Published online 19 June 2020

Nucleic Acids Research, 2020, Vol. 48, No. 13 ***7099–7118***
doi: 10.1093/nar/gkaa530

DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning

Alexander Gulliver Bjørnholt Grønning[1,2,3,†], Thomas Koed Doktor[1,2,*,†], Simon Jonas Larsen[3], Ulrika Simone Spangsberg Petersen[1,2], Lise Lolle Holm[1,2], Gitte Hoffmann Bruun[1,2], Michael Birkerod Hansen[1,2], Anne-Mette Hartung[1,2], Jan Baumbach[3,4] and Brage Storstein Andresen[1,2]

[1]Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense M, Denmark, [2]Villum Center for Bioanalytical Sciences, University of Southern Denmark, 5230 Odense M, Denmark, [3]Department of Mathematics and Computer Science, University of Southern Denmark, 5230 Odense M, Denmark and [4]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany

Received October 23, 2019; Revised May 11, 2020; Editorial Decision June 02, 2020; Accepted June 10, 2020

ABSTRACT

Nucleotide variants can cause functional changes by altering protein–RNA binding in various ways that are not easy to predict. This can affect processes such as splicing, nuclear shuttling, and stability of the transcript. Therefore, correct modeling of protein–RNA binding is critical when predicting the effects of sequence variations. Many RNA-binding proteins recognize a diverse set of motifs and binding is typically also dependent on the genomic context, making this task particularly challenging. Here, we present DeepCLIP, the first method for context-aware modeling and predicting protein binding to RNA nucleic acids using exclusively sequence data as input. We show that DeepCLIP outperforms existing methods for modeling RNA-protein binding. Importantly, we demonstrate that DeepCLIP predictions correlate with the functional outcomes of nucleotide variants in independent wet lab experiments. Furthermore, we show how DeepCLIP binding profiles can be used in the design of therapeutically relevant antisense oligonucleotides, and to uncover possible position-dependent regulation in a tissue-specific manner. DeepCLIP is freely available as a stand-alone application and as a webtool at http://deepclip.compbio.sdu.dk.

INTRODUCTION

The massive technological progress in next generation sequencing (NGS) technologies has made sequencing affordable in the context of precision medicine and personalized health care. NGS analysis enables identification of millions of sequence variants in each patient sample, increasing the need for *in silico* prediction of the functional consequences of a diverse range of variations. In particular, the effect of deep intronic sequence variants at the mRNA level through altered binding to RNA-binding proteins (RBPs) is difficult to predict *in silico* as existing tools' predictions of functional outcomes of splicing are primarily based on the analysis of point mutations within or near exons (1–3). While some existing binding site prediction tools can work on sequences of any type, there is an unmet need for improved modeling of contextual dependencies other than structure that are important for correctly estimating the *in vivo* functionality of the binding sites. Extracted contextual information may form the basis for design of antisense oligonucleotide based therapies, which modulate RBP activity, such as splice-switching oligonucleotides (SSOs) (4–6). Thus, improving information on whether contexts act positively or negatively with regard to binding is an important area of research that will ultimately enable the development of novel therapeutic options in personalized medicine.

Sequencing technologies have also vastly expanded the wealth of information concerning protein binding to RNA when combined with cross-linking and immunoprecipitation (CLIP) techniques (7–9), which allow accurate mapping of protein binding sites in functional *in vivo* contexts. Classically, binding preferences or binding motifs have been represented by position frequency matrices (PFMs). Well-known *de novo* motif discovery tools such as MEME (10) and HOMER (11) output PFMs and base their motif detection and identification on the PFM concept. This approach to motif discovery implicitly assumes that such fixed-length

*To whom correspondence should be addressed. Tel: +45 6550 2416; Email: thomaskd@bmb.sdu.dk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

motifs exist and that they function in a context-independent manner regarding the surrounding sequences. They further assume pairwise independence of the nucleotide frequencies within the motifs.

However, proteins that bind RNA typically do so in a context dependent manner. In particular, secondary structure may influence the binding of some RBPs (12). Information about double-stranded or single-stranded structure has been incorporated into MEMERIS (13), which is an extension of the MEME algorithm. Further structural dependencies have been incorporated into RNAcontext (12), which expands the information about secondary structure from simple double or single-stranded structures into paired, hairpin loops, bulges and internal or multiloops, and unstructured contexts in order to further optimize the modeling of binding preference of RBPs. More recently, a graph-based modeling of structural and sequence binding preferences was introduced in the GraphProt (14) software, which out-performed RNAcontext on a set of diverse CLIP datasets using different CLIP methods. GraphProt uses RNAshapes (15) to predict the structures of RNA-sequences, which are then encoded into a hypergraph from which important structural features can be extracted. To improve the structure estimations, GraphProt extends the CLIP-derived sequences by 150 bp in each direction. Together with sequence features extracted only from the CLIP-derived binding sites, an overall model of binding preference is generated using support vector machines.

While inclusion of structural preferences may increase accuracy in prediction, these models still fail to capture other contextual dependencies affecting the *in vivo* functionality, such as a high density of protein binding sites nearby or localization within a specific functional region of the transcript, such as proximity to splice sites. For instance, exonic splicing enhancers (ESEs) that enhance splicing of exons by binding to SR proteins are enriched in exons, while exonic splicing silencers (ESSs) are underrepresented in exons. These observations have been used to generate ESE and ESS motifs from sequences enriched (16,17) or depleted (16) in exons. Such contextual dependencies were recently introduced in the iONMF software (18), which uses integrative orthogonality-regularized nonnegative matrix factorization to incorporate multimodal information about CLIP-derived binding sites such as their position within the gene (5′UTR, CDS, exon, intron, 3′UTR), gene ontology, and presence of other protein binding sites determined by other CLIP studies, in addition to structural information, which improved performance for some datasets.

In recent years, deep learning techniques have been used to model protein binding. Deep learning has proven successful in various difficult classification tasks such as natural language processing (19), object recognition (20) and reconstructing brain circuits (21). Deep learning allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction (22). Deep learning models can identify dependencies and complex structures in very high-dimensional data—such as CLIP data - and have been used, for example, for predicting the effects of mutations in non-coding DNA on gene expression and disease (1,23), predicting DNA function (24), mRNA coding potential (25) and prediction of subcellular locations of proteins (26). Starting with DeepBind (27), which is trained on *in vitro* RNAcompete data, convolutional neural networks (CNN) have been used to estimate binding affinity, and later methods such as DLPRB (28) have expanded on this to also use secondary structure as well as recurrent neural networks trained on *in vitro* RNAcompete data.

Using *in vivo* CLIP training data, Deepnet, a multimodal deep belief network incorporating 2D structure information (mDBN-) or both secondary and tertiary structure information (mDBN+) along with a CNN architecture was introduced in the deepnet-rbp software (29), while more recently the iDeep framework combines the annotation data used by iONMF with a CNN into a multimodal neural network with increased accuracy in classification compared to iONMF(30). Even more recently, iDeepS was introduced as a replacement for iDeep to include analysis of 2D structural motifs much like GraphProt, using a combination of CNN and bidirectional LSTM (Long Short-Term Memory) layers akin to DeepCLIP's architecture (31).

Previous models for RBP binding properties that consider contextual clues are focused either specifically on structural dependencies, which may fail to capture other important contextual dependencies, or on the presence of annotation data to aid in the task of classification. However, static annotations will not contribute to determining the effect of a mutation on the binding activity of proteins. For instance, a model, which relies heavily on clues from annotation data about the genomic region, such as location within an exon, will be unable to use this level of information to ascertain the effect of an exonic point mutation in which the context is maintained. Only iDeepS has a general-purpose LSTM layer able to model general context dependency, but it is supplemented with structural predictions from an external program and thus does not work on sequence data alone.

Understanding binding preferences is important for evaluation of the phenotypic impact of sequence variations. Mutations may alter the phenotype at several different levels, as in the case of missense mutations, which in addition to altering the amino-acid sequence, may also change the splicing pattern (32). Other mutations with less *visible* deleterious effects may abolish healthy splicing by altering the binding of RBPs, sometimes at somewhat distant sites. Splicing and the overall binding activity of RBPs is the result of a balance between positively and negatively acting elements that cooperate or compete for binding (33,34), so even minor changes in RBP binding sites can change the outcome of splicing events.

Importantly, before they can be applied in predicting clinically important changes or functional elements to be targeted, binding models needs to be validated in the laboratory, using *in vitro* techniques such as RNA-protein affinity measurements and *in vivo* techniques such as minigene transfections and predictions need to be consistent with effects reported in clinically affected patients.

In this paper, we present DeepCLIP, a novel deep learning based tool for discovering protein–RNA binding sites and for characterizing binding preferences of RBPs. We demonstrate how it outperforms current state-of-the-art RBP binding analysis tools, and we show that DeepCLIP's

predictions provide information about high-affinity RBP binding sites and that it successfully predicts alterations of the RBP affinity for RNA-sequences when single nucleotide polymorphisms (SNP) or disease-causing mutations are introduced. This is reflected both in the binding profiles that show the region(s) important for RBP binding, and in the predictions of the sequences which indicate whether they are more similar to the 'consensus' CLIP-sequence or to the genomic background. Last but not least, we have made DeepCLIP available as an online tool for training and application of protein–RNA binding deep learning models and prediction of the potential effects of clinically detected sequence variations (http://deepclip.compbio.sdu.dk/). We also provide DeepCLIP as a configurable stand-alone program (http://www.github.com/deepclip).

MATERIALS AND METHODS

DeepCLIP: more than just a motif discoverer

DeepCLIP is essentially a deep neural network that uses shallow 1D convolutional layers to find and enhance features of a set of presented sequences (26,35,36). This is followed by a Bidirectional Long Short Term Memory (BLSTM) layer (37,38) which uses the extracted features and contextual information of the sequences to find areas of the RNA-sequences associated with RBP binding (Figure 1). Initially, the convolutional layers of DeepCLIP can be regarded as a collection of randomly generated PFMs of user-defined sizes that, as training progresses, learn to recognize important nucleotide patterns in the input data. When predicting, the convolutional layers score sequence segments according to their importance for the classification task. Pseudo-PFMs can be generated by collecting scored sequence patterns and counting the frequencies of different nucleotides at each possible position. We use the term pseudo-PFMs because each sequence used for the PFM generation is weighted by the squared output score given by the convolutional layers. In this way, the pseudo-PFMs will depict important class-specific nucleotide patterns. The BLSTM layer of DeepCLIP is used to generate a binding profile at the nucleotide level. The BLSTM layer consist of two LSTM layers that analyze 'hidden' sequence representations (modified outputs of the convolutional layers) in a bidirectional manner (Supplementary Figure S1).

The DeepCLIP tool takes a single or more RNA oligonucleotides (short RNA sequences) as input and predicts binding probability and calculates a binding profile. The main purpose of DeepCLIP is to identify binding sites of proteins in novel untested sequences using trained models that have extracted binding site information provided by CLIP data, to predict the effect of sequence variants on the binding, and to identify the importance of individual nucleotides for protein binding affinity. DeepCLIP can be run on a standard PC with Linux installed as operating system. Training of smaller datasets require at least 4GB RAM, but larger datasets will require more memory. The largest dataset used for training in this study consisted of a total of 1 256 372 input sequences, with a memory footprint of ∼43GB RAM. DeepCLIP is fast enough to run online on a web server (http://deepclip.compbio.sdu.dk/) and its Python code is also publicly available (http://www.github.com/deepclip).

Training workflow of DeepCLIP

The core of DeepCLIP is a convolutional BLSTM network implemented in Theano (39) using the Lasagne library (40) and a few customized network layers and functions. DeepCLIP is a binary classifier that uses supervised learning to distinguish between unbound sequences and bound sequences derived from CLIP-experiments. The input to DeepCLIP consists of positive (bound) sequences, which are assigned to class 1, and negative (unbound) sequences assigned to class 0. These input sequences are converted into linearized one-hot encoded vectors, which serve as the actual input to the neural network layers. The DeepCLIP architecture is shown in Figure 1.

By default, 80% of the input data is used for training while 10% is used for validation and the last 10% for testing (Supplementary Figure S1d). DeepCLIP is trained by iterating over the training and validation sets several times (also called epochs). While training, performance is measured on the validation set after every epoch and the best performing model is saved. Early stopping can be applied to prevent training after a likely maximum performance has been obtained. The final performance of the saved model is measured on the test set, which contain data that have not previously been introduced to the model. When running in 10-fold cross-validation the input sequences are divided into 10 equal sized bins, and each bin is used once as a test set, once as a validation set and 8 times as part of the training set (Supplementary Figure S1e).

Encoding of sequence data

DeepCLIP processes sequence data as linearized one-hot encoded vectors. In the one-hot representation, the items of the vocabulary, $v = (A, C, G, U)$, are represented by vectors with lengths equal to the length of the vocabulary that each have a 1 in unique dimensions. The one-hot encoded bases are therefore independent of one another and are equally similar or dis-similar. It signals that no prior correlations between the bases are known. In this way, the network will determine correlations between the bases on its own (41). All input sequences are zero-padded until they have identical lengths and until the largest filter can conduct 'full convolutions' as it is defined in the Lasagne documentation (40). Following vectorization of the bases of the sequences, the combined vectors are linearized, and the resulting one-dimensional data used as input to the neural network.

Convolutional neural network layer

Convolutional layers consist of nodes that are only sensitive to a defined receptive field referred to as kernels or filters. Nodes of convolutional layers apply weight-sharing and sparse-connectivity, which means that the same filter can 'view' all possible filter-sized segments of the input individually (42).

As in previous work (26,27,29,30), the filters of the convolutional layers in DeepCLIP can be interpreted as motif detectors. The sizes of the filters of the convolutional layers are optional but we used ranges from four to eight one-hot encoded bases. DeepCLIP only applies a single filter of each

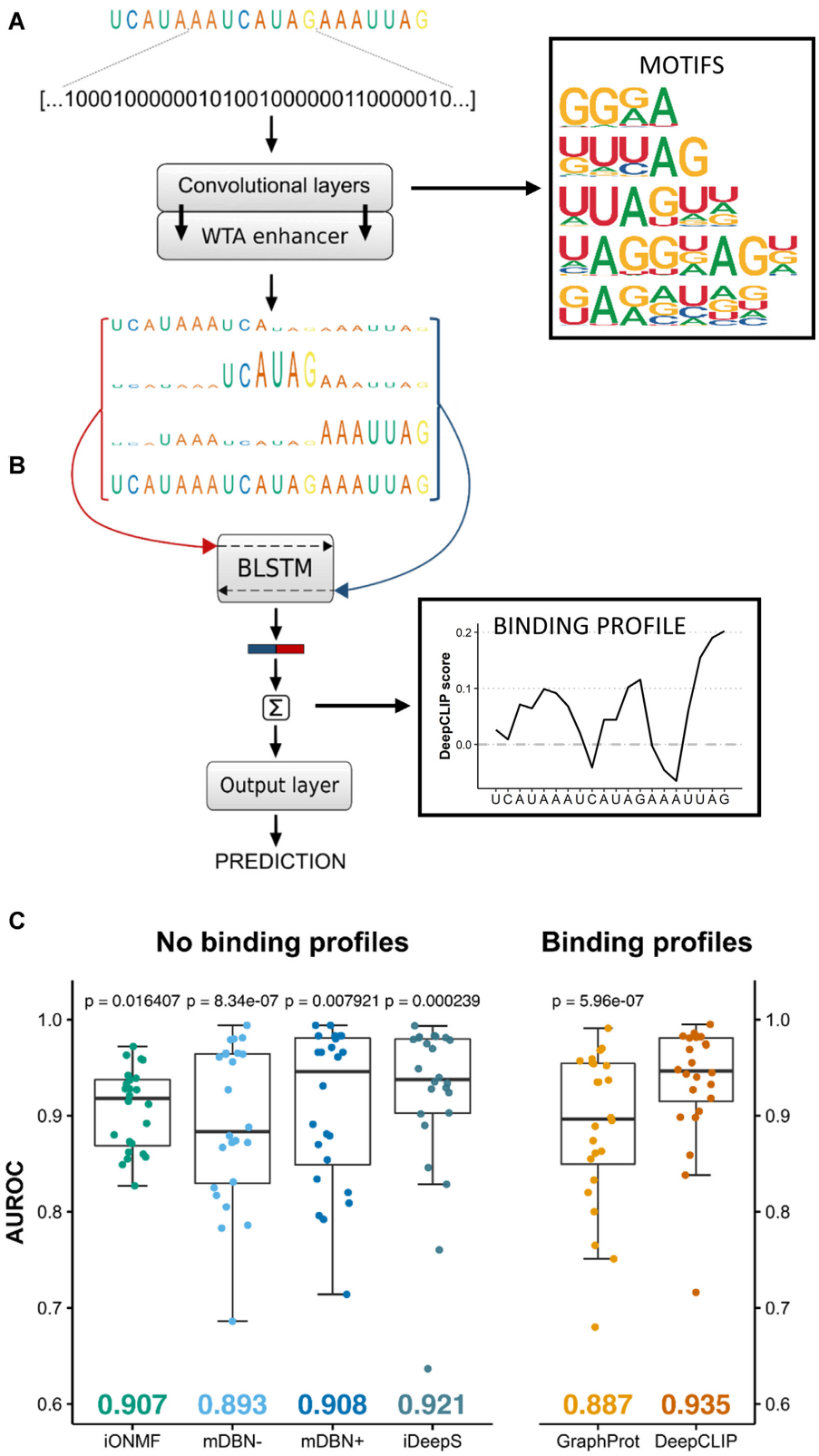

Figure 1. Classification performance of DeepCLIP surpasses competing methods. (**A**) One-hot embedded RNA-sequences function as input to the neural network and the 1D convolutional layers, which are further enhanced by a winner-takes-all (WTA) layer. (**B**) The outputs are concatenated with the input sequence and each segment of the concatenated arrays that corresponds to aligned bases is introduced to the following BLSTM layer as individual time-steps. The WTA-enhanced RNA-sequences are the three uppermost and the input sequence is in the bottom. DeepCLIP produces a prediction score, which is used during training, as well as binding motifs and a binding profile. (**C**) Boxplot of comparative analyses of DeepCLIP classification performance against other state-of-the-art tools. Area under receiver operator characteristic curve (AUROC) were measured in 10-fold cross-validation and statistical significance was computed through an Wilcoxon signed-rank test. *P*-values above individual tools are from pair-wise comparisons with DeepCLIP. Mean AUROC score is indicated in the bottom.

size. The strides of the filters are $|v|$, so the filters only convolve patterns consisting of whole one-hot encoded bases. The convolutional layers apply the rectifier activation function (43). In this context, it means that only patterns that receive a score above zero can be assumed important for sequence identification.

The bias parameters of the convolutional layers are removed to ensure that only areas in the sequence input containing one-hot encoded bases produce outputs above 0, preventing recognition of zero-padded areas alone. As the nodes in the convolutional layers are rectifying linear units, only sequence-segments that are deemed important by the neural network will produce convolutional-outputs above 0. The initial weights of the convolutional nodes are set to 0.01. The filter sizes allow for diverse sequence patterns of various length to be incorporated into the model. The output vectors of convolutional layers in DeepCLIP are vectors containing values between 0 and ∞. Before the output vectors are passed to the BLSTM layer they undergo a so-called WTA-enhancement (Winner Take All-enhancement), which is described below.

The single highest values of the different output vectors (44) are multiplied by 2 which is followed by a squaring of the vectors to enhance differences between high and low values. These squared vectors have now been WTA-enhanced, where the 'winners' are the highest values in the output vectors. The WTA-enhanced vectors are used for a recreation of the original one-hot embedded sequences where the one-hot values are defined by the WTA-enhanced vector elements. For each convolutional layer, a WTA-enhanced sequence is created and concatenated in a manner that makes it possible to process each numerical base representation of the sequences as individual steps in the following BLSTM layer (see Figure 1).

In this way, the convolutional layers help guide the attention of the BLSTM layer. The pseudo-PFMs created by the DeepCLIP tool that depict the important patterns in the RBP-bound sequences, are based on the specific sequential areas that only relate to class 1. Meaning, if an area of a sequence is, by the BLSTM layer, predicted as being associated with class 0, any convolutional output values in the sequential area will be zeroed out and therefore will not be a part of the pseudo-PFM calculation. The filters of the convolutional layers of the model with the best performance in the 10-fold cross validation were extracted and used for creation of pseudo-PFMs. The top 1000 sequences with respect to the predictions were used for the pseudo-PFM calculation.

Bidirectional LSTM layer

Long short-term memory (LSTM) networks have already proven successful in biological sequence analysis (26,45). DeepCLIP uses a single BLSTM layer that processes the WTA-enhanced sequences (Figure 1). In BLSTM layers, the input sequences are presented forwards and backwards in two separate LSTM layers that are connected to the same output layer (38). The implementation of a single LSTM layer in DeepCLIP is given by Equations (1-9):

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \quad (1)$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \quad (2)$$

$$\boldsymbol{g}_t = \tanh\left(\boldsymbol{W}_{xg}\boldsymbol{x}_t + \boldsymbol{W}_{hg}\boldsymbol{h}_{t-1} + \boldsymbol{b}_g\right) \quad (3)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t \quad (4)$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \quad (5)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh\left(\boldsymbol{c}_t\right) \quad (6)$$

$$\sigma\left(z\right) = \frac{1}{1+e^{-z}} \quad (7)$$

$$\tanh\left(z\right) = \frac{e^{2z}-1}{e^{2z}+1} \quad (8)$$

$$\text{Hadamard product} = \odot \quad (9)$$

where $\boldsymbol{i}$, $\boldsymbol{f}$, $\boldsymbol{g}$, $\boldsymbol{o}$ and $\boldsymbol{c}$ are the input gate, forget gate, modulatory gate, output gate and cell, respectively. $\boldsymbol{x}_t$ is the input vector at timestep t, $\boldsymbol{W}_{xi}$ is the input-input gate weight matrix, $\boldsymbol{W}_{hi}$ is the hidden-input gate weight matrix, $\boldsymbol{b}_i$ is the bias of the input gate and $\boldsymbol{h}_{t-1}$ is the hidden output vector from timestep $t-1$. The same logic applies for the remaining gates. σ is the sigmoid activation function, *tanh* is the hyperbolic tangent activation function and the Hadamard product indicates elementwise multiplication. The hidden output vector of a LSTM memory block is $\boldsymbol{h}_t$ at timestep t.

The forward LSTM reads an input sequence with length T from $\boldsymbol{x}_1$ to $\boldsymbol{x}_T$ and the backward LSTM reads the same input sequence from $\boldsymbol{x}_T$ to $\boldsymbol{x}_1$. The forward LSTM layer produces forward hidden vectors, $\overrightarrow{\boldsymbol{h}_1}, \overrightarrow{\boldsymbol{h}_2} \ldots \overrightarrow{\boldsymbol{h}_T}$ and the backward LSTM layer produces backward hidden vectors $\overleftarrow{\boldsymbol{h}_1}, \overleftarrow{\boldsymbol{h}_2} \ldots \overleftarrow{\boldsymbol{h}_T}$. Here, the output of the backwards LSTM layer has been reversed, so the outputs of forward and backward LSTM layers go from $\boldsymbol{x}_1$ to $\boldsymbol{x}_T$. The hidden vector of the BLSTM layer at time step t, $\boldsymbol{h}_t$, is given by the concatenation of the forward hidden vector and the backward hidden vector $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}]$ (46).

The output sequence of a BLSTM layer given an input sequences of length T can be seen as a matrix, $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots \boldsymbol{h}_T)$, where each $\boldsymbol{h}_t$ is a row. Each $\boldsymbol{h}_t$ contains information about the whole input sequence with a strong focus on the parts surrounding the t^{th} input vector (46). In DeepCLIP, each $\boldsymbol{h}_t$ represents a base of a sequence that has knowledge of the surrounding sequence. These context-aware representations of sequences function as input to the output layer where the final prediction is calculated. Dropout is applied on $\boldsymbol{H}$, which means that recurrent connections are not affected.

Output layer, binding profile and prediction

The output layer consists of a single fixed node without a bias parameter that uses the sigmoid activation function. By 'fixed' we mean that the parameters of the node do not update when training. The initial weight of the node is set to 1.0, which means that the node is forced to associate positive values with class 1 and negative values with class 0. The

input to the output layer, given a single input sequence, is the vector that results from a row summation of the matrix $\boldsymbol{H}$ where segments based on zero-paddings are zeroed out. By zeroing out these segments it is ensured that only areas that contain one-hot embedded bases are used for the prediction of the given sequence. Basically, the prediction of a given RNA-sequence is the sum of the output values of the BLSTM layer inserted into a sigmoid activation function.

In terms of the BLSTM output, if a $\boldsymbol{h}_t$ is mainly positive the base at its specific position is associated with the 'consensus patterns' of the sequences derived from the CLIP experiments. If a $\boldsymbol{h}_t$ is primarily negative, the base at its specific position is more associated with random background sequences derived from the genome. And if the values of a $\boldsymbol{h}_t$ sums to ∼0, the base at its position could belong to both classes. This approach makes DeepCLIP able to highlight sequential areas that differ from genomic background and thereby able to identify *in vivo* binding sites. The binding profiles are constructed using the input vectors of the output layer, where all the zero-padding has been removed.

DeepCLIP default settings

DeepCLIP uses ADAM (47) ($\alpha = 0.0002,\ \beta_1 = 0.9,\ \beta_2 = 0.999, \epsilon = 10^{-8}$) for gradient descent optimization. For the BLSTM layer, the parameters are sampled from a Gaussian distribution with $\mu = 0.0$ and $std = 0.01$. DeepCLIP uses binary cross entropy as loss function and employs dropout in order to avoid overfitting. The most optimal network weights based on Area Under Receiver Operator Curve (AUROC) performance on the validation set are saved during training. Dropout is applied to the BLSTM layer (10%).

Generation of background sequences

Background sequences can be either supplied by the user as sequences or generated automatically by DeepCLIP from positive binding sites in one of two ways. The default way is to supply positive binding site in BED format and then generate a random set of identically sized genomic regions as the positive binding site and randomly placing each within the same gene as the corresponding positive binding site such that no background regions overlap either a positive binding site or another background region. Alternatively, background sequences can be generated from positive binding sites by scrambling the input sequences. When positive binding sites are supplied in BED format, they can optionally be expanded on each size, or fixed to a certain width, or a combination of these. In this study, we have used the random genomic background method to most accurately obtain *in vivo* non-bound sites.

Analysis of area under receiver operator curve classification performance

Comparison of area under receiver operator curve (AUROC) classification performance was performed on a dataset compiled in a previous study paper (14). In order to minimize computational complexity, we took the performance numbers of alternative models on this dataset as they were reported in previous studies (14,18,29). We could not compare to iDeep (30), as they did not provide numbers for the same datasets, and did not provide any way of producing the multimodal data required as input. We ran iDeepS on 101 nt sequences from the GraphProt dataset by expanding the peak-areas on either side until the sequence was 101 nt. We ran iDeepS on these sequences with the same number of epochs that we used to train DeepCLIP models. We generated 10-fold cross-validation sets where one set was held out for testing one time, and used in training the other 9 times, in order to obtain comparable performance measures across the full datasets. We ran DeepCLIP on the peak-area sequences of the datasets in a 10-fold cross-validation, such that each site was held out exactly once for validation during training, and once for final testing, while being used for actual training the remaining eight times. Model performance was measured for each dataset using the performance measure tool 'perf' as used by GraphProt on the combined predictions from all CV cycles. Additionally, AUROC confidence intervals were estimated using the DeLong algorithm as implemented in the 'pROC' R package (48).

Benchmark analysis on eCLIP data

Binding sites for hg38 from ENCODE (49) were downloaded and replicate experiments were combined into the overlapping regions of both replicates to obtain high-confidence binding sites from which we extracted sites with lengths 12–75 nt plus 150 nt padding to enable GraphProt compatibility. A balanced negative background set was constructed using DeepCLIP's implementation, namely shifting each positive binding site to a random site within the same gene that does not overlap positive binding sites or other previously allocated negative binding sites. Similarly, we downloaded eCLIP binding sites from the POSTAR2 database (50), merged all sites to remove any duplicates and extracted sites in the range 12–75 nt with additional 150 nt padding, which we used to construct a balanced background set. Sequences were trimmed to obtain an equal length of 101 nt for use with iDeepS, and padding removed to obtain sequences of 12–75 nt to use with DeepCLIP models. After creating these balanced datasets we used either the published models (GraphProt) or models we had trained ourselves (iDeepS and DeepCLIP) to measure classification performance using the AUROC metric.

Additional models from public CLIP data

Binding sites from an eCLIP study of SRSF1 (49) in K562 cells were downloaded and the overlap between two replicates were extracted to create a set of non-redundant binding sites. These were used for constructing the bound dataset, with a matched genomic background as control. We then trained an SRSF1 DeepCLIP model on this dataset using same running parameters as previous models, with 50 training epochs and early stopping after five epochs. TDP-43 binding sites were downloaded from POSTAR2 (50) and non-redundant input sites were used to train a DeepCLIP model, again using identical running parameters as previous models, but adjusting the number of training epochs to 50 and early stopping after five epochs to account for the

much larger training set. Similarly, hnRNP A1 binding sites (51), were used to train a DeepCLIP model on binding sites with *P*-values below 0.01 using default parameters with 200 training epochs and early stopping after 20 epochs. In all cases, 10-fold cross-validation was used to identify the best performing model.

Additional RNAcompete based models

We trained binary classification models from RNAcompete data (52) by sorting the sequences by the normalized intensities and setting the 1000 highest scoring as the positive class, and the 1000 lowest scoring as the negative class. Sequences were zero-padded to allow models trained on the shorter RNAcompete sequences to predict class of the 12–75 nt GraphProt benchmark datasets. We then used the models to obtain AUROC estimates for the corresponding GraphProt dataset.

Analysis of NM_032776.1 transcript with PUM2 and QKI models

DeepCLIP was run in long-prediction mode with the PUM2 and QKI models trained on the GP dataset to produce binding profiles across the length of the transcript (8762 nt total). A sliding window of 9 nt was then used to identify regions with a minimum mean profile score >0.3. Overlapping 9 nt windows were combined to produce a non-redundant set of predicted binding sites. Mapping locations relative to the transcript of predicted as well as observed binding sites from eCLIP (49) and PAR-CLIP (8) were visualized with Gviz (53).

Analysis of TDP-43 repressed pseudoexons

Pseudoexons activated by conditional knock-out of TDP-43 in mice (54) were analyzed with DeepCLIP by first extracting the sequence of the pseudoexon along with 100 nt of the neighboring intronic sequences. These sequences were then used to produce binding profiles by using a sliding window approach to produce raw DeepCLIP profiles of smaller segments, taking the value of the central nucleotide to build a binding profile covering the entire length of the sequence. Subsequently, regions corresponding to the 25 first and last nucleotides of the exons along with the 50 first and last nucleotides of the neighboring introns were extract in order to analyze TDP-43 binding to the acceptor and donor splice site regions.

Minigene generation

ACADM exon 5 minigenes were identical to the previously used wt *ACADM* minigene (34), with the exception of nucleotide variants at positions corresponding to c.361, c.362 and c.363 with exon 5. These variants were introduced as previously described (34).

The *ACADM* exon 6 wt minigene was generated from genomic DNA by amplifying the complete exon 6 (81 bp) along with 864 bp of intron 5 and 603 bp of intron 6 and subsequent cloning into the pSPL3 vector (Gibco BRL) using the BamHI and XhoI restriction sites. For amplification we used the forward primer 5′-TCGAGAATTCAGGAGCA-3′ and the reverse primer 5′-CTCCACTAAATAGAGC-3′. The IVS6+7A>G mutation was introduced by GenScript (GenScript, Piscataway, NJ, USA).

ACADM exon 5 minigene transfections and RT-PCR

HEK-293 cells were seeded in 3.5 cm^2 12-well plates (Nunc) at a density of 4×10^5 cells/well 24 h prior to transfection. In each well, cells were transiently transfected using X-tremeGENE 9 DNA Transfection Reagent (Merck): 0.3 μg of one of the *ACADM* exon 5 minigenes c.362C (wildtype), c.361C, c.361G, c.361T, c.362A, c.362G, c.362T, c.363A, c.363C or c.363G. After 48 h of incubation following minigene transfection, cells were harvested using QIAzol Lysis Reagent (Qiagen), followed by phenol/chloroform extraction of total RNA. Reverse transcription was performed using the High Capacity cDNA Reverse Transcription Kit (Thermo Scientific). Splicing patterns were analyzed by PCR amplification, using TEMPase Hot Start DNA Polymerase (Ampliqon), and agarose gel electrophoresis. We used the *ACADM* exon 5 minigene specific primers: MCTEST2AS (5′-AGACTCGAGTTACTATTAATTACACATC-3′) and MC242S (5′-CCTGGAACTTGGTTTAATG-3′). PCR products were quantified according to molar ratios by capillary gel electrophoresis on a Fragment Analyzer™ instrument (Agilent), and visualized on 1.5% agarose gels. Experiments were performed in triplicates.

ACADM exon 6 minigene transfections with siRNA mediated knock-down of TDP-43 and RT-PCR

Knockdown of TDP-43 was obtained by performing reverse transfection during initial seeding of cells and another transfection 48 h later. Both transfections were performed using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) and 40 nM of siRNA targeting *TARDBP* (L-012394-00-0020, Dharmacon) or non-targeting siRNA (D-001810–10-20, Dharmacon). HeLa cells were seeded in 3.5 cm^2 12-well plates (Nunc) at a density of 1.5×10^5 cells/well 24 h prior to minigene transfection. In each well, cells were transiently transfected using X-tremeGENE 9 DNA Transfection Reagent (Merck): 0.4 μg of one the two *ACADM* exon 6 minigenes: WT or +7A>G. After 48 h of incubation following minigene transfection, cells were harvested using QIAzol Lysis Reagent (Qiagen), followed by phenol/chloroform extraction of total RNA. Reverse transcription was performed using the High Capacity cDNA Reverse Transcription Kit (Thermo Scientific). Splicing patterns were analyzed by PCR amplification, using TEMPase Hot Start DNA Polymerase (Ampliqon), and agarose gel electrophoresis. We used the minigene specific primers: SD6 (5′-TCTGAGTCACCTGGACAACC-3′) and SA2 (5′-ATCTCAGTGGTATTTGTGAGC-3′). PCR products were quantified according to molar ratios by capillary gel electrophoresis on a Fragment Analyzer™ instrument (Agilent), and visualized on 1.5% agarose gels. Knockdown of TDP-43 was validated by SDS-PAGE and

Western Blotting and membranes were probed with antibodies anti-TDP-43 (10782-2-AP, ProteinTech) and as a loading control anti-HPRT (HPA006360, Merck). Experiments were performed in triplicates.

SSO co-transfection with *ACADM* exon 6 minigenes

HeLa cells were reverse transfected in duplicates with 40 nM SSO using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) according to the manufacturer's protocol, and seeded in 3.5 cm^2 12-well plates (Nunc) at a density of 2×10^5 cells/well 24 h prior to minigene transfection. SSOs were phosphorothioate oligonucleotides with 2′-*O*-methyl modifications on each sugar moiety (LGC Biosearch Technologies): SSO1 (5′-UAAGUGUGAAAUAAAGCGGCAGUUA-3′), SSO2 (5′-AGUGUGAAAUAAAGCGGCAGUUACA-3′), or a control SSO without any human target sites: 5′-GCUCAAUAUGCUACUGCCAUGCUUG-3′. Cells were transiently transfected using X-tremeGENE 9 DNA Transfection Reagent (Merck): 0.4 μg of one of the two *ACADM* exon 6 minigenes: WT, or +7A>G. After 24 h of incubation following minigene transfection, cells were harvested using QIAzol Lysis Reagent (Qiagen), followed by phenol/chloroform extraction of total RNA. Reverse transcription was performed using the High Capacity cDNA Reverse Transcription Kit (Thermo Scientific). Splicing patterns were analyzed by PCR amplification, using TEMPase Hot Start DNA Polymerase (Ampliqon), and agarose gel electrophoresis. We used the minigene specific primers: SD6 (5′-TCTGAGTCACCTGGACAACC-3′) and SA2 (5′-ATCTCAGTGGTATTTGTGAGC-3′). PCR products were quantified according to molar ratios by capillary gel electrophoresis on a Fragment Analyzer™ instrument (Agilent), and visualized on 1.5% agarose gels. Experiments were performed in triplicates.

Surface plasmon resonance imaging method

Biotinylated oligonucleotides were immobilized on a Senseye G strep (SSENS) sensorchip in a $2 \times 4 \times 12$ array by continuous flow in a CFM 2.0 printer (Wasatch microfluidics). The oligonucleotides were diluted in 1XTBS to a concentration of 1 μM and spotted for 20 min followed by 5 min washing with TBS + 0.05% Tween-20. The sensor chip was transferred to the MX-96 (IBIS technologies), and the system was primed with SPR buffer (10 mM HEPES/KOH pH 7.9, 150 mM KCl, 10 mM $MgCl_2$ and 0.075% Tween-80). Surface plasmon resonance imaging (SPRi) by IBIS MX-96 was used to measure the kinetics of recombinant hnRNP A1 (ab224866, Abcam), SRSF1 (GenScript, Piscataway, NJ, USA) and TDP-43 (R&Dsystems, AP-190) binding to the immobilized RNA oligonucleotides. Binding was measured in real time by following changes of the SPR angles at all printed positions of the array during 10 min. injections of recombinant protein over the entire surface. Seven injections of a 2-fold titration series from 6.25 to 400 nM protein was injected in sequence from the lowest concentration to the highest. Before adding protein to the chip, residual background binding was blocked by injecting 20mg/ml BSA in SPR buffer onto the chip for 10 min. A continuous flow of SPR buffer flowed over the surface before, between and after the protein injections, to measure baseline and dissociation kinetics. Dissociation was measured for 8 min, by injecting SPR buffer over the chip at a rate of 4 μl/s. Responses for a calibration curve were created after the concentration series by measuring SPR responses from defined dilutions of glycerol in running buffer (ranging from 5 to 0% glycerol) and of pure water as defined by the automated calibration routine of IBIS MX-96.

Data analysis: The SPRi data was imported into SPRINTX software (v. 2.1.1.0, IBIS technologies), calibrated, reference subtracted, and the baseline of the responses before all injections were zeroed. The time starting point was aligned at the beginning of each new injection. Then the data were exported to Scrubber 2 (v 2.0c, Biologics Inc.). Binding curves for all chip positions where binding was observed were fitted globally to the integrated rate equation that describes simple first order 1:1 binding kinetics to obtain kinetic association rate (k_a), dissociation rate (k_d) as well as the R_{max} for the binding model. For hnRNPA1 and TDP-43 a 1:2 biphasic model was calculated and fitted. ClampXP (version 3.50, Biosensor Data Analysis) was used with a bimodal model to fit the binding data. The secondary K_a and K_d parameters were fixed to 1e−5 M, due to very low secondary association and dissociation. Primary binding parameters and ligand concentration (R_{max}) were set to float. SPRi measurements were performed twice, with technical duplicates being used for model fitting each time.

Statistical analyses

All statistical analyses were performed in R (version 3.5.3). We used the default wilcox.test() for two-tailed Wilcoxon rank sum and Wilcoxon signed rank tests. Linear regression was carried out using ggplot2 (version 3.2.1) and geom_smooth(method = lm, colour = 'red', se = TRUE), with correlation significance analysis using the default cor.test() with method = 'spearman'.

RESULTS

DeepCLIP outperforms structural and multimodal models from sequence data alone

DeepCLIP is a neural network that combines shallow convolutional layers with a small bidirectional long short-term memory network to produce both a binding profile and a classification score ranging from 0 to 1 (Figure 1, Supplementary Figure S1a–c). Models are created by training a network on a set of known binding sites and a set of background genomic sequences (Supplementary Figure S1d, e), which can optionally be generated by DeepCLIP by providing binding locations instead of raw binding sequences.

To ascertain DeepCLIP's classification performance on a standardized dataset, we generated models from the curated CLIP datasets (8,9,55–63) used in the GraphProt publication (14), which has previously been used in other studies (18,29). First, we trained DeepCLIP models in a 10-fold cross-validation scheme using 50–500 epochs depending on the size of the individual dataset with early stopping after 10% of the maximum number of epochs (Supple-

mentary Table S1). Next, we measured area under receiver operator characteristic curve (AUROC) using the standard method of 10-fold cross-validation and the combined performance across the 10 different sets (Supplementary Figure S1e, Table S2). Importantly, DeepCLIP does not directly model structure. Consequently, we used the peak area alone, which is akin to the viewpoint mechanism as adopted in GraphProt. We compared the performance of our models with the performance numbers reported in the earlier studies describing GraphProt, iONMF, and deepnet-rbp (mDBN– and mDBN+). To obtain AUROC values for iDeepS, we performed a 10-fold cross-validation using the curated CLIP-datasets, as this was omitted in the iDeepS paper, by extending the peak area to 101 nt per the input requirement for iDeepS. We found that DeepCLIP was the overall best classifier in every pair-wise comparison and when looking at the mean AUROC score, underscoring that DeepCLIP performs well on a broad set of data. Furthermore, DeepCLIP had more narrow distributions of scores with fewer low-scoring datasets and a majority of datasets scoring above 0.9 (Figure 1C). DeepCLIP consistently ranked among the best classifiers on individual datasets (Supplementary Figure S2a, Table S3), even without additional contextual data, i.e. working on the sequence input data alone. In total, DeepCLIP had the best performance for 14 of the 24 datasets, and second best for 6 datasets. For TAF15, mDBN+ and DeepCLIP had identical scores when rounding to the third decimal. DeepCLIP was also a better classifier than DeepBind ((pseudo)median 15.16%-points (CI-95%: 7.76–24.17), Wilcoxon signed-rank $P = 0.0004883$), on the 12 datasets for which DeepBind models are available (Supplementary Figure S2b,c).

We chose the best model based on AUROC measure on the validation set, but a more conventional choice is to use the validation loss metric to select the best model. In a *post-hoc* analysis, we also implemented loss-based model selection and found that AUROC performance on the held out test set improved, but only to a small degree (Supplementary Figure S3).

We did not observe any significant differences in DeepCLIP AUROC scores between the CLIP methods (Supplementary Figure S4a), a significant correlation to the number of bound sites (Supplementary Figure S4b, $P = 0.647$), or the mean input length of the training data (Supplementary Figure S4c, $P = 0.118$), but a tendency towards improved performance on the nucleotide-resolution CLIP datasets (iCLIP and PAR-CLIP) versus HITS-CLIP did appear. We did, however, observe a significant negative correlation between GC-content and AUROC performance (Supplementary Figure S4d, $P = 7.89 \times 10^{-6}$), which seems to be primarily driven by a positive correlation with U-content (Supplementary Figure S4e, $P = 0.000356$) and a negative correlation with G-content (Supplementary Figure S4h, $P = 1.04 \times 10^{-5}$). Since many of the proteins in the dataset bind U-rich motifs, and U is known to cross-link more efficiently than other bases, this indicates a general CLIP bias in the models, where the model recognizes not a specific binding site, but a CLIP site in general. In particular, PAR-CLIP is known to display U-bias due to protocol specific treatment of the samples.

We therefore benchmarked the models against completely independent eCLIP datasets from ENCODE (49) and the POSTAR2 database (50). We were not able to run iONMF or mDBN models for these datasets, but compared to GraphProt and iDeepS, DeepCLIP performed favorably on both with both types of model selection (Supplementary Figure S5a, b, Table S4). In general, there were many models that performed poorly with AUROC values close to or <0.5. We therefore also filtered out models scoring <0.6 in all datasets, to focus on just the models with reasonable performance by at least one method. All methods were more similar in this analysis, with no method having a clear advantage over the others. Notably, in other studies using *in vitro* binding data to train models, a better performance was demonstrated on the GraphProt benchmark dataset. We therefore trained DeepCLIP models on corresponding RNAcompete datasets by setting the 1000 highest scoring RNAcompete sequences as the positive class, and the 1000 lowest scoring sequences as the negative class. We then compared the performance of these models on the GraphProt dataset to the performance of RCK (64), RNAcontext (12), DeepBind (27), DLPRB-CNN and DLPRB-LSTM (28) (Supplementary Figure S5c, Table S5). We found that DeepCLIP had the overall highest mean performance with a more narrow distribution of scores, indicating excellent performance of DeepCLIP trained on *in vitro* data compared to existing state-of-the-art methods.

DeepCLIP performed well on all CLIP datasets regardless of size and CLIP method, with the exception of ALKBH5, which is a problematic dataset for all methods that do not rely on additional metadata, presumably due to non-specific binding that may take place in cooperation with a number of other factors that target the factor to specific regions within the transcript. DeepCLIP is thus a robust classifier of *in vivo* binding sites using only sequence data. It compares favorably to models employing external structural information and annotation data in addition to sequence data.

DeepCLIP models predict binding motifs

Although the classifications of DeepCLIP are based entirely on the values of the binding profiles, motifs can be assessed from the CNN filters incorporated in the network architecture. The motif of each filter was generated using the patterns from the 1,000 input sequences that produced the highest DeepCLIP classification score. We found that DeepCLIP produces motifs that are visually similar to previously published motifs (65–70), illustrating that DeepCLIP's classification performance is not simply a result of learning how to recognize the background sequences, but depends on the binding preferences of the RBP in question (Supplementary Figure S6). Additional model performance metrics and CNN filter motifs are available (Supplementary Table S2, Figure S7–S30). Alternatively, filters can also be produced from sequences scoring above a certain score, such as 0.5 (Supplementary Figure S31).

DeepCLIP predictions and binding profiles explain splicing mutations

Splicing of mRNA is regulated by binding of RBPs to the nascent pre-mRNA. To test DeepCLIP's ability to predict effects of nucleotide variants on splicing, we generated

new models for the splicing factors hnRNP A1 and SRSF1 based on previous CLIP studies (49,51) with DeepCLIP-generated background sequences in order to demonstrate DeepCLIP's performance on novel datasets. We ran 10-fold cross-validation (Supplementary Figures S32 and S33) using the same input parameters as previously, and extracted the two best performing models. Their pseudo-PFMs are shown in Figure 2A and B and they show high visual agreement with previously published motifs of hnRNP A1 and SRSF1 (60,69,71,72).

We used the models to predict binding of hnRNP A1 and SRSF1, respectively. We then used the best performing model to predict binding of hnRNP A1 to a set of exonic point mutations (2), grouped into mutations known to cause skipping and mutations known to not cause skipping (Figure 2C, D, Supplementary Table S6). In Figure 2C, we show the distributions of predictions given the 15-mer sequences used by Raponi *et al.* in their work describing splicing (2). This specific sequence length represents the length of a typical RNA oligonucleotide used in affinity purification experiments to measure the binding of protein to RNA. Interestingly, we observe no significant difference in the overall change of hnRNP A1 scores for skipping and non-skipping mutations (Figure 2C, left, Wilcoxon signed rank $P = 0.29$), suggesting that hnRNP A1 is not a general regulator of these splicing events. This is to be expected, since only a subset of these events is likely to be meditated by altered binding of hnRNP A1.

When scoring the same 15 nt oligonucleotides with the best performing SRSF1 model, we observe that the SRSF1 scores are significantly different between the two groups (Figure 2C, middle, Wilcoxon signed rank $P = 0.026$). Also, when combining the hnRNP A1 scores with the SRSF1 scores, we find that the groups were significantly different (Figure 2C, right, Wilcoxon signed rank $P = 0.033$). Importantly, the scores of the mutations known to cause skipping were decreased, consistent with the known role of SRSF1 as a positive regulator of exon inclusion. When we used all trained models to analyze the dataset and adjusting *P*-values for multiple testing, none of the models showed a significant difference in score change between skipping and non-skipping exonic mutations. However, the two SRSF1 models had the lowest adjusted *P*-values (Supplementary Figure S34a).

To investigate whether the hnRNP A1 and SRSF1 models improve with extended sequence context, we expanded the 15-mer sequences from the middle and out to a length of 75 nt, the maximum sequence length used during model training. This resulted in less pronounced changes that were not statistically significant (Figure 2D, Supplementary Table S6), although the combined score indicated that the combined effects of losing SRSF1 binding and gaining hnRNP A1 binding was retained to a higher degree. This is likely caused by DeepCLIP's classification being based on the total binding profile resulting in diminished differences as the sequence is expanded. This is exemplified by the *ATP7A* c.3904G>A exon skipping mutation located at +103 in exon 20 of *ATP7A,* which results in an overall score change of +0.00166 between the wt (0.00056) and mutant (0.00222) 75 nt long sequences (Figure 2E), but a score change of +0.28484 (from 0.34651 to 0.63135) when the 15 nt long sequence is used (Figure 2F). Importantly, both binding profiles predict a localized increase in hnRNP A1 binding to the mutant, showcasing the relevance of using binding profiles when analyzing sequence data and not simply an overall prediction score.

DeepCLIP hnRNP A1 and SRSF1 prediction scores correlate with exon inclusion levels of a known SRSF1-dependent exon

We have previously characterized splicing of *ACADM* exon 5, which shares sequence similarity with *SMN1* exon 7 and we identified a similar regulation with splicing ultimately relying on the balance of SRSF1 and hnRNP A1 binding (34). A prevalent disease-causing c.362C>T mutation reduces the strength of a SRSF1 binding ESE, allows hnRNPA1 binding and causes exon 5 skipping. To test DeepCLIP models of hnRNP A1 and SRSF1 in relation to splicing of *ACADM* exon 5, we generated minigenes with all possible variants at positions c.361, c.362 and c.363 located down-stream of the CAG core motif (Figure 3A). DeepCLIP scores were obtained from the sequences using a window of 36 nt on each side of the three positions, totaling 75 nt (Supplementary Table S7). We then transfected the minigenes in HEK293 cells and measured exon inclusion levels (PSI) using RT-PCR and gel-electrophoresis (Figure 3B). We observe a strong negative correlation (Spearman's $\rho = -0.939$, $P < 2e{-}16$, Figure 3C) between the hnRNP A1 prediction score and the observed inclusion of *ACADM* exon 5, and a strong correlation between the SRSF1 prediction score and the observed inclusions (Spearman's $\rho = 0.770$, $P = 0.0137$, Figure 3D). We did not observe a significant correlation with the SRSF1 model trained on the GraphProt benchmark dataset (Supplementary Figure S35a), which illustrates that the training data impacts the quality of the models. The observed correlation is also very strong for the combined scores (Spearman's $\rho = 0.915$, $P = 0.000467$, Figure 3E), in agreement with the hypothesis that the overall inclusion level is a result of the balance between positive and negative factors. The same is true when we use the SRSF1 model generated from the GraphProt dataset (Supplementary Figure S35b), but not when we perform the same analysis with EX-SKIP (2) on the full exon (Spearman's $\rho = -0.177$, $P = 0.625$, Supplementary Figure S36a). When we use SPANR (1) we do observe a stronger positive correlation than with EX-SKIP (Spearman's $\rho = 0.552$, $P = 0.104$, Supplementary Figure S36b), but all variants are predicted to have an inclusion level between 81.9% and 82.8%. This conflicts directly with the observed level of exon skipping induced by the disease-causing c.362C>T mutation in patient cells (34) as well as with the observed splicing pattern of the minigenes tested.

Using GraphProt and iDeepS models trained on the SRSF1 eCLIP and hnRNP A1 iCLIP datasets, we could obtain similar but less pronounced correlations, with GraphProt (Supplementary Figure S36c) performing better than iDeepS (Supplementary Figure S36d).

Overall, when the hnRNP A1 DeepCLIP model predicts an increase in hnRNP A1 binding, there was a decrease in exon 5 inclusion (Figure 3F). In particular, the high degree of exon skipping of all c.362 variants relative to wt were re-

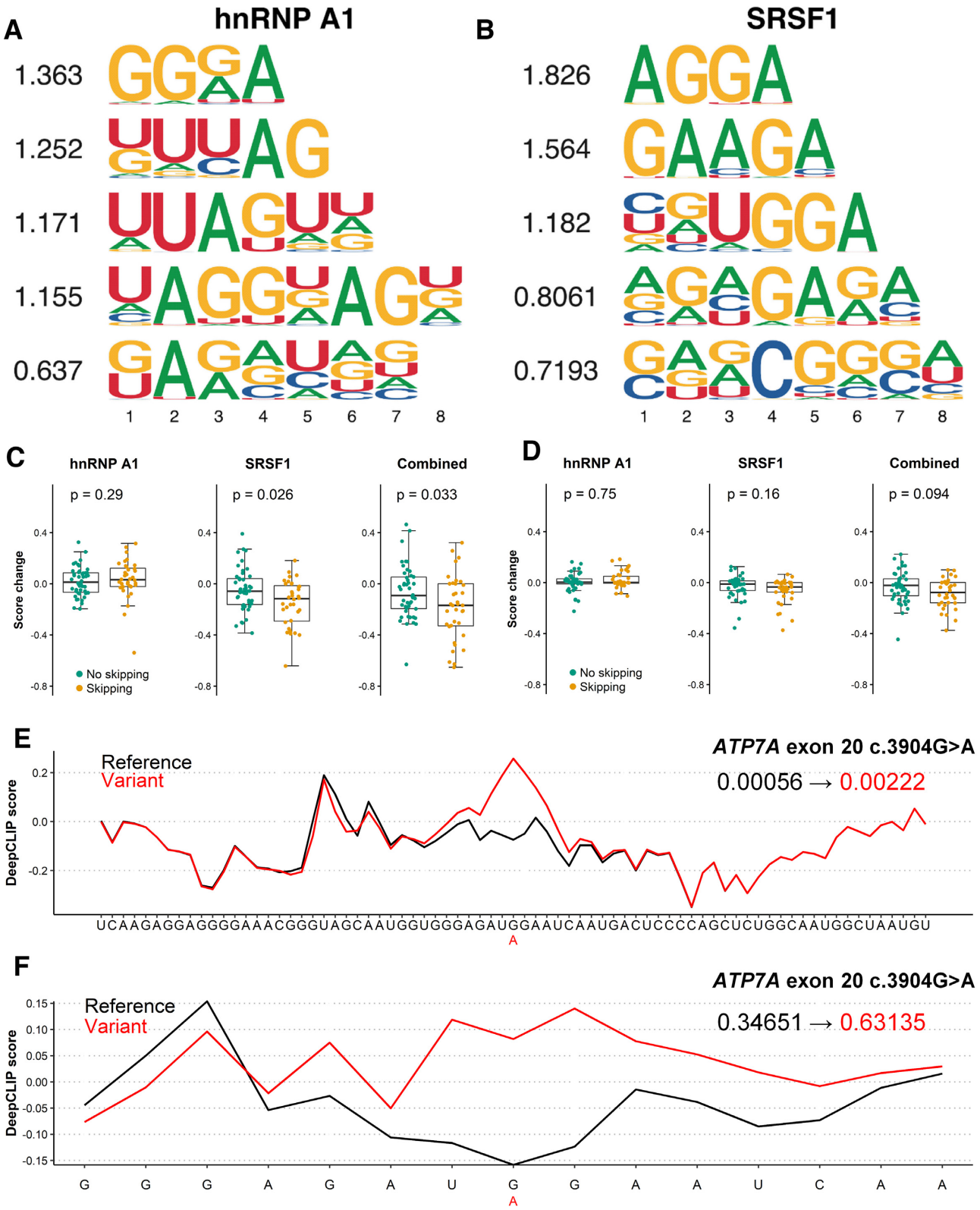

Figure 2. DeepCLIP models of hnRNP A1 and SRSF1 used to analyze splicing mutations. (**A**) The CNN filters trained in the hnRNP A1 DeepCLIP model. (**B**) The CNN filters trained in the SRSF1 model. (**C**) Box-plots of the distributions of predictions on 15 nt sequences representing wt and mutant versions exons that are skipped upon mutation (yellow, $n = 37$ sequence-pairs) and exons that remain included in the mutant version (green, $n = 46$ sequence-pairs) using of the hnRNP A1 model (left), SRSF1 model (middle) and combined (right). The combined change in DeepCLIP scores is obtained by subtracting the hnRNP A1 scores from the SRSF1 scores. Two-tailed Wilcoxon rank sum test *P*-value is indicated above. Box-plot elements are defined as center line: median, box limits: upper and lower quartiles, whiskers: 1.5× interquartile range. All data points are shown, outliers are not highlighted. (**D**) Same as (C) but with 75 nt sequences. (**E**) DeepCLIP binding profile of wt (black) and mutant (red) of the 75 nt sequence representing the *ATP7A* exon 20 +103G>A mutation. The overall DeepCLIP prediction scores are indicated in bold within the plot. (**F**) Same as (E), but with 15 nt input sequence.

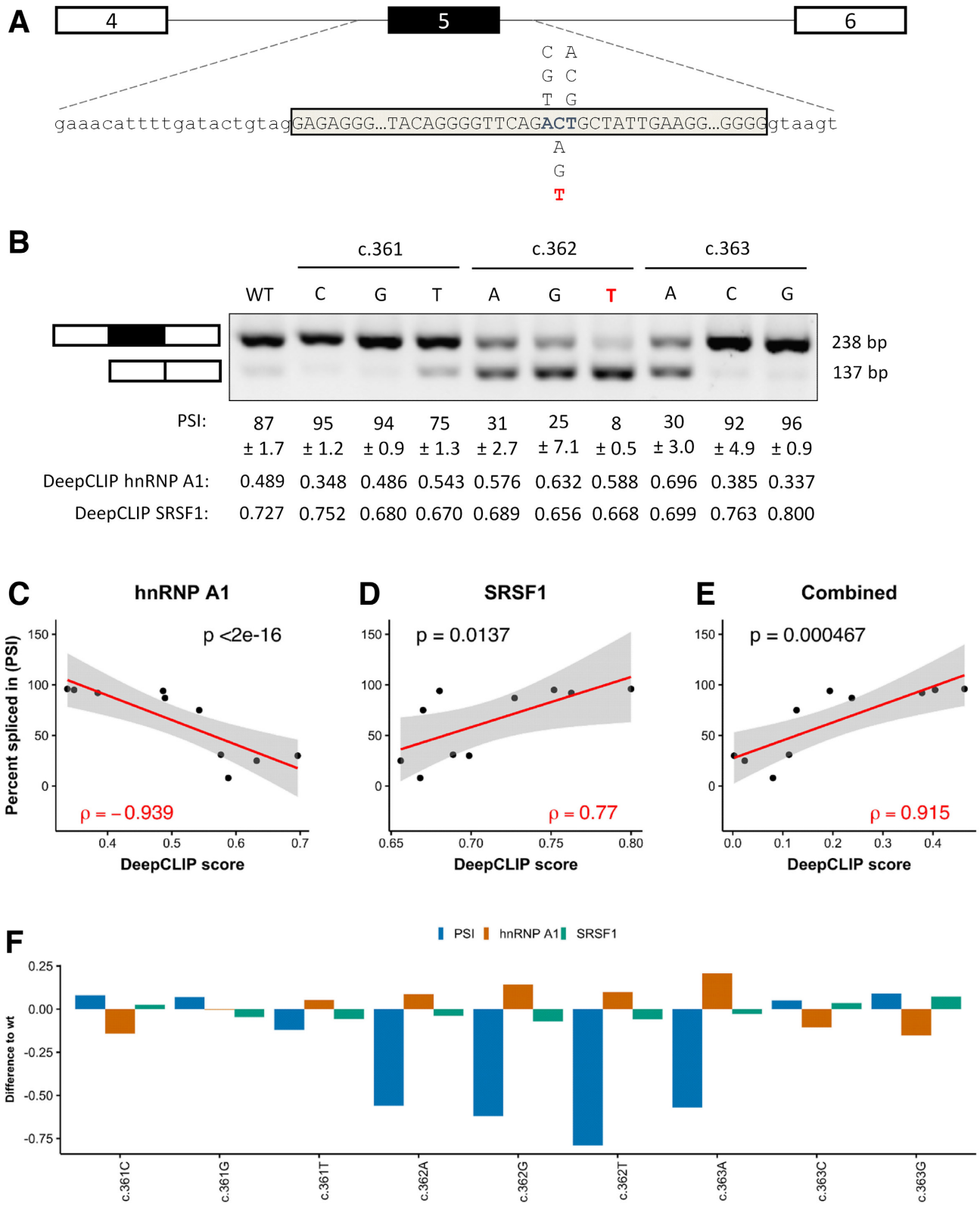

Figure 3. DeepCLIP models successfully model *ACADM* minigene splicing results. (**A**) Minigene schematic and location of variants tested, reference in blue. The disease-causing mutation is indicated in red. (**B**) Splicing of minigenes determined by RT-PCR. Estimates of mean PSI ($n = 3$) is indicated below, along with 95% CI size. (**C**) Scatter plot of PSI and DeepCLIP hnRNP A1 score with linear regression (red line, $n = 10$) and 95% confidence interval (shaded area). (**D**) Same as (C) but with DeepCLIP SRSF1 score instead. (**E**) Same as (C) and (D), but showing the DeepCLIP SRSF1 score minus the DeepCLIP hnRNP A1 score. (**F**) Barplot showing the difference to wt for the minigene PSI and DeepCLIP prediction scores for hnRNP A1 and SRSF1. Spearman's correlation coefficient is indicated in (C), (D) and (E).

flected by increases in hnRNP A1 scores (Figure 3B and F). The c.363A variant is predicted to abolish binding of SRSF1 and increase binding of hnRNP A1 and the minigene analysis demonstrates predominant skipping in agreement with this. Like hnRNP A1, many of the variants predicted to lose SRSF1 binding show increased skipping consistent with a loss of ESE activity.

The data indicate that while single DeepCLIP models capture binding preferences of individual proteins, the scores are additive and can be used to model effects of multiple proteins interacting in antagonistic and synergistic ways. Because splicing is complex, multiple other factors may also contribute to the outcome of splicing, which may explain why there is not a complete correlation between scores and observed inclusion levels.

DeepCLIP binding profiles can guide the design of therapeutic antisense oligonucleotides

DeepCLIP models produce binding profiles, which are directly used for prediction calculations. We wanted to test how the binding profiles reflect *in vivo* sequence-protein binding dynamics and see if the profiles can help locating sites where splice-switching oligonucleotides (SSOs) can be applied for correction of splicing. Thus, we analyzed a known disease-causing mutation (73) in the *ACADM* gene, c.468+7A>G. The mutation is located outside the core U1 snRNP binding motif but is located within a larger GT-rich region, which is extended by the A>G mutation suggesting that it could be generating or strengthening a TDP-43 binding site. We selected wt and mutant sequences 75 nt of length by including the first 37 nt from either side of the position of the A>G mutation. We then generated a TDP-43 DeepCLIP model based on publicly available binding sites from the POSTAR2 database (50) using the same model parameters as previously (Supplementary Figure S37), and scored the wt and mutant sequences. We found that DeepCLIP predicts increased binding of TDP-43 to the mutant relative to the wt (Figure 4A).

Next, we designed a minigene harboring *ACADM* exon 6 and part of the flanking introns to test whether the c.468+7A>G mutation affects splicing of exon 6. We found that the mutation caused dramatic skipping of exon 6 from the minigene (Figure 4b). We hypothesized that this was caused by an increase of TDP-43 binding to the mutant sequence, and that exon skipping therefore could be reversed by treating the cells containing the minigenes with siRNA targeting TDP-43 mRNA. Indeed, TDP-43 siRNA treatment resulted in increased exon inclusion (Figure 4B), corroborating that the c.468+7A>G mutation generates a TDP-43 binding site that causes skipping of *ACADM* exon 6.

Splice-switching oligonucleotides (SSOs) are a type of antisense oligonucleotide (ASO) that can be used to modulate splicing by sterically preventing binding of splicing regulatory factors to the RNA. Because of the close proximity of the mutation to the 5′ splice site, directly blocking the mutant position with an SSO most likely would not result in increased exon inclusion. Interestingly, DeepCLIP finds sites important for binding in a region downstream of the GT-rich core binding motif, which suggests that blocking these sites could prevent TDP-43 binding to the core motif and restore splicing of exon 6. We tested this hypothesis using two different SSO molecules that targeted this downstream region and which had small overlaps with the end of the GT-rich region (Figure 4D). Strikingly, both SSOs proved very efficacious and almost completely restored splicing from the mutant minigene, indicating that blocking of the downstream motif prevented binding of TDP-43. This indicates that TDP-43 may exhibit context dependent binding modularity, and that the DeepCLIP model is able to detect these context-dependent signatures from the sequence alone.

To validate that TDP-43 binding is directly affected by the mutation, we first analyzed a set of 23 nt oligonucleotides with DeepCLIP (Figure 4e,f) showing that in this shorter context the mutation is still predicted to increase. We then used Surface Plasmon Resonance imaging (SPRi) to measure binding to the 3′ biotin labeled RNA oligonucleotides and observed a pronounced increase in TDP-43 binding to the mutant (Figure 4G, H) in agreement with DeepCLIP predictions.

DeepCLIP binding scores correlate with *in vitro* binding affinities

One of the most important tasks of a model that predicts presence of RBP binding sites is to accurately estimate the effects of mutations on binding affinity. Therefore, we analyzed 6 sets of wt and mutant exonic variants from the Raponi et al 15-mer set (2) employing SPRi using recombinant hnRNP A1 and SRSF1 as input-proteins (Supplementary Figures S38–S43, Table S8). These measurements allow quantification of the binding to wt and mutant oligonucleotides, allowing confirmation of DeepCLIP predictions, such as the increase in hnRNP A1 binding to the *ATP7A* exon 20 c.3904G>A mutant (Figure 5A, B). The maximum affinity values obtained by fitting binding models to the measured response by the different SPRi-models correlated well with both hnRNP A1 and SRSF1 DeepCLIP models (Figure 5C, D), across the diverse set of sequences in the dataset (hnRNP A1: Spearman correlation $\rho = 0.874$, $P = 0.000309$; SRSF1: Spearman correlation $\rho = 0.77$, $P = 0.0137$). This was also true when we compared DeepCLIP predictions with the Rmax value (Supplementary Figure S44). This demonstrates that despite being trained on *in vivo* data, the modeling approach of DeepCLIP is also applicable with short *in vitro* sequences, which can be used to examine and validate specific changes in binding to target sites identified by DeepCLIP.

DeepCLIP analysis of TDP-43-repressed pseudoexons indicates that tissue-specificity is position-dependent

In addition to analyzing sequence variations, DeepCLIP can also be used on a global scale to conduct larger analyses of binding preferences of RBPs. TDP-43 is depleted in the nucleus of motor neurons in patients suffering from amyotrophic lateral sclerosis (ALS) (74,75). TDP-43 has been reported to repress the inclusion of pseudoexons, and these are then erroneously activated following nuclear depletion, potentially leading to development of ALS symptoms (76). A conditional TDP-43 knock-out mouse-model

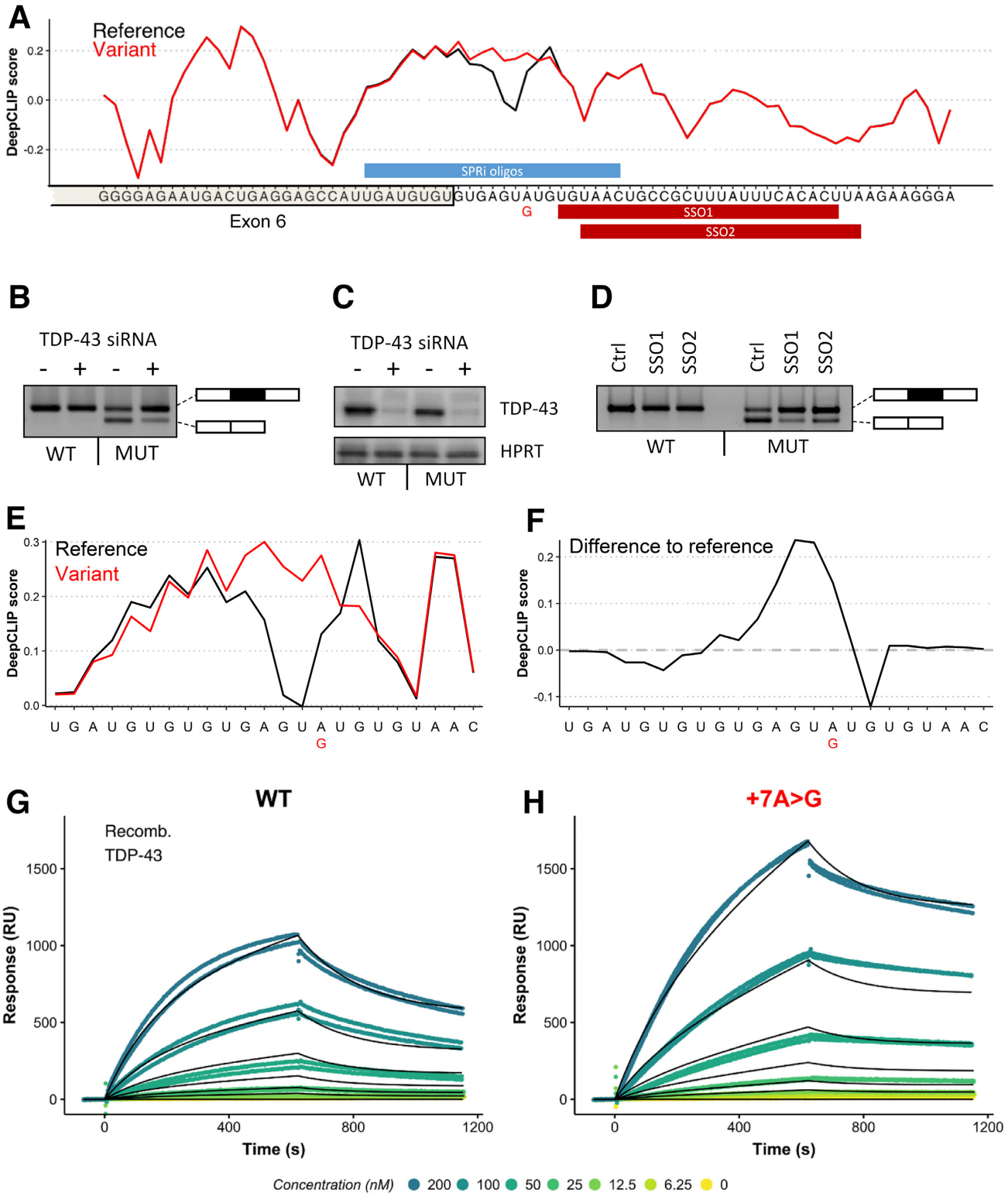

Figure 4. DeepCLIP predicts increased TDP-43 binding as mechanism behind *ACADM* exon 6 skipping. (**A**) DeepCLIP TDP-43 profile across the 5′ss of *ACADM* exon 6 with wt indicated in black and patient mutation indicated in red. Along the first axis the sequence is shown and along the second axis the DeepCLIP BLSTM values are shown. SPRi oligo location and SSO locations are indicated in blue and red bars above and below the sequence, respectively. (**B**) Splicing of wt and mutant minigenes with either TDP-43 targeting siRNA or non-targeting siRNA determined by RT-PCR. (**C**) Western blot of TDP-43 and HPRT from siRNA and minigene transfected samples. (**D**) Splicing of wt and mutant minigenes treated with either a control SSO (Ctrl-SSO), SSO1, or SSO2 determined by RT-PCR. (**E**) DeepCLIP profile of short RNA oligos used in SPRi measurement, reference in black and +7A>G variant in red. (**F**) The difference in DeepCLIP binding profiles in (E) between reference and variant. Positive score indicates higher score in variant. (**G**) SPRi measurements of TDP-43 binding to the wt oligo in (E). (**H**) SPRi measurements of TDP-43 binding to the variant oligo in (E). In both (G) and (H), the black line indicates the fitted binding model.

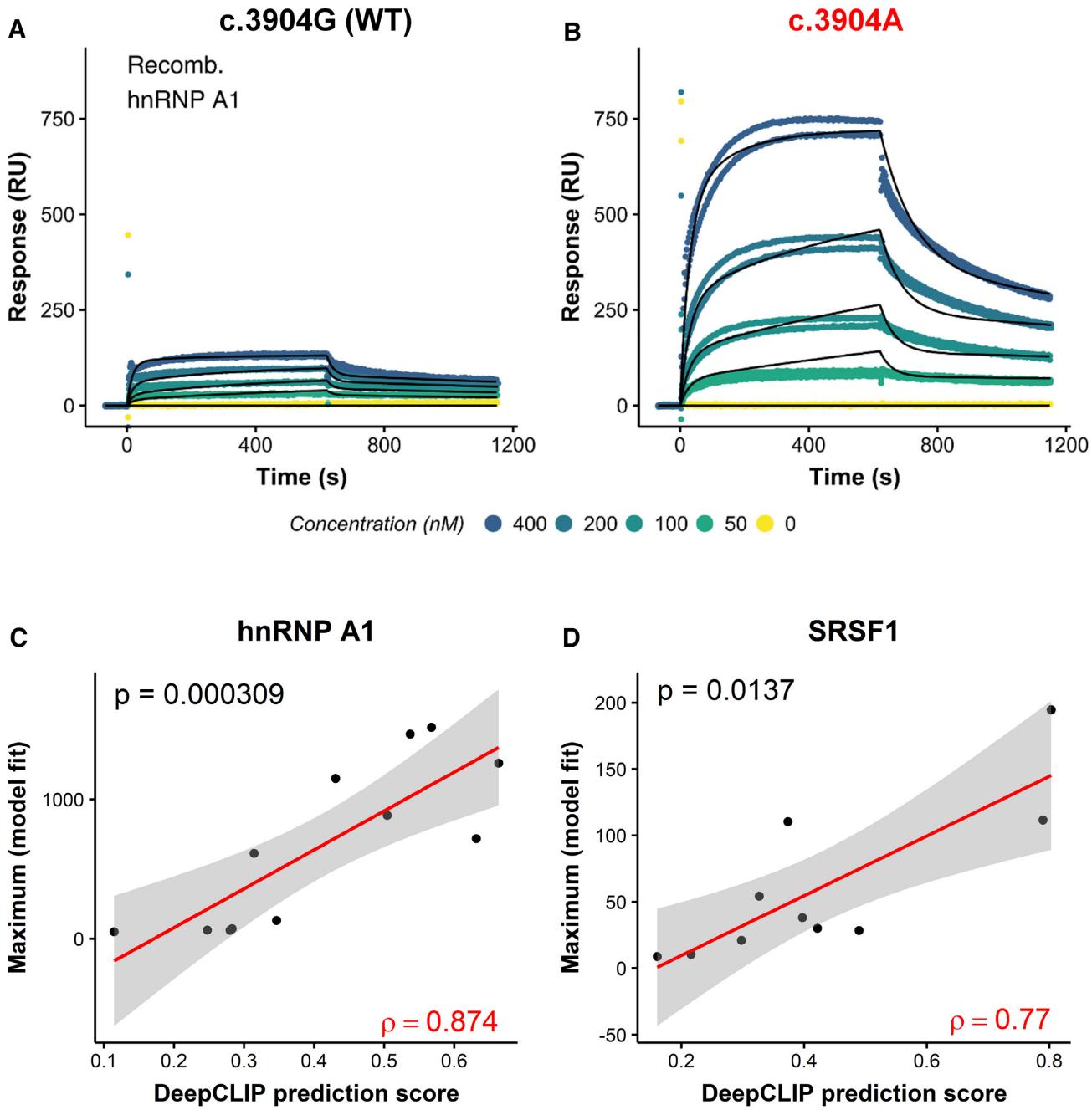

Figure 5. DeepCLIP predictions correlate with binding affinity studies. (**A, B**) Scatter plots of raw hnRNP A1 SPRi measurements (dots) and the fitted models (black lines) to wt (A) and mutant (B) *ATP7A* exon 20 15 nt oligonucleotide. (**C, D**) Scatter plots showing DeepCLIP predictions of hnRNP A1 binding (C) and SRSF1 binding (D) to 15 nt oligonucleotides corresponding to wt and mutant pairs from Raponi *et al.* against the maximum of the binding model fitted to SPRi measurements. The 95% confidence intervals of fitted linear regression models (red line) are shown in gray. Spearman's rho is show in red in lower right corner, and the *P*-value in the upper left.

displays increased pseudoexon inclusion, some of which are muscle and neuron-specific (54). Because these pseudoexons are not necessarily conserved in humans, they may not directly relate to ALS, but they may nevertheless improve our understanding of how some pseudoexons are selectively up-regulated in motor neurons. This can prove important to the understanding of the underlying molecular pathology of ALS. We therefore used DeepCLIP to analyze TDP-43-repressed pseudoexons in mice to examine the tissue specific differences in TDP-43 binding. We found that DeepCLIP overall predicted decreased binding to the region down-stream of the 5′ splice site of pseudoexons that are neuron specific compared to pseudoexons that are muscle-specific (Figure 6, Supplementary Figure S45), while neuron-specific pseudoexons were predicted to bind more TDP-43 in the region covering the poly-pyrimidine tract compared to muscle-specific pseudoexons. This might reflect interplay between TDP-43 and tissue-specific factors

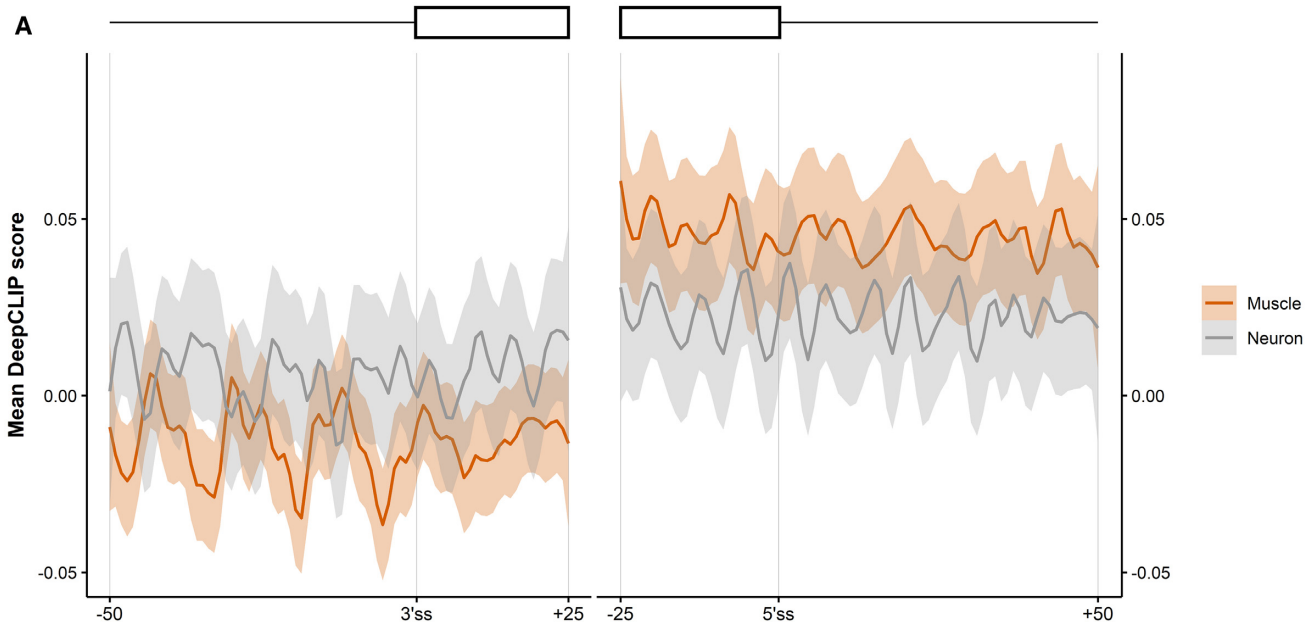

Figure 6. DeepCLIP analysis of TDP-43 repressed pseudoexons indicate position-dependent tissue-specificity. (**A**) The average DeepCLIP TDP-43 profile scores of 58 neuron-specific and 79 muscle-specific pseudoexons activated in TDP-43-null mice in the areas covering the 25 first and last nucleotides of the pseudoexon, and the 50 nt spanning intronic regions. 95%-confidence intervals are indicated by shaded areas.

interacting with these regions in a position-dependent manner. These results indicate that sequence analysis of known pseudoexons can lead to discovery of neuron-specific pseudoexons involved in ALS pathology in humans.

DISCUSSION

We present DeepCLIP, a novel deep learning approach to modeling RNA-binding protein sites using a shallow neural network composed of CNN and LSTM layers to capture context-dependent binding. DeepCLIP generalizes well across a diverse set of sequences in both *in vitro* and *in vivo* settings, and produces a profile of the sequence, which indicates sequence elements important for the binding of the RNA binding protein in question.

Previous RNA–protein binding classifiers attempted to improve their performance by incorporating context dependencies in a number of different ways, e.g. secondary and tertiary structure, known binding sites of other RNA-binding proteins, and annotated gene regions such as exons, introns and UTRs.

With DeepCLIP, we demonstrate that a neural network, in which context dependency is not pre-defined, but modeled implicitly by a BLSTM layer, is competitive or outperforming existing classifiers that, in addition to the RNA-sequences, depend on one or more predefined data sets containing different categories of contextual information. This allows DeepCLIP to be agnostic with regard to other inputs and makes it robust towards any limitations in e.g. the modeling of the structure, or the level and quality of annotation of gene structure and other protein binding sites.

Another benefit is the omission of a lengthy structure modeling step. We compared the runtime of DeepCLIP, iDeepS and GraphProt using the positive sequences from the GraphProt benchmark sets as input to their respective models and found that DeepCLIP was ∼250 faster than GraphProt, and ∼50 times faster than iDeepS based on linear regression of their running times (Supplementary Figure S46). This greatly improves the scalability of RBP analysis, enabling larger transcriptomic regions to be analyzed by DeepCLIP in long prediction mode.

Secondary RNA structure modeling was previously shown to improve model accuracy for the datasets Ago1-4, CAPRIN1, IGF2BP1-3, MOV10 and ZC3H7B (14), and in general RBPs preferentially bind to structured RNA (77). While DeepCLIP does not directly include predictions of secondary structures when classifying, DeepCLIP AUROC measures for these proteins were the highest of all classifiers except for IGF2BP1-3, where iONMF, which also does not model structure, had a higher AUROC score. This indicates that the BLSTM layer of DeepCLIP captures contextual dependencies that are as important as secondary RNA structure modeling. The inclusion of tertiary structure modeling improved the performance of mDBN+ on the hnRNP C, PTBP1 and TDP-43 datasets over DeepCLIP, indicating that more advanced structure modeling provides an improvement in the prediction of some proteins.

DeepCLIP produces motifs of varying sizes ranked by the average information content (Figure 2A, B). The top-ranking motifs of DeepCLIP for the analyzed proteins were visually remarkably similar to core binding sites as described in literature (67,68,72,78–82) (Supplementary Figure S6). The motifs depicted in Supplementary Figure S6 are based on the top-1000 scoring sequences from the positive and negative input dataset and represent the two pseudo-PFMs with the highest mean information score among the five pseudo-PFMs produced. In addition, DeepCLIP allows pseudo-PFM generation based on sequences

scoring above a given threshold, e.g. sequences with score higher than 0.5 (see Supplementary Figure S31), which may produce slightly different pseudo-PFMs. This is a result of the more heterogeneous sequence input when using more of the training input, which results in visualizations with less conserved motifs, but the underlying CNN filters remain the same. DeepCLIP does not model or generate motifs describing structural preferences of RBPs. However, this may be obtained by using structure modeling software on high-scoring sites identified by DeepCLIP.

When searching for binding sites in longer sequences, the information contained in the DeepCLIP binding profiles becomes invaluable, since it unravels interesting areas that are important for protein binding (Figure 2E, F). In the case where a longer sequence contains mainly strong background patterns and only a small segment with binding site potential, DeepCLIP and other tools will be prone to classify this sequence as a background sequence. However, DeepCLIP is able to identify the foreground segment and highlight it on the binding profile of the sequence.

DeepCLIP binding profiles can be used for estimating high- and low-affinity regions of sequences (Figure 4). The prediction scores, which are directly based on the binding profile values, display a strong correlation with affinity studies (Figure 5) suggesting that DeepCLIP successfully captures binding preferences of RBPs. To this end, the binding profiles produced by DeepCLIP can be used to identify splicing regulatory sites that can be targeted by SSOs (Figure 4A), which is an important novel and missing functionality of existing binding site discovery tools. Thus, DeepCLIP greatly facilitates the design of new drugs based on blocking protein–RNA binding sites, which is a very promising new therapeutic approach, as illustrated for instance by the recent success of the SpinrazaTM SSO in treating SMA (83,84). DeepCLIP could potentially be used to predict binding of RBPs to SSOs themselves (Supplementary Table S9). However, as SSOs are chemically modified, their binding properties are not directly comparable to RNA and the predictions based on studies of RNA–protein binding are only very uncertain estimates.

DeepCLIP models are trained on protein–RNA binding data from studies on specific proteins and cannot be used to predict binding preferences of proteins where no such data yet exists. Because the binding properties of both proteins and RNA depend on their sequence, it may be possible to train networks that predict binding based on protein and RNA sequences in pairs, such that novel interactions could be predicted from unknown RBPs. Several approaches for this exist, such as autoencoders or generative adversarial networks (GAN). However, some RBPs bind RNA sequence patterns and structures in a more unspecific manner, making it difficult to accurately predict possible binding sites. The utility of this approach needs thorough and rigorous investigation and testing.

DeepCLIP models also depend on the quality of the training data. Significant methodological biases can result in models that recognize more generally CLIP sites than protein specific sites. This is, however, a common trait for the methods and not just DeepCLIP as evidenced by overall similar performance on independent datasets (Supplementary Figure S5a and b). An example of this is the PUM2 model trained on PAR-CLIP data, which performed relatively poorly on the POSTAR2 and ENCODE eCLIP datasets, although in general DeepCLIP had the best performance for PUM2 out of the three methods compared (Supplementary Table S4). This model also produces many high prediction scores when used on the positive sequences from the PAR-CLIP traning set used to train the QKI model (AUROC = 0.73197), while the QKI model is much less biased (AUROC = 0.93383) (Supplementary Figure S47a and b). This kind of test represents the most challenging as both input classes contain the biases specific to PAR-CLIP. In our benchmarks, we kept training parameters the same for all models and used the backgrounds originally used by GraphProt. If we instead adjust parameters of the model to use only two CNN filters of length 7 and 8, set batch size to 64 and train the PUM2 model with an equal mix of the other PAR-CLIP datasets within the GraphProt benchmark set, but excluding the QKI training set, we obtain a model which is able to better distinguish between PUM2 and QKI (AUROC increase from 0.73197 to 0.90104), while improving performance on the eCLIP datasets from POSTAR2 (AUROC increasing from 0.64414 to 0.69185) and ENCODE (AUROC increase from 0.63616 to 0.66780). However, because DeepCLIP produces binding profiles we can still use the more biased models to identify potential targets. Using DeepCLIP's long prediction mode we analyzed a transcript, which is known to bind both QKI and PUM2, and defined binding sites as 9 nt windows or more with a mean profile score >0.3. We then mapped these sites onto the transcript along with known PAR-CLIP and eCLIP sites and found that the original PUM2 model identified four binding site that were all overlapped by known CLIP sites, either PAR-CLIP or eCLIP. None of them overlapped the QKI sites, and were all located within the 3′UTR, consistent with PUM2's known binding preference for 3′UTR regions. This demonstrates that despite bias in the model, the profile produced still allows for specific identification of PUM2 sites, in the presence of known QKI sites. This type of analysis also demonstrates again how binding profiles are much more informative than overall prediction scores and are necessary when comparing sequences of different lengths, as DeepCLIP's architecture is not suitable for comparing prediction scores directly between sequences of different lengths. Similarly, mutations at the edge of sequences should not be analyzed with DeepCLIP, as the contextual information in these areas is insufficient and edge-specific effects may lead to misleading results.

In summary, DeepCLIP models provide valuable insight into the functional consequences of sequence variants. Both *in vitro* binding assays and *in vivo* splicing assays as well as observed splicing of disease-causing mutations in patients cells correlate well with DeepCLIP predictions. This demonstrates that an *in silico* analysis with DeepCLIP can serve as a valuable tool for assessing the functional effects of potentially pathogenic sequence variants, providing an important tool for clinical diagnosis. Finally, we demonstrate that DeepCLIP can serve as tool for designing efficient SSOs for correcting aberrant splicing caused by disease-causing mutations. DeepCLIP is freely available, both as stand-alone and as a webtool. The webtool allows analysis

of sequences and sequence variants with multiple models in one analysis, as many RBPs compete for binding site positions on RNA molecules, and identifying the most likely change or binding partner is important for further experimental analysis.

DATA AVAILABILITY

All data, including raw data for all figures, in this study is available either through http://github.com/deepclip or http://deepclip.compbio.sdu.dk or direct communication with the authors.

CODE AVAILABILITY

All code is available at http://github.com/deepclip.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to laboratory technicians Kasper Stonor Werner, Susanne Ruszczycka and Gabriella Jensen for their assistance with experiments.
Author contributions: T.K.D. conceived and designed the project. The core algorithm (intrinsic neural network functionalities, layer inter-connectivities and general network composition) was designed and implemented by A.G.B.G. DeepCLIP was implemented by A.G.B.G. and T.K.D. The DeepCLIP web-interface was designed and implemented by S.J.L. Experiments were designed by T.K.D., A.G.B.G. and B.S.A. and performed by A.G.B.G., U.S.S.P., L.L.H., G.H.B., M.B.H. and A.M.H. All computational analyses were designed by A.G.B.G., T.K.D., B.S.A. and J.B., and performed by A.G.B.G. and T.K.D. All authors read and contributed to the manuscript.

FUNDING

Lundbeckfonden [R231-2016-2823 to T.K.D.]; Muskelsvindfonden; Natur og Univers, Det Frie Forskningsråd [4181-00515 to B.S.A.]; Novo Nordisk Fonden [NNF17OC0029240 to B.S.A.]; ODEx at SDU; the work of A.G.B.G. and J.B. has been supported by VILLUM Young Investigator [73528]; Parts of J.B.'s work was also funded by H2020 project RepoTrial [777111]; Node hours on the Abacus 2.0 HPC was provided by the DeiC National HPC Centre, SDU. Funding for open access charge: ODEx at SDU.
Conflict of interest statement. None declared.

REFERENCES

1. Xiong,H.Y., Alipanahi,B., Lee,L.J., Bretschneider,H., Merico,D., Yuen,R.K., Hua,Y., Gueroussov,S., Najafabadi,H.S., Hughes,T.R. *et al.* (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806.
2. Raponi,M., Kralovicova,J., Copson,E., Divina,P., Eccles,D., Johnson,P., Baralle,D. and Vorechovsky,I. (2011) Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Hum. Mutat.*, **32**, 436–444.
3. Shibata,A., Okuno,T., Rahman,M.A., Azuma,Y., Takeda,J., Masuda,A., Selcen,D., Engel,A.G. and Ohno,K. (2016) IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome. *J. Hum. Genet.*, **61**, 633–640.
4. Bruun,G.H., Bang,J.M.V., Christensen,L.L., Broner,S., Petersen,U.S.S., Guerra,B., Gronning,A.G.B., Doktor,T.K. and Andresen,B.S. (2018) Blocking of an intronic splicing silencer completely rescues IKBKAP exon 20 splicing in familial dysautonomia patient cells. *Nucleic Acids Res.*, **46**, 7938–7952.
5. Sinha,R., Kim,Y.J., Nomakuchi,T., Sahashi,K., Hua,Y., Rigo,F., Bennett,C.F. and Krainer,A.R. (2018) Antisense oligonucleotides correct the familial dysautonomia splicing defect in IKBKAP transgenic mice. *Nucleic Acids Res.*, **46**, 4833–4844.
6. Hua,Y., Vickers,T.A., Okunola,H.L., Bennett,C.F. and Krainer,A.R. (2008) Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. *Am. J. Hum. Genet.*, **82**, 834–848.
7. Ule,J., Jensen,K., Mele,A. and Darnell,R.B. (2005) CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods*, **37**, 376–386.
8. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr., Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
9. König,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
10. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
11. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
12. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.
13. Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.
14. Maticzka,D., Lange,S.J., Costa,F. and Backofen,R. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**, R17.
15. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
16. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
17. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
18. Strazar,M., Zitnik,M., Zupan,B., Ule,J. and Curk,T. (2016) Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, **32**, 1527–1535.
19. Henaff,M., Weston,J., Szlam,A., Bordes,A. and LeCun,Y. (2016) Tracking the world state with recurrent entity networks. arXiv doi: https://arxiv.org/abs/1612.03969, 12 December 2016, preprint: not peer reviewed.
20. Redmon,J. and Farhadi,A. (2016) YOLO9000: better, faster, stronger. arXiv doi: https://arxiv.org/abs/1612.08242, 25 December 2016, preprint: not peer reviewed.
21. Helmstaedter,M., Briggman,K.L., Turaga,S.C., Jain,V., Seung,H.S. and Denk,W. (2013) Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, **500**, 168–174.

22. LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning. *Nature*, **521**, 436–444.
23. Leung,M.K., Xiong,H.Y., Lee,L.J. and Frey,B.J. (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics*, **30**, i121–i129.
24. Quang,D. and Xie,X. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**, e107.
25. Hill,S.T., Kuintzle,R., Teegarden,A., Merrill,E. 3rd, Danaee,P. and Hendrix,D.A. (2018) A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.*, **46**, 8105–8113.
26. Almagro Armenteros,J.J., Sønderby,C.K., Sønderby,S.K., Nielsen,H. and Winther,O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.
27. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
28. Ben-Bassat,I., Chor,B. and Orenstein,Y. (2018) A deep neural network approach for learning intrinsic protein–RNA binding preferences. *Bioinformatics*, **34**, i638–i646.
29. Zhang,S., Zhou,J., Hu,H., Gong,H., Chen,L., Cheng,C. and Zeng,J. (2016) A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic. Acids. Res.*, **44**, e32.
30. Pan,X. and Shen,H.B. (2017) RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics*, **18**, 136.
31. Pan,X., Rijnbeek,P., Yan,J. and Shen,H.B. (2018) Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics*, **19**, 511.
32. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.
33. Huppertz,I., Attig,J., D'Ambrogio,A., Easton,L.E., Sibley,C.R., Sugimoto,Y., Tajnik,M., Konig,J. and Ule,J. (2014) iCLIP: protein–RNA interactions at nucleotide resolution. *Methods*, **65**, 274–287.
34. Nielsen,K.B., Sorensen,S., Cartegni,L., Corydon,T.J., Doktor,T.K., Schroeder,L.D., Reinert,L.S., Elpeleg,O., Krainer,A.R., Gregersen,N. *et al.* (2007) Seemingly neutral polymorphic variants may confer immunity to splicing-inactivating mutations: a synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer. *Am. J. Hum. Genet.*, **80**, 416–432.
35. Krizhevsky,A., Sutskever,I. and Hinton,G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, **60**, 84–90.
36. LeCun,Y., Boser,B., Denker,J., Henderson,D., Howard,R., Hubbard,W. and Jackel,L. (1990) Handwritten Digit Recognition with a Back-Propagation Network. In: *Advances in Neural Information Processing Systems 2*. Morgan-Kaufmann, pp. 396–404.
37. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
38. Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
39. Al-Rfou,R., Alain,G., Almahairi,A., Angermueller,C., Bahdanau,D., Ballas,N., Bastien,F., Bayer,J. and Team,T.D. (2016) Theano: a {Python} framework for fast computation of mathematical expressions. arXiv doi: https://arxiv.org/abs/1605.02688, 09 May 2016, preprint: not peer reviewed.
40. Dieleman,S., Schlüter,J., Raffel,C., Olson,E., Sønderby,S.K., Nouri,D., Maturana,D., Thoma,M., Battenberg,E., Kelly,J. *et al.* (2015) Lasagne: First release. Zenodo doi: http://dx.doi.org/10.5281/zenodo.27878, 13 August 2015, preprint: not peer reviewed.
41. Goldberg,Y. (2016) A primer on neural network models for natural language processing. *J. Artif. Intell. Res.*, **57**, 345–420.
42. Goodfellow,I., Bengio,Y. and Courville,A. (2016) In: *Deep Learning*. MIT Press.
43. Nair,V. and Hinton,G.E. (2010), Rectified Linear Units Improve Restricted Boltzmann Machines. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. Scientific Research Publishing, Haifa. pp. 807–814.
44. Srivastava,R.K., Masci,J., Kazerounian,S., Gomez,F. and Schmidhuber,J.R. (2013) Compete to Compute. In: *Advances in neural information processing systems*. pp. 2310–2318
45. Park,S., Min,S., Choi,H. and Yoon,S. (2016) deepMiRGene: deep neural network based precursor microRNA prediction. arXiv doi: https://arxiv.org/abs/1605.00017, 29 April 2016, preprint: not peer reviewed.
46. Bahdanau,D., Cho,K. and Bengio,Y. (2014) Neural machine translation by jointly learning to align and translate. arXiv doi: https://arxiv.org/abs/1409.0473, 19 May 2016, preprint: not peer reviewed.
47. Kingma,D. and Ba,J. (2014) Adam: A method for stochastic optimization. arXiv doi: https://arxiv.org/abs/1412.6980, 30 Jan 2017, preprint: not peer reviewed.
48. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
49. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
50. Zhu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2019) POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.*, **47**, D203–D211.
51. Bruun,G.H., Doktor,T.K., Borch-Jensen,J., Masuda,A., Krainer,A.R., Ohno,K. and Andresen,B.S. (2016) Global identification of hnRNP A1 binding sites for SSO-based splicing modulation. *BMC Biol.*, **14**, 54.
52. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
53. Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.*, **1418**, 335–351.
54. Jeong,Y.H., Ling,J.P., Lin,S.Z., Donde,A.N., Braunstein,K.E., Majounie,E., Traynor,B.J., LaClair,K.D., Lloyd,T.E. and Wong,P.C. (2017) Tdp-43 cryptic exons are highly variable between cell types. *Mol Neurodegener*, **12**, 13.
55. Baltz,A.G., Munschauer,M., Schwanhausser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K., Milek,M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.
56. Hoell,J.I., Larsson,E., Runge,S., Nusbaum,J.D., Duggimpudi,S., Farazi,T.A., Hafner,M., Borkhardt,A., Sander,C. and Tuschl,T. (2011) RNA targets of wild-type and mutant FET family proteins. *Nat. Struct. Mol. Biol.*, **18**, 1428–1431.
57. Kishore,S., Jaskiewicz,L., Burger,L., Hausser,J., Khorshid,M. and Zavolan,M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.
58. Lebedeva,S., Jens,M., Theil,K., Schwanhausser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
59. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M. Jr., Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
60. Sanford,J.R., Wang,X., Mort,M., Vanduyn,N., Cooper,D.N., Mooney,S.D., Edenberg,H.J. and Liu,Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.
61. Sievers,C., Schlumpf,T., Sawarkar,R., Comoglio,F. and Paro,R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160.
62. Wang,Z., Kayikci,M., Briese,M., Zarnack,K., Luscombe,N.M., Rot,G., Zupan,B., Curk,T. and Ule,J. (2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.*, **8**, e1000530.
63. Xue,Y., Zhou,Y., Wu,T., Zhu,T., Ji,X., Kwon,Y.S., Zhang,C., Yeo,G., Black,D.L., Sun,H. *et al.* (2009) Genome-wide analysis of PTB-RNA

interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, **36**, 996–1006.
64. Orenstein,Y., Wang,Y. and Berger,B. (2016) RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data. *Bioinformatics*, **32**, i351–i359.
65. Buratti,E., Brindisi,A., Pagani,F. and Baralle,F.E. (2004) Nuclear factor TDP-43 binds to the polymorphic TG repeats in CFTR intron 8 and causes skipping of exon 9: a functional link with disease penetrance. *Am. J. Hum. Genet.*, **74**, 1322–1325.
66. Barker,A., Epis,M.R., Porter,C.J., Hopkins,B.R., Wilce,M.C., Wilce,J.A., Giles,K.M. and Leedman,P.J. (2012) Sequence requirements for RNA binding by HuR and AUF1. *J. Biochem. (Tokyo)*, **151**, 423–437.
67. Gregersen,L.H., Schueler,M., Munschauer,M., Mastrobuoni,G., Chen,W., Kempa,S., Dieterich,C. and Landthaler,M. (2014) MOV10 Is a 5′ to 3′ RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3′ UTRs. *Mol. Cell*, **54**, 573–585.
68. Perez,I., Lin,C.H., McAfee,J.G. and Patton,J.G. (1997) Mutation of PTB binding sites causes misregulation of alternative 3′ splice site selection in vivo. *RNA*, **3**, 764–778.
69. Wang,X., Juan,L., Lv,J., Wang,K., Sanford,J.R. and Liu,Y. (2011) Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. *BMC Genomics*, **12**(Suppl. 5), S8.
70. Dember,L.M., Kim,N.D., Liu,K.Q. and Anderson,P. (1996) Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities. *J. Biol. Chem.*, **271**, 2783–2788.
71. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
72. Feng,H., Bao,S., Rahman,M.A., Weyn-Vanhentenryck,S.M., Khan,A., Wong,J., Shah,A., Flynn,E.D., Krainer,A.R. and Zhang,C. (2019) Modeling RNA-binding protein specificity in vivo by precisely registering protein–RNA crosslink sites. *Mol. Cell*,**74**, 1189–120.
73. Waddell,L., Wiley,V., Carpenter,K., Bennetts,B., Angel,L., Andresen,B.S. and Wilcken,B. (2006) Medium-chain acyl-CoA dehydrogenase deficiency: genotype-biochemical phenotype correlations. *Mol. Genet. Metab.*, **87**, 32–39.
74. Arai,T., Hasegawa,M., Akiyama,H., Ikeda,K., Nonaka,T., Mori,H., Mann,D., Tsuchiya,K., Yoshida,M., Hashizume,Y. *et al.* (2006) TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochem. Biophys. Res. Commun.*, **351**, 602–611.
75. Neumann,M., Sampathu,D.M., Kwong,L.K., Truax,A.C., Micsenyi,M.C., Chou,T.T., Bruce,J., Schuck,T., Grossman,M., Clark,C.M. *et al.* (2006) Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, **314**, 130–133.
76. Ling,J.P., Pletnikova,O., Troncoso,J.C. and Wong,P.C. (2015) TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, **349**, 650–655.
77. Sanchez de Groot,N., Armaos,A., Grana-Montes,R., Alriquet,M., Calloni,G., Vabulas,R.M. and Tartaglia,G.G. (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
78. Gao,F.B., Carson,C.C., Levine,T. and Keene,J.D. (1994) Selection of a subset of mRNAs from combinatorial 3′ untranslated region libraries using neuronal RNA-binding protein Hel-N1. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11207–11211.
79. Munteanu,A., Mukherjee,N. and Ohler,U. (2018) SSMART: sequence-structure motif identification for RNA-binding proteins. *Bioinformatics*, **34**, 3990–3998.
80. Colombrita,C., Onesto,E., Megiorni,F., Pizzuti,A., Baralle,F.E., Buratti,E., Silani,V. and Ratti,A. (2012) TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. *J. Biol. Chem.*, **287**, 15635–15647.
81. Adinolfi,M., Pietrosanto,M., Parca,L., Ausiello,G., Ferre,F. and Helmer-Citterich,M. (2019) Discovering sequence and structure landscapes in RNA interaction motifs. *Nucleic. Acids. Res.*, **47**, 4958–4969.
82. Meyer,C., Garzia,A., Mazzola,M., Gerstberger,S., Molina,H. and Tuschl,T. (2018) The TIA1 RNA-binding protein family regulates EIF2AK2-mediated stress response and cell cycle progression. *Mol. Cell*, **69**, 622–635 .
83. Finkel,R.S., Mercuri,E., Darras,B.T., Connolly,A.M., Kuntz,N.L., Kirschner,J., Chiriboga,C.A., Saito,K., Servais,L., Tizzano,E. *et al.* (2017) Nusinersen versus sham control in infantile-onset spinal muscular atrophy. *N. Engl. J. Med.*, **377**, 1723–1732.
84. Mercuri,E., Darras,B.T., Chiriboga,C.A., Day,J.W., Campbell,C., Connolly,A.M., Iannaccone,S.T., Kirschner,J., Kuntz,N.L., Saito,K. *et al.* (2018) Nusinersen versus sham control in later-onset spinal muscular atrophy. *N. Engl. J. Med.*, **378**, 625–635.