Immunology and Inflammation
Computational and Systems Biology

Reviewed Preprint
v2 • April 23, 2025
Revised by authors

Reviewed Preprint
v1 • May 9, 2024

Deconstructing Complexity: A Computational Topology Approach to Trajectory Inference in the Human Thymus with *tviblindi*

Jan Stuchly, David Novak, Nadezda Brdickova, Petra Hadlova, Vojen Sadilek, Ahmad Iksi, Daniela Kuzilkova, Michael Svaton, George Alehandro Saad, Pablo Engel, Herve Luche, Ana E Sousa, Afonso RM Almeida, Tomas Kalina

Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic • Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium • Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium • Centre d'Immunophénomique - CIPHE (PHENOMIN), Aix Marseille Université (UMS3367), Inserm (US012), CNRS (UAR3367), Marseille, France • Department of Biomedical Sciences, Medical School, University of Barcelona, Barcelona, Spain • Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

eLife Assessment

The authors present an algorithm and workflow for the inference of developmental trajectories from single-cell data, including a mathematical approach to increase computational efficiency. In this latest version, the authors addressed the benchmarking of the novel method, but the absence of quantitative comparisons to state-of-the-art methods still make this study **incomplete**. Based on the shown validation approaches, one can neither ultimately judge if the shown method will be an advance over previous work nor whether the approach will be of general **useful** applicability.

https://doi.org/10.7554/eLife.95861.2.sa2

Abstract

Understanding complex, organ-level single-cell datasets represents a formidable interdisciplinary challenge. This study aims to describe developmental trajectories of thymocytes and mature T cells. We developed *tviblindi*, a trajectory inference algorithm that integrates several autonomous modules - pseudotime inference, random walk simulations, real-time topological classification using persistent homology, and autoencoder-based 2D visualization using the *vaevictis* algorithm. This integration facilitates interactive exploration of developmental trajectories, revealing not only the canonical CD4 and CD8 development but also offering insights into checkpoints such as TCRβ selection and positive/negative selection. Furthermore, *tviblindi* allowed us to thoroughly characterize thymic regulatory T cells,

tracing their development passed the negative selection stage to mature thymic regulatory T cells. At the very end of the developmental trajectory we discovered a previously undescribed subpopulation of thymic regulatory T cells. Experimentally, we confirmed its extensive proliferation history and an immunophenotype characteristic of activated and recirculating cells. *tviblindi* represents a new class of methods that is complementary to fully automated trajectory inference tools. It offers a semi-automated tool that leverages features derived from data in an unbiased and mathematically rigorous manner. These features include pseudotime, homology classes, and appropriate low-dimensional representations. These features can be integrated with expert knowledge to formulate hypotheses regarding the underlying dynamics, tailored to the specific trajectory or biological process under investigation.

Teaser

tviblindi reveals αβ T-cell development checkpoints and places activated phenotype $CD4^{+}CD8^{dim}$ Tregs at the end of Treg trajectory.

1. Introduction

Exploring the development of mature, fully functional cells from their progenitors has been the focus of researchers for many decades. While single-cell techniques such as flow cytometry were fundamental to the description of developmental stages, recent advances in single-cell methods measuring large sets of markers (mass cytometry, single-cell RNA-seq) have enabled truly comprehensive descriptions of entire tissues and organs. In parallel, computational methods that capture and describe possible dynamics in the data, as well as topological relationships among cell populations, were developed. After the seminal Wanderlust methodological study (1), trajectory inference (TI) and pseudotime data analysis became an area of intensive research, producing gradually more and more sophisticated tools for a growing number of general topologies (2–7). Ideally, bioinformatic tools should reveal the structure of the data to an expert who can interpret it with a focus on his or her research question.

However, certain limitations prevent widespread adoption of these tools in analytical workflows of real-world data. First, the computational complexity of existing TI methods is often prohibitive when analyzing large datasets from current single-cell RNA methods (especially when the data comprise cell atlases of entire organisms, embryos, and human organs) or from flow and mass cytometry with datasets of millions of cells (3,8–11). Second, TI methods often use multiple steps of dimensionality reduction and/or clustering, inadvertently introducing bias. The choice of hyperparameters also fixes the a priori resolution in a way that is difficult to predict. Third, the current TI methods work well on artificial datasets but lack a straightforward approach to control the effect of noise (technical artifacts or unexpected events) in multiscale topologies of real-world data. The fourth, and perhaps the major obstacle is that existing TI tools do not offer the necessary interaction with the analytical process. We believe that such interaction, which helps the researcher understand the particularities of a sample, is crucial for exploratory analyses of unknown data.

Motivated to develop a generic TI solution useful to a biologist investigating a real-world dataset, we developed *tviblindi. tviblindi* is a modular TI method designed to tackle large datasets, significant noise, technical artifacts, and unequal distribution of cells along the time axis. As a proof of concept, we investigate αβ T-cell development in human thymus and the transition of mature T cells to peripheral blood.

Human T cells develop in the thymus, which is seeded by bone marrow derived $CD45^{pos}$ $CD44^{pos}CD34^{hi}CD7^{neg}$ and Notch primed $CD45^{pos}CD44^{pos}CD34^{hi}CD7^{pos}$ progenitors (12–15). Large $CD34^{hi}CD1a^{neg}$ immature thymocytes develop into small $CD34^{dim}CD1a^{pos}$ immature thymocytes (13), becoming gradually restricted to the T-cell lineage (16). After losing CD44 expression, right before gaining CD1a, they become committed to a T-cell fate (17).

These early human $CD34^{pos}$ thymocytes then progress through three stages (18) (1) $CD34^{pos}CD38^{neg}CD1a^{neg}$, (2) $CD34^{pos}CD38^{pos}CD1a^{neg}$ and (3) $CD34^{pos}CD38^{pos}CD1a^{pos}$. The first TCRβ D-J rearrangements are detected in the $CD34^{pos}CD38^{pos}CD1a^{neg}$ population. In-frame TCRβ are mostly selected at the transition from $CD34^{pos}CD38^{pos}CD1a^{pos}$ to the next stage: the $CD4^{pos}$ immature single positive (ISP) population. Concordantly, pre-TCRα (pTα) expression peaks in the $CD34^{pos}CD38^{pos}CD1a^{pos}$ and ISP stages, after which it declines. TCRα rearrangements are initiated when thymocytes progress from $CD34^{pos}CD38^{pos}CD1a^{pos}$ toward the ISP stage and continue until the $CD3^{pos}$CD4 and CD8 double positive (DP) stage (18).

The complete TCRαβ receptors are checked for their binding to MHC-self peptide complexes on thymic epithelial cells and dendritic cells. Failure to engage the MHC-self peptide leads to death by neglect (19). Efficient TCR binding leads to positive selection and further development along the CD4/CD8 axis. Positively selected DP cells express CD69 on their surface and start to downregulate the CD8 co-receptor. If they are selected via the MHC-II peptide complex, the TCR signal persists. When the duration of this signal exceeds a certain time limit, the transcription factor ThPOK is induced, and the cells become CD4 cells (20). If selected on the MHC-I peptide complex, the signal diminishes, the transcription factor RUNX3 is induced, expression of the CD8 co-receptor is reactivated and CD4 eliminated, leading to the CD8 cell fate (20). On the other hand, strong TCR binding leads to negative selection and apoptosis or to the so-called agonist selection and development of regulatory T cells (Tregs) (21). During these processes the cells acquire the chemokine receptor CCR7 and move from the thymic cortex to the medulla, where tissue-specific self-antigens are ectopically expressed on the surface of medullary epithelial cells (22,23). The current models suggest that both the strength/duration of TCR signal (24–27) and the integration of TCR signals from consecutive T cell – antigen-presenting cell encounters (28,29) help to determine the autoreactive T-cell fate. Costimulatory signals (30–32) as well as cytokines (33–39) contribute to the process. The upregulation of the high-affinity IL-2 receptor CD25 and its signaling are known to stabilize FOXP3 expression (36,37). Importantly, JAK/STAT signaling triggered by γ chain cytokines promote $FOXP3^{pos}$ Treg proliferation (38) and induction of BCL-2 and its anti-apoptotic effect (39). While the TCR repertoire of conventional T cells and Tregs are largely distinct and non-overlapping (40), the exact branching point of these two main lineages is still a matter of intensive research (41,42) also employing human data based on single-cell transcriptomics (43–45).

When investigated using single-cell methods (15,43–48), the process of T-cell development translates into high-dimensional single-cell data with complex topology which need to be interrogated computationally.

We developed *tviblindi,* a new tool which breaks new ground in several aspects:

- (1) It offers a highly scalable, linear complexity framework that works at a single-cell level.
- (2) Analysis is performed in the original high-dimensional space, avoiding artifacts of dimensionality reduction.
- (3) The framework is adapted to discover features taking into consideration their varying scales (along the time axis or in different cell types).
- (4) Crucially, our method allows the user to interact with the analytical process and to set the appropriate level of resolution in real time.

In this paper, we first describe our new computational approach to exploration of developmental trajectories. Then we present an analysis of real-world datasets of T-cell development to illustrate the power of our analytical method and to describe the development of human T regulatory cells in detail. Last, we describe novel discrete stages in the development of Tregs and cell surface markers used in their study.

2. Results

2.1. Computational method for the TI and interrogation – *tviblindi*

The objective of *tviblindi* is to offer a flexible framework for TI and topological data analysis (TDA) interrogation of single-cell data. *tviblindi* facilitates efficient interpretation of data as well as sensitive discovery of minor trajectories. These two competing goals are resolved by interactive grouping of random walks, which are used as probes in the high-dimensional space, into trajectories, allowing for real-time adjustment of trajectory-specific resolution. This interactive approach provides an exhaustive description of the data, while allowing the researcher to introduce expert knowledge into the analytical process and to gradually gain insight into a particular dataset.

tviblindi was designed as an integrated framework implementing several independent modules to facilitate the use of diverse approaches to pseudotime estimation and data visualization, without the need to rerun the analysis. *tviblindi* allows to compute pseudotime which is resilient to unequal distributions of cells (e.g., accumulation of double positive T cells during T-cell development or in case of the presence of a developmental block). It also keeps the resolution on the single-cell level, while processing data in a reasonable time frame. Random walks, directed by the pseudotime, are then simulated, thus creating a set of probes in the high-dimensional space. To assemble random walks into meaningful trajectories, *tviblindi* employs witness complex triangulation (49) to capture the topology of the original data without the need for dimensionality reduction. It implements a representation of homology classes, which accounts for incomplete coverage of the high-dimensional points by the triangulation, navigating around the potential pitfall of this computational approach.

Contrary to other approaches, *tviblindi* considers each random walk a separate entity, which allows for probing major as well as minor trajectories while keeping track of the feature dispersion along the time axis. This approach enables sensitive detection of irregularities, such as hubs and discontinuities, in the topology of the underlying graph. While pseudotime captures local geometry of the data, persistent homology (50) aims to capture significant non-local features such as sparse regions (holes), which persist over a range of scales. In simplified terms, the persistence of a sparse region may be understood as the difference between the density of points at its boundary and in its interior.

Most persistent features can be extracted and used to capture the differences between random walks, inducing a multiscale clustering (51). We generalized and extended this idea to high-dimensional data. We took advantage of the natural hierarchical structure of persistent homology classes to achieve real-time, on-demand aggregation of random walks into trajectories respecting both local geometry and global topology of the point cloud.

2.1.1. Introducing tviblindi on an artificial data set

We showcase our methodology using an artificial dataset created with the *dyntoy* package (for further details see Supplementary note, subsection 1.1 *dyntoy* dataset). **Figure 1A** shows a schematic representation of the data, capturing the basic topology with sparse regions labeled α and β.

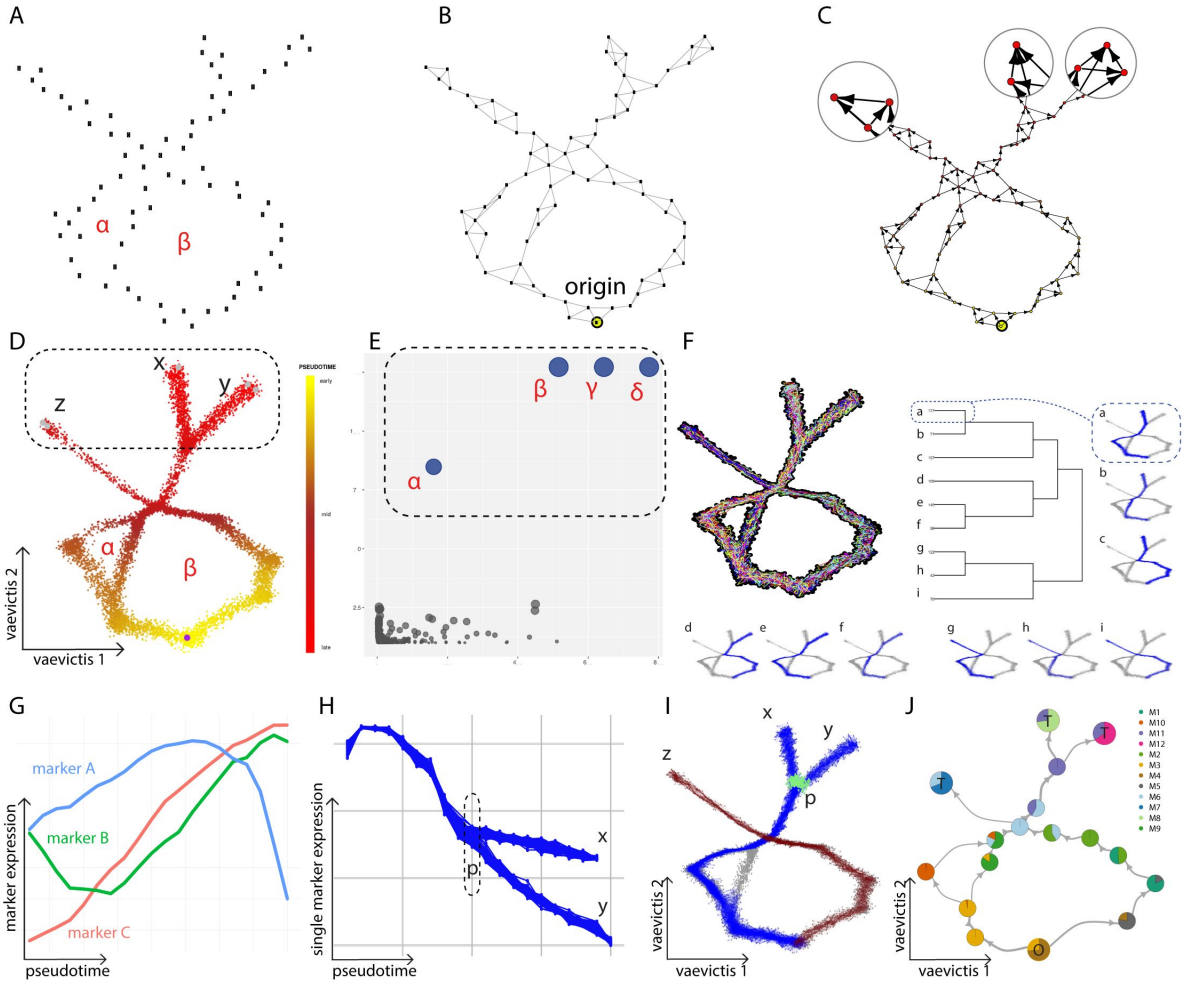

Figure 1

Analytical process on the artificial *dyntoy* dataset.

Dashed lines mark selections made by the user, panels A-C show a scheme of the whole dataset, panels D-I correspond to the interactive GUI for this dataset: (A) The scheme of a point cloud representing single-cell data, where sparse regions α and β are present. (B) The dataset is represented as a k-NNG and the cell of origin is selected. (C) Pseudotime is calculated as expected hitting distance and the edges of the k-NNG are oriented according to the pseudotime. Candidate endpoints are automatically suggested as vertices without outgoing edges (circled). (D) Estimated pseudotime is represented by a yellow-to-red color gradient, the origin is shown as the purple dot. Potential developmental fates (x, y, z) are shown as gray points on the *vaevictis* plot. Endpoints can be selected for further investigation (in this example, all endpoints are selected, dashed line). (E) Sparse regions are represented on a persistence diagram, which enables the selection of the significant sparse regions (dashed line). The sparse regions γ and δ correspond to essential classes created by the selection of disparate endpoints. This way, random walks are classified using the selected ends and sparse regions. (F) All random walks are shown on the *vaevictis* plot and their classification into trajectories is depicted in the hierarchical clustering dendrogram. (G) By selecting branch (a) leading to the endpoint (x) of this dendrogram, the given trajectory can be investigated in detail, including viewing the average expression of multiple markers along pseudotime. (H) By selecting two endpoints (x and y) but only a single relevant marker, its dispersion and possible branching points can be examined. Trajectories (a and e) from panel F are visualized and the branching region (p) is selected. (I) Multiple trajectories (trajectory i in red and a and e in blue) can be visualized on the 2D plot. Points in the selected branching region (p) are highlighted in green. (J) Connectome representing a basic structure of the data and the simulated random walks. The pie charts indicate the distributions of cell populations in each vertex (cluster). The arrows show the direction of pseudotime. Vertex containing the cell of origin (O) and vertices containing endpoints (T).

First, the data are represented as an undirected k-nearest neighbor graph (k-NNG; **Figure 1B**). The cell of origin is defined by the user (either directly, or a population of origin is specified and the point closest to the centroid of this population is then taken) and used to estimate pseudotemporal ordering (pseudotime) directed away from the origin (**Figure 1B**). Edges of the k-NNG are oriented with respect to the pseudotime forming a directed acyclic graph (DAG).

Next, a large number (typically thousands) of random walks is simulated on the DAG. These walks are finite and their final vertices are candidates to become developmental endpoints (**Figure 1C**). The user can select one or more potential endpoints for further investigation (**Figure 1D** labeled x, y, z).

A persistence diagram, which captures the prominence of sparse regions within the point cloud, then allows the user to select significant sparse regions (**Figure 1E**), and random walks are organized into trajectories by means of hierarchical clustering. This clustering respects the global geometry and classifies trajectories based on how they navigate around the sparse regions (**Figure 1F**). In the presented artificial dataset, *tviblindi* correctly identified three putative endpoints (x, y, z) and organized simulated random walks into three distinct trajectories for each endpoint (a, b and c, for end x; d, e and f for end y and g, h and i for end z).

Our interactive framework allows the user to inspect the evolution of specific markers (**Figure 1G**), track key points in their development (e.g., branching point p) and focus on cells at such key points (**Figure 1H**, I). For a quick overview, a "connectome" summarizing the basic structure of the point cloud and of simulated random walks can be plotted (**Figure 1J**). We describe each *tviblindi* module below. Detailed algorithmic descriptions of particular modules can be found in the Supplementary note, sections 2-4.

2.1.2. Visualization

Visualization of high-dimensional data is essential for the initial overview and for the informed interactions with the data during downstream analysis. However, any dimensionality reduction introduces simplification, which can lead to misinterpretation. In *tviblindi* framework, the dimensionality reduction step is independent of all other modules (which work directly in the original high-dimensional space) and is used solely for a visual representation of the dataset.

While dimensionality reduction has been extensively and successfully used for single-cell data (e.g., UMAP (52,53) and viSNE (54,55)), interpreting developmental trajectories requires not only the local, but also the global relationships of cells to be captured (often distorted in the lower-dimensional embedding by current dimensionality reduction algorithms). For this purpose, we created *vaevictis*, a variational autoencoder model based on ideas in *scvis* (56) and *ivis* (57) fitted with naive importance sampling. This technique creates a continuous representation of the data rather than isolated islands of populations (Supplementary note, section 2 Dimensionality reduction, **Figure 1B**, C). Due to the naive importance sampling step, numerically large compact populations are not overrepresented. This way, we obtain a clearer picture of the developmental dynamics in the data (Supplementary note, **section 2** Dimensionality reduction, **Figure 1D**, E). Moreover, once calculated, it can be applied to new datasets with the same measured parameters (protein markers or detected transcripts), facilitating integration of new results.

2.1.3. Pseudotime and random walks simulation

Correct pseudotemporal ordering is an essential and computationally intensive part of TI algorithms. Here we use a particular formulation (and modification) of the idea of *expected hitting time* (58). We calculate the expected number of random steps necessary to reach any cell in a dataset from the cell of origin. To leverage the sparsity of the k-NNG and to calculate the hitting times on a single-cell level, we use the following formulation: given an undirected k-NNG $G = (V, E,$

p), with vertices V, edges E, weights p: $E \rightarrow [0,1]$ representing the probability of transition of each edge respectively, and an origin v_0, the expected hitting time of a cell is a weighted average of hitting times of its neighbors increased by 1 $\tau(y) = \sum_{x:\{x,y\}\in E} p_{xy}(\tau(x)+1), y \neq v_0$ and $\tau(v_o) = 0$. This formula translates into a sparse linear system, which can be solved efficiently by numerical methods (such as conjugate gradients) to any precision even for graphs with millions of vertices. Consequently, we are able to calculate the pseudotime directly in the original space at a single-cell level. Furthermore, we obtain a measure of success: the relative error reported by the numerical solver.

A key application of *tviblindi* is the comparison of trajectories between normally developing tissues and those with developmental abnormalities, manifesting a block and/or pathological proliferation resulting in an overabundant intermediate population. In such cases, the hitting time calculation would assign the highest pseudotime to this intermediate population, which would hamper the topology of simulated random walks (Supplementary note, section 3, Pseudotime estimation & random walks simulation). To mitigate this effect, the formula above can be modified in the following way: suppose that apart from $G = (V, E, p)$, we have a sparse matrix D, recording mutual distances for all vertices in the graph connected by an edge. Then the definition of the pseudotime as the *expected hitting distance is* $\tau(y) = \sum_{x:\{x,y\}\in E} p_{xy}(\tau(x)+D_{xy}), y \neq v_0$ and $\tau(v_o) = 0$.

In other words, the expected distance to travel before reaching vertex y. Both methods are also very robust with respect to the choice of the number of neighbors in the k-NNG (Supplementary note, section 8 Performance evaluation).

Once pseudotime is computed, the edges in the graph $G = (V, E, p)$ get oriented (**Figure 1B**) and a large number of random walks is simulated on the DAG (**Figure 1F**). Importantly, all random walks on the graph are finite. The set of terminal vertices of these walks represents estimates of terminal biological fates (for further details see Supplementary note, **section 3** Pseudotime estimation & random walks stimulation).

2.1.4. Real-time interactive topological classification

The most distinctive feature of *tviblindi* is the interactive aggregation of simulated random walks into trajectories with respect to the global topology of the point cloud. This procedure is based on the idea of multi-scale topological classification using persistent homology (59). Random walks are considered distinct entities, and they can be aggregated into trajectories respecting the structure of the point cloud.

Walks are clustered based on how they circumnavigate significant sparse regions identified using persistent homology. The appropriate level of significance can be adjusted by the user by the selection of significant sparse regions on a persistence diagram (**Figure 1E**). Classification by the means of persistent homology has a natural hierarchical structure induced by filtration (59,60) (see Supplementary note, section 4 Topological clustering of random walks for details), which translates into a dendrogram of random walks (**Figure 1F**).

Applying classification by persistent homology to a large high-dimensional dataset with considerable technical noise requires several improvements of previously published work (59). First, as noted above, the selection of significant sparse regions is interactive, which allows the user to choose an appropriate level of detail for each dataset or trajectory. Second, we use witness complex triangulation (61) to adapt for high dimensionality of single-cell data (for which the original methodology was intractable). Third, since the witness complex triangulation does not guarantee filling the space completely, we introduce the detection and representation of sparse regions with infinite persistence. This last modification also allows us to classify random walks into trajectories based on their different endpoints and consequently to study developmental branching. (This is impossible with the original formulation.) See Supplementary Videos and Supplementary note for a presentation of the *tviblindi* graphical user interface (GUI).

2.1.5. Connectome – a fully automated pipeline

For a quick overview of a given dataset, we have implemented a "connectome" functionality inspired by PAGA (5) (**Figure 1J**). This is a fully automated pipeline to create a directed graph, in which the data is clustered (by default using Louvain community detection (62)). Random walks are contracted to the identified clusters, providing a lower-resolution estimate of developmental trajectories. The clusters are plotted as pie charts of the represented cell populations. The graph layout employs the same 2D representation(s) as *tviblindi* GUI to facilitate the interpretation. The clusters containing the cell of origin and endpoints are automatically detected and denoted O and T respectively. At the same time the orientations and widths of edges reflect the direction and number of underlying random walks. Connectome suffers from some of the limitations of other, fully automated, end-to-end methods. The resolution is predetermined at the cell-population level by the choice of clustering. In addition, there is no direct way to interactively choose a resolution at the level of random walks, nor to detect and exclude clear artifacts. Therefore, we recommend using this functionality only to get an overview of a given single-cell dataset with suggested trajectories before performing the full interactive *tviblindi* analysis.

2.1.6 Performance evaluation

We evaluated *tviblindi* against several state-of-the-art trajectory inference tools (Monocle 3(63), Stream(64), Palantir(65), Via(66), PAGA(67), CellRank 2(68) and StaVia(69)) and found it to be superior in terms of speed and resolution with respect to typical challenges. Specifically, it was better at dealing with locally abundant clusters, developmental loops, and non-dominant trajectories (shown and discussed in detail in Supplementary Note, Section 8). To further evaluate the performance of *tviblindi* on complex single-cell RNA sequencing data, we analyzed in detail the development of immune cells in the bone marrow of a published dataset(70). We also tested *tviblindi* on mouse gastrulation data(71) to assess its ability to interrogate large cell atlas datasets. We compared the results of *tviblindi analysis* of these datasets with the two best-performing methods (StaVia, CellRank2) from the performance evaluation. This comparison has shown that *tviblindi* is superior when dealing with distinct trajectories converging to a common fate (shown and discussed in detail in Supplementary Note, Section 10).

Sensitivity to hyperparameters is discussed in Supplementary Note, Section 8.1.

2.2. Using tviblindi on mass cytometry and single-cell RNA-seq data sets

2.2.1. tviblindi exhaustively dissects human T-cell development

We applied *tviblindi* to the 34-parameter mass cytometry data obtained in our human T-cell development study (Supplementary Table Mass panel 1). The data contained barcoded human thymocytes and human peripheral blood mononuclear cells (PBMC). Dead cells and cells not belonging to the αβ T cell developmental lineage were excluded, yielding a total of 1,182,802 stained cells for analysis (Supplementary Figure 1). When the cells were visualized using a CD4xCD8 dot plot, we observed the expected patterns of double negative (DN), double positive (DP), CD4 single positive (SP) and CD8 SP populations for a cryopreserved thymus (**Figure 2A**). We then used the default *vaevictis* dimensionality reduction to reveal the basic structure of the T-cell compartment. Thymic and peripheral blood T cells were projected into distinct areas of the *vaevictis* embedding with a minor overlap (**Figure 2B**). Localization of the CD8, CD4 and Annexin V positive cells can be viewed on the plot using a heatmap color scale for expression levels of these markers. This provides an interpretable image of the T-cell compartment (**Figure 2C**).

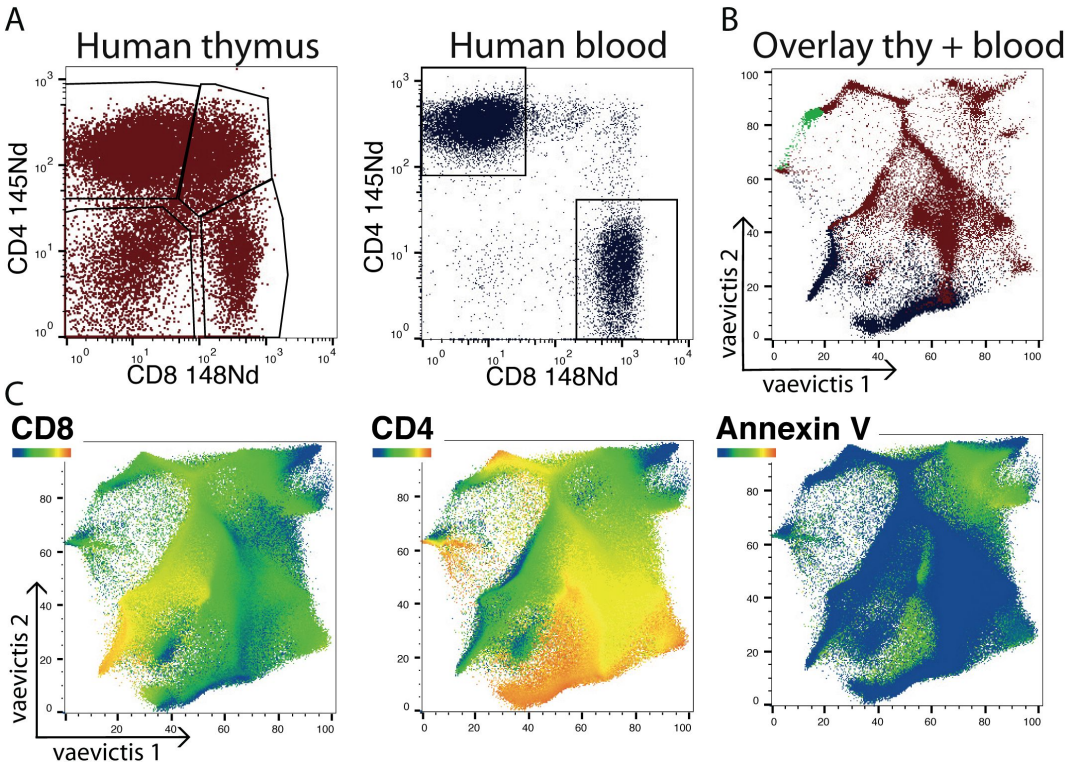

Figure 2

Visualization of the basic structure of the T-cell compartment using *vaevictis.*

(A) Bivariate CD4 x CD8 dot plot of cells from human thymus (brown) and human peripheral blood (blue) acquired using mass cytometry. The cells were gated as DN, DP, CD4 SP and CD8 SP. (B) *vaevictis* plot of 37-parameter (including barcode) mass cytometry panel measurement showing the positions of human thymus (brown) and human peripheral blood (blue), with the $CD34^{pos}$ progenitors shown in green. (C) Expression of CD8, CD4 and Annexin V shown using a blue-green-yellow-red color gradient on the *vaevictis* plot. Blue color indicates the lowest expression and red color indicates the highest.

Next, we gated the population of origin as $CD34^{hi}$ $CD1a^{neg}$ DN cells and analyzed the data with *tviblindi*. We simulated 7500 random walks and discovered eleven groups of endpoints labeled #1 - #11 (**Figure 3A**). Five of them (#1 to #5) were located in the region of peripheral CD4 T cells (where #5 corresponded to mature naive CD4 T cells transiting to peripheral blood). Endpoint #6 found among the thymic CD4 SP T cells is further described in the section *Non-conventional end interpretation.* Three groups of endpoints were located in the peripheral CD8 T cell compartment (where #7 corresponded to cells of CD8 TEMRA phenotype, #8 to CD8 central memory T cells and #9 to mature naive CD8 T cells transiting to the periphery). Two remaining endpoints (#10 and #11) belonged to the apoptotic thymocytes (compare to **Figure 2B** and **2C**).

The connectome shows a basic overview of dynamics in the data on a cluster level (**Figure 3B**). The subsequent interactive analysis of simulated random walks gives biological meaning to the observed connections and allows for the detailed interpretation of data on a single-cell level.

First, we chose endpoint #1 representing peripheral CD4 effector memory T cells for further analysis. We applied persistent homology to select the significant sparse regions in the data and used these to cluster individual simulated random walks (**Figure 3C**). The topology of the point cloud representing human T-cell development is more complex than that of the illustrative artificial dataset shown in **Figure 1** and does not offer a clear cutoff for the choice of significant sparse regions. Interactive selection allows the user to vary the resolution and to investigate specific sparse regions in the data iteratively.

Clustering the random walks resulted in a dendrogram, where leaves with abundant walks represent the dominant developmental trajectories. Nonetheless even the less abundant groups of walks can be selected for in-depth investigation (**Figure 3D**).

Leaf I. contains the largest number (452) of random walks (**Figure 3D** and **3E**). Of note, leaf I. faithfully represents a larger supercluster of random walks II. (**Figure 3D** and **3E**). For the leaf I. trajectory, we displayed the intensity of expression of key markers along the calculated pseudotime, the pseudotime lineplot (**Figure 3F** top). We then mapped the key developmental stages onto the *vaevictis* plot (**Figure 3F** bottom).

The CD4 ISP stage was found where CD4 and CD1a expression begins, but prior to CD8, intracellular TCRβ and CD3 expression. This is followed by a $CD3^{neg}$ DP stage, where intracellular TCRβ increases. Eventually, the DP cells become $CD3^{pos}$, reflecting the successful expression of the TCRαβ receptor. Next, they lose CD8 and later they lose the cortical thymocyte marker CD1a. Finally, the CD4 SP thymocytes gain CD45RA and are ready to egress from the thymus as naive CD4 SP cells. In the peripheral blood they lose the CD45RA during maturation only to gain it at their terminal stage (**Figure 3F**). Similarly, if endpoint #7 is chosen, the canonical trajectory leading to peripheral CD8 cells can be visualized and described (Supplementary Figure 2).

This knowledge of the topology of key points of canonical development helped us to interpret additional developmental endpoints and the trajectories leading to them.

2.2.2. Variants of trajectories including selection processes

We then focused on alternative trajectories leading to peripheral CD4 cells corresponding to endpoint #1. Leaf IV., comprising 103 walks (**Figure 4A**) exhibits a clear artifact connecting highly dividing progenitor cells with doublets, which could invalidate a fully automated analysis (Supplementary Figure 3).

For further in-depth investigation we selected the two remaining larger groups of random walks represented in the dendrogram by leaves highlighted as III. and V. (**Figure 4A**). After the $CD3^{neg}$ DP stage, these trajectories diverted from the canonical course described above. The cells passed

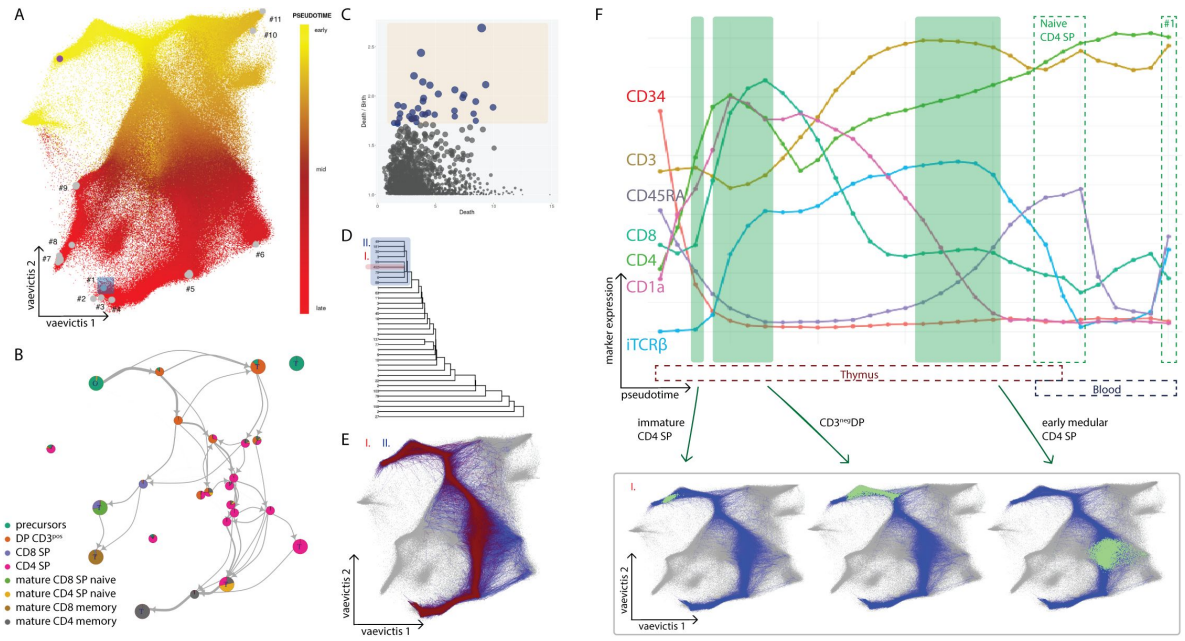

Figure 3

Developmental endpoints and detailed analysis of major trajectory leading to endpoint #1 by *tviblindi*.

(A) *vaevictis* plot of T-cell development in thymus and peripheral blood. Estimated pseudotime is represented by a yellow-to-red color gradient with CD34^{pos} progenitors as the population of origin (purple dot). Gray dots indicate the discovered developmental endpoints. Endpoint #1 highlighted by a blue rectangle represents mature CD4^{pos} effector memory T cells selected for further exploration. (B) Connectome shows low resolution structure of the data and of simulated random walks. (C) Persistence diagram representing sparse regions detected within the point cloud of measured cells. The orange rectangle marks a user-defined selection of sparse regions. (D) Dendrogram of clustered trajectories shows a subcluster of 452 walks (red rectangle, labeled I) within a larger group of random walks (blue rectangle, labeled II). The number to the left of each leaf indicates the number of random walks in the leaf. (E) *vaevictis* plot displaying the above selected trajectories in corresponding colors. (F) Pseudotime line plot showing the average expression of selected individual markers along the trajectory to endpoint #1 (top). The selected areas of interest corresponding to T-cell developmental stages (green rectangles) are shown in green (as indicated by the arrows) on the *vaevictis* plots (below).

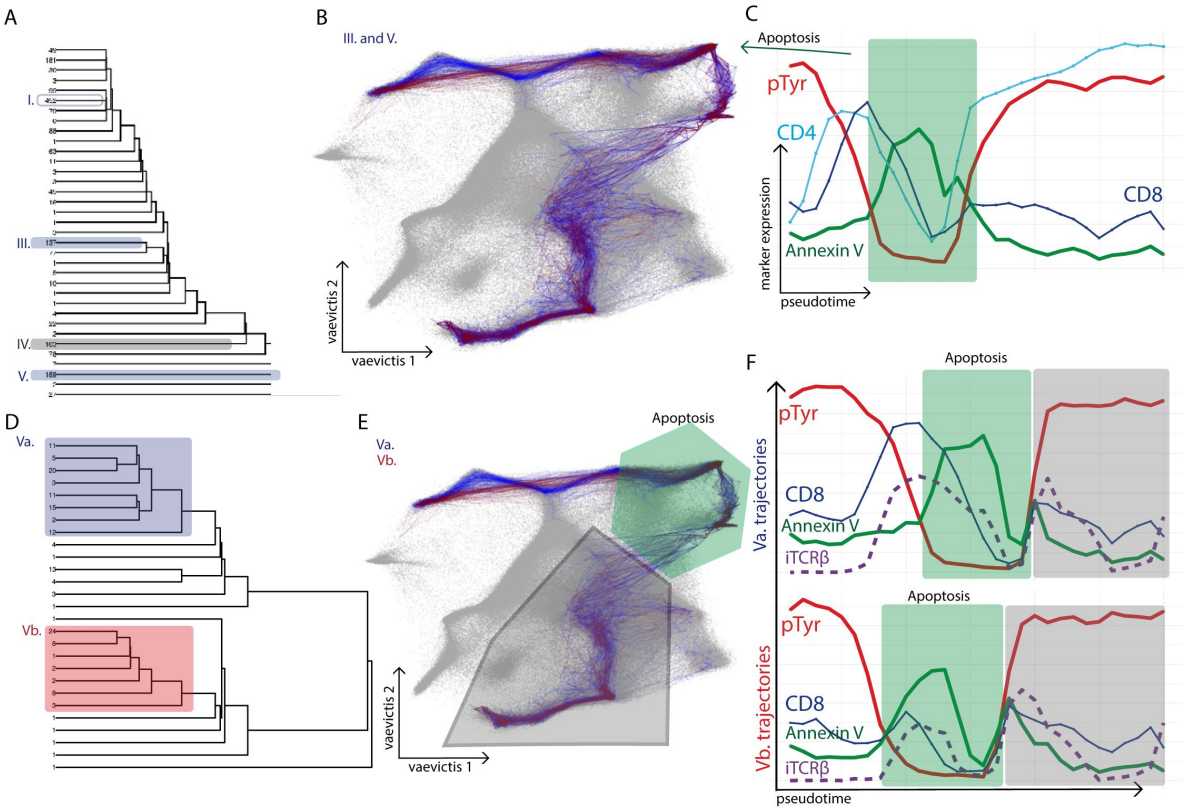

Figure 4

tviblindi analysis of trajectories leading to apoptosis.

(A) Dendrogram identical to **Figure 3D** with additional trajectories selected for closer investigation (blue rectangles) labeled III (137 random walks) and V (159 random walks). (B) *vaevictis* plot showing the topology of trajectories in leaves III and V. Apoptotic cells are shown in green. (C) Pseudotime line plot, which shows the average expression of selected markers along calculated pseudotime. The green rectangle highlights the region with increased expression of apoptotic marker Annexin V and decreased expression of phosphotyrosine. Events in the selected region are displayed in panel 4B in green (see the arrow). (D) A detailed dendrogram of leaf V from panel 4A. Two distinct trajectories were selected for further analysis (blue rectangle, Va and red rectangle, Vb). (E) *vaevictis* plot showing the topology of trajectories in leaves Va and Vb. The green polygon marks the region of apoptosis, the gray polygon marks the region of more advanced stages of thymic and peripheral CD4 T-cell development. (F) Pseudotime line plot, which depicts the average expression of selected markers along the trajectories Va and Vb. The region of apoptosis and the region of more advanced stages of CD4 T-cell development are highlighted (green and gray rectangle respectively).

through a region with an overall decrease in phosphorylated tyrosine expression and an increase in Annexin V expression, indicating apoptosis (**Figure 4B** and **4C**). Of note, endpoints #10 and #11 were situated in this apoptotic region (see **Figure 3A** for their precise location). We therefore interpreted the first part of these trajectories as leading to apoptosis upon failure of DP thymocytes to pass either positive or negative selection checkpoints in the thymic cortex (**Figure 4B** and **4C**).

Trajectories grouped in the leaves III. and V. continued through the apoptotic stage (indicated by a green hexagon in **Figure 4E**) and reconnected to the canonical CD4 T-cell maturation trajectory (indicated by a gray hexagon in **Figure 4E**). While this is an accurate description of the topology of the single cell data points, expert interpretation finds it counterintuitive since apoptotic death of thymocytes is the biological end-point for the unselected thymocytes. t*viblindi* interface offers interaction with the data in the form of pseudotime lineplot showing the evolution of each marker versus pseudotime (**Figure 4C**), where in the stage following the apoptotic region, we observe a gradual gain of CD4 (but not CD8). Thus, the biological explanation for this counterintuitive observation is that this stage is actually described in the reverse direction. It corresponds to the negative selection of CD4SP thymocytes in the thymic medulla where cells lose CD4 en route to apoptotic thymocytes. Topologically, trajectories form a triangular shape with corners in the DP thymocytes, CD4SP thymocytes and apoptotic thymocytes. Since the pseudotime distance from DP to apoptotic thymocytes is shorter than the real negative selection (DP to CD4SP to apoptotic cells), the calculation of the pseudotime inference cannot properly interpret the biological direction of the CD4SP cells to apoptotic cells. This highlights the value of the data-driven expert interpretation approach of *tviblindi*, which can solve even such complex topologies.

One of the most important questions in exploratory data analysis is whether the chosen level of detail is sufficient to exploit the information present in the data, or whether resolution needs to be increased. The interactive design of *tviblindi* allowed us to change the level of resolution on the persistence diagram and to investigate the random walks heading to apoptosis in more detail. After increasing the resolution, we observed that leaf V random walks split into two clear trajectories (Va and Vb, **Figure 4D**). The Va trajectories corresponded to the aforementioned failure to pass the positive/negative selection checkpoint. The Vb trajectories diverted from the canonical trajectory even earlier (**Figure 4E**), before the DP stage, and headed directly toward apoptosis.

Contrary to Va where the cells expressed intracellular TCRβ prior to Annexin V, the Vb cells failed to express TCRβ before entering the apoptotic trajectory (**Figure 4F**). We interpret this as failure to pass the β selection checkpoint due to unsuccessful TCRβ chain rearrangement. The later appearance of TCR β, after expression of Annexin V (**Figure 4F**), is caused by mixing with cells entering apoptosis at later developmental stages.

Similarly, if we focus on CD8 T cells, the same checkpoints can be discovered and interrogated (Supplementary Figure 4).

In summary, we show that *tviblindi* organized individual cells expressing various levels of 34 markers into trajectories leading to the expected ends of αβ T-cell development and discovered the corresponding checkpoints. Biological interpretation was enabled by plotting the expression of key markers along the developmental pseudotime. We were able to describe the canonical development of conventional CD4 T cells and CD8 T cells from their progenitors in the thymus to effectors in peripheral blood. We also found trajectories leading to apoptosis, when thymocytes failed to pass TCRβ selection or positive selection or when they were negatively selected (See Supplementary Video 2).

2.2.3. Non-conventional end interpretation

The last analyzed trajectory remaining to be interpreted was the one leading to endpoint #6 (**Figure 3A**, **Figure 5A**). The corresponding population is hereafter referred to as End#6. End#6 was found in the thymic region (compare **Figure 2B** and **Figure 3A** for precise location in the *vaevictis* plot) and contained CD4 SP cells as well as a smaller fraction of cells which fall within the DP gate (Supplementary Figure 5A). The pseudotime line plot of CD8 expression shows that the End#6 cells gain dim expression of CD8 at the very end of the trajectory (Supplementary Figure 5B, **Figure 5B**). The possibility that $CD4^{pos}CD8^{dim}$ cells from the End#6 population are doublets consisting of a developing DP cell and a more mature CD4 SP T cell, as previously suggested (72), was ruled out by comparing the DNA content of End#6 SP and DP populations with the DNA content of dividing $CD34^{pos}$ precursor cells (Supplementary Figure 5C).

The trajectory leading to End#6 diverged from the conventional CD4 trajectory at the branching point of negative selection (**Figure 5C**, compare with **Figure 4E**) and continued through a distinct intermediate subset before reaching End#6. The End#6 cells have a $CD25^{hi}$, $CD127^{pos}$ phenotype, which distinguishes them from the preceding stage reminiscent of T regulatory (Treg) cells ($CD25^{pos}CD127^{neg}$), which could represent the immature Treg stage (**Figure 5B** and **5D**). This observation led us to investigate the expression of Treg-specific markers in End#6 cells using a modified antibody panel (Supplementary Table Mass panel 2). After we gated the relevant populations and visualized them in the *vaevictis* plot (Supplementary Figure 6), we observed that FOXP3 (transcription factor of Treg cells) and Helios (transcription factor shown to be upregulated upon negative selection in mice (21,73,74)) is expressed by the $CD25^{pos}CD127^{neg}$ immature Treg cells as well as by the $CD25^{hi}CD127^{pos}$ End#6 cells (**Figure 5E**). Therefore, we hypothesized that End#6 cells constitute a specific Treg population (**Figure 5D** and **5E**).

Some published reports describe the divergence of Treg cell development at the DP stage, with FOXP3 expression already detectable at the DP stage (38,39,72,75–78). Therefore, we wanted to ensure that positioning of the $CD8^{dim}$ cells at the end of the trajectory is not an artifact of the TI. We decided to sort the populations of interest by fluorescence-activated cell sorting (FACS) and estimate their developmental order based on the analysis of the presence of T-cell receptor excision circles (TRECs).

We used a specific panel (Supplementary Table Mass panel 3) to select surface markers distinguishing between End#6 cells, immature Tregs, CD4 SP and DP cells (Supplementary Figure 7) and to design a gating strategy for sorting these populations (Supplementary Figure 8). To confirm that the sorting strategy led to identification of the correct populations, we overlaid the conventionally gated populations over the *vaevictis* plot (Supplementary Figure 9). Based on this analysis, we designed an 11-color panel for FACS sorting (Supplementary Table FACS panel) and sorted the key populations from thymus and from pediatric and adult peripheral blood (Supplementary Figure 10). We measured the presence of TRECs and T cell receptor alpha constant gene region (TCRAC) in the sorted populations to estimate the number of cell divisions undergone since the recombination of the TCRα chain (**Figure 6A**). The number of cell divisions, as estimated by the calculated TREC/TCRAC, ratio was lower than two in immature $CD3^{pos}$ DP cells and immature Treg (DP as well as CD4 SP) cells. Thymic conventional mature naive CD4 SP did not feature a significant number of divisions. The TREC/TCRAC ratio in conventional naive CD4 SP cells in pediatric peripheral blood corresponded to 1-2 cell divisions, while in the adult peripheral blood it corresponded to 3-4 cell divisions. For $CD45RA^{neg}$ Treg cells sorted from pediatric as well as adult peripheral blood, this was more than 5 cell divisions. For the End#6 $CD4^{pos}CD8^{dim}$ and End#6 CD4 SP cells, the TREC/TCRAC ratio corresponded to more than 5 divisions. These results confirm that $CD4^{pos}CD8^{dim}$ cells at the End#6 of the trajectory leading through immature Treg CD4 SP cells are not the true thymic DP cells. In agreement with our TI algorithm, we confirmed that End#6 cells represent a more developmentally advanced stage.

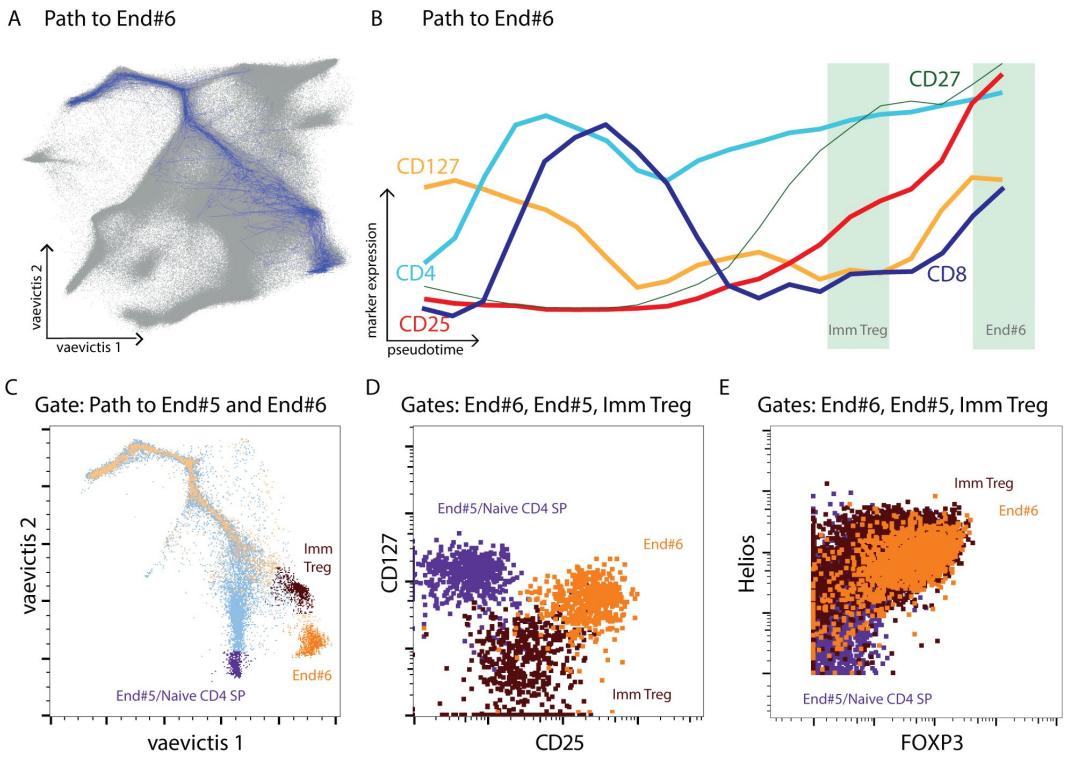

Figure 5

tviblindi analysis of the trajectory leading to End#6.

(A) A trajectory leading to End#6 located in the thymic portion of the *vaevictis* plot (compare to **Figure 2B**). (B) Pseudotime line plot showing average expression of selected individual markers along developmental pseudotime. (C) *vaevictis* plot showing the trajectory to the conventional naive CD4 SP T cells (blue) and the trajectory to End#6 (orange). Conventional naive CD4 SP T cells are shown in purple, the immature Treg stage in brown and End#6 in orange. (D) Bivariate plot of CD25 and CD127 expression on gated conventional naive CD4 SP T cell (purple), immature Treg stage (brown) and End#6 (orange). (E) Bivariate plot of Helios and FOXP3 expression on conventional naive CD4 SP T cell (purple), immature Treg stage (brown) and End#6 (orange) in a validation experiment.

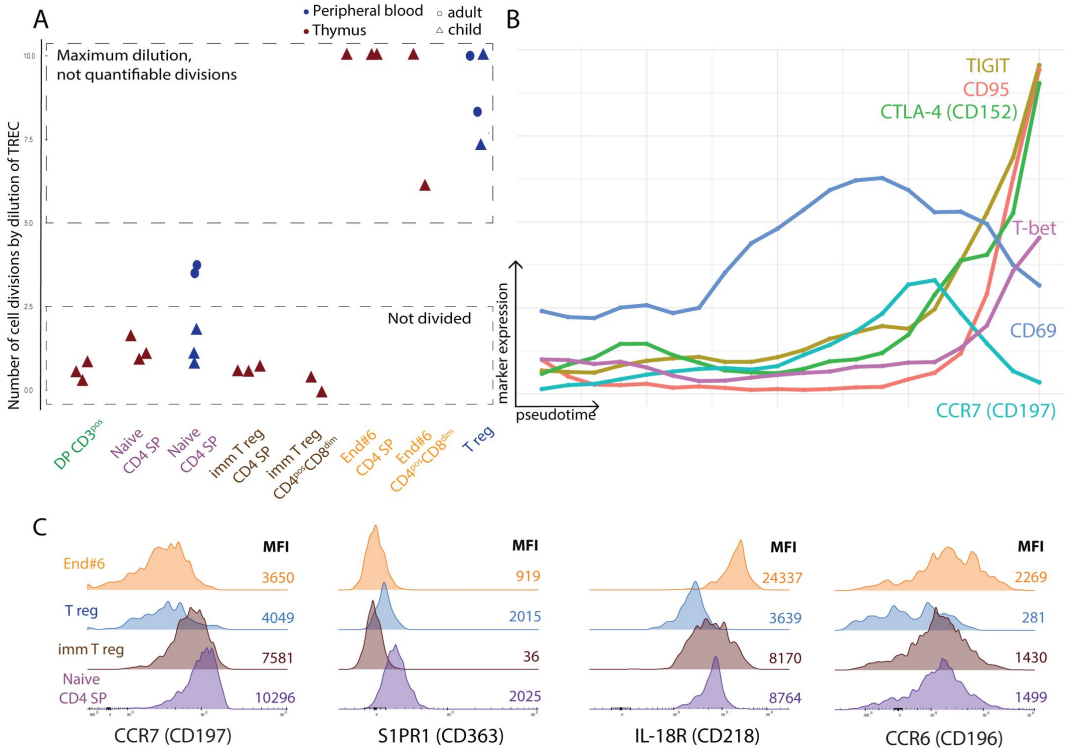

Figure 6

Detailed analysis of End#6.

(A) Visualization of the TREC/TCRAC ratio, depicted as the calculated number of divisions for each population. Symbol code: cells from peripheral blood (blue), cells from thymus (brown), cells from adult donor (circle), cells from pediatric donor (triangle). (B) Pseudotime line plot depicting the expression of TIGIT, CD95, CD152, T-bet, CD69 and CD197 markers by cells from trajectory to End#6. (C) Overlaid histograms showing the expression of chemokine and cytokine receptors CD197, CD363, CD218 and CD196 in the respective populations. Symbol code: End#6 cells from thymus (orange), Tregs from peripheral blood (blue), immature Tregs from thymus (brown) and naive CD4 SP T cell from peripheral blood (purple).

Further investigation revealed an increase in the expression of several other markers at the end of the trajectory: apart from CD127 and CD8 the cells gained expression of T-bet, CD95, TIGIT and CD152 (CTLA-4) (**Figure 6B**). In fact, End#6 cells, irrespective of CD8 expression, phenotypically resembled the previously described long-lived thymic regulatory cells, which have been reported to recirculate from the periphery back to the thymus (44,79–83).

To test the hypothesis that End#6 cells are recirculating Tregs homing into thymus, we measured the expression of key chemokine receptors (80,81,84–86) in the populations of interest (**Figure 6C**). In contrast with the thymic mature naive CD4 SP (84), End#6 cells as well as immature Treg cells lack the S1PR1 (CD363) necessary to egress from the thymus. On the other hand, End#6 cells express IL-18R (CD218), which has been reported to cause upregulation of CCR6 (CD196) upon stimulation with IL-18 (81). CCR6 can then mediate the entry of Treg cells into the thymus (81,85). End#6 cells expressed more CCR6 compared to CD4 SP naive T cells and immature Treg cells. The last measured chemokine receptor, CCR7 (CD197) has been previously reported to distinguish the newly formed Treg cells, which do express CCR7, from the recirculating Treg cells which do not express CCR7(80). Consistent with this, while our immature Treg cells express CCR7, the End#6 cells do not (**Figure 6C**).

Since the thymic Treg population identified in the study of Park et al. (45), interpreted by the authors as long-residing in the thymus (here referred to as Treg-atlas), was placed in a terminal developmental position similar to End#6 (**Figure 7A**), we reanalyzed their single-cell RNA-seq data using *tviblindi* (See Supplementary note, section 9, Analysis of human thymus single-cell RNA-seq data). See **Figure 7B** for the localization of Treg-atlas on the *vaevictis* plot and **Figure 7C** for the trajectory leading to Treg-atlas. The RNA expression profile of Treg-atlas corresponds to the protein profile of End#6 suggesting that we are analyzing the same population (**Figure 7D-F**, compare to **Figure 5B** and **Figure 6B**, C). Apart from markers already presented in our mass cytometry panels, Treg-atlas show a sharp increase in *IL-1R2* RNA (**Figure 7E**) further reinforcing the hypothesis that these cells are recirculating from the periphery (83,87–91). Furthermore, **Figure 7G** shows the expression of additional chemokine receptors known to play a role in guiding Treg cells to peripheral sites (tumors and other sites of inflammation). Thus, *tviblindi* detected the trajectory leading to developmentally advanced Treg cells in single-cell RNA-seq data as well as in mass cytometry protein data.

3. Discussion

We have developed *tviblindi*, a modular and interactive TI tool which allows a biologist working with real-world high-dimensional datasets to explore the development of cells from progenitors toward differentiated stages. *tviblindi* simulates possible random walks from a user-defined point of origin and highlights their endpoints in the dataset. When used in an automated mode, it creates a “connectome” graph that visually represents the dataset with all ends and the trajectories connecting them to the point of origin. While most current TI tools stop here, we prefer to follow up with expert interaction with the data, which allows us to focus in depth on selected discovered ends and trajectories. *tviblindi* allows selection and grouping of similar random walks into trajectories based on visual interaction with the data. For each trajectory, marker intensity changes can be shown as the trajectory pseudotime progresses. Also, the key points (e.g., branching) can be projected back onto the two-dimensional visual representation of the dataset. *tviblindi* does not anticipate a single endpoint nor a single branching point with two endpoints. Importantly, it can be used for both mass cytometry (millions of cells but only dozens of parameters) as well as single-cell RNA-seq (thousands of parameters) data.

By integrating several modules (random walks simulation, pseudotime inference, real-time interactive topological classification using persistent homology and two-dimensional visualization using the *vaevictis* algorithm), *tviblindi* presents a feasible algorithmic solution for TI, which

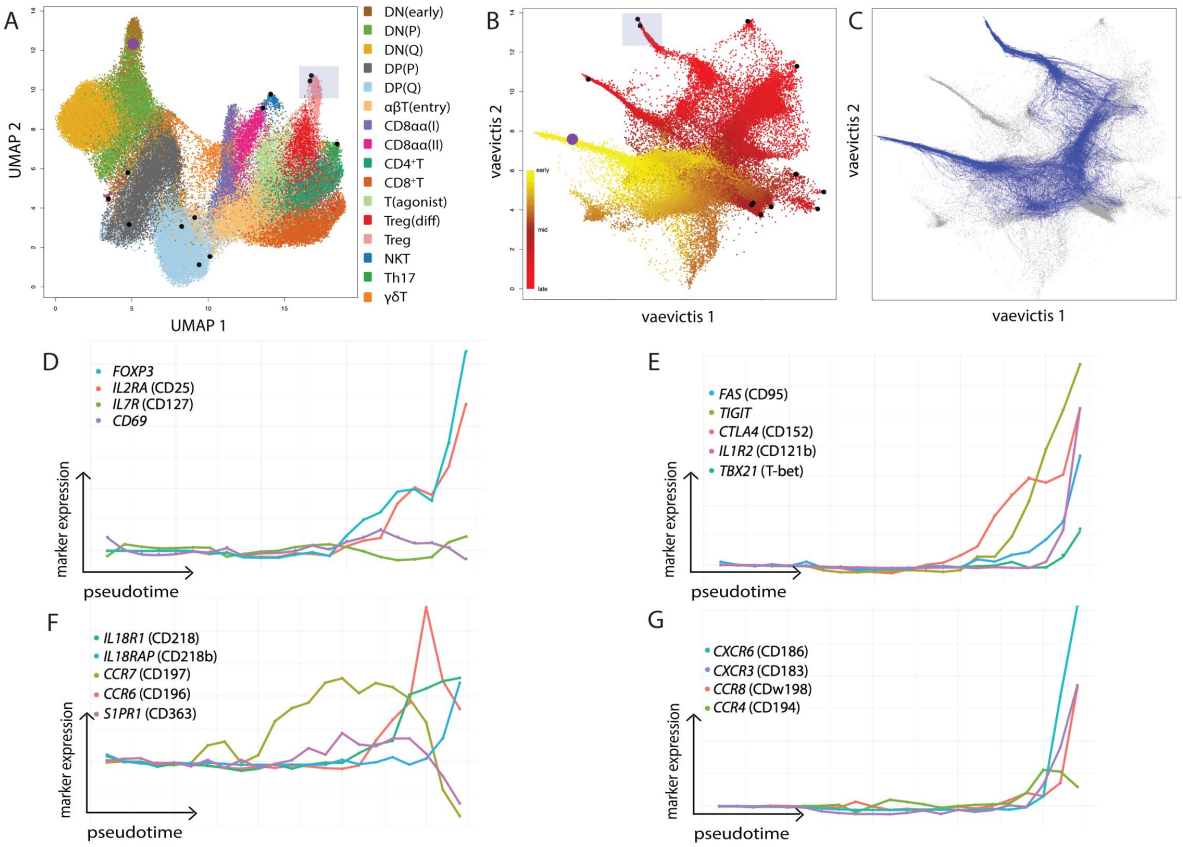

Figure 7

tviblindi analysis of single-cell RNA data from the study of Park et al. (45) showing the trajectory to Treg population corresponding to End#6.

For A and B, the origin (centroid of DN (early) population) is marked by a purple dot, the candidate endpoints with more than 1% of simulated random walks are indicated by black dots and the endpoints corresponding to the Treg-atlas are highlighted by a blue rectangle. (A) UMAP plot identical to the one shown in **Figure 3A** of published research by Park et al. (45) showing the populations using the annotations from the authors, legend at the bottom left. (B) *vaevictis* plot of the same data as in A with shown pseudotime represented by yellow-to-red color gradient. (C) *vaevictis* plot showing the trajectory leading to Treg-atlas. For (D-G), the individual markers are labeled in accordance with the original research by their gene names and the CD designated markers are shown in parentheses. (D) Pseudotime line plot of markers canonical to the Treg population. (E) Pseudotime line plot of markers associated with Treg activation overlapping with our mass cytometry data from End#6 cells shown in **Figure 6B**. (F) Pseudotime line plot of cytokine and chemokine receptors overlapping with our cytometry data from End#6 cells shown in **Figure 6C**. (G) Pseudotime line plot showing the expression of additional chemokine markers measured on Treg-atlas.

significantly extends upon existing TI algorithms. Its versatility and ability to interact with the user are its main contributions, creating a unique tool for a biologist to intuitively understand the dataset, visualize the data, and to focus the analysis on meaningful TI endpoints. An efficient discovery of the underlying dynamics is made possible by using persistent homology to aggregate random walks into trajectories, project them immediately onto a low-dimensional embedding (created using a custom dimensionality reduction algorithm: *vaevictis*) and visualize marker-vs-pseudotime line plots.

Importantly, the complete workflow has linear time complexity (See Supplementary note, subsection 8.3 Running times) and thus is manageable on a desktop computer in a reasonable computation time. We extensively compared its performance (See Supplementary Note, Sections 8 and 10 for details) to current TI algorithms and we found it to be superior in most criteria, as well as versatile in dealing with different datasets with widely varying numbers of cells and parameters investigated (single-cell RNA-seq or mass cytometry). It has been applied to publicly available datasets with success as well as to new data on B-cell development in health and immunodeficiency(92).

The dimensionality reduction is only used to visualize the data, and is not involved in any computation within the other modules. *vaevictis* reduction is designed to show gradual changes of single-cell phenotypes, suppressing the effect of abundant populations and showing local as well as global relationships of cells. This is in contrast to UMAP (52,53) and t-SNE (54,55), which prioritize clustering of locally similar cells at the expense of global relationships.

The *vaevictis* plot of mass cytometry data from thymus and peripheral blood faithfully depicted the basic scheme of T-cell development in this study and was also used to depict gradually changing data of T-ALL signaling (93)and B-cell development (92)(94).

When $CD34^{pos}$ progenitors were selected as the population of origin in our dataset (thymocytes and PBMC), the endpoints of trajectories were found within peripheral CD4 and CD8 cells, the apoptotic cells, and the population of $CD4^{pos}CD8^{dim}$ thymic cells which we further explored in this paper and which we refer to as End#6 cells.

The endpoints at the thymus/peripheral blood transition (#5 for CD4 SP and #9 for CD8 SP) reflect the accumulation of cells at the stage of mature naive T cells. From this stage activated T cell may progress to become memory T cells. Of note, certain biological endpoints might be skipped if the final population is too small (the endpoint cells are not sufficiently represented) or if their phenotyping markers are missing (e.g., Th1-type, Th2-type markers are not included in any of our mass cytometry panels).

Furthermore, in the absence of markers resolving similar but developmentally unrelated cells, a random walk can create a connection between these unrelated cells and skip from one trajectory to another. When these immunophenotypically undistinguishable cells themselves represent developmental fate of two distinct trajectories, the two trajectories may combine into a single trajectory, skipping the endpoint in the middle. This phenomenon has been observed in the apoptotic region where some random walks leading to cells dying due to the failure of DP cells to pass the positive/negative selection checkpoint continued past the apoptotic stage in a reverse direction towards SP cells undergoing negative selection. Here the algorithm correctly found the trajectories, but expert input (understanding that apoptosis is a one-way process) was essential for the interpretation of the direction.

While the *vaevictis* embedding plots the cells on a 2D plot in an intuitive manner, some inherent limitations of 2D embeddings to comprehensively capture complex topologies were manifested. An example is the failure to visualize positive/negative selection of CD8 SP T cells, which is clearly detected by *tviblindi.* Note that when focusing on the trajectories of cells undergoing positive

selection (V. Trajectories in **Figure 4**) in more detail, we have identified two groups of trajectories, one of them corresponding to positive selection (Va. trajectory) and one of them corresponding to an earlier β selection (cells entered the apoptotic pathway before expressing TCR β chain, Vb. trajectory). This highlights the benefit of expert-guided adjustment of the level of detail.

Investigation of the canonical CD4 and CD8 T-cell development reassured us that we can reconstruct the known sequence of maturation steps extensively described by CDMaps (95), connecting the thymic and peripheral T-cell compartment. We were able to dismiss artifacts caused by cell doublets. Putting expression levels of markers on a pseudotime axis showed us the proper sequence and intensity of expression changes. Note that, for the trajectory leading to CD4 SP cells, both CD4 and CD8 reach an expression peak at the DP stage, both decrease their intensity afterwards and then CD4 steadily increases towards the CD4 SP stage. CD4 expression level in peripheral SP cells is twice as high as in thymic SP cells. Key points of marker intensity changes can be traced back to the *vaevictis* 2D plots and groups of trajectories can be selected for further investigation (expression levels and 2D plot location).

A good example of the power of the presented approach was found when interpreting thymic Treg populations. The trajectory leading towards thymic End#6 cells diverges from the conventional CD4 trajectory at the location of negative selection. It goes through a CD45RA^{neg} immature Treg stage, characterized by the expression of CD25 in the absence of CD127 (conventional Treg immunophenotype), and continues towards the End#6 cells highly expressing CD25 and also expressing CD127. Both the immature Treg cells and the End#6 cells express Treg markers FOXP3 and Helios, confirming their Treg identity. The End#6 cells gain an expression of T-bet, CTLA-4, TIGIT, while they lose CD69, CD38 and CCR7. Surprisingly, they also gain a dim expression of CD8.

It has been shown that the ThPOK expression can be reduced in CD4 T cells, leading to reacquisition of CD8, and thus mimicking the DP immunophenotype. This was shown for CD8αα CD4 intraepithelial lymphocytes and CD8αβ CD4 cells in human blood as well as DP cells found in various tissues and under various disease settings. These have been reported to have enhanced cytotoxic functions or suppressive regulatory functions (96,97).

Although some of End#6 cells can fall within the conventional DP gate, the CD8 expression level is lower than that of the "bona fide" DP thymocytes and *tviblindi* positioned these CD4^{pos}CD8^{dim} thymocytes at the end of the developmental trajectory. We wanted to make sure that their terminal end position is correct and that they are truly developmentally more advanced than the conventional immature DP cells and the immature CD4 SP Treg cells. Indeed, the proliferation history of End#6 cells (both CD8^{neg} and CD8^{dim}) showed many more cell divisions than any other thymic subset, more than naive T cells in the periphery (by a measured dilution of TREC DNA in the sorted subsets). Our End#6 CD8^{dim} population may have been included in the FOXP3^{pos} DP cells placed at the beginning of Treg development in previous studies (39,78). Our approach provides a tool which can overcome the limitations of conventional gating and help to distinguish these cells.

However, we cannot claim that End#6 cells follow an uninterrupted developmental path from the precursors to End#6. In fact, the basic assumption of any TI method is that all developmental stages are present in the analyzed sample. If part of the development occurs in a peripheral organ which is not analyzed then an artificial shortcut can be mistaken for the whole path.

Indeed, the surface immunophenotype of End#6 cells agrees with the previously described population of peripheral Treg cells recirculating back to the thymus (42,82). These cells were reported to have the ability to regulate the development of their precursors either negatively, by competing for the limiting amount of IL-2 (82) or positively by blocking potentially harmful effects of the inflammatory cytokine IL-1 via their IL-1R2 decoy receptor (83). Furthermore,

End#6 cells correspond to the Cluster 3 Tregs from the previously described thymus (44), where scRNA and surface protein expression is suggestive of their recirculating mature effector Treg phenotype. This is in line with the low TREC DNA content in End#6 cells. Future experimental studies specifically designed to investigate this Treg population should address whether these cells result from recirculation or represent an additional path of intrathymic Treg effector maturation. The future studies could benefit from a *tviblindi* analysis of their immunophenotyping datasets.

In conclusion, we have built and validated a TI method that has generic use for mass cytometry, flow cytometry and single-cell RNA-seq. *tviblindi* is modular and interactive, thus combining the information strength of rich single-cell datasets with the power of expert knowledge for the interpretation of developmental trajectories.

4. Methods

4.1. Human cells

Thymocytes were obtained from healthy donor thymi of pediatric patients undergoing heart surgery at the Motol University Hospital, after informed consent was given by the patient guardians, in accordance with the Declaration of Helsinki. Peripheral blood mononuclear cells (PBMC) were obtained by density gradient centrifugation from buffy coats of healthy adult donors, at Institute of Hematology and Blood Transfusion in Prague. The pediatric PBMC were obtained from healthy pediatric patients undergoing pre-surgical examination for non-malignant and otherwise unrelated medical issue after consent was given by patients' guardians. Cells were stored in liquid nitrogen and thawed when needed for the experiment. After thawing, the cells were quickly transferred into 10 ml of pre-warmed RPMI supplemented with 10% FBS and 100µg of Pulmozyme (dornase alpha, Roche, Basel, Switzerland) for 30 min at 37°C. In order to remove dead cells and avoid clumping of the thymocytes, the cells were centrifuged at low speed 108g for 15 minutes. Subsequent steps were performed at 4°C.

4.2. Mass cytometry

4.2.1. Reagents

Unless otherwise stated, buffers and isotope labeled monoclonal antibodies (moAbs) were purchased from Standard BioTools Inc. (South San Francisco, USA). In-house antibody conjugations were performed using MaxPar labeling kits (Standard BioTools Inc.) according to the manufacturer's instructions. Carrier-protein-free moAbs were purchased from Exbio (Prague, Czech Republic), BioLegend (San Diego, USA), R&D Systems (Minneapolis, USA), eBioscience (San Diego, USA, part of Thermo Fisher Scientific Waltham, USA), Miltenyi Biotec (Bergisch Gladbach, Germany), Cell Signalling Technology (Danvers, USA) (for details see Supplementary Table Mass panels).

Antibody cocktails

The amount of labeled moAbs needed for one test was determined by titration. Two sets of antibody cocktails were prepared and frozen as described previously (98): 1) the antibody cocktail containing moAbs against surface antigens, which was applied prior to fixation and permeabilization steps, and 2) antibody cocktail containing moAbs against intracellular antigens, which was applied after the permeabilization.

4.2.2. Staining

Cells were washed in Annexin binding buffer, ABB (Exbio), and counted, approximately 5 million thymocytes and 1 million PBMC were used for one experiment.

To minimize technical variability, the individual tissues were barcoded with CD45 Abs conjugated with different isotopes (30 min, 4 °C). Biotin labeled Annexin V was added together with the barcoding Abs, to distinguish apoptotic cells with exposed phosphatidylserine residues. After this incubation, the sample volume was risen to 1 ml and cisplatin 198Pt was added to distinguish dead cells (5 min, 4 °C). Cells were then washed twice in ABB (92 g, 15 min, 4 °C) and the samples were pooled together, spun (92 g, 15 min, 4 °C) and processed further as one sample.

The surface Ab cocktail, including the isotope labeled anti-biotin mAb to visualize the annexin, was added to the cell pellet (total volume of surface Abs was approx. 100 µl, 30 min, 4 °C).

Cells were washed twice in ABB (92 g, 15 min, 4 °C), prefixed with 1 ml freshly diluted 1,6% PFA (10 min, RT) and washed once more in Maxpar® Cell Staining Buffer, CSB (1050 g, 5 min, 4 °C). Then, the cells were fixed and permeabilized using Maxpar® Nuclear Antigen Staining buffer (30 min, 4 °C) followed by two washings in Maxpar® Nuclear Antigen Staining Perm (1050 g, 5 min, 4 °C).

In cases when the FOXP3 antibody was used, the cells stained with the surface Abs and washed in ABB were fixed and permeabilized in FOXP3 fixation/permeabilization buffer (30 min, 4 °C) followed by two washings in FOXP3 permeabilization buffer (600 g, 5 min, 4 °C).

The intracellular Ab cocktail was added to the cell pellet, together with anti-TCRb frame antibody labeled with APC (it was not possible for us to get this antibody in the unlabeled carrier-free form allowing in-house isotope conjugation) (the total volume of intracellular Abs was approx. 60 µl, 30 min, 4°C). The cells were washed twice in CSB (1050 g, 5 min, 4 °C). Finally, an APC directed isotope labeled mAb was added (30 min, 4 °C), cells were washed twice in CSB (1050 g, 5 min, 4 °C), fixed in 1 ml freshly diluted 1.6% formaldehyde (10 min, RT) and centrifuged (1050 g, 5 min, 4 °C).

4.2.3. Mass cytometry samples acquisition

After incubation in 125nM 191/193Ir in Maxpar® Fix and Perm Buffer (at least 12 hours, up to 2 weeks), the cells were washed twice in CSB (1050 g, 5 min, RT) and once in Maxpar® Cell Acquisition Solution (CAS, 1050 g, 5 min, RT). Then they were diluted up to 1×10^6 cells/ml in 15 % EQ^{TM} Four Element Calibration Beads (Standard BioTools Inc.) in CAS and filtered through a 35 µm nylon mesh cell-strainer cap (BD Biosciences, San Jose, CA). The samples were acquired using Helios (Standard BioTools Inc.), CyTOF software version 6.7.1014. The instrument was tuned for acquisition using Tuning Solution (Standard BioTools Inc.) according to the manufacturer's instructions. The noise reduction (the cell length 7–150, lower convolution threshold 200) was applied during the acquisition. The signal was normalized using a Standard BioTools Inc. algorithm, which is based on the 'Bead Passport' concept.

4.2.4. Analysis of mass cytometry data

After export to listmode data (FCS format), files were further processed using FlowJo (v10, BD Biosciences). First, nucleated cells were gated by metal-tagged DNA intercalator 191/193Ir, and thymus versus peripheral blood was debarcoded by manual gating (Supplementary Figure 1). Next, PBMC were gated on CD3^{pos} lineage negative and TCR γδ negative T cells. The thymus cells were gated on lineage negative, viable (platinum negative) and TCR γδ negative thymocytes. Thus, only TCR αβ T cells and their progenitors in the thymus were used for further processing by the

tviblindi algorithm described later. When manual gating was performed prior to *tviblindi*, the FlowJo workspace file was used to save manual gate positions, so that the gated cells could be displayed in the GUI, to help the interpret the *vaevictis* plot.

The *tviblindi* algorithm was programmed in R, C++ and Python. The source code is deposited on GitHub (*https://github.com/stuchly/tviblindi*).

The listmode data was transformed using *arcsinh* with cofactor 5 for subsequent analysis (99). After data interrogation by *tviblindi*, an 'enhanced' FCS file was created, containing artificial channels with newly computed variables (2-dimensional embedding coordinates, pseudotime, pathway membership identifier, labels and selected points on trajectories) and a graphical output for figures was created using FlowJo layouts.

4.3. FACS sorting

Human thymocytes and PBMC obtained as described above for mass cytometry experiments were stained with appropriate amount of fluorescently labeled monoclonal antibodies, at 1 million cells/50 µl for 30 minutes on ice in the dark (see Supplementary Table FACS panel). After incubation, cells were washed in ice-cold PBS and sorted using BD FACS Aria III cell sorter into tubes with 10 µl of ice-cold Platinum Direct cell lysis buffer with Proteinase K (ThermoFisher Scientific). The obtained subsets (see Supplementary Figure 10) were then directly processed according to the manufacturer's instructions to minimize any loss of material.

4.4. TREC analysis

Extracted DNA was used for the quantification of T-cell receptor excision circles (TREC) and T cell receptor alpha constant gene region (TCRAC) via previously established real-time PCR assay (100,101). Estimated number of cell divisions was calculated based on the ratio between TREC and TCRAC copies in these populations.

All the background information on *tviblindi* design including used versions of R packages, dataset used for testing, dimensionality reduction, pseudotime estimation, topological clustering and biological interpretation of re-used datasets is available in the Supplementary note. Please, see the supplementary note also for detailed description and guide through the user interface and comparison with other state-of-the-art-methods (102–136).

Acknowledgements

T.K. and J.S. were financially supported by project NU23-07-00170 of the Czech Republic Ministry of Health. Institutional support was provided by the project National Institute for Cancer Research (Project No. LX22NPO5102), funded by the European Union–Next Generation EU. D.N. is a fellow of FWO (Fonds Wetenschappelijk Onderzoek – Vlaanderen), supported by Strategic Basic research grant 1S40423N.

Additional information

Author Contributions

Conceptualization, T.K., N.B., J.S.; Methodology, T.K., D.N., N.B., P.H., J.S., A.E.S., A.R.M.A; Software, D.N., V.S., A.I., G.A.S., J.S.; Formal Analysis, D.N., J.S.; Investigation, N.B., P.H., D.K., M.S., H.L.; Resources, P.E.; Writing – Original Draft, T.K., N.B., P.H., J.S.; Writing – Review & Editing, D.N., D.K., A.E.S., A.R.M.A.; Supervision, T.K., J.S.

Additional files

Supplementary Table Mass panel 1

Supplementary Video 1

Supplementary Video 2

Supplementary figures

Supplementary note

References

1. Bendall S. C., et al. (2014) **Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development** *Cell* **157**:714–725

2. Saelens W., Cannoodt R., Todorov H., Saeys Y (2019) **A comparison of single-cell trajectory inference methods** *Nature Biotechnology 2019* **37**:547–554

3. Cao J., et al. (2019) **The single-cell transcriptional landscape of mammalian organogenesis** *Nature 2019* **566**:496–502

4. Chen H., et al. (2019) **Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM** *Nature Communications 2019* **10**:1–14

5. Wolf F. A., et al. (2019) **PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells** *Genome Biol* **20**:1–9

6. Setty M., et al. (2019) **Characterization of cell fate probabilities in single-cell data with Palantir** *Nature Biotechnology 2019* **37**:451–460

7. Stassen S. V., Yip G. G. K., Wong K. K. Y., Ho J. W. K., Tsia K. K (2021) **Generalized and scalable trajectory inference in single-cell omics data with VIA** *Nature Communications 2021* **12**:1–18

8. Packer J. S., et al. (2019) **A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution** *Science (1979)* **365**

9. Briggs J. A., et al. (2018) **The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution** *Science (1979)* **360**

10. Cao J., et al. (2017) **Comprehensive single-cell transcriptional profiling of a multicellular organism** *Science (1979)* **357**:661–667

11. Litviňuková M., et al. (2020) **Cells of the adult human heart** *Nature 2020* **588**:466–472

12. Farley A. M., et al. (2013) **Dynamics of thymus organogenesis and colonization in early human development** *Development* **140**:2015–2026

13. Haddad R., et al. (2006) **Dynamics of Thymus-Colonizing Cells during Human Development** *Immunity* **24**:217–230

14. Lavaert M., et al. (2020) **Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes** *Immunity* **52**:1088–1104

15. Le J., et al. (2020) **Single-Cell RNA-Seq Mapping of Human Thymopoiesis Reveals Lineage Specification Trajectories and a Commitment Spectrum in T Cell Development** *Immunity* **52**:1105–1118

16. Hosokawa H., Rothenberg E. V (2020) **How transcription factors drive choice of the T cell fate** *Nature Reviews Immunology 2020* **21**:162–176

17. Canté-Barrett K., et al. (2017) **Loss of CD44dim expression from early progenitor cells marks T-cell lineage commitment in the human thymus** *Front Immunol* **8**

18. Dik W. A., et al. (2005) **New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling** *Journal of Experimental Medicine* **201**:1715–1723

19. Surh C. D., Sprent J (1994) **T-cell apoptosis detected in situ during positive and negative selection in the thymus** *Nature 1994* **372**:100–103

20. Kimura M. Y., et al. (2016) **Timing and duration of MHC I positive selection signals are adjusted in the thymus to prevent lineage errors** *Nature Immunology 2016* **17**:1415–1423

21. Klein L., Robey E. A., Hsieh C. S (2018) **Central CD4+ T cell tolerance: deletion versus regulatory T cell differentiation** *Nature Reviews Immunology 2018* **19**:7–18

22. Derbinski J., Schulte A., Kyewski B., Klein L (2001) **Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self** *Nature Immunology 2001* **2**:1032–1039

23. Anderson M. S., et al. (2002) **Projection of an immunological self shadow within the thymus by the aire protein** *Science (1979)* **298**:1395–1401

24. Moran A. E., et al. (2011) **T cell receptor signal strength in Treg and iNKT cell development demonstrated by a novel fluorescent reporter mouse** *Journal of Experimental Medicine* **208**:1279–1289

25. Caton A. J., et al. (2014) **Strength of TCR signal from self-peptide modulates autoreactive thymocyte deletion and Foxp3+ Treg-cell formation** *Eur J Immunol* **44**:785–793

26. Lee H. M., Bautista J. L., Scott-Browne J., Mohan J. F., Hsieh C. S (2012) **A Broad Range of Self-Reactivity Drives Thymic Regulatory T Cell Selection to Limit Responses to Self** *Immunity* **37**:475–486

27. Tai X., et al. (2023) **How autoreactive thymocytes differentiate into regulatory versus effector CD4+ T cells after avoiding clonal deletion** *Nature Immunology 2023* **24**:637–651

28. Khailaie S., Robert P. A., Toker A., Huehn J., Meyer-Hermann M (2014) **A signal integration model of thymic selection and natural regulatory T cell commitment** *J Immunol* **193**:5983–5996

29. Le Borgne M., et al. (2009) **The impact of negative selection on thymocyte migration in the medulla** *Nature Immunology 2009* **10**:823–830

30. Hsieh C. S., Lee H. M., Lio C. W. J (2012) **Selection of regulatory T cells in the thymus** *Nature Reviews Immunology 2012* **12**:157–167

31. Coquet J. M., et al. (2013) **Epithelial and dendritic cells in the thymic medulla promote CD4+Foxp3+ regulatory T cell development via the CD27–CD70 pathway** *Journal of Experimental Medicine* **210**:715–728

32. Mahmud S. A., et al. (2014) **Costimulation via the tumor-necrosis factor receptor superfamily couples TCR signal strength to the thymic differentiation of regulatory T cells** *Nature Immunology 2014* **15**:473–481

33. Liu Y., et al. (2008) **A critical function for TGF-β signaling in the development of natural CD4+CD25+Foxp3+ regulatory T cells** *Nature Immunology 2008* **9**:632–640

34. D'Cruz L. M., Klein L (2005) **Development and function of agonist-induced CD25+Foxp3+ regulatory T cells in the absence of interleukin 2 signaling** *Nature Immunology 2005* **6**:1152–1159

35. Owen D. L., Sjaastad L. E., Farrar M. A (2019) **Regulatory T Cell Development in the Thymus** *The Journal of Immunology* **203**:2031–2041

36. Burchill M. A., Yang J., Vogtenhuber C., Blazar B. R., Farrar M. A (2007) **IL-2 Receptor β-Dependent STAT5 Activation Is Required for the Development of Foxp3+ Regulatory T Cells** *The Journal of Immunology* **178**:280–290

37. Lio C. W. J., Hsieh C. S (2008) **A Two-Step Process for Thymic Regulatory T Cell Development** *Immunity* **28**:100–111

38. Caramalho I., et al. (2015) **Human regulatory T-cell development is dictated by Interleukin-2 and −15 expressed in a non-overlapping pattern in the thymus** *J Autoimmun* **56**:98–110

39. Vanhanen R., Tuulasvaara A., Mattila J., Pätilä T., Arstila T. P (2018) **Common gamma chain cytokines promote regulatory T cell development and survival at the CD4+ CD8+ stage in the human thymus** *Scand J Immunol* **88**:e12681

40. Golding A., Darko S., Wylie W. H., Douek D. C., Shevach E. M (2017) **Deep sequencing of the TCR-β repertoire of human forkhead box protein 3 (FoxP3)+ and FoxP3– T cells suggests that they are completely distinct and non-overlapping** *Clin Exp Immunol* **188**:12–21

41. Owen D. L., et al. (2019) **Thymic regulatory T cells arise via two distinct developmental programs** *Nature Immunology 2019* **20**:195–205

42. Santamaria J. C., Borelli A., Irla M (2021) **Regulatory T Cell Heterogeneity in the Thymus: Impact on Their Functional Activities** *Front Immunol* **12**

43. Heimli M., et al. (2023) **Multimodal human thymic profiling reveals trajectories and cellular milieu for T agonist selection** *Front Immunol* **13**

44. Morgana F., et al. (2022) **Single-Cell Transcriptomics Reveals Discrete Steps in Regulatory T Cell Development in the Human Thymus** *The Journal of Immunology* **208**:384–395

45. Park J. E., et al. (2020) **A cell atlas of human thymic development defines T cell repertoire formation** *Science (1979)* **367**

46. Setty M., et al. (2016) **Wishbone identifies bifurcating developmental trajectories from single-cell data** *Nature Biotechnology 2016* **34**:637–645

47. Wei S. C., et al. (2019) **Negative Co-stimulation Constrains T Cell Differentiation by Imposing Boundaries on Possible Cell States** *Immunity* **50**:1084–1098

48. Chopp L. B., et al. (2020) **An Integrated Epigenomic and Transcriptomic Map of Mouse and Human αβ T Cell Development** *Immunity* **53**:1182–1201

49. Silva V. de, Carlsson G. (2004) **Topological estimation using witness complexes** In: *Symposium on Point Based Graphics 04* https://doi.org/10.2312/SPBG/SPBG04/157-166

50. Edelsbrunner Herbert, Harer John L. (2010) **Computational Topology: An Introduction** American Mathematical Society

51. Pokorny F., Hawasly M., Ramamoorthy S (2014) **Multiscale Topological Trajectory Classification with Persistent Homology** In: *Robotics: Science and Systems 2014* https://doi.org/10.15607/RSS.2014.X.054

52. Becht E., et al. (2018) **Dimensionality reduction for visualizing single-cell data using UMAP** *Nature Biotechnology 2018* **37**:38–44

53. McInnes L., Healy J., Saul N., Großberger L (2018) **UMAP: Uniform Manifold Approximation and Projection** *J Open Source Softw* **3**

54. Amir E. A. D., et al. (2013) **viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia** *Nat Biotechnol* **31**

55. Maaten L. van der, Hinton G (2008) **Visualizing Data using t-SNE** *Journal of Machine Learning Research* **9**:2579–2605

56. Ding J., Condon A., Shah S. P (2018) **Interpretable dimensionality reduction of single cell transcriptome data with deep generative models** *Nat Commun* **9**

57. Szubert B., Cole J. E., Monaco C., Drozdov I (2019) **Structure-preserving visualisation of high dimensional single-cell datasets** *Sci Rep* **9**

58. Förster Y. P., Gamberi L., Tzanis E., Vivo P., Annibale A (2022) **Exact and approximate mean first passage times on trees and other necklace structures: a local equilibrium approach** *J Phys A Math Theor* **55**

59. Pokorny F., Hawasly M., Ramamoorthy S (2014) **Multiscale Topological Trajectory Classification with Persistent Homology** In: *Robotics: Science and Systems 2014* https://doi.org/10.15607/RSS.2014.X.054

60. Edelsbrunner H., Harer J. (2010) **Computational Topology: An Introduction. Computational Topology** Miscellaneous Books

61. de Silva V, Carlsson G (2004) **Topological estimation using witness complexes** In: *Symposium on Point Based Graphics 04* http://diglib.eg.org/handle/10.2312/SPBG.SPBG04.157-166

62. Blondel V. D., Guillaume J. L., Lambiotte R., Lefebvre E (2008) **Fast unfolding of communities in large networks** *Journal of Statistical Mechanics: Theory and Experiment 2008*

63. Cao J., et al. (2019) **The single-cell transcriptional landscape of mammalian organogenesis** *Nature* **566**:496–502

64. Chen H., et al. (2019) **Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM** *Nat Commun* **10**:1903

65. Setty M., et al. (2019) **Characterization of cell fate probabilities in single-cell data with Palantir** *Nat Biotechnol* **37**:451–460

66. Stassen S. V., Yip G. G. K., Wong K. K. Y., Ho J. W. K., Tsia K. K (2021) **Generalized and scalable trajectory inference in single-cell omics data with VIA** *Nat Commun* **12**:5528

67. Wolf F. A., et al. (2019) **PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells** *Genome Biol* **20**

68. Weiler P., Lange M., Klein M., Pe'er D., Theis F (2024) **CellRank 2: unified fate mapping in multiview single-cell data** *Nat Methods* **21**:1196–1205

69. Stassen S. V., et al. (2024) **StaVia: spatially and temporally aware cartography with higher-order random walks for cell atlases** *Genome Biol* **25**

70. Zhang X., et al. (2024) **An immunophenotype-coupled transcriptomic atlas of human hematopoietic progenitors** *Nat Immunol* **25**:703–715

71. Pijuan-Sala B., et al. (2019) **A single-cell molecular map of mouse gastrulation and early organogenesis** *Nature* **566**:490–495

72. Lee H. M., Hsieh C.-S (2009) **Rare Development of Foxp3+ Thymocytes in the CD4+CD8+ Subset** *The Journal of Immunology* **183**:2261–2266

73. Mittelstadt P. R., Taves M. D., Ashwell J. D (2019) **Glucocorticoids Oppose Thymocyte Negative Selection by Inhibiting Helios and Nur77** *The Journal of Immunology* **203**:2163–2170

74. Daley S. R., Hu D. Y., Goodnow C. C (2013) **Helios marks strongly autoreactive CD4+ T cells in two major waves of thymic deletion distinguished by induction of PD-1 or NF-κB** *J Exp Med* **210**:269–285

75. Tuovinen H., Pekkarinen P. T., Rossi L. H., Mattila I., Arstila T. P (2008) **The FOXP3+ subset of human CD4+CD8+ thymocytes is immature and subject to intrathymic selection** *Immunol Cell Biol* **86**:523–529

76. Lehtoviita A., Rossi L. H., Kekäläinen E., Sairanen H., Arstila T. P (2009) **The CD4+CD8+ and CD4+ Subsets of FOXP3+ Thymocytes Differ in their Response to Growth Factor Deprivation or Stimulation** *Scand J Immunol* **70**:377–383

77. Vanhanen R., Leskinen K., Mattila I. P., Saavalainen P., Arstila T. P (2020) **Epigenetic and transcriptional analysis supports human regulatory T cell commitment at the CD4+CD8+ thymocyte stage** *Cell Immunol* **347**

78. Nunes-Cabaço H., Caramalho Í., Sepúlveda N., Sousa A. E (2011) **Differentiation of human thymic regulatory T cells at the double positive stage** *Eur J Immunol* **41**:3604–3614

79. Yang E. J., Zou T., Leichner T. M., Zhang S. L., Kambayashi T (2014) **Both retention and recirculation contribute to long-lived regulatory T-cell accumulation in the thymus** *Eur J Immunol* **44**:2712–2720

80. Cowan J. E., McCarthy N. I., Anderson G (2016) **CCR7 Controls Thymus Recirculation, but Not Production and Emigration, of Foxp3+ T Cells** *Cell Rep* **14**:1041–1048

81. Peligero-Cruz C., et al. (2020) **IL-18 signaling promotes homing of mature tregs into the thymus** *eLife* **9**:1–23

82. Thiault N., et al. (2015) **Peripheral regulatory T lymphocytes recirculating to the thymus suppress the development of their precursors** *Nature Immunology 2015* **16**:628–634

83. Nikolouli E., et al. (2020) **Recirculating IL-1R2+ Tregs fine-tune intrathymic Treg development under inflammatory conditions** *Cellular & Molecular Immunology 2020* **18**:182–193

84. Allende M. L., Dreier J. L., Mandala S., Proia R. L (2004) **Expression of the Sphingosine 1-Phosphate Receptor, S1P1, on T-cells Controls Thymic Emigration** *Journal of Biological Chemistry* **279**:15396–15401

85. Cowan J. E., et al. (2018) **Aire controls the recirculation of murine Foxp3+ regulatory T-cells back to the thymus** *Eur J Immunol* **48**:844–854

86. Matloubian M., et al. (2004) **Lymphocyte egress from thymus and peripheral lymphoid organs is dependent on S1P receptor 1** *Nature 2004* **427**:355–360

87. Weaver J. D., et al. (2022) **Differential expression of CCR8 in tumors versus normal tissue allows specific depletion of tumor-infiltrating T regulatory cells by GS-1811, a novel Fc-optimized anti-CCR8 antibody** *Oncoimmunology* **11**

88. De Simone M., et al. (2016) **Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells** *Immunity* **45**:1135–1147

89. Vila-Caballer M., et al. (2019) **Disruption of the CCL1-CCR8 axis inhibits vascular Treg recruitment and function and promotes atherosclerosis in mice** *J Mol Cell Cardiol* **132**:154–163

90. Gao Y., et al. (2022) **Intratumoral stem-like CCR4+ regulatory T cells orchestrate the immunosuppressive microenvironment in HCC associated with hepatitis B** *J Hepatol* **76**:148–159

91. Oldham K. A., et al. (2012) **T Lymphocyte Recruitment into Renal Cell Carcinoma Tissue: A Role for Chemokine Receptors CXCR3, CXCR6, CCR5, and CCR6** *Eur Urol* **61**:385–394

92. Bakardjieva M., et al. (2024) **Tviblindi algorithm identifies branching developmental trajectories of human B-cell development and describes abnormalities in RAG-1 and WAS patients** *Eur J Immunol* :e2451004 https://doi.org/10.1002/eji.202451004

93. Kuzilková D., et al. (2022) **Either IL-7 activation of JAK-STAT or BEZ inhibition of PI3K-AKT-mTOR pathways dominates the single-cell phosphosignature of ex vivo treated pediatric T-cell acute lymphoblastic leukemia cells** *Haematologica* **107**:1293–1310

94. Thiel J., et al. (2024) **Defects in B-lymphopoiesis and B-cell maturation underlie prolonged B-cell depletion in ANCA-associated vasculitis** *Ann Rheum Dis ard-2024-225587* https://doi.org/10.1136/ard-2024-225587

95. Kalina T., et al. (2019) **CD maps - dynamic profiling of CD1–CD100 surface expression on human leukocyte and lymphocyte subsets** *Front Immunol* **10**

96. Overgaard N. H., Jung J.-W., Steptoe R. J., Wells J. W (2015) **CD4+/CD8+ double-positive T cells: more than just a developmental stage?** *J Leukoc Biol* **97**:31–38

97. Taniuchi I (2018) **CD4 Helper and CD8 Cytotoxic T Cell Differentiation** *Annu Rev Immunol* **36**:579–601

98. Schulz A. R., et al. (2019) **Stabilizing Antibody Cocktails for Mass Cytometry** *Cytometry A* **95**:910–916

99. Nowicka M., et al. (2019) **CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets** *F1000Research 2019* **6**:748

100. Froňková E., et al. (2014) **The TREC/KREC assay for the diagnosis and monitoring of patients with DiGeorge syndrome** *PLoS One* **9**

101. Douek D. C., et al. (2000) **Assessment of thymic output in adults after haematopoietic stemcell transplantation and prediction of T-cell reconstitution** *The Lancet* **355**:1875–1881

102. Bailey E. (2022) **CRAN - Package shinyBS** https://cran.r-project.org/web/packages/shinyBS/index.html

103. Bates D., Eddelbuettel D (2013) **Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package** *J Stat Softw* **52**:1–24

104. Bates D., Maechler M., Jagan M., Davis T. A. (2023) **Matrix: Sparse and Dense Matrix Classes and Methods** https://CRAN.R-project.org/package=Matrix

105. Bauer U., Kerber M., Reininghaus J., Wagner H (2017) **Phat - Persistent Homology Algorithms Toolbox** *J. Symb. Comput* **78**:76–90

106. Cannoodt R., Saelens W. (2022) **dynverse/dyntoy: Generating simple toy data of cellular differentiation version 0.9.9 from GitHub** https://rdrr.io/github/dynverse/dyntoy/

107. Carlsson G., Ishkhanov T., De Silva V., Zomorodian A (2008) **On the local behavior of spaces of natural images** *Int J Comput Vis* **76**:1–12

108. Csárdi G., Nepusz T. (2006) **The igraph software package for complex network research**

109. de Vries A., Ripley B. D. (2022) **ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'** *CRAN* https://doi.org/10.32614/CRAN.package.ggdendro

110. Ding J., Condon A., Shah S. P (2018) **Interpretable dimensionality reduction of single cell transcriptome data with deep generative models** *Nature Communications 2018* **9**:1–13

111. Eddelbuettel D., Emerson J. W., Kane M. J. (2023) **CRAN - Package BH** https://cran.r-project.org/web/packages/BH/index.html

112. Eddelbuettel D., Sanderson C (2014) **RcppArmadillo: Accelerating R with high-performance C++ linear algebra** *Comput Stat Data Anal* **71**:1054–1063

113. Eddelbuettel D., et al. (2023) **BH: Boost C++ Header Files** https://cran.r-project.org/web/packages/Rcpp/index.html

114. Ellis B., et al. (2023) **flowCore: flowCore: Basic structures for flow cytometry data** *Bioconductor* https://doi.org/10.18129/B9.bioc.flowCore

115. Galili T. (2015) **dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering** *Bioinformatics* **31**:3718–3720

116. Hao Y., et al. (2021) **Integrated analysis of multimodal single-cell data** *Cell* **184**:3573–3587

117. Hatcher A (2000) **Algebraic topology** Cambridge University Press

118. Huber W., et al. (2015) **Orchestrating high-throughput genomic analysis with Bioconductor** *Nature Methods 2015* **12**:115–121

119. Kachanovich S. (2015) **Witness complex. GUDHI User and Reference Manual, GUDHI Editorial Board** http://gudhi.gforge.inria.fr/doc/latest/groupwitnesscomplex.html

120. Kulichova T., Kratochvil M. (2023) **Scatterplots with More Points** *CRAN* 1.2 https://doi.org/10.32614/CRAN.package.scattermore

121. Maechler M. (2022) **cluster: Cluster Analysis Basics and Extensions** https://cran.r-project.org/web/packages/Matrix/index.html

122. Melville J., Lun A., Djekidel M. N., Hao Y., Eddelbuettel D. (2023) **uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction, r package version 0.1.16** https://cran.r-project.org/web/packages/uwot/index.html

123. Meyer F., Perrier V. (2022) **shinybusy: Busy Indicators and Notifications for 'Shiny' Applications** https://cran.r-project.org/web/packages/shinybusy/index.html

124. Pedersen T. L., Nijs V., Schaffner T., Nantz E. (2022) **shinyFiles: A Server-Side File System Viewer for Shiny** https://cran.r-project.org/web/packages/shinyFiles/index.html

125. Perrier V., Meyer F., Granjon D., Fellows I., Matthews S. (2023) **shinyWidgets: Custom Inputs Widgets for Shiny** https://cran.r-project.org/web/packages/shinyWidgets/index.html

126. Rouvreau V. (2015) **Aplha complex. GUDHI User and Reference Manual, GUDHI Editorial Board**

127. Sali A., Hass L., Attalli D. (2020) **shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating** https://cran.r-project.org/web/packages/shinycssloaders/index.html

128. Seel M. (2019) **CGAL 5.6 - dD Geometry Kernel: User Manual** https://doc.cgal.org/latest/Kernel_d/index.html

129. The CGAL Project (2023) **CGAL User and Reference Manual**

130. Urbanek S. (2022) **jpeg: Read and write JPEG images** https://cran.r-project.org/web/packages/jpeg/index.html

131. Ushey K., et al. (2023) **reticulate: Interface to 'Python'** *CRAN* https://doi.org/10.32614/CRAN.package.reticulate

132. van Dijk D., et al. (2018) **Recovering Gene Interactions from Single-Cell Data Using Data Diffusion** *Cell* **174**:716–729

133. Warnes G. R., et al. (2022) **gplots: Various R Programming Tools for Plotting Data** https://cran.r-project.org/web/packages/gplots/index.html

134. Wickham H., Seidel D. (2022) **scales: Scale Functions for Visualization** https://cran.r-project.org/web/packages/scales/index.html

135. Wickham H., et al. (2019) **Welcome to the Tidyverse** *J Open Source Softw* **4**:1686

136. Wilke C. O. (2020) **cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'** *CRAN* https://doi.org/10.32614/CRAN.package.cowplot

Author information

Jan Stuchly[$]
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

For correspondence: jan.stuchly@lfmotol.cuni.cz
[$]Contributed equally

David Novak[$]
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic, Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[$]Contributed equally

Nadezda Brdickova[$]
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

[$]Contributed equally

Petra Hadlova
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

Vojen Sadilek
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

Ahmad Iksi
Centre d'Immunophénomique - CIPHE (PHENOMIN), Aix Marseille Université (UMS3367), Inserm (US012), CNRS (UAR3367), Marseille, France

Daniela Kuzilkova
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

Michael Svaton
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

George Alehandro Saad
Centre d'Immunophénomique - CIPHE (PHENOMIN), Aix Marseille Université (UMS3367), Inserm (US012), CNRS (UAR3367), Marseille, France

Pablo Engel
Department of Biomedical Sciences, Medical School, University of Barcelona, Barcelona, Spain

Herve Luche
Centre d'Immunophénomique - CIPHE (PHENOMIN), Aix Marseille Université (UMS3367), Inserm (US012), CNRS (UAR3367), Marseille, France

Ana E Sousa
Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

Afonso RM Almeida
Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

Tomas Kalina
Childhood Leukaemia Investigation Prague (CLIP), Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

For correspondence: tomas.kalina@lfmotol.cuni.cz

Editors

Reviewing Editor
Frederik Graw
Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany

Senior Editor
Aleksandra Walczak
CNRS, Paris, France

Reviewer #1 (Public review):

The authors present tviblindi, an algorithm to infer cell development trajectories from single-cell molecular data. The paper is well-written and the algorithm is conceptually interesting. However, the validation is incomplete as the comparison against existing trajectory inference

methods is weak: although the lack of a proper benchmark was pointed out as the main weakness of the original version of the manuscript, the revised version still only contains qualitative comparisons against state-of-the-art methods.

Both me and Reviewer 2 pointed out that the lack of a proper benchmark against state-of-the-art methods on a wider variety of datasets (including scRNA-seq data) was a major weakness of the original version of the manuscript. In response to this criticism, the authors now did the following:

- They ran various competitor methods on the datasets that were used already for the previous version of the manuscript.
- They ran tviblindi and two of the competitors on two public scRNA-seq datasets.
- For all datasets, they qualitatively assessed the trajectories computed by tviblindi and its competitors and argued that tviblindi's trajectories better reflect the biological signal in the data.
- The results of all of these additional analyses are reported in the supplement, which has now become very lengthy (88 pages).

In my opinion, this is insufficient to establish that tviblindi is comparable or even superior to the state of the art in the field. To show that this is the case, the authors would have to carry out a systematic benchmark study which relies on quantitative evaluation metrics rather than on qualitative intepretations of trajectories. As method developers, we are all susceptive to confirmation bias when comparing our new algorithms to the state of the art. To avoid this pitfall, reporting quantitative performance metrics is required. At the moment, the only quantitative metric reported by the authors is runtime, which is insufficient.

Moreover, the results of a benchmark study should be reported in the main manuscript, not in the supplement. When presenting a new algorithm in a field as crowded as trajectory inference, a benchmark against the state of the art serves to establish trust in the new algorithm and to provide the readers with a rationale to use it for their research. For this, the results of the benchmark have to be presented prominently and should not be hidden in the supplement.

A second major criticism raised in Reviewer 2's review of the original version of the manuscript is that tviblindi invites cherry picking due to its inherently interactive design. In response to this, the authors now argue at length that "the data-driven expert interpretation approach of tviblindi" (quote from Section 2.2.2) is a strength rather than a weakness. If we concede for the sake of the argument that tviblindi's "expert interpretation approach" is indeed a strength of the method (although I tend to agree with Reviewer 2 that it is rather a limitation), usability for biologists becomes critical. However, given the current implementation of tviblindi, its usability is far from optimal. The authors do not provide tviblindi as a web interface that is directly usable for domain experts without programming experience and not even as a package that is installable via some widely used package manager such as conda. Instead, they implemented tviblindi as an R package with a Shiny GUI that can either run in a Docker container or requires the installation of several dependencies. I therefore strongly doubt that many biologists will be able or willing to run tviblindi, which substantially limits the value of its "expert interpretation approach". Moreover, tviblindi does not support Apple silicon, which prevented also myself from testing the tool.

https://doi.org/10.7554/eLife.95861.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***eLife Assessment***
>
> *The authors present an algorithm and workflow for the inference of developmental trajectories from single-cell data, including a mathematical approach to increase computational efficiency. While such efforts are in principle useful, the absence of benchmarking against synthetic data and a wide range of different single-cell data sets make this study incomplete. Based on what is presented, one can neither ultimately judge if this will be an advance over previous work nor whether the approach will be of general applicability.*

We thank the eLife editor for the valuable feedback. Both benchmarking against other methods and validation on a synthetic dataset ("dyntoy") are indeed presented in the Supplementary Note, although this was not sufficiently highlighted in the main text, which has now been improved.

Our manuscript contains benchmarking against a challenging synthetic dataset in Figure 1; furthermore, both the synthetic dataset and the real-world thymus dataset have been analyzed in parallel using currently available TI tools (as detailed in the Supplementary Note). z other single-cell datasets (single-cell RNA-seq) were added in response to the reviewers' comments.

One of the reviewers correctly points out that *tviblindi* goes against the philosophy of automated trajectory inference. This is correct; we believe that a new class of methods, complementary to fully automated approaches, is needed to explore datasets with unknown biology. *tviblindi* is meant to be a representative of this class of methods—a semi-automated framework that builds on features inferred from the data in an unbiased and mathematically well-founded fashion (pseudotime, homology classes, suitable low-dimensional representation), which can be used in concert with expert knowledge to generate hypotheses about the underlying dynamics at an appropriate level of detail for the particular trajectory or biological process.

We would also like to mention that the algorithm and the workflow are not the sole results of the paper. We have thoroughly characterized human thymocyte development, where, in addition to expected biological endpoints, we found and characterized an unexpected activated thymic T-reg endpoint.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *The authors present tviblindi, a computational workflow for trajectory inference from molecular data at single-cell resolution. The method is based on (i) pseudo-time inference via expecting hitting time, (ii) sampling of random walks in a directed acyclic k-NN where edges are oriented away from a cell of origin w.r.t. the involved nodes' expected hitting times, and (iii) clustering of the random walks via persistent homology. An extended use case on mass cytometry data shows that tviblindi can be used elucidate the biology of T cell development.*
>
> *Strengths:*
>
> *- Overall, the paper is very well written and most (but not all, see below) steps of the tviblindi algorithm are explained well.*

> *- The T cell biology use case is convincing (at least to me: I'm not an immunologist, only a bioinformatician with a strong interest in immunology).*

We thank the reviewer for feedback and suggestions that we will accommodate, we respond point-by-point below

> *Weaknesses:*
>
> *- The main weakness of the paper is that a systematic comparison of tviblindi against other tools for trajectory inference (there are many) is entirely missing. Even though I really like the algorithmic approach underlying tviblindi, I would therefore not recommend to our wet-lab collaborators that they should use tviblindi to analyze their data. The only validation in the manuscript is the T cell development use case. Although this use case is convincing, it does not suffice for showing that the algorithms's results are systematically trustworthy and more meaningful (at least in some dimension) than trajectories inferred with one of the many existing methods.*

We have compared *tviblindi* to several trajectory inference methods (Supplementary note section 8.2: Comparison to state-of-the-art methods, namely Monocle3 (v1.3.1) Cao et al. (2019), Stream (v1.1) Chen et al. (2019), Palantir (v1.0.0) Setty et al. (2019), VIA (v0.1.89) Stassen et al. (2021), StaVia (Via 2.0) Stassen et al. (2024), CellRank 2 (v2.06) Weiler et al. (2024) and PAGA (scanpy==1.9.3) Wolf et al. (2019). We added thorough and systematic comparisons to the other algorithms mentioned by reviewers. We included extended evaluation on publicly available datasets (Supplementary Note section 10).

Also, in the meantime we have successfully used *tviblindi* to investigate human B-cell development in primary immunodeficiency (Bakardjieva M, et al. Tviblindi algorithm identifies branching developmental trajectories of human B-cell development and describes abnormalities in RAG-1 and WAS patients. Eur J Immunol. 2024 Dec;54(12):e2451004. doi: 10.1002/eji.202451004.).

> *- The authors' explanation of the random walk clustering via persistent homology in the Results (subsection "Real-time topological interactive clustering") is not detailed enough, essentially only concept dropping. What does "sparse regions" mean here and what does it mean that "persistent homology" is used? The authors should try to better describe this step such that the reader has a chance to get an intuition how the random walk clustering actually works. This is especially important because the selection of sparse regions is done interactively. Therefore, it's crucial that the users understand how this selection affects the results. For this, the authors must manage to provide a better intuition of the maths behind clustering of random walks via persistent homology.*

In order to satisfy both reader types: the biologist and the mathematician, we explain the mathematics in detail in the Supplementary Note, section 4. We improved the Results text to better point the reader to the mathematical foundations in the Supplementary Note.

> *- To motivate their work, the authors write in the introduction that "TI methods often use multiple steps of dimensionality reduction and/or clustering, inadvertently introducing bias. The choice of hyperparameters also fixes the a priori resolution in a way that is difficult to predict." They claim that tviblindi is better than the original methods because "analysis is performed in the original high-dimensional space, avoiding artifacts of dimensionality reduction." However, in the manuscript, tviblindi is tested only on mass cytometry data which has a much lower dimensionality than scRNA-seq data for which most existing trajectory inference methods are designed. Since tviblindi works on a k-NN graph representation of the input data, it is unclear if it could be run on scRNA-seq data*

> *without prior dimensionality reduction. For this, cell-cell distances would have to be computed in the original high-dimensional space, which is problematic due to the very high dimensionality of scRNA-seq data. Of course, the authors could explicitly reduce the scope of tviblindi to data of lower dimensionality, but this would have to be stated explicitly.*

In the manuscript we tested the framework on the scRNA-seq data from Park et al 2020 (DOI: 10.1126/science.aay3224). To illustrate that *tviblindi* can work directly in the high-dimensional space, we applied the framework successfully on imputed 2000 dimensional data. Furthermore we successfully used *tviblindi* to investigate bone marrow atlas scRNA-Seq dataset Zhang et al. (2024) and atlas of mouse gastrulation Pijuan-Sala et al. (2019). The idea behind *tviblindi* is to be able to work without the necessity to use non-linear dimensionality reduction techniques, which reduce the dimensionality to a very low number of dimensions and whose effects on the data distribution are difficult to predict. On the other hand the use of (linear) dimensionality reduction techniques which effectively suppress noise in the data such as PCA is a good practice (see also response to reviewer 2). We have emphasized this in the revised version and added the results of the corresponding analysis (see Supplementary note, section 9).

> *- Also tviblindi has at least one hyper-parameter, the number k used to construct the k-NN graphs (there are probably more hidden in the algorithm's subroutines). I did not find a systematic evaluation of the effect of this hyper-parameter.*

Detailed discussion of the topic is presented in the Supplementary Note, section 8.1, where Spearman correlation coefficient between pseudotime estimated using k=10 and k=50 nearest neighbors was 0.997. The number k however does affect the number of candidate endpoints. But even when larger k causes spurious connection between unrelated cell fates, the topological clustering of random walks allows for the separation of different trajectories. We have expanded the “sensitivity to hyperparameters” section 8.1 also in response to reviewer 2.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *In Deconstructing Complexity: A Computational Topology Approach to Trajectory Inference in the Human Thymus with tviblindi, Stuchly et al. propose a new trajectory inference algorithm called tviblindi and a visualization algorithm called vaevictis for single-cell data. The paper utilizes novel and exciting ideas from computational topology coupled with random walk simulations to align single cells onto a continuum. The authors validate the utility of their approach largely using simulated data and establish known protein expression dynamics along CD4/CD8 T cell development in thymus using mass cytometry data. The authors also apply their method to track Treg development in single-cell RNA-sequencing data of human thymus.*
>
> *The technical crux of the method is as follows: The authors provide an interactive tool to align single cells along a continuum axis. The method uses expected hitting time (given a user input start cell) to obtain a pseudotime alignment of cells. The pseudotime gives an orientation/direction for each cell, which is then used to simulate random walks. The random walks are then arranged/clustered based on the sparse region in the data they navigate using persistent homology.*

We thank the reviewer for feedback and suggestions that we have accommodated, we responded point-by-point below.

> *The notion of using persistent homology to group random walks to identify trajectories in the data is novel.*
>
> *The strength of the method lies in the implementation details that make computationally demanding ideas such as persistent homology more tractable for large scale single-cell data. This enables the authors to make the method more user friendly and interactive allowing real-time user query with the data.*
>
> *Weaknesses:*
>
> *The interactive nature of the tool is also a weakness, by allowing for user bias leading to possible overfitting for a specific data.*

tviblindi is not designed as a fully automated TI tool (although it implements a fully automated module), but as a data driven framework for exploratory analysis of unknown data. There is always a risk of possible bias in this type of analysis - starting with experimental design, choice of hyperparameters in the downstream analysis, and an expert interpretation of the results. The successful analysis of new biological data involves a great deal of expert knowledge which is difficult to a priori include in the computational models.

tvilblindi tries to solve this challenge by intentionally overfitting the data and keeping the level of resolution on a single random walk. In this way we aim to capture all putative local relationships in the data. The on-demand aggregation of the walks using the global topology of the data allows researchers to use their expert knowledge to choose the right level of detail (as demonstrated in the Figure 4 of the manuscript) while relying on the topological structure of the high dimensional point cloud. At all times *tviblindi* allows to inspect the composition of the trajectory to assess the variance in the development, possible hubs on the KNN-graph etc.

> *The main weakness of the method is lack of benchmarking the method on real data and comparison to other methods. Trajectory inference is a very crowded field with many highly successful and widely used algorithms, the two most relevant ones (closest to this manuscript) are not only not benchmarked against, but also not sited. Including those that specifically use persistent homology to discover trajectories (Rizvi et.al. published Nat Biotech 2017). Including those that specifically implement the idea of simulating random walks to identify stable states in single-cell data (e.g. CellRank published in Lange et.al Nat Meth 2022), as well as many trajectory algorithms that take alternative approaches. The paper has much less benchmarking, demonstration on real data and comparison to the very many other previous trajectory algorithms published before it. Generally speaking, in a crowded field of previously published trajectory methods, I do not think this one approach will compete well against prior work (especially due to its inability to handle the noise typical in real world data (as was even demonstrated in the little bit of application to real world data provided).*

We provided comparisons of *tviblindi* and *vaevictis* in the Supplementary Note, section 8.2, where we compare it to Monocle3 (v1.3.1) Cao et al. (2019), Stream (v1.1) Chen et al. (2019), Palantir (v1.0.0) Setty et al. (2019), VIA (v0.1.89) Stassen et al. (2021), StaVia (Via 2.0) Stassen et al. (2024), CellRank 2 (v2.06) Weiler et al. (2024) and PAGA (scanpy==1.9.3) Wolf et al. (2019). We added thorough and systematic comparisons to the other algorithms mentioned by reviewers. We included extended evaluation on publicly available datasets (Supplementary Note section 10).

> *Beyond general lack of benchmarking there are two issues that give me particular concern. As previously mentioned, the algorithm is highly susceptible to user bias and overfitting. The paper gives the example (Figure 4) of a trajectory which mistakenly shows that cells may pass from an apoptotic phase to a different developmental stage. To*

> *circumvent this mistake, the authors propose the interactive version of tviblindi that allows users to zoom in (increase resolution) and identify that there are in fact two trajectories in one. In this case, the authors show how the author can fix a mistake when the answer is known. However, the point of trajectory inference is to discover the unknown. With so much interactive options for the user to guide the result, the method is more user/bias driven than data-driven. So a rigorous and quantitative discussion of robustness of the method, as well as how to ensure data-driven inference and avoid over-fitting would be useful.*

Local directionality in expression data is a challenge which is not, to our knowledge, solved. And we are not sure it can be solved entirely, even theoretically. The random walks passing "through" the apoptotic phase are biologically infeasible, but it is an (unbiased) representation of what the data look like based on the diffusion model. It is a property of the data (or of the panel design), which has to be interpreted properly rather than a mistake. Of note, except for Monocle3 (which does not provide the directionality) other tested methods did not discover this trajectory at all.

The "zoom in" has in fact nothing to do with "passing through the apoptosis". We show how the researcher can investigate the suggested trajectory to see if there is an additional structure of interest and/or relevance. This investigation is still data driven (although not fully automated). Anecdotally in this particular case this branching was discovered by a bioinformatician, who knew nothing about the presence of beta-selection in the data.

We show that the trajectory of apoptosis of cortical thymocytes consists of 2 trajectories corresponding to 2 different checkpoints (beta-selection and positive/negative selection). This type of a structure, where 2 (or more) trajectories share the same path for most of the time, then diverge only to be connected at a later moment (immediately from the point of view of the beta-selection failure trajectory) is a challenge for TI algorithms and none of tested methods gave a correct result. More importantly there seems to be no clear way to focus on these kinds of structures (common origin and common fate) in TI methods.

Of note, the "zoom in" is a recommended and convenient method to look for an inner structure, but it does not necessarily mean addition of further homological classes. Indeed, in this case the reason that the structure is not visible directly is the limitation of the dendrogram complexity (only branches containing at least 10% of simulated random walks are shown by default). In summary, *tviblindi* effectively handled all noise in the data that obscured biologically valid trajectories for other methods. We have improved the discussion of the robustness in the current version.

> *Second, the paper discusses the benefit of tviblindi operating in the original high dimensions of the data. This is perhaps adequate for mass cytometry data where there is less of an issue of dropouts and the proteins may be chosen to be large independent. But in the context of single-cell RNA-sequencing data, the massive undersampling of mRNA, as well as high degree of noise (e.g. ambient RNA), introduces very large degree of noise so that modeling data in the original high dimensions leads to methods being fit to the noise. Therefore ALL other methods for trajectory inference work in a lower dimension, for very good reason, otherwise one is learning noise rather than signal. It would be great to have a discussion on the feasibility of the method as is for such noisy data and provide users with guidance. We note that the example scRNA-seq data included in the paper is denoised using imputation, which will likely result in the trajectory inference being oversmoothed as well.*

We agree with the reviewer. In our manuscript we wanted to showcase that *tviblindi* can directly operate in high-dimensional space (thousands of dimensions) and we used MAGIC imputation for this purpose. This was not ideal. More standard approach, which uses 30-50

PCs as input to the algorithm resulted in equivalent trajectories. We have added this analysis to the study (Supplementary note, section 9).

In summary, the fact that *tviblindi* scales well with dimensionality of the data and is able to work in the original space does not mean that it is always the best option. We have added a corresponding comment into the Supplementary note.

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *Stuchly et al. proposed a single-cell trajectory inference tool, tviblindi, which was built on a sequential implementation of the k-nearest neighbor graph, random walk, persistent homology and clustering, and interactive visualization. The paper was organized around the detailed illustration of the usage and interpretation of results through the human thymus system.*
>
> *Strengths:*
>
> *Overall, I found the paper and method to be practical and needed in the field. Especially the in-depth, step-by-step demonstration of the application of tviblindi in numerous T cell development trajectories and how to interpret and validate the findings can be a template for many basic science and disease-related studies. The videos are also very helpful in showcasing how the tool works.*
>
> *Weaknesses:*
>
> *I only have a few minor suggestions that hopefully can make the paper easier to follow and the advantage of the method to be more convincing.*
>
> *(1) The "Computational method for the TI and interrogation - tviblindi" subsection under the Results is a little hard to follow without having a thorough understanding of the tviblindi algorithm procedures. I would suggest that the authors discuss the uniqueness and advantages of the tool after the detailed introduction of the method (moving it after the "Connectome - a fully automated pipeline".*

We thank the reviewer for the suggestion and we have accommodated it to improve readability of the text.

> *Also, considering it is a computational tool paper, inevitably, readers are curious about how it functions compared to other popular trajectory inference approaches. I did not find any formal discussion until almost the end of the supplementary note (even that is not cited anywhere in the main text). Authors may consider improving the summary of the advantages of tviblindi by incorporating concrete quantitative comparisons with other trajectory tools.*

We provided comparisons of *tviblindi* and *vaevictis* in the Supplementary Note, section 8.2, where we compare it to Monocle3 (v1.3.1) Cao et al. (2019), Stream (v1.1) Chen et al. (2019), Palantir (v1.0.0) Setty et al. (2019), VIA (v0.1.89) Stassen et al. (2021), StaVia (Via 2.0) Stassen et al. (2024), CellRank 2 (v2.06) Weiler et al. (2024) and PAGA (scanpy==1.9.3) Wolf et al. (2019). We added thorough and systematic comparisons to the other algorithms mentioned by reviewers. We included extended evaluation on publicly available datasets (Supplementary Note section 10).

> *(2) Regarding the discussion in Figure 4 the trajectory goes through the apoptotic stage and reconnects back to the canonical trajectory with counterintuitive directionality, it can be a checkpoint as authors interpret using their expert knowledge, or maybe a false*

> *discovery of the tool. Maybe authors can consider running other algorithms on those cells and see which tracks they identify and if the directionality matches with the tviblindi.*

We have indeed used the thymus dataset for comparison of all TI algorithms listed above. Except for Monocle 3 they failed to discover the negative selection branch (Monocle 3 does not offer directionality information). Therefore, a valid topological trajectory with incorrect (expert-corrected) directionality was partly or entirely missed by other algorithms.

> *(3) The paper mainly focused on mass cytometry data and had a brief discussion on scRNA-seq. Can the tool be applied to multimodality data such as CITE-seq data that have both protein markers and gene expression? Any suggestions if users want to adapt to scATAC-seq or other epigenomic data?*

The analysis of multimodal data is the logical next step and is the topic of our current research. At this moment *tviblindi* cannot be applied directly to multimodal data. It is possible to use the KNN-graph based on multimodal data (such as weighted nearest neighbor graph implemented in Seurat) for pseudotime calculation and random walk simulation. However, we do not have a fully developed triangulation for the multimodal case yet.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Suggestions for improved or additional experiments, data or analyses:*
>
> *- Benchmark against existing trajectory inference methods.*
>
> *- Benchmark on scRNA-seq data or an explicit statement that, unlike existing methods, tviblindi is not designed for such data.*

We provided comparisons of *tviblindi* and *vaevictis* in the Supplementary Note, section 8.2, where we compare it to Monocle3 (v1.3.1) Cao et al. (2019), Stream (v1.1) Chen et al. (2019), Palantir (v1.0.0) Setty et al. (2019), VIA (v0.1.89) Stassen et al. (2021), StaVia (Via 2.0) Stassen et al. (2024), CellRank 2 (v2.06) Weiler et al. (2024) and PAGA (scanpy==1.9.3) Wolf et al. (2019). We added thorough and systematic comparisons to the other algorithms mentioned by reviewers. We included extended evaluation on publicly available datasets (Supplementary Note section 10).

> *- Systematic evaluation of the effetcs of hyper-parameters on the performance of tviblindi (as mentioned above, there is at least one hyper-parameter, the number k to construct the k-NN graphs).*

This is described in Supplementary Note section 8.1

> *Recommendations for improving the writing and presentation:*
>
> *- The GitHub link to the algorithm which is currently hidden in the Methods should be moved to the abstract and/or a dedicated section on code availability.*
>
> *- The presentation of the persistent homology approach used for random walk clustering should be improved (see public comment above).*

This is described extensively in Supplementary Note

> *- A very minor point (can be ignored by the authors): consider renaming the algorithm. At least for me, it's extremely difficult to remember.*

We choose to keep the original name

> *Minor corrections to the text and figures:*
>
> *- Labels and legend texts are too small in almost all figures.*
>
> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(1) On page 3: "(2) Analysis is performed in the original high-dimensional space avoiding artifacts of dimensionality reduction." In mass cytometry data where there is no issue of dropouts, one may choose proteins such that they are not correlated with each other making dimensionality reduction techniques less relevant. But in the context of an unbiased assays such as single-cell RNA-sequencing (scRNA-seq), one measures all the genes in a cell so dimensionality reduction can help resolve the redundancy in the feature space due to correlated/co-regulated gene expression patterns. This assumption forms the basis of most methods in scRNA-seq. More importantly, in scRNA-seq data the dropouts and ambient molecules in mRNA counts result in so much noise that modeling cells in the full gene expression is highly problematic. So the authors are requested to discuss in detail how they would propose to deal with noise in scRNA-seq data.*
>
> *On this note, the authors mention in Supplementary Note 9 (Analysis of human thymus single-cell RNA-seq data): "Imputed data are used as the input for the trajectory inference, scaled counts (no imputation) are shown in line plots". The line plots indicate the gene expression trends along the obtained pseudotime. The authors use MAGIC to impute the data, and we request the authors to mention this in the Methods section (currently one must look through the code on Supplementary Note 1.3 to find this). Data imputation in single-cell RNA-seq data are intended to enable quantification of individual gene expression distribution or pairwise gene associations. But when all the genes in an imputed data are used for visualization, clustering or trajectory inference, the averaging effect will compound and result in severely smoothed data that misses important differences between cell states. Especially, in the case of MAGIC, which uses a transition matrix raised to a power, it is over-smoothing of the data to use a transition matrix smoothed data to obtain another transition matrix to calculate the hitting time (or simulate random walks). Second, the authors' proposal to use scaled counts to study gene trends cannot be generalized to other settings due to drop out issue. Given the few genes (and only one branch) that are highlighted in Figure 7D-G and Figure 31 in Supplementary Note, it is hard to say if scaling raw values would pick up meaningful biology robustly here for other branches.*
>
> *We recommend that this data be reanalyzed with non-imputed data used for trajectory inference and imputed gene expression used for line plots.*

As stated above in the public review, we reanalyzed the scRNA Seq data using a more standard approach (first 50 principal components). We have also analyzed two additional scRNA Seq datasets (Section 1 and section 10 of Supplementary Note)

> *On the same note, the authors use Seurat's CellCycleScoring to obtain the cell cycle phase of each cell and later use ScaleData to regress them out. While we agree that it is valuable to remove cell cycle effect from the data for trajectory inference (and has been used previously in other methods), the regression approach employed in Seurat's ScaleData is not appropriate. It is an aggressive approach that severely changes expression pattern of many genes and can result in new artifacts (false positives) in the data. We recommend the authors to explore this more and consider using a more principled alternatives such as fscLVM (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1334-8).*

Cell cycle correction is an open problem (Heumos, Nat Rev Genetics, 2023)

Here we use an (arguably aggressive) approach to make the presentation more straightforward. The cells we are interested here (end #6) are not dividing and the regression does not change the conclusion drawn in the paper

> *(2) The figures provided are extremely low in resolution that it is practically impossible to correctly interpret a lot of the conclusion and references made in the figure (especially Figure 3 in the main text).*

Resolution of the Figures was improved

> *(3) There are many aspects of the method that enable easy user biases and can lead to substantial overfitting of the data.*
>
> *a. On page 7: "The topology of the point cloud representing human T-cell development is more complex ... and does not offer a clear cutoff for the choice of significant sparse regions. Interactive selection allows the user to vary the resolution and to investigate specific sparse regions in the data iteratively." This implies that the method enables user biases to be introduced into the data analysis. While perhaps useful for exploration, quantitative trajectory assessment using such approach can be faulty when the user (A) may not know the underlying dynamics (B) forces preconceived notion of trajectory.*
>
> *The authors should consider making the trajectory inference approach less dependent on interactive user input and show that the trajectory results are robust to any choices the user may make. It may also help if the authors provide an effective guide and mention clearly what issues could result due to the use of such thresholds.*

As explained in the response in public reviews, *tviblindi* is not designed as a fully automated TI tool, but as a data driven framework for exploratory analysis of unknown data.

There is always a risk of possible bias in this type of analysis - starting with experimental design, choice of hyperparameters in the downstream analysis, and an expert interpretation of the results. The successful analysis of new biological data involves a great deal of expert knowledge which is difficult to a priori include in the computational models. To specifically address the points raised by the reviewer:

“(A) may not know the underlying dynamics” - *tviblindi* is designed to perform exploratory analysis of the unknown underlying dynamics. We showcase in the study how this can be performed and we highlight possible cases which can be resolved expertly (spurious connections (doublets), different scales of resolution (beta selection)). Crucially, compared to other TI methods, *tviblindi* offers a clear mechanism on how to discover, focus and resolve these issues which would (and do) contaminate the trajectories discovered fully automatically by tested methods (cf. the beta selection, or the development of plasmacytoid dendritic cells (PDCs) (Supplementary note, section 10.1).

“(B) forces preconceived notion of trajectory” - user interaction in *tviblindi* does not force a preconceived notion of the trajectory. The random walks are simulated before the interactive step in an unbiased manner. During the interactive step the user adjusts trajectory specific resolution - incorrect choice of the resolution may result in either merging distinct trajectories into one or over separating the trajectories (which is arguably much less serious). However the interactive step is designed to deal with exactly this kind of challenge. We showcase (e.g. beta selection, or PDCs development) how to address the issue - *tviblindi* allows us to investigate deeper structure in any considered trajectory.

Thus, *tviblindi* represents a new class of methods that is complementary to fully automated trajectory inference tools. It offers a semi-automated tool that leverages features derived from data in an unbiased and mathematically rigorous manner, including pseudotime, homology classes, and appropriate low-dimensional representations. These can be integrated with expert knowledge to formulate hypotheses regarding the underlying dynamics, tailored to the specific trajectory or biological process under investigation.

> *b. In Figure 4, the authors discuss the trajectory of cells emanating from CD3 negative double positive stage and entering apoptotic phase and mention tviblindi may give "the false impression that cells may pass through an apoptotic phase into a later developmental stage" and propose that the interactive version of tviblindi can help user zoom into (increase resolution) this phenomenon and identify that there are in fact two trajectories in one. Given this, how do the other trajectories in the data change if a user manually adjusts the resolution? A quantification of the robustness is important. Also, it appears that a more careful data clean up could avoid such pitfalls where the algorithm infers trajectory based on mixed phenotype and the user would not have to manually adjust the resolution to obtain clear biological conclusion. We not that the original publication of this data did such "data clean up" using simple diffusion map based dimensionality reduction which the authors boast they avoid. There is a reason for this dimensionality reduction (distinguishing signal from noise), even in CyTOF data, let alone its importance in single cell data.*

The reviewer is concerned about two different, but intertwined issues we wish to untangle here. First, data clean-up is typically done on the premise that dead cells are irrelevant and they are a source of false signals. In the case of the thymocytes in the human thymus this premise is not true. Apoptotic cells are a legitimate (actually dominant) fate of the development and thus need to be represented in the TI dataset. Their biological behavior is however complex as they stop expressing proteins and thus lose their surface markers gradually, as dictated by the particular protein degradation kinetics. So can we clean up dead and dying cells better? Yes, but we don't want to do it since we would lose cells we want to analyze. Second, do trajectories change when we zoom into the data? No, only the level of detail presented visually changes. Since we calculate 5000 trajectories in the dataset, we need to aggregate them already for the hierarchical clustering visualization. Note that Figure 4, panel A highlights 159 trajectories selected in V. group. Zooming in means that the hierarchy of trajectories within V. group is revealed (panel D, groups V.a and Vb.) and can be interpreted on the *vaevictis* and lineplot graphs (panel E, F).

> *c. In the discussion, the authors write "[tviblindi] allows the selection and grouping of similar random walks into trajectories based on visual interaction with the data". This counters the idea of automated trajectory inference and can lead to severe overfitting.*

As explained in reply to Q3, our aim was NOT to create a fully automated trajectory inference tool. Even more, in our experience we realized that all current tools are taking this fully automated approach with a search for an “ideal” set of hyperparameters. This, in our experience, leads to a “blackbox” tool that is difficult to interpret for the expert in the biological field. To respond to this need we designed a modular approach where the results of the TI are presented and the expert can interact with them to focus the visualization and to derive interpretation. Our interactive concept is based on 15 years of experience with the data analysis in flow cytometry, where neither manual gating nor full automation is the ultimate solution but smart integration of both approaches eventually wins the game.

Thus, tviblindi represents a new class of methods that is complementary to fully automated trajectory inference tools. It offers a semi-automated tool that leverages features derived from data in an unbiased and mathematically rigorous manner. These features include

pseudotime, homology classes, and appropriate low-dimensional representations. These features can be integrated with expert knowledge to formulate hypotheses regarding the underlying dynamics, tailored to the specific trajectory or biological process under investigation.

> *d. The authors provide some comment on the robustness to the relaxation parameter for witness complex construction in Supplementary Note Section 8.1.2 but it is limited given the importance of this parameter and a more thorough investigation is recommended. We request the authors to provide concrete examples with figures of how changing alpha2 parameter leads to simplicial complexes of different sizes and an assessment of contexts in which the parameter is robust and when not (in both simulated and publicly available real data). Of note, giving the users a proper guide for parameter choice based on these examples and offering them ways to quantify robustness of their results may also be valuable.*

Section 8 in Supplementary Note was extended as requested.

> *e. The authors are requested for an assessment of possible short-circuits (e.g. cells of two distantly related phenotypes that get connected erroneously in the trajectory) in the data, and how their approach based on persistent homology deals with it.*

If a short circuit results in a (spurious) alternative trajectory, the persistent homology approach allows us to distinguish it from genuine trajectories that do not follow the short circuit. This prevents contamination of the inferred evolution by erroneous connections. The ability to distinguish and separate distinct trajectories with the same fate is a major strength of this approach (e.g., the trajectory through doublets or the trajectories around checkpoints in thymocytes' evolution).

> *(4) The authors propose vaevictis as a new visualization tool and show its performance compared to the standard UMAP algorithm on a simulated data set (Figure 1 in Supplementary Notes). We recommend a more comprehensive comparison between the two algorithms on a wide array of publicly available single-cell datasets. As well as comparison to other popular dimensionality reduction approaches like force directed layouts, which are the most widely used tool specifically to visualize trajectories.*

We added Section 10 to Supplementary Note that presents multiple comparisons of this kind. It is important to note that tviblindi works independently of visualization and any preferred visualization can be used in the interactive phase (multiple visualisation methods are implemented).

> *(5) In Supplementary Note 8.2, the authors compare tviblindi against the other methods. We recommend the authors to quantify the comparison or expand on their assesments in real biological data. For example, in comparison against Palantir and VIA the authors mention "... discovers candidate endpoints in the biological dataset but lacks toolbox to interrogate subtle features such as complex branching" and "fails to discover subtle features (such as Beta selection)" respectively. We recommend the authors to make these comparisons more precise or provide quantification. While the added benefit of interactive sessions of tviblindi may make it more user friendly, the way tviblindi appears to enable analysis of subtle features (e.g. Figure 1H) should be possible in Palantir or VIA as well.*

We extended the comparisons and presented them in Section 8 and 10 in Supplementary Note.

> *(6) The notion of using random walk simulations to identify terminal (and initial states) has been previously used in single-cell data (CellRank algorithm: https://www.nature.com/articles/s41592-021-01346-6). We request the authors to compare their approach to CellRank.*

We compared our algorithm to the CellRank successor CellRank 2 (see section 8.2, Supplementary Note)

> *(7) The notion of using persistent homology to discover trajectories has been previously used in single cell data https://pubmed.ncbi.nlm.nih.gov/28459448/. we request a comparison to this approach*

The proposed algorithm was not able to accommodate the large datasets we used.

scTDA (Rizvi, Camara et al. Nat. Biotechnol. 2017) has not been updated for 6 years. It is not suited for complex atlas-sized datasets both in terms of performance and utility, with its limited visualization tools. It also lacks capabilities to analyze individual trajectories.

> *(8) In Figure 3B, the authors visualize the endpoints and simulated random walks using the connectome. There is no edge from start to the apoptotic cells here. It is not clear why? If they are not relevant based on random walks, can the user remove them from analysis? Same for the small group of pink cells below initial point.*

The connectome is a fully automated approach (similar to PAGA) which gives a basic overview of the data. It is not expected to be able to compete with the interactive pipeline of tviblindi for the same reasons as the fully automated methods (difficult to predict the effect of hyperparameters).

> *(9) In Supplementary Figure 3, in relation to "Variants of trajectories including selection processes" the author mention that there is a spurious connection between CD4 single positive, and the doublet set of cells. The authors mention that the presence of dividing cells makes it difficult to remove the doublets. We request the authors to discuss why. For example, the authors seem to have cell cycle markers (e.g. Ki67, pH3, Cyclin) and one would think that coupled with DNA intercalator 191/193Ir one could further clean-up the data. Can the authors employ alternative toolkits such as doublet detection methods?*

To address this issue, we do remove doublets with illegitimate cell barcodes (e.g. we remove any two cells from two samples with different barcode which present with double barcode). Although there are computational doublet removal approaches for mass cytometry (Bagwell, Cytometry A 2020), mostly applied to peripheral blood samples (where cell division is not present under steady state immune system conditions), these are however not well suited for situations where dividing samples occur (Rybakowska P, Comput Struct Biotechnol J. 2021), which is the case of our thymocyte samples. Furthermore, there are other situations where doublet formation is not an accident, but rather a biological response (Burel JG, Cytometry A (2020). Thus, the doublet cell problem is similar to the apoptotic cell problem discussed earlier.

We could remove cells with the double DNA signal, but this would remove not only accidental doublets but also the legitimate (dividing) cells. So the question is how to remove the illegitimate doublets but not the legitimate?

Of note, the trajectory going through doublets does not affect the interpretation of other trajectories as it is readily discriminated by persistent homology and thus random walks

passing through this (spurious) trajectory do not contaminate the markers' evolution inferred for legitimate trajectories.

We therefore prefer to remove only the barcode illegitimate and keep all others in analysis, using the expert analysis step also to identify (using the cell cycle markers plus other features) the artificially formed doublets and thus spurious connections.

> *(10) The authors should discuss how the gene expression trend plots are made (e.g. how are the expression averaged? Rolling mean?).*

The development of those markers is shown as a line plot connecting the average values of a specific marker within a pseudotime segment. By default, the pseudotime values are divided into uniform segments (each containing the same number of points) whose number can be changed in the GUI. To focus on either early or late stages of the development, the segment division can be adjusted in GUI. See section 6 of the Supplementary Note.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *The overall figures quality needs to be improved. For example, I can barely see the text in Figure 3c.*

Resolution of the Figures was improved

https://doi.org/10.7554/eLife.95861.2.sa0