

[Artificia](https://doi.org/10.1016/j.ailsci.2023.100075)l[IntelligenceintheLifeSciences3(2023)100075](https://doi.org/10.1016/j.ailsci.2023.100075)

|  |  |  |
| --- | --- | --- |
|  | Contentslistsavailableat[ScienceDirect](http://www.ScienceDirect.com) |  |
| ArtificialIntelligenceintheLifeSciences |
| journalhomepage:[www.elsevier.com/locate/ailsci](http://www.elsevier.com/locate/ailsci) |

ResearchArticle   
piCRISPR:PhysicallyinformeddeeplearningmodelsforCRISPR/Cas9 off-targetcleavageprediction   
FlorianStörtz,Jeffrey K.Mak,PeterMinary∗  
*DepartmentofComputerScience,UniversityofOxford,ParksRoad,OxfordOX13QD,UK*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  CRISPR  Cas9  Deeplearning  Cleavageprediction  Nucleosomeorganisation | | CRISPR/Casprogrammablenucleasesystemshavebecomeubiquitousinthefieldofgeneediting.Withpro-gressingdevelopment,applicationsin*invivo*therapeuticgeneeditingareincreasinglywithinreach,yetlimited bypossibleadversesideeffectsfromunwantededits.Recentyearshavethusseencontinuousdevelopmentof off-targetpredictionalgorithmstrainedon*invitro*cleavageassaydatagainedfromimmortalisedcelllines.It hasbeenshownthatincontrasttoexperimentalepigeneticfeatures,computedphysicallyinformedfeaturesare sofarunderutiliseddespitebearingconsiderablylargercorrelationwithcleavageactivity.Here,weimplement state-of-the-artdeeplearningalgorithmsandfeatureencodingsforoff-targetpredictionwithemphasison*physi-callyinformed*featuresthatcapturethebiologicalenvironmentofthecleavagesite,hencetermingourapproach piCRISPR.Featuresweregainedfromthelarge,diversecrisprSQLoff-targetcleavagedataset.Wefindthatour best-performingmodelshighlighttheimportanceofsequencecontextandchromatinaccessibilityforcleavage predictionandcomparefavourablywithliteraturestandardpredictionperformance.Wefurthershowthatour novel,environmentallysensitivefeaturesarecrucialtoaccuratepredictiononsequence-identicallocuspairs, makingthemhighlyrelevantforclinicalguidedesign.Thesourcecodeandtrainedmodelscanbefoundready touseat[github.com/florianst/picrisp](https://github.com/florianst/picrispr)r. |

**Introduction**

Theclusteredregularlyinterspacedshortpalindromicrepeats (CRISPR)sequencefamilywasfirstdescribedin*E.coli*in1987[1],but ittookuntil2007torecogniseitasapartoftheviraldefensesystemof mostarchaeaandbacteria[2].ExogenousviralDNAiscleavedoff by specialisednucleaseenzymes,codedforongenomicregionswhichare oftenadjacenttoCRISPRandhencenamedCRISPR-associated(Cas). Cleaved-off regionsaresubsequentlyincorporatedintotheCRISPRse-quences,whichactasaviralhistoryoftherespectivecell,stabilised bythepalindromicnatureoftheirsavedstateswhichresultsinstable secondarystructures[3].FromtheretheycanbetranscribedtocrRNA andinvadingcopiesofthemcansubsequentlyberenderedinactive.Re-searchershaveusedthisabilityforprogrammablegenomeeditingin manyeukaryoticspecies,complementingstrategiessuchaszinc-finger nucleases(ZFNs,[4])andtranscriptionactivator-likeeffectornucleases (TALENs,[5]).

Weconcentrateontheeffectsofthewild-typeCas9proteingained from*Streptococcuspyogenes*.ThecrRNAwhichisoriginallyresponsible forrecognitionofa20bpviralsequenceformsanactivecomplexwith thetracrRNA,calledsingleguideRNA(sgRNA),ofabout50bplength [6].HomologyofthecrRNApartwitha20bpregioninthegenomere-

∗Corresponding[author.](mailto:peter.minary@cs.ox.ac.uk)

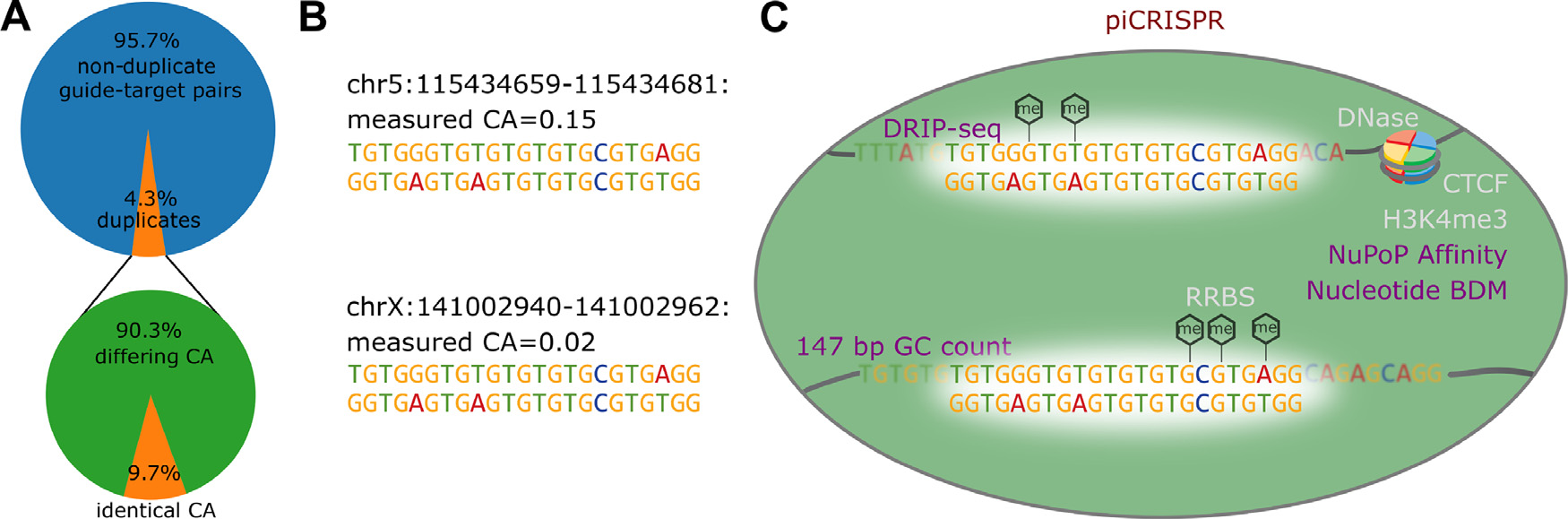
*E-mailaddress:*[peter.minary@cs.ox.ac.uk](mailto:peter.minary@cs.ox.ac.uk)(P.Minary).

sultsinannealingofthesgRNAwithonestrandofthisregion,which wecall‘targetstrand’.Bindinghappenswhentheinteractionofthe3bp protospacer-adjacentmotif(PAM)ontheopposite,non-targetstrand withtheCas9proteinisfavourable[7].For*S.pyogenes*Cas9,thisis thecaseforan’NGG’PAMwhereNstandsforanarbitrarynucleobase (A,T,C,G).

TertiaryDNAstructure,suchasnucleosomeoctamers,canoccludeor exposedifferentregionsofDNAandhinderCas9access[8].Afterbind-inghastakenplace,nuclease-activeenzymeswithinCas9cancleavethe double-strandedDNA3bpupstreamofthePAM.Duetothestochastic, energy-drivennatureofboththebindingandthecleavageprocess,we expectadistributionofcutsoverthewholegenome,includingundesired off-targeteffectswhichcouldpossiblyhavecatastrophicconsequences, suchasknockingouttumorsuppressorgeneslikep53andRb[9]. Wenoticedthatrepositoriesofoff-targetcleavagedatacontaina significantamountofdatapointswhichmatchinbothguideand(off-)targetsequenceanddifferonlyinthebiologicalenvironmentofthe respectiveloci(seeFig.1).Capturingthisenvironmentisthereforein-strumentalinprovidingaccuratepredictionsofcleavageactivity.We recentlyfoundthatcomputednucleosomeorganisation-relatedfeatures correlatebetterwithcleavagefrequencyvaluesthanexperimentalepi-geneticmarkers(Deoxyribonuclease-Ihypersensitivesitessequencing

<https://doi.org/10.1016/j.ailsci.2023.100075>  
[Received29December2022;Receivedinrevi](https://doi.org/10.1016/j.ailsci.2023.100075)sedform2April2023;Accepted30April2023   
Availableonline15May2023   
[2667-3185/© 2023TheAuthors.PublishedbyElsevi](http://creativecommons.org/licenses/by-nc-nd/4.0/)erB.V.ThisisanopenaccessarticleundertheCCBY-NC-NDlicense (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

*F.Störtz,J.K.MakandP.Minary*  *ArtificialIntelligenceintheLifeSciences3(2023)100075*



**Fig.1.A**ThecrisprSQLdatasetcontainsanappreciableamountofperfectguide-targetduplicates.Weonlyconsiderdatagainedfromhumancelllinesandputative off-targetswhichwegeneratedbasedonsequencesimilarity.8922of230,274datapointshaveatleastoneguide-targetduplicatewithinthissetwhichdiffersin cleavageactivity(CA).Astheexamplefromourdatasetinpanel**B**shows,suchapairlooksidenticaltopurelysequence-basedpredictionalgorithms.Theymight thereforenotpredictdangerousoff-targeteffects.**C**piCRISPRremediesthisbytakingintoaccountthebiologicalenvironmentofthecleavagesitebasedonarange offeaturesbeyondguideand(off-)targetsequence.Sofar,predictionalgorithmshaveusedfeaturesrelatedtochromatinorganisation(CTCF,[20]),chromatin accessibility(DNase-Seq),DNAmethylation(RRBS)andhistonemethylation(H3K4me3,[21]).Weontheotherhandusefeaturespertainingtothe147bpsequence contextaroundeach(off-)targetnucleotide:GCcount,sequencecomplexity(BDM,[19])andnucleosomepositioninginformation(NuPoP,[8])whichintroduce unprecedentedsensitivitytothebiologicalenvironmentofthecleavagesite.Usingthese,piCRISPRcancorrectlyrankthetwoexamplelocigivenhere.

/DNase,reducedrepresentationbisulfitesequencing/RRBS,CCCTC-bindingfactor/CTCF,histone-3lysine-4trimethylation/H3K4me3) [10]whichhaveheretoforebeentheliteraturestandardforcleavage predictionmodels.Thesecomputedfeaturesalsosurpassedepigenetic markersintermsoftheirfeatureimportanceinpreliminarycleavage activitypredictionmodelswhichhadaccesstobothcomputedandex-perimentalepigeneticfeatures.Wethereforeaimtomakefulluseofthis novelclassoffeaturesbyembeddingtheminarichfeatureset,including DNA/RNAsequenceandcontextsequence-basedfeatures.

Withaconsiderableamountofcleavagepredictionalgorithms presentinliterature[11–15],wepresenthereachoiceoftwomodel architectures,twoencodingsandtwosetsoffeatures,yieldingatotal ofeightcombinations.Wescrutinisetheseaccordingtobothprediction performanceandinterpretability.Besidesimprovingpredictionaccu-racyandcapturingoff-targeteffectsthatmightsofarhavegoneunno-ticed,thiswillalsogenerateinsightintothebiologicalenvironmentthat influencesCRISPRcleavage.

**Methods**

*DataSource*

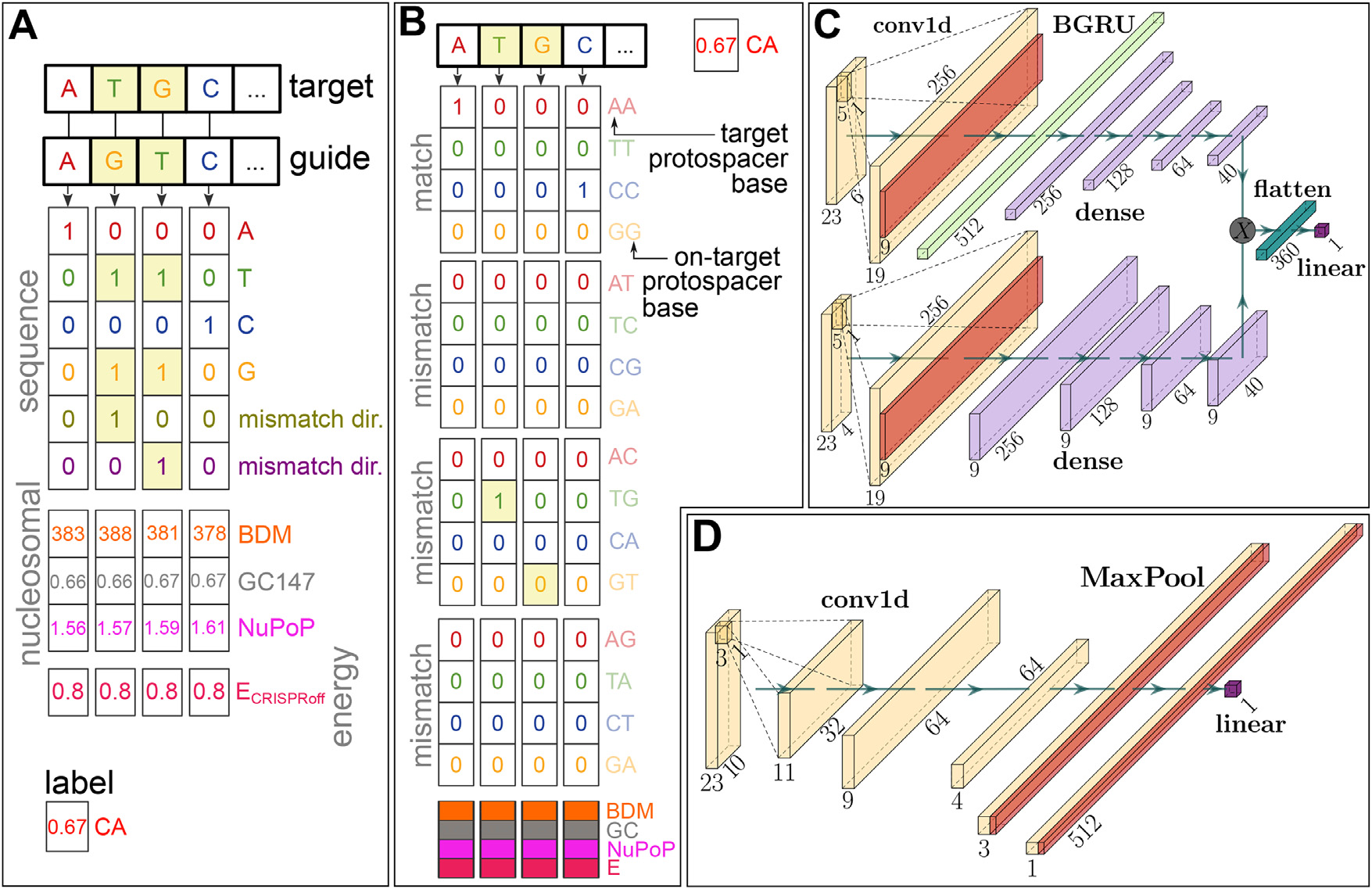
Inordertoachievemaximumtransparencyandcomparability,we useguide-targetpairsfromthecrisprSQLdataset[16]curatedbyour group.Itisacollectionof17base-pairresolvedoff-targetcleavagestud-iesonSp-Cas9,comprising25,632datapointsandislargerthanmost datasetsusedtotrainpredictionalgorithmstodate.Itcontainsdataon variouscelllines,mainlyU2OS,HEK293andK562.Wehavechosento useversion26/05/2020ofthedatabasewhichdoesnotincludeT-cell datafromLazzarotto*etal.*[17]inordertoavoidintroducingaconsider-ablecelllineimbalance.Furthermore,theevaluationofourmodelling on-targetdatasetsisbeyondthescopeofthisworkduetotheirdifferent underlyingexperimentaltechniquesandcleavagequantificationmea-sures.

Experimentaldatapointscontainingguideandtargetloci,sequence, cellline,assaytypeandcleavagefrequencyhavebeencompletedand enrichedbysequencecontextaswellastheCRISPRoffscore,anem-piricalestimateofthe(off-)targetbindingenergy[18].

Besidestheseestablishedfeatures,weproposetheusageofnucle-osomeorganisation-relatedfeatures[10]whichaddanunprecedented levelofsensitivitytowardsthebiologicalenvironmentofthecleavage site(seeFig.1C).Inthatpublication,wetrainedapreliminarycleavage predictionmodelon13distinctnucleosomeorganisation-relatedscores

2

*F.Störtz,J.K.MakandP.Minary*  *ArtificialIntelligenceintheLifeSciences3(2023)100075*



**Fig.2.**Overviewofourtwofeatureencodingsandmodelarchitectures.**A**6× 23encoding:(Off-)targetandguide(on-targetprotospacer)sequencesareone-hot encodedandcopiedtogetherusingabitwiseORoperation.Inordertomakethisencodinglossless,twochannelsareaddedthatencodewhichnucleotideisonthe guideandwhichonthetarget,respectively(inthecaseofamismatchedinterface).Thisisidenticaltotheencodingusedin[13].Thenucleosomepositioningchannels (147bpGCcount,NucleotideBDM[19],NuPoPAffinity[8])ofthetargetaswellastheCRISPRofffreeenergyestimatescore[18,27]areconcatenatedtothis matrix.TheBox–Coxtransformedcleavageactivity(CA)isusedasalabel.**B**The16× 23encodingusesa16-letteralphabetwhichexplicitlycontainsinformation abouttheprecisenatureofthemismatch.**C**BidirectionalgatedrecurrentunitarchitectureasusedinourRNNmodel,modelledafterthenetworkin[28].Upperand lowerarmofthenetworkcontainthesequenceandnucleosomal/energyinformation,respectively.**D**ConvolutionalneuralnetworkarchitectureusedinourCNN model,comparablewiththemodelin[11].Dimensionsinbothmodelarchitecturesarevalidforthe6× 23encoding(panelA).

*FeatureEncoding*  *ModelArchitectures*

Weemploytwodifferentfeatureencodingschemeswhichoccupy differentpointsinthetradeoff betweensparsityandinterpretability. Thefirstwasintroducedin[13]andconsistsofone-hotencodedrepre-sentationsoftheguideandtargetsequencewhichhavebeencombined usingabitwiseORoperation.Inordertomakeupforthelossofinfor-mationthatthisoperationcausesintermsofmismatches,twoadditional channelsareaddedcontaininginformationaboutthedirectionalityof thebasesinvolvedinthemismatch,i.e.whichofthetwoentriesde-scribesthetargetandwhichtheguidenucleotide.Notethattheguide nucleotideisfirsttranslatedintoitscorrespondingtargetprotospacer nucleotide.Wecallthisencodingthe6× 23encodingbasedonthere-sultingshapeofthesequencematrix(seeFig.2A).

Basedontheenergy-drivennatureofbindingandcleavage,wehy-pothesisethatmismatchedinterfacesaffectbindinginatotallydiffer-entwaythanmatchedinterfaces.Thishassofarnotbeenrecognised indetailbyoff-targetpredictionalgorithms.Sincethe6× 23encoding containstheinformationabouttheprecisenatureofagiveninterface onlyimplicitly,wedecidedtoincludeafurtherencodingwhichdoesso explicitly.Thisusesaone-hotrepresentationusingthe16lettercross productbetweenguideand(off-)targetnucleotide,andishencetermed the16× 23encoding(seeFig.2B).Thisissimilartotheencodingscheme in[26].

Bothmatricesarethenconcatenatedwithamatrixofbase-pairre-solvednucleosomalfeatures,aswellastheCRISPRoffvalueofthe giventarget-guideinterfacerepeatedalongthesequenceaxis.Explor-inglatentrepresentationsofguideortargetisnotwithinthescopeof thiswork,giventhatitfurthercomplicatescomparisonbetweenmodels.

3

*F.Störtz,J.K.MakandP.Minary*

*ModelTraining&Evaluation*

Given the imbalance of validated/measured and non-validated/augmenteddatapoints,weemployabootstrappingstrategy assuggestedin[30],wheretrainingbatchesonaveragecontainequal numbersofbothclasses.Forregression(classification),earlystopping isbasedonthemeansquarederror(binarycross-entropy)lossonhalf ofthetestset,wheretheotherhalfisreservedforevaluation.

TheCNNmodelsaretrainedinthesameway,withhyperparameters ofbatchnorm\_momentum=0.01,Gaussiannoisewith*𝜇* =0*,𝜎* =0*.*01 andAdamlearningrate10−3.

TheRNNmodelsaretrainedfor100epochs,wherebatchesof 10,000pointsaresampledeachepochoutofaclass-balancedsampleof 50,000datapoints.Wereplicatethetransferlearningapproachtakenin [28]withadjustmentstoincreasetrainingstabilityandgeneralisation performanceasdetailedintheSupplementaryText.Dropoutprobability was0.2andtheAdamlearningratewas10−3.

*TestingScenario1:heldoutstudies*   
 Inthisscenarioweholdoutstudies[31–33]fromthetrainingset. Thesestudieshavenotbeenincludedinthetrainingsetforthestate-of-the-artoff-targetpredictionalgorithmCRISPR-Net[13],suchthatthey remainanindependenttestsettocompareCRISPR-NetandpiCRISPR sidebyside.Theinherentclassimbalanceinthistestsetis1:103.96. 22%oftheuniqueguideswithinthetrainingsethaveatleastonecorre-spondingguideinthetestsetwithfiveorfewermismatches,indicating asatisfactoryindependencebetweentrainingandtestset.

*TestingScenario2:literaturecomparison*   
 InthisscenarioweusetheCIRCLE-seq[34]datasetastheheldout testset,aswasdonein[15].Theexacttestsethasbeenreplicatedusing thecodeprovidedbytheauthors,suchthatcomparisonvaluescouldbe takenstraightfrompublication[15].Nucleosomalandempiricalenergy datawasfilledinusingthecrisprSQLdataset.

*TestingScenario3:setofduplicatepairs*   
 Inthisscenario,wescrutiniseourhypothesisthatanenvironmen-tallysensitivefeaturesetisfittonotonlyincreasepredictionperfor-manceoverall,butespeciallyforgivengroupsofidenticalguide-target sequencepairs.Tothisendwecalculatetwoquantities:First,themean squarederror(MSE)betweenthepredictedregressionscoresandthe groundtruthcleavagefrequencieswithineachofthe2703groups.Sec-ond,theaverageproportionofthetruecleavageactivitydifferencefor twopointswithinagivengroupwhichthemodelpredicts.Thisiszero forpurelysequence-basedmodelsandunityforanidealpredictor.This

*ArtificialIntelligenceintheLifeSciences3(2023)100075*

quantifieshowfaithfulamodelistothedifferencesinbiologicalenvi-ronmentforagivenpair.Inordertoemphasisesmalldeviationswhich preservetherankofpredictedcleavageactivities,weusethecubicroot asasign-preservingnonlinearityandtermthisquantity*relativediffer-ence*.Weconsidertheresultingdistributionsofbothofthesequantities fordifferentfeaturesets.

*ModelExplanation*

Weobtainfeatureimportancesusingthemodel-agnosticShapleyAd-ditiveExplanations(SHAP)library[35].SincepiCRISPRwrapsthefea-tureencodinginsideagivenmodel,weretainfullexplainabilityofinput featuresevenfornon-invertibleencodings.Inthisway,usingthetwoen-codingsdetailedabove,weobtainanunprecedented,context-sensitive resolutionofsequence-basedfeatures.

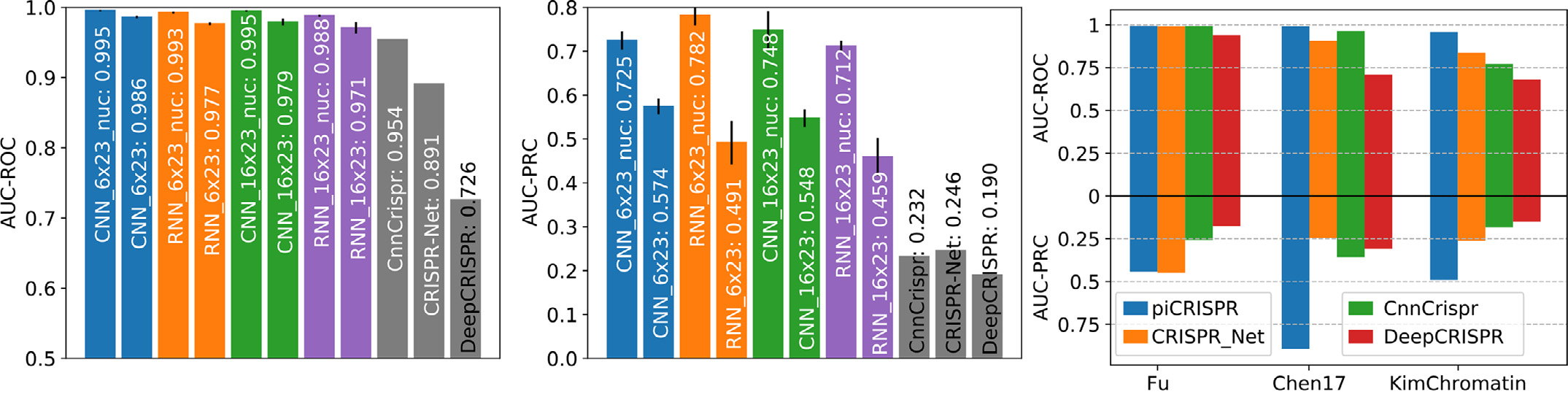
StickingwiththeconventionsetbytheSHAPlibrary[35],wecal-culateglobalSHAPvaluesasthemeanoftheabsolutevalueofSHAP valuesacrossdatapointsintheexplanationset,whichisarandomsub-setof500pointsfromtheheldouttestset.Inordertoshownotonly themagnitudebutthedirectioninwhichagivenfeatureinfluencesthe model’sprediction,wemultiplyeachfeature’sglobalSHAPvaluebythe signoftheaverageSHAPvalueofalldatapointswhosevalueislarger thanthemedianofthatfeature.

*Commandlineusageofourmodels*

Wehaveimplementedacommandlineinterfacewithwhichpi-CRISPRpredictionscanreadilybeobtained.Formaximumusability, themodelautomaticallyusesdefaultfeaturevaluesincaseacertain featurewasnotprovided,therebyloweringpredictionperformance(see FigureS4).Thedefaultvalueofagivenfeatureisdefinedastheaverage featurevalueofthesetofthosecrisprSQLdatapointswhichliewithin a20%intervalaroundthemeanSHAPvalue.Thismeansthathigh-accuracypiCRISPRpredictionscanbeobtainedinauser-friendlyway, evenwhenprovidingonlyguideand(off-)targetsequence.Ouronline repositorycontainshands-onexamplesonthis.

**Results&Discussion**

*TestingScenario1*   
 Fig.3showstheregressionandclassificationperformanceofour piCRISPR-implementedmodels,withthe6× 23RNNmodelyieldingthe highestbenchmarks.Asmentionedin[30],theareaunderprecision-recallcurve(AU-PRC)isamuchmoresuitablemeasurethanthearea underreceiveroperatingcurve(AU-ROC)foroff-targetprediction,since



**Fig.3.**ComparisonofpiCRISPRmodelswithpublishedalgorithms.Allmodelsweretestedonheldoutstudies[31–33](testingscenario1).Non-validateddata pointshavebeenoversampledinthetestsettomatchtheclassimbalanceof1:79.35foundinthedatasetI-1from[13].piCRISPRmodelshavebeentrainedonthe remainingdatapointswithinthecrisprSQLdataset.**Lefttwopanels**:Comparisonwiththreepublishedoff-targetpredictionalgorithms[11,13,26]thatwererunon thistestset.Withinamodelfamilyofthesamecolour,themodellabelled“nuc” containsnucleosomalfeatureswhereastheotherdoesnot.piCRISPRtrainingand testinghavebeenrepeated5timestoobtainmeanandstandarddeviationasshown.FortheunderlyingROCandPRCcurvesseeFigureS1.**Rightpanel**:AUC-ROC andAUC-PRCbenchmarksfortheRNN6× 23modelwithnucleosomalfeatures,resolvedbyindividualstudywithintheheldouttestset.

4

*F.Störtz,J.K.MakandP.Minary*

inclinicalapplication,falsenegativeshavefarmoreadverseeffectsthan falsepositives.Theadditionofthenucleosomalfeaturesconsiderably improvesmodelperformanceaccordingtoallbenchmarks,supporting ourhypothesisthatnucleosomalfeaturescanserveasakeyingredient tocleavageprediction.

Adirectcomparisonwithpredictionresultsobtainedfromthepub-lishedversionsofCnnCrispr,CRISPR-NetandDeepCRISPRontheiden-ticalheldouttestsetshowsthatpiCRISPRachieveshigherclassification benchmarksintermsofareasunderROCandPRCcurveforallthreein-dividualstudiescontainedinthetestset,exceptforstudy[31]forwhich piCRISPRandCRISPR-Netachievecomparablebenchmarks.

*TestingScenario2*   
 Fig.6showsthatwhentestingontheCIRCLE-seqdataset[34],pi-CRISPRperformancedropsslightlyascomparedtotestingscenario1. EspeciallyRNNmodelsgeneraliseslightlyworsetothisstudy.Westill observethatnucleosomalfeaturesenhancetheperformanceofamodel, andpiCRISPRstilloutperformsbothCRISPR-IPandCRISPR-Netmodels accordingtoareaunderPRCcurve.

*TestingScenario3*   
 Table1showsthatthemodelperformance,measuredbythemean squarederrorofpredictionswithinagroupofdatapointsthatshare bothguideandtargetsequence,isconsiderablyimprovedbyintroduc-ingfeaturesbeyondsequenceinformation(leftcolumn).Theresulting distributionofMSEsisshowninFigureS9.

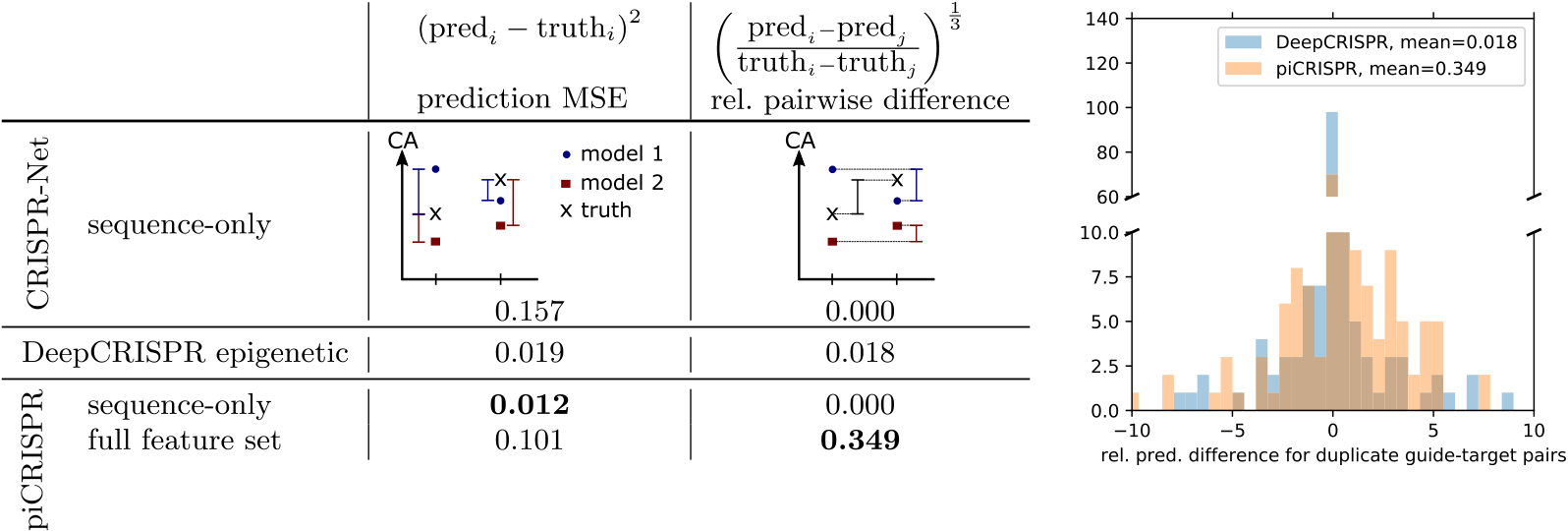
Lookingattherelativepairwisedifference,weobservethatintro-ducingfeaturesbeyondsequenceleadstoanincreaseoftheaverage proportionoftruecleavagefrequencydifferencesbetweenpointsofdif-feringbiologicalenvironmentwhichiscapturedbythemodel.Thisis trueforbothDeepCRISPRandpiCRISPR.Whilstthefullfeatureset inpiCRISPRachievesthehighestproportionincomparison,itisthe piCRISPRsequence-onlymodelthatachievesthelowestoverallmean squarederror.Thisindicatesthatalowmeansquarederrordoesnot necessarilygohandinhandwiththemodeldrawingthecorrectconclu-sionsfromenvironmentallysensitivefeatures.Thiscanbeseenaswell whenconsideringthecomparablysmallrelativepairwisedifferencethat isrecoveredbytheDeepCRISPRmodelfromtheliterature-standardepi-geneticschannelstowhichithasaccess.

*ArtificialIntelligenceintheLifeSciences3(2023)100075*

*Featureimportance*

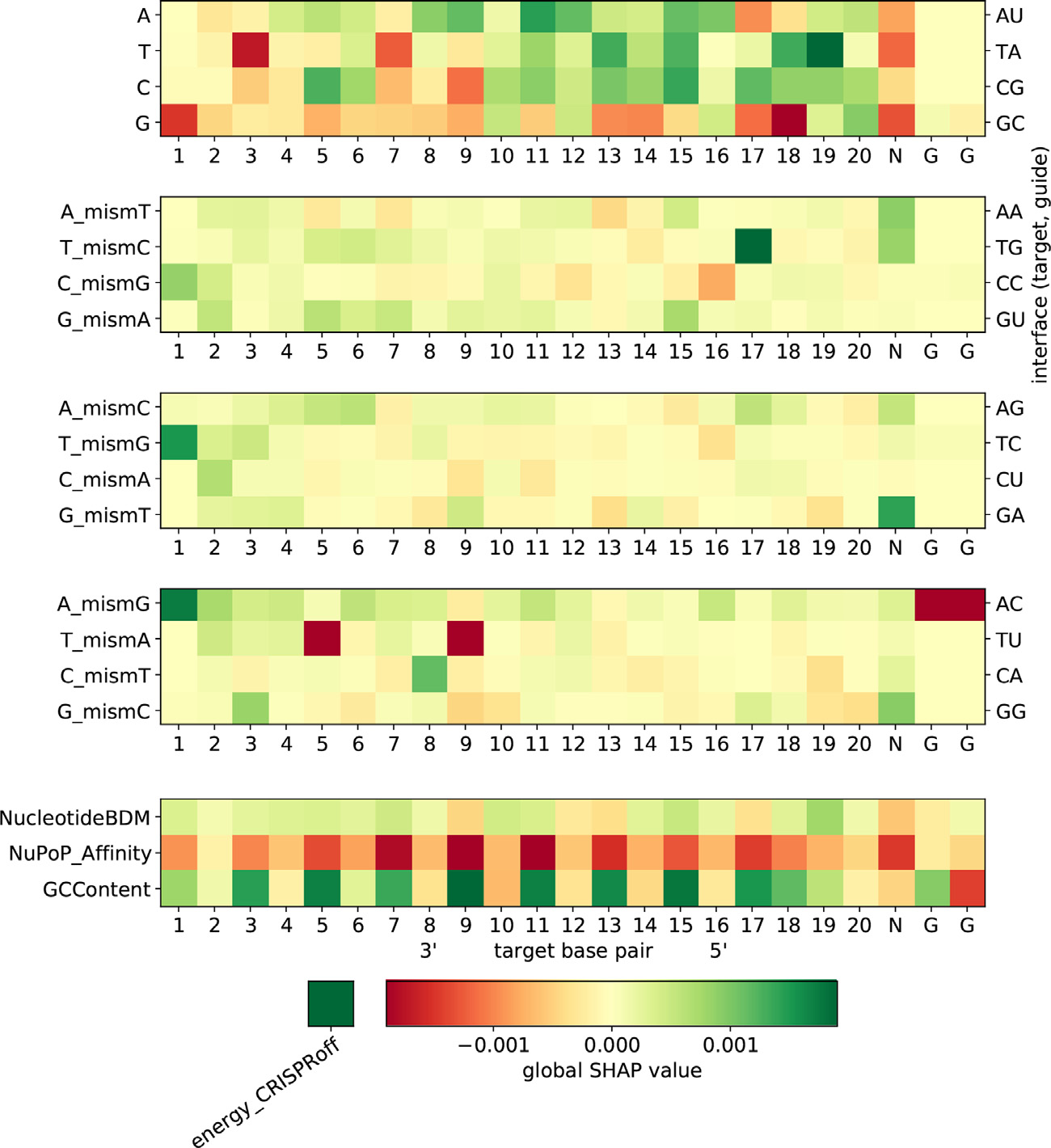
Duetoitscomparativelystrongerpredictionbenchmarksbetween testingscenarios1and2,weusethe16× 23CNNclassificationmodel intestingscenario1toextractfeatureimportancevaluesofunprece-dentedresolution.Fig.4showsthatthemodeldrawsonsequencefea-tureswhichstemfrommismatchedinterfacesdifferentlythanonthose frommatchedinterfaces,supportingourhypothesisthatthisdifferen-tiationisnotonlyphysicallyindicatedbutalsobackedbythemodel’s behaviour.GlobalSHAPvaluessuggestthatthepreferenceofthevari-ablePAMnucleotideatposition21iscontingentonthespecificsgRNA–DNAinterfaceformed.Werecoverthepreferenceforcytosineatposi-tion17[11,29,36]aswellasposition20[11,28,29]foundinliterature formatchedinterfaces.However,formismatchedinterfaces,cytosine isdisfavoured.Whilstwecannotrecoverastrongpreferenceforthe variablePAMnucleotideatposition21formatchedinterfaces,weob-servethepreferenceforguaninereportedinliterature[11,28,36]for mismatchedinterfaces.Thissupportsthenotionthataconcentrationon guide-targetinterfacesratherthanpurebaseidentitiesisnecessaryfor off-targetprediction,andthatdeeperinsightisrequiredthanthenotion ofapreferredbaseataspecificposition.Itthereforeappearsnecessary toconsidermismatchinterfacestogetherwithsequencesinthedesired genome,notjustthesequenceoftheputativeguide,forsgRNAdesign. Notethatduetothelowprevalenceofnon-NGGPAMsinourdataset, ashasbeenourchoicewhenaugmentingitwithputativeoff-targets,the modelattributeslittleimportancetothetwo5′ GGbasepairs.Weob-servetheblindspotofmismatchdiscriminationbytheREC3domainof Cas9aroundnucleotide7(seealsoFigureS5)whichhasbeenreported inarecentcryo-EMstructuralstudy[37]andresultsinreducedim-portanceofsequencefeaturespertainingmismatchedinterfacesinthis region.Atnucleotides3–5and9–11,wheremismatchdetectionbythe REC3domainofCas9ishigh,weobserveamismatch-inducedreduc-tionincleavageactivity.WefurtherobserveaPAM-distal’preference zone’andaPAM-proximal’avoidingzone’ofmismatcheswhenaver-agingoverfeatureimportancevaluesbynucleotide,whichhasbeen observedincomputational[11]aswellascryo-EM[37]studies. Themodeldrawsheavilyontheempiricalenergyestimatefeature *𝐸*CRISPRoff whichyieldsthelargestglobalSHAPvalue.Wefurtherob-serveaconsiderablecorrelationbetweenitsvalueandtheSHAPvalue attributedtoitbythemodel(FigureS8).Anenergyscoreof*𝐸*CRISPRoff =

**Table1**   
Benchmarkquantitiesgainedonthesubsetofduplicateguide-targetsequencepairs(testingscenario3)usingour6× 23CNNmodel aswellastheCRISPR-Net[13]andDeepCRISPR[11]modelsforcomparison.ForpiCRISPR,wealsogiveasequence-onlyversion ofthemodelinwhichnucleosomeorganisationrelatedfeaturesandtheempiricalenergyestimatehavebeensettoadefaultvalue acrossalldatapoints.**Leftcolumn**:meansquarederror(MSE)betweenpredictedcleavagescoreandgroundtruthcleavageactivity, averagedoverallgroupsofidenticalguide-targetsequencepairs.**Rightcolumn**:Howfaithfulamodelistothedifferencesin biologicalenvironmentforagivenpairwithinsuchagroupismeasuredbytheaverageproportionofthetruecleavageactivity differencewhichthemodelpredicts.Thisiszeroforpurelysequence-basedmodelsandunityforanidealpredictor.Toemphasise smalldeviationswhichpreservetherankofpredictedcleavageactivities,weusethecubicrootasasign-preservingnonlinearity andtermthisquantity*relativedifference*.**Rightpanel**:Exampledistributionsofrelativepairwisedifferenceforthetwomodels.All underlyingdistributionsareshowninFigureS9.



5

*F.Störtz,J.K.MakandP.Minary*



*ArtificialIntelligenceintheLifeSciences3(2023)100075*

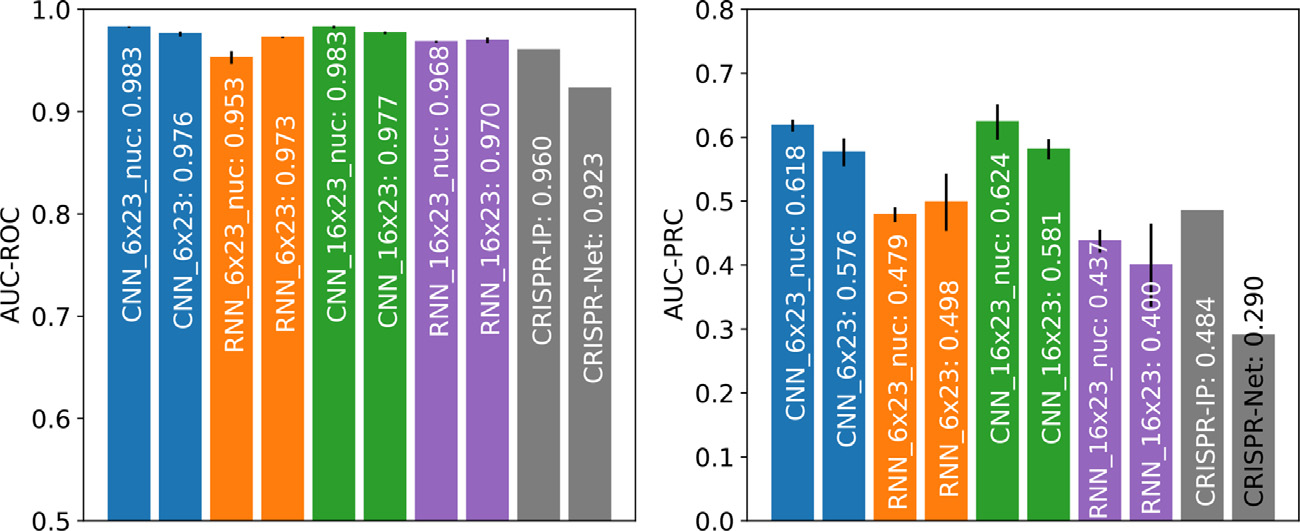
**Fig.4.**GlobalSHAPvaluesforthe16× 23CNNclassifica-tionmodel.NegativeglobalSHAPvalues(red)indicateanav-eragepredicteddecreaseinguideactivityfortherespective feature.Mismatchchannels(middlethreeheatmaps)canbe representedbythe(off-)targetandon-targetprotospacernu-cleotides(leftverticalaxis)aswellasthephysicalbasepair interfaces(rightverticalaxis),suchthatA\_mismTdescribes allconfigurationsinwhichanadenineonthetargetstrand facesanadenineonthesgRNA.Thebottomheatmapvisualises theinfluenceofourchosensetofnucleosomalorganisation featuresoncleavageactivity.Abarrepresentationofthiscan befoundinFigureS5.(Forinterpretationofthereferencesto colourinthisfigurelegend,thereaderisreferredtotheweb versionofthisarticle.)



**Fig.5.**Base-pairresolvedSHAPvaluesforthe16× 23(leftpanel,seeFig.4)and6× 23(rightpanel,seeFigureS7)CNNclassificationmodels.SHAPvalueshavebeen obtainedontheheldoutstudies[31–33]fromthecrisprSQLdataset.NotethathighvaluesoftheNuPoPAffinityfeature(reddots),i.e.highlypositionednucleosomes, alwaysinfluencethemodeltowardsreducedcleavageactivity(negativeSHAPvalue).(Forinterpretationofthereferencestocolourinthisfigurelegend,thereader isreferredtothewebversionofthisarticle.)

6

*F.Störtz,J.K.MakandP.Minary*



*ArtificialIntelligenceintheLifeSciences3(2023)100075*

**Fig.6.**ComparisonbetweenpiCRISPR,CRISPR-Net[13]and CRISPR-IP[15],withcomparisonvaluesforthelattertwo modelstakenfrom[15].Allmodelsweretestedonathesub-setoftheCIRCLE-seqdataset[34]asgivenin[15](testing scenario2).NotetheslightlyreducedperformanceoftheRNN modelscomparedtotestingscenario1(Fig.3).

|  |  |
| --- | --- |
| −1*.*15hasaneutralinfluenceoncleavageactivityinthe16× 23CNN model,withhigher(lower)valuesyieldinglarger(smaller)SHAPval-ues,i.e.apositive(negative)influenceoncleavageactivity.  Whenconsideringnucleosomepositioning-relatedfeaturechannels, weseethatthe147bpGCcontentaroundeachnucleotidehasanet positiveinfluenceoncleavageactivity.Similartotheargumentin[10], thiscanbeattributedtotheincreasedbendabilityofGC-richDNA [38]whichisbeneficialtoCas9sequencereadoutduringbinding[39]. WefurtherobservethattheNuPoPAffinityscorerankshigherinterms ofglobalSHAPvaluethanmostsequencefeatures.Thenegativeinflu-enceofnucleosomeaffinitycanbeexplainedbythereducedaccessi-bilityofhigh-affinityDNAregions,andisobservedstronglybetween nucleotides5and19.Thiseffecthasbeenobservedin[10]aswell.We furtherobserveanoverallnegativeinfluenceoflowNucleotideBDM valuesoncleavageactivity,supportingwhathasbeenobservedinpre-liminary,non-sequencebasedmodelsin[10].  Thisalsodemonstratestheimportanceofnucleosome-relatedfea-turesforcleavageprediction,andalsosupportsthenotionofchromatin accessibilityinfluencingcleavageactivityfoundin[40].Toourknowl-edge,thisstrongeffectofamorethan10bpwidesequencecontexton genome-wideoff-targetcleavagepredictionhasnotbeendemonstrated yet.Hintsofithavebeenseenonlyforsmallercontextsandon-target efficacyprediction[41,42].Inaddition,ourfindingspresentanunprece-dentedexampleinwhichinformationinthe147bpsequencecontext hasconsiderableimpactonthemodel.  Asimilaranalysisforthe6× 23CNNmodelcanbefoundinFigureS6 andforthe16× 23RNNmodelinFigureS7.Figure5showstheunder-lyingSHAPvaluesforbothCNNclassificationmodels.Notethatwithin thenucleosomalfeatureclass,theRNNmodelsattributemoreimpor-tancetotheNucleotideBDMfeaturethantheCNNmodelsscrutinised here.Thiscouldinpartexplaintheirslightdifferenceinperformance betweentestingscenarios1and2.  **Conclusion**  Throughtheintroductionofanewfeatureclassandthecareful adjustmentofmodelarchitectures,wehaveidentifiedasetofmod-elswhichmatchtheperformanceofstate-of-the-artoff-targetcleavage predictionalgorithmsindirectcomparison.Allmodelsarehighlyin-fluencedbynucleosomeorganisation-relatedfeaturessuchashistone bindingaffinity,whichemphasisestheimportanceofcapturingthebio-logicalenvironmentaroundthecleavagesitewhenmodellingcleavage activity.Ourapproachhasshownthatthesecomputedphysicallyin-formedfeaturesarefittoenhancethepredictivepowerofcleavagepre-dictionmodelsandtoreplaceexperimentalepigeneticmarkersinfuture modellingefforts.Wehavefurtherprovidedanaccessible,user-friendly commandlineinterfacethatallowsusersofvariousdisciplinestoutilise allourmodels,evenwithoutprovidingacompletesetoffeatures.This allpavesthewaytowardsthepredictionofoff-targetsiteswhichwould sofarhavegoneunnoticed. | Ourenvironmentallysensitivesetoffeaturesrevealsseveralnovel, promisingpathwaystowardsfurtherimprovementofoff-targetcleav-ageprediction.Goingforward,itcouldbefruitfultoincreasemodel complexity,e.g.usinga2Dconvolutionalkerneltocaptureinteraction betweenfeaturesofasinglenucleotide.A2Dconvolutionkernelwould beabletocapturethebase-pairresolvedinteractionbetweensequence andnucleosomalmarkersaswellasbetweensequencek-mers.Further thanthis,ourmultimodaldatacouldbefusedatdifferentstages,such thatsequence,nucleosomalandenergyfeaturesinteractatdifferentlev-elsofrepresentationofeachother.  Wefurtherenvisiontoreplacetheepigeneticinformationofthe guide,whichsofaronlycopiestheepigeneticinformationofthetar-getDNA.Thisisclearlyanunphysicalchoice.Giventhatasynthetic sgRNAdoesbydesignnotcarryepigeneticmarkers,aone-hotencoded dot-bracketrepresentationofthesgRNAfoldingwouldbeamoresuit-ablechoicetocaptureitsinformativeproperties.  **Funding**  ThisresearchwasfundedinwholeorinpartbytheBiotechnology andBiologicalSciencesResearchCouncil(BBSRC)[BB/M011224/1, BB/S507593/1].ForthepurposeofOpenAccess,theauthorhas appliedaCCBYpubliccopyrightlicencetoanyAuthorAc-ceptedManuscript(AAM)versionarisingfromthissubmission. SomeofthepresentedresultshavebeenobtainedusingtheUni-[versityofOxfordAdvancedResearc](https://doi.org/10.5281/zenodo.22558)hComputing(ARC)facility (<https://doi.org/10.5281/zenodo.22558>).Theauthorsdeclarenocon-[flic](https://doi.org/10.5281/zenodo.22558)t[ofinterest.](https://doi.org/10.5281/zenodo.22558)  **DeclarationofCompetingInterest**  Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.  **Dataavailability**  Thedataset,sourcecodeandtrainedmodelscanbefoundreadyto useatgithub.com/florianst/picrispr.  **Supplementarymaterial**  Supplementaryma[terialassociatedwiththisarti](https://doi.org/10.1016/j.ailsci.2023.100075)clecanbefound,in theonlineversion,at[10.1016/j.ailsci.2023.100075](https://doi.org/10.1016/j.ailsci.2023.100075)  **References**  [1][Ishino,etal.JBacteriol1987;169:5429.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0001)  [2][Horvath,etal.Science2010;327:167.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0003)  [3][Kunin,etal.GenomeBiol2007;8:R61.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0004)  [4][Urnov,etal.NatRevGenet2010;11:636.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0005)  [5][Joung,etal.NatRevMolCellBiol2013;14:49.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0005) [6][Ran,etal.NatProtoc2013;8:2281.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0005) |

7

*F.Störtz,J.K.MakandP.Minary*

[7][Wang,etal.AnnuRevBiochem2016;85:227.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0007)

[8][Xi,etal.BMCBioinformatics2010;11:346.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0007)

[9][Ozaki,etal.Cancers(Basel)2011;3:994.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0009)

[10][Mak,etal.BMCGenomics2022;23:805.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0009)

[11][Chuai,etal.GenomeBiol2018;19:80.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0012)

[12][Liu,etal.PLoSComputBiol2019;15:e1007480.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0012)

[13][Lin,etal.AdvSci2020;7:1903562.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0012)

[14][Charlier,etal.Bioinformatics2021](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0013)[;37:2299.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0015)

[15][Zhang,etal.ComputStructBiotechnolJ2022;20:650.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0015) [16][Störtz,etal.NucleicAcidsRes2020;49:855.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0015)

[17][Lazzarotto,etal.NatBiotechnol2020;3](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0018)[8:13](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0016)[17.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0017)

[18][Alkan,etal.GenomeBiol2018;19:177.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0018)

[19][Zenil,etal.NucleicAcidsRes2019](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0020)[;47:](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0018)[e129.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0019)

[20][Franco,etal.BiolReprod2014;91.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0020)

[21][Sims,etal.TrendsGenet2003;19:](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0020)[6](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0022)[29.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0021)

[22][Anders,etal.Nature2014;513:569.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0023)

[23][Kim,etal.NatMethods2015;12:237.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0023)

[24][Box,etal.JRStatSocB1964;26:211](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0023)[.](http://refhub.elsevier.com/S2667-3185(23)00019-3/sbref0024)

8