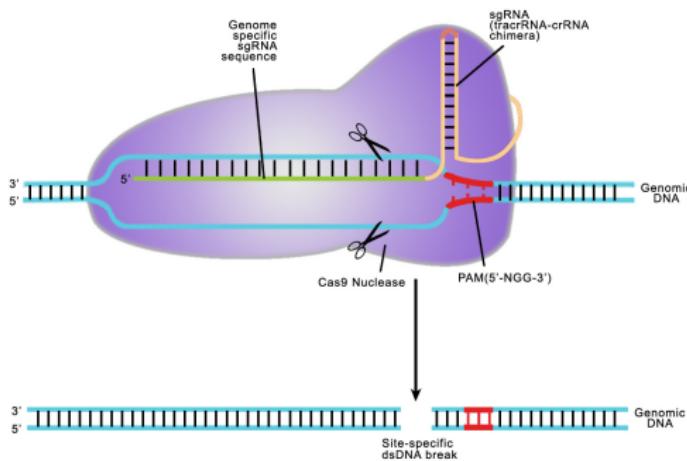




PAVOOC - Prediction and visualization of on- and off-targets for CRISPR

Moritz Schäfer | Technische Universität Berlin & Bayer Pharma | Master thesis

Background



- Knockout experiments used in drug target validation
- Sequence (partially) determines efficacy



Problem

- Guide prediction scores still vary in performance



Problem

- Guide prediction scores still vary in performance
- A lot of manual labor for guide selection

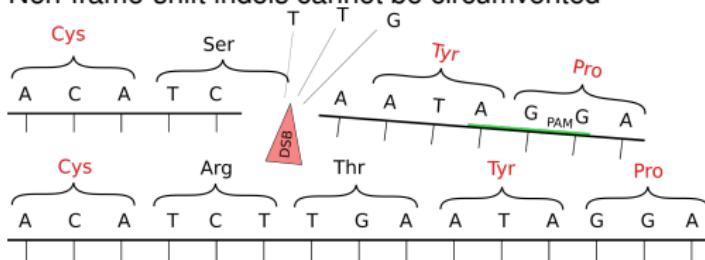
19	PULRZL	28,5	28	-3,00/5211818282
20	SPC24	29,5	42	-2,83403119872683
21	THAP1	31,5	27	-3,01155447143932
22	CDC123	32,5	44	-2,81699463079959
23	WDD74	37	68	2,52033320817648

Problem

- Guide prediction scores still vary in performance
- A lot of manual labor for guide selection

19	PULRZL	28,5	28	-3,00/5211818282
20	SPC24	29,5	42	-2,83403119872683
21	THAP1	31,5	27	-3,01155447143932
22	CDC123	32,5	44	-2,81699463079959
23	WDND7A	27	68	2,58033320817648

- Non-frame-shift indels cannot be circumvented

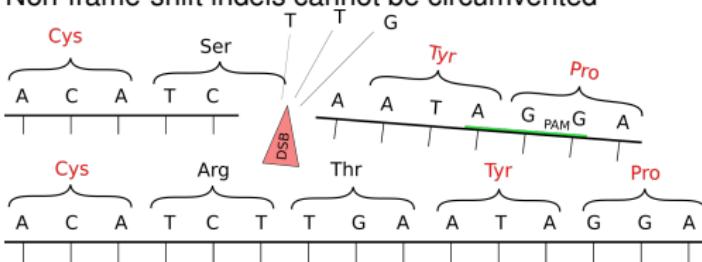


Problem

- Guide prediction scores still vary in performance
- A lot of manual labor for guide selection

19	PULRZL	28,5	28	-3,00/5211818282
20	SPC24	29,5	42	-2,83403119872683
21	THAP1	31,5	27	-3,01155447143932
22	CDC123	32,5	44	-2,81699463079959
23	WDND7A	27	68	2,58033320817648

- Non-frame-shift-indels cannot be circumvented



- Cancer cellline data affects certain guides
- AGCATCGTAAGT GAATTACGG
- + PAM

DMS153 lung cancer cellline SNP



Solution

- Cutting-edge guide efficacy scoring
- All-in-one guide design tool
- Web-based
- Functional domain-aware
- Incorporate cancer cellline data

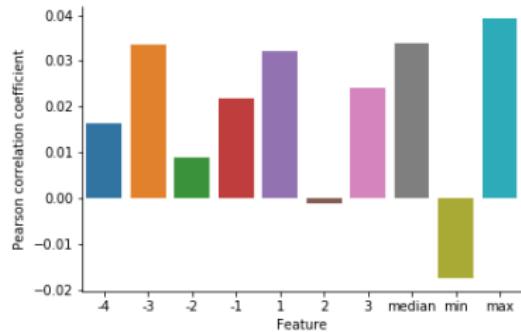
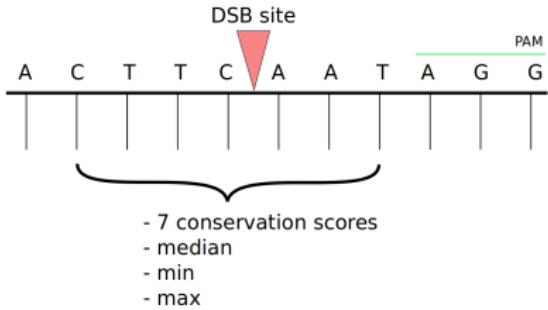
Guide efficacy prediction – Dataset

Guide	Measured efficacy
GTTAGGGTCCGTACTCAGCAAGG	0.86
ACACTGCCGAGCGATGAGGATGG	0.42
AAGGTGAAGGAGGATGCGGCGGG	0.53
GAAAAGATAAGGTCACTGACCCGG	0.12
GCAAGTCACTGAGTGCAGAACGG	0.73
GCATTGGTAAGCGCACAGGAAGG	0.70
AAGACTGGCGCATGGTCCACTGG	0.57
...	...

- 5310 data rows
- Efficacy relates to cell proliferation after CRISPR application

"Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9", 2016, John G. Doench et al.

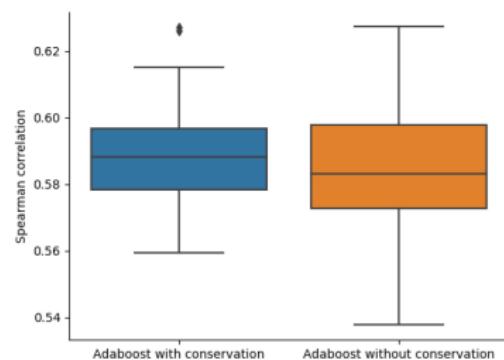
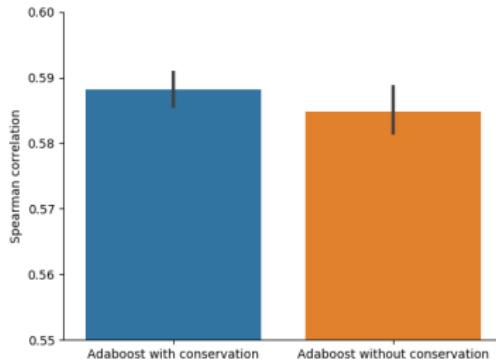
Conservation features



$$p_{max} = 0.0043$$

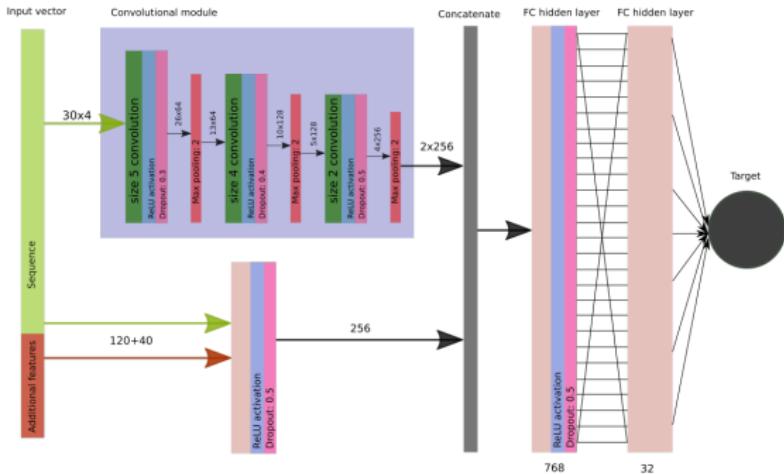
Conservation feature results

Comparison of 100 runs on adaboost



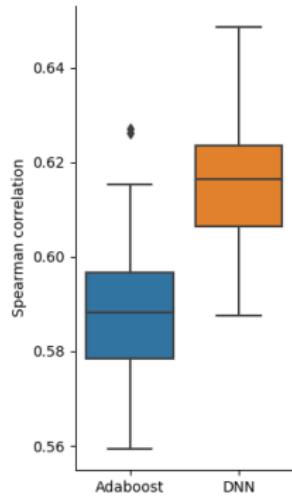
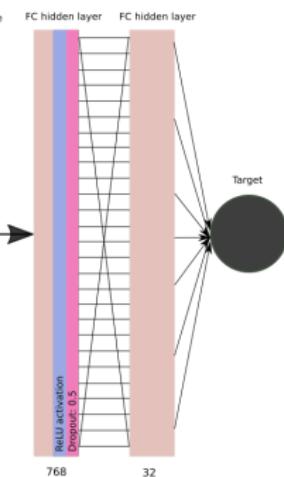
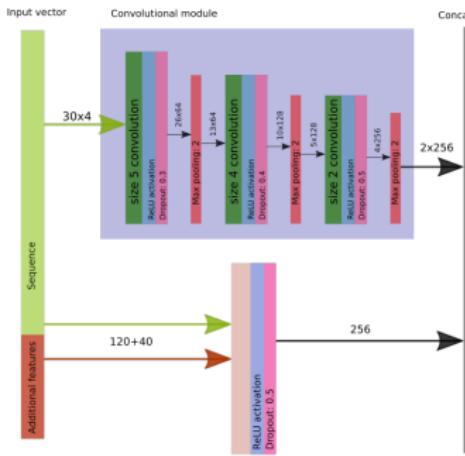
Deep Learning architecture

"Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning", 2015, B. Alipanahi et al.



Deep Learning architecture

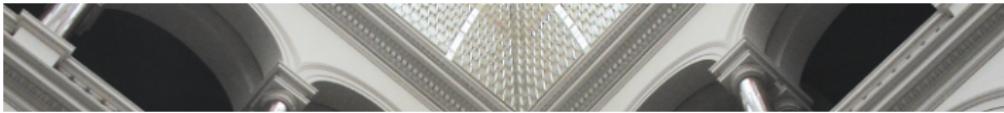
"Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning", 2015, B. Alipanahi et al.



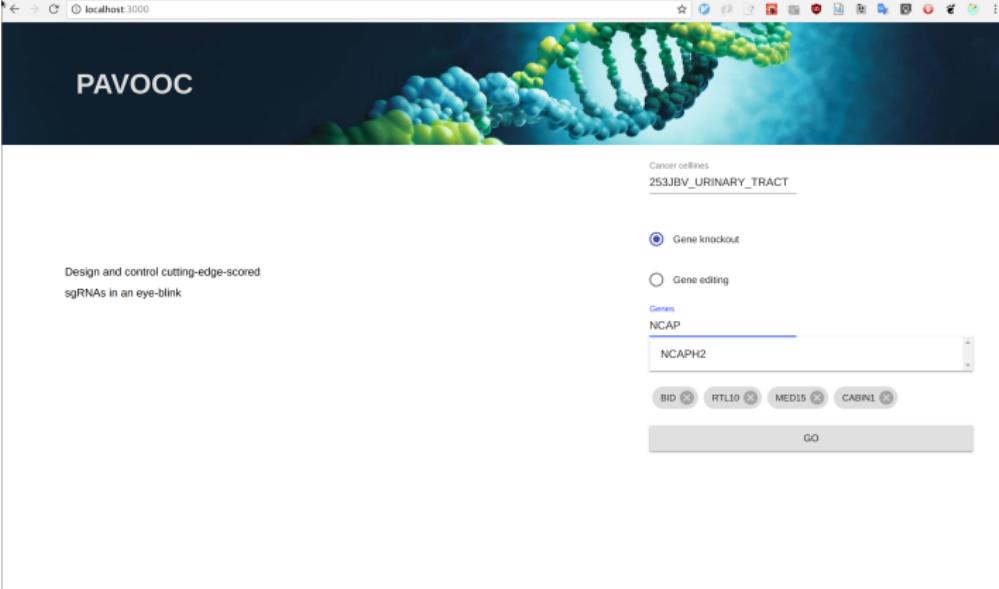
Conclusion: Deep Learning can improve guide efficacy prediction

Architecture





Application overview



The screenshot shows the PAVOOC application interface running on a local host at port 3000. The title "PAVOOC" is visible on the left. The background features a 3D rendering of a DNA double helix.

Cancer cell lines:
253JBV_URINARY_TRACT

Gene knockout
 Gene editing

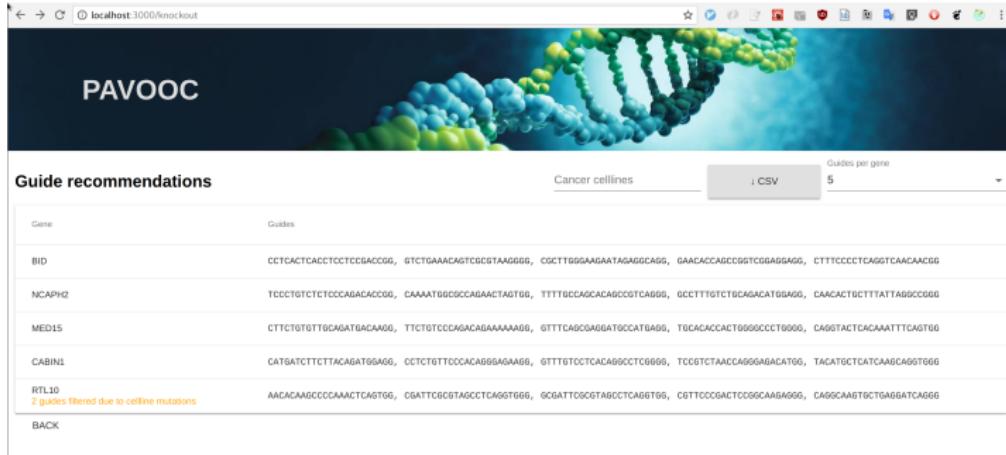
Genes:
NCAP
NCAPH2

BID RTL10 MED15 CABIN1

GO



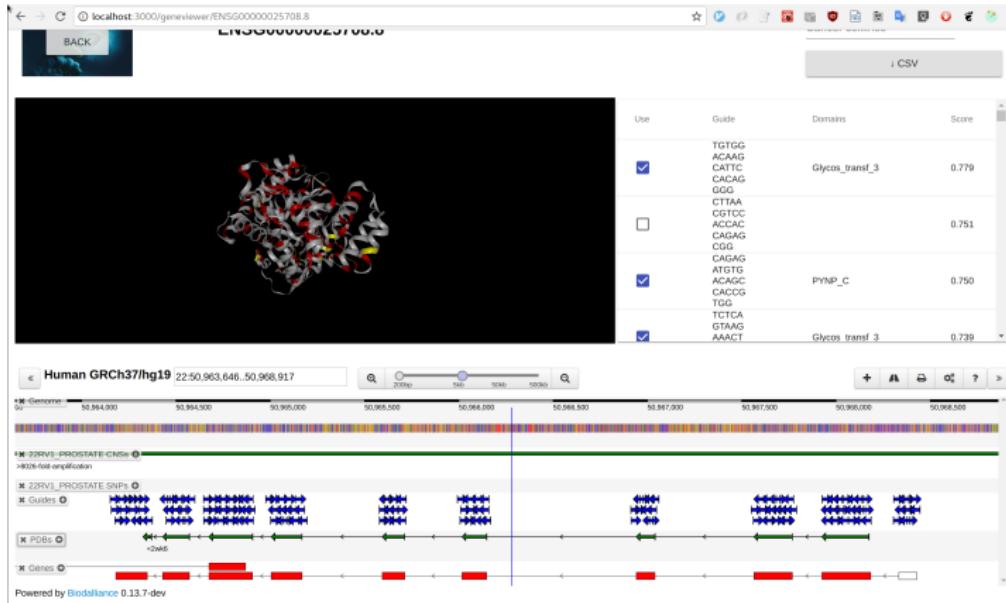
Application overview



The screenshot shows a web-based application titled "PAVOOC". The main header features the word "PAVOOC" in white on a dark blue background with a 3D DNA model. Below the header, there's a search bar with the URL "localhost:3000/knockout". The main content area is titled "Guide recommendations". It includes a table with columns for "Gene" and "Guides". The table lists several genes with their corresponding CRISPR guide sequences. A dropdown menu "Guides per gene" is set to 5. There are also tabs for "Cancer cellines" and a "CSV" download button. At the bottom left, there's a note about filtering guides due to celline mutations, and a "BACK" button.

Gene	Guides
BID	CCTCACTCACCTCCCTCCGACCGG, GTCGTAAACAGTCGCGTAAGGG, CCCTTGGGAAAGAATAGAGGCAGG, GAAACACCAGCCGGTCGGAGGAGG, CTTTCCCCTAGGGTCAAACAACGG
NCAPH2	TCCCTGTCTCTCCCAGAACCCGG, CAAAATBGCAGGAAACTAGTGG, TTTTBCCAGCACAGCGTCAGGG, GCCTTGTCTBAGAGCATGGAGG, CAACACACTGTTTATTAGGGCCGGG
MED15	CTTCTGTGTTGGAGATGACAAGG, TTCTGTCCCCAGACABAAGGGAGG, GTTTCAAGCGAGGAGTGGCATBAGG, TGCACACCCACTGGGGCCCTGGGG, CAAGGTACTCACAAATTTCAGGG
CABIN1	GATGATCTCTTACAGATGGAGG, CCTCTGTCCCCACAGGGAGAAGG, GTTTGTCTCACAGGGCCCTGGGG, TCCGCTTAACCAAGGGAGACATGG, TAGATGCTCATCAAAGGGTGGG
RTL10	AACACAAAGCCCCAAACTCAAGTGG, CGATTTCGCGTAAGCTCAAGGTGGG, CGGATTGGCGTAAGCTCAAGGTGG, CGTTCGCCAACTCCGGCAAGGG, DAAGGCAAGTGGCTGAAGGGATCAAGGG
2 guides filtered due to celline mutations	

Application overview



The screenshot displays the PAVOOC application interface. At the top, there is a navigation bar with a back button and a search field containing "ENSG00000025708.8". To the right of the search field are various icons for file operations like CSV export and zoom.

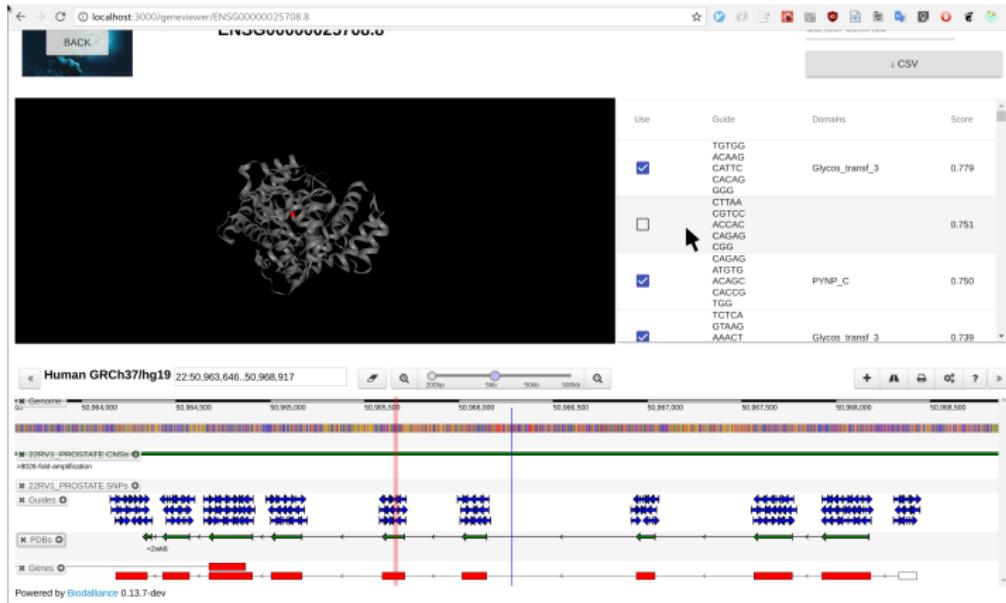
The main content area features a 3D ribbon diagram of a protein structure, primarily white with some red and yellow residues. Below the protein is a genomic track for Human GRCh37/hg19, showing a zoomed-in region from 50,964,000 to 50,968,500. The track includes a scale bar from 200kb to 10kb, a search bar, and a zoom tool.

To the right of the genomic track is a table titled "Use", listing several guide RNAs (gRNAs) along with their target domains and scores:

Use	Guide	Domains	Score
<input checked="" type="checkbox"/>	TGTGG ACAGG CATTC CACAG CGG CTTAA CGTCC ACCAAC CAGAG CGG CAGAG ATGTG ACAGC CACCG TGG TCTCTCA GTAAAG AAACT	Glycos_transf_3	0.779
<input type="checkbox"/>			0.751
<input checked="" type="checkbox"/>	PYNP_C		0.750
<input checked="" type="checkbox"/>		Glycos_transf_3	0.739

Below the genomic track, there are several panels: "22RNV1-PROSTATE-CNSe", "22RNV1-PROSTATE SNPs", "Guides", "PDBs", and "Genes". The "Guides" panel shows a grid of blue arrows indicating target sites across the genomic region. The "Genes" panel shows red boxes representing genes. At the bottom left, it says "Powered by Bioldalliance 0.13.7-dev".

Application overview



The screenshot displays the PAVOOC application interface, which integrates protein visualization, genomic data, and target sequence analysis.

Protein Structure: A large ribbon diagram of a protein structure is shown against a black background. A small red dot highlights a specific residue or region of interest.

Genomic View: Below the protein structure, a genomic track for Human GRCh37/hg19 is displayed. The track shows a zoomed-in region from 50,964,900 to 50,968,500. The track includes:

- Annotations:** A green line labeled "22RNU1-PROSTATE-CNSe" and a blue line labeled "22RNU1-PROSTATE-SNPs".
- SNPs:** A grid of blue arrows indicating SNP locations and orientations.
- Genes:** Red boxes representing gene structures.

Target Sequences: On the right, a table lists target sequences with their corresponding scores:

Use	Guide	Domains	Score
<input checked="" type="checkbox"/>	TGTGG ACGAG CATTG CACAG GCG CTTAA CGTCC ACCAAC CAGAG CGG ATGAG ATGTG ACAGC CACCG TGG TCTCA GTAAG AAACT	Glycos_transf_3	0.779
<input type="checkbox"/>			0.751
<input checked="" type="checkbox"/>	PYNP_C		0.750
<input checked="" type="checkbox"/>		Glycos_transf_3	0.739

Powered by Bioldalliance 0.13.7-dev



Acknowledgements

- Dr. Andreas Steffen (supervisor at Bayer)
- Prof. Manfred Opper (supervisor at TU Berlin)
- Dr. Djork-Arne Clevert (machine learning scientist at Bayer)
- Robin Winter (PhD student at Bayer)
- Bayer Pharma AG