scientific reports

OPEN

Inherently interpretable position-aware convolutional motif kernel networks for biological sequencing data

Jonas C. Ditz[1]✉, Bernhard Reuter[1,2] & Nico Pfeifer[1,2]✉

Artificial neural networks show promising performance in detecting correlations within data that are associated with specific outcomes. However, the black-box nature of such models can hinder the knowledge advancement in research fields by obscuring the decision process and preventing scientist to fully conceptualize predicted outcomes. Furthermore, domain experts like healthcare providers need explainable predictions to assess whether a predicted outcome can be trusted in high stakes scenarios and to help them integrating a model into their own routine. Therefore, interpretable models play a crucial role for the incorporation of machine learning into high stakes scenarios like healthcare. In this paper we introduce Convolutional Motif Kernel Networks, a neural network architecture that involves learning a feature representation within a subspace of the reproducing kernel Hilbert space of the position-aware motif kernel function. The resulting model enables to directly interpret and evaluate prediction outcomes by providing a biologically and medically meaningful explanation without the need for additional post-hoc analysis. We show that our model is able to robustly learn on small datasets and reaches state-of-the-art performance on relevant healthcare prediction tasks. Our proposed method can be utilized on DNA and protein sequences. Furthermore, we show that the proposed method learns biologically meaningful concepts directly from data using an end-to-end learning scheme.

Biological sequences contain valuable information for a wide variety of biological processes. While this property makes them crucial for advances in related research fields, it also provides the potential to improve diagnosis and treatment decisions in healthcare systems. For this reason, a large amount of machine learning approaches that solve learning tasks on biological sequences were developed over the last years. Among others, these approaches include the prediction of splice sites[1] and translation initiation sites[2], predicting binding affinity between proteins and DNA/RNA[3,4], drug resistance prediction[5], or the denoising of biological sequence data[6]. However, trained models can only be safely incorporated into medical routines if their prediction outcomes can be thoroughly interpreted and understood even by domain experts, e.g., healthcare providers like medical practitioners, without strong knowledge in the foundations of machine learning. Kernel methods and statistical models provide the possibility to interpret results within the data's domain, hence, allowing domain experts to judge outcomes using their own expertise. Yet, scalability issues in terms of data size limit their utility considering the rapid increase of available data in medical and biological research. On the other hand, gradient-based learning approaches like neural networks can handle huge data pools with relative ease but are normally developed as black-box models. Although there are model-agnostic techniques to interpret these models, e.g., saliency maps[7] or Shapley additive explanations (SHAP)[8], recent work by Rudin[9] advises the use of inherently interpretable models for high stakes scenarios over post-hoc explaining black-box models. One problem of post-hoc ML explanation models identified by Rudin is their unfaithfulness regarding the original model's computation, which can result in misleading explanations. Sixt and colleagues showed this unfaithfulness for attribution methods by proving that most methods ignored later layers of a model when computing explanations[10]. Furthermore, Bordt and colleagues showed the limitations of *post-hoc* explanations in adversarial contexts[11]. Lipton warned about the danger of optimizing post-hoc methods to produce plausible but misleading explanations[12]. In high stakes scenarios like healthcare,

[1]Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Sand 14, Tübingen 72076, Germany. [2]These authors jointly supervised this work: Bernhard Reuter and Nico Pfeifer. ✉email: jonas.ditz@uni-tuebingen.de; nico.pfeifer@uni-tuebingen.de

decisions made on misleading or wrong explanations can cause dangerous situations with the potential to further harm patients or other vulnerable groups.

In recent years, several efforts were published to combine kernel functions and neural networks[13–16]. Combining these two approaches enhances neural networks with the interpretability and robustness of kernel methods. On the other hand, it allows to extend learning within a reproducing kernel Hilbert space (RKHS) to problems with massive numbers of data points. Recently, Chen and colleagues introduced these efforts into data mining on biological sequences by developing convolutional kernel networks based on a continuous relaxation of the mismatch kernel[17]. Although these models show promising performance, the choice of kernel resulted in the necessity of a post-hoc model for interpretation. Another limitation results from the fact that the mismatch kernel restricts considered k-mer occurrences to a position-independent representation[18,19]. In many medical tasks, however, positional and compositional variability provide key information. One kernel network approach that utilizes positional information is the recurrent kernel network (RKN) proposed by Chen et al.[20]. Another recent approach to incorporate positional information was proposed by Mialon et al.[22]. They utilizes a fixed matrix to introduce positional information. While these architectures showed promising performance capabilities, the chosen architectures resulted once again in black-box models with the need for post-hoc interpretation.

The oligo kernel proposed by Meinicke and colleagues is able to model positional variability and can additionally provide traditional monomer-based representations as well as position-independent k-mer representations as limiting cases[21]. Furthermore, the oligo kernel allows for intuitive and simple interpretation of k-mer relevance and positional variability. However, the oligo kernel cannot be directly incorporated into a convolutional network architecture and does not take into account information provided by compositional variability of motifs. While k-mers are short sequences with fixed letters at each position, motifs are short sequence patterns that can represent more than one possible letter at each position. The above mentioned limitations motivated our work presented here.

This work is structured in the following way. "Methods" section introduces the position-aware motif kernel function and details how to incorporate the position-aware motif kernel into a convolutional kernel layer and how to interpret a trained CMKN model. "Experiments" section provides details regarding the conducted experiments on synthetic and real-world data and the results. Finally, "Discussion" section provides a discussion of presented prediction and interpretation results and "Conclusion" section completes this work with a conclusion.

In summary, our manuscript provides the following contributions:

- We extend convolutional kernel network models for biological sequences to incorporate positional information and make them inherently interpretable, which removes the necessity for post-hoc explanation models. The new models are called convolutional motif kernel networks (CMKNs).
- This extension is achieved by introducing a new kernel function, called position-aware motif kernel, that quantifies the position dependent similarity of motif occurrences.
- We use one synthetic and two real-world datasets to show how our method can be used as a research tool to gain insight into biological sequence data and how CMKNs can provide local interpretation that can help domain experts, e.g., healthcare providers, to quickly interpret and validate prediction outcomes of a trained CMKN model with their domain expertise.

Methods

In the following section, we will introduce our new kernel function and show how this kernel can be used to create inherently interpretable kernel networks.

Position-aware motif kernel

We introduce a new kernel function that incorporates the positional uncertainty of the oligo kernel[21] but is defined for arbitrary sequence motifs. Furthermore, our kernel function can be used to construct a convolutional kernel layer as described by Mairal[16]. Our kernel function is based on two main ideas: First, we introduce a mapping of sequence positions onto the unit circle, which allows us to represent the position comparison term by a linear operation followed by a non-linear activation function. Second, we introduce a k-mer comparison term. This extension enables the kernel function to deal with inexact k-mer matching, which capacitates our kernel function to handle arbitrary sequence motifs. We call our new kernel function position-aware motif kernel (PAM).

The first part of our position-aware motif kernel compares sequence positions. In prior work, e.g., Meinicke et al.[21] or Mialon et al.[22], a quadratic term is usually employed to measure the similarity of positions. We utilize a linear comparison term instead. First, all positions are mapped onto the upper half of the unit circle to create unit ℓ_2-norm vectors: $\tilde{p} = \left((\cos\left(\frac{p}{|\mathbf{x}|}\pi\right), \sin\left(\frac{p}{|\mathbf{x}|}\pi\right)\right)^T$, where $|\mathbf{x}|$ denotes the length of the corresponding sequence. Due to the position vectors now having unit ℓ_2-norm, the position comparison term can be written as follows: $-\frac{1}{4\sigma}\left\|\tilde{p}-\tilde{q}\right\|_2^2 = \frac{1}{2\sigma}(\tilde{p}^T\tilde{q}-1)$. This allows us to define the following position comparison kernel function over pairs of sequence positions:

$$K_{\text{position}}(p, q) = \exp\left(\frac{\beta}{2\sigma^2}\left(\tilde{p}^T\tilde{q}-1\right)\right), \tag{1}$$

where β is a scaling parameter that compensates for the reduced absolute distance between sequence positions due to the introduced mapping and σ is a positional uncertainty parameter similar to the homonymous σ parameter of the oligo kernel.

The second part of our position-aware motif kernel compares sequence motifs. For biological sequences, a motif describes a nucleotide or amino acid pattern of a certain length. Sequence motifs can be written in form of a normalized position frequency matrix (nPFM), which is a matrix in $\mathbb{R}_{+}^{|\mathbf{A}|\times k}$ with $|\mathbf{A}|$ being the size of the alphabet over which the motif is created and k being the length of the motif. An nPFM has to fulfill the additional constraint that each column has unit ℓ_2-norm (see Supplement for more details). For two motifs ω and ω' of length k given as flattened nPFMs, i.e., the columns are concatenated to convert the matrix into a vector, we define the following motif comparison kernel function:

$$K_{\text{nPFM}}(\omega, \omega') = \exp\left(\alpha\left(\omega^T\omega' - k\right)\right). \tag{2}$$

This function will become one if the two motifs match exactly and will approach zero with increasing difference of the two motifs. The parameter α determines how fast the function approaches zero and, hence, specifies the influence of inexact matching motifs.

We define our position-aware motif kernel by forming the product kernel using the functions introduced in Eqs. (1) and (2) and aggregating the kernel evaluation of all motif-position pairs with a sum. In other words, the position-aware motif kernel for pairs of sequences $\mathbf{x}$ and $\mathbf{x}'$ over an alphabet $\mathbf{A}$ is given by:

$$K_{\text{PAM}}(\mathbf{x}, \mathbf{x}') = C\sum_{p=1}^{|\mathbf{x}|}\sum_{q=1}^{|\mathbf{x}'|} K_0((\boldsymbol{\omega}_p, p), (\boldsymbol{\omega}_q, q)), \tag{3}$$

with

$$K_0((\omega_p, p), (\omega_q, q)) = K_{\text{nPFM}}(\omega_p, \omega_q)\cdot K_{\text{position}}(p, q) = \exp\left(\alpha\left(\omega_p^T\omega_q - k\right) + \frac{\beta}{2\sigma^2}\left(\tilde{p}^T\tilde{q} - 1\right)\right).$$

Here, $|\mathbf{x}|$ and $|\mathbf{x}'|$ are the lengths of the respective sequences, ω_p is the motif of length k starting at position p in sequence $\mathbf{x}$ represented as a flattened nPFM, and ω_q is defined analogously to ω_p but for sequence $\mathbf{x}'$. The constant $C = \sqrt{\frac{\pi^2\sigma^2}{2\alpha\beta}}$ results from the derivation of the motif kernel matrix elements as the inner product of two sequence representatives $\phi_{\mathbf{x}}, \phi_{\mathbf{x}'}$ in the feature space of all motifs as detailed in the Supplement.

Extracting a feasible kernel layer using Nyström's method

Mairal et al. showed that a variant of the Nyström method[23,24] can be used to incorporate learning within a reproducing kernel Hilbert space (RKHS) into neural networks[15,16]. We use the same approach to construct a finite-dimensional subspace of the RKHS $\mathscr{H}$ over motif-position pairs that is implicitly defined by K_0 and incorporate learning within this subspace into a neural network architecture.

Consider a set of n anchor points $z_1, \ldots, z_n$, where each anchor point is a motif-position pair $z_i = (\omega_{z_i}, p_{z_i})$. We define an n-dimensional subspace $\mathscr{E}$ of $\mathscr{H}$ that is spanned by a set of anchor points, i.e.

$$\mathscr{E} \quad = \text{Span}\left(\phi_{z_1}, \ldots, \phi_{z_n}\right), \tag{4}$$

where ϕ_{z_i} denotes the projection of each anchor point into the RKHS $\mathscr{H}$. Utilizing the kernel trick, a motif-position pair can be projected onto $\mathscr{E}$ without explicitly calculating the images of the anchor points $\phi_{z_1}, \ldots, \phi_{z_n}$. This natural parametrization is given by[16]

$$\psi((\omega, p)) = K_{ZZ}^{-\frac{1}{2}} K_Z((\omega, p)). \tag{5}$$

Here, $K_{ZZ} = (K_0(z_i, z_j))_{i=1,\ldots,n;j=1,\ldots,n}$ is the Gram matrix formed over the anchor points, $K_{ZZ}^{-\frac{1}{2}}$ is the (pseudo)-inverse square root of the Gram matrix, and $K_Z((\omega, p)) = (K_0(z_1, (\omega, p)), \ldots, K_0(z_n, (\omega, p)))^T$. We follow the procedure proposed in prior work[16,17] to initialize anchor points. First, we sample a set of $m >> n$ motif-position pairs from the training data. Afterwards, we perform k-means clustering with euclidean distance metric using k-means++ initialization to get n cluster centers of the sampled set. After convergence, we enforce the nPFM constraints onto the cluster centers. With initialized anchor points, CMKN models can be trained by a simple end-to-end learning scheme. A schematic overview over a CMKN model for DNA input together with a visualization of the information flow within the network is shown in Fig. 1.

Interpreting a CMKN model

The main intuition behind the position-aware motif kernel is to detect similarities between motifs, even if they occur at a certain distance from each other and even if the nPFM underlying the motif is different to a certain degree. In this way, our approach extends previous approaches like the oligo kernel[21] and the weighted degree kernel with shifts[25], which only evaluated exact k-mer matches. However, our kernel is based on a concept that we call motif functions which are extensions of the oligo functions introduced by Meinicke et al.[21] (see Supplement for details). A motif function represents the nPFM and position(s) of occurrence of the corresponding motif with a smoothing of the position to account for positional uncertainty. Apart from providing a biologically meaningful feature representation, the use of a kernel based on motif functions allows for a direct interpretation of a trained CMKN model without the need for post-hoc methods. If the CMKN model consists only of linear fully-connected layers after the kernel layer, as strictly applied throughout this study, important sequence positions and corresponding motifs can be directly inferred from the learned weights and anchor points, since this

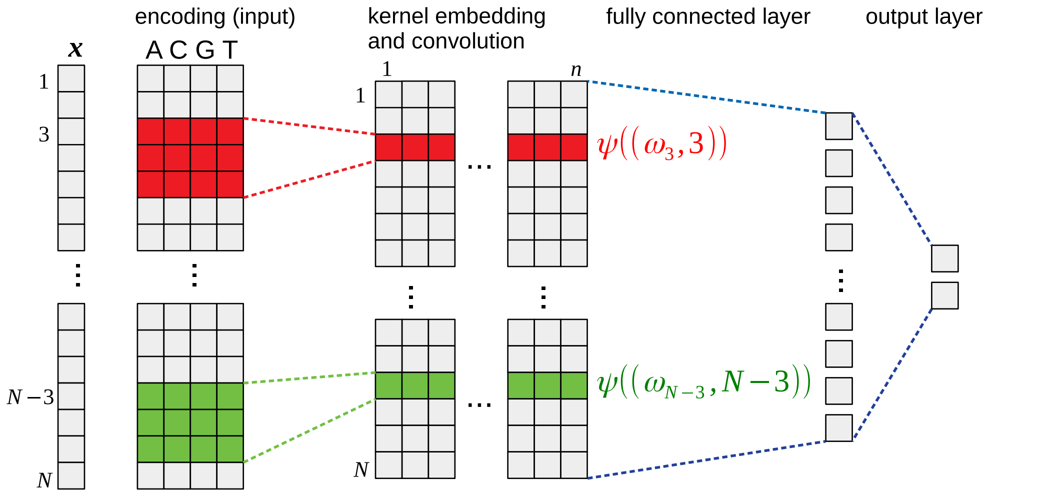

Figure 1. Schematic overview of an CMKN model. Each motif-position pair of the input is projected onto the subspace of the RKHS by the kernel layer. Afterwards, the projected input is classified using one or several linear fully-connected layers.

ensures that only linear combinations of the learned feature representations are considered. The importance of a sequence position for a certain class can be assessed by calculating the mean positive weight of the edges that connect the position with the output state that corresponds to the class. The importance ι of position p for class c can thereby be expressed as:

$$\iota_c^p = \frac{1}{|N_p|} \sum_{n \in N_p} \tilde{\iota}_{n,c} \ , \qquad \tilde{\iota}_{n,c} = \begin{cases} \sum_{\{m | m \in N^{(n)}\}} w_{n,m} \tilde{\iota}_{m,c}, & \text{if } N^{(n)} \cap N^{(O)} = \emptyset \\ 1, & \text{if } N^{(n)} \cap N^{(O)} = o_c \ . \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Here, N_p denotes the set of neurons contributing to the importance of position p, $N^{(n)}$ denotes the set of neurons from the next layer connected by an edge with positive weight to neuron $n \in N_p$, $w_{n,m}$ denotes the weight of the edge connecting neuron n with neuron $m \in N^{(n)}$, and $N^{(O)}$ denotes the set of $|c|$ neurons o_c, each representing a single class, in the output layer. Furthermore, the motif associated with the class at that position is retrieved by identifying all learned motifs with positive weights and calculating the weighted mean motif using the learned weights. This procedure is similar to inferring feature importances from the primal representation using the learned parameters of a SVM. Said utilization of the primal representation is possible for linear kernels and most string kernels[21]. The importance of each amino acid at each position of the motif can be directly accessed by sorting the rows of each column of the associated nPFM in decreasing order. Additionally, motif functions enhance CMKN models with the ability to compute local interpretations, i.e., an explanation of prediction results for single inputs within the data's domain. For an input sequence and a learned motif-position pair, we can estimate the importance of that pair by calculating the ℓ_2-norm of the corresponding motif function. To assess the class that a model associates with an important position, the class-specific motifs that were learned by a model at that position can be retrieved and ranked by the ℓ_2-norm of the motif functions on the input sequence. The motif with the highest ℓ_2-norm determines which class a model assigns to the position. We show an exemplary visualization for domain experts of this procedure in Fig. 3b.

Experiments

We used synthetic data to evaluate CMKN's ability to recover meaningful sequence patterns. Furthermore, we evaluated the performance capability of our proposed method on two different prediction tasks: antiretroviral drug resistance prediction and splice site recognition.

Recovering meaningful patterns in synthetic data

In order to assess whether CMKN models can reliably recover distinct biological patterns from sequences, we created a synthetic dataset containing 1000 randomly generated DNA sequences of length 100. The set was equally split into negative and positive sequences, with a distinct motif embedded into each class of sequences at a specific position (see Fig. 2a for the embedded motifs). For negative sequences, the motif was embedded at position 20 with a positional uncertainty of ± 5 positions. For positive sequences, the motif was embedded at position 80 with a positional uncertainty of ± 5 positions. The compositional variability shown in Fig. 2a can be understood in a way that one-third of the 5-mers embedded into negative sequences had a thymine at position 2 while two-third of the k-mers had a guanine. This is equivalent for the other motif positions with compositional variability. By creating a synthetic dataset with this procedure, we made sure that the data contains positional and compositional variability that are important for the prediction task. We trained a CMKN model using a motif length of 5 and a positional uncertainty parameter of 4. For the kernel layer, we chose 50 anchor points.

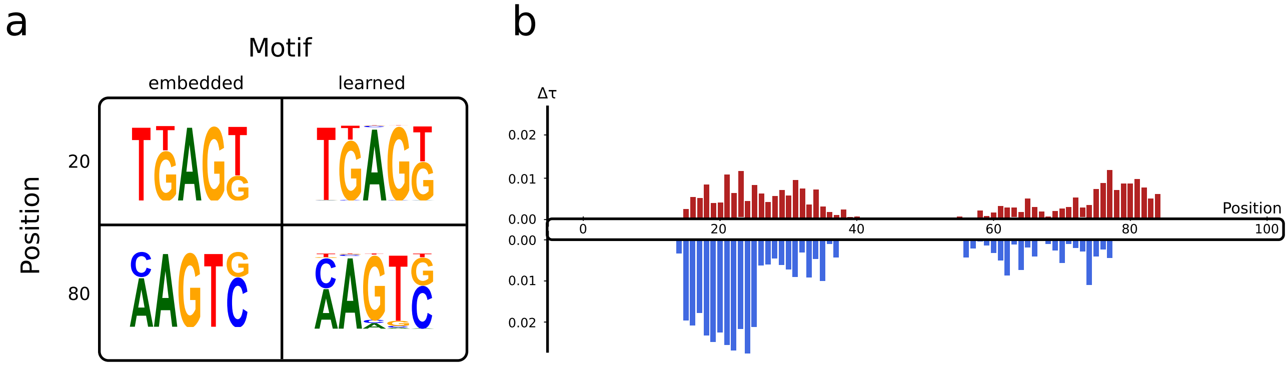

Figure 2. Evaluation of the interpretation capabilities of CMKN using synthetic data. (**a**) The matrix shows the embedded motifs (left column) and the motifs learned by CMKN (right column). The first row shows the motif at position 20 which was only embedded into negative sequences. The second row shows the motif at position 80 which was only embedded into positive sequences. (**b**) Positional feature importance of CMKN on the synthetic data. Each bar shows the derivation from the mean positional feature importance for the corresponding sequence position. Red bars indicate importance for the positive class and blue bars indicate importance for the negative class.

The other kernel hyperparameters were set to $\alpha = 1$ and $\beta = 1000$. The model was trained for 50 epochs using the binary cross-entropy with logits loss function.

Figure 2 shows the results of our experiment with the synthetic dataset. We recovered the positional feature importance values as well as the learned motifs at position 20 and 80 using the procedure described in "Interpreting a CMKN model" section. As clearly visible on the left side, CMKN is able to recover the two embedded motifs with high similarity using simple end-to-end learning without post-hoc model optimization. Furthermore, the right side of Fig. 2 shows that CMKN is able to detect the relevant areas of biological sequences.

Prediction of antiretroviral drug resistance

When choosing a personalized treatment combination for HIV-infected people, it is crucial to know the resistance profile of the viral variants against available drugs. It has been shown that the genetic sequence of a virus can be used to predict resistance against certain antiretroviral drugs[5]. We performed resistance prediction for drugs representing the three most commonly used antiretroviral drug classes against HIV infections: Nucleoside reverse-transcriptase inhibitors (NRTIs), non-nucleoside reverse-transcriptase inhibitors (NNRTIs), and protease inhibitors (PIs). This prediction task was chosen for evaluation of the proposed method, since it remains an highly important problem in the treatment of HIV infections and the acquired immune deficiency syndrome (AIDS) and is often considered as a role model for precision medicine.

Amino acid sequences of virus protein variants with corresponding drug resistance information were extracted from Stanford University's HIV drug resistance database (HIVdb)[26,27]. An overview of the available data for each of the drugs included in the evaluation can be found in the Supplement. The network architecture used for HIV drug resistance prediction consists of a single convolutional motif kernel layer followed by two fully-connected layers. The first fully-connected layer projected the flattened output of the kernel layer onto 200 nodes and the second fully-connected layer had two output states, one for the susceptible class and one for the resistant class. The motif length and the hyperparameter α of the kernel function were both fixed to 1 based on prior biological knowledge (for details see Supplement material). The scaling hyperparameter β was fixed to $\frac{|\mathbf{x}|^2}{10}$ with $|\mathbf{x}| = 99$ for PI datasets and $|\mathbf{x}| = 240$ for NRTI/NNRTI datasets. This compensates for the transformation of sequence positions (for details see Supplement Material). The number of anchor points and the positional uncertainty parameter σ were optimized using a grid search (for details see Supplement Material). Due to the limited number of available samples, each model was trained using a fivefold stratified cross-validation. The data splits for each fold were fixed across models to ensure the same training environment for each hyperparameter combination. Training success was evaluated using the performance measures accuracy, F1 score, and area under the receiver operating characteristic curve (auROC). Due to the fact that some datasets were highly unbalanced, we also included the Matthew's correlation coefficient (MCC)[28] in the performance assessment.

Mean performances achieved for each of the three investigated drug classes can be found in Table 1. Our method was able to achieve high accuracy, F1 score, and auROC values for each drug class. Even though the classification problem is highly imbalanced for some of the tested drugs, our model is still able to achieve a high Matthew's correlation coefficient (MCC) value with mean MCC performance exceeding 0.75 for each of the three investigated drug classes. We compared CMKN's performance to previously used models for HIV drug resistance prediction: SVMs with polynomial kernel[5] and random forest (RF) classifiers[29]. Furthermore, we included a SVM utilizing the oligo kernel and the CKN_{seq} model[17] into our analysis. Additionally, we performed an ablation test by replacing the kernel layer with a standard convolutional layer to investigate the influence of our kernel architecture onto prediction performance (denoted by CNN in Table 1). The results for all models can be found in Table 1. Our method either outperformed the competitors or achieved similar performance.

Drug class	Model	Accuracy	F1 score	auROC	MCC
PI	SVM_{poly}	0.90 ± 0.04	0.83 ± 0.09	0.95 ± 0.03	0.75 ± 0.10
	$\text{SVM}_{\text{oligo}}$	**0.92 ± 0.03**	0.86 ± 0.09	**0.97 ± 0.03**	**0.81 ± 0.09**
	RF	**0.92 ± 0.04**	0.85 ± 0.13	**0.97 ± 0.03**	0.79 ± 0.13
	CNN	0.91 ± 0.3	0.84 ± 0.11	0.94 ± 0.05	0.77 ± 0.11
	CKN_{seq}	0.84 ± 0.05	0.72 ± 0.12	0.88 ± 0.05	0.60 ± 0.11
	CMKN	**0.92 ± 0.03**	**0.87 ± 0.09**	0.96 ± 0.03	**0.81 ± 0.10**
NRTI	SVM_{poly}	0.86 ± 0.06	0.82 ± 0.09	0.90 ± 0.05	0.70 ± 0.12
	$\text{SVM}_{\text{oligo}}$	0.88 ± 0.05	0.85 ± 0.09	**0.94 ± 0.03**	0.75 ± 0.10
	RF	0.88 ± 0.06	0.84 ± 0.12	**0.94 ± 0.04**	0.74 ± 0.15
	CNN	0.88 ± 0.05	0.85 ± 0.09	0.93 ± 0.04	0.74 ± 0.12
	CKN_{seq}	0.79 ± 0.06	0.73 ± 0.12	0.85 ± 0.05	0.54 ± 0.13
	CMKN	**0.89 ± 0.05**	**0.86 ± 0.09**	0.93 ± 0.05	**0.76 ± 0.11**
NNRTI	SVM_{poly}	0.82 ± 0.06	0.76 ± 0.11	0.84 ± 0.06	0.63 ± 0.14
	$\text{SVM}_{\text{oligo}}$	0.89 ± 0.05	0.86 ± 0.11	0.94 ± 0.05	0.79 ± 0.12
	RF	0.88 ± 0.05	0.85 ± 0.09	0.93 ± 0.07	0.75 ± 0.12
	CNN	0.89 ± 0.04	0.86 ± 0.08	0.94 ± 0.06	0.78 ± 0.10
	CKN_{seq}	0.73 ± 0.06	0.63 ± 0.16	0.78 ± 0.08	0.42 ± 0.15
	CMKN	**0.91 ± 0.03**	**0.89 ± 0.06**	**0.95 ± 0.05**	**0.81 ± 0.08**

Table 1. Mean performance and standard derivation of prediction models for three different HIV drug classes: PIs, NRTIs, NNRTIs. Models include polynomial kernel SVMs (SVM_{poly}), oligo kernel SVMs ($\text{SVM}_{\text{oligo}}$), random forests (RF), convolutional neural networks (CNN), convolutional kernel networks (CKN_{seq}), and convolutional motif kernel networks (CMKN). Highest values are displayed in bold.

Utilizing CMKN's interpretation capabilities to identify resistance mutation positions and motifs
Apart from assessing CMKN's prediction performance, we investigated how well our models were able to learn biologically meaningful patterns from drug resistance data. For each sequence position, we calculated the position importance for each class as described in "Interpreting a CMKN model" section and identified peaks with a sliding window approach, i.e., the mean importance of a window of length 11 around each position was calculated and subtracted from the position importance. We selected the 10 highest peaks identified using this sliding window approach. For each peak position, the associated mean motif (of length one) as well as the two most important amino acids of this mean motif were retrieved using the approach described in "Interpreting a CMKN model" section. To get position importance and mean motifs for one of the three investigated drug classes (PIs, NRTIs, and NNRTIs), we averaged the importance values as well as the mean motifs over all models that belong to drugs of the same drug class (8 models for PIs, 6 models for NRTIs, and 3 models for NNRTIs). Figure 3a displays the top ten position of the resistant and susceptible class together with the top two amino acids of the corresponding mean motif for each of the three investigated drug classes. The results indicate that CMKN models are able to learn biologically meaningful patterns from real-world datasets. The most important positions identified by CMKN models correspond mainly to known drug resistance mutation (DRM) positions while the corresponding learned motifs are focused on DRMs. This result is consistent for all three tested drug types. However, CMKN models provide more than a global interpretation. Figure 3b shows the result of CMKN's local interpretation capabilities (as described in "Interpreting a CMKN model" section) for the model trained on nelfinavir (NFV) data and three randomly selected isolates. First we identified the ten most important positions learned by the model. Afterwards, we retrieved the resistant and susceptible motifs for each position from the trained model. Using the motif functions, we were able to identify which positions the model indicated to be informative for the susceptible class and which positions were indicated to be informative for the resistant class using the procedure described in "Interpreting a CMKN model" section. This local interpretation shows biologically meaningful patterns and can be used by domain experts to verify a prediction made by the model. For a more detailed discussion of the visualization results, see "Discussion" section.

Splice site prediction

The recognition of splice sites is an important task in healthcare, since it can uncover genetic variants and differences in protein composition in individual patients. It consists of two classification problems: distinguishing decoys from true targets for acceptor sites and for donor sites.

We used two benchmarks to assess performance of our model on the splice site recognition task: NN269[33] and DGSplicer[34]. Both benchmarks provide test sets and are highly imbalanced. Details on training and test sets for both benchmarks can be found in the Supplement. For splice site recognition, we used the same architecture that was used for the HIV drug resistance prediction. The hyperparameter α was again fixed to 1. We similarly fixed the scaling parameter to $\beta = \frac{|\mathbf{x}|^2}{10}$ with $|\mathbf{x}| = 90$ for acceptor sequences and $|\mathbf{x}| = 15$ for donor sequences on the NN269 benchmark and $|\mathbf{x}| = 36$ for acceptor sequences and $|\mathbf{x}| = 18$ for donor sequences on the DGSplicer benchmark. The number of anchor points, the motif length k, and the positional uncertainty parameter were optimized using a grid search with 5-fold stratified cross-validation on the training data (details can be found in

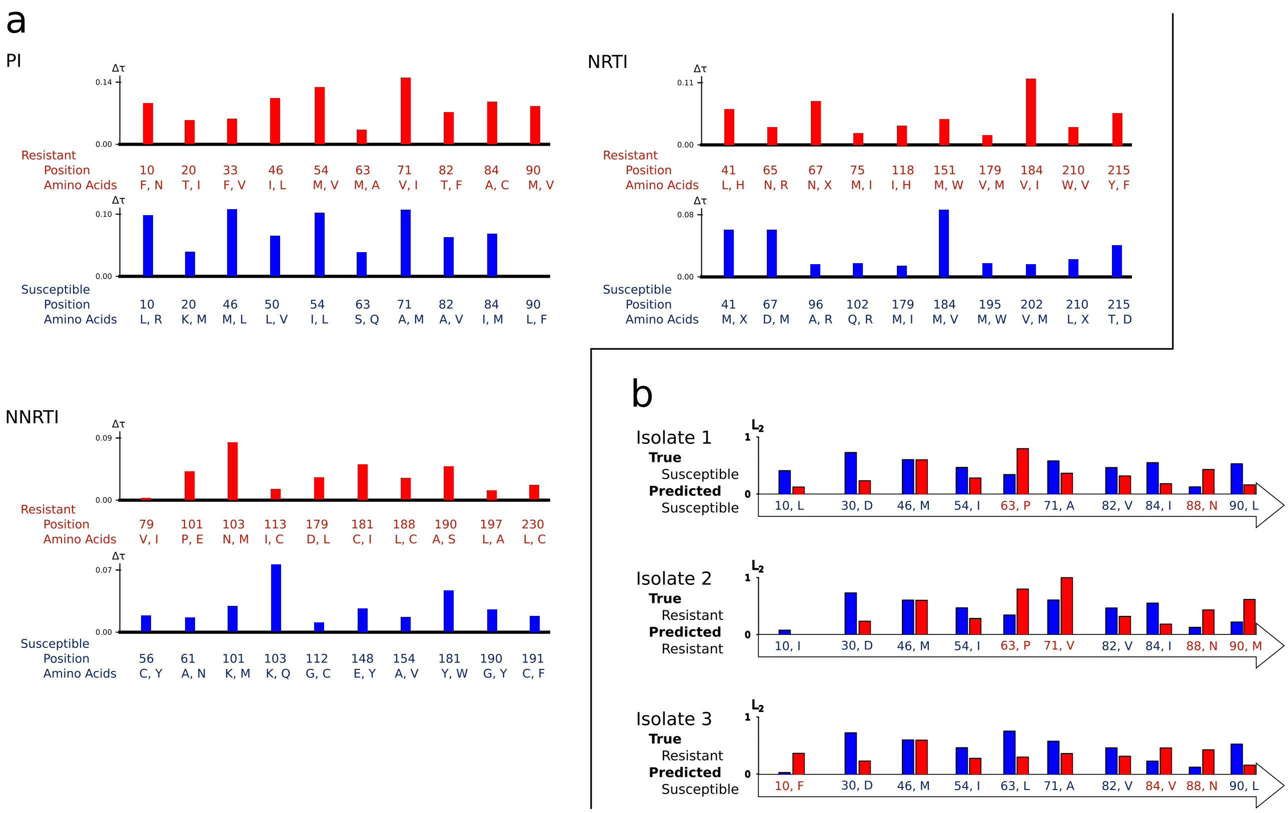

Figure 3. (**a**) (Global Interpretation): CMKNs can be used for data mining on biological sequences. The ten most important positions learned by the model, together with the top two contributing amino acids, are displayed. The height of the bar plot at each position indicates the normalized feature importance of that position, i.e., the mean position feature importance was subtracted from the feature importance of the specific position. Higher bars indicate more important positions. The importance of each sequence position was calculated as described in "Interpreting a CMKN model" section and peaks were identified using a sliding window approach with a window length of 11. Afterwards, the model's learned motifs associated with the ten highest peaks were calculated (see "Interpreting a CMKN model" section) and the two amino acids with the highest contribution to these motifs were selected. Positions displayed in red (blue) are associated with the resistant (susceptible) class. (**b**) (Local Interpretation): We created an exemplary visualization of CMKN's explanation capabilities. Prediction results of the nelfinavir (NFV) model for three randomly chosen input sequences are visualized by showing the learned top ten positions together with the amino acid occurring at the respective position in the input. For each position, the motif functions of the learned motifs are evaluated to identify the one with the highest ℓ_2-norm on the input (see "Interpreting a CMKN model" section). If the corresponding motif is a learned resistance (susceptibility) associated motif, the position-amino-acid pair is highlighted in red (blue). The height of the bars above each position corresponds to the ℓ_2-norm of the corresponding susceptible (blue) and resistant (red) motif functions (scaled between 0 and 1). For each isolate, the true and predicted label is displayed.

the Supplement). The model with the best hyperparameter combination was retrained on the whole training set and evaluated using the test set. Training success was evaluated using the area under the precision-recall curve (auPRC), to account for class imbalance, and the auROC to enable comparison with previously published models.

We compared our method to several methods that were previously applied on splice site recognition. These included higher order Markov Chain (MC) classifiers, SVMs with the locality improved kernel (LIK), the weighted degree kernel (WD), and the weighted degree kernel with shifts (WDS) published in[30], a method combining higher order Markov Chains and SVMs with polynomial kernel (MC-SVM) published in[31], and a CNN architecture called SpliceRover[32]. On the NN269 benchmark, our method performed comparable to other methods in terms of auROC and outperformed almost all competitors in terms of auPRC (see Table 2). On the DGSplicer benchmark, our method performed comparable to other methods in terms of auROC, while substantially outperforming all competitors in terms of auPRC (see Table 2). An evaluation of CMKN's interpretation on the splice site prediction task can be found in the Supplement.

Discussion

In this work, we introduced convolutional motif kernel networks (CMKNs), a convolutional network architecture that allows for end-to-end learning within a subspace of our proposed position-aware motif kernel's RKHS.

	NN269				DGSplicer			
	Acceptor		Donor		Acceptor		Donor	
Model	auROC	auPRC	auROC	auPRC	auROC	auPRC	auROC	auPRC
MC	0.97	0.88	0.98	0.92	0.97	0.31	**0.98**	0.42
MC-SVM	0.97	0.88	0.98	0.90	0.95	–	0.95	–
LIK	0.98	0.92	0.98	0.93	–	–	–	–
WD	0.98	0.93	**0.99**	0.93	**0.98**	0.32	**0.98**	0.40
WDS	**0.99**	**0.94**	0.98	0.93	0.97	0.29	0.97	0.36
SpliceRover	**0.99**	–	0.98	–	–	–	–	–
CMKN	0.97	**0.94**	0.98	**0.96**	0.97	**0.65**	**0.98**	**0.65**

Table 2. Test performance on splice site benchmarks. The displayed methods include higher order Markov Chain (MC) classifiers[30], a combination of higher order Markov Chains and SVMs with polynomial kernel (MC-SVM)[31], SVMs with the locality improved kernel (LIK)[30], SVMs with the weighted degree kernel (WD)[30], SVMs with the weighted degree kernel with shifts (WDS)[30], SpliceRover[32], and our CMKN. Highest numbers are shown in bold. Dashes indicate missing values in the original manuscripts.

By combining a convolutional network architecture with a kernel function, our model is able to perform robust end-to-end learning on relatively small datasets as was shown on data from Standford's HIVdb. Our model was able to generalize to validation data with only a few hundred training samples even in highly unbalanced scenarios. However, due to the fact that our model is based on a standard convolutional network architecture, CMKNs can easily be used on datasets with several hundreds of thousands of samples, as shown on the splice site prediction benchmarks. This allows to utilize our proposed kernel function on very large datasets, something that would be notoriously hard using standard kernel methods like SVMs, since the calculation of a large Gram matrix for our position-aware motif kernel is computationally very demanding.

We included accuracy and auROC as performance measures in our evaluation, since both measures are often used in the ML literature. However, on imbalanced data their informative value is decreased due to a bias towards the majority class[35] as can be seen by considering the auROC vs. auPRC performances on the DGSplicer benchmark in Table 2. Therefore, we included measures that provide better insights on imbalanced data with few positives: F1 and MCC for HIV drug resistance prediction and auPRC for splice site prediction. Considering F1, MCC, and auPRC, our model performed similar or better compared to all other models.

Another advantage of introducing kernel function evaluation into a neural architecture is the possibility to overcome the black-box nature of neural networks. Since learning within our proposed kernel layer admits a projection onto a subspace of the RKHS of our position-aware motif kernel, each output node of the kernel layer is associated with a position-motif pair. This allows for a biological interpretation of the learned weights associated with each node of the kernel layer. With these global interpretation capabilities, our model can be used as a tool for data mining on biological sequence data. We showed on HIV drug resistance data that our model is able to learn biologically meaningful patterns using standard end-to-end learning methods (see Fig. 3a). The majority of the ten most important positions correspond to known DRM positions (nine for PI drugs, eight for NRTI drugs, seven for NNRTI drugs). Furthermore, the top amino acids in the learned resistant motifs reflect known DRMs while the top amino acids in the learned susceptible motifs either reflect the wildtype or none DRMs. There are three exceptions where the susceptible motif features amino acids that lead to an increased drug resistance. These exceptions are leucine (L) and valine (V) at position 50 for PI drugs, valine (V) at position 184 for NRTI drug, and aspartic acid (D) at position 215 for NRTI drugs. However, these exceptions appear to occur due to the averaging of motifs over all drugs for a specific drug class. While all of the four mentioned mutations cause an increase resistance against a subset of drugs[36–38], they are also a cause of increased susceptibility or have no effect for other drugs[26,39,40]. Valine at position 50 reduces susceptibility to fosamprenavir (FPV), lopinavir (LPV), and darunavir (DRV) but increases susceptibility to tipranavir (TPV). Leucine at position 50 confers high-level resistance to atazanavir (ATV) but increases susceptibility to all other PI drugs. For NRTI drugs, valine at position 184 reduces susceptibility to lamivudine (3TC) but increases susceptibility to zidovudine (AZT), stavudine (d4T), and tenofovir (TDF). At position 215, a mutation to aspartic acid is a so-called thymidine analog mutation that reduces susceptibility to AZT and d4T but has no effect on susceptibility to all other NRTI drugs.

Apart from the data mining capabilities of our proposed CMKN model, the motif functions enrich our model with the capability to provide local interpretations for prediction results within the data's domain. Figure 3b shows an example of the visualization capabilities of our CMKN model using nelfinavir (NFV) data, one of the PI drugs. The figure was created with the following steps. First, the trained NFV model was used to build the susceptible and resistant motifs for each of the ten most informative resistance positions learned for the NFV drug, as described in "Interpreting a CMKN model" and "Utilizing CMKN's interpretation capabilities to identify resistance mutation positions and motifs" sections. Afterwards, we assessed for each position if the model relates the position to the susceptible or resistant calss, as described in "Interpreting a CMKN model" section. For the first input, which was correctly classified as susceptible, the visualization shows that the model associated susceptible motifs with each of the positions except for position 63 and 88. However, a domain expert can quickly verify that the model falsely classified that the amino acid asparagine (N) at position 88 indicates resistance, since asparagine corresponds to the wildtype and is therefore in accordance with a susceptible isolate. Furthermore, there is no experimental evidence supporting that position 63 is associated with a resistance causing mutation. Using this knowledge, a domain expert can make an educated decision that the prediction is correct. For the correctly classified resistant input, the model associates resistant motifs with positions 63, 71, 90, and, again falsely, with position 88. Since a mutation to methionine (M) at position 90 causes a strong resistance against NFV[36,41–43], a domain expert could again directly validate the prediction result. The interpretation capabilities gain importance in case of a wrongly classified input as shown in the bottom part of Fig. 3b. Here a domain expert would see that a susceptibility to NFV was predicted while three positions, 10, 84, and 88, are associated with resistant motifs. We again have the previously described, apparently systematic, error at position 88, but a mutation to valine (V) at position 84 causes a moderate resistance against NFV[44]. Additionally, a mutation to phenylalanine (F) at position 10 is known to be associated with reduced in vitro susceptibility to NFV[36,45]. Thus, the visualization provides the domain expert with all information needed to treat the prediction outcome with the adequate caution. This shows that utilizing the proposed kernel formulation in our model's architecture, together with the proposed motif functions, can provide a visualization of a trained model's output that helps domain experts to validate the predictions.

Conclusion

Our convolutional motif kernel network architecture provides inherently interpretable end-to-end learning on biological sequence data and achieves state-of-the-art performance on relevant healthcare prediction tasks, namely predicting antiretroviral drug resistance of HIV isolates and distinguishing decoys from real splice sites.

We show that CMKN is able to learn biologically meaningful motif and position patterns on synthetic and real-world datasets. CMKN's global interpretation can foster data mining and knowledge advancement on biological sequence data. On the other hand, CMKN's local interpretation can be utilized by domain experts to judge the validity of a prediction.

Possible future improvements include investigating a combination of different motif kernel layers to combine different motif lengths and extend the architecture to utilize meaningful combinations of motifs. Another improvement that we want to explore in future work is the extension of the kernel formulation to multi-layer networks while securing the interpretation capabilities.

Data availability

Source code, pre-processing scripts, and experimental scripts are available on GitHub at https://github.com/jditz/CMKN. HIV drug resistance data can be found online on https://hivdb.stanford.edu/pages/genopheno.dataset.html. The NN269 benchmark can be downloaded from the George Manson University (https://cs.gmu.edu/~ashehu/sites/default/files/tools/EFFECT_2013/data.html). The DGSplicer benchmark was provided by the original author, Prof. Chung-Chin Lu (https://icannwiki.org/Chung-Chin_Lu).

Received: 16 May 2023; Accepted: 4 October 2023
Published online: 11 October 2023

References

1. Degroeve, S., De Baets, B., Van de Peer, Y. & Rouzé, P. Feature subset selection for splice site prediction. *Bioinformatics* **18**, S75–S83 (2002).
2. Zien, A. *et al.* Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* **16**, 799–807 (2000).
3. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
4. Yang, W. & Deng, L. Predba: A heterogeneous ensemble approach for predicting protein-dna binding affinity. *Sci. Rep.* **10**, 1–11 (2020).
5. Döring, M. *et al.* geno2pheno [ngs-freq]: A genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Res.* **46**, W271–W277 (2018).
6. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell rna-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 1–14 (2019).
7. Kadir, T. & Brady, M. Saliency, scale and image description. *Int. J. Comput. Vis.* **45**, 83–105 (2001).
8. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems* 4768–4777 (2017).
9. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
10. Sixt, L., Granz, M. & Landgraf, T. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning* 9046–9057 (PMLR, 2020).
11. Bordt, S., Finck, M., Raidl, E. & von Luxburg, U. Post-hoc explanations fail to achieve their purpose in adversarial contexts. Preprint at http://arxiv.org/abs/2201.10295 (2022).
12. Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **16**, 31–57 (2018).
13. Cho, Y. & Saul, L. Kernel methods for deep learning. *Adv. Neural Inf. Process. Syst.* **22**, 342–350 (2009).
14. Bo, L., Lai, K., Ren, X. & Fox, D. Object recognition with hierarchical kernel descriptors. In *CVPR 2011* 1729–1736 (IEEE, 2011).
15. Mairal, J., Koniusz, P., Harchaoui, Z. & Schmid, C. Convolutional kernel networks. In *Advances in Neural Information Processing Systems* 2627–2635 (2014).
16. Mairal, J. End-to-end kernel learning with supervised convolutional kernel networks. In *Advances in Neural Information Processing Systems* 1399–1407 (2016).
17. Chen, D., Jacob, L. & Mairal, J. Biological sequence modeling with convolutional kernel networks. *Bioinformatics* **35**, 3294–3302 (2019).
18. Eskin, E., Weston, J., Noble, W. S. & Leslie, C. S. Mismatch string kernels for svm protein classification. In *Advances in Neural Information Processing Systems* 1441–1448 (2003).
19. Leslie, C. & Kuang, R. Fast string kernels using inexact matching for protein sequences. *J. Mach. Learn. Res.* **5**, 1435–1455 (2004).

20. Chen, D., Jacob, L. & Mairal, J. Recurrent kernel networks. In *Advances in Neural Information Processing Systems* 13431–13442 (2019).
21. Meinicke, P., Tech, M., Morgenstern, B. & Merkl, R. Oligo kernels for datamining on biological sequences: A case study on prokaryotic translation initiation sites. *BMC Bioinform.* **5**, 169 (2004).
22. Mialon, G., Chen, D., d'Aspremont, A. & Mairal, J. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021* (OpenReview.net, 2021).
23. Williams, C. & Seeger, M. Using the nyström method to speed up kernel machines. In *Proc. 14th Annual Conference on Neural Information Processing Systems, CONF* 682–688 (2001).
24. Zhang, K., Tsang, I. W. & Kwok, J. T. Improved nyström low-rank approximation and error analysis. In *Proc. 25th International Conference on Machine Learning* 1232–1239 (2008).
25. Rätsch, G., Sonnenburg, S. & Schölkopf, B. Rase: Recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics* **21**, 369–377. https://doi.org/10.1093/bioinformatics/bti1053 (2005).
26. Rhee, S.-Y. *et al.* Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31**, 298–303 (2003).
27. Shafer, R. W. Rationale and uses of a public hiv drug-resistance database. *J. Infect. Dis.* **194**, S51–S58 (2006).
28. Matthews, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **405**, 442–451 (1975).
29. Raposo, L. M., Rosa, P. T. C. & Nobre, F. F. Random forest algorithm for prediction of hiv drug resistance. In *Pattern Recognition Techniques Applied to Biomedical Problems* 109–127 (Springer, 2020).
30. Sonnenburg, S., Schweikert, G., Philips, P., Behr, J. & Rätsch, G. Accurate splice site prediction using support vector machines. *BMC Bioinform.* **8**, 1–16 (2007).
31. Baten, A. K., Chang, B. C., Halgamuge, S. K. & Li, J. Splice site identification using probabilistic parameters and svm classification. *BMC Bioinform.* **7**, 1–15 (2006).
32. Zuallaert, J. *et al.* Splicerover: Interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **34**, 4180–4188 (2018).
33. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in genie. *J. Comput. Biol.* **4**, 311–323 (1997).
34. Chen, T.-M., Lu, C.-C. & Li, W.-H. Prediction of splice sites with dependency graphs and their expanded bayesian networks. *Bioinformatics* **21**, 471–482 (2005).
35. Chicco, D. T. quick tips for machine learning in computational biology. *BioData Mining* **10**, 1–17 (2017).
36. Rhee, S.-Y. *et al.* Hiv-1 protease mutations and protease inhibitor cross-resistance. *Antimicrob. Agents Chemother.* **54**, 4253–4261 (2010).
37. Colonno, R. *et al.* Identification of i50l as the signature atazanavir (atv)-resistance mutation in treatment-naive hiv-1-infected patients receiving atv-containing regimens. *J. Infect. Dis.* **189**, 1802–1810 (2004).
38. Goudsmit, J., De Ronde, A., Ho, D. D. & Perelson, A. S. Human immunodeficiency virus fitness in vivo: Calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. *J. Virol.* **70**, 5662–5664 (1996).
39. Bethell, R. *et al.* Phenotypic protease inhibitor resistance and cross-resistance in the clinic from 2006 to 2008 and mutational prevalences in hiv from patients with discordant tipranavir and darunavir susceptibility phenotypes. *AIDS Res. Hum. Retroviruses* **28**, 1019–1024 (2012).
40. Larder, B. A., Kemp, S. D. & Harrigan, P. R. Potential mechanism for sustained antiretroviral efficacy of azt-3tc combination therapy. *Science* **269**, 696–699 (1995).
41. Schapiro, J. M. *et al.* The effect of high-dose saquinavir on viral load and cd4+ t-cell counts in hiv-infected patients. *Ann. Intern. Med.* **124**, 1039–1050 (1996).
42. Craig, C. *et al.* Hiv protease genotype and viral sensitivity to hiv protease inhibitors following saquinavir therapy. *AIDS* **12**, 1611–1618 (1998).
43. Zolopa, A. R. *et al.* Hiv-1 genotypic resistance patterns predict response to saquinavir–ritonavir therapy in patients in whom previous protease inhibitor therapy had failed. *Ann. Intern. Med.* **131**, 813–821 (1999).
44. Kempf, D. J. *et al.* Identification of genotypic changes in human immunodeficiency virus protease that correlate with reduced susceptibility to the protease inhibitor lopinavir among viral isolates from protease inhibitor-experienced patients. *J. Virol.* **75**, 7462–7469 (2001).
45. Van Marck, H. *et al.* The impact of individual human immunodeficiency virus type 1 protease mutations on drug susceptibility is highly influenced by complex interactions with the background protease sequence. *J. Virol.* **83**, 9512–9520 (2009).

Acknowledgements

The authors would like to thank Prof. Chung-Chin Lu for providing the DGSplicer benchmark. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project Number 390727645. This research was supported by the German Federal Ministry of Education and Research (BMBF) project 'Training Center Machine Learning, Tübingen' with Grant Number 01|S17054. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. This work was support by the Open Access Publication Fund of the University of Tübingen.

Author contributions

J.C.D., B.R., and N.P. developed the method. B.R. worked out the mathematical formulation of the method based on initial work by J.C.D. and N.P.; J.C.D. implemented the method. J.C.D. conducted and analysed the experiment(s) supported by B.R. and N.P.; J.C.D. wrote the manuscript supported by B.R. and N.P.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-44175-7.

Correspondence and requests for materials should be addressed to J.C.D. or N.P.