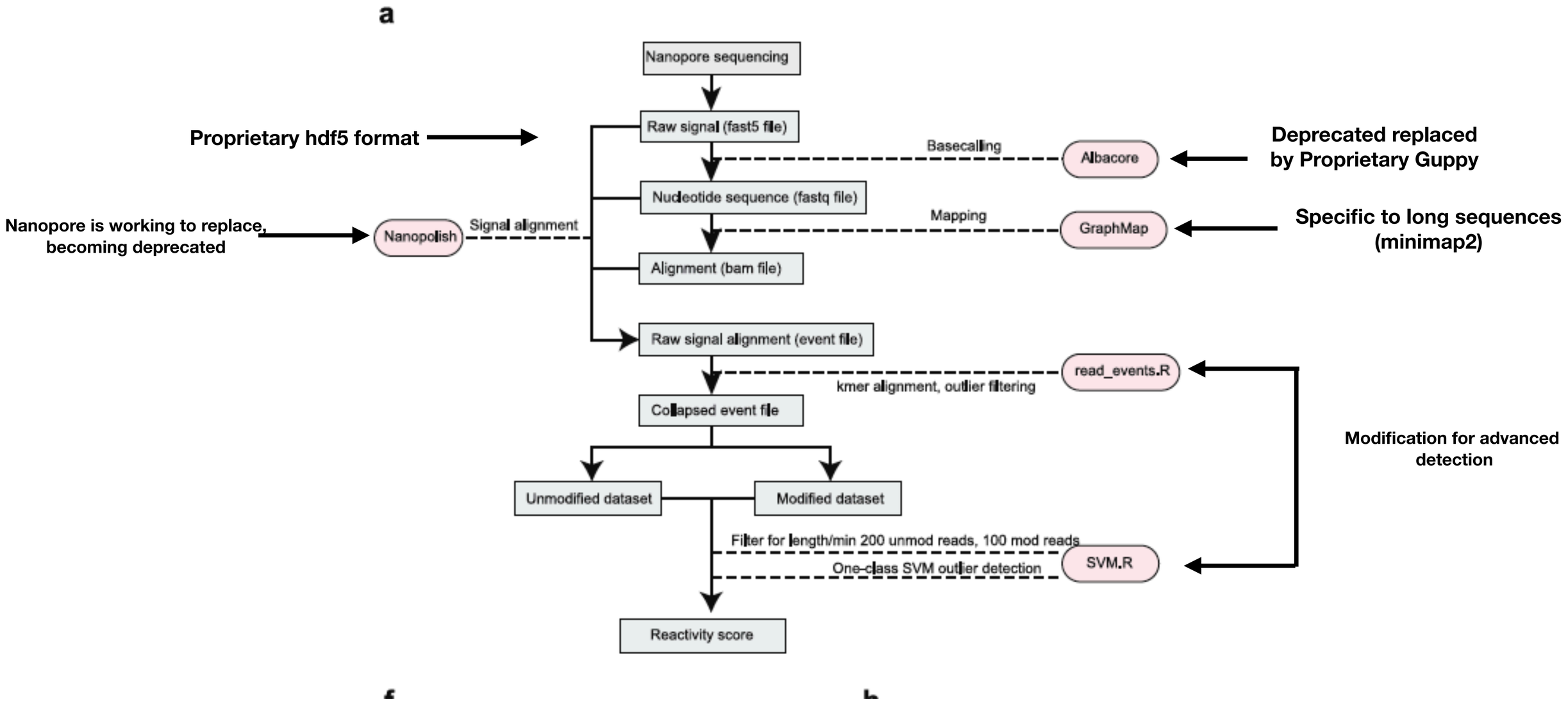
nanoShape: Nanopore based sequence and structure detection on novel RNA

J. White Bear, Grégoire De Bisschop, Juan-Carlos Padilla, Eric Lécuyer, Jérôme Waldispühl

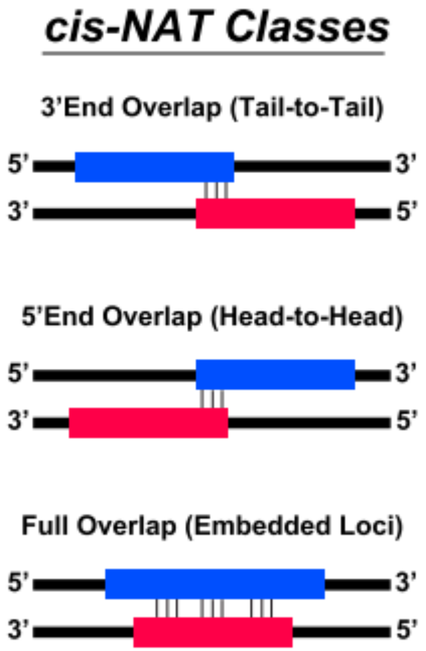
nanoShape Workflow Overview



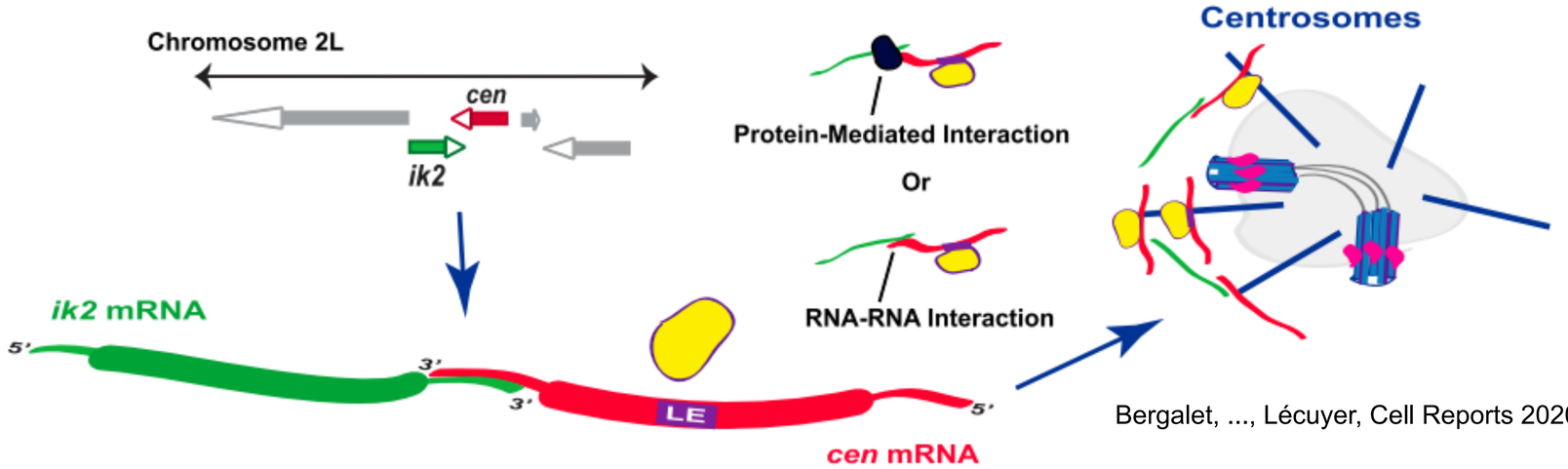
Aw, J.G.A., Lim, S.W., Wang, J.X. *et al.* Determination of isoform-specific RNA structure with nanopore long reads. *Nat Biotechnol* 39, 336–346 (2021). https://doi.org/10.1038/s41587-02

Novel Antisense RNA

Cis-NAT RNA

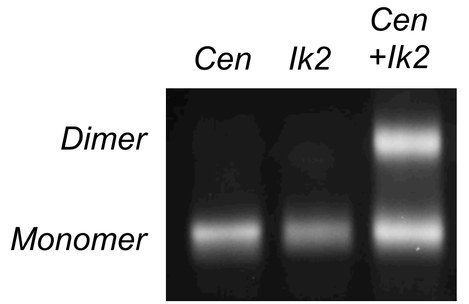


- cis-Natural Antisense Transcripts are overlapping transcripts found in every domains of life
- Assumption that they form long intermolecular duplexes is based on limited structural characterization



Bergalet, ..., Lécuyer, Cell Reports 2020

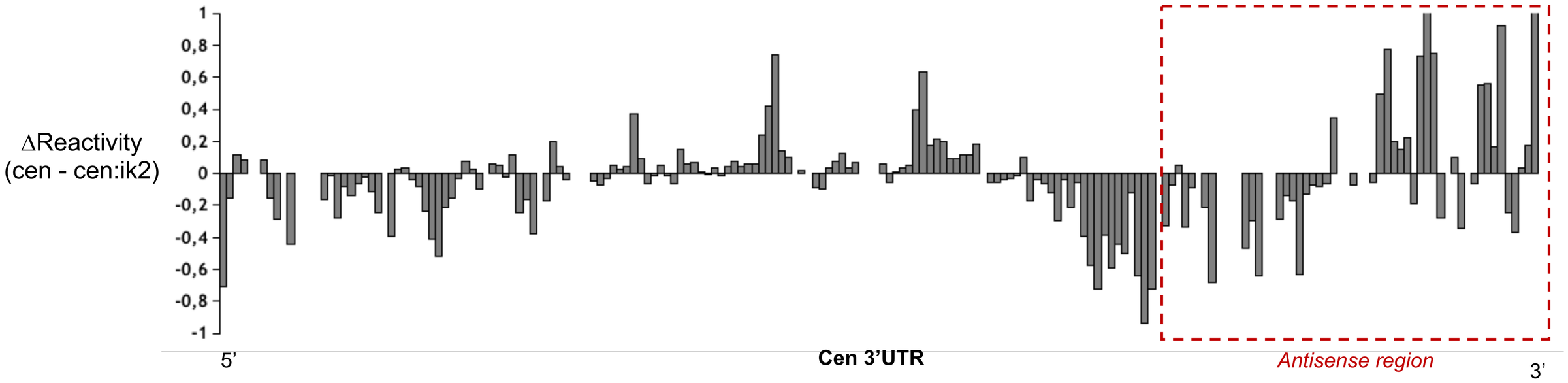
- Cen and Ik2 are cis-NAT involved in mitosis in *Drosophila* embryos
- They share a 59-nucleotide antisense region in their 3'UTR
- Antisense region is important for their interaction and localized translation



10.1093/plcell/koab010<https://doi-org.insb.bib.cnrs.fr/10.1093/plcell/koab010>

Cis-NAT RNA

- Preliminary structure probing indicate a complex re-organization exceeding the mere antisense region:



- Conventional structure probing outputs an **average reactivity signal** resulting from an undefined set of different conformations.

- Our goal is to perform **single molecule structural profiling** using the Nanopore platform to further characterize the interaction between cis-NAT

Experimental Set

- >cen 3'utr
- ACTTGTTTAGAGAATGTAAATAAGCAATTAAACAGTGCATTCTAGCCATAGGGCATTCTACCATTTTTAAATTGTGTGTGCCATGCAGTCTAGTCCGCTTTTTCATGTATAGACAGTTAAATAACAATAACTA
  AATAACTATAATCGGAAATTTAATTTTATTTCAGCATGATAAAATAAATAATTTAATGACCTACAG

- >ik2 3'utr
- ACGGGCATATCATGAAAGTGCAAGAATATTTTATTTGCCTTTACTTTGTAAGTTAACAATAAATGTTTACTTTTTTATATCTGAATTTGTAAAGCAACTACATATATTCCTATTGAAACTTGGCTGAATATCGTG
  AAAGAGTAAGATTTCTGTAGGTCATTAAATTATTTATTTTATCATGCTGAAATAAAATTAAATTTCCGATTAT

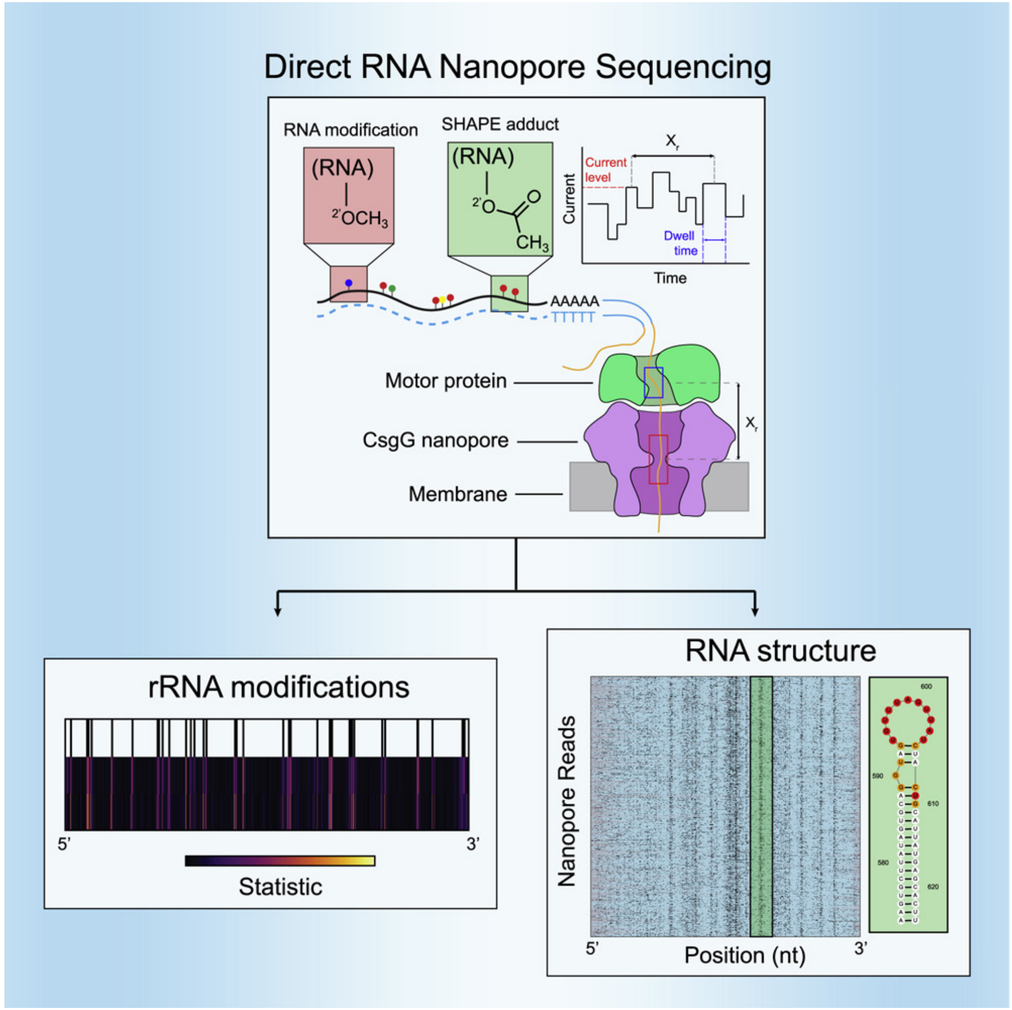
- >E coli tmRNA
- GGGGCTGATTCTGGATTCGACGGGATTTGCGAAACCCAAGGTGCATGCCGAGGGGCGGTTGGCCTCGTAAAAAGCCGCAAAAAATAGTCGCAAACGACGAAAACTACGCTTTAGCAGCTTAATAACC
  TGCTTAGAGCCCTCTCTCCCTAGCCTCCGCTCTTAGGACGGGGATCAAGAGAGGTCAAACCCAAAAGAGATCGCGTGGAAGCCCTGCCTGGGGTTGAAGCGTTAAAACTTAATCAGGCTAGTTTGTT
  AGTGGCGTGTCCGTCCGCAGCTGGCAAGCGAATGTAAAGACTGACTAAGCATGTAGTACCGAGGATGTAGGAATTTCGGACGCGGGTTCAACTCCCGCCAGCTCCACCA

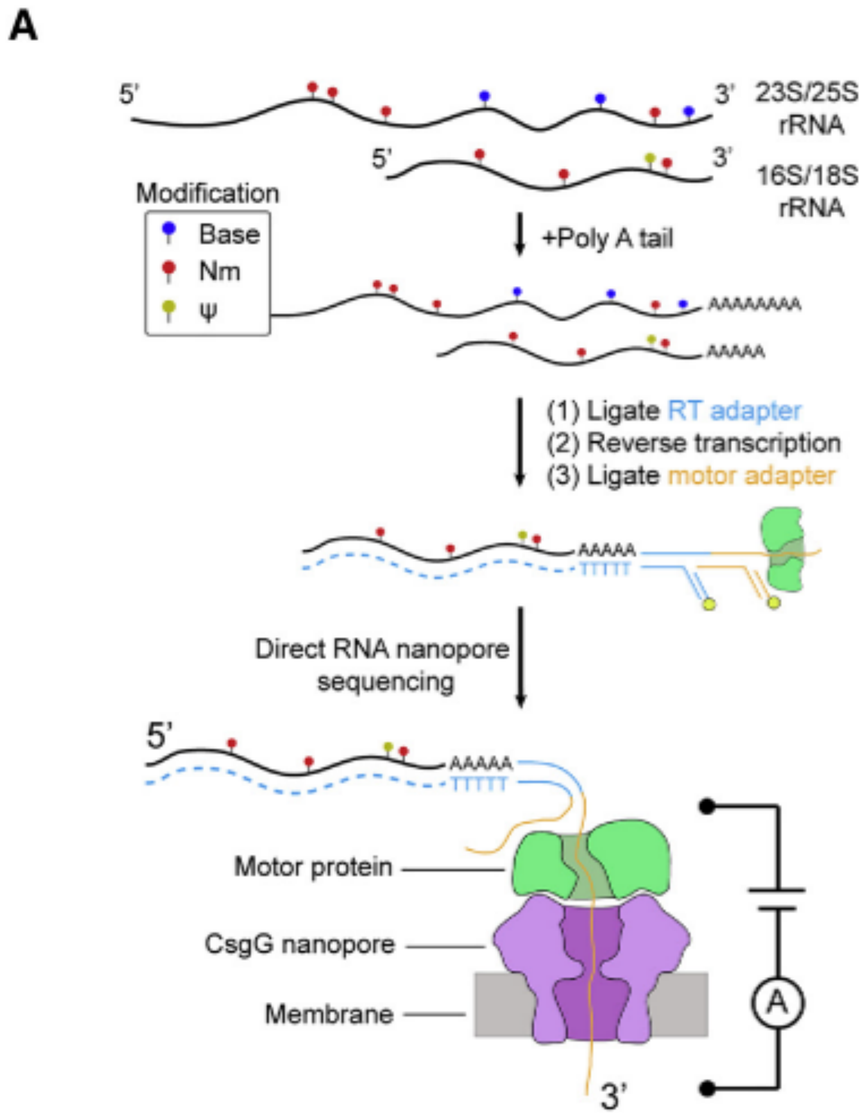
- >T thermophila self splicing intron
- GGAGGGAAAAGTTATCAGGCATGCACCTGGTAGCTAGTCTTTAAACCAATAGATTGCATCGGTTTAAAAGGCAAGACCGTCAAATTGCGGGAAAGGGGTCAACAGCCGTTCAGTACCAAGTCTCAGGG
  GAAACTTTGAGATGGCCTTGCAAAGGGTATGGTAATAAGCTGACGGACATGGTCCTAACCACGCAGCCAAGTCCTAAGTCAACAGATCTTCTGTTGATATGGATGCAGTTCACAGACTAAATGTCGGTC
  GGGGAAGATGTATTCTTCTCATAAGATATAGTCGGACCTCTCCTTAATGGGAGCTAGCGGATGAAGTGATGCAACACTGGAGCCGCTGGGAACTAATTTGTATGCGAAAGTATATTGATTAGTTTTGGAG

- >HCV internal entry site
- CGAGTTGGGGGCGACACTCCACCATAGATCACTCCCCTGTGAGGAACTACTGTCTTCACGCAGAAAGCGTCTAGCCATGGCGTTAGTATGAGTGTCGTGCAGCCTCCAGGACCCCCCCTCCCGGGA
  GAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCGGAATTGCCAGGACGACCGGGTCCTTTCTTGGATCAACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCGAGACTGCTAGCCGAGTA

Nanopore Basecalling

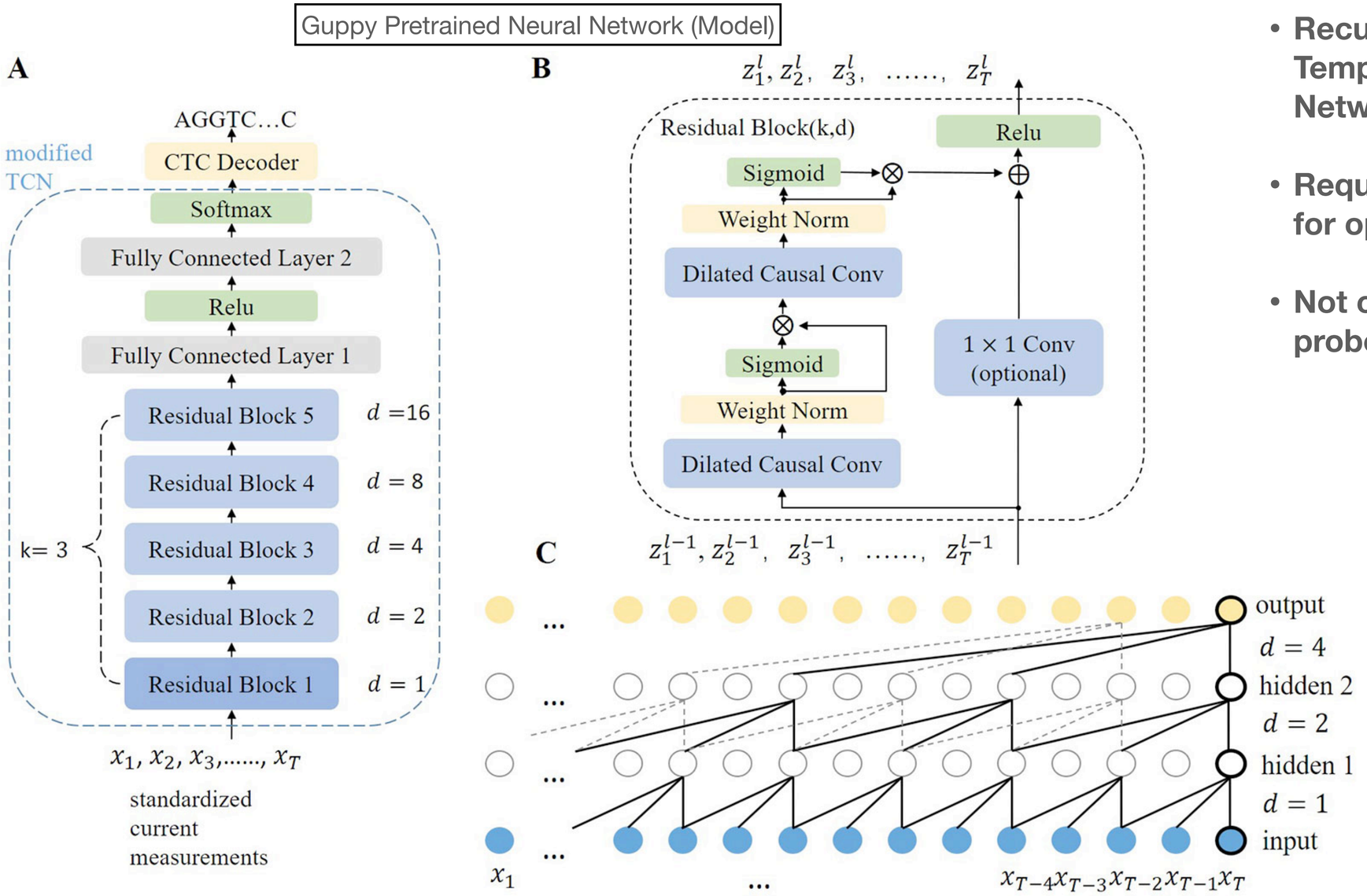
Nanopore Overview





W. Stephenson, R. Razaghi, S. Busan, K. M. Weeks, W. Timp, and P. Smibert, “Direct detection of RNA modifications and structure using single-molecule nanopore sequencing,” *Cell Genomics*, vol. 2, no. 2, p. 100097, Feb. 2022, doi: 10.1016/j.xgen.2022.100097.

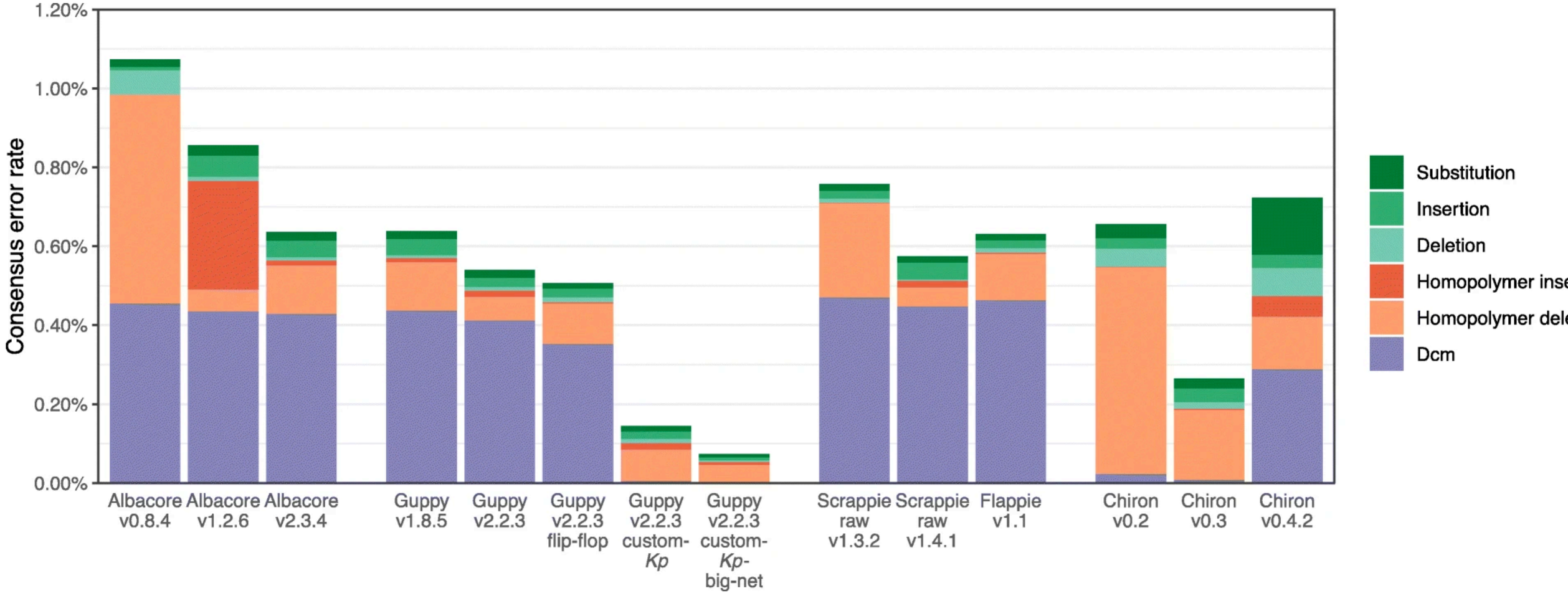
Nanopore: Proprietary Basecalling



- Recurrent Neural Network (RNN)/ Temporal Convolutional Neural Network (TCN)
- Requires custom trained networks for optimal performance
- Not characterized across structure probes

An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling Shaojie Bai, J. Zico Kolter, Vladlen Koltun https://arxiv.org/abs/1803.01271

Basecalling: Error Characterization

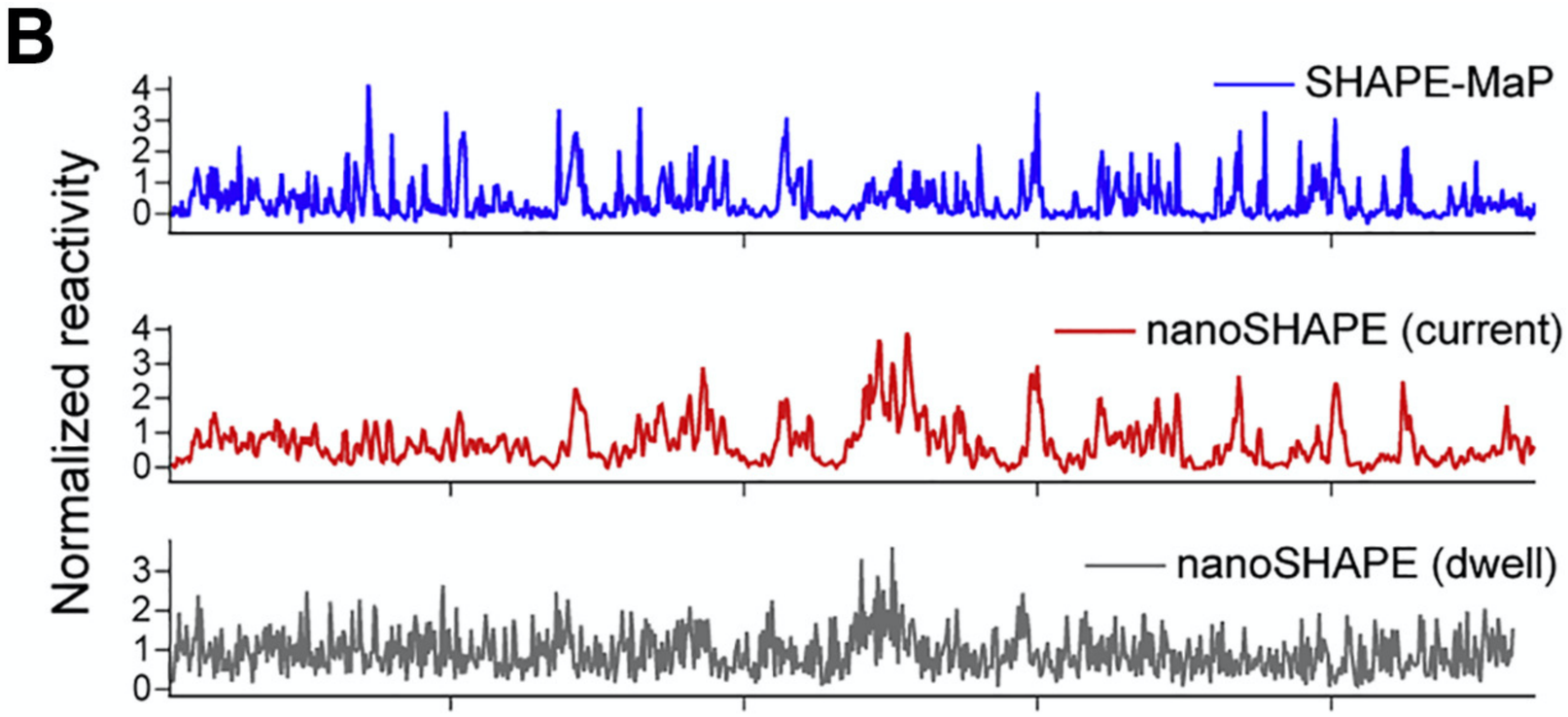


Consensus errors per basecaller for the *K. pneumoniae* benchmarking set, broken down by type. Dcm refers to errors occurring in the CCAGG/CCTGG Dcm motif. Homopolymer errors are changes in the length of a homopolymer three or more bases in length (in the reference). This plot is limited to basecallers/versions with less than 1.2% consensus error and excludes redundant results from similar versions

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1727-y

Reactivity Profiles

Calculating Reactivities



Normalized SHAPE-MaP reactivity (blue) and nanoSHAPE reactivity detected by changes in current (red) and dwell time (gray) for the pri-miR-17~92 RNA.

W. Stephenson, R. Razaghi, S. Busan, K. M. Weeks, W. Timp, and P. Smibert, "Direct detection of RNA modifications and structure using single-molecule nanopore sequencing," *Cell Genomics*, 2022, doi: 10.1016/j.xgen.2022.100097.

Basecall Error Profile

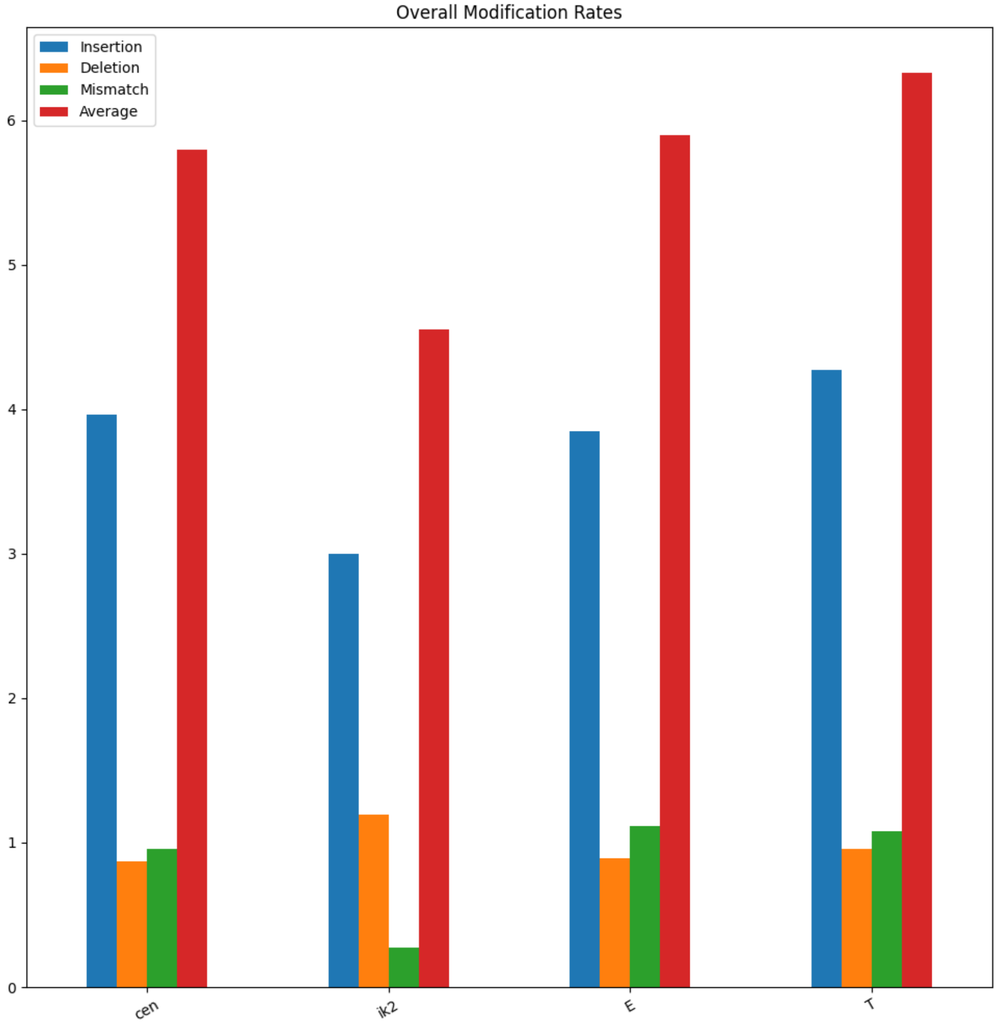
DMSO_mod_rates

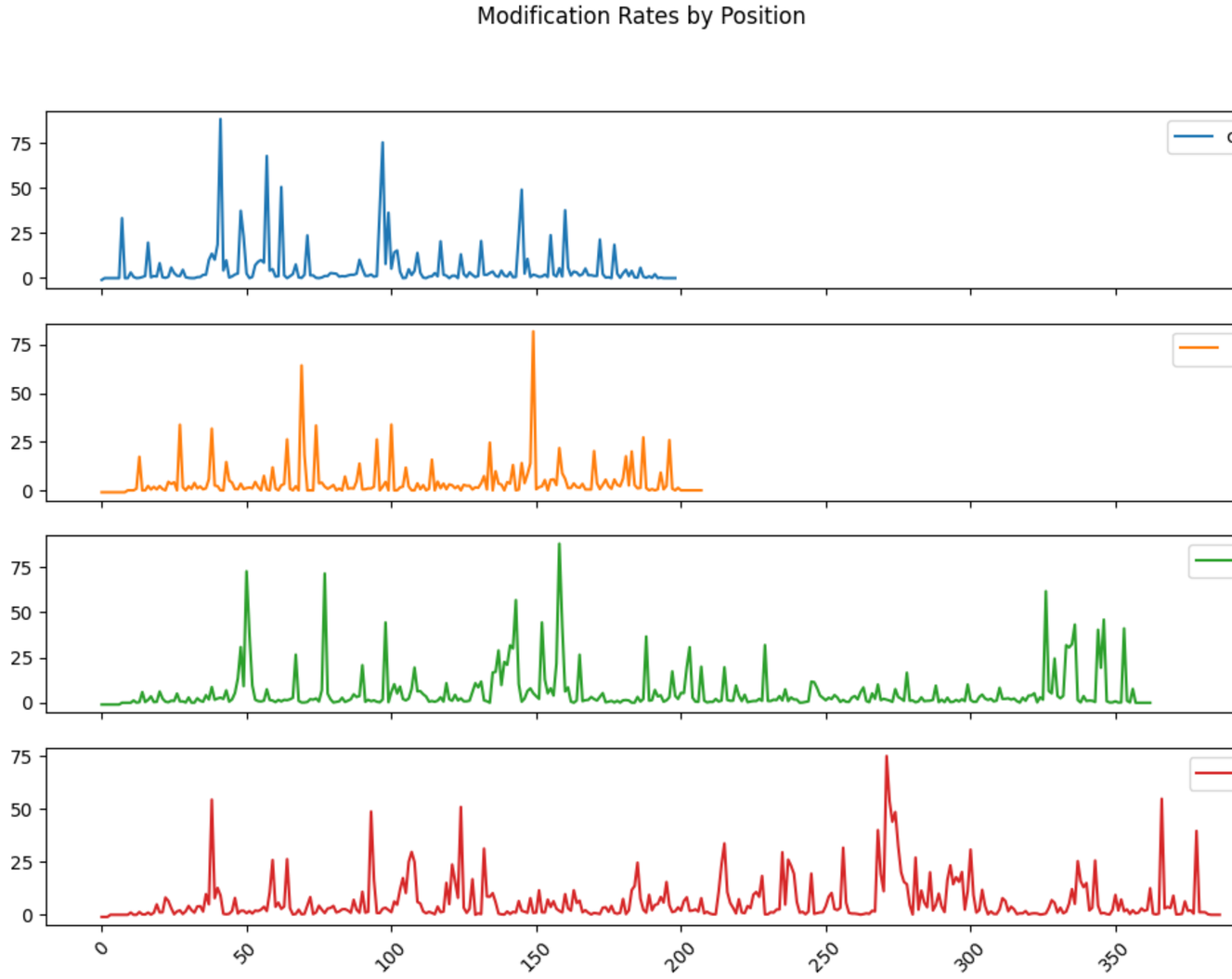
	cen	ik2	E	T
0	-1.0	-1.0	-1.0	-1.0
1	0.0	-1.0	-1.0	-1.0
2	0.0	-1.0	-1.0	-1.0
3	0.0	-1.0	-1.0	0.0
4	0.0	-1.0	-1.0	0.0
5	0.0	-1.0	-1.0	0.0
6	0.0	-1.0	-1.0	0.0
7	33.33333	-1.0	0.0	0.0
8	0.0	-1.0	0.0	0.0
9	0.0	0.0	0.0	0.0
10	3.125	0.0	0.0	1.041666
11	0.675675	0.0	1.2779552	0.0
12	0.0	0.8695652	0.0761614	0.0
13	0.147492	17.261904	0.3490401	1.311188
14	0.592592	0.0	5.8845861	0.208768
15	1.217656	0.0	0.3959683	0.153022

DMSO_ave_mod_rates

	Condition	Insertion	Deletion	Mismatch	Average
cen	DMSO	3.9593924522324800	0.8714612548359300	0.9578974378748570	5.7988013961995500
ik2	DMSO	2.999816460831550	1.1911793176931700	0.27227237590318100	4.549806615966360
E	DMSO	3.8482626986964700	0.889682800559347	1.1160871116077800	5.892600103976550
T	DMSO	4.27167100379748	0.9589208789876810	1.0813439460927700	6.327439704846920

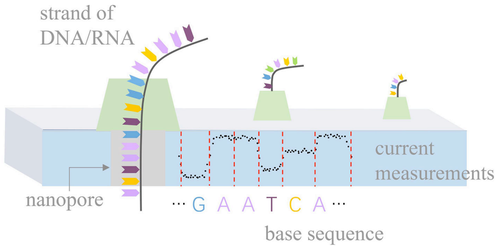
Basecall Error Profile





Signal Analysis

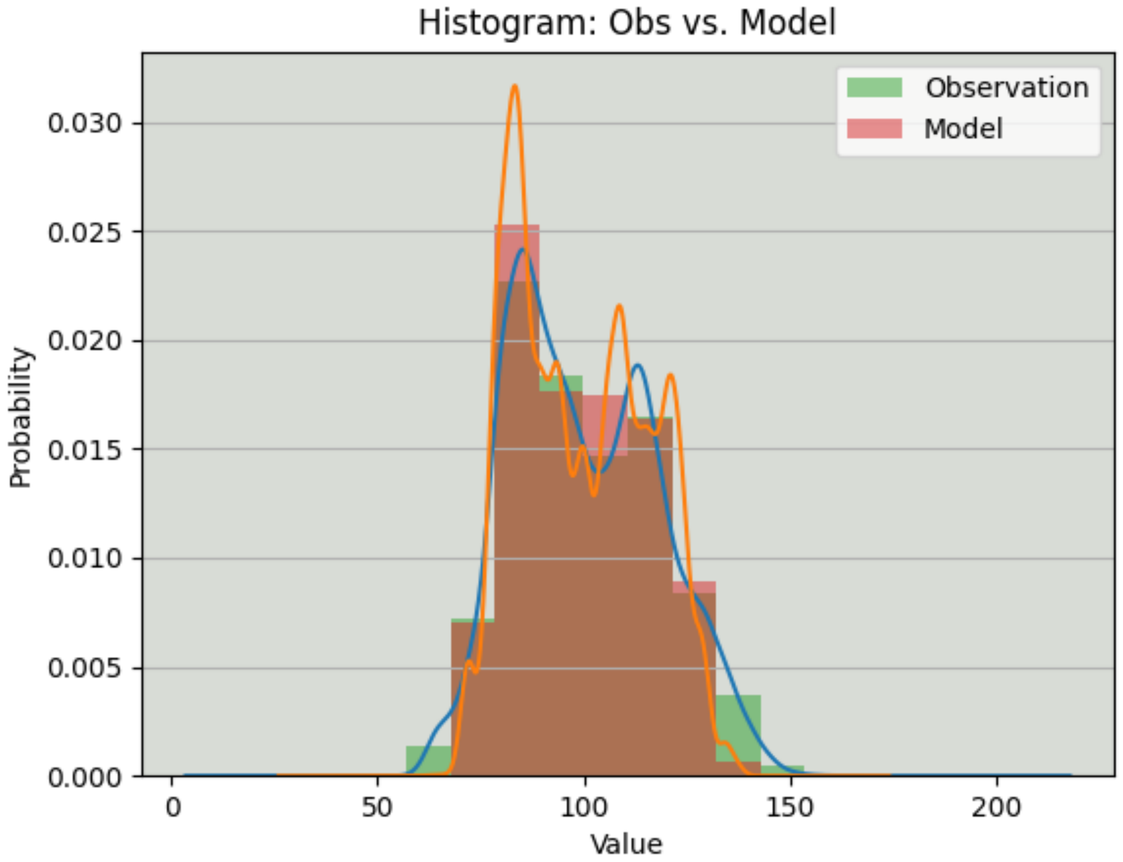
Signal Analysis Error Profile

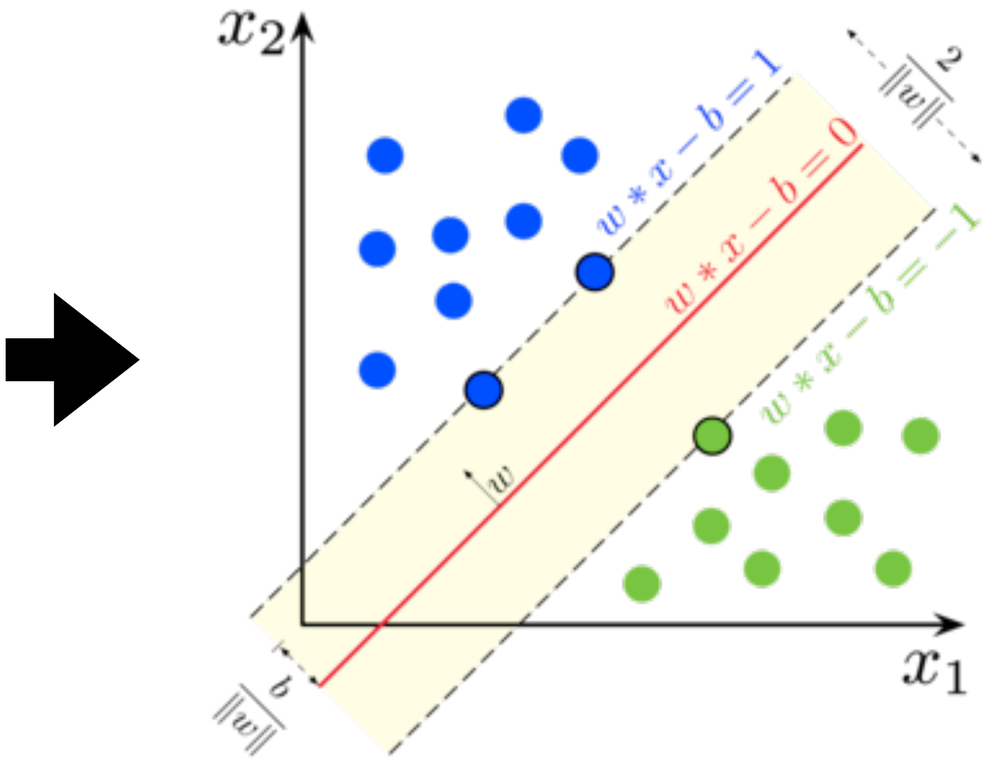


Gene Events

contig	position	reference_kmer	read_index	strand	event_index	vent_level_mea	event_stdv	event_length	model_kmer	model_mean	model_stdv	andardized_lev
cen	1	CTTGT	0	t	527	79.85	2.509	0.00465	ACAAG	78.46	3.02	0.36
cen	2	TTGTT	0	t	526	96.35	2.844	0.00531	AACAA	87.82	3.18	2.10
cen	3	TGTTT	0	t	525	103.47	1.585	0.00896	AAACA	99.38	3.45	0.93
cen	4	GTTTA	0	t	524	102.41	1.181	0.00232	TAAAC	103.45	2.68	-0.31
cen	4	GTTTA	0	t	523	104.88	1.859	0.00797	TAAAC	103.45	2.68	0.42
cen	4	GTTTA	0	t	522	98.75	1.705	0.00232	TAAAC	103.45	2.68	-1.38
cen	5	TTTAG	0	t	521	90.94	4.412	0.00432	CTAAA	94.00	3.04	-0.79
cen	6	TTAGA	0	t	520	80.95	5.200	0.01096	TCTAA	84.15	2.63	-0.95
cen	7	TAGAG	0	t	519	79.07	2.177	0.02988	CTCTA	79.85	2.07	-0.30
cen	8	AGAGA	0	t	518	86.03	1.772	0.01295	TCTCT	79.44	2.85	1.82
cen	9	GAGAA	0	t	517	79.96	1.758	0.01129	TTCTC	77.30	2.07	1.01
cen	10	AGAAT	0	t	516	78.60	1.407	0.01162	ATTCT	81.64	2.07	-1.15
cen	10	AGAAT	0	t	515	70.89	1.415	0.00398	NNNNN	0.00	0.00	inf
cen	16	TAAAT	0	t	514	75.37	2.945	0.00398	ATTTA	77.58	1.97	-0.88
cen	17	AAATA	0	t	513	95.42	2.218	0.01726	TATTT	94.78	5.53	0.09
cen	18	AATAA	0	t	512	85.11	1.536	0.00465	TTATT	94.20	2.63	-2.71
cen	19	ATAAG	0	t	511	81.96	20.145	0.00498	CTTAT	84.65	2.46	-0.86

Basecalling w/ Signal Analysis





Signal Analysis Error Profile

Aw SVM Analysis vs Basecall Error Rates

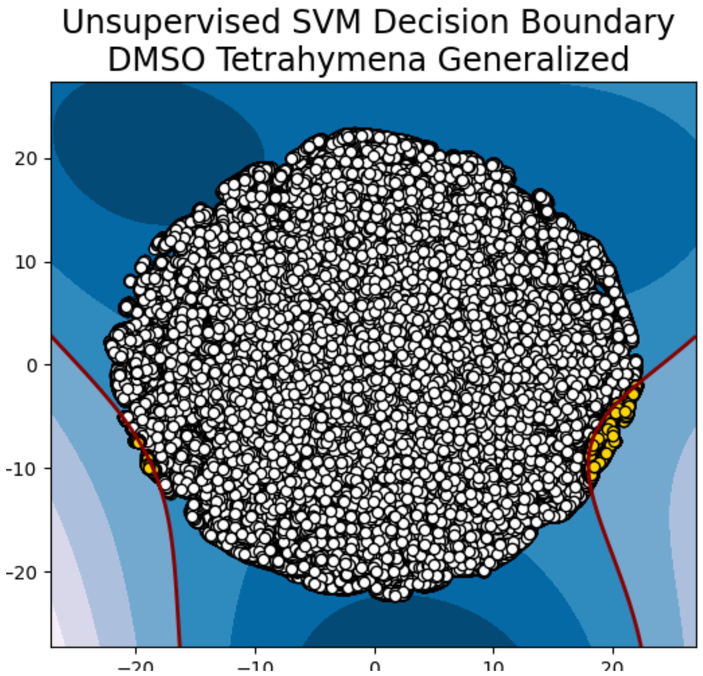
cen___error_comparison_norm__signal_out

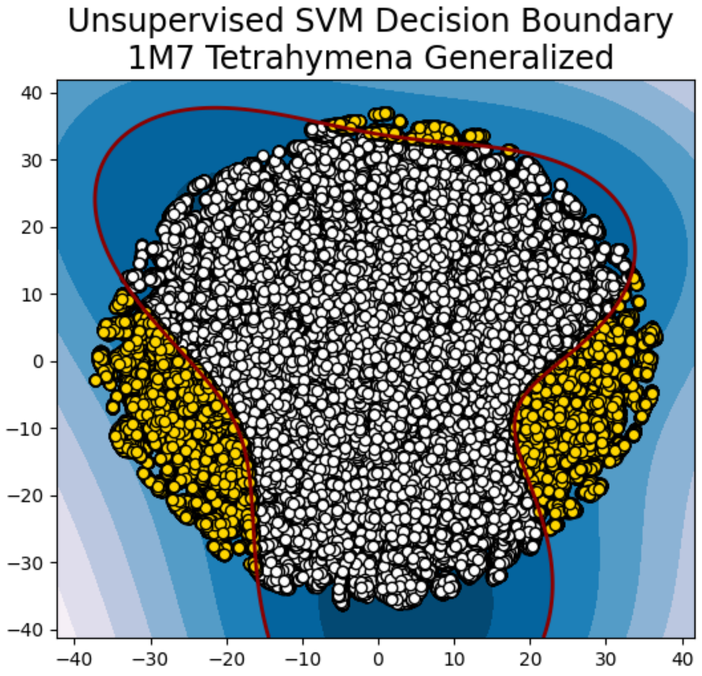
Position	Num_Reads	SVM_DMSO	DMSO BC Error Rate
0	0	0.0	0.0
1	1	0.0	0.0
2	6	0.0	0.0
3	5	0.0	0.0
4	4	0.0	0.0
5	5	0.0	0.0
6	19	0.0	0.028860991446715600
7	3	0.0	0.0
8	3	0.0	0.0
9	69	0.039113483289612200	0.002705717948129590 0
10	112	0.04819339905327220	0.0005850200968928840
11	30	0.0	0.0
12	76	0.0	0.0001277035019766180
13	67	0.0	0.0005130842923860550
14	78	0.0	0.0010542827925740900
15	83	0.0	0.017085364372758900

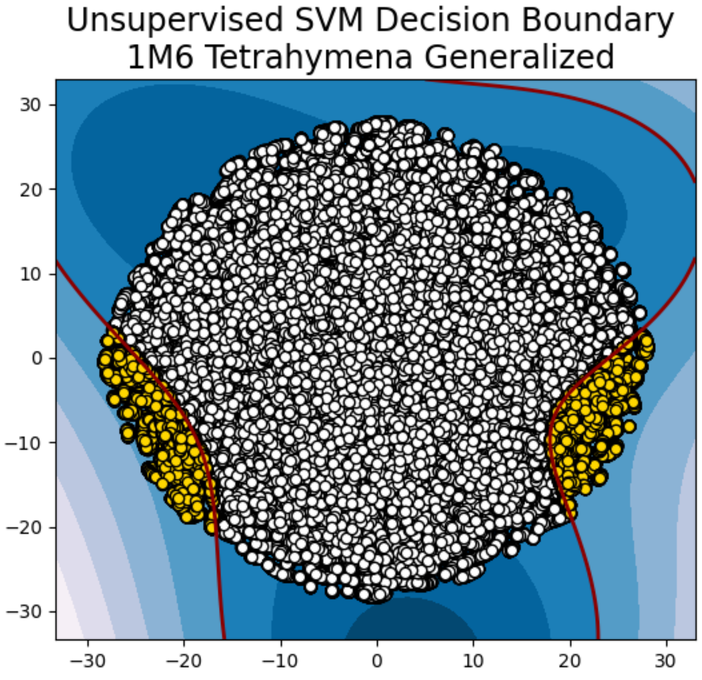
Signal vs Basecall Error Correlation:

cen:[('DMSO'	**MannwhitneyuResult(statistic=12047.5**	**pvalue=1.2235595514447803e-10))]**
ik2:[('DMSO'	MannwhitneyuResult(statistic=7622.0	pvalue=2.4365222662534698e-29))]
E:[('DMSO'	MannwhitneyuResult(statistic=60241.0	pvalue=0.08134386144338046))]
T:[('DMSO'	MannwhitneyuResult(statistic=50487.0	pvalue=3.684209884609389e-14))]

Optimizing for error rather than truth.
What does an outlier really tell you?

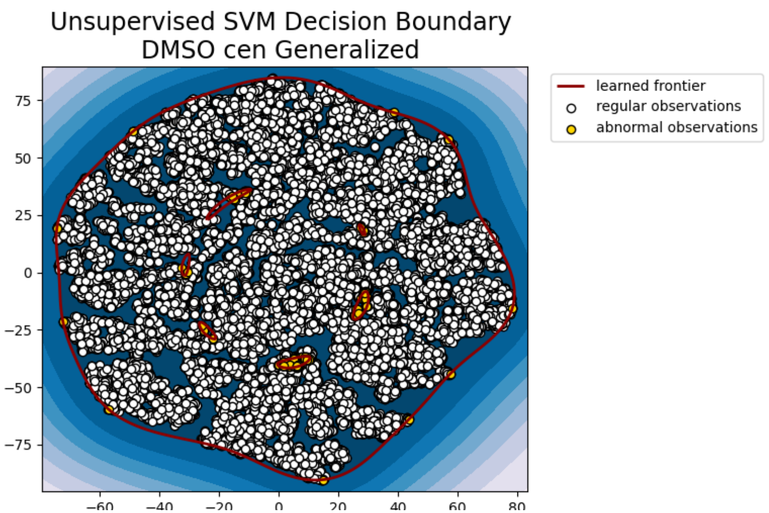


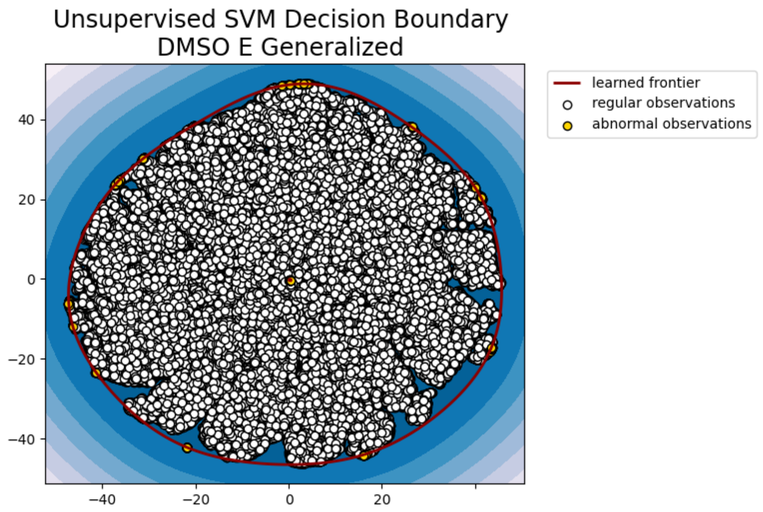


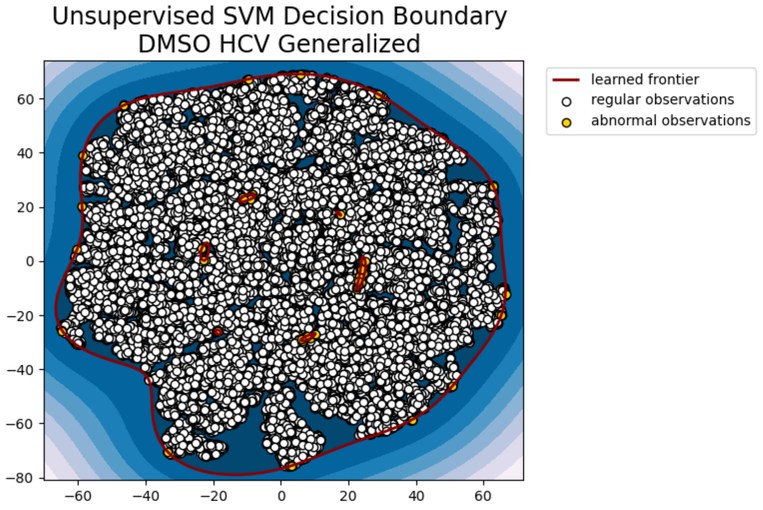


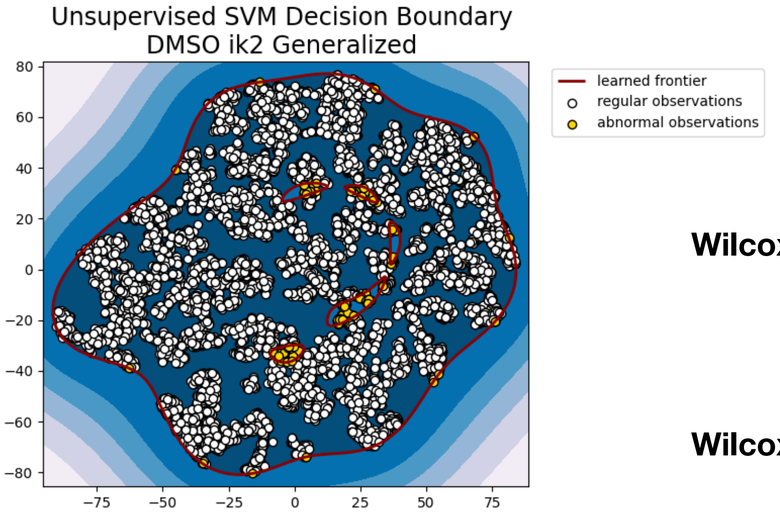
Loss of Structural Correlation, Aw et. al. focused only on SS loci and chose best performers

Signal Analysis Error Profile



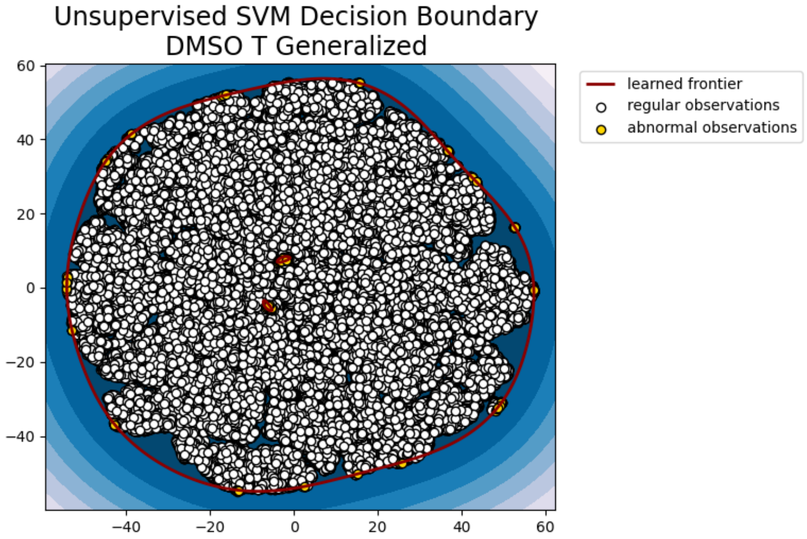






Aw_SVM
WilcoxonResult(pvalue=0.017500)

Expanded_SVM
WilcoxonResult(pvalue=0.001340)



Aw, J.G.A., Lim, S.W., Wang, J.X. *et al.* Determination of isoform-specific RNA structure with nanopore long reads. *Nat Biotechnol* 39, 336–346 (2021). https://doi.org/10.1038/s41587-020-0712-z

Deep Learning for Modification and Structure Prediction

Adding Structural Information

Improving performance on SS and BP pairs

dmso_signal_out

Sequer	Reference N	Position	SVM Signal A	Number of Mis	Mismatc	Unmodifie
Tetrahy	G	0	-1			
Tetrahy	A	1	-1			
Tetrahy	C	2	-1			
Tetrahy	C	3	-1			
Tetrahy	G	4	-1			
Tetrahy	T	5	-1			
Tetrahy	C	6	1			
Tetrahy	A	7	-1	0.052770448548	AC	BP
Tetrahy	A	8	1	0.052770448548	AG	S
Tetrahy	A	9	-1			
Tetrahy	T	10	1			
Tetrahy	T	11	1	0.052770448548	TA	BP
Tetrahy	G	12	1			
Tetrahy	C	13	1			
Tetrahy	G	14	1	0.738786279683	GA	BP
Tetrahy	G	15	-1	0.211081794195	GA	BP
Tetrahy	G	16	1	0.369393139841	GA	BP
Tetrahy	A	17	1	0.158311345646	AG	S
Tetrahy	A	18	1			

Tetrahymena_DMSO_modification_bias

Sequence	Mismatch	Number of Misma	Total Mismatches in Aligne	Modified Bp Freq	Position
Tetrahyme	AC	50	1895	2.638522427440630	204
Tetrahyme	AG	81	1895	4.274406332453830	229
Tetrahyme	TA	83	1895	4.379947229551450	213
Tetrahyme	GA	810	1895	42.74406332453830	237
Tetrahyme	GC	15	1895	0.79155672823219	179
Tetrahyme	GT	36	1895	1.8997361477572600	216
Tetrahyme	TC	149	1895	7.862796833773090	236
Tetrahyme	CT	336	1895	17.730870712401100	225
Tetrahyme	AT	157	1895	8.284960422163590	239
Tetrahyme	CA	111	1895	5.857519788918210	224
Tetrahyme	CG	16	1895	0.8443271767810030	201
Tetrahyme	TG	51	1895	2.691292875989450	210

Error and Modification Rates are Influenced by:

Sequence
Position
Structure
Probe
Stoichiometry

Supervised Learning (MultiClass)

Multi-class Neural network with Expanded Features

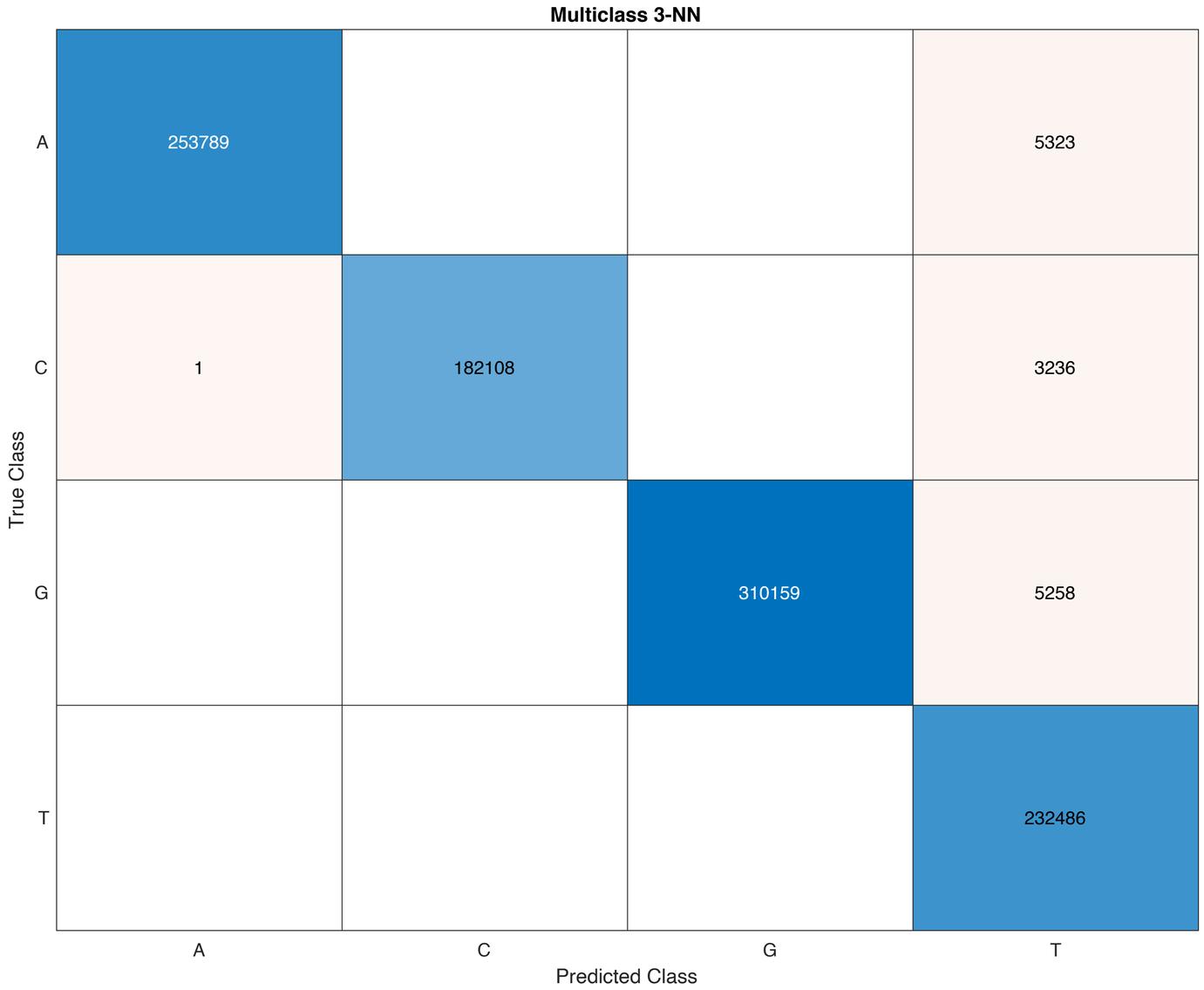
Multi-Class Data

0	'GACCG'	3319	't'	8	700000000	400000000	990000000	'GACCG'	700000000	000000000	000000000	'G'
1	'ACCGT'	3319	't'	9	900000000	700000000	640000000	'ACCGT'	300000000	000000000	000000000	'A'
1	'ACCGT'	3319	't'	10	700000000	500000000	820000000	'ACCGT'	300000000	000000000	000000000	'A'
1	'ACCGT'	3319	't'	11	000000000	800000000	980000000	'ACCGT'	300000000	000000000	000000000	'A'
2	'CCGTC'	3319	't'	12	900000000	100000000	650000000	'CCGTC'	700000000	000000000	000000000	'C'
3	'CGTCA'	3319	't'	13	030000000	300000000	320000000	'CGTCA'	000000000	000000000	000000000	'C'
3	'CGTCA'	3319	't'	14	500000000	600000000	570000000	'CGTCA'	000000000	000000000	000000000	'C'
4	'GTCAA'	3319	't'	15	900000000	500000000	920000000	'GTCAA'	600000000	000000000	000000000	'G'
5	'TCAAA'	3319	't'	16	700000000	000000000	980000000	'TCAAA'	100000000	000000000	000000000	'T'
6	'CAAAT'	3319	't'	17	700000000	100000000	650000000	'CAAAT'	070000000	000000000	200000000	'C'
7	'AAATT'	3319	't'	18	320000000	600000000	970000000	'AAATT'	980000000	000000000	000000000	'A'
7	'AAATT'	3319	't'	19	720000000	600000000	320000000	'AAATT'	980000000	000000000	000000000	'A'
7	'AAATT'	3319	't'	20	820000000	200000000	980000000	'AAATT'	980000000	000000000	000000000	'A'

96.7% Accuracy

Predicting base based on position and expanded features

Currently Tetrahymena only due to data size and training time



Adding Structural Information

(Multi-class and Multi-output)- Tetrahymena

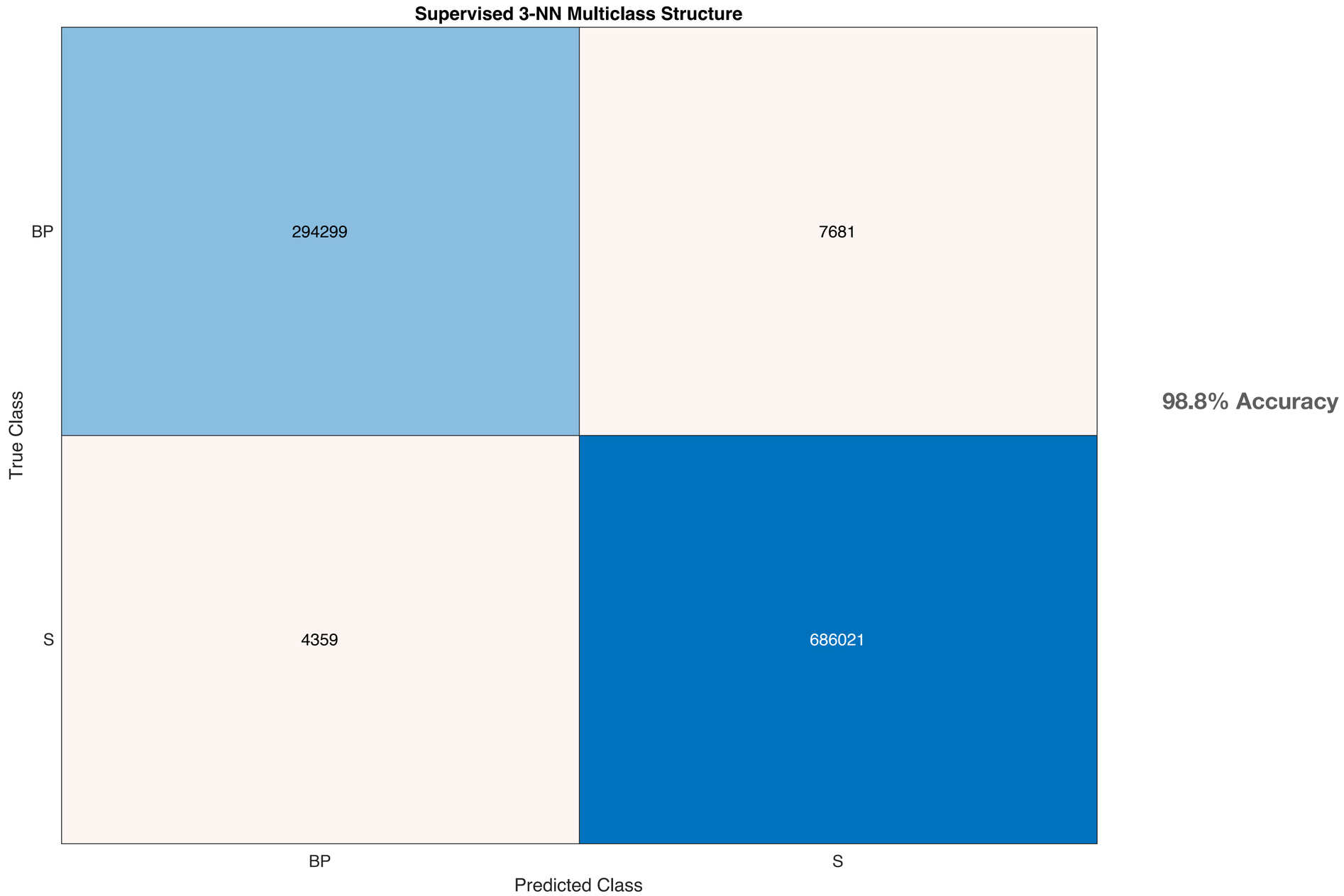
Neural Network with multi-output

Multi-Class & Multi-Output

1	'ACCGT'	3319	't'	9	.190000000004	70000000000	564000000000	'ACCGT'	.730000000007	800000000000	500000000000	ABP
1	'ACCGT'	3319	't'	10	.770000000006	250000000000	232000000000	'ACCGT'	.730000000007	800000000000	400000000000	ABP
1	'ACCGT'	3319	't'	11	.900000000003	880000000000	598000000000	'ACCGT'	.730000000007	800000000000	490000000000	ABP
2	'CCGTC'	3319	't'	12	.590000000001	110000000000	465000000000	'CCGTC'	.270000000000	500000000000	500000000000	CBP
3	'CGTCA'	3319	't'	13	0.030000000002	030000000000	232000000000	'CGTCA'	.900000000000	700000000000	230000000000	CBP
3	'CGTCA'	3319	't'	14	.050000000004	160000000000	245700000000	'CGTCA'	.900000000000	700000000000	590000000000	CBP
4	'GTCAA'	3319	't'	15	.790000000006	650000000000	209200000000	'GTCAA'	.560000000003	000000000000	680000000000	GBP
5	'TCAAA'	3319	't'	16	.270000000008	600000000000	498000000000	'TCAAA'	.610000000000	200000000000	880000000000	TS
6	'CAAAT'	3319	't'	17	.970000000003	310000000000	365000000000	'CAAAT'	0.070000000006	800000000000	620000000000	CBP
7	'AAATT'	3319	't'	18	1.320000000003	160000000000	697000000000	'AAATT'	5.980000000011	000000000000	270000000000	ABP
7	'AAATT'	3319	't'	19	5.720000000003	460000000000	232000000000	'AAATT'	5.980000000011	000000000000	320000000000	ABP
7	'AAATT'	3319	't'	20	7.820000000002	020000000000	598000000000	'AAATT'	5.980000000011	000000000000	320000000000	ABP
7	'AAATT'	3319	't'	21	9.830000000009	300000000000	232000000000	'AAATT'	5.980000000011	000000000000	300000000000	ABP
7	'AAATT'	3319	't'	22	3.380000000006	060000000000	432000000000	'AAATT'	5.980000000011	000000000000	140000000000	ABP

Multi-output represented as concatenated y-hat additional ensemble/chained representation required

Adding Structural Information



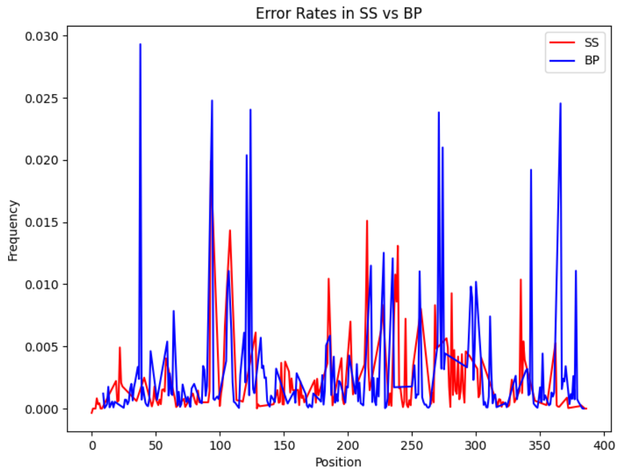
98.8% Accuracy

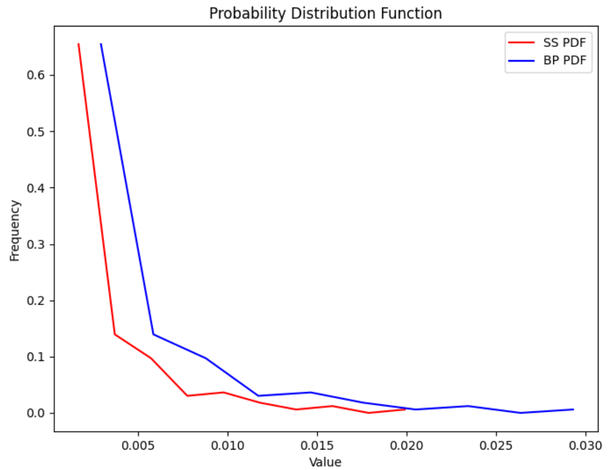
Adding Structural Information

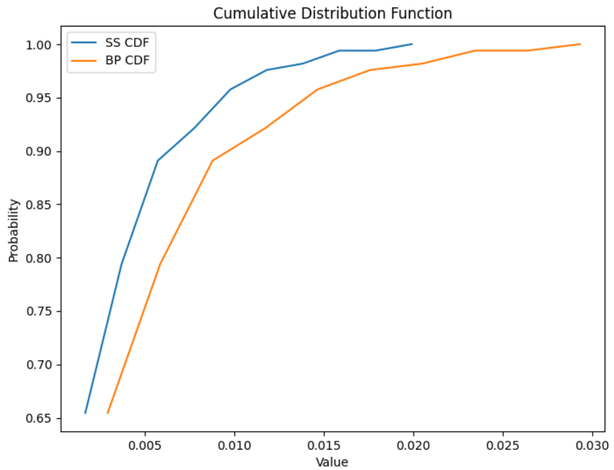
97.5% Accuracy

Encoded Structural Information

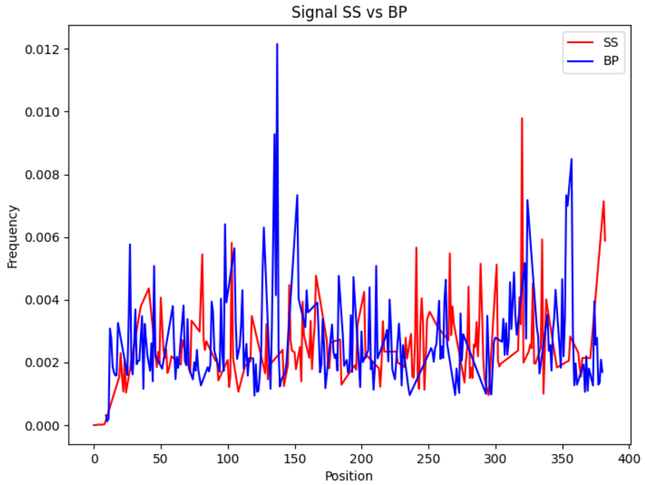
Basecall:

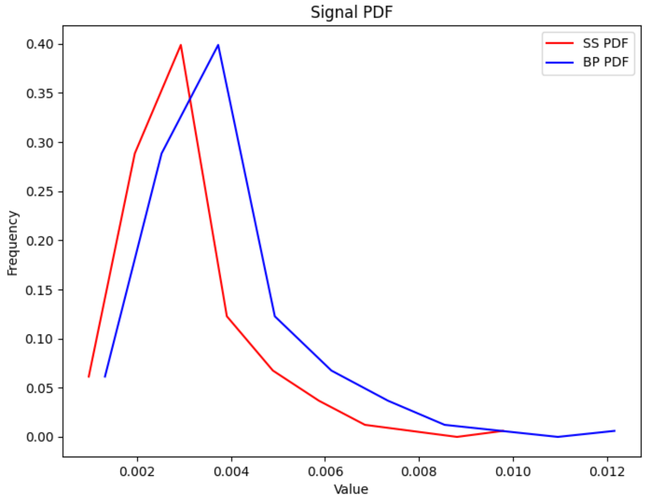


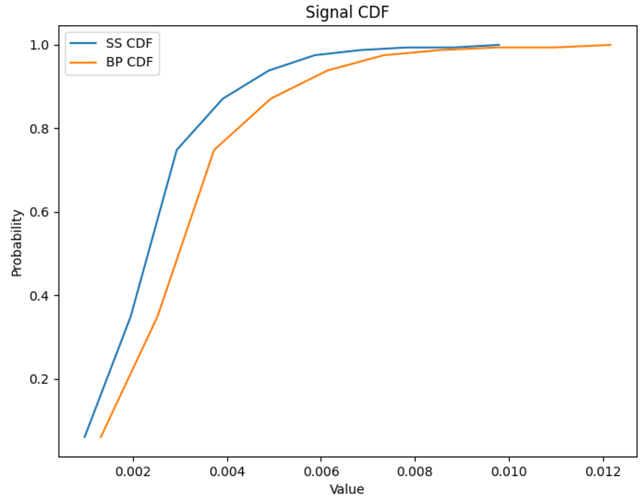




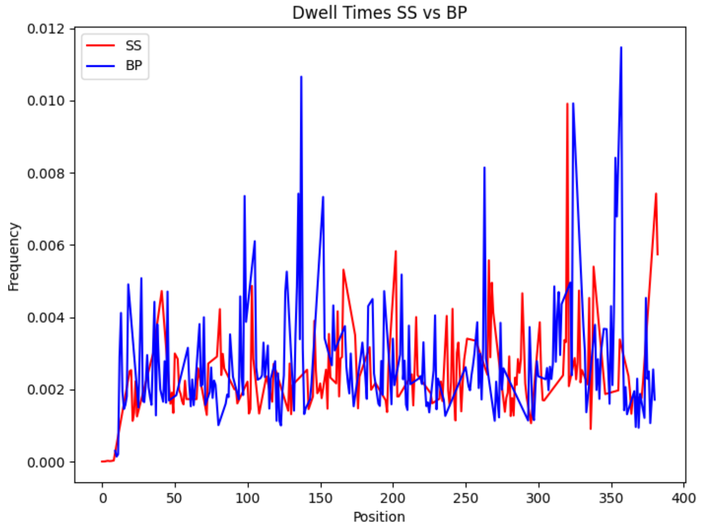
Signal Fluctuation:

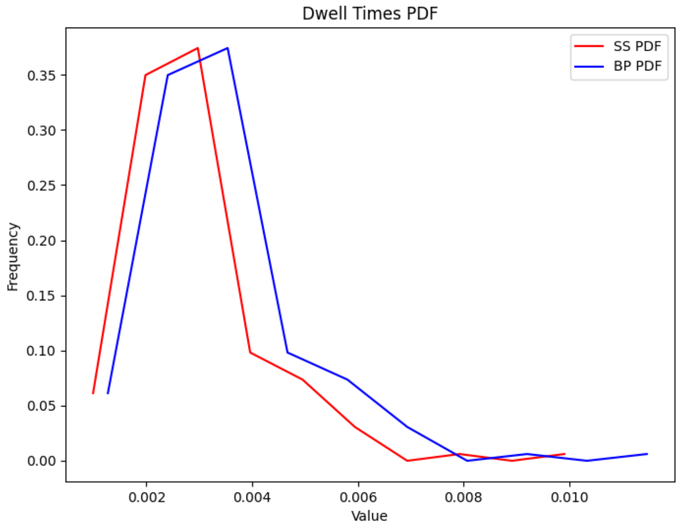


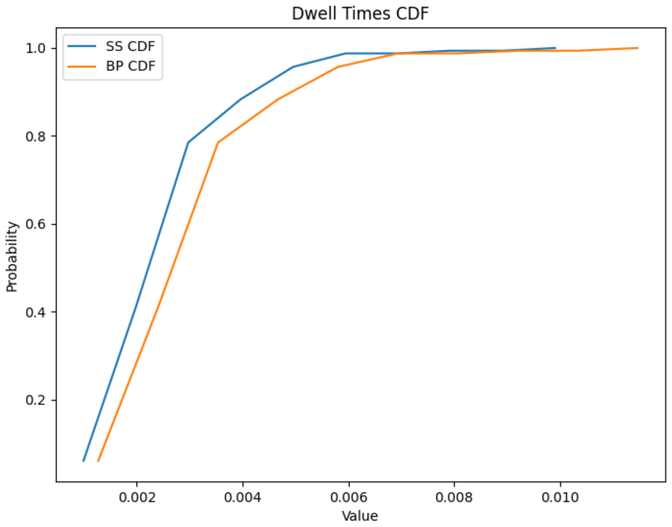




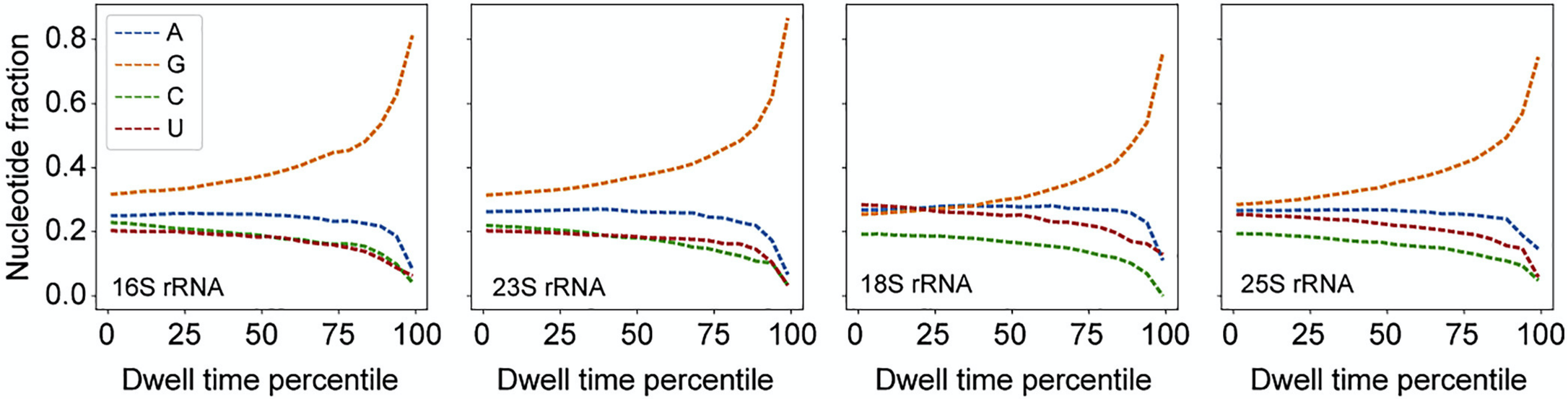
Encoded Structural Information







C

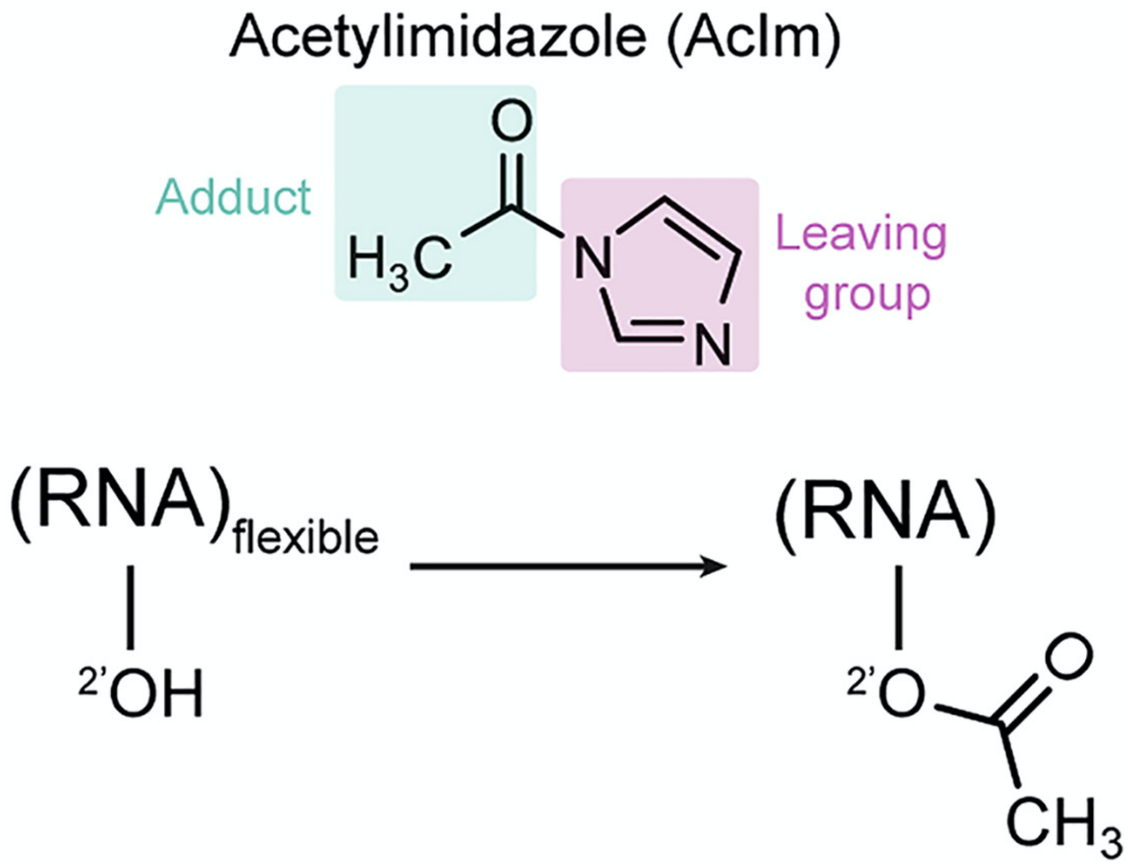


W. Stephenson, R. Razaghi, S. Busan, K. M. Weeks, W. Timp, and P. Smibert, “Direct detection of RNA modifications and structure using single-molecule nanopore sequencing,” *Cell Genomics*, vol. 2, no. 2, p. 100097, Feb. 2022, doi: 10.1016/j.xgen.2022.100097.

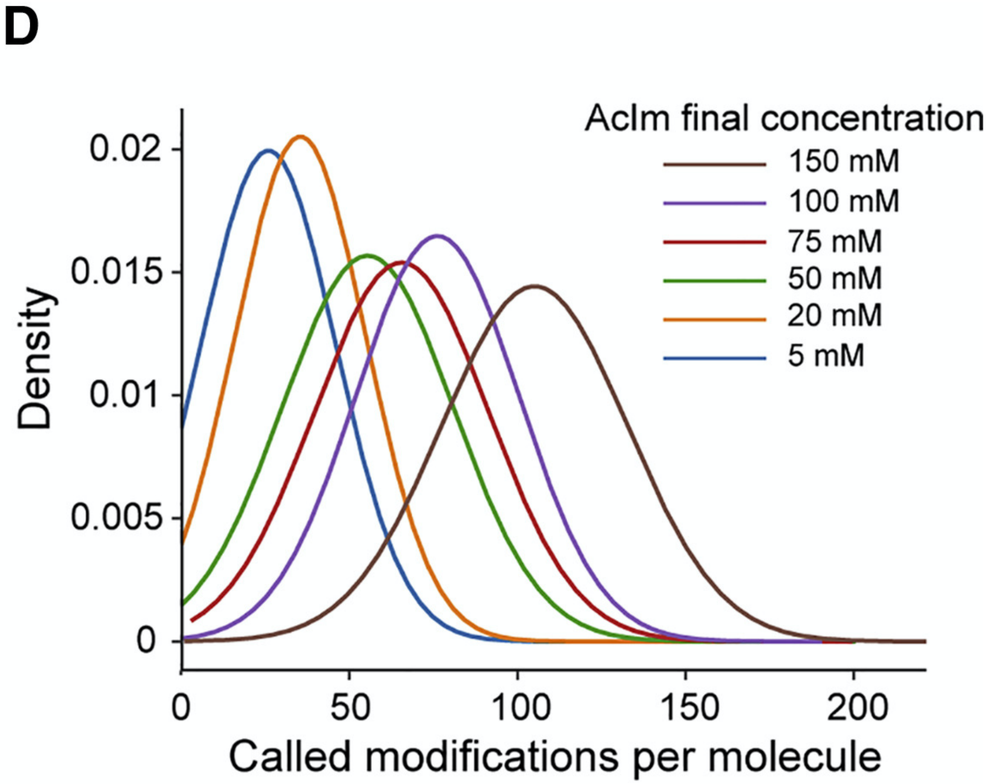
Next Steps

Probe Selection and Concentration

A



D



Acknowledgements

Direct Collaborators @IRCM:
Grégoire De Bisschop (RNA, Nanopore, and Shape Experiments)
Juan-Carlos Padilla (Nanopore Sequencing)

Supervisor: Jérôme Waldispühl

Waldispühl Group

Esteemed Collaborators:
Eric Lécuyer
Lécuyer Group
Yann Ponty
Vladimir Reinharz

Benasque Conference

McGill
School of Computer Science

IRCM
Montreal Clinical
Research Institute
Foundation

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture