scientific reports

Check for updates

OPEN

Understanding learner behaviour in online courses with Bayesian modelling and time series characterisation

Robert L. Peach[1,5]✉, Sam F. Greenbury[1,2], Iain G. Johnston[3], Sophia N. Yaliraki[4], David J. Lefevre[5] & Mauricio Barahona[1]✉

The intrinsic temporality of learning demands the adoption of methodologies capable of exploiting time-series information. In this study we leverage the sequence data framework and show how data-driven analysis of temporal sequences of task completion in online courses can be used to characterise personal and group learners' behaviors, and to identify critical tasks and course sessions in a given course design. We also introduce a recently developed probabilistic Bayesian model to learn sequential behaviours of students and predict student performance. The application of our data-driven sequence-based analyses to data from learners undertaking an on-line Business Management course reveals distinct behaviors within the cohort of learners, identifying learners or groups of learners that deviate from the nominal order expected in the course. Using course grades a posteriori, we explore differences in behavior between high and low performing learners. We find that high performing learners follow the progression between weekly sessions more regularly than low performing learners, yet within each weekly session high performing learners are less tied to the nominal task order. We then model the sequences of high and low performance students using the probablistic Bayesian model and show that we can learn engagement behaviors associated with performance. We also show that the data sequence framework can be used for task-centric analysis; we identify critical junctures and differences among types of tasks within the course design. We find that non-rote learning tasks, such as interactive tasks or discussion posts, are correlated with higher performance. We discuss the application of such analytical techniques as an aid to course design, intervention, and student supervision.

Learning is a process that occurs over time[1]: new knowledge is built upon existing knowledge, suggesting that we should incorporate a temporal dimension in the analysis of learning data. The availability of time-stamped data has improved substantially with the advent of computer-based education platforms[2], coupled with advances in educational data mining and learning analytics that provide new and innovative tools for automated analysis. However, despite this increase in available tools, temporal analyses are still understudied in education research[1,3]. It is therefore important to expand the toolbox of temporal methodologies that are constructed to extract actionable information directly from the observed data without over-simplifying the learning process. For instance, the superiority of distributed learning over massed learning inherently posits the benefits of a consistent learning behaviour over time instead of an irregular, syncopated approach to learning[4]. However, the distinction into only two such broad behaviours is an over-simplification, and a number of subtly different temporal behaviours have been shown to exist in real data[5]. In general, the wide availability of data opens the possibility to introduce more powerful methodologies that do not reduce observations to predetermined, broad categories, and aim to extract more nuanced results and conclusions.

In educational data mining, the most common approach to temporal analysis is to describe each student through a *feature vector* composed of a small selection of aggregated features from time-series data (e.g., time-gap

[1]Department of Mathematics, Imperial College London, London, UK. [2]NIHR Imperial Biomedical Research Centre, ITMAT Data Science Group, Imperial College London, London, UK. [3]Department of Mathematics, University of Bergen, Bergen, Norway. [4]Department of Chemistry, Imperial College London, London, UK. [5]Imperial College Business School, Imperial College London, London, UK. ✉email: r.peach13@imperial.ac.uk; m.barahona@imperial.ac.uk

between sessions, participation time, number of sessions, post views)[6–8]. These feature vectors can be combined with additional non-temporal information, such as grades or demographic information. The position of each vector (i.e., each student) in feature space can be further analysed using supervised or unsupervised learning algorithms to make predictions or to investigate the structure of the data[9]. For instance[10], extracted features aggregated over the course of a week (video views and active learning days combined with features such as country of origin) to predict drop-out rates. Similarly, one study selected four predictive features (video-skipping, assignment skipping, lag, and assignment performance) to predict drop out rates[11]. Feature analyses have also been highly successful at predicting stop out 1-week in advance[12], yet the analysis of posting patterns of students on a forum over the course of a semester. did not reveal anything about learning outcomes and student engagement with forum threads[13]. Recently, unsupervised clustering of features from learners following a blended course on two platforms (face-to-face, online)[14] identified four behavioural groups according to their differing levels of engagement across the two platforms. Other examples of temporal feature extraction include the number of times a student logged into their account[15], temporal trace data (e.g., time to answer correctly)[16], and the time to half-way completion and total frequency of completion[17]. Each of these approaches aimed to develop tools for course managers or learning designers to understand and visualise participant performance.

Whilst feature vector analyses can be predictive of drop-out or performance in particular scenarios, these approaches are limited by the choice of features ('feature engineering'), itself limited by the granularity of the data. Indeed, extending the feature space to include higher resolution temporal features (e.g, timestamps of lecture viewings or start-dates of peer-graded assessments) has been shown to improve prediction accuracies for drop-out and final performance[18,19]. Despite such improvements, temporal feature vectors lead to a 'temporally averaged' analysis of student trajectories, as they do not capture longitudinal (dynamical) changes, and therefore may miss important events or temporal dependencies.

Beyond feature vector representations, there have been developments guided by theoretical considerations[20]. For example, a temporal analysis of conversational data used segmentation (or 'stanzas') to construct a network of micro-scale elements to reveal connections across timescales[21]. A study of communal knowledge in online knowledge building discourse used temporal analytics combined with graph theory to identify ideas and their mobility over time[22]. Graph centrality has been used to explore temporal and contextual profiles of shared epistemic agency on discourse transcripts[23], identifying pivotal points of discourse exchanges. One study emphasised the importance of temporality as the main component of learning analytics through main path analysis of Wikiversity domains to model the flow of ideas[24]. The method they developed could be used to support teachers in coordinating knowledge building, yet the method is not suited to sequential completion of tasks and would instead be interesting when applied to courses with no distinct task order. Another approach of temporal analysis is to perform so-called micro-analyses, where singular 'extreme' learners are examined in depth to explore differing behaviours[8]. Whilst such approaches provide insight into particular behaviours, the conclusions are difficult to generalise and justify statistically[25].

Whilst each temporal analysis requires a theoretically justified model, there is clearly also a need for high-resolution data that can reveal nuanced longitudinal relationships. Considering this, another study used a coarse grained approach across knowledge components to identify focal points that were worthy of more in-depth analysis using high-resolution data and analysis[26]. This method integrated the temporal and data resolution scales that often evade temporal analyses, however, this approach may ignore potentially important focal points that can only be identified using high resolution analysis from the start. An alternative study used dynamic time warping on high-resolution task data to calculate similarities between the raw time-series of a cohort of students undertaking an on-line degree, and implemented a quasi-hierarchical clustering algorithm to identify data-driven groupings of students that exhibited similar temporal behaviours[5]. The behavioural clusters were shown to be more predictive of final performance than classifiers based on temporal feature extraction method.

More recently, sequence-based methods have been shown to provide meaningful insights. For example, one study used optimal sequence matching to analyse the sequence of hand movements in a multi-modal learning environment[27]. They found that incorporating temporality into their analysis was crucial to find a correlation between sensorimotor coordination and learning gains. A similar study used sequence-based method such as Lag-sequential analysis (LSA) and Frequent Sequence Mining (FSM) for the purpose of discovering sequential patterns in collaborative learning[28]. They found that both LSA and FSM were able to uncover productive threads for collaborative learning. Data sequence models aim to model within- and between-semester temporal patterns[29], emphasising the importance of incorporating temporal relationships of heterogeneous data to more accurately predict success or failure of students relative to non-temporal models with the same data.

The recent success of sequence data analyses, as described above, suggests that using a sequence data framework can provide insightful analyses into learning analytics. We also highlight that the temporal sequence data framework is flexible: it provides an opportunity to analyse student trajectories as well as course learning design. One study concluded that using analytical techniques, including temporal analyses such as sequence mining, could highlight interesting anomalies or irregularities that can be the object of a more focused investigation by individuals with deep knowledge of the course[30] . Accordingly, we use the data sequence framework to take a task-centric perspective, and identify common irregularities in task sequences across students.

Here, we develop a framework to model and analyse temporal sequences of task completion and task-to-task transitions by students taking a course, and allows us to compare them to the expected (i.e., designed) task order in the course. Our work introduces novel data-driven metrics and a Bayesian probabilistic model for analysis and prediction of sequence data, and shows that the order of task completion at fine-grained resolution facilitates improved prediction of performance and can be used to inform task-level course design. Unlike methods that construct feature vectors, our temporal data sequence framework[29] directly incorporates temporal relationships between tasks. In the paper, we present the mathematical formulation and structure of the temporal data sequence framework, and we show how the data-driven analyses and Bayesian model can be used to predict

Degree Course	Online MBA (2 years)
Module	**Corporate Finance**
Number of sessions	10
Number of tasks, T	123
Number of students, N	81
Male/Female	57/24
Age range (years)	28–53
Mean grade (std)	68.4% (9.3%)
Mean grade bottom 25% (std)	55.6% (4.6%)
Mean grade top 25% (std)	79.3% (3.3%)

Table 1. Information about course and summary statistics for the dataset used in this paper.

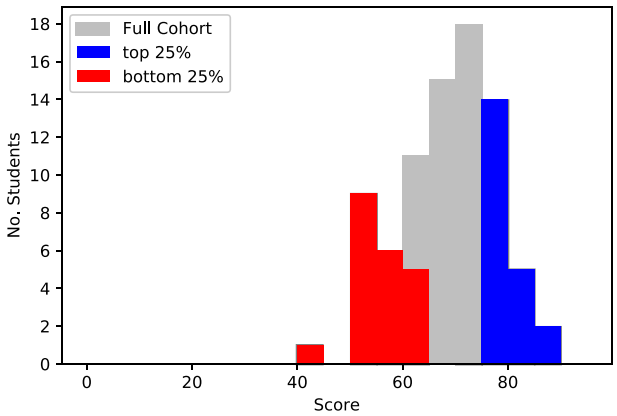

Figure 1. The grade distribution for the entire cohort with the bottom and top 25% groups highlighted in red and blue respectively.

student success, to reveal student confidence at a task level, and, from a task-centric perspective, to aid with learning re-design. We illustrate our work through data of student task completion collected from an online course taken over one semester as part of a 2-year MBA programme.

Methods

In this section we first describe and summarise the high-resolution dataset used to exemplify our analytical methods. Secondly, we describe the structure of the sequence data framework which forms the basis of the analytical methods and models introduced in this paper for the analysis of temporal sequence data. We then highlight key data-driven quantities that can be extracted from our data sequence model representation for the analysis of high-resolution LMS data. We then introduce a Bayesian probabilistic model for modeling sequence trajectories of learners and use it for predicting performance. Finally, we show that sequence data framework can be used to perform task-centric analyses and introduce data-driven metrics to reveal insights into course learning design.

Course information. The subjects in this study were $N = 81$ post-experience learners pursuing a 2-year post-graduate part-time Management degree at a selective research orientated institution, a summary of the data is displayed in Table 1. The cohort ranged in age from 28 to 53 years old (57 males, 24 females) and resided in 18 geographically disparate countries. Although the subjects met face-to-face at the start of each academic year, the course was studied completely online without any further physical interaction. We gathered computer interaction data of the learners undertaking the online management course 'Corporate Finance' during the second semester of the first academic year. The anticipated study load was 5–7 h per week for the single course we analyzed. This course was run in parallel with a second course (not analyzed here) with a similar workload. An online session was delivered to the students each week, with the course running over 10 weeks (for a total of 10 sessions). The course was assessed via a combination of coursework and a final exam. Coursework was undertaken at three points during the course: Sessions 3, 4 and 9. The performance distribution of the entire cohort had a median of 69.7 (mean 68.4), with a longer tail towards low performance, see Fig. 1.

Types of tasks. As students progress through the course, they are faced with *tasks* that are either marked (e.g., coursework) or non-marked (e.g., watching a video). The tasks fall into six main categories:

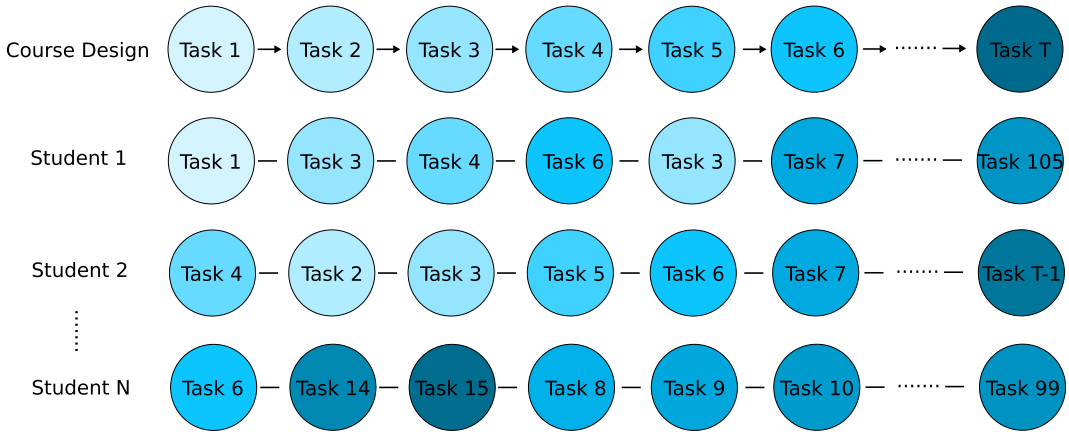

Figure 2. The sequence data framework represents each task as a node which is ordered according to the course learning design. Each student is represented by a sequence of tasks. This framework forms the basis for application of various data-driven analyses and Bayesian sequence modeling.

1. Coursework: Marked. Contribution (30%) towards final grade. The coursework is normally introduced 2 weeks before the deadline. Completion defined as user submitted.
2. Quiz: Marked. Contribution (10%) towards final grade. The quiz would be a multi-choice response and would be automatically graded. Completion defined as grade awarded.
3. Multi-response Polls. Not marked. No contribution towards final grade. This task type generally includes a variety of questions which require text input as an answer. The final answers are not seen by their peers but can be seen by the tutors and course leader. Completion defined as user input.
4. Reading/video. Not marked. No contribution towards final grade. Completion of the task is defined as having watched the video or having opened the reading link.
5. G-chart. Not marked. No contribution towards final grade. The learner is asked a question and offered a set of options. Their answer is added to a chart that displays the distribution of learner answers. Completion defined as user input.
6. Discussion post. Not marked. No contribution towards final grade. Discussion posts are often questions posed by the course leader which require thoughtful response from the learners. The learners responses can be seen by their peers after they have completed the task. Completion defined as user input.

Note that, whilst some tasks do not directly contribute towards the final grades, they do have an indirect effect on the final grade—a 10% participation grade is awarded to each student based on their overall task completion to incentivise task completion. The most up-to-date version of our learning management system now includes a much larger and more diverse set of tasks. However, at the time of data collection, the set of tasks was limited to those described above. As a further note, the students take an exam at the end of the year that contributes (50%) of their grade for a given module. The exam is not included as a task during the course, but does contribute towards the final grade. In the remainder of the paper, a 'task' is any one of the above six types.

Sequence data framework. The 10 sessions of the course are made up of a total of $T = 123$ tasks identified by their 'Task ID', an integer from 1 to 123 representing the *nominal order* of the course layout, i.e., the order in which tasks appear to learners as they navigate the online course website. This nominal task order g is represented as the ordered set $s^{(g)} = \{1, 2, \ldots, 123\}$. Each learner k is described by the ordered set of completed tasks:

$$s^{(k)} = \left\{ t_1^{(k)}, \ldots, t_n^{(k)} \right\}$$

where $t_j^{(k)}$ is the task ID of the jth task completed by learner k. Note that the sets may be of different lengths for different learners since task completion is not compulsory and the number of completed tasks may vary from individual to individual. Figure 2 displays some example sequences for students using the framework described here.

We can also coarse grain our sequence framework. Given that tasks are grouped into sessions we compute the transition probability $p(\text{next session } s+1 | \text{ previous session } s)$, by substituting the task IDs for session IDs in the ensemble S. We use all these quantities to analyze the behaviors and critical juncture tasks associated with the course.

Data-driven learner-centric analysis. Given our data sequence framework we introduce two key methods for analyzing the sequence data to investigate learner behaviours. The whole cohort of learners is then described by the ensemble of ordered sets of tasks, $S = \{s^{(k)}\}_{k=1}^{81}$. From this ensemble, we can compute the following two key quantities:

1. The number of times that task i appears as the jth element across the ensemble S is denoted as n_{ij}. The probability p_{ij} that task i appears as the jth completed task across S is thus given by $p_{ij} = n_{ij}/\sum_j n_{ij}$. These probabilities are compiled in the $T \times T$ matrix $P = (p_{ij})$.
2. To quantify the transitions between tasks, we count the number of times m_{ij} that task j is immediately preceded by task i across the ensemble S, and normalise it across all transitions to obtain the probability that task j is preceded by task i across the ensemble: $p_{i\to j} = m_{ij}/\sum_{i,j} m_{ij}$.

   To obtain the conditional probability, we need to normalize across all tasks: $\pi_{ij} = p(\text{next task}j|\text{ previous task }i) = p_{i\to j}/\sum_j p_{i\to j}$. These conditional probabilities are compiled in the $T \times T$ matrix $\pi = (\pi_{ij})$.

   It should be noted that this matrix provides a summary of transitions to any task given a previous task acquisition, but π is not a stochastic transition matrix for task acquisition due to tasks being irreversibly acquired.

The sequence analysis methods were adapted from[31].

Bayesian probabilistic model: a generative model for task sequences. In the description of a learner's task sequence, no assumptions about an underlying model that may have generated the sequences. The most general probabilistic model assumes that task completion is determined by all the tasks previously completed. If we assume that the order in which previous tasks were completed can be ignored, this model is a first-order Markov chain with a hypercubic state space defined by 2^T states and an associated $T\,2^{T-1}$ edges representing the transition matrix between these states. As $T = 123$, the state space and associated parameter space of the transition matrix is too large to be tractable.

Here we introduce the application of HyperTraPS, a recently developed probabilistic Bayesian model[31–33]. We fit this model to derive parameterisations that can be used for student classification. The model reduces the full hypercubic state space to a tractable $T^2 = 15{,}129$ state space by assuming that the probability of the next task in the sequence is proportional to a basal rate of acquisition for that task and independent contributions from tasks already completed. These are fitted from the aggregate set of observed transitions (the student task sequences).

As we have complete information on the order in which tasks are completed, the utility of a generative model in this case may be found in its use as a classifier. Utilising and extending the Naïve Bayes classifier introduced in[33], we consider the probability that a learner k belongs to the high performing group g_1 and the low performing group g_2 after completing n tasks with task set s^k. We can write the odds ratio relating to this probability as:

$$\frac{P(k \in g_1|s^{(k)})}{P(k \in g_2|s^{(k)})} = \frac{P(s^{(k)}|\pi(g_1))P(g_1)}{P(s^{(k)}|\pi(g_2))P(g_2)}$$

where $\pi(g_j)$ are posterior samples drawn from the HyperTraPS model fitted to a set of sequences from labeled group j. In this paper, the two groups considered are the top 25% (g_1) and the bottom 25% (g_2) based upon their score in the course.

Data-driven task-centric analysis. Using our sequence data framework, we compute statistical properties of every task: (1) the frequency with which it is completed and (2) the mean rank at which the task was completed across all learner sequences. The difference in these values could be calculated between two groups, giving the difference in completion frequency and difference in completion order respectively.

Student confidence analysis. To investigate the effectiveness of our analytical methods and model, we use a student confidence analysis to introduce a 'story-telling' or causal inference aspect . At the end of each session, students are asked to reflect on the tasks they had been given to complete. There were given three options to choose from for each task: (a) 'Yes I feel confident I can do this', (b) 'I need to revisit this', and (c) 'I need more support'. For each task we calculate the proportion of students that felt confident (stated confidence) or did not feel confident (needed to revisit or requested support).We use this data to quantify when a student was not confident on an individual task, and we measure the confidence of a student across the set of tasks as:

$$\mathscr{C} = \frac{\text{`}confident\text{'}}{\text{`}revisit\text{'} + \text{`}support\text{'} + \text{`}confident\text{'}} \tag{1}$$

where a larger value indicates a more confident student and a lower value indicates a less confident student. Using this additional student-provided data, we can associate anomalous sequence orderings with measures of student confidence at both individual and group levels.

Ethics and consent. All methods were carried out in accordance with relevant guidelines and regulations. Ethical approval from the Education Ethics Review Process (EERP) at Imperial College London was attained (EERP 1718-032b) and a waiver for informed consent was granted (also contained within EERP 1718-032b) for this study.

Results

Data-driven learner-centric analysis. *Ensemble task completion behaviors of student cohort relative to course design.* As defined above, the completion sequences of N learners over T tasks can be summarized as a $T \times T$ 2D-histogram (or matrix of numbers) P, where each element p_{ij} corresponds to the probability that task i

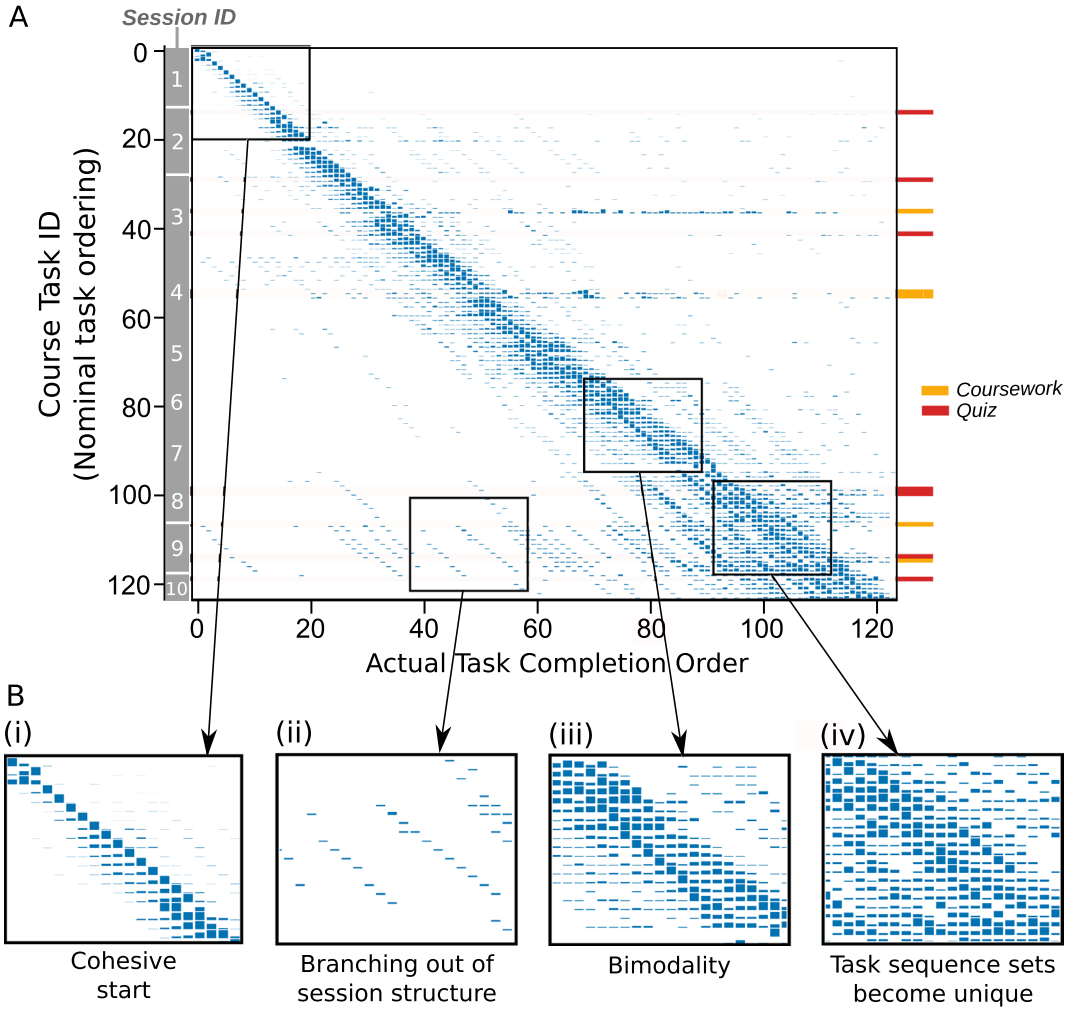

Figure 3. (**A**) 2D-histogram compiling the probabilities of task completion for the 81 learners. The *y*-axis denotes the course task ID, i.e., the nominal completion order. Each task is part of a weekly session. The *x*-axis displays the point in the learners' sequence in which that task was actually completed. Some of the key task types (coursework, quiz) are indicated by colours. (**B**) Each insert shows a magnification of a different part of A: (i) the beginning of the course; (ii) large deviations from the nominal task order later on; (iii) bimodality of responses for particular tasks; and (iv) the broad distributions of task completion towards the end. Note also the wider spread in coursework and quiz tasks.

was completed in the *j*th position across the ordered sequences of the cohort. Figure 3A displays this histogram for our learner cohort, with tasks ordered according to their nominal task order g (1–123) along the vertical axis and plotted against the actual completion order in the learners sequences on the horizontal axis. The histogram for each task across each row summarizes the spread of the actual order of completion across the learners in the cohort. If all learners had completed *all tasks in the nominal order g*, then P would have been the identity matrix, with all the probability located on the diagonal. Therefore, the off-diagonal spread signals departure from the nominal order across the cohort. For example, learners completed task 1 as their first task with high coincidence (77%, $p_{11} = 0.77$), yet, even for this first task, there was a sizeable proportion of learners (21%, $p_{12} = 0.21$) that completed task 1 as their second task. Overall, the strong diagonal component of Fig. 3A suggests that learners generally follow the nominal task order, with deviations evidenced by the presence of such off-diagonal elements.

At the beginning of the course, we observe a cohesive start in which learners generally follow the expected course structure as evidenced by the sequential peaks of each histogram up to task 15 (Fig. 3B(i)). Yet, more in detail, we observe some deviation from the nominal course structure. For example, very early on task 2 is most commonly completed in third position, whilst task 3 is completed second (Fig. 3B(i)). Considering these two tasks in more detail, task 2 requires the learner to complete a vast amount of reading whilst task 3 requires the learner to submit an estimate for a quantitative question. Task 3 is located on the next screen requiring the learner to actively leave the web page containing task 2 to access task 3. Since task 3 is inherently related to task 2 (both explore financial inter-mediation), it could be that learners did not feel they comfortably understood the material in task 2 until they completed task 3. Another deviation occurs between task 15 and task 16 (Fig. 3B(i)), whereby task 16 is completed earlier than task 15 on average. Task 15 requires the learner to complete a quantitative quiz

with several complex financial questions. Although the quiz did not contribute to their final course grade, the learners may have wanted to leave the quiz until the end of the learning session, or they might have believed that additional material may become available later to aid them in the quiz.

Beyond such small deviations from the nominal task order, we also identify much larger deviations. Figure 3B(ii) highlights tasks that have been completed much earlier that expected from the nominal task order. Interestingly, these tasks are completed in sequential order within their session, signalling a jump-ahead from the learners to an out-of-order later session. Such jump-ahead deviations appear as diagonal streaks in the lower triangle of Fig. 3A, but note that similar deviations are also observed in the upper triangle of Fig. 3A indicating tasks that were completed later relative to the nominal task order.

There are several features of the cohort dynamics that become more prominent as task completion progresses. In Fig. 3B(iii), we observe an example of bimodality (two peaks) in task completion indicating that there are two groups of learners completing these tasks systematically, but at different points in their sequences (relative to the nominal task order). Note that the presence of bimodality is not simply an artifact of learners randomly missing tasks, if so we would observe a single broad peak, but the emergence of subgroups of learners following different strategies. Indeed, this instance of bimodality appears at the beginning of Session 7 (task 84–task 95) and becomes less pronounced after Session 7 is finished. After a more thorough analysis, we find that a large group of learners appear to skip both task 78 and task 81 whilst the remaining learners complete these two tasks, resulting in a branching of two sets of learners corresponding to left-hand and right-hand peaks. Both of task 78 and 81 are interactive tasks that required an application of learned knowledge, which may have caused the split between those that were able to complete the tasks and those that were not.

As task completion progresses, we observe a broadening of the task distributions (Fig. 3B(iv)). Such broadening is expected when completion of tasks is not mandatory; if a learner misses a task and continues to follow the nominal task ordering of the course they will have a shift in their actual completion sequence relative to other learners. The broadening is amplified towards the end of the course, as a consequence of learners skipping an increased number of tasks (variable across the cohort of learners), in order to finish the course before the final exam.

An examination of the deviation of the different learners from the nominal task order is presented in Fig. 4. Figure 4A shows two distinct task sequences for two learners: learner 17, who follows closely the nominal course order almost all the way to completion, and learner 29, who completes only about half of the total tasks in a more idiosyncratic order, with several jumps between sessions and completing only a little over 50% of the tasks of the course. We compile this information for all 81 learners in Fig. 4B, where we show the (normalized) distance between the actual completion of tasks and the nominal task order for each learner ordered in descending order of performance from top to bottom. Tasks in blue were completed earlier in the sequence relative to the nominal order, whilst tasks in red indicates later completion in the sequence relative to the nominal order. The learners of Fig. 4A are indicated by dots (green for learner 17, purple for learner 29). The observed deviations of learner 29 are seen as blocks of tasks (in blue and red) completed out of sync with the nominal task order. In contrast, the profile of learner 17 is almost totally white, indicating a constant progression in accordance with the nominal order. Figure 4B also shows that low performing learners tend to exhibit large deviations from the nominal task order more often than high performing learners (although certainly not exclusively) and a lower fraction of completed tasks (as evidenced by the larger black blocks). Note that the figure shows sequential blocks of tasks, commonly corresponding to a single session, which have been completed out of sequence relative to the nominal order. This indicates that analyzing sequences not only at the level of tasks, but also at the level of weekly sessions can provide a meaningful temporal basis for the analysis of learners' behaviors[21], as we explore below.

A comparison between high-resolution and coarse-grained temporal sequences. Above, we extracted quantities that revealed deviations from the designed course sequence for both individual students and the ensemble of students. In this section we explore the high resolution task-level data and compare it against a coarse grained version of the data (coarse grained by week).

Figure 5A plots the task transition probability matrix π defined above in two different (but equivalent) ways: as a transition graph (i) and as a transition matrix (ii). As expected, we find strong transitions that follow the nominal structure both within each session as well as clockwise from session to session in Fig. 5A(i). This is similarly shown in Fig. 5A(ii) by the strong concentration on the diagonal and upper diagonal, which indicates the high probability that learners follow the nominal order of the course. However, we also find a large number of non-sequential transitions between tasks as a consequence of learners jumping forwards or backwards in the course. These are shown as non-sequential edges in Fig. 5A(i) and off-diagonal elements in Fig. 5A(ii). Specifically, we observe a large number of non-sequential transitions within sessions 2, 3 and 5. For example, within session 3 there is a large probability that learners complete task 42 and then return to task 38. Task 38 is a discussion post where learners were required to publicly write a response to a tutor question whereas task 42 is a quiz with various questions regarding equity markets and valuations. It appears that a large number of learners completed the quiz before the discussion post, perhaps because they did not feel confident completing the public discussion post until they had gained additional information or understanding from the quiz. To explore this further we analysed the percentage of students that felt confident undertaking this task. We found that only 60% (59/81) of students were confident in this task (self-assessed at the end of each session), whereas across the remaining tasks within session 3 the average number of students that were confident was higher (69%, std. 7%) and the average task confidence across the module was (68% std. 8%).

For comparison, we take a coarse-grained approach and calculate the transition probabilities between the 10 weekly sessions (Fig. 5B) represented as: the inter-session transition graph in (i) and the coarse-grained transition matrix in (ii). As in the task level analysis, there is a strong clock-wise sequential component in the graph

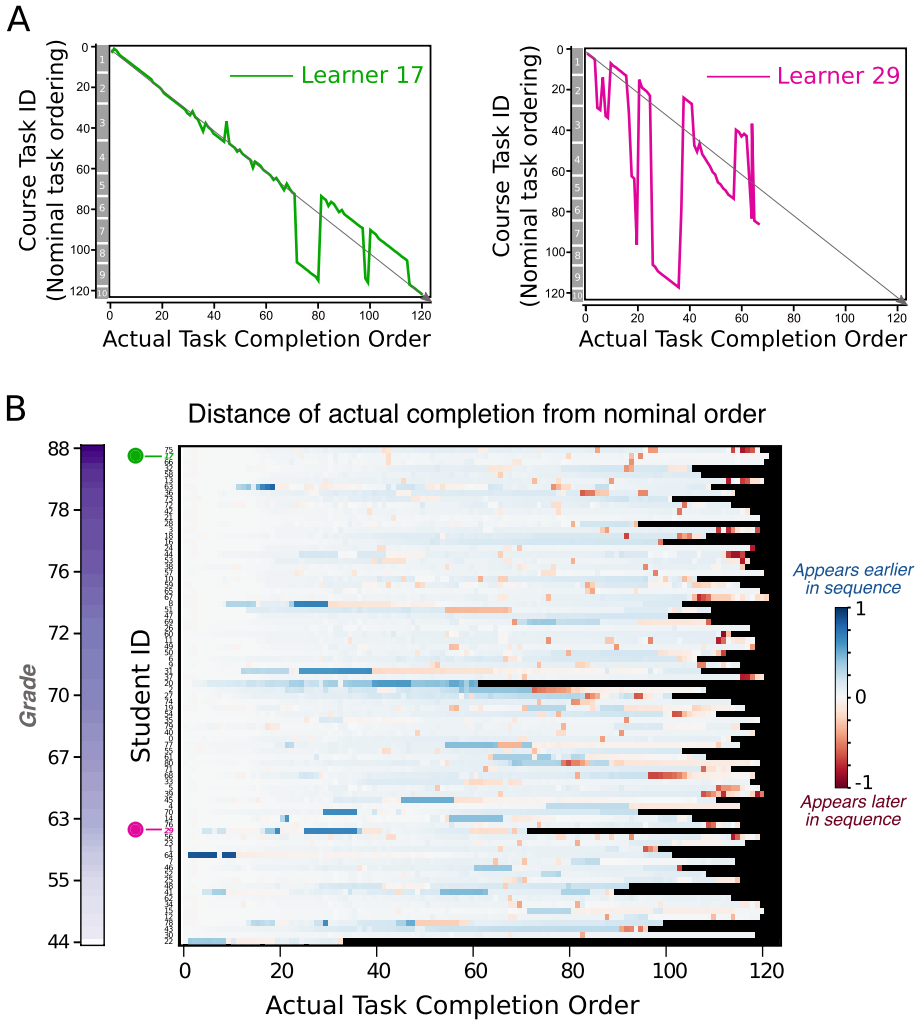

Figure 4. (**A**) Sample trajectories of two learners (17 and 29) as they undertook the course. Learner 17 followed closely the nominal task ordering for the majority of the course only deviating towards the latter tasks. Learner 29 exhibited large deviations from the nominal task ordering throughout and only completed half the course. (**B**) A heat map displaying the relative distance between the actual completion order and the nominal order for all learners (rows). The learners are ordered by descending performance from top to bottom, as indicated by the grade colourmap. If a learner had completed the course exactly as the nominal order dictates then the row would be white. Tasks completed earlier relative to the nominal order appear blue, whilst tasks completed later than the nominal order appear red. The region after the student has completed the course is colored in black. The learners in A are indicated by a green dot (learner 17) and a purple dot (learner 29).

(i) and a concentration of probability on the main and upper diagonal in the matrix (ii), indicating that learners generally follow the session sequence. However, there are obvious deviations. For example, we observe that there is a high probability that learners will transition back to session 3 from several sessions (6, 7, 8, 9, 10). Similarly, learners at session 10 (the final session) will often transition to previous sessions (sessions 8 and 9), maybe in an effort to revise and complete parts of the course they had skipped over while completing the course.

These results suggest that task-level (high resolution) and session-level (coarse-grained) sequence data provide true but alternative descriptions of the data that can lead to distinct but complementary conclusions. Within the high resolution analysis we are able to identify individual anomalous tasks (such as task 38), whilst within the coarse-grained analysis we identified an anomalous week (e.g. week 3). This section highlights the importance of using high-resolution temporal data coupled with an appropriate temporal model for analysis.

Comparing sequence completion patterns between high and low performance learners. Given the differences between the high resolution and coarse grained analyses, we perform a similar comparison whilst exploring student performance. As suggested by Fig. 4B, the pathways and patterns of task completion may differ between high and low performing learners. To explore this hypothesis in more detail, we divide our cohort of learners into two groups: those in the top 25% (Group 1) and bottom (Group 2) 25% of performers according to course

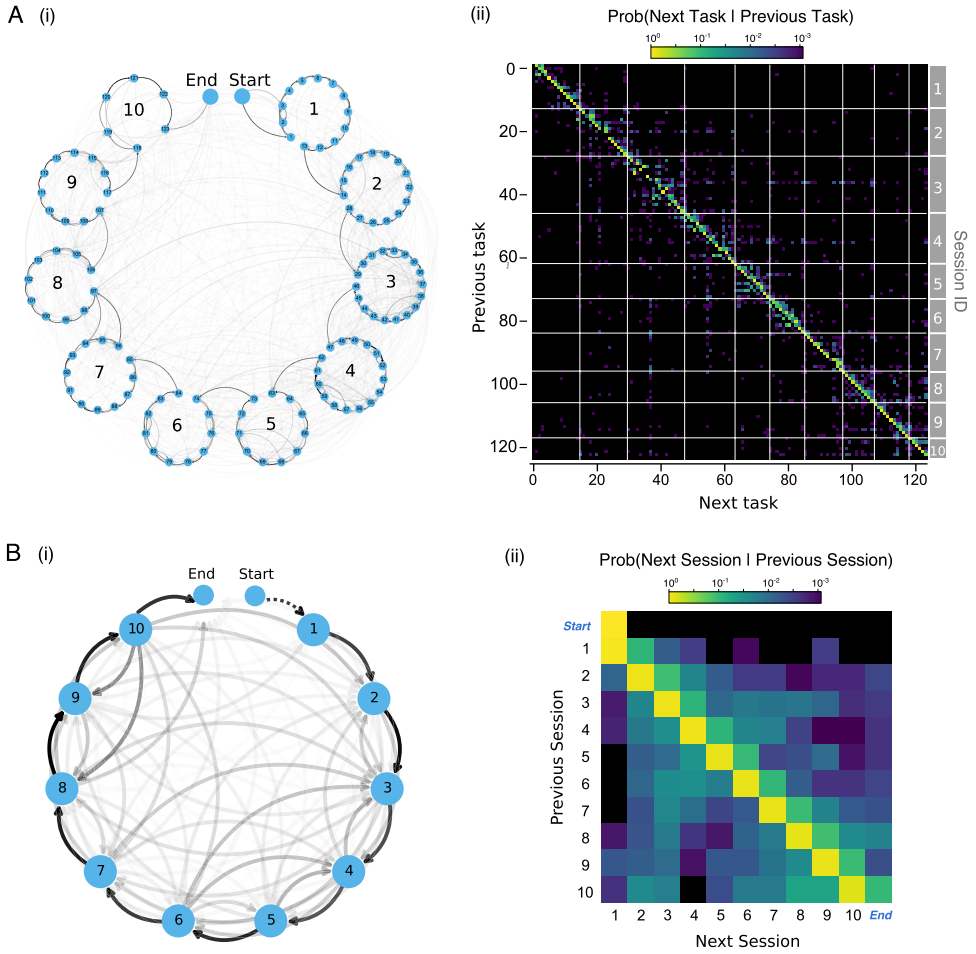

Figure 5. (**A**(i)) The transition probability between tasks π is represented as a transition graph. Each blue node corresponds to a task, each sub-circle corresponds to a weekly session of tasks (numbered from 1-10), and the entire circle corresponds to the full course. The edge thickness between tasks corresponds to the probability π_{ij} of transitioning to task j given the previously acquired task i. Generally, there is a high probability of transitioning between sequential tasks; however, there are often large non-sequential transition probabilities. (**A**(ii)) A heatmap of the matrix π (in logarithmic scale). Off-diagonal elements correspond to deviations from the course structure. (**B**(i)) A transition probability graph between weekly sessions, where each node corresponds to a session and the edges between sessions i and j are weighted by the probability of transitioning between any task in session i to any task in session j. Although there is a strong probability of sequentially completing the sessions, there are is a significant number of deviations from the session sequence (e.g., from session 10 to sessions 9 and 8, or from session 6 to session 3). The transitions within a session (self-loops) and the transition from start to session 1 are not shown for clarity of visualisation (**B**(ii)) A heatmap of the transition probability between sessions. The diagonal exhibits the highest probability given that a learner is most likely to transition between tasks within a session followed by the upper diagonal, representing transitions to the next session in the nominal sequence.

grade. Although more nuanced performance groups could be created, we opted for a simple split into the bottom 25% and top 25% to ensure a bimodal performance distribution.

We compute two separate transition matrices π_{HP} and π_{LP} from the sequences of learners in Group 1 (high performers) and in Group 2 (low performers), respectively, and obtain the difference between both: $\Delta\pi = \pi_{HP} - \pi_{LP}$. In Fig. 6A we show the transition graph (i) and transition matrix (ii) where the high performers have larger probability (i.e., the elements of $\Delta\pi > 0$). Similarly, Fig. 6B presents the transition graph (i) and transition matrix where the low performers have larger probability (i.e., the elements $-\Delta\pi > 0$). To compare the transition probabilities between session of high vs. low performers, we compute the Jensen–Shannon (J–S) distance (Figure S1). Our results show that the mean J–S distance for outgoing session transitions (rows of the transition matrix) was 0.12 and the mean J–S distance of incoming session transitions (columns of the transition matrix) was 0.16, where a J–S distance of 0 corresponds to identical distributions and a value of 1 to maximally different). The relatively low dissimilarity between the two groups reflects the fact that both groups of students follow overall the intended course structure. We also note that the dissimilarity between the transition patterns of the two groups (high and low performers) increases as the course progresses towards the later sessions.

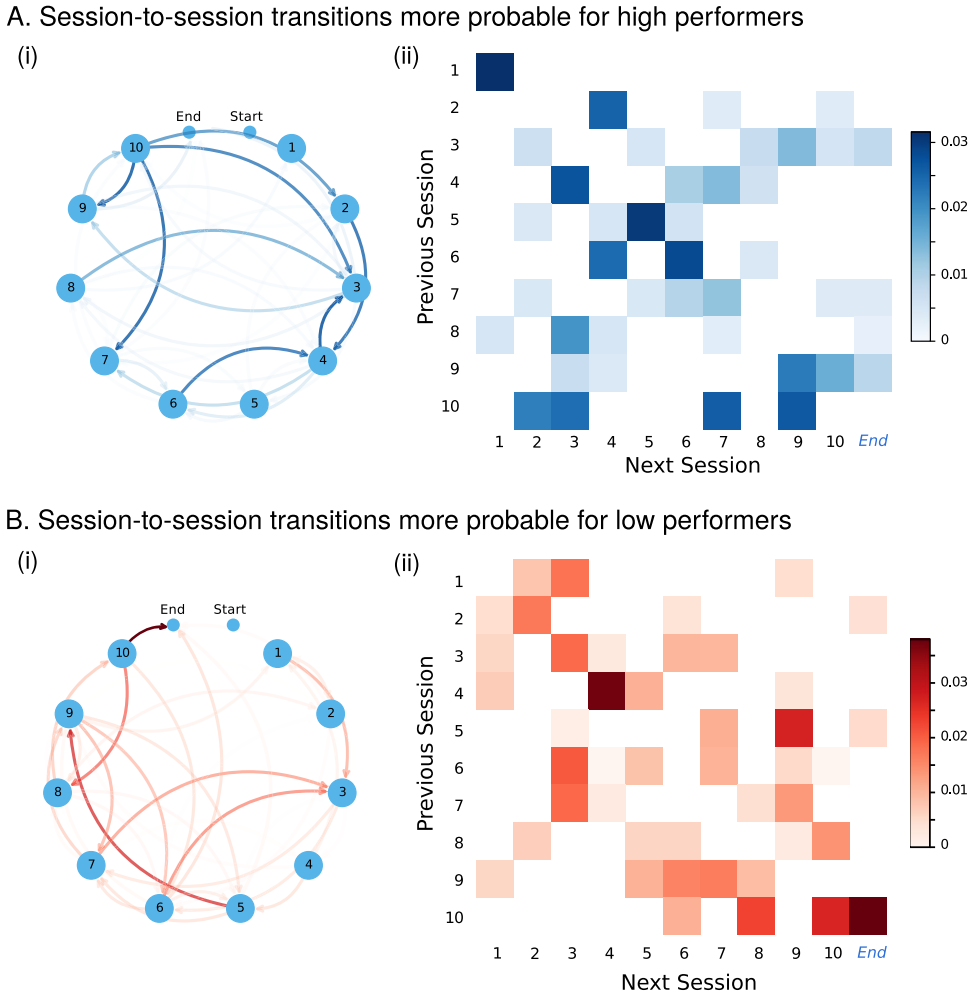

Figure 6. The transition probability between weekly sessions is calculated separately for the top 25% performers and the bottom 25% performers and the difference between the probabilities in the two groups is calculated. (**A**) The session-to-session transitions that are more probable for high performance learners (i) mapped onto a network structure and (ii) visualized as a transition matrix heatmap. (**B**) The session-to-session transitions that are more probable for low performance learners (i) mapped onto a network structure and (ii) visualised as a transition matrix heatmap.

However, we also identify a number of differences in the transitions when comparing high and low performers. For example, low performers were more likely to transition directly from session 5 to session 9, which contained a coursework task that needed to be completed, and after completion of session 9 they were likely to transition back to session 6. In general, low performing learners have a higher probability of transitioning to a task within the same session (diagonal of Fig. 6B(ii)) but also a higher rate of transitions between non-sequential sessions (Fig. 6B(i)). These two views are only seemingly contradictory: low performers tend to follow the nominal task order within sessions but do not follow the nominal session order within the course. We also observe that the only point where high performers are more likely to depart from the course session structure than low performers is in the jump back transitions form session 10, an indication of an effort to revise and complete missing tasks just before the end of the course. These results reinforce the need for high-resolution data and analysis. Using a coarse-grained approach we would observe low performing students following the course session structure in a less consistent manner, yet we would not observe their higher probability of following the nominal task order within each session.

Using the student confidence $\mathscr{C}$ defined in Eq. (1), we found that the low performance group were significantly (paired t-test, p=0.041, alpha=0.05) less confident (no. tasks confident, mean=0.65, std=0.27) than the high performance group (no. tasks confident, mean=0.76, std=0.19) across the entire module. At the group level, the confidence values correlate with performance as we would expect.

In addition to comparing high and low performers, we can also compare groups of students characterised by other factors, such as demographic information. As another illustration of the applicability of the framework, we have carried out similar analyses for groups based on gender and age (see Supplementary Information). Figure S2A shows that female learners are more likely to follow the nominal course order (as indicated by

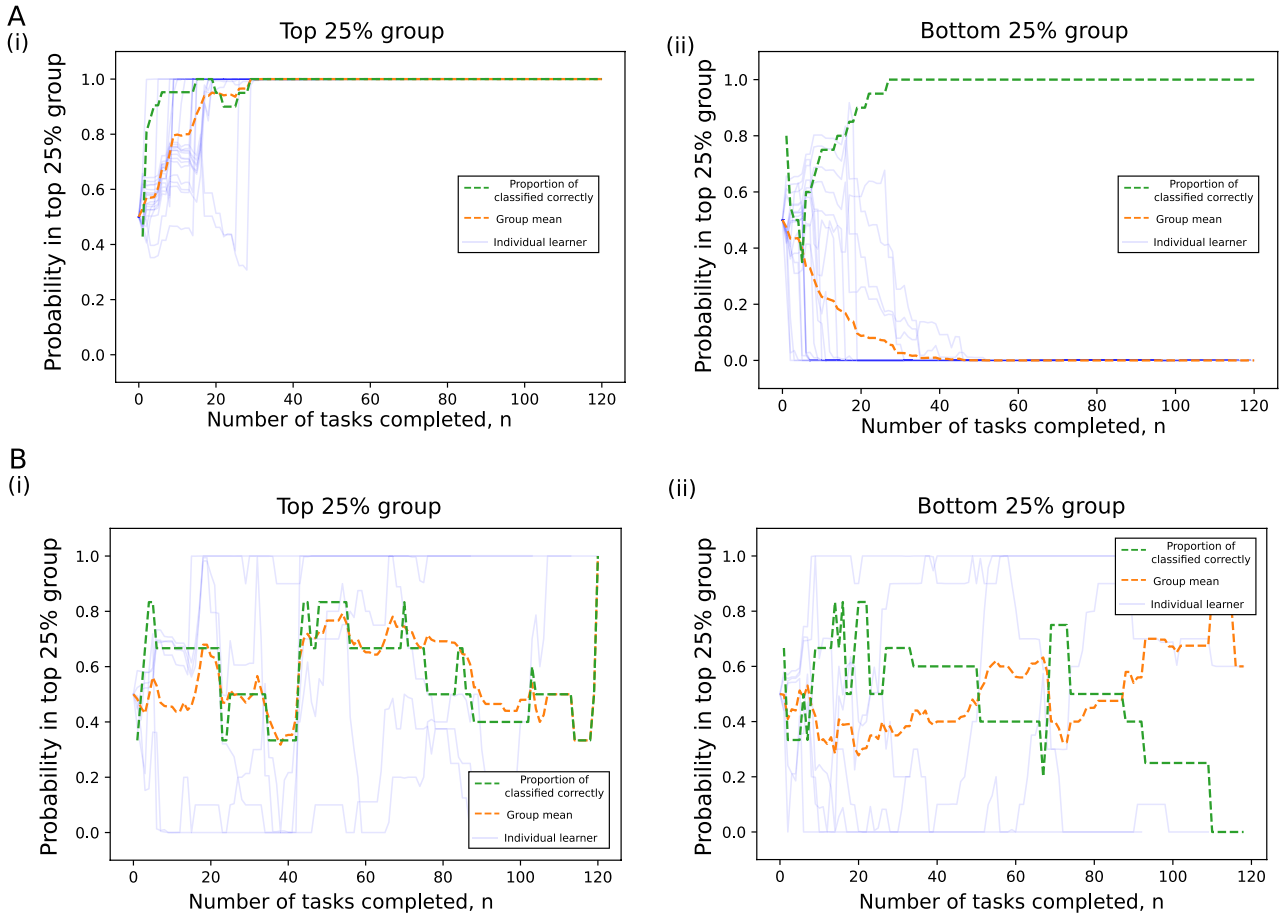

Figure 7. Application of Bayesian probabilistic model to learn sequence trajectories. (**A**) Probability that each learner in top 25% belongs to top 25% (i) and the probability that each learner in the bottom 25% belongs to top 25% (ii) using averages across posterior samples trained on data from the respective groups. Each learner is a blue line, the mean across all learners an orange dashed line and the proportion over 0.5 as a green dashed line. After around 30 tasks the probability goes to unity and zero respectively for the two test groups providing indication that the probabilistic model is able to learn differences between the two groups. (**B**) The top plots utilized data included in the training dataset for testing, while the bottom plots train with 70% of the learners from each group and test on the remaining 30% from each group. This indicates whether the learned model can differentiate unseen learners giving it true predictive utility. There is much greater variability in the position of the each learner's blue line, but on average the orange and green lines lie above and below the 0.5 threshold indicating there is some predictive power. Interestingly, the strongest predictive region seems to be in the gap where coursework is completed later after completing more of the course, a trend observed in the higher performing group. The predictive power weakens at large n due to fewer training and test samples completing that many tasks.

the higher probability of on-diagonal transitions) when compared to male learners, who exhibit more erratic sequences. Interestingly, we find that the larger probability of transitioning from session 10 to earlier sessions, which was associated with efforts to revise or complete skipped sections, is dominated by male learners. Regarding age groups, Figure S3A shows the analysis that compares the highest 25% and lowest 25% by age. We find that younger learners exhibit a stronger adherence to the expected ordering of the course, whereas older learners depart from the nominal order and have a stronger tendency to revise earlier sessions (especially 2, 3 and 4).

Bayesian model for sequence trajectories. In this section, we use the probabilistic generative model to examine whether we can predict the group (top 25%, g_1 or bottom 25%, g_2) to which a student belongs after completing n tasks with task set $s^{(k)}$. We perform two experiments to consider the plausibility of such a model for the data.

In our first numerical experiment, we test whether the probabilistic model can learn parameterisations that can distinguish two groups. We draw posterior samples $\pi(g_1)$ and $\pi(g_2)$ from HyperTraPS for the entire set of g_1 and g_2. For each learner in the training datasets g_1 and g_2, we plot the probability that they belong to their respective groups. Figure 7 (top) shows the result of this when the test learners that we are trying to predict group identity are including in the training dataset. In this case, the left-hand plot (for the top 25% learners) and right-hand plot (for the bottom 25% learners) show strong predictive ability after only a few tasks have been

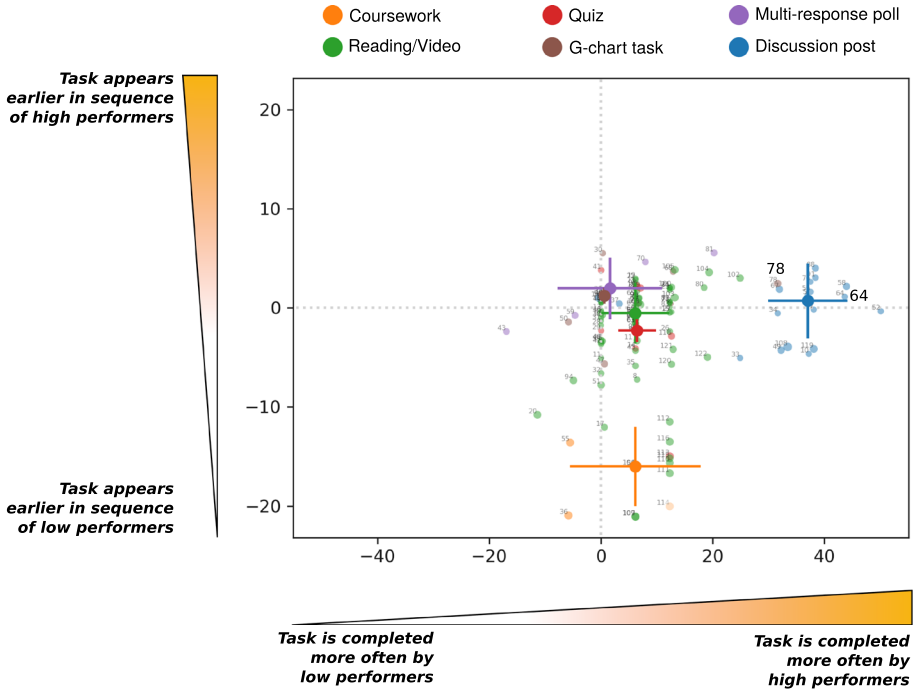

Figure 8. Comparing the high and low performing groups: The difference between the two groups in completion frequency versus the difference between the two groups in mean completion order. The tasks are identified by their task ID and colored by type (the six types are listed in the legend). For each task type, the median and inter-quartile range is plotted. If a task appears in the upper half, the task has been completed earlier on average by the group of high performers. If a task is located in the right half, the task has been completed more frequently by the group of high performers. Discussion posts, multi-response polls and G-chart tasks are completed significantly earlier and more frequently by high performers suggesting the importance of active learning. Coursework tasks appear earlier in the sequence of low performers because low performers have completed less of the course by the time the coursework is due for submission.

completed in the set of all tasks with the probability going to unity and zero for the groups respectively after around 30 tasks. This indicates the probabilistic model is able to learn differences between the two groups in its parameterisations , an important prerequisite for establishing whether learners task set $s^{(k)}$ in conjunction with parameterisations of a generative model can be used to predict performance.

In our second numerical experiment, we test whether the probabilistic model can use learned parameterisations to predict which group unseen learners belong to. We do this by splitting our groups g_1 and g_2 into a 70% training and 30% test split. We draw posterior samples from $\pi_{\text{train}}(g_1)$ and $\pi_{\text{train}}(g_2)$, and then perform the same method to calculate the probability that learners from the test split belong to g_1 and g_2 (given we know their true labels). Figure 7 (bottom) illustrates the results for the test split of learners (split by their true label g_1 on left-hand side and g_2 on right-hand side). The results indicate much less certainty for individual learners, with misclassification occurring. However, the averages across learners in each group indicate some predictive power with average values mainly above 0.5 for g_1 and below 0.5 for g_2. Additionally, the region of strongest predictive power where the averages are farthest from 0.5 is around the region where we have observed learners being particularly divergent in their choice of task suggesting the results may be robust. Given such small dataset sizes of only around 15 learners for training, this result would be able to be validated on courses with many more learners available.

Data-driven task-centric analysis. Using a data sequence framework, we have shown that we can analyse individual or ensemble trajectories of students. However, one of the benefits of a data sequence framework is that it can be used for task-centric analysis, i.e. each individual task can be analysed with respect to the sequences of a group of students to study particular outcomes such as performance. In this section we introduce task-centric quantities that can be extracted using a data sequence framework. In particular, we explore the differences between high and low performance in completion behaviour for individual tasks and task types.

The tasks are of six types: Coursework; Reading/Video; Quiz; G-chart task, Multi-response poll, and Discussion post. For each task, we calculate: (i) the difference in the frequency of completion between high and low performers; (ii) the difference in the mean position in the completion sequence between high and low performers. The results for all tasks are shown in Fig. 8 identified by their Task ID and colored by their task type. Deviations from the (0,0) point in the centre of Fig. 8 correspond to differences between low and high performers. If a task appears in the upper half, the task has been completed earlier on average by the group of high performing learners. If a task is located on the right half, the task has been completed more frequently by the group of high

performing learners. Therefore, a task located in the upper right quadrant is completed earlier (in the learner sequence - not necessarily in time) and more frequently by high performance learners.

In Fig. 8 we also show the median and interquartile range for each task type. The median for the 'Reading/Video' tasks appears almost directly at the centre of the plot, indicating that completing a video or a reading task is not correlated with high or low performance. In contrast, other task types show significant deviations between the two groups. Specifically, discussion posts or interactive tasks such as multi-response submissions (usually requires a learner to make a calculation) and G-chart tasks (learners must submit a choice from a selection of answers) have a median that is located in the upper right quadrant, i.e., they are completed more frequently and earlier by high performers. In addition, coursework tasks appear significantly earlier in the sequences of low performers. Although this might seem counter intuitive at first sight, it is consistent with the fact that low performers tend to skip more tasks but do complete courseworks.

Beyond general task types, we can identify specific tasks that may be correlated with high performance. For example, tasks 64 and 78 (highlighted in Fig. 8) exhibit high frequency and early completion by the high performing learners. Task 64 was an open discussion where learners were asked to perform two calculations and then discuss the results; task 78 required learners to make a decision about picking stocks for a portfolio. Both of these tasks require the learner to understand the previous content. The public nature of the answers among their peers might put off learners that lack confidence in their answers. In the case of discussion tasks, where a learner must submit a public answer to a tutor's question, 19/20 discussion tasks are found in the right half, where high performers complete these tasks more frequently, and 15/20 are found in the upper right quadrant, where they also appear earlier in the sequence of high performers.

The bottom half contains a number of tasks such as the coursework tasks 36, 54, 55 and 115. As discussed above, coursework tasks appear significantly earlier in the sequence of low performers as a reflection of the fact that they have completed fewer tasks by the time the coursework is due for submission. Hence low performance learners skip ahead to reach the coursework. Other tasks completed earlier and more often by low performing students include a number of tasks from session 2 (tasks 17, 20, and 29) and session 9 (tasks 108 and 109) which appear because low performing students skipped the beginning of both sessions.

For completeness, we have also extended the task-centric analysis to compare learners by gender and age (see Supplementary Information). We find that the female group completed a larger number of tasks and complete them later in their sequence (Figure S2B). Regarding age groups, only discussion posts and multi-response polls are completed more often by older learners relative to younger learners, but older learners complete quizzes and coursework later in their sequence of tasks (Figure S3B).

Discussion

In general, current temporal analyses of student data tend to obfuscate the longitudinal relationships between events through averaging, coarse-graining or feature extraction. In this paper, we have used a data sequence framework to develop a set of data-driven analyses and a Bayesian sequence model to investigate high-resolution task completion data of learners. We show that using this framework we can introduce a task-centric analysis which could help inform learning design. In general, we identify aggregate and individual learner behaviors; explore the relationship between course structure and learner trajectories; reveal the connection between completion patterns and task types; and identify points in the course where learners might benefit from intervention.

The first section used the data sequence model to analyze the ensemble of learner trajectories relative to the nominal task order. We highlighted the fact that although learners generally follow the nominal course structure, a number of exceptions occur, for different reasons, at different junctures in the course. We identified various behaviors such as bimodality and branching from nominal task ordering that occurred as a consequence of groups of learners taking alternative approaches to task completion. Our comparison between learner trajectories and nominal course structure revealed large deviations in different parts of the course, and increasingly as the course progresses. Whilst not pursued here, analyzing the causal effects of such behaviors could provide the educator with an improved understanding of their course design, and about individual learners undertaking the course.

In the second section, we compared high-resolution and coarse-grained temporal analyses. The transition probability between tasks and sessions was used to study the deviations when learners transition between non-sequential tasks. The analysis at the coarse-grained level of inter-session transitions revealed specific sessions in the course structure where deviations from the nominal course structure occurred, highlighting the importance of an analysis at different levels of time resolution to understand the interactions of learners with the material[21].

In the third section, we furthered our comparison of high-resolution and coarse-grained approaches to evaluate differences between high and low performance learners. We began our exploration of the differences in completion behaviors between high and low performing learners. Having split the learners into two groups according to the median grade, we showed that low performing learners followed more strictly the task structure within a session but were less likely to follow the session structure throughout the course. The analysis also showed that low performers tend to skip a number of sessions in order to complete coursework tasks. Similar comparisons between gender and age groups were also presented.

In the penultimate results section, we introduced a Bayesian probabilistic model to learn the trajectories of task completion sequences. Using the top 25% and bottom 25% groups of students (as a simple bimodal performance distribution) we showed that the task completion sequences could be learned and used for predicting student performance.

In the final section, we showed that our data sequence model could be used for task-centric analyzes. We examined the link between the completion of individual tasks and learner performance. Our analysis showed that completion of task types related to active learning (such as discussion posts, multi-response polls, or G-chart

tasks) were correlated with above median performance. The completion of such tasks, some of which are also visible to the peer learners, might indicate a level of understanding and confidence in the knowledge of the material around that task. An educator may consider encouraging the completion of some of those key tasks (or making them compulsory) in order to encourage the learner to learn the material in order to complete that task.

We found a clear distinction between high and low performers as regards to the types of tasks, how frequently they were completed, and at what point in their sequence. These findings could be considered in the context of Bloom's Taxonomy of Learning[34]. For example, discussion-based tasks require skills toward the higher end of the taxonomy, whereas tasks that comprise reading texts and watching videos require skills towards the lower end. It is then not surprising that tasks in the former category were more commonly completed and earlier in the study sequence of high performing students. Identifying these types of tasks could help tutors structure the courses to reinforce the importance and build-up towards such tasks, as well as identifying differences between groups of learners. Through a small illustrative exemplar study, these results highlight the power of relatively simple sequence analysis methods applied to learning data, and the potential for their use on available data to reveal the connections between learning patterns, performance, and course design.

Whilst the majority of on-line courses have a designed linear structure, it is clear that cognitive development does not progress through a fixed sequence of events. The alternative is to present each course component in parallel, without a displayed intended structure, and students can begin at any point and transition between any components of the course. In such a system, our Bayesian sequence model could still be applied given that it is able to characterise the complete state-space of possible trajectories and is not dependent on the ground truth course structure. We intend to pursue this use of our Bayesian model in further work. A second direction for future methodological research would be to look at higher order network models which do not make Markovian assumptions. Indeed, considering the full trajectories of each student could yield higher order temporal dependencies that may better model the transitions between tasks[35].

Finally, it would be important to investigate the practical implications of our framework. Firstly, we are able to make real-time predictions of student success given limited observed sequences; therefore, we could implement our framework into a live course and make live student interventions to measure changes in student performance. This experiment would provide interesting insights into precision/personalised learning. Secondly, our framework provides the potential for in-depth analysis into course design by educators and learning designers to re-organise a course based on task importance, which could then be benchmarked against the original course. An experiment of this form would confirm the practicality of our framework and provide evidence for analytically driven decision making in education.

Received: 24 July 2020; Accepted: 1 January 2021
Published online: 02 February 2021

References

1. Knight, S., Wise, A. F. & Chen, B. Time for change: Why learning analytics needs temporal analysis. *J. Learn. Anal.* **4**, 7–17. https://doi.org/10.18608/jla.2017.43.2 (2017).
2. Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open university learning analytics dataset. *Sci. Data* **4**, 170171. https://doi.org/10.1038/sdata.2017.171 (2017).
3. Barbera, E., Gros, B. & Kirschner, P. Paradox of time in research on educational technology. *Time Soc.* **24**, 96–108. https://doi.org/10.1177/0961463X14522178 (2015).
4. Bloom, K. & Shuell, T. J. Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *J. Educ. Res.* **74**, 245–248. https://doi.org/10.1080/00220671.1981.10885317 (1981).
5. Peach, R. L., Yaliraki, S. N., Lefevre, D. & Barahona, M. Data-driven unsupervised clustering of online learner behaviour. *NPJ Sci. Learn.* **4**, 1–11. https://doi.org/10.1038/s41539-019-0054-0 (2019).
6. Kapur, M., Voiklis, J. & Kinzer, C. K. Sensitivities to early exchange in synchronous computer-supported collaborative learning (CSCL) groups. *Comput. Educ.* **51**, 54–66. https://doi.org/10.1016/j.compedu.2007.04.007 (2008).
7. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T. & Rohrer, D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychol. Bull.* **132**, 354. https://doi.org/10.1037/0033-2909.132.3.354 (2006).
8. Wise, A. F., Perera, N., Hsiao, Y.-T., Speer, J. & Marbouti, F. Microanalytic case studies of individual participation patterns in an asynchronous online discussion in an undergraduate blended course. *Internet High. Educ.* **15**, 108–117. https://doi.org/10.1016/j.iheduc.2011.11.007 (2012).
9. Munch, E. A user's guide to topological data analysis. *J. Learn. Anal.* **4**, 47–61. https://doi.org/10.18608/jla.2017.42.6 (2017).
10. Kloft, M., Stiehler, F., Zheng, Z. & Pinkwart, N. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, 60–65. https://doi.org/10.3115/v1/w14-4111 (2014).
11. Halawa, S., Greene, D. & Mitchell, J. Dropout prediction in MOOCS using learner activity features. In *Proceedings of the Second European MOOC Stakeholder Summit*, Vol. 37, pp. 58–65 (2014).
12. Taylor, C., Veeramachaneni, K. & O'Reilly, U.-M. Likely to stop? Predicting stopout in massive open online courses. *arXiv preprint*. arXiv:1408.3382 (2014).
13. Haythornthwaite, C. & Gruzd, A. Exploring patterns and configurations in networked learning texts. In *2012 45th Hawaii International Conference on System Sciences*, pp. 3358–3367. https://doi.org/10.1109/HICSS.2012.268 (IEEE, 2012).
14. Carroll, P. & White, A. Identifying patterns of learner behaviour: What business statistics students do with learning resources. *INFORMS Trans. Educ.* **18**, 1–13. https://doi.org/10.1287/ited.2016.0169 (2017).
15. Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A. & Goodrich, V. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, pp. 103–112. https://doi.org/10.1145/2567574.2567589 (2014).
16. Papamitsiou, Z. & Economides, A. A. Temporal learning analytics for adaptive assessment. *J. Learn. Anal.* **1**, 165–168. https://doi.org/10.18608/jla.2014.13.13 (2014).
17. Riel, J., Lawless, K. A. & Brown, S. W. Timing matters: Approaches for measuring and visualizing behaviours of timing and spacing of work in self-paced online teacher professional development courses. *J. Learn. Anal.* **5**, 25–40. https://doi.org/10.18608/jla.2018.51.3 (2018).

18. Ye, C. & Biswas, G. Early prediction of student dropout and performance in MOOCS using higher granularity temporal information. *J. Learn. Anal.* **1**, 169–172 (2014) (**10.18608/jla.2014.13.14**).
19. Ye, C. *et al.* Behavior prediction in MOOCs using higher granularity temporal information. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*, pp. 335–338. https://doi.org/10.1145/2724660.2728687 (2015).
20. Wise, A. F. & Shaffer, D. W. Why theory matters more than ever in the age of big data. *J. Learn. Anal.* **2**, 5–13. https://doi.org/10.18608/jla.2015.22.2 (2015).
21. Lund, K., Quignard, M. & Shaffer, D. W. Gaining insight by transforming between temporal representations of human interaction. *J. Learn. Anal.* **4**, 102–122. https://doi.org/10.18608/jla.2017.43.6 (2017).
22. Lee, A. V. Y. & Tan, S. C. Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *J. Learn. Anal.* **4**, 76–101. https://doi.org/10.18608/jla.2017.43.5 (2017).
23. Oshima, J., Oshima, R. & Fujita, W. A mixed-methods approach to analyze shared epistemic agency in jigsaw instruction at multiple scales of temporality. *J. Learn. Anal.* **5**, 10–24. https://doi.org/10.18608/jla.2018.51.2 (2018).
24. Halatchliyski, I., Hecking, T., Goehnert, T. & Hoppe, H. U. Analyzing the main paths of knowledge evolution and contributor roles in an open learning community. *J. Learn. Anal.* **1**, 72–93. https://doi.org/10.1145/2460296.2460311 (2014).
25. Reimann, P. Time is precious: Variable-and event-centred approaches to process analysis in CSCL research. *Int. J. Comput. Support. Collab. Learn.* **4**, 239–257. https://doi.org/10.1007/s11412-009-9070-z (2009).
26. Liu, R., Stamper, J. & Davenport, J. A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. *J. Learn. Anal.* **5**, 41–54. https://doi.org/10.18608/jla.2018.51.4 (2018).
27. Andrade, A., Danish, J. A. & Maltese, A. V. A measurement model of gestures in an embodied learning environment: Accounting for temporal dependencies. *J. Learn. Anal.* **4**, 18–46. https://doi.org/10.18608/jla.2017.43.3 (2017).
28. Chen, B., Resendes, M., Chai, C. S. & Hong, H.-Y. Two tales of time: Uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interact. Learn. Environ.* **25**, 162–175. https://doi.org/10.1080/10494820.2016.1276081 (2017).
29. Mahzoon, M. J., Maher, M. L., Eltayeby, O. & Dou, W. A sequence data model for analyzing temporal patterns of student data. *J. Learn. Anal.* **5**, 55–74. https://doi.org/10.18608/jla.2018.51.5 (2018).
30. Mendez, G., Ochoa, X., Chiluiza, K. & De Wever, B. Curricular design analysis: A data-driven perspective. *J. Learn. Anal.* **1**, 84–119. https://doi.org/10.18608/jla.2014.13.6 (2014).
31. Greenbury, S. F., Barahona, M. & Johnston, I. G. HyperTraPS: Inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways. *Cell Syst.* **10**, 39-51.e10. https://doi.org/10.1016/j.cels.2019.10.009 (2020).
32. Johnston, I. G. & Williams, B. P. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Syst.* **2**, 101–111. https://doi.org/10.1016/j.cels.2016.01.013 (2016).
33. Johnston, I. G. *et al.* Precision identification of high-risk phenotypes and progression pathways in severe malaria without requiring longitudinal data. *npj Digit. Med.* **2**. https://doi.org/10.1038/s41746-019-0140-y (2019).
34. Bloom, B. S., Englehard, M., Furst, E., Hill, W. & Krathwohl, D. *Taxonomy of educational objectives: The classification of educational goals* (1956). arXiv:9511007v1.
35. Myall, A. C. *et al.* Network memory in the movement of hospital patients carrying drug-resistant bacteria. *arXiv preprint*arXiv :2009.14480 *(2020).*

Acknowledgements

We would like to thank Dr. Nai Li for assistance with data collection and interpretation. We would also like to thank Prof Alan Spivey for helping promote the project and attain funding from Imperial College London. This research has been funded by a President's Excellence Award from Imperial College London. M.B. and S.N.Y. acknowledge support from EPSRC award EP/N014529/1 funding the EPSRC Centre for Mathematics of Precision Healthcare at Imperial.

Author contributions

R.L.P., D.J.L., S.N.Y. and M.B. conceived the study, R.L.P. collected the data, S.F.G. developed the source code, R.L.P. and S.F.G. conducted the analysis, R.L.P., S.F.G., I.G.J. and M.B. wrote the manuscript and generated figures. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-81709-3.

Correspondence and requests for materials should be addressed to R.L.P. or M.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.