Practical Recommendations and Limitations

From our analyses and experience with the deepCRE toolkit we have summarized the following section, providing users with expert knowledge for best practice. The underlying models of the deepCRE toolkit are trained on DNA sequence for the prediction of low and high gene transcript levels.

- Note that the deepCRE model is trained on DNA sequence and RNAseq coverage. Consequently, the deepCRE model identifies sequence features relevant for regulating gene expression, including RNA turnover and RNA stabilisation, too (Peleke et al. 2024). The model cannot identify sequences that are shared among all transcripts, e.g. general transcript initiation and sequence features relevant for translation initiation.
- The estimation of gene transcript levels for training used bulk transcriptome/RNAseq. Accordingly, the deepCRE model can only identify sequence features relevant for regulating transcript levels within the respective bulk experiment. In the deepCRE toolkit version 1 models are available for bulk leaf and root experiments.

We recommend using the *Explanation* tab to analyse above mentioned cases with care. Within this version, the deepCRE toolkit does not provide qualitative annotations of *cis*-regularoty elements. Externally, researchers may use public databases for transcription factor binding sites (TFBS) like JASPAR or RNA-binding proteins (RBP) like RBP2GO (Castro-Mondragon et al. 2022; Caudron-Herger et al. 2021) and consolidate extended literature.

The predicted probabilities are highly dependent on gene start (transcript start sites) and end sites (transcript termination sites) due to extraction affecting spacing of CREs.

- Note that differences in annotation versions may change predicted probabilities due to change in extraction windows.
- Note that differences in predicted probabilities between closely related genotypes/species may occur due to structural variation (IN/DEL)

mutations in the extracted windows and a change of gene start and end sites.

- Note in the deepCRE toolkit that alternative transcription start sites or splicing variants are ignored by default during the processing of genes for prediction.

We recommend using the *Mutation* mode to analyse the above mentioned cases. Externally, researchers may use sequence alignment tools and check provided versions when using the deepCRE toolkit.

The predicted probabilities are highly interdependent for individual gene flanking regions.

- Note that predicted probabilities result from a combination of different sequence features. To our knowledge these combinations can depend on short context, e.g. reflecting the motif of a transcription factor binding site, but can also exist for longer ranges across different segments.
- Note that the model identifies features with positional preferences that affect the predictions and importance score depending on its position relation to transcript start or end sites (Peleke et al. 2024).

We recommend using the *Explanation* and *Mutation* functions to analyse above mentioned cases. Within this version, the deepCRE toolkit does not provide qualitative annotations of cis-regulatory elements. Externally, researchers may use public databases for TFBS like JASPAR or RBP like RBP2GO (Castro-Mondragon et al. 2022; Caudron-Herger et al. 2021) consolidate related literature on regulatory positional dependencies (Voichek et al. 2024; Duttke et al. 2024; Hardy and Balcerowicz 2024).

The deepCRE models' accuracy depends on species and sequence divergence affecting prediction results.

- Note that single species reference accuracies are available for four selected model organisms (*Arabidopsis thaliana, Solanum lycopersicum, Sorghum bicolor* and *Zea mays*) inferred from cross-prediction

experiments become lower with species and sequence divergence (Peleke et al. 2024).

- Note that prediction accuracies of genes with tissue specific gene transcript levels (i.e. tested for root and leaf samples) are much lower compared to genes with similar gene transcript levels across multiple tissues (Peleke et al. 2024).

We recommend that researchers use models trained for the same organism, if available or choose a model from an organism that is most closely related. While the predicted probabilities may vary across different models, at least leaf and root models per species model should provide congruent results, reflecting a prediction accuracy of more than 90%.

The Interpretations results can vary with user-dependent handling and data availability.

- Note that mutation like structural variations and SNPs change the sequence composition of the gene regions extracted windows.
- Note that uploading VCFs only works for SNPs.
- Note that missing data, e.g. represented as empty (“N”s, “X”s or “-”s), is assigned with an importance score of zero.
- Note that genes smaller than 1 kbp will always contain empty sequences in the extraction windows center to avoid sequence duplication.

We recommend avoiding the use of “N”s and gaps in the analyses using the deepCRE toolkit. Missing sequence data should be substituted by flanking sequence data if available. To achieve best results, researchers should substitute empty sequences in the gUR or the gTDR with upstream sequences. If there are empty sequences in the gTUR or the gDR, downstream sequences should be used for substitution. If the researchers intend to characterize SVs, we highly recommend using a corresponding gene feature annotation file.

The deepCRE model's accuracy depends on a state of genome completion, experimental variation and annotation quality which may affect prediction results.

- Note that deepCRE models were trained on specific versions and datasets available to the publication of (Peleke et al. 2024). The deepCRE toolkit team observed that the model accuracies change with training data used to some extent (unpublished observation).

We recommend documenting the models version by citing this publication (deepCRE toolkit version 1). Upcoming updates and releases will be documented here: https://deepcre.ipk-gatersleben.de/