
Wave-LSTM: Multi-scale analysis of somatic whole genome copy number profiles

Charles Gadd
University of Oxford

Christopher Yau
University of Oxford
Health Data Research UK

Abstract

Changes in the number of copies of certain parts of the genome, known as copy number alterations (CNAs), due to somatic mutation processes are a hallmark of many cancers. This genomic complexity is known to be associated with poorer outcomes for patients but describing its contribution in detail has been difficult. Copy number alterations can affect large regions spanning whole chromosomes or the entire genome itself but can also be localised to only small segments of the genome and no methods exist that allow this multi-scale nature to be quantified. In this paper, we address this using Wave-LSTM, a signal decomposition approach designed to capture the multi-scale structure of complex whole genome copy number profiles. Using wavelet-based source separation in combination with deep learning-based attention mechanisms. We show that Wave-LSTM can be used to derive multi-scale representations from copy number profiles which can be used to decipher sub-clonal structures from single-cell copy number data and to improve survival prediction performance from patient tumour profiles.

1 Introduction

Somatic copy number alterations (SCNAs) are a hallmark of cancers (Beroukhim et al., 2010). The normal human genome typically contains two copies of every gene (Figure 1a), one copy derived from each biological parent, and a normal complement of 44 non-sex chromosomes. In a cancer genome, errors in DNA repair and dysregulation of cellular processes due to mutations can alter the genomic content leading to the loss or gain of genomic segments (Figure 1b). This can lead to a phenomenon known as *chromosomal instability* (CIN) in which large-scale whole genome changes occur (Sansregret et al., 2018). Copy number changes affecting certain genes can lead to an oncogenic effect in which additional copies of a gene (oncogene) promote cancer development or the loss of a gene (tumour suppressor gene) stops a suppressive role that normally occurs.

Modern molecular analysis technologies, such as DNA sequencing, together with bioinformatics (such as TITAN (Ha et al., 2014) or ASCAT (Van Loo et al., 2010)), allow the number of total copies of each part of the genome to be measured and counted. This work is not concerned with the low-level processing of raw sequencing data. Instead, we focused on the analysis of the derived copy number profiles from the processed sequenced data which consists of *major* and *minor* copy numbers or so-called *unphased* copy number counts (Figure 1c). Therefore, the data we consider consists of the following: chromosome, position along chromosome, major copy number and minor copy number (note it is normally not possible to resolve to the level of parentally derived copy number without special experimental techniques). There are typically millions of genomic positions measured.

Copy number alterations (CNAs) occupy multiple genomic scales. Some alterations affect the whole genome itself – such as *whole genome doubling* (WGD) (Bielski et al., 2018) – where failed cellular division causes daughter cells to gain a whole extra copy of the genome. Other changes may only affect individual or partial chromosomes, or are localised to small parts of the genome. Evolutionary forces and biological constraints (i.e. cellular viability) often drive cancers from different individuals and types to acquire recurrent patterns of SCNAs. Recent studies have sought to understand their underlying aetiology (Drews et al., 2022) and have discovered associations with severity of disease and clinical outcomes (Beroukhim et al., 2010). However, it remains a challenge to understand how the mixture of large and fine-scale CNAs lead to certain outcomes (e.g. survival).

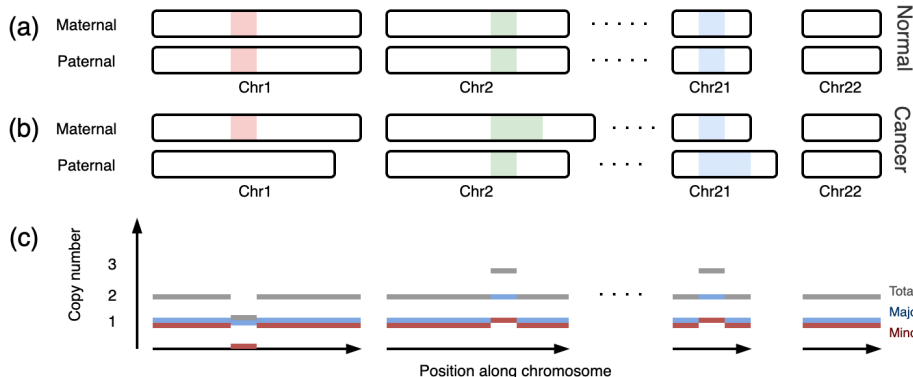


Figure 1: **An example of a somatic copy number alteration.** (a) A normal cell. (b) A cancerous cell which has undergone a deletion on the paternal strand of chromosome (chr) 1, and an insertion on the maternal and paternal strands of chromosome 2 and 21, respectively. (c) These cancerous mutations are quantified by the copy number.

Contributions. In this paper, we present an approach to learn multi-scale representations of high-dimensional, structured cancer copy number profiles called Wave-LSTM. We demonstrate through examples that Wave-LSTM provides a means for exploratory and predictive analyses that enable us to understand the multi-scale contribution of copy number alterations to cancer subtypes and survival.

Illustration. We first illustrate using a toy example in Figure 2a. Here, we simulated six signal classes which are displayed at low, mid and full scale (raw signal). Notice that at low scales, classes 3/4 and 5/6 are indistinguishable but, at higher scales, these classes can be separated using the fine scale signal features. Figure 2c shows a series of scale-dependent Wave-LSTM embeddings derived from the simulated data. The low-scale Wave-LSTM embeddings contain only four distinct groups corresponding to the four distinguishable signal classes. While the higher-scale Wave-LSTM embeddings contain increasing numbers of classes from 4 to 6. Figure 2b shows an overall *multi-scale embedding* in which Wave-LSTM combines information across all scales. When applied to copy number data we expect this capability to allow us to separate the effects of large-scale chromosomal alterations from finer-scale, focal aberrations in unsupervised and supervised learning tasks.

Reproducibility. We make our code available as a PyTorch package with accompanying notebooks: <https://github.com/cwlgadd/waveLSTM>.

2 Methods

Overview. Wave-LSTM consists of three main integrated components: (i) wavelet-based source separation, (ii) source-specific convolutional filters and (iii) a self-attention block that examines information across scales. This leads to an output representation of the input signal which can be used directly for unsupervised analyses or integrated into a prediction/classification module for supervised learning (Figure 3).

Notation. We briefly define some notation that will be of use in the understanding of the next sections. Let $X_i \in \mathbb{R}^{C \times W}$ denote the i -th training sequence which, in generality could be composed of C channels, each of width W . We index each channel as $x^{(c)} \in \mathbb{R}^W$, excluding sample index i for notational clarity. We then denote by n the position along each channel, and for brevity and notational convenience we assume channels are binned to a width that is shared across channels.

In our application, each channel then corresponds to the division of the genome across chromosomes and strands (i.e. major and minor copy number). For the human genome, this then corresponds to $C = 46$ channels from 23 chromosomes (when including non-autosomes), across 2 strands.

Consequently, $x^{(c)}[n]$ then denotes the sequence of binned copy numbers at each (proportional) position along the respective chromosome and strand. We refer to these positions along a channel as the loci.

Wavelet-based source separation. We use the discrete wavelet (Haar) transform as a cascading filter bank. Each level of the filter leads to a corresponding low-pass decomposition u (which approximates the signal), and high-pass decomposition w (which details the difference between the current level and the approximation), with impulse responses $u[\cdot]$ and $w[\cdot]$, respectively. Wavelet coefficients are obtained by repeatedly applying these low-pass, and high-pass

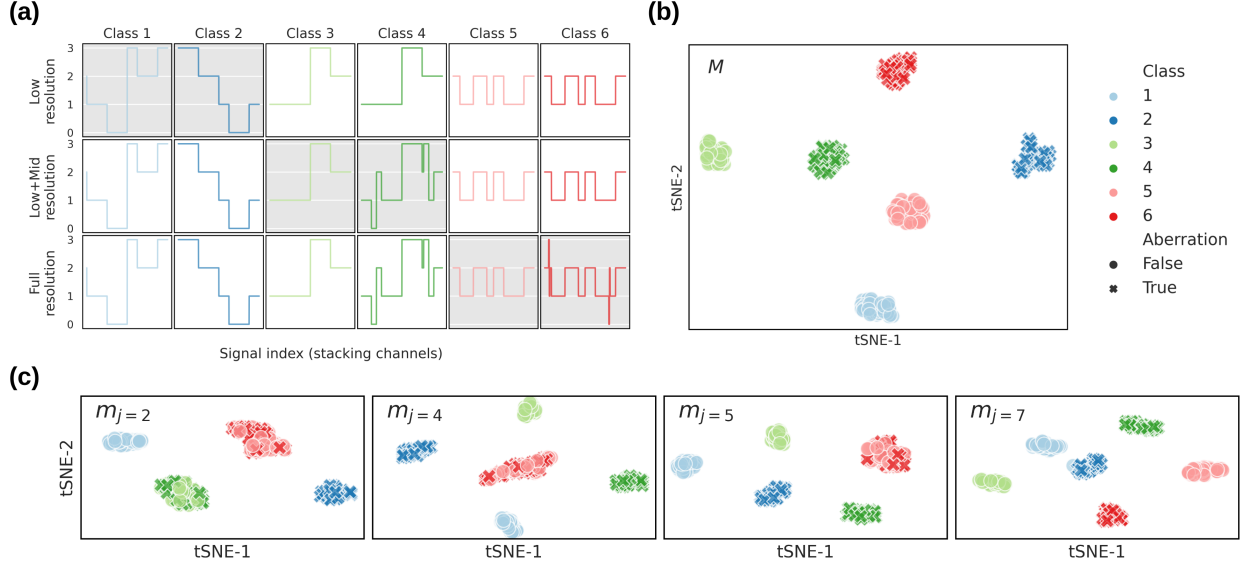


Figure 2: **Illustrative Example.** Source-separated signals are adaptively filtered to obtain scale embeddings. These scale embeddings are then combined to obtain a single scale-attentive representation. (a) Simulated noise-free signals. As the scale increases, more pairs of classes become distinguishable (highlighted). Only the first pair of classes are distinguishable at the low-scale (top row), the second two at medium scale (middle row), whilst the final two become distinguishable only at high scales (bottom row). (b) tSNE projection of our multi-scale embedding (M). (c) tSNE projection of our adaptively learnt scale specific embeddings (m_j).

filters along each channel $x^{(c)}[n]$, and then down-sampling. At the first level, this is then equivalent to performing the convolutions $y_{\text{high}}[n] := (x^{(c)} * w) \downarrow Q = \sum_k x^{(c)}[k]w[Qn-k]$, and $y_{\text{low}}[n] := (x^{(c)} * u) \downarrow Q = \sum_k x^{(c)}[k]u[Qn-k]$ where Q is known as the decimation factor and \downarrow is the subsampling operator.

This obtains detail coefficients $y_{\text{high}}[n] := \alpha_j^{(c)}$, and approximation coefficients $y_{\text{low}}[n]$. As depicted in Figure 3, we repeat this procedure on each subsequent set of approximation coefficients, at each stage obtaining the next set of detail coefficients, $\alpha_j^{(c)}$, and a new set of approximation coefficients. This is performed independently across each input channel, obtaining a sequence of wavelet coefficients for each. This sequence of coefficients details information of increasing scale within a channel, and we index j over these scales. We can use this sequence of coefficients to recover scale specific components of the original input signal using the Inverse Wavelet Transform (IWT). This is achieved by zero-masking the $\alpha_j^{(c)}$ components to obtain $\hat{x}_j^{(c)} = \text{IWT}([\mathbf{0}, \dots, \mathbf{0}, \alpha_j^{(c)}, \mathbf{0}, \dots, \mathbf{0}])$. Repeating this for each channel at scale j , we denote \hat{X}_j as the concatenation of channels $\hat{x}_j^{(c)}$, and obtain a sequence of source separated, scale specific components

$$\{\hat{X}_j\}_{j=1}^J := \left\{ \left[\hat{x}_j^{(1)}, \dots, \hat{x}_j^{(C)} \right] \right\}_{j=1}^J.$$

We can choose J to be either the maximum depth, or a truncation of the sequence which then applies a de-noising effect.

An additional advantage of truncation is that it also applies a form of average-pooling on our input when components $j > J$ are excluded from the IWT (Mallat, 1999). Consequently, each source separated signal produced needs to only be long enough to capture the finest details, specified through the chosen wavelet's decomposition length and the choice of J . For example, given a wavelet with a decomposition length of 2 and a recursion depth of $J = 6$, we can accurately capture artifacts longer than $2^{-J} = 1/64$ -th of the input signal, with each source-separated input component being 64 elements long.

Convolutional-LSTM cell. Long Short-Term Memory (LSTM) networks are a type of recurrent artificial neural network which are specifically designed towards processing sequential data. Shi et al. (2015) introduced a convolutional LSTM network for channelised signals, which is defined through a set of recurrence relations. We extend upon this by adding a convolutional *recurrent projection layer* to the output gate. Our contribution then simply replaces the typical

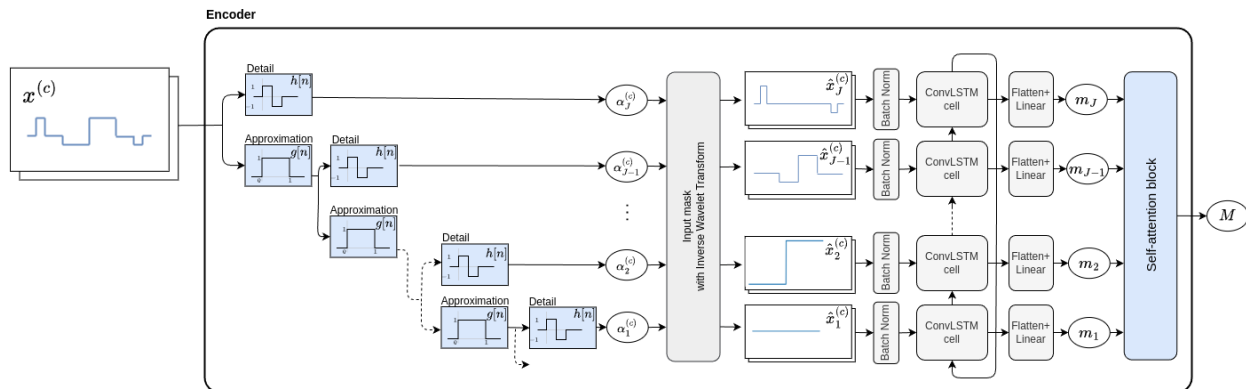


Figure 3: **The Wave-LSTM Encoder.** Source separation is achieved through zero-masking of the multi-resolution wavelet cascading filter bank. This is then adaptively filtered and combined to output a scale-attentive encoding.

output gate with

$$H_j = \tanh(W_{oh} * (o_j \odot \tanh(C_j))), \quad (1)$$

where C_j is the tensor of cell states composed of h_c channels, H_j is the tensor of hidden states now composed of $h_{\text{proj}} < h_c$ channels, o_j is the response of the output gate, whilst W_{oh} and convolutional operator $*$ define the recurrent convolutional projection. This is a modified version to the recurrent projection proposed by Sak et al. (2014), and reduces the number of channels of the hidden short-term memory state vector whilst maintaining the ability to stack LSTM layers. This was originally proposed for large vocabulary models which necessitated a larger long-term memory cell that led to prohibitive memory requirements. With this projection, the cell and hidden state dimensions are uncoupled and the flow of information from coarse-scale embeddings to fine-scale embeddings can be controlled. We couple this with a linear non-recurrent output projection to obtain the *scale embedding*, $m_j = W_j F(H_j)$ where $F(\cdot)$ denotes a flattening layer which concatenates each hidden channel and $\mathbf{m}_j \in \mathbb{R}^D$.

Self-attention block. Having obtained a sequence of scale-specific embeddings, we now aim to find a single multi-scale representation that can be used for prediction tasks. Towards this end we employ the self-attentive embedding of Lin et al. (2017). Originally developed to convert a sequence of word embeddings to a sentence embedding, we instead convert our sequence of *scale embeddings* to a single *multi-scale embedding*. We achieve this by taking a linear combination of the J LSTM hidden vectors in $\mathbf{S} = [\mathbf{m}_1^T, \dots, \mathbf{m}_J^T]$ of size D -by- J , using a self-attention mechanism. The attention mechanism takes the scale embeddings \mathbf{S} as input, and outputs weights

$$\mathbf{a} = \text{softmax}(\mathbf{w}_{s2} \tanh(\mathbf{W}_{s1} \mathbf{S})), \quad (2)$$

where \mathbf{W}_{s1} is a weight matrix of shape d_a -by- D , and \mathbf{w}_{s2} is a row vector of parameters with size 1-by- d_a , and we obtain an attention vector with J elements, whose weights sum to one. This attention vector can explain a singular semantic meaning between different scales through $\mathbf{m} = \mathbf{a} \mathbf{S}^T$. For example, in some outcome prediction tasks in

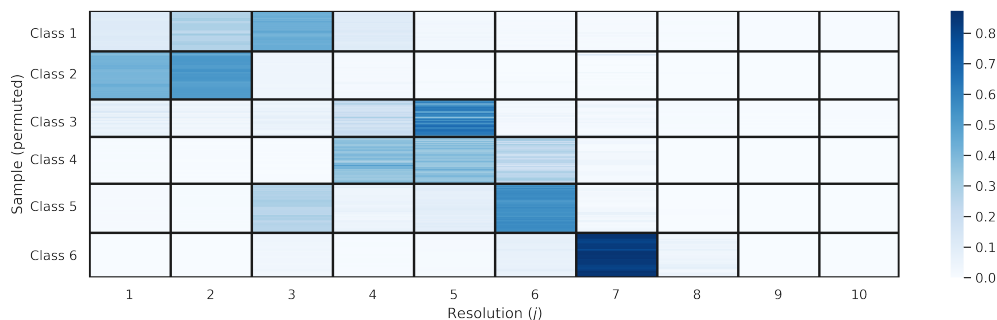


Figure 4: **Self-attention of the Illustrative Example** ($\mathbf{A} \in \mathbb{R}^{1 \times J}$), for each increasing scale. Samples are ordered by label, then permuted via spectral bi-clustering. We observe that attention is given to higher scales in the presence of transient signals such as finer-scale, focal aberrations.

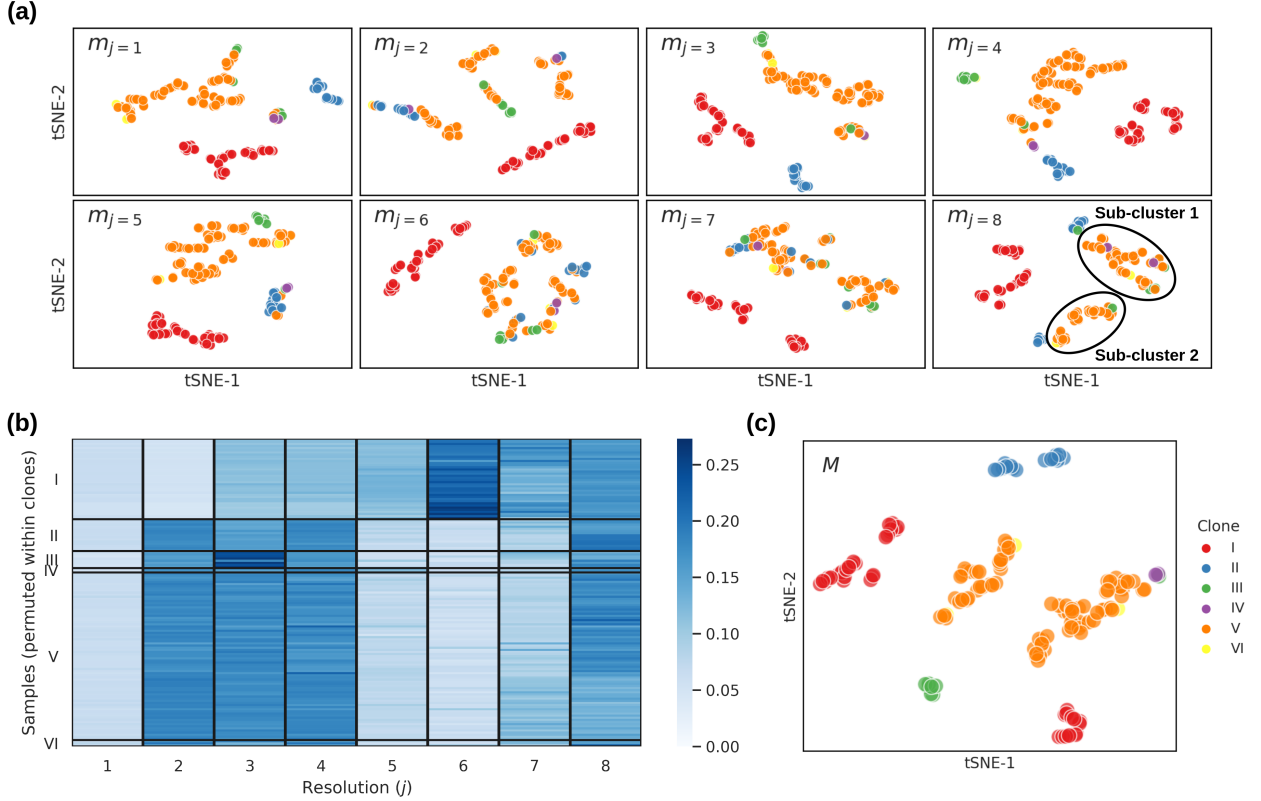


Figure 5: **Single-cell copy number scale embeddings.** (a) t-SNE plots of scale embeddings ($\mathbf{m}_j \in \mathbb{R}^3$). (b) Attention ($\mathbf{A} \in \mathbb{R}^{1 \times J}$), for each class and scale. (c) tSNE projection of multi-scale embedding ($\mathbf{M} \in \mathbb{R}^{1 \times 3}$).

cancer genomics, the presence of whole genome doubling (at a coarsest scale) and a gene mutation (at a fine scale) may have some semantic meaning when occurring together. However, there may exist many meaningful combinations across other scales, and so we extend \mathbf{w}_{s2} to an r -by- d_a matrix \mathbf{W}_{s2} , where r is the maximum number of semantic combinations we can model, and obtain the attention matrix

$$\mathbf{A} = \text{softmax}(\mathbf{W}_{s2} \tanh(\mathbf{W}_{s1} \mathbf{S})). \quad (3)$$

The *multi-scale embedding* matrix for a single sample is then the r -by- D matrix, $\mathbf{M} = \mathbf{A} \mathbf{S}^T$. To utilise this multi-scale embedding, we flatten each row of \mathbf{M} and use this as input to a decoder/task-specific head, training end-to-end. This is demonstrated by our running example using a simple fully connected classifier (previously shown in Figure 2). Figure 4 shows the corresponding attention matrix \mathbf{A} . Notice that attention is paid to resolution 7 for Class 6 which is only different from Class 5 in terms of fine detailed structures. Attention is paid to mid-resolution scales 4/5 for Classes 3 and 4 where mid-sized differences are present. While attention is only needed to be paid to resolutions 1-3 for classes 1 and 2 which differ by large, broad alterations.

where we correctly identify a multi-scale embedding that is attentive to the finer scales in the presence of aberrations and coarse scales otherwise, as shown in .

3 Experiments

Single cell analysis. We first consider the task of detecting clonal structures from single-cell profiling of copy number alterations. Cancers evolve through an iterative process whereby successive generations of cancer cells can accumulate somatic mutations. Over time, this process leads to related but genetically distinguishable descendant sub-populations within a single tumour. These are referred to as sub-clones, and their collective clonal structure reveals key aspects of tumour development which can be useful for targeted treatment and early intervention.

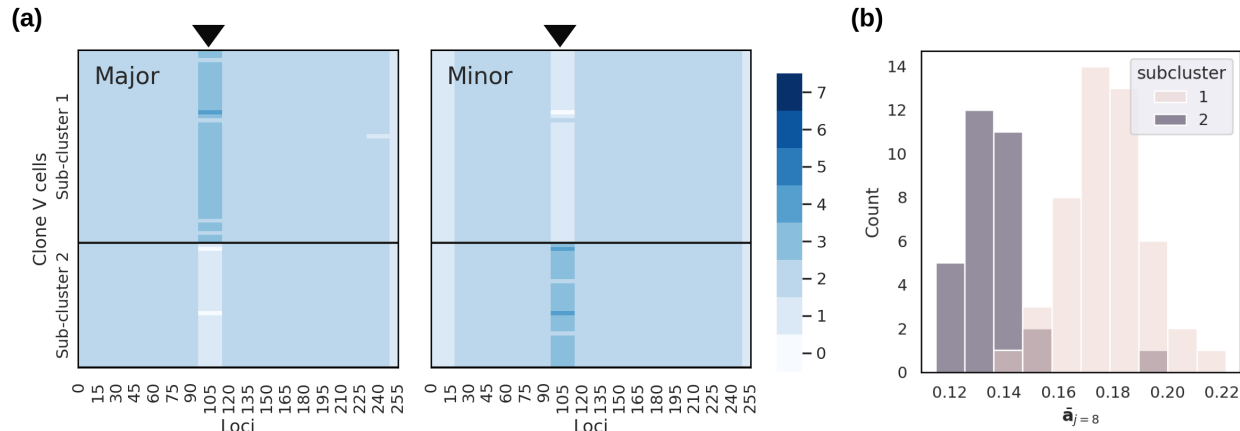


Figure 6: **Novel sub-clone discovery.** (a) Copy number profile of clone V’s major (left) and minor (right) strands of chromosome 1. Samples are split into two sub-clusters in the embedding $h_{j=8}$ which are segregated by a systematic copy number difference at a single locus indicated by the black arrows. (b) The amount of attention placed on the $j = 8$ scale differs for each sub-cluster.

We obtained a 10x Genomics Chromium single-cell DNA sequencing dataset which contains cells obtained from frozen breast tumour tissue from two patients, at five tumour sections and sequenced with a low coverage between 0.02x and 0.05x¹. However, for our experiments we follow previous work and consider only site E of patient S0, comprising of 2,075 sequenced cells. Zaccaria and Raphael (2021) used the CHISEL algorithm to allocate 1,448 of the 2,075 tumour cells to 6 clones, including one diploid clone (labelled I) comprising mostly normal cells, and 5 aneuploid clones which contain varying degrees of copy number alterations (labelled II-VI) (the remaining cells were unclassified and we excluded them from the analysis).

Using the CHISEL labels as an *assumed* soft ground-truth², we compared these labels to the embeddings obtained on an auto-encoding task using the Wave-LSTM encoder. A similar task was also performed by Yu et al. (2023) using a Convolutional Neural Network (CNN) auto-encoder, and we use the same decoding architecture for comparison. Following Yu et al. (2023), we encode to 3-dimensional embedding spaces, s.t. $\mathbf{m}_j \in \mathbb{R}^{D=3}$, and use a single attention hop, s.t. $\mathbf{M} \in \mathbb{R}^{1 \times 3}$. Additionally, we choose to stack chromosomes such that $W = 256 \times 22$ (using only autosomes), $C = 2$, and choose $J = 8$. This is both in keeping with Yu et al. (2023), and allows us to view coarser scale embeddings.

Two-dimensional t-SNE plots of the three-dimensional scale and multi-scale embeddings are shown in Figure 5a and 5c, respectively, coloured by CHISEL labels. Scales from $j = 1$ to $j = 5$ showed latent structures (clusters) that are broadly consistent with the CHISEL clonal structure but sub-structures not previously captured by CHISEL or rCAE also emerge. For example, consider clone V at $j = 8$ which is split into two groups. In Figure 6a, we observe that this is due to the cells in clone V containing a small copy number gain on opposite strands. This separation is also reflected in the self-attention mechanism, shown in Figure 5b where cell samples are permuted through clustering of the attention matrix, and in Figure 6b where we see each sub-cluster places a different level of attention on this scale. This is also clearly demonstrated in clone II at scale $j = 6$, which is grouped into two clusters. Upon inspection, this distinction is based on deletions on chromosomes 6, 17 and 22.

Overall, Wave-LSTM is able to recover clonal patterns from single-cell sequencing data in an unsupervised fashion that are broadly consistent with current state-of-the-art methods such as CHISEL. However, it also provides enhanced exploratory capabilities to look at multi-scale structures which identified novel lineages of cancer cells.

Cancer survival prediction. Copy number alterations are known to be predictive of survival outcome in cancer (Hieronymus et al., 2018). In this section we consider the use of Wave-LSTM as a means of providing multi-scale information from copy number profiles, with the goal of predicting cancer survival outcomes.

The inclusion of CNAs in survival modelling can be challenging due to the multi-channelled, spatially correlated sequence data. We used Wave-LSTM to encode the CNA profiles as an input into a deep survival model, DeSurv (Danks and Yau, 2022), alongside other features, to predict survival outcomes. Note that the choice of DeSurv is for convenience as it is a flexible neural network based approach for modelling survival distributions in continuous time.

¹URL: <https://github.com/raphael-group/chisel-data>

²There is no objective ground truth in such settings.

Study	Encoder	$C_{td}(\uparrow)$	IBS (\downarrow)	INBLL (\downarrow)
Simulation	Average CN	0.54 ± 0.004	0.16 ± 0.002	0.49 ± 0.004
	rcCAE	0.71 ± 0.003	0.06 ± 0.003	0.23 ± 0.040
	LSTM	0.68 ± 0.002	0.06 ± 0.004	0.22 ± 0.010
	Wave-LSTM	0.72 ± 0.005	0.04 ± 0.006	0.14 ± 0.017
TCGA	Average CN	0.62 ± 0.002	0.13 ± 0.002	0.40 ± 0.005
	rcCAE	0.75 ± 0.011	0.11 ± 0.007	0.34 ± 0.016
	LSTM	0.72 ± 0.007	0.11 ± 0.005	0.34 ± 0.015
	Wave-LSTM	0.78 ± 0.008	0.10 ± 0.010	0.31 ± 0.024

Table 1: **Survival performance metrics**, with one standard deviation across five random seeds.

Cox-based survival models (e.g. PyCox (Kvamme et al., 2019)) or alternative neural approaches (e.g. SumoNet (Rindt et al., 2022), Neural Fine-Gray (Jeanselme et al., 2023)) could also be used.

Simulation study. We simulate two cancer types, each described by a different underlying linear Cox proportional hazard models with Gompertz distributed survival functions. Consequently survival times are generated by

$$T = \frac{1}{\beta} \ln \left(1 - \frac{\beta \ln(u)}{\alpha \exp\{\sum_i X_i\}} \right),$$

where u is a random variable uniformly distributed on the unit interval, and $\{X_i\}_{i=1}^{12}$ is the sequence of covariates, including standard Gaussian age, Bernoulli distributed gender, and 10 uniformly distributed insertion locations of equal width across 2 channels. Using this model we sample survival outcomes originating from two cancer types, described by $\alpha = 0.1$ and 1.5 respectively. For each, the rate of censoring was kept equal and we choose $\beta = 1$. For us to test the capacity of Wave-LSTM to identify small aberrations, the CNA profiles from one model were given an additional CNA gain event distinguishing it from the other cancer type. These were placed in the same position, at $1/64$ of a channel width. Consequently, the only information which then distinguishes cancer types is this additional gain event.

We compared predictive performance to three baselines. The first was a simple baseline summary statistic, the genome-wide average copy number, which has been historically used as a measure of overall tumour genomic complexity (Average CN). Additionally, we also compared to a CNN encoder architecture of Yu et al. (2023) (rcCAE), and an LSTM-based encoder in which recurrence is over loci. These alternative baselines account for the sequential and spatially correlated nature of the data but not the multi-scale aspect. All use the same DeSurv survival model. We again reiterate and emphasise that we are unaware of *any* previously developed representation learning approach for copy number data in the literature. Hence, we created these baselines to test the added benefit of the wavelet and attention components of Wave-LSTM over these simpler encoding mechanisms.

We report three metrics: Time-dependent Concordance Index (C_{td} , higher is better) measures how well a model distinguishes risk between samples at different times. The Integrated Brier Score (IBS, lower is better) measures if the model can accurately predict the probability of an event occurring over a period, whilst the Integrated Negative Binomial Log-Likelihood (INBLL, lower is better) measures how well the model prediction matches the actual occurrence of the event.

Table 1 shows that the predictive performance of Wave-LSTM exceeds all other baselines on all three measures. In particular, the IBS and INBLL measures were significantly improved, indicating that the calibration of Wave-LSTM predictions was superior.

The Cancer Genome Atlas We next obtained copy number data from multiple cohorts from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) to test the utility of multi-scale representations in improving survival prediction accuracy. Specifically, we consider the survival outcomes of patients with Thyroid Cancer (THCA), Breast Cancer (BRCA), Ovarian cancer (OV), Glioblastoma Multiforme (GBM), and Head and Neck Squamous cell Carcinomas (HNSC), listed in order of typically decreasing expected survival outcome. In total, we examined 3021 samples. This selection was used as a control to determine if Wave-LSTM was able to recover latent quantities reflecting this survival heterogeneity. Copy number profiles were pre-computed using Allele-Specific Copy number Analysis of Tumors (ASCAT) (Van Loo et al., 2010)³. In this example we channelise chromosomes.

The learnt multi-scale representations are plotted in Figure 7 using kernel density estimation and grouped by cancer type. We observe increasing values correspond to worsening survival outcome showing that Wave-LSTM was able to identify a latent dimension associated with cancer survival. Additionally, we find two modes to thyroid cancer, and

³URL: https://github.com/VanLoo-lab/ascat/tree/master/ReleasedData/TCGA_SNP6_hg19

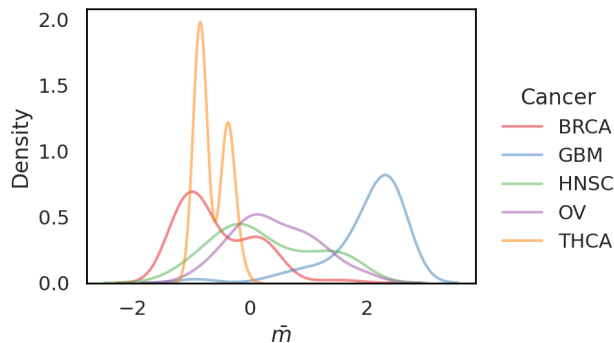


Figure 7: **The Cancer Genome Atlas Survival analysis.** Wave-LSTM maps different cancer types along the multi-scale embedding dimension (\bar{m}) in order of expected survival rates. Note that some cancer types (e.g. thyroid cancer, THCA; breast cancer, BRCA) appear multimodal in distribution which corresponds to different molecular subtypes within the respective cancer type.

upon inspection we identified that Wave-LSTM had identified the previously known prognostic significance of male gender which is associated with poorer survival outcomes in papillary thyroid cancer than females (Liu et al., 2017).

We again compared our approach to alternative encoder baselines in Table 1, where the Wave-LSTM encoder significantly outperforms the competing encoding approaches on all scores, suggesting that its specific design is encapsulating more information from the data.

4 Discussion

Within this work we introduced Wave-LSTM, a multi-scale signal representation learning module, motivated by the analysis of cancer sequencing data in which mutations may occur across entire chromosomal regions, or isolated to a small number of genes. Wave-LSTM decomposes signals into multiple frequency bands, producing a series of band-pass outputs which are able to represent local transient structures at different scales. For each, we sequentially learn a scale-specific embedding using a novel LSTM framework. These are then combined using a self-attention mechanism to obtain a multi-scale embedded representation which can then be used in downstream tasks.

Alongside a demonstrative classification example, two examples analysing cancer sequencing data are presented. We first validated Wave-LSTM’s ability to recover genetically distinguishable descendant sub-populations, recovering clonal patterns that are consistent with state-of-the-art specialist bioinformatics methods, whilst providing enhanced exploratory capabilities which identified novel lineages of cancer cells. Second, we used Wave-LSTM to learn multi-scale information of copy number profiles with the goal of predicting cancer survival times. We demonstrated significant improvements on a number of benchmark metrics against competing methods.

Alternatives to the LSTM architecture could be employed, such as Transformers (Vaswani et al., 2017) or more recently introduced alternatives such as selective state-space models (Gu and Dao, 2023). These architectures have demonstrated powerful capabilities for a number of applications and data types. However, the complexity of these models relative to our LSTM architecture may preclude their use in genomic data settings where sample sizes are substantially more limited than in natural language, image and video data applications. Wave-LSTM can also be generally applied to obtain multi-scale embeddings of complex structured, high-dimensional multi-channel input objects. We have tested its utility for genomic copy number analysis but its wider utility for other data types has not been examined here.

Acknowledgements

CG and CY are supported by an EPSRC Turing AI Acceleration Fellowship (Grant Ref: EP/V023233/1).

References

- Rameen Beroukhi, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010.
- Craig M Bielski, Ahmet Zehir, Alexander V Penson, Mark TA Donoghue, Walid Chatila, Joshua Armenia, Matthew T Chang, Alison M Schram, Philip Jonsson, Chaitanya Bandlamudi, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics*, 50(8):1189–1195, 2018.
- Dominic Danks and Christopher Yau. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 7240–7256. PMLR, 2022.
- Ruben M Drews, Barbara Hernando, Maxime Tarabichi, Kerstin Haase, Tom Lesluyes, Philip S Smith, Lena Morrill Gavarró, Dominique-Laurent Couturier, Lydia Liu, Michael Schneider, et al. A pan-cancer compendium of chromosomal instability. *Nature*, 606(7916):976–983, 2022.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, 2014.
- Haley Hieronymus, Rajmohan Murali, Amy Tin, Kamlesh Yadav, Wassim Abida, Henrik Moller, Daniel Berney, Howard Scher, Brett Carver, Peter Scardino, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife*, 7:e37294, 2018.
- Vincent Jeanselme, Chang Ho Yoon, Brian Tom, and Jessica Barrett. Neural fine-gray: Monotonic neural networks for competing risks. In *Conference on Health, Inference, and Learning*, pages 379–392. PMLR, 2023.
- Havard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30, 2019.
- Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
- Chunping Liu, Tianwen Chen, Wen Zeng, Shuntao Wang, Yiquan Xiong, Zeming Liu, and Tao Huang. Reevaluating the prognostic significance of male gender for papillary thyroid carcinoma and microcarcinoma: a seer database analysis. *Scientific Reports*, 7(1):11412, 2017.
- Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.
- David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1190–1205. PMLR, 2022.
- Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.
- Laurent Sansregret, Bart Vanhaesebroeck, and Charles Swanton. Determinants and clinical implications of chromosomal instability in cancer. *Nature Reviews Clinical Oncology*, 15(3):139–150, 2018.
- Xingjian Shi, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015.
- Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, and Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113, 2013.
- Zhenhua Yu, Furui Liu, Fangyuan Shi, and Fang Du. rccae: a convolutional autoencoder method for detecting intra-tumor heterogeneity and single-cell copy number alterations. *Briefings in Bioinformatics*, page bbad108, 2023.
- Simone Zaccaria and Benjamin J Raphael. Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature Biotechnology*, 39(2):207–214, 2021.