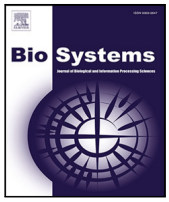
BioSystems 235 (2024) 105095

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems



SetQuence & SetOmic: Deep set transformers for whole genome and exome tumour analysis

Neringa Jurenaite [a,*,1], Daniel León-Periñán [a,b,1], Veronika Donath [a], Sunna Torge [a], René Jäkel [a]

[a] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), TU Dresden, Chemnitzer Str 46b, Dresden, 01187, Saxony, Germany
[b] Max-Delbrück-Centrum für Molekulare Medizin, Hannoversche Str. 28, Berlin, 10115, Germany

ARTICLE INFO

Dataset link: https://github.com/danilexn/setquence

Keywords:
Multi-omics
Language model
Deep neural network
Set representations
Whole-genome motifs

ABSTRACT

In oncology, Deep Learning has shown great potential to personalise tasks such as tumour type classification, based on per-patient omics data-sets. Being high dimensional, incorporation of such data in one model is a challenge, often leading to one-dimensional studies and, therefore, information loss. Instead, we first propose relying on *non-fixed sets* of whole genome or whole exome variant-associated sequences, which can be used for supervised learning of oncology-relevant tasks by our Set Transformer based Deep Neural Network, SetQuence. We optimise this architecture to improve its efficiency. This allows for exploration of not just coding but also non-coding variants, from large datasets. Second, we extend the model to incorporate these representations together with multiple other sources of omics data in a flexible way with SetOmic. Evaluation, using these representations, shows improved robustness and reduced information loss compared to previous approaches, while still being computationally tractable. By means of Explainable Artificial Intelligence methods, our models are able to recapitulate the biological contribution of highly attributed features in the tumours studied. This validation opens the door to novel directions in multi-faceted genome and exome wide biomarker discovery and personalised treatment among other presently clinically relevant tasks.

1. Introduction

The 40 trillion cells that constitute human bodies follow precise instructions coded in their (mostly identical) genomes, focusing on certain regions in order to meet unique, specific roles (Bianconi et al., 2013). Multiple factors can alter this information in a specific tissue, leading to a spectrum of cancer diagnoses (Hirata and Sahai, 2017). Deep learning (DL) methods, adopted from Natural Language Processing (NLP), support a better understanding of biological sequences (Elnaggar et al., 2021; Ji et al., 2021). In this work, we explore genome sequences by applying NLP-inspired DL methods on mutomes (variants, or genome sequences representing differences between a patient's genome and a reference) and transcriptome counts (expression quantification of RNA-sequencing data), and their corresponding clinical annotations across 33 tumour types.

The main challenges which hinder the clinical applicability of Deep Neural Networks (DNN) on omics data-sets are the large amount of model parameters, their limiting methodology, and lack of model transparency. First, we address the issue of the amount of parameters by using somatic variants, a subset of sequences from the genome. Second, they consider omics as fixed-dimensionality inputs (i.e., messenger RNA (mRNA) expression matrices, DNA methylation profiles, or Single Nucleotide Polymorphisms (SNP) arrays). If the number of experimentally measured and quantified *loci* increases, the new input dimension most likely does not fit to the model deeming the model unusable without manual intervention. Moreover, the hidden and output states of such models are sensitive to input ordering and require adaptation of specific pre-processing pipelines. We attempt to overcome this issue by allowing a *set* representation for input data. Third, to incorporate larger datasets or non-coding data into our model, the computational complexity of the model has to be addressed. We do so by making architectural changes, such as knowledge distillation, to the model with the focus on robustness and accuracy metrics of the model in mind. Fourth, the interpretability and transparency of DNNs is scarcely explored in previous research, regardless of how well the DNNs perform under controlled testing conditions (Tjoa and Guan, 2021). The need for explainability and robustness investigations of biomedical AI models has been highlighted in previous studies, in particular to progress the newly-introduced model's usability for it is intended clinical application (Tran et al., 2021; Amann et al., 2020). We undertake this

* Correspondence to: ScaDS.AI, Technische Universität Dresden, Bürokomplex Falkenbrunnen, Chemnitzer Str. 46b, 01187, Dresden, Germany.
E-mail addresses: neringa.jurenaite@tu-dresden.de (N. Jurenaite), daniel.leonperinan@mdc-berlin.de (D. León-Periñán), veronika.donath@tu-dresden.de (V. Donath), sunna.torge@tu-dresden.de (S. Torge), rene.jaekel@tu-dresden.de (R. Jäkel).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.biosystems.2023.105095
Received 15 January 2023; Received in revised form 17 October 2023; Accepted 28 November 2023
Available online 6 December 2023

challenge by exploring Explainable Artificial Intelligence (XAI) method applications which yield an increase of trust in DNNs. Finally, models are rarely generalised. We tackle this by linking multiple (qualitative and quantitative) omic sources into a single, multi-dimensional model and making the model applicable for use cases such as cancer type, and therefore cancer sub-type, classification, among other downstream tasks.

In this feasibility study, we explore the advantages of applying NLP-based DL methods on multi-omics data as an alternative method to existing techniques, with a focus on their explainability. We concentrate on model robustness (minimal changes in input resulting in minimal changes in output), reproducibility (with other data), immutability (with increase in data), and interpretability (input explainability). In this study, we provide a proof of concept, showing that attention-based approaches accurately extract features from omic datasets. More specifically, we introduce SetQuence, a DNN built on set representations of mutomes which encodes variably many, arbitrarily long sequences in a permutation invariant manner. This is done on a small, publicly available *The Cancer Genome Atlas* (TCGA) dataset containing somatic variants data of only coding regions. We show the biological meaningfulness of our model using attribution methods and therefore provide clinically-relevant, nucleotide-level explanations of applicability of our models in cancer research.

Furthermore, larger and noisier data which includes unfiltered, Whole-Genome somatic mutation data provides more detailed insight into the tumour types being studied than coding-only somatic variant TCGA data. Therefore, using the Catalogue Of Somatic Mutations In Cancer (COSMIC) dataset, we are working with more than four times the patient samples and investigate how the non-coding genome influences the classification of tumour types. We do so by first architecturally optimising the original implementation of SetQuence to reduce its computational limitations and to improve usability of the model in comparison to other approaches to answer multiple clinically relevant questions. This enabled training our model on larger omics datasets, improving tumour type classification metrics by including non-coding variants into the analysis and exploring the robustness of the model on multiple clinical applicabilty tasks.

We further generalise the network and data representation to incorporate transcriptome counts in SetOmic, a DNN which can integrate different patient-wise omics sources. Thanks to set representations and our architectural approach, our approach allows the incorporation of more granular data compared to state-of-the-art, the applicability on multiple downstream tasks, and offers nucleotide-level insights into cancer-associated motifs, while showing improved robustness and comparable model performance to state-of-the-art approaches. We would expect that with additional datasets, our results would improve even further, such as via integration of methylation data, see Section 8.

In Section 2, related work is discussed. The background behind our approach is described in Section 3. In the following Section 4 the methods detail the construction of the SetQuence model, its optimisation strategies we have chosen, the SetOmic model, and the explainability approaches. Architectural model results are presented in Section 5 and the clinical applicability results in 6. Both results sections are discussed in Section 7 and the paper concludes with an overview of possible future directions is Section 8.

2. Related works

Both supervised and unsupervised Machine Learning (ML) methods have been widely used for omics data analysis (Remli et al., 2017; Ming et al., 2019; Petegrosso et al., 2019; Gal et al., 2020) and integration of multiple omic sources into a single model has shown improved accuracy in multiple oncology-relevant classification tasks (Picard et al., 2021; Sharifi-Noghabi et al., 2019). More recently, Variational Autoencoder (VAE) architectures such as DeepT2Vec (Yuan et al., 2020) helped to better recapitulate tumour expression patterns compared to previous clustering techniques. Thus, an initial unsupervised training phase helps to better capture non-linear interactions in omics (Mazlan et al., 2021), which then are fine-tuned in a supervised manner for a downstream task. This strategy was successfully implemented by the VAE-based OmiEmbed architecture (Zhang et al., 2021), a state-of-the-art multi-omics hybrid method with a tumour type prediction accuracy ~96%. However, OmiEmbed's input format is a gene-level fixed-dimensionality vector input, where the pipeline is set up for a specific set of genes and is, therefore, sensitive to input ordering.

Hybrid training of VAEs for multi-omics integration is a reoccurring strategy in DL for cancer omics (Zhang et al., 2019, 2021; Simidjievski et al., 2019). However, comparable models such as OmiEmbed (Zhang et al., 2021), either do not consider sequence-level Whole Exome and Whole Genome Sequence (WES and WGS respectively) data at all, or map coding variants to a quantitative representation, such as CPEM (Lee et al., 2019b). These approaches, firstly, lose the granularity for the sake of reducing the curse of dimensionality of multi-omics data, secondly, do not incorporate non-coding data which has been proved to provide important insights into tumour activity (Ling et al., 2015), and thirdly, do not consider variant-associated sequences in multi-omics settings.

As an alternative, learnable feature extractors (e.g., via CNNs, Recurrent Neural Networks (RNN) and Neural Embedding methods Shaheen et al., 2016; Young et al., 2018) are being gradually adopted for omics data, via Language Models (LMs) (Song et al., 2021). In Natural Language Processing, attention-based networks (i.e., the Transformer Vaswani et al., 2017a) are used for pre-training and fine-tuning LMs (Devlin et al., 2019a; Dai et al., 2019), from large text corpora. In an omics setting, the attention-based networks (i.e., the Transformer Vaswani et al., 2017a) enables a more efficient parallel capture of long range interactions (Bahdanau et al., 2015) between features. DNABERT (Ji et al., 2021), a BERT-based LM pre-trained on human genomes, encodes DNA sequences to predict regulatory motifs and detect genomic variants (among other tasks) and outperforms previous baselines such as CNNs, which we therefore choose as our *encoder*.

Mainly, three types of genome or exome representations have been explored in the past: (**i**) per-patient lists of IDs such as a list of mutated genes, (**ii**) one-hot disease-associated vector representations, and (**iii**) using features appended to the one-hot vectors in **ii**. In **i**, the presence of an ID indicates that a disease associated event, such as a mutation, was observed at this *locus* (location within a genome), as explored in Mut2Vec (Kim et al., 2018) and *Genome Impact Transformer* (GIT) (Tao et al., 2019). GIT, as a baseline, is comparable to our network since it considers the input as a set representation (of gene IDs), while also being based on multi-head attention. However, multiple events (e.g., mutations) are left unrepresented if they occur at the same *locus*. This challenge is overcome by using (**ii**) one-hot vector representations which indicate the presence/absence of M number of events (e.g., mutations), where M is the number of all previously observed events in a database (usually used together with Convolutional Neural Network (CNN) architectures which are mostly limited to local pattern detection) (Yuan et al., 2016; Chakraborty et al., 2021a). (**iii**) uses features, such as mutation rates, appended to the one-hot vectors (e.g., *Cancer Predictor using an Ensemble Model* (CPEM) Lee et al., 2019b), which, similarly to **ii**, are restricted to mutation events that are known *a priori*, precluding generalisation to unobserved events (i.e., to the largely unexplored non-coding regions of the genome). Also, the sparsity of vectors in **ii-iii** might be a major obstacle when training such models (Evci et al., 2019).

In the past, Marquard et al. (2015) introduced *TumorTracer* along with feature extractors Φ that yield, apart from one-hot encoding of variants, three features from prior knowledge of genomes: (**a**) non-synonymous mutations (i.e., point mutation status of annotated genes); (**b**) the base substitution frequency (i.e., the frequency of 6 possible SNPs (C > A, C > G, C > T, T > A, T > C and T > G)); and (**c**) the trinucleotide base substitution frequency (i.e., the relative frequency

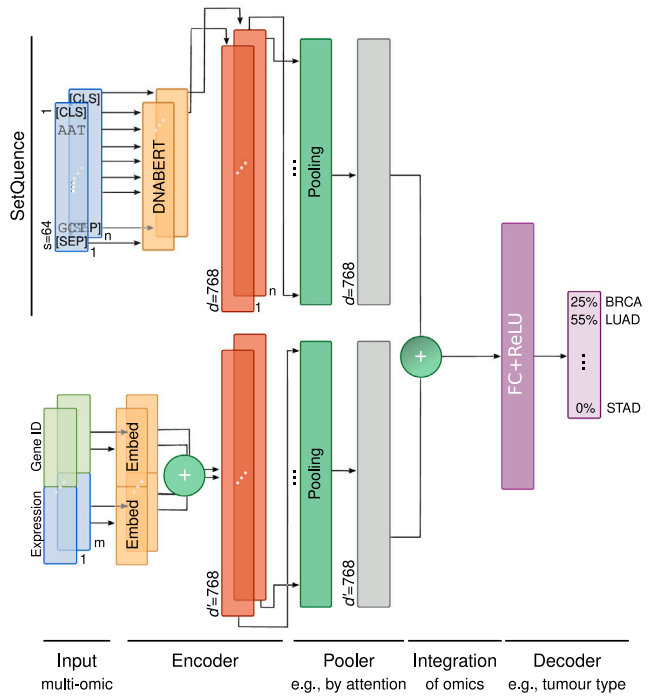


Fig. 1. Architecture of SetQuence and SetOmic, where SetQuence, built of set representations of variant-associated sequences, is part of the architecture of SetOmic, built for multiple quantitative and non-quantitative omic sources.

of each of the substitutions in **b** surrounded by a sequence context (96 possibilities of 3 bp long)). In practice, *TumorTracer* is Random Forest (RF) trained on features **a–c**, able to yield an average ~70% prediction accuracy on 10 tumour types from the TCGA pan-cancer cohort, compared to the < 30% accuracy of using one-hot encoding using architecturally similar RFs. Later, Lee et al. (2019b) introduced with *Cancer Predictor using an Ensemble Model* (CPEM) an additional feature, namely, (**d**) mutation rates (i.e., number of SNPs, *indels*, or mutated genes per genomic region). This, together with the use of Neural Network (NN) and RF ensembles, improved model quality compared to *TumorTracer* and other prior methods. More recently, Chakraborty et al. (2021b) proposed a Hierarchical Bayesian perspective based on all these previous features (**a–c**), plus (**e**) regional mutational density (RMD) (i.e., the regional densities of mutations at contiguous regions along each chromosome – previously introduced by Soh et al. (2017)). In general, **a–e** are known as *hidden genome* features, which together improve classification quality with respect to previous individual features. In particular, Chakraborty et al. (2021b) showed that each feature had different importance for tumour type classification from coding and non-coding genome variants.

3. Background

3.1. Language models

Language Models (LMs) assess the conditional probabilities of words in a language (e.g., given $n-1$ previous (and following) words in a sentence by imposing the nth order Markov property Shannon, 1948). Current LMs are built upon NNs which build an embedding space of the vocabulary, such that *words* are mapped into a lower-dimensional latent semantic space containing the *statistical properties* and *semantics* of languages (Wang et al., 2019). In the case of LMs of biological sequences, *semantic* refers to the underlying, low-level functional or physiochemical properties of biological sequences, the so-called *language of life* (Elnaggar et al., 2021). Recent LMs of biological sequences are inspired by their NLP counterparts: LMs are obtained via unsupervised pre-training of a DNN architecture on large amounts of text data, as in the case of *BERT* (Devlin et al., 2019b), *T5* (Raffel et al., 2020), or *GPT-3* (Brown et al., 2020). In a biological context, this translates into training upon large collections of biological sequences, at omics scale. This is the approach followed for *DNABERT* (Ji et al., 2021) and *ProtTrans* (Elnaggar et al., 2021), LMs of DNA and protein sequences, respectively.

3.2. Attention is all genomes need

Most recent LMs are built on the *Transformer* (Vaswani et al., 2017b; Devlin et al., 2019b), thus, centrally based on the idea of the *attention* mechanism, to efficiently capture pairwise long-range interactions between input or latent space vectors (e.g., of word embeddings in a sentence Bahdanau et al., 2015). Attention maps a query Q and a set of key–value pairs K, V, as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

where $Q, K, B \in \mathbb{R}^{N\times d}$ (i.e., across N-many d-dimensional vectors (in the case of self-attention, $Q = K = V$)). Attention values are weights for pairwise interactions, computed via the dot product QK^T; thus, higher-level interactions are captured by stacking attention blocks. In practice, attention yields more expressive latent spaces for a wide variety of tasks compared to CNN and RNN counterparts (Vaswani et al., 2017b; Chaudhari et al., 2021).

3.3. Set neural network

An architecture which can simultaneously integrate multiple omic sequences as sets is a Set Neural Network (SNN). Invariance, with respect to input ordering and cardinality, is ensured via a *pooling* operation that reduces sets to fixed representations (Lee et al., 2018). Apart from the practical benefits for input flexibility, SNNs have been shown to improve model generalisation and robustness (Tang and Ha, 2021). Hence in this paper, we will extend upon SNNs.

4. Methods

4.1. Data and its representation

We refer Appendix A for the detailed description of the data and its pre-processing together with the way it is represented in our model. A total of 32 tumour classes were investigated.

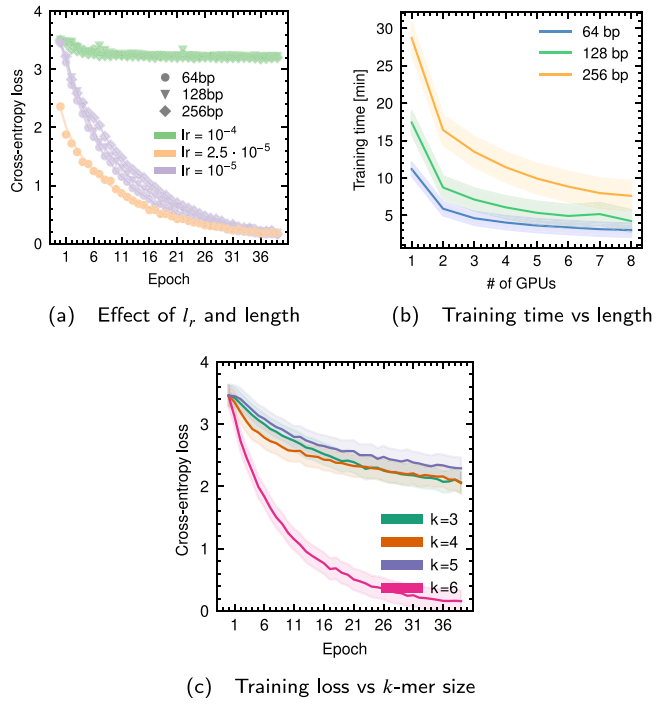
4.2. SetQuence : Set representations of sequences

SetQuence is a Neural Network $f_\theta : \mathbb{N}^{n\times l} \to \mathbb{R}^c$, where n is the number of sequences for a patient, l is the maximum length of input sequences represented as integer tokens, and c is the number of output classes ($c = 1$ for binary classification; or $c > 1$ for multi-class classification), defined as Eq. (2). SetQuence constructs fixed-dimensionality representations from a set $\{x_{i,j}\}_{j=1}^{n_i}$, consisting of as many n_i sequences as somatic variants are called for a sample i (e.g., a patient's tumour).

$$f_\theta(\cdot) = f_{\text{decoder}}(f_{\text{pooler}}(f_{\text{encoder}}(\cdot))) \tag{2}$$

Given a patient's mutome, SetQuence (Fig. 1) uses DNABERT (Ji et al., 2021) to *encode* each of the n_i *variant-associated* sequences into as many d-dimensional vectors $e_{i,j} \in \mathbb{R}^d$ which correspond to the output of the [CLS] token.

Subsequently, a *pooling* operation reduces the set of sequence encodings $\{e_{i,j}\}_{j=1}^{n_i}$ into a single representation, that is, a single d-dimensional vector that can be used for any downstream task, independent of n_i.



(a) Effect of l_r and length (b) Training time vs length

(c) Training loss vs k-mer size

Fig. 2. Effect of training hyperparameters in model and computational performance. In (a), per-epoch cross-entropy loss is assessed for different learning rates (l_r) and sequence lengths (s) in base-pairs (bp), across 40 training epochs. In (b), the average time per epoch (and 95% confidence interval) are assessed for training distributed across 1 to 8 GPUs, for different sequence lengths (s). In (c), the average (and 95% confidence interval) cross-entropy training loss per epoch are assessed for different k-mer sizes, $k \in 3, 4, 5, 6$.

Lastly, a decoder module then projects the d-dimensional $\hat{e}'_i$ into an output space to perform a downstream task (e.g., multi-class or binary classification of tumour types).

Learning rates (l_r) had the greatest impact on training convergence, see Fig. 2(a). Choosing $l_r \in [10^{-5}, 2.5 \cdot 10^{-5}]$ for SetQuence is shown to provide the lowest cross-entropy loss after 40 epochs which we chose for further study. Changes in sequence length (for $s \geq 64$), have little impact on loss curves during training. Choosing $l_r \geq 10^{-4}$ precluded learning, possibly due to catastrophic forgetting of DNABERT's parameter space, a commonly reported issue when fine-tuning BERT-based models (Xu et al., 2020).

We additionally assessed the average training time per epoch on 1 to 8 GPUs for different learning rates and sequence lengths (s) across 40 training epochs, see Fig. 2(b). With a 95% confidence interval, sequence length $s = 64$ showed the lowest training times for any number of GPUs in a single node which we chose for further study.

The k-mer length $k = 6$ was chosen throughout this study, since it provided the lowest cross-entropy loss after 40 epochs, see Fig. 2(c). This agrees with the exponential increase of dictionary size, scaling as 4^k.

We compared multiple *pooling* strategies, which are input permutation independent: (*maximum*, *minimum* or *mean* pooling, and pooling by attention).

In the case of pooling by attention, encodings are subjected to Multi-Head Self-Attention (SAB) to assess interactions between elements of the set and are further reduced into a single vector by multi-head attention (PMA) (Lee et al., 2018); The SAB operation has a complexity $\sim O(n_i^2)$, which might be a limitation for large mutomes, therefore we also explored Induced Set Attention Block (ISAB) (Lee et al., 2018) of much lower time complexity.

Training and test data, were structured upon the `hdf5` standard (HDF Group, 1997), and these `hdf5` files are asynchronously parsed as a GPU-compatible PyTorch Tensor. Patient labels are parsed only once when instancing `Dataset`, into an 8-bit integer tensor with shape (P), where P is the number of patients. Therefore, for variant data, a tuple of `torch.tensor` structures is generated for the requested sample.

4.3. Implementation of optimisation strategies

4.3.1. Knowledge distillation

The protocol for model knowledge distillation introduced by Sanh et al. (2019) was adapted to the DNABERT (Ji et al., 2021) model (teacher). It was performed over 1–11 Transformer Encoder blocks (students), compared to the 12 blocks in the teacher model.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{distil}} + (1 - \alpha) \cdot \mathcal{L}_{\text{MLM}} \tag{3}$$

Student models were trained to minimise the loss function $\mathcal{L}$ in Eq. (3) with a linear combination parameter $\alpha = 0.5$, temperature $T = 2$, cross-entropy loss for Masked Language Modelling loss ($\mathcal{L}_{\text{MLM}}$), and *Kullback–Leibler* divergence for distillation loss ($\mathcal{L}_{\text{distil}}$) (Cover and Thomas, 2005; Hinton et al., 2015). Instead of relying on Python's `pickle`, we used a custom tokenizer that stored the 16-bit integer k-mer genome sub-sequences in a `hdf5` format (i.e., as a dataset containing an $(N \times s_{\max})$ matrix of N genomic sequences of maximally $s_{\max}$ tokens). Therefore, data loading and random sequence masking occur asynchronously to GPU-side compute.

Training was performed on a single GPU worker using the Adam optimiser (Kingma and Ba, 2015) with $l_r = 2.5 \cdot 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, with a batch size of 64 sequences across 115,435 training steps, after which $\mathcal{L}$ reached a plateau. Student model quality was compared by fine-tuning the *prom-core* 2-class classification task – to predict proximal and core promoter regions, see Ji et al. (2022).

4.3.2. Encoder with q-sequence freezing

Prior to encoding any (ordered) sequence set $\{x_{i,j}\}_{j=1}^{n_i}$, let us randomly reassign the indices j to shuffle them, and let us define $q_g \in \mathbb{N}$ such that $0 \leq q_g \leq n_{\max}$. The q-sequence freezing method (Algorithm 1) implies that the first $q_g \leq n_i$ sequences are passed to the encoder, to yield e_i^{graph}. On the other hand, the remaining $n_i - q_g$ sequences are encoded but not attached to the computational graph, to yield $e_i^{\text{no-graph}}$. Additionally, computation can be performed in chunks of maximally q_f sequences. For the TCGA dataset, $q_f + q_g$ is maximally 500. Finally, all sequence encodings are collected into a single set e'_i by concatenating e_i^{graph} and $e_i^{\text{no-graph}}$, such that only q_g sequences contribute to the gradient calculation at the encoder. Because PMA is permutation-invariant and all sequences are considered (also in its computational graph), pooling and downstream operations remain unchanged.

4.4. SetOmic : Set representations of multi-omics

SetQuence is built to extract features from sequence data. We built SetOmic, (Fig. 1), to incorporate further omic data types. The *encoder-pooler-decoder* architecture of SetQuence is generalised for a multi-omics data-set, $D_{\text{multi}} = \{(x_i, y_i)\}_{i=1}^{n}$, if $x_i = \{x_i^{\omega}\}_{\omega \in \Omega}$, where $x_i^{\omega} \in \{\mathbb{N}^{n_i}, \mathbb{R}^{m_i}\}$, with $\|\Omega\| > 1$. The processing of sequences takes place through SetQuence; meanwhile, *locus-exp* token pairs, are embedded via a linear layer $f_e : \mathbb{N}^G \to \mathbb{R}^{G \times d}$, where G is the maximum number of quantified *loci*. These embeddings encode, qualitatively, the functional aspects of the gene $g \in G$ at which expression has been measured, and quantitatively, the expression level. Formally, a per-patient *locus-exp* embedding $e_i^{\omega_{\text{exp}}}$ is the sum of *loci* and *expression* embeddings, as:

$$e_i^{\omega_{\text{exp}}} = \{f_{e_{\text{loci}}}(g) + f_{e_{\text{exp}}}(f_d(x_{i,g}^{\omega}))\}_{g \in G} \tag{4}$$

Analogous to $e_i^{\omega_{\text{mut}}}$, *locus-expression* embeddings constitute a set representation suitable for the previous *pooling* architecture. For instance,

in the case of mutomes and transcriptome counts, the *encoder* part of SetOmic is defined as $(e_i^{\omega_{\text{mut}}}, e_i^{\omega_{\text{exp}}}) = \text{SetOmic}_{\text{ENC}}(x_i)$ such that:

$$e_i^{\omega_{\text{mut}}} = \text{PMA}(S^{\omega_{\text{mut}}}, \text{SAB}(\text{DNABERT}(x_i^{\omega_{\text{mut}}})))$$
$$e_i^{\omega_{\text{exp}}} = \text{PMA}(S^{\omega_{\text{exp}}}, \text{ENC}(\{f_{e_{\text{loci}}}(g) + f_{e_{\text{exp}}}(f_d(x_{i,g}^{\omega}))\}_{g\in G}))$$

where S are the k-seeds for PMA, and ENC is an *encoder* such as SAB or ISAB. Then, embeddings e_i^{ω} are summed (sum-*pooling*) to yield a multi-omic representation:

$$e_i^{\text{multi-omic}} = \sum_{\omega\in\Omega} e_i^{\omega} \tag{5}$$

4.5. Classification of tumour types

The supervised classification module of tumour types consists of a Fully Connected (FC) block with dropout ($p = 0.3$), consisting of a Linear layer with a Rectified Linear Unit (ReLU) activation function:

$$\hat{y}_i = \text{ReLU}(\text{Dropout}_p(e_i^{\text{multi-omic}})) * W^T + B \tag{6}$$

where $\hat{y}_i$ is a vector of predicted class logits for patient i, $e_i^{\text{multi-omic}}$ is an encoding with dimension d, W and B are weight and bias matrices, and the output dimension c_s is the number of classes to be classified (i.e., tumour types and healthy controls). The networks were trained by back-propagation of the cross-entropy loss function:

$$\mathcal{L}(\hat{y}, y) = \{l_1, \dots, l_N\}^T, l_n = -w_{y_n} \log \frac{\exp(\hat{y}_{n,y_n})}{\sum_{c=1}^{C} \exp(\hat{y}_{n,c})} \tag{7}$$

where each $\hat{y}_{n,c}$ is the *logit* of class c out of C classes, for the instance n in a batch. Equivalently, y is the one-hot vector of the target class (label), and w_{y_n} is the weighting of the loss for class y_n. During model training, model parameters were updated with the Adam optimiser (Kingma and Ba, 2015), parameterised with $\beta_1 = 0.9, \beta_2 = 0.999$, learning rates in the interval $l_r = [0.001, 10^{-6}]$, and a linear learning schedule, with 10% warm-up steps over the total number of training steps, with a peak learning rate $l_r^{\max} = l_r$. We implemented layer normalisation at the *encoder-pooler* to provide regularisation effects (Ba et al., 2016).

Data in the form $D = \{(x_i, y_i)\}_{i=1}^n$ was randomly split into training and testing sets $D = D_{\text{train}} \bigcup D_{\text{test}}$, where $\|D_{\text{train}}\| = \lceil 0.8 * \|D\| \rceil$, $\|D_{\text{test}}\| = \|D\| - \|D_{\text{train}}\|$, in a stratified manner. We accounted for class imbalance during training by weighting:

$$w_{y_n} = \frac{\|\{y_i = y_n | \forall (x_i, y_i) \in D\}\|}{\|D\|} \tag{8}$$

4.6. Classification metrics

The final model quality was assessed on the D_{test} subset using five metrics: precision – ratio of classifications that are true positives across all positive predictions; recall – ratio of classifications that are positives across all ground-truth positives; accuracy – ratio of correct classifications; and Area Under the Receiver Operating Characteristic (ROC_{AUC}) or Precision–Recall (PR_{AUC}) curves, – higher values are better (interval $[0, 1]$). In multi-class cases, metrics were macro-averaged.

Table 1
Comparison of pooling methods in Baseline SetQuence, with and without disabling of the fine tuning during training.

Pooling strategy	Accuracy
Max	0.199
Min	0.045
Mean	0.271
PMA ($s = 1$, $h = 12$)	0.404
SAB + PMA	0.415
ISAB + PMA ($I = 50$)	0.413
PMA_{dis}	0.297
PMA_{dis} > epoch 10	**0.475**

4.7. Explainability through Primary Attribution Methods

We discuss the Primary Attribution Methods together with how they were used to measure the contribution of each feature in Appendix B.

5. Results: Architecture

In this section, we assess the architectural choices for SetQuence. Furthermore, we assess the effectiveness of the applied optimisations. Furthermore, Then we discuss the direction taken with SetOmic. For the purpose of clarity, we refer to the original implementation of SetQuence, Sections Section 4.2, as Baseline SetQuence, and when referring to Optimised SetQuence, we refer to the model that includes all of the optimisations applied in Section 4.3.

5.1. Baseline SetQuence *pooling strategies*

We assess the classification performance of three different *pooling* strategies: (**i**) *max*, *min* and *mean*-pooling; (**ii**) pooling by attention (PMA); and (**iii**) self attention, via ISAB or SAB operations, followed by PMA, as indicated in Table 1. Classification performance is expressed as the macro-averaged accuracy after 40 epochs, for the validation subset. Specific combinations of architectural hyperparameters are indicated (i.e., number of attention heads (h), number of seeds for PMA (s) and number of inducing elements for ISAB (I)). The accuracy was lowest for *mean*, *max*, *min* and PMA pooling strategies. SAB and ISAB showed marginally better classification accuracy than for PMA. Learning was expected to fail under *min*-pooling, as the latent space mostly consists of zeros after the ReLU activation layer at the classifier.

We explored disabling the fine-tuning (PMA_{dis}) during training of SetQuence to investigate the load balance problem we faced due to the high memory requirements of DNABERT. Disabling fine-tuning after 10 training epochs (PMA_{dis} > epoch 10) yielded the best overall classification performance, even compared to enabled fine-tuning throughout all training epochs. Showing the best balance between efficiency and accuracy, we chose PMA pooling and epoch-selective fine-tuning as the preferred strategy for training SetQuence.

Algorithm 1 Encoding step with q-sequence freezing

Input: q_g, q_f, S_i ▷ Sequences in graph; chunk size; sequence set
Output: $e_i^{\dagger}$ ▷ Sequence encoding set
  $S_i^{\text{shuffled}} \leftarrow \text{shuffle}(S_i)$
  $S_i^{\text{finetune}} \leftarrow \{x_{i,j} \in S_i \mid \forall j, j \le q_g\}$
  $e_i^{\dagger} \leftarrow \text{DNABERT}(S_i^{\text{finetune}})$
  for $q_i = q_g$; $q_i \le n_s$; $q_i \leftarrow q_i + q_f$ **do, with** `freeze()`: ▷ Get chunk of sequences, freeze graph
    $S_i^{\text{non-finetune}} \leftarrow \{x_{i,j} \in S_i \mid \forall j, q_i \le j < \min(q_i, n_i)\}$
    $e_i^{\text{non-finetune}} \leftarrow \text{DNABERT}(S_i^{\text{non-finetune}})$
    $e_i^{\dagger} \leftarrow e_i^{\dagger} \cup e_i^{\text{non-finetune}}$
  end for

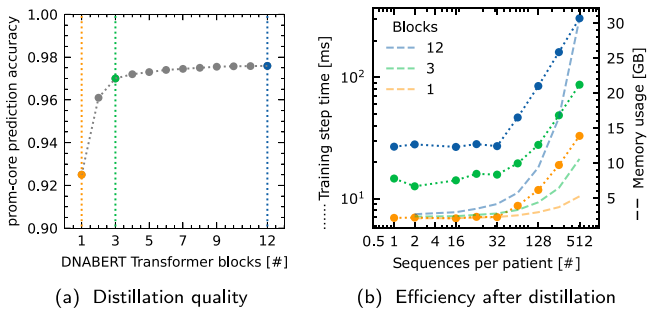


Fig. 3. Efficiency and benchmark quality after DNABERT distillation, with 1 up to 11 transformer blocks, compared to the model with 12 blocks. In (a), accuracy after 5 fine-tuning epochs on the *prom-core* task, as a benchmark used to compare the different block configurations. In (b), the time needed for forward–backward operations on DNABERT model with 1, 3 and 12 transformer blocks, as a function of the number of sequences to encode in a batch.

5.2. Analysis of the optimisation strategies

We addressed two efficiency aspects: the memory/time limitations of the DNABERT encoder and kernel execution sparsity during pooling. More than 100 million parameters of Baseline SetQuence mostly correspond (85%) to the DNABERT encoder. Therefore,

the following optimisations were implemented upon the Baseline SetQuence: (**1**) asynchronous data-loading based on the `hdf5` standard, (**2**) diffusive load balancing, (**3**) lower number of transformer blocks at the encoder (after knowledge distillation), (**4**) JIT execution at the pooler, and (**5**) q-sequence freezing, with $q_g = 256$.

5.2.1. Knowledge distillation

Knowledge distillation was performed on the original DNABERT model (*teacher*) to yield models (*students*) with 1–11 transformer blocks (see Section 4.3.1). To achieve this goal, all distillations with 3 or more transformer Encoder blocks yield metrics as good as the largest model with 12 blocks (e.g., accuracy shown in Fig. 3(a)).

Distillation did not impact average metrics when training end-to-end SetQuence on TCGA for tumour type classification: precision (0.395), recall (0.361) and accuracy (0.478) remained similar to the Baseline SetQuence model built on with a 12 block model (0.375, 0.362, and 0.475, respectively). As expected, forward and backward times and memory usage were reduced when using fewer transformer blocks.

5.2.2. q-sequence freezing: Reducing computational graph size

Reducing the number of transformer blocks improved the efficiency of the encoder; however, there is an upper bound on the number of sequences that can be encoded during a training step (see Fig. 3(b)), an important limitation if the number of available data sample sequences is large (e.g., $n_i > 512$ on an encoder with 12 layers).

Training the SetQuence architecture with default hyperparameters on the TCGA dataset for tumour type classification, different numbers of q_f sequences (from 0 to 500, where $q = \frac{q_g}{q_g+q_f}$ ranges from 0 to 1) led to changes in test classification metrics (i.e., accuracy shown in Fig. 4).

Specifically, the configuration $q_f = 0$ corresponds to not updating encoder parameters (complete freezing), whereas $q_f = 500$ means that at most 500 sequences are used in the computational graph; for TCGA, it corresponds to all sequences available per patient. Among these, using a $q_f = 256$ shows the highest test accuracy.

5.2.3. Memory-efficient attention and JIT execution improve pooling

Distillation and q-sequence freezing can reduce the space–time complexities of the encoder. Similarly, we aimed to improve the efficiency of the pooling by attention (PMA) implemented by Lee et al. (2019a). Indeed, PMA shows low GPU usage with low values of $k \cdot n_{\max}$, that

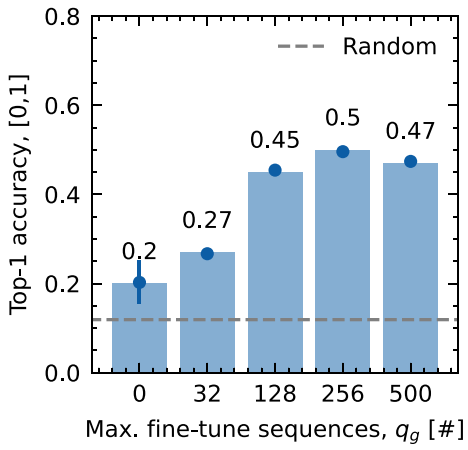


Fig. 4. Maximum model accuracy when training the SetQuence architecture with default hyperparameters on the TCGA dataset for tumour type classification.

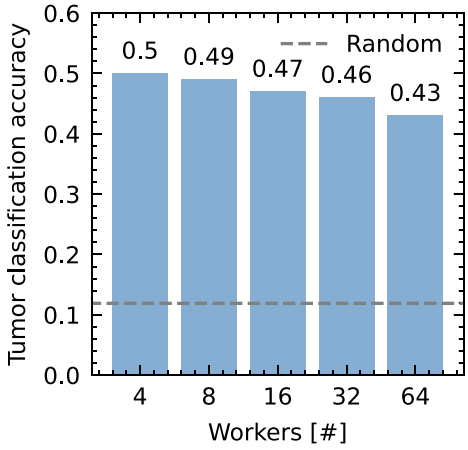


Fig. 5. The final model test accuracy across workers of SetQuence, trained on TCGA dataset for tumour type classification from variant-associated sequences after all optimisation strategies have been applied. Metrics for 32 and 64 workers are the same as training on 4 and 8 nodes with gradient synchronisation only 1 out of 8 steps, respectively.

is, low number of pooler seeds and sequences Therefore, we alternatively relied on TorchScript JIT for module serialisation instead of the *eager* execution at the PMA module in SetQuence. The original vanilla attention could not be performed on more than 20,000 sequences and 4000 seeds (thus, not shown), owing to GPU memory limitations. This restricts the applicability of SetQuence to future datasets with large numbers of sequences, regardless of the efficiency of the encoder, if the pooler has many k seeds. This would be useful (e.g., as increasing k is positive for model generalisation), see Section 6.1.2.

To circumvent this, we replaced the vanilla attention module (`nn.MultiHeadAttention`) in PyTorch, used for the Baseline SetQuence implementation, by Top-k attention, a more memory-efficient, chunked-attention-based mechanism (Gupta et al., 2021).

Chunked attention is preferred whenever $k \cdot n_{\max} > 8 \cdot 10^7$; no other configuration justifies the additional overhead of Gupta's implementation. Indeed, vanilla attention was used in Section 6, as no dataset or model configuration involved >20,000 sequences or required >4000 seeds.

5.3. Optimised SetQuence : All strategies put together

Combining all optimisation strategies (see Fig. 5), training epoch time on TCGA were reduced from ~11 min to ~3 min on 1 GPU, a ~3.7x

Table 2
Tumour type classification performance: SetQuence, vs. GIT, a comparable method, and random classifiers.

Classifier	Precision	Recall	F1-score	Accuracy	ROC_{AUC}
SetQuence	**0.375**	**0.362**	**0.359**	**0.475**	**0.910**
GIT	0.312	0.248	0.229	0.247	0.740
Strat. Random	0.035	0.034	0.034	0.041	0.500
Prior Random	0.004	0.031	0.007	0.118	0.500

speedup. These optimisations did not affect model quality (test metrics for tumour type classification, after 25 epochs) when comparing against the unoptimised model on same resources. However, increasing the number of workers affected model metrics – (e.g., test accuracy, reduced from 0.5 to 0.43 when scaling from 4 to 64 DDP workers). This was caused by design and was to be expected, since we kept the number of iterations constant for the purpose of the experiment. This is a efficiency vs. model quality trade-off, as larger batch sizes lead to lower final test metrics (e.g., accuracy) after the same number of training steps. We would expect, with longer training times, especially with these distributed settings, this can be compensated and the models could reach or surpass the baseline model quality.

5.4. SetOmic : *Integration of other omics data*

SetOmic, built on Baseline SetQuence for the feasibility study, supports sequence and non-sequence inputs via multi-omics late integration. Specifically, we focused on integrating transcriptome counts with variant-associated sequences. We trained two models with PMA pooling for tumour type prediction. The first learnt from transcriptome counts only (SetOmic-exp), the second with a combination of variant-associated sequences and transcriptome counts of the pan-cancer TCGA data-set (SetOmic).

5.5. *HPC implementation*

Efficiency analysis and model training experiments were carried out on nodes, interconnected via InfiniBand HDR200, equipped with 2x AMD EPYC CPU 7352 (24 cores, multi-threading capable), 1 TB of RAM and 8x NVIDIA A100-SXM4 GPUs (40 GB HBM2 vRAM), in a fully connected intra-node topology (8 × 8 links, 3rd generation NVLink). We used PyTorch 1.8.2 and OmniOpt (Winkler et al., 2021) for our DNN implementations.

6. Results: Applicability

This section covers the clinical applicability of our models and how the applied optimisations impact their effectiveness (Section 6.1). Furthermore, we explore how SetOmic (Section 6.2), a generalisation of Baseline SetQuence, is able to support multiple omics inputs. Then model explainability was explored in Section 6.3.

6.1. SetQuence

6.1.1. *Tumour type classification task*

We constructed a confusion matrix (Fig. 6) for 7 out of the 32 tumour types: Lung Adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Esophageal carcinoma (ESCA), Sarcoma (SARC), Breast invasive carcinoma (BRCA), Colon Adenocarcinoma (COAD) and Stomach Adenocarcinoma (STAD). Predictions for lung tumours, LUAD and LUSC (53 and 91 respectively) are mostly correctly predicted; misclassifications (39 samples) are caused by similarities between the two types, potentially due to similar sub-types, or same primary site. The same holds for the digestive tract tumours, COAD and STAD.

We explored how accurate SetQuence is in deciding between all of the tumour classes (i.e., non-binary classification) and compared it to

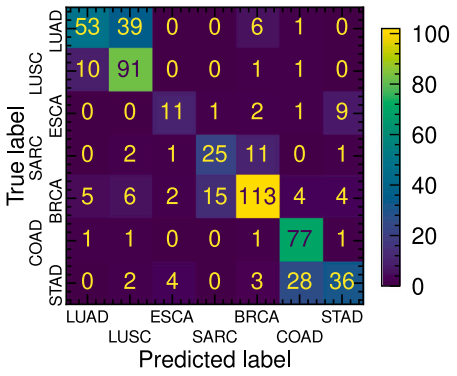


Fig. 6. SetQuence confusion matrix for the seven selected tumour types.

GIT (Tao et al., 2019), which is described in Section 2. It performed better than GIT in all four macro-averaged metrics: precision, recall, F1, accuracy, and also one-vs.-rest ROC_{AUC} Curve, showing improved generalisation with SetQuence (Table 2). The *Stratified Random* dummy classifier returns class labels randomly sampled from a distribution parameterised by the empirical label distribution. *Prior Random* returns the most frequent label in the empirical distribution.

6.1.2. *TCGA vs. COSMIC: Somatic, exome-only variants*

The TCGA variant dataset mainly provides somatic variants at exons, where COSMIC *Catalogue Of Somatic Mutations In Cancer* dataset is a larger Whole-Genome dataset. We show that Optimised SetQuence is trainable on such larger dataset, and how model quality compares on Cosmic respect to TCGA for the same goal and data types (e.g., somatic, exome-only variants). Furthermore, we show whether the inclusion of non-coding variation further improves classification quality and provides additional biological insights. In Section 6.1.4, we show the impact of alternative DNA-LMs on SetQuence (i.e., pre-trained on other genome references) to better understand the projection of sequences into the latent feature space.

Let us compare the TCGA dataset to the larger Cosmic, specifically to a subset of coding variants across 31,677 patients across 32 tumour types comparable to TCGA labels, in order to better explore the benefits of the previous optimisations when training SetQuence for yielding a single multi-class tumour type classifier. By choice, classification quality is subsequently shown via macro-averaged metrics: precision, recall, accuracy, (ROC_{AUC}), and (PR_{AUC}) Curve, see Section 4.6.

Optimisations in Section 5.2 need to be enabled when training on Cosmic. Otherwise, the model was untrainable due to a lack of GPU memory. Therefore, we trained Optimised SetQuence on 8 GPUs, which took ∼220 min when trained on Cosmic until convergence (40 epochs, ∼160 GB of sequences accumulatively encoded), compared to the around 18 min taken on the TCGA dataset (25 epochs, ∼3 GB of accumulate sequence data); the higher average amount of sequences per patient in Cosmic (1000 s vs. 100 s) leads to better saturation of GPU resources, thus, better training efficiency. Regarding model quality, tumour type classification test metrics on Cosmic were similar when using an architecture analogous to Baseline SetQuence approach on TCGA, that is, with one pooling seed, $k = 1$ (Table 3). This configuration led to similar metrics respect to *hidden-genome* approaches (e.g., NN and RF built on non-synonymous mutation frequencies as input Chakraborty et al., 2021b).

Then, we studied the effect of increasing the number of pooling seeds (k). This improved all classification metrics on Cosmic, significantly surpassing the baselines. These improvements correlate with the converged parameter space being found at flatter and wider local minima of the loss landscape (Fig. 7). This is an argument in favour of model generalisation (Hinton and van Camp, 1993; Li et al., 2018), supporting that large models (proportionally to k) lead to better

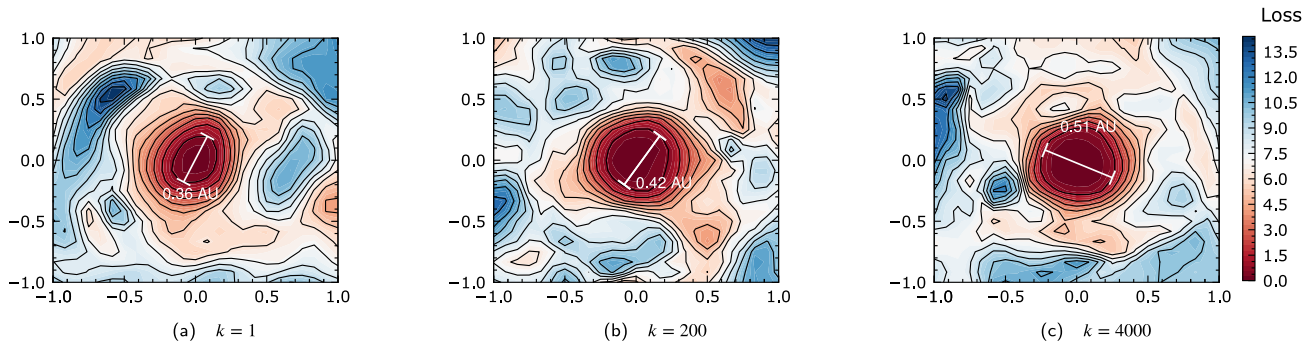


Fig. 7. Loss landscapes for Optimised SetQuence trained on Cosmic, for the training set, at three different architectural configurations (respect to number of pooling seeds, k). The widths of local minima basins are indicated in Arbitrary Units. X,Y axes correspond to parameters α and β, scaling parameters to gradually add i.i.d. Gaussian noise δ, η to each parameter at the analysed model, located at the $(0, 0)$ coordinate.

Table 3
Classification quality of SetQuence (SetQuence) and baseline models trained on all TCGA or Cosmic coding variants (CV), across 32 tumour types. The NN and RF baselines use non-synonymous mutation frequencies as input data, and are parameterised as in Chakraborty et al. (2021b). The *Stratified* baseline is a dummy classifier that returns class labels sampled from the distribution of training labels. Classification metrics are as macro averages across tumour types for the test cohort.

Dataset	Model	Precision	Recall	Accuracy		ROC_{AUC}	PR_{AUC}
				Top-1	Top-2		
TCGA	SetQuence ($k = 1$)	0.375	0.362	0.475	0.627	0.910	0.376
	Stratified random	0.004	0.031	0.118	0.354	0.500	0.031
Cosmic (CV)	SetQuence ($k = 1$)	0.344	0.312	0.483	0.604	0.878	0.395
	SetQuence ($k = 200$)	0.381	0.350	0.521	0.636	0.881	0.416
	SetQuence ($k = 4000$)	**0.502**	**0.402**	**0.567**	**0.657**	**0.891**	**0.468**
	NN	0.320	0.274	0.519	0.623	0.838	0.269
	RF	0.408	0.263	0.472	0.581	0.852	0.313
	Stratified random	0.030	0.030	0.060	0.066	0.499	0.000

Table 4
Classification quality (macro-averaged metrics) of SetQuence and Random Forest (RF), see Chakraborty et al. (2021b). $-_{CV}$: coding variants; $-_{NCV}$: non-coding variants; $-_{GeneID}$: non-synonymous mutation frequencies; $-_{RMD}$: Regional Mutational Density, per 100 kb. Ensemble models represented as ∩. In $-_{CV \to NCV}$, CV were used to predict NCV, and viceversa.

Model	Precision	Recall	Accuracy		ROC_{AUC}	PR_{AUC}
			Top-1	Top-2		
$SetQuence_{CV+NCV}$	**0.605**	0.514	**0.709**	0.812	0.906	0.603
$SetQuence_{CV}$	0.501	0.493	0.682	0.762	0.891	0.529
$SetQuence_{NCV}$	0.521	0.456	0.649	0.787	0.883	0.531
RF_{RMD}	0.442	**0.522**	0.694	0.800	0.908	**0.624**
RF_{GeneID}	0.451	0.494	0.701	**0.816**	**0.910**	0.569
$SetQuence_{CV \to NCV}$	0.480	0.501	0.609	0.749	0.843	0.478
$SetQuence_{NCV \to CV}$	0.516	0.435	0.632	0.781	0.877	0.509
$SetQuence_{CV+NCV} \cap RF_{RMD}$	0.552	0.631	**0.787**	**0.868**	**0.926**	0.679
$SetQuence_{CV} \cap RF_{RMD}$	**0.556**	0.631	0.768	0.867	0.924	**0.699**
$RF_{GeneID} \cap RF_{RMD}$	0.461	**0.681**	0.719	0.803	0.913	0.613
Stratified random	0.106	0.106	0.134	0.411	0.496	0.061

generalisation, not just memorisation of training data. For this reason, the number of pooler seeds was set to $k = 4000$ for the following experiments on the Cosmic dataset.

6.1.3. *Including non-coding variation improves tumour classification*

A subset of 4314 Cosmic samples was tagged as *Whole-Genome Sequencing*, providing both coding (CV) and non-coding variant (NCV) data – notice that the intersection of these and the TCGA public cohort is the empty set. More specifically, Optimised SetQuence was trained on 2913 samples belonging to 10 tumour types comparable to analysed by Chakraborty et al. (2021b), with at least 10 variants per sample — ranging from 4 cases of ovarian cancer to 865 of breast cancer. In Table 4, Optimised SetQuence models separately trained on CV and NCV show metrics comparable to baselines respectively trained on non-synonymous mutation frequencies and RMD signatures. Classification metrics were higher when simultaneously training on CV and NCV compared to using these separately, although Optimised SetQuence did not outperform a baseline model ensemble of two RFs, for non-synonymous variation frequencies and RMD signatures. Also, classification on unseen data with NCV only is possible with a model trained on CV only, and vice-versa. Lastly, a model ensemble of Optimised SetQuence and a RF for RMD showed the highest test metrics. This ensemble model yielded good separation of tumour types at two-dimensional embedding space (Fig. 8(b)). Test metrics were different across individual tumour

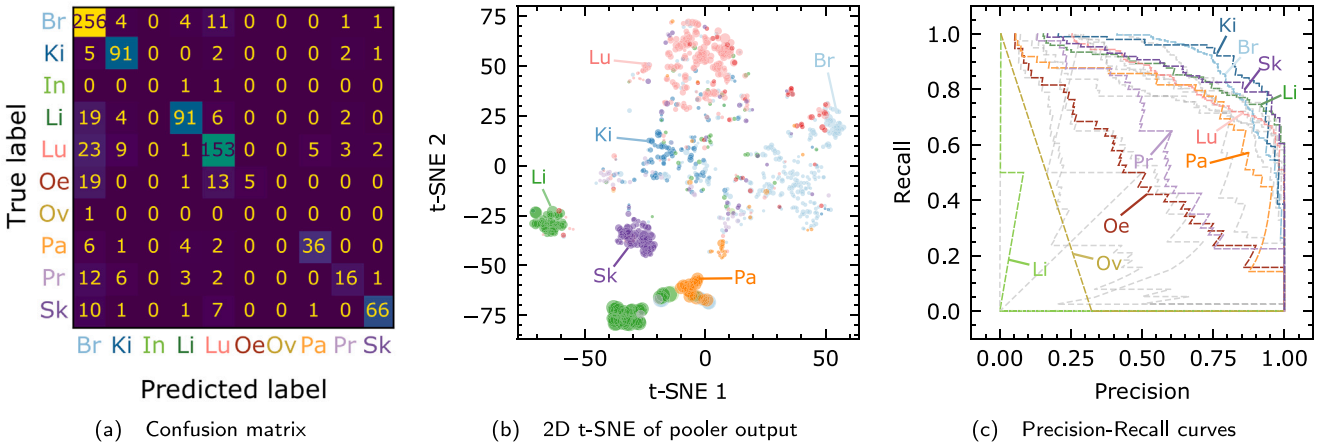


Fig. 8. Tumour type classification from coding and non-coding variants. For a OPTIMISED SETQUENCE$_{CV+NCV}$ ∩ RF$_{RMD}$ model: (a) Confusion Matrix; (b) 2D t-SNE of pooler outputs, dot size proportional to n_i per patient; (c) Precision–Recall curves for OPTIMISED SETQUENCE (coloured) and RF baseline (grey). Types: **Br**east, **Ki**dney, Large **In**testine, **Li**ver, **Lu**ng, **Oe**sophagus, **Ov**arian, **Pa**ncreas, **Pr**ostate, **Sk**in.

types — highest for Breast, Kidney and Skin cancers, and lowest for Large Intestine and Ovarian cancer (Figs. 8(a) and 8(c)).

6.1.4. Robustness to changes in the language model encoder

The majority of the human genome is non-coding (Piovesan et al., 2019); for this reason, differences between genome reference assemblies and individual's genomes mostly happen at those regions (Nurk et al., 2022). To further test the robustness of encoding to pre-processing, we trained OPTIMISED SETQUENCE with an alternative sequence encoder that we pre-trained on the same DNABERT architecture with 3 transformer blocks, using the T2T assembly (Nurk et al., 2022) instead of the GRCh38 assembly as training data — for the same number of steps, hyperparameters and objective as specified by Ji et al. (2021). With this T2T pre-trained encoder, tumour classification metrics decreased upon non-coding and coding variants, compared to using the original DNABERT encoder pre-trained on GRCh38 (Table 5). The classifier with coding variants was the only showing an increase on test precision, although it had the greatest decrease on accuracy. On the other hand, the classifier with both coding and non-coding variants had the best robustness across the three, with same test precision, and some of the lowest decreases in test recall and accuracy. Moreover, we tested whether the sequence encodings from GChR38 and T2T LMs are informed by sequences themselves, or by other aspects such as GC-content. To explore so, we shuffled the order of the nucleotides of each variant-associated sequence, before k-merization. This decreased classification metrics in both cases (marked as 'a' in Table 5), leading to models with classification properties similar to counterparts based on one-hot encoding of variants (i.e., the baseline model in Chakraborty et al., 2021b). These results confirm the essential role of LMs in feature extraction, as a ϕ-function, from variant-associated sequences.

6.2. SETOMIC

For the purpose of proof of concept, we used the BASELINE SETQUENCE model together with expression count data from the TCGA dataset to compare how incorporation of further omics data features impact the classification metrics. We refer to this model as SETOMIC. We constructed a confusion matrix, Fig. 9, for SETOMIC, to compare it to Fig. 6. The same 7 tumour types are showing a higher proportion of correct classifications as well as a reduction of confusion between sub-types.

Indeed, the ROC curve graph, where each ROC curve corresponds to each tumour type in Fig. 10 shows how the use of both expression and mutation data improved the ability to detect all tumour types (red for BASELINE SETQUENCE, blue for SETOMIC, where bright red and bright blue lines are average ROC curves across all tumour types for

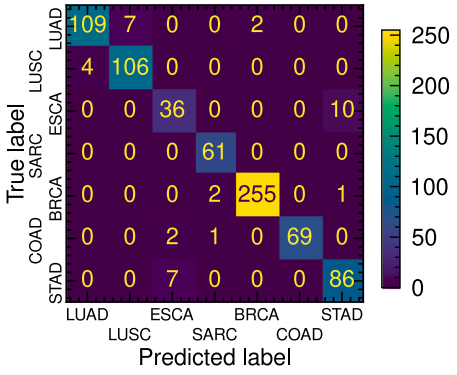


Fig. 9. SETOMIC confusion matrix for the seven selected tumour types.

BASELINE SETQUENCE and SETOMIC respectively). Uterine Carcinosarcoma performed worst using BASELINE SETQUENCE, second worst when using SETOMIC, following closely behind Kidney Chromophobe. While the sample sizes of both were low, clear correlation between ROC and sample size was not observed.

Table 6 provides a comparison of classification performance between our models and OMIEMBED, a VAE architecture for multi-omics integration (Zhang et al., 2021) trained on transcriptomes. Performance metrics are macro-averages across all classes, for the test subset. When trained on transcriptome counts, SETOMIC-EXP achieves close to OMIEMBED's state-of-the-art performance. In the multi-omics scenario (SETOMIC), classification performance significantly improves over our single-omics model and OMIEMBED.

We assessed the robustness of our models and previous baselines after randomly excising parts of the available data points per patient, for transcriptome counts and variant-associated sequences. Classification performance on SETOMIC, compared to OMIEMBED, was affected to a lesser extent when removing half of the input omics (17% vs. 22% decrease in accuracy), although with similar robustness in the transcriptome-only case (24% decrease in accuracy). In the case of variant-associated sequence data, the sequence encoding approach introduced with BASELINE SETQUENCE showed a lower penalisation when randomly removing half of the available mutation events, with a 9.7% drop in accuracy, compared to using Gene IDs with the GIT method, with a 22.3% drop.

6.3. XAI recapitulates the biological meaningfulness of our models

Explainability was introduced as essential for ML/DL trustworthiness in oncology. Primary attribution methods served us to obtain

Table 5
Using alternative LMs at the SetQuence encoder. SetQuence with $-_{CV}$ and $-_{NCV}$ with our T2T encoder is compared to counterparts in Table 4 (with GRCh38 encoder). Macro-averaged test metrics are precision, recall, accuracy, ROC_{AUC} and PR_{AUC}, for tumour type classification on Cosmic, 10 types.

Model	Precision	Recall	Accuracy	ROC_{AUC}	PR_{AUC}
$SetQuence_{CV+NCV}$	0.607, ↑0.3%	0.482, ↓6.2%	0.585, ↓17.5%	0.881, ↓2.6%	0.555, ↓8.0%
$SetQuence_{NCV}$	0.477, ↓8.4%	0.371, ↓18.6%	0.561, ↓13.6%	0.826, ↓6.5%	0.465, ↓12.4%
$SetQuence_{CV}$	0.563, ↑56.3%	0.451, ↓8.51%	0.545, ↓20.1%	0.797, ↓10.5%	0.506, ↓4.3%
$SetQuence^{a}{}_{CV}$ T2T	0.408, ↓18.6%	0.363, ↓26.4%	0.483, ↓29.2%	0.648, ↓27.3%	0.345, ↓34.8%
$SetQuence^{a}{}_{CV}$ Ch38	0.330, ↓34.1%	0.267, ↓45.8%	0.436, ↓36.1%	0.671, ↓24.7%	0.263, ↓50.3%

[a] The effect of *shuffling* the position of nucleotides at variant-associated sequences is shown for both encoders.

Table 6
Tumour type classification performance of SetOmic-exp, expression count data only model vs. SetOmic, sequence data and expression count data model, to investigate if expression count data model sufficiently produces well performing metrics. We also consider a comparable model OmiEmbed and stratified and prior random classifiers.

Classifier	Precision	Recall	F1-score	Accuracy	ROC_{AUC}
SetOmic-exp	0.876	0.887	0.880	0.923	0.976
SetOmic	**0.945**	0.909	**0.921**	**0.950**	**0.997**
OmiEmbed	0.932	**0.911**	0.914	0.942	**0.997**
Strat. Random	0.035	0.034	0.034	0.041	0.500
Prior Random	0.004	0.031	0.007	0.118	0.500

Table 7
Performance comparison of four attribution methods for tumour type prediction (single run on the test subset).

Attribution method	$ROC_{AUC}(A_+)$	$ROC_{AUC}(A_-)$
Input X Grad	**0.704**	**0.247**
DeepLIFT	0.698	0.295
Integrated Gradients	0.619	0.376
Gradient SHAP	0.581	0.371
Random attribution	0.500	0.500

attribution scores $A(x^{\omega}_{i,j})$, per patient i and feature j, for specific omics ω.

We learn that only a few features are important to determine tumour type by calculating the attribution scores of each omics using the Input X Grad attribution method, from Optimised SetQuence and from SetOmic. Attribution scores followed the power-law distribution.

6.3.1. SetOmic : Expression and mutation attributions

We obtained a sorted collection of variant-associated sequence and transcriptome count features (for Breast Cancer, BRCA) to evaluate whether attribution scores are intuitively related to biological functions. In the case of mutation data, all top five genes with highest attribution scores are known to have a significant influence of cancer occurrence: PIK3CA is mutated in 35.7% of BRCA patients (Martínez-Sáez et al., 2020); MYADM regulates the Rac1 targeting and is required for cell migration (Aranda et al., 2011); CD200 inhibits metastatic growth of tumour cells (Erin et al., 2014); FABP3 is a tumour suppressor gene in BRCA (Tang et al., 2016); and XKR6 determines the responsiveness to drugs in BRCA (Coyle et al., 2018). Respect to expression data, all top five genes have been proposed to have implications on BRCA samples when differentially expressed respect to healthy tissues: S100A11 is a S100 calcium binding protein A11 and plays a role in cancer cell growth, associated with poor survival (Zhang et al., 2017); the differential expression of SLC39A6 is associated with different prognosis (Cui et al., 2015); the low expression of PRKAR1 A is indicative of poor survival in basal-like and HER2 tumours (Beristain et al., 2014); RAD21 correlates with resistance to chemotherapy in various BRCA sub-types (Xu et al., 2011); and COL5A2 is associated with migration and invasiveness through extracellular matrix (Vargas et al., 2012). Similarly, the transcriptome features with lowest attribution scores correspond to pseudo-genes and non-coding RNAs.

More systematically, all attribution methods, except Gradient SHAP, showed $ROC_{AUC}(A_+)$ significantly higher than for a random attribution method (Table 7). This supports how SetOmic is able to map inputs to outputs by recapitulating high-level and meaningful biological information, not exclusively by memorisation. Input X Grad yielded the best attribution performance, with highest $ROC_{AUC}(A_+)$ and lowest $ROC_{AUC}(A_-)$, followed by DeepLIFT and Integrated Gradients. Moreover, Input X Grad and DeepLIFT scores for SetOmic are consistent with previous XAI implementations for pan-cancer TCGA such as XOmiVAE (based on a DeepLIFT approximation of SHAP values), with a reported $ROC_{AUC}(A_+) = 0.690$ for the TCGA-BRCA cohort (Withnell et al., 2021).

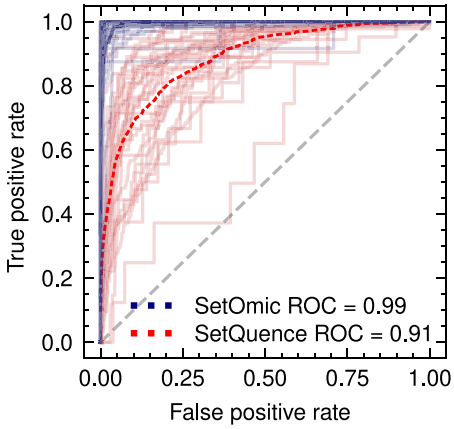


Fig. 10. ROC curve plot for each of the 32 TCGA tumour types, where the *blue* curves were produced by SetOmic and the *red* by Baseline SetQuence on the test subset.

Additionally, we analysed the most relevant biological annotations via Gene Ontology (GO) enrichment with *g:Profiler* (Reimand et al., 2007), for the Top-100 expression *loci* sorted by Input X Grad attribution (in the TCGA-BRCA data-set). The most significant GO biological processes were *extracellular matrix (ECM) organisation* ($p_{adj} = 1.346 \cdot 10^{-8}$), *collagen fibril organisation* ($p_{adj} = 2.480 \cdot 10^{-8}$) and *supra-molecular fibre organisation* ($p_{adj} = 1.246 \cdot 10^{-6}$). These processes have been widely studied as potential targets for Breast Cancer treatment (Vargas et al., 2012; Wang et al., 2018; Henke et al., 2020).

6.3.2. Nucleotide-level attribution discovers cancer-relevant motifs

We extracted the top five statistically significant motifs ($p_{adj} < 0.005$) using Input X Grad from sequences with highest attribution scores, for TCGA-BRCA mutomes (see Table 8). They are: the GAGA site promoting tumour proliferation, metastasis (Saux et al., 1999; Ferreira et al., 2021), HMG box motif, lymphocyte transcriptional activator (van de Wetering et al., 1993), E2F1 regulatory element, BRCA-1 promoter activator region (Pignatelli et al., 2003), and Metal responsive elements (MREs), found in the core promoter of various human genes; associated with malignant progression (Murphy et al., 1999). These motifs do not necessarily correspond to the mutation events, but to sequences that might be important for the function of a genomic region, as for example, the motif TCTGAG.

Table 8
Top 5 sequence motifs detected from Input X Grad attribution scores using SetOmic on the TCGA-BRCA test cohort.

Motif	Function
AAGAGAG	GAGA site at the regulatory region of genes such as LOXL2, promoting tumour proliferation and metastasis (Saux et al., 1999; Ferreira et al., 2021)
TCAAAG	HMG box motif, transcriptional activator in lymphocytes. van de Wetering et al. (1993)
TCTGAG	E2F1 regulatory element, BRCA-1 promoter activator region (Pignatelli et al., 2003)
TTATCTG	GATA-regulatory element; over-expression of hPTTG1 to promote cell invasion and metastasis in Breast Cancer (Pei, 2001; Bagu and Santos, 2011; Liao et al., 2011)
TGCACGT	MREs, found in the core promoter of various human genes; associated with malignant progression (Murphy et al., 1999)

Table 9
Top 5 non-coding sequences with highest Input X Grad attribution values from the COSMIC breast cancer subset using Optimised SetQuence. All variants occur only once at the test set and were unseen during training. Features filtered by attribution scores are assessed via *RegulomeDB* (Boyle et al., 2012), to detect motifs and regulatory annotations; Rank **2b**: TF binding + any motif + DNase Footprint + DNase peak; Rank **3a**: TF binding + any motif + DNase peak.

Motif	RegulomeDB target	Rank
CTTACCTGT	ZEB1, associated with poor survival, resistance and metastatic risk (Wu et al., 2020)	2b
GTTGGGAGG	IKZF1, chromatin remodelling and the regulation of lymphocyte differentiation	2b
TTTGGGAAT	TBK1, pharmacological target; activation of T-cell immunity (Runde et al., 2022)	3a
TATTTATAG	MEF2A, mediates metastasis via TGF-β upregulation of MMP10 (Xiao et al., 2021)	2b
ACAGATTGT	NR3C1, tumour suppression in estrogen receptor-positive (ER+) (Snider et al., 2019)	2b

6.3.3. The attribution of non-coding features

The Optimised SetQuence was analysed on non-coding variant-associated sequences with high Input X Grad attribution. Table 9 shows the top 5 sequences with highest attribution scores. Upon *RegulomeDB* analysis (Boyle et al., 2012), the SNPs underlying these high score sequences map to regions and motifs with previously known roles in breast cancer physiology. We further tested whether these features, often unique in the dataset, were indeed important for classification, via per-patient feature removal: omitting the 2 or 4 features with highest positive attribution (in favour of breast cancer classification), Top-1 accuracy of $\text{SetQuence}_{\text{NCV}}$ dropped from 0.649 to 0.497 and 0.422; removing the 2 and 4 features with most negative attribution (against breast cancer classification) improved accuracy to 0.761 and 0.814.

7. Discussion

The exponential growth of cancer patient omic data has led to an increase in the need for models that are robust, reproducible, immutable, and interpretable, in particular for personalised treatments. Keeping these in mind, we constructed a feasibility study; we took advantage of NLP-inspired DL techniques, investigated their performance, improved their efficiency and explored their explainability and generalisation in an omics context. For the purpose of demonstrating the capabilities of our models, we chose tumour type classification as our use case.

More precisely, we introduced Baseline SetQuence, an *encoder-pooler-decoder* DNN, for arbitrarily many variant-associated sequences. We explored various pooling strategies where we discovered permutation and cardinality (order and measure of a set respectively) invariant PMA and a selective fine-tuning strategy produced highest testing accuracy. Then, by means of a DNABERT *encoder* and a FC+ReLU *decoder*, Baseline SetQuence showed significantly better performance than GIT, a comparable method to perform oncology-relevant downstream tasks from variant-associated sequences. In contrast to prior methods for mutome representation mentioned in Section 2, our approach is more expressive and robust, and has the potential to better generalise to unseen mutations at coding and non-coding regions. From the confusion matrices, we observe that SetQuence learns biologically meaningful information from patient's variant-associated sequences, enough to predict primary sites.

We first performed the feasibility study on sequence variant data from the TCGA database. We concluded that it is limiting to analyse coding only data, especially since the non-coding genome in a significant player in tumour analysis (Ling et al., 2015). We set out to overcome the clinically-relevant gaps discussed in Section 2 by considering whole-exome and whole-genome sequence variants themselves. The use of non-fixed sets of sequences as input via a transformers-based architecture allowed us to represent long range interactions between tokens across the genome/exome as well as their contextual information. From the hardware optimisation point of view, it also allows for parallel processing (Vaswani et al., 2017b), and thus opens up the possibility of processing large amounts of data.

To do so, we first identified and addressed the architectural limitations that came with introducing this additional data, namely by applying load balancing of patient sequences, knowledge distillation of the encoder module from 12 to 3 transformer blocks, a q-sequence freezing method that reduced memory constraints at the encoder, and a JIT-mode pooler/decoder to reduce CPU bottlenecks (Section 5.2). Combining these optimisations, with respect to the original implementation, wall times improved >3x and scalability became near linear when distributing training on up to 8 GPU workers on a single node (Section 5.3). This was shown to impact final model metrics (Fig. 5) if the same number of iterations are applied during training. Therefore, while there is a trade-off between computational efficiency at scale and final model quality, increasing the number of iterations could lead to comparable or even improved accuracy. Regardless, the new optimisations led to the ability to investigate the large dataset of non-coding variant-associated sequences. Looking at the applicability of Optimised SetQuence on the COSMIC dataset, increasing the number of pooler seeds to $k = 4000$ produced the best model performance, which are located at wider local minima in the loss landscape over the parameter space indicating better model generalisation to unseen data as noted in previous theoretical and empirical research on vanilla NNs (Hochreiter and Schmidhuber, 1997; Dinh et al., 2017; Keskar et al., 2017) and transformer-like models (Yang et al., 2021; Caillon and Cerisara, 2021).

When exploring how our model behaves when trained on either coding or non-coding variants, we show in Table 4 that an ensemble of Optimised SetQuence models (either CV or CV+NCV) and an RF for RMD signatures yielded the model performance of all SetQuence and baseline models. RMD signatures contain information beyond the mere sequence space, as not just the sequence, but also its topological position, drive the functional implications of a genomic region. We hypothesise that our latent-space features from coding variant-associated sequences are comparable to the frequency of non-synonymous mutations as input, and complementary to RMDs in the case of non-coding variants.

It was possible to classify test patient data consisting of coding variants with a model trained only on non-coding variants, and vice-versa; on the other hand, attribution values were independent from the

number of occurrences of a variant in the dataset (Table 9). Moreover, we show that using our replacement LM pre-trained on T2T assembly changes model quality compared to the original DNABERT, although both have the same architecture (Table 5). This further corroborates, first, how the encoder (DNABERT) in SetQuence represents genomic sequences into a *functional* space, similar to reported in other LM for biological sequences and in the NLP field for Sentence-BERT-like models (Reimers and Gurevych, 2019; Iuchi et al., 2021); second, that the LM itself is critical for the *quality* of latent-space representations of variant-associated sequences.

To incorporate set representations of a patient's multi-omics, we introduced SetOmic. We provide a proof of concept approach where we investigate the inclusion of expression count data, shown to be extensively used in tumour type classification (Lu and Han, 2003), together with our model built on variant-associated sequences. Based on comparisons between Figs. 6 and 9, classification of tumour types improved with SetOmic compared to SetQuence, signalling that differences between tumour types is clearer after incorporation of multiple sources of omics data. This is further shown by the ROC curve analysis in 6.1. Based on macro-averaged metrics of SetOmic-exp we also notice that a model built on expression data performs well, however, the incorporation of both variant-associated sequence data and expression count data resulted in the model performance. This shows the importance of a multi-omics approach to a model's architecture and therefore to precision oncology, in line with previous studies (Nicora et al., 2020). In conclusion, SetOmic reaches comparable results to a more coarse and limited VAE approach, OmiEmbed (Zhang et al., 2021), while providing more granularity as well as insights into long range interactions between features for a specific classification at the nucleotide level.

We illustrated the fulfilment of the *explainability* design principle by the use of XAI techniques at individual genes (*loci*) and at individual nucleotides. With our approach, we recapitulated subsets of input feature sets with a significant enrichment on biological functions important for tumour development; this illustrated the potential of our tool for a better genome-wide understanding of tumours at coding and non-coding regions.

8. Future directions

8.1. Model improvement

First, we propose further exploration of load imbalance and training strategies. Model complexity has potential to be reduced using various computational optimisation strategies. To address the trade-off between computational efficiency at scale and final model quality, two directions can be explored. On the one hand, by implementing strategies focused on reducing communication volume, such as those introduced by Alistarh et al. (2017), Bernstein et al. (2018), Rajbhandari et al. (2021), Li and Hoefler (2022) and Dettmers et al. (2022), although these may further impact model quality. On the other hand, by performing additional hyperparameter optimisation to compensate for the impact on model metrics upon increasing the batch size or number of parallel steps per training.

Second, architectural improvements, for example, implementation of the latest LM models can be explored, such as LOGO (Yang et al., 2022), which relies on convolution modules for feature extraction and tokenization based on byte-pair encoding, or SNP2Vec (Cahyawijaya et al., 2022). In addition, the PMA configuration, although improved over the original $k = 1$, still consists of only one block; thus, it only models low-order interactions (pairwise) between encoded sequences. Consequently, future work on pooling strategies would further improve model generalisation.

Third, model validation and further generalisation can be explored by using alternative data-sets investigating the same questions as explored in this paper and also other clinically relevant questions such as how germline mutations effect tumour type classification. This approach evaluates the quantification and cohort bias, however, is challenging from the data integration point of view; each database has differing data representation and applies unique data pre-processing or processing pipelines.

In Section 6.1.2, we compare our method (a DNN) against RFs which, together with Support-Vector Machines and other Bayesian approaches, are known to generally perform better than DNN for these classification goals (Lee et al., 2019b; Chakraborty et al., 2021b). Thus, the fourth direction is to explore alternative downstream classifiers from the feature space encoded and pooled using SetQuence. In addition, we report that classification quality is affected by class imbalance (i.e., as in the low predictive quality for Large Intestine and Ovarian cancer). Although this study relied on loss weighting, other direct approaches, such as over- or under-sampling (Chawla et al., 2002), as well as indirect approaches, such as noise stability regularisation (Hua et al., 2021), which could help reduce overfitting to overrepresented classes, or the lack of convergence to underrepresented classes.

Finally, some interpretability perspectives still need to be addressed, such as through clustering and correlation analysis of embedding and latent spaces to biological features. The embedding and attention mechanisms have been used in other fields to more systematically explain input and latent spaces (i.e., to recapitulate graph learning in bioinformatics, or to identify clusters of similar shapes in biomedical imaging data Nelson et al., 2019; Hagenah et al., 2019). In addition, alternative XAI approaches such as the Interaction Detection, could further explain the role of features as an ensemble.

8.2. Model applicability

Taking our paper from the feasibility study level to the applicability and predictive level, multiple further clinically-relevant downstream tasks can be performed. One of particular importance is improved tumour sub-type classification. Our model can be used to indicate existing and novel coding and non-coding regions which are significantly associated to a certain subtype. This can lead to clearer distinction between tumour sub-types which are often misidentified, mistreated, or understudied, and improved personalised treatments (Richards et al., 2022; Yao et al., 2018; Heo et al., 2021). However, patient data availability per certain tumour types tends to be sparse, especially rare tumour types or sub-types. Therefore, the main condition to be satisfied for state-of-the-art classification metrics of our model is the availability of enough samples per tumour type, as the larger the feature space, the more patient samples are required.

The second applicability direction is germline variant exploration. Due to how our architecture is built, it is relatively easy to apply it to a different use case and even a different dataset such as germline variant data. As a proof of concept study, we investigated the Pancreatic Cancer (PDAC) dataset, provided by Al-Fatlawi et al. (2021), and replicated their classification study using our model. We distinguished between the Pancreatic Cancer (either as resectable Pancreatic Adenocarcinoma, or non-resectable Pancreatic Carcinoma), from general Chronic Pancreatitis, motivating another clinically relevant use-case, distinguishing between cancerous vs. non-cancerous diseases.

We were able to extract Pancreatic Cancer specific variant-associated sequences per patient with high attribution scores, such as STAT3 and B4GALT5 (Corcoran et al., 2011; Indellicato et al., 2020). This shows the capabilities of our models when applied on germline variants and the potential of it to be used in their studies. In addition, SetQuence can distinguish between Pancreatic Cancer and Chronic Pancreatitis on mutually exclusive feature spaces (i.e., when variants across training and testing data are in mutually exclusive genome regions). This makes SetQuence more advantageous to use in

the direction of Precision Oncology than the original study (Al-Fatlawi et al., 2021), which relies on roughly 70 variants as input features, such that any of the fixed input variants must be present in a patient sample to enable classification.

Third, to improve SetQuence, the prior *hidden genome* features (Chakraborty et al., 2021b) can be integrated into the LM encoder. These feature spaces could provide complementary views on genomic variation.

Fourth, variant sequence embeddings and gene expression data are good indicators for cancer patient survival analysis (Beristain et al., 2014), and has been explored by comparative models, OmiEmbed and GIT. Since our model outperforms OmiEmbed and GIT on the classification task, it also has potential to reveal more insights into patient survival probabilities.

Fifth, inclusion of further Omic datasets in SetOmic could bring more insight into the multidimensionality of tumours (e.g., using DNA methylation data or Copy Number Variation profiles). Previous studies show that the omics approach provides a more multi-faceted insight into individual tumour types and is of clear interest in the current cancer research (Joshi et al., 2020; Hasin et al., 2017). However, what should be carefully considered is the type of data used and the information about the disease it provides as well as the downstream task that is explored. We investigated adjusting SetOmic to predict tumour type from sets of methylated sequences and their β-values measured at each probe yielding state-of-the-art results. The possibility of classifying from mutually exclusive features for training and testing with models supports how SetOmic maps the sequence space into a latent space that does not only reduce sparseness, but generalises to other *raw* input spaces.

Finally, the data we used are bulk tumour data which contains a mixture of normal and tumour cells. Incorporating tumour purity data or single-cell omics into the model can reveal further improvements in accuracy, since beyond purity, heterogeneity of clones within the tumour need to be accounted for. This is of particular importance in studies of aggressive tumour sub-types such as the Triple Negative Breast Cancer (Deepak et al., 2020).

Data availability

Our models are available at https://github.com/danilexn/setquence.

Acknowledgements

The authors gratefully acknowledge the GWK support for this project by providing computing time through the Center for Information Services and High Performance Computing (ZIH) at TU Dresden, Germany. The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga and COSMIC, the Catalogue Of Somatic Mutations In Cancer: https://cancer.sanger.ac.uk/cosmic/. The authors thank the TUD BIOTEC Schroeder group, in particular Ali Al-Fatlawi, for the scientific exchange, Andreas Gocht-Zech, for the insightful HPC expertise and Lutz Brusch, for the valuable comments and suggestions on the earlier versions of this manuscript. There are no relevant financial or non-financial competing interests to report.

Declaration of competing interest

There are no Conflicts of Interest to note.

Appendix A. Data and its representation

A.1. Sequence representations

In our investigations we use variant-associated WES and WGS, where a variant can be defined as alternative allele detected against a reference. We take the sequence around this mutated allele, which therefore indicates the location of this variant in the genome. More specifically, for each patient sample i, the genomic coordinates (l_s, l_e) of each variant j are queried against the corresponding reference genome $a_1 ... a_{l_{\text{ref}}}$. This allows defining a sequence $x_{i,j}$ as the so-called *variant-associated* sequence: a string of length $s \cdot 2$ basepairs $a_{l_s-s} ... a_{l_s,l_e} ... a_{l_e+s}$, such that a_{l_s,l_e} contains sample's i allele for a variant j, in tuple $v_{i,j}$. In practice, such string sets are later mapped into integer token matrices such that $\mathbb{X} \subseteq \mathbb{N}^{n_i \times s_{\max}}$, where each string over the dimension n_i is the integer tokenization of the k-merized string (Ji et al., 2021).

Each sequence $x_{i,j}^{\omega_{\text{mut}}}$ of nucleotides $\in \{A, T, C, G\}$ is k-mer coded (i.e., tokenised into an ordered list of $s - k$ sub-strings obtained by concatenating k consecutive nucleotides in a DNA sequence of length s). In each of these sets, a `[CLS]` token is appended to the head and a `[SEP]` token to the tail, and `[PAD]` tokens to zero-pad sequences shorter than 72 k-mers. Then, each k-mer (token) is converted to a numeric, integer value by means of a dictionary d_k. Dictionary size depends on k-mer size, scaling as $4^k + 2$ (k-mers + `[CLS]` + `[SEP]`), with zero values being used for right-padding. Therefore, a context j has $t_{i,j} = n_j - k + 2$ tokens, where n_j is the number of nucleotides and k is the k-mer size, plus the additional delimiting tokens. $x_i^{\omega_{\text{mut}}}$ then has $T_i = \sum_{j=0}^{N}(n_j - k + 2)$ total tokens, maximally bounded by $T_i^{max} = \sum_{j=0}^{N}(n_{max} - k + 2) = N(n_{max} - k + 2)$, where N is the number of annotated mutations for patient i, and n_{max} is the number of base pairs at the longest mutation context, j_{max}. Every element in a mutome set is mapped to the Ensembl ID of the corresponding genomic region for later processing (see Appendices A and B.2).

Individual sequences built using a sequence context of $s \cdot 2 = 64$, yielding sequences of maximally 72 k-mers with $k = 6$, that is, SNPs and short *indels* (i.e., insertions of maximally $72-64 = 8$ bp) were considered as variants. Sequences are treated as sets (i.e., were appended a `[CLS]` token at the head, a `[SEP]` token at the tail, and `[PAD]` tokens to zero-pad sequences shorter than 72 k-mers).

A.2. TCGA: The cancer genome atlas

The Cancer Genome Atlas (TCGA) dataset was used for the somatic mutation data from Whole Exome Sequencing (MAF files) and their corresponding clinical annotations across 33 tumour types (Weinstein et al., 2013) as well as transcriptome expression data in the SetOmic model. This dataset was chosen as a baseline for tumour type classification from omic data (Sections 4.2 and 6.2) and for the implementation of optimisation strategies (Section 4.3). Individual data points (TCGA barcodes) $d \in D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i = \{x_i^{\omega}\}_{\omega \in \Omega} \subset \mathbb{X}$ is the input for a patient i (out of n), from omic sources Ω and $y_i \in \mathbb{Y}$ is a clinically relevant predictor. The somatic mutation (mut) data obtained were used to build, per patient code i, *variant-associated* sequence sets $\{x_{i,j}\}_{j=1}^{n_i}$, mapping to the GRCh38 genome reference assembly coordinates (GenBank accession GCA_000001405.29). In this dataset, $n_{\max} = 500$; that is, a sequence set maximally contains 500 variants per patient. In cases where $n_i > 500$, only the 500 variants with highest frequency in tumour vs. paired normal tissue were kept. Tumour type labels were used as the response variable $y_i \in \mathbb{Y}$ for classification.

The gene expression (exp) count data of 60,483 genomic *loci* per patient were processed into $[0, 1]$ normalised, log2-transformed Fragments Per Kilobase of transcript per Million mapped reads. `NaN` entries were mapped to zeros. It can be represented as a set $\{x_{i,g}^{\omega_{\text{exp}}}\}_{j=1}^{G}$, where

G is maximally 60,483, and each g maps to an Ensembl gene ID. We discretised it into 50 expression levels, such that $x_{i,g}^{\omega_{\text{exp}}} \in \{0, \ldots, 50\}$ via $f_d : \mathbb{R} \to \mathbb{N}$. Each discrete expression level is paired to a corresponding *locus* token g, as the tuple $(x_{i,g}^{\omega_{\text{exp}}}, g)$.

We kept the sequences which overlap between filtered somatic mutation data and transcriptome count data, excluding Acute Myeloid Leukemia cohort due to minimal overlap. 544 healthy and 7518 tumour samples across 32 tumour types remained.

A.3. Cosmic: Catalogue of somatic mutations in cancer

The v95 release (No. 2021) Cosmic (Tate et al., 2019) dataset was used to analyse the impact on tumour type classification upon including Whole-Genome (vs. Whole-Exome) somatic variation data. It provides more than 20 million unique coding and non-coding somatic variants, stratified into 49 tumour types. These data were collected from over 28,000 peer reviewed studies, and more than 39,000 screenings from TCGA and ICGC, among others, providing WGS, WES and epigenomic profiling data. For this study, coding and non-coding variants from WGS/WES were downloaded from https://cancer.sanger.ac.uk. These data were processed into (per-patient code i) variant-associated sequence sets, queried against the GRCh37 reference (GenBank accession GCA_000001405.1). A maximum set cardinality of $n_{\max} = 20{,}000$ was chosen and patients with less than 10 mutation events and tumour types with less than 10 patient samples were filtered out to reduce noise. Therefore, a total of 31,677 unique patient samples across 32 tumour types were used. Furthermore, the work presented in Section 6.1.2 uses only a subset of coding variants. Section 6.1.3 uses a subset of cases for which whole genome sequencing was performed, thus, containing coding and non-coding variants. This subset contained 2913 patients across same 10 tumour types analysed by Chakraborty et al. (2021b), for comparison purposes, after filtering for a minimum of 10 variants per patient.

Appendix B. Explainability through primary attribution methods

We introduce an additional objective: finding a function $A : \mathbb{X} \times \theta \times \mathbb{Y} \to \mathbb{R}$ that quantifies the contribution of inputs x_i (or parameters θ) when evaluating $f_\theta(x_i)$. For our study on tumour omics, we focus on attribution methods that aim to generate a representation of a model's decisions from the input (or other hidden layers) that is easily interpretable by humans.

B.1. Attribution methods

Primary attribution is defined through functions

$$A_{f_\theta}(x_{i,j}, x'_i) = (a_i^1, \ldots, a_i^n) \in \mathbb{R}^n \tag{9}$$

that measure the contribution of input feature j to the output $\hat{y}_i = f_\theta(x_i)$ (with dimension n), for a sample i with respect to a baseline input $x'_i = \{0\}^n$. Attribution was assessed via four different back-propagation-based methods: Integrated Gradients (Sundararajan et al., 2017), Input X Grad (Simonyan et al., 2014), DeepLIFT (Shrikumar et al., 2017) and SHAP (Lundberg and Lee, 2017). In particular, primary attribution was studied from two perspectives: (**i**) at the token level, for a sequence $x_{i,j}^{\text{mut}} = a_1 \ldots a_S$ with S integer tokens a_s, attributions are obtained by back-propagation using any of the attribution methods up to the DNABERT embedding layer and (**ii**) for an input consisting of a set of n elements $\{x_{i,j}\}_{j=1}^n$, attribution is a measure of the relevance of each element represented by an intermediate encoding $e'_{i,j}$ (e.g., individual [CLS] outputs for DNABERT), calculated at an intermediate encoding layer l as $A_l(e'_{i,j})$.

B.2. Measuring the biological significance of attribution scores

To assess the biological significance of attribution scores, a list of the identifiers mapping to each feature j of omic ω was retrieved from inputs $x_{i,j}^{\omega}$ across all patients i. These (Ensembl) IDs are subsequently split into A_-, A_+ and A_{random} lists, depending on the positive, negative or random attribution score at different thresholds $A_{\text{thr}} \in [A_{\min}, A_{\max}]$, such that $A(x_{i,j}^{\omega} \in A_+) \geq A_{\text{thr}}$ and $A(x_{i,j}^{\omega} \in A_-) < A_{\text{thr}}$, $\forall i, j$. Then, attribution values are averaged across patients i for each ω. The Ensembl ID of genes with known impact on cancer were retrieved from GeneCards (Stelzer et al., 2016). TP (true positives), TN (true negatives), FN (false negatives), FP (false positives), and the Area under the $\text{ROC}_{\text{AUC}}(\cdot)$ Curve metric were measured with respect to the gene lists A_-, A_+ and A_{random} from the database.

In the case of mutomes, we additionally measured the relevance of individual nucleotides in tumour type prediction. From the attribution scores at each input token (which can be mapped back to nucleotides), relevant sequence motifs were obtained, using the tools described in Ji et al. (2021).

References

Al-Fatlawi, A., Malekian, N., García, S., Henschel, A., Kim, I., Dahl, A., Jahnke, B., Bailey, P., Bolz, S.N., Poetsch, A.R., Mahler, S., Grützmann, R., Pilarsky, C., Schroeder, M., 2021. Deep learning improves pancreatic cancer diagnosis using RNA-based variants. Cancers 13 (11), 2654. http://dx.doi.org/10.3390/cancers13112654.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., Vojnovic, M., 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. 20 (1), 310.

Aranda, J.F., Reglero-Real, N., Kremer, L., Marcos-Ramiro, B., Ruiz-Sáenz, A., Calvo, M., Enrich, C., Correas, I., Millán, J., Alonso, M.A., 2011. MYADM regulates Rac1 targeting to ordered membranes required for cell spreading and migration. Mol. Biol. Cell 22 (8), 1252–1262. http://dx.doi.org/10.1091/mbc.e10-11-0910.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv:1607.06450.

Bagu, E.T., Santos, M.M., 2011. Friend of GATA suppresses the GATA-induced transcription of hepcidin in hepatocytes through a GATA-regulatory element in the HAMP promoter. J. Mol. Endocrinol. 47 (3), 299–313. http://dx.doi.org/10.1530/jme-11-0060.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1409.0473.

Beristain, A.G., Molyneux, S.D., Joshi, P.A., Pomroy, N.C., Grappa, M.A.D., Chang, M.C., Kirschner, L.S., Privé, G.G., Pujana, M.A., Khokha, R., 2014. PKA signaling drives mammary tumorigenesis through src. Oncogene 34 (9), 1160–1173. http://dx.doi.org/10.1038/onc.2014.41.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., Anandkumar, A., 2018. SignSGD: Compressed optimisation for non-convex problems. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80, PMLR, pp. 560–569, URL: https://proceedings.mlr.press/v80/bernstein18a.html.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., Canaider, S., 2013. An estimation of the number of cells in the human body. Ann. Hum. Biol. 40 (6), 463–471. http://dx.doi.org/10.3109/03014460.2013.807878.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M., Snyder, M., 2012. Annotation of functional variation in personal genomes using regulomedb. Genome Res 22 (9), 1790–1797. http://dx.doi.org/10.1101/gr.137323.112.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., pp. 1877–1901, URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cahyawijaya, S., Yu, T., Liu, Z., Zhou, X., Mak, T.W.T., Ip, Y.Y.N., Fung, P., 2022. SNP2vec: Scalable self-supervised pre-training for genome-wide association study. In: Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, pp. 140–154. http://dx.doi.org/10.18653/v1/2022.bionlp-1.14, URL: https://aclanthology.org/2022.bionlp-1.14.

Caillon, P., Cerisara, C., 2021. Growing neural networks achieve flatter minima. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2021. Springer International Publishing, pp. 222–234.

Chakraborty, S., Martin, A., Guan, Z., Begg, C.B., Shen, R., 2021a. Mining mutation contexts across the cancer genome to map tumor site of origin. Nature Commun. 12 (1), http://dx.doi.org/10.1038/s41467-021-23094-z.

Chakraborty, S., Martin, A., Guan, Z., Begg, C.B., Shen, R., 2021b. Mining mutation contexts across the cancer genome to map tumor site of origin. Nature Commun. 12 (1), 3051. http://dx.doi.org/10.1038/s41467-021-23094-z.

Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R., 2021. An attentive survey of attention models. ACM Trans. Intell. Syst. Technol. 12 (5), http://dx.doi.org/10.1145/3465055.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Int. Res. 16 (1), 321–357.

Corcoran, R.B., Contino, G., Deshpande, V., Tzatsos, A., Conrad, C., Benes, C.H., Levy, D.E., Settleman, J., Engelman, J.A., Bardeesy, N., 2011. STAT3 plays a critical role in KRAS-induced pancreatic tumorigenesis. Cancer Res. 71 (14), 5020–5029. http://dx.doi.org/10.1158/0008-5472.CAN-11-0908.

Cover, T.M., Thomas, J.A., 2005. Entropy, relative entropy, and mutual information. In: Elements of Information Theory. John Wiley & Sons, Ltd, pp. 13–55. http://dx.doi.org/10.1002/047174882X.ch2, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/047174882X.ch2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch2. chapter 2.

Coyle, K., Dean, C., Thomas, M., Vidovic, D., Giacomantonio, C., Helyer, L., Marcato, P., 2018. DNA methylation predicts the response of triple-negative breast cancers to all-trans retinoic acid. Cancers 10 (11), 397. http://dx.doi.org/10.3390/cancers10110397.

Cui, X.-B., yuan Shen, Y., ting Jin, T., Li, S., ting Li, T., mao Zhang, S., Peng, H., xia Liu, C., gang Li, S., Yang, L., Li, N., ming Hu, J., Jiang, J.-F., Li, M., hua Liang, W., Li, Y., tao Wei, Y., zhu Sun, Z., yue Wu, C., Chen, Y.-Z., Li, F., 2015. SLC39A6: a potential target for diagnosis and therapy of esophageal carcinoma. J. Transl. Med. 13 (1), http://dx.doi.org/10.1186/s12967-015-0681-z.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R., 2019. Transformer-XL: Attentive language models beyond a fixed-length context. CoRR abs/1901.02860. arXiv:1901.02860. URL: http://arxiv.org/abs/1901.02860.

Deepak, K., Vempati, R., Nagaraju, G.P., Dasari, V.R., S., N., Rao, D., Malla, R.R., 2020. Tumor microenvironment: Challenges and opportunities in targeting metastasis of triple negative breast cancer. Pharmacol. Res. 153, 104683. http://dx.doi.org/10.1016/j.phrs.2020.104683, URL: https://www.sciencedirect.com/science/article/pii/S1043661819322303.

Dettmers, T., Lewis, M., Shleifer, S., Zettlemoyer, L., 2022. 8-bit optimizers via block-wise quantization. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=shpkpVXzo3h.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv abs/1810.04805.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

Dinh, L., Pascanu, R., Bengio, S., Bengio, Y., 2017. Sharp minima can generalize for deep nets. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML '17, JMLR.org, pp. 1019–1028.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B., 2021. ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach. Intell. PP, 1. http://dx.doi.org/10.1109/TPAMI.2021.3095381.

Erin, N., Podnos, A., Tanriover, G., Duymuş, Ö., Cote, E., Khatri, I., Gorczynski, R.M., 2014. Bidirectional effect of CD200 on breast cancer development and metastasis, with ultimate outcome determined by tumor aggressiveness and a cancer-induced inflammatory response. Oncogene 34 (29), 3860–3870. http://dx.doi.org/10.1038/onc.2014.317.

Evci, U., Pedregosa, F., Gomez, A.N., Elsen, E., 2019. The difficulty of training sparse neural networks. CoRR abs/1906.10732. arXiv:1906.10732. URL: http://arxiv.org/abs/1906.10732.

Ferreira, S., Saraiva, N., Rijo, P., Fernandes, A.S., 2021. LOXL2 inhibitors and breast cancer progression. Antioxidants 10 (2), 312. http://dx.doi.org/10.3390/antiox10020312.

Gal, J., Bailleux, C., Chardin, D., Pourcher, T., Gilhodes, J., Jing, L., Guigonis, J.-M., Ferrero, J.-M., Milano, G., Mograbi, B., Brest, P., Chateau, Y., Humbert, O., Chamorey, E., 2020. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. Comput. Struct. Biotechnol. J. 18, 1509–1524. http://dx.doi.org/10.1016/j.csbj.2020.05.021.

Gupta, A., Dar, G., Goodman, S., Ciprut, D., Berant, J., 2021. Memory-efficient transformers via top-k attention. In: Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing. Association for Computational Linguistics, pp. 39–52. http://dx.doi.org/10.18653/v1/2021.sustainlp-1.5, URL: https://aclanthology.org/2021.sustainlp-1.5.

Hagenah, J., Kühl, K., Scharfschwerdt, M., Ernst, F., 2019. Cluster analysis in latent space: Identifying personalized aortic valve prosthesis shapes using deep representations. In: Cardoso, M.J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., Vercauteren, T. (Eds.), Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning. In: Proceedings of Machine Learning Research, vol. 102, PMLR, pp. 236–249, URL: https://proceedings.mlr.press/v102/hagenah19a.html.

Hasin, Y., Seldin, M., Lusis, A., 2017. Multi-omics approaches to disease. Genome Biol. 18 (1), http://dx.doi.org/10.1186/s13059-017-1215-1.

HDF Group, 1997. Hierarchical data format, version 5. URL: https://www.hdfgroup.org/HDF5/.

Henke, E., Nandigama, R., Ergün, S., 2020. Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. Front. Mol. Biosci. 6, http://dx.doi.org/10.3389/fmolb.2019.00160.

Heo, Y.J., Hwa, C., Lee, G.-H., Park, J.-M., An, J.-Y., 2021. Integrative multi-omics approaches in cancer research: From biological networks to clinical subtypes. Mol. Cells 44 (7), 433–443. http://dx.doi.org/10.14348/molcells.2021.0042.

Hinton, G.E., van Camp, D., 1993. Keeping the neural networks simple by minimizing the description length of the weights. In: Proceedings of the Sixth Annual Conference on Computational Learning Theory. COLT '93, Association for Computing Machinery, pp. 5–13. http://dx.doi.org/10.1145/168304.168306.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop. URL: http://arxiv.org/abs/1503.02531.

Hirata, E., Sahai, E., 2017. Tumor microenvironment and differential responses to therapy. Cold Spring Harb. Perspect. Med. 7 (7), a026781. http://dx.doi.org/10.1101/cshperspect.a026781.

Hochreiter, S., Schmidhuber, J., 1997. Flat minima. Neural Comput. 9 (1), 1–42. http://dx.doi.org/10.1162/neco.1997.9.1.1.

Hua, H., Li, X., Dou, D., Xu, C., Luo, J., 2021. Noise stability regularization for improving BERT fine-tuning. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 3229–3241. http://dx.doi.org/10.18653/v1/2021.naacl-main.258, URL: https://aclanthology.org/2021.naacl-main.258.

Indellicato, R., Zulueta, A., Caretti, A., Trinchera, M., 2020. Complementary use of carbohydrate antigens lewis a, lewis b, and sialyl-lewis a (CA19.9 epitope) in gastrointestinal cancers: Biological rationale towards a personalized clinical application. Cancers 12 (6), http://dx.doi.org/10.3390/cancers12061509, URL: https://www.mdpi.com/2072-6694/12/6/1509.

Iuchi, H., Matsutani, T., Yamada, K., Iwano, N., Sumi, S., Hosoda, S., Zhao, S., Fukunaga, T., Hamada, M., 2021. Representation learning applications in biological sequence analysis. Comput. Struct. Biotechnol. J. 19, 3198–3208. http://dx.doi.org/10.1016/j.csbj.2021.05.039, URL: https://www.sciencedirect.com/science/article/pii/S2001037021002208.

Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V., 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 37 (15), 2112–2120. http://dx.doi.org/10.1093/bioinformatics/btab083.

Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V., 2022. Jerryji1993/DNABERT: DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. https://github.com/jerryji1993/DNABERT. Accessed on 30/05/2022.

Joshi, A., Rienks, M., Theofilatos, K., Mayr, M., 2020. Systems biology in cardiovascular disease: a multiomics approach. Nat. Rev. Cardiol. 18 (5), 313–330. http://dx.doi.org/10.1038/s41569-020-00477-1.

Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=H1oyRlYgg.

Kim, S., Lee, H., Kim, K., Kang, J., 2018. Mut2vec: distributed representation of cancerous mutations. BMC Med. Genom. 11 (S2), http://dx.doi.org/10.1186/s12920-018-0349-7.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6980.

Lee, K., Jeong, H.-o., Lee, S., Jeong, W.-K., 2019b. CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. Sci. Rep. 9 (1), 16927. http://dx.doi.org/10.1038/s41598-019-53034-3.

Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W., 2018. Set transformer. ArXiv abs/1810.00825.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W., 2019a. Set transformer: A framework for attention-based permutation-invariant neural networks. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97, PMLR, pp. 3744–3753, URL: https://proceedings.mlr.press/v97/lee19d.html.

Li, S., Hoefler, T., 2022. Near-optimal sparse allreduce for distributed deep learning. In: Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. PPoPP '22, Association for Computing Machinery, pp. 135–149. http://dx.doi.org/10.1145/3503221.3508399.

Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T., 2018. Visualizing the loss landscape of neural nets. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 31. Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Liao, Y.C., Ruan, J.W., Lua, I., Li, M.H., Chen, W.L., Wang, J.R.Y., Kao, R.H., Chen, J.H., 2011. Overexpressed hPTTG1 promotes breast cancer cell invasion and metastasis by regulating GEF-H1/RhoA signalling. Oncogene 31 (25), 3086–3097. http://dx.doi.org/10.1038/onc.2011.476.

Ling, H., Vincent, K., Pichler, M., Fodde, R., Berindan-Neagoe, I., Slack, F.J., Calin, G.A., 2015. Junk DNA and the long non-coding RNA twist in cancer genetics. Oncogene 34 (39), 5003–5011. http://dx.doi.org/10.1038/onc.2014.456, URL: https://www.nature.com/articles/onc2014456.

Lu, Y., Han, J., 2003. Cancer classification using gene expression data. Inf. Syst. 28 (4), 243–268. http://dx.doi.org/10.1016/S0306-4379(02)00072-8, URL: https://www.sciencedirect.com/science/article/pii/S0306437902000728. Data Management in Bioinformatics.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Marquard, A.M., Birkbak, N.J., Thomas, C.E., Favero, F., Krzystanek, M., Lefebvre, C., Ferté, C., Jamal-Hanjani, M., Wilson, G.A., Shafi, S., Swanton, C., André, F., Szallasi, Z., Eklund, A.C., 2015. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med. Genom. 8 (1), 58. http://dx.doi.org/10.1186/s12920-015-0130-0.

Martínez-Sáez, O., Chic, N., Pascual, T., Adamo, B., Vidal, M., González-Farré, B., Sanfeliu, E., Schettini, F., Conte, B., Brasó-Maristany, F., Rodríguez, A., Martínez, D., Galván, P., Rodríguez, A.B., Martinez, A., Muñoz, M., Prat, A., 2020. Frequency and spectrum of PIK3CA somatic mutations in breast cancer. Breast Cancer Res. 22 (1), http://dx.doi.org/10.1186/s13058-020-01284-9.

Mazlan, A.U., Sahabudin, N.A., Remli, M.A., Ismail, N.S.N., Mohamad, M.S., Nies, H.W., Warif, N.B.A., 2021. A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data. Processes 9 (8), 1466. http://dx.doi.org/10.3390/pr9081466.

Ming, W., Xie, H., Hu, Z., Chen, Y., Zhu, Y., Bai, Y., Liu, H., Sun, X., Liu, Y., Gu, W., 2019. Two distinct subtypes revealed in blood transcriptome of breast cancer patients with an unsupervised analysis. Front. Oncol. 9, http://dx.doi.org/10.3389/fonc.2019.00985.

Murphy, B.J., Andrews, G.K., Bittel, D., Discher, D.J., McCue, J., Green, C.J., Yanovsky, M., Giaccia, A., Sutherland, R.M., Laderoute, K.R., Webster, K.A., 1999. Activation of metallothionein gene expression by hypoxia involves metal response elements and metal transcription factor-1. Cancer Res. 59 (6), 1315–1322.

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., Sharan, R., 2019. To embed or not: Network embedding as a paradigm in computational biology. Front. Genet. 10, http://dx.doi.org/10.3389/fgene.2019.00381.

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., Bellazzi, R., 2020. Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. Front. Oncol. 10, http://dx.doi.org/10.3389/fonc.2020.01030.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S.L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A., Gerton, J.L., O'Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376 (6588), 44–53. http://dx.doi.org/10.1126/science.abj6987, arXiv:https://www.science.org/doi/pdf/10.1126/science.abj6987. URL: https://www.science.org/doi/abs/10.1126/science.abj6987.

Pei, L., 2001. Identification of c-myc as a down-stream target for pituitary tumor-transforming gene. J. Biol. Chem. 276 (11), 8484–8491. http://dx.doi.org/10.1074/jbc.m009654200.

Petegrosso, R., Li, Z., Kuang, R., 2019. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief. Bioinform. 21 (4), 1209–1223. http://dx.doi.org/10.1093/bib/bbz063.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., Droit, A., 2021. Integration strategies of multi-omics data for machine learning analysis. Comput. Struct. Biotechnol. J. 19, 3735–3746. http://dx.doi.org/10.1016/j.csbj.2021.06.030.

Pignatelli, M., Cocca, C., Santos, A., Perez-Castillo, A., 2003. Enhancement of BRCA1 gene expression by the peroxisome proliferator-activated receptor γ in the MCF-7 breast cancer cell line. Oncogene 22 (35), 5446–5450. http://dx.doi.org/10.1038/sj.onc.1206824.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., Caracausi, M., 2019. Human protein-coding genes and gene feature statistics in 2019. BMC Res. Notes 12 (1), 315. http://dx.doi.org/10.1186/s13104-019-4343-8.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67, URL: http://jmlr.org/papers/v21/20-074.html.

Rajbhandari, S., Ruwase, O., Rasley, J., Smith, S., He, Y., 2021. Zero-infinity: Breaking the GPU memory wall for extreme scale deep learning. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC '21, Association for Computing Machinery, http://dx.doi.org/10.1145/3458817.3476205.

Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J., 2007. G:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35 (suppl_2), W193–W200. http://dx.doi.org/10.1093/nar/gkm226.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, URL: https://arxiv.org/abs/1908.10084.

Remli, M.A., Daud, K.M., Nies, H.W., Mohamad, M.S., Deris, S., Omatu, S., Kasim, S., Sulong, G., 2017. K-means clustering with infinite feature selection for classification tasks in gene expression data. In: Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 50–57. http://dx.doi.org/10.1007/978-3-319-60816-7_7.

Richards, R., Jour, G., Tafe, L.J., Pinto, A., Brčić, I., Linos, K., Kerr, D.A., 2022. Primary pulmonary round cell sarcomas: multiple potential pitfalls for the pathologist. Int. J. Surgic. Pathol. 30 (8), 844–852, https://doi.org/10.1177/10668969221091586.

Runde, A.P., Mack, R., Breslin, S.J.P., Zhang, J., 2022. The role of TBK1 in cancer pathogenesis and anticancer immunity. J. Exp. Clin. Cancer Res. 41 (1), 135. http://dx.doi.org/10.1186/s13046-022-02352-y.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108.

Saux, C.J.-L., Tronecker, H., Bogic, L., Bryant-Greenwood, G.D., Boyd, C.D., Csiszar, K., 1999. The LOXL2 gene encodes a new lysyl oxidase-like protein and is expressed at high levels in reproductive tissues. J. Biol. Chem. 274 (18), 12939–12944. http://dx.doi.org/10.1074/jbc.274.18.12939.

Shaheen, F., Verma, B., Asafuddoula, M., 2016. Impact of automatic feature extraction in deep learning architecture. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. http://dx.doi.org/10.1109/DICTA.2016.7797053.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, URL: http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C.C., Ester, M., 2019. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 35 (14), i501–i509. http://dx.doi.org/10.1093/bioinformatics/btz318.

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML '17, JMLR.org, pp. 3145–3153.

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Terre, H.A., Shams, Z., Jamnik, M., Liò, P., 2019. Variational autoencoders for cancer data integration: Design principles and computational practice. Front. Genet. 10, http://dx.doi.org/10.3389/fgene.2019.01205.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio, Y., LeCun, Y. (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings. URL: http://arxiv.org/abs/1312.6034.

Snider, H., Villavarajan, B., Peng, Y., Shepherd, L.E., Robinson, A.C., Mueller, C.R., 2019. Region-specific glucocorticoid receptor promoter methylation has both positive and negative prognostic value in patients with estrogen receptor-positive breast cancer. Clin. Epigenetics 11 (1), 155. http://dx.doi.org/10.1186/s13148-019-0750-x.

Soh, K.P., Szczurek, E., Sakoparnig, T., Beerenwinkel, N., 2017. Predicting cancer type from tumour DNA signatures. Genome Med. 9 (1), 104. http://dx.doi.org/10.1186/s13073-017-0493-2.

Song, B., Li, Z., Lin, X., Wang, J., Wang, T., Fu, X., 2021. Pretraining model for biological sequence data. Brief Funct. Genom. 20 (3), 181–195. http://dx.doi.org/10.1093/bfgp/elab025.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., Lancet, D., 2016. The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr. Protoc. Bioinform. 54 (1), http://dx.doi.org/10.1002/cpbi.5.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML '17, JMLR.org, pp. 3319–3328.

Tang, Y., Ha, D., 2021. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. arXiv:2109.02869.

Tang, Z., Shen, Q., Xie, H., Zhou, X., Li, J., Feng, J., Liu, H., Wang, W., Zhang, S., Ni, S., 2016. Elevated expression of FABP3 and FABP4 cooperatively correlates with poor prognosis in non-small cell lung cancer (NSCLC). Oncotarget 7 (29), 46253–46262. http://dx.doi.org/10.18632/oncotarget.10086.

Tao, Y., Cai, C., Cohen, W.W., Lu, X., 2019. From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. In: Biocomputing 2020. WORLD SCIENTIFIC, http://dx.doi.org/10.1142/9789811215636_0008.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A., 2019. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 47 (D1), D941–D947. http://dx.doi.org/10.1093/nar/gky1015.

Tjoa, E., Guan, C., 2021. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. 32 (11), 4793–4813. http://dx.doi.org/10.1109/tnnls.2020.3027314.

Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V., Waddell, N., 2021. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 13 (1), 152.

van de Wetering, M., Oosterwegel, M., van Norren, K., Clevers, H., 1993. Sox-4, an sry-like HMG box protein, is a transcriptional activator in lymphocytes. EMBO J. 12 (10), 3847–3854. http://dx.doi.org/10.1002/j.1460-2075.1993.tb06063.x.

Vargas, A.C., Reed, A.E.M., Waddell, N., Lane, A., Reid, L.E., Smart, C.E., Cocciardi, S., da Silva, L., Song, S., Chenevix-Trench, G., Simpson, P.T., Lakhani, S.R., 2012. Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. Breast Cancer Res. Treat. 135 (1), 153–165. http://dx.doi.org/10.1007/s10549-012-2123-4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017a. Attention is all you need. CoRR abs/1706.03762. arXiv:1706.03762. URL: http://arxiv.org/abs/1706.03762.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017b. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, B., Wang, A., Chen, F., Wang, Y., Kuo, C.-C.J., 2019. Evaluating word embedding models: methods and experimental results. APSIPA Trans. Signal Inf. Process. 8, e19. http://dx.doi.org/10.1017/ATSIP.2019.12.

Wang, Y., Xu, H., Zhu, B., Qiu, Z., Lin, Z., 2018. Systematic identification of the key candidate genes in breast cancer stroma. Cell. Mol. Biol. Lett. 23 (1), http://dx.doi.org/10.1186/s11658-018-0110-4.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The cancer genome atlas pan-cancer analysis project. Nature Genet. 45 (10), 1113–1120. http://dx.doi.org/10.1038/ng.2764.

Winkler, P., Koch, N., Hornig, A., Gerritzen, J., 2021. OmniOpt – a tool for hyperparameter optimization on HPC. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 285–296. http://dx.doi.org/10.1007/978-3-030-90539-2_19.

Withnell, E., Zhang, X., Sun, K., Guo, Y., 2021. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. Brief. Bioinform. 22 (6), http://dx.doi.org/10.1093/bib/bbab315.

Wu, H.-T., Zhong, H.-T., Li, G.-W., Shen, J.-X., Ye, Q.-Q., Zhang, M.-L., Liu, J., 2020. Oncogenic functions of the EMT-related transcription factor ZEB1 in breast cancer. J. Transl. Med. 18 (1), 51. http://dx.doi.org/10.1186/s12967-020-02240-z.

Xiao, Q., Gan, Y., Li, Y., Fan, L., Liu, J., Lu, P., Liu, J., Chen, A., Shu, G., Yin, G., 2021. MEF2A transcriptionally upregulates the expression of ZEB2 and CTNNB1 in colorectal cancer to promote tumor progression. Oncogene 40 (19), 3364–3377. http://dx.doi.org/10.1038/s41388-021-01774-w.

Xu, H., Yan, M., Patra, J., Natrajan, R., Yan, Y., Swagemakers, S., Tomaszewski, J.M., Verschoor, S., Millar, E.K., van der Spek, P., Reis-Filho, J.S., Ramsay, R.G., O'Toole, S.A., McNeil, C.M., Sutherland, R.L., McKay, M.J., Fox, S.B., 2011. Enhanced RAD21 cohesin expression confers poor prognosis and resistance to chemotherapy in high grade luminal, basal and HER2 breast cancers. Breast Cancer Res. 13 (1), http://dx.doi.org/10.1186/bcr2814.

Xu, Y., Zhong, X., Yepes, A.J.J., Lau, J.H., 2020. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, http://dx.doi.org/10.1109/ijcnn48605.2020.9206891.

Yang, Y., Hodgkinson, L., Theisen, R., Zou, J., Gonzalez, J.E., Ramchandran, K., Mahoney, M.W., 2021. Taxonomizing local versus global structure in neural network loss landscapes. CoRR abs/2107.11228. arXiv:2107.11228. URL: https://arxiv.org/abs/2107.11228.

Yang, M., Huang, L., Huang, H., Tang, H., Zhang, N., Yang, H., Wu, J., Mu, F., 2022. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gkac326.

Yao, D., Zhang, L., Wu, P.L., Gu, X.L., Chen, Y.F., Wang, L.X., Huang, X.Y., 2018. Clinical and misdiagnosed analysis of primary pulmonary lymphoma: a retrospective study. BMC Cancer 18 (1), http://dx.doi.org/10.1186/s12885-018-4184-1.

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018. Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. 13, 55–75.

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., Feng, D.D., 2016. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinform. 17 (S17), http://dx.doi.org/10.1186/s12859-016-1334-9.

Yuan, B., Yang, D., Rothberg, B.E.G., Chang, H., Xu, T., 2020. Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. Sci. Rep. 10 (1), http://dx.doi.org/10.1038/s41598-020-75715-0.

Zhang, S., Wang, Z., Liu, W., Lei, R., Shan, J., Li, L., Wang, X., 2017. Distinct prognostic values of S100 mRNA expression in breast cancer. Sci. Rep. 7 (1), http://dx.doi.org/10.1038/srep39786.

Zhang, X., Xing, Y., Sun, K., Guo, Y., 2021. OmiEmbed: A unified multi-task deep learning framework for multi-omics data. Cancers 13 (12), 3047. http://dx.doi.org/10.3390/cancers13123047.

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., Guo, Y., 2019. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, http://dx.doi.org/10.1109/bibm47256.2019.8983228.