nature portfolio

Peer Review File

Bonsai reconstructs tree representations for distortion-free visualization and exploration of high-dimensional data

Corresponding Author: Professor Erik van Nimwegen

Parts of this Peer Review File have been redacted as indicated as we could not obtain permission to publish the reports from one of the peer reviewers.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors propose an alternative approach to represent the high-dimensional space of single-cell expression data in a tree structure. I agree that this representation naturally offers several advantages for downstream analyses, such as providing an intuitive framework for data partitioning and lineage reconstruction.
The main axis of comparison throughout the manuscript concerns the representation of cell–cell distances in two-dimensional embeddings. The authors demonstrate that commonly used approaches, such as UMAP and PCA, tend to distort true cell–cell distances.

The authors rely heavily on simulated data to illustrate these distortions. However, the simulations appear to model only Poisson noise, and it remains unclear whether other biologically realistic sources of variation were considered. My understanding is that methods like PCA are designed to down-weight uninformative variance at the initial dimension reduction stage. Therefore, it is unsurprising that a tree-based method such as Bonsai would better preserve simulated cell–cell distances than PCA or UMAP.
A more critical and informative analysis would be to evaluate whether the tree representation produces biologically meaningful clustering. For instance, in the cord blood dataset, do Bonsai-derived clusters achieve higher cell-type purity compared to standard clustering approaches such as Louvain or Leiden?
Such comparisons would strengthen the biological relevance of the method.

In the cord blood data, the authors discover a novel myeloid derived NK-cell population. I am not a hematology expert and can only speak to potential technical artefacts. To strengthen this conclusion, it would be helpful to clarify whether the markers used to define the myeloid and NK cell lineages correspond to surface markers (e.g., CD56 and CD16) available in CITE-seq datasets. Additionally, it would be important to confirm that doublets and ambient RNA contamination were appropriately controlled for, as these factors could affect lineage inference.

I am not particularly convinced by the football example, which feels somewhat tangential. It would be far more interesting to include a developmental dataset. The authors suggest that the tree structure inherently captures gene regulatory trajectories along its branches. The authors could use differentiation data to show that this claim has merit by showing that precursor cells are indeed positioned closer to ancestral nodes.
Additional detail on the initial preprocessing steps and how these choices affect the resulting tree would also be valuable. For example, would be interesting to see how Bonsai performs without the imputation step implemented in Sanity, using a simple and fast normalization and variance stabilization.

Moreover, I am missing a discussion of batch effects would be far more relevant, especially since large atlas datasets typically integrate cells from multiple samples and experiments.

Minor Comments
* The text and especially the introduction are somewhat lengthy and repetitive; it could be shortened for clarity.
* Although the authors mention runtime, it would be useful to include a direct runtime comparison with UMAP for the same

datasets.
* Figure 4A suffers from overplotting, which makes it difficult to interpret.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Please see attached file.

(Remarks on code availability)
I have not reviewed the code, but I have tried out the method through the website. Overall, the website is very user-friendly and it is a big plus for the paper.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Biotechnology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Author Rebuttal letter:

Dear Anahita,

In response to your decision letter from October 30 2025, we are delighted to send you an extensively revised version of our manuscript `Bonsai : Tree representations for distortion-free visualization and exploratory analysis of single-cell omics data' that addresses all the comments of the reviewers as well as the issues you mentioned in your decision letter.

We have added a large number of new analyses and before providing detailed point-by-point responses to all reviewer comments below, we here first summarize the main changes and additions to the manuscript:

● Scalability: A key concern was Bonsai's scalability, i.e. its ability to be applied to very large datasets such as cell atlases from major consortia. To address this concern we have now developed an extension of Bonsai that is specifically designed to analyze very large datasets. This extension of Bonsai, which we dubbed Backbone-Based-Bonsai, follows an approach often used to build large trees of DNA sequences: a guide tree is constructed from a random subset of the cells, after which the other cells are iteratively placed into the guide tree. Backbone-Based-Bonsai allows analysis of very large datasets which we illustrate with an application to a retinal cell atlas with 1.2 million cells (Fig 3). We also provide updated plots of the runtime scaling for both the original Bonsai and Backbone-Based-Bonsai (Fig 3 and Suppl Figs. S18 and S19).
● Demonstration of Bonsai's ability to accurately recover differentiation trajectories on real data: From several reviewer comments, e.g. that we relied heavily on synthetic data and that application to developmental data might better highlight Bonsai's advantages, we realized that we failed to make clear that Bonsai's results on the cord-blood dataset in fact precisely show Bonsai's ability to recover differentiation trajectories. For the revision we have rewritten this section of the paper and added several additional analyses to highlight Bonsai's performance on these data. In particular, we now directly compare with the current view in the literature on the differentiation hierarchy of blood cell types and show that Bonsai virtually perfectly recovers this hierarchy without using any prior biological information (Fig 4). In addition, to illustrate how vastly Bonsai outperforms other visualization tools, we visualized the same data with PCA, UMAP, PHATE, and DTNE, showing that none of the other methods comes anywhere near to capturing these relationships (Fig S31).
● Demonstration that Bonsai accurately places progenitor states in the tree structure: In response to reviewers asking whether Bonsai can correctly identify cells in progenitor states or different developmental stages, we added two additional analyses. First, we applied Bonsai to simulated data that includes not only the fully differentiated cells at the leaves of the developmental tree, but also cells from progenitor states and find that Bonsai accurately places the progenitor states in its tree with the distance from the root reflecting developmental stage (Suppl Fig S24). To show that this also applies to real data we applied Bonsai to a dataset from C. elegans for which the authors of the original study provided an annotation of the

developmental stage of each cell. We again find that the distance of the cells from the root of the Bonsai tree reflects the developmental stage of the cells (Suppl Fig S25).
● Clades in the Bonsai tree match cell-type annotations: In response to comments from both reviewers, we have now included an extensive analysis of the extent to which the clades in Bonsai's trees match cell-type annotations and how this compares with standard Leiden and Louvain clustering methods. Apart from the unsupervised clustering provided with Bonsai-scout, we now also developed a supervised method which identifies the clades in a Bonsai tree that best match a provided cell type annotation. We tested both methods on a collection of scRNA-seq datasets for which cell-type annotation was available, compared with results from Leiden and Louvain clustering. The results show that Bonsai's clades are generally highly consistent with provided reference annotations (Suppl Fig. S22).
● Negative controls: In response to a very insightful comment of reviewer 2, we have now included analysis of structureless random data to show that Bonsai far more accurately identifies the absence of structure in the data than other methods (Suppl Fig S23).
● Conservation of Euclidean distances: A question of reviewer 2 regarding the relationship between Euclidean distances and distances along differentiation trajectories made us realize our presentation was confusing and we have substantially reorganized the corresponding sections in the paper.

Additional revisions are described in the point-by-point responses below. We are confident that we have successfully addressed all the reviewers' comments and thank them for their highly constructive reviews which led to important additions and significant strengthening of the paper. If there are any questions or additional information we could provide, please do not hesitate to contact us.

Sincerely,

Erik van Nimwegen, on behalf of all authors.
Point-by-point responses to the Reviewers
Reviewer #1 (Remarks to the Author):

The authors propose an alternative approach to represent the high-dimensional space of single-cell expression data in a tree structure. I agree that this representation naturally offers several advantages for downstream analyses, such as providing an intuitive framework for data partitioning and lineage reconstruction.

We thank the reviewer for recognizing the fundamental advantages that tree reconstruction brings.

The main axis of comparison throughout the manuscript concerns the representation of cell–cell distances in two-dimensional embeddings. The authors demonstrate that commonly used approaches, such as UMAP and PCA, tend to distort true cell–cell distances.

1 The authors rely heavily on simulated data to illustrate these distortions. However, the simulations appear to model only Poisson noise, and it remains unclear whether other biologically realistic sources of variation were considered.

We interpret this comment to imply the following two concerns:
1. That we rely mostly on simulated data to argue for Bonsai's improvement over existing methods.
2. That the noise model used for our simulated data might not be sufficiently realistic.

With regards to the first concern, we note that we also presented an in-depth analysis of a cord-blood dataset to show that Bonsai can accurately recover lineage relationships in real data for which lineage information is available and even uncovers compelling evidence of a novel cell type (i.e., myeloid-like NK cells). However, based on comments from both reviewers we now realize that we focused too much on the discovery (and validation) of this novel cell type, and did not sufficiently clarify that Bonsai's results on this dataset also clearly illustrate its superior performance over other visualization tools. In the revision we now show in much more detail how Bonsai's tree accurately recovers the known lineage relationships between blood cell types (Fig 4B and C) and how other visualization methods are unable to recover such lineage relationships (Suppl Fig S31).

However, since ground truth distances between cells are virtually almost unknown for real data, we have to rely on synthetic data to be able to quantify just how much less Bonsai's representations distort the data-structure compared to those of other methods. Nonetheless, in the revision we now also note that, in the cord blood dataset, Bonsai correctly identifies that the gene expression states of erythrocytes and megakaryocytes are especially far removed from those of other blood cell types (Fig 4A and B), whereas other methods do not (Suppl Fig S31). Finally, for the revision we have now also added an analysis of a developmental dataset from C. elegans, demonstrating that Bonsai correctly infers the developmental stages of cells (Suppl Fig S25).

Regarding the second concern, we respectfully disagree that the noise model of our synthetic data is unrealistic. We in fact made significant efforts to make these synthetic datasets as realistic as possible, i.e. we not only matched total UMI counts across cells, absolute expression levels across genes, distributions of log fold-changes across cells, and even the covariance structure of gene expression changes to the statistics in real data, we also made sure to add realistic

noise to the data. We note that we have previously published an in-depth study of the noise properties of single-cell RNA-seq data, and how to properly deal with this noise (Breda et al, Nature Biotechnology 2021). In addition, experiments with aliquots from homogeneous RNA pools confirm that this is a realistic model of the noise in scRNA-seq experiments (Grün et al Nature Methods 2014, Svensson et al Nature Methods 2017). Finally, we note that the same type of noise is routinely used in other platforms for simulated scRNA-seq data.

2 My understanding is that methods like PCA are designed to down-weight uninformative variance at the initial dimension reduction stage. Therefore, it is unsurprising that a tree-based method such as Bonsai would better preserve simulated cell–cell distances than PCA or UMAP.

The reviewer is correct that, in most standard scRNA-seq processing pipelines, after log-transformation the data are projected onto the first 10-100 PCA components with the aim of removing noise. However, it is a misunderstanding that this `down-weighing of uninformative variance' favors Bonsai in preserving cell-to-cell distances. Crucially, the cell-to-cell distances on which we quantify the tool's performances are not the cell-to-cell distances in the raw data but the ground truth distances before any noise was added to the data. To obtain good performance, methods must thus remove the measurement noise that was added. Indeed, in our tests we perform such projecting on top PCA dimensions for UMAP, PHATE and DTNE, and this improves their performance. If we were to skip this projection on the top PCA components, these methods would perform even worse than they already do now.

Finally, we stress that Bonsai also down-weighs uninformative variance, i.e. the pre-processing with Sanity assigns error-bars to all gene expression estimates and Bonsai automatically down-weighs measurements with large error-bars. We thank the reviewer for pointing out that these issues were unclear and in the revision we have now also added an analysis showing that if instead of using Sanity (that downweighs noisy measurements) we use standard logp1 processing, the distance estimations become much less accurate (Suppl Fig S10).

3 A more critical and informative analysis would be to evaluate whether the tree representation produces biologically meaningful clustering. For instance, in the cord blood dataset, do Bonsai-derived clusters achieve higher cell-type purity compared to standard clustering approaches such as Louvain or Leiden?
Such comparisons would strengthen the biological relevance of the method.

Although Bonsai is not a clustering tool, we agree with the reviewer that it is important to assess to what extent the structures in the Bonsai tree also recapitulate cell-type clusters that would be identified with standard approaches such as Louvain or Leiden. To address this issue we have added analyses in the revision to answer the following two questions:
1. Can Bonsai automatically detect clusters in the tree and to what extent do these clusters match clusters from standard clustering tools?
2. Given a reference cluster/cell-type annotation provided with a given dataset, to what extent is the structure in the Bonsai tree consistent with this annotation, i.e. to what extent do the annotated cell-types correspond to clades of the Bonsai tree?
For the first question, we use the unsupervised clustering method that was already provided with the original version of the manuscript that identifies clusters in the Bonsai tree that minimize the pairwise distances between cells in the resulting clades. This unsupervised method is appropriate for users that do not have an annotation of their cells into clusters/cell-types. In contrast, the second question is most relevant for users that already have produced a trusted annotation of their cells into clusters/cell-types and want to know which clades on the Bonsai tree best match these annotated cell types. To address this second question, we developed a new supervised clustering method that cuts the Bonsai tree into clades so as to maximize the normalized mutual information (NMI) between the clades and a given annotation of clusters/cell types.

To assess the performance of these two Bonsai-guided clustering methods and compare them with the performance of the standard Leiden and Louvain clustering methods, we gathered scRNA-seq datasets from the BGee database (Bastian et al, Nucleic Acids Research, 2021) that come with cell type annotations and ran Bonsai on all these datasets. We ran Bonsai's two clustering methods on these datasets in addition to Louvain and Leiden clustering, and then compared how well these 4 different clustering methods matched the provided annotation of each dataset (Suppl Fig S22 in the revision).

We find that,our supervised clustering matches the reference annotation better than either Leiden, Louvain, or our unsupervised clustering for every single dataset. That is, the clades in the Bonsai tree are generally consistent with the provided annotation, and more so than the results of standard Leiden or Louvain clustering. For roughly the top 50% of datasets with clearest clusters (i.e. the datasets with highest average-NMI across methods in the right half of the figure), the unsupervised Bonsai clustering, Leiden, and Louvain clustering all have virtually identical performance.
Bonsai's unsupervised clustering generally performs a little bit worse than Leiden and Louvain on datasets where the average NMI is low. However, since many of the reference annotations were in fact produced using Leiden. Louvain, or similar kNN-based methods, it is debatable whether this can be interpreted as reflecting better performance of Louvain or Leiden, or whether it simply reflects some circularity of the test.

Finally, we want to stress that Bonsai's tree reconstructions go far beyond what clustering methods provide, i.e. it aims to reconstruct differentiation trajectories and accurately capture all structure in the data on all scales. For example, rather than assuming that clusters must exist in the data, Bonsai's results can be used to assess to what extent there is evidence of clear cluster structure, or whether a description in terms of continuous trajectories is more appropriate. The

comments of the reviewer (which are also repeated to some extent by reviewer 2) made us realize that we did not make this strength sufficiently clear in the original manuscript. To address this, we have now extended our discussion of the cord-blood dataset to show that, whereas Bonsai's tree reconstruction accurately recovers known lineage relationships between blood cell types, this type of structure cannot be recovered by other visualization tools (Fig 4B+C and Suppl Fig S31 in the revision). This further illustrates how, in contrast to other methods, Bonsai's visualization goes well beyond identifying clusters of cells.

4 In the cord blood data, the authors discover a novel myeloid derived NK-cell population. I am not a hematology expert and can only speak to potential technical artefacts. To strengthen this conclusion, it would be helpful to clarify whether the markers used to define the myeloid and NK cell lineages correspond to surface markers (e.g., CD56 and CD16) available in CITE-seq datasets.

We thank the reviewer for making us realize that we didn't make sufficiently clear how we validated Bonsai's prediction of the novel myeloid-derived subpopulation of NK cells. Indeed we confirm both the identity of these cells as NK cells and the fact that they carry myeloid lineage markers using the cell surface marker counts. This provides a completely independent validation because these surface marker counts were not used in the Bonsai reconstruction, nor in the author-provided annotation shown in Fig 4; both of these were solely based on the scRNA-seq modality. In addition, we show that similar subpopulations are also observed in a completely different blood cell dataset from bone marrow (Suppl Fig S28). In the revision we have tried to make these points clearer.

5 Additionally, it would be important to confirm that doublets and ambient RNA contamination were appropriately controlled for, as these factors could affect lineage inference.

We agree with the reviewer that this is an important point. To ensure that our conclusions do not depend on the presence of doublets, we have now used the broadly-adopted Scrublet method to remove potential doublets from the dataset. As shown in Suppl Fig S27, removal of doublets does not substantially change the Bonsai tree and does not affect our conclusions regarding the myeloid-derived NK cells.

As to possible contamination of our samples due to ambient RNA: in our standard scRNA-sequence preprocessing pipeline, we have quite strict filters on which cells we allow in the dataset. For example, we discard cells with too few mRNA counts, and we remove cells that look like empty droplets (i.e. having only ambient RNA). In addition, we only took the cells that also made it through the filtering steps of the original publication. Moreover, the expression patterns of the two subpopulations of NK cells do not look like ambient RNA at all but express both clear NK markers plus either markers of the myeloid or lymphoid lineage (Fig 4E+F and Suppl Figs S28-S30).

6 I am not particularly convinced by the football example, which feels somewhat tangential. It would be far more interesting to include a developmental dataset. The authors suggest that the tree structure inherently captures gene regulatory trajectories along its branches. The authors could use differentiation data to show that this claim has merit by showing that precursor cells are indeed positioned closer to ancestral nodes.

The football dataset was only intended to illustrate the broad applicability of Bonsai and we agree with the reviewer that it sheds little additional light on Bonsai's performance for single-cell datasets. We had chosen the cord blood dataset to illustrate Bonsai's ability to reconstruct regulatory trajectories as it is a system with a lot of prior knowledge regarding lineage relationships, and because cells at various stages of differentiation are available within the data. However, due to our focus on the discovery of the novel NK-cell subpopulation, these aspects were not sufficiently highlighted in the original manuscript. In the revision, we have revised this section to show more explicitly that, without using any prior biological knowledge, Bonsai's tree reconstruction accurately recovers the known lineage relationships between blood cell types (Fig 4B+C) and that other visualization tools cannot recover these lineage relationships (Suppl Fig S31).

Beyond this extension of our analysis of the cord-blood dataset, we also agree with the reviewer that it would strengthen our case to provide additional demonstration that Bonsai indeed positions precursor cells closer to ancestral nodes, including on a developmental dataset. To address this, we now perform two additional analyses in the revision. First, we created a synthetic dataset with a known cell division tree, and instead of providing only the fully differentiated cells at the leaves, we also included cells for all internal nodes of the tree, so that we could directly test where Bonsai places such precursor cells. As shown in Suppl Fig. S24, Bonsai performs as well as could be hoped for on this test, with each precursor cell being placed on a short terminal branch from the internal node that it corresponds to.
Second, we identified a developmental dataset for which it would be possible to perform a very similar test. In particular, we obtained a developmental dataset from C. elegans for which the authors of the initial publication provided an annotation of the developmental stage of each cell (Packer et al., Science, 2019). We find that the distance of each cell from the root in the Bonsai tree correlates well with its annotated developmental stage (Suppl Fig S25), confirming that Bonsai can indeed properly place precursor cells in developmental datasets.

We thank the reviewer for these comments, because we think these two additional analyses have considerably strengthened the presentation.

7 Additional detail on the initial preprocessing steps and how these choices affect the resulting tree would also be valuable. For example, would be interesting to see how Bonsai performs without the imputation step implemented in Sanity, using a simple and fast normalization and variance stabilization.

We agree with the reviewer that it is instructive to compare the default Bonsai tree (which uses Sanity to preprocess the raw data) with a more basic normalization procedure. For the revision we now provide results on the simulated data where we use a conventional normalization pipeline (i.e following the Scanpy-tutorial, https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html, including log1p-transformation + highly variable gene selection).
The results, shown in Suppl Fig S10, are indeed instructive.
First, the topologies of the trees that Bonsai builds on data that were processed in a conventional way are fairly close to the ground truth and the visualization of the overall structure in the data is still more accurate than what is provided by other tools (i.e. PCA, UMAP, Phate, and DTNE). This shows that Bonsai's tree reconstructions are highly robust.
On the other hand, the trees obtained with conventional normalization have unrealistically long terminal branches. This is a direct consequence of the fact that standard preprocessing does not explicitly take the noise in the data into account. As a result, every cell appears far from every other cell even though in reality some cells are very close. This is not a detail because, as shown in Suppl Fig S10, the estimated distances are much less accurate with the basic normalization procedure than with our default Sanity preprocessing. Moreover, nearest-neighbor identification is also much less accurate when using the basic normalization procedure.
We thank the reviewer for bringing up this issue because it underscores that the use of Sanity is important for the quantitative accuracy of the results, while the overall qualitative visualization of the structure in the data is quite robust to changes in the preprocessing.

8 Moreover, I am missing a discussion of batch effects would be far more relevant, especially since large atlas datasets typically integrate cells from multiple samples and experiments.

'Batch effect' is a collective term for a number of different phenomena that include both variation introduced by technical variability in the experimental protocols as well as true biological variation, for example between identical tissues from different animals. We agree with the reviewer that these types of variability are often a challenge in analysis of scRNA-seq data and that we should have provided a discussion of this issue. In the revision we now provide such a discussion in the Discussion section of the manuscript. In addition, we also provide code for a basic batch correction procedure and have added a description on this procedure in the methods section.

Because batch effects can correspond to a wide range of different effects, what kind of `correction' is appropriate highly depends on the experimental design and the specific scientific questions of the study (e.g. whereas variation across samples from different animals/patients can be a nuisance that one wants to remove in one setting, it might be the main signal one wants to identify in another). Providing solutions for all imaginable scenarios is beyond thus the scope of our manuscript. In the revision we have provided a simple batch correction method with Bonsai in which systematic differences in average gene expression levels across batches are corrected for.

Minor Comments
* The text and especially the introduction are somewhat lengthy and repetitive; it could be shortened for clarity.

We agree with the reviewer and have gone over the entire text with the aim of removing redundancies. This resulted, for example, in significant reductions of both the introduction and discussion sections.

* Although the authors mention runtime, it would be useful to include a direct runtime comparison with UMAP for the same datasets.

We agree that users should have a good idea of Bonsai's runtime, and we have revised Fig 3A to make Bonsai's runtimes clear. However, we respectfully disagree that a direct side-by-side comparison of the runtimes of Bonsai and UMAP is helpful. First, runtime is not a valid performance criterion in the way that accuracy of the results is. We are providing a number of rigorous quantitative assessments of the accuracy of different tools and we feel it would be misleading to provide a side-by-side comparison of runtimes as if this were an equally relevant performance criterion.
Second, UMAP is typically used in a very different manner from the way Bonsai is used. In contrast to Bonsai, UMAP has many parameters and is often run many times in a trial-and-error fashion with different parameter settings, while Bonsai needs to be run only once.
However, we agree with the reviewer that it is important that Bonsai runs fast enough, even on llarge datasets, to not cause the computational analysis to become the bottleneck. While the original version of Bonsai was already sufficiently fast for datasets of typical sizes (Fig 3A and Suppl Fig S16), for very large datasets the runtime did become unacceptably large.
For the revision (and also in response to comments of reviewer 2) we developed a new version of Bonsai that is especially suited for large datasets, and describe this new version of the algorithm, which we dubbed Backbone-Based-Bonsai, in the Methods. Briefly, Backbone-Based-Bonsai follows a similar approach to the way very large phylogenetic trees are constructed from DNA sequence data. First a guide-tree is constructed on a random subset of the cells, after which all other cells are iteratively placed on the backbone of this guide tree (which is further iteratively improved). In Figure 3, we now also show the (vastly improved) scaling in runtime of Backbone-Based-Bonsai and showcase Bonsai results on an atlas of 1.2M retinal cells. Finally, more details on runtimes and memory usage of both versions are in Suppl Figs S18-S19 and in Suppl Fig S17 we show that Backbone-Based-Bonsai is still highly accurate.

* Figure 4A suffers from overplotting, which makes it difficult to interpret.

We thank the reviewer for this feedback. For the revision we have improved the figure, although it is virtually impossible

to avoid all over-plotting when there are many thousands of cells. Moreover, since it is generally difficult to convey the full structure of the Bonsai results in a single picture we developed the Bonsai-scout app, with which one can zoom in on specific parts, view the tree in different layouts, and even cluster the cells and find marker genes for specific clades. For example, the cord blood dataset of Fig 4A is available under https://bonsai.unibas.ch/bonsai-scout/?dir=Cord_blood_cells_CITE-seq.

Reviewer #2 (Remarks to the Author):

Here, de Groot et al present bonsai which is a method for visualizing high dimensional data. The main motivation and application is scRNAseq data, where one typically deals with ~20k dimensions (even though the actual manifold is much smaller), and sometimes millions of data points. Visualizing these datasets remains a challenge, and in recent years several methods including UMAP and PHATE have been developed for this purpose. However, accurately representing structures and distances when projecting to a plane is a highly challenging problem, and there is a consensus in the field that existing methods have various shortcomings. Bonsai tries to address many of these issues by trying to identify a tree structure for the data.

We thank the reviewer for these comments which indeed reflect some of our motivations for developing Bonsai.

The authors argue that this approach is able to overcome the curse of dimensionality by exploiting structures that are present in the data. Despite the need for better visualization methods, I believe that there are several issues that needs to be resolved:

1) The scalability analysis in Figure 3 focuses solely on computational complexity. However, a more meaningful assessment should include actual computational time comparisons. I recommend the authors to provide wall-clock time comparisons with other visualization tools.

We agree that it is important that users get a realistic sense of Bonsai's wall clock runtime for realistic datasets, and that it scales so that it can be run in a reasonable time even on large datasets. For the revised version of the manuscript, we remade Figure 3 to show concrete wall clock runtimes as a function of dataset size (with more detail added in Suppl Fig S18). In addition, as we detail in our answer to the reviewer's question 7, we have developed a new version of Bonsai that is specifically designed to scale to very large datasets. This Backbone-Based-Bonsai first builds a guide-tree based on a random subset of the cells and then iteratively places the remaining cells on this guide-tree, while also iteratively improving the final tree. Figure 3 also shows concrete wall clock runtimes for this new version of Bonsai.

As in our response to a minor comment from reviewer 1, we respectfully disagree that a direct side-by-side comparison of the runtimes of Bonsai and other tools is helpful. While it is important that a tool can be run in a reasonable amount of time, runtime is not a valid performance criterion in the way that accuracy of the results is. We therefore feel it would be misleading to present both accuracy and runtime comparisons as if these were of equal importance. Moreover, many other visualization tools have many tunable parameters and are often rerun many times in a trial-and-error fashion, whereas Bonsai has no free parameters and thus needs to be run only once, making a direct comparison not very meaningful.

2) In addition to runtime, memory usage is important and it would be useful to report how this scales with the size of the data.

We agree with the reviewer that memory usage is also important and now present an analysis of Bonsai's memory usage in Suppl Fig S19.

3) Although the visualizations are compelling, the authors should also evaluate how well the tree structure matches up with known cell type annotations in Figure 3 quantitatively.

We agree that evaluation of the extent that Bonsai's clade structure is consistent with cell type annotations is important. We note that this comment is very similar to comment number 3 of Reviewer 1 and we refer the reviewer to our extensive answer to that comment. Briefly, in the original manuscript we already provided an unsupervised clustering method that identifies well-separated clades in a Bonsai tree. For the revision we have now also developed a supervised clustering method that takes a given cell-type annotation and identifies which clades of Bonsai's tree best match this annotation. Furthermore, using a collection of scRNA-seq datasets with cell-type annotations we compared the performance of both Bonsai's supervised and unsupervised clustering methods with that of standard Louvain and Leiden clustering. We find that Bonsai's supervised clustering generally matches the provided annotations very well, and in particular better than standard Louvain or Leiden clustering, while the unsupervised clustering matches the reference annotations similarly to those of Leiden and Louvain clustering (Suppl Fig S22 of the revision).

4) While the authors claimed that trees are a natural representation for lineage structure in single-cell data, the biological interpretation of the tree remains unclear. Specifically, the relationship between tree topology and developmental relationships is ambiguous. For instance, two cells located on different branches could represent (1) different developmental stages along the same differentiation trajectory (temporal progression), or (2) distinct cell fates on separate lineage branches (lineage divergence). Bonsai does not distinguish between these two scenarios. Adding a

discussion of this aspect would improve understanding.

Bonsai indeed aims to reconstruct the most likely trajectories by which the gene expression states of the cells have diverged. We agree with the reviewer that, because the measured cells are per definition placed at the leaves of the tree, it is a priori unclear how Bonsai would represent cells in precursor states or from different developmental stages. In the revision, we have now added two new analyses to address this question directly. We first created synthetic data where the cells derive from different developmental stages. For our original paper, we tested Bonsai and other tools on synthetic datasets where the gene expression states of cells evolved along a cell division tree, but in those tests our datasets included only the cells at the leaves of the tree. In the revision we now also created simulated data where not only the `fully differentiated cells' are included, but all progenitor cells from the internal nodes of the tree are also included in the dataset. We ran Bonsai on this data to see where progenitor cells would be mapped. As shown in Suppl Fig S24, progenitor cells are mapped about as accurately as could be hoped for: each precursor cell is attached by a short terminal branch to the internal node of the tree that it corresponds to. Moreover, the distance of the cells from the root of the tree accurately reflects their developmental stage.

Next, we looked for a developmental dataset where we could perform an analogous test, i.e. a dataset for which the developmental stages of different cells is known. We obtained a dataset of cells from C. elegans development for which the authors of the original publication provided annotations of the developmental stage of each cell. We reconstructed a Bonsai tree on this data and found that, also in this real dataset, the distance of each cell from the root of the tree correlates well with its annotated developmental stage (Suppl Fig S25 in the revision). This confirms that for developmental datasets, Bonsai can successfully identify the developmental stage of the cells, with earlier cells placed closer to the root.

We have added a discussion of these results to the main text and thank the reviewer for making this suggestion, which we feel has significantly improved the manuscript.

5) A bigger conceptual question: even if Euclidean distances are perfectly preserved on the tree, Euclidean distance in gene expression space may not reflect true developmental distance. Using Euclidean distance basically creates shortcuts across the true underlying manifold, ignoring the actual developmental trajectory. It could be helpful to discuss when Euclidean distance preservation is biologically meaningful versus when geodesic distance along developmental manifolds would be more appropriate.

These are questions that we have been deeply thinking about in our research group for years and the reviewer's question makes us realize that, in our original manuscript, we essentially passed over a discussion of these subtle conceptual issues, and simply assumed readers would agree that it is per definition desirable for the visualization to preserve Euclidean distances. In the revision we have now completely rewritten the presentation in an attempt to present these conceptual issues more clearly. The most relevant sections are at the start of the Results section and in the section 'the blessing of dimensionality', but we will try to summarize the argument here.

The main aim of the Bonsai algorithm is to construct a tree that gives the most likely gene expression trajectories between all pairs of cells. Or equivalently, that following branches from the root the tree gives the hierarchy of differentiation trajectories along which the cells have diverged.
We agree with the reviewer that it is reasonable to think of the gene expression dynamics of cells during differentiation as moving along `geodesics' of some manifold embedded in high-dimensional gene expression space. However, we do not know what this manifold looks like, not even its dimensionality, let alone its potentially very complex structure.
In the absence of knowledge about the manifold, we make use of the fact that, locally, every smooth manifold is isomorphic to flat Euclidean space so that, locally, Euclidean distances should reflect distances along the manifold. The key assumption that our Bonsai algorithm makes is that cells have been sampled sufficiently densely that, at least along each branch of the tree, the Euclidean distance along the branch is a good approximation to the true distance along the manifold.
This assumption, in combination with the assumption that the process is Markov and that there is a priori no preferred direction in gene expression space, uniquely defines the likelihood along the branches and defines the entire algorithm.

That is, the Bonsai algorithm was designed to reconstruct the most likely trajectories and not to preserve Euclidean distances between cells. However, we find that, in practice, Bonsai does preserve the Euclidean distances between all cells. Why is this? This turns out to be a consequence of features of high-dimensional spaces. First, in high-dimensional spaces almost all directions are orthogonal. Consequently, when the gene expression states of cells have diverged along the branches of a tree in a high-dimensional space, the movement along each branch of this tree is orthogonal to the movement along each other branch, so that (because of the Pythagorean theorem) squared Euclidean distances simply sum along the branches of the tree. As a consequence, the squared Euclidean distance between any pair of cells at the leaves accurately matches the sum of the branch lengths along the tree.
Beyond this, we find that even if the data do not derive from a tree structure, as long as it is high-dimensional, the trees that Bonsai reconstructs still accurately preserve all pairwise Euclidean distances. This surprising observation, which we dubbed `the blessing of dimensionality', likely results from the same properties of high-dimensional spaces, but we have no mathematical proof for why this has to be the case. It is simply an observation: for pseudo-bulk clusters, and random scatters of points, as well as real scRNA-seq data, Bonsai's tree accurately conserves Euclidean distances between all points.
Although Bonsai was not designed to do this, it is highly fortuitous, because it implies that Bonsai's trees not only recover the most likely differentiation trajectories (when the data derive from an underlying differentiation process) but

also accurately capture the structure of the data by preserving all pairwise distances.

In the revision we have completely rewritten the presentation in an attempt to make these subtle conceptual points more clear.

6) There are no negative controls. To evaluate bonsai, I created a synthetic example with 512 genes and 1,024 cells. Each element was a random integer in the range [0, 1000]. As expected, PCA visualization shows the cells as a uniform cloud with no apparent structure. By contrast, Bonsai identified two clusters as shown below. To be fair, closer inspection of the dendrogram clearly indicates that the two clusters are not well separated, but the lack of negative controls and strategies for identifying this scenario remains a challenge.

We thank the reviewer for raising this point because it made us realize that Bonsai is in fact far more powerful in 'performing negative controls' than other tools, and in the revision we now illustrate this point using a random dataset created using the reviewer's instructions.

First, we understand that the reviewer was confused by the Bonsai-scout app appearing to suggest there are two clusters in their random dataset. This was a bug that we have now fixed. Originally, when the user did not provide any annotation with their dataset, Bonsai-scout defaulted to showing the first clustering results on the dataset, which happens to be dividing the data into two groups, even when there is no evidence for group structure in the data. This was clearly misleading and we apologize for this bug which has now been fixed.

However, as the reviewer also noticed, the Bonsai tree clearly showed that there is virtually no structure in this data, i.e. with the tree corresponding to a star-tree in which all cells are close to equidistant. Importantly, this is in fact the correct answer for this random dataset. Random points in high-dimensional spaces are all close to equidistant and Bonsai correctly shows that this is the structure of this random dataset. In contrast, PCA and UMAP hallucinate that some cells are much closer together than others. For the revision we have now included an explicit analysis of this random dataset (Suppl Fig S23) to illustrate that only Bonsai can correctly identify that there is no structure at all in this random dataset, i.e. that all cells are equidistant.

7) The authors mention in the discussion that bonsai could potentially be sped up by leveraging ideas from evolutionary biology. Given the current limitations relative to the size of datasets that many of the readers will be working with, I think that this is a critical limitation. The poor scaling of bonsai severely limits its usefulness.

This is a very valid concern. Although we believe that the accuracy of the method is far more important than its precise running time, we completely agree that it is important that the tool can be run in a reasonable time, even on very large datasets. While our initial version of Bonsai does run in a reasonable time for most scRNA-seq datasets generated by individual labs (Fig 3A and Suppl Fig S16), it indeed did not scale to very large datasets. For the revision of our paper we have now developed a new mode of running Bonsai specifically designed to run on very large datasets. This Backbone-Based-Bonsai uses an approach similar to approaches for reconstructing phylogenetic trees of very large numbers of DNA sequences: We first build a guide-tree on a random subset of cells, and then iteratively place all other cells on this guide tree. As shown in an updated version of Figure 3, Backbone-Based-Bonsai has vastly improved scaling of runtime with dataset size and we showcase Bonsai results on an atlas of 1.2M retinal cells. More details on runtimes and memory usage of both Bonsai versions are in Suppl Figs S18-S19, and in Suppl Fig S17 we show that Backbone-Based-Bonsai is still highly accurate.

8) The Bonsai visualization of the temporal preimplantation data effectively captures the known developmental structure in human embryonic data (data from "Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos"). The bifurcating pattern in the Bonsai output correctly reflects the ICM/TE segregation that occurs at E5, which is consistent with the original study's findings (shown in the t-SNE plots in Figure 1E and Figure 2A of original study).
However, UMAP (the second plot) also successfully identifies this bifurcation structure. The authors claim Bonsai is superior for hierarchical data because it explicitly maps to a tree structure, but testing on this example doesn't strongly support that claim since other methods capture the same biological structure. To demonstrate Bonsai's advantages, it would be more convincing to show cases with real-world datasets where the tree structure reveals hierarchical relationships that other methods miss or obscure. With these comparisons, it's more clear what practical benefit the tree-based approach provides beyond what existing methods already achieve..

We completely agree with the reviewer that Bonsai's power can be best demonstrated by application to real-world data with known hierarchical structures or lineage relationships. It is precisely for this reason that we presented an in-depth analysis of a cord-blood dataset. That is, there is much knowledge in the literature about the lineage relationships between blood cell types, allowing us to test whether Bonsai recovers these relationships. Moreover, this CITE-seq dataset also contained independent cell surface marker measurements, allowing for independent validation of cell type annotations.

However, it is clear from the comments of both reviewers, that we failed to properly highlight how Bonsai's results on this data vastly outperform other methods. For the revision, we have rewritten this section to highlight more clearly how accurately Bonsai recovers known lineage relationships. As shown in Fig 4B+C of the revision, we show that the lineage relations recovered by Bonsai are in remarkable agreement with the current view in the literature, i.e. with hematopoietic

stem cells at the root, the first splitting off of the erythrocyte lineage, followed by the major split into the myeloid and lymphoid lineages, further splits into B- and T-cells in the lymphoid lineage, and splitting of different subtypes of monocytes and dendritic cells in the myeloid lineage. In contrast, as shown in Suppl Fig S31 of the revision, other visualization methods completely fail to recover these lineage relationships. Finally, we note that Bonsai's analysis of this data also provides the first clear evidence for a new subtype of cells: NK cells deriving from myeloid precursors, which we validate with independent cell surface marker measurements. In addition, we show this same novel subtype of NK cells exists in other blood cell data from bone marrow as well.

Reviewer #2 (Remarks on code availability):

I have not reviewed the code, but I have tried out the method through the website. Overall, the website is very user-friendly and it is a big plus for the paper.

We thank the reviewer for expressing their appreciation of this aspect of our work.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Biotechnology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all my concerns. I particularly like the additional analysis of the developmental dataset of C.elegans. Also the honest comparison of the cell type cluster/clade correspondence is helpful.
All in all, I think the manuscript has improved enough for publication.
Having an alternative method for scRNA-seq data-visualisation is already of great value on its own.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have done a very nice job at addressing all of my major concerns. There is only one minor issue that I would like them to address in the discussion section:

The claim about the manifold being approximated by Euclidean distances for short distances is of course true. Moreover, the results that they achieve using Bonsai implies that this could be happening in practice. However, as far as I understand, their paper does not show explicitly that the sampling is dense enough to fulfill this criterion. Hence, I would like the authors to acknowledge that this is only a conjecture and that demonstrating that it is true would likely improve confidence that the representation obtained by Bonsai does indeed represent the true manifold.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Biotechnology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Author Rebuttal letter:

Dear Anahita,

Thank you for your decision letter of April 21 2026. We are delighted that our manuscript is accepted in principle. As

requested, and following your instructions, we have prepared a final revision of our manuscript. Besides including more discussion regarding the final comment of reviewer #2, we have included your edits, made some further edits ourselves (especially to streamline and avoid redundancy between Introduction and Discussion sections), and followed the checklist in reformatting our manuscript.

Below we provide a `point-by-point' response to the final comment of reviewer #2. We would like to generally thank the reviewers for their highly constructive reviews which led to important additions and significant strengthening of the paper, and thank you for editing our manuscript. If there are any questions or additional information we could provide, please do not hesitate to contact us.

Sincerely,

Erik van Nimwegen, on behalf of all authors.

Point-by-point responses to the Reviewers
Reviewer #1 (Remarks to the Author):

The authors have addressed all my concerns. I particularly like the additional analysis of the developmental dataset of C.elegans. Also the honest comparison of the cell type cluster/clade correspondence is helpful. All in all, I think the manuscript has improved enough for publication. Having an alternative method for scRNA-seq data-visualisation is already of great value on its own.

We thank the reviewer for the gracious comments and thank them for their constructive reviews that improved the paper.

Reviewer #2 (Remarks to the Author):

The authors have done a very nice job at addressing all of my major concerns.

We thank the reviewer for their constructive comments that improved the paper.

There is only one minor issue that I would like them to address in the discussion section:

The claim about the manifold being approximated by Euclidean distances for short distances is of course true. Moreover, the results that they achieve using Bonsai implies that this could be happening in practice. However, as far as I understand, their paper does not show explicitly that the sampling is dense enough to fulfill this criterion. Hence, I would like the authors to acknowledge that this is only a conjecture and that demonstrating that it is true would likely improve confidence that the representation obtained by Bonsai does indeed represent the true manifold.

We agree that, at the moment, we simply assume that the sampling is dense enough and have no proof that this is indeed the case. In our latest revision, we have made sure to stress that this is just an assumption, both in the introduction section and in the results section where the general approach is introduced.
In addition, we have now also added a discussion of this general point in the Discussion section. In particular, we note that, besides assuming that Euclidean distance is a reasonable metric for the distances between neighboring nodes in the tree, we point out that Bonsai also assumes that movements in any direction in gene expression space are a priori equally likely. We now discuss that, because gene expression changes are driven by activity changes in regulators, the movement is not only expected to be constrained to a lower-dimensional manifold but we also expect there to be a non-trivial correlation structure, i.e. biasing which directions are more or less likely. In the Discussion we note that we already have evidence that Bonsai could likely be improved by learning this correlation structure from the data and that this is a future avenue for extension of Bonsai.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature

Biotechnology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.