Multinomial belief networks for healthcare data

H. C. Donker H.C.DONKER@UMCG.NL
Department of Epidemiology
University Medical Center Groningen
University of Groningen, the Netherlands.

D. Neijzen D.NEIJZEN@UMCG.NL
Department of Epidemiology
University Medical Center Groningen
University of Groningen, the Netherlands.

J. de Jong JOHANN.DE_JONG@BOEHRINGER-INGELHEIM.COM
Statistical Modeling
Global Computational Biology and Digital Sciences
Boehringer Ingelheim Pharma GmbH & Co. KG
Biberach an der Riß, Germany.

G. A. Lunter G.A.LUNTER@UMCG.NL
Department of Epidemiology
University Medical Center Groningen
University of Groningen, the Netherlands.

Abstract

Healthcare data from patient or population cohorts are often characterized by sparsity, high missingness and relatively small sample sizes. In addition, being able to quantify uncertainty is often important in a medical context. To address these analytical requirements we propose a deep generative Bayesian model for multinomial count data. We develop a collapsed Gibbs sampling procedure that takes advantage of a series of augmentation relations, inspired by the Zhou–Cong–Chen model. We visualise the model's ability to identify coherent substructures in the data using a dataset of handwritten digits. We then apply it to a large experimental dataset of DNA mutations in cancer and show that we can identify biologically meaningful clusters of mutational signatures in a fully data-driven way.

1. Introduction

Healthcare data is expensive and limited. In addition, events of interest may be infrequent, and high missingness is a feature of most practical data sets. These features, along with the need to handle uncertainty, present challenges to traditional maximum likelihood-based machine learning methodologies, which often give rise to biased results, as highlighted by several studies (Smith and Naylor, 1987; Beerli, 2005; Alzubaidi et al., 2023). In addition, the predictions from these methodologies are often overconfident on out-of-distribution data and fail to adequately address situations where data incompleteness or uncertainty plays an important role (Acharya et al., 2015; Emmanuel et al., 2021; Murphy, 2023).

As was argued persuasively in Papamarkou et al. (2024), an approach that promises to overcome these limitations is to use generative and fully Bayesian methods. Examples

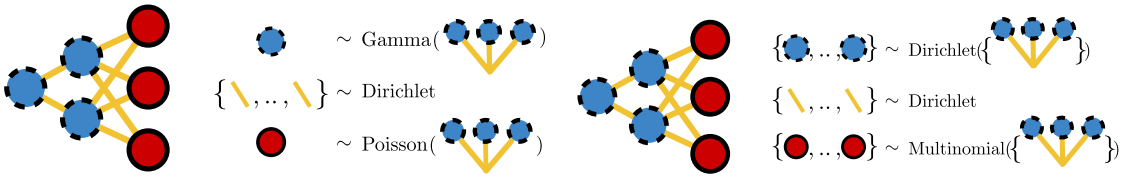


Figure 1: Schematic representation of the two belief networks. Red nodes are observations, blue dashed circles are latent hidden units, and edges are latent weights.

include latent Dirichlet (LDA) allocation (Blei et al., 2003), deep belief nets (Hinton et al., 2006; Zhou et al., 2016), and variational autoencoders (Kingma and Welling, 2014). In principle, Bayesian methods are data efficient, guard against overtraining, account for uncertainty, and deal with missing data in a principled way. However, current implementations fall short of this ideal. One class of methods uses variational approaches to enable learning, which involves approximations such as fixing the posterior form beforehand or by making a mean field assumption (Ranganath et al., 2015, 2016; Ferreira et al., 2022; Soleimani et al., 2017; Murphy, 2023). Other methods use exact sampling but are limited in their representational power by using binary variables, a shallow architecture, or use Poisson-distributed variables as (intermediate) output (Welling et al., 2004; Hinton et al., 2006; Zhou et al., 2016; Zhao et al., 2018; Panda et al., 2019; Donker and Groen, 2021).

In this paper, we tailor Bayesian belief networks to healthcare data by choosing a multinomial distribution for the output variable (Fig. 1). The versatility of the multinomial distribution makes it well-suited for modelling data types commonly found in healthcare data, from categorical variables found in patient questionnaires to text documents (Griffiths and Steyvers, 2004) and DNA mutations in cancer. Using augmentation techniques real-valued, ordinal, and survival (i.e., censored) data can be modelled as well (note that "augmentation" refers to the addition of new hidden variables to support inference, rather than augmentation of training data). Importantly, it is straightforward to model several multinomials simultaneously (each with their own dimension and observation count) which naturally enables modelling of heterogeneous data, while missingness by setting relevant observation counts to 0. We will develop those extensions elsewhere; here we focus on modelling a single multinomial-distributed observation.

A prominent example of a model with multinomial output variables is LDA (Blei et al., 2003). In one representation of the model, the latent parameter matrix of the multinomial distributions across samples is factorized as a low-rank product of (sample-topic and topic-feature) matrices whose rows are drawn from Dirichlet distributions. The low-rank structure, the sparsity induced by the Dirichlet priors, and the existence of effective inference algorithms have resulted in numerous applications and extensions of LDA. Despite this success, LDA has some limitations. One is that inference of the Dirichlet hyperparameters is often ignored or implemented using relatively slow maximum likelihood methods (Minka, 2003; George and Doss, 2017). This was elegantly addressed by Teh et al. (2006) by endowing the Dirichlet distribution with another Dirichlet prior in a hierarchical structure, allowing information to be borrowed across samples. Another limitation is that LDA ignores

any correlation structure among the topic weights across samples, which for higher latent (topic) dimensions becomes increasingly informative. An effective approach that addresses this issue was developed by Zhou et al. (2016), who developed a multi-layer fully-connected Bayesian network using gamma variates and Poisson, rather than multinomial, observables. Here we combine and extend these two approaches in the context of multinomial observations, resulting in a model whose structure resembles a fully connected multi-layer neural network but retains the efficient inference properties of LDA. This paper is structured as follows. In section 2 we review Zhou–Cong–Chen's Poisson-gamma belief network (PGBN) and subsequently introduce our model in section 3. We apply the model to handwritten digits and mutations in cancer in Sec. 4. We end with conclusions and a discussion (Sec. 5).

Generalizable Insights about Machine Learning in the Context of Healthcare

We introduce a Bayesian machine learning model, the multinomial belief network (MBN), to analyze healthcare data by decomposition. This unsupervised approach is well suited for analysing 'omics data such as gene expression and mutation profiles or to model the heterogeneity in clinical presentation, risk factors, and the underlying disease mechanisms of patient populations. Compared to e.g. non-negative matrix factorisation, this method is less prone to overtraining, and inferences come with uncertainty estimates. Unlike traditional topic modelling methods, MBNs can capture topic interactions across multiple layers. Initiatives like the 100k genomes project, which aim to integrate signature analysis into standard patient care (Everall et al., 2023), highlight the need for reliable deconvolution and uncertainty quantification to support treatment decisions.

2. Poisson gamma belief network

2.1. Generative model

We first review the Gamma belief network of Zhou et al. (2016). The backbone of the model is a stack of Gamma-distributed hidden units $\theta_{vj}^{(t)}$ (K_t per sample j), where the last unit parameterizes a Poisson distribution generating observed counts x_{vj}, one for each sample j and feature v. The generative model is

$$
\begin{aligned}
a_{vj}^{(T+1)} &= r_v \\
\theta_{vj}^{(t)} &\sim \mathrm{Gam}(a_{vj}^{(t+1)}, c_j^{(t+1)}), && t = T, \ldots, 1 \\
a_{vj}^{(t)} &= \sum_{k=1}^{K_t} \phi_{vk}^{(t)} \theta_{kj}^{(t)}, && t = T, \ldots, 1 \\
x_{vj} &\sim \mathrm{Pois}(a_{vj}^{(1)}).
\end{aligned}
$$

For $T = 1$ we only have one layer, and the model reduces to Poisson Factor Analysis, $x_{vj} = \mathrm{Pois}([\boldsymbol{\phi}\boldsymbol{\theta}]_{vj})$ (Zhou et al., 2012). For multiple layers, the features $\boldsymbol{\theta}^{(t+1)}$ on layer $t+1$ determine the shape parameters of the gamma distributions on layer t through a non-negative connection weight matrix $\boldsymbol{\phi}^{(t+1)} \in \mathbb{R}_+^{K_t \times K_{t+1}}$, so that $\boldsymbol{\phi}^{(t+1)}$ induces correlations between features on level t. The rate parameter for $\boldsymbol{\theta}^{(t)}$ is $c_j^{(t)} \sim \mathrm{Gam}(e_0, f_0)$, one for each

sample j and layer t; for $t > 1$ the $c_j^{(t)}$ act as concentration parameters for the activations below. The weights $\boldsymbol{\phi}^{(t)}$ that connect latent states between layers are normalised as $\sum_v \phi_{vk}^{(t)} = 1$ owing to their Dirichlet priors $\phi_{vk}^{(t)} \sim \mathrm{Dir}(\{\eta_v^{(t)}\}_v)$; here we use curly braces to denote vectors, with the subscript indicating the index variable; we drop the subscript if the index variable is unambiguous. The top-level activation is controlled by $r_v \sim \mathrm{Gam}(\gamma_0/K_T, c_0)$ where hyperparameters γ_0 and c_0 determine the typical number and scale, respectively, of active top-level hidden units. The lowest-level activations $\boldsymbol{a}^{(1)}$ parameterize Poisson distributions that generate the observed count variables x_{vj} for sample (individual, observation) j. This completes the specification of the generative model.

This model architecture is similar to a T-layer neural network, with activations $\boldsymbol{a}^{(t)}$ representing the activity of features (topics, factors) of increasing complexity as t increases.

2.2. Deep Poisson representation

An alternative and equivalent representation is obtained by integrating out the hidden units $\boldsymbol{\theta}$ and augmenting with a sequence of latent counts $\boldsymbol{x}^{(t+1)} \to \boldsymbol{m}^{(t)} \to \boldsymbol{y}^{(t)} \to \boldsymbol{x}^{(t)}$. Specifically, let $\mathrm{Log}(p)$ be the logarithmic distribution, with probability mass function $\mathrm{Log}(k; p) \propto p^k/k$ where $0 < p < 1$, and define $n \sim \mathrm{SumLog}(l, p)$ by $n = \sum_{i=1}^{l} u_i$ where each $u_i \sim \mathrm{Log}(p)$. Henceforth, underlined indices denote summation, so that $x_{\underline{j}} := \sum_j x_j$. Augmenting each layer with counts $x_{kj}^{(t)} \sim \mathrm{Pois}(q_j^{(t)} a_{kj}^{(t)})$, it turns out these can be generated as

$$
\begin{aligned}
&m_{jk}^{(t)} \sim \mathrm{SumLog}(x_{kj}^{(t+1)}, 1 - e^{-q_j^{(t+1)}});\\
&\{y_{vjk}^{(t)}\}_v \sim \mathrm{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v);\\
&x_{vj}^{(t)} := y_{vj\underline{k}}^{(t)},
\end{aligned}
$$

where $q_j^{(t+1)} = \ln[1 + q_j^{(t)}/c_j^{(t+1)}]$; see Zhou et al. (2016) and Supplementary Material sections S2.2, S2.3. Starting with $t = T$ this shows how to eventually generate the observed counts $x_{vj}^{(1)}$ using only count variables, with all $\boldsymbol{\theta}^{(t)}$ integrated out. The two alternative schemes are shown graphically in Fig. 2.

2.3. Inference

At a high level, inference consists of repeatedly moving from the first representation to the second, and back. This is achieved by swapping, layer-by-layer, the direction of the arrows to sample upward $\boldsymbol{x}^{(t)} \to \boldsymbol{y}^{(t)} \to \boldsymbol{m}^{(t)} \to \boldsymbol{x}^{(t+1)}$; these counts are then used to sample $\boldsymbol{\phi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$, after which the procedure starts again. To propagate latent counts upwards, we use the identity from Theorem 1 of Zhou and Carin (2015):

$$
\mathrm{Pois}(x|qa)\,\mathrm{SumLog}(m|x, 1 - e^{-q}) = \mathrm{NB}(m|a, 1 - e^{-q})\,\mathrm{CRT}(x|m, a), \tag{1}
$$

to turn $\boldsymbol{m} \to \boldsymbol{x}$ into $\boldsymbol{x} \to \boldsymbol{m}$; here $\mathrm{CRT}(x|m, a)$ is the number of occupied tables in a Chinese restaurant table distribution over m customers with concentration parameter a, and $\mathrm{NB}(k|r, p)$ is the negative-binomial distribution with r successes of probability p. Finally, we use that independent Poisson variates conditioned on their sum are multinomially distributed with probabilities proportional to the individual Poisson rates, to convert $\boldsymbol{y} \to \boldsymbol{x}$ to

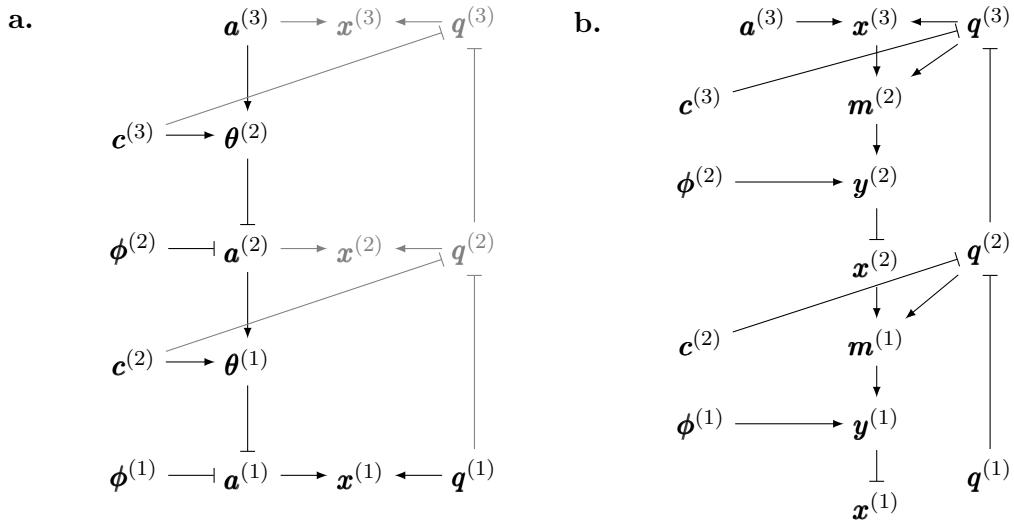


Figure 2: Two equivalent generative models for a count variable $\boldsymbol{x}^{(1)}$ from the Poisson gamma belief network, using a tower of (**a**) real-valued latent variables $\boldsymbol{\theta}$, $\boldsymbol{a}$, or (**b**) latent counts $\boldsymbol{m}$, $\boldsymbol{y}$, $\boldsymbol{x}$. Blunt arrows indicate deterministic relationships. The variable $\boldsymbol{q}^{(1)}$ is a dummy and has a fixed value 1. In representation (**a**) the grayed-out counts $\boldsymbol{x}^{(t)}$ and variables $\boldsymbol{q}^{(t)}$, $t > 1$, are included for clarity (and have the same distribution as the variables in the right model) but are not used to generate the outcome $\boldsymbol{x}^{(1)}$, and so can be marginalized out.

$\boldsymbol{x} \to \boldsymbol{y}$ as well as $\boldsymbol{m} \to \boldsymbol{y}$ to $\boldsymbol{y} \to \boldsymbol{m}$. To Gibbs sample the variables, we first make an upward pass from $\boldsymbol{x}^{(1)} \to \cdots \to \boldsymbol{x}^{(T+1)}$ followed by a downward pass where the multinomial-Dirichlet conjugacy is used to update $\boldsymbol{\phi}^{(t)}$, the gamma-gamma rate conjugacy to update $\boldsymbol{c}^{(t)}$ and the Poisson-gamma conjugacy to update $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{r}$. Details are provided in the supplement.

3. Multinomial belief network

3.1. Generative model

We now introduce the multinomial belief network (MBN). To model multinomial observations, we replace Poisson observables with multinomials, and we replace the gamma-distributed hidden activations with Dirichlet-distributed weights $\{\theta_{vj}^{(t)}\}_v$. The generative model is

$$a_{vj}^{(T+1)} = r_v, \tag{2}$$

$$\{\theta_{vj}^{(t)}\}_v \sim \mathrm{Dir}(\{c^{(t+1)} a_{vj}^{(t+1)}\}_v), \qquad t = T, \ldots, 1 \tag{3}$$

$$a_{vj}^{(t)} = \sum_{k=1}^{K_t} \phi_{vk}^{(t)} \theta_{kj}^{(t)}, \qquad t = T, \ldots, 1 \tag{4}$$

$$\{x_{vj}\}_v \sim \mathrm{Mult}(n_j, \{a_{vj}^{(1)}\}_v). \tag{5}$$

As before, the weights are Dirichlet distributed $\phi_{vk}^{(t)} \sim \mathrm{Dir}(\{\eta_v^{(t)}\}_v)$ with hyperparameters $\{\eta_v^{(t)}\}_v$. Different from the PGBN model we choose one $c^{(t)} \sim \mathrm{Gam}(e_0, f_0)$ per dataset (and per layer) instead of one per sample j, reducing the number of free parameters per sample, and allowing the variance across samples to inform the $c^{(t)}$. Finally we let the top-level activations r_v be Dirichlet distributed, $\{r_v\}_v \sim \mathrm{Dir}(\{\gamma_0 / K_T\}_v)$ with γ_0 a hyperparameter. This completes the definition of the generative model.

3.2. Deep multinomial representation

Integrating out $\boldsymbol{\theta}^{(t)}$, the generative model can be alternatively represented as a deep multinomial factor model, as follows (Supplementary Material, Sec. S3.1,):

$$\begin{aligned} n_j^{(t+1)} &\sim \mathrm{CRT}(n_j^{(t)}, c^{(t+1)}); \\ \{x_{kj}^{(T+1)}\}_k &\sim \mathrm{Mult}(n_j^{(T+1)}, \{r_k\}); \\ \{m_{kj}^{(t)}\}_k &\sim \mathrm{Polya}(n_j^{(t)}, \{x_{kj}^{(t+1)}\}_k); \\ \{y_{vjk}^{(t)}\}_v &\sim \mathrm{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v); \\ x_{vj}^{(t)} &= y_{vj\underline{k}}^{(t)}. \end{aligned} \tag{6}$$

Here underlined subscripts denote summation; and $\mathrm{Polya}(n, \{y_k\})$ is the distribution of the contents of an urn after running a Polya scheme: starting with y_k balls of color k, repeatedly drawing a ball, returning the drawn ball and a new identically colored one each time, until the urn contains n balls. The two representations of the model are structurally identical to the two representations of the PGBN shown in Fig. 2, except that the $q^{(t)}$ are replaced by $n^{(t)}$, and the relationship between successive $n^{(t)}$ is stochastic instead of deterministic.

3.3. Inference

Similar to the PGBN, we reverse the direction of $\boldsymbol{x}^{(t+1)} \to \boldsymbol{m}^{(t)} \to \boldsymbol{y}^{(t)} \to \boldsymbol{x}^{(t)}$ to propagate information upward. To reverse $\boldsymbol{x}^{(t+1)} \to \boldsymbol{m}^{(t)}$ into $\boldsymbol{m}^{(t)} \to \boldsymbol{x}^{(t+1)}$ we use the following:

Theorem 1 *The joint distributions over n, $\{x_k\}$ and $\{m_k\}$ below are identical:*

$$\mathrm{DirMult}(\{m_k\}|n_0, \{ca_k\}) \left[\prod_k \mathrm{CRT}(x_k|m_k, ca_k)\right] \delta_{n,x_{\underline{k}}} =$$
$$\mathrm{CRT}(n|n_0, c)\,\mathrm{Mult}(\{x_k\}|n, \{a_k\})\,\mathrm{Polya}(\{m_k\}|n_0, \{x_k\}).$$

(For the proof see Sec. S1.1, Supplementary Material.) Here $\mathrm{DirMult}(\{x_k\}|n, \{ca_k\})$ is the Dirichlet-multinomial distribution of n draws with concentration parameters $\{ca_k\}$, and $\delta_{i,j}$ denotes the Kronecker delta function that is 1 when $i = j$ and zero otherwise; here it expresses that $n = x_{\underline{k}}$. Note that $a_{\underline{k}} = 1$, as it parameterizes a multinomial.

In words, the theorem states that observing $\{m_k\}$ from a multinomial parameterized by probabilities from a Dirichlet distribution that itself has parameters $\{c\alpha_k\}$, provides information about the probabilities $\{a_k\}$ through an (augmented) multinomial-distributed observation $\{x_k\}$, and information about the concentration parameter c through a CRT-distributed observation $n = x_{\underline{k}}$. Therefore, by augmenting with $\{x_k\}$ and choosing conjugate priors to the multinomial and CRT distributions for $\{a_k\}$ and c respectively, we can obtain posteriors for these parameters.

The remaining arrows can be swapped by augmenting and marginalizing multinomial distributions. Taken together, to sample in the MBN we use augmentation and the Dirichlet-multinomial conjugacy to update $\boldsymbol{\phi}^{(t)}$, $\boldsymbol{\theta}^{(t)}$, and $\boldsymbol{r}$, while to update $c^{(t)}$ we sample from the Chinese restaurant table conjugate prior $\mathrm{CRTCP}(\alpha|m, \{n_j\}_j, a, b) \propto \mathrm{Gam}(\alpha|a, b)\alpha^m \prod_j \frac{\Gamma(\alpha)}{\Gamma(\alpha+n_j)}$ using the method described by Teh et al. (2006). In more detail, sampling proceeds as follows. Identifying $x_{vj}^{(1)} \equiv x_{vj}$, $n_j^{(1)} \equiv x_{\underline{v}j}$ and $a_{kj}^{(T+1)} \equiv r_v$:

Algorithm 3.1 Procedure to Gibbs sample weights $\boldsymbol{\phi}^{(t)}$, hidden units $\boldsymbol{\theta}^{(t)}$, concentration parameters $c^{(t)}$, and activations $\boldsymbol{r}$ of an MBN given training data $x_{vj}^{(1)}$.

For $t = 1, \ldots, T$:
  Sample $\{y_{vjk}^{(t)}\}_k \sim \mathrm{Mult}(x_{vj}^{(t)}, \{\phi_{vk}^{(t)}\theta_{kj}^{(t)}\}_k)$
  Compute $m_{jk}^{(t)} = y_{\underline{v}jk}^{(t)}$,
  Sample $x_{kj}^{(t+1)} \sim \mathrm{CRT}(m_{jk}^{(t)}, c^{(t+1)}a_{kj}^{(t+1)})$;
  Compute $n_j^{(t+1)} = x_{\underline{v}j}^{(t+1)}$,
  Sample $\{\phi_{vk}^{(t)}\}_v \sim \mathrm{Dir}(\{\eta_v^{(t)} + y_{v\underline{j}k}^{(t)}\}_v)$,
Sample $\{r_v\} \sim \mathrm{Dir}(\{\gamma_0/K_T + x_{v\underline{j}}^{(T+1)}\}_v)$
For $t = T, \ldots, 1$:
  Sample $\{\theta_{kj}^{(t)}\}_k \sim \mathrm{Dir}(\{c^{(t+1)}a_{kj}^{(t+1)} + m_{jk}^{(t)}\}_k)$
  Sample $c^{(t+1)} \sim \mathrm{CRTCP}(n_{\underline{j}}^{(t+1)}, \{n_j^{(t)}\}_j, e_0, f_0)$

In practice, sampling might proceed per observation j, and resampling of global parameters $\boldsymbol{\phi}^{(t)}$ and $c^{(t)}$, which involve summing over j, is done once all observations have been processed. For details see Supplement section S3.

4. Experiments

4.1. Performance evaluation

Having reviewed the PGBN and introduced our model, we now illustrate its application on small images of handwritten digits and on DNA point mutations in cancer. To evaluate performance, we hold out 50% of the pixel quanta (resp. mutations) from the images (resp. patients) to form a test set, $\mathbf{x}^{\text{test}}$, and evaluate the held-out perplexity as:

$$\mathcal{L}(\mathbf{x}^{\text{test}}) = \exp\left(-\frac{1}{J}\sum_{j=1}^{J}\sum_{v=1}^{V}\frac{x_{vj}^{\text{test}}\ln p_{vj}}{x_{\underline{v}j}^{\text{test}}}\right), \tag{7}$$

where p_{vj} is the probability of feature v in example j. For non-negative matrix factorisation (NMF) trained on a Kullback-Leibler (KL) loss (which is equivalent to a Poisson likelihood, Lee and Seung (1999)), the probability $p_{vj} = a_{vj}/a_{\underline{v}j}$ is the training set reconstruction $a_{vj} \equiv \sum_k \phi_{vk}\theta_{kj}$ (so that $a_{vj} \approx x_{vj}$) normalised across features v.

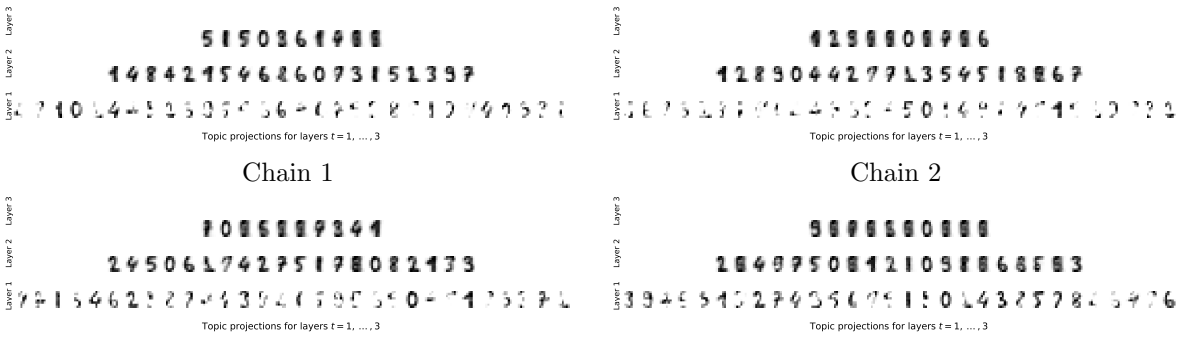
Similarly, $p_{vj} = \sum_{\sigma=1}^{S} a_{vj}^{(\sigma)}/a_{\underline{v}j}^{(\sigma)}$ for the PGBN and MBN where $a_{vj} = \sum_{k=1}^{K_1}\phi_{vk}^{(1)}\theta_{kj}^{(1)}$ is the bottom layer activation normalised and averaged over S posterior samples $\sigma = 1, \ldots, S$ (for the MBN a_{vj} is normalised by construction) similar to Zhou et al. (2012); Ranganath et al. (2015).

Unlike in Zhou et al. (2016), for the PGBN we use a gamma distribution to model the scale $c_j \sim \text{Gam}(e_0, f_0)$ for all layers, instead of a separate beta-distributed $p_j^{(2)}$ to set $c_j^{(2)} = p_j^{(2)}/(1-p_j^{(2)})$ for the first layer only. In addition, we consider γ_0 a fixed hyperparameter for both belief networks. For each experiment, four Markov chains were initialised using the prior to ensure overdispersion relative to the posterior (as suggested in Gelman et al. (2013)). Each chain was run in parallel on a separate nVidia A40 device.

4.2. UCI ML handwritten digits

We considered the Optical Recognition of Handwritten Digits dataset from the UC Irvine Machine Learning Repository (Alpaydin and Kaynak, 1998) containing 1797 images of handwritten digits. Each pixel in the 8×8 images had a discrete intensity ranging from 0 to 15 which we modelled as counts. Three separate models ($[K_1, K_2, K_3] = [30, 20, 10]$ latent components; $\gamma_0 = e_0 = f_0 = 1$, $\eta = 0.05$) with one through three layers were run for 10^5 Gibbs burn-in steps and 1280 samples were collected from each chain. In terms of likelihood, all three MBN models converged within 5000 iterations of burn-in. Extensive simulation-based calibration tests (Talts et al., 2018) for various network configurations on small datasets (we tested up to $J = 19$ examples) indicate that our sampler was correctly implemented.

The perplexity on holdout pixel intensity quanta were $31.0^{+0.1}_{-0.1}$, $30.7^{+0.1}_{-0.1}$, and $30.7^{+0.1}_{-0.1}$ for one to three layers (lower is better, bootstrapped 95% confidence intervals), respectively.



Chain 1 Chain 2

Chain 3 Chain 4

Figure 3: Hierarchy of topics learned by a three-layer MBN (with $[K_1, K_2, K_3] = [30, 20, 10]$ latent components) after training on the Optical Recognition of Handwritten Digits dataset from the UC Irvine Machine Learning Repository (Alpaydin and Kaynak, 1998). Topics are represented by their projection $\prod_{l=1}^{t} \boldsymbol{\phi}^{(l)}$ onto the pixels. Separate panels refer to individual Markov chains that were run in parallel.

Note that latent Dirichlet allocation is a special case of the MBN with one layer [see Eqs. (3-5)]. The relatively modest improvement with depth is typical for these models and also observed e.g. in the PGBN on 20 newsgroup data (Zhou et al., 2016) For comparison, we trained NMF (from Sci-kit learn (Pedregosa et al., 2011)) using a KL loss and achieved a significantly larger (worse) perplexity of $34.2^{+0.3}_{-0.3}$ [1]. Importantly, the digits highlight the interpretability of the network, which hierarchically learns topics from the specific to the general (Fig. 3). The sum-of-parts representations enforced by the network cause the lowest layer in the network to learn digit patches. These are then combined in higher layers to form increasingly general digit representations (Fig. 3).

4.3. Mutational signature attribution

Next, we turned to mutational signatures imprinted in the DNA of cancer cells. Mutational signatures arise from observed patterns of DNA mutations that are influenced by various factors including mutagenesis, DNA damage sensing, repair pathways and chromatin context (Singh et al., 2020). To compute mutational signatures, observed DNA mutation profiles are typically first summarized into a 96-feature vector of counts $\boldsymbol{x}$ of different types of mutations (Alexandrov et al., 2013), and then approximately factorized into "mutational signatures" $\boldsymbol{\phi}$ and sample-specific attributions of mutations to each signature $\boldsymbol{\theta}$. The signatures ϕ_{vk}^{COSM} reported in the COSMIC database (Tate et al., 2019; Cosmic, 2023) are the *de facto* standard in the field, relating mutation spectra across $v = 1, \ldots, 96$ possible point mutations types to $k = 1, \ldots, 78$ signatures named SBS1, . . . ,SBS94. Attributions based on this curated set of signatures are increasingly used to guide therapeutic decisions

1. Technically, the perplexity of NMF was infinite because zero probability was assigned to non-zero intensity. Samples where NMF attributed zero probability were removed from the perplexity calculation.

Table 1: Perplexity of held-out mutations for inferred mutational signature attributions. Signatures were based on COSMIC v3.3 signature weights. Super/subscripts indicate ninety-five per cent confidence intervals, computed by bootstrapping.

Method		**Hold out perplexity**
SigProfilerExtractor[1]		$64.5^{+0.7}_{-0.7}$
Zhou–Cong–Chen	(1 layer)	$62.0^{+0.7}_{-0.7}$
	(2 layers)	$61.9^{+0.7}_{-0.7}$
This work	(1 layer)	$62.0^{+0.7}_{-0.7}$
	(2 layers)	$61.9^{+0.7}_{-0.7}$

in cancer (Brady et al., 2022; Patterson et al., 2023), underscoring the need for accurate attribution and quantification of uncertainty.

In short, our goal was to infer for each patient the proportion of mutations, $\boldsymbol{\theta}$, corresponding to specific signatures ($\boldsymbol{\phi}^{\text{COSM}}$, COSMIC v3.3) given their mutation profiles. We tested our model on the mutation dataset of Alexandrov et al. (2020) comprising ∼85 million mutations from 4,645 patients and compared with SigProfilerExtractor, considered the state of art for de novo extraction of mutational signatures (Islam et al., 2022), and the Zhou–Cong–Chen model. Since we expected around 5–10 signatures to be present per sample, we set $\gamma_0 = 10$ and other hyperparameters $\eta = e_0 = f_0 = 1$ for both the Zhou–Cong–Chen and our model and used the greedy layer-wise training procedure (Sec. S4.1, Supplementary Material). Although the test-set likelihood indicated that the chains of both models had not yet fully converged, we halted computation due to the large computation time (a total of 77 and 78 GPU days for MBN and PGBN, respectively). Since the chains were initialised with overdispersed values (compared to the posterior), pre-mature termination of the Markov chains overestimates between-chain variance compared to the "true" posterior. That is, our uncertainty estimates are conservative. Nevertheless, both the Zhou–Cong–Chen and our model more accurately attribute mutations than SigProfilerExtractor (Table 1). As expected, both belief networks score comparably with similar architecture.

Next, we constructed robust consensus meta-mutational signatures (i.e., topics $\phi^{(2)}_{vk}$ from the second layer) from the MBN (Appendix S4.2) that capture co-occurrence of mutational signatures in patients. This resulted in four meta-signatures denoted $k = \text{M}_1, \ldots, \text{M}_4$ (Fig. 4).

In brief, the following co-occurrence patterns were identified (an in-depth analysis is provided in the Supplementary Material, Sec. S4.3). M_1 describes the combination of replicative DNA polymerase ϵ (POLE) damage and mismatch-repair deficiency (MMR) (Fig. 4, first row, left column). Tumours with an ultra-hypermutated phenotype (≥ 100 mutations Mb^{-1}) are often characterised by these joint disruptions in MMR and POLE (Hodel et al., 2020). Meta-signature M_2 primarily captures, presumably, oxidative stress. To a lesser extent, M_2 also captures a signature implicated in BRCA1 and BRCA2 dysfunction in breast cancer (Nik-Zainal et al., 2016) which are believed to originate from (uncorrected) replication errors (Singh et al., 2020). Meta-signature M_3 is marked by components with a

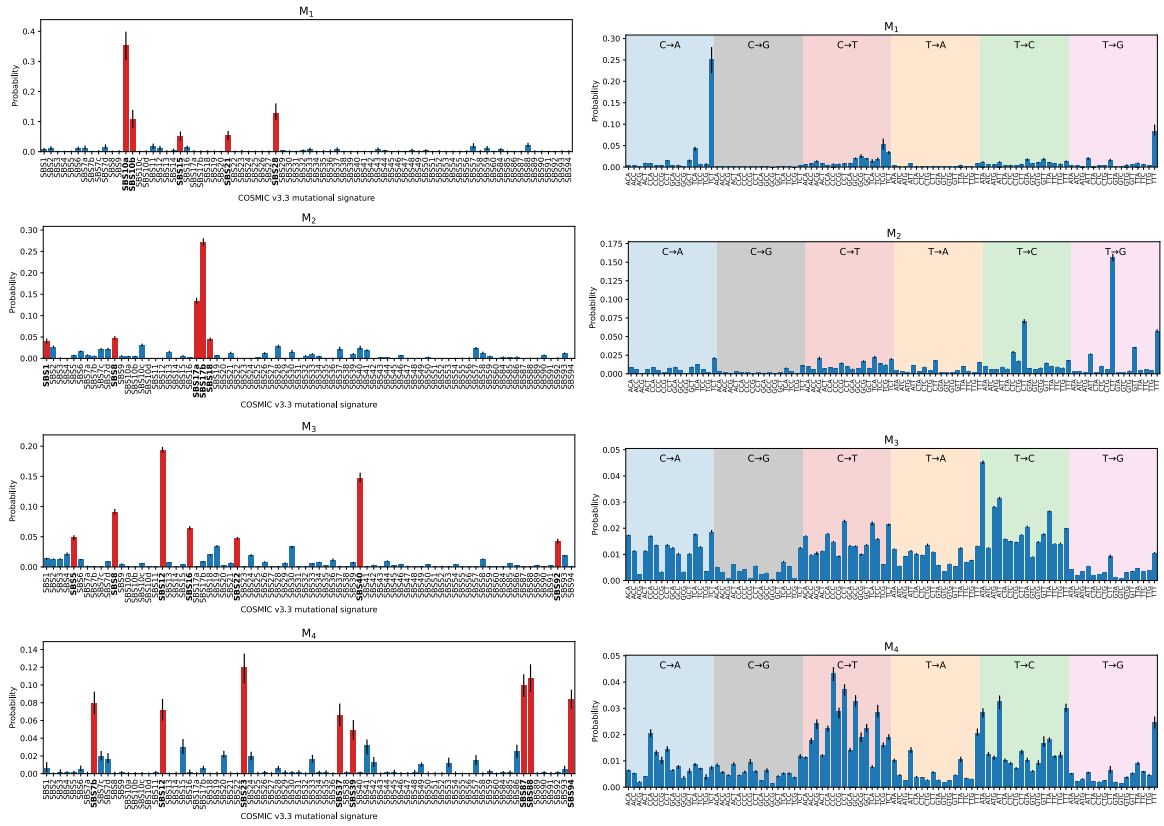


Figure 4: Posterior of four meta-mutational signatures $\phi^{(2)}_{vk}$ (labelled $k = \mathrm{M}_1, \ldots, \mathrm{M}_4$) in terms of COSMIC v3.3 mutational signatures $v = \mathrm{SBS1}, \ldots, \mathrm{SBS94}$ (left column) and its projection $\sum_{v=\mathrm{SBS1}}^{\mathrm{SBS94}} \phi^{(1)}_{lv} \phi^{(2)}_{vk}$ onto tri-nucleotide single base substitutions l (right column). Bars indicate the average and 95% quantile range of the posterior samples. On the left, mutational signatures exceeding three times the uniform probability have been marked in bold red.

pronounced transcriptional strand bias. It contains mutation patterns believed to be related to transcription-coupled nucleotide excision repair (Alexandrov et al., 2020), aging (Alexandrov et al., 2015), and aristolochic acid exposure (Hoang et al., 2013; Poon et al., 2013; Nik-Zainal et al., 2015). Finally, M_4 describes the co-occurrence of several, seemingly disparate, mutational signatures. Some are of known aetiology such as ultraviolet light (Nik-Zainal et al., 2015; Hayward et al., 2017), thiopurine chemotherapy exposure (Li et al., 2020a), and damage by the Escherichia coli bacterium (Pleguezuelos-Manzano et al., 2020; Boot et al., 2020). But M_4 also describes the co-occurrence with several signatures of unknown aetiology.

5. Discussion & Conclusion

In this paper, we proposed a multi-layer ("deep") extension of latent Dirichlet allocation designed to model healthcare data, extending the Zhou–Cong–Chen PGBN model (Zhou et al., 2016). The principal difference is that the observed data are multinomial instead of Poisson distributed: the number of observations per sample is considered fixed (i.e., conditioned on). In addition to being a practical choice for modelling common data types encountered in healthcare data, it trivially allows for modelling missingness. The model's weights are generated by a Dirichlet and therefore normalized as probabilities, while the dispersion at each layer is controlled by a single concentration parameter to allow sharing of statistical strength between samples, similar to the approach taken in the hierarchical Dirichlet process (Teh et al., 2006). In the Zhou–Cong–Chen model, Gibbs sampling was achieved by augmentation with Poisson counts throughout the network; the posterior is then sampled by exploiting an alternative factorization of distributions involving an overdispersed Poisson [negative binomial distribution, Eq. (1)]. In our model, we instead augment with multinomials. Similar to the Zhou–Cong-Chen case, this results in overdispersed (namely Dirichlet-)multinomials, and we achieve posterior sampling by developing an analogous factorization, separating the overdispersed distribution's mean and dispersion parameters and representing their posterior evidence as latent observations from a multinomial and a Chinese restaurant table distribution respectively (Theorem 1). In this way, we can treat the multinomial variable as a (latent) observation generated by the layer above so that the process can continue upwards, while we obtain the posterior of the CRT governing the dispersion for this layer using techniques introduced by Escobar and West (1995) and Teh et al. (2006).

As we show using handwritten digits, the hierarchical setup allows the model to discover different levels of abstractions in the data. We then applied our model to mutations in cancer and identified four meta-signatures describing the co-occurrence of mutagenic processes in cancer. To our knowledge, this is the first time such a first-principles characterisation of mutagenic processes in cancer has been described (two recent works Serrano Colome et al. (2023); Yaacov et al. (2024) explored an alternative approach, by building neural inference networks).

Our work focuses on the multinomial-distributed outcome variables, given their natural fit for modelling categorical and count variables frequently encountered in the domain of healthcare. As a versatile building block, the multinomial can be used to model real-valued, ordinal, and censored observations using augmentation techniques (unpublished

work), enabling it to accommodate the heterogeneous nature of health data within a unified framework. By taking a fully Bayesian approach our model is inherently robust against overfitting, which is of particular interest for applications in healthcare where data is often sparse, data collection is expensive, and avoiding spurious associations is crucial. In complex diseases, like type-1 diabetes and heart disease, patient populations are often highly heterogeneous in clinical presentation and underlying disease mechanisms (Cordell and Todd, 1995; Poulter, 1999). Topic modelling approaches such as latent Dirichlet allocation have been frequently used to model such complexities (Lu et al., 2016; Li et al., 2020b; Ahuja et al., 2020; Breuninger et al., 2021; Zhang et al., 2023), but disregard interactions between topics. MBNs can model topics and their interactions across multiple layers, capturing more intricate relationships between patient (sub)populations in terms of their risk factors, disease mechanisms, and clinical presentation. Our approach is also well suited for decomposing 'omics data such as gene expression data, metabolic profiles, and mutation profiles, all representable as counts. While these data are usually analysed using non-negative matrix factorisation (NMF), like in cancer (Hamamoto et al., 2022), MBNs are able to capture the layered complexity of these datasets. In addition to avoiding overtraining, inferences under the MBN model come with uncertainty estimates. Initiatives like the 100k genomes project aim to integrate signature analysis into standard patient care (Everall et al., 2023), underscoring the importance of reliable deconvolution and uncertainty quantification to support treatment decisions.

A technical contribution of this paper is the relation between the Dirichlet-multinomial distribution, the Chinese restaurant table distribution and the Polya urn scheme presented in Theorem 1, which to the best of our knowledge is novel. In addition, in the Supplement we present a comprehensive review of the Zhou–Cong–Chen model, which otherwise is described across several technical papers; we hope this is useful as an introduction to this elegant model.

Limitations Scaling up to large datasets remains challenging using our Gibbs sampling approach, despite our GPU implementation that can run on multiple accelerators using JAX (Bradbury et al., 2018). Approximate Markov chain Monte Carlo (Ma et al., 2015) and hybrid approaches (Zhang et al., 2018, 2020) are an attractive middle-ground between exact and approximate inference that can scale deep probabilistic models to large datasets. We leave this problem for future work.

Code and Data availability Code is available from `https://gitlab.com/hylkedonker/mubelnet` under the MIT license.

References

Ayan Acharya, Dean Teffer, Jette Henderson, Marcus Tyler, Mingyuan Zhou, and Joydeep Ghosh. Gamma process poisson factorization for joint modeling of network and documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 283–299. Springer, 2015.

Yuri Ahuja, Doudou Zhou, Zeling He, Jiehuan Sun, Victor M Castro, Vivian Gainer, Shawn N Murphy, Chuan Hong, and Tianxi Cai. surelda: A multidisease automated

phenotyping method for the electronic health record. *J. Am. Med. Inform. Assoc.*, 27(8): 1235–1243, 2020.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500 (7463):415–421, 2013.

Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407, 2015.

Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

E. Alpaydin and C. Kaynak. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. doi: 10.24432/C50P49.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A.s Albahri, Bashar Al-dabbagh, Mohammed Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data*, 10, 04 2023.

Peter Beerli. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3):341–345, 11 2005. ISSN 1367-4803.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, mar 2003.

Arnoud Boot, Alvin WT Ng, Fui Teen Chong, Szu-Chi Ho, Willie Yu, Daniel SW Tan, N Gopalakrishna Iyer, and Steven G Rozen. Characterization of colibactin-associated mutational signature in an asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res.*, 30(6):803–813, 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Samuel W Brady, Alexander M Gout, and Jinghui Zhang. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. *Trends Genet.*, 38(2):194–208, 2022.

Taylor A Breuninger, Nina Wawro, Jakob Breuninger, Sandra Reitmeier, Thomas Clavel, Julia Six-Merker, Giulia Pestoni, Sabine Rohrmann, Wolfgang Rathmann, Annette Peters, et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent dirichlet allocation. *Microbiome*, 9(1):1–18, 2021.

Heather J Cordell and John A Todd. Multifactorial inheritance in type 1 diabetes. *Trends Genet.*, 11(12):499–504, 1995.

Cosmic. Cosmic - catalogue of somatic mutations in cancer, May 2023. URL https://cancer.sanger.ac.uk/cosmic. Accessed: 2023-10-12.

Hylke C Donker and Harry JM Groen. Energy-based survival modelling using harmoniums. *arXiv preprint arXiv:2110.01960*, 2021.

Tlamelo Emmanuel, Thabiso M. Maupong, Dimane Mpoeleng, Thabo Semong, Mphago Banyatsang, and Oteng Tabona. A survey on missing data in machine learning. *J. Big Data*, 8, 2021.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90(430):577–588, 1995.

Andrew Everall, Avraam Tapinos, Aliah Hawari, Alex Cornish, Amit Sud, Daniel Chubb, Ben Kinnersley, Anna Frangou, Miguel Barquin, Josephine Jung, et al. Comprehensive repertoire of the chromosomal alteration and mutational signatures across 16 cancer types from 10,983 cancer patients. *medRxiv*, pages 2023–06, 2023.

Pedro F. Ferreira, Jack Kuipers, and Niko Beerenwinkel. Deep exponential families for single-cell data analysis. *bioRxiv*, 2022.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 2013.

Clint P. George and Hani Doss. Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model. *J. Mach. Learn. Res.*, 18:162:1–162:38, 2017.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.*, 101(suppl_1):5228–5235, 2004.

Ryuji Hamamoto, Ken Takasawa, Hidenori Machino, Kazuma Kobayashi, Satoshi Takahashi, Amina Bolatkan, Norio Shinkai, Akira Sakai, Rina Aoyama, Masayoshi Yamada, et al. Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. *Brief. Bioinform.*, 23(4):bbac246, 2022.

Nicholas K Hayward, James S Wilmott, Nicola Waddell, Peter A Johansson, Matthew A Field, Katia Nones, Ann-Marie Patch, Hojabr Kakavand, Ludmil B Alexandrov, Hazel Burke, et al. Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653): 175–180, 2017.

Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.

Margaret L Hoang, Chung-Hsin Chen, Viktoriya S Sidorenko, Jian He, Kathleen G Dickman, Byeong Hwa Yun, Masaaki Moriya, Noushin Niknafs, Christopher Douville, Rachel Karchin, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.*, 5(197):197ra102, 2013.

Karl P Hodel, Meijuan JS Sun, Nathan Ungerleider, Vivian S Park, Leonard G Williams, David L Bauer, Victoria E Immethun, Jieqiong Wang, Zucai Suo, Hua Lu, et al. POLE mutation spectra are shaped by the mutant allele identity, its abundance, and mismatch repair status. *Mol. Cell*, 78(6):1166–1177, 2020.

SM Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W Teague, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.*, 2 (11), 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Benshang Li, Samuel W Brady, Xiaotu Ma, Shuhong Shen, Yingchi Zhang, Yongjin Li, Karol Szlachta, Li Dong, Yu Liu, Fan Yang, et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood-J. Hematol.*, 135(1): 41–55, 2020a.

Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M Biernacka, et al. Inferring multimodal latent topics from electronic health records. *Nat. Commun.*, 11(1):2536, 2020b.

Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. Modeling healthcare data using multiple-channel latent dirichlet allocation. *J. Biomed. Inform.*, 60:210–223, 2016.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

T.P. Minka. Estimating a Dirichlet distribution. *Ann. Phys.*, 2000(8):1–13, 2003.

Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Serena Nik-Zainal, Jill E Kucab, Sandro Morganella, Dominik Glodzik, Ludmil B Alexandrov, Volker M Arlt, Annette Weninger, Monica Hollstein, Michael R Stratton, and David H Phillips. The genome as a record of environmental exposure. *Mutagenesis*, 30 (6):763–770, 2015.

Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

Rajat Panda, Ankit Pensia, Nikhil Mehta, Mingyuan Zhou, and Piyush Rai. Deep topic models for multi-label learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2849–2857. PMLR, 16–18 Apr 2019.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, Jose Miguel Hernandez Lobato, Aliaksandr Hubin, Alexander Immer, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Agustinus Kristiadi, Yingzhen Li, Stephan Mandt, Christopher Nemeth, Michael A. Osborne, Tim G. J. Rudner, David Rügamer, Yee Whye Teh, Max Welling, Andrew Gordon Wilson, and Ruqi Zhang. Position paper: Bayesian deep learning in the age of large-scale ai, 2024.

Andrew Patterson, Abdurrahman Elbasir, Bin Tian, and Noam Auslander. Computational methods summarizing mutational patterns in cancer: Promise and limitations for clinical applications. *Cancers*, 15(7):1958, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(85):2825–2830, 2011.

Cayetano Pleguezuelos-Manzano, Jens Puschhof, Axel Rosendahl Huber, Arne van Hoeck, Henry M Wood, Jason Nomburg, Carino Gurjao, Freek Manders, Guillaume Dalmasso, Paul B Stege, et al. Mutational signature in colorectal cancer caused by genotoxic pks^+ E. coli. *Nature*, 580(7802):269–273, 2020.

Song Ling Poon, See-Tong Pang, John R McPherson, Willie Yu, Kie Kyon Huang, Peiyong Guan, Wen-Hui Weng, Ee Yan Siew, Yujing Liu, Hong Lee Heng, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science Transl. Med.*, 5(197):197ra101, 2013.

N Poulter. Coronary heart disease is a multifactorial disease. *Am. J. Hypertens.*, 12(S6): 92S–95S, 1999.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep Exponential Families. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 762–771, San Diego, California, USA, 09–12 May 2015. PMLR.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 324–333, New York, New York, USA, 20–22 Jun 2016. PMLR.

Claudia Serrano Colome, Oleguer Canal Anton, Vladimir Seplyarskiy, and Donate Weghorn. Mutational signature decomposition with deep neural networks reveals origins of clock-like processes and hypoxia dependencies. *bioRxiv*, pages 2023–12, 2023.

Vinod Kumar Singh, Arnav Rastogi, Xiaoju Hu, Yaqun Wang, and Subhajyoti De. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.*, 3(1):421, 2020.

Richard L. Smith and J. C. Naylor. A comparison of maximum likelihood and bayesian estimators for the three- parameter weibull distribution. *J. R. Stat. Soc. C: Appl. Stat.)*, 36(3):358–369, 1987. ISSN 00359254, 14679876.

Hossein Soleimani, James Hensman, and Suchi Saria. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE T. Pattern Anal.*, 40(8):1948–1963, 2017.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 47(D1):D941–D947, 2019.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *J. Am. Stat. Assoc.*, 101(476):1566–1581, 2006.

Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. *Adv. Neural Inf. Process. Syst.*, 17, 2004.

Adar Yaacov, Gil Ben Cohen, Jakob Landau, Tom Hope, Itamar Simon, and Shai Rosenberg. Cancer mutational signatures identification in clinical assays using neural embedding-based representations. *Cell Rep. Med.*, 2024.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. Whai: Weibull hybrid autoencoding inference for deep topic modeling. In *ICLR*, 2018.

Hao Zhang, Bo Chen, Yulai Cong, Dandan Guo, Hongwei Liu, and Mingyuan Zhou. Deep autoencoding topic model with scalable hybrid bayesian inference. *IEEE T. Pattern Anal.*, 43(12):4306–4322, 2020.

Yidong Zhang, Xilin Jiang, Alexander J Mentzer, Gil McVean, and Gerton Lunter. Topic modeling identifies novel genetic loci associated with multimorbidities in uk biobank. *Cell Genom.*, 3(8), 2023.

He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Dirichlet belief networks for topic structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE T. Pattern Anal.*, 37(2):307–320, 2015.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *J. Mach. Learn. Res.*, 17(163):1–44, 2016.

Supplemental Materials for Multinomial belief networks for healthcare data

S1. Preliminaries

A sampling strategy for a Bayesian network involves a series of marginalization and augmentation steps, with relations between distributions that can be summarized by factorizations such as

$$p(x)p(y|x) = p(y)p(x|y), \tag{S1}$$

which implies $p(x) = \int p(y)p(x|y)\mathrm{d}y$, a relation that can be used either to marginalize y or to augment with y. The first factorization we use involves the Poisson distribution. Let

$$x_j \sim \mathrm{Pois}(\lambda_j); \qquad y = x_{\underline{j}}, \tag{S2}$$

where underlined indices denote summation, $x_{\underline{j}} := \sum_j x_j$, and we write vectors as $\{x_j\}_j$, dropping the outer index j when there is no ambiguity. Then y is also Poisson distributed, and conditional on y the x_j have a multinomial distribution:

$$y \sim \mathrm{Pois}(\lambda_{\underline{j}}); \qquad \{x_j\} \sim \mathrm{Mult}(y, \{\lambda_j/\lambda_{\underline{j}}\}). \tag{S3}$$

This is an instance of (S1) if the deterministic relationship $y = x_{\underline{j}}$ is interpreted as the degenerate distribution $p(y|\{x_j\}) = \delta_{x_{\underline{j}},y}$. Distributions hold conditional on fixed values of variables that appear on the right-hand side; for instance the distribution of $\{x_j\}$ in (S3) is conditional on both $\{\lambda_j\}$ and y, while in (S2) x_j is conditioned on λ_j only.

The negative binomial distribution can be seen as an overdispersed version of the Poisson distribution, in two ways. First, we can write it as a gamma-Poisson mixture. The joint distribution defined by

$$\lambda \sim \mathrm{Gam}(a, c); \qquad x \sim \mathrm{Pois}(q\lambda), \tag{S4}$$

is the same as the joint distribution defined by

$$x \sim \mathrm{NB}(a, \frac{q}{q+c}); \qquad \lambda \sim \mathrm{Gam}(a + x, c + q), \tag{S5}$$

also showing that the gamma distribution is a conjugate prior for the Poisson distribution. Note that we use the shape-and-rate parameterization of the gamma distribution.

The negative binomial can also be written as a Poisson-Logarithmic mixture Zhou and Carin (2015). Let $\mathrm{Log}(p)$ be the distribution with probability mass function

$$\mathrm{Log}(k; p) = \frac{-1}{\ln(1-p)}\frac{p^k}{k},$$

where $0 < p < 1$, and define $n \sim \mathrm{SumLog}(l, p)$ by $u_i \sim \mathrm{Log}(p)$ for $i = 1, \ldots, l$, and $n = \sum_{i=1}^{l} u_i$. Then, the joint distribution over l and n defined by

$$n \sim \mathrm{NB}(a, p); \qquad l \sim \mathrm{CRT}(n, a), \tag{S6}$$

is the same as

$$l \sim \mathrm{Pois}(-a\ln(1-p)); \qquad n \sim \mathrm{SumLog}(l, p), \tag{S7}$$

where CRT is the Chinese restaurant table distribution Antoniak (1974). This factorization allows augmenting a gamma-Poisson mixture (the negative binomial n) with a pure Poisson variate l, which is a crucial step in the deep Poisson factor analysis model. For an extension of the model we will need a similar augmentation of a Dirichlet-multinomial mixture with a pure multinomial. It can be shown (see S1.1 below) that the joint distribution over $\{x_k\}$ and $\{y_k\}$ defined by

$$\{x_k\} \sim \text{DirMult}(n, \{\lambda_k\}); \qquad y_k \sim \text{CRT}(x_k, \lambda_k); \qquad m = y_{\underline{k}} \tag{S8}$$

is the same as the joint distribution over $\{x_k\}$ and $\{y_k\}$ defined by

$$m \sim \text{CRT}(n, \lambda_{\underline{k}}); \qquad \{y_k\} \sim \text{Mult}(m, \{\lambda_k\}); \qquad \{x_k\} \sim \text{Polya}(n, \{y_k\}), \tag{S9}$$

Here $\text{Polya}(n, \{y_k\})$ is the distribution of the contents of an urn after running a Polya scheme (drawing a ball, returning the drawn ball and a new identically colored one each time, until the urn contains n balls), where the urn initially contains y_k balls of color k. It is straightforward to see that $\text{Polya}(n, \{y_k\}) = \{y_k\} + \text{DirMult}(n - m, \{y_k\})$.

S1.1. Proof of Dirichlet-multinomial-CRT factorization (Theorem 1; eqs. (S8)-(S9))

A draw from a Dirichlet-multinomial is defined by

$$\{p_j\} \sim \text{Dir}(\{\lambda_j\}); \qquad \{x_j\} \sim \text{Mult}(n, \{p_j\}); \qquad \{x_j\} \sim \text{DirMult}(n, \{\lambda_j\})$$

By building up a draw from the multinomial as n draws from a categorical distribution and using Dirichlet-multinomial conjugacy we get the Polya urn scheme,

$$\{x_j^{(1)}\} \sim \text{Mult}(1, \{\lambda_j\}); \qquad \{x_j^{(i+1)}\} \sim \{x_j^{(i)}\} + \text{Mult}(1, \{\lambda_j + x_j^{(i)}\}),$$

where $x_j = x_j^{(n)}$ and to simplify notation we dropped the normalization of the multinomial's probability parameter. This scheme highlights the overdispersed or "rich get richer" character of the Dirichlet-multinomial mixture distribution.

For the proof of (S8)-(S9), recall that a draw from the Chinese Restaurant Table distribution $t \sim \text{CRT}(n, \lambda)$ is generated by a similar scheme. Starting with an urn containing a single special ball with weight λ, balls are drawn n times, and each time the drawn ball is returned together with a new, ordinary ball of weight 1. The outcome t is the number of times the special ball was drawn.

Now return to the Polya urn scheme above and let the initial j-colored balls of weight λ_j be made of iron, let j-colored balls that are added because an iron j-colored ball was drawn be made of oak, and let other balls be made of pine. Wooden balls have weight 1 and all balls are drawn with probability proportional to their weight. Let x_j be the final number of j-colored wooden balls in the urn, let y_j be the final number of j-colored oak balls, and m the final number of oak balls of any color.

Using the equivalence between the Polya urn scheme and the Dirichlet multinomial we see that $\{x_k\}$ follow a Dirichlet multinomial distribution with parameters n and $\{\lambda_j\}$. By focusing on material and ignoring color, we see that m follows a CRT distribution with

parameters n and $\lambda_{\underline{j}}$. Similarly, focusing only on j-colored balls shows that conditional on x_j, y_j again follows a CRT distribution, with parameters x_j and λ_j, since the only events of interests are drawing a j-colored iron or wooden ball, which have probabilities proportional to λ_j and 1 respectively. Since iron balls are drawn with probability proportional to their weight, conditional on m the distribution over colors among the m oak balls is multinomial with parameters m and $\{\lambda_j\}$. Finally, conditional on knowing the number and color of the oak balls $\{y_j\}$, the process of inserting the remaining pine balls is still a Polya process except that events involving drawing iron balls are now forbidden, so that the distribution of pine balls $\{x_j - y_j\}$ is again a Dirichlet-multinomial but with parameters $n-m$ and $\{y_j\}$. This proves (S8) and (S9).

S1.2. Sampling the concentration parameters of a Dirichlet distribution

In models similar to the one considered here, the concentration parameters of a Dirichlet distribution are often kept fixed Blei et al. (2003); Zhou et al. (2012, 2016) or inferred by maximum likelihood Minka and Lafferty (2002). The factorization above makes it possible to efficiently generate posterior samples from the concentration parameters, under an appropriate prior and given multinomial observations driven by draws from the Dirichlet. The setup is

$$\begin{aligned} \alpha \sim \mathrm{Gam}(a,b); \quad \{\eta_k\} \sim \mathrm{Dir}(\{\eta_k^0\}); \\ \{x_{jk}\}_k \sim \mathrm{DirMult}(n_j, \{\alpha\eta_k\}_k); \quad y_{jk} \sim \mathrm{CRT}(x_{jk}, \alpha\eta_k); \quad m_j = y_{j\underline{k}}. \end{aligned} \tag{S10}$$

where we have written the concentration parameters as the product of a probability vector $\{\eta_k\}$ and a scalar α; these will be given Dirichlet and Gamma priors respectively. By the factorization above this is the same joint distribution as

$$\begin{aligned} \alpha \sim \mathrm{Gam}(a,b); \quad \{\eta_k\} \sim \mathrm{Dir}(\{\eta_k^0\}); \\ m_j \sim \mathrm{CRT}(n_j, \alpha); \quad \{y_{jk}\}_k \sim \mathrm{Mult}(m_j, \{\eta_k\}); \quad \{x_{jk}\}_k \sim \mathrm{Polya}(n_j, \{y_{jk}\}_k). \end{aligned} \tag{S11}$$

The numbers m_j represent the total number of distinct groups in a draw from a Dirichlet process, given the concentration parameter α Teh et al. (2006). Evidence for the value of α is encoded in the (unobserved) m_j, which in turn are determined by the (unobserved) y_{jk} which sort the unobserved groups into K separate subgroups conditional on the observed counts x_{jk}. The likelihood of the number of distinct groups m in a draw from a Dirichlet process given the concentration parameter α and total number of draws n is the probability mass function of the CRT distribution,

$$p(m|\alpha, n) = s(n,m)\alpha^m \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}, \tag{S12}$$

where $s(n,m)$ are unsigned Stirling numbers of the first kind Antoniak (1974); Teh et al. (2006). By multiplying over observations j a similar likelihood is obtained for multiple observations, together with a gamma prior on α results in a posterior distribution that we refer to as the CRT-gamma posterior:

$$\alpha \sim \mathrm{Gam}(a,b); \quad m_j \sim \mathrm{CRT}(n_j, \alpha); \quad \alpha \sim \mathrm{CRTCP}(m_{\underline{j}}, \{n_j\}_j, a, b), \tag{S13}$$

where $\mathrm{CRTCP}(\alpha|m,\{n_j\}_j,a,b) \propto \mathrm{Gam}(\alpha|a,b)\alpha^m \prod_j \frac{\Gamma(\alpha)}{\Gamma(\alpha+n_j)}$ Teh et al. (2006). A sampling scheme for α for the likelihood (S12) and a Gamma prior was devised by Escobar and West (1995), and was extended by Teh et al. (2006) to the case of multiple observations (S13). Finally, the multinomial distribution of $\{y_{jk}\}_k$ is conjugate to the Dirichlet prior on $\{\eta_k\}$ leading to

$$\{\eta_k\} \sim \mathrm{Dir}(\{\eta_k^0\}); \qquad \{y_{jk}\}_k \sim \mathrm{Mult}(m_j,\{\eta_k\}_k); \qquad \{\eta_k\} \sim \mathrm{Dir}(\{\eta_k^0 + y_{\underline{j}k}\}). \tag{S14}$$

S1.3. Sampling parameters of the gamma distribution

If a Poisson-distributed observation with rate proportional to a gamma-distributed variable is available, we can use conditional conjugacy to sample the posterior of the gamma parameters. Suppose that

$$\alpha \sim \mathrm{Gam}(a_0,b_0); \quad \beta \sim \mathrm{Gam}(e_0,f_0); \quad \theta \sim \mathrm{Gam}(\alpha,\beta); \quad m \sim \mathrm{Pois}(q\theta),$$

and that θ is not observed, but the count m is. Marginalizing θ we get $m \sim \mathrm{NB}(\alpha, q/(q+\beta))$. Augmenting with $x \sim \mathrm{CRT}(m,\alpha)$ and using (S6) and (S7) we find that $x \sim \mathrm{Pois}[\alpha \ln(1+q/\beta)]$. Using gamma-Poisson conjugacy gives the mutually dependent update equations

$$x \sim \mathrm{CRT}(m,\alpha); \quad \alpha \sim \mathrm{Gam}(a_0+x, b_0+\ln(1+q/\beta)). \tag{S15}$$

As posterior for θ given m we get $\theta \sim \mathrm{Gam}(\alpha+m,\beta+q)$; augmenting with θ and using the gamma-gamma conjugacy

$$\beta \sim \mathrm{Gam}(e_0,f_0); \qquad \theta \sim \mathrm{Gam}(\alpha,\beta); \qquad \beta \sim \mathrm{Gam}(e_0+\alpha, f_0+\theta). \tag{S16}$$

results in the mutually dependent update equations

$$\theta \sim \mathrm{Gam}(\alpha+m,\beta+q); \quad \beta \sim \mathrm{Gam}(e_0+\alpha, f_0+\theta). \tag{S17}$$

S2. Gamma belief network

Since many of the techniques of the Gamma belief network of Zhou et al. Zhou et al. (2016) apply to the multinomial belief network, we start by summarising and reviewing their method in some detail. We stay close to their notation, but have made some modifications where this simplifies the future connection to the multinomial belief network.

S2.1. Backbone of feature activations

The backbone of the model is a stack of Gamma-distributed hidden units $\theta_{vj}^{(t)}$, where the last unit parameterizes observed counts x_{vj} following the Poisson distribution, one for each sample j and feature v. The generative model is

$$a_{vj}^{(T+1)} = r_v, \tag{S18}$$

$$\theta_{vj}^{(t)} \sim \mathrm{Gam}(a_{vj}^{(t+1)}, c_j^{(t+1)}), \qquad t = T,\dots,1 \tag{S19}$$

$$a_{vj}^{(t)} = \sum_{k=1}^{K_t} \phi_{vk}^{(t)} \theta_{kj}^{(t)}, \qquad t = T,\dots,1 \tag{S20}$$

$$x_{vj} \sim \mathrm{Pois}(a_{vj}^{(1)}). \tag{S21}$$

For $T = 1$ we only have one layer, and the model reduces to $x_{vj} = \text{Pois}([\boldsymbol{\phi}\boldsymbol{\theta}]_{vj})$, called Poisson Factor Analysis Zhou et al. (2012). For multiple layers, the features $\boldsymbol{\theta}^{(t+1)}$ on layer $t+1$ determine the shape parameters of the gamma distributions on layer t through a connection weight matrix $\boldsymbol{\phi}^{(t+1)} \in \mathbb{R}^{K_t \times K_{t+1}}$, so that $\boldsymbol{\phi}^{(t+1)}$ induces correlations between features on level t. The lowest-level activations $\boldsymbol{a}^{(1)}$ are used to parameterize a Poisson distribution, which generates the observed count variables x_{vj} for individual (document, observation) j. Below we will treat r_v and $c_j^{(t)}$ as random variables and targets for inference, but for now, we consider them as fixed parameters and focus on inference of $\boldsymbol{\phi}^{(t)}$ and $\boldsymbol{\theta}_j^{(t)}$. We will assume that $\sum_v \phi_{vk}^{(t)} = 1$, which later on is enforced by Dirichlet priors on $\boldsymbol{\phi}_k$.

This model architecture is similar to a T-layer neural network, with $\boldsymbol{a}^{(t)}$ playing the role of activations that represent the activity of features (topics, factors) of increasing complexity as t increases. In the remainder, we use the language of topic models, so that x_{vj} is the number of times word v is used in document j, and $\phi_{vk}^{(1)}$ is the probability that word v occurs in topic k. This is for the lowest level 1; we will similarly refer to level-t topics and level-t "words", the latter representing the activity of corresponding topics on level $t-1$.

Different from Zhou et al. (2016) we use a Gamma-distributed variate $c_j^{(2)}$ as rate parameter of the gamma distribution for $\theta^{(1)}$, instead of $p_j^{(2)}/1 - p_j^{(2)}$ where $p_j^{(2)}$ has a Beta distribution; we will come back to this choice below.

S2.2. Augmentation with latent counts

We review Zhou's augmentation and marginalization scheme that enables efficient inference for this model. First, introduce new variables

$$x_{vj}^{(t)} \sim \text{Pois}(q_j^{(t)} a_{vj}^{(t)}), \qquad t = 1, \dots, T+1 \tag{S22}$$

where we set $q_j^{(1)} = 1$ so that we can identify $x_{vj}^{(1)}$ with the observed counts x_{vj}; the $q_j^{(t)}$ for $t > 1$ will be defined below. Using (S2)-(S3) we can augment $x_{vj}^{(t)}$ as

$$y_{vjk}^{(t)} \sim \text{Pois}(q_j^{(t)} \phi_{vk}^{(t)} \theta_{kj}^{(t)}); \tag{S23}$$

$$x_{vj}^{(t)} = y_{vj\underline{k}}^{(t)}. \tag{S24}$$

The counts $y_{vjk}^{(t)}$ represent a possible assignment of level-t words to level-t topics. Marginalizing over v and using (S2)-(S3) again we get the augmentation

$$m_{jk}^{(t)} := y_{\underline{v}jk}^{(t)} \sim \text{Pois}(q_j^{(t)} \theta_{kj}^{(t)}); \tag{S25}$$

$$y_{vjk}^{(t)} = \text{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v). \tag{S26}$$

since $\phi_{\underline{v}k}^{(t)} = 1$. These counts represent level-t topic usage in document j. Now, marginalizing $\boldsymbol{\theta}^{(t)}$ turns $m_{jk}^{(t)}$ into an overdispersed Poisson distribution; from (S19) we see that $\boldsymbol{\theta}^{(t)}$ is Gamma distributed, so it becomes a negative binomial:

$$m_{jk}^{(t)} \sim \text{NB}(a_{kj}^{(t+1)}, q_j^{(t)}/(q_j^{(t)} + c_j^{(t+1)})), \tag{S27}$$

using (S4) and (S5). This gives us a count variable that is parameterized by the activation of the layer above t, but one which follows a negative binomial distribution rather than a Poisson distribution (S22). However, using (S6) and (S7) we can augment once more to write the negative binomial as a Poisson-Logarithmic mixture:

$$x_{kj}^{(t+1)} \sim \text{Pois}(a_{kj}^{(t+1)} \ln \frac{q_j^{(t)} + c_j^{(t+1)}}{c_j^{(t+1)}}); \qquad m_{jk}^{(t)} \sim \text{SumLog}(x_{kj}^{(t+1)}, \frac{q_j^{(t)}}{q_j^{(t)} + c_j^{(t+1)}}), \tag{S28}$$

so that $x_{kj}^{(t+1)}$ agrees with (S22) if we choose

$$q_j^{(t+1)} := \ln \frac{q_j^{(t)} + c_j^{(t+1)}}{c_j^{(t+1)}}. \tag{S29}$$

This allows us to continue the procedure for layer $y + 1$, and so on until $t = T$, sampling augmented variables $y_{vjk}^{(t)}$, $m_{jk}^{(t)}$ and $x_{kj}^{(t+1)}$ for $t = 1, \ldots, T$.

S2.3. Alternative representation as Deep Poisson Factor model

The procedure described in section S2.2 not only augments the model with new counts but also integrates out $\boldsymbol{\theta}^{(t)}$. This provides an alternative and equivalent representation as a generative model. Starting from $a_{kj}^{(t+1)}$ we can use (S28), (S26) and (S24) to sample $x_{kj}^{(t+1)}$, $m_{jk}^{(t)}$, $y_{vjk}^{(t)}$, and finally $x_{vj}^{(t)}$. Continuing downwards this shows how to eventually generate the observed counts $x_{vj}^{(1)}$ using count variables, while $\boldsymbol{\theta}^{(t)}$ is integrated out. Explicitly, the generative model becomes

$$x_{kj}^{(t+1)} \sim \text{Pois}(q_j^{(t+1)} a_{kj}^{(t+1)});$$
$$m_{jk}^{(t)} \sim \text{SumLog}(x_{kj}^{(t+1)}, 1 - e^{-q_j^{(t+1)}}); \qquad \{y_{vjk}^{(t)}\}_v \sim \text{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v); \qquad x_{vj}^{(t)} := y_{vj\underline{k}}^{(t)}.$$

(Note that throughout we condition on $q_j^{(t)}$ for all t, and therefore on all $c_j^{(t)}$ as well; we also haven't specified how to sample $\phi_{vk}^{(t)}$ yet.) This equivalent generative process motivates the name Deep Poisson Factor Analysis. The two alternative schemes are shown graphically in figure S1.

S2.4. Sampling per-document latent variables

The derivation above can be used to sample the latent counts conditional on observations (and parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ and $\boldsymbol{r}$), from layer 1 upwards. These steps are,

$$y_{vjk}^{(t)} \sim \text{Mult}(x_{vj}^{(t)}, \{\phi_{vk}^{(t)} \theta_{kj}^{(t)}\}_k); \tag{S30}$$
$$m_{jk}^{(t)} = y_{\underline{v}jk}^{(t)}; \tag{S31}$$
$$x_{kj}^{(t+1)} \sim \text{CRT}(m_{jk}^{(t)}, a_{kj}^{(t+1)}), \tag{S32}$$

where for (S30) we used (S2)-(S3) and (S23)-(S24); and for (S32) we used (S6)-(S7) and (S27)-(S28). Note that after the last step $x_{kj}^{(t+1)}$ is no longer conditioned on $\theta_{kj}^{(t)}$ because it

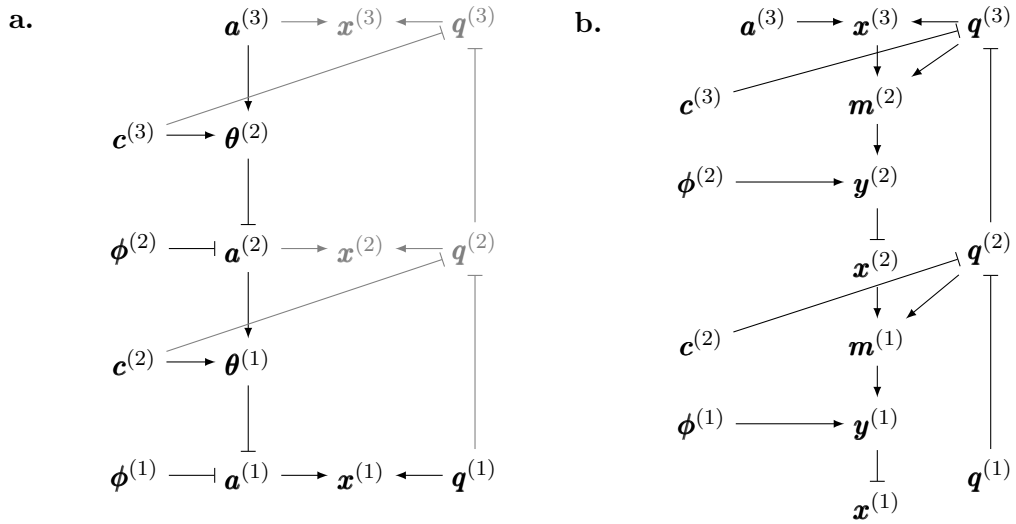


Figure S1: Two equivalent generative models for a count variable $\boldsymbol{x}^{(1)}$ from the Poisson gamma belief network, using (**a**) a tower of real-valued latent variables $\boldsymbol{\theta}$, $\boldsymbol{a}$, or (**b**) latent counts $\boldsymbol{m}, \boldsymbol{y}, \boldsymbol{x}$. Blunt arrows indicate deterministic relationships. The variable $\boldsymbol{q}^{(1)}$ is a dummy and has a fixed value 1. The counts $\boldsymbol{x}^{(t)}$ and variables $\boldsymbol{q}^{(t)}$, $t > 1$, in the left representation, are included for clarity (and have the same distribution as the variables in the right model) but are not used to generate the outcome $\boldsymbol{x}^{(1)}$, and so can be marginalized out.

is integrated out, however, explicit values for $\theta_{kj}^{(t)}$ *are* used in (S30). To sample new values for $\theta_{kj}^{(t)}$, we use Gamma-Poisson conjugacy (S4)-(S5) on (S19) and (S25) to get

$$\theta_{kj}^{(t)} \sim \text{Gam}(a_{kj}^{(t+1)} + m_{kj}^{(t)}, c_j^{(t+1)} + q_j^{(t)}). \tag{S33}$$

In order to sample the final set of per-document variables, the inverse scaling parameters $c_j^{(t)}$, we first need to integrate out $q_j^{(t)}$ since it depends on $c_j^{(t)}$ via (S29). That means we also need to integrate out $x_{vj}^{(t)}$ and $m_{jk}^{(t-1)}$ which both depend on $q_j^{(t)}$, as well as $y_{vjk}^{(t-1)}$ because of the deterministic relationship (S25). We do not need to marginalize other variables as $\boldsymbol{x}^{(t-1)}$ and its dependents are conditionally independent of $\boldsymbol{x}^{(t)}$ given $\boldsymbol{\theta}^{(t-1)}$, as can be seen from figure 2a. Once $\boldsymbol{x}^{(t)}$, $\boldsymbol{m}^{(t-1)}$ and $\boldsymbol{y}^{(t-1)}$ are integrated out, $\theta_{vj}^{(t-1)}$ is related to $c_j^{(t)}$ solely through (S19), and marginalizing this over v we get

$$\theta_{\underline{v}j}^{(t-1)} \sim \text{Gam}(a_{\underline{v}j}^{(t)}, c_j^{(t)}), \qquad (t = 2, \ldots, T+1).$$

where $a_{\underline{v}j}^{(t)} = \theta_{\underline{k}j}^{(t)}$ for $t = 2, \ldots, T$, and $a_{\underline{v}j}^{(T+1)} = r_{\underline{v}}$. The conjugate prior for a gamma likelihood with fixed shape parameter is a gamma distribution again. This gives the following prior and posterior distributions for $c_j^{(t)}$:

$$c_j^{(t)} \sim \text{Gam}(e_0, f_0); \qquad c_j^{(t)} \sim \text{Gam}(e_0 + a_{\underline{v}j}^{(t)}, f_0 + \theta_{\underline{k}j}^{(t-1)}). \qquad (t = 2, \ldots, T+1) \tag{S34}$$

As an aside, note that it is possible to integrate out $\boldsymbol{\phi}^{(1)}$ and $\boldsymbol{\theta}^{(1)}$, in the same way as is done in the collapsed Gibbs sampler for the LDA model. This is done in Zhou et al. (2016) and may lead to faster mixing. However, if we do that no $\boldsymbol{\theta}^{(1)}$ is available to construct a posterior for $\boldsymbol{c}^{(2)}$ as in (S34). Instead, (S27) can be used:

$$m_{jk}^{(1)} \sim \text{NB}(a_{kj}^{(2)}, 1/(1 + c_j^{(2)})),$$

using that $q_j^{(1)} = 1$. Using the beta-negative binomial conjugacy, we can write

$$p^{(2)} \sim \text{Beta}(a_0, b_0); \qquad p^{(2)} \sim \text{Beta}(a_0 + a_{\underline{k}j}^{(2)}, b_0 + m_{j\underline{k}}^{(1)}),$$

where we defined

$$p^{(2)} := (1 + c_j^{(2)})^{-1}, \qquad \text{so that} \qquad c_j^{(2)} = (1 - p^{(2)})/p^{(2)},$$

which gives $\boldsymbol{c}^{(2)}$ a Beta distribution of the second kind. We cannot similarly integrate out $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\phi}^{(t)}$ for $t > 1$ as already for collapsed Gibbs sampling, values for $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\phi}^{(2)}$ are necessary to define the prior on $\boldsymbol{\theta}^{(1)}$.

S2.5. Sampling $\boldsymbol{\phi}$ and $\boldsymbol{r}$

It remains to sample the model-level parameters $\phi_{vk}^{(t)}$, $\eta_v^{(t)}$, and r_v. For $\boldsymbol{\phi}$ we marginalize (S26) over j and, noting that the multinomial is conjugate to a Dirichlet, we use a Dirichlet prior for $\boldsymbol{\phi}$ to obtain the update equations

$$\{\phi_{vk}^{(t)}\}_v \sim \text{Dir}(\{\eta_v^{(t)}\}_v); \qquad \{y_{v\underline{j}k}^{(t)}\}_v \sim \text{Mult}(m_{\underline{j}k}^{(t)}, \{\phi_{vk}^{(t)}\}_v); \qquad \{\phi_{vk}^{(t)}\}_v \sim \text{Dir}(\{\eta_v^{(t)} + y_{v\underline{j}k}^{(t)}\}_v). \tag{S35}$$

To sample $\{\eta_v^{(t)}\}$, we integrate out $\phi_{vk}^{(t)}$ and consider $x_{v\underline{j}k}^{(t)}$ as a draw from a Dirichlet-multinomial distribution with parameters $\alpha^{(t)}\eta_v^{(t)}$ where the factors $\alpha^{(t)}$ and $\{\eta_v^{(t)}\}$ have Gamma and Dirichlet priors respectively, as in (S10). This results in the following update equations,

$$\alpha^{(t)} \sim \text{Gam}(a,b); \quad \{\eta_v^{(t)}\}_v \sim \text{Dir}(\{\eta_{v,0}\}_v); \quad \{y_{v\underline{j}k}^{(t)}\}_v \sim \text{DirMult}(m_{\underline{j}k}^{(t)}, \{\alpha^{(t)}\eta_v^t\}_v);$$
$$z_{vk}^{(t)} \sim \text{CRT}(y_{v\underline{j}k}^{(t)}, \alpha^{(t)}\eta_v^{(t)}); \quad \alpha \sim \text{CRTCP}(z_{v\underline{k}}^{(t)}, \{m_{\underline{j}k}^{(t)}\}_k, a, b); \quad \{\eta_v^{(t)}\}_v \sim \text{Dir}(\{\eta_{v,0} + z_{v\underline{k}}^{(t)}\}_v). \tag{S36}$$

For r_v we use (S28) and (S29) for $t = T$, marginalize j, and use gamma-Poisson conjugacy to get update equations

$$r_v \sim \text{Gam}(\gamma_0/K_T, c_0); \qquad x_{v\underline{j}}^{(T+1)} \sim \text{Pois}(r_v q_{\underline{j}}^{(T+1)}); \qquad r_v \sim \text{Gam}(\gamma_0/K_T + x_{v\underline{j}}^{(T+1)}, c_0 + q_{\underline{j}}^{(T+1)}) \tag{S37}$$

S2.6. Sampling strategy

It is helpful to think of the model as organised as an alternating stack of layers, one taking inputs $\boldsymbol{a}$ and $\boldsymbol{c}$ and using the gamma distribution to produce an output $\boldsymbol{\theta}$; and one taking $\boldsymbol{\theta}$ and edge weights $\boldsymbol{\phi}$ to produce an activation $\boldsymbol{a}$. During inference, the model also uses latent variables x_{vj}, y_{vjk}, m_{jk} and q_j. Inference proceeds in two main stages. First, the $\boldsymbol{q}^{(t)}$ are calculated, followed by augmentation with the $\boldsymbol{y}$, $\boldsymbol{m}$ and $\boldsymbol{x}^{(t)}$ ($t \geq 2$) count variables going up the stack, while marginalising $\boldsymbol{\phi}$, and also updating the $\boldsymbol{\phi}$ variables. After updating the parameter $\boldsymbol{r}$, the second stage involves updating $\boldsymbol{c}$ and $\boldsymbol{\theta}$ going down the stack, while the augmented count variables are dropped again. Table S1 provides a detailed overview.

S3. Multinomial observables

S3.1. Deep Multinomial Factor Analysis

To model multinomial observations, we replace the Poisson observables with a multinomial and, for each sample j, we swap out the gamma-distributed hidden activations for Dirichlet samples $\{\theta_{vj}^{(t)}\}_v$. The generative model is

$$a_{vj}^{(T+1)} = r_v, \tag{S38}$$
$$\{\theta_{vj}^{(t)}\}_v \sim \text{Dir}(\{c^{(t+1)}a_{vj}^{(t+1)}\}_v), \qquad t = T, \ldots, 1 \tag{S39}$$
$$a_{vj}^{(t)} = \sum_{k=1}^{K_t} \phi_{vk}^{(t)}\theta_{kj}^{(t)}, \qquad t = T, \ldots, 1 \tag{S40}$$
$$\{x_{vj}\}_v \sim \text{Mult}(n_j, \{a_{vj}^{(1)}\}_v). \tag{S41}$$

Missing from this model definition are the specification of the prior distributions of r_v and $c^{(t)}$; these are introduced in section S3.2 but are considered fixed in this section. The variables $c^{(t)}$, $t = 2, \ldots, T+1$ set the scale of the Dirichlet's concentration parameters which modulate the variance of $\theta_{vj}^{(t-1)}$ across documents j. Different from the PGBN model

			factor-layer 1										factor-layer 2										
								gamma layer 1									gamma layer 2						
Stage	Eq.	(Eq.)	x^1	q^1	a^1	ϕ^1	y^1	θ^1	m^1	c^2	x^2	q^2	a^2	ϕ^2	y^2	θ^2	m^2	c^3	x^3	q^3	r
			O	1	x	x	-	x	-	x	-	-	x	x	-	x	-	x	-	-	x
0 (all t)	(S29)	(S55)	O	1	x	x	-	x	-	x	-	**U**	x	x	-	x	-	x	-	**U**	x
1 (t=1)	(S30)	(S52)	O	1	x	x	**S**	x	-	x	-	x	x	x	-	x	-	x	-	x	x
2 (t=1)	(S35)	(S35)	O	1	-	**S**	x	x	-	x	-	x	x	x	-	x	-	x	-	x	x
3 (t=1)	(S31)	(S53)	O	1	-	x	x	x	**U**	x	-	x	x	x	-	x	-	x	-	x	x
4 (t=1)	(S32)	(S54)	O	1	-	x	x	-	x	x	**S**	x	x	x	-	x	-	x	-	x	x
1 (t=2)	(S30)	(S52)	O	1	-	x	x	-	x	x	x	x	x	x	**S**	x	-	x	-	x	x
2 (t=2)	(S35)	(S35)	O	1	-	x	x	-	x	x	x	x	-	**S**	x	x	-	x	-	x	x
3 (t=2)	(S31)	(S53)	O	1	-	x	x	-	x	x	x	x	-	x	x	x	**U**	x	-	x	x
4 (t=2)	(S32)	(S54)	O	1	-	x	x	-	x	x	x	x	-	x	x	-	x	x	**S**	x	x
r	(S37)	(S58)	O	1	-	x	x	-	x	x	x	x	-	x	x	-	x	x	x	**S**	
5 (t=2)	(S33)	(S56)	O	1	-	x	x	-	x	x	x	x	-	x	x	**S**	x	x	x	x	x
6 (t=2)	(S34)	(S57)	O	1	-	x	x	-	x	x	x	x	-	x	-	x	-	**S**	-	-	x
7 (t=2)	(S20)	(S40)	O	1	-	x	x	-	x	x	x	x	**U**	x	-	x	-	x	-	-	x
5 (t=1)	(S33)	(S56)	O	1	-	x	x	**S**	x	x	x	x	x	x	-	x	-	x	-	-	x
6 (t=1)	(S34)	(S57)	O	1	-	x	-	x	-	**S**	-	-	x	x	-	x	-	x	-	-	x
7 (t=1)	(S20)	(S40)	O	1	**U**	x	-	x	-	x	-	-	x	x	-	x	-	x	-	-	x

Table S1: Variable instantiation and marginalisation during inference. Upper indices denote layer number (without parentheses). Lower indices are suppressed. All variables include the observation index j except for r and ϕ. The symbols 1, −, O, x, S, and U denote the value 1; marginalized; observed; instantiated; sampled; and deterministically updated respectively. Boxed symbols denote dependencies (either on an instantiated value, or on the corresponding variable having been marginalized). Inference of $\eta^{(t)}$, the parameter of the Dirichlet prior for $\phi^{(t)}$, not shown. Equation numbers refer to Gamma-Poisson model (left column) and multinomial belief network (right column; replace columns labeled "q" with "n").

we choose one $c^{(t)}$ per dataset instead of one per sample j, reducing the number of free parameters per sample, and allowing the variance across samples to inform the $c^{(t)}$. Similar to the PGBN, the first step towards a posterior sampling procedure involves augmentation and marginalization. First, we introduce new variables

$$\{x_{vj}^{(t)}\}_v \sim \text{Mult}(n_j^{(t)}, \{a_{vj}^{(t)}\}_v), \qquad t = 1, \ldots, T+1 \tag{S42}$$

where we set $n_j^{(1)} := n_j$ and we identify x_{vj} with $x_{vj}^{(1)}$; below we define $n_j^{(t)}$ for $t > 1$. We can augment $x_{vj}^{(t)}$ as

$$\{y_{vjk}^{(t)}\}_{vk} \sim \text{Mult}(n_j^{(t)}, \{\phi_{vk}^{(t)}\theta_{kj}^{(t)}\}_{vk}); \qquad x_{vj}^{(t)} = y_{vj\underline{k}}^{(t)}; \tag{S43}$$

Marginalizing $y_{vjk}^{(t)}$ over v results in the augmentation

$$\{m_{jk}^{(t)}\}_k := y_{\underline{v}jk}^{(t)} \sim \text{Mult}(n_j^{(t)}, \{\theta_{kj}^{(t)}\}_k); \tag{S44}$$

$$\{y_{vjk}^{(t)}\}_v \sim \text{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v). \tag{S45}$$

where (S44) holds since $\phi_{\underline{v}k}^{(t)} = 1$. Now marginalizing over $\theta_{kj}^{(t)}$ in (S44) results in a Dirichlet-multinomial, an overdispersed multinomial that plays a role similar to the negative binomial as an overdispersed Poisson for the PGBN:

$$\{m_{jk}^{(t)}\}_k \sim \text{DirMult}(n_j^{(t)}, \{c^{(t+1)}a_{kj}^{(t+1)}\}_k). \tag{S46}$$

To augment this overdispersed multinomial with a pure multinomial, so that we can continue the augmentation in the layer above, we use (S8)–(S9):

$$n_j^{(t+1)} \sim \text{CRT}(n_j^{(t)}, c^{(t+1)}); \tag{S47}$$

$$\{x_{kj}^{(t+1)}\}_k \sim \text{Mult}(n_j^{(t+1)}, \{a_{kj}^{(t+1)}\}_k); \tag{S48}$$

$$\{m_{jk}^{(t)}\}_k \sim \text{Polya}(n_j^{(t)}, \{x_{kj}^{(t+1)}\}_k), \tag{S49}$$

where in the first line we used that $a_{\underline{k}j}^{(t+1)} = 1$, because $\phi_{\underline{v}k}^{(t+1)} = \theta_{\underline{k}j}^{(t+1)} = 1$. Equation (S47) recursively defines the distribution of the scale counts $n_j^{(t)}$ for $t > 1$, which play a role analogous to the $q_j^{(t)}$ in the PGBN; these counts depend only on n_j and are independent of the observations $x_{vj}^{(1)}$ that we condition on, depending only on the $c^{(t')}$, $t' \leq t$. Continuing this procedure results in augmented counts $y_{vjk}^{(t)}$, $m_{jk}^{(t)}$ and $x_{kj}^{(t+1)}$ for $t = 1, \ldots, T$.

With $\theta^{(t)}$ integrated out, the emerging alternative generative model representation is a deep multinomial factor model, as follows:

$$n_j^{(t+1)} \sim \text{CRT}(n_j^{(t)}, c^{(t+1)}); \qquad \{x_{kj}^{(t+1)}\}_k \sim \text{Mult}(n_j^{(t+1)}, \{a_{kj}^{(t+1)}\}_k); \tag{S50}$$

$$\{m_{kj}^{(t)}\}_k \sim \text{Polya}(n_j^{(t)}, \{x_{kj}^{(t+1)}\}_k); \qquad \{y_{vjk}^{(t)}\} \sim \text{Mult}(m_{jk}^{(t)}, \{\phi_{vk}^{(t)}\}_v); \qquad x_{vj}^{(t)} = y_{vj\underline{k}}^{(t)}. \tag{S51}$$

The two representations of the model are structurally identical to the two representations of the PGBN shown in figure S1, except that $q^{(t)}$ are replaced by $n^{(t)}$.

S3.2. Sampling posterior variables

To sample counts conditional on observations we use

$$\{y_{vjk}^{(t)}\}_k \sim \text{Mult}(x_{vj}^{(t)}, \{\phi_{vk}^{(t)}\theta_{kj}^{(t)}\}_k); \tag{S52}$$

$$m_{jk}^{(t)} = y_{\underline{v}jk}^{(t)}; \tag{S53}$$

$$x_{kj}^{(t+1)} \sim \text{CRT}(m_{jk}^{(t)}, c^{(t+1)}a_{kj}^{(t+1)}); \tag{S54}$$

$$n_j^{(t+1)} = x_{\underline{k}j}^{(t+1)} \tag{S55}$$

similar to (S30)-(S31); for (S54) we used (S46)-(S49) and (S8)–(S9). To sample θ, use (S44), (S39) and Dirichlet-multinomial conjugacy to get

$$\{\theta_{kj}^{(t)}\}_k \sim \text{Dir}(\{c^{(t+1)}a_{kj}^{(t+1)} + m_{jk}^{(t)}\}_k), \tag{S56}$$

To sample the scaling factor $c^{(t)}$, we use the Chinese restaurant representation of $n_j^{(t)}$ together with (S13):

$$c^{(t)} \sim \text{Gam}(e_0, f_0); \quad n_j^{(t)} \sim \text{CRT}(n_j^{(t-1)}, c^{(t)}); \quad c^{(t)} \sim \text{CRTCP}(n_{\underline{j}}^{(t)}, \{n_j^{(t-1)}\}_j, e_0, f_0). \tag{S57}$$

where $n_j^{(t)} = x_{\underline{k}j}^{(t)}$. Because the relationship (S45) between $y_{vjk}^{(t)}$ and $\phi_{vk}^{(t)}$ is as in the PGBN model, we sample $\phi^{(t)}$ and its prior parameters using (S35)-(S36) as before, using a Dirichlet prior on $\phi^{(t)}$, and a gamma-Dirichlet prior on its concentration parameters. Finally, using a Dirichlet prior for r_v we have the update equations

$$\{r_v\}_v \sim \text{Dir}(\{\gamma_0/K_T\}_v); \qquad \{x_{vj}^{(T+1)}\}_v \sim \text{Mult}(n_j^{(T+1)}, \{r_v\}_v); \qquad \{r_v\}_v \sim \text{Dir}(\{\gamma_0/K_t + x_{v\underline{j}}^{(T+1)}\}_v). \tag{S58}$$

S4. Experiments

S4.1. Greedy layer-wise training on mutational signatures

For both the single layer PGBN and MBN, four chains were run for 1700 Gibbs steps each. Samples from the last 250 iterations, thinned every fifth sample, were collected for analysis (leaving 50 samples per chain). Thereafter, an additional $K_2 = 78$ latent component layer was added on top of each respective model and the chains were run for 500 additional steps. For the PGBN we inferred 38 latent components on the second layer (that is, out of all four chains, the smallest number of empty signatures $m_{\underline{jk}}^{(2)} = 0$). The top layer was subsequently pruned back to 38 latent components and the chains were run for an additional 550 steps collecting the last 250 samples thinned to 50 samples. The MBN, was (accidentally) run slightly longer, for 750 steps, and we inferred 41 latent components. After pruning the empty topics, 250 additional steps were collected and thinned for analysis. Overall, a total of 77 days (78 days) of GPU time—divided across four nVidia A40 GPU devices—were used to execute 2700 Markov steps per chain for the MBN (2800 steps for the PGBN).

S4.2. Meta-signature construction

Consensus meta-signatures were determined by matching the topics of different chains to its' centroid by repeatedly solving the optimal transport problem Murphy (2023) for the Jensen-Shannon distance (JSD) Murphy (2023) using the Hungarian algorithm Crouse (2016) until the centroid converged in terms of silhouette score Rousseeuw (1987), similar to Ref. Alexandrov et al. (2020). The centroid was initialised with restarting points coming from different chains and the consensus meta-signatures that gave the best silhouette score were selected. Finally, we selected robust meta-signatures by choosing those centroids where the JSD between the closest signature was no less than 0.25 across all chains, leaving four meta-signatures in total (named, M_1 through M_4). For completeness, we list all 37 other meta signatures in Figs. S3–S5. While completely inactive meta signatures were pruned, seven (out of the 41) signatures with a very small topic activity remained (to wit, M_{23}, M_{26}, M_{35}, M_{37}-M_{40}).

S4.3. Interpretation meta signatures $\mathrm{M}_1, \ldots, \mathrm{M}_4$

Here, we characterise the four meta signatures named M_1 through M_4. Summarising the meta-signatures by entropy $s(k) = -\sum_{v=1}^{78} \phi_{vk}^{(2)} \ln \phi_{vk}^{(2)}$, we found that the posterior coverage was low with an entropy-based effective sample size (Vehtari et al., 2021) of 11, 5, 6, and 6, respectively.

Next, we describe, per meta-signature, the ($K_1 = 78$) mutational signatures, v, exceeding three times uniform probability (i.e., $\phi_{vk}^{(2)} \geq 3/K_1$, analogous to Zhou et al. (2016)) and their biological interpretation.

M_1 describes the co-occurrence of replicative DNA polymerase ϵ (POLE) damage (SBS10a, SBS10b, and SBS28 (Li et al., 2018; Hodel et al., 2020), but *not* POLE associated SBS14 (Hodel et al., 2020)) and mismatch-repair deficiency (MMR, SBS15 and SSB21 (Meier et al., 2018)) (Fig. 4, first row, left column). Tumours with an ultra-hypermutated phenotype (≥ 100 mutations Mb^{-1}) are often characterised by these joint disruptions in MMR and POLE (Hodel et al., 2020). Combined, M_1 describes a preference for altering C→A and T→G when flanked by a T on either side (Fig. 4, first row, right column).

Meta-signature M_2 primarily captures, presumably, oxidative stress. Its constituents SBS17a/b (Secrier et al., 2016) and SBS18 are thought to be related to guanine oxidation, resulting in the formation of 8-Oxo-2'-deoxyguanosine (8-oxo-dG) (Nones et al., 2014; Tomkova et al., 2018; Poetsch et al., 2018; Christensen et al., 2019); SBS18 is additionally linked to hydroxyl radicals in culture (Kucab et al., 2019). Similar to clock-like signature SBS1 (describing spontaneous deamination of 5-methylcytosine (Nik-Zainal et al., 2012; Alexandrov et al., 2015)), damage due to 8-oxo-dG accumulates in the course of life (Nie et al., 2013). To a lesser extent, M_2 also captures SBS8, which is implicated in BRCA1 and BRCA2 dysfunction in breast cancer (Nik-Zainal et al., 2016) and believed to be (uncorrected) replication errors (Singh et al., 2020). Characteristically, M_2 prefers T→G and T→A singlets with a contextual T on the right-hand side (Fig. 4, second row, right column).

Meta-signature M_3 is marked by a pronounced transcriptional strand bias, including signatures such as SBS5, SBS8, SBS12, SBS16 (Alexandrov et al., 2020), along with SBS92 (Lawson et al., 2020) and SBS22 (Cosmic, 2023), with SBS40 being the exception. Its primary constituent SBS12 is believed to be related to transcription-coupled nucleotide exci-

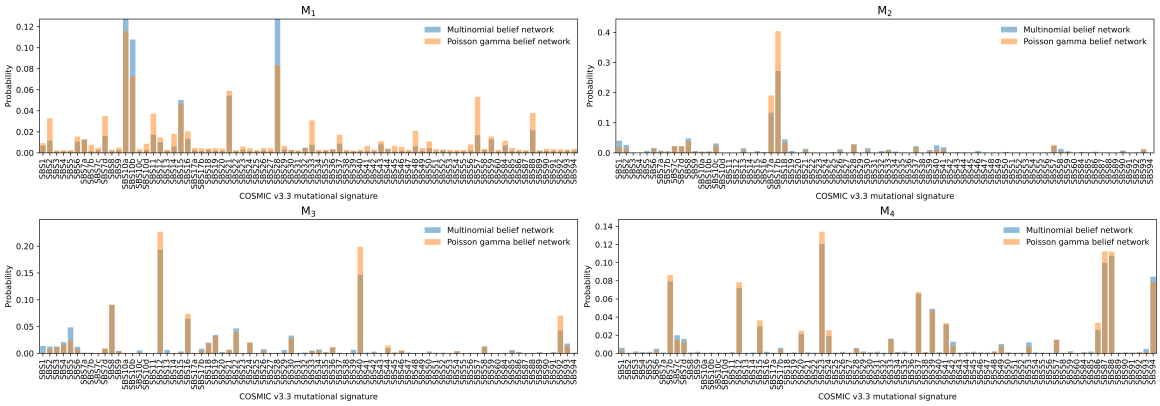


Figure S2: Meta signatures M_1–M_4 replicate in the Poisson gamma belief network (PGBN). The data shows the posterior average of the closest matching meta signatures extracted from the PGBN.

sion repair (Alexandrov et al., 2020). The second largest contributor, SBS40, is a spectrally flat, late-replicating (Singh et al., 2020), signature with spectral similarities to SBS5 (both are related to age (Alexandrov et al., 2015)) and is believed to be linked to SBS8 (Singh et al., 2020). According to COSMIC, some contamination between SBS5 and SBS16 may be present (Tate et al., 2019; Cosmic, 2023); M_3 is consistent with this observation. Finally, M_3 also captures the co-occurrence with SBS22, which is canonically attributed to aristolochic acid exposure (Hoang et al., 2013; Poon et al., 2013; Nik-Zainal et al., 2015). Overall, M_3 gives rise to a dense spectrum and inherits the quintessential depletion of C substitutions when right-flanked by a G from SBS40 (Fig. 4, third row, right column).

Finally, M_4 describes the co-occurrence of several, seemingly disparate, mutational signatures of known and unknown aetiology. Of known cause are, SBS7b, linked to ultraviolet light (Nik-Zainal et al., 2015; Hayward et al., 2017), SBS87 to thiopurine chemotherapy exposure (Li et al., 2020) (although its presence has been reported in a thiopurine-naive population (Donker et al., 2023)) and SBS88, related to colibactin-induced damage from the Escherichia coli bacterium (Pleguezuelos-Manzano et al., 2020; Boot et al., 2020) (found in various tissue types (Boot et al., 2020; Lee-Six et al., 2019; Pleguezuelos-Manzano et al., 2020)). Concurrently, M_4 comprises SBS12 (Alexandrov et al., 2020), SBS23 (Alexandrov et al., 2015; Nik-Zainal et al., 2016), SBS37 (Alexandrov et al., 2020), SBS39 (Alexandrov et al., 2020) and SBS94 (Islam et al., 2022), all of unknown cause. Jointly, these signatures describe a dense spectrum with a slight tendency for C→T substitutions. Reassuringly, meta-signatures $M_1, \ldots, M_4$ replicated independently in the PGBN (Fig. S2, Supplementary Material). To our knowledge, this is the first time a first-principles characterisation of the organising principles of mutagenic processes in cancer has been carried out.

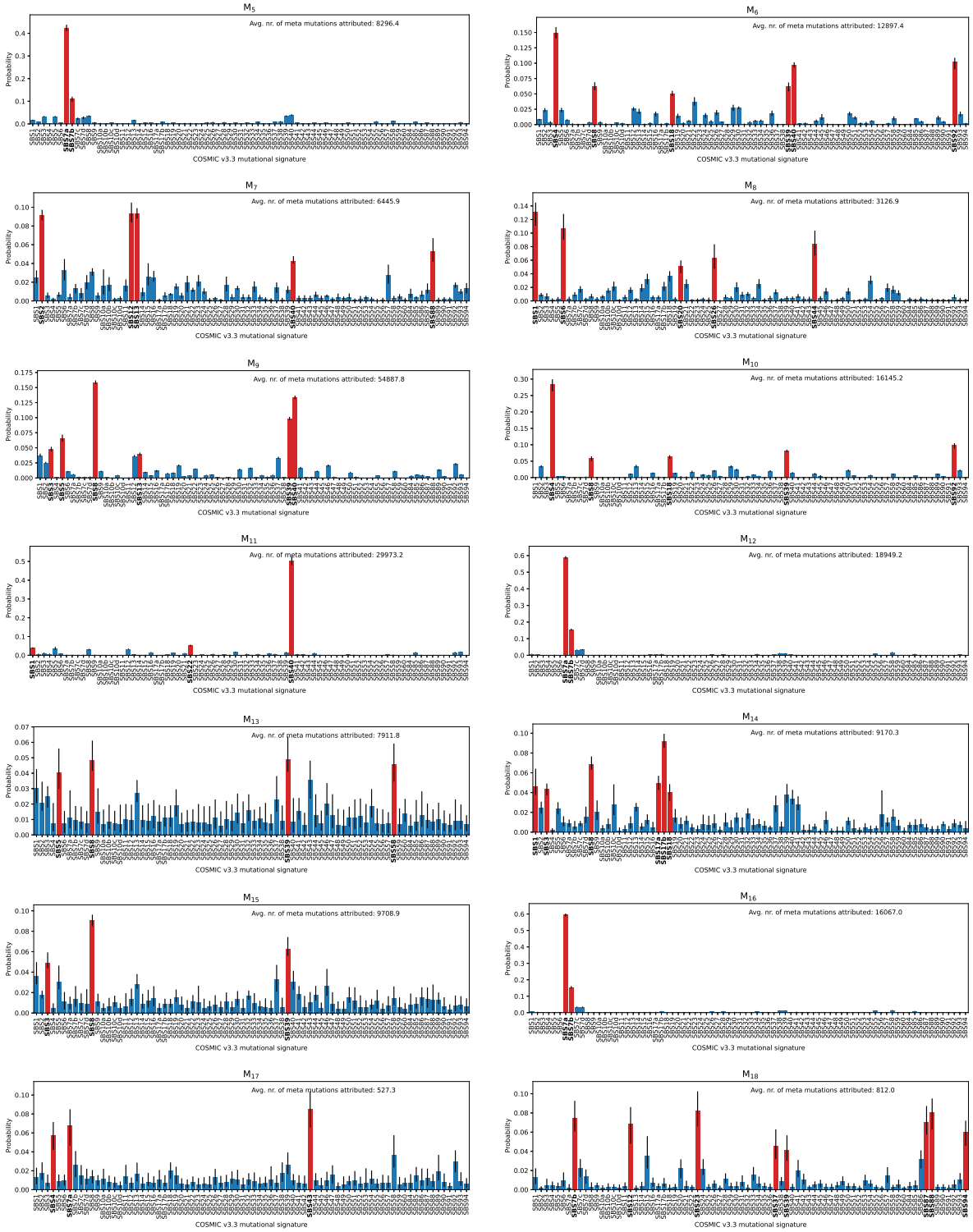


Figure S3: Posterior of meta-mutational signatures $k = \mathrm{M}_5, \ldots, \mathrm{M}_{18}$ for the multinomial belief network (meta signatures $k = \mathrm{M}_{19}, \ldots, \mathrm{M}_{41}$ are listed in subsequent figures). In each panel, the total number of meta signature k counts $m^{(2)}_{k\underline{j}}$ (averaged over the posterior samples) is indicated as a measure of topic loading.

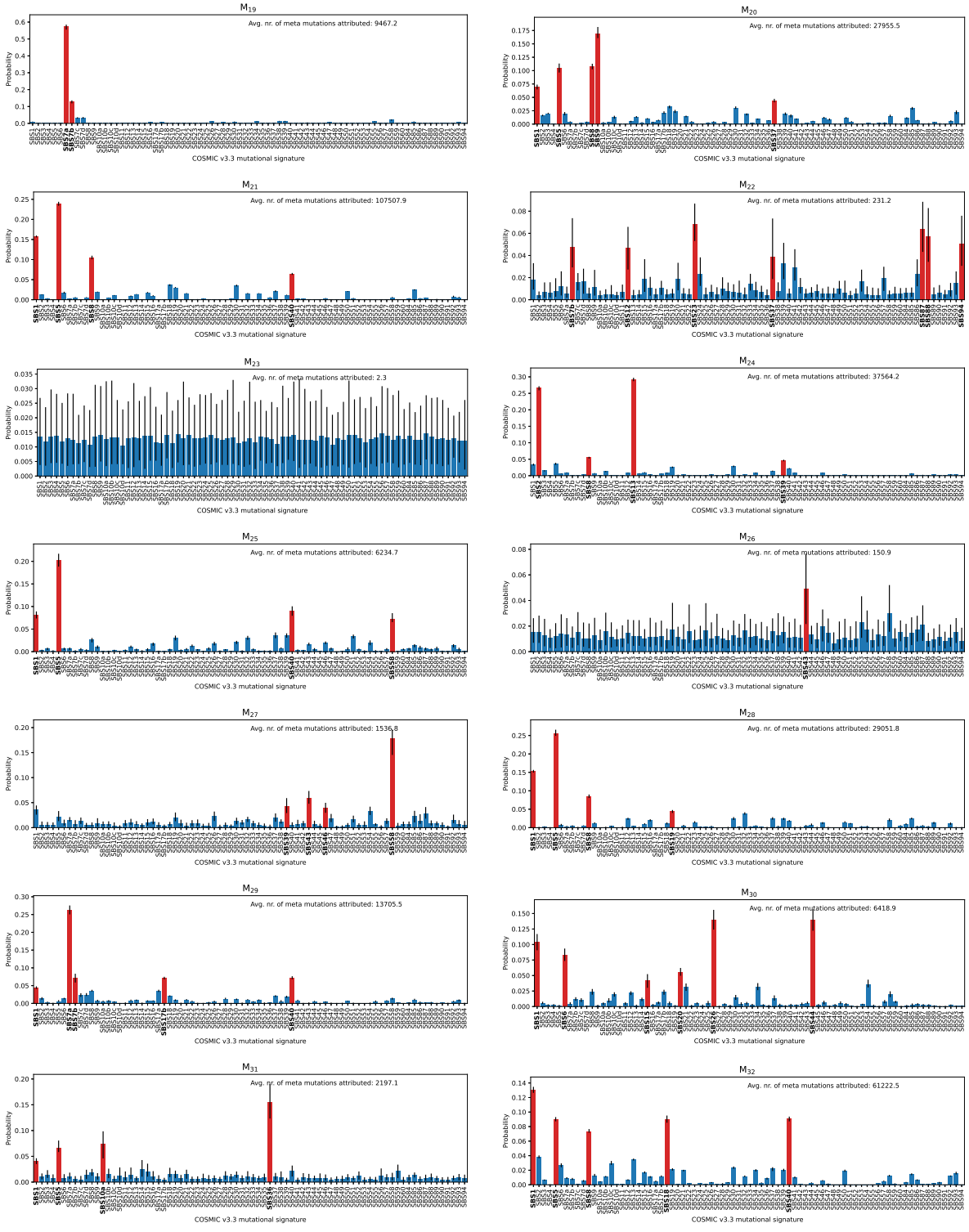


Figure S4: Continuation of Fig. S3 listing $k = \mathrm{M}_{19}, \ldots, \mathrm{M}_{32}$ (meta signatures $k = \mathrm{M}_{33}, \ldots, \mathrm{M}_{41}$ are listed in the subsequent figure).

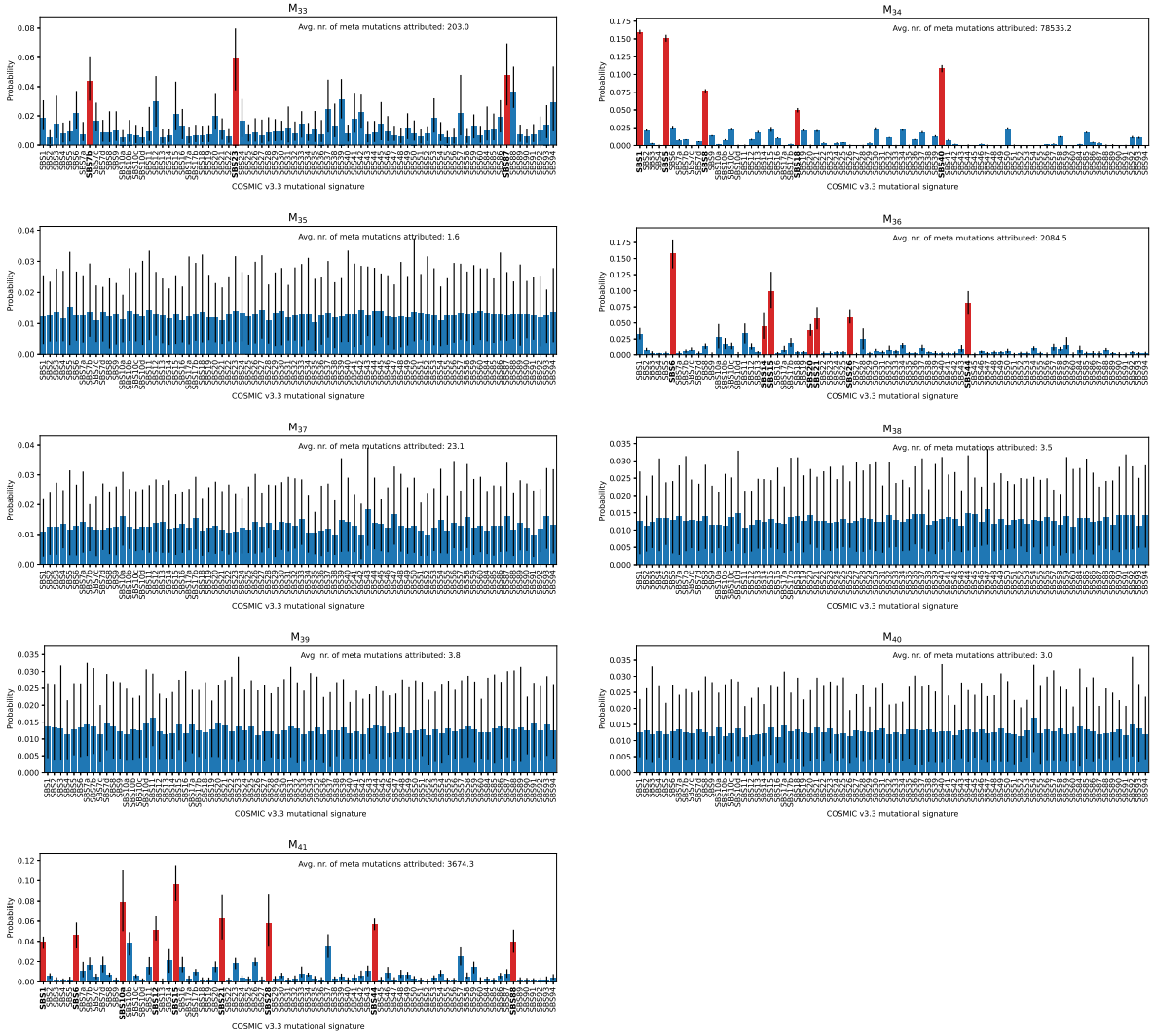


Figure S5: Continuation of Figs. S3 and S4 listing $k = \mathrm{M}_{33}, \ldots, \mathrm{M}_{41}$.

Table S2: Summary statistics of the hyperparameter $c^{(t)}$.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$c^{(3)}$	0.298	0.036	0.230	0.351	0.016	0.012	5.0	12.0	3.10
$c^{(2)}$	33.978	4.545	26.305	39.319	2.125	1.611	5.0	13.0	3.73

Table S3: Summary statistics of the hyperparameter $\{r_k\}$ (top-level activation), per meta signature k.

Meta-signature	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
M_1	0.0005	0.0003	0.0000	0.0011	0.0001	0.0001	7.8437	45.1667	1.5635
M_2	0.0274	0.0171	0.0051	0.0489	0.0080	0.0060	5.2076	25.3881	2.9207
M_3	0.0462	0.0275	0.0180	0.0934	0.0129	0.0098	5.2138	35.0239	2.8226
M_4	0.0023	0.0009	0.0008	0.0037	0.0003	0.0002	7.2637	57.5069	1.6368
M_5	0.0090	0.0091	0.0014	0.0255	0.0043	0.0032	5.6255	42.9691	2.3436
M_6	0.0174	0.0141	0.0025	0.0425	0.0066	0.0050	5.1255	18.2017	3.0118
M_7	0.0086	0.0077	0.0009	0.0208	0.0036	0.0027	5.9037	37.2047	2.1290
M_8	0.0047	0.0056	0.0006	0.0154	0.0026	0.0020	7.2067	56.4231	1.6564
M_9	0.0807	0.0395	0.0297	0.1444	0.0185	0.0140	5.1425	33.8033	2.9455
M_{10}	0.0173	0.0102	0.0033	0.0298	0.0047	0.0036	5.6516	25.2606	2.3022
M_{11}	0.0424	0.0294	0.0070	0.0889	0.0137	0.0104	5.2041	39.3170	2.8706
M_{12}	0.0125	0.0045	0.0049	0.0189	0.0020	0.0015	5.4742	18.1715	2.4575
M_{13}	0.0119	0.0190	0.0000	0.0463	0.0089	0.0067	5.3522	42.5339	2.6005
M_{14}	0.0121	0.0108	0.0000	0.0307	0.0051	0.0038	5.1627	23.8320	2.9830
M_{15}	0.0139	0.0141	0.0000	0.0335	0.0066	0.0050	5.3330	24.4206	2.6387
M_{16}	0.0094	0.0015	0.0066	0.0121	0.0005	0.0004	8.5465	70.0544	1.4770
M_{17}	0.0007	0.0005	0.0000	0.0017	0.0002	0.0002	6.3390	41.6786	1.8971
M_{18}	0.0015	0.0009	0.0001	0.0029	0.0004	0.0003	6.1458	22.9947	1.9816
M_{19}	0.0069	0.0036	0.0015	0.0117	0.0016	0.0012	5.9197	36.6124	2.0946
M_{20}	0.0441	0.0335	0.0124	0.1036	0.0157	0.0119	5.0797	15.1183	3.1108
M_{21}	0.2091	0.0914	0.1374	0.3724	0.0428	0.0324	5.2223	40.6697	2.7940
M_{22}	0.0003	0.0003	0.0000	0.0009	0.0001	0.0001	8.9105	41.9669	1.4437
M_{23}	0.0001	0.0001	0.0000	0.0003	0.0000	0.0000	17.0880	69.2559	1.1846
M_{24}	0.0590	0.0462	0.0209	0.1413	0.0216	0.0164	5.4445	25.3216	2.4660
M_{25}	0.0124	0.0071	0.0048	0.0249	0.0033	0.0025	5.8507	32.4049	2.1428
M_{26}	0.0003	0.0003	0.0000	0.0010	0.0001	0.0001	8.8749	41.5220	1.4720
M_{27}	0.0030	0.0019	0.0001	0.0056	0.0008	0.0006	5.8643	17.1875	2.1408
M_{28}	0.0626	0.0459	0.0172	0.1442	0.0215	0.0163	5.1546	24.8252	2.9731
M_{29}	0.0221	0.0235	0.0007	0.0601	0.0110	0.0083	5.5399	45.9551	2.4353
M_{30}	0.0056	0.0035	0.0004	0.0101	0.0016	0.0012	5.3785	26.2501	2.5512
M_{31}	0.0044	0.0070	0.0000	0.0175	0.0033	0.0025	6.0873	22.1597	2.0126
M_{32}	0.0991	0.0535	0.0183	0.1604	0.0251	0.0190	5.2073	48.6473	2.8941
M_{33}	0.0004	0.0003	0.0000	0.0010	0.0001	0.0001	8.3580	80.6760	1.4931
M_{34}	0.1459	0.0950	0.0333	0.2602	0.0445	0.0338	5.1767	33.8603	2.9398
M_{35}	0.0000	0.0001	0.0000	0.0002	0.0000	0.0000	115.4890	95.5287	1.0455
M_{36}	0.0020	0.0017	0.0000	0.0050	0.0008	0.0006	5.4310	59.6550	2.5273
M_{37}	0.0001	0.0001	0.0000	0.0003	0.0000	0.0000	13.4681	93.1817	1.2433
M_{38}	0.0001	0.0001	0.0000	0.0002	0.0000	0.0000	78.6768	160.1208	1.0418
M_{39}	0.0001	0.0001	0.0000	0.0003	0.0000	0.0000	20.6372	145.3196	1.1420
M_{40}	0.0001	0.0001	0.0000	0.0002	0.0000	0.0000	14.2321	22.7601	1.2294
M_{41}	0.0042	0.0057	0.0000	0.0150	0.0027	0.0020	5.4181	21.5255	2.5456

References

Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407, 2015.

Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Ann. Stat.*, 2(6):1152 – 1174, 1974.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, mar 2003.

Arnoud Boot, Alvin WT Ng, Fui Teen Chong, Szu-Chi Ho, Willie Yu, Daniel SW Tan, N Gopalakrishna Iyer, and Steven G Rozen. Characterization of colibactin-associated mutational signature in an asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res.*, 30(6):803–813, 2020.

Sharon Christensen, Bastiaan Van der Roest, Nicolle Besselink, Roel Janssen, Sander Boymans, John WM Martens, Marie-Laure Yaspo, Peter Priestley, Ewart Kuijk, Edwin Cuppen, et al. 5-Fluorouracil treatment induces characteristic T¿G mutations in human cancer. *Nature Commun.*, 10(1):4571, 2019.

Cosmic. Cosmic - catalogue of somatic mutations in cancer, May 2023. URL https://cancer.sanger.ac.uk/cosmic. Accessed: 2023-10-12.

David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE T. Aero. Elec. Sys.*, 52(4):1679–1696, 2016.

H. C. Donker, B van Es, M Tamminga, G. A. Lunter, L. C. L. T. van Kempen, E Schuuring, T. J. N. Hiltermann, and H. J. M. Groen. Using genomic scars to select immunotherapy beneficiaries in advanced non-small cell lung cancer. *Sci. Rep.*, 13(1):6581, 2023.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90(430):577–588, 1995.

Nicholas K Hayward, James S Wilmott, Nicola Waddell, Peter A Johansson, Matthew A Field, Katia Nones, Ann-Marie Patch, Hojabr Kakavand, Ludmil B Alexandrov, Hazel Burke, et al. Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653): 175–180, 2017.

Margaret L Hoang, Chung-Hsin Chen, Viktoriya S Sidorenko, Jian He, Kathleen G Dickman, Byeong Hwa Yun, Masaaki Moriya, Noushin Niknafs, Christopher Douville, Rachel Karchin, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.*, 5(197):197ra102, 2013.

Karl P Hodel, Meijuan JS Sun, Nathan Ungerleider, Vivian S Park, Leonard G Williams, David L Bauer, Victoria E Immethun, Jieqiong Wang, Zucai Suo, Hua Lu, et al. POLE mutation spectra are shaped by the mutant allele identity, its abundance, and mismatch repair status. *Mol. Cell*, 78(6):1166–1177, 2020.

SM Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W Teague, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.*, 2 (11), 2022.

Jill E Kucab, Xueqing Zou, Sandro Morganella, Madeleine Joel, A Scott Nanda, Eszter Nagy, Celine Gomez, Andrea Degasperi, Rebecca Harris, Stephen P Jackson, et al. A compendium of mutational signatures of environmental agents. *Cell*, 177(4):821–836, 2019.

Andrew RJ Lawson, Federico Abascal, Tim HH Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A Sanders, Anne Y Warren, Krishnaa TA Mahbubani, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, 370(6512):75–82, 2020.

Henry Lee-Six, Sigurgeir Olafsson, Peter Ellis, Robert J Osborne, Mathijs A Sanders, Luiza Moore, Nikitas Georgakopoulos, Franco Torrente, Ayesha Noorani, Martin Goddard, et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574(7779):532–537, 2019.

Benshang Li, Samuel W Brady, Xiaotu Ma, Shuhong Shen, Yingchi Zhang, Yongjin Li, Karol Szlachta, Li Dong, Yu Liu, Fan Yang, et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood-J. Hematol.*, 135(1): 41–55, 2020.

Hao-Dong Li, Ileana Cuevas, Musi Zhang, Changzheng Lu, Md Maksudul Alam, Yang-Xin Fu, M James You, Esra A Akbay, He Zhang, Diego H Castrillon, et al. Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load. *J. Clin. Invest.*, 128(9):4179–4191, 2018.

Bettina Meier, Nadezda V Volkova, Ye Hong, Pieta Schofield, Peter J Campbell, Moritz Gerstung, and Anton Gartner. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Res.*, 28(5):666–675, 2018.

T. Minka and J. Lafferty. Expectation-Propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, San Francisco, CA, 2002.

Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Ben Nie, Wei Gan, Fei Shi, Guo-Xin Hu, Lian-Guo Chen, Hiroshi Hayakawa, Mutsuo Sekiguchi, Jian-Ping Cai, et al. Age-dependent accumulation of 8-oxoguanine in the dna and rna in various rat tissues. *Oxid. Med. Cell. Longev.*, 2013, 2013.

Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149 (5):979–993, 2012.

Serena Nik-Zainal, Jill E Kucab, Sandro Morganella, Dominik Glodzik, Ludmil B Alexandrov, Volker M Arlt, Annette Weninger, Monica Hollstein, Michael R Stratton, and David H Phillips. The genome as a record of environmental exposure. *Mutagenesis*, 30 (6):763–770, 2015.

Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

Katia Nones, Nicola Waddell, Nicci Wayte, Ann-Marie Patch, Peter Bailey, Felicity Newell, Oliver Holmes, J Lynn Fink, Michael CJ Quinn, Yue Hang Tang, et al. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.*, 5(1):5224, 2014.

Cayetano Pleguezuelos-Manzano, Jens Puschhof, Axel Rosendahl Huber, Arne van Hoeck, Henry M Wood, Jason Nomburg, Carino Gurjao, Freek Manders, Guillaume Dalmasso, Paul B Stege, et al. Mutational signature in colorectal cancer caused by genotoxic pks^+ E. coli. *Nature*, 580(7802):269–273, 2020.

Anna R Poetsch, Simon J Boulton, and Nicholas M Luscombe. Genomic landscape of oxidative dna damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.*, 19(1):1–23, 2018.

Song Ling Poon, See-Tong Pang, John R McPherson, Willie Yu, Kie Kyon Huang, Peiyong Guan, Wen-Hui Weng, Ee Yan Siew, Yujing Liu, Hong Lee Heng, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science Transl. Med.*, 5(197):197ra101, 2013.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

Maria Secrier, Xiaodun Li, Nadeera De Silva, Matthew D Eldridge, Gianmarco Contino, Jan Bornschein, Shona MacRae, Nicola Grehan, Maria O'Donovan, Ahmad Miremadi, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.*, 48(10):1131–1141, 2016.

Vinod Kumar Singh, Arnav Rastogi, Xiaoju Hu, Yaqun Wang, and Subhajyoti De. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.*, 3(1):421, 2020.

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 47(D1):D941–D947, 2019.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *J. Am. Stat. Assoc.*, 101(476):1566–1581, 2006.

Marketa Tomkova, Jakub Tomek, Skirmantas Kriaucionis, and Benjamin Schuster-Böckler. Mutational signature distribution varies with dna replication timing and strand asymmetry. *Genome Biol.*, 19(1):1–12, 2018.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved r-hat for assessing convergence of mcmc (with discussion). *Bayesian Anal.*, 16(2):667–718, 2021.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE T. Pattern Anal.*, 37(2):307–320, 2015. doi: 10.1109/TPAMI.2013.211.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *J. Mach. Learn. Res.*, 17(163):1–44, 2016.