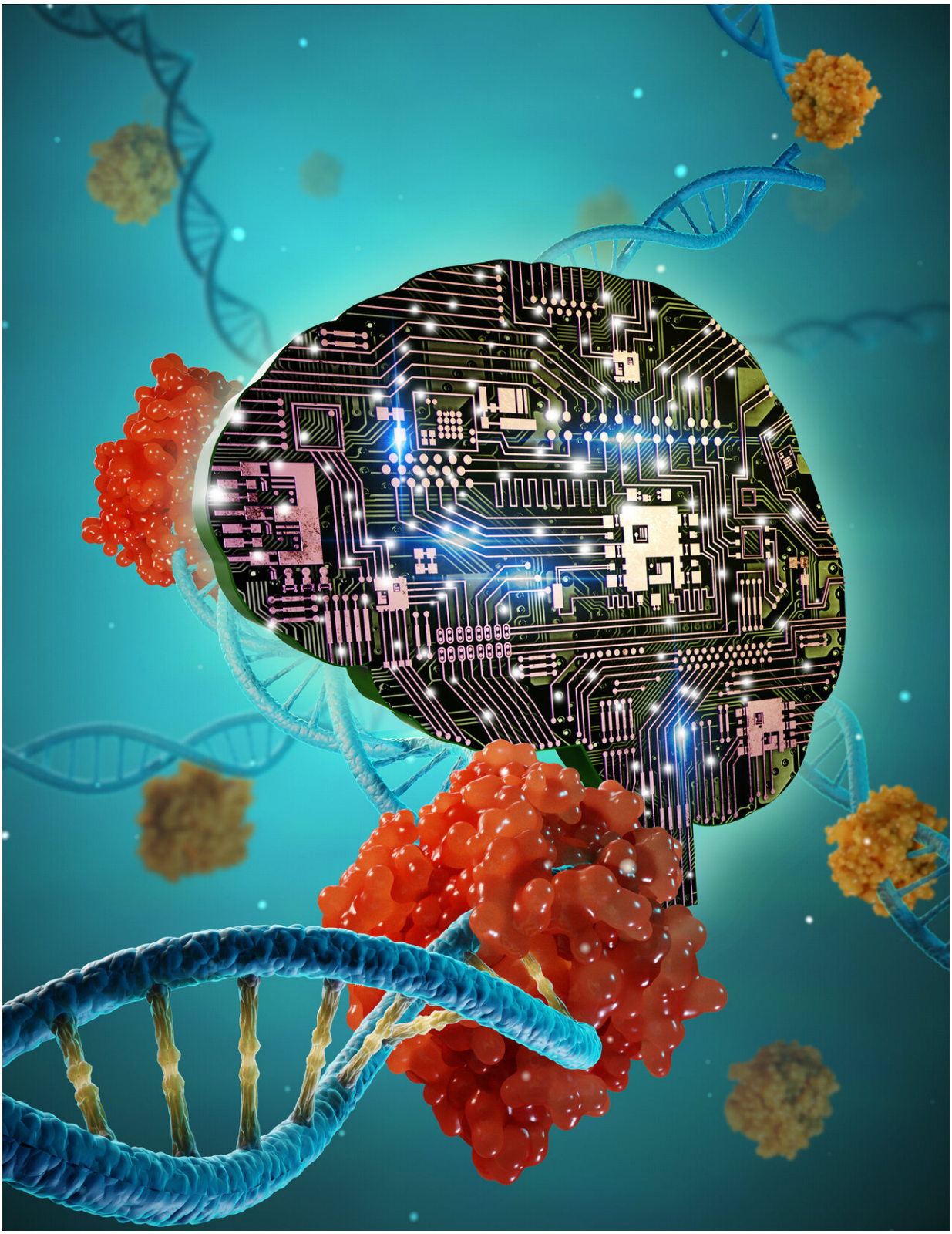
A deep learning-based model DeepSpCas9 to predict SpCas9 activity

November 22 2019, by Thamarasee Jeewandara



SpCas9 activity prediction using DeepSpCas9, a deep learning-based model with

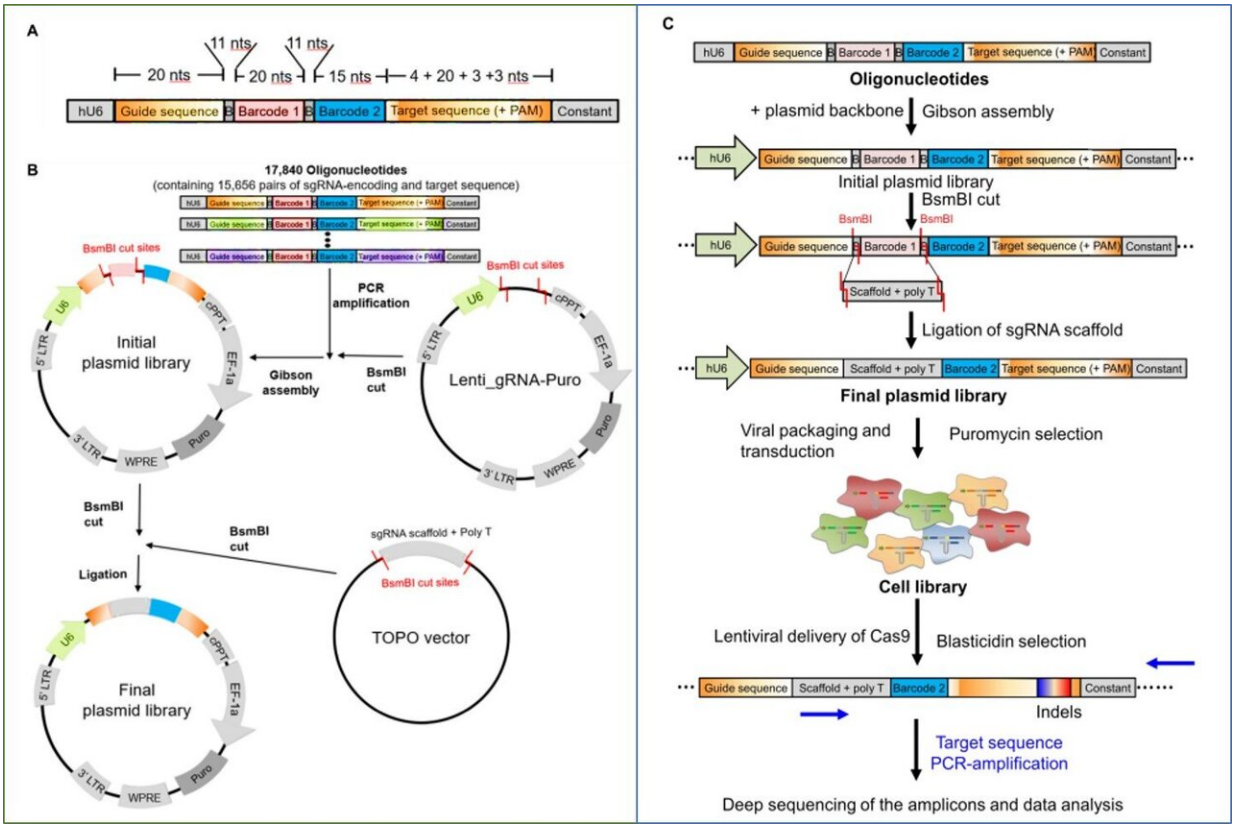
high generalization performance. Credit: Dongsu Jang, Yonsei University College of Medicine. Science Advances, doi: 10.1126/sciadv.aax9249

In a new report on *Science Advances*, Hui Kwon Kim and interdisciplinary researchers at the departments of Pharmacology, Electrical and Computer Engineering, Medical Sciences, Nanomedicine and Bioinformatics in the Republic of Korea, evaluated the activities of SpCas9; a bacterial RNA-guided Cas9 endonuclease variant (a bacterial enzyme that cuts DNA for genome editing) from Streptococcus pyogenes. They used a high-throughput approach with 12,832 target sequences based on a human cell library to build a deep learning model and predict the activity of SpCas9.

The data contained oligonucleotides (nucleotides or building blocks) containing target sequence pairs and a corresponding guide sequence to encode single-guide RNA (sgRNA), which can direct the Cas9 protein to bind and cleave a specific DNA sequence for genome editing. They implemented deep learning-based training on the large dataset of SpCas9-induced indel (insertion or deletion) frequencies to develop an SpCas9 activity predicting model named DeepSpCas9 now available online. When the team tested the software against independently generated datasets, the results showed high generalization performance, i.e. the model could properly adapt to new, previously unseen data.

The CRISPR-Cas prokaryotic adaptive immune system functions as a genome editing tool with translational research potential in a variety of species and cell types including human cells, where the capacity to accurately predict SpCas9 enzyme activity is important. Researchers had previously developed several computational models to predict SpCas9 activity based on datasets of phenotypic changes of gene-edited cells or based on medium-sized datasets of plasmid-based (vehicles that transfer

genes between bacteria and other cells) library-on-library approaches. However, the generalization performance of these models were limited, since the quality and size of the datasets were not ideal. For instance, model-predicted gene insertions and deletions (indels) to create functional knockout models (a method to inactivate genes in an experimental animal model in lab) resulted in false negatives. Additionally, these SpCas9-induced indel frequency datasets were also only medium-sized.

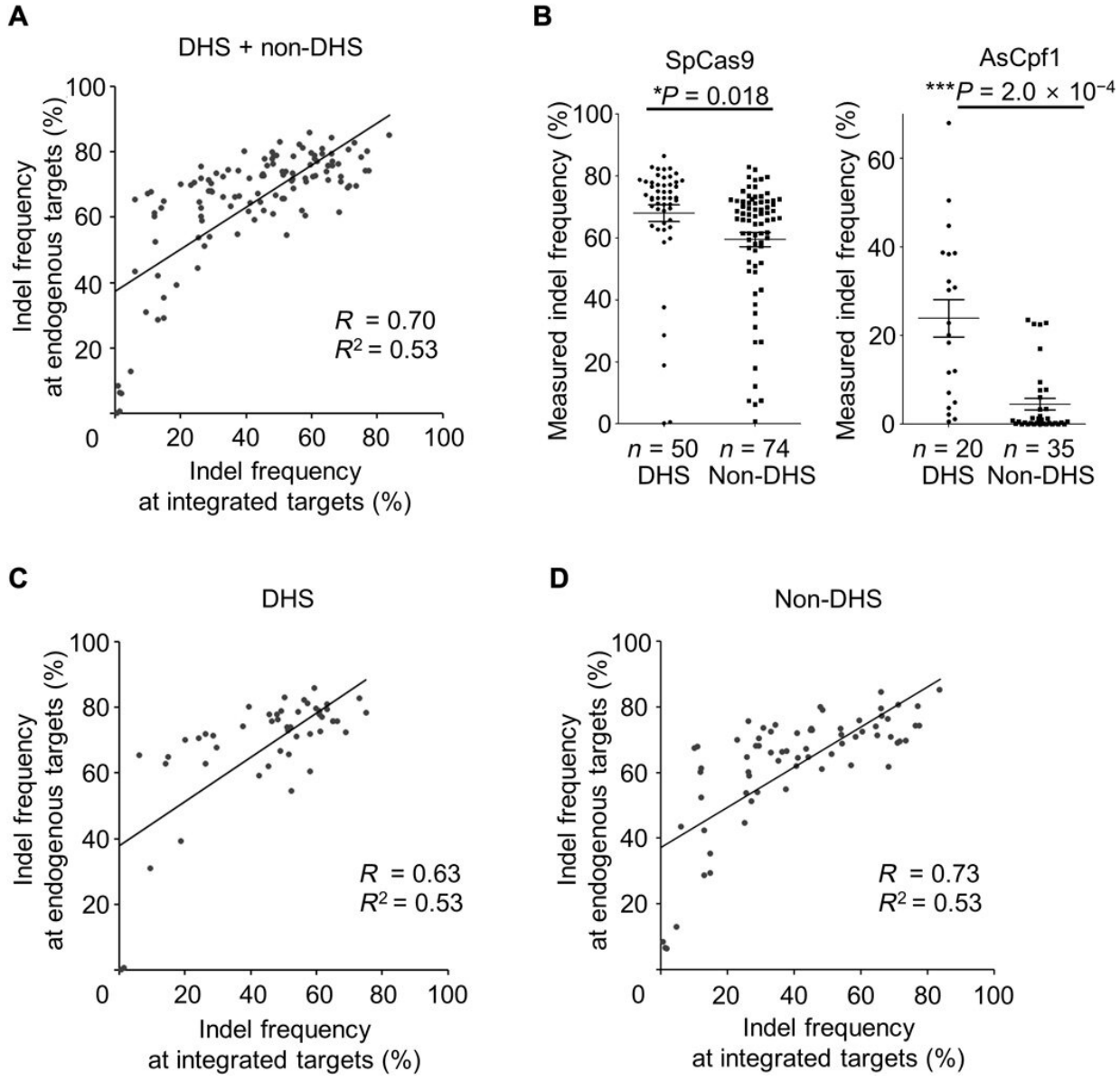


Development of a high-throughput evaluation system for Cas9-induced indel frequencies. (A) Oligonucleotide structure. Each oligonucleotide contained a 20-nt sgRNA guide sequence, a BsmBI restriction site, a 20-nt barcode sequence (barcode 1), a second BsmBI restriction site, a 15-nt barcode sequence (barcode 2), and the 30-nt corresponding target sequence including a PAM. (B) Overview of the cloning strategy for generating the plasmid library of sgRNA-encoding

and target sequence pairs. A pool of 17,840 oligonucleotides, each containing a guide RNA-target sequence pair, was PCR amplified and cloned into the Lenti_gRNA-Puro vector using Gibson assembly. This initial plasmid library was linearized using BsmBI digestion and ligated with a BsmBI digested guide RNA scaffold fragment to generate the final plasmid library. (C) A schematic of the high-throughput evaluation system used in this study. The pool of oligonucleotides was PCR-amplified and cloned into a plasmid using Gibson assembly. The sgRNA scaffold sequence was inserted into this initial plasmid library using BsmBI induced cutting and subsequent ligation. The resulting final plasmid library was used to generate a lentiviral library, which was in turn used to treat HEK293T cells to create a cell library. Lentiviral delivery of Cas9 into this cell library induced indels at the integrated target sequences with frequencies that depended on the sgRNA activity. Credit: *Science Advances*, doi: 10.1126/sciadv.aax9249

Kim et al. had previously reported on a deep learning-based computational model named DeepCpf1 to predict the activity of a different endonuclease (AsCpf1 from Acidaminococcus species) with high generalization performance. For this, they used lentiviral libraries of guide-RNA-encoding, target sequence pairs to generate a large training dataset known as DeepCpf1. While similar library-based methods were used to develop computational models that predicted indel frequencies generated by the Cas9 enzyme, a large dataset of Cas9-induced frequencies remains to be formed.

Scientists must therefore develop Cas9 activity-predicting computational models with high generalization performance. In this work, Kim et al. generated a high-throughput model to test SpCas9-induced indel frequencies at tens of thousands of target sequences by modifying their previously developed DeepCpf1 method to form DeepSpCas9. The DeepSpCas9 web tool is a deep learning-based model that can accurately predict the activities of SpCas9 with high generalization performance.



Correlations between indel frequencies at endogenous and integrated sites and effect of chromatin accessibility on indel frequencies. (A) Correlation between indel frequencies at 120 endogenous and corresponding integrated target sequences. The Spearman correlation coefficients (R) and squared Pearson correlation coefficients (R2) are shown. (B) Effect of chromatin accessibility on the activities of SpCas9 (left) and AsCpf1 (right) at endogenous sites. Indel frequencies at endogenous sites were evaluated after transfection of plasmids encoding SpCas9 or AsCpf1 and guide RNAs. Indel frequencies at the target

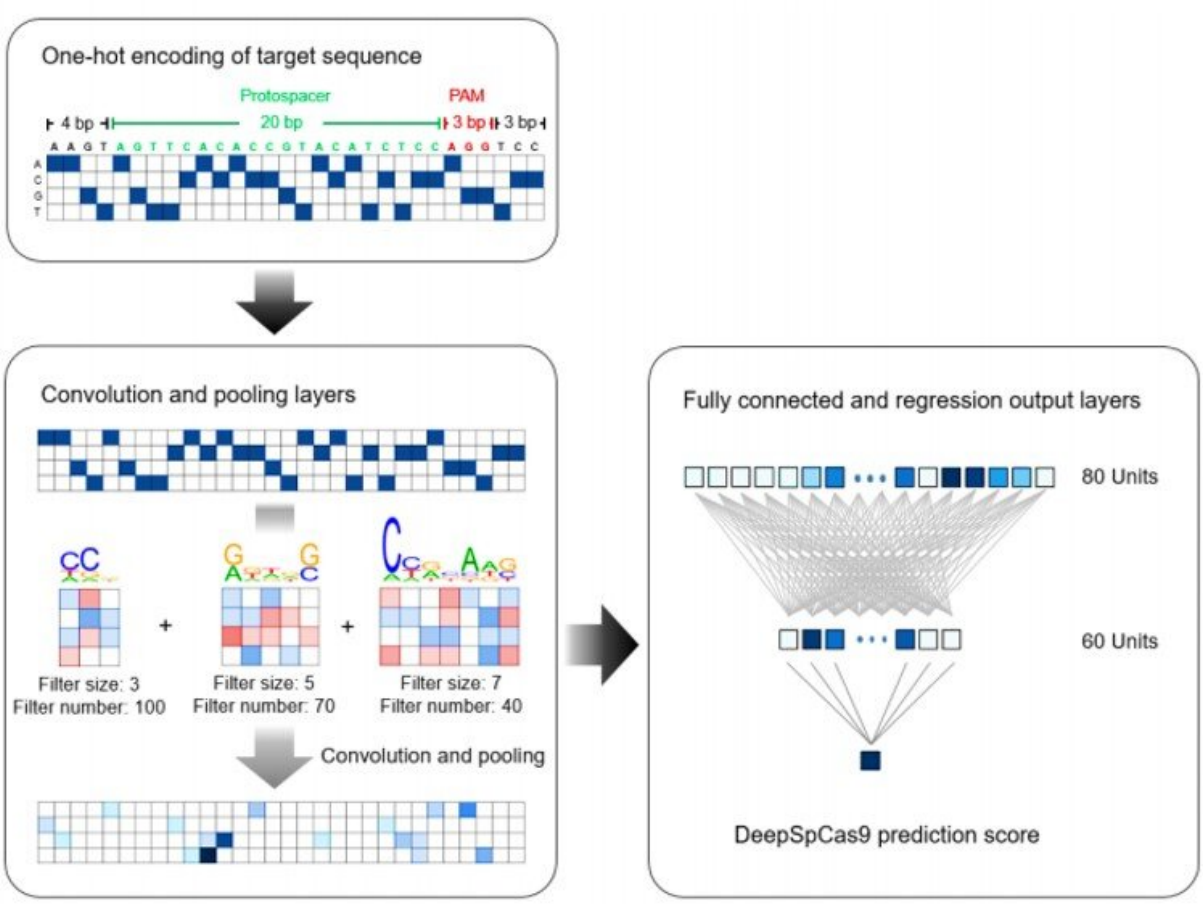
sites were compared after being divided into two groups, DHS (deoxyribonuclease I (DNase I) hypersensitive) sites and other sites (non-DHS). The numbers of analyzed target sites are as follows: SpCas9, n = 50 for DHS target sites and n = 74 for non-DHS target sites; AsCpf1, n = 20 for DHS target sites and n = 35 for non-DHS target sites. The HEK-plasmid dataset was used for drawing this graph. Error bars represent SEM. Statistical significances determined by Student's t test are shown. (C and D) Correlation between indel frequencies at endogenous and corresponding integrated target sequences at 50 DHS sites (C) and 70 non-DHS sites (D). The Spearman correlation coefficients (R) and squared Pearson correlation coefficients (R2) are shown. Credit: *Science Advances*, doi: 10.1126/sciadv.aax9249

Kim et al. first prepared a lentiviral (a complex retrovirus subfamily that can incorporate foreign DNA) library of 15,656 guide RNA (gRNA)-encoding and target sequence pairs, for high-throughput assessment of SpCas9 activities. The research team amplified the pool of oligonucleotides containing pairs of guide and target sequences using the polymerase chain reaction (PCR) and cloned them into a lentiviral plasmid (transgene delivery system to transfer genetic material between cells) using the Gibson DNA assembly technique.

In a two-step approach, the researchers cut plasmids and inserted the sgRNA scaffold sequence at the cut site to generate plasmid libraries. To subsequently form a cell library, the scientists treated human embryonic kidney cells (HEK 293T) with lentivirus generated from the plasmid library. Each cell now contained a synthetic target sequence in its genome and expressed the corresponding sgRNA. The scientists then treated the cell library with the SpCas9-encoding lentivirus to cause sgRNA-directed cleavage and indel formation at the target sequences with frequencies that depended on the sgRNA activity. To measure the indel frequencies, the scientists PCR-amplified the target sequences and subjected them to deep sequencing. Based on the high throughput

experiments, Kim et al. generated two datasets for training and testing purposes of the DeepSpCas9 model.

The scientists selected SpCas9 activities at 124 endogenous target sites with different properties of chromatin accessibility (effect of chromatin structure modifications on gene transcription) to test if the indel frequencies at the integrated synthetic target sequence correlated with those at the corresponding endogenous site. They observed a strong correlation between indel frequencies at the ingrained target sites and at the endogenous locations within the HEK cells.



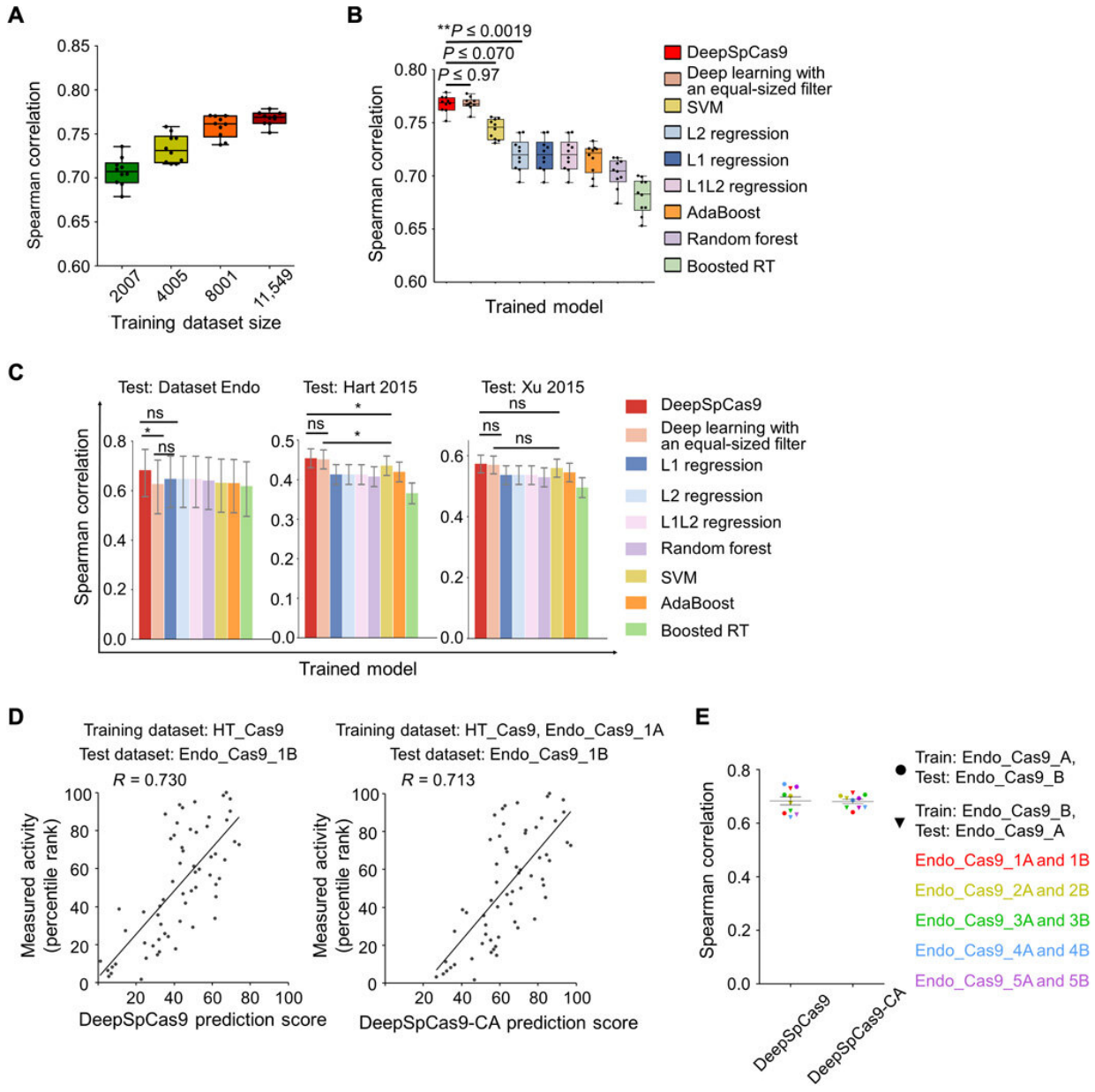
Overview of DeepSpCas9 development. DeepSpCas9 is based on a deep learning

framework using robust convolutional neural networks (CNNs). DeepSpCas9 development involved the following steps. (1) 30 base pair (bp) input sequences that include the target and neighboring sequences were converted into a four-dimensional binary matrix. (2) A total of 210 filters (100, 70, and 40 filters at 3, 5, and 7 nts in length, respectively) were shifted through the four-dimensional binary matrix to determine the position weight matrices. The maximum values were pooled from the local features calculated from the previous convolution layer to "pool out" those that were informative. (3) Pooled features were then combined according to the weighted sum and rectified linear unit non-linear function in the fully connected layers. (4) The output layer performs linear regression and predicts the activity score for each SpCas9 guide RNA. Credit: Science Advances, doi: 10.1126/sciadv.aax9249.

The research team next developed an accurate computational model to predict SpCas9 activity on a large dataset using an end-to-end deep learning framework to form DeepSpCas9 and predict the SpCas9 activity. For the base model architecture, they used a convolutional neural network (CNN, similar to ordinary neural networks) and for the input sequence they used a 30-nucleotide sequence, which they converted into a four-dimensional binary matrix using one-hot encoding (splitting columns containing numerical categorical data to many columns). To understand the generalization performance of model selection and training, the team conducted 10-fold cross-validation using Spearman correlation coefficients between experimental measurements and predicted Cas9 activity levels.

When they increased the size of the training dataset for cross-validation, the average Spearman correlation coefficients between the experimental indel frequencies and predicted scores from the DeepSpCas9 model steadily increased up to 0.77. Compared to conventional machine learning algorithms such as support vector machine (SVM), AdaBoost (adaptive boosting), random forest and gradient-boosted regression trees,

previously used for SpCas9 activity prediction, Spearman correlations of the DeepSpCas9 model were significantly higher. In total, DeepSpCas9 exhibited the best performance among all models.



Evaluation of machine learning–based computational models predicting Cas9 activities. (A) Cross-validation of DeepSpCas9 models trained on datasets of varying sizes. Each dot represents the Spearman correlation coefficient between

the measured indel frequency and the predicted activity from 10-fold cross-validation (total n = 10 correlation coefficients). (B) Cross-validation of SpCas9 activity prediction models based on previously reported machine learning–based approaches. Each dot represents the Spearman correlation coefficient between the measured indel frequency and the predicted activity from 10-fold cross-validation (total n = 10 correlation coefficients). Statistical significances between the best, next-best, and third-best models are shown (Steiger's test). In (A) and (B), the top, middle, and bottom lines in the boxes represent the 25th, 50th, and 75th percentiles, respectively. Whiskers indicate the minimum and maximum values. The confidence intervals are described in table S6. RT, regression trees. (C) Performance comparison of DeepSpCas9 with other prediction models using dataset Endo_Cas9 (n = 124 independent target sites) and two published datasets (n = 4207 and 2060 independent target sites for datasets Hart 2015 and Xu 2015, respectively) as the test datasets. Error bars represent 95% confidence intervals, which are described in detail in table S6. For clarity, results from statistical testing are shown only for DeepSpCas9 versus deep learning with an equal-sized filter, DeepSpCas9 versus the best conventional machine learning–based model, and deep learning with an equal-sized filter versus the best conventional machine learning–based model (left to right: *P = 1.4 × 10^−2, DeepSpCas9 versus deep learning with an equal-sized filter; *P = 1.1 × 10^−2, DeepSpCas9 versus SVM; *P = 4.6 × 10−2, deep learning with an equal-sized filter versus SVM; Steiger's test). ns, not significant. (D) Performance comparison of DeepSpCas9 and DeepSpCas9-CA (chromatin accessibility). The DeepSpCas9-CA model was developed by fine-tuning the DeepSpCas9 model using the Endo-1A dataset. DeepSpCas9 (left) and DeepSpCas9-CA (right) models were evaluated with the Endo-1B dataset. The Spearman correlation coefficients (R) are shown. (E) Results from 10 iterations of fine-tuning and evaluation. Each dot represents the Spearman correlation coefficient between the measured indel frequency and the predicted activity. A total of 10 (= 2 × 5) rounds of fine-tuning and subsequent testing results are shown. Credit: *Science Advances*, doi: 10.1126/sciadv.aax9249

In previous work, Kim et al. considered chromatin accessibility information to improve the prediction of AsCpf1 enzyme activities at endogenous target sites. They sought to determine if such considerations

would also improve SpCas9 activity predictions. The results implied that fine-tuning with chromatin accessibility information barely improved the accuracy of DeepSpCas9 to predict indel frequencies at endogenous sites compared to their previous efforts with AsCpf1. The SpCas9 activity was only therefore slightly affected by chromatin accessibility in strong contrast to the previously developed DeepCpf1 algorithm.

To understand the generalization performance of DeepSpCas9, the research team tested the model using sufficiently large, published datasets derived from diverse research studies as test data. They compared the results with those of other SpCas9 activity predicting programs such as DeepCRISPR. The results suggested DeepSpCas9 to maintain the highest generalization function among nine published models used to predict SpCas9 activity. In this way, Hui Kwon Kim and research team extensively validated the potential to accurately predict SpCas9 activity using the DeepSpCas9 web tool, now available online, alongside supplementary code provided for research scientists to incorporate DeepSpCas9 into existing models. Based on the high generalization performance of DeepSpCas9, the research team expect to facilitate higher accuracy for SpCas9-based genome editing.

More information: Hui Kwon Kim et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance, *Science Advances* (2019). DOI: 10.1126/sciadv.aax9249

Hui Kwon Kim et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity, *Nature Biotechnology* (2018). DOI: 10.1038/nbt.4061

Hui K Kim et al. In vivo high-throughput profiling of CRISPR–Cpf1 activity, *Nature Methods* (2016). DOI: 10.1038/nmeth.4104