Repeated Decision Stumping Distils Simple Rules from Single Cell Data

Ivan A. Croydon Veleslavov & Michael P.H. Stumpf^{1,2}

¹ Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

² School of BioSciences and School of Mathematics and Statistics, University of Melbourne, Parkville VIC 3010, Australia
mstumpf@unimelb.edu.au

Abstract

Here we introduce repeated decision stumping, to distill simple models from single cell data. We develop decision trees of depth one – hence ‘stumps’ – to identify in an inductive manner, gene products involved in driving cell fate transitions, and in applications to published data we are able to discover the key-players involved in these processes in an unbiased manner without prior knowledge. The approach is computationally efficient, has remarkable predictive power, and yields robust and statistically stable predictors: the same set of candidates is generated by applying the algorithm to different subsamples of the data.

Background

The development, testing, and iterative improvements of models and hypotheses are cornerstones of scientific practice. There is, however, an inevitable trade-off between predictive power and interpretability in any modelling of real-world data and processes[1]. This is especially true for deep neural networks or reinforcement learning which are capable of capturing complex and non-linear relationships among large training data. But understanding the basis of these predictions – or distilling mechanistic insights from them – has often proved hard[2]. While recent applications to single-cell data have demonstrated the predictive power of these methods[3, 4], applications of machine and statistical learning to this type of data are also revealing how difficult it is to gain concrete mechanistic biological insights from these black-box approaches.

A model with high predictive performance in assigning cell types, for example, may in some contexts be less useful or interesting than a model with worse predictive performance that instead offers some mechanistic insights as to why a certain cell type is realised[5, 1]. Even a small insight, a “rule of thumb” or engineering heuristic, of what drives a cell-fate transition[6, 7, 8, 9, 10] can provide useful and interpretable insights into the fundamental biological processes. Distilling such simple rules from complex data is our primary motivation here.

Single cell technologies have opened up new avenues for the analysis of cellular processes[11, 12, 13, 2]. Importantly, they have revealed widespread cell-to-cell variability of apparently phenotypically identical cells at the molecular level. Transcriptomic analyses account for the majority of single cell data, and single cell transcriptomics has become widely deployed in e.g. stem cell and developmental biology, but also in immunology and cancer research. Changes at the transcriptome level are widely seen as indicative of, and even causal for, cell fate decision making and progression of cells along the differentiation path[14, 15].

Figure 1: Identifying informative cell fate markers from single-cell transcriptomic data using Repeated Decision Stumping. Cells sampled across a developmental boundary are profiled and decision stump models generated to classify cell stages, using the measured expression values for each gene as potential thresholds. The resulting decision stumps are then ranked by the information gain associated with dividing the set of classes along the corresponding boundary, reporting the topmost informative genes.

The riches contained in these data, do not reveal themselves easily, however, as a host of problems plague analysis and interpretation[12, 2]: experimental (undesirable) noise interferes with the molecular noise (which holds signatures of molecular processes); the high dimensionality of the data poses considerable statistical challenges. Models that learn to make robust predictions for cell fates under such noise carry clear motivation[16]. Further, if we are able to design these models to also provide basic biological insight, we will have a pragmatic tool, based on decision stumps[17, 18, 19], with which to (i) study high-dimensional developmental dynamics; or (ii) develop engineering strategies e.g. to control cellular behaviour [20, 21, 22].

Here we introduce a new algorithm, *Repeated Decision Stumping* (ReDX), that exploits the simple structure of decision tree models to distill highly interpretable but still predictive insights from single cell data: single genes or sets of genes that show particular statistical association with developmental events. This approach complements traditional studies that e.g. follow sets of genes in a hypothesis driven manner; and data-driven approaches that approach transcriptomic data without a specific hypothesis. ReDX sits between these two types of algorithms and allows us to generate new interpretable hypotheses and mechanistic models in a data-driven framework.

Results

In the following, we demonstrate the effectiveness and pragmatism of this approach in identifying testable hypotheses for a range of developmental contexts. We are particularly interested in finding genes with expression patterns that align well with changes in cell state. These transitions are, of course, initiated and orchestrated by gene regulatory interactions [23, 11, 12]; but identification of genes that show a particularly strong association with a cell fate decision making process or transition can help with improving our mechanistic understanding, and the development or improvement of models describing such processes. We show

Table 1: Results of ReDX applied to murine neuronal cell fate commitment, with 5 iterations. The feature name and chosen threshold is presented, alongside each model's performance on withheld data. The direction of the prediction is also provided, with respect to the established lineage $ES \rightarrow EPI \rightarrow NPC$.

Transition	Rank	Name	Threshold	Accuracy	F1	Progress
ESC→EPI	1	Klf4	10.94	0.9	0.79	$X \leq T$
	2	Esrrb	10.69	0.97	0.93	$X \leq T$
	3	Zfp42	13.96	0.93	0.88	$X \leq T$
	4	Nr5a2	8.99	0.9	0.79	$X \leq T$
	5	Tbx3	10.97	0.9	0.77	$X \leq T$
EPI→NPC	1	Gdf3	7.78	0.87	0.86	$X \leq T$
	2	Pou5f1	7.33	0.9	0.9	$X \leq T$
	3	Fgf4	1.9	0.93	0.92	$X \leq T$
	4	Dnmt3b	16.35	0.88	0.86	$X \leq T$
	5	Dppa4	9.35	0.77	0.77	$X \leq T$

that the models learnt by ReDX exhibit high predictive performance on withheld test data; furthermore the simplicity of the model structures allows us to obtain interpretable insights into the complex developmental dynamics. We also devise a metric, γ , for quantifying the stability [24, 25] of this feature selection under random permutations of the training data, and use it to demonstrate the robustness of the resulting predictions.

Identifying genes associated with mouse neuronal precursors

We first apply ReDX to the Stumpf *et al.* [26, 23] data on murine neural differentiation. Here mouse embryonic stem (ES) cells were profiled using RT-q-PCR along their commitment to neuronal precursor (NP) cells, via an epiblast-like intermediate cell stage (EPI). For each successive developmental boundary captured in the data ($ES \rightarrow EPI$; $EPI \rightarrow NP$), we apply ReDX to identify the topmost informative genes for discriminating between the cell classes.

The results show that ReDX is able to learn simple models that characterise each transition well - in terms of predictive performance on the withheld test data - whilst incorporating genes already implicated in murine neuronal commitment [Table 1].

For example, for the $ESC \rightarrow EPI$ transition, ReDX identifies *Klf4*, *Esrrb*, *Zfp42*, *Nr5a2* and *Tbx3* as the genes best able to discriminate between the cell classes; all of which have been shown to regulate naive pluripotency. After training, the learnt model identifies that thresholding for *Klf4* at 10.94 splits the cell classes such that samples whose *Klf4* expression satisfies $Klf4 \leq 10.94$ are classified as embryonic stem cells and those for which $Klf4 > 10.94$ are classified as epiblast like cells. The directionality learnt by the model is consistent with our understanding of *Klf4*'s role in murine neuronal development, whereby it regulates key transcription factors during embryonic development and prevents the differentiation of embryonic stem cell populations [27]. *Esrrb* is known to prevent the transition from ES cell to epiblast like cells through co-recruitment of *Sox2* and downstream positive regulation of NROB1 inhibiting the epi-blast like transcriptional program [28]. This supports the features identified by RedX for characterising boundaries between cell types, and highlights the pragmatism that the simple, interpretable model structure offers for hypothesis generation.

Across both transitions, we find that the simple decision stump models exhibit high predictive performance on withheld test data. We observe a mean classification accuracy of 92% across the top 5 ranked decision stumps for the $ES \rightarrow EPI$ boundary and 87% on the $EPI \rightarrow NP$ boundary. As detailed below, section , ReDX ranks the candidate decision stumps by their prospective information gain, i.e. how effectively they segregate classes. An alternative approach would be to simply rank models using performance metrics such as the misclassification rate, evaluated on some withheld test or validation set. However, ranking by information gain alone is more intuitive and eliminates concerns with constructing a representative validation set, to which the models might otherwise overfit. As expected, we observe that high ranking models (those with high information gain) also have high predictive performance on withheld test data.

To this end, we demonstrate our method's ability to learn simple, interpretable models from single-cell gene expression data. Further, we emphasise the utility in selecting maximally discriminant features for differentiating (in the common sense) between cell types – without the need for prior expert information e.g. feature

Figure 2: Repeated Decision Stumping identifies top features and associated thresholds for cell fate choice. a. Top five features identified for each boundary in the linear trajectory of murine neural cell fate commitment, **b.** Top five features identified for branching trajectories in early *Xenopus tropicalis* embryogenesis.

importance or underlying noise distributions. The low-dimensional models learnt in this approach offer a framework from which we can begin to learn aspects of developmental dynamics from transcriptomic data.

Characterising early embryogenesis in *Xenopus tropicalis*

In addition to linear developmental processes, we are also able to characterise *branching* cell fate decisions [15] using RedX. We illustrate this in the context of early *Xenopus tropicalis* embryogenesis [29]. Here ‘Stage O8 blastula’ cells commit to one of six more specialised ‘Stage 10’ progenitors {‘Stage 10 neuroectoderm’, ‘non-neural ectoderm’, ‘marginal zone’, ‘Spemann organizer (mesoderm)’, ‘Spemann organizer (endoderm)’ and ‘endoderm’}. Applying RedX, we are able to identify the topmost informative features and thresholds for each developmental trajectory [Table 2].

Table 2: Repeated Decision Stumping characterises cell fate commitment of stage 08 *Xenopus tropicalis* blastula cells to each of the known stage 10 progenitor cell types.

Transition	Rank	Name	Threshold	Accuracy	F1	Progress
So8-blastula→S10-neuroectoderm	1	<i>Ptmap12</i>	19.23	0.93	0.9	$X > T$
	2	<i>Eef1a1</i>	12.12	0.89	0.83	$X > T$
	3	<i>Cirbp</i>	23.37	0.89	0.83	$X > T$
	4	<i>Gsk3b</i>	158.96	0.85	0.77	$X \leq T$
	5	<i>Sox2</i>	2.21	0.84	0.77	$X > T$
So8-blastula→S10-non-neural ectoderm	1	<i>Ptmap12</i>	17.51	0.94	0.9	$X > T$
	2	<i>Eef1a1</i>	12.06	0.89	0.83	$X > T$
	3	<i>Cirbp</i>	21.39	0.87	0.79	$X > T$
	4	<i>Muc5b</i>	4.69	0.89	0.85	$X > T$
	5	<i>Hnrnpa1</i>	8.76	0.88	0.81	$X > T$
So8-blastula→S10-marginal zone	1	<i>Eef1a1</i>	13.42	0.85	0.79	$X > T$
	2	<i>Gsk3b</i>	147.08	0.84	0.81	$X \leq T$
	3	<i>Ccnb1</i>	0.69	0.83	0.73	$X \leq T$
	4	<i>Cirbp</i>	21.27	0.82	0.75	$X > T$
	5	<i>Fth1b</i>	1.87	0.76	0.7	$X > T$
So8-blastula→S10-Spemann organizer (mesoderm)	1	<i>Chrd1</i>	6.42	0.95	0.97	$X > T$
	2	<i>Lhx1</i>	3.73	0.96	0.98	$X > T$
	3	<i>Opn7b</i>	0.81	0.95	0.97	$X > T$
	4	<i>Eef1a1</i>	16.39	0.94	0.96	$X > T$
	5	<i>Gsk3b</i>	141.24	0.91	0.95	$X \leq T$
So8-blastula→S10-Spemann organizer (endoderm)	1	<i>Frzb2</i>	1.43	0.94	0.96	$X > T$
	2	<i>Gsk3b</i>	132.51	0.91	0.94	$X \leq T$
	3	<i>Frzb1</i>	2.61	0.88	0.92	$X > T$
	4	<i>Eef1a1</i>	15.55	0.91	0.94	$X > T$
	5	<i>Opn7b</i>	1.82	0.87	0.92	$X > T$
So8-blastula→S10-endoderm	1	<i>Frzb2</i>	3.39	0.97	0.99	$X > T$
	2	<i>Frzb1</i>	8.07	0.96	0.98	$X > T$
	3	<i>Atp1a1</i>	9.5	0.97	0.98	$X > T$
	4	<i>Tmem150b</i>	2.45	0.98	0.99	$X > T$
	5	<i>Sox17a</i>	5.61	0.97	0.98	$X > T$

As with the murine results, models trained on the *Xenopus* data and ranked by information gain perform well on withheld test data. The learnt decision stumps show high mean predictive performance across 5-fold cross-validation on unseen test data, with the top models averaging 92.5% accuracy. While the performance of the highest ranked models is generally high, for some boundaries the drop-off in performance - even within the top 5 ranked models - is more pronounced. The stage 08 blastula to stage 10 marginal zone boundary, for instance, remains well characterised by *EEF1A1* (86% accuracy), but by rank 5, thresholding for *FTH1B* classifies the withheld data with only 76% accuracy. This may indicate that only a few genes robustly drive (or reflect at the level of mRNA) cell fate commitment; perhaps even more likely is that such boundaries between some cell-types are more complex than those between others. As the suggested genes are derived directly from classifiers learnt from data, we can use the model performance on unseen data to inform our belief in a gene's involvement in a given process.

Looking for candidate developmental genes, we see that both the Neuroectoderm and Non-neural ectoderm branches returned *ptmap12* as the most informative gene (albeit with varying thresholds); this gene has long been known to be involved in *Xenopus* development [30, 31]. Thus a role in this transition is not unexpected. The overlap in discriminatory genes between the two branches, extending to *EEF1A1* - an elongation factor with known developmental roles in different organisms - and *CIRBP* - required for gastrulation, cell migration during embryogenesis, and neural development in *Xenopus* - suggests shared, or partially shared, regulatory mechanisms for the commitment of stage 9 blastula cells to stage 10 neuroectoderm and non-neural ectoderm cell types. From previous studies[32, 33], the involvement of *gsk3b*, in several transitions is also not surprising. It appears to be required at high mRNA levels in the stage 08 blastula. and a decrease in expression is observed in several S10 stages. Because of its complex interactions[34], *GSK3B*, can attain mul-

tiply stable states [35], but here (see Table 2 and Figure 2), at the transcriptional level only two states can be discerned. Similarly, *frzb1*, *frzb2*, and *sox17a* show strong prior association with developmental processes and embryogenesis. The approach developed here yields these targets in a purely inductive, data-driven, manner without any further intervention or the requirement for domain expertise.

Feature rankings are stable under random training-test partitions

We also demonstrate the robustness of predictions made by ReDX to perturbations in, or systematic problems with, the training data. To do so, we devise a stability metric[25], which describes how consistently and how highly a given feature is ranked over many random partitions of the training data (see methods).

We calculate this gamma score for each feature over 100 random splits of the data into 80% training and 20% test partitions. Sorting the features by γ shows that ReDX produces consistent gene rankings when trained on these different portions of the data [Figure 3]. Moreover, the feature reproduction is stable[24, 25], and approaches consistent reproduction across the random partitions. Together, these findings increase our confidence in the quality of the embedded feature selection for discerning developmental boundaries from such data.

Conclusions

We have developed ReDX as an approach that allows us to learn *interpretable* models from high-dimensional and potentially complex data, such as current single-cell data. It is aimed to augment the impressive state-of-the-art machine learning approaches[36], such as deep neural networks and reinforcement learning, by focusing on the generation of deliberately simple mechanistic hypotheses[37, 26]. Despite the focus on informative classifiers, the features learned by ReDX also retain high predictive power. This makes them immediately applicable for informing mechanistic model development, and for generating interpretable and testable hypotheses about e.g. the underlying developmental processes. The opportunities presented by ReDX are two-fold: to generate predictive models from data (which themselves can be used to annotate unlabelled samples, or identify potentially mislabelled samples); and to provide an effective, scalable means of feature selection for guiding further, (including hypothesis-driven) experimental design and mechanistic modelling.

One common critique[18] of decision trees and stumps is their instability to variations in the training data. To address this concern, we devised a quantitative stability measure, γ [24, 25]. Calculating it over random partitions of the training data, we are able to show that ReDX generates consistent and stable feature rankings for our developmental case studies. Whilst this may not be true of data with highly intra-heterogeneous classes, γ scores provide us with an objective and intuitive means of assessing the quality of predictions made using ReDX in new contexts with little available prior information.

Whilst each learnt model is simple and 1-dimensional, our findings demonstrate their ability to segregate cells along diverse developmental boundaries with great precision. Further, as a pragmatic method for gaining insights, maybe even intuition, about a high-dimensional system – all without prior expert knowledge – the merits of such an approach seem clear, especially if deployed alongside other, predictive data-driven methods.

ReDX can be applied in many contexts where labelled, single-cell data are available. Our method robustly proposes features and thresholds that optimally discriminate between the target labels (here cell states). We demonstrate the approach’s effectiveness for characterising cell-stage boundaries during healthy ontogenesis across linear and branching trajectories. The method can also be applied to other, e.g. multi-omics[38, 39] without further modification (proteomic, metabolic and transcriptomic) and carries promise in biological contexts, where suitable (labelled) data and features are available but for which mechanistic models remain elusive. We can also apply ReDX to discern discriminant features in healthy v.s. diseased samples, or between control and perturbation conditions given appropriate experimental design.

The simple candidate mechanisms generated by ReDX will then serve as seeds for developing better mechanistic models.

Figure 3: Robustness of ReDX feature rankings over 100 random partitions of the MacArthur data set into 80% training and 20% test. Features reported along the x-axis in decreasing γ . The ‘Naive’ line shows the expected γ scores if ReDX was ranking features randomly at each partition. The ‘Unpeturbed’ line shows how γ would decay if ReDX returned identical feature rankings from all 100 partitions.

Methods

Learning genes associated with developmental processes from single-cell data

Decision trees are a commonly used class of predictive statistical model capable of learning decision boundaries from labelled data[40, 41, 42, 43]. One salient characteristic of these models is their inbuilt capacity for feature selection during model training. Given high-dimensional input, a low-dimensional model is returned; and they are capable of classifying samples according to a systematically reduced subset of features. We can exploit the structure learnt by these models to not only make predictions for unclassified samples, but also to better understand the boundaries between cell classes, and to gain meaningful biological insight[11, 14].

Consider a single cell, i , whose expression profile for J genes is given by the vector $\mathbf{x}_i$. Each element of the J -dimensional vector $\mathbf{x}$ represents the gene expression value of the j th gene: $\mathbf{x} = (x_1, x_2, \dots, x_{n-1}, x_n)$. For N such cells, consider the matrix X , such that $X_{i,j}$ is similarly the expression of gene j measured in cell i . The developmental stage of each cell is given by $\mathbf{y}$, such that y_i is the class label of the i th cell.

The first bifurcation of a decision tree splits these N samples into two groups, q_1 and q_2 , using a feature, j and threshold t , such that $j \leq t$ for all samples in q_1 and $j > t$ for all samples in q_2 ; these trees are referred to as *decision stumps*[17, 40]. We wish to identify the best gene and corresponding expression threshold, j^* and t^* , respectively, with which to divide the samples such that class labels in the resulting groups are maximally segregated. If we are able to learn informative feature-threshold pairs from data, we will can gain insights into the underlying developmental dynamics of our system.

To quantify the reduction in entropy over the class labels $\mathbf{y}$ following a split at the proposed gene and threshold pair (j, t) we use the *information gain* (IG) measure,

$$\text{IG}(\mathbf{y}, j, t) = H(\mathbf{y}) - \sum_{q \in Q} p(q)H(q), \quad (1)$$

where $H(\mathbf{y})$ is the entropy associated with the class labels $\mathbf{y}$, Q is the set of subsets q proposed by splitting $\mathbf{y}$ on the feature j and threshold t , which thus satisfies $\mathbf{y} = \bigcup_{q \in Q} q$, $p(q)$ is the fraction of elements contained in the subset q compared to the set $\mathbf{y}$ and $H(q)$ is the entropy over the class labels contained in subset q (see methods for details).

Following established procedure for growing decision trees [43] we calculate IG over all possible thresholds for each gene ($t_j \in X_j$), and select the gene and threshold for which IG is maximised (j^* and t^* respectively); optimising over all pairs (j, t) according to

$$(j^*, t^*) = \arg \min_{j \in \{1, \dots, J\}} \min_{t \in T_j} IG(\{x_i, y_i : x_{ij} \leq t\}) + cost(\{x_i, y_i : x_{ij} > t\}), \quad (2)$$

where T_j is the set of possible thresholds for feature j obtained by sorting the unique values of X_j .

This first bifurcation of a decision tree represents the most informative axis-parallel split in the feature data for segregating the set of class labels. Intuitively, this implicit feature selection can be exploited to identify those features to which the classification problem is most sensitive.

While traditional decision tree algorithms [43], such as ID3 and C4.5, recurse on the resultant leaf nodes to grow a full decision tree [44], here we interrupt the process after the first node is grown. The feature associated with the resulting *decision stump* [17] is recorded and removed from the set of available features. A new stump grown on the *full* set of samples – *excluding* the previously selected feature – thus identifies the second most informative feature/threshold pair with respect to the cell labels. This process can be repeated n times to identify the top n most discriminatory features associated with a binary classification problem. This forms the basis of our approach, for which a visual summary is provided in Figure 1.

Learning and ranking decision stumps

To learn each decision stump, we identify the best feature, j^* , and best threshold for that feature, t^* , from labelled single-cell gene expression data. To do so, we follow the established procedure for growing a decision tree [43, 41] and optimise over all pairs (j, t) according to

$$(j^*, t^*) = \arg \min_{j \in \{1, \dots, D\}} \min_{t \in T_j} cost(\{x_i, y_i : x_{ij} \leq t\}) + cost(\{x_i, y_i : x_{ij} > t\}), \quad (3)$$

where T_j is the set of possible thresholds for feature j obtained by sorting the unique values of x_{ij} . The cost function corresponds to a chosen impurity measure used to quantify the quality of a split with regards to purity of the daughter node labels. In our case, we use information gain (see below, section).

For continuous features, such as the gene expression values derived by qPCR, we consider thresholds at values for that feature already present in the training data. For each feature j , the cost function is calculated for all possible thresholds $t \in T_j$, corresponding to the unique expression values for gene j across all the cell samples in the training data. As shown in equation 2, the best attribute to split on is the gene and expression threshold (j^*, t^*) that maximises label purity over the proposed daughter nodes according to the chosen cost function.

Information Gain

Information Gain is a measure of the reduction in entropy [45, 46] in a data set given some transformation. In our case, we calculate the information gain in the set of class labels, S , with respect to a candidate split, A , according to

$$IG(A, S) = H(S) - \sum_{q \in Q} p(q)H(q), \quad (4)$$

where $H(S)$ is the entropy of the parent set S , Q is the set of subsets q proposed by splitting S on attribute A which thus satisfies $S = \bigcup_{q \in Q} q$, $p(q)$ is the proportion of elements in the subset q compared to the set S and $H(q)$ is the entropy over the subset q . For clarity, the entropy of any set S containing C unique classes is

$$H(S) = \sum_{c=1}^C -p(c) \log(p(c)). \quad (5)$$

Model evaluation

Misclassification rate and accuracy

After training our models, we have to evaluate the quality of their predictions on unseen data. Principally, we quantify the difference between the predicted labels and true (hidden) labels using the misclassification rate and accuracy. Let $\hat{y}_i$ be the class prediction for the i th sample in a test set of n samples. Then the test set misclassification rate is simply the fraction of misclassified cases, given by,

$$r = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i \neq \hat{y}_i), \quad (6)$$

where $\mathbb{I}$ is the indicator function such that

$$\mathbb{I}(y_i \neq \hat{y}_i) = \begin{cases} 1, & \text{if } y_i \neq \hat{y}_i \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Following equation (6), the model's accuracy on the test data is given by

$$a = 1 - r. \quad (8)$$

F_1 score

Whilst accuracy defined in this way provides an immediately intuitive single-number evaluation metric, it fails to take into account unequal label frequencies in the test data. For instance, consider a binary classifier which simply returns label *A* every time, regardless of the input features. Using accuracy as the sole metric, we might end up considering this classifier as highly effective if the test data labels consisted of 99% label *A* and 1% label *B* (resulting in a model accuracy of 99%). However, a model which completely ignores its input features will not generalise well to new data, and should intuitively be avoided.

To address this, we also report the F_1 score for each learnt model [43]. This is defined as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (9)$$

where precision is $\frac{TP}{TP+FP}$ and recall is $\frac{TP}{TP+FN}$.

More generally, the F_1 score is a specific instance of F_β scores

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}, \quad (10)$$

where β is real-valued, positive and chosen such to make recall is β times as important than precision. In our use cases, we have no reason to attribute more weight to either recall and precision, and report the F_1 score as standard ($\beta = 1$).

Feature stability analysis

One established limitation of decision trees is their instability [40]. Small variation in training data can affect the topology of the resulting tree, as well as affecting which features are selected for the discrimination of the same target variable. Decision stumps, and the order in which they are selected in ReDX, can also be affected by small changes in the training data. This is problematic if we are aiming to propose a list of the top-most informative genes to act as biomarkers, or to direct experimental and mechanistic modelling effort.

To investigate and quantify the stability of features proposed by ReDX in the context of single-cell data, we devise an associated method for calculating feature stability, following Kirk *et al.* [25], over random partitions

of the training data. This stability metric, γ , is calculated for each feature in order to quantify how consistently genes are ranked given perturbed training data. Given a gene of interest, for N random partitions of the data and q top features for each iteration of ReDX,

$$\gamma = \frac{1}{Nq} \sum_{n=1}^N \max(0, (q+1) - r_n), \quad (11)$$

where r_n is the rank of that gene at the n^{th} partition.

As such, γ is confined to $\{\gamma \in \mathbb{R} : 0 \leq \gamma \leq 1\}$. A value of 1 would mean that the gene was selected as the most informative over every random partition of the data and a value of 0 indicates that the gene was never selected to be in the top q stumps over the random partitions. A value of 0.5 might indicate that the feature was selected in half of the partitions at rank 1 and unranked in the other half, or the feature may have been chosen for the middle rank in all of the random partitions, or some equivalent thereof.

Datasets and preprocessing

Linear Differentiation Pathway

The first data set contains single-cell gene expression data for mouse embryonic stem cells undergoing differentiation to neural progenitor cells [26]. In the initial study, the mouse embryonic stem cells in the E14tg2a and R1 embryonic mouse cell lines were profiled during cell fate commitment, identifying a previously unknown epiblast like, intermediary cell state.

In total, 547 cells were sampled over the course of seven days; at 24h, 48h, 72h, 120h and 168h after transfer from leukemia inhibitory factor (LIF) + 2i conditions to N2B27 neural basal medium; following existing protocol [47]. Gene expression was reported for 96 genes affiliated with ES cell development using a high-throughput RT-PCR array. The resulting 547x96 feature matrix was supplemented by additional annotations provided in the initial work for each cell sampled. These included the sampling time for each cell, alongside a classification label for the cell; “ESC”, “EPI” or “NPC”; representing populations of embryonic stem cells, epiblast-like intermediate cells and neural progenitor cells respectively. As discussed in the supplementary material of the original work, these cell class labels were assigned through k-means clustering. Together with the gene expression data, these annotations provide us with a gold standard reference with which to train the models in our own work and to evaluate their predictive performance.

Branching Differentiation Pathway

The second data is derived from a larger high-throughput sequencing effort, focused on early embryo development in *Xenopus tropicalis* [29]. A total of 37,136 cells were sampled from zygotic genome activation (stage 8, 5 hours postfertilization) through to early organogenesis (stage 22, 22 hours postfertilization). Single cells were profiled with sc-RNAseq, following a previously established high-throughput single-cell droplet barcoding pipeline [48]. The resulting data describes the underlying scRNA-seq counts for 25,087 genes in each of the sampled cells. Cell stage annotations were derived in the original work, using hierarchical clustering on the scRNA-seq data and cluster specific genes used to draw cluster labels from known embryonic cell types documented in the Xenopus Bioinformatics database – Xenbase [49].

The full scRNA-seq count data and raw FASTQ files remain publicly accessible at tinyurl.com/scXen2018.

Availability of data and materials

The datasets analysed during the current study and the associated code are available in the RepeatedDecisionStumping GitHub repository, <https://doi.org/10.5281/zenodo.4017262>

Competing interests

The authors declare that they have no competing interests.

Author's contributions

IACV and MPHS conceived the methodology; IACV developed the algorithm and performed the analysis; IACV and MPHS wrote the paper.

Acknowledgements

We gratefully acknowledge the support from the members of the *Theoretical Systems Biology Group* at Imperial College London and the University of Melbourne. IACV is funded through a Wellcome Trust PhD Studentship. MPHS gratefully acknowledges funding from the Volkswagen Foundation and the Deputy Vice Chancellor of Research Fund.

References

- [1] Huang, Sui: The Tension Between Big Data and Theory in the "Omics" Era of Biomedical Research. *Perspectives in biology and medicine* **61**(4), 472–488 (2018). doi:10.1353/pbm.2018.0058
- [2] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B.d., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.-H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J.d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A.: Eleven grand challenges in single-cell data science. *Genome biology* **21**(1), 31 (2020). doi:10.1186/s13059-020-1926-6
- [3] Grønbech, C.H., Vording, M.F., Timshel, P., Sønderby, C.K., Pers, T.H., Winther, O.: scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* (Oxford, England) (2020). doi:10.1093/bioinformatics/btaa293
- [4] Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G., Li, M.: Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature communications* **11**(1), 2338 (2020). doi:10.1038/s41467-020-15851-3
- [5] Kirk, P.D.W., Babbie, A.C., Stumpf, M.P.H.: Systems biology (un)certainties. *Science* **350**(6259), 386–388 (2015). doi:10.1126/science.aac9505
- [6] Moris, N., Pina, C., Martinez Arias, A.: Transition states and cell fate decisions in epigenetic landscapes. *Nature reviews. Genetics* **17**(11), 693–703 (2016). doi:10.1038/nrg.2016.98
- [7] Huang, S., Li, F., Zhou, J.X., Qian, H.: Processes on the emergent landscapes of biochemical reaction networks and heterogeneous cell population dynamics: differentiation in living matters. *Journal of the Royal Society Interface* **14**(130), 20170097 (2017). doi:10.1098/rsif.2017.0097
- [8] Brackston, R.D., Lakatos, E., Stumpf, M.P.H.: Transition state characteristics during cell differentiation. *PLoS computational biology* **14**(9), 1006405 (2018). doi:10.1371/journal.pcbi.1006405
- [9] Wang, S., Karikomi, M., Maclean, A.L., Nie, Q.: Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic acids research* **47**(11), 66–66 (2019). doi:10.1093/nar/gkz204

- [10] Tyson, J.J., Novak, B.: A Dynamical Paradigm for Molecular Cell Biology. *Trends in Cell Biology* **30**(7), 504–515 (2020). doi:10.1016/j.tcb.2020.04.002
- [11] Babbie, A.C., Chan, T.E., Stumpf, M.P.H.: Learning regulatory models for cell development from single cell transcriptomic data. *Current Opinion in Systems Biology* **5**, 72–81 (2017)
- [12] Stuart, T., Satija, R.: Integrative single-cell analysis. *Nature reviews. Genetics* **20**(5), 257–272 (2019)
- [13] Kiselev, V.Y., Andrews, T.S., Hemberg, M.: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature reviews. Genetics* **6**, 1 (2019). doi:10.1038/s41576-018-0088-9
- [14] Griffiths, Jonathan A, Scialdone, Antonio, Marioni, John C: Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology* **14**(4), 8046 (2018). doi:10.15252/msb.20178046
- [15] Gao, N.P., Gandrillon, O., Paldi, A., Herbach, U., bioRxiv, R.G., 2020: Universality of cell differentiation trajectories revealed by a reconstruction of transcriptional uncertainty landscapes from single-cell transcriptomic data. *bioRxiv.org*. doi:10.1101/2020.04.23.056069v1.abstract
- [16] Jetka, T., Nienaltowski, K., Filippi, S., Stumpf, M.P.H., Komorowski, M.: An information-theoretic framework for deciphering pleiotropic and noisy biochemical signaling. *Nature communications* **9**(1), 4591 (2018). doi:10.1038/s41467-018-07085-1
- [17] Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.* **11**, 63–91 (1993)
- [18] Andriushchenko, M., Hein, M.: Provably Robust Boosted Decision Stumps and Trees against Adversarial Attacks. *arXiv.org* (2019). 1906.03526v2
- [19] Klusowski, J.M.: Sparse learning with CART. *arXiv.org* (2020). 2006.04266v1
- [20] Ruess, J., Parise, F., Milias-Aregetis, A., Khammash, M., Lygeros, J.: Iterative experiment design guides the characterization of a light-inducible gene expression circuit. *Proceedings of the National Academy of Sciences of the United States of America* **112**(26), 8148–8153 (2015). doi:10.1073/pnas.1423947112
- [21] Lakatos, E., Stumpf, M.P.H.: Control mechanisms for stochastic biochemical systems via computation of reachable sets. *Royal Society Open Science* **4**(8), 160790 (2017). doi:10.1098/rsos.160790
- [22] Aoki, S.K., Lillacci, G., Gupta, A., Baumschlager, A., Schweingruber, D., Khammash, M.: A universal biomolecular integral feedback controller for robust perfect adaptation. *Nature* **570**(7762), 533–537 (2019). doi:10.1038/s41586-019-1321-1
- [23] Chan, T.E., Stumpf, M.P.H., Babbie, A.C.: Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems* **5**(3), 251–2673 (2017). doi:10.1016/j.cels.2017.08.014
- [24] Meinshausen, N., Bühlmann, P.: Stability selection. *Journal of the Royal Statistical Society Series B-Methodological* **72**(4), 417–473 (2010). doi:10.1111/j.1467-9868.2010.00740.x
- [25] Kirk, P.D.W., Witkov, A., Bangham, C.R.M., Richardson, S., Lewin, A.M., Stumpf, M.P.H.: Balancing the Robustness and Predictive Performance of Biomarkers. *Journal of Computational Biology* **20**(12), 979–989 (2013). doi:10.1089/cmb.2013.0018
- [26] Stumpf, P.S., Smith, R.C.G., Lenz, M., Schuppert, A., Müller, F.-J., Babbie, A.C., Chan, T.E., Stumpf, M.P.H., Please, C.P., Howison, S.D., Arai, F., Macarthur, B.D.: Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Systems* **5**(3), 268–2827 (2017). doi:10.1016/j.cels.2017.08.009
- [27] Li, Y., McClintick, J., Zhong, L., Edenberg, H.J., Yoder, M.C., Chan, R.J.: Murine embryonic stem cell differentiation is promoted by sox-3 and inhibited by the zinc finger transcription factor klf4. *Blood* **105**(2), 635–637 (2005)
- [28] Zhou, W., Lo, S.-C., Liu, J.-H., Hannink, M., Lubahn, D.B.: *Errβ*: a potent inhibitor of nrf2 transcriptional activity. *Molecular and cellular endocrinology* **278**(1-2), 52–62 (2007)

- [29] Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., Klein, A.M.: The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **5780**, 1–16 (2018). doi:10.1126/science.aar5780
- [30] Gawantka, V., Pollet, N., Delius, H., Vingron, M., Pfister, R., Nitsch, R., Blumenstock, C., Niehrs, C.: Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mechanisms of development* **77**(2), 95–141 (1998). doi:10.1016/S0925-4773(98)00115-4
- [31] Parain, K., Mazurier, N., Bronchain, O., Borday, C., Cabochette, P., Chesneau, A., Colozza, G., El Yakoubi, W., Hamdache, J., Locker, M., Gilchrist, M.J., Pollet, N., Perron, M.: A large scale screen for neural stem cell markers in *xenopus* retina. *Dev Neurobiol* **72**(4), 491–506 (2012). doi:10.1002/dneu.20973
- [32] Pierce, S.B., Kimelman, D.: Regulation of spemann organizer formation by the intracellular kinase *xgsk-3*. *Development* **121**(3), 755–65 (1995)
- [33] Kim, K., Cho, J., Hilzinger, T.S., Nunns, H., Liu, A., Ryba, B.E., Goentoro, L.: Two-element transcriptional regulation in the canonical wnt pathway. *Curr Biol* **27**(15), 2357–23645 (2017). doi:10.1016/j.cub.2017.06.037
- [34] Peuchen, E.H., Cox, O.F., Sun, L., Hebert, A.S., Coon, J.J., Champion, M.M., Dovichi, N.J., Huber, P.W.: Phosphorylation dynamics dominate the regulated proteome during early *xenopus* development. *Sci Rep* **7**(1), 15647 (2017). doi:10.1038/s41598-017-15936-y
- [35] Harrington, H.A., Feliu, E., Wiuf, C., Stumpf, M.P.H., Stumpf, M.P.H.: Cellular compartments cause multistability and allow cells to process more information. *Biophysical journal* **104**(8), 1824–1831 (2013). doi:10.1016/j.bpj.2013.02.028
- [36] Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J.: Next-Generation Machine Learning for Biological Networks. *Cell* **173**(7), 1581–1592 (2018). doi:10.1016/j.cell.2018.05.015
- [37] Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J., Arnaud, O., Kupiec, J.-J., Espinasse, T., Gonin-Giraud, S., Gandrillon, O.: Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLoS biology* **14**(12), 1002585 (2016). doi:10.1371/journal.pbio.1002585
- [38] Bludau, Isabell, Aebersold, Ruedi: Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nature reviews. Molecular cell biology* **27**, 499–14 (2020). doi:10.1038/s41580-020-0231-2
- [39] Lederer, Alex R, La Manno, Gioele: The emergence and promise of single-cell temporal-omics approaches. *Current opinion in biotechnology* **63**, 70–78 (2020). doi:10.1016/j.copbio.2019.12.005
- [40] Dwyer, K., Holte, R.: Decision tree instability and active learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **4701 LNAI**, 128–139 (2007)
- [41] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements Of Statistical Learning: Data Mining, Inference, And Prediction*. Springer, ??? (2009)
- [42] Efron, B.: *Large-Scale Inference: Empirical Bayes Methods For Estimation, Testing, And Prediction*. Cambridge University Press, ??? (2012)
- [43] Murphy, K.P.: *Machine Learning: A Probabilistic Perspective*, p. 545. The MIT Press, ??? (2012)
- [44] Quinlan, J.R.: *Induction of decision trees*. *Machine Learning* **1**(1), 81–106 (1986). doi:10.1007/BF00116251
- [45] Cover, T.M., Thomas, J.A.: *Elements Of Information Theory*. J. Wiley, ??? (2005)
- [46] Mc Mahon, S.S., Sim, A., Sim, A., Filippi, S., Johnson, R., Liepe, J., Smith, D., Stumpf, M.P.H.: Information theory and signal transduction systems: from molecular information processing to network inference. *Seminars in Cell & Developmental Biology* **35**, 98–108 (2014). doi:10.1016/j.semcdb.2014.06.011

- [47] Ying, Q.-L., Stavridis, M., Griffiths, D., Li, M., Smith, A.: Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nature biotechnology* **21**(2), 183–186 (2003)
- [48] Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W.: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**(5), 1187–1201 (2015). doi:10.1016/j.cell.2015.04.044
- [49] Karimi, K., Fortriede, J.D., Lotay, V.S., Burns, K.A., Wang, D.Z., Fisher, M.E., Pells, T.J., James-Zorn, C., Wang, Y., Ponferrada, V.G., Chu, S., Chaturvedi, P., Zorn, A.M., Vize, P.D.: Xenbase: A genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Research* **46**(D1), 861–868 (2018). doi:10.1093/nar/gkx936

Supplementary Material

Concordance with Wilcoxon Rank Sum tests

The Wilcoxon Rank Sum is a commonly used method in gene set enrichment analysis. Similarly to our method of repeated stumping, it is used to identify differentially expressed genes between groups of cells. It also shares its salient non-parametric form and calculations based on ordinal transformations with our decision stump approach, which combine to make the methods particularly amenable for gene expression data with potentially many outliers and whose underlying distributions are unknown.

To compare how similarly the two approaches rank enriched features, we also performed a Wilcoxon Rank Sum test for each gene across each of the developmental boundaries considered in the main body of this work. Ranking these genes by the significance level, and applying a Benjamini-Hochberg correction for multiple testing[42] using a false discovery rate of 0.01, we were then able to compare these lists of ranked surviving feature with those produced by the decision stumps.

Figure 4: Comparison of the top 50 ranked features in the murine data following gene set enrichment analysis by Wilcoxon Rank Sum tests and Repeated Decision Stumping.

The concordance between the established Wilcoxon Rank Sum approach and our decision stump implementation improves our confidence in the validity and potential of decision stumps for identifying gatekeeper genes. This is especially true for the *Xenopus* data, where both approaches select a highly similar set of the 50 topmost significant genes from 25,087 possible features. Further, we believe that the mechanistic interpretation offered by the learnt decision stump models – in terms of their learnt expression threshold/feature pair – brings additional value to initial gene set enrichment analyses than the conventional statistical enrichment methods. That the learnt models can also behave as classifiers in their own right is also of significant interest and utility e.g. labelling unclassified samples, generating binary networks.

Figure 5: Comparison of the top 50 ranked features in the *Xenopus tropicalis* data following gene set enrichment analysis by Wilcoxon Rank Sum tests and Repeated Decision Stumping.