HyperTraPS: Inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways

Sam F. Greenbury[a,c,e], Mauricio Barahona[a,c], Iain G. Johnston[b,c,d,*]

[a]Department of Mathematics, Imperial College London, UK
[b]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bergen, Norway
[c]EPSRC Centre for the Mathematics of Precision Healthcare, Imperial College London, UK
[d]Alan Turing Institute, London, UK
[e]ITMAT Data Science Group, Imperial College London, UK

Abstract

The explosion of data throughout the biomedical sciences provides unprecedented opportunities to learn about the dynamics of evolution and disease progression, but harnessing these large and diverse datasets remains challenging. Here, we describe a highly generalisable statistical platform to infer the dynamic pathways by which many, potentially interacting, discrete traits are acquired or lost over time in biomedical systems. The platform uses HyperTraPS (hypercubic transition path sampling) to learn progression pathways from cross-sectional, longitudinal, or phylogenetically-linked data with unprecedented efficiency, readily distinguishing multiple competing pathways, and identifying the most parsimonious mechanisms underlying given observations. Its Bayesian structure quantifies uncertainty in pathway structure and allows interpretable predictions of behaviours, such as which symptom a patient will acquire next. We exploit the model's topology to provide visualisation tools for intuitive assessment of multiple, variable pathways. We apply the method to ovarian cancer progression and the evolution of multidrug resistance in tuberculosis, demonstrating its power to reveal previously undetected dynamic pathways.

Keywords: HyperTraPS, trait evolution, phylogenetic character mapping, Bayesian inference, precision healthcare, cancer progression models

1. Introduction

Many problems in biology, medicine, and throughout the sciences involve the serial stochastic acquisition of discrete features or traits. These traits may be, for example, the symptoms experienced by a patient during progressive diseases, the genetic and physiological features underlying cancer progression, or the acquisition of drug-resistance traits in pathogens. Understanding the dynamics of these processes has the potential to inform targetted therapies, reveal biological mechanisms, and predict future behaviours, and has been an open challenge throughout the data explosion in biomedical sciences (Colijn et al., 2017).

Existing methods to reconstruct the past, and predict the future, of processes involving discrete trait acquisition have emerged from both the cancer science and evolutionary literatures. In the cancer field, disease-related alterations are classified as progressive 'hallmarks' (Hanahan and Weinberg, 2000, 2011). Several approaches, reviewed in Beerenwinkel et al. (2015), utilise computational methods for understanding the way in which cancer progresses via hallmarks at the genetic level (Schwartz and Schäffer, 2017). These methods range from stochastic models employing Markov chains for acquisition on graphs such as in Hjelm et al. (2006), to Bayesian network approaches where trees, forests or directed acyclic graphs (DAGs) are to be inferred from the data (Szabo and Boucher, 2002; Beerenwinkel et al., 2007; Gerstung et al., 2009; Loohuis et al., 2014; Ramazzotti et al., 2015). This field often focusses on independent samples exhibiting differing presence of alterations (cross-sectional data) for reconstructing oncogenetic models (Beerenwinkel et al., 2015) to discover progression pathways, or potentially causal relationships between markers, in patients.

Evolutionary and phylogenetic approaches for inferring trait dynamics, by contrast, must account for the relatedness of individuals and the possibility that a given state in a progressive system is inherited from an ancestor. Notable models that have attempted to solve this problem have included Simmap (Bollback, 2006), a Markov Chain Monte Carlo (MCMC) approach sampling character mappings on a phylogeny, Ordermutation (Youn and Simon, 2012), and Reversible Jump MCMC (RJ-MCMC) methodology also applied to a master equation formulation of character dynamics (Pagel and Meade, 2006). Such approaches have been utilised for understanding

[*] Lead contact
Email address: iain.johnston@uib.no (Iain G. Johnston)

the evolution of phenotypic traits in populations (Mahler et al., 2010; Watts et al., 2015). In connection with cancer progression, recent modelling approaches aim to reconstructing 'phylogenetic' cancer models from sources such as single-cell sequencing data (Beerenwinkel et al., 2015; Ross and Markowetz, 2016; Zafar et al., 2017; Ramazzotti et al., 2017).

Challenges remain in applying these algorithms to dissect the dynamics of systems involving many, potentially coupled, traits. Existing methods may assume a limited number of, or limited interactions between, traits. Computational runtime often scales exponentially with the underlying number of traits, and frequently exhibits challenging scaling with the number of observations. This scaling limits the applicability of some approaches to many forms of biomedical data, particular given modern trends of increasing data volumes and heterogeneity. Further, several approaches for inferring disease or evolutionary pathways are rather system-specific. In other words, they can process, for example, data on chromosomal aberrations in cancer progression, but are not readily generalised to other (or mixed) data types or diseases. This specificity can be a strength, allowing a more targetted interpretation, but relies on there being specific interest and funding in a particular disease to design a tailored approach for it.

A recent approach, HyperTraPS (hypercubic transition path sampling) (Johnston and Williams, 2016), aimed to address these issues, allowing the inference of the dynamics of many coupled traits from general observational data following arbitrary (but known) phylogenetic relationships. HyperTraPS represents progressive dynamics as paths on a hypercubic space connecting all possible patterns of trait presence and absence, and uses observations of intermediate states to learn the most likely pathways of progress through this space. In this way, snapshot data can be used to learn the probabilistic structure of dynamic pathways, which have in turn been used to identify the mechanisms underlying the evolutionary dynamics of $L = 65$ mtDNA genes (Johnston and Williams, 2016) and C_3 to C_4 photosynthesis (Williams et al., 2013).

To date, HyperTraPS has only been used to address these specific evolutionary questions. However, in the current era of large-scale scientific and biomedical data, questions about the structure of dynamic pathways are expanding and becoming increasingly pertinent to evolutionary biology and precision medicine. Hypercubic inference represents a powerful new way of addressing these questions, but a general platform for its application, interpretation, and visualisation remains absent. Such a platform would provide many advantages over the current state of the art: large-scale datasets can be readily analysed, different types of observational data can be used (cross-sectional, longitudinal, and/or phylogenetically coupled observations); Bayesian quantification of uncertainty and a completely unrestricted set of states and transitions can be applied, and competing pathways and their detailed structure can be resolved and characterised, facilitating the identification of progression mechanisms. In principle, any dataset where the relationship of the samples is known or can be inferred is amenable to this detailed analytic approach.

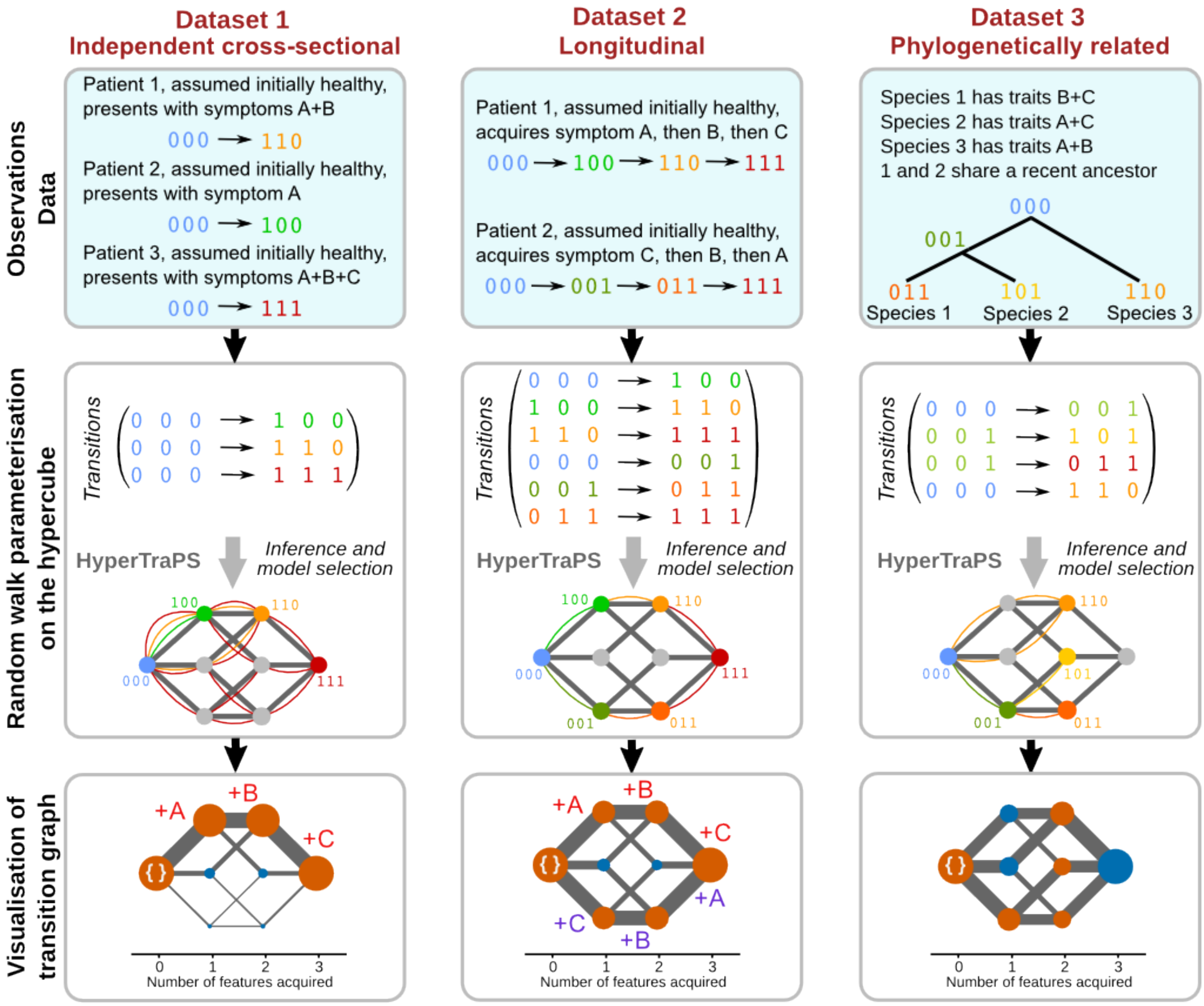
Here, we address this target, presenting a novel and expansive set of methodological developments to allow the inference of dynamic pathways from highly general datasets. We embed HyperTraPS in a new and efficient platform for parametric inference and model selection, simultaneously allowing Bayesian inference of dynamic pathways and the identification of model structures that best describe the dynamics and interactions contained within a given set of observations. This model selection simultaneously guards against overfitting and reveal mechanistic insights, namely the extent to which interactions between features dictate the dynamics of the observed system. Models identified in this way have the strongest power to predict out-of-sample observations, which we demonstrate with synthetic and real-world examples, illustrating the predictive power of the approach. To further facilitate interpretation of the inference outcomes, we introduce approaches for intuitively visualising and comparing the high-dimensional pathways inferred from complex datasets, which may include multiple distinct orderings for the acquired traits. While this overall approach is thus highly general, its Bayesian nature means that domain-specific knowledge constraining a system's behaviour can be readily included for a specific application. This could include, for example, insight into biological mechanisms that forbids feature A appearing before feature B, or that suggests the presence of feature C makes feature D twice as likely.

We illustrate the performance of these methods in three different scenarios: with synthetic datasets; with two datasets on different scales on the progressive acquisition of genetic alterations in ovarian cancer; and with a recent large-scale dataset on drug-resistant tuberculosis. In these final two cases we demonstrate and discuss several new insights into progression dynamics that the HyperTraPS platform provides. We compare this platform to other approaches from the disease progression and evolutionary literatures for trait inference, highlighting its intersection between these fields and consequent general power and applicability. We conclude by discussing the breadth of applications in the expanding fields of precision medicine, data science, and evolutionary inference, and provide an open source package for the code.

2. Results

2.1. Inferring dynamic pathways involving coupled traits on general state spaces

HyperTraPS represents every possible state of a system with L features or traits (we use these terms synonymously here) as a binary string of length L, where 0 and 1 at the ith position correspond respectively to



Patient 1, assumed initially healthy, presents with symptoms A+B

000 → 110

Patient 2, assumed initially healthy, presents with symptom A

000 → 100

Patient 3, assumed initially healthy, presents with symptoms A+B+C

000 → 111

Patient 1, assumed initially healthy, acquires symptom A, then B, then C

000 → 100 → 110 → 111

Patient 2, assumed initially healthy, acquires symptom C, then B, then A

000 → 001 → 011 → 111

Species 1 has traits B+C
Species 2 has traits A+C
Species 3 has traits A+B
1 and 2 share a recent ancestor

Fig. 1: The HyperTraPS pipeline for learning dynamics underlying cross-sectional and/or longitudinal observations. HyperTraPS allows dynamic inference with three classes of input data. In each case, presence/absence of traits are labelled with a binary marker, and temporal relationships between observations (if present) are invoked to represent observed samples as observed transitions. The likelihood that a given set of edge weights on the underlying hypercubic transition network will give rise to the observed transitions can be calculated efficiently using a path sampling approach (coloured lines). Each illustrative hypercube corresponds to a dataset, with colour coded curved edges and states showing the possible paths that can be taken to reach observed samples. Embedding this likelihood calculation in a Bayesian inference scheme allows posterior weights on inferred transition graphs to be computed, constituting a complete characterisation of the dynamic systems. In the final, visualisation step, the inferred transition graph is embedded and plotted, with edge widths and vertex areas are proportional to the posterior weighting, vertices coloured according to whether they reflect observed (orange) or hidden (blue) states, and paths labelled by the progressive acquisition of features.

absence or presence of the ith trait. Traits are acquired stochastically and irreversibly, according to transition probabilities linking states on a hypercubic transition graph (Fig. 1). We consider instances of an evolving or progressing system as an ensemble of random walkers on this graph. As in a hidden Markov model Murphy (2012), observations are assumed to arise through signals randomly emitted by these walkers; a signal corresponds to the current set of acquired traits of the random walker. The task at the core of HyperTraPS is to compute the likelihood of observing a set of emissions that match the transitions in a dataset, given a parameterisation W describing the transition probabilities on the edges of the hypercube.

In STAR Methods, Fig. 1, and Supplementary Figure S1 , we outline the HyperTraPS algorithm to estimate this likelihood given a set of observations. As Fig. 1 illustrates, these observations can be independent and cross-sectional (for example, single snapshots of symptom presence/absence in independent patients), longitudinal (for example, time series of symptom presence/absence in the same patients over time), and/or phylogenetically related (for example, evolving traits which may be inherited from ancestor to descendent). Cross-sectional and longitudinal data structures involve many independent evolutionary processes running in parallel; phylogenetic data structures involve an initially single process that may branch, with different branches subsequently evolving independently. In contrast to previous approaches (Johnston and Williams, 2016; Williams et al., 2013), we embed the core likelihood calculation in an auxilary pseudo-marginal MCMC (APM MCMC) framework (Murray and Graham, 2015) to allow more efficient Bayesian inference of the hypercubic transition network supporting the observed dynamics. The APM MCMC embedding overcomes potential issues arising from uncertainty in the likelihood estimates for long pathway calculations (STAR Methods), better guaranteeing that the MCMC process will mix well and converge to a consistent posterior in the case of large, sparse inference challenges. For example, in the ovarian cancer inference presented below, the APM embedding reduced the characteristic MCMC mixing time by a factor of 5. APM MCMC makes it possible to address systems involving dozens of sparsely sampled traits, as we demonstrate below.

The next important consideration in this inference process is how this transition network is parameterised. Individually parameterising each of $L2^{L-1}$ hypercubic edges represents a substantial inference challenge for (likely) very little model fit reward. Instead, we propose a hierarchy of parameter representations (Supplementary Figure S2 ; STAR Methods). For the *zero order* model every feature has equal probability of acquisition. All edges on the transition network thus have the same weight, requiring no parameters. In the *first order* model, every feature has an independent acquisition probability regardless of current state. Transition edge weights between two states are thus exclusively determined by the trait that distinguished the two states (requiring $k = L$ parameters). In the *second order* model, every feature's acquisition probability depends independently on the presence of each other feature. Transition edge weights between two states thus depend on the distinguishing trait and the presence/absence of each other trait (requiring L^2 parameters; as in (Johnston and Williams, 2016)). Higher order models, including the full $L2^{L-1}$ set naturally follow, introducing more complex interactions between the co-occurrence of features (as in Williams et al. (2013)). The appropriate choice of parameterisation is dictated by the generative processes underlying the observed data; if trait acquisitions are independent, the parsimonious first-order model is more appropriate; if traits interact pairwise, the second-order model will be required to capture the dynamics. A given dataset may be best described by an intermediate representation between two of these cases.

To identify the optimal parameter representation for a given dataset, we introduce methods for regularising the inferred model parameterisations (see STAR Methods), allowing the appropriate choice of model structure to describe the observed data and a means of generating maximum likelihood parameterisations without overfitting. As we demonstrate below, the regularisation process allows us to distinguish simple cases, where all dynamics can be described by traits behaving independently, from more complex cases where the acquisition of one or more traits influences the probability of acquisition of other traits. This combination of an efficient and general inference platform, a process for model selection, and a new toolbox for visualising and interpreting inferred posteriors, allows us for the first time to apply HyperTraPS to a dramatically expanded range of biomedical questions.

2.2. Inference of pathways from synthetic data

To illustrate the ability of HyperTraPS to characterise dynamics from independent cross-sectional samples, we constructed two cross-sectional datasets with different underlying progressions. The first (D_1; Fig. 2A(i)) involves samples taken uniformly from each state along a single trajectory, where features are accumulated from *left to right*. For example, for $L = 3$, the sequence of acquisition would be $000 \rightarrow 100 \rightarrow 110 \rightarrow 111$. The second ($D_2$; Fig. 2A(ii)) involves samples taken uniformly from states along two distinct progression pathways with exactly opposing temporal ordering of acquisition: one where features are acquired from *left to right* and the other where features are acquired from *right to left*. For example, for $L = 3$, this would correspond to the two trajectories $000 \rightarrow 100 \rightarrow 110 \rightarrow 111$ and $000 \rightarrow 001 \rightarrow 011 \rightarrow 111$.

We chose these structures to illustrate HyperTraPS' ability to infer both single and multiple competing pathways. For the single pathway, traits can be independent – a suitable ordering of the 'basal rates' is sufficient to

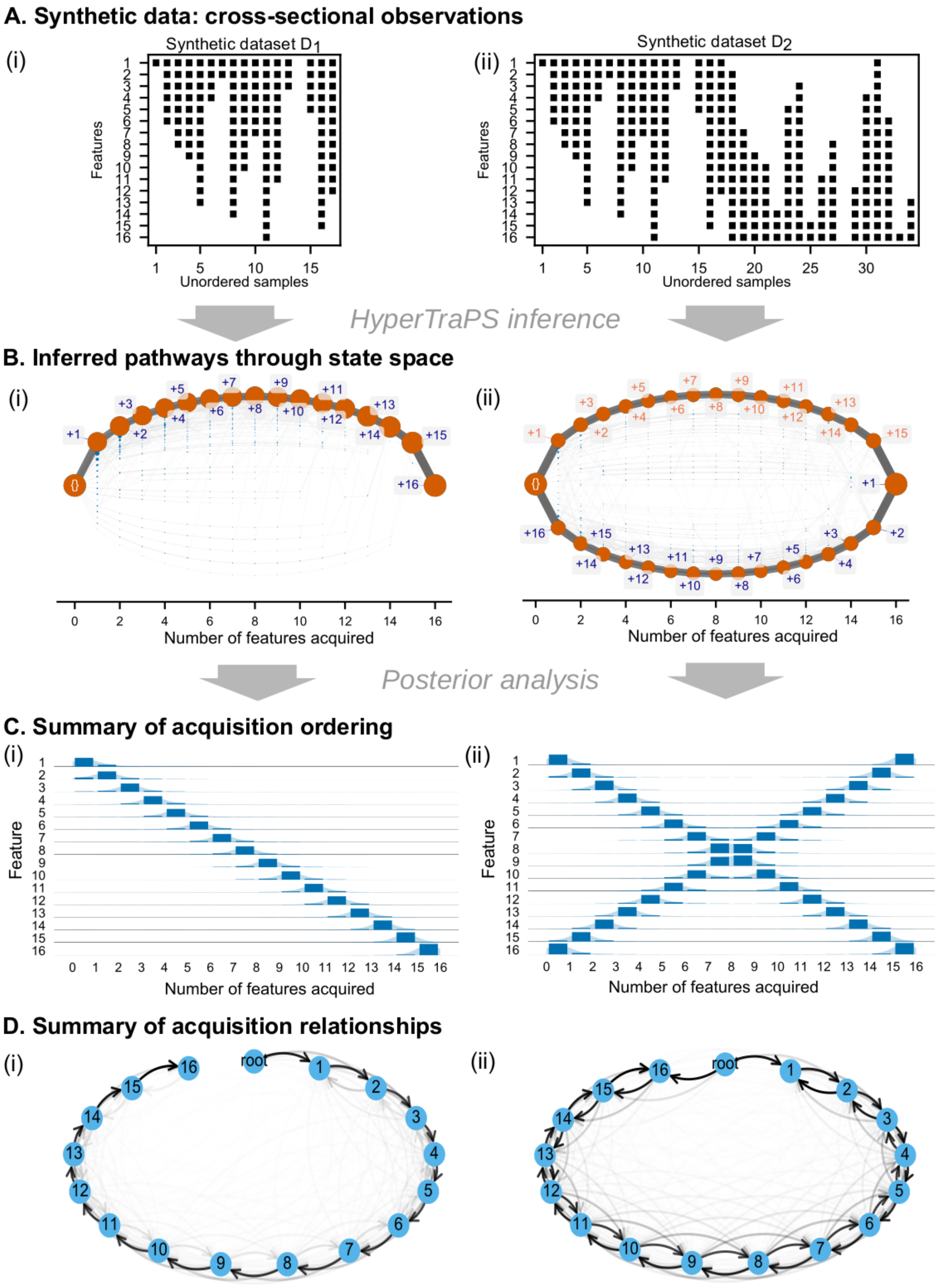


Fig. 2: HyperTraPS inference for two synthetic datasets. (A) Synthetic datasets used in the inference process. (i) Dataset D_1 supports only a single pathway; (ii) dataset D_2 supports two competing pathways with features acquired in opposing orders. **(B)** Inferred dynamics on the hypercubic transition graph. Edge widths and node areas are proportional to the number of times edges/nodes are encountered. States are plotted from left to right in order of the number of features acquired (embedding and labelling procedure described in STAR methods). The single pathway clearly dominates in (i), while the two competing pathways are clearly observable in (ii). **(C)** Inferred dynamics represented as the posterior probability that a feature (horizontal axis) is acquired at a given step (vertical axis). Bimodality in ordering posteriors (ii) reflect the presence of distinct progressions that exist in the underlying dynamics. **(D)** Inferred dynamics represented as a directed graph (edges run from left to right in these embeddings) summarising trait acquisition relationships to the previous acquisition. Paths on these graphs reflect possible acquisition ordering inferred by HyperTraPS: respectively a single pathway (i) and two pathways (clockwise and anti-clockwise) in opposite directions (ii).

generate the observations. By contrast, competing pathways require traits to interact – acquisition of traits on one pathway must repress acquisition of traits on the other pathway.

In Fig. 2, we show the structure of the data and the outcomes of the inference process. To visualise the learned dynamic behaviour, we use a customised algorithm (described in further detail in STAR methods) to project the inferred hypercubic transition network into two dimensions, arranging states with increasing numbers of features from left to right (Fig. 2B). A single dominant progression is clear for Fig. 2B(i), while the two progressions are clearly shown in Fig. 2B(ii). Fig. 2C shows an alternative representation: the posterior probabilities with which each trait is acquired in each possible ordering. Again, the dynamics corresponding to the simple single pathway and the more complex competing two-pathway model are clearly visible.

In this extreme example, the inferred ordering distributions for all but the central traits in the multiple-pathway case (ii) exhibit *bimodality*. Generally in such histograms from HyperTraPS posteriors, bimodality (and multimodality more generally) reflects structurally distinct progression pathways (for example, where a feature can be acquired early or late, but not at intermediate stages), while unimodal distributions reflect sets of pathways with a consistent structural trend. The width of such modes reflects the amount of variability in the order for which a feature is acquired in the progression associated with the mode. Multimodal distributions in these plots provide a suggestive signature of distinct dynamic pathways of the system. In STAR Methods and Supplementary Table S1 , we compare this inference of competing pathways to existing alternative approaches and show that HyperTraPS has a unique ability to resolve and characterise multiple progressive pathways.

In Fig. 2D, we represent dynamics from the inference process as *probabilistic feature graphs* (PFGs), allowing more direct comparison with existing approaches. These PFGs summarise the probability the feature Y is acquired next, given that feature X was last to be acquired (see STAR Methods). Once more, in Fig. 2D, the single monotonic path in (i) and two paths for (ii) are clearly visible.

In Fig. 3 and Supplementary Figure S3 - Supplementary Figure S5 , we demonstrate the performance of the inference process under availabilities and structures of source data, and in the presence of prior knowledge about pathways. Supplementary Figure S3 shows that characteristic pathways can readily be identified under each of the three different types of data from Fig. 1. The resulting posterior distributions are sharper for cross-sectional data than for longitudinal and phylogenetic data, reflecting the fact that the independent samples from cross-sectional data provide more evidence for corresponding pathways than the coupled data in the other cases. Fig. 3A shows the ability of HyperTraPS to identify pathways given limited data ($N = 10$ observation are sufficient to broadly characterise a single pathway for an $L = 16$ system; $N = 50$ gives near-perfect reconstruction). Even for competing pathways, $N \geq 20$ serves to provide information on pathway structure in this case. Fig. 3B and Supplementary Figure S4 demonstrates that HyperTraPS can readily discern several completely independent pathways (8 pathways can be readily identified for the $L = 16$ system; 16 completely independent pathways pose more of a challenge). Finally, Fig. 3C and Supplementary Figure S5 highlight the Bayesian nature of HyperTraPS by demonstrating how the inclusion of prior knowledge about pathway structure can help resolve degeneracy in the identified solutions, for repeated and/or incomplete observations. To summarise, HyperTraPS can readily identify pathway structure including multiple, competing, independent pathways, using limited volumes of data, and can readily harness prior knowledge.

This final point is particularly pertinent when applying HyperTraPS to specific scientific questions. When uninformative priors are used, HyperTraPS is a highly general approach, where mechanistic inference is guided by the data alone. For domain-specific cases – for example, particular diseases, or particular metabolic pathways – subject-specific knowledge may constrain the allowed pathways (for example, mechanistic insight may forbid or favour transitions between particular states). In these cases, the inclusion of this knowledge via prior distributions as in Fig. 3C and Supplementary Figure S5 can readily and generally be used to constrain the posterior dynamics supported by HyperTraPS.

In STAR Methods and Supplementary Figure S6 - Supplementary Figure S11 , we further expand upon these test cases (Supplementary Figure S6 - Supplementary Figure S7) and the interpretation of pathway dynamics (Supplementary Figure S8), and demonstrate that HyperTraPS successfully learns pathways in the case of partial (Supplementary Figure S9), noisy (Supplementary Figure S10), and non-uniform (Supplementary Figure S11) sampling.

2.3. Model regularisation and validation

We next demonstrate how regularisation can be used to determine the optimal model structure required to describe and predict features of the two synthetic datasets. D_1 is produced by a model with no trait interactions, and hence requires only L independent parameters to reproduce its dynamics. D_2 requires interactions between traits: progress along one pathway must suppress progress along the other. More parameters are thus required to encode these interactions to adequately match the data. We therefore asked if, given a range of starting model representations, the regularisation process could identify the appropriate number of parameters for each case.

Fig. 4A demonstrates this regularisation process. For D_1, the first-order model remains intact with its original L parameters, and the second-order model is reduced from L^2 to $\sim L$ parameters, reflecting the fact that L (and

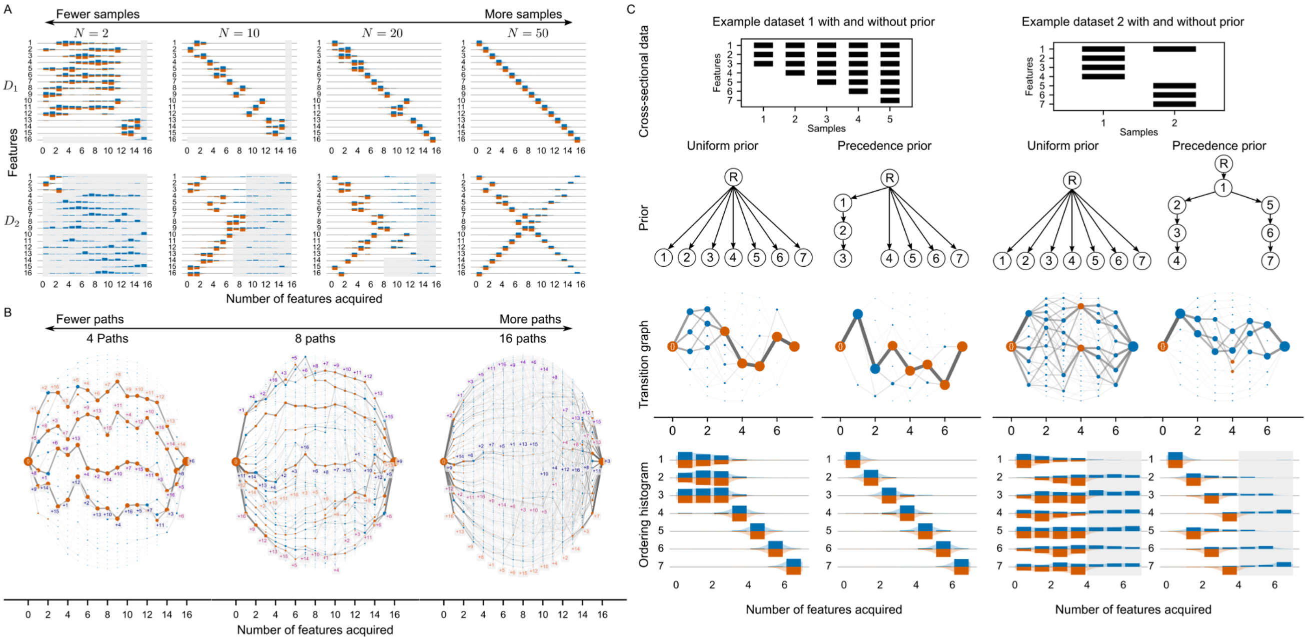


Fig. 3: HyperTraPS inference of pathway structures under different conditions. **(A)** Inference of D_1 (single pathway) and D_2 (two competing pathway) structures with $L = 16$ and increasing number of sample states N. Both pathway structures are readily identified for $N \geq 20$. Lower N challenges reconstruction of pathways, although the outline of the single pathway is still visible for $N = 10$. At lower N, the posteriors tend towards the uniform priors. **(B)** Inference of p competing pathways with $L = 16$ and $N = 16p$ samples. 8 completely distinct pathways are readily identified with clearly distinguished posterior density; 16 independent pathways pose more of a challenge but are still identified. **(C)** Including prior information in HyperTraPS inference. (left) Single pathway dataset without any observations with fewer than three acquisitions. Inference without prior information of these first three features leads to a uniform inference of acquisition order; including a prior tree (see text) recovers the true ordering. (right) Two competing trees of acquisition order for prior information. Without prior information there is large heterogeneity in the order and precedence of feature acquisition following inference. Including prior information canalises the inferred pathways and recovers the original structure.

only L) parameters are required to capture the single-pathway dynamics. This regularised second-order model performs equally well to the first-order model.

For D_2, with two competing pathways, the first-order model fails to capture the observed behaviour even with its full set of L parameters. The regularisation process reduces the first-order model to ~ 0 parameters: as no instance of model 1 can adequately describe the observations, the parameter set is minimised for parsimony. By contrast, the second-order model is reduced to $\sim 2L$ parameters, which provides an optimal description of the data. The requirement for higher-order terms here is a consequence of the trait-interaction terms in the second-order model allowing the required cross-repression of pathways, making it a better explanatory model in this case.

To validate these findings and explore the predictive power of our inference platform, we split the data into two halves to form a training and test dataset. We obtained posteriors from the training set for each model, and computed the likelihood associated with the test set for these inferred posteriors. Fig. 4B shows the AIC scores for the full model, and the log-likelihoods for training and validation datasets. For the single-pathway dataset D_1, the first order model and second order model provide similar explanatory power in the full model, and predictive power in the validation experiment, both improved over the zero order model (null model). For the two-pathway dataset D_2, the second order model enhances predictive power compared to both the first order and null models, and regularisation improves the parsimony of this model with no cost to model fit ($p < 0.001$ for the a likelihood ratio test against the null model).

2.4. *Comparison with existing inference approaches*

We next sought to compare the outputs of the HyperTraPS inference process to existing approaches to infer dynamic pathways from data (Fig. 5). We highlight here that HyperTraPS is, to our knowledge, the only inference approach that attempts to learn the transition rates (with uncertainties) between every possible state of a system. Other approaches typically focus on a reduced subset of states. The full, high-dimensional posteriors inferred by HyperTraPS therefore cannot be readily compared with the outputs of other approaches. However, summaries of these posteriors, losing some information, can more naturally be compared with lower-dimensional alternatives.

To this end, we compared reduced summaries of the dynamics learned by HyperTraPS with the Bayesian networks derived from the Capri algorithm Ramazzotti et al. (2015) and Conjunctive Bayes Network approaches (Montazeri et al., 2016) (using MC-CBN, the most recent CBN package for large or small scale inference), two commonly used Bayesian network methods in the literature, using synthetic datasets (Fig. 5). These approaches produce directed acyclic graphs (DAGs) on the set of features, where an edge between X and

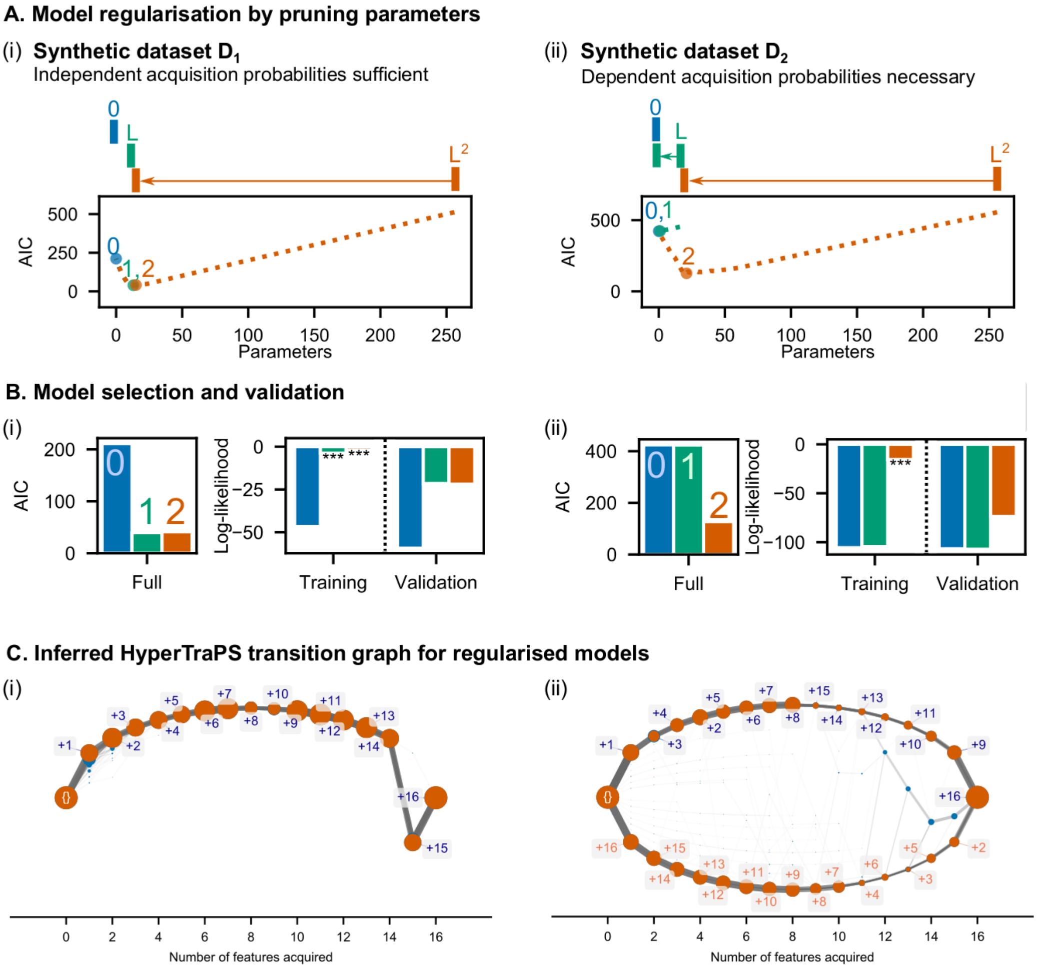


Fig. 4: Regularisation and model comparison for HyperTraPS inference. We compare models '0' (zeroth order, 0 parameters, all traits are acquired with the same independent probability); '1' (first order, L parameters, acquisition probabilities are independent but may differ); and '2' (second order, L^2 parameters, pairwise interactions between trait acquisition probabilities) for the datasets D_1 and D_2 in Fig. 2. **(A)** Model regularisation. Parameters are greedily pruned from each inferred model to identify a reduced parameter set that minimises AIC. The turning points illustrating an optimally sparse parameterisation are marked for each model. **(B)** Model selection and validation. (left) AIC scores for the regularised version of each model; (right) likelihoods for the training and validation datasets (see text). For the full and training dataset, stars give the p-value from a likelihood ratio in comparison to the zero order model (the null model) with significance levels of $*** < 0.001$, $** < 0.01$ and $* < 0.05$. **(C)** Inferred dynamics on the hypercubic transition graph for the regularised first order model for D_1 and for the regularised second order model for D_2. Each corresponding pathway is still well captured despite substantial parameter reduction.

Y denotes an inferred causal relationship between X and Y. Such representations allow for possible causal relationships between features to be found, but *a priori* impose that such relationships exist and are monotonic. For example, if trait X influences the presence of trait Y, trait Y may not influence trait X. Overall ordering of feature acquisition may not be unique (a joint probability distribution of events may have underlying degeneracy in the order of those events), but the monotonic relationship between features does impose partial ordering. In HyperTraPS, no monotonic precedence is imposed between features: X may influence Y and Y may influence X. This relaxation allows, for example, cross-repression of traits, as we shall see for dataset D_2. For comparison with other approaches, we condense the full output of inference (DAGs in state space, i.e. on the hypercube) into graphs in feature space.

As seen in Fig. 5, for the single-pathway case of dataset D_1, all graphs have the same structure and therefore are in agreement over the single pathway that most likely explains the data. For the competing pathway case of dataset D_2, the outputs are different in each case. The HyperTraPS feature graph captures the dual pathways, with directed edges between each pair of non-root nodes. Capri is unable to resolve a meaningful relationship between features, because the competing pathways frustrate the assignment of temporal priority between the features. The outputted graph is therefore unable to recover a significant relationship between features representing precedence relationships. The Conjunctive Bayes Network is able to resolve one of the directed paths but not the other.

These comparisons have been performed with cross-sectional synthetic observations. As discussed above (Fig. 1A), HyperTraPS can also infer dynamic pathways given longitudinal and phylogenetically coupled data. Ref. Johnston and Williams (2016) demonstrated that HyperTraPS has several advantages over existing ap-

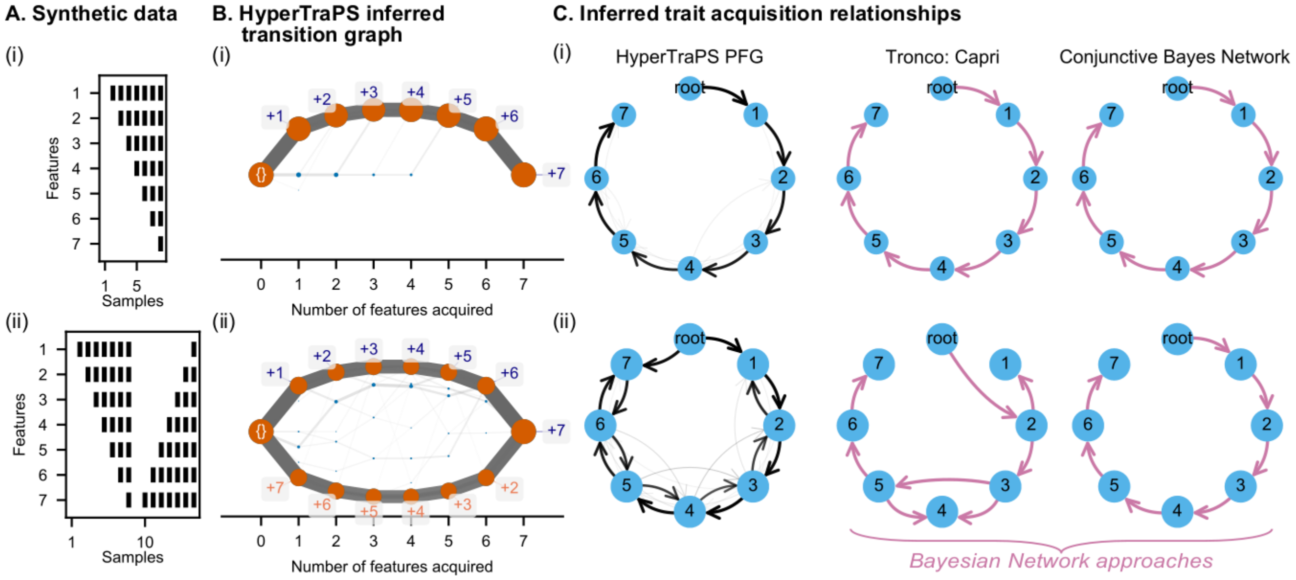


Fig. 5: Comparison between HyperTraPS and alternative inference approaches. (A) Synthetic datasets (i) and (ii) following the forms from Fig. 2. **(B)** Full inferred transition graphs from HyperTraPS. **(C)** Inferred transition graphs from HyperTraPS represented as probabilistic feature graphs, compared to alternative inference approaches. All approaches agree for the simple, single pathway (i). For the competing pathways (ii), the full inferred transition graph from HyperTraPS, and its structure summarised into trait contingencies, capture the two alternative pathway structures, while alternative approaches (highlighted) are more challenged, either presenting a combination of steps from both pathways or exclusively reporting one.

proaches for trait inference on phylogenies. In STAR Methods and Discussion, we pursue these comparisons further and show that HyperTraPS presents several scaling and performance advantages over alternative methods, again reflecting its ability to resolve independent pathways involving many coupled traits.

Taken together, these results provide support for our platform's ability to learn single progression pathways efficiently and also dissect competing progression pathways more directly than alternative approaches. We reiterate that, in addition to these coarse-grained readouts, HyperTraPS learns explicit probabilities for transitions between every state of a system, allowing a still finer resolution of dynamics.

2.5. Application to cross-sectional ovarian cancer data

To demonstrate HyperTraPS' ability to elucidate dynamic pathways of biomedical importance, we next asked whether our approach could be used to infer pathways of cancer progression. The field of cancer progression models is diverse, with many methods designed for performing inference with different types of data (Beerenwinkel et al., 2015; Schwartz and Schäffer, 2017). As Schwartz and Schäffer (2017) discuss, data relating to alterations in cancer broadly belong to three categories: bulk tumour samples from different patients, bulk tumour samples from different tumours within a single patient, or single cell data typically from a single tumour. Computational methods can broadly be categorised into those inferring the phylogenetic relationship of samples (their history and genealogy), and those inferring direct relationships between the features suggestive of precedence or progressions relating to feature acquisitions. We discuss the methods within the cancer progression model literature further in STAR Methods.

As illustrated in Fig. 1, HyperTraPS can both handle independent and arbitrarily dependent samples, and so can be used with any of the above types of dataset. We here focus on the case of independent bulk samples from different patients where there is no phylogenetic relationship between samples, as it is assumed that features are acquired during a patient's lifetime. Existing approaches for this problem (Beerenwinkel et al., 2015) focus on the reconstruction of different types of Bayesian network relating the acquisition of genetic alterations relating to the progression of cancer. As cancer is directly related to the acquisition of driver mutations that provide fitness advantage for the cells in which they are acquired, recent work such as Diaz-Uriarte (2018) has argued for the need to consider cancer progression from a different perspective in which features may have multiple orderings due to the high-dimensional structure of fitness landscapes and the potential presence of epistatic effects. The HyperTraPS platform directly allows this inference of multiple paths.

We first applied HyperTraPS to the well-studied dataset for chromosomal alterations in ovarian cancer, recovered through Comparative Genomic Hybridization (CGH) (Knutsen et al., 2005). This dataset is included in the Oncotrees package (Szabo and Boucher, 2002) and utilised in comparisons with the Caprese algorithm (Loohuis et al., 2014). The data consist of a sample of $N = 87$ patients for $L = 7$ chromosomal alterations associated with ovarian cancer, with the assumption that none of the alterations were present in the individual at birth.

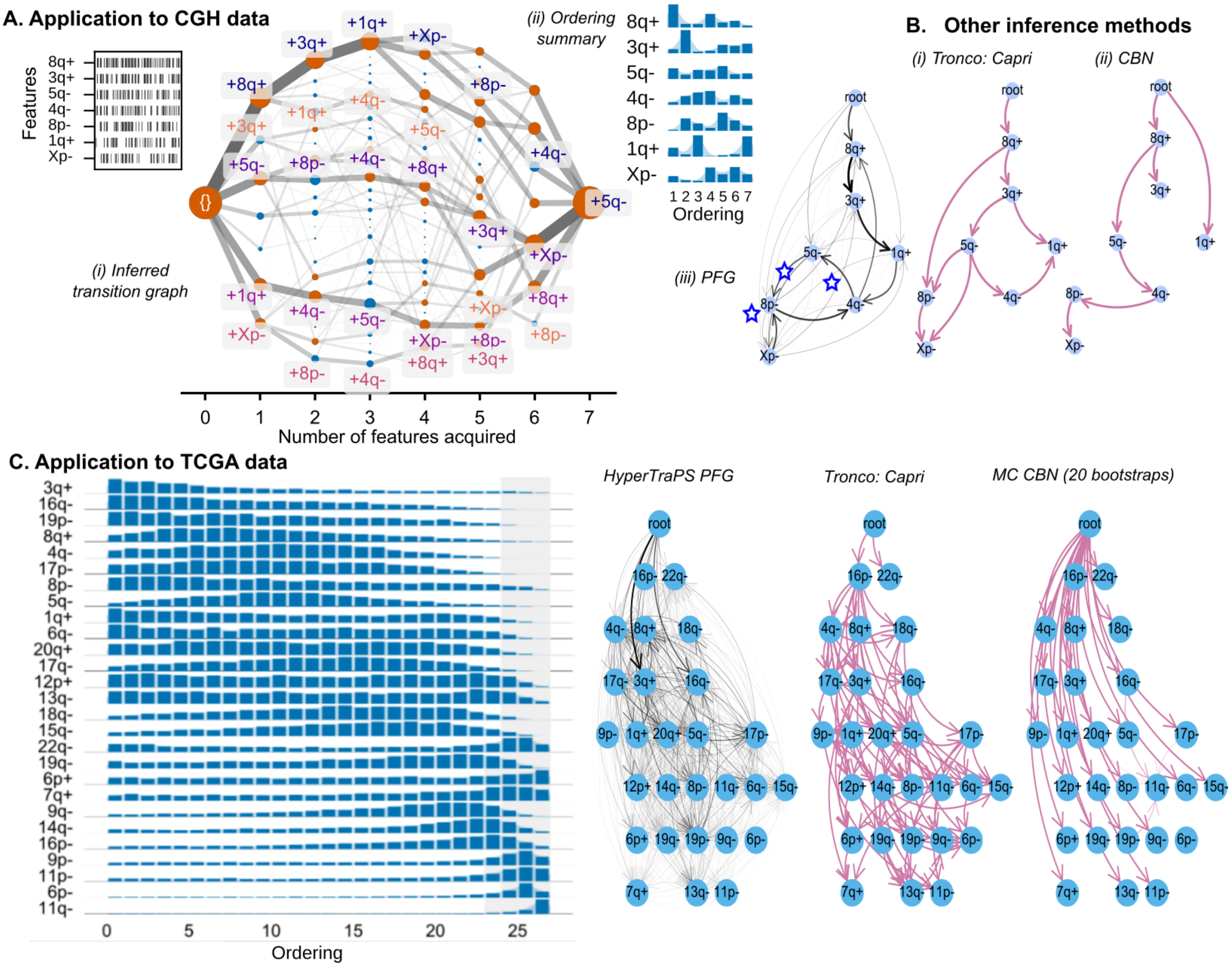


Fig. 6: HyperTraPS inference for ovarian cancer progression reveals canalised progression pathways and new transition information. (A) HyperTraPS inference applied to a dataset (inset) of cross-sectional observations of chromosomal aberrations in an ovarian cancer dataset. The inference process produces transition graph (i), summary ordering posterior (ii), and corresponding probabilistic feature graph (iii), reflecting the inferred dynamics of cancer progression. Progression pathways are substantially canalised, with the first acquired aberration feature largely determining the subsequent dynamics of the disease. Starred edges in (iii) correspond to edges present in the probabilistic feature graph, related to the $4q- \to 5q-$ system discussed in the text, that are absent in other approaches. **(B)** Trait relationships inferred with alternative computational approaches. HyperTraPS largely agrees with the core structure of alternative approaches (especially with Capri where there is less strict constraints on precedence) but reveals several additional features (illustrated with stars in (B)). For example, the $4q- \to 5q-$ pathway from is omitted and directly opposed in alternative approaches where only monotonic relationships between *5q-* and other features are permitted. Further, the canalised structure present in B is not naturally captured by the inferred outputs of the alternative approaches. **(C)** Inferred orderings of chromosomal changes in ovarian cancer progression using observations from the cancer genome atlas (TCGA) dataset, and corresponding inferred transition graphs from the TCGA inference compared to alternative approaches as in (B).

Fig. 6A and Supplementary Figure S12 provide a visual representation of the dataset, showing the presence/absence of each genetic alteration in each patient. Fig. 6A(i) shows the recorded transitions following parameter inference on the hypercube. A set of several constrained, well-defined paths are visible, with flexible ordering in the acquisition of initial features being apparent. Interestingly, the feature that is acquired first has substantial influence over the subsequent pathway structure, visible as the tightly constrained individual pathways in Fig. 6A(i) with rather few transitions between pathways, and as bimodal structure in the posterior summary plot in Fig. 6A(ii). This canalisation suggests substantial memory effects in the later stages of cancer progression.

To further examine the multiple non-monotonic pathways that the data may contain, we make use of the probabilistic feature graphs described above and in STAR Methods. Fig. 6A(iii) shows the probabilistic feature graph between each pair of features and Fig. 6B shows Bayes network representations of feature relationships from alternative approaches. Here, as above, each edge is directed and has a weight in proportion to the probability of acquiring feature Y having just acquired feature X. Elements of the core structure are shared between the HyperTraPS, Capri, and CBN approaches.

To demonstrate another example of where HyperTraPS' increased detail allows new insight into multiple pathways, we focus on several transitions that have strong edges in the HyperTraPS PFG that are missing from the other approaches. In both alternative Bayesian network approaches, an edge is present from *5q-* to *4q-* but never the other way around. The precedence in these models is due to the fact that *5q-* is more frequent than *4q-* and the need to ensure a monotonicity between features to construct the desired Bayesian network output. As HyperTraPS places no such restriction, it is capable of finding additional pathways in which *4q-* is acquired prior to *5q-*. As seen in Fig. 6B(i), this ordering may be achieved in several ways through the acquisition of *8p-*, *3q+* or *1q+* and, given the acquisition of those features, is in fact more likely to be acquired prior to *5q-*. The acquisition of *4q-* prior to *5q-* is indeed observed in 10 of 87 (11.5%) samples in the data.

Having gained substantial insight from this comparatively simple dataset, we next asked whether HyperTraPS could be used with the larger volumes of data that emerge from more modern genome-scale studies. To this end, we obtained raw data from the cancer genome atlas (TCGA) project (Bell et al., 2011). We converted these raw data into feature 'barcodes' over a variety of scales, yielding a set of cross-sectional datasets (see STAR Methods). First, we constructed a dataset describing the chromosomal regions in which each of the $N = 489$ patients had amplifications/deletions above the significance threshold defined in the study. This gave a dataset describing each patient's presence or absence of aberration in $L = 55$ regions. Secondly, we considered the subset of the $L = 27$ chromosomal regions marked as of particular interest in Fig. 1c of Bell et al. (2011).

Our APM MCMC embedding of HyperTraPS allowed the algorithm to readily produce posterior distributions in each case. In the first, larger, case, posteriors show a clear ordering in the acquisition propensity for different chromosomal features (Supplementary Figure S13). However, the large dynamic space associated with these $L = 55$ features makes more detailed interpretation of these posteriors rather laborious. This reflects a challenge in the application of HyperTraPS: while posteriors can readily be obtained for large numbers of features (Johnston and Williams, 2016), the interpretation of these posteriors can be challenged by the output volume.

Consistent with this, the results from the subset of regions are more interpretable (Fig. 6C). Here, clearly converged posterior distributions are visible, with some bimodality (for example, in features *1q+, 13q-* and *22q-*) suggesting the presence of competing pathways. In particular, bimodality in the *1q+* posterior reflects the multiple associated pathways in the previous CGH dataset (Fig. 6A). The orderings of other features from the CGH dataset are consistently reflected in HyperTraPS' treatment of the TCGA data, with the additional volume of data in the TCGA case helping to further detail posterior structure. Interpreted as a PFG (Fig. 6C), these posteriors highlight both the heterogeneity of, and strong structures within, the associated progression pathways. Strong early edges, for example, surround the *3q+* feature, linking $\emptyset \rightarrow$ *3q+* and *3q+* $\rightarrow$ *8q+*, and the *16q-* feature.

Other approaches do not capture several of these transitions. For example, 70 samples in the dataset possess the $3q+$ feature but not $8q+$ (compared to 71 which possess $8q+$ but not $3q+$), while the Capri Bayesian network is only able to identify a single causal relationship from $8q+ \rightarrow 3q+$ and the CBN approach does not identify any edge between the pair (due to this large proportion of conflicting samples).

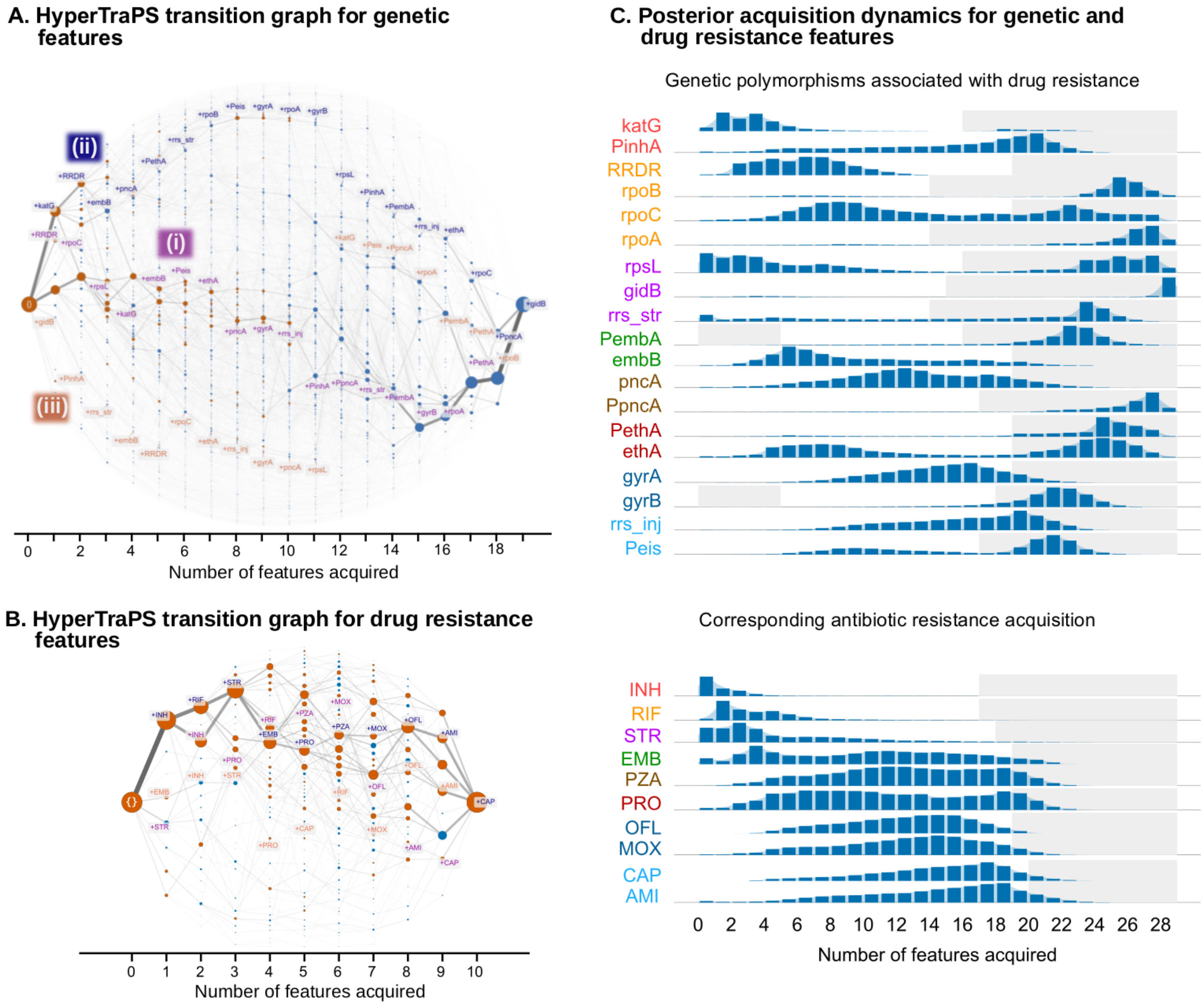
These biomedical examples serve to illustrate the power of the HyperTraPS to infer multiple competing pathways providing interpretable representations of such paths, and further the shortcomings of alternative approaches that restrict the output of learnt networks to be of the Bayesian network variety.

2.6. Application to the evolution of multi-drug resistant tuberculosis

We next asked whether our HyperTraPS approach could efficiently characterise dynamics in a system where observations are phylogenetically related. To this end, we consider the case of pathways of genetic polymorphisms that underpin drug-resistant tuberculosis isolates reported in Casali et al. (2014). In this study, the authors considered the sequences of 1000 drug-resistant tuberculosis isolates from Samara in Russia. The data consists of presence/absence markers of polymorphisms at 16 key genes/promoter regions that confer drug-resistance, as well as mutations in three RNA polymerase genes, and susceptibility or resistance to ten drugs for each of 395 isolates. These observed isolates are linked by a phylogeny, which Casali *et al.* constructed from genome-wide information (importantly, consisting of a much wider set of genomic regions than just those involved in drug resistance). As in Fig. 1, the source data then consists of the states on the leaves of a phylogeny and a phylogenetic structure that is previously, and essentially independently, constructed.

We assume that mutations are sufficiently rare such that convergent evolution is not a leading-order dynamic process between descendant and parent nodes in the phylogeny. With this assumption, we work backwards through the phylogeny parsimoniously to estimate unobserved parent states. From these estimates, we can reconstruct the transitions from parent nodes to descendant nodes on the phylogeny. These transitions then form the observations used by the HyperTraPS platform (Fig. 1). In Supplementary Figure S14 , we characterise the effects of this phylogeny on our posteriors, showing that its detailed structure has only limited quantitative influence on the general pathways we identify.

In Fig. 7A we show the inferred hypercubic transition graph for the dataset with $L = 19$ genetic sites alone, highlighting the genetic pathways by which polymorphisms may be acquired. Once more, a collection of previously unreported dynamic pathways are immediately observed, illustrated by the differential density of edges in different regions of the plot. In contrast to the large number of highly focussed paths inferred from the ovarian cancer data, this transition graph demonstrates a smaller number of looser – but still distinct in structure – paths


A. HyperTraPS transition graph for genetic features

B. HyperTraPS transition graph for drug resistance features

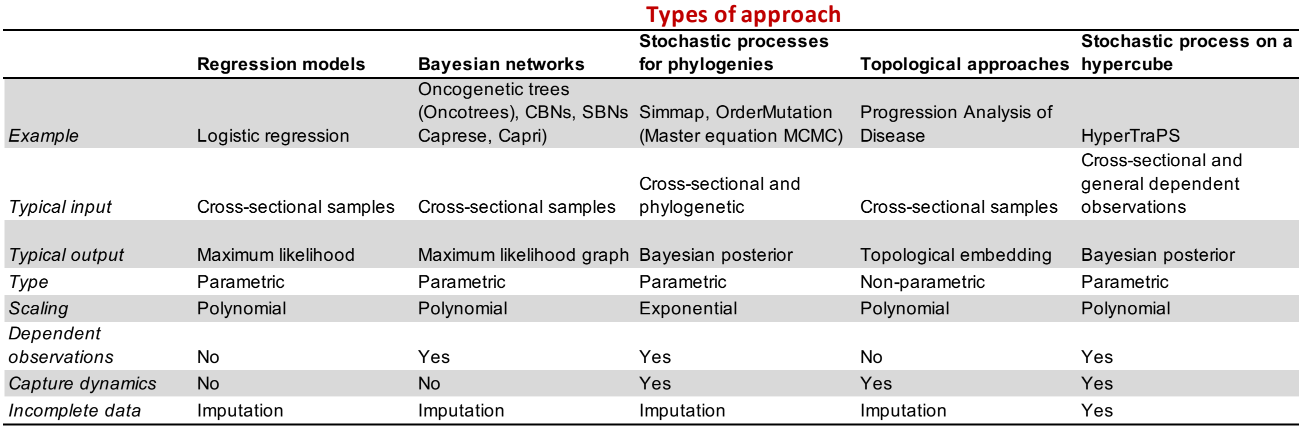
C. Posterior acquisition dynamics for genetic and drug resistance features

Fig. 7: HyperTraPS inference for multidrug resistance in tuberculosis identifies different pathway structures linking genetic and drug resistance features. (A) Dynamic pathways of the acquisition of genetic features leading to drug resistance, inferred from the dataset of $L = 19$ genetic sites across 395 phylogenetically related isolates. Multiple pathways through the genetic space associated with drug resistance are highlighted by regions of different density. Three distinct classes of pathway (i)-(iii) are highlighted and discussed in the text. **(B)** Dynamic pathways of the acquisition of resistance to specific drugs, inferred from the dataset of $L = 10$ antibiotics across 395 phylogenetically related isolates. **(C)** Posterior orderings of genetic and drug resistance features. Rows are colour-coded to link known genetic polymorphisms with the specific drug to which they confer resistance. The genetic sites occupy the first 19 rows, followed by the five 'first-line' drugs with the five 'second-line' drugs in the last five rows. Density in the grey regions corresponds to acquisitions that do not directly affect the likelihood, as features are not observed to be acquired in these regions in the dataset.

across the hypercube, each with a 'cloud' of variability indicating some flexibility in specific orderings within these pathways. We highlight this diversity with three specific pathways: (i) a central common pathway with a rifamycin resistance mutation *RRDR* is acquired first along with a fitness compensatory mutation *rpoC* second; (ii) an alternative path where no genetic correlates of streptomycin resistance (usually acquired early) are acquired until the sixth acquisition; (iii) a third pathway where the most common polymorphism *katG* is acquired late (the twelfth acquisition).

In Fig. 7B we show the inferred hypercubic transition graph for the dataset labelling resistance or susceptibility to each of the $L = 10$ antibiotics. The corresponding transition graph reports phenotypic pathways, existing in parallel with the genetic pathways in Fig. 7A. Notably, these phenotypic pathways are more canalised than the inferred genetic pathways. Resistance to 'first-line' drugs – those that are first used in treatment – dominate the initial dynamics, with comparatively little variation in ordering (isoniazid-rifamycin-streptomycin-ethambuol being a common pathway). There is more variation in dynamics of resistance acquisition to the remaining 'second-line' drugs, with acquisition subsequently progressing through several different pathways.

Fig. 7C shows the acquisition ordering plot for the combined genetic and phenotypic state of strains. Competition between different genetic pathways is reflected in the multimodality of several polymorphism acquisition distributions. Notably, *katG*, *rpoC*, and *rpsL* display ordering bimodality, evidencing several different pathways in which these features may be acquired early or late but not at intermediate orderings. This structural flexibility



Types of approach					
	Regression models	**Bayesian networks**	**Stochastic processes for phylogenies**	**Topological approaches**	**Stochastic process on a hypercube**
Example	Logistic regression	Oncogenetic trees (Oncotrees), CBNs, SBNs Caprese, Capri)	Simmap, OrderMutation (Master equation MCMC)	Progression Analysis of Disease	HyperTraPS
Typical input	Cross-sectional samples	Cross-sectional samples	Cross-sectional and phylogenetic	Cross-sectional samples	Cross-sectional and general dependent observations
Typical output	Maximum likelihood	Maximum likelihood graph	Bayesian posterior	Topological embedding	Bayesian posterior
Type	Parametric	Parametric	Parametric	Non-parametric	Parametric
Scaling	Polynomial	Polynomial	Exponential	Polynomial	Polynomial
Dependent observations	No	Yes	Yes	No	Yes
Capture dynamics	No	No	Yes	Yes	Yes
Incomplete data	Imputation	Imputation	Imputation	Imputation	Yes

Table 1: Comparison of HyperTraPS with other methods for inference from state space observations. We consider some of the key properties that HyperTraPS introduces. The following abbreviations are used: Suppes-Bayes Network (SBN), Conjunctive Bayes Network (CBN) and Markov chain Monte Carlo (MCMC).

gives rise to the separated pathways discussed above for Fig. 7A. The flexibility in genetic pathways corresponding to first-line drug resistance (for example, *katG*-*PinhA* and *PinhA*-*katG*, both leading to isoniazid (*INH*) resistance) provides a potential explanation for the early acquisition of resistance to these drugs.

Consistent with the more canalised phenotypic pathways in Fig. 7B, there is less multimodality in the ordering distributions of drug resistance features. Resistance to the first line drugs typically occurs before the second line drugs with a more precise order, likely indicative of the more widespread and increased time that tuberculosis has been treated with first line drugs. The ordering in which second line drug resistance is acquired is more broad, agreeing with the flexible phenotypic pathways seen above. Further, despite some heterogeneity, notable dynamic correlations may be observed between drugs and their known genetic correlates. The gene *katG* and drug isoniazid (*INH*), *rpsL* and streptomycin (*STR*), *embB* and ethambuol (*EMB*), illustrate clear examples of such links, providing a predictive and probabilistic connection between the dynamic acquisition of polymorphisms and the acquisition of specific drug resistance phenotypes.

Taken together, this dynamic pathway inference yields several new insights into the structure and variability of the evolutionary trajectories by which drug resistance is acquired. We discuss some specific evolutionary implications in STAR Methods, and compare with outputs of the approach of Bollback (2006) in STAR Methods (Supplementary Figure S15). Broadly, the joint polymorphism and drug resistance dynamics results suggest a consistent, convergent dynamic adaptation to first-line drugs, followed by more heterogeneity in the adaptation to second-line drugs. This convergence in first-line adaptation is likely facilitated, at least in part, by the flexible genetic pathways corresponding to these phenotypes (as found in other convergent evolution examples Williams et al. (2013)). These separate pathways (for example, those involving early vs late polymorphisms in *katG*, *rpoC*, or *rpsL*) are naturally distinguishable from the structures in Fig. 7A and multimodality in Fig. 7C. The HyperTraPS posteriors further provide a predictive framework which in future can be applied, for example, to predict the next likely drug resistance acquisitions given that a strain is in a particular state.

3. Discussion

We have introduced a powerful and highly generalisable statistical platform for inferring probabilistic, coupled dynamics from samples in a binary state space. The generality of this question is illustrated by the diversity of existing approaches that have some bearing on the corresponding inference problem. Table 1 illustrates several broad classes of these approaches, including regression models, Bayesian network models, stochastic processes on phylogenies, topological approaches and finite state space models (HyperTraPS).

Regression models are applied widely across the statistical and biomedical community, but are usually reliant on a linear underlying model and do not attempt to capture dynamics in which variables evolve. Additionally, they require a clear dichotomy between predictors and response variables to be imposed *a priori*, when such a distinction may not be appropriate, especially from the perspective of the inference of pathways. Bayesian networks provide a common platform for the relationships between features to be learned, with two examples being Conjunctive Bayes Networks (Beerenwinkel et al., 2007) and Suppes-Bayes networks (Loohuis et al., 2014). These are commonly used in oncogenetic inference problems, and have proved successful at unpicking causal relationships between features. We have shown that HyperTraPS aligns with the outputs of these approaches in simple cases. In more general settings, the stochastic model underlying HyperTraPS has the potential to reveal more detailed dynamic structure, including the identification of competing stochastic pathways, complex sets of interactions between coupled traits, and the quantification of uncertainty in the pathway structures that are revealed.

Dimensionality reduction approaches have been considered for finding representations of temporal dynamics from samples. Such methods are powerful and have been applied to vast data collected in whole genome single cell RNA experiments (Campbell and Yau, 2016) and also to disease (Nicolau et al., 2011). While highly flexible, these approaches often rely on specific assumptions about the quantitative details of the dimensionality reduction, leading to variability from method to method, and have yet to be considered in detail for finite space state models like the presence/absence structures we consider here.

Modelling trait evolution on phylogenies is the closest group of models to which HyperTraPS is related, and typically requires computation of master equation rate matrices that do not place restrictions on the transitions that may occur in the state space (Bollback, 2006; O'Meara, 2012). By embedding transitions on a hypercubic graph, HyperTraPS has the ability to handle orders of magnitude more features without noticeable loss of generality (simultaneous transitions are represented as equally weighted, temporally adjacent, transitions). Additionally, these methods are designed specifically for phylogenies, while HyperTraPS has applicability to generic sample dependency.

The HyperTraPS framework presented here has several advantages: (a) its polynomial scaling allows it to deal with large (many observations and many traits) datasets; (b) the regularisation processes we outline allow it not only to reveal and deal with arbitrary coupling between traits, but to select good and statistically significant parametric representations of these couplings to yield sparse models (thus applying Occam's razor); (c) it yields general and readily interpretable predictions; (d) it simultaneously provides inferred pathway structure, mechanistic insight, and uncertainty quantification; (e) the ability to include prior information about pathway structure when existing knowledge about biological mechanisms forbids, disfavours, or enhances the probability associated with particular transitions. Despite these advantages, there are of course some limitations to the platform's capabilities. Incomplete data currently provides a challenge for inference with HyperTraPS. There is nothing in principle preventing *hypercubic inference* with incomplete data: unbiased random walks can be simulated on a hypercube and their ability to recapitulate observations can be computed. Indeed, HyperTraPS can be applied in the case of uncertain *end points* of observed transitions (representing an advantage over existing methods). However, the sampling algorithm that allows HyperTraPS' efficient sampling of high-dimensional spaces currently does not translate to incompletely described *start points* of observed transitions, requiring future work is needed for further generalisations. Further, our approach for regularisation, while successfully implemented above, relies on an imperfect greedy algorithm and on the subjective use of the Akaike Information Criterion (AIC) for finding such sparser models. A multitude of methods are available for performing model selection within a full Bayesian setting (O'Hara and Sillanpää, 2009; Murphy, 2012) and exploration of alternative approaches for exploration of mappings from $W \to \pi$ and regularisation of HyperTraPS models is an important future avenue of research.

Our platform occupies the under-explored intersection between methods for inferring dynamics from uncoupled and/or longitudinal observations (as in cancer progression) and from phylogenetically linked observations (as in evolutionary inference). We have shown that HyperTraPS has a unique power to dissect multiple competitive dynamic pathways (yielding new insight in two biomedical case studies), and demonstrated how the processes of regularisation can be used to identify the best model structures for a given scientific setting. We underline that HyperTraPS requires no domain-specific knowledge, but can readily include such knowledge in the form of priors and in posterior interpretation. The platform is therefore ideal for contexts where mechanistic insight and modelling are less developed, and hence may also find valuable use in the wide range of progressive diseases that are less studied than cancer. We anticipate that this flexibility, and the abilities of HyperTraPS to naturally quantify uncertainty and form probabilistic predictions about future behaviours, will be of use across biomedical, evolutionary, and other scientific disciplines as volumes of available data continue to increase.

4. Lead Contact and Materials Availability

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Iain Johnston (iain.johnston@uib.no). This study did not generate new unique reagents.

5. Data and Software Availability

All computational work was performed with custom-written software in C++ and Python. The code for the HyperTraPS package is freely available at https://github.com/sgreenbury/HyperTraPS (DOI 10.5281/zenodo.3478290) and usable under the creative commons licence.

6. Acknowledgements

All authors acknowledge support from EPSRC grant EP/N014529/1. IGJ additionally acknowledges support from a Turing Fellowship from the Alan Turing Institute.

7. Author Contributions

Study concept and design: SFG, MB, IGJ; Development of source code: SFG, IGJ; Analysis and interpretation of data: SFG, MB, IGJ; Writing and revision of the manuscript: SFG, MB, IGJ; Study supervision: MB, IGJ.

8. Declaration of interests

The authors declare that there is no conflict of interest.

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli*, 13(4):893–909.

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25.

Beerenwinkel, N. and Sullivant, S. (2009). Markov models for accumulating mutations. *Biometrika*, 96(3):645–661.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., Dhir, R., Disaia, P., Gabra, H., Glenn, P., Godwin, A. K., Gross, J., Hartmann, L., Huang, M., Huntsman, D. G., Iacocca, M., Imielinski, M., Kalloger, S., Karlan, B. Y., Levine, D. A., Mills, G. B., Morrison, C., Mutch, D., Olvera, N., Orsulic, S., Park, K., Petrelli, N., Rabeno, B., Rader, J. S., Sikic, B. I., Smith-Mccune, K., Sood, A. K., Bowtell, D., Penny, R., Testa, J. R., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Gunaratne, P., Hawes, A. C., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Santibanez, J., Reid, J. G., Trevino, L. R., Wu, Y. Q., Wang, M., Muzny, D. M., Wheeler, D. A., Gibbs, R. A., Getz, G., Lawrence, M. S., Cibulskis, K., Sivachenko, A. Y., Sougnez, C., Voet, D., Wilkinson, J., Bloom, T., Ardlie, K., Fennell, T., Baldwin, J., Gabriel, S., Lander, E. S., Ding, L., Fulton, R. S., Koboldt, D. C., McLellan, M. D., Wylie, T., Walker, J., O'Laughlin, M., Dooling, D. J., Fulton, L., Abbott, R., Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M., Schierding, W., Shen, D., Harris, C. C., Schmidt, H., Kalicki, J., Delehaunty, K. D., Fronick, C. C., Demeter, R., Cook, L., Wallis, J. W., Lin, L., Magrini, V. J., Hodges, J. S., Eldred, J. M., Smith, S. M., Pohl, C. S., Vandin, F., Raphael, B. J., Weinstock, G. M., Mardis, E. R., Wilson, R. K., Meyerson, M., Winckler, W., Verhaak, R. G., Carter, S. L., Mermel, C. H., Saksena, G., Nguyen, H., Onofrio, R. C., Hubbard, D., Gupta, S., Crenshaw, A., Ramos, A. H., Chin, L., Protopopov, A., Zhang, J., Kim, T. M., Perna, I., Xiao, Y., Zhang, H., Ren, G., Sathiamoorthy, N., Park, R. W., Lee, E., Park, P. J., Kucherlapati, R., Absher, D. M., Waite, L., Sherlock, G., Brooks, J. D., Li, J. Z., Xu, J., Myers, R. M., Laird, P. W., Cope, L., Herman, J. G., Shen, H., Weisenberger, D. J., Noushmehr, H., Pan, F., Triche, T., Berman, B. P., Van Den Berg, D. J., Buckley, J., Baylin, S. B., Spellman, P. T., Purdom, E., Neuvial, P., Bengtsson, H., Jakkula, L. R., Durinck, S., Han, J., Dorton, S., Marr, H., Choi, Y. G., Wang, V., Wang, N. J., Ngai, J., Conboy, J. G., Parvin, B., Feiler, H. S., Speed, T. P., Gray, J. W., Socci, N. D., Liang, Y., Taylor, B. S., Schultz, N., Borsu, L., Lash, A. E., Brennan, C., Viale, A., Sander, C., Ladanyi, M., Hoadley, K. A., Meng, S., Du, Y., Shi, Y., Li, L., Turman, Y. J., Zang, D., Helms, E. B., Balu, S., Zhou, X., Wu, J., Topal, M. D., Hayes, D. N., Perou, C. M., Wu, C. J., Shukla, S., Sivachenko, A., Jing, R., Liu, Y., Noble, M., Carter, H., Kim, D., Karchin, R., Korkola, J. E., Heiser, L. M., Cho, R. J., Hu, Z., Cerami, E., Olshen, A., Reva, B., Antipin, Y., Shen, R., Mankoo, P., Sheridan, R., Ciriello, G., Chang, W. K., Bernanke, J. A., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Sanborn, J. Z., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Wilkerson, M. D., Zhang, N., Akbani, R., Baggerly, K. A., Yung, W. K., Weinstein, J. N., Shelton, T., Grimm, D., Hatfield, M., Morris, S., Yena, P., Rhodes, P., Sherman, M., Paulauskis, J., Millis, S., Kahn, A., Greene, J. M., Sfeir, R., Jensen, M. A., Chen, J., Whitmore, J., Alonso, S., Jordan, J., Chu, A., Barker, A., Compton, C., Eley, G., Ferguson, M., Fielding, P., Gerhard, D. S., Myles, R., Schaefer, C., Mills Shaw, K. R., Vaught, J., Vockley, J. B., Good, P. J., Guyer, M. S., Ozenberger, B., Peterson, J., and Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Bollback, J. P. (2006). SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC bioinformatics*, 7:88.

Campbell, K. R. and Yau, C. (2016). Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLoS Computational Biology*, 12(11):1–20.

Caravagna, G., Giarratano, Y., Ramazzotti, D., Tomlinson, I., Graham, T. A., Sanguinetti, G., and Sottoriva, A. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, 15(9):707–714.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R. D., Brown, T., and Drobniewski, F. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics*, 46(3):279–286.

Colijn, C., Jones, N., Johnston, I. G., Yaliraki, S., and Barahona, M. (2017). Toward precision healthcare: Context and mathematical challenges. *Frontiers in Physiology*, 8(MAR):1–10.

De Sano, L., Caravagna, G., Ramazzotti, D., Graudenzi, A., Mauri, G., Mishra, B., and Antoniotti, M. (2016). TRONCO: An R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, 32(12):1911–1913.

Desper, R., Jiang, F., Kallioniemi, O. P., Moch, H., Papadimitriou, C. H., and Schäffer, a. a. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology : a journal of computational molecular cell biology*, 6(1):37–51.

Diaz-Uriarte, R. (2018). Cancer progression models and fitness landscapes: A many-to-many relationship. *Bioinformatics*, 34(5):836–844.

Gerstung, M., Baudis, M., Moch, H., and Beerenwinkel, N. (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25(21):2809–2815.

Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B., and Beerenwinkel, N. (2011). The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS ONE*, 6(10).

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

Hjelm, M., Höglund, M., and Lagergren, J. (2006). New Probabilistic Network Models and Algorithms for Oncogenesis. *Journal of Computational Biology*, 13(4):853–865.

Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biology*, 17(1):86.

Johnston, I. G. and Williams, B. P. (2016). Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Systems*, 2(2):101–111.

Knutsen, T., Gobu, V., Knaus, R., Padilla-Nash, H., Augustus, M., Strausberg, R. L., Kirsch, I. R., Sirotkin, K., and Ried, T. (2005). The interactive online SKY/M-FISH & CGH database and the Entrez Cancer Chromosomes search database: Linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes and Cancer*, 44(1):52–64.

Loohuis, L. O., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M., and Mishra, B. (2014). Inferring tree causal models of cancer progression with probability raising. *PLoS ONE*, 9(10).

Mahler, D. L., Revell, L. J., Glor, R. E., and Losos, J. B. (2010). Ecological opportunity and the rate of morphological evolution in the diversification of greater Antillean anoles. *Evolution*, 64(9):2731–2745.

Montazeri, H., Kuipers, J., Kouyos, R., Böni, J., Yerly, S., Klimkait, T., Aubert, V., Günthard, H. F., and Beerenwinkel, N. (2016). Large-scale inference of conjunctive Bayesian networks. *Bioinformatics*, 32(17):i727–i735.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, MA.

Murray, I. and Graham, M. M. (2015). Pseudo-Marginal Slice Sampling. page 9.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–118.

O'Meara, B. C. (2012). Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):267–285.

Pagel, M. and Meade, A. (2006). Bayesian Analysis of Correlated Evolution of Discrete Characters by ReversibleJump Markov Chain Monte Carlo. *The American Naturalist*, 167(6):808–825.

Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., and Mishra, B. (2015). CAPRI: Efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026.

Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M., and Caravagna, G. (2017). Learning mutational graphs of individual tumor evolution from multi-sample sequencing data.

Ross, E. M. and Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1):69.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229.

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk metropolis algorithms. *Annals of Statistics*, 43(1):238–275.

Szabo, A. and Boucher, K. (2002). Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical Biosciences*, 176(2):219–236.

Watts, J., Greenhill, S. J., Atkinson, Q. D., Currie, T. E., Bulbulia, J., and Gray, R. D. (2015). Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proceedings of the Royal Society B: Biological Sciences*, 282(1804):20142556–20142556.

Williams, B. P., Johnston, I. G., Covshoff, S., and Hibberd, J. M. (2013). Phenotypic landscape inference reveals multiple evolutionary paths to C4photosynthesis. *eLife*, 2:1–19.

Youn, A. and Simon, R. (2012). Estimating the order of mutations during tumorigenesis from tumor genome sequencing data. *Bioinformatics*, 28(12):1555–1561.

Zafar, H., Navin, N., Nakhleh, L., and Chen, K. (2018). Computational approaches for inferring tumor evolution from single-cell genomic data. *Current Opinion in Systems Biology*, 7:16–25.

Zafar, H., Tzen, A., Navin, N., Chen, K., and Nakhleh, L. (2017). SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, 18(1):178.

STAR Methods

HyperTraPS pipeline

In Supplementary Figure S1 , we provide a diagrammatic overview of the HyperTraPS pipeline. The different elements are described below. As described in Fig. 1, the first step is to convert cross-sectional, longitudinal, or phylogenetically linked observations to a set of transitions, which we will represent as $D = \{s_i, t_i\}$, where s_i is the ith source state and t_i the ith target state, and there are n_D observations in total.

Bayesian framework and likelihood of transition dataset

As introduced in Johnston and Williams (2016), we choose a Bayesian framework for inferring parameters for the set of edge weights W on the hypercubic transition graph that explain the data D.

As such we are concerned with drawing samples from the posterior:

$$P(W|D) = \frac{P(D|W)}{\int P(D|W)P(W)dW} P(W)$$

which is proportional to the product of our prior probability density $P(W)$ on edge parameterisations and the likelihood $\mathcal{L}(W|D) = P(D|W)$, such that we have $P(W|D) \propto \mathcal{L}(W|D)P(W)$. Throughout this work we choose a uniform prior distribution on $P(W)$ and therefore only need to consider the calculation of $\mathcal{L}(W|D)$ in order to derive samples from the posterior probability distribution.

From this transition set, we can decompose the likelihood into the following form (regardless of whether the source data was cross-sectional, longitudinal, or phylogenetically coupled Johnston and Williams (2016)):

$$\mathcal{L}(W|D) = \prod_{i=1}^{n_D} P_{\text{observe}}(s_i \to t_i)$$

where n_D is the size of the transition dataset. P_{observe}, the probability of observing such a transition requires a signal to be emitted by our system at both the source and target states, with the system having reached the source state and then made the transition to the target state via any possible walk on the hypercube. Therefore, the probability of observing such a transition can be written as:

$$P_{\text{observe}}(s_i \to t_i) = P_{\text{emit}}(s_i, t_i) P_{\text{reach}}(s_i|W) P(t_i|s_i, W)$$

We assume that signal emission in a given state is a random process that independent of the state. Given the term $P_{\text{emit}}(s_i, t_i)$ is also independent of W, and that we deal only with complete data here, P_{emit} yields a constant multiplicative factor which can be ignored in the inference process. In Johnston and Williams (2016), it is shown that the remaining log-likelihood can be written as:

$$\log \mathcal{L}(W|D) = \sum_{i=1}^{n_D} \log P(t_i|s_i, W) := l(W|D) \quad (1)$$

where the only computation required is the probability of making the transition to t_i from s_i for a given parameterisation of W.

In order to calculate $P(t_i|s_i, W)$, a sum over all possible paths between s_i and t_i is required. Given that the number of paths between s_i and t_i scales as the factorial of the Hamming distance, the problem of deriving the rate matrix becomes intractable for systems of dimensions around $L \gtrsim 10$. Instead we tackle the problem by way of performing biased random walks restricted to pathways that end in t_i. This method of sampling was introduced in Johnston and Williams (2016) and allows systems with more features to be considered than previously has been the case. This HyperTraPS algorithm that forms the key part of the HyperTraPS framework is captured in Algorithm 1.

Tractable parameterisations of hypercube

The transition graph linking states with L features has $L2^{L-1}$ edges that we aim to parameterise. As L grows, we require a way of reducing this number of parameters k without compromising our ability to describe the dynamics of a system. Shrinkage and model selection tools may be used to achieve this reduction: we explore a simple approach for this process later. However, given the potentially large number of parameters in the default model, we also consider methods to reduce parameter space before the inference process.

One intuitive approach is based around considering the factors that may influence a given transition. The full parameterisation allows independent rates between any two states. In this picture, the probability $P(i)$ of acquiring the ith trait can take arbitrary and independent values for every possible combination of the other

Algorithm 1: HyperTraPS algorithm for complete data: Hypercubic Transition Path Sampling was first introduced by Johnston and Williams (2016) to sample random walks on a hypercube across a restricted set of compatible states between a source and target state.

```
Data: $D^{\text{transitions}} = \{s_i \to t_i\}_{i=1}^{n_D}$
Result: Estimate of $P(D^{\text{transitions}}|W)$
begin
    for $(s \to t) \in D^{\textit{transitions}}$ do
        $s_c \leftarrow s$
        Initialise $N_h$ trajectories starting at state $s$
        for $i \in N_h$ do
            $s_c \leftarrow s$
            $\alpha_i \leftarrow 1$
            while $t$-compatible move possible for $s_c$ do
                Calculate the probability of making a $t$-compatible move, record as $\alpha_i'$
                $\alpha_i \leftarrow \alpha_i \alpha_i'$
                Choose a $t$-compatible move at random in proportion to its transition probability
                Make move and update $s_c$ accordingly
        $\hat{P}(s \to t) = N_h^{-1} \sum_i \alpha_i$
        $P(D^{\text{transitions}}|W) \leftarrow P(D^{\text{transitions}}|W) + \hat{P}(s \to t|W)$
```

$L-1$ traits. As an alternative, we can restrict the dependence of $P(i)$ on the *coupling* of other trait patterns. For example, if we assume that each of the $L-1$ other traits influence $P(i)$ independently (no synergistic interactions), we need only L^2 parameters: a 'basal rate' of acquisition for each trait i, and the amount by which this basal rate is modified by the presence of trait $j \neq i$. This reduction is analogous, for example, to Generalised Linear Models where response variables can be considered a function of independent variables and interaction terms between the independent variables, neglecting higher order interaction terms.

From this perspective a hierarchy of models may be constructed (Supplementary Figure S2). For the 'zero order' model every feature has equal probability of acquisition ($k = 0$ parameters). In the 'first order' model, every feature has an independent acquisition probability ($k = L$ parameters). In the 'second order' model, every feature's basal acquisition probability is independently modulated by the presence of each other feature ($k = (1 + (L-1)) \times L = L^2$ parameters). Higher order models, including the full $L2^{L-1}$ set can be envisaged, introducing more complex interactions between the co-occurrence of features.

To illustrate these parameterisations, consider the weight $w_{s \to t}$ of the edge from state s to state t. These edge weights are nonzero only for pairs s, t where t differs from s by the acquisition of exactly one feature, with a hypercubic network remaining. Then, for the zero-order model, every edge in the hypercube is equally weighted, and we can set this weighting to unity, $w_{s \to t} = 1$. For the first-order model, the weight of an edge is completely specified by the feature that the edge corresponds to acquiring, $w_{s \to t} = p_i$, where i is the feature that distinguishes s from t. The first-order parameterisation is thus described by the vector $\mathbf{p}$ with L elements, one for each feature. For the second-order model:

$$w_{s \to t} = p_{ii} \prod_{j \neq i} q(s, j, i), \tag{2}$$

where i is the feature that distinguishes s from t, s_j is the presence/absence of feature j in state s, and $q(s, j, i) = 1$ if $s_j = 0$ and p_{ji} otherwise. The second-order parameterisation is thus described by the matrix p with L^2 elements, where diagonal element p_{ii} gives the 'basal' rate associated with feature i, and off-diagonal elements p_{ji} give the influence that the presence of feature j in source state s has on this basal rate.

For numerical convenience, we implement Eqn. 2 via a logarithmic transformation, such that $p_{ij} = \ln \pi_{ij}$, and work with π_{ij} as the parameterisation of the model. We will use π generally to refer to edge weight parameters in the inference process.

Monte Carlo sampling methods

The complexity of the inference problem challenges analytic or uniform sampling approaches to compute Eq. (1). Instead, we employ Markov Chain Monte Carlo (MCMC) in order to generate samples from the posterior on edge weights W. As the HyperTraPS algorithm generates an estimate of the likelihood (with the same expected value as the exact likelihood), this is in fact a pseudo-marginal MCMC sampler which has been shown to yield the same stationarity properties as if it were exact (Andrieu and Roberts, 2009).

Previous approaches for specific scientific questions (Williams et al., 2013; Johnston and Williams, 2016) found this pseudo-marginal MCMC sampler to demonstrate good mixing. However, there are cases where this simple approach produces poor mixing, specifically when the Hamming distance between a source and target state becomes large. This is because Algorithm 1 generates an estimate of the likelihood with increased variance around its exact value due to the greater number of acquisitions made during path sampling. This can lead to poor mixing of sampler chains, if the sampler draws a high value for the likelihood estimate which subsequent random draws have a high probability of having a lower likelihood for the same parameterisation. This occurs when the variance of the total log-likelihood has a variance with magnitude greater than unity (Sherlock et al., 2015).

To address this issue and generalise to more diverse datasets, we embedded HyperTraPS within an auxiliary pseudo-marginal MCMC algorithm (APM MCMC), which also satisfies the same convergence properties as MCMC (Murray and Graham, 2015). By making the likelihood a joint density $l(\pi, u)$ over the parameters of the model and also the random variable from which our estimate is drawn, alternate Metropolis-Hastings steps can be performed by keeping π and u alternately fixed during the proposed update to the chain. For HyperTraPS, a new proposal for the random variable u is a new set of random trajectories across the hypercube over which each observations's likelihood is estimated. We make use of this scheme throughout this work, as little computational overhead is introduced, and mixing times are dramatically improved.

As discussed, we have assumed a uniform prior for the parameterisations of the hypercube. For our choice of mapping π, this means we choose $P(\pi) \in U(-m, m)$ where $m = 10$ and $m = 20$ are used in this work to cover several orders of magnitude of relative size across the inferred parameters.

We begin MC sampling runs with the parsimonious initial condition $\pi = \mathbf{0}$. This is equivalent to the zero order model where there is no directionality pre-supposed. This facilitates the avoidance of local traps in the parameter landscape while remaining agnostic in introducing directionality into the inferred parameterisations for a particular dataset. A burn-in period occurs before expected convergence of an MCMC chain. For each of datasets in the main text, over 10^6 iterations are performed along the chain, ensuring samples are used only when convergence is apparent. We consider convergence to be reached when the chain shows stability in average likelihood for a sustained period with the ratio of accepted parameterisations that yield increased or decreased likelihoods to be in equal proportion.

Simulated walks to illustrate order of acquisition

The inference process above yields inferred posterior distributions on the hypercubic edge weights W. We can query these posteriors in a number of ways to gain descriptive and predictive information about the mechanisms generating observed states. First, we produce a parsimonious and intuitive representation of the dynamic pathways supported by the inferred posteriors. Here, we simulate an ensemble of random walkers generating complete trajectories on hypercubes with sets of transition probabilities sampled from the inferred posterior. This ensemble reflects the likely dynamic pathways supported by the dynamic transition model after parameterisation. We simulate an ensemble of random walks in two ways: *Walk Simulation 1 (WS1)*, with walkers that run from $\{0\}^L$ to $\{1\}^L$ where a feature is acquired at every time step and *Walk Simulation 2 (WS2)* only simulates trajectories corresponding to transitions observed in the dataset. In each case, we record every transition between states allowing the construction of a weighted directed graph of all states and transitions encountered. From this graph, the frequency f_{ij} with which feature i is gained at step j.

Graph embedding and visualisation for dynamic acquisition on the hypercube

With each simulated random walk, L transitions occur between states on the hypercube. Across a large sample of random walks, we define this set of states as $\mathcal{S} = \{s_i\}$ and we can represent the number of transitions between any two states by a directed, weighted graph with adjacency matrix a_{ij}.

In order to visualise this graph to reveal characteristic progressions across the hypercube resulting from a given parameterisation, we use a custom embedding to project the high-dimensional graph into two dimensions. First, we project the hypercube on to the surface of a sphere and optimise the projection by making the following choices:

- Every state is given the same radial coordinate, $r = 1$.

- The number of features acquired in the dataset is a measure of the how far the state is along the progression from 0^L to 1^L. Therefore, for every state $\mathcal{S}$, we count the number of acquired features (n out of L) and assign a polar angle θ such that $\sin\theta = n/L$.

- The azimuthal angle ϕ on the interval $0 \leq \phi \leq \pi$ is assigned by considering the *mean angle of the states from all incoming edges*, therefore attempting to maximise the potential spread of the most common distinct paths across the hemisphere. A final assumption involves choosing all states with a single acquisition (L states) to be uniformly spread on the cosine of the interval $[0, \pi]$.

With the embedding, the plot of the adjacency matrix a_{ij} is augmented by choosing node sizes and edge widths in proportion to the number of times the state and the transition are respectively encountered by the ensemble of random walks. Three examples of plots generated from this embedding with parameterisations of the hypercube are shown in Fig. 2B(i)-(iii), illustrating the ability to display different underlying progressions inferred with HyperTraPS.

In presenting the embedding, we adjust the graphical depiction to highlight the features of the graph in the following way:

- Vertex area is in proportion to the number of times the vertex is visited by WS1 simulated random walks.
- Edge widths and opacity are in proportion to the number of times the transition between states is made with a random walker under WS1.
- States encountered are coloured blue if $s \notin D^{\text{transitions}}$ and orange if $s \in D^{\text{transitions}}$
- To highlight representative paths across the hypercube, we employ a labelling scheme as follows. As walkers start from the empty "$\{\}$" state ($\{0\}^L$), we can consider the addition of single features as each edge is traversed. In the plots, we use a greedy mechanism for determining which edges to label. Starting from 0^L, we take the most probable outgoing edge at each vertex encountered and label the feature acquired across that edge at the resulting vertex until the 1^L state is reached, giving is the first greedy path. The following n greedy paths make use of the same approach but disregard any previously labelled edges, taking the next most probable available. We use the approach to clearly identify the left-right and right-left paths in Fig. 2B(i)-(iii).
- Finally, an optional transform to remove vertex overlap may be applied to remove overlap of vertices with a given number of features, while retaining the relative area of each vertex that is determined by the number of times the vertex is encountered.

Probabilistic feature graph representation

Using either WS1 or WS2, the set of states encountered may be considered as a directed weighted acyclic graph through *sample space*/*state space*, due to the irreversible acquisition of features. As paths through state space involve the acquisition of a feature with each incoming and outgoing edge, a different representative graph may also be constructed relating the observed consecutive feature acquisitions producing a graph in *feature space*.

To this end, we consider the ensemble of observed $P(Y_{out}, X_{in}; s)$ derived from a set of simulated walks across sample space, which gives the probability that feature Y is acquired leaving state s, with feature X having been acquired to reach state s. An average joint relationship can then be written as the following:

$$P(Y_{out}, X_{in}) = \sum_{s} P(Y_{out}, X_{in}; s) P(s)$$

where $P(s)$ is the proportion of times state s is encountered. $P(Y_{out}, X_{in})$ gives the edge weight between X and Y for the probabilistic feature graphs in this article.

Regularisation

We previously discussed approaches to reduce the parameter space of the HyperTraPS model while retaining dynamic information. We can *a priori* also employ model reduction approaches to identify supported parameter structures given a particular dataset. This regularisation helps identify more interpretable, parsimonious models and to guard against over-fitting.

One approach to model selection would be a fully Bayesian exploration of the joint space of model structures and parameters. However, the combinatorial explosion of search space with L currently makes this approach unfeasible for all but the simplest systems. Instead, we sacrifice a full exploration of this complicated space in favour of a tractable but principled approach to balance the reduction of model complexity against the ability to fit the data. This illustrative metric can indicate the amount of redundancy present in the parameterised π that can be removed in order to reduce the potential for over-fitting. To this end, we introduce a cost function to penalise the log-likelihood and then perform a algorithmic search to optimise this function.

We note that the number of parameters k required to adequately describe a given dynamic system is deeply related to the mechanisms underlying that system. If features are acquired independently, the first order model with L parameters should be sufficient to capture the dynamics (as seen in Section 2.1 of the main text for dataset D_1), and the features may be completely ordered for the average trajectory. If a higher order model with more parameters is required, it suggests that interactions exist between features, such that one feature may influence the acquisition propensity of another. Identifying the sparsest model that can account for observations therefore also reveals mechanistic insight into the system.

For simplicity, we use the Akaike Information Criterion (AIC) (Murphy, 2012) to introduce sparsity. The AIC score for a model can be written as:

$$\text{AIC} = 2(k - \hat{l})$$

where k are the number of parameters in the model, and $\hat{l}$ is the maximum log-likelihood. The score comprises the log likelihood and a penalty for lack of sparsity, in this case, the number of non-zero elements included in the maximum likelihood parameterisation π. Other options for regularization scoring include the Bayesian Information Criterion (BIC), but we refrain from exploring different metrics here, focussing firstly on illustrating how such regularisation can be performed within the HyperTraPS framework. A more general model selection approach will be the subject of future work.

To find parameterisations that optimise the AIC, we take a *greedy backward selection* approach (Murphy, 2012) to reduce the number of parameters k for a given model type. The process can be applied to both the first- and second- order models. An issue with such a greedy approach is that each single greedy backward step is unable to account for interactions between multiple parameters that lead to lower scores. Therefore, given a set of potentially distinct approximately maximum likelihood parameterisations, different backward selection processes from different starting maximum likelihood models may yield different minimum AIC scores for a given value of k. In an attempt, to bypass this problem, we take an ensemble of the top 100 maximum likelihood parameterisations from an MCMC sampling procedure (top 1000 for the ovarian cancer datsets) and perform the greedy backward selection process to each one. Across the ensemble, for a given parameter number k, we take the minimum AIC score as a proxy for the minimum model at this level of parameterisation. The global minimum with respect to AIC is taken as the *first order regularised* or *second order regularised* model for the a first order and second order starting point respectively. The regularised models are then taken used in the subsequent section to perform model validation.

In Fig. 2D(i)-(ii), we show the regularisation process described above for the minimum of the ensemble at each value of k for the two synthetic datasets D_1 and D_2 and, later in STAR Methods, the process for a third synthetic dataset and the ovarian and tuberculosis datasets respectively.

Validation

Importantly, the inferred parameterisations from our approach can be used to predict future behaviour for a given state. We have described two procedures for generating parameterisations: sampling from the full posterior for a given model (first- or second- order) or regularised parameterisations constructed by the procedure in the previous section. In this section, we perform model validation through using the regularised parameterisations in order to identify the strength of evidence for the first- or second- order models. Using the outcome of this procedure, either samples from the full posteriors of the identified model or from the corresponding regularised parameterisation can be used for prediction.

We validate this predictive power through two methods: firstly, through basic model comparison between the regularised first- and second- order models; and subsequently, by calculating the likelihood of observing data not used in the inference part of the method as a proxy for the predictive capability of each model. As a simple procedure to illustrate this, we split the $D^{\text{transitions}}$ dataset into two halves: a training dataset D_{train} on which samples from the posterior are drawn and model comparisons can be made, and a testing dataset D_{test} with which the likelihood can be calculated using samples from the posterior for D_{train}.

For model comparisons, we choose the zero order model as a null model. For comparisons between the different order models, we find the regularised first- and second- order model for the training dataset and denote this likelihood as $\hat{l}(\pi|D_{\text{train}})$. We then perform a likelihood ratio test, using the log-likelihood ratio statistic (LLR):

$$LLR = 2\hat{l}(\pi_r^{(j)}|D_{\text{train}}) - 2\hat{l}(\pi_r^{(0)}|D_{\text{train}})$$

where $\pi_r^{(j)}$ is regularised j^{th} order model. We compare to the χ^2 distribution for the number of non-zero parameters in π. With regard to the test dataset, we then use HyperTraPS to estimate $\log P(D_{\text{test}}|\pi_r^{(j)})$ providing a measure of predictive capability of the j^{th} order regularised model. This is an intuitive option for measuring performance as it is not guaranteed that a given transition from $s \to t$ should end at t – there may be multiple pathways. Therefore, the overall largest likelihood $(\log P(D_{\text{test}}|\pi_r^{(j)}))$ across competing j models for the test dataset will be monotonically related with better parameterisations.

Testing and validating HyperTraPS with differing data structures, volumes, and priors

In this section we investigate HyperTraPS' capacity to learn pathway structures by varying several features of the synthetic datasets used in the main text. Fig. 3 in the main text provides central aspects of this investigation;

Supplementary Figure S3 involves different relatednesses of observations; Supplementary Figure S4 provides posteriors for the investigation of different numbers of competing pathways; Supplementary Figure S5 provides probabilistic feature graphs for the use of prior information.

Both quantitative and structural prior information about models can be included in HyperTraPS. Quantitative information (for example, the acquisition of one feature scaling the acquisition probability of another) can readily be included through applying an appropriate prior distribution on the corresponding element of the transition matrix. Simple structural information, such as forbidding one transition before another, can also readily be captured by setting priors on the corresponding parameters.

Prior information can also be incorporated where an underlying tree structure of precedence between features is known. We denote this the *prior tree*. In order to incorporate such information, we wish to avoid parameterisations of π that would violate the ordering described within a prior tree. In order to prohibit transitions in the second-order L^2 parameterisation system, for a given edge in the prior tree $a \to b$, we enforce a prior with low basal probabilities for the acquisition of b (in proportion to the depth of b from the root in the prior tree). That is, spontaneous acquisition of states below the root in the prior tree is enforced to be highly unlikely. We then enforce a prior with off-diagonal elements so that the acquisition of a compensates this low basal probability on the acquisition of b. Hence, other features aside from a remain unable to affect the acquisition of b, but once the precedent feature a is acquired, then b may be acquired.

To demonstrate this approach, consider a prior tree with edges: $R \to 1 \to 2; R \to 3$, for $L = 3$, where R corresponds to the root of the prior tree. Starting from a uniform prior $U(-m, m)$ on all elements of π, we enforce three prior requirements. First, $\pi_{22} < \pi_{ii} - \Delta$ for all $i \neq 2$ (enforcing low basal acquisition for feature 2). Second, $\pi_{12} \geq \Delta$ (allowing the acquisition of feature 1 to 'rescue' this low basal rate). Third, $\pi_{i2} = 0$ for all $i \neq 1$ (allowing no other acquisitions to 'rescue' the low basal rate). In this way, we ensure an acquisition probability of 2 prior to 1 or 3 is suppressed by a factor of e^{Δ}. In practise we have used $\Delta = dm/l$, where d is the depth of a feature in the prior tree, m as above is the range of the original uniform prior, and l is the maximum depth of the prior tree.

Additional synthetic cross-sectional dataset

In this section, we illustrate the inference, regularisation and predictions with a third cross-sectional dataset D_3. This dataset can be considered a composite of previous synthetic sets D_1 and D_2, such that new set D_3 is the linear combination $D_3 = 2D_1 + D_2$. In this case, we have a dominant progression underlying the dataset but with a substantial minority contribution from an alternative pathway.

In Supplementary Figure S6 A, the structure of this additional cross-sectional dataset is depicted. Supplementary Figure S6 B, C and D indicate that HyperTraPS can infer the two distinct progressions and the proportion with which these progressions occur within the data. For example, in the density plots, feature $i = 0$ is acquired three times as frequently as feature $i = 7$ in for step $j = 0$.

In Supplementary Figure S7 A and Supplementary Figure S7 B, we show the results of regularisation and the outputs of the validation methodology: the first order model can be observed to be a better predictor than the null model (it captures the dominant progression) as seen with larger and significant log-likelihoods for the full and training datasets. The second order and regularised second order models perform much better still by having the ability to capture both the dominant and secondary progressions present in the dataset, as illustrated in the validation methodology by the much larger associated likelihoods.

Alternative interpretation of inferred acquisition orderings ('Walk Simulation 2')

In STAR Methods above we introduced a protocol for using samples from the posterior of $\mathcal{L}(\pi|D)$ to illustrate the order in which features are acquired. We denoted this process Walk Simulation 1 (WS1) as simulations from $\{0\}^L$ to $\{1\}^L$ are performed with the feature i acquired at step j being recorded as a proportion f_{ij}. As a feature is a always gained in each step, and all features are gained at some stage during this simulation process, the two properties $\sum_k f_{kj} = 1$ and $\sum_k f_{ik} = 1$ both hold. We illustrated the result of this simulation using a histogram for the matrix f_{ij} with kernel density estimates overlaid for each feature.

An alternative simulation protocol is to only simulate trajectories corresponding to transitions that are observed in the dataset. In other words, rather than assuming random walkers proceed from 0^L to 1^L, we simulate a set of walkers between each pair of source and target states s_i, t_i in the dataset, relaxing the requirement that walkers start at 0^L and end at 1^L. We denote this process Walk Simulation 2 (WS2). For WS2, we can consider f_{ij} as the probability:

$$f_{ij} \approx P(\text{feature } i \text{ is gained at step } j | s = \{0\}^L \to t = \{1\}^L)$$

where s is the source state and t is the target state of the set of random walks. Summation over the rows or columns of f_{ij} no longer hold as there is no guarantee in the data that a feature is acquired at a given step j or that every feature i is acquired in each random walk.

The main distinction between WS1 and WS2 is the following: WS1 infers trajectories, informed by data, that start at $\{0\}^L$ and acquire all features to reach $\{1\}^L$. WS2 restricts the inference to the region 'covered' by the set of transitions observed in the dataset. Therefore, WS1 provides a readout of a complete process of acquisition (so may be more appropriate for analysis in systems where this is the expected outcome), while WS2 gives a readout of trajectories without extrapolating beyond the limits of observed states (and may be more appropriate if the walks are not believed to go to completion).

We plot the densities for WS1 and WS2 in Supplementary Figure S8 for two datasets from the main text, synthetic set (ii) and the tuberculosis dataset. As a result of this different approach, there are three key differences. First, posterior probabilities are rescaled according to how much a trait is 'covered' by observations. This is seen, for example, in feature 1 (and feature 16) in Supplementary Figure S8 A. Here, under WS1, early and late acquisitions of the feature are inferred to be equally likely, as walks are inferred to always run to completion. Under WS2, the number of walks that run to completion is lower (only some observations include 'complete' acquisition). The early acquisition mode is then inferred to be more likely, with a balancing probability that the feature is *not* acquired.

Secondly, with WS1, as the process starts from $\{0\}^L$, for a single random walk, the transitions observed in the dataset are not guaranteed to be reached by random walkers. This means that the overall inferred parameterisations across the entire dataset may not lead to transitions in the dataset being encountered for a finite ensemble of random walks. As a result, the WS1 process does not allow us to directly consider solely the acquisitions between states in the original transition datasets. By exactly considering these transitions, WS2 allows this data to be examined using the parameterisations that have been sampled across the entire dataset allowing for a different type of inference. A clear example of this is seen in Supplementary Figure S8 B for feature *PembA* or *PethA* that are rarely encountered in the window of acquisition where they are acquired in the dataset, illustrated by the strikingly different distributions for WS1 and WS2.

Thirdly, there is no density observed in the grey regions for WS2 due to there being no transitions in the dataset 'covering' these regions, so no transitions performed with WS2 record any density there. In Supplementary Figure S8 B, in application to the tuberculosis dataset, the lack of WS2 density in the grey regions is apparent. In addition, there is clearly observable multimodality in WS2. Multimodality in WS1 is indicative of a feature belonging to multiple progressions that may include an absence of acquisition if the trajectory does not terminate. In contrast, multimodality in WS2 is indicative of multiple progressions where multiple orders of acquisition of a given feature are directly observed in the data. A striking example is *PethA* where in WS1 the predominant visible mode of acquisition is in the grey region towards the end of all possible acquisitions, while in WS2, the acquisition is observed in two distinct regions at step $j = 5$ and step $j = 10$, suggesting that the transition data contains multiple types of progression where *PethA* is acquired. This is also clearly the case for other features such as *PembA*, *PinhA*, *ethA* and *RRDR*.

We introduced WS2 here as a supplementary form of enquiry of the posteriors that can potentially reveal additional inferences about the underlying progressions from which the data may be derived. In the next section, we look in more detail at the assumptions, types of progressions and the outputs in the plots we have used for the inference in order to motivate intuition further.

Implicit assumptions and interpretation of parameterisations

Here we consider several features of datasets that may be considered challenges to inference with HyperTraPS, and illustrate the corresponding outcomes of our approach:

1. *No structure*: only in the case of independent feature acquisition and identical frequencies will no suggestive progression be found, in which case the prior distribution (in this article, uniform across all trajectories) will be recovered by the inference process.
2. *Samples from complete and partial progressions*: If one or more of the underlying progressions does not correspond to a complete walk across the hypercube, transition density in unsampled regions will be dictated by extrapolated dynamics or the prior, depending on whether WS1 or WS2 is used. In Supplementary Figure S9 (i) we illustrate the synthetic dataset (i) for $L = 8$ but for a progression that now stops after gaining feature $i = 4$. In this case, with no other progressions present in the dataset, we find that the remaining features gained in the grey region do so with a uniform distribution over remaining orderings (recovering the prior). In Supplementary Figure S10 (ii) we examine the case where there is a complete right-left path and a partial left-right path (that ends with feature $i = 8$ being acquired, which is the start of the complete trajectory). Trajectories belonging to the left-right transition in WS1 may be interpreted as joining the full right-left path. WS2 does not clearly disambiguate these dynamics – it is not clear whether features 5-8 are acquired. WS1, in the bottom right quadrant of the plot, shows some support for the beginning of the complete progression beginning after the partial progression ends. Supplementary Figure S9 (iii) looks at two partial progressions again illustrating that in the grey region (acquisitions without support in the dataset), there can be a mixed signal from the two partial progressions.

3. *Noisy observations*: We consider the influence of noise in observations in Supplementary Figure S10 by looking at the single left-right progression conflated with noisy observations (from a cross-sectional dataset made up of 10 randomly sampled trajectories). From Supplementary Figure S10 (i)-(iii), the number of noisy (random acquisition of traits) observations increases, introducing breadth into the inferred posterior around the modal pathway (Supplementary Figure S10 for example). However, even with 50% noisy observations in Supplementary Figure S10 (ii), it is possible to clearly recover the modal progression. Even for the extreme case, the non-noisy pathway is almost exactly reproduced with the first greedy path across the hypercube.
4. *Repeated uniform sampling*: When repeated sampling occurs, it can strengthen the inference around where traits are acquired. For example, comparing the first four traits of Supplementary Figure S9 (i) and Supplementary Figure S10 (i), we can see that the repeated sampling afforded by 10 repeated trajectories almost completely removes any density for acquisition off the progression.
5. *Non-uniform sampling across the progression*: We consider this assumption in Supplementary Figure S11 . When some states are sampled a greater number of times, parameterisations that lead to this state will have a stronger 'signal' than those where the observation just occurs once. We illustrate this important effect with several examples. In all cases we consider the complete left-right progression but with the state $s = 11110000$ sampled 100 times more than the others. In Supplementary Figure S11 (i) we see this state acts as a 'gateway' by removing uncertainty for the acquisition of features present in s after s is encountered, and removing uncertainty in acquisition of features absent in s before s is encountered. In Supplementary Figure S11 (ii), the right-left progression is also included but with uniform sampling. The non-uniform sampling leads to a much greater representation of the left-right progression. In Supplementary Figure S11 (iii), two noisy trajectories are now included (only uniform sampling for the noisy trajectories). As the noise is uniform, acquisitions before s still clearly resemble the progression, while features not present in s become affected by the noise.

HyperTraPS and cancer progression models

Understanding pathways of cancer progression is highly complex due to widespread genetic heterogeneity at inter-patient, intra-patient and intra-tumour levels. Several methods aim to infer progression dynamics given different types and structures of observations (Schwartz and Schäffer, 2017). Additionally, cancer progression models can broadly be split into two classes: (i) approaches that consider the multitude of raw 'omic alterations that occur during carcinogenesis and (ii) approaches that take such alterations as absent or present (binarised data), and utilise description of the data at this level to consider progression. Our work fits within the second type of approach where relevant feature subsets have been identified and the presence of absence of such features is a measured aspect in samples.

For understanding variation between patients, no phylogenetic relationship is generally assumed to exist in the accumulation of genetic alterations. Key inference methods applied to binary data at the inter-patient level that determine feature relationships include Conjunctive Bayes network approaches (Gerstung et al., 2009; Beerenwinkel and Sullivant, 2009; Gerstung et al., 2011; Montazeri et al., 2016) and the Tronco packages (Loohuis et al., 2014; De Sano et al., 2016), among a wide-range of similar approaches (Beerenwinkel et al., 2015; Schwartz and Schäffer, 2017) and date back to oncogenetic tree models introduced by (Desper et al., 1999). Recent work by Diaz-Uriarte (2018) suggests that, where complexity in the fitness landscape is present such as with the presence of reciprocal sign epistasis, Bayesian network type approaches in feature space may have shortcomings in being able to represent genetic pathways effectively due to the assumption of monotonicity. As we show in the main text, in contrast to other methods that work with absence/presence data, HyperTraPS focusses on the process of dynamic acquisition in the full space of binary states. This removes the restrictive prior assumption of monotonicity in feature relationships, while presenting tractable parameterisations that include interactions between features. The HyperTraPS platform provides a new means for exploring oncogenetic data at large-scale, lifting this assumption.

In this article, we focus on inter-patient observations, for which established and well-studied datasets allow 'benchmark' comparisons between approaches (as in the main text). However, we note that HyperTraPS' ability to infer dynamics from phylogenetically coupled observations also makes it an appropriate platform for the emerging field of intra-patient cancer study, where 'phylogenetic' with somatic mutations as opposed to solely germline relationships between cells must be considered. Recent methods for understanding feature relationships in single-cell data include SCITE Jahn et al. (2016) and SiFit Zafar et al. (2017), while methods for relating the samples phylogenetically in single-cell data and evaluating clonal clusters include OncoNEM (Ross and Markowetz, 2016). Zafar et al. (2018) discuss these methods in the context of single cell cancer observations. At the intermediate level of attempting to find common relationships in feature space across multiple cancer samples in different patients and different tissues, the recent Revolver platform attempts to provide a unifying interpretative approach via the method of transfer learning Caravagna et al. (2018), and note that HyperTraPS could be readily applied to compilations of patient specific somatic trees too. In Section 2.6 and STAR Methods, we demonstrate that HyperTraPS allows efficient inference of many traits on phylogenies; application of HyperTraPS to these cancer 'phylogenies', and comparison to these alternative approaches, will be the subject of future work.

Regularisation and model validation for ovarian and tuberculosis datasets

In STAR methods, we introduced a greedy backward selection process for inducing parsimonious parameterisations from samples of maximum likelihood models and demonstrated the process for an ensemble for the synthetic datasets (Fig. 2D). In Supplementary Figure S12 A, plots for the ovarian and tuberculosis datasets are also shown with the minimum AIC score at each k from 1000 and 100 (for ovarian and tuberculosis respectively) unique greedy backward selection procedures for different maximum likelihood parameterisations. The AIC score is observed to decrease to a global minimum for each model. First-order models may only have a few parameters removed before reaching a minimum, while second order models, depending on the number of interactions in the underlying dataset, can have a greater proportion of parameters removed.

For the ovarian dataset, the global minimum is sharply found at $k = 30$ following an initial approximately linear decrease. Non-monotonic increase in AIC may then be seen, indicating the interacting nature of parameters to facilitate inference in this model, and is purely an artefact of the greedy backward selection process. For the $L = 19$ genetic features of the complete tuberculosis dataset, a smoother increase in AIC is observed following a global minimum at $k = 149$ parameters, indicating less strong direct interactions between parameter combinations.

Validation calculations for the model (Supplementary Figure S12 B(i)-(ii)) further support this message. All models experience statistically significant support over the null model in terms of the log-likelihood ratio. While the first order regularised model has improved predictive power over the null model, the second order regularised model provides around twice the increase in log-likelihood compared with the first order model. For the test dataset, the second order model has a marginal advantage over the first order model, both producing greater likelihoods than the null model. The lack of the same level of improvement from the second order model for the test dataset, indicates that the parameters remaining for the minimum AIC model from the validation set are not sufficient to capture the full heterogeneity of the datasets in these two specific cases.

Analysis for specific biological datasets

For synthetic, CGH, and tuberculosis datasets, the original data naturally takes the form of presence/absence 'barcodes' with defined features, and can therefore immediately be used in HyperTraPS.

The TCGA study (Bell et al., 2011) includes data on somatic copy-number alterations (SCNAs) from $N = 489$ ovarian carcinoma DNA samples. The authors utilised a focal GISTIC methodology to identify significant peaks of amplification and deletion, and 'key regions' of the genome where these SCNAs occurred. For a given observation, GISTIC analysis assigns an amplitude score and a significance level based on comparison to a control observation. We used these data to build a dataset describing whether or not a significant SCNA was found in each of $L = 55$ chromosomal regions for each patient. We used the authors' GISTIC-derived magnitudes and significance levels, marking an SCNA as present in region R if an observation was found overlapping with region R for which the GISTIC magnitude exceeded 0.2, the associated p-value was under a conservative genome-wide corrected value of 10^{-10}, and the sign of the SCNA (deletion or amplification) agreed with that found in the original key region analysis. A range of changes in these thresholds for magnitude and significance did not have strong qualitative effects on the structure of the inferred pathways. For the PFG analysis with TCGA data we used the WS2 protocol as described in STAR Methods.

In order to consider the data at different coarse grained levels from this full binary dataset, we created the following feature subsets:

- Chromosomal-level *TCGA-C1*: the union of presence/absence aberrations across a given chromosomal arm is considered, leading to $L = 55$ chromosomal features. These are represented as chromosome number (integer), chromosome arm (p/q) and amplification or deletion (+/-)
- High significance chromosomal-level *TCGA-C2*: where we consider the subset of chromosomal positions reported in Fig. 1c of Bell et al. (2011) in particular due to the authors indication that these were of greater significance. This led to a dataset with $L = 27$ features.

We consider HyperTraPS and Bayesian network analysis of TCGA-C2 in the main text. In Supplementary Figure S13 we demonstrate HyperTraPS inferences with TCGA-C1, across all chromosomal arms and key amplifications. Ordering histograms for the features for random walks with WS1 (blue) and WS2 (orange) are depicted with features ordered vertically by mean acquisition step. The inferred order of acquisition is highly heterogeneous, with early acquisitions observed in previously well known chromosomal regions (for example, *8q+, 3q+, 5q-*). There is some multimodality observed in the WS1 and WS2 indicating multiple competing pathways. However, the dominant inferences are with respect to early and late acquisition at this large-scale level of description.

In Supplementary Figure S14 , as described in the Main Text, we demonstrate the limited effect of phylogenetic structure in the tuberculosis dataset on the overall posterior structure.

Likelihood comparison of HyperTraPS with alternative Bayesian network approaches

In this section, we make a direct comparison of the likelihoods computed by the Bayesian network models compared with HyperTraPS. For the likelihoods to be comparable, we must include the additional probability of a random walk that leads to a target state emitting a signal in that target state by incorporating the $P_{\text{emit}}(\{0\}^L, t_i)$ for each $t_i \in D^{\text{transitions}}$. In this case, if signal emission is equally probable across all states, for every sample an additional factor of $1/(L+1)$ must be included for each sample given an irreversible walk from $\{0\}^L$ to $\{1\}^L$ may occupy $(L+1)$ states. This is discussed in further detail in Johnston and Williams (2016).

Supplementary Table S1 provides a comparison of the maximum likelihood output of each model. HyperTraPS produces a similar maximum likelihood to the trained Bayesian network models for dataset D_1, while attaining greater likelihoods for datasets D_2 and D_3 from the ability to capture the competing pathways present in this dataset. For the ovarian dataset, the regularised (AIC criterion) maximum likelihoods are provided for Capri and HyperTraPS, while the maximum likelihood for CBN output is shown. HyperTraPS again attains the largest maximum likelihood. However, it should be noted that Capri model records a lower model complexity making the AIC scores of similar magnitude.

Additional interpretation of findings for tuberculosis dataset

Additional comparisons can be made between the inferred order of polymorphism acquisition in Fig. 7 and Supplementary Figure S8 B and the findings of by Casali et al. (2014). Of the $L = 19$ features used for the analysis, we pick a subset here that provide interesting discussion points with regard to co-associations discussed by the authors. These points demonstrate the ability of HyperTraPS to provide quantitative support for existing hypotheses, and to suggest new avenues of mechanistic research, in complex biological systems.

- *Drug-resistance and fitness compensatory mutations*: Of the $L = 19$ features, the first 16 correspond to the drug-resistant polymorphisms within genes or in the promoter regions. The last three (*rpoA*, *rpoB* and *rpoC* are nonsynonymous SNPs within RNA polymerase genes. The authors considered the occurrence of compensatory mutations in *rpoA* and *rpoC* in response to drug-resistance polymorphism in *rpoB*. WS2 reveals an acquisition ordering with *rpoB* and *RRDR* being acquired prior to *rpoC*, suggesting a compensatory effect follows drug-resistance mutations in this case, while *rpoA* is acquired primarily in some cases and then typically later with similar acquisition patterns to *rpoC*.

- *Genetic sites particularly associated with adaptive selection*: Highly polymorphic genes conferring resistance are known to be *embB*, *pncA*, *ethA* (Casali et al., 2014). Interestingly these polymorphisms occur at a wide range of orderings within the inferred orderings, illustrative of their flexibility and why they may be particularly polymorphic – they can play different roles in different progressions.

- *Transmissibility of drug-resistance*: With respect to transmissibility Casali et al. (2014) suggest that *katG* is prior to *RRDR*, which is supported in the top two greedy paths highlighted in the hypercube plot in the main text Fig. 7.

Here we make a direct comparison of the order in which mutations are acquired with Simmap, which takes the form of a continuous time Markov model with mater equation approach to acquiring characters that belong to leaves on a phylogeny. This approach runs into computational issues when the number of states under evolution grows large (only tractable in short run times for the tuberculosis up to $L \approx 5$). This is in contrast to HyperTraPS which can handle the full $L = 19$ traits.

As an illustration of compatibility with this alternative approach, we restrict the tuberculosis dataset to $L = 3$ features (*katG*, *PinhA* and *RRDR*) with the full set of isolates and enforce single irreversible acquisitions as transitions within the Simmap model in order to make direct comparisons with HyperTraPS. In Supplementary Figure S15 A, we show the output for the density of order of acquisition from simulated rate matrices outputted by Simmap with the hypercubic restriction imposed and irreversibility. Alongside in Supplementary Figure S15 we show the result for WS2 with HyperTraPS (as the transitions performed with Simmap are to the sample data and do not fully acquire all features as is the case with WS1). The plots are in close agreement, providing good validation that HyperTraPS generates results consistent with current platforms.

HyperTraPS: Inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways

Supplementary Figures & Tables

Dataset	Maximum regularized likelihood with Capri	Maximum likelihood with CBN	Maximum likelihood with HyperTraPS
Synthetic D_1	-16.64	-17.52	-16.64
Synthetic D_2	-46.97	-48.96	-41.59
Synthetic D_3	-88.81	-86.05	-80.04
Ovarian CGH	-356.57*	-380.01	-347.72*

Table S1: Maximum likelihood values for Capri, CBN and HyperTraPS outputs with each synthetic cross-sectional dataset and ovarian CGH dataset. Where there is a single progression (dataset D_1) all models reproduce the similar maximum likelihoods. Where there is more than a single progression (datasets D_2 and D_3), the additional stochastic flexibility available in HyperTraPS parameterisations allows models with larger maximum likelihoods to be recovered. For the ovarian dataset, HyperTraPS and Capri both have likelihoods compared in regularised forms (denoted with asterisks), with HyperTraPS again attaining the largest likelihood. It should be noted however, that the model complexity of the Capri model is less than that for HyperTraPS in this case, leading to a lower AIC score (not shown above).

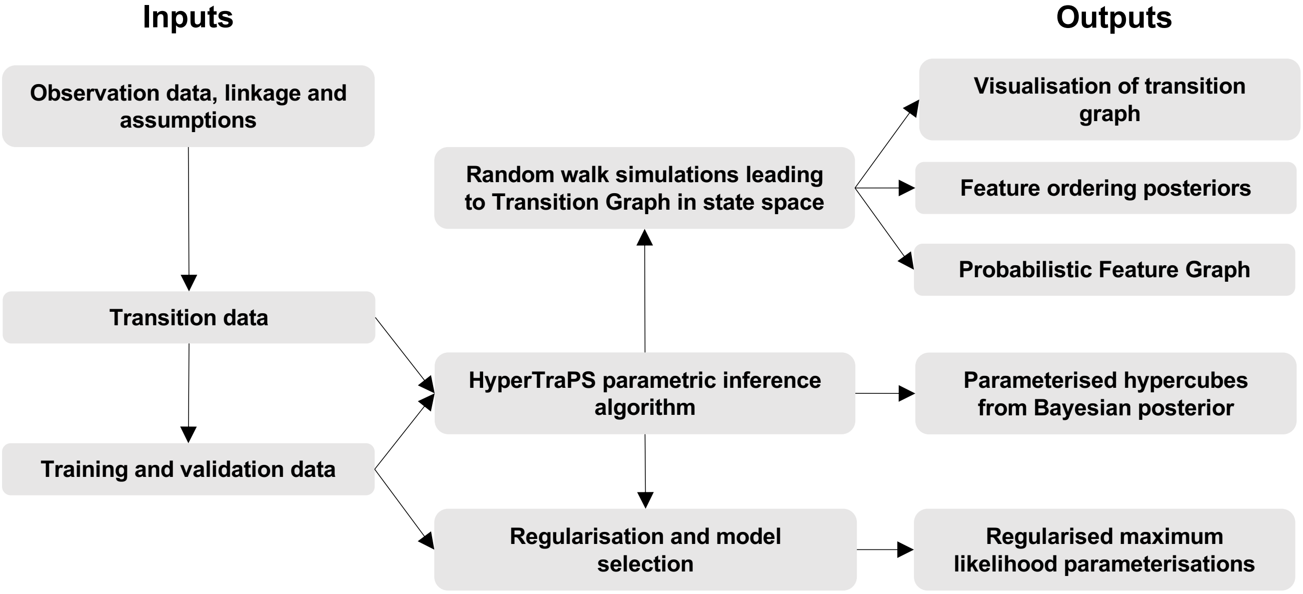

Inputs
Outputs
Observation data, linkage and assumptions
Transition data
Training and validation data
Random walk simulations leading to Transition Graph in state space
HyperTraPS parametric inference algorithm
Regularisation and model selection
Visualisation of transition graph
Feature ordering posteriors
Probabilistic Feature Graph
Parameterised hypercubes from Bayesian posterior
Regularised maximum likelihood parameterisations

Fig. S1: (related to ‘HyperTraPS pipeline’) **An illustration of the pipeline from inputs to outputs with the underlying inference, application and description methods within HyperTraPS.**

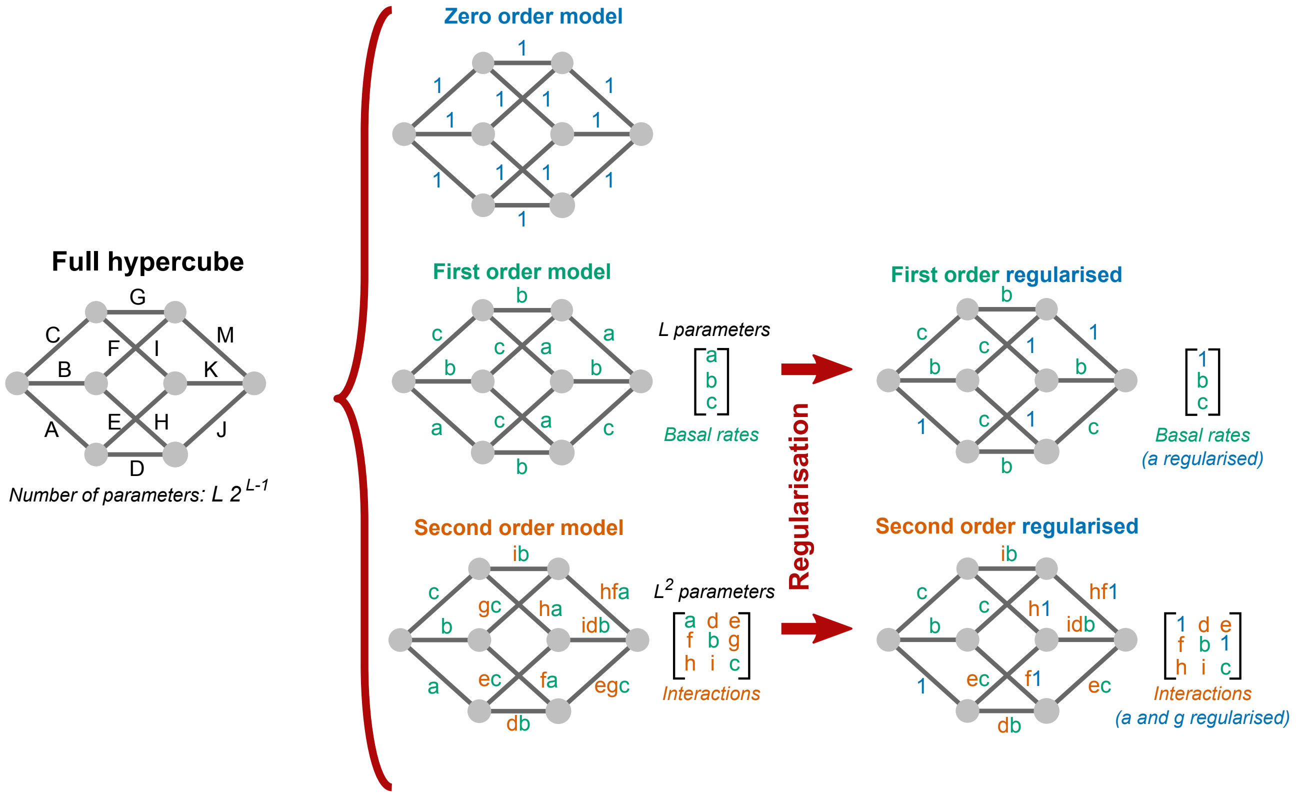


Fig. S2: (related to 'Tractable parameterisations of hypercube') **Tractable parameterisations and regularisation.** A full irreversible directed hypercube is parameterised by edge set W and contains $L2^{L-1}$ edges. We define three orders of model (*zero order*, *first order* and *second order*) for reducing the parameter space and regularised models (*first order regularised and second order regularised*). The zero-, first- and second- order models are nested in the sense that a second order model can capture the first order model (interaction terms all set to unity) and the first order model can capture the zero order model (all basal terms set to unity). In the example above, for $L = 3$, the 12 edges of the full hypercube (A-K) are reduced down to combinations of a set of 9 parameters (a-i). The advantage becomes clear for larger L. At $L = 16$, over 500,000 edge weights are reduced to just 256 parameters for the second order model. Regularisation harnesses structure in the data to further reduce model complexity. We utilise a greedy backward selection process to identify which parameters may be removed (set to the value of the zero order model, unity) and decrease a criterion, which we choose to be the Akaike Information Criterion. In the illustration above, for the first order regularised model, parameter a is set to unity and, for the second order regularised model, parameters a and g are both set to unity (as would be the case in a zero order model) with the consequent impact on the hypercube edge weights shown.

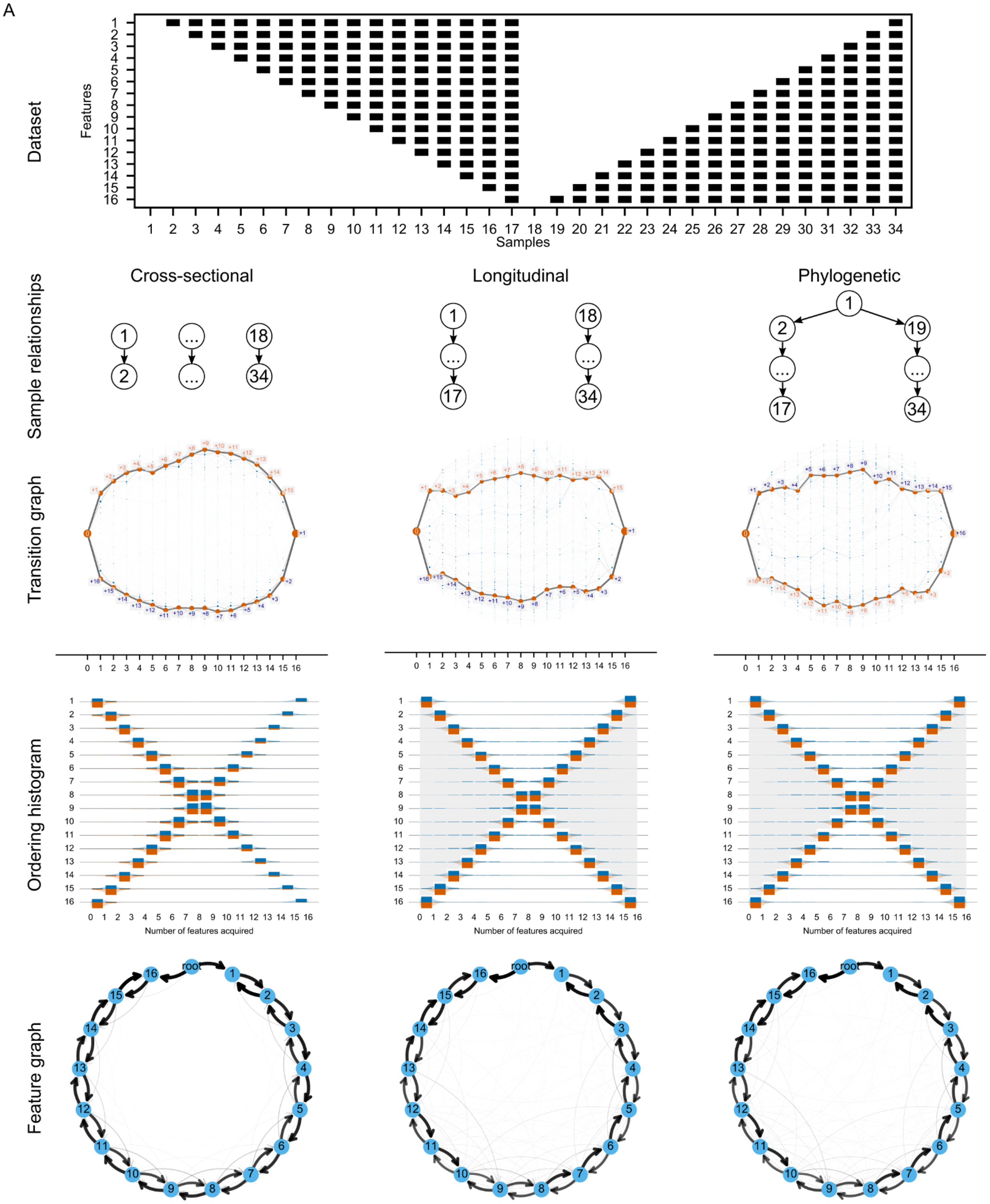


Fig. S3: (related to 'Testing and validating HyperTraPS with differing data structures, volumes, and priors') **HyperTraPS inference with different data types.** The results of HyperTraPS inference using the $L = 16$ two-pathway synthetic system in the main text, where observations are cross-sectional, longitudinal, or phylogenetically linked. Pathways are readily recovered; posteriors are slightly sharper for cross-sectional data, as each observation is independent and thus provides more evidence than the coupled observations under the other two modes.

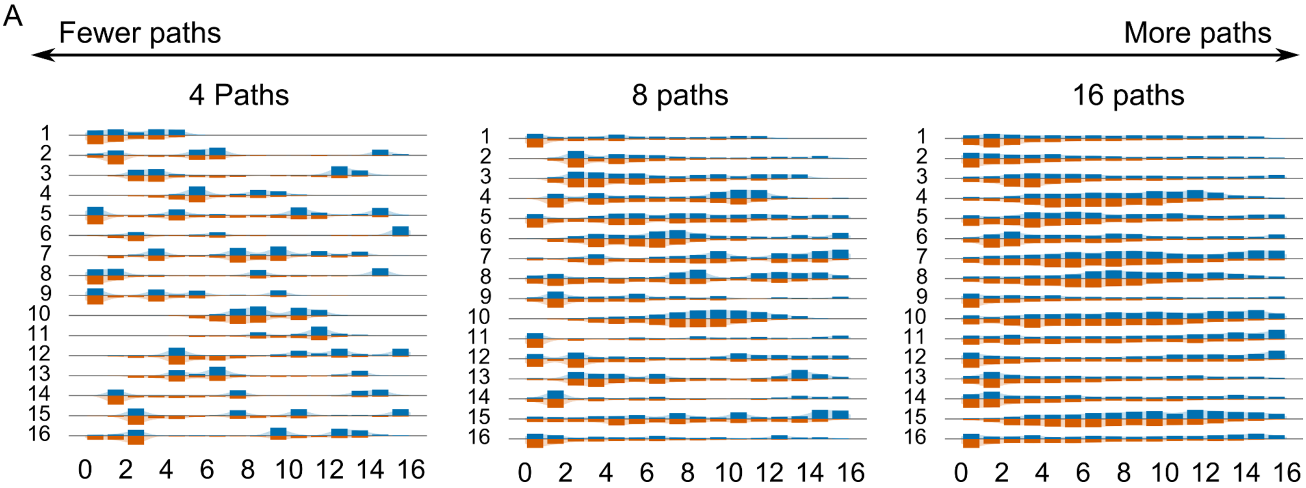


Fig. S4: (related to 'Testing and validating HyperTraPS with differing data structures, volumes, and priors') **HyperTraPS inference with different numbers of competing pathways.** Posteriors corresponding to the hypercube plots in Fig. 3A, using different $L = 16$ synthetic systems like those in the main text, but supporting different numbers $p > 2$ of competing paths, with $N = 16p$ observations. Four and eight pathways are readily discerned; sixteen independent pathways poses more of a challenge, although posterior density is still aligned with the synthetic pathways.

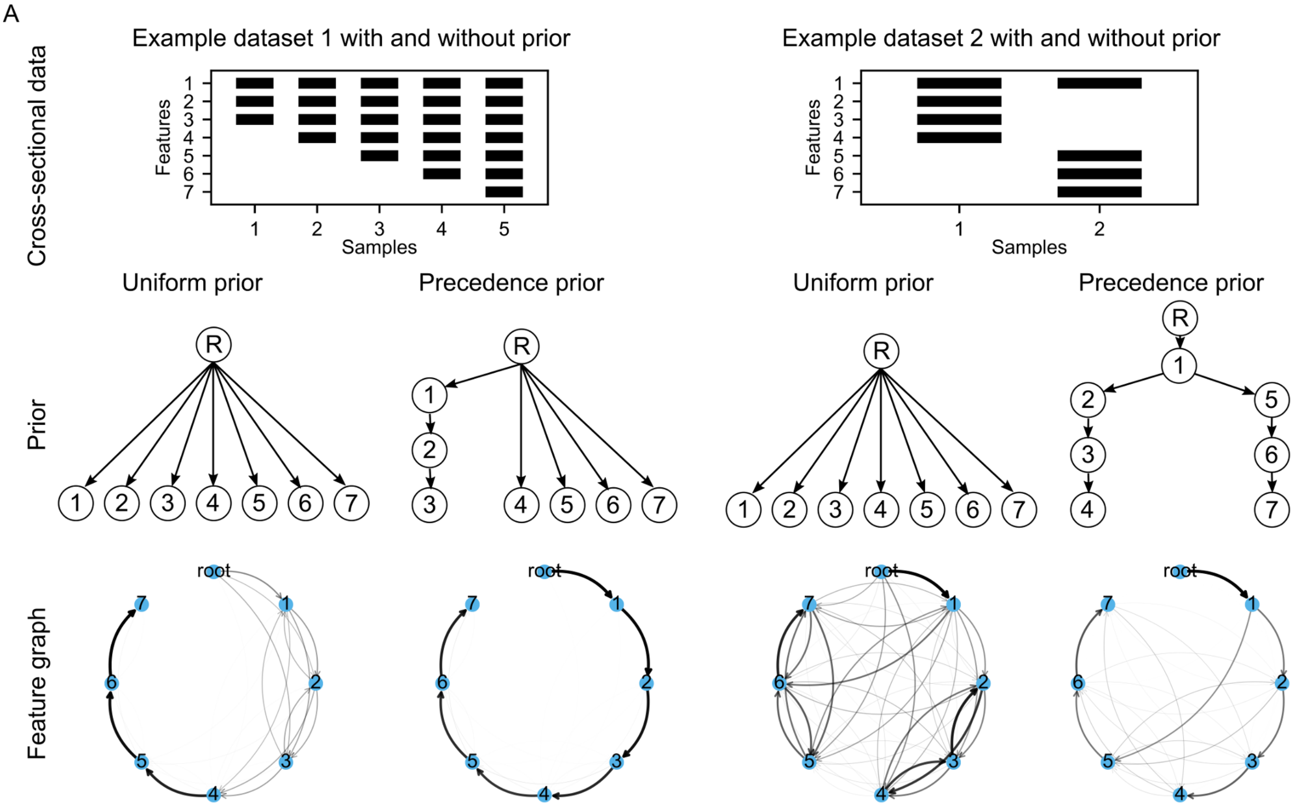
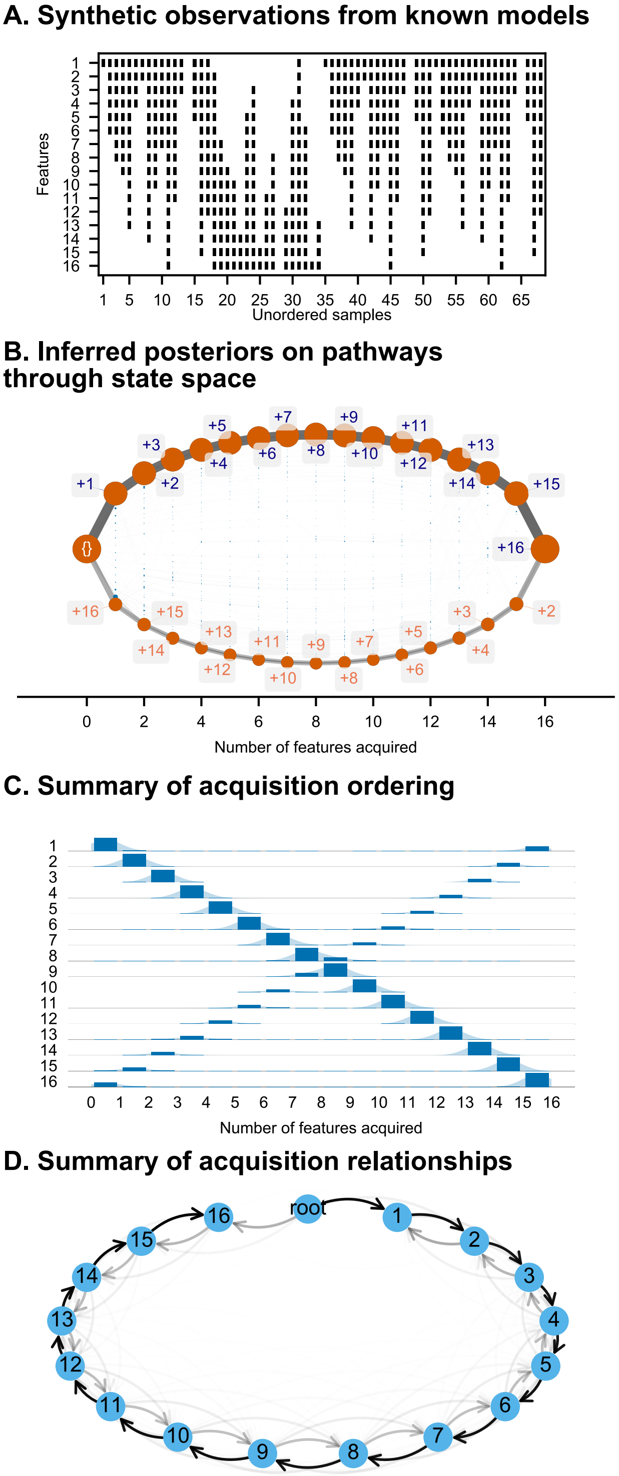


Fig. S5: (related to 'Testing and validating HyperTraPS with differing data structures, volumes, and priors') **HyperTraPS inference including prior information on pathway structure.** Probabilistic feature graphs corresponding to the inclusion of prior knowledge in Fig. 3C.



A. Synthetic observations from known models

B. Inferred posteriors on pathways through state space

C. Summary of acquisition ordering

D. Summary of acquisition relationships

Fig. S6: (related to ‘Additional synthetic cross-sectional dataset’) **HyperTraPS inference with additional synthetic dataset.** **(A)** The structure of this synthetic dataset, supporting two competing pathways with features in different orders and with a likelihood ratio of 3:1 between the opposing orderings. **(B)** Inferred dynamics on the hypercubic transition graph. The two competing pathways are recovered in proportion to the amount they are observed in the dataset (the ratio of 3:1). **(C)** Inferred dynamics represented as the posterior probability that a feature (horizontal axis) is acquired at a given step (vertical axis), with bi-modality in proportion to the prevalence of each pathway. **(D)** Inferred dynamics represented as a graph summarising trait acquisition relationships. An edge from node i to node j suggests that trait i is acquired in the previous step before the acquisition of trait j. Again, two clear directions oaf acquisition can be seen with edge weights in proportion to their frequency in the underlying cross-sectional datasets.

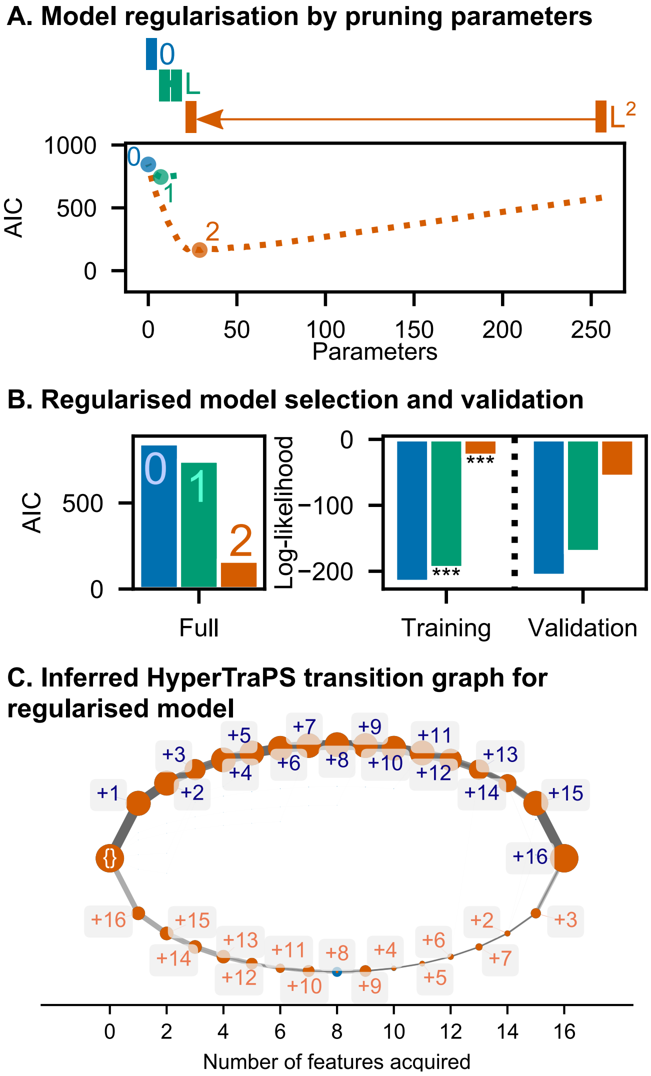


Fig. S7: (related to 'Additional synthetic cross-sectional dataset') **Regularisation and validation for the additional synthetic dataset. (A)** Regularisation of the second order model (orange) leads to many fewer parameters than the full L^2 but still greater than the first order model's $L = 16$ due to the two paths being present, necessitating interactions between features. **(B)** Regularised model selection and validation illustrates that the regularised first order model does better than the null model due to the full ordering of the dominant pathway that it is able to capture. The regularised second order model, however, results in much larger likelihoods still as it is able to capture both paths from the data. **(C)** Pathway structure remains well captured by the regularised model.

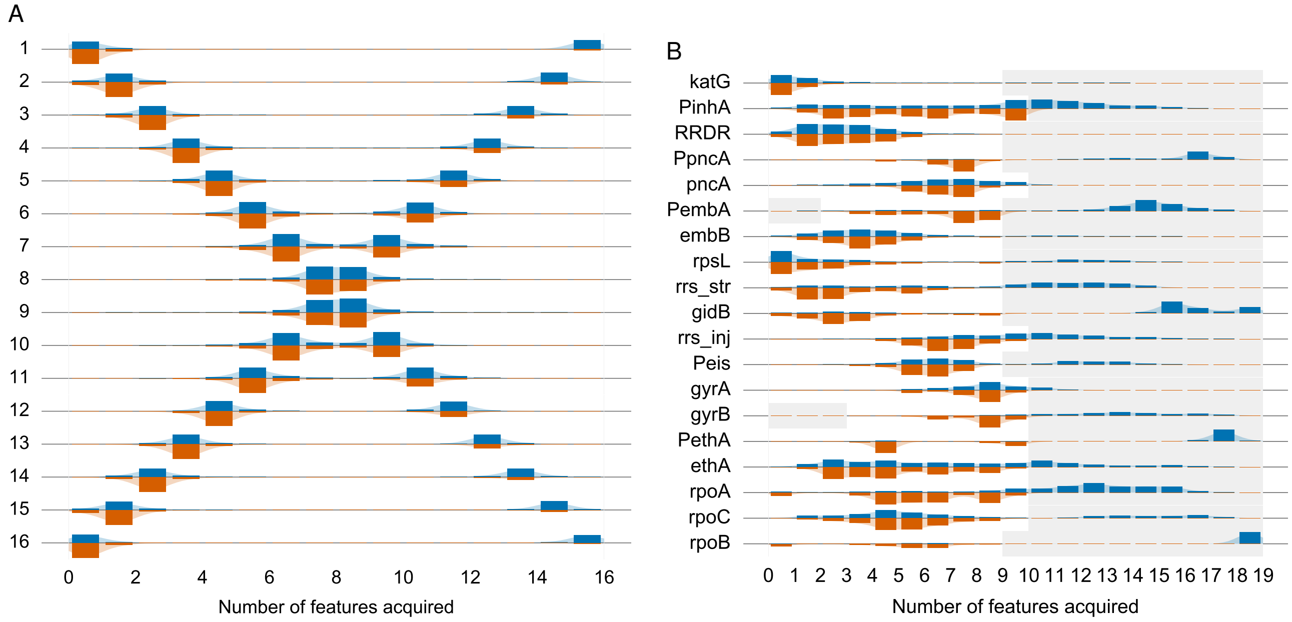


Fig. S8: (related to 'Alternative interpretation of inferred acquisition orderings') **Comparison between WS1 and WS2 represented for the cross-sectional dataset (ii) (A) and tuberculosis (B) from the main text.** The blue bars are illustrate density corresponding to acquisitions with WS1 and the orange bars density for acquisitions with WS2. Kernel density estimates are overlaid to guide the eye.

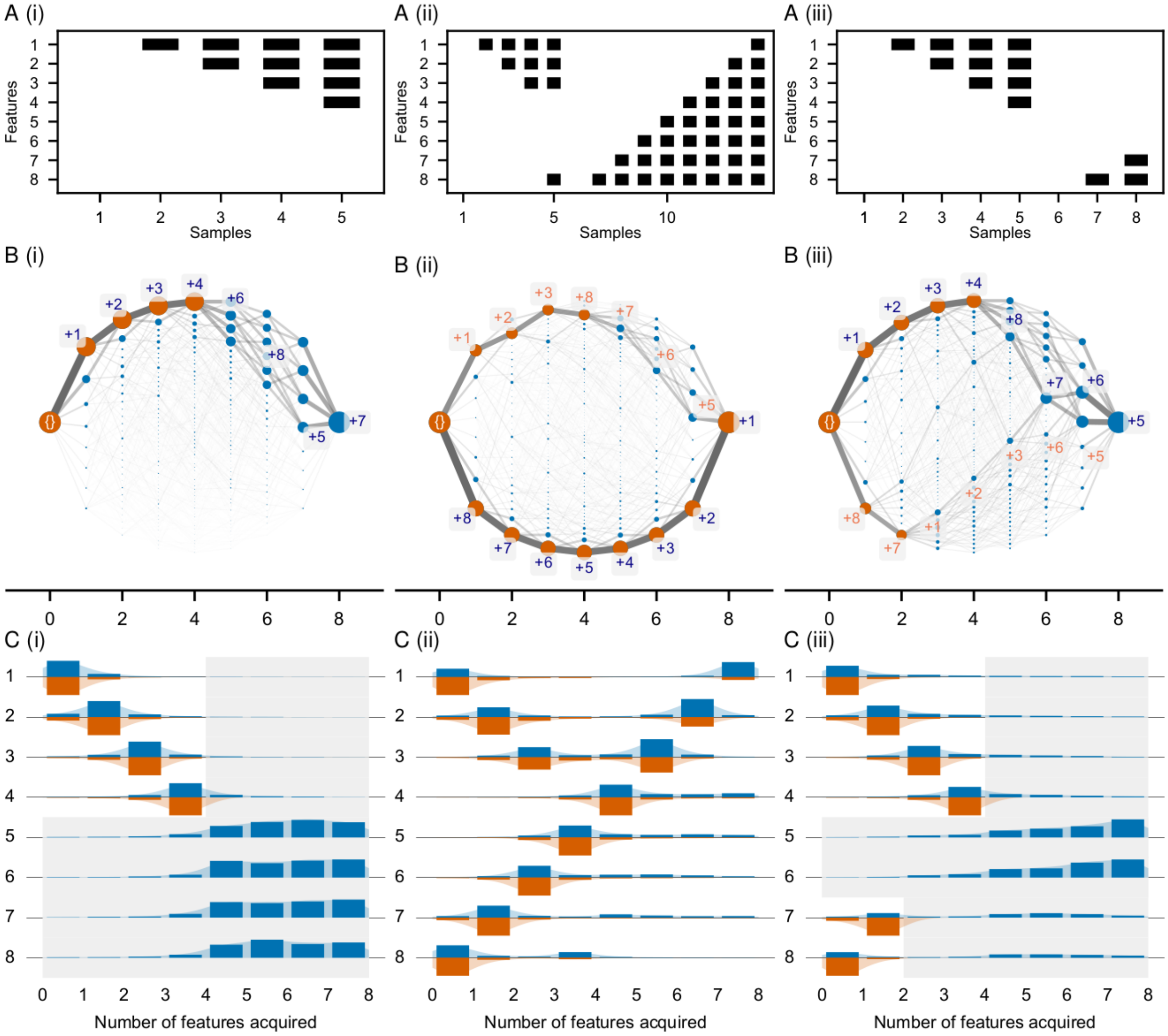


Fig. S9: (related to 'Implicit assumptions and interpretation of parameterisations') **HyperTraPS inference in the presence of partial and multiple progressions.** Three datasets are considered: (i) A single partial progression (1,2,3,4); (ii) A single partial progression (1,2,3,8) and a complete second progression (8,7,6,5,4,3,2,1); and (iii) Two partial progressions (1,2,3,4) and (8,7). In each case: **(A)** shows the dataset structure (*dataset plots*); **(B)** The inferred paths on the hypercube with samples from the second order posterior and WS1 simulations (*hypercube plots*). Orange vertices are observed in the dataset, while blue ones are not; and **(C)** The corresponding histograms for WS1 and WS2 (*histogram plots*). For (i), the partial progression is inferred following by uniform acquisitions in line with the prior expectation. In the hypercube plots, paths on the hypercube are seen to diverge with equal proportion in this region illustrating this point. For (ii), the hypercube plot highlights the ability to infer both progressions. The longer path has greater weight due to an increased number of observations associated. The greedily labelled paths show an interesting feature where at the end of the partial progression, as the last feature is the first feature of the complete progression, the pattern of acquisition seen in the second progression is 'predicted' to occur in continued acquisition. This is visible in the histogram plot by the asymmetric density in WS1 flowing from feature $i = 7$ for the fifth feature acquired onwards. For (iii) with two partial progressions, the two paths are clearly distinguished in the hypercube plot with the same property of the progressions continuing on from each other after each partial progression is completed, eventually joining together after the sixth feature is acquired. The spread of other states encountered highlights the stochastic nature of the platform's predictions.

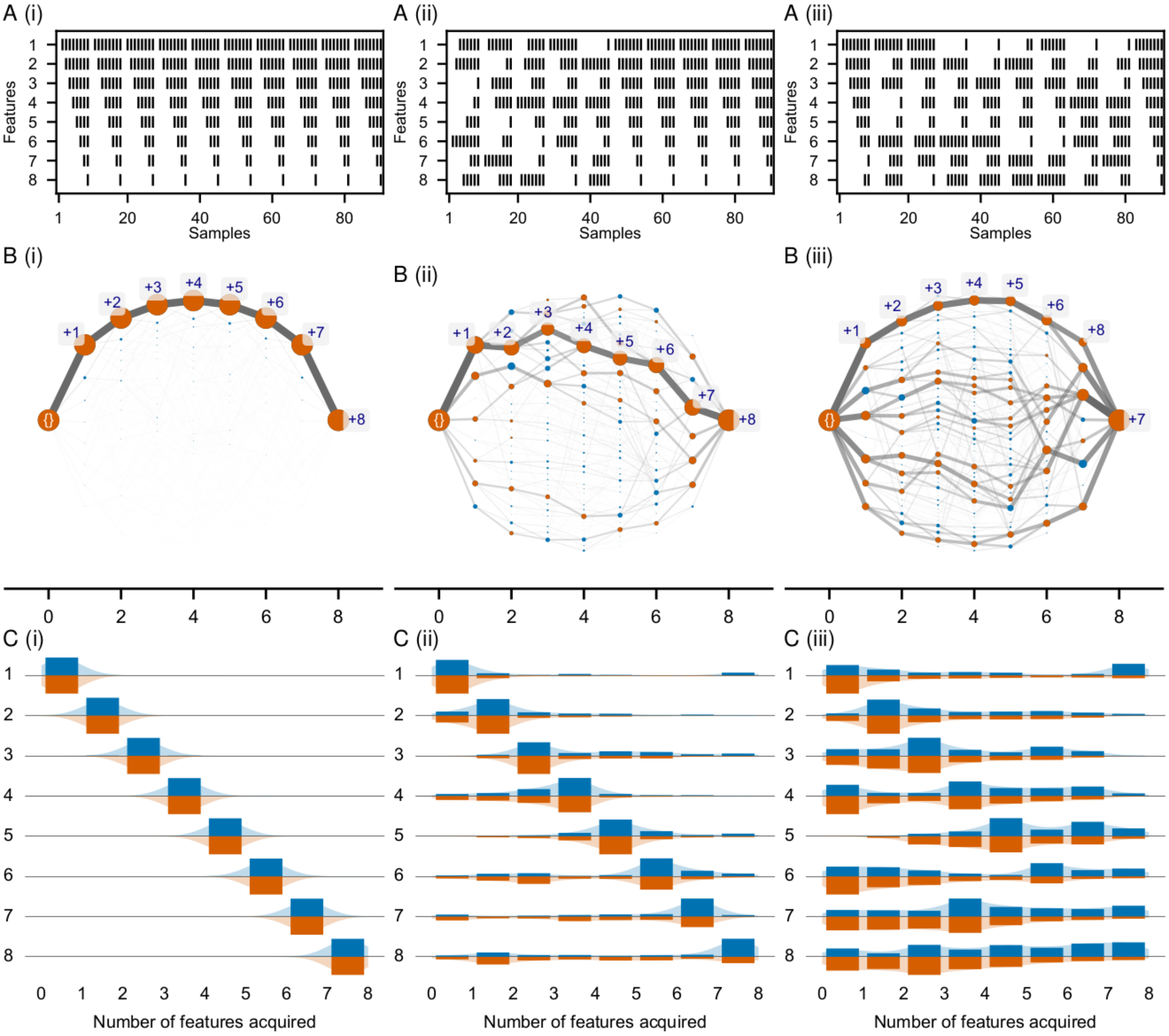


Fig. S10: (related to 'Implicit assumptions and interpretation of parameterisations') **HyperTraPS inference in the presence of noisy samples.** (i) One complete progression with ten samples from each state instead of a single sample ($|D| = 10L$ compared to $|D| = L$). (ii) Five out of the ten trajectories part of the dataset involved the features being randomly acquired instead of the left-right progression. (iii) Nine out of the ten trajectories part of the dataset involved the features being randomly acquired instead of the left-right progression. The figure structure mirrors that of Supplementary Figure S9 . For (i), the hypercube plot and histogram plot shows more tightly defined paths due the ten-fold increase in data supporting the primary pathway, pushing the posterior towards the maximum likelihood parameterisation. In (ii), the introduction of this noise is visible but does not obscure the dominant non-noisy progression from being disambiguated. (iii) For (iii), the introduction of the uniform noise has a significant effect on the nature of paths observed across the hypercube, although even in this case it should be noted the appearance of the first greedy path being almost identical in structure to the non-noisy path structure.

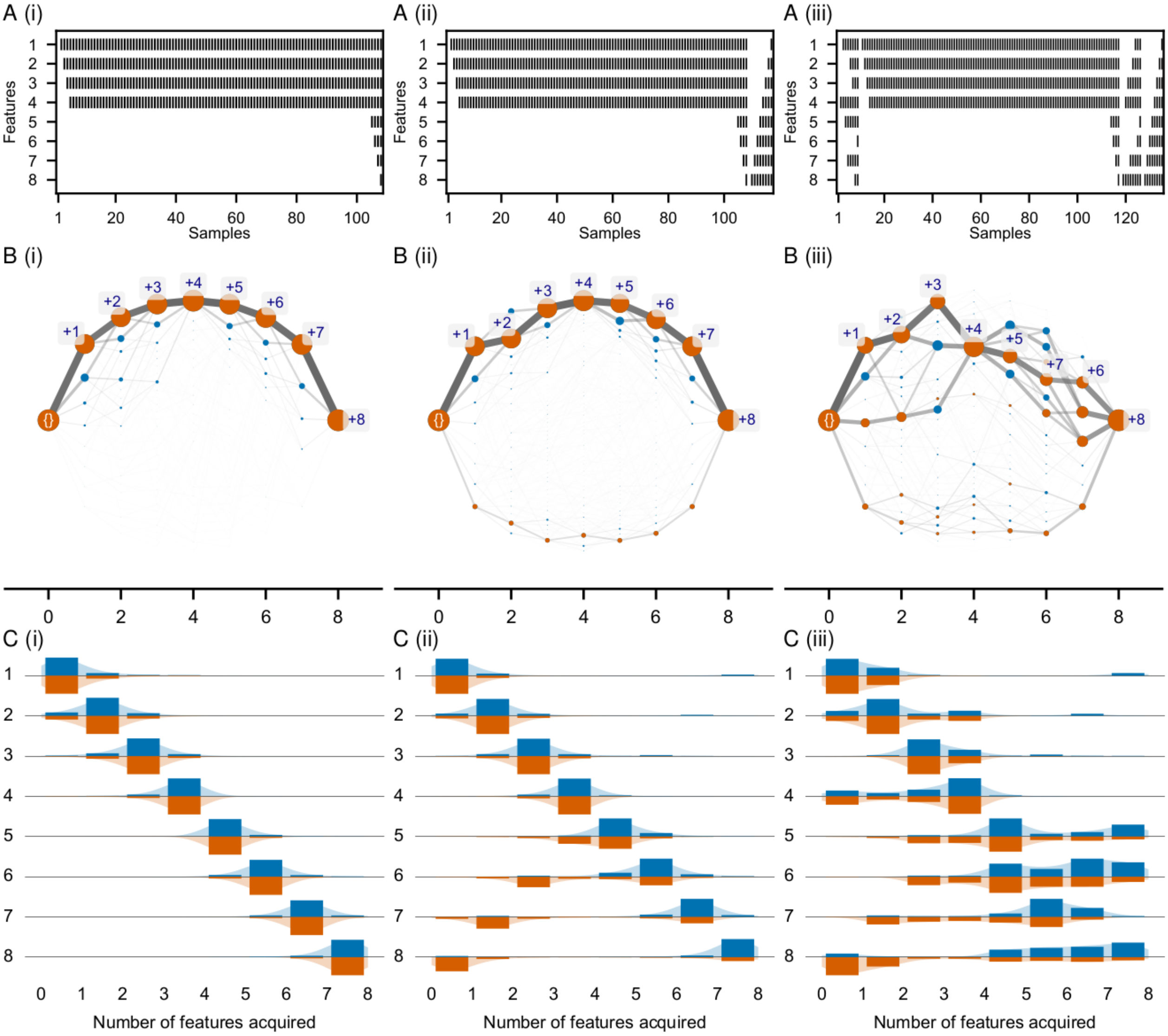
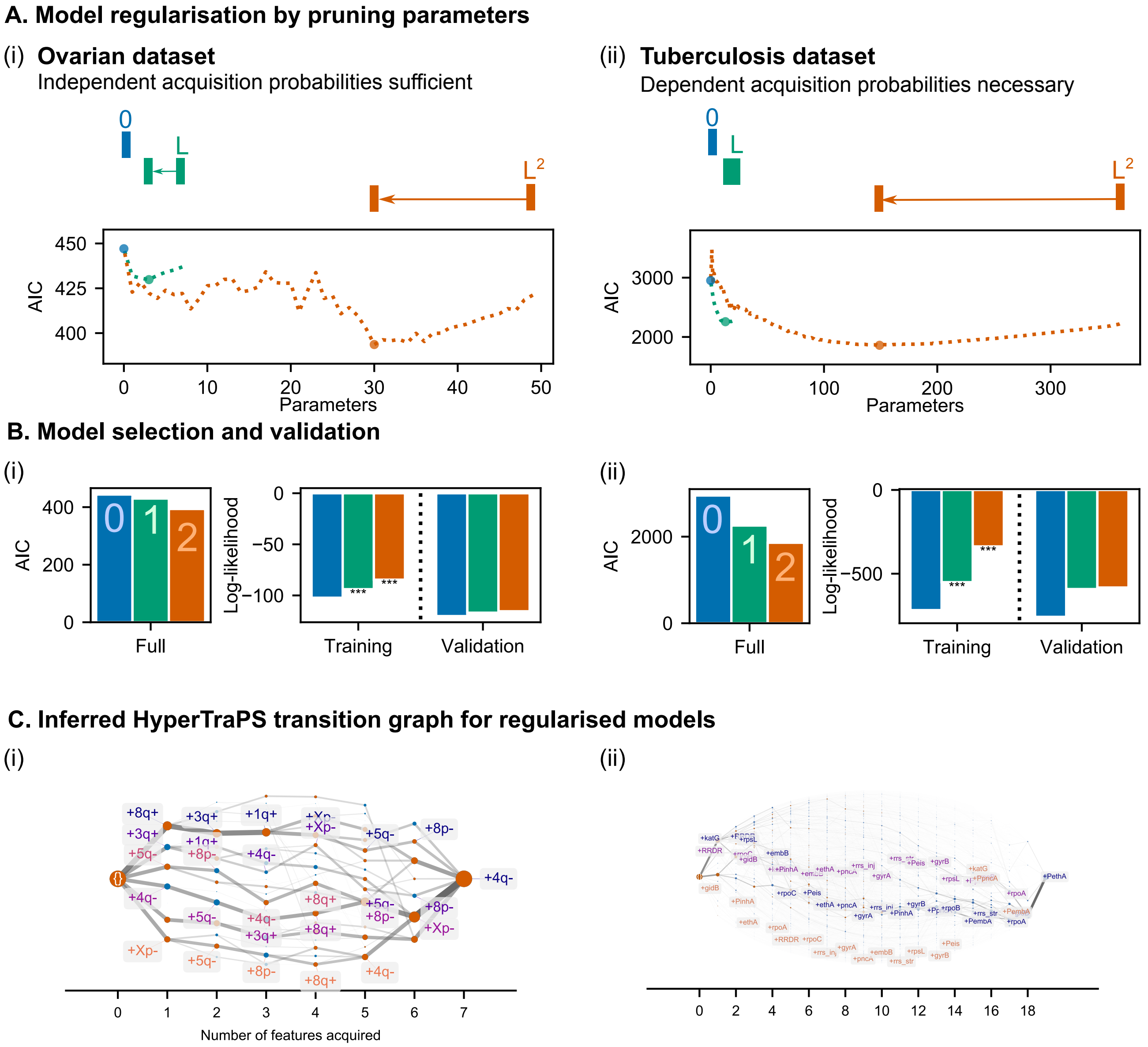


Fig. S11: (related to 'Implicit assumptions and interpretation of parameterisations') **HyperTraPS inference in the presence of non-uniform sampling.** In each of (i)-(iii) the state $s = 11110000$ is sampled 100 times more than all other samples. (i) Only the single left-right progression. (ii) The single left-right progression with the non-uniform sampled middle state is present and a second progression with uniform sampling from right-left. (iii) Same as (ii) but a single noisy progression is added in each direction. The figure structure mirrors that of Supplementary Figure S9 and Supplementary Figure S10 . For (i), the oversampled state acts as a gateway with uncertainty remaining in the regions where acquisition occurs before and after the gate. For example, $f_{45} \approx 0$ in contrast to Supplementary Figure S9 (a), while $f_{43} \neq 0$ as for the uniform case. For (ii), where two progressions are present but only the left-right has oversampling in the middle, due to the oversampling in the left-right path there is a large bias towards random walks from 0^L following this path, as seen by the strength of corresponding path in the hypercube plot. WS2 allows for this to be accounted for illustrating the other pathway more clearly as the simulations ensure the right-left progression is visited. For (iii), noise is now introduced for both progressions. As the noise is uniform, acquisitions before the oversampled state s still resemble the dominant progression, while subsequently the noise clearly affects the order of acquisition increasing the uniformity of feature acquisition. The right-left progression becomes difficult to distinguish at all due to a lack of random walks beginning at 0^L following this progression. However, the ability for the inference to perform random walks that take this weaker and noisy second progression is remarkable as observed by the fact orange states from the data associated with the progression are still encountered.


A. Model regularisation by pruning parameters

(i) Ovarian dataset
Independent acquisition probabilities sufficient

(ii) Tuberculosis dataset
Dependent acquisition probabilities necessary

B. Model selection and validation

C. Inferred HyperTraPS transition graph for regularised models

Fig. S12: (related to ‘Regularisation and model validation for ovarian and tuberculosis datasets’) **Regularisation and model validation for the ovarian and genetic tuberculosis dataset. (A)** Regularisation of the parameterisations for the (i) ovarian dataset and (ii) tuberculosis genetic dataset ($L = 19$ genetic sites). Dashed green and orange lines illustrate the minimum AIC found at each value of k over the ensemble of backward selection processes. Circles illustrate the minimum for each order of model. The second order model is favoured produces lower AIC scores in both cases. **(B)** Model validation for the (i) ovarian dataset and (ii) tuberculosis dataset. In each of B(i) and B(ii), the left-hand plot depicts lower AIC scores for the second order models. The right-hand plots show highly significant second order regularised models compared to the null model and much larger log-likelihoods on the validation datasets. **(C)** Transition graphs constructed from WS1 random walks with the minimum AIC second order regularised models for each dataset.

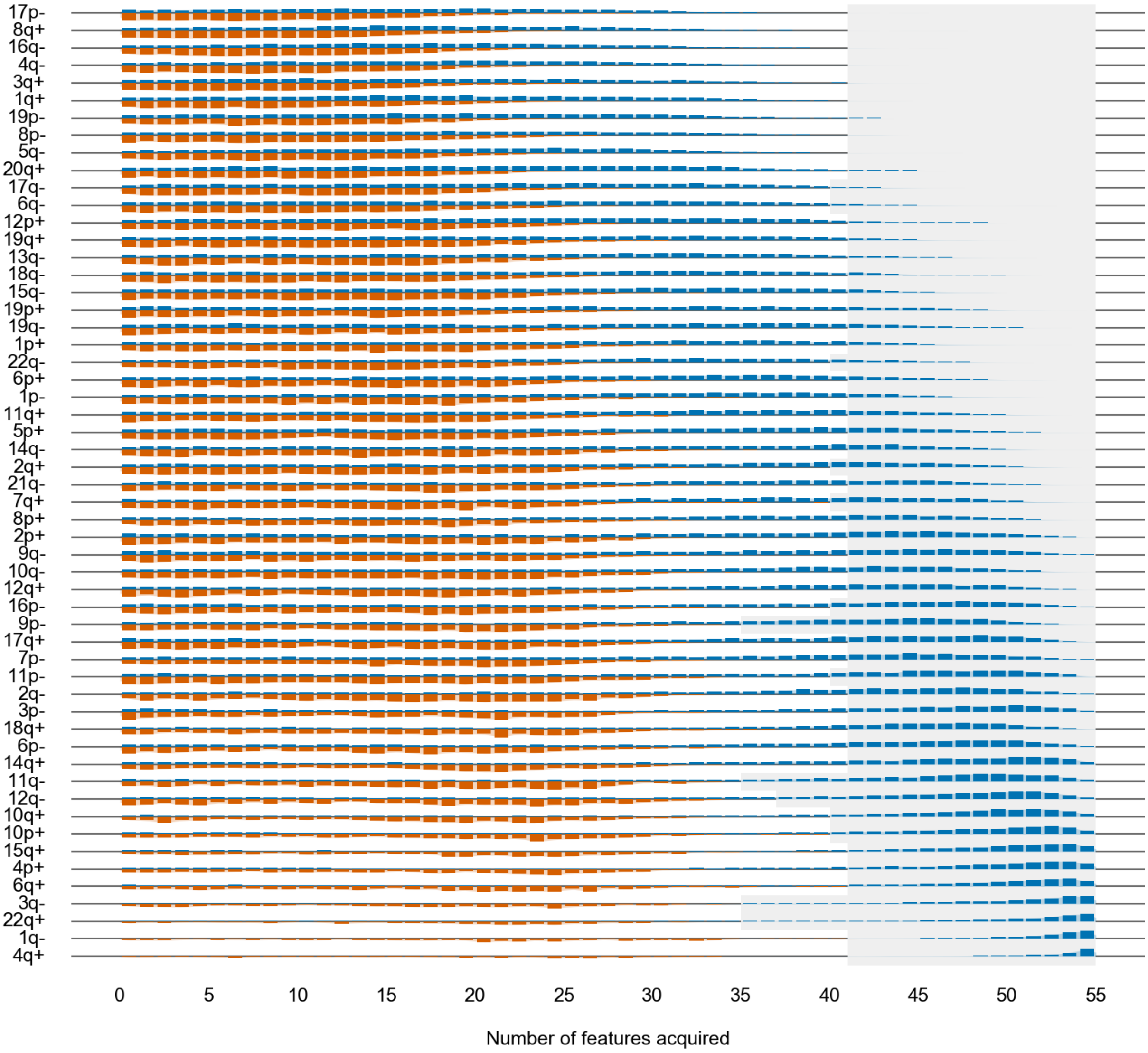


Fig. S13: (related to ‘Analysis for specific biological datasets’) **Ordering histograms for random walks from posterior samples for the TCGA-C1 dataset are depicted.** Random walks with WS1 (blue) and WS2 (orange) are summarised into feature acquisition proportions at a given time. The features are ordered by mean acquisition time from WS1. The order of acquisition is highly heterogeneous, with general trends of early and late acquisition being clearly attributable to each feature. However, there is wide dispersion in the exact time of acquisition in almost all cases. There is some multimodality observed in the WS1 and WS2 indicating multiple competing pathways.

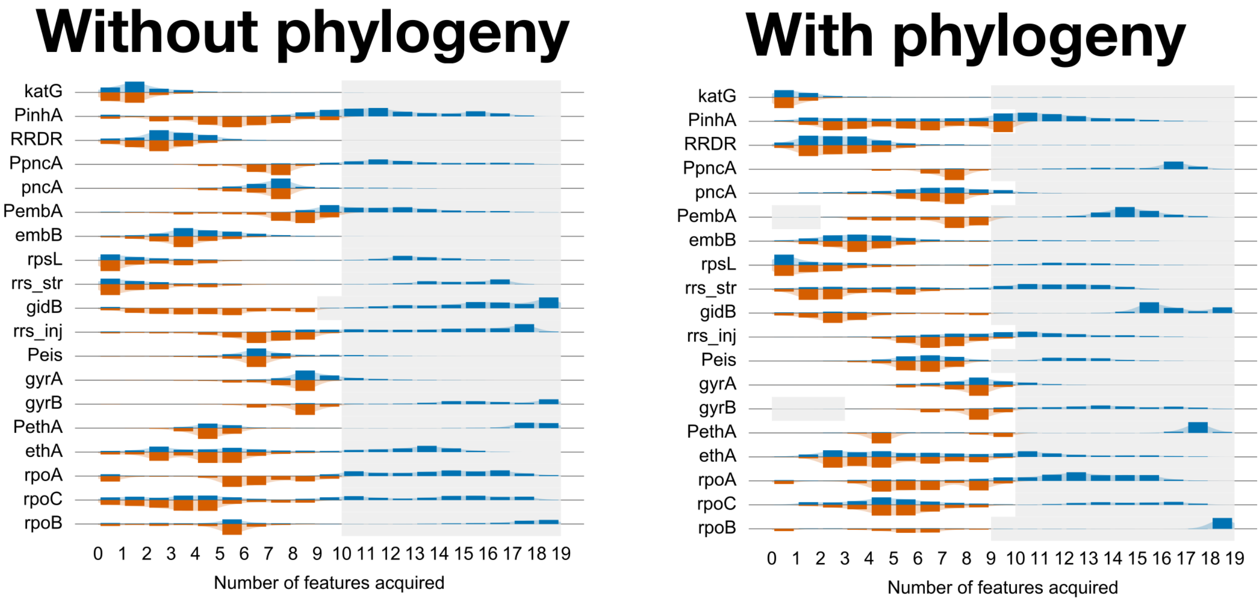


Fig. S14: (related to 'Additional interpretation of findings for tuberculosis dataset') **Tuberculosis pathway inference and phylogenetic information.** (left) The inferred structure of tuberculosis feature acquisitions, given the phylogeny used in the main text. (right) The inferred structure in the absence of the phylogeny, treating each observation as independent. Most ordering posteriors remain qualitatively similar to those inferred with phylogenetic information, illustrating their robustness to errors in phylogenetic structure.

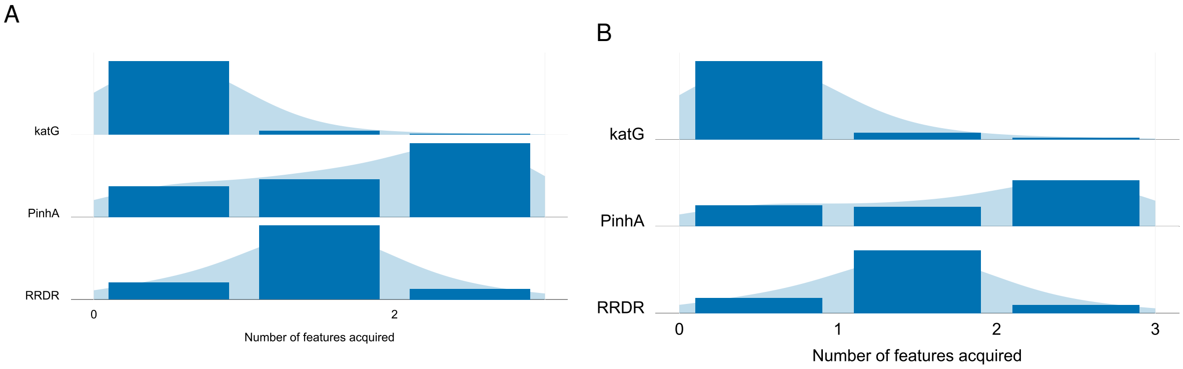


Fig. S15: (related to 'Additional interpretation of findings for tuberculosis dataset') Comparison of the tuberculosis dataset analysed with both HyperTraPS (A) and Simmap (B) on the restricted, tractable set of genetic sites: *katG*, *PinhA* and *RRDR*.