Reconstructing unobserved cellular states from paired single-cell lineage tracing and transcriptomics data

Khalil Ouardini [1 2] **Romain Lopez** [1] **Matthew G. Jones** [3 4] **Sebastian Prillo** [1] **Richard Zhang** [1]
Michael I. Jordan [1] **Nir Yosef** [1]

Recent advances in CRISPR/Cas9-based genome engineering and single-cell sequencing assays have enabled the simultaneous measurement of lineage information and transcriptomic state at the single-cell resolution (Raj et al., 2018; Chan et al., 2019). Already, these technologies have been used to study mammalian embryogenesis (Chan et al., 2019) and cancer metastasis (Quinn et al., 2021), amongst other applications. In these studies, researchers use phylogenies - tree structures that describe relationships between observed cells - to infer transcriptomic determinants of dynamic phylogenetic patterns. A key limitation thus far, however, derives from the fact that only the leaves of these trees are directly observed whereas the internal nodes of the tree represent unobserved, *ancestral* states. While rich insights can be yielded from the relationships between leaves alone, accurately inferring these ancestral states of a tree would allow researchers to formulate far more sophisticated "evolutionary" models of biological processes (Wagner & Klein, 2020).

The variational autoencoder (VAE) (Kingma & Welling, 2014) is a powerful framework for fitting flexible generative models to data in a scalable fashion. In particular, such tools have been applied in several areas of molecular biology (Lopez et al., 2020), including modeling of single-cell RNA sequencing (scRNA-seq) data. Most of the subsequent research focuses on the case where each datapoint is an independent replicate of the same generative process, providing better variational distributions or more flexible models (Kingma et al., 2016; Louizos et al., 2016; Burda et al., 2016). In this work, we seek to exploit the tree structure generated from lineage tracing as prior information about the sample-sample covariance structure. In doing so, we seek to fit a prescribed model that can be used to predict ancestral expression across the tree in a principled fashion. Here we introduce TreeVAE, a fully-probabilistic approach that builds on previous work, such as the time-marginalized coalescent VAE (Vikram et al., 2019), by tailoring the inference to bespoke observation models for scRNA-seq and rich phylogenies inferred from CRISPR/Cas9-based lineage tracing data. After describing our generative model and an inference procedure for it (Section 1), we compare TreeVAE to alternative methods on simulated and real datasets (Section 2).

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley, USA [2]École CentraleSupélec, Gif-sur-Yvette, France [3]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, USA [4]Biological and Medical Informatics Graduate Program, University of California, Berkeley, USA. Correspondence to: Romain Lopez <romain_lopez@berkeley.edu>, Matthew G. Jones <mattjones315@berkeley.edu>, Nir Yosef <niryosef@berkeley.edu>.

The 2021 ICML Workshop on Computational Biology.

1. Tree Variational Auto-encoder (TreeVAE)

1.1. Formal Description of the Generative Model

We assume that we know a phylogeny $\mathcal{T} = (V, E, b)$, a directed rooted tree with vertex set V, edge set E and edge length function b. We note the weight $b(e)$ of an edge $e = (u, v)$ as $b_{u,v}$. The vertex set V is partitioned into L leaf vertices $V_L = \{1, \dots, L\}$ (a set of cells) and I internal vertices $V_I = \{L+1, \dots, L+I\}$ (ancestral cell states) such that $V = V_L \cup V_I$. Phylogenies may be inferred from single-cell lineage tracing mutation data using publicly available methods such as Neighbor Joining (Saitou & Nei, 1987) or Cassiopeia (Jones et al., 2020).

We introduce a probabilistic model describing the evolution of latent random variables z_v at every vertex v along the phylogeny $\mathcal{T}$. Those latent variables correspond to a low-dimensional embedding of cell states, as explored in previous work (Lopez et al., 2018). Unlike in most applications of generative models to single-cell data, where those variables are independently sampled for every cell, we model correlation between cells with a Gaussian Random Walk (GRW) on $\mathcal{T}$. For the root node r, we sample from an isotropic Gaussian distribution:

$$z_r \sim \text{Normal}(0, I_d), \tag{1}$$

where d denotes the dimension of z_r. Then, every vertex $v \in V \setminus \{r\}$, z_v is sampled according to an isotropic Gaussian distribution centered at its parent's location and with covariance scaled by the edge length:

$$z_v \mid z_{\pi(v)} \sim \text{Normal}\left(z_{\pi(v)}, b_{v,\pi(v)} I_d\right), \tag{2}$$

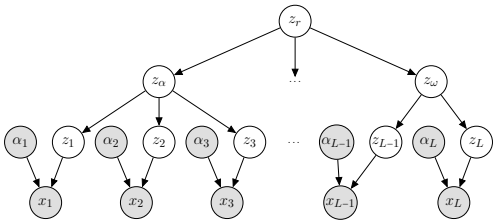

Figure 1. The proposed graphical model. Shaded vertices represent observed random variables. Empty vertices represent latent random variables. Edges signify conditional dependency.

where $\pi(v)$ is the unique parent of v in $\mathcal{T}$.

We observe the scRNA-seq measurements $x_{1:L} = \{x_1, \dots x_L\}$, as well as the library size (i.e., the number of gene counts per cell) $\alpha_{1:L} = \{\alpha_1, \dots \alpha_L\}$ at each of the leaves. We propose to build on an observation model previously used for single-cell transcriptomics data (Lopez et al., 2018) whereby for each cell l and for every gene g, the gene expression x_{lg} is generated as:

$$x_{lg} \sim \text{NegativeBinomial}(\alpha_l \rho_{lg}, \theta_g), \tag{3}$$

where $\rho_{lg} = f^g(z_l)$ is the output of a neural network (f has a softmax output layer for normalization purposes) and θ_g is a gene-wise global parameter learned by variational Bayes.

This model assumes that the correlation between cells is fully-characterized by the GRW as illustrated in Figure 1 and described in the following likelihood decomposition:

$$p_\theta(x_{1:L} \mid \alpha_{1:L}) = \int \prod_{l=1}^{L} p_\theta(x_l \mid z_l, \alpha_l) dp(z_{1:L}). \tag{4}$$

This model is an extension of the SVAE (Johnson et al., 2016), and is a particular instance of the TMC-VAE (Vikram et al., 2019), for which the phylogeny is known *a priori*. We refer to those models as TreeVAEs in this manuscript.

1.2. Inference

As with standard VAEs, the marginal likelihood of (4) is intractable. We therefore develop a variational inference recipe to (i) learn the parameters θ of the model and (ii) approximate the posterior distribution $p(z_{1:L} \mid x_{1:L})$. Because our model couples a complex non-linear observation model with a more simple correlation model in latent space (any marginal distribution for the GRW is tractable), we may build upon previous work (SVAE and TMC-VAE) to derive a variational inference recipe.

We introduce a mean-field variational approximation to the posterior $p(z_{1:L} \mid x_{1:L})$ which we assume factorizes as: $\bar{q_\phi} = \prod_{l=1}^{L} q_\phi(z_l \mid x_l)$, and we derive the evidence lower bound (ELBO):

$$\log p_\theta(x_{1:L}) \geq \mathbb{E}_{\bar{q_\phi}} \left[\sum_{l=1}^{L} \log \frac{p_\theta(x_l \mid z_l)}{q_\phi(z_l \mid x_l)} + \log p(z_{1:L}) \right]. \tag{5}$$

Provided that one can calculate the marginal likelihood $p(z_{1:L})$ of the latent variables over the leaf nodes, and the gradient of the ELBO with respect to the parameters of the variational distribution, one may then learn the parameters of the generative model (θ) and the inference model (ϕ) via stochastic gradient ascent on the ELBO (Kingma & Welling, 2014). However, the complexity of naive calculations of this marginal distribution is cubic in the number of leaves, and therefore unsuitable. Fortunately, the marginalization of these latent variables is tractable in linear time, using a message passing algorithm on the tree (Teh et al., 2008). Although usually those algorithms are only derived for bifurcating trees, we propose here an extension for multifurcating trees. Below, we derive the marginals for the base case of bifurcating trees (a triplet). We describe the general recursive algorithm in Appendix A.

Triplet example Let (x_r, x_a, x_b) be a triplet of random variables, following a local Gaussian diffusion model:

$$x_r \sim \text{Normal}(0, I_d) \tag{6}$$

$$x_a \mid x_r \sim \text{Normal}(x_r, b_{r,a} I_d) \tag{7}$$

$$x_b \mid x_r \sim \text{Normal}(x_r, b_{r,b} I_d). \tag{8}$$

We note as $\phi(\cdot\,; \mu, \sigma^2)$ the probability density function of the multivariate Gaussian distribution with mean μ and covariance $\sigma^2 I$. Completing the square, we have that $p(x_a, x_b \mid x_r)$ identifies as an unnormalized Gaussian density:

$$p(x_a, x_b \mid x_r) = Z \phi(x_r; \mu, \nu), \tag{9}$$

$$\text{with} \quad \nu^{-1} = b_{r,a}^{-1} + b_{r,b}^{-1} \quad \text{and} \quad \mu = \nu \left(\frac{x_a}{b_{r,a}} + \frac{x_b}{b_{r,b}} \right), \tag{10}$$

and a tractable normalization constant Z. The marginal likelihood $p(x_a, x_b)$ simply follows from integrating out the prior $p(x_r)$.

1.3. Posterior Predictive Density

Unlike the vanilla VAE, our TreeVAE model explicitly helps define posterior predictive densities for latent variables z_i and gene expression measurements x_i at each internal node $i \in V_I$ of the phylogeny. On the latent space, the posterior predictive is approximated as:

$$p(z_i \mid x_{1:L}) \approx \int p(z_i \mid z_{1:L}) d\bar{q_\phi}. \tag{11}$$

Similarly, on the feature space, the posterior predictive is approximated as:

$$p(x_i \mid x_{1:L}) \approx \int p_\theta(x_i \mid z_i) dp(z_i \mid z_{1:L}) d\bar{q_\phi}. \tag{12}$$

All quantities appearing in (11) and (12) are readily available after training, except for the conditional distribution $p(z_i \mid z_{1:L})$ for which we use another message passing algorithm, that also has linear time complexity (Appendix A).

2. Performance Benchmarks

To assess the performance of our approach, we evaluate the accuracy of the posterior densities. This is a sensible approach, as we seek to correctly estimate the cellular states at ancestral nodes of the phylogeny. Towards this end, we report several metrics to evaluate the quality of $p_\theta(z_i \mid x_{1:L})$ (on latent space), as well as $p_\theta(x_i \mid x_{1:L})$ (on feature space). Because the ground truth for these densities is generally intractable, we propose three sets of experiments. First, we benchmark the quality of our approximate posterior predictive density on a Gaussian process factor analysis model (Section 2.1), for which the ground truth is tractable. Second, we provide a benchmark on a simulated single-cell RNA sequencing data, using the prior predictive density as (approximate) ground truth (Section 2.2). Finally, we explore the application of TreeVAE to real single-cell data from cancer metastasis (Section 2.3).

For the experiments in Section 2.1 and 2.2, we simulate ground truth tree topologies using a generalized birth-death model. A tree is simulated by beginning with a single node. We use two exponential distributions, parameterized by α and β, to model the time until a cell divides (i.e., birth of a new lineage) and the time until a cell dies, respectively. We repeat the birth-death process until a desired number of leaves is reached.

Throughout, we compare the imputation accuracy of TreeVAE to two baselines. The first one is a naive approach based on averaging gene expression at all the leaves beneath an internal node. The second one is based on averaging the latent space for leaves below an internal node from a fitted VAE before decoding. Such heuristics have been used in several scRNA-seq data analysis scenarios (Lotfollahi et al., 2019). We ran VAE and TreeVAE on a NVIDIA TITAN Xp, and fitting the data took less than a few minutes for every dataset.

2.1. Gaussian Process Factor Analysis Simulations

As a first test-bed, we consider a simulation framework based on the Gaussian process factor analysis model (Yu et al., 2009), for which posterior predictive densities are tractable (fully-specified in Appendix B, with derivations for the ground truth). This model is obtained by using a linear Gaussian conditional distribution in place of the negative binomial decoder in (3).We adapt the observation model of the VAE and the TreeVAE accordingly. Each simulated dataset is based on a tree with 100 leaves (cells) and 100 genes. All results are averaged across ten simulations.

In this setting, we compare the posterior predictive on the latent variables $p_\theta(z_i \mid x_{1:L})$ as well as the one on the features $p_\theta(x_i \mid x_{1:L})$ to their respective ground truth. On the latent space, we report the mean square error of the mean of the posterior predictive $p_\theta(z_i \mid x_{1:L})$ (MSE; lower is better), the k-nearest neighbors purity (Purity; higher is better) (Xu et al., 2021) and the cross entropy between the prior distribution $p_\theta(z_{1:L})$ and the approximate posterior $\bar{q_\phi}$ (CE; lower is better). (To note, latent space metrics do not apply for the average baseline.) On the feature space, we report the average Pearson correlation for imputation of ancestral gene expression across all genes (r; higher is better), as well as the Spearman correlation (ρ; higher is better), and the mean-square error for gene expression (MSE; lower is better).

	LATENT SPACE			FEATURE SPACE		
	MSE	Purity	CE	r	ρ	MSE
Average	-	-	-	0.828	0.803	0.768
VAE	2.28	0.372	2,515	0.846	0.821	0.863
TreeVAE	**1.89**	**0.450**	**281**	**0.869**	**0.844**	**0.541**

Table 1. Results on the Gaussian process factor analysis simulations (averaged across ten different simulations).

	LATENT SPACE		FEATURE SPACE		
	Purity	CE	r	ρ	MSE
Average	-	-	0.350	0.314	7.53
VAE	0.523	8,481	0.402	0.324	5.81
TreeVAE	**0.615**	**1,577**	**0.413**	**0.327**	**5.80**

Table 2. Results on the Gaussian process Poisson Log-normal simulations (averaged across ten different simulations). MSE is reported on normalized counts.

We report the results in Table 1. On latent space, TreeVAE outperforms the VAE with respect to all three metrics. In particular, the MSE metric suggests that TreeVAE learned a more accurate posterior approximation compared to VAE. Moreover, the k-NN purity and cross entropy metric show that the latent space inferred by TreeVAE is more reflective of the phylogeny compared to VAE (as expected). On the feature space, TreeVAE also outperforms baselines on all three metrics. Although all the correlation scores are rather high for all methods in this setting, TreeVAE still provides a substantial improvement over the naive baselines.

2.2. Gaussian Process Poisson Log-normal Simulations

Although the linear Gaussian system described previously helps diagnose the effectiveness of our inference procedure, it remains an unrealistic model for describing scRNA-seq data. Therefore, we instead use a Poisson Log-normal observation model as a more realistic simulation framework. More precisely, in place of the observation model described in (3), we generate x_l as:

$$y_l \mid z_l \sim \text{Poisson}\left(\exp\{W z_l + \beta\}\right) \tag{13}$$

$$x_l \mid y_l \sim \text{Binomial}\left(y_l, p\right), \tag{14}$$

in which p was adjusted to bring the frequency of zeros to 80% in the dataset. Each simulated dataset is based on

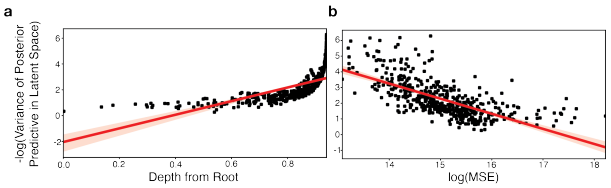

Figure 2. Relationship between uncertainty and error in the Gaussian Process Poisson Log-normal simulation experiments with the TreeVAE model. (a) Variance of posterior predictive density on latent space for each internal node compared to the depth (Pearson's $r = 0.6761$). (b) Error and uncertainty of each prediction (Pearson's $r = -0.6765$).

	CE	ELBO
scVI	21,348	268.43
TreeVAE	**3,051**	**266.09**

Table 3. Results on the cancer metastasis data. ELBO denotes the evidence lower-bound on observed data.

a tree with 500 leaves (cells) and 1000 genes. All results are averaged across ten simulations. Because the ground truth distribution for both posterior predictive densities are not accessible, we slightly modify the evaluation metrics presented previously. First, we do not report MSE on the latent space because it cannot be evaluated. Second, because the gene expression posterior predictive $p_\theta(x_i \mid x_{1:L})$ is no longer tractable, here we use the prior predictive distribution $p_\theta(x_i)$ as a proxy for ground truth. We report the results in Table 2. On the latent space, TreeVAE outperforms the VAE on both metrics. This emphasizes again that the TreeVAE produces representation at internal nodes that are more reflective of the tree structure. On the feature space, we observe that the correlation scores are generally low compared to the previous experiment, likely due to the high level of noise from binomial sub-sampling. However, even in this setting we observe similar trends as from our previous experiment: the TreeVAE outperforms both metrics. Finally, we investigate the relationship between certainty and accuracy in our TreeVAE model. As expected, our estimates become more uncertain closer to the root and this in turn affects model accuracy (Figure 2).

2.3. Analysis of Cancer Metastasis Data

We next assess the performance of the TreeVAE model on real CRISPR/Cas9 single-cell lineage tracing data. As a proof of concept, we analyzed a single clone of 603 cells from a recent dataset that traced the lineages of lung cancer tumors as they metastasized throughout a mouse (Quinn et al., 2021). Here, we leverage the tree reconstructed from CRISPR/Cas9 barcodes with Cassiopeia (Jones et al., 2020) as used in the original study. Because Cassiopeia does not explicitly model the edge weights of the phylogeny, we separately inferred these based on the assumption of an ultrametric tree. Finally, because a large fraction of genes detected

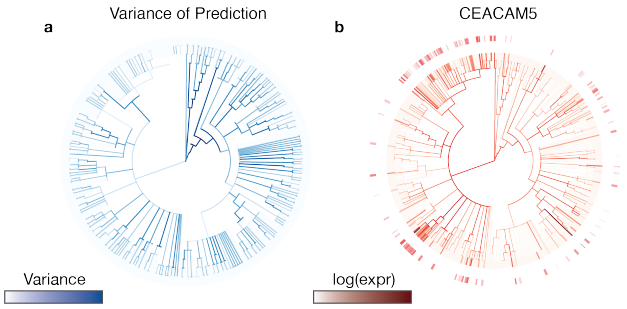

Figure 3. Behavior of TreeVAE internal node predictions. (a) Variance of posterior predictive density on latent space for each internal node. Uncertainty is negatively correlated with depth (Pearson's $r = -0.60$). (b) Predicted expression of *CEACAM5* for each internal node. Color gradient at the leaves indicates observed gene expression.

by scRNA-seq do not have a strong relationship to the lineage, we only considered the top 100 genes autocorrelated with the phylogeny, as evaluated by *Hotspot* (DeTomaso & Yosef, 2021).

We fit the TreeVAE with the observation model described in (3). In this experiment, we have limited ground truth because neither latent variables nor gene expression is known at the ancestral nodes of the phylogeny. We still propose to compare to scVI (Lopez et al., 2018), a VAE with identical observation model as the TreeVAE considered in this experiment. As a first metric, we report the cross-entropy score on the latent space (CE, as in previous experiments). Then, we propose a comparison of the evidence lower bound (ELBO) on observed samples. Although in this experiment we do not have held-out data because we may only observe one cell at each leaf of the tree, we expect these numbers to be comparable because the same neural architecture, noise models, and hyperparameters were used for fitting both models. Our results, reported in Table 3, suggest that TreeVAE better fits the data and proposes an approximate posterior that is more reflective of the tree structure.

Finally, we investigate the behavior of TreeVAE's predictions of ancestral gene expression. As expected, TreeVAE's imputations are more certain closer to the observations at the leaves and becomes more uncertain for nodes closer to the root (Figure 3a). We next predict the ancestral gene expression of *CEACAM5*, an important cell adhesion molecule that is associated with metastatic invasion (Minami et al., 2001). We observe a predicted pattern that broadly agrees with tree structure and observations at the leaves (Figure 3b) and offers rich hypotheses on the subclonal dynamics of *CEACAM5* expression. Critically, because our Bayesian model quantifies uncertainty, we can directly evaluate the stability of a given hypothesis, unlike naive averaging. Overall, these results underscore the promise of the TreeVAE model for gene expression prediction.

References

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Jones, M. G., Khodaverdian, A., Yosef, N., Meissner, A., and Weissman, J. S. Molecular recording of mammalian embryogenesis. *Nature*, 570(7759):77–82, Jun 2019.

DeTomaso, D. and Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Systems*, May 2021.

Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, pp. 2946–2954, 2016.

Jones, M. G., Khodaverdian, A., Quinn, J. J., Chan, M. M., Hussmann, J. A., Wang, R., Xu, C., Weissman, J. S., and Yosef, N. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biology*, 21(1):92, Apr 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053, 2018.

Lopez, R., Gayoso, A., and Yosef, N. Enhancing scientific discoveries in molecular biology with deep generative models. *Molecular Systems Biology*, 16(9):e9198, 2020.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods*, 16 (8):715–721, 2019.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. In *International Conference on Learning Representations*, 2016.

Minami, S., Furui, J., and Kanematsu, T. Role of carcinoembryonic antigen in the progression of colon cancer cells that express carbohydrate antigen. *Cancer Research*, 61 (6):2732–2735, 2001.

Quinn, J. J., Jones, M. G., Okimoto, R. A., Nanjo, S., Chan, M. M., Yosef, N., Bivona, T. G., and Weissman, J. S. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 371(6532), 2021.

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., and Schier, A. F. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36 (5):442–450, May 2018.

Saitou, M. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, July 1987.

Teh, Y. W., Daume III, H., and Roy, D. M. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*, pp. 1473–1480, 2008.

Vikram, S., Hoffman, M. D., and Johnson, M. J. The LORACs prior for VAEs: Letting the trees speak for the data. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Wagner, D. E. and Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, Jul 2020.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 2021.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S., Shenoy, K. V., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in Neural Information Processing Systems*, 2009.

A. Message Passing Algorithms

In the main text of this manuscript, we presented a simple marginalization procedure provided for a triplet of nodes, based on completing the square. The procedure may be used recursively as part of a well-established message passing algorithm for computing marginal likelihood of leaf observations in the context of binary trees, and in particular explained in the TMC-VAE manuscript (Vikram et al., 2019). In this appendix, we present a generic message passing algorithm for multifurcating trees (Section A.1). In particular, we propose a generalization of the normalizing constant formula for multifurcating trees, a key contribution for dealing with phylogenetic information inferred from Cassiopeia and other algorithms that do not necessarily produce binary trees. We then demonstrate how to use this algorithm for computations of the marginal likelihood of the leaves $p(z_{1:L})$ (Section A.2) and for posterior predictive densities $p(z_i \mid z_{1:L})$ (Section A.3).

A.1. The base message passing algorithm

Let $\mathcal{T} = (V, E, b)$ be a phylogeny with vertex set V, edge set E and edge length function b. We note the weight $b(e)$ of an edge $e = (u, v)$ as $b_{u,v}$. Again, the vertex set V is partitioned into L leaf vertices $V_L = \{1, \ldots, L\}$ (a set of cells) and I internal vertices $V_I = \{L+1, \ldots, L+I\}$ (ancestral cell states) such that $V = V_L \cup V_I$. For a node i let $I = i_1, \ldots, i_n$ denote the indices of its n children.

Message passing is defined recursively, starting from a source node s (always an internal node), and requesting messages from each of its neighbors. We initialize the message of each the leaf l with the following content:

$$\log Z_l = 0, \quad \nu_l = 0, \quad \mu_l = z_l, \tag{15}$$

where z_l is the evidence at leaf l.

Then, we use the following update rules to propagate messages from a set of child nodes $(i_1, \ldots, i_n)$ to a parent node i:

$$\nu_i^{-1} = \sum_{j=1}^{n} \frac{1}{\nu_{i_j} + b_{i,i_j}}, \quad \mu_i = \nu_i \sum_{j=1}^{n} \frac{\mu_{i_j}}{\nu_{i_j} + b_{i,i_j}} \tag{16}$$

and for the normalizing constant:

$$\log Z_i = -\frac{d(n-1)}{2} \log(2\pi) - \frac{d}{2} \log T - \frac{1}{2T} \left(\sum_{j \neq l} \left(\prod_{k \in (1,\ldots,n) \setminus (j,l)} (\nu_{i_k} + b_{i,i_k}) \right) \|\mu_{i_j} - \mu_{i_l}\|^2 \right) \tag{17}$$

with $T = \sum_{i=1}^{n} \left(\prod_{j \neq i} (\nu_{i_j} + b_{i,i_j}) \right)$.

A.2. Computing marginals

In order to calculate the marginal likelihood of leaf observations $p(z_{1:L})$, we run the message passing algorithm using the root node as the source, and using the leaf observations $(z_1, \ldots, z_L)$ as local evidence. This is described in Vikram et al. (2019), although our last steps differ from it, as they did not integrate out the prior in their calculations. For this, we compute the integral of the last message with the prior:

$$\log Z_\infty = \log Z_r - \frac{\|\mu_r\|^2}{2(1+\nu_r)} - \frac{d}{2} \log 2\pi(1+\nu_r), \tag{18}$$

where r indicates the root. We then calculate the desired marginal distribution as:

$$\log p(z_{1:L}) = \log Z_\infty + \sum_{l=1}^{L} \log Z_l. \tag{19}$$

A.3. Computing posterior predictive densities

In order to calculate the conditional distribution $p(z_i \mid z_{1:L})$ for an arbitrary internal node i, we run the message passing algorithm using the query node i as the source and using the leaf observations $(z_1, \ldots, z_L)$ as local evidence. In contrast to Vikram et al. (2019), we take into account the prior information during the message passing protocol. To do this, we add a dummy node attached to the root with null evidence and unit diagonal variance.

A.4. Implementation and unit tests

We implemented this message passing information in vanilla PyTorch. In order to check our calculations, we have developed a suite of unit tests for random trees, with ground truth based on the Gaussian conditioning formula.

B. Gaussian Process Factor Analysis Model

In this appendix, we describe the Gaussian process factor analysis model used for simulating data, and define a tractable ground truth for both posterior predictive densities.

B.1. Full specification of the generative model

Let $\tau = (E, V, b)$ be a phylogeny with N nodes. We index the vertex set $V = \mathcal{L} \bigcup \mathcal{I}$ by leaves $\mathcal{L} = \{1 \dots L\}$ and internal nodes $\mathcal{I} = \{L+1 \dots N\}$. As for the TreeVAE model, latent variables z_v for each vertex v form a multivariate Gaussian vector:

$$z_{1:N} \sim \text{Normal}(0, \Sigma), \tag{20}$$

where Σ has for shape $Nd \times Nd$ (the latent space is d-dimensional). For every leaf n, observation x_n is generated as:

$$x_n \mid z_n \sim \text{Normal}\left(Wz_n, \sigma^2 I_p\right), \tag{21}$$

with $W \in \mathbb{R}^{p \times d}$ and $\sigma > 0$ (the feature space is p-dimensional).

B.2. Posterior distribution

The posterior distribution $p(z_{1:L} \mid x_{1:L})$ is tractable via Gaussian conditioning formula. Indeed, it is easy to see that $(z_{1:L}, x_{1:L})$ is a Gaussian vector:

$$\begin{pmatrix} z_{1:L} \\ x_{1:L} \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_L \\ x_1 \\ \vdots \\ x_L \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_L \\ Wz_1 + \sigma e_1 \\ \vdots \\ Wz_L + \sigma e_L \end{pmatrix} \in \mathbb{R}^{(Ld+Lp)}, \tag{22}$$

where $e_{1:L}$ is sampled from a Gaussian isotropic distribution. Consequently, we can characterize the distribution of this vector by its mean and covariance. The random vector is centered, as $z_{1:L}$ and $e_{1:L}$ are both centered. Let us denote the covariance matrix of $(z_{1:L}, x_{1:L})$ by $\Lambda \in \mathbb{R}^{(Ld+Lp)\times(Ld+Lp)}$. We decompose Λ with a block structure:

$$\Lambda = \begin{pmatrix} \Sigma_{z_{1:L}} & \Lambda_{z_{1:L}, x_{1:L}} \\ \Lambda^T_{z_{1:L}, x_{1:L}} & \Lambda_{x_{1:L}} \end{pmatrix}, \tag{23}$$

and where each term can be calculated as follows:

Marginalized latent covariance $\Sigma_{z_{1:L}} \in \mathbb{R}^{(Ld)\times(Ld)}$ is the marginalized covariance Σ of the leaves (computed by adequately slicing Σ).

Marginalized feature covariance $\Lambda_{x_{1:L}} \in \mathbb{R}^{(Lp)\times(Lp)}$ such that for all pairs of leaves (i, j), $\Lambda^{i,j}_{x_{1:L}} \in \mathbb{R}^{p\times p}$ is the block encoding correlations between x_i and x_j:

$$\Lambda^{k,l}_{x_{1:L}} = \text{Cov}(x_k, x_l) = \text{Cov}(Wz_k + e_k, Wz_l + e_l) = W\Sigma_{k,l}W^T + \sigma^2 I_p. \tag{24}$$

Correlation term The matrix $\Lambda_{z_{1:L}, x_{1:L}} \in \mathbb{R}^{(Ld)\times(Lp)}$ encodes correlations between $z_{1:L}$ and $x_{1:L}$ such that:

$$\Lambda_{z_{1:L}, x_{1:L}} = \begin{pmatrix} \Sigma_{1,1}W^T & \dots & \Sigma_{1,L}W^T \\ \vdots & \ddots & \vdots \\ \Sigma_{L,1}W^T & \dots & \Sigma_{L,L}W^T \end{pmatrix} = \{\Sigma_{k,l}W^T\}^L_{k,l=1}, \tag{25}$$

where the previous result comes from the simple fact that for a pair of leaves (k, l), we have:

$$\mathrm{Cov}(z_k, x_l) = \mathbb{E}\left[z_k (W z_l + e_l)^T\right] = \Sigma_{k,l} W^T, \tag{26}$$

where $\Sigma_{k,l}$ is the marginalized Σ corresponding to the correlations between variables z_k and z_l.

Finally, we may use the Gaussian conditioning formulas to derive the posterior distribution $p(z_{1:L} \mid x_{1:L})$:

$$\mathbb{E}\left[z_{1:L} | x_{1:L}\right] = \Lambda_{z_{1:L}, x_{1:L}} \left(\Lambda_{x_{1:L}}^{-1}\right) x_{1:L} \tag{27}$$

$$\mathrm{Var}\left[z_{1:L} | x_{1:L}\right] = \Sigma_{z_{1:L}} - \Lambda_{z_{1:L}, x_{1:L}} \left(\Lambda_{x_{1:L}}^{-1}\right) \Lambda_{z_{1:L}, x_{1:L}}^T \tag{28}$$

B.3. Posterior predictive densities at internal nodes

In all our imputation experiments, we are interested in the posterior predictive density of the internal nodes. For both quantities, we integrate over a suitable set of latent variables. On latent space, we utilize the following decomposition:

$$p(z_i \mid x_{1:L}) = \int p(z_i \mid z_{1:L}) dp(z_{1:L} \mid x_{1:L}), \tag{29}$$

where $p(z_i \mid z_1 \dots z_L)$ may be computed exactly using the message passing algorithm (as in Section A.3) for any internal node i. On feature space, we similarly proceed and integrate out latent variables:

$$p(x_i \mid x_{1:L}) = \int p(x_i \mid z_i) dp(z_i \mid x_{1:L}). \tag{30}$$

Therefore, we can efficiently compute the posterior predictive of any internal node through MCMC sampling in two steps. First, we sample from the posterior predictive distribution on latent space at node i: $p(z_i \mid x_{1:L})$. Then, we generate $x_i \mid z_i$ with the generative model specified in Section B.1.