Gene selection for prediction of transcriptome signal based on a machine learning approach

Discover Applied Sciences

RESEARCH Open Access

Gene selection for prediction of transcriptome signal based on a machine learning approach

Pietro Demurtas[1], Jacopo Bertozzi[1], Irene Di Silvestro[1], Kevin Carlin[2], Andre Ghetti[2], Brian Krause[2], Giovanni Perini[1], Ferdinando Zanchetta[1†] and Rita Fioresi[1*†]

† Giovanni Perini and Rita Fioresi have contributed equally to this work.

*Correspondence:
Rita Fioresi
rita.fioresi@unibo.it
[1]FaBiT, University of Bologna, Via San Donato 15, 41127 Bologna, BO, Italy
[2]Anabios Organization, 1155 Island Ave Suite 200, San Diego, CA 92101, USA

Abstract

Background In recent years, RNA-seq technology has gained widespread use in diverse research and clinical applications. Alongside this expansion, machine learning techniques have enabled accurate reconstruction of full transcriptomic signals from a considerably reduced set of highly informative genes (e.g., S1500+).

Results We employ machine learning methods, specifically XGBoost (eXtreme Gradient Boosting) a decision tree approach, to perform RNA-seq and transcriptomic analyses across multiple tissues. Our goal is to identify a small subset of expressed genes that can capture the complete tissue-specific transcriptomic profile. Using public GTEx (Genotype Tissue Expression) data, we analyze each tissue separately and discover the key fact that taking into account just the top 500 genes per tissue (ranked by XGBoost feature importance) are sufficient to provide transcriptomic signatures that match the performance of state-of-the-art gene sets (e.g., S1500+). To further validate our approach, we apply it to neuronal tissues by comparing samples from individuals with neuropathic pain versus those without pain. In dorsal root ganglia (DRG) RNA-seq data from patients experiencing varying levels of pain, our method suggests EGR1 as a factor in radicular/neuropathic pain, thereby opening avenues on the development of therapies that may alleviate pain by targeting EGR1 pathway.

Conclusions We demonstrate how to train and apply the XGBoost algorithm to select a small gene set that can approximate the full transcriptomic signal with varying accuracy depending on tissue type, generally achieving performance comparable to S1500+ gene sets in GTEx data. This method focusses on one tissue at a time, a different list of genes is selected for each tissue and ranked according to the importance of each gene into the reconstruction of the transcriptomic signal. This ranking also aids in highlighting specific genes that may be critical in predicting tissue-specific pathologies.

1 Introduction

RNA sequencing new technologies (RNA-seq) and transcriptomic analysis are now widely used to diagnose hereditary diseases and to assess general body health conditions [1, 2]. Though in recent years the cost of genome expression profiles has greatly decreased, the question of whether a reduced number of selected gene expression is able

to successfully reconstruct the whole transcriptomic signal, via machine learning methods, remains interesting from both a practical and theoretical point of view [3–5]. In particular, the growing focus on explainability in the machine learning algorithms [6] is now providing new tools to understand the logic behind the choice of the reduced set of genes effectively used in the predictions of the full signal [5]. Consequently, with this added knowledge, we can address tissue-specific questions and gain insight into the role and quantitative importance of each gene in the reconstruction of the full transcriptomic signal in a tissue-specific manner. Moreover, our analysis and the resulting gene ranking, based on their importance, can provide support in formulating new hypotheses and validating findings.

In this paper, we draw inspiration from the seminal work [3] in which only about 2600 genes, known as S1500+, out of a set of approximately 18K were shown to be effective in predicting adverse cellular responses induced by exposures to chemicals, drugs, and other agents. It is now widely accepted that the transcriptional profiles of many genes can be derived from the expression of a few key ones. This is indeed the main hypothesis driving the study [3], where the authors combine a machine learning approach for gene selection with a knowledge-driven one, leading to the identification of the sentinel genes, known as S1500+. This perspective was later pursued by others, in particular in [5], where the machine learning algorithm XGBoost (eXtreme-Gradient Boosting [7]) was employed to furtherly reduce the selection, from S1500+ to a set of about 1000 genes. XGBoost is indeed a strong learner model based on multiple decision tree models, on the principle of the celebrated Random Forest [8]. Generally it surpasses the performance of both Random Forest and the more standard methods including deep neural networks and KNN for transcriptomic analysis, as detailed in the introduction of [5]. It is now widely used in tandem with other methods for a variety of purposes [9–11], (see also [12] for a generative AI approach to similar biomedical questions).

In our work, we begin by conducting the same investigation as in [3], but we separately analyze the RNA-seq data from five different tissues: brain, skin, blood, esophagus, and blood vessels. These tissues were selected based on the number of available samples in the new publicly available dataset GTEx (Genotype-Tissue Expression) [13], i.e. we choose the tissues with the most abundant samples. The GTex portal provides RNA-seq mostly with Illumina technologies to create a uniform and reliable dataset of human gene expressions. As described in [13], GTex is a visionary project collecting multiple tissue types (up to 54) from about 960 deceased donors to study the relationship between genetic variation and gene expression and provides the whole genome sequence and RNA-seq data generated from these tissues. We take the V10 version as described in [14], where all protocols are clearly explained.

We train an XGBoost model, starting from the S1500+ information, to reconstruct the whole transcriptomic signal with varying accuracy, depending on the chosen tissue, but comparable with S1500+ selection as in [3]. Once the training is over, we take advantage of the "feature importance", an internal function of XGBoost, for the full explainability of its inner working, and we rank the S1500+ according to it (see also [12] for a recent report on such methods). We then show that the top 500 genes, ranked according to their feature importance, separately for each tissue, can reconstruct, through a new training the whole transcriptomic signal with accuracies comparable to the S1500+ one, thus showing that our tissue specific selection of genes allows a reduction of roughly

80%, from the 2600 genes as in [3] to 500 genes. These genes can successfully reconstruct the full transcriptomic signal in a tissue-specific manner, thus represent a substantial improvement with respect to both [3] and [5]. We measure accuracy according to several popular metrics as Mean Absolute Error (MAE), Mean Square Error (MSE), etc. see Sect. 2. The selection of genes varies significantly across tissues, and the number of selected common genes between two tissues (see also our Appendix).

We validate our selection by using the 500 genes we selected for brain tissue to predict the presence of radicular neuropathic pain, employing a completely different dataset from GTEx. In [15] the authors showed, through a proof of concept study, how a machine learning approach to RNA transcriptome analysis holds the potential for such prediction. They first perform RNA-sequencing on dorsal root ganglia (DRG) samples, taken from patients with variable presence of radicular/neuropathic pain. Then, using such labeled data (pain versus no-pain), they applied a Random Forest model to classify whether a sample belonged to the "pain" or "no-pain" category, achieving an accuracy of approximately 90% on a small set of samples. We take a similar, yet subtly different approach: we train an XGBoost model, starting with information from the top 100 genes in our ranking, then adding the information of one gene at a time as the experiment progresses, repeating the training as the information of each gene is added. We observe that maximum accuracy is reached only after including the EGR1 gene, suggesting that it plays a key role in the development of chronic pain, as independently confirmed by the study in [16].

In summary, we developed a novel approach to introduce a data-driven concept of *gene importance* in transcriptomic analysis, tailored to specific tissues. We demonstrated that a relatively small set of 500 genes can successfully reconstruct the full transcriptomic signal, achieving accuracies comparable to the S1500+ recognized set. Even using a fraction of this set (250 genes) still yields a reasonable level of accuracy. We validated our findings on a different dataset with binary labeling for the presence of radicular/neuropathic pain. Our approach highlights the EGR1 gene, confirming independent findings, and shows potential of our proof of concept method for future significant biomedical applications (e.g. cancer studies [17]), that should also be taking into account possible bias inherent to transcriptomic data [18, 19].

2 Materials and methods

We describe in this section the datasets we use for our gene selection and predictions, the machine learning algorithm for the gene selection (XGBoost) and the metrics we use to establish the accuracy of our prediction.

2.1 Datasets

GTEx dataset. The Adult Genotype Tissue Expression (GTEx) Project [13, 14] is an extensive public resource for researching human gene expression and regulation, as well as its connection to genetic variation across various diverse tissues and individuals. All the included individuals were densely genotyped to check for genetic variation by Whole Genome Sequencing (WGS) [20]. Gene expression of each tissue was assessed by RNA sequencing (bulk RNA-seq). In the present work, we focus on the Gene Expression Transcripts Per Million (TPM) data available for the GTEx V10 release (december 2024). The GTEx V10 release spans 31 generic tissue types divided in 54 specific tissues,

for about 1000 adult individuals. For our analysis we focused on the most abundant 5 tissues: brain, skin, blood, esophagus and blood vessels; their numerosity is given in Table 1.

S1500+ Genes. The S1500+ gene set [3] is a set of sentinel genes that adequately represents all canonical pathways from the Molecular Signature Database (MSigDB v4.0) [21, 22] and can be used to infer expression changes for the remainder of the transcriptome. The S1500+ set was obtained using first a data-driven computational model, then augmenting it by the addition of a knowledge-driven selection of supplementary genes. In this work, we use the S1500+ gene set as a starting point to rank gene importance for each of the five tissues. We then select the top 500 genes from each list and, after a second training, we measure the accuracy of full transcriptomic signal reconstruction, using only the information from these 500 selected genes. We always focus on one tissue at a time.

2.2 Algorithm and accuracy evaluation metrics

We briefly describe the algorithm XGBoost and the evaluation the performance of the trained models using the metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), R2 score and Fold metric. We shall describe just the R2 score and Fold metric, MAE and MSE being standard.

Algorithm XGBoost. The Extreme Gradient Boosting (XGBoost), is a scalable distributed gradient-boosted decision tree (GBDT) machine learning algorithm [7]. XGBoost is an ensemble classifier that uses the power of many weak learners working together, similarly to the well known Random Forest algorithm [8]. XGBoost is considered the gold standard machine learning algorithm for handling tabular data outperforming neural networks [23] and other boosting methods [24]. Furthermore, being a Gradient Boosting Method (GBM), during training, XGBoost calculates the relative importance of each input feature. The importance of features computed by XGBoost can be used to rank them by their impact on the output. We employed XGBoost native feature importance function to rank the importance of each gene in the reconstruction of the full transcriptome signal for each of the five examined tissues.

R2 **score**. The coefficient of determination, also known as R2 score and denoted by R2 is defined as:

$$\mathrm{R2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where y_i, $\hat{y}_i$ and $\bar{y}$ are the ground truth, the predicted label and the sample average respectively. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes with respect to the model prediction. Notice that the closer the R2 score is to 1, the better the prediction is. More importantly, it indicates greater reliability in statistical terms. In fact, a prediction close to the average of sample values will result in a large negative R2 value. Indeed, this metric can also take on a negative value, unlike more popular metrics such as MAE and MSE, highlighting, in such case, the poor statistical meaning of the prediction.

Table 1 Samples per tissue, we use from GTEx

	Blood	Brain	Skin	Esophagus	Blood vessel
Number of samples	1130	3234	2057	1578	1431

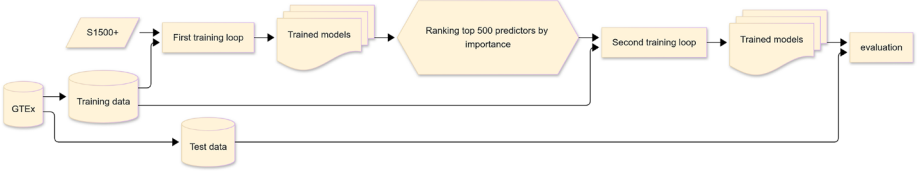

Fig. 1 Flow chart of our algorithmic procedure to reach the transcriptomic signal prediction

Table 2 Metrics for the brain tissue: first training (S1500+)

Fold	MSE	MAE	R2	FM	Mean	Variance
0	225.789	1.902	0.770	0.853	18.848	6995.647
1	287.346	1.913	0.529	0.853	18.948	6793.331
2	369.551	1.900	0.733	0.852	18.599	7401.000
3	292.287	1.902	0.746	0.853	18.880	7618.306
4	289.341	1.888	0.769	0.854	18.829	7536.330
Mean	292.863	1.901	0.709	0.853	18.821	7268.923

Fold Metric. The Fold Metric is defined as:

$$\mathrm{FM}(y,\hat{y}) = \begin{cases} \frac{\max(y,\hat{y})}{\min(y,\hat{y})+\epsilon}, \text{if } y > 0 \\ 1, \quad \text{if } y = 0 \quad \text{or} \quad \hat{y} < t \end{cases}$$

where t is the first non-zero percentile of data, e.g. if the first percentile of the data has transcription expression value equal to zero we set $t = 0$. This parameter takes into account the fact that our dataset is sparse.

For simplicity we report the mean across samples of how many prediction fall within 2 folds. In other words we check, on average, on how many genes $\mathrm{FM}(y,\hat{y}) < 2$. The final metric is normalized to obtain a number between 0 and 1 representing the fraction of genes whose prediction is within two folds from ground truth. The closer this index is to 1, the better the prediction is.

3 Results and discussion

Our results are obtained as follows and the procedure is summarized in the flow chart in Fig. 1. We first perform a standard training of the XGBoost algorithm, aiming at the reconstruction of the full transcriptomic signal, starting from the information contained in the S1500+ dataset. For this training, we take a standard 5-fold cross-validation on each of the GTEx datasets of the five tissues as in Table 1. After such training, we achieve the reconstruction of the full transcriptomic signal, as in the work [3], with similar accuracy. The accuracies resulting from this first training are reported for each fold and for each tissue in Table 2 (brain), Table 4 (skin), Table 6 (blood), Table 7 (esophagus), Table 8 (blood vessels).

Once this first training is over, we compute the importance of each gene, using the feature importance function of XGBoost and taking the average of such importance across the five folds, accounting also for the statistical relevance of the results in each fold, via the R2 score. Our formula to compute the *importance* of a given gene G is:

$$\mathrm{Imp_G} = \frac{1}{5}\sum_{n=1}^{5} \mathrm{Imp}_{\mathrm{G},n} \times \mathrm{R2}_n \tag{1}$$

where $\mathrm{Imp}_{\mathrm{G},n}$ is the value of the feature importance function of XGBoost in the n^{th} fold for a given gene G. We multiply $\mathrm{Imp}_{\mathrm{G},n}$ by $\mathrm{R2}_n$, the R2 score in the n^{th} fold, to give more weight to the statistically relevant folds. The feature importances obtained by XGBoost give a quantitative way to understand which are the feature that are more valuable in constructing the boosted decision trees part of the XGBoost model. In our case, genes with an higher importance are deemed by the algorithm to have an higher expressive power to predict the TPM values of the target genes. As these importances are more informative the more effective the model is, we have weighted them using the R2 scores obtained for each fold. This mechanism can be interpreted as an attention mechanism: our modified importances pay more attention to the importances of the best performing models. We remark that this is very similar to XGBoost feature importance: it just rewards the folds with more statistical significance.

We then proceed to rank genes according to their importance as computed in (1). For each tissue we take the top 500 genes, that we shall call, from now on, the *top predictors* and we retrain XGBoost to reconstruct the full transcriptomic signal, starting from this information. This second training is conceptually the same as the first one, except that we replace the S1500+ gene set with the 500 top predictors for each tissue. We report the results of the second training in Table 3 (brain), Table 5 (skin), Table 9 (blood), Table 10 (esophagus), Table 11 (blood vessels), and we discuss them in detail below.

We explain graphically in Fig. 1 how our algorithm arrives to the predictions of the full transcriptomic signal for each tissue using only the top 500 predictors.

We trained each XGBoost model using the hyperparameters shown below, resorting to the default XGBoost library hyperparameters, when not specified. In particular, we take as hyperparameters for XGBoost the ones found in [5], where the authors perform a grid search to determine the optimal hyperparamenters to use for a gene expression value prediction task very similar to ours. The only exception is the number of estimators: as our dataset is smaller than the one used in [5], we used 150 instead of 300, to reduce overfitting as we found that with both choices the algorithms achieve similar performances on the validation set.

No. of estimators	Learning rate	Max depth	Training subsample ratio	Columns subsample ratio	Gamma	Importance type
150	0.100	8	0.800	0.800	0.100	Total Gain

3.1 Transcriptome signal prediction in tissues

We now examine the predictive power of our trained XGBoost algorithm on the transcriptome signal reconstruction after the first and then after the second training, when we use just the top 500 predictors. We report our results separately for the various tissues: brain, skin, blood, esophagus and blood vessels.

Table 3 Metrics for the brain tissue: second training (top 500 predictors)

Predictors	MSE	MAE	R2	FM	Mean	Variance
S1500+	292.863 ± 45.658	1.901 ± 0.008	0.709 ± 0.091	0.853 ± 0.001	18.821 ± 0.118	7268.923 ± 319.964
top 500	784.818 ± 345.112	2.310 ± 0.029	0.641 ± 0.082	0.854 ± 0.001	20.279 ± 0.173	8199.891 ± 650.727
top 500*	409.656 ± 43.633	2.113 ± 0.024	0.709 ± 0.028	0.854 ± 0.001	20.279 ± 0.173	8199.891 ± 650.727

Brain tissue. Brain tissue is one of the most represented tissue type in the Adult GTEx dataset and the study of its transcriptome is a crucial step to answer several biological and clinical questions (see also Sect. 3.2). In the first training, the model is trained with the information of the S1500+ gene set and, at the end of the training, it achieves good results, in reconstructing the full transcriptomic signal, as expressed in Table 2, for each of the five folds according to the metrics described in Sects. 2, 2.2.

In fact, the model scores a median R2 of 0.7 and a mean MAE of 1.9 compared with a mean testing data variance of 7268.92. Overall, the performances are stable across the five folds. It is worth noting that, while the MAE is quite small across all the folds, the MSE is significantly larger. This relation between MAE and MSE suggest that the model performs quite well in regressing the expression profiles of most genes, hence the low MAE, but it makes some significant errors on few genes, giving an higher MSE.

We then determine the top 500 predictors of the S1500+ ranked by importance as computed in formula (1). Figure 2 expresses in a graph the importance of the top predictor genes for the brain tissue.

Once the top 500 predictor list is determined, we proceed and perform the second training. The average performance of the model trained with the information coming just from the top 500 predictors in reconstructing the full transcriptomic signal is very similar to the one obtained the S1500+ set as summarized by Table 3. This similarity in performance clearly confirms that it is possible to reduce the number of predictors without decreasing the performance on tissue-specific data. Hence the top 500 predictors hold the capability to reconstruct the full transcriptomic signal.

In Table 3, we denote with 500* the predictions of the transcriptomic signal *excluding* the S1500+ genes, to make the comparison between the S1500+ and the top 500 predictors performances more fair. In any case we report in the table below the results of the top 500 predictors on both the full transcriptomic signal (500 label) and the transcriptomic signal where we remove the S1500+ set (500* label). More in general we denote with N^* the N top predictor genes excluding those in S1500+, to make a fair comparison with the S1500+ performance.

Skin tissue. After the first training, the XGBoost model trained on skin data achieves a lower performance than the one trained on brain data in the task of reconstructing the full transcriptomic signal using the information of the S1500+. In fact in Table 4 we notice higher MSE and MAE than the one reported in Table 2 on brain data, however it must be noted tht the variance of skin data is significantly larger. Furthermore, the

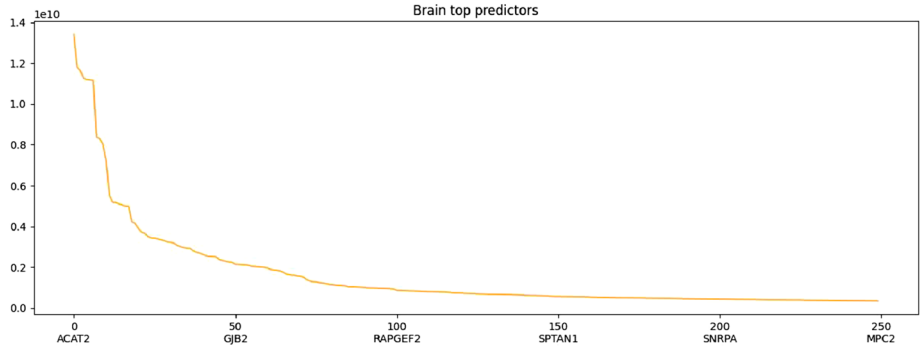

Fig. 2 Importance of the top predictor genes for brain tissue

Table 4 Metrics for the skin tissue: first training

Fold	MSE	MAE	R2	FM	Mean	Variance
0	2519.287	3.636	0.595	0.821	30.902	72573.096
1	2624.706	3.631	0.600	0.826	30.290	70782.972
2	2200.413	3.522	0.592	0.822	30.681	70723.796
3	2395.010	3.613	0.361	0.823	30.585	69129.999
4	2209.257	3.600	0.201	0.820	30.836	72261.590
Mean	2389.735	3.600	0.470	0.823	30.659	71094.291

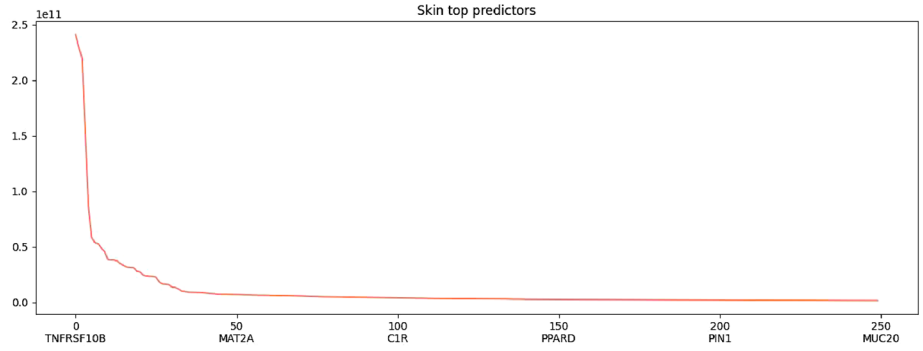

Fig. 3 Importance of the top 250 predictor genes for skin tissue

Table 5 Metrics for the skin tissue: second training (top 500 predictors)

Predictors	MSE	MAE	R2	FM	Mean	Variance
S1500+	2389.735 ± 167.593	3.600 ± 0.041	0.470 ± 0.162	0.823 ± 0.002	30.659 ± 0.216	71094.291 ± 1236.291
top 500	3823.821 ± 101.667	4.803 ± 0.064	0.435 ± 0.154	0.824 ± 0.001	35.770 ± 0.107	87331.877 ± 1568.660
top 500*	2946.035 ± 160.183	4.023 ± 0.058	0.463 ± 0.127	0.824 ± 0.001	35.770 ± 0.107	87331.877 ± 1568.660

model achieves lower R2 score and similarly lower FM with respect to the one trained on brain tissue.

We report the results regarding the first training, that is the transcriptomic signal reconstruction with the S1500+ information in Table 4 and the graph of the gene importance in Fig. 3.

As for the second training, we notice that, similarly to the brain tissue model, the model trained using just the top 500 predictors achieved again comparable performances with respect to the S1500+, as reported in Table 5. Thus, we confirm that the information contained in the top 500 predictors can successfully reconstruct the full transcriptomic signal also for the skin tissue.

Blood, esophagus, blood vessel tissues. In this section we report the results relative to the first and second training of the model for the tissues: blood, esophagus, blood vessels (Tables 6, 7, 8). Similar considerations as in the previous exposition of skin tissue result hold for these results. In particular Tables 9, 10, 11, confirm the capability of the top 500 predictors to reconstruct the full transcriptomic signal (See Figs. 4, 5, 6).

Notice that, as our tables show, the metrics for blood, esophagus and blood vessel tissues appear more unstable, with respect to those regarding brain and skin tissues. This is due to a reduced number of samples available for the training as our Table 1 shows. Since

Table 6 Metrics for the blood tissue: first training

Fold	MSE	MAE	R2	FM	Mean	Variance
0	1317.454	3.130	0.323	0.740	21.920	31269.100
1	1338.857	3.175	0.297	0.739	22.146	31492.946
2	1232.518	3.124	0.471	0.748	23.105	31687.549
3	1495.385	3.047	0.116	0.742	22.139	30260.275
4	1381.304	3.239	−0.043	0.741	23.319	29799.153
Mean	1353.104	3.143	0.233	0.742	22.526	30901.804

Table 7 Metrics for the esophagus tissue: first training

Fold	MSE	MAE	R2	FM	Mean	Variance
0	2315.095	3.557	−0.251	0.826	27.487	60729.925
1	1814.509	3.465	−9.999	0.827	27.316	62694.249
2	4557.481	3.575	0.378	0.826	27.442	65742.101
3	1915.733	3.443	0.591	0.826	27.018	55770.666
4	2078.051	3.573	0.430	0.825	27.416	58762.818
Mean	2536.174	3.523	−1.770	0.826	27.336	60739.952

Table 8 Metrics for the blood vessel tissue: first training

Fold	MSE	MAE	R2	FM	Mean	Variance
0	1542.074	4.572	0.237	0.805	33.653	34300.505
1	1676.169	4.460	−1.088	0.808	33.524	35020.225
2	1598.693	4.544	0.311	0.807	33.554	33921.032
3	1385.905	4.387	0.347	0.808	33.474	33455.230
4	1493.017	4.472	−1.547	0.808	33.789	34373.410
Mean	1539.172	4.487	−0.348	0.807	33.599	34214.080

Table 9 Metrics for the blood tissue: second training (top 500 predictors)

Predictors	MSE	MAE	R2	FM	Mean	Variance
S1500+	1353.104	3.143	0.233	0.742	22.526	30901.804
std	113.818	0.035	0.211	0.002	0.285	768.249
top 500*	1541.806	3.651	0.092	0.744	24.631	33704.971
std	139.471	0.030	0.374	0.002	0.294	721.852
top 500*tbf	292.863	1.901	0.709	0.853	18.821	7268.923

Table 10 Metrics for the esophagus tissue: second training (top 500 predictors)

Predictors	MSE	MAE	R2	FM	Mean	Variance
S1500+	2536.174 ± 1024.713	3.523 ± 0.057	−1.770 ± 4.125	0.826 ± 0.001	27.336 ± 0.168	60739.952 ± 3389.505
top 500	2367.105 ± 951.715	3.960 ± 0.060	−1.253 ± 2.701	0.828 ± 0.001	29.026 ± 0.185	57517.187 ± 3021.023
top 500*	2409.652 ± 1078.885	3.662 ± 0.054	1.479 ± 3.038	0.828 ± 0.001	29.026 ± 0.185	57517.187 ± 3021.023

the performance correlates with sample size, we expect a better performance with larger datasets. Moreover the brain samples are more diverse, besides being more numerous, thus helping XGBoost in the learning process and reaching a better generalization capability, for such tissue.

Table 11 Metrics for the blood vessel tissue: second training (top 500 predictors)

Predictors	MSE	MAE	R2	FM	Mean	Variance
S1500+	1539.172 ± 97.919	4.487 ± 0.066	−0.348 ± 0.806	0.807 ± 0.001	33.599 ± 0.112	34214.080 ± 518.547
top 500	1588.899 ± 113.820	4.830 ± 0.074	−0.599 ± 0.973	0.809 ± 0.001	34.730 ± 0.115	32529.391 ± 438.951
top 500*	1468.360 ± 88.372	4.607 ± 0.064	−0.516 ± 0.903	0.809 ± 0.001	34.730 ± 0.115	32529.391 ± 438.951

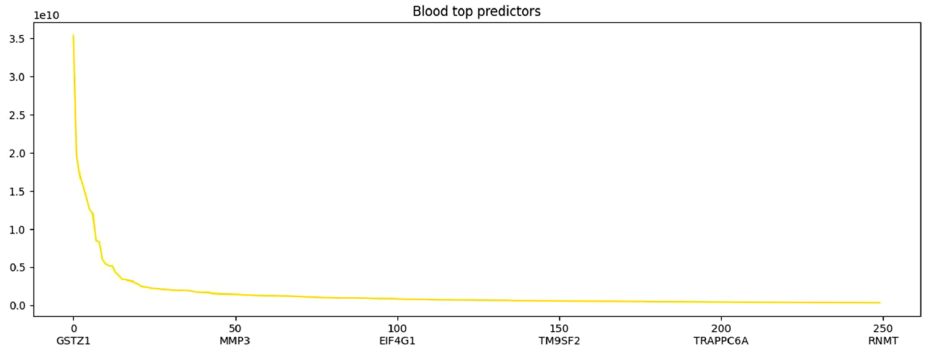

Fig. 4 Importance of the top predictor genes for blood tissue

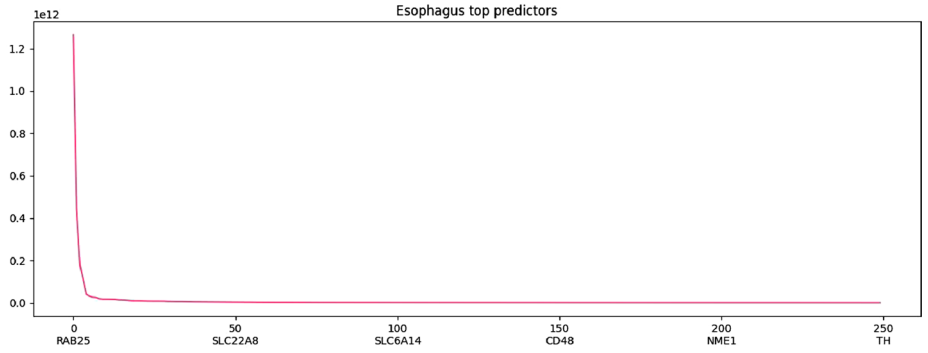

Fig. 5 Importance of the top predictor genes for esophagus tissue

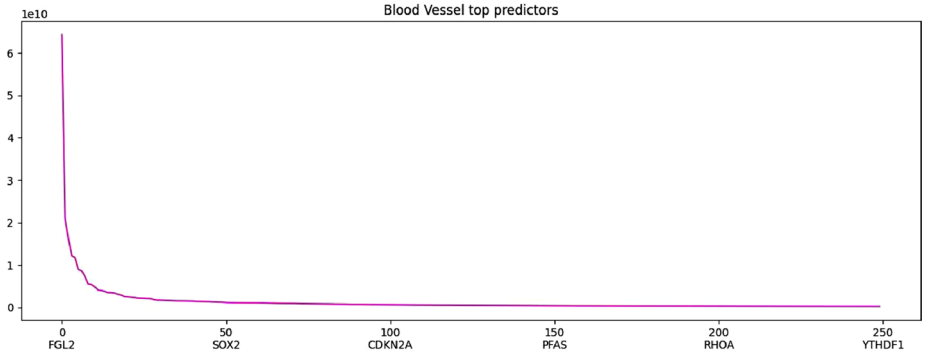

Fig. 6 Importance of the top predictor genes for blood vessel tissue

3.2 Application for pain versus no pain prediction

We now examine a database consisting of the RNA-seq data from 21 DRG (Dorsal Root Ganglia) samples from 15 patients as in [15]. We perform several experiments with XGBoost, following the same pattern as in [15], where they train a Random Forest algorithm with cross-validation method that leaves out one sample at each run. We take advantage of our 500 top importance gene selection of the brain tissue, as explained in the previous section, see Table 3, since we expect the gene ranking for this tissue to contain the key information for the pain versus no pain prediction for DRG samples of selected patients [15]. These are the top 500 genes according to our ranking, and they allowed us to obtain a full transcriptomic signal reconstruction with accuracy comparable to the S1500+ as expressed in Table 3.

As for the present question, we perform via XGBoost a binary pain versus no-pain classification by taking a varying number of genes starting from 100 up to the full S1500+ set, following the order of importance (Fig. 7). We describe in more detail our procedure, exemplified in the flow chart given below. To start with, we rank the S1500+ gene list to obtain an ordered list. We then choose the fist 100 genes of this ordered list as predictor genes and we perform a Leave One Out Cross Validation (LOOCV) of an XGBoost binary classifier using as features only the seleected predictor genes. The performance of the LOOCV is evaluated computing the accuracy, precision, recall and F1 score metrics, (see also the procedure found in [15] for more explaination). The process is then repeated selecting as predictor genes the first 101, 102, 103, ...genes of our ordered list until we use as predictor genes all the S1500+ genes for which we have data. For our XGBoost classifier, we set the n_estimators, max_leaves and max_depth to be equal to 25, 10 and 5 respectively following the choices made in [15] for the parameters of the random forest algorithm they employ. We used a learning rate of 0.01 and left the other hyparparameters as the default ones for this task. Finally, in each LOOCV fold,

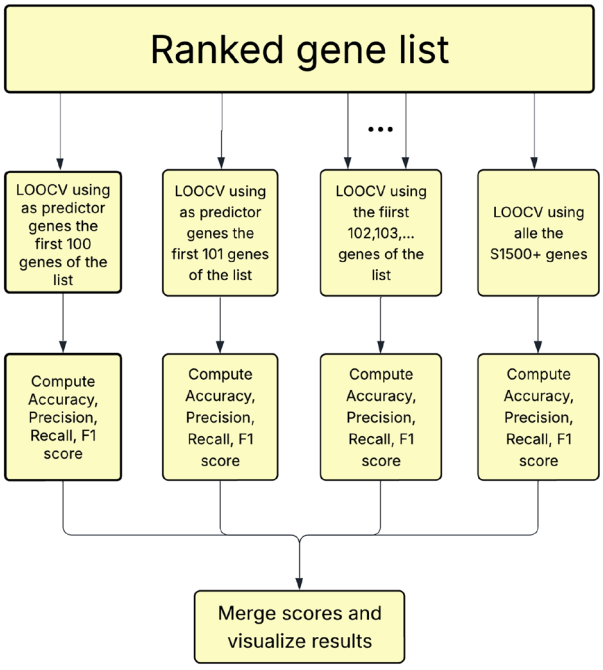

Fig. 7 Flow chart of our gene addition procedure. Starting from the S1500+ gene list ranked using the feature importances we perform and evaluate a LOOCV on our dataset using the first 100, 101, 102, ... genes of our list

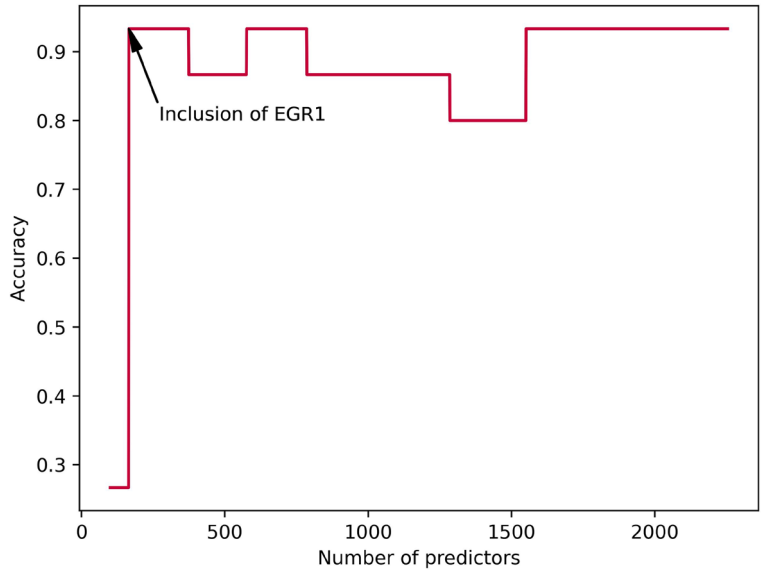

Fig. 8 Pain classification accuracy

we modified the parameter scale_pos_weight of XGBoost as described in the XGBoost documentation to address the class imbalance. We report accuracies in Fig. 8 and also in a section in the appendix.

Notice also, that for this part, our training method is substantially different than the one used in the previous section. It is also important to take into account that our database is not balanced and contains only 25% of no pain labels. Consequently, the accuracy achieved by a weighted random guess is 62.5%, while the accuracy achieved by a naive majority-class prediction is 75%. We can use these two accuracies as a baseline for our evaluations, [25]. Indeed, the accuracy of a binary classifier on an unbalanced set is obtained by the formula

$$\text{accuracy} = P(\text{class} = 0) * P(\text{prediction} = 0) + P(\text{class} = 1) * P(\text{prediction} = 1) = 0.62$$

since in our case,

$$P(\text{class} = 0) = P(\text{prediction} = 0) = 0.75$$

and

$$P(\text{class} = 1) = P(\text{prediction} = 1) = 0.25$$

This is the accuracy obtained by a weighted guess on a data with the given imbalance (0.75 and 0.25) and we take it as baseline for our reasoning on accuracies.

Figure 8 represents the accuracy of the model on the y axis and the number of predictors on the x axis. We also report in our appendix more graphs regarding other metrics for our result.

We notice the following remarkable fact: we reach maximum accuracy, once we include the gene EGR1 (ranked as the 164^{th} most important regressor for the brain tissue). Indeed, there is a growing consensus [26] on the fact the EGR1 gene may play a critical role in the development of chronic pain. In both animal and human studies [16], EGR1 has been found to be upregulated in pain-processing brain regions and the spinal

cord following nerve injury or inflammation [27–29]. Despite the fact we cannot claim EGR1 to be a definite biomarker for chronic pain based solely on our study, our method supports the conclusions and it is in accordance with the findings of the above mentioned bibliography.

Hence, our machine learning method demonstrates merit in the following ways:

- It successfully predicts pain versus no pain by analyzing a small subset of genes, thereby confirming the findings of [15], which, although limited in statistical power, address a highly relevant topic.
- It suggests a potential correlation between the EGR1 gene and the presence of chronic pain in a limited number of patients, as also indicated in [15].

Although such a small sample size cannot provide statistically significant results, our gene-ranking approach offers valuable insights into potentially relevant genes for the biomedical question and calls for a further investigation on this topic with larger and more statistically balanced datasets. Furthermore, it corroborates the findings of other studies [15] based on the same limited dataset, thereby supporting the potential for further investigation using similarly reduced datasets.

4 Conclusions

We show that the XGBoost algorithm can be effectively trained to reconstruct with accuracy comparable to the state of the art algorithms, the full transcriptome gene expression in selected tissues, using the information relative to our selection of top 500 predictor genes, ranked via the importance feature in XGBoost. The ranking of the genes and the accuracies are strictly tissue dependent. By exploiting the information regarding the order of importance of the top 500 predictor genes from the brain tissue, we examine a pain versus no-pain labelled dataset. Our method is highlighting the EGR1 gene as a possible indicator for pain versus no-pain conditions, hence confirming preliminary in vitro studies with our in silico methods. The pain case study exemplifies how our ranking based on gene importance in XGBoost can drive hypotheses (e.g. EGR1 related to pain mechanism), however larger validations are essential to confirm our proof-of-concept method.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s42452-025-07841-1.

Supplementary file 1.

Acknowledgements
We thank the anonymous Referees for comments that greatly helped us clarify the results in our paper.

Author contributions
PD, JB, IDS wrote the software. RF, FZ, GP, AG, BK, KC reviewed the manuscript and discussed the main concept. RF, PD wrote the manuscript.

Funding
This research was funded by CaLISTA CA 21109, CaLIGOLA MSCA-2021-SE-01- 101086123, MSCA-DN CaLiForNIA - 101119552, PNRR MNESYS, PNRR National Center for HPC, Big Data and Quantum Computing, SimQuSec, INFN Sezione Bologna.

Data availability
The datasets used and/or analysed during the current study available https://www.gtexportal.org/home/downloads/adult-gtex. The algorithms and their implementations are available upon request to the authors.

Declarations

Ethics approval and consent to participate
All authors declare that this work complies with ethical guidelines set by the journal. Each author has approved of and agreed to participate in the article.

Consent for publication
Each author approves and agrees to submit and publish this article.

Competing interests
The authors declare no Conflict of interest. The funders (EU) had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Received: 25 July 2025 / Accepted: 29 September 2025
Published online: 07 November 2025

References

1. Wang Z, Gerstein MB, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
2. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. Nucleic Acids Res. 2012;40:10084–97.
3. Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, Collins JB, et al. A hybrid gene selection approach to create the s1500+ targeted gene sets for use in high-throughput transcriptomics. PLoS ONE. 2018;13:e0191105. https://doi.org/10.1371/journal.pone.0191105.
4. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. bioRxiv. 2015.
5. Li W, Yin Y, Quan X, Zhang H. Gene expression value prediction based on xgboost algorithm. Front Genet. 2019;10:1077. https://doi.org/10.3389/fgene.2019.01077.
6. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med. 2022;133:102423. https://doi.org/10.1016/j.artmed.2022.102423.
7. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining, 2016: 785–794. https://doi.org/10.1145/2939672.2939785.
8. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
9. Chandragandhi S, Srihari K, Arvind C. Advanced predictive disease modeling in biomedical IoT using the temporal adaptive neural evolutionary algorithm. Sci Rep. 2025. https://doi.org/10.1038/s41598-025-08426-z.
10. Shah SM, Chandrasekar V, Aboumarzouk O, Singh A, Dakua S. Leveraging machine and deep learning algorithms for herg blocker prediction. IEEE Access. 2025. https://doi.org/10.1109/ACCESS.2025.3566440.
11. Morgoeva A, Morgoev I, Klyuev R, Kochkovskaya S. Forecasting hourly electricity generation by a solar power plant using machine learning algorithms. Bull Tomsk Polytech Univ Geo Assets Eng. 2023;334:7–19. https://doi.org/10.18799/24131830/2023/12/4253.
12. Ai X, Smith MC, Feltus FA. Generative adversarial networks applied to gene expression analysis: an interdisciplinary perspective. Comput Syst Oncol. 2023;3(3):e1050. https://doi.org/10.1002/cso2.1050.
13. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (gtex) project. Nat Genet. 2013;45(6):580–5. https://doi.org/10.1038/ng.2653.
14. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60. https://doi.org/10.1126/science.1262110.
15. North RY, Li Y, Ray PR, Rhines LD, Tatsui CE, Rao G, et al. Electrophysiological and transcriptomic correlates of neuropathic pain in human dorsal root ganglion neurons. Brain J Neurol. 2019;142(5):1215–26. https://doi.org/10.1093/brain/awz063.
16. Ko SW, Vadakkan KI, Ao H, Gallitano-Mendel A, Wei F, Milbrandt JD, et al. Selective contribution of egr1 (zif/268) to persistent inflammatory pain. J Pain Off J Am Pain Soc. 2005;6(1):12–20. https://doi.org/10.1016/j.jpain.2004.10.001.
17. Sonkin D, Thomas A, Teicher B. Cancer treatments: past, present, and future. https://doi.org/10.20944/preprints202401.1989.v1.
18. Liu H, Guo Z, Wang P. Genetic expression in cancer research: challenges and complexity. Gene Reports. 2024;37:102042. https://doi.org/10.1016/j.genrep.2024.102042.
19. Liu H, Li Y, Karsidag M, Tu T, Wang P. Technical and biological biases in bulk transcriptomic data mining for cancer research. J Cancer. 2025;16:34–43. https://doi.org/10.7150/jca.100922.
20. El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, et al. Whole-genome sequencing in health care recommendations of the European society of human genetics. Eur J Hum Genet EJHG. 2013;21(Suppl 1):1–5.
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
22. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004.
23. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? 2022. https://arxiv.org/abs/2207.08815.
24. Florek P, Zagdański A. Benchmarking state-of-the-art gradient boosting algorithms for classification. 2023. https://arxiv.org/abs/2305.17094.
25. Bertsekas DP, Tsitsiklis JN. Introduction to probability. 2002.
26. Hecke O, Austin SK, Khan RA, Smith BH, Torrance N. Neuropathic pain in the general population: a systematic review of epidemiological studies. PAIN. 2014;155:654–62. https://doi.org/10.1016/j.pain.2013.11.013.

27. Jiang MEA. Microrna-124-3p attenuates the development of nerve injury-induced neuropathic pain by targeting early growth response 1 in the dorsal root ganglia and spinal dorsal horn. J Neurochem. 2021;158:928–42. https://doi.org/10.1111/jnc.15433.
28. Xie J, Herr SA, Ma D, Wu S, Zhao H, Sun S, et al. Acute transcriptomic and epigenetic alterations at t12 after rat t10 spinal cord contusive injury. Mol Neurobiol. 2023;60:2937–53.
29. Wang K, Wang S, Chen Y, Wu D, Hu X, Lu Y, et al. Publisher correction: single-cell transcriptomic analysis of somatosensory neurons uncovers temporal development of neuropathic pain. Cell Res. 2021;31:939–40.

Publisher's Note

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com