **Figure 2. Integrated analysis of patient clustering and miRNA and metabolite feature**  
 520 **signatures.** (a) Three-dimensional illustration (latent dimension 1-3) of the three clusters obtained  
 521 using a Variational Autoencoder-based deep learning framework. Cluster 1 (red) comprises 146  
 522 patients, Cluster 2 (green) consists of 138 patients, and Cluster 3 (blue) includes 130 patients. (b)  
 523 Molecular signature profiles for clusters 1, 2, and 3 respectively (left to right) derived through the  
 524 intersection of the most significant variables ( $p < 0.05$ ) identified from both standardized mean  
 525 differences analysis and differential expression analysis within the plasma metabolites and  
 526 miRNA domains, differentiating each cluster.

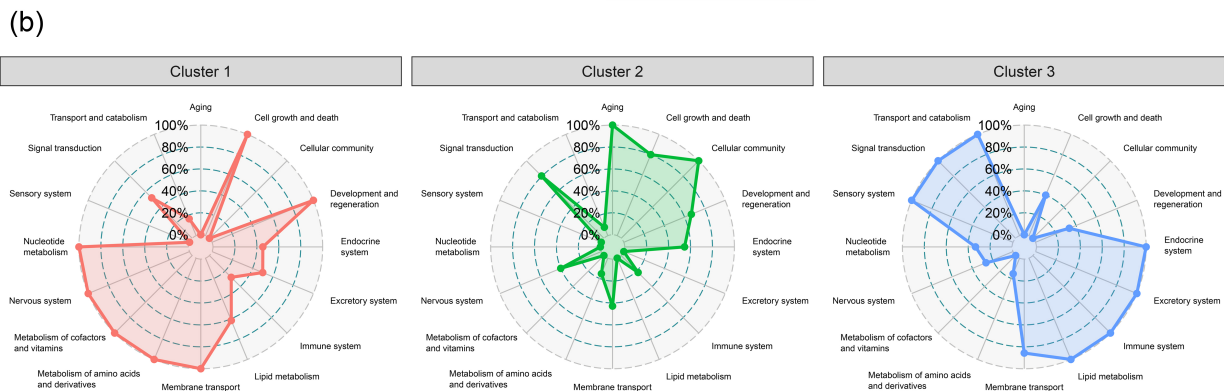
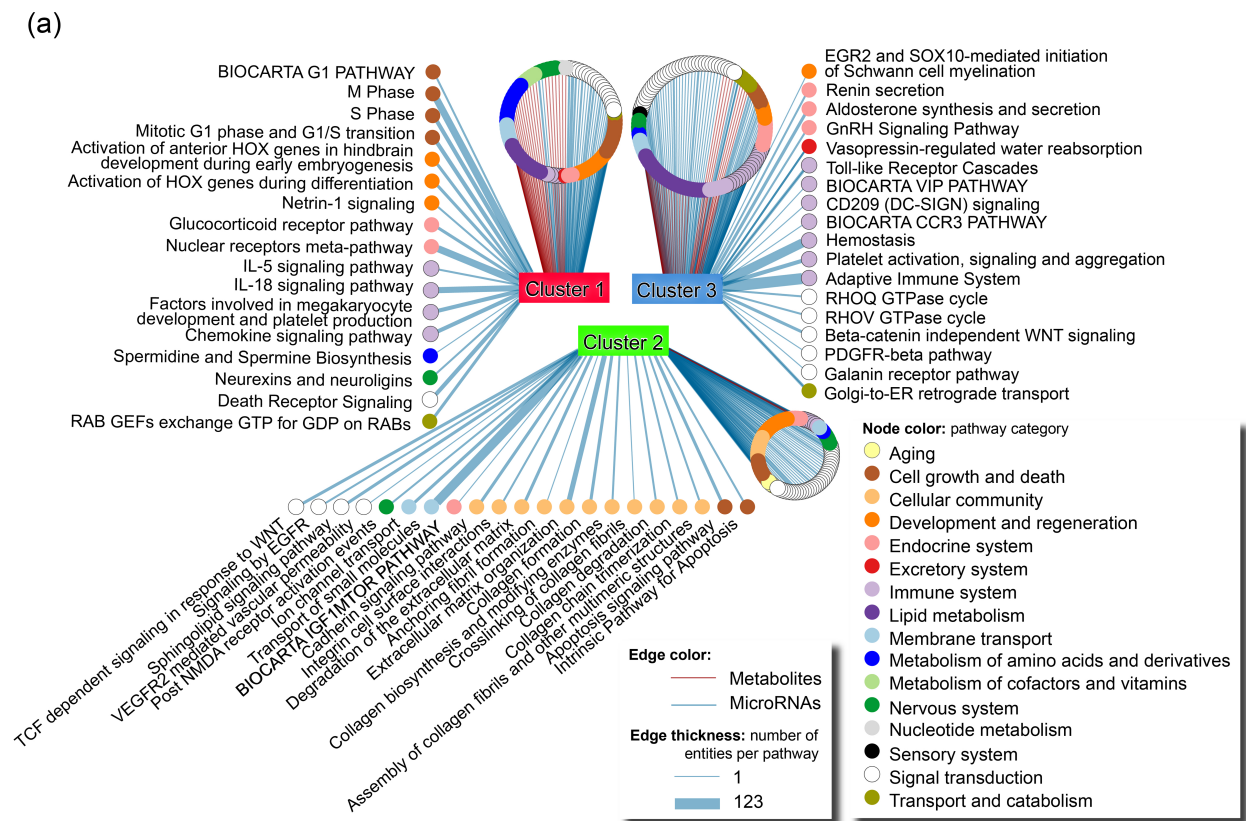
527

**Table 1. Summary statistics of clinical variables within each cluster identified from variational autoencoder machine-learning modeling and K-means clustering.** Frequency (percentage) are provided for categorical variables while median (Quartile 1, Quartile 3) values are presented for continuous variables according to patients within each cluster. p value is computed using wilcoxon rank sum test for continuous variables and chi-square or fisher's test as appropriate for categorical variables.

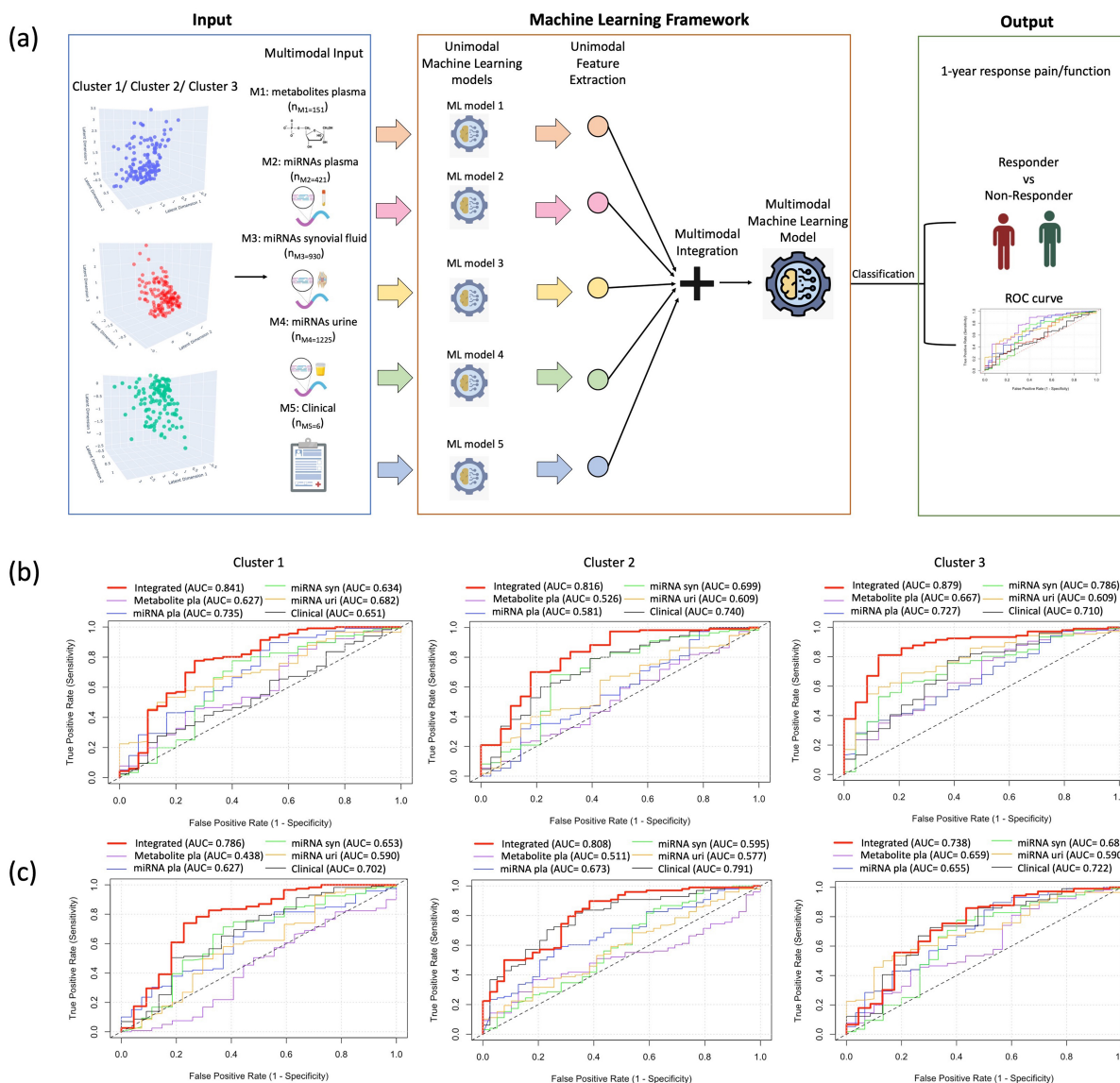
	Cluster 1 (n=146)	Cluster 2 (n=138)	Cluster 3 (n=130)	p-value between clusters
<b>Sex</b>				0.9
Female [N, (%)]	84 (58)	76 (55)	72 (55)	
Male [N, (%)]	62 (42)	62 (45)	58 (45)	
<b>Age</b>				0.37
Mean (sd)	66.2 (8.2)	65.4 (8.0)	65.5 (8.9)	
Median (Q1,Q3)	67 (62, 71)	65.0 (60.0, 70.8)	66 (60, 72)	
<b>BMI</b>				0.55
Mean (sd)	30.7 (6.1)	31.9 (7.7)	31.4 (7.3)	
Median (Q1,Q3)	29.2 (26.4, 33.8)	30.0 (26.5, 35.7)	31.1 (25.3, 34.7)	
<b>HADS anxiety</b>				0.72
normal [N, (%)]	103 (72)	98 (73)	100 (79)	
borderline [N, (%)]	21 (15)	20 (15)	13 (10)	
case [N, (%)]	19 (13)	17 (13)	14 (11)	
Missing (N)	3	3	3	
<b>HADS depression</b>				<b>0.037</b>
normal [N, (%)]	107 (74)	105 (78)	112 (87)	
borderline [N, (%)]	23 (16)	19 (14)	6 (5)	
case [N, (%)]	15 (10)	11 (8)	11 (9)	
Missing (N)	1	3	1	
<b>painDETECT</b>				<b>0.011</b>
neuropathic [N, (%)]	30 (21)	14 (10)	11 (8)	
nociceptive [N, (%)]	86 (59)	82 (59)	87 (67)	
unclear [N, (%)]	27 (18)	31 (22)	23 (18)	
Missing [N, (%)]	3 (2)	11 (8)	9 (7)	
<b>WOMAC pain baseline</b>				0.98
Mean (sd)	10.2 (3.5)	10.0 (3.2)	10.0 (3.7)	
Median (Q1,Q3)	10 (8, 12)	10 (8, 12)	10.0 (8.0, 12.4)	
<b>WOMAC pain 1yr</b>				0.34
Mean (sd)	3.7 (4.0)	3.7 (3.6)	3.2 (3.6)	
Median (Q1,Q3)	3.0 (1.0, 5.8)	3 (1, 6)	2.0 (0.2, 5.0)	
<b>WOMAC pain change (1 yr - baseline)</b>				0.55
Mean (sd)	-6.5 (4.1)	-6.3 (3.9)	-6.8 (4.1)	
Median (Q1,Q3)	-7 (-9, -4)	-6.0 (-9.0, -3.2)	-7.0 (-9.8, -4.0)	
<b>WOMAC pain change categorical</b>				0.48
≤33.3% [N, (%)]	27 (18)	33 (24)	22 (17)	
>33.3% [N, (%)]	119 (82)	105 (76)	108 (83)	
<b>WOMAC function baseline</b>				0.74
Mean (sd)	34.7 (11.9)	34.9 (11.4)	33.8 (12.3)	
Median (Q1,Q3)	36 (26, 41)	35.0 (26.2, 43.0)	34 (26, 41)	
Missing (N)	1	0	1	
<b>WOMAC function 1 yr</b>				0.24
Mean (sd)	14.8 (13.7)	14.7 (12.9)	12.8 (13.0)	
Median (Q1,Q3)	12.0 (4.2, 21.0)	11.5 (4.0, 20.8)	9.0 (2.2, 19.0)	
<b>WOMAC function change (1 year- baseline)</b>				0.76
Mean (sd)	-20.3 (13.4)	-20.0 (13.8)	-20.2 (13.3)	
Median (Q1,Q3)	-20 (-30, -11)	-18.4 (-31.0, -10.0)	-19.0 (-28.8, -11.0)	
<b>WOMAC function change categorical</b>				0.24
≤33.3% [N, (%)]	33 (23)	32 (23)	24 (19)	
>33.3% [N, (%)]	112 (77)	106 (77)	105 (81)	
Missing (N)	1	0	1	

BMI, body mass index; HADS, hospital anxiety and depression scale; Q, quartile; sd, standard deviation.

528  
529  
530



531  
 532  
 533 **Figure 3. Cluster endotype signatures are associated with different physiological pathways.**  
 534 (a) Network depicting unique enriched pathways from miRNA-gene targets and annotated  
 535 pathways from metabolites associated with individual cluster endotype signatures. Labels show  
 536 pathways with lowest q-value or highest number of annotated genes. (b) Radar plots of pathway  
 537 categorizations from enriched pathways from miRNA-gene targets and annotated pathways from  
 538 metabolites indicating categories most associated to individual clusters endotypes.  
 539



540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552

**Figure 4: Machine Learning modeling for classifying response to 1 year pain and function within each cluster** (a) Comprehensive two-step machine learning (ML) framework wherein, initially, unimodal ML models extract features from each domain, including metabolites, miRNA from plasma, synovial fluid, urine, and clinical data. A second-level multimodal machine learning classifier integrates these features to efficiently classify response vs. non-response at 1 year. (b) Receiver Operating Characteristic (ROC) plots illustrating the Area Under the Curve (AUC) for individual domains, alongside the integrated AUC (in red) combining all five domains to classify WOMAC pain response. (c) Receiver Operating Characteristic (ROC) plots illustrating the Area Under the Curve (AUC) for individual domains, alongside the integrated AUC (in red) combining all five domains to classify WOMAC function response. Notably, the integrated AUC outperforms individual AUCs in Cluster 1, Cluster 2, and Cluster 3.