In [1]:
```
config_file = "/users/kcochran/projects/new_procap_models/modisco_out/procap/K562/strand
```

In [2]:
```
# Parameters
config_file = "/users/kcochran/projects/new_procap_models/modisco_out/procap/K562/strand
```

In [3]:
```
import os
import numpy as np
import sys
sys.path.append("../2_train_models")
from utils import load_json
from report_utils import load_coords, load_modisco_results, report_motifs, plot_all_meta

config = load_json(config_file)

proj_dir = config["proj_dir"]

cell_type = config["cell_type"]
model_type = config["model_type"]
timestamp = config["timestamp"]
data_type = config["data_type"]

genome_path = config["genome_path"]
chrom_sizes = config["chrom_sizes"]

in_window = config["in_window"]
out_window = config["out_window"]

slice_len = config["slice"]

peak_path = config["train_val_peak_path"]

scores_path = config["scores_path"]

modisco_results_path = config["results_save_path"]
```

In [4]:
```
from modiscolite_utils import load_sequences, load_scores, load_observed_profiles

coords = load_coords(peak_path, in_window=in_window)

onehot_seqs = load_sequences(genome_path,
                             chrom_sizes,
                             peak_path,
                             slice_len=slice_len,
                             in_window=in_window)

scores = load_scores(scores_path,
                     slice_len=slice_len,
                     in_window=in_window)
```

```
OMP: Info #276: omp_set_nested routine deprecated, please use omp_set_max_active_levels
instead.
/users/kcochran/miniconda3/envs/procap_A100/lib/python3.9/site-packages/tqdm/auto.py:22:
TqdmWarning: IProgress not found. Please update jupyter and ipywidgets. See https://ipyw
idgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
Loading genome sequence from /mnt/lab_data2/kcochran/new_procap_models/genomes/hg38.with
rDNA.fasta
== In Extract Sequences ==
Peak filepath: /mnt/lab_data2/kcochran/new_procap_models/deepshap_out/procap/K562/strand
_merged_umap/2022-10-05_03-39-32_in/peaks_uni_and_bi_train_and_val.bed.gz
Sequence length: 2114
Num. Examples: 27000
```

In [5]:
```
modisco_results = load_modisco_results(modisco_results_path)
```

In [6]:
```
from file_configs import ValFilesConfig, TrainFilesConfig

# TODO: assert that we use the same peak file across all of these

val_config = ValFilesConfig(cell_type, model_type, timestamp, data_type)
train_config = TrainFilesConfig(cell_type, model_type, timestamp, data_type)

true_profs = load_observed_profiles(train_config.plus_bw_path,
                                    train_config.minus_bw_path,
                                    peak_path,
                                    slice_len=slice_len,
                                    out_window=out_window)

pred_profs = np.exp(np.load(val_config.pred_profiles_train_val_path))
```

```
Timestamp: 2022-10-05_03-39-32
Timestamp: 2022-10-05_03-39-32
== In Extract Profiles ==
Peak filepath: /mnt/lab_data2/kcochran/new_procap_models/deepshap_out/procap/K562/strand
_merged_umap/2022-10-05_03-39-32_in/peaks_uni_and_bi_train_and_val.bed.gz
Profile length: 1000
Num. Examples: 27000
```

In [7]:
```
len(coords), onehot_seqs.shape, scores.shape, true_profs.shape, pred_profs.shape
```

Out[7]:
```
(27000, (27000, 1000, 4), (27000, 1000, 4), (27000, 2, 1000), (27000, 2, 1000))
```

In [8]:
```
from IPython.display import HTML

report_html = report_motifs(modisco_results, proj_dir,
                            os.path.dirname(modisco_results_path))
HTML(report_html)
```

```
findfont: Font family ['Arial Rounded'] not found. Falling back to DejaVu Sans.
```

Out[8]:

pattern	num_seqlets	modisco_cwm_fwd	modisco_cwm_rev	match0
pos_patterns.pattern_0	8650			KLF12_HUMAN.H11MO.0.C
pos_patterns.pattern_1	8063			SIX2_MA1119.1
pos_patterns.pattern_2	5862			ELK4_MA0076.2
pos_patterns.pattern_3	4465			NFYA_MA0060.3
pos_patterns.pattern_4	3488			NRF1_MA0506.1
pos_patterns.pattern_5	2249			ATF3_HUMAN.H11MO.0.A
pos_patterns.pattern_6	1560			SP2_HUMAN.H11MO.0.A
pos_patterns.pattern_7	1050			THAP1_HUMAN.H11MO.0.C
pos_patterns.pattern_8	1011			NaN
pos_patterns.pattern_9	879			TBP_HUMAN.H11MO.0.A
pos_patterns.pattern_10	838			ZNF76_HUMAN.H11MO.0.C
pos_patterns.pattern_11	833			SP2_HUMAN.H11MO.0.A
pos_patterns.pattern_12	760			THAP1_HUMAN.H11MO.0.C
pos_patterns.pattern_13	747			ATF3_MOUSE.H11MO.0.A

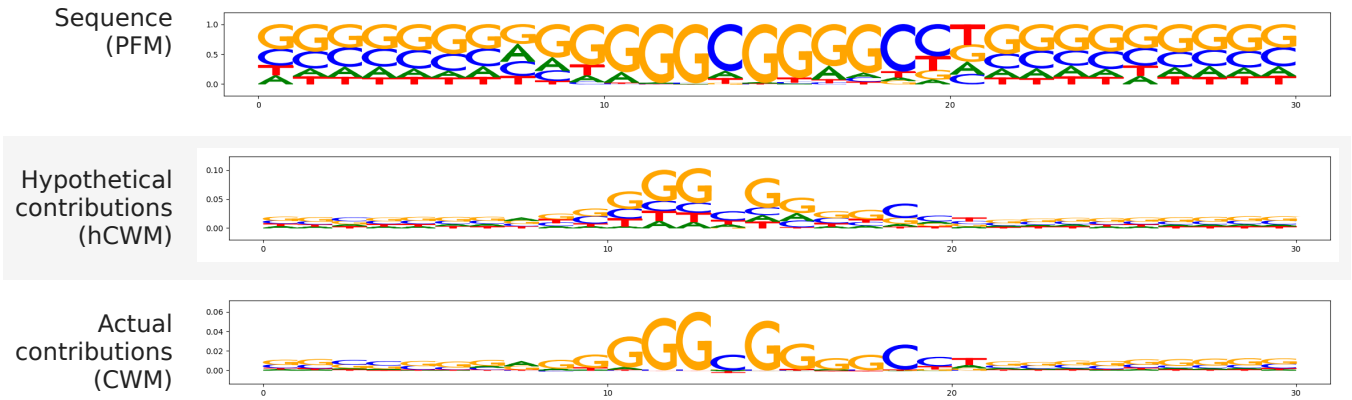
pos_patterns.pattern_14	535		CTCF_MOUSE.H11MO.0.A
pos_patterns.pattern_15	410		ZBTB33_MA0527.1
pos_patterns.pattern_16	401		NRF1_MOUSE.H11MO.0.A
pos_patterns.pattern_17	219		MYBL1_MYB_1
pos_patterns.pattern_18	198		CPEB1_RRM_1
pos_patterns.pattern_19	190		CTCFL_HUMAN.H11MO.0.A
pos_patterns.pattern_20	163		ZN770_HUMAN.H11MO.0.C
pos_patterns.pattern_21	153		ELF2_MOUSE.H11MO.0.C
pos_patterns.pattern_22	148		ZFX_MOUSE.H11MO.0.B
pos_patterns.pattern_23	135		ZN816_HUMAN.H11MO.0.C
pos_patterns.pattern_24	130		GATA5_GATA_1
pos_patterns.pattern_25	113		ZBTB33_MA0527.1
pos_patterns.pattern_26	107		ZNF76_HUMAN.H11MO.0.C
pos_patterns.pattern_27	105		SP1_MOUSE.H11MO.0.A
pos_patterns.pattern_28	105		ZN770_HUMAN.H11MO.0.C
pos_patterns.pattern_29	82		THAP1_HUMAN.H11MO.0.C
pos_patterns.pattern_30	67		SREBF1_MA0595.1
pos_patterns.pattern_31	63		RFX3_MOUSE.H11MO.0.C
pos_patterns.pattern_32	54		THAP1_HUMAN.H11MO.0.C
pos_patterns.pattern_33	49		RUNX2_RUNX_1
pos_patterns.pattern_34	41		TEAD1_HUMAN.H11MO.0.A
pos_patterns.pattern_35	40		PRDM1_MA0508.2
pos_patterns.pattern_36	39		NaN
pos_patterns.pattern_37	34		ATF1_HUMAN.H11MO.0.B
pos_patterns.pattern_38	24		NaN
neg_patterns.pattern_0	87		KLF12_HUMAN.H11MO.0.C
neg_patterns.pattern_1	61		Gabpa_MA0062.2
neg_patterns.pattern_2	42		JUND_MA0491.1
neg_patterns.pattern_3	25		NRF1_MOUSE.H11MO.0.A

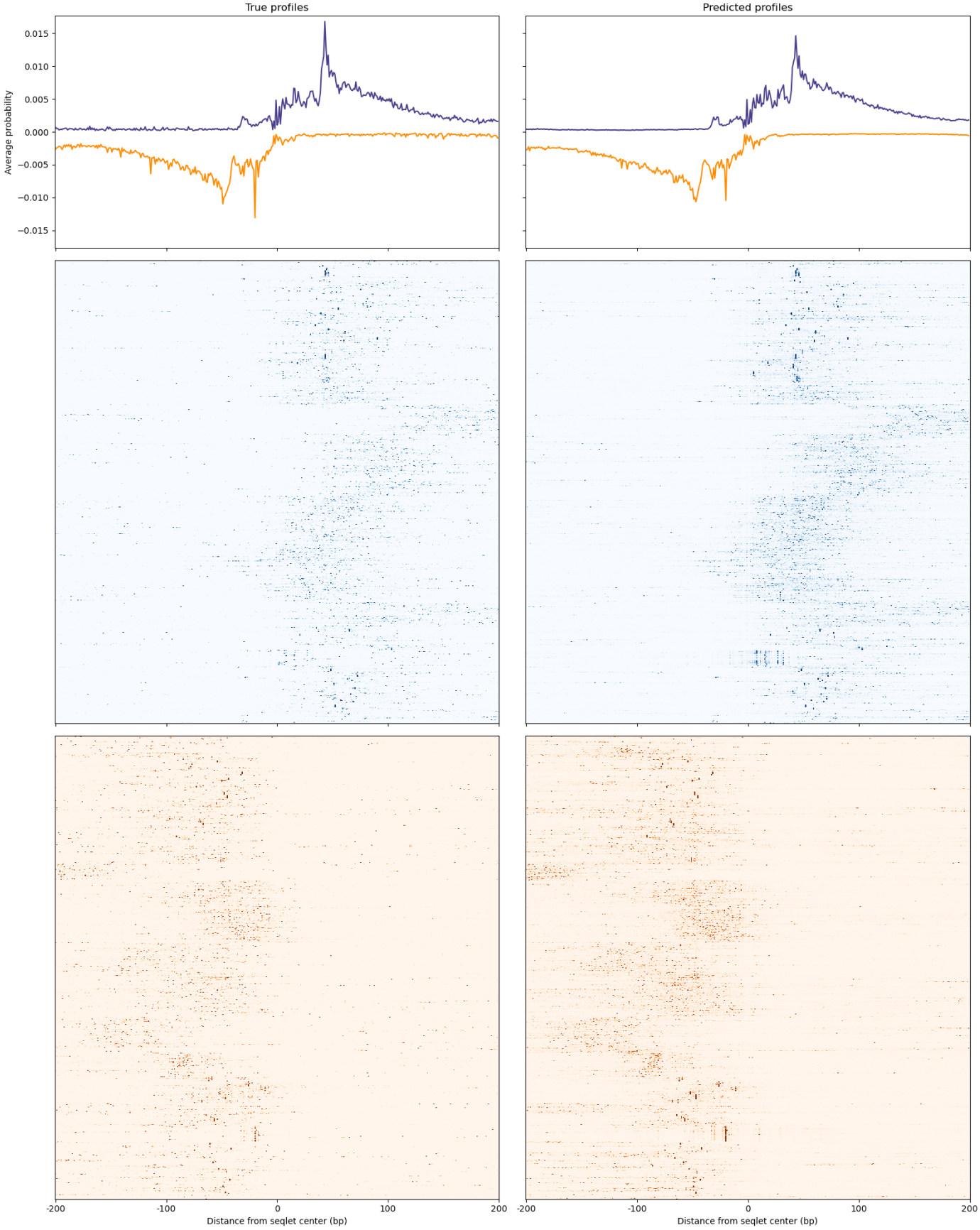
In [9]:

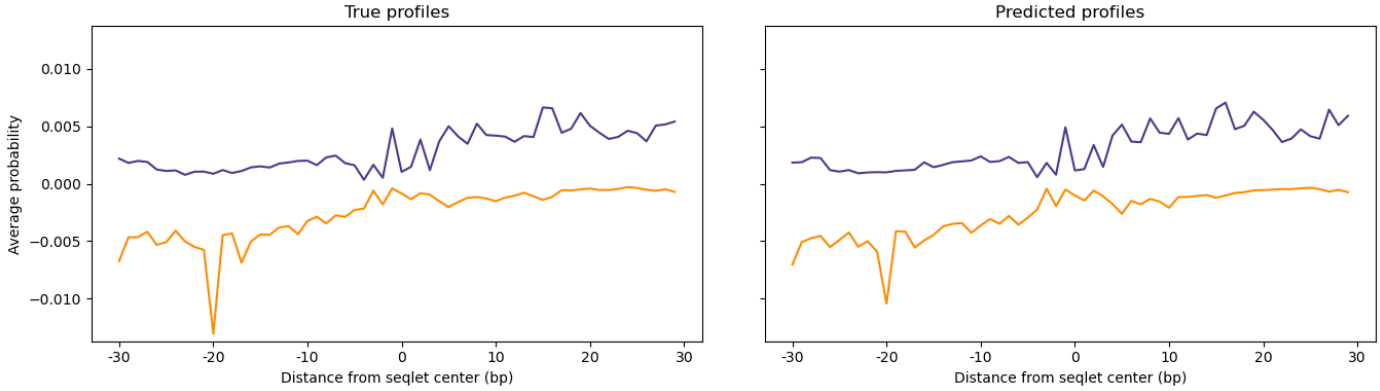
```
%matplotlib inline

plot_all_metaclusters(modisco_results, onehot_seqs, scores, true_profs, pred_profs, coor
                 in_window, out_window, slice_len, 400)
```

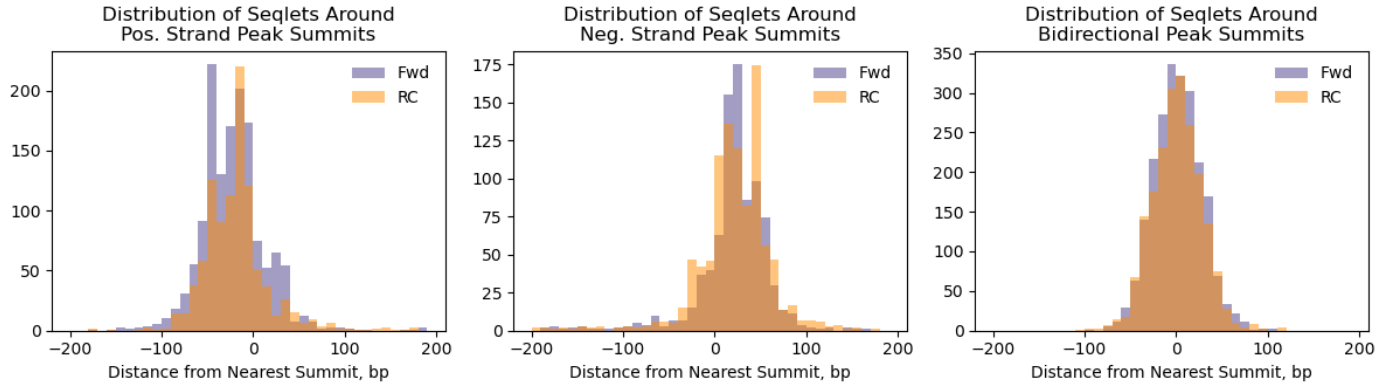
Pattern 0/39

8650 seqlets


Sequence (PFM)
Hypothetical contributions (hCWM)
Actual contributions (CWM)


True profiles
Predicted profiles
Average probability
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
-200
-100
0
100
200
Distance from seqlet center (bp)
Distance from seqlet center (bp)

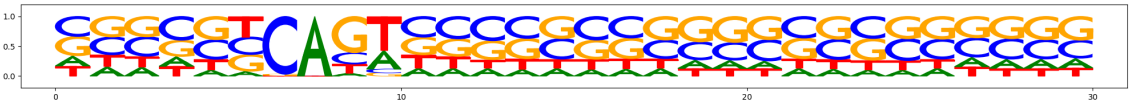




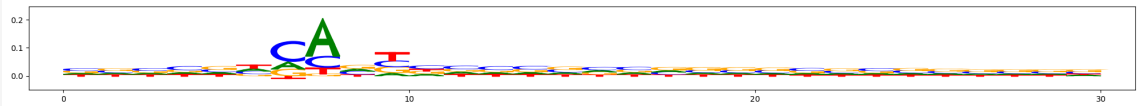
Pattern 1/39

8063 seqlets

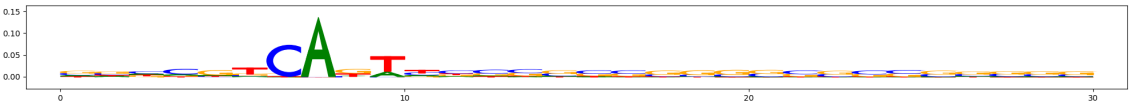
Sequence (PFM)

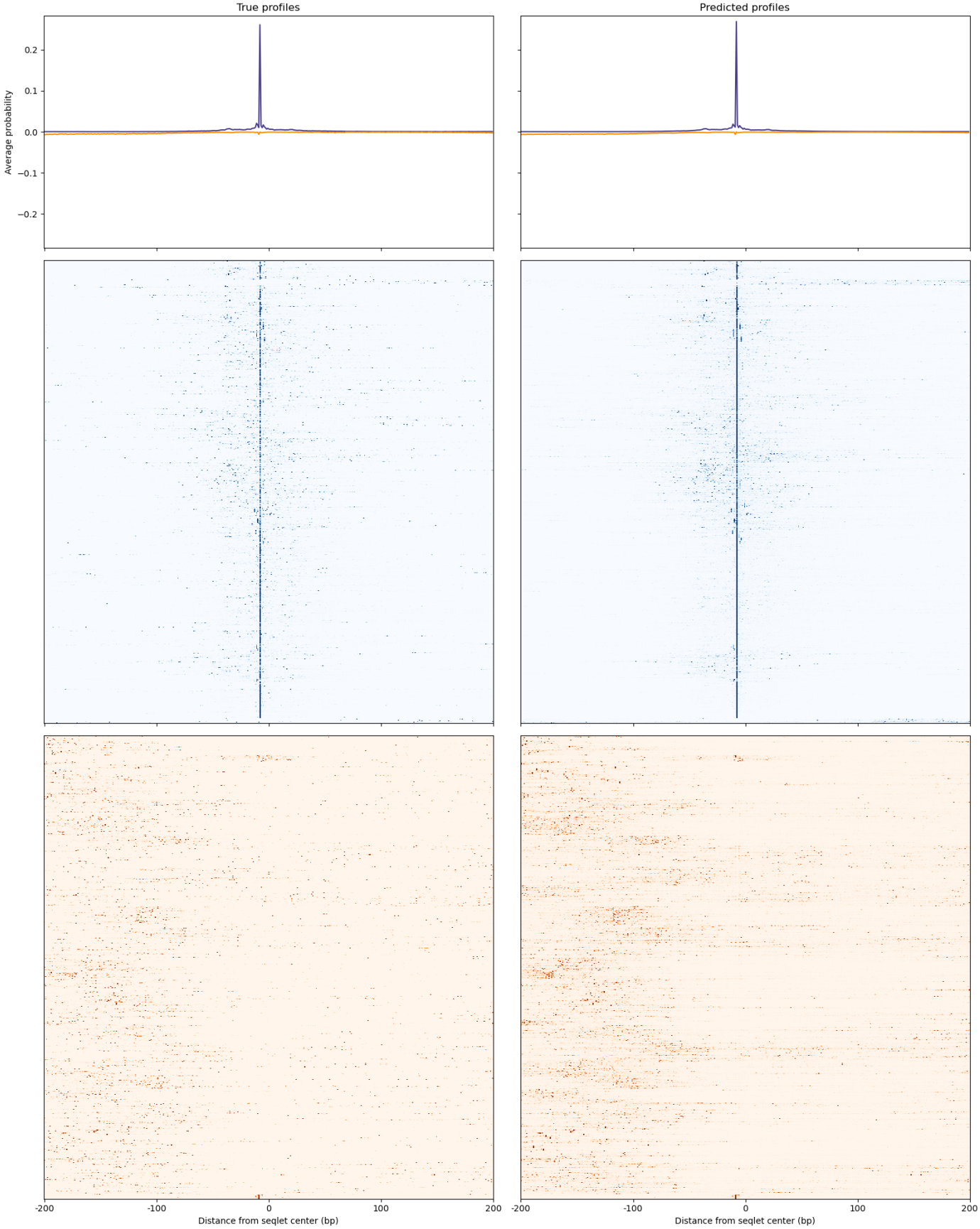


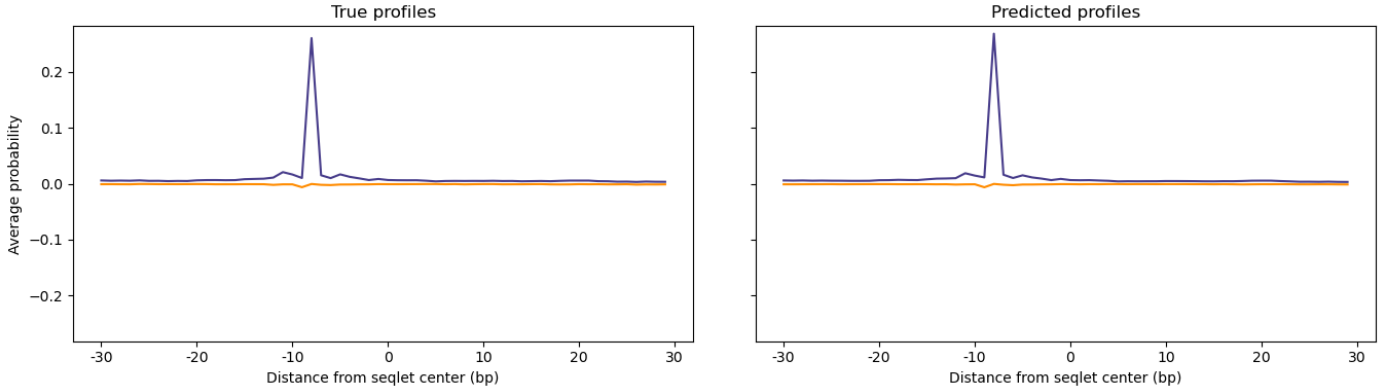
Hypothetical contributions (hCWM)

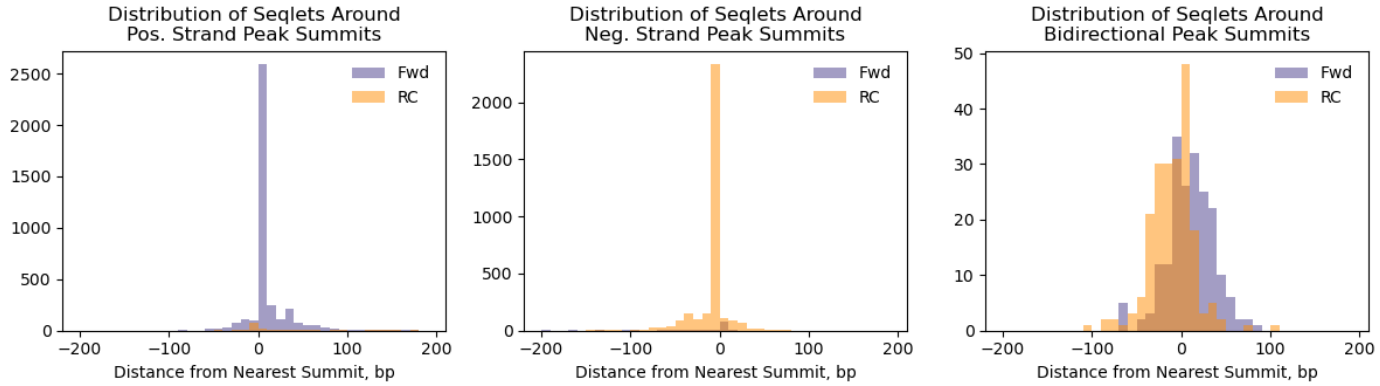


Actual contributions (CWM)




True profiles
Predicted profiles
Average probability
0.2
0.1
0.0
−0.1
−0.2
-200
-100
0
100
200
Distance from seqlet center (bp)





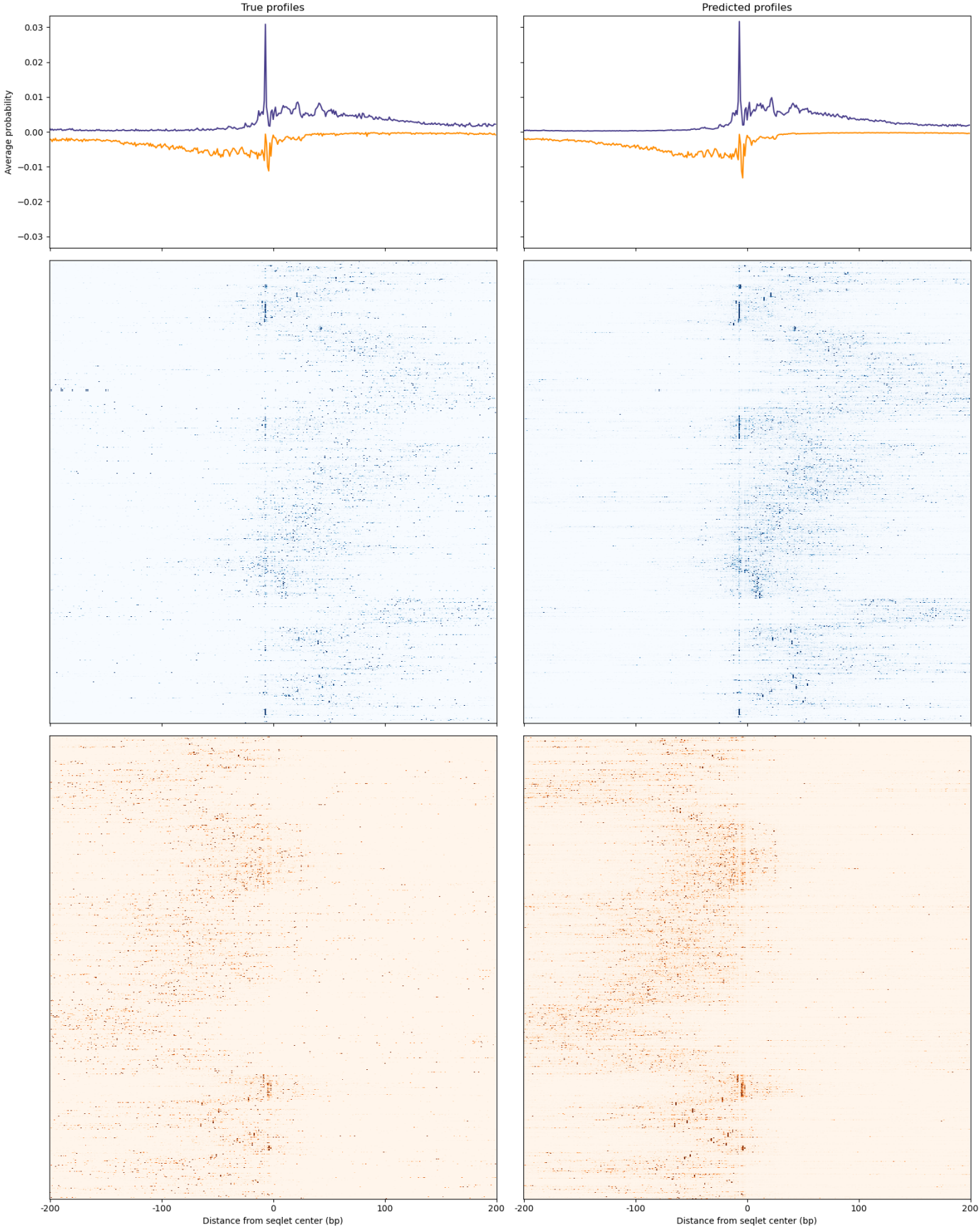
Pattern 2/39

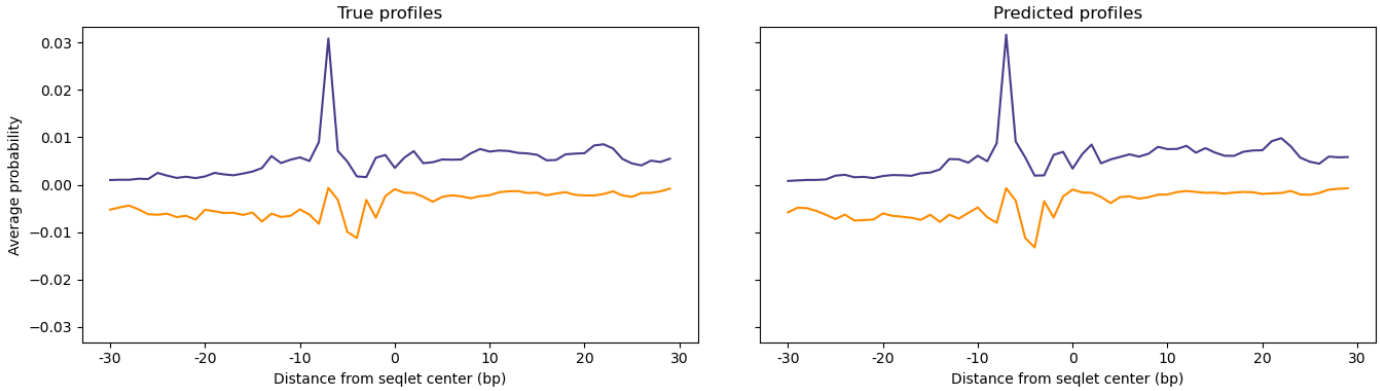
5862 seqlets

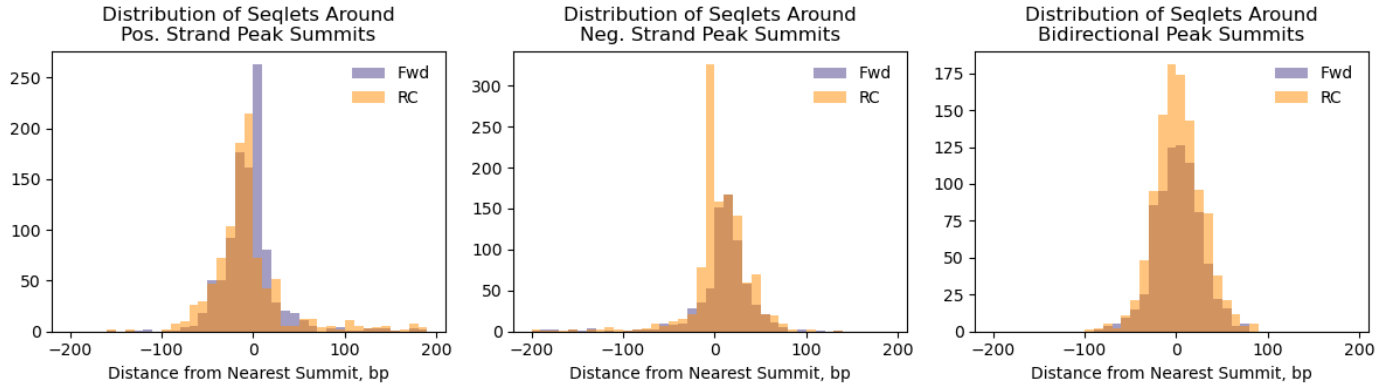
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)


True profiles
Predicted profiles
Average probability
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
-200
-100
0
100
200
Distance from seqlet center (bp)
Distance from seqlet center (bp)



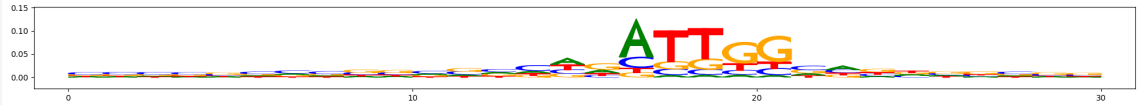


Pattern 3/39

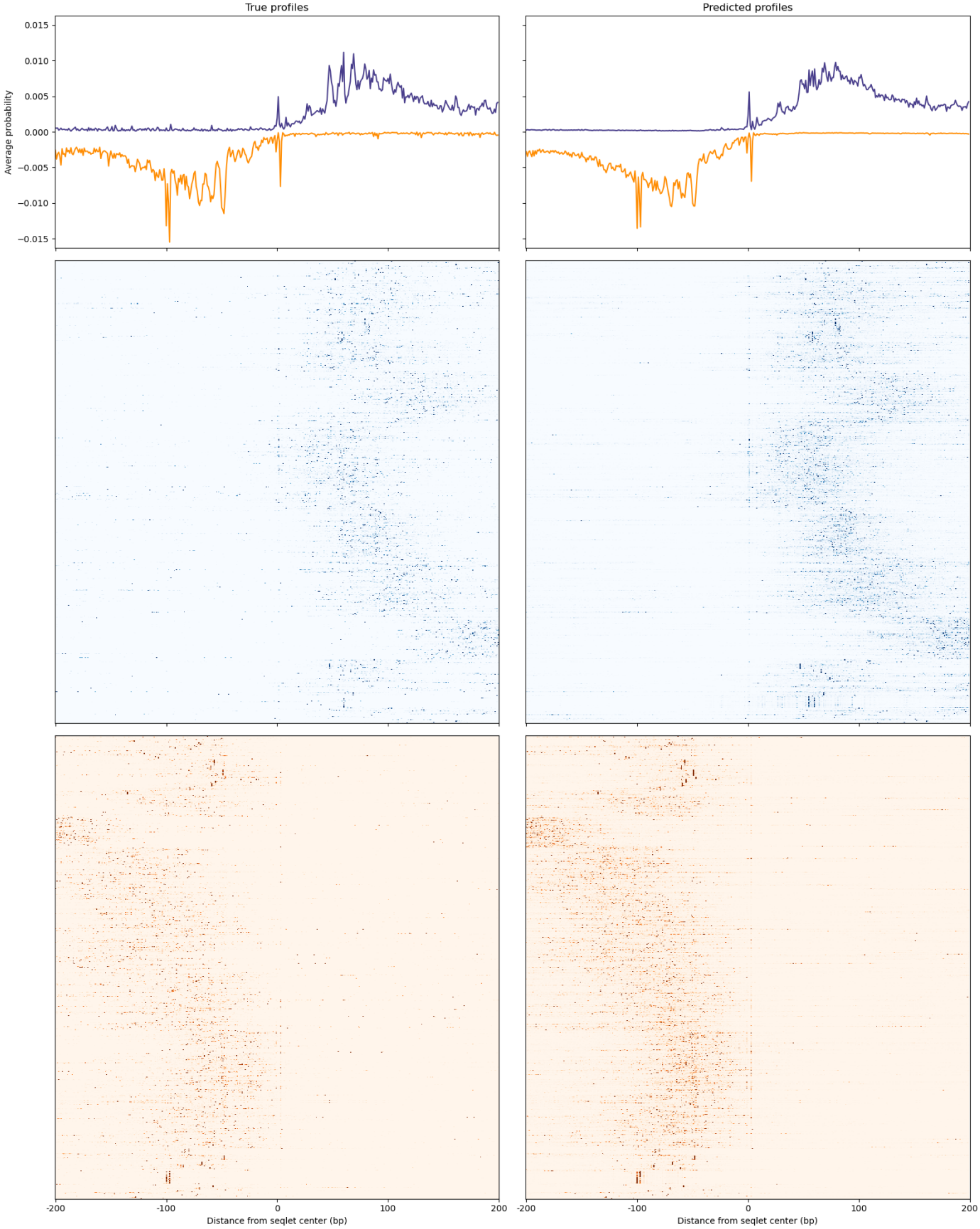
4465 seqlets

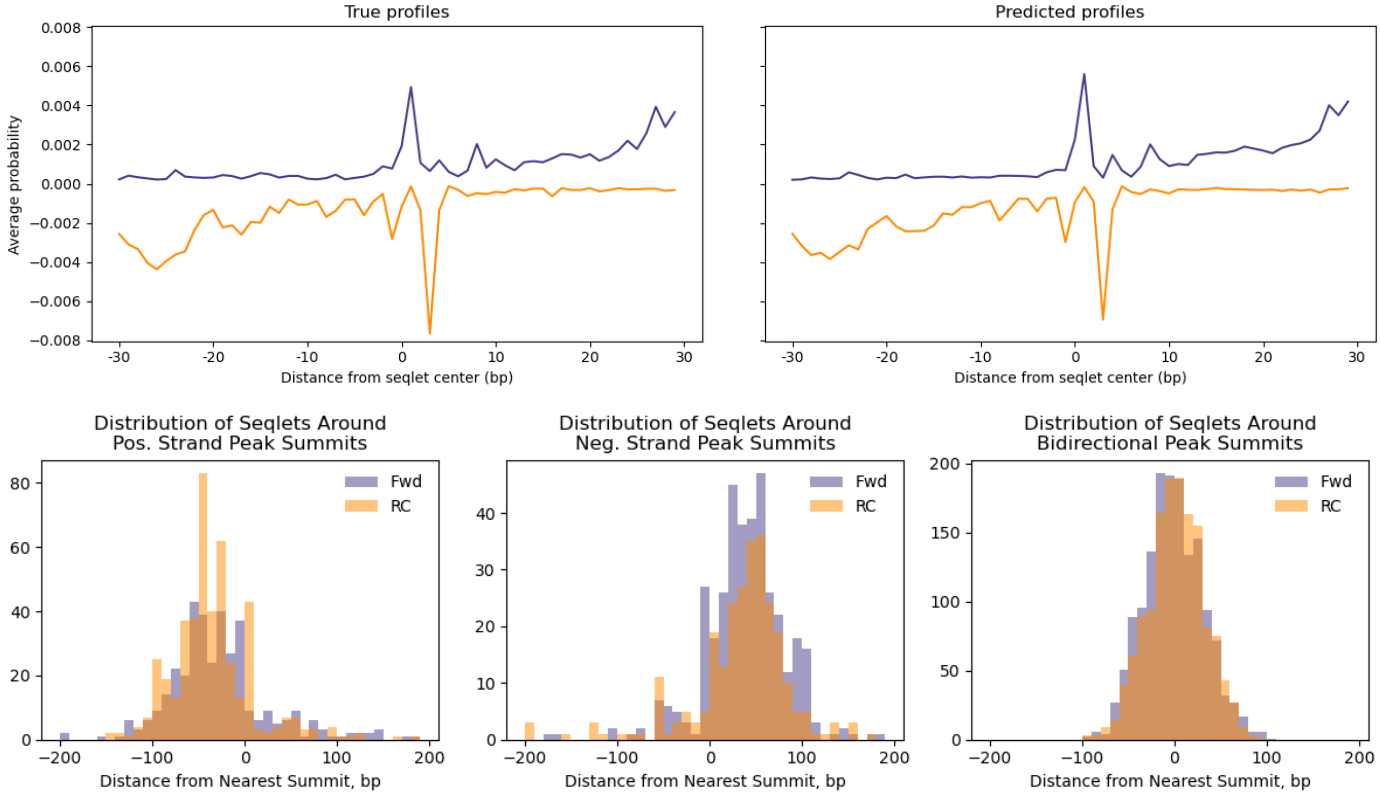
Sequence (PFM)

Hypothetical contributions (hCWM)



Actual contributions (CWM)


True profiles
Predicted profiles
Average probability
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
-200
-100
0
100
200
Distance from seqlet center (bp)
Distance from seqlet center (bp)



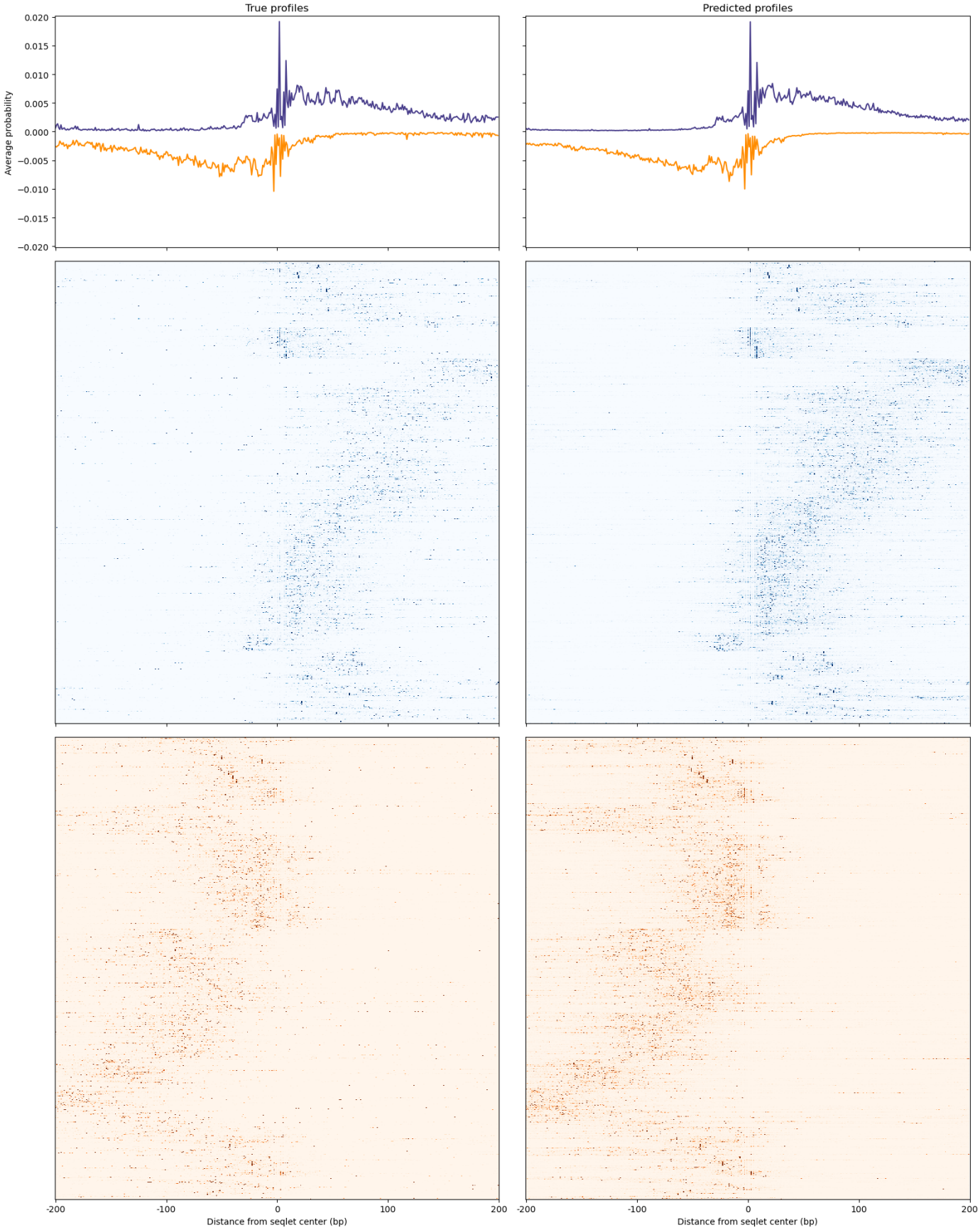
Pattern 4/39

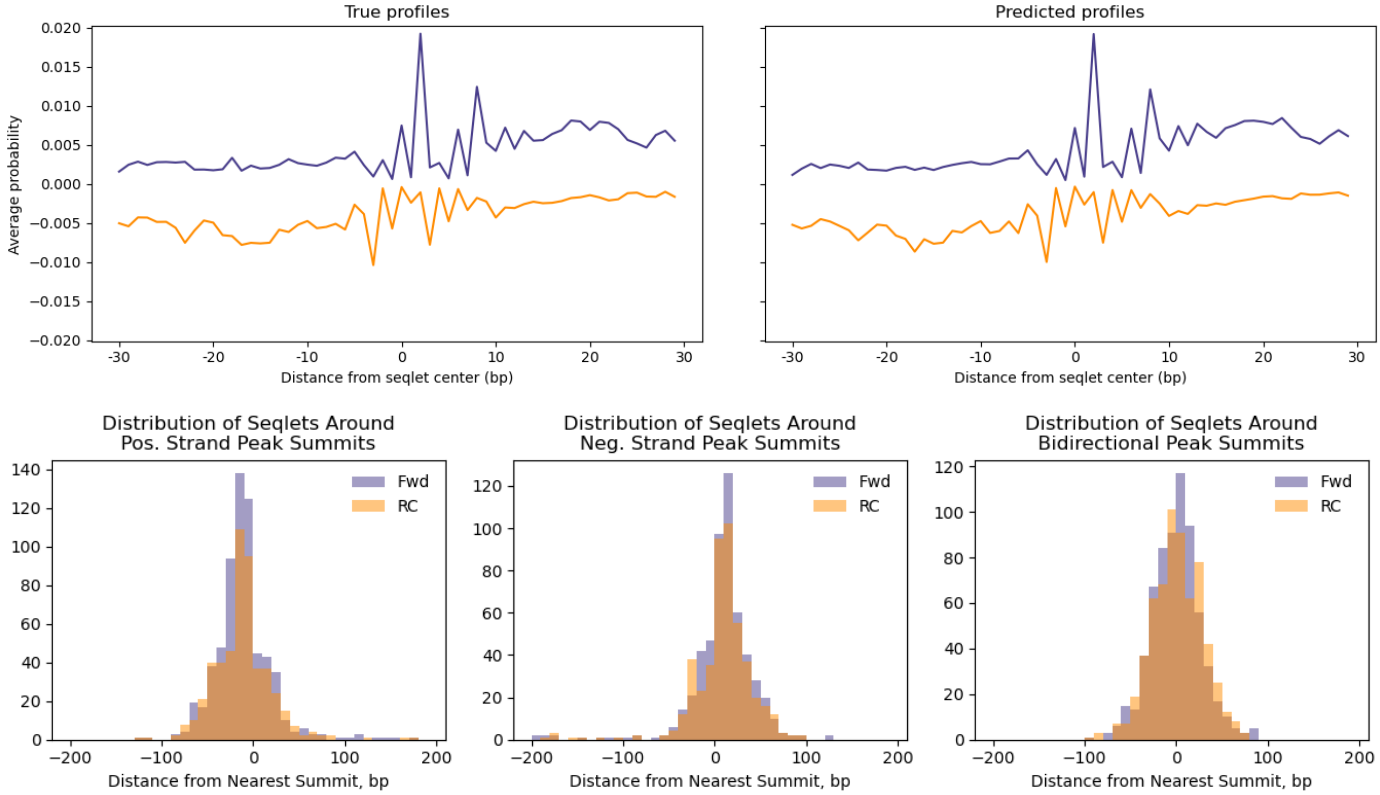
3488 seqlets

Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)


True profiles
Predicted profiles
Average probability
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
-200
-100
0
100
200
Distance from seqlet center (bp)
Distance from seqlet center (bp)

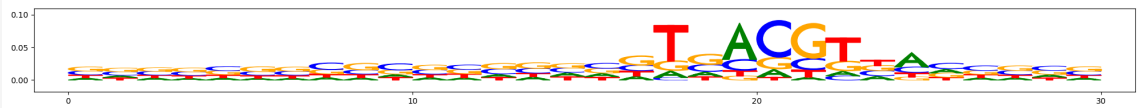


Pattern 5/39

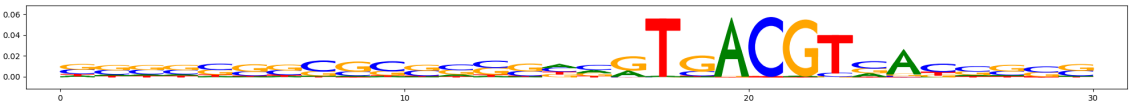
2249 seqlets

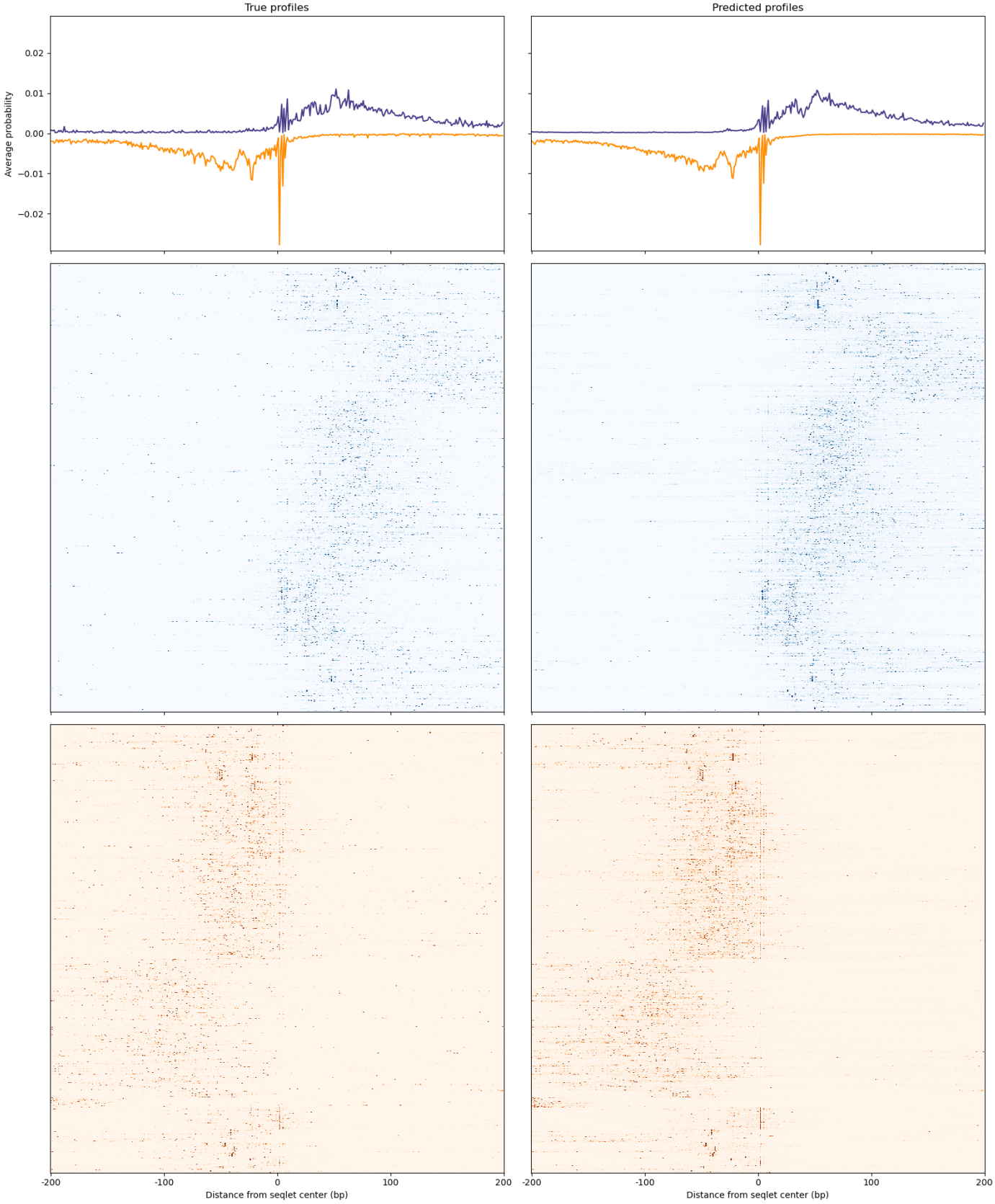
Sequence (PFM)

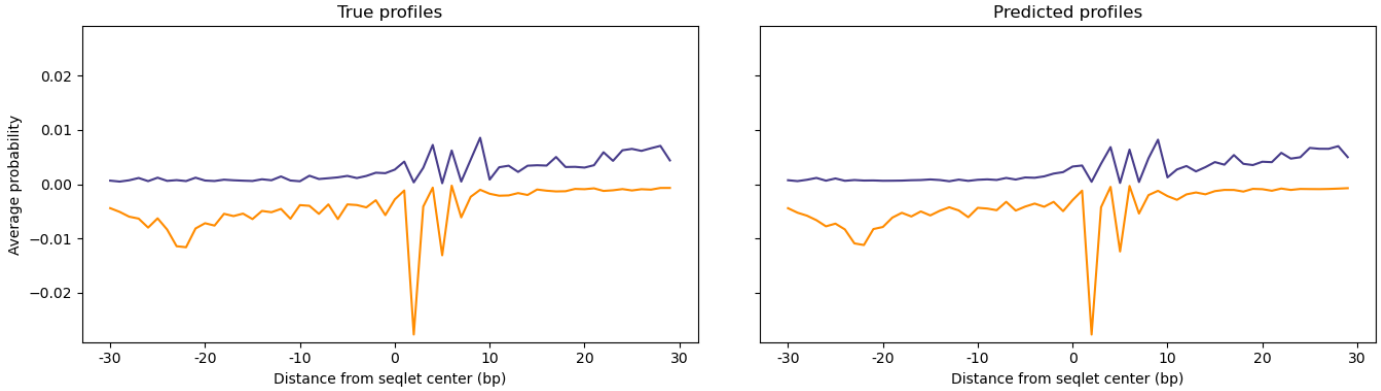
Hypothetical contributions (hCWM)

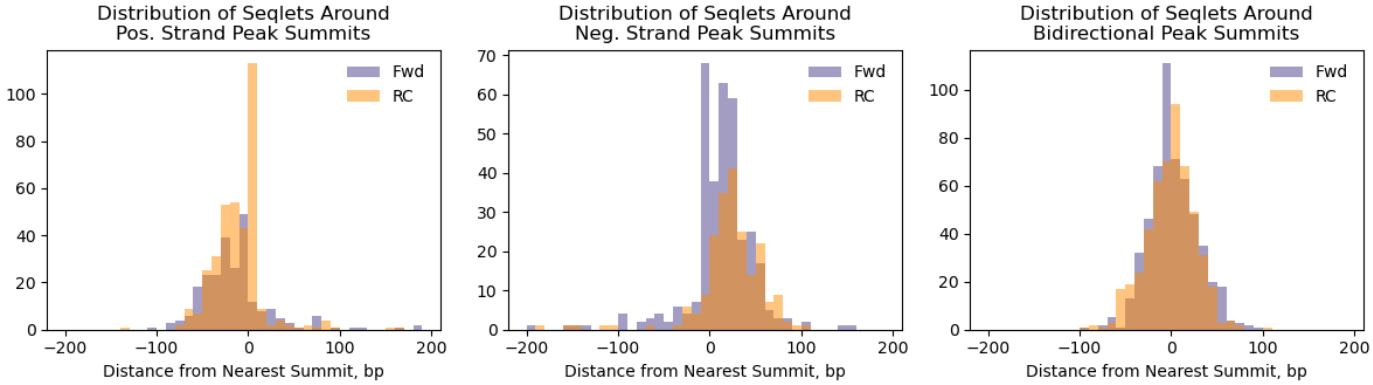


Actual contributions (CWM)



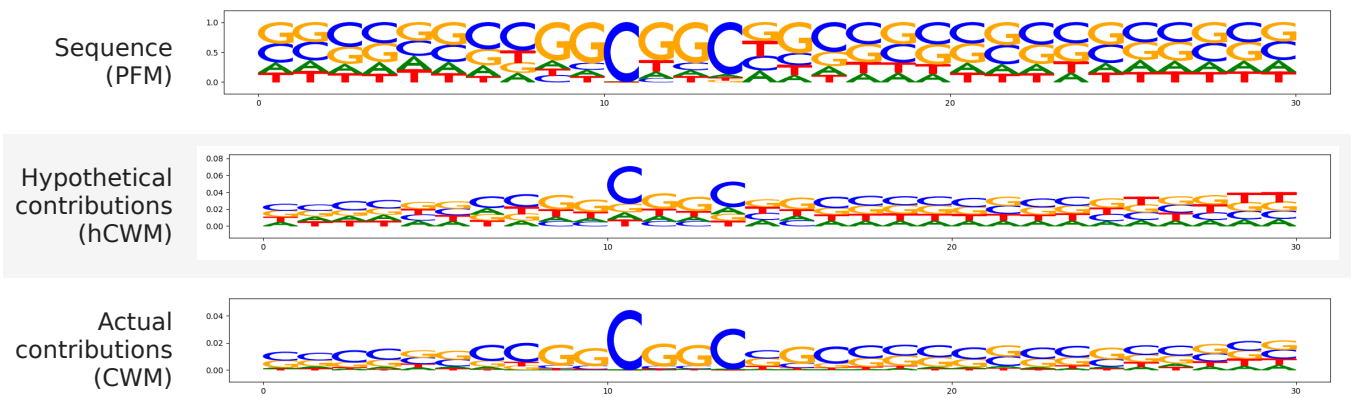

True profiles
Predicted profiles
Average probability
0.02
0.01
0.00
−0.01
−0.02
-200
-100
0
100
200
Distance from seqlet center (bp)

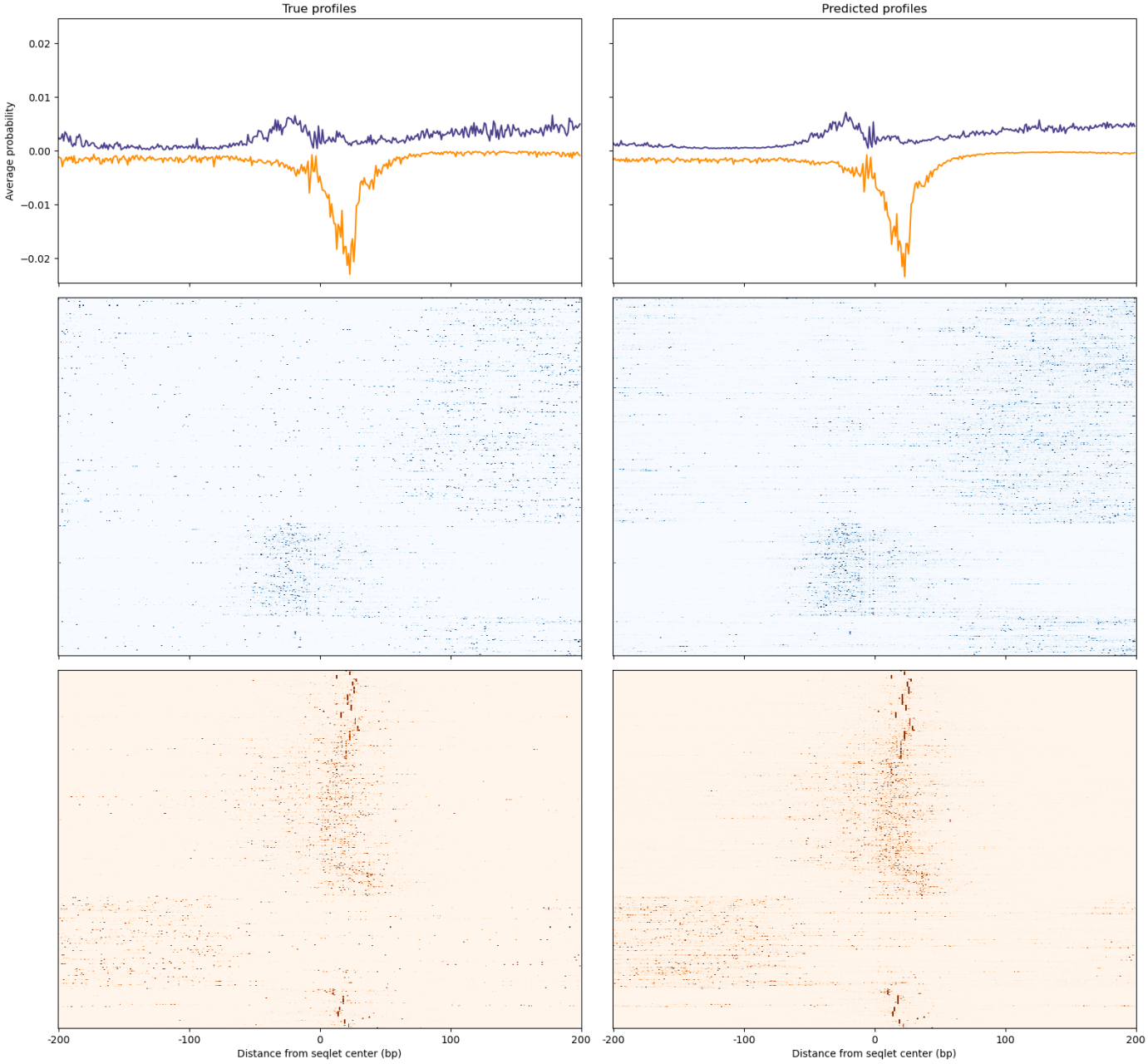

True profiles
Predicted profiles
Average probability
0.02
0.01
0.00
−0.01
−0.02
-30
-20
-10
0
10
20
30
Distance from seqlet center (bp)

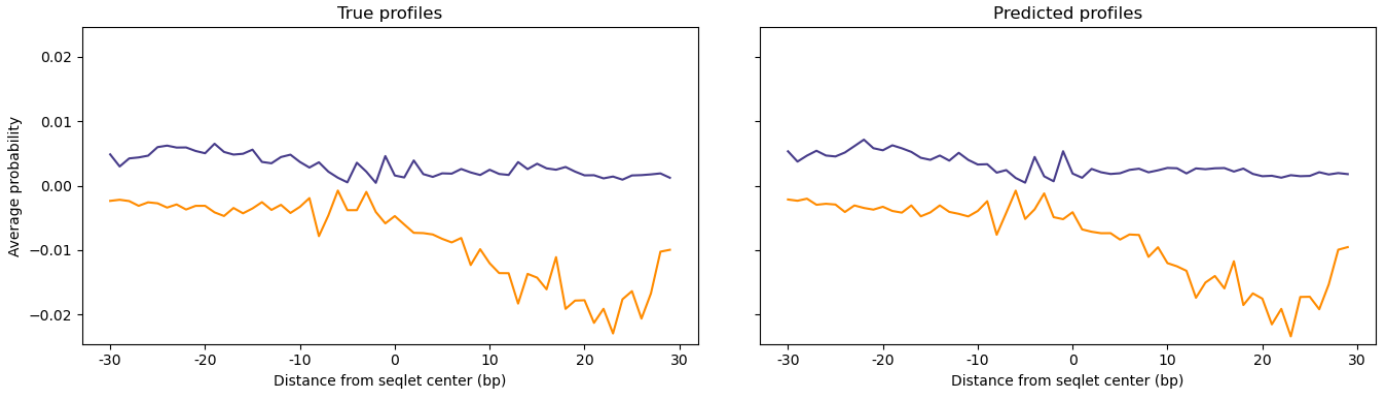


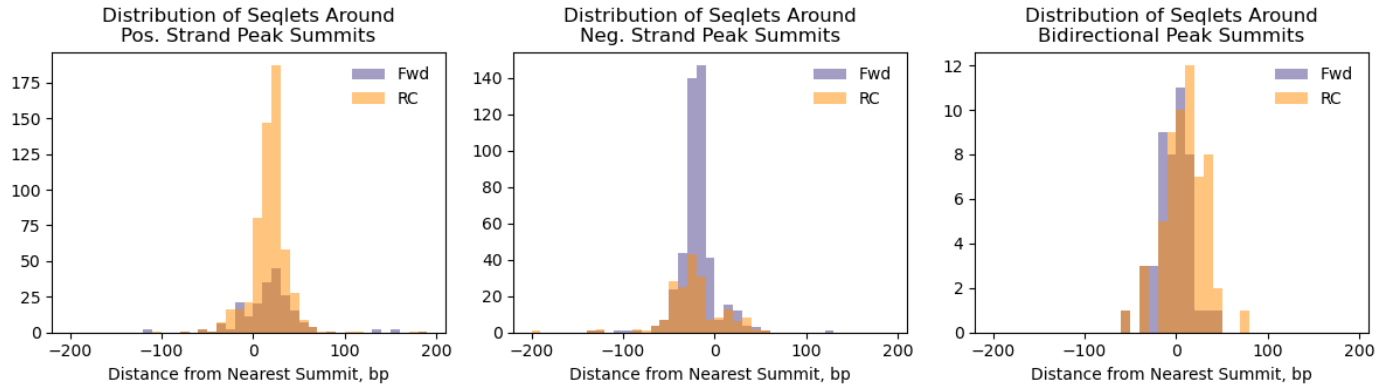
Pattern 6/39

1560 seqlets




True profiles
Predicted profiles
Average probability
0.02
0.01
0.00
−0.01
−0.02
-200
-100
0
100
200
Distance from seqlet center (bp)


True profiles
Predicted profiles
Average probability
0.02
0.01
0.00
−0.01
−0.02
-30
-20
-10
0
10
20
30
Distance from seqlet center (bp)


Distribution of Seqlets Around Pos. Strand Peak Summits
Distribution of Seqlets Around Neg. Strand Peak Summits
Distribution of Seqlets Around Bidirectional Peak Summits
Fwd
RC
−200
−100
0
100
200
Distance from Nearest Summit, bp

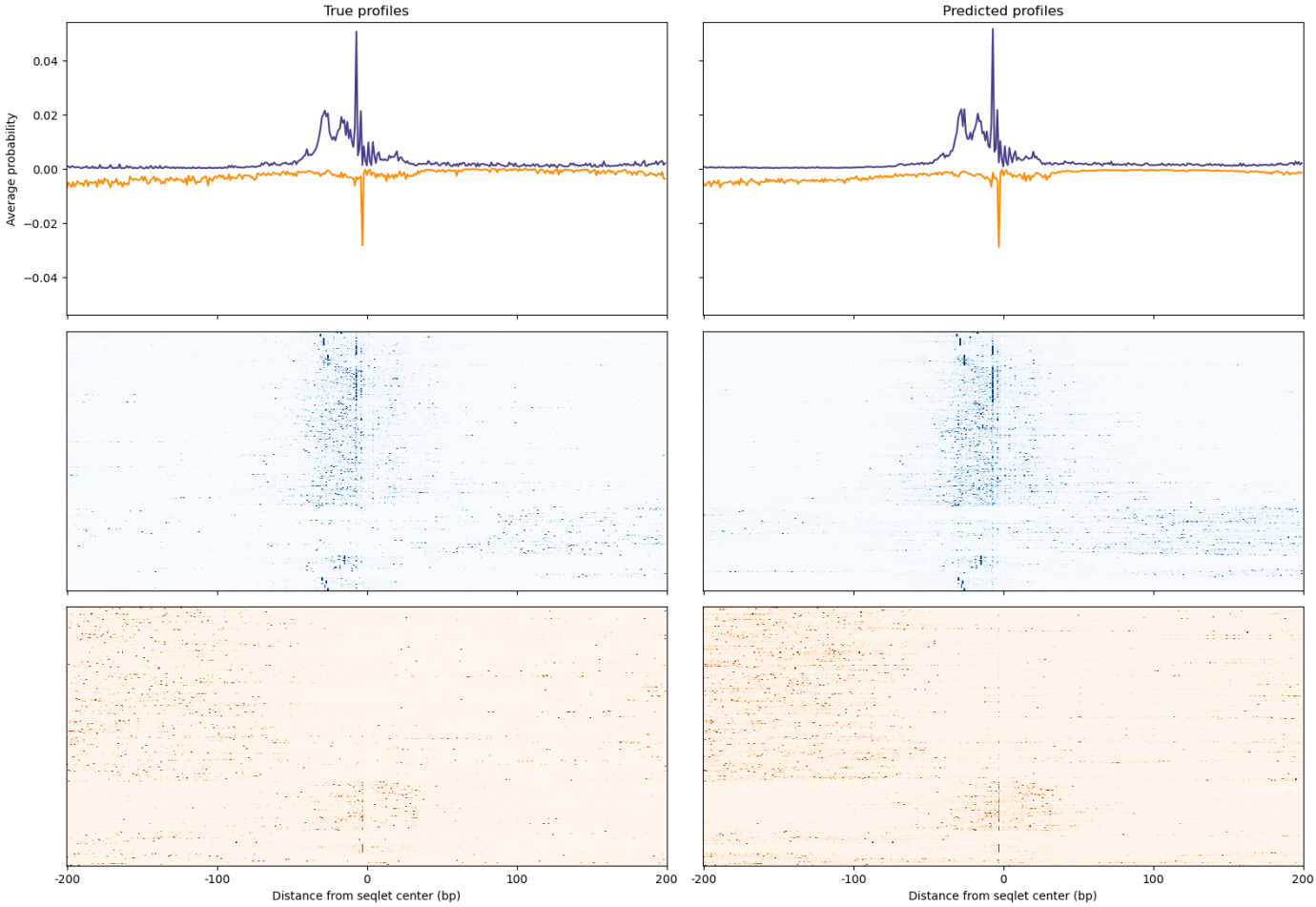
Pattern 7/39

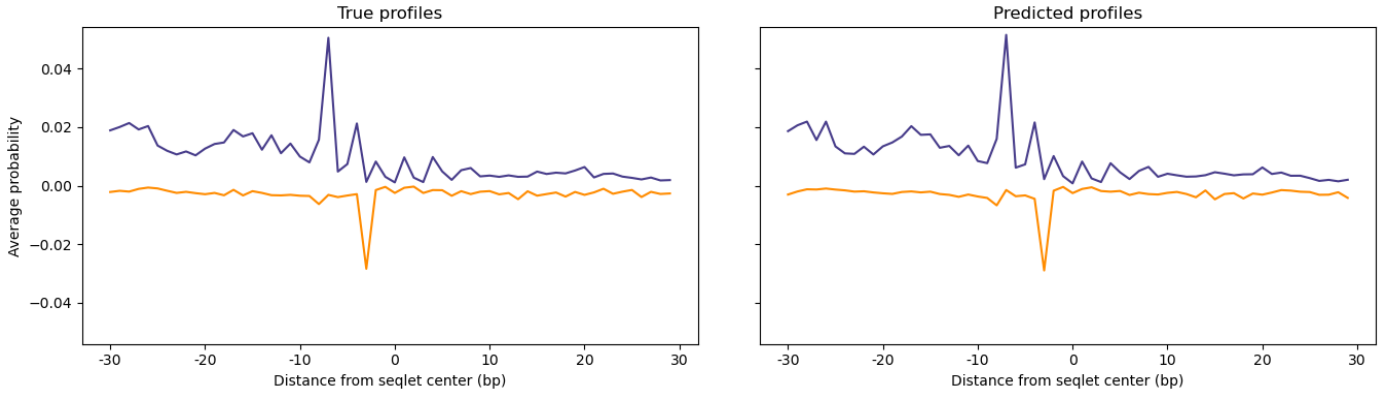
1050 seqlets

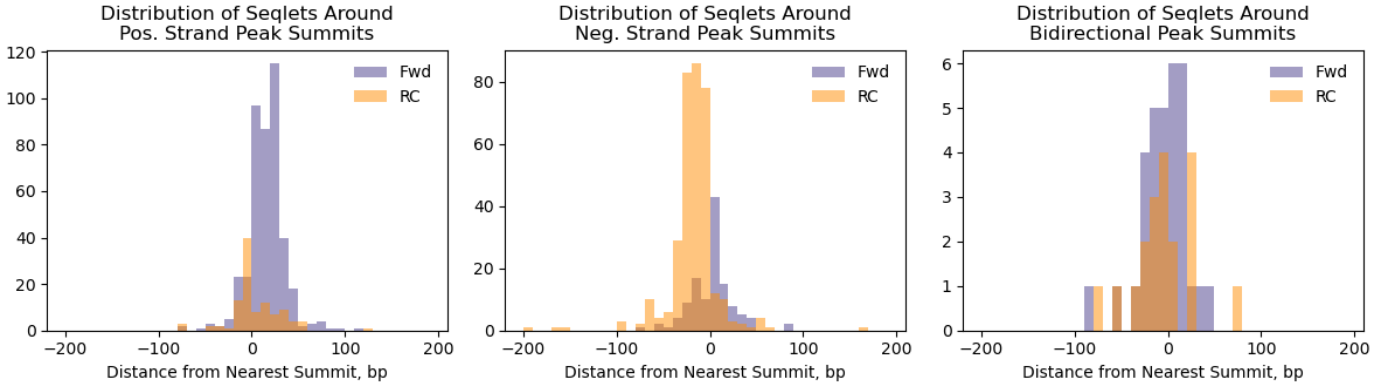
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





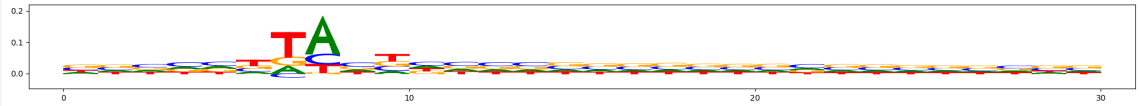


Pattern 8/39

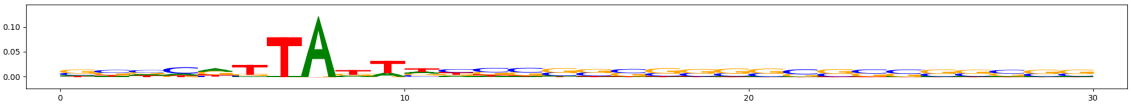
1011 seqlets

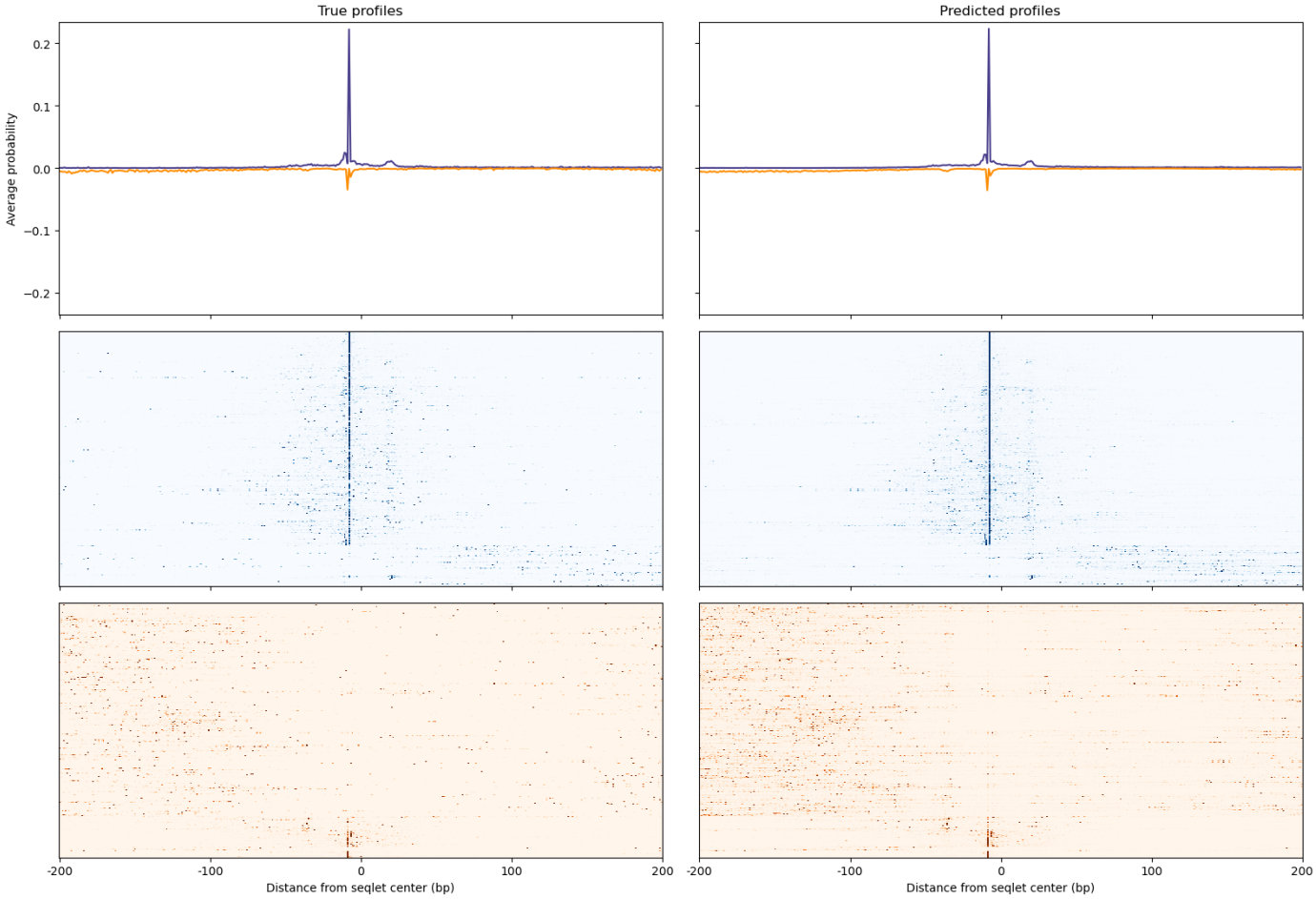
Sequence (PFM)

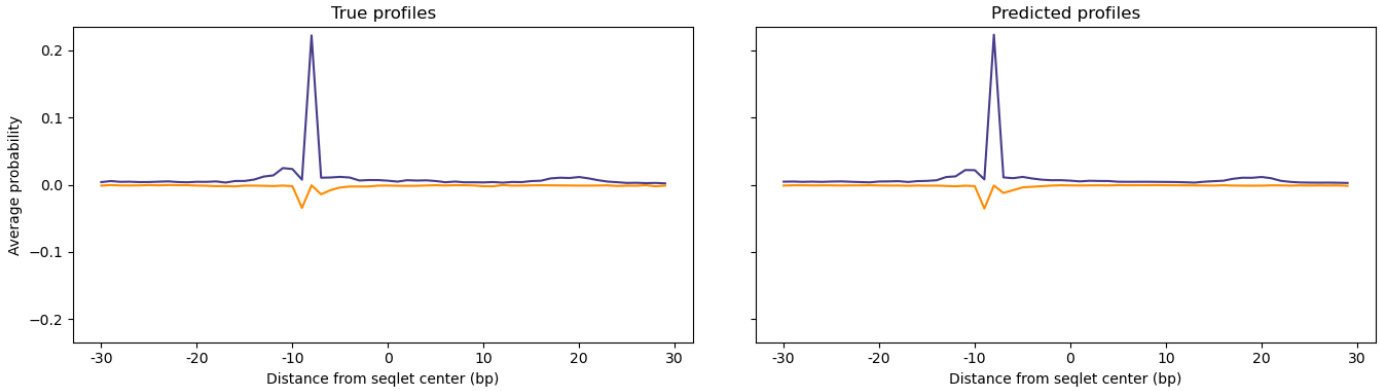
Hypothetical contributions (hCWM)

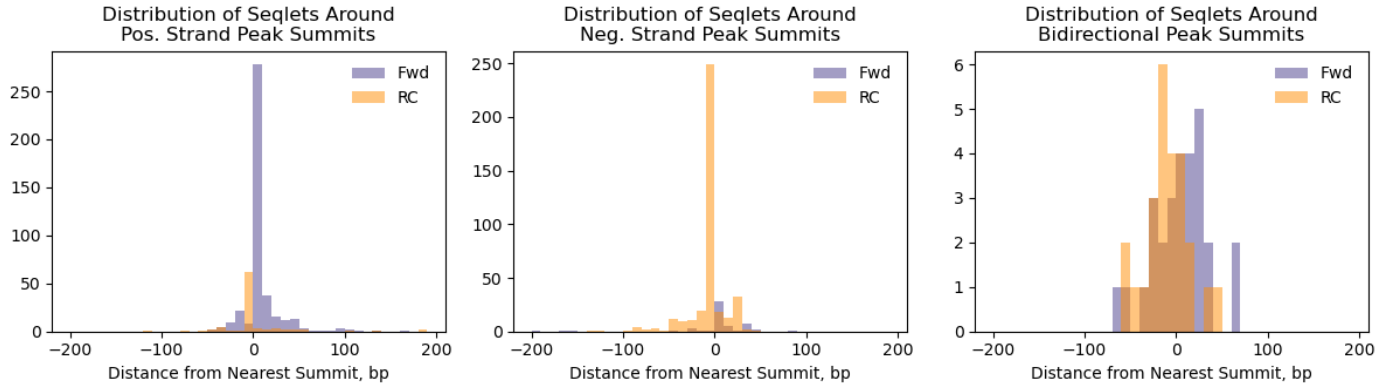


Actual contributions (CWM)







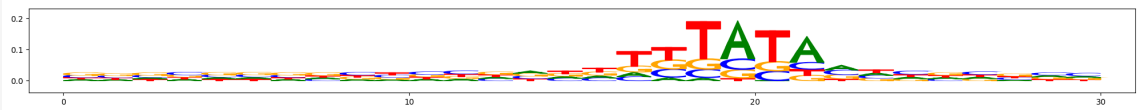


Pattern 9/39

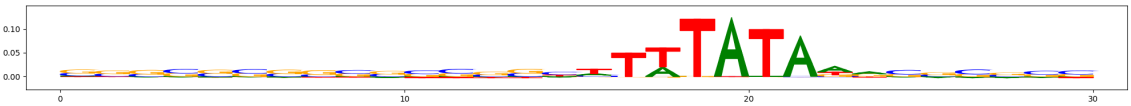
879 seqlets

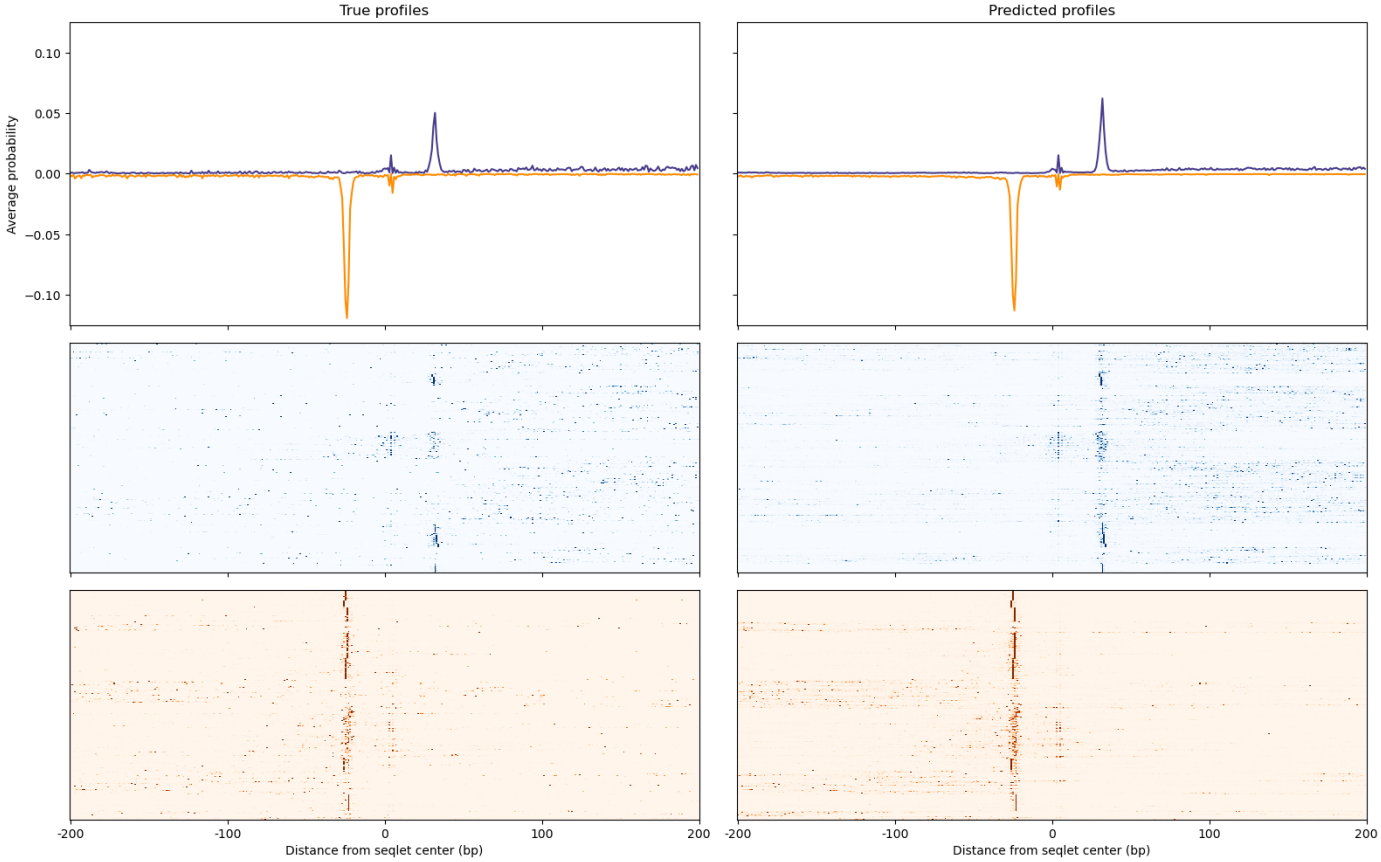
Sequence (PFM)

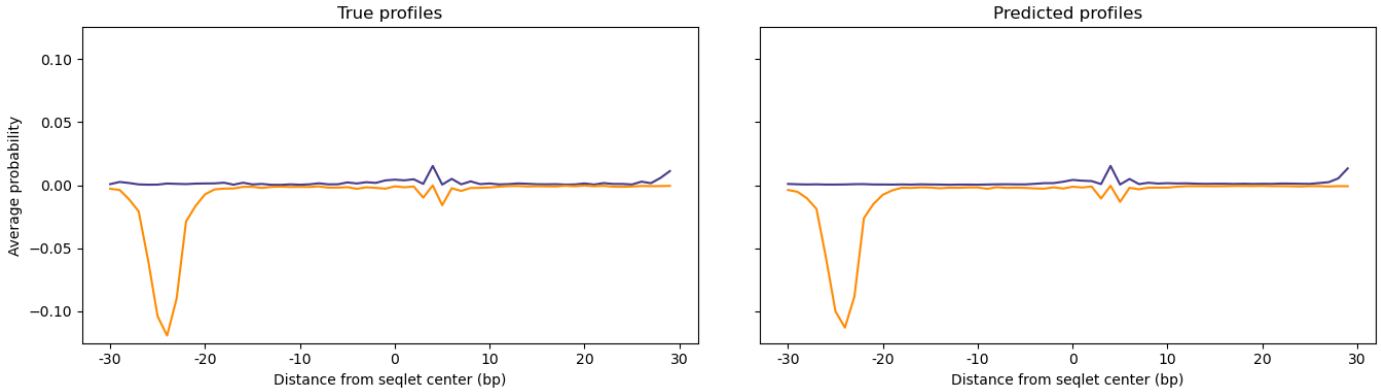
Hypothetical contributions (hCWM)

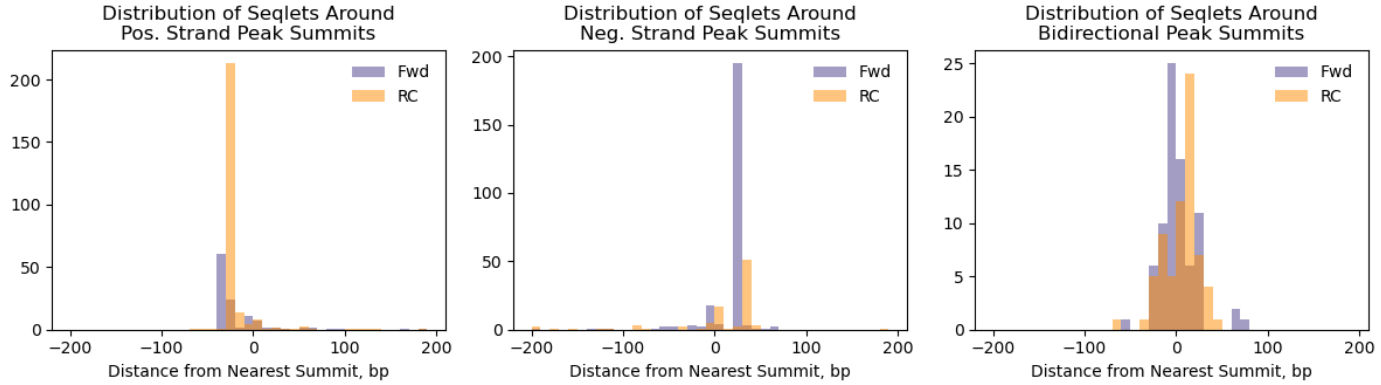


Actual contributions (CWM)









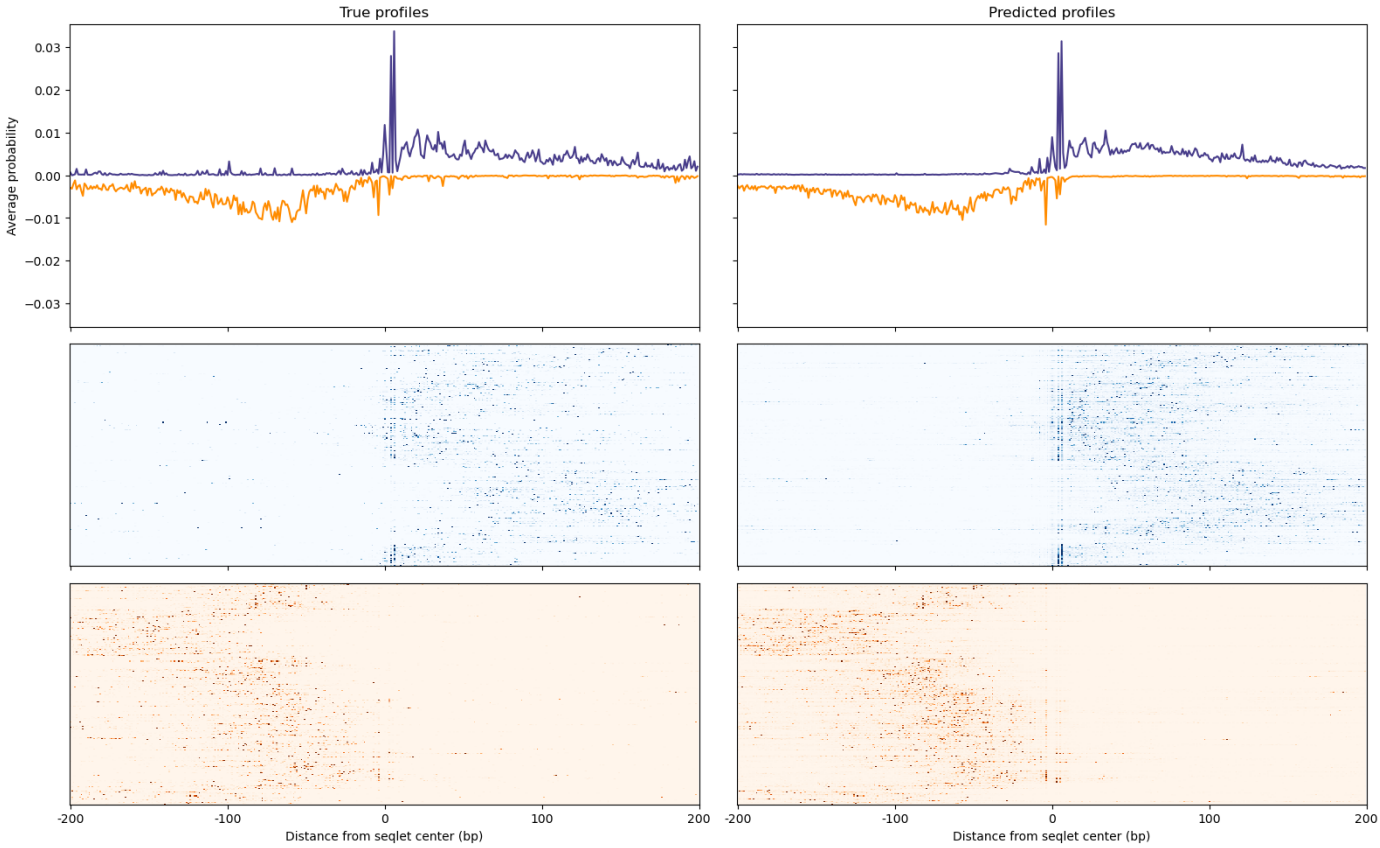
Pattern 10/39

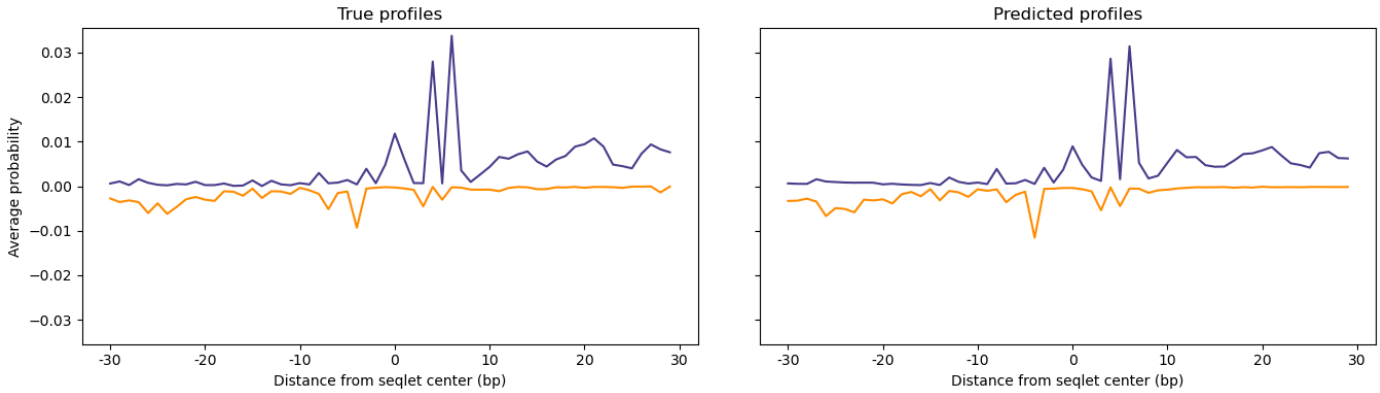
838 seqlets

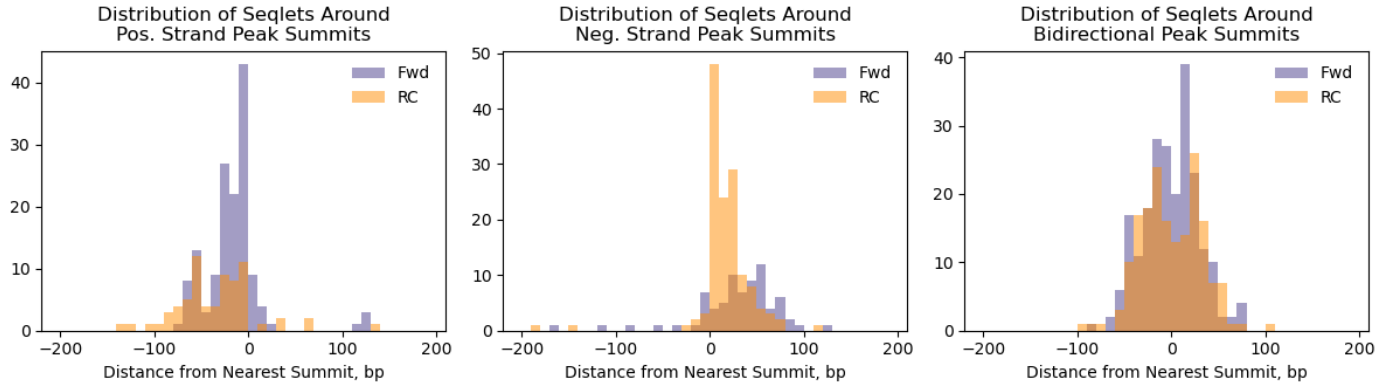
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







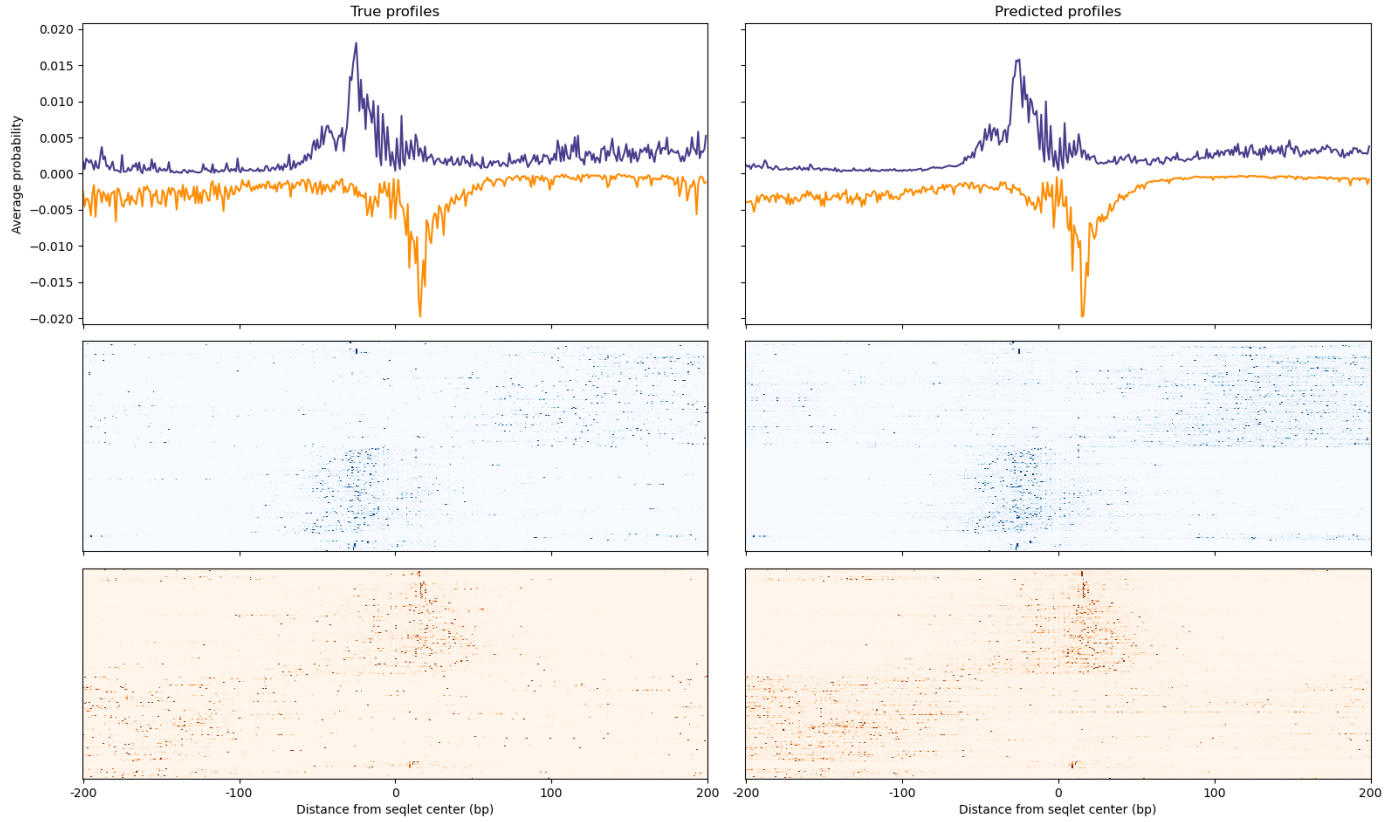
Pattern 11/39

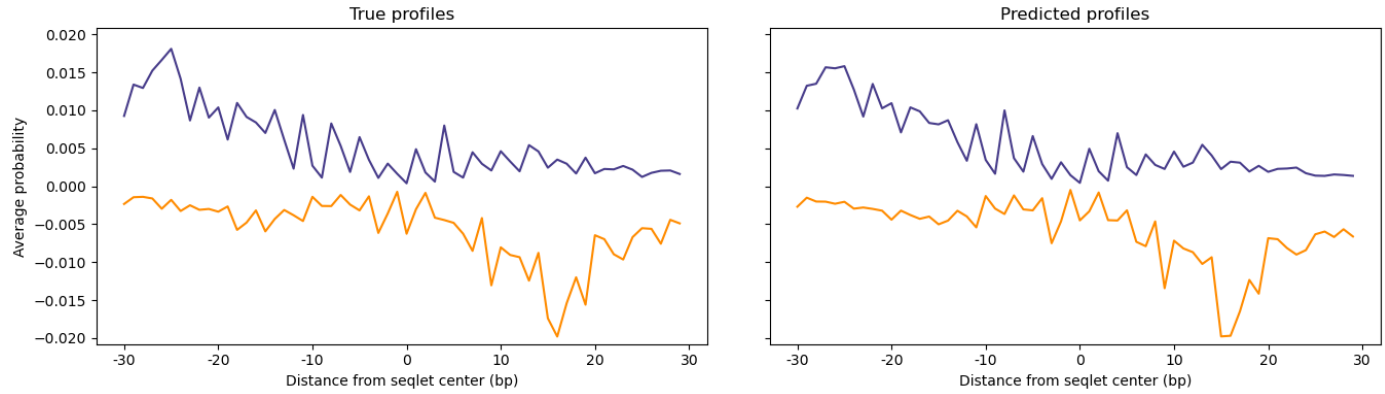
833 seqlets

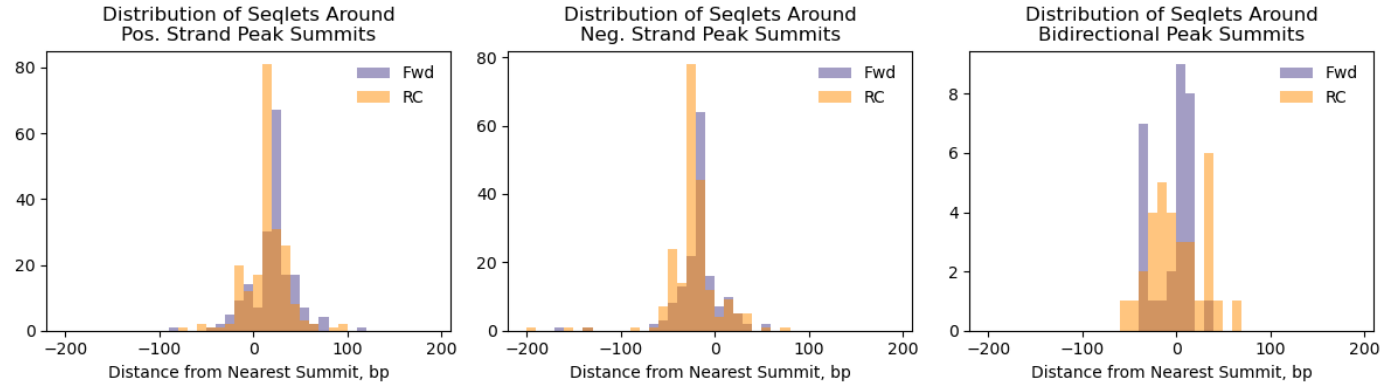
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







Pattern 12/39

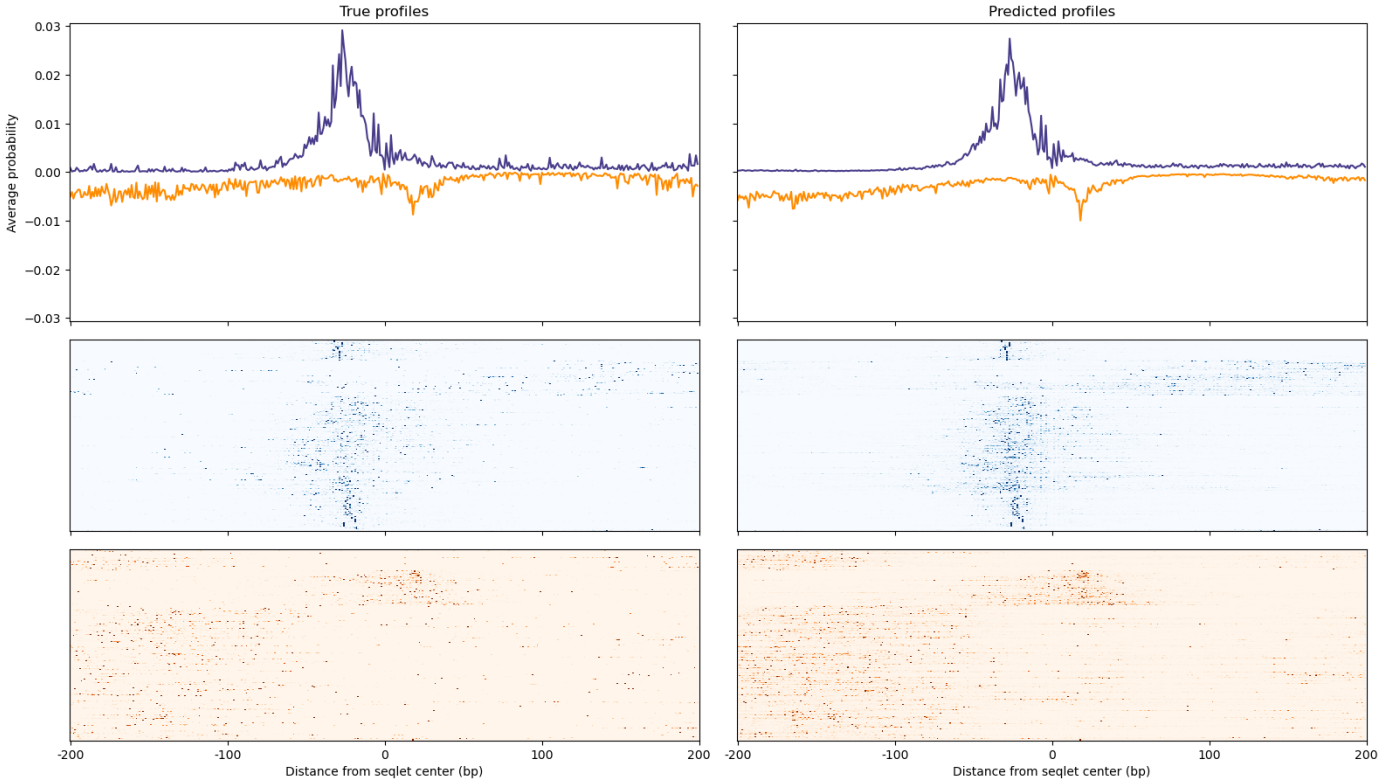
760 seqlets

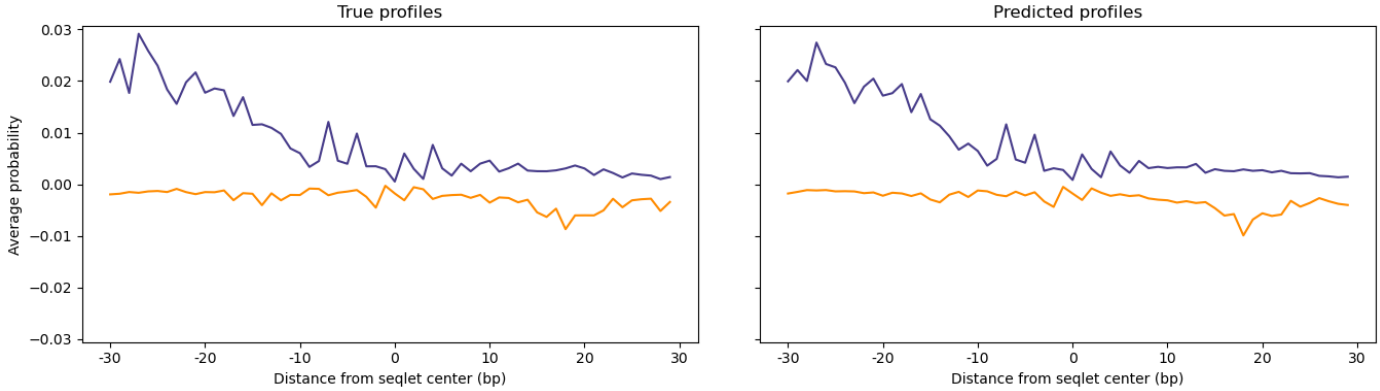
Sequence

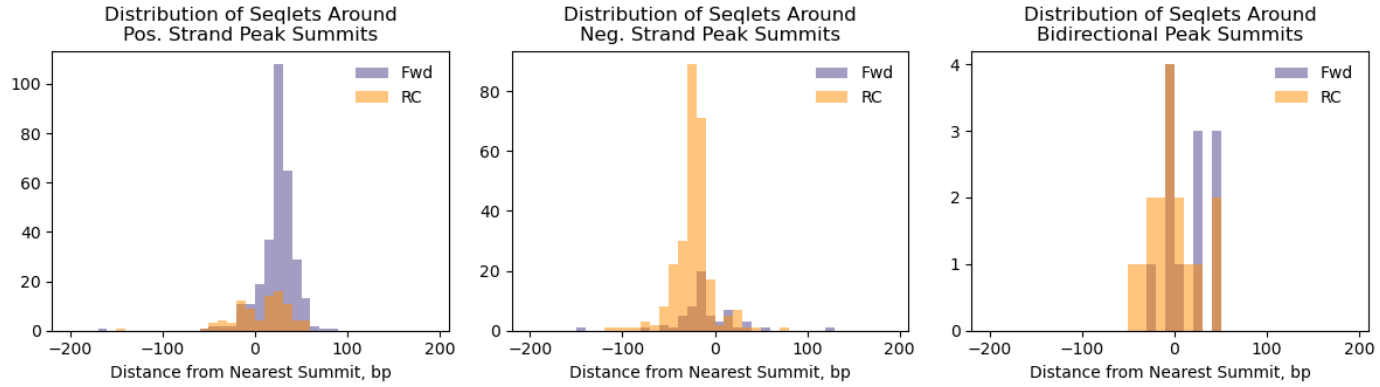
(PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)

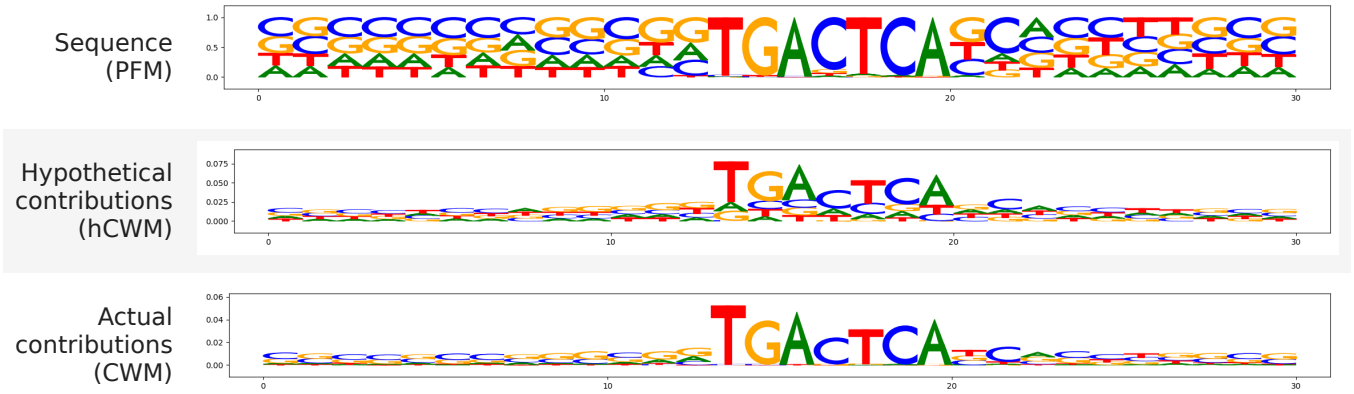


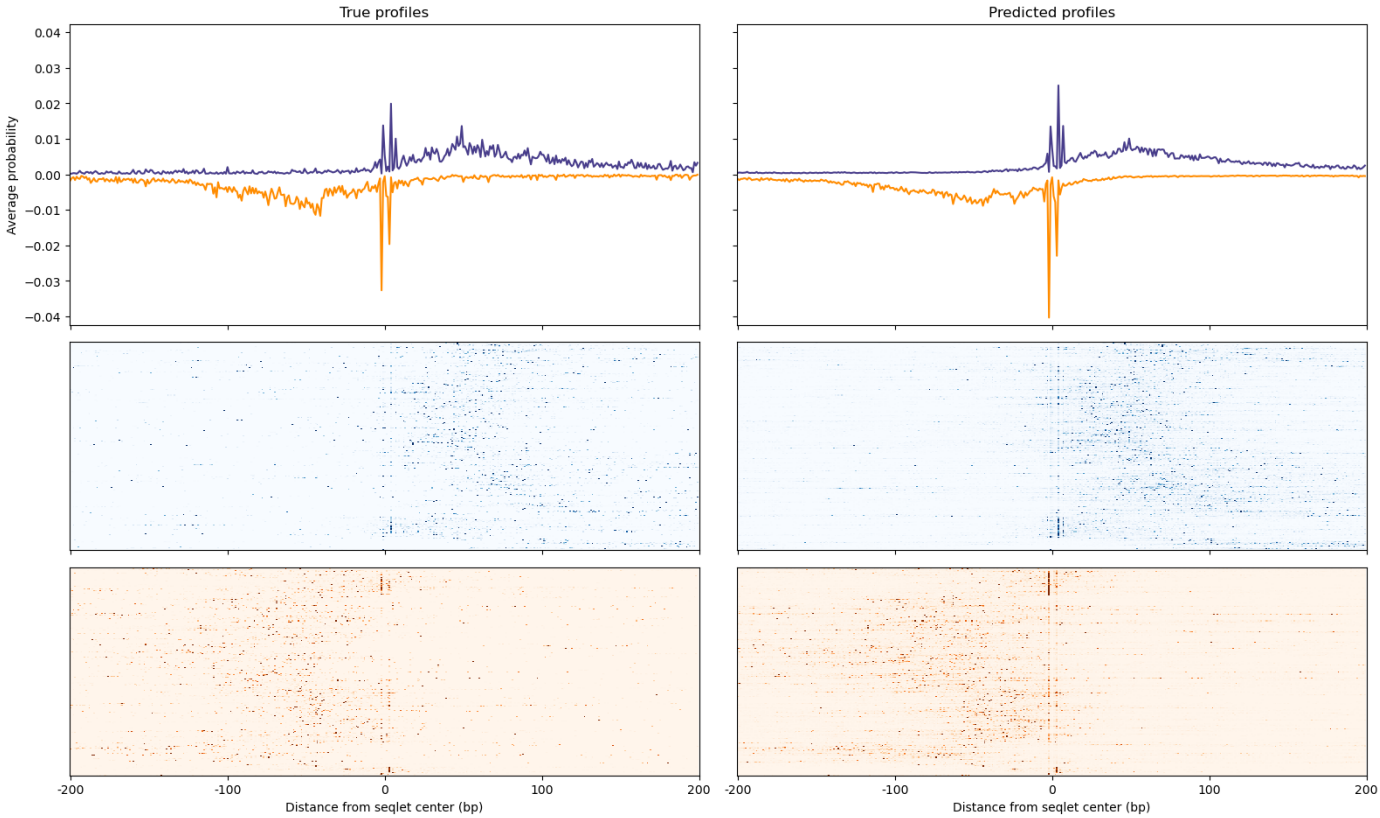


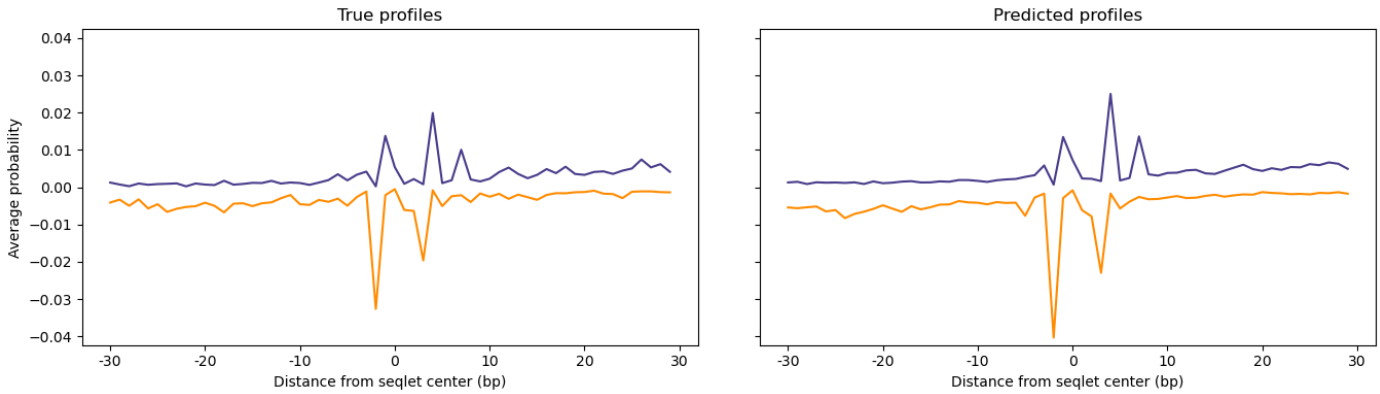


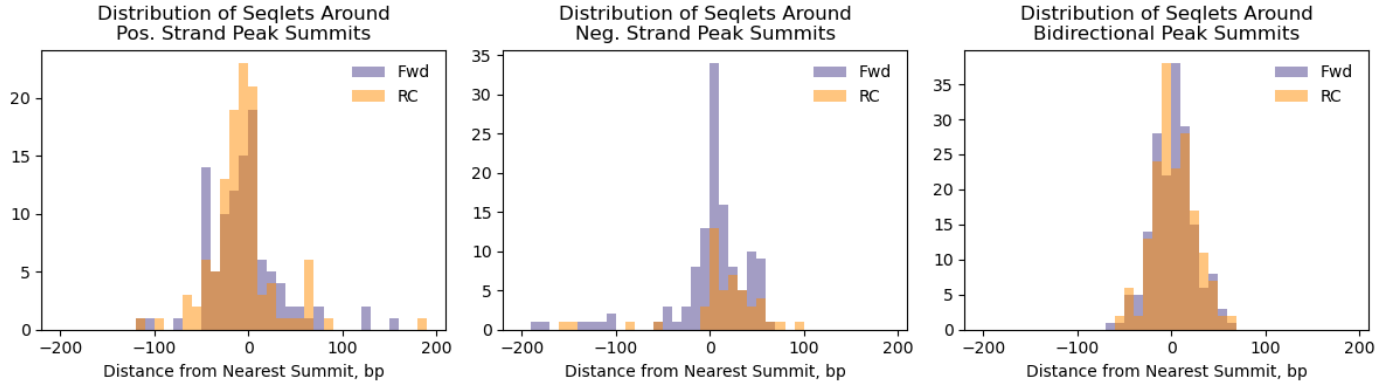
Pattern 13/39

747 seqlets


Sequence (PFM)
Hypothetical contributions (hCWM)
Actual contributions (CWM)


True profiles
Predicted profiles
Average probability
Distance from seqlet center (bp)


True profiles
Predicted profiles
Average probability
Distance from seqlet center (bp)



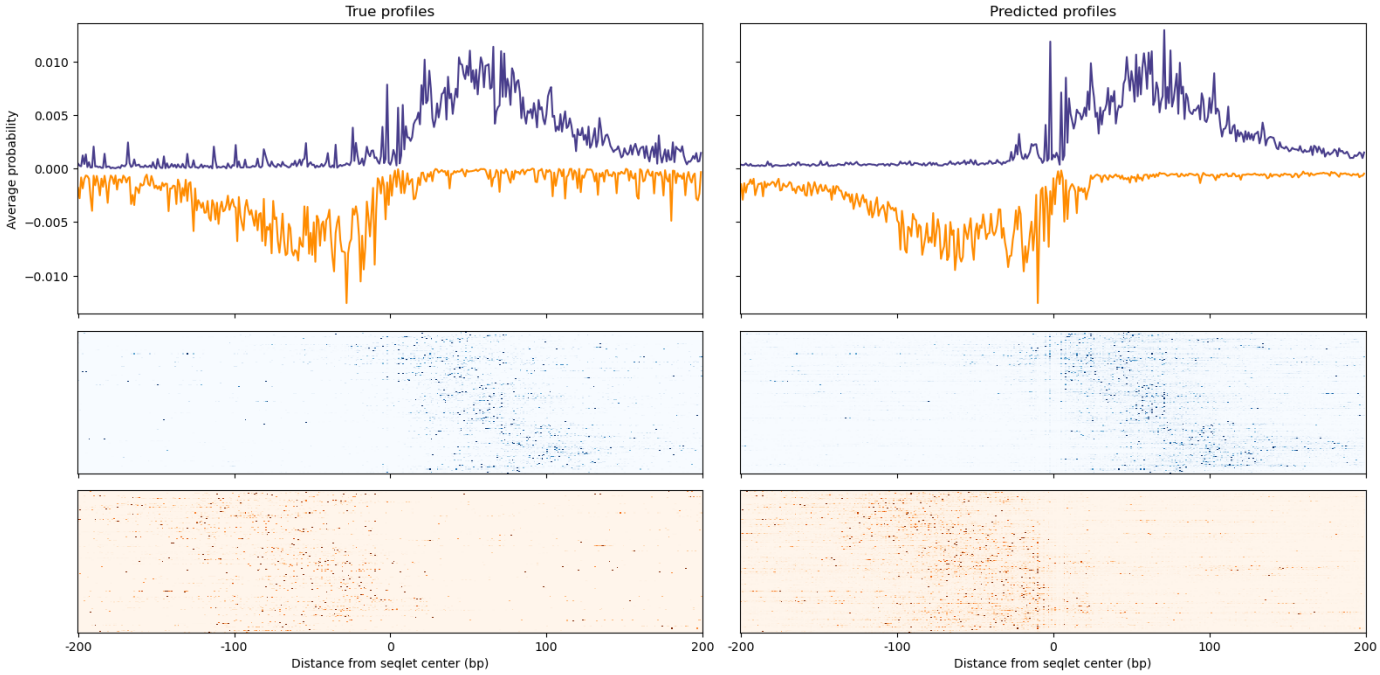
Pattern 14/39

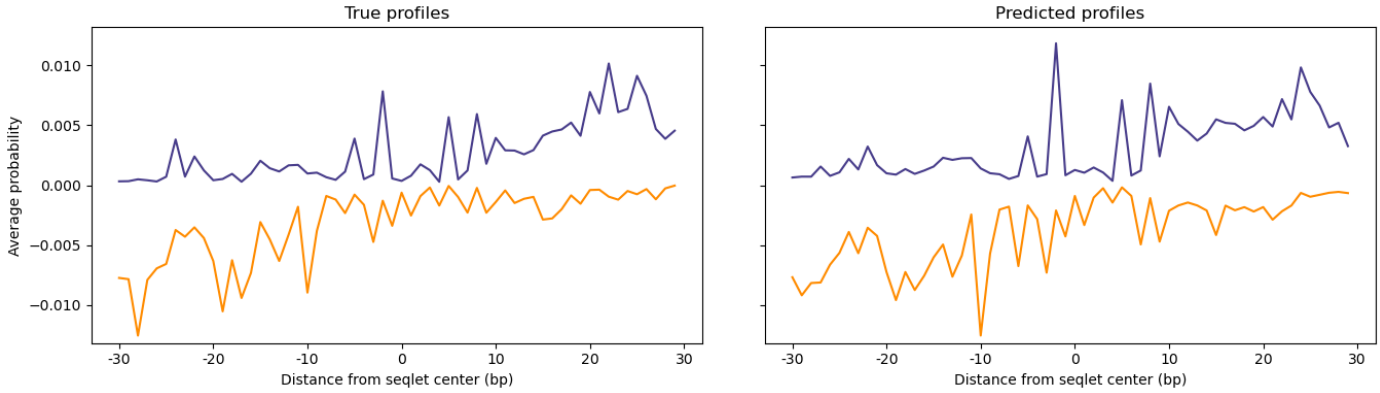
535 seqlets

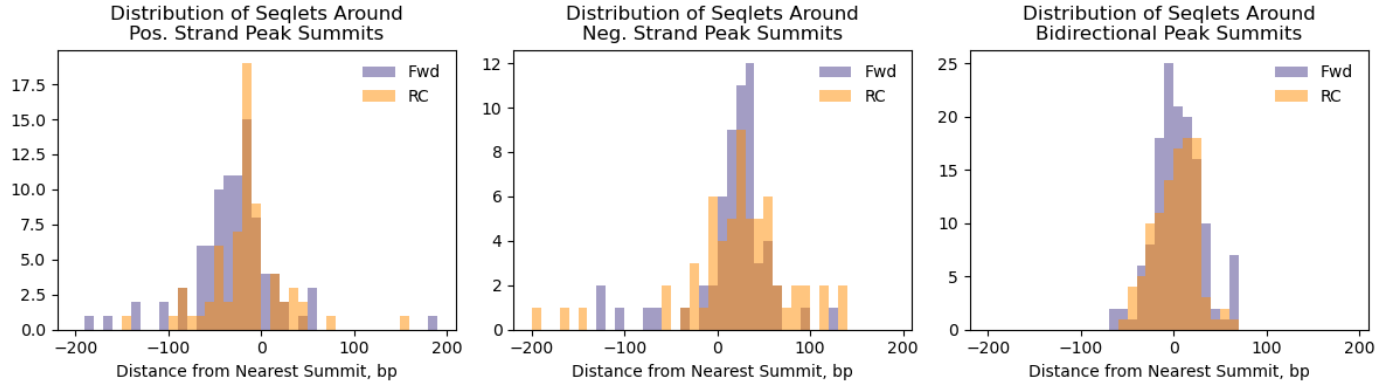
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







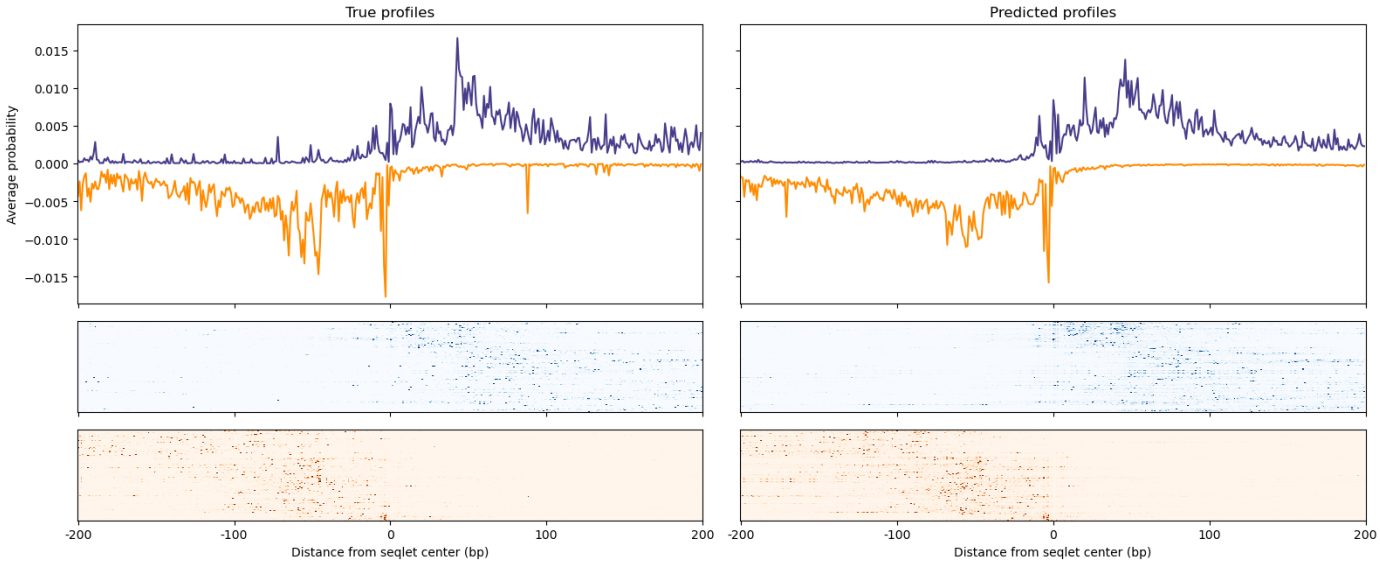
Pattern 15/39

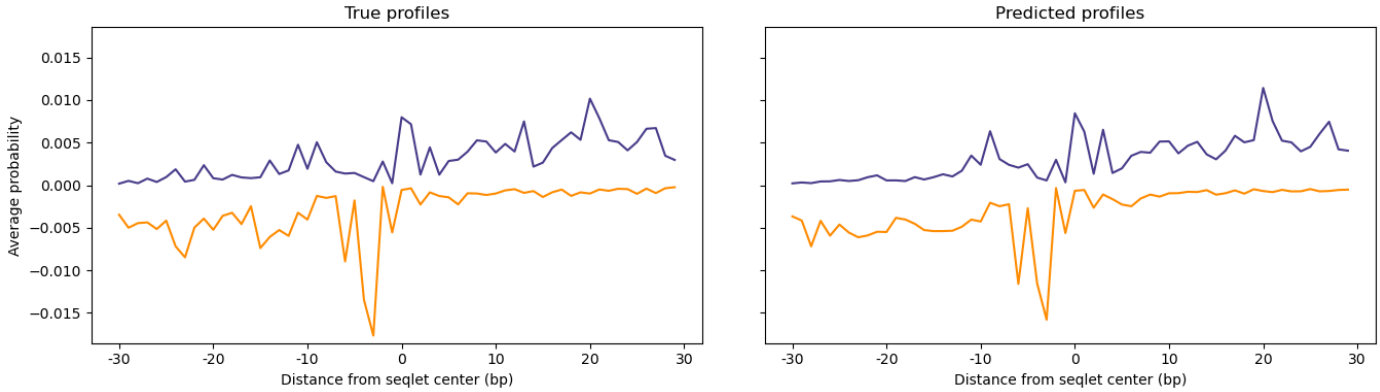
410 seqlets

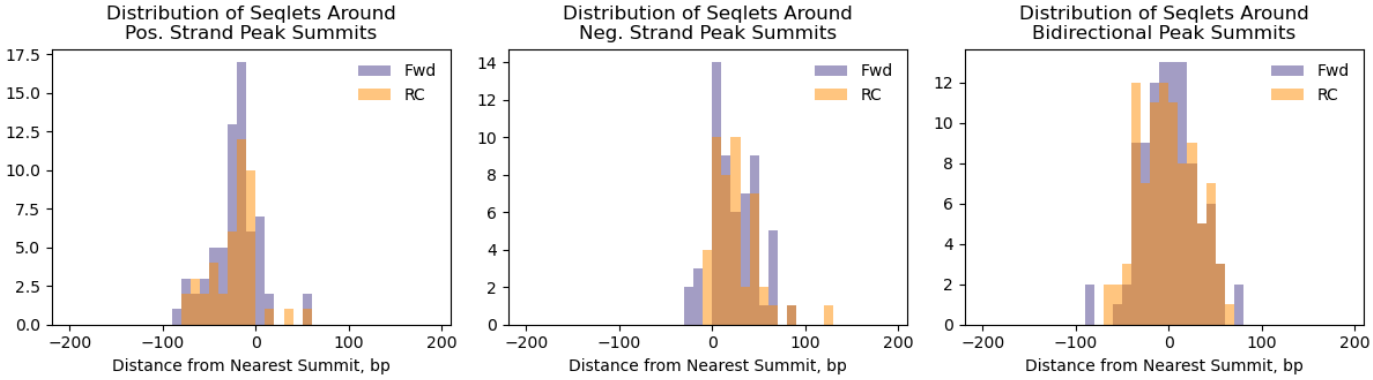
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







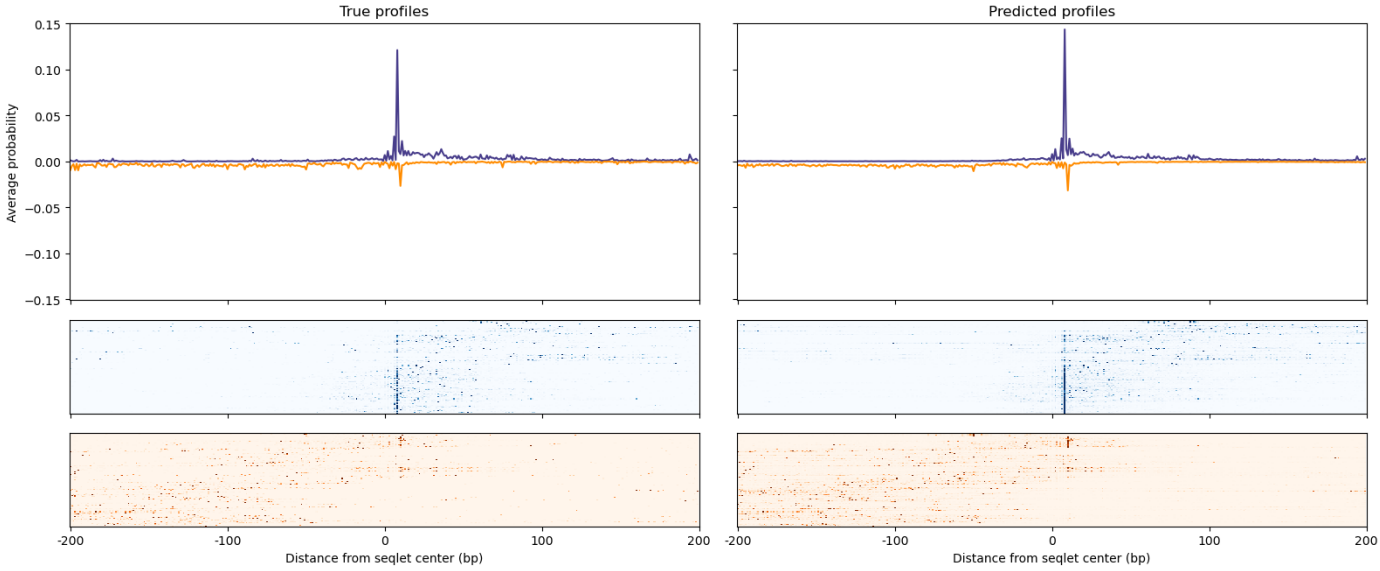
Pattern 16/39

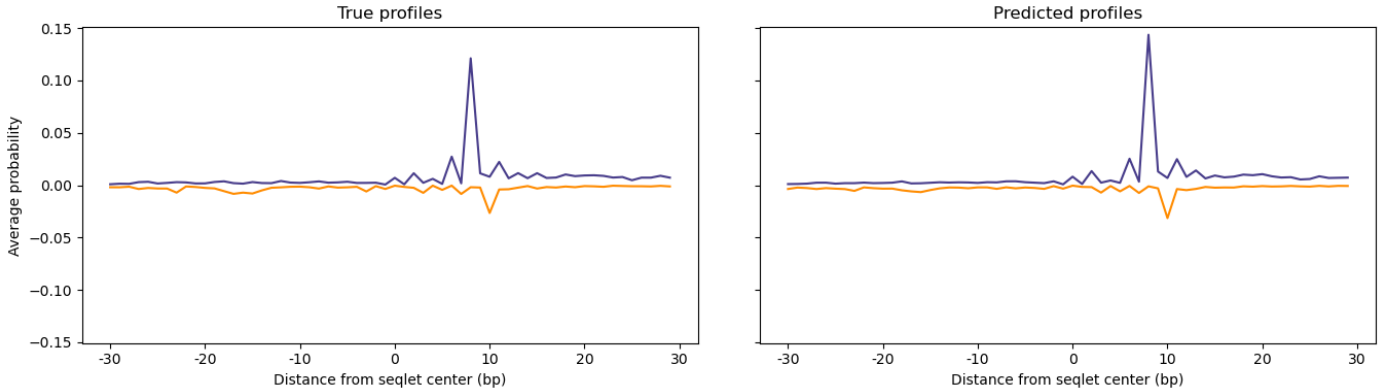
401 seqlets

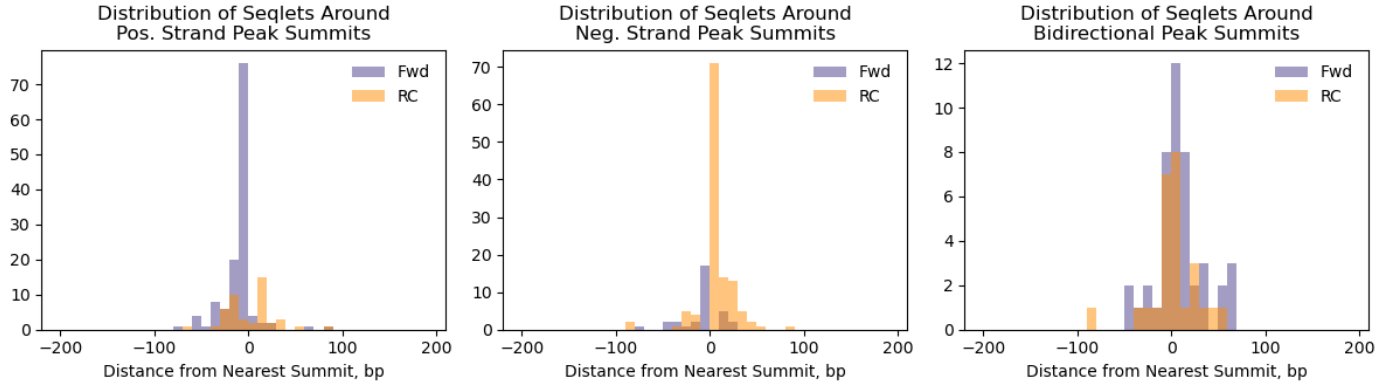
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





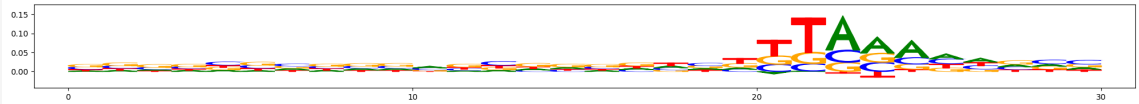


Pattern 17/39

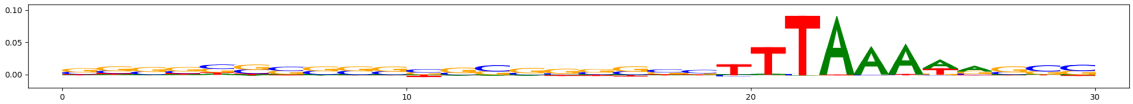
219 seqlets

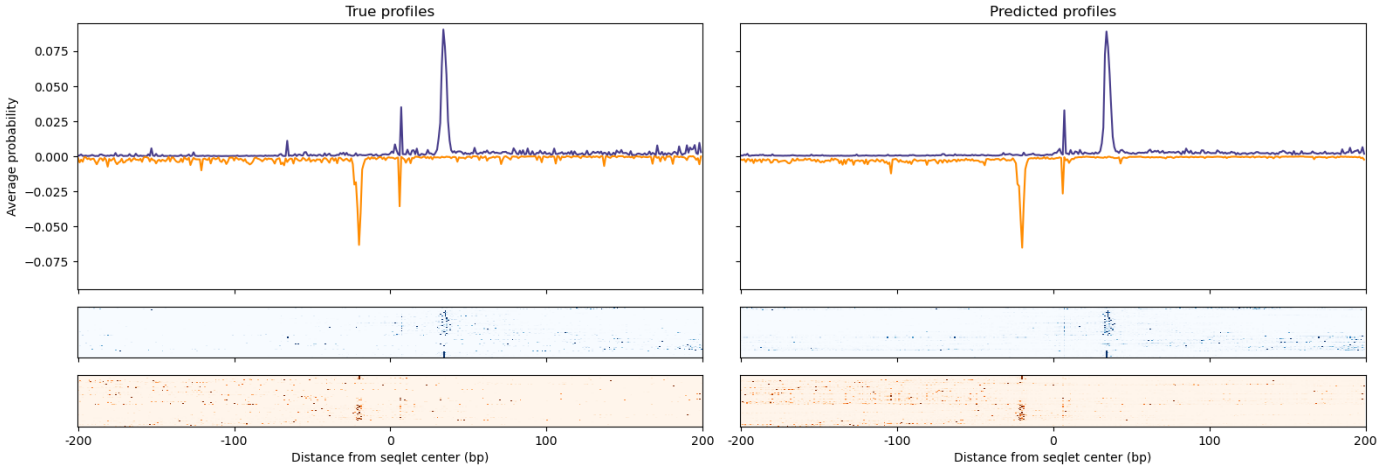
Sequence (PFM)

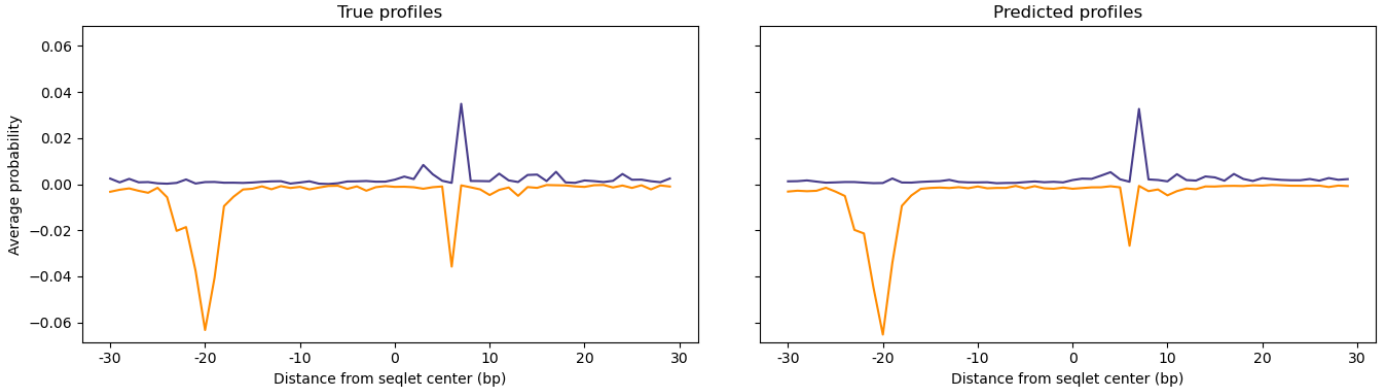
Hypothetical contributions (hCWM)

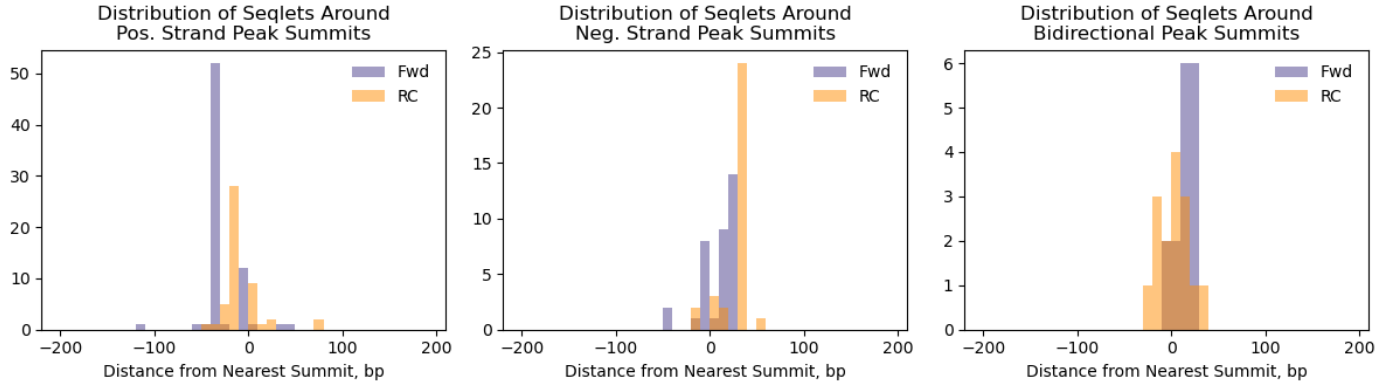


Actual contributions (CWM)







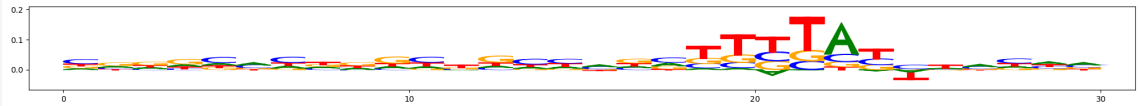


Pattern 18/39

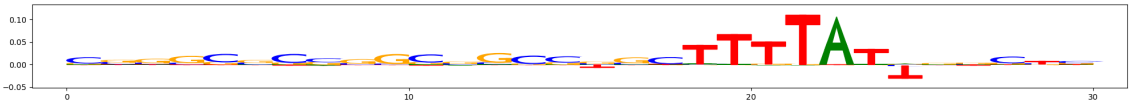
198 seqlets

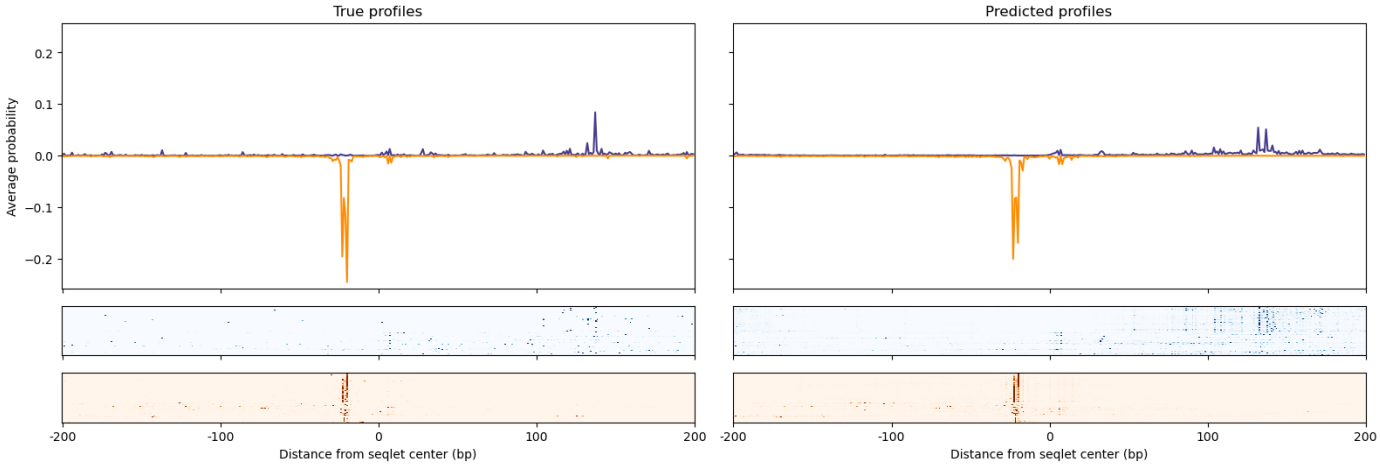
Sequence (PFM)

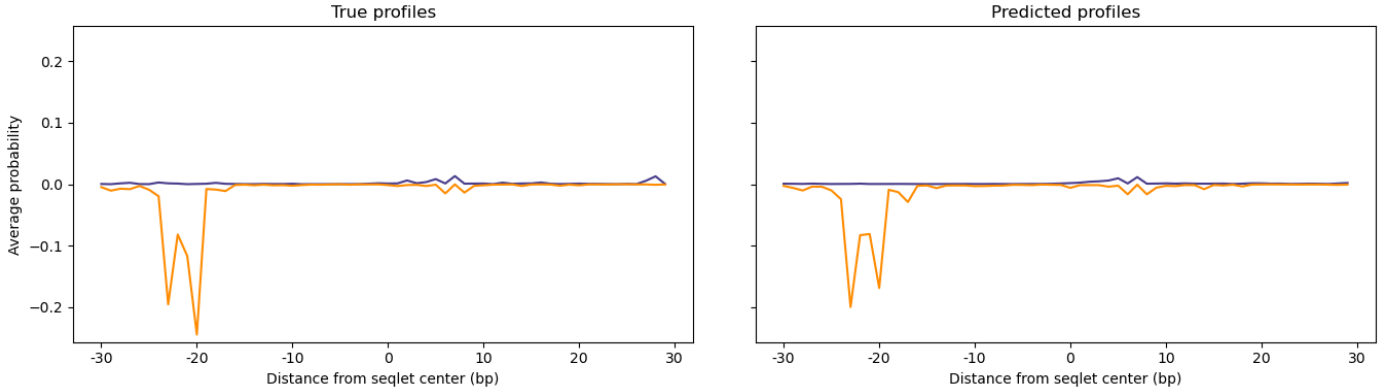
Hypothetical contributions (hCWM)

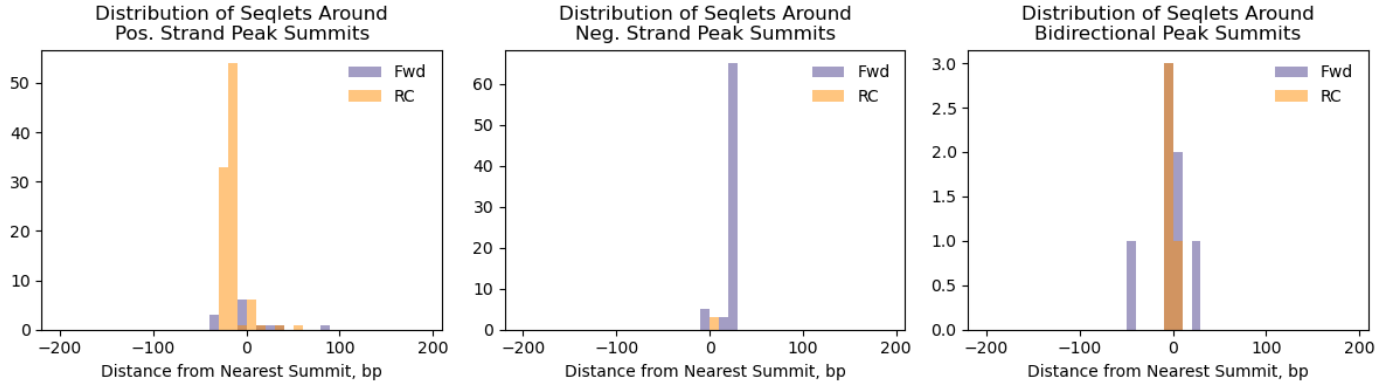


Actual contributions (CWM)









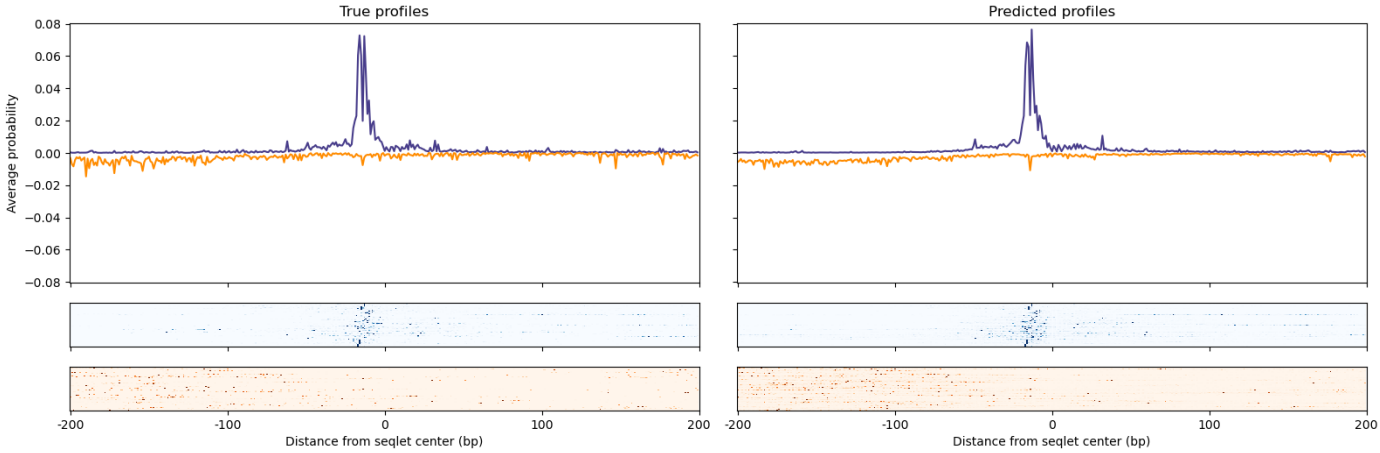
Pattern 19/39

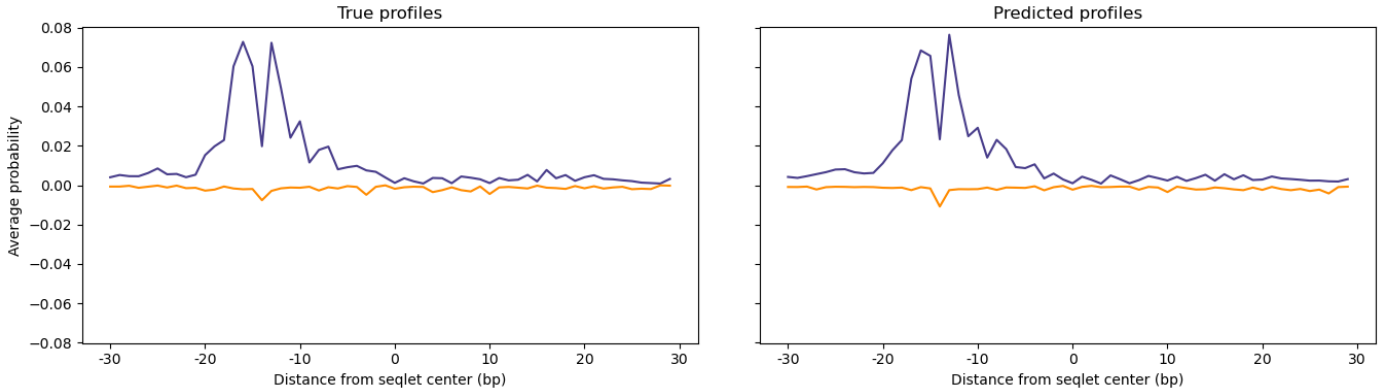
190 seqlets

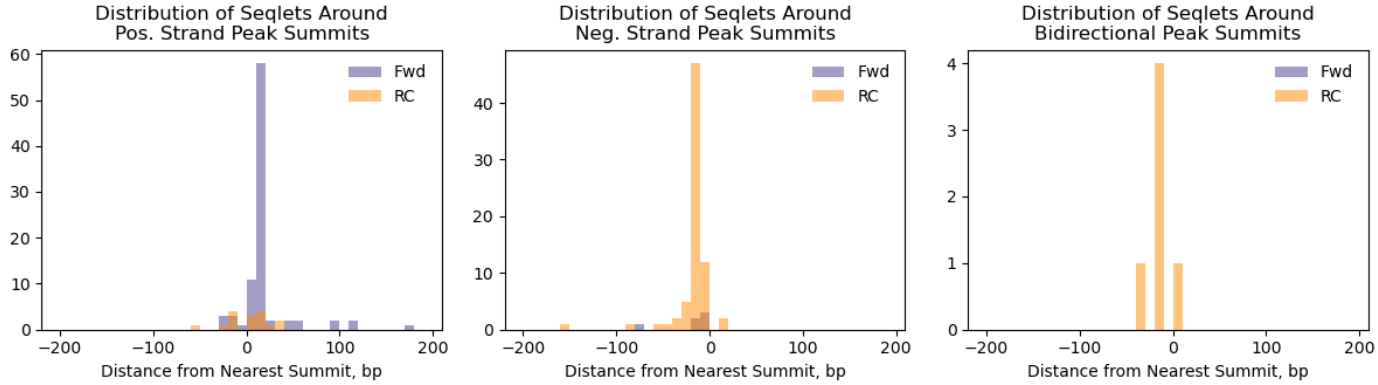
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





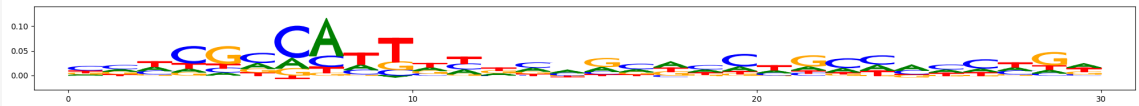


Pattern 20/39

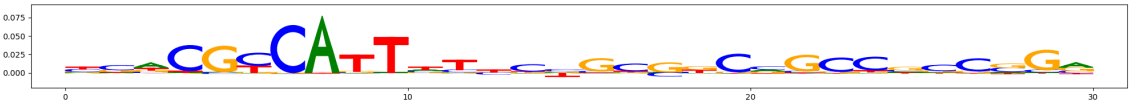
163 seqlets

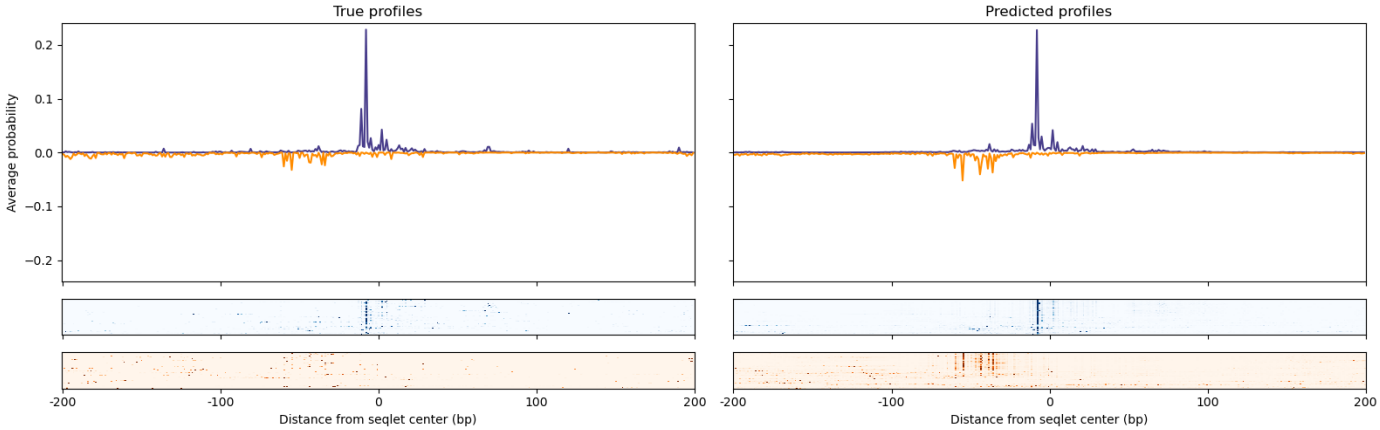
Sequence (PFM)

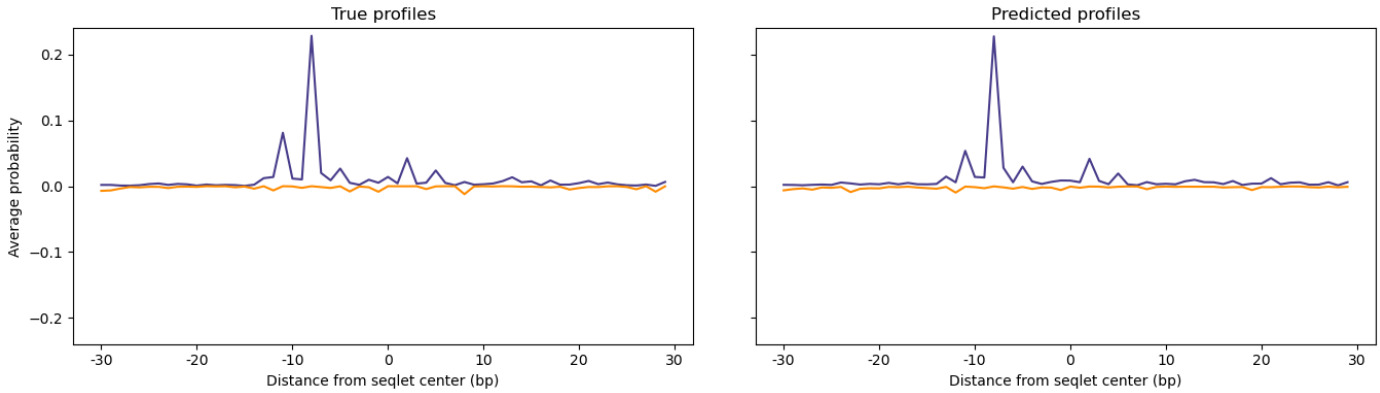
Hypothetical contributions (hCWM)

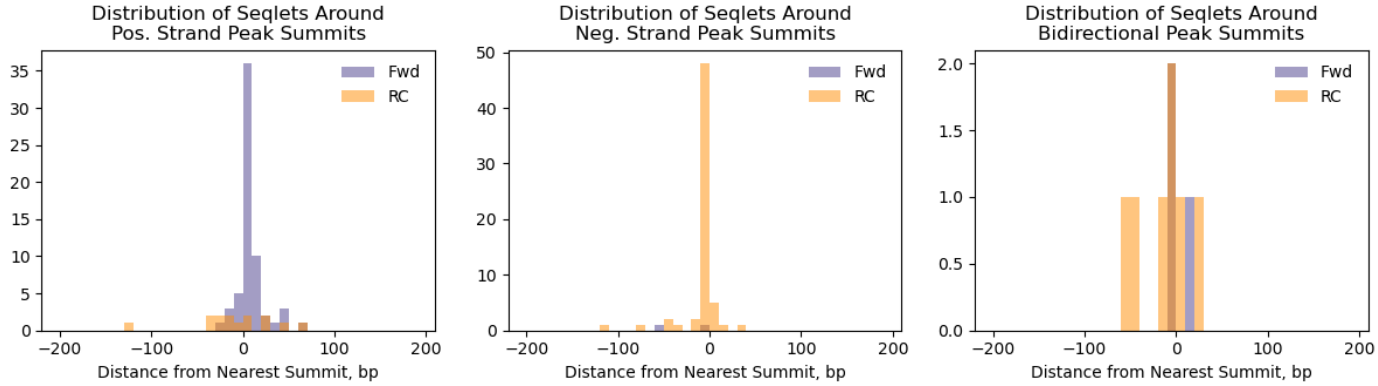


Actual contributions (CWM)







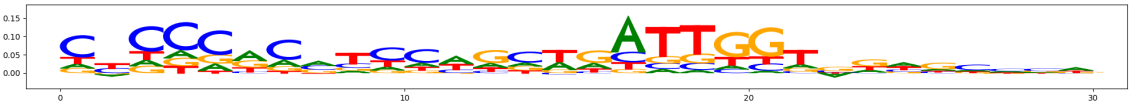


Pattern 21/39

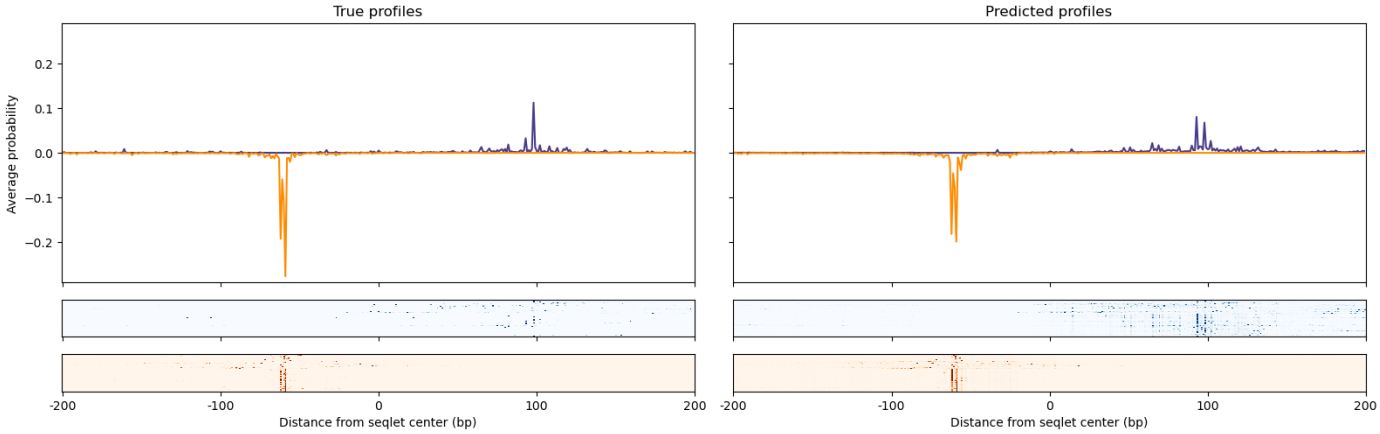
153 seqlets

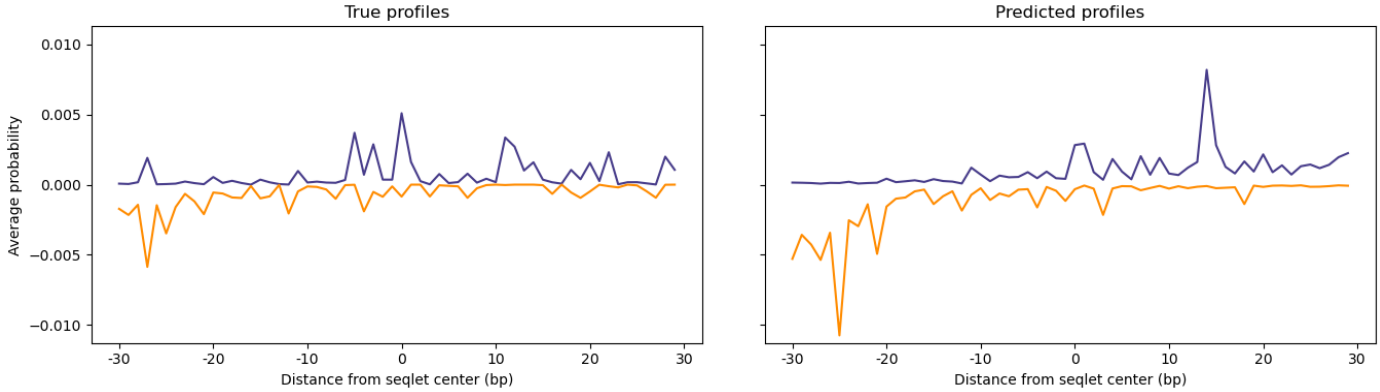
Sequence (PFM)

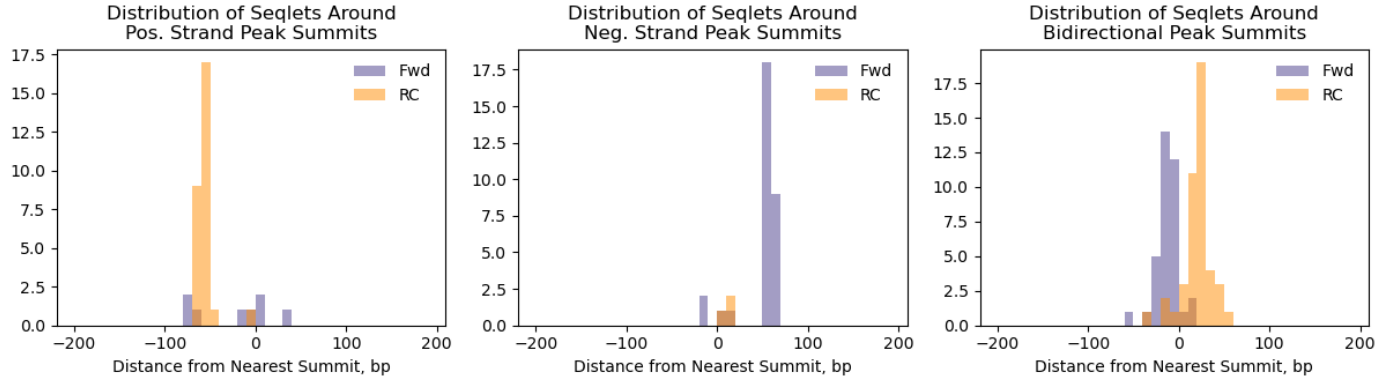
Hypothetical contributions (hCWM)



Actual contributions (CWM)





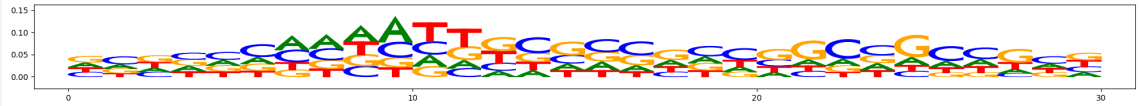


Pattern 22/39

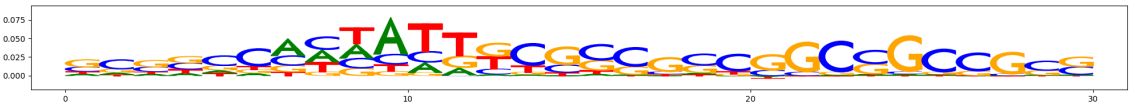
148 seqlets

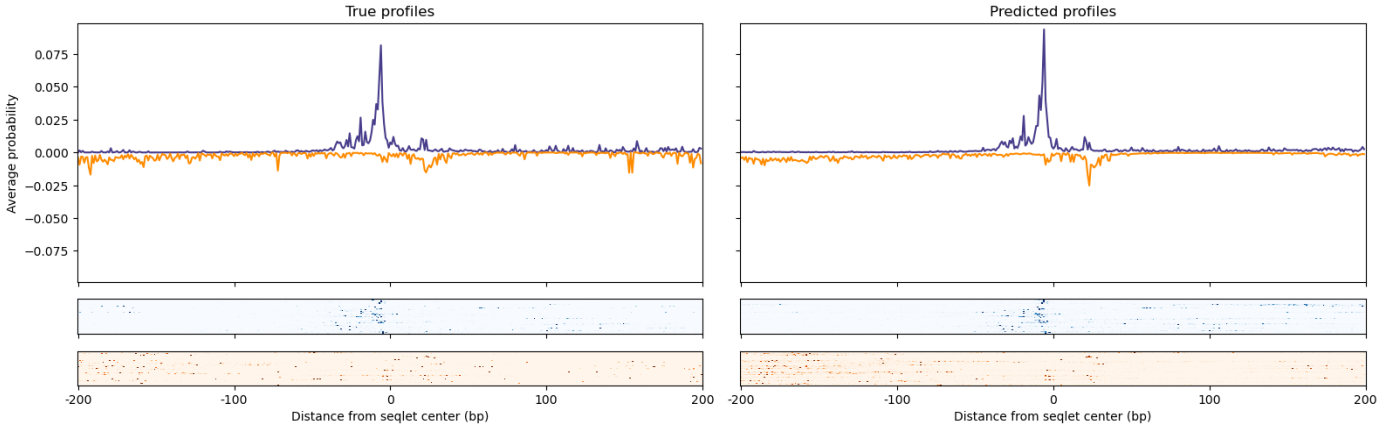
Sequence (PFM)

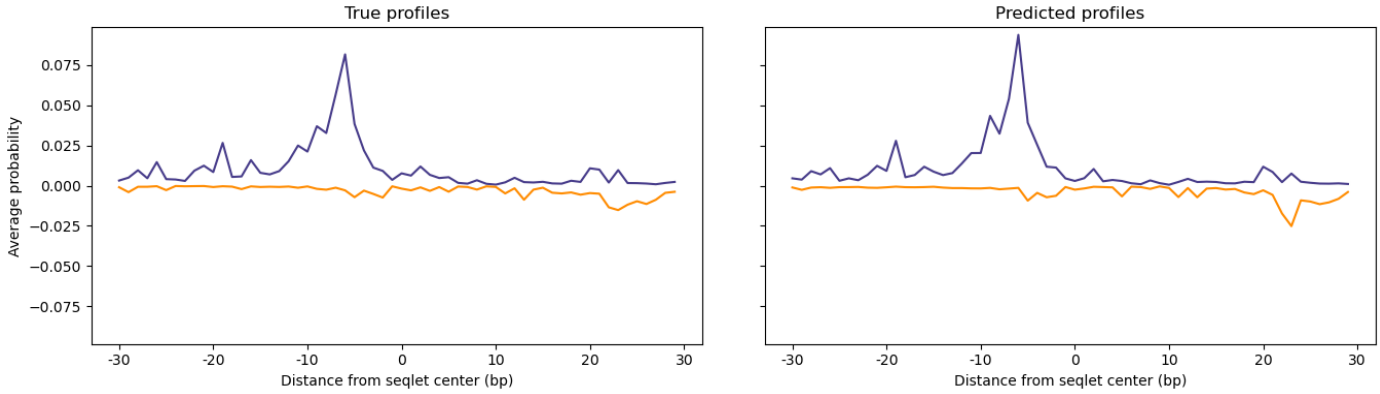
Hypothetical contributions (hCWM)

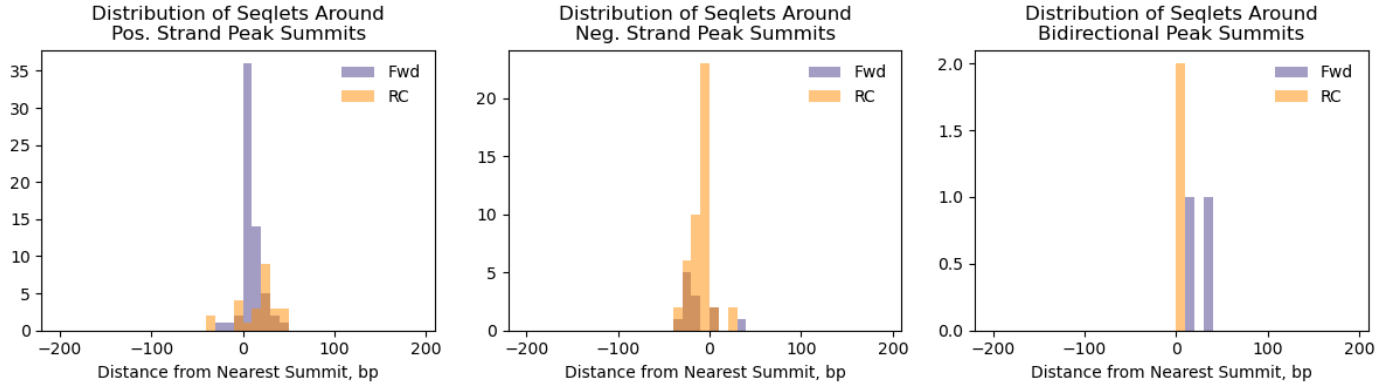


Actual contributions (CWM)







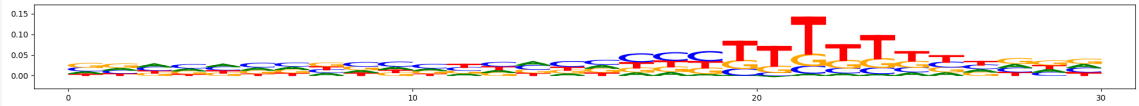


Pattern 23/39

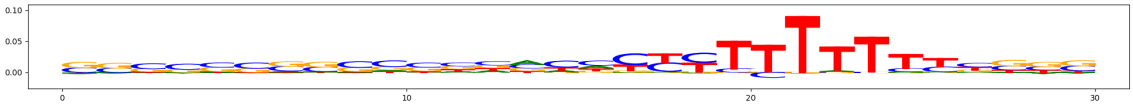
135 seqlets

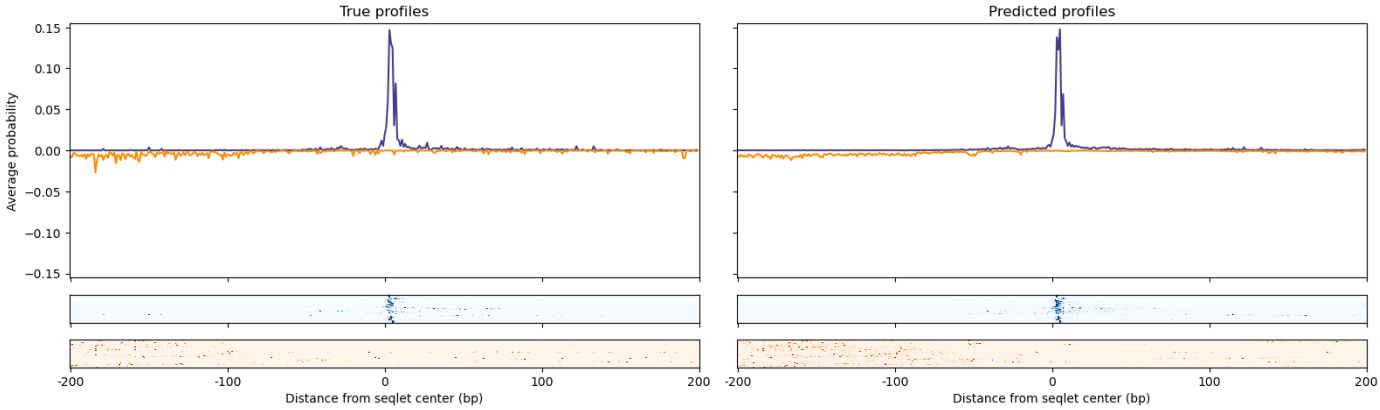
Sequence (PFM)

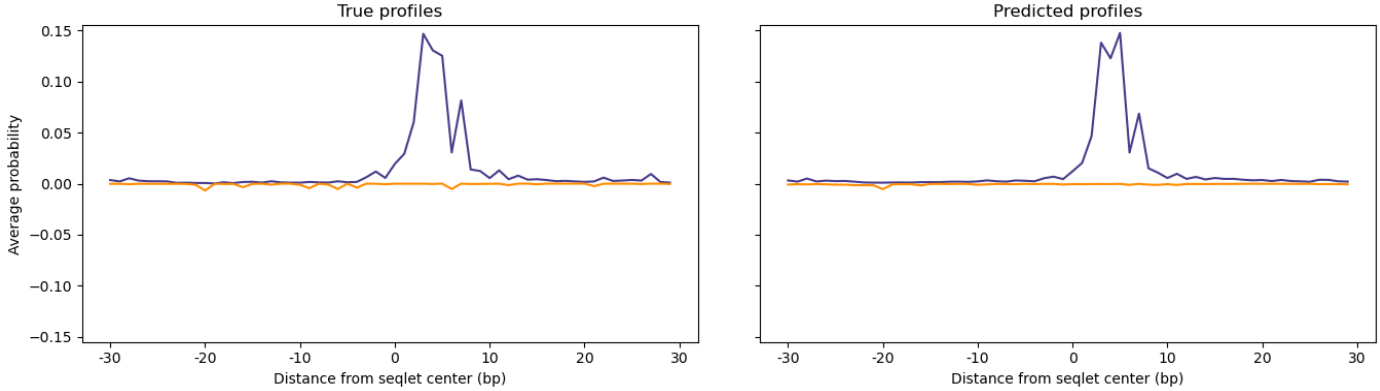
Hypothetical contributions (hCWM)

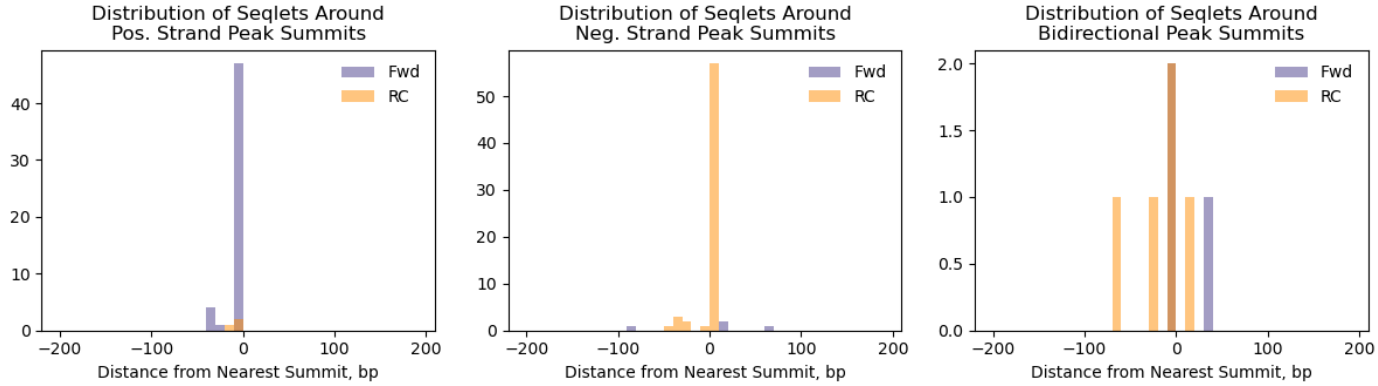


Actual contributions (CWM)



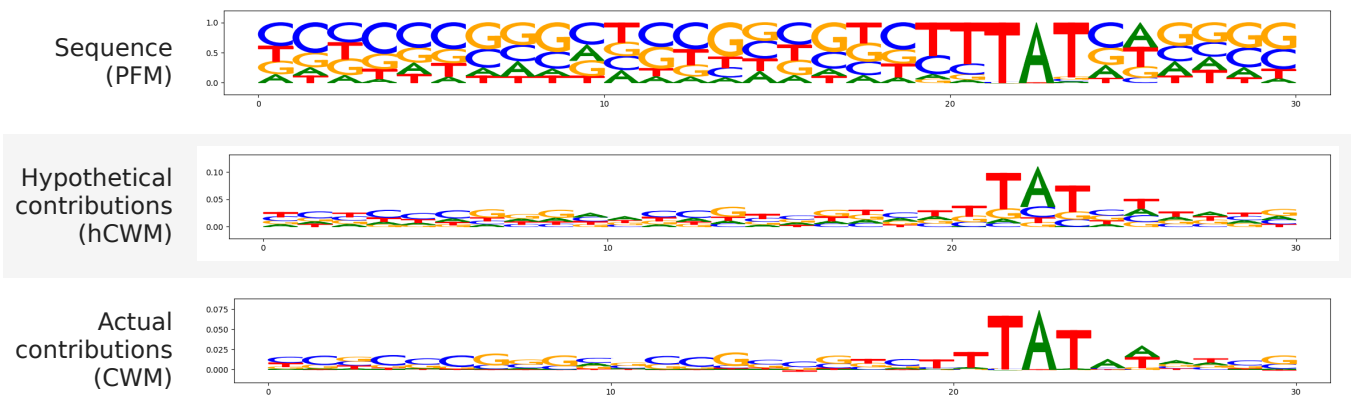


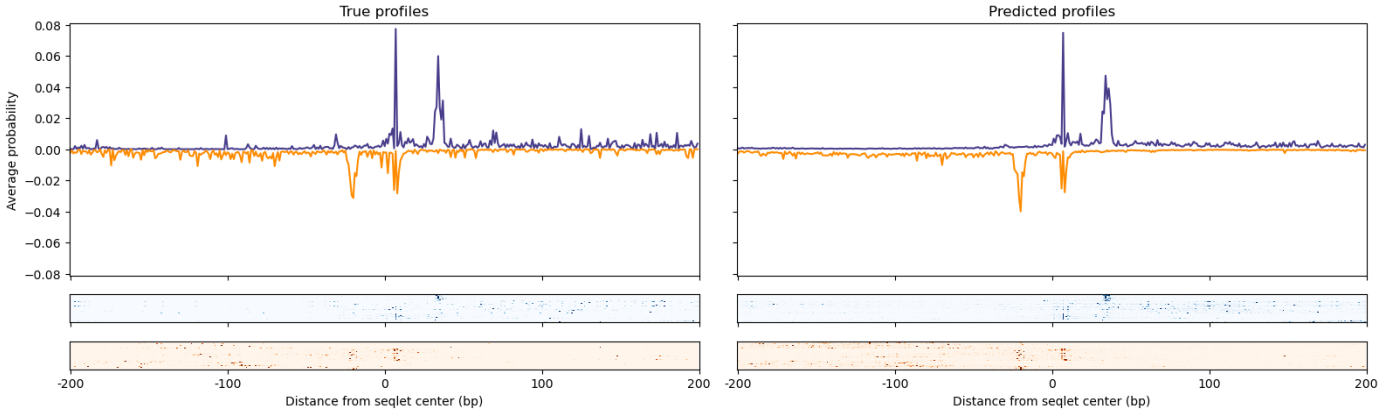


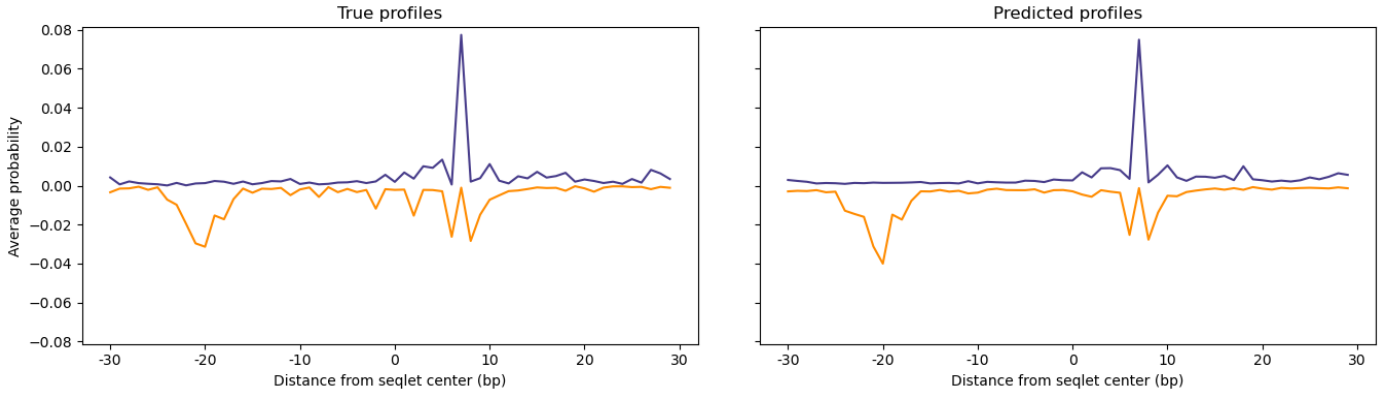


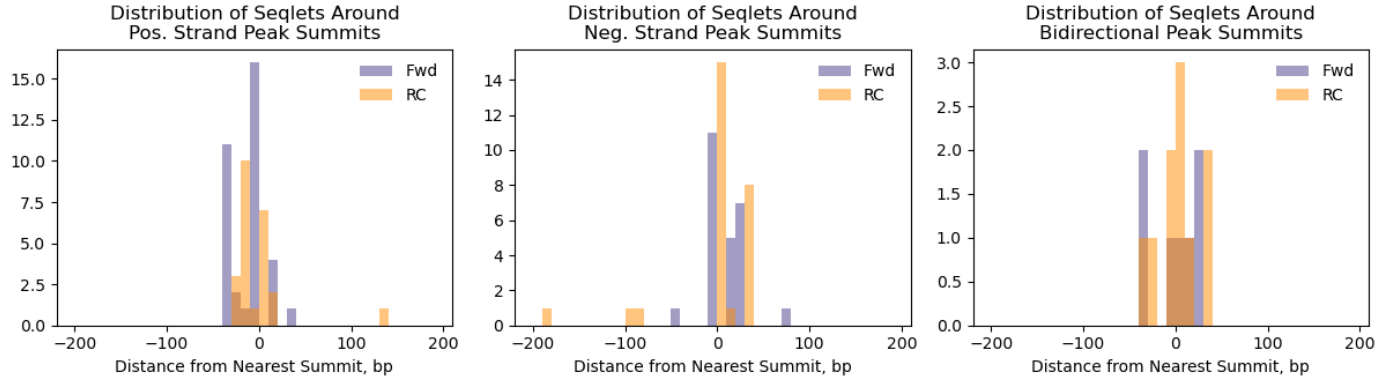
Pattern 24/39

130 seqlets









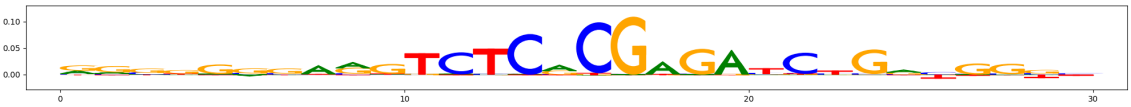
Pattern 25/39

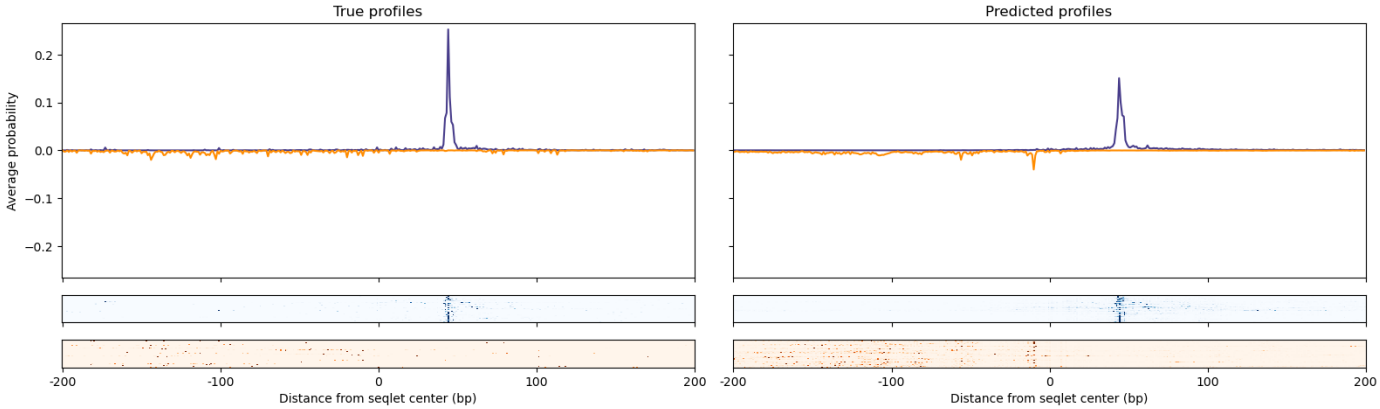
113 seqlets

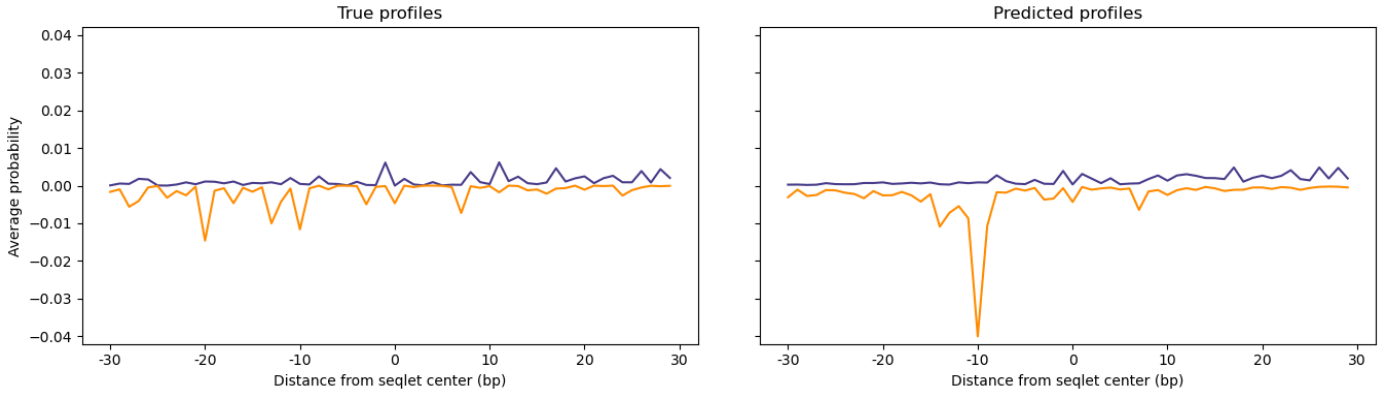
Sequence (PFM)

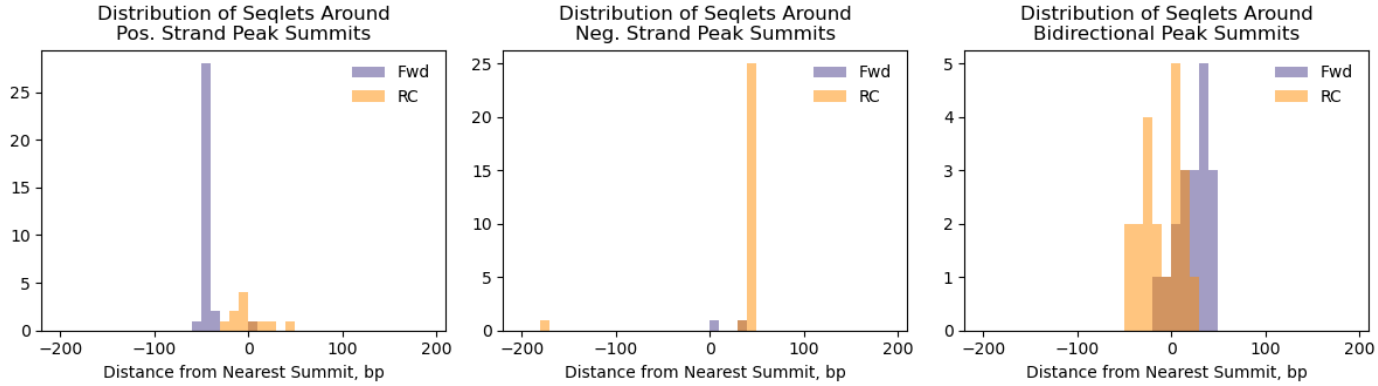
Hypothetical contributions (hCWM)

Actual contributions (CWM)









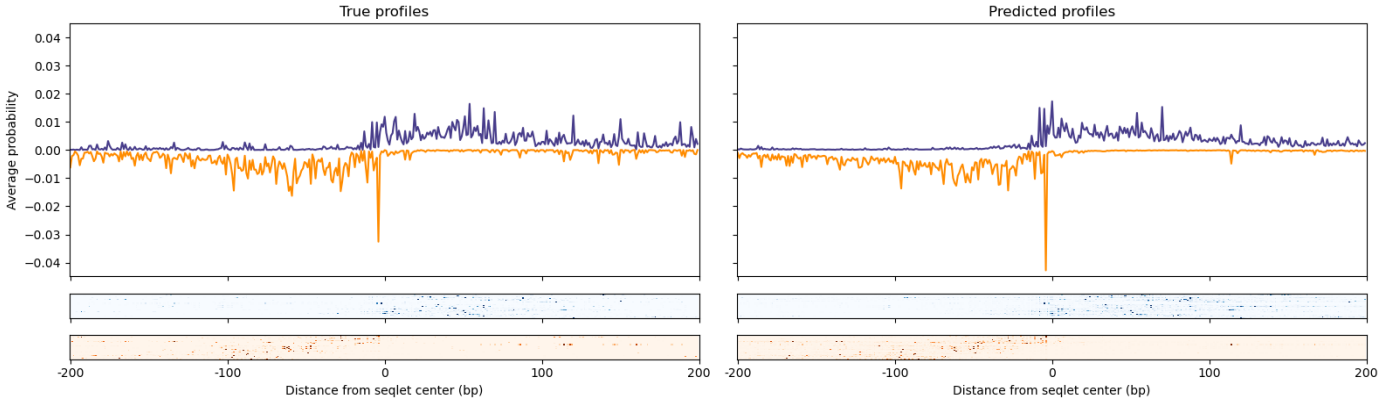
Pattern 26/39

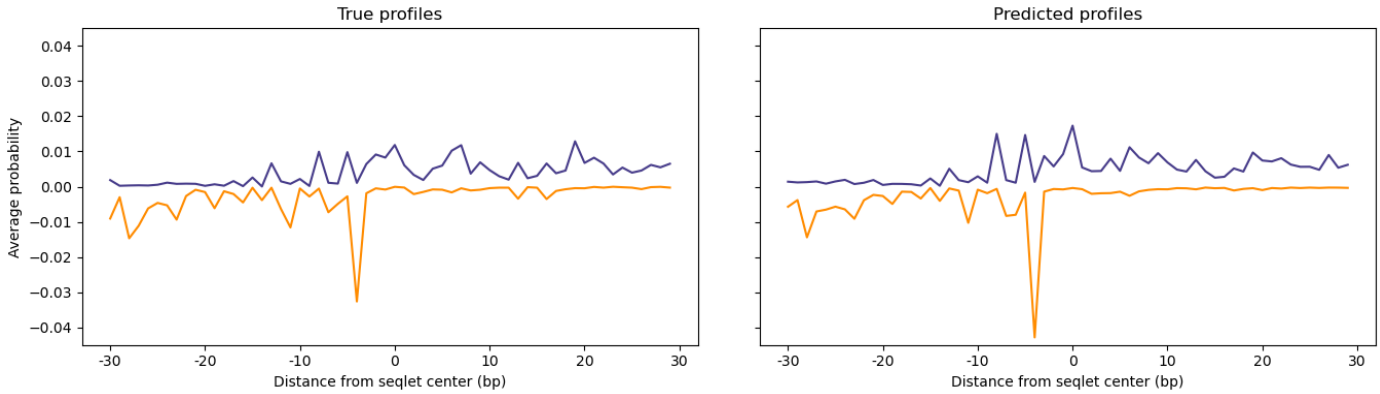
107 seqlets

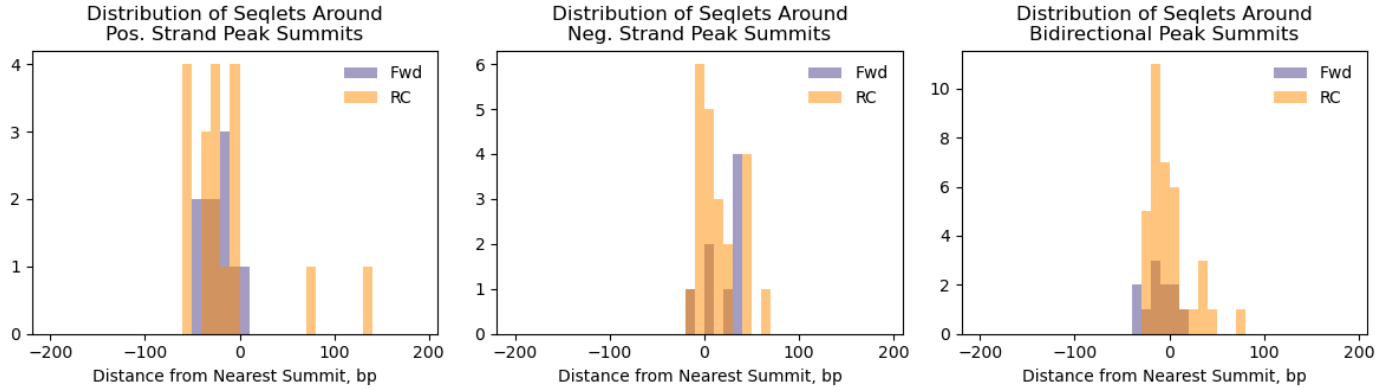
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







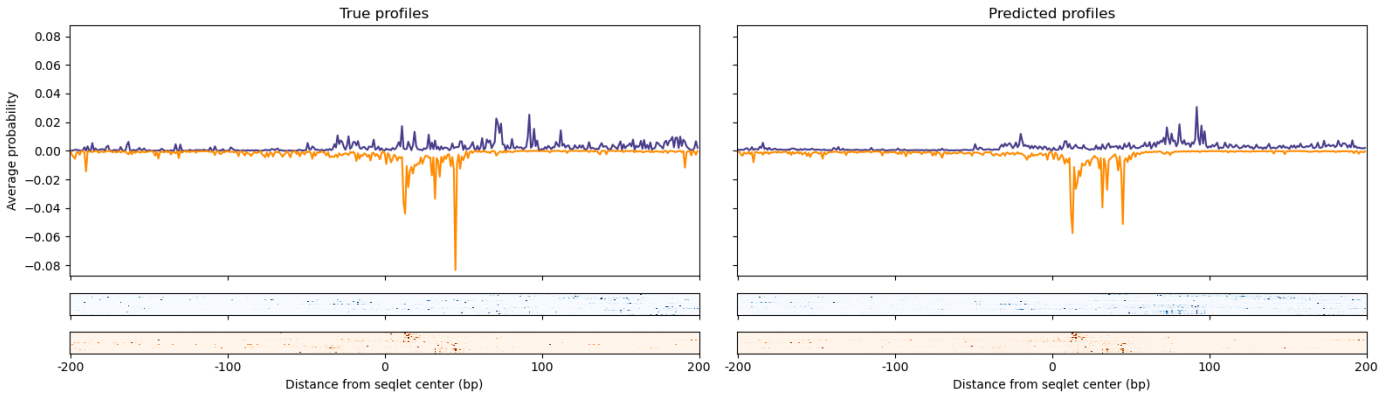
Pattern 27/39

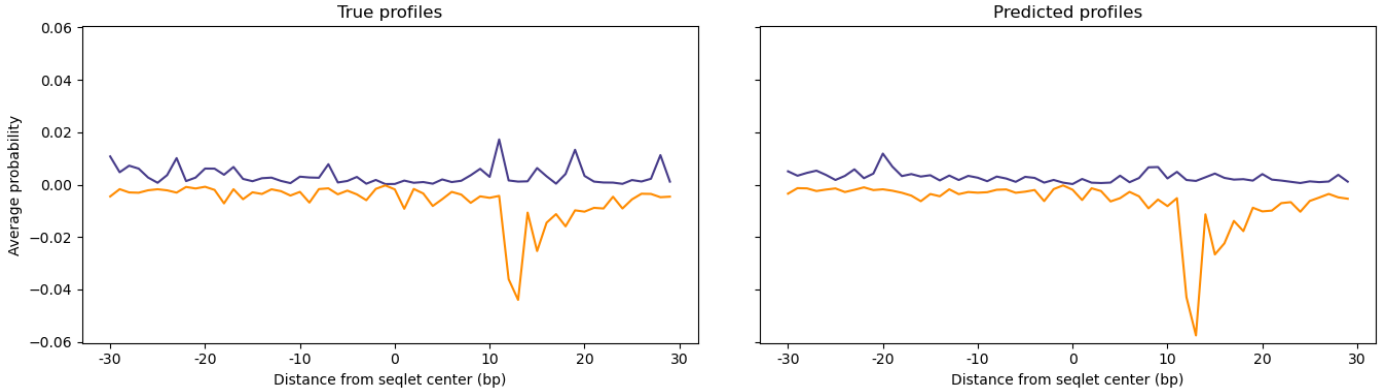
105 seqlets

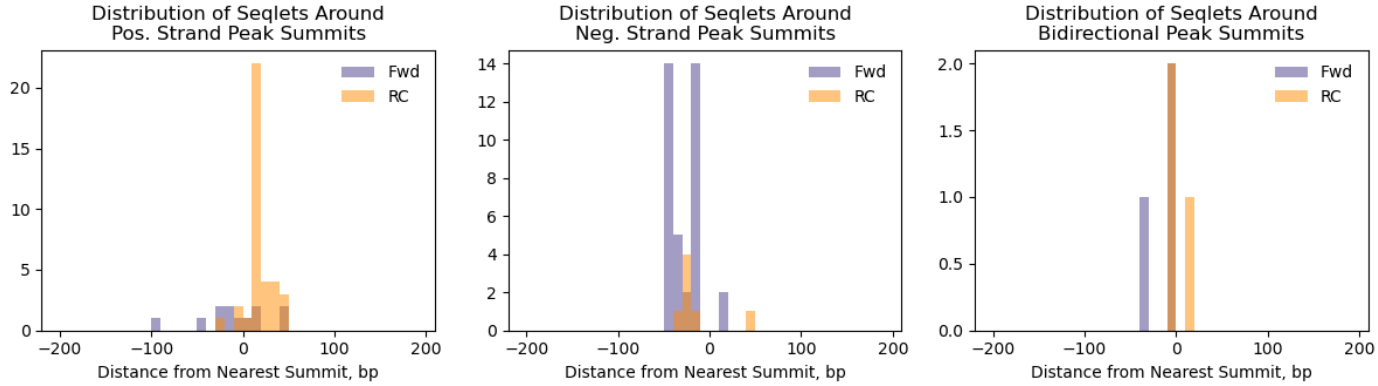
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







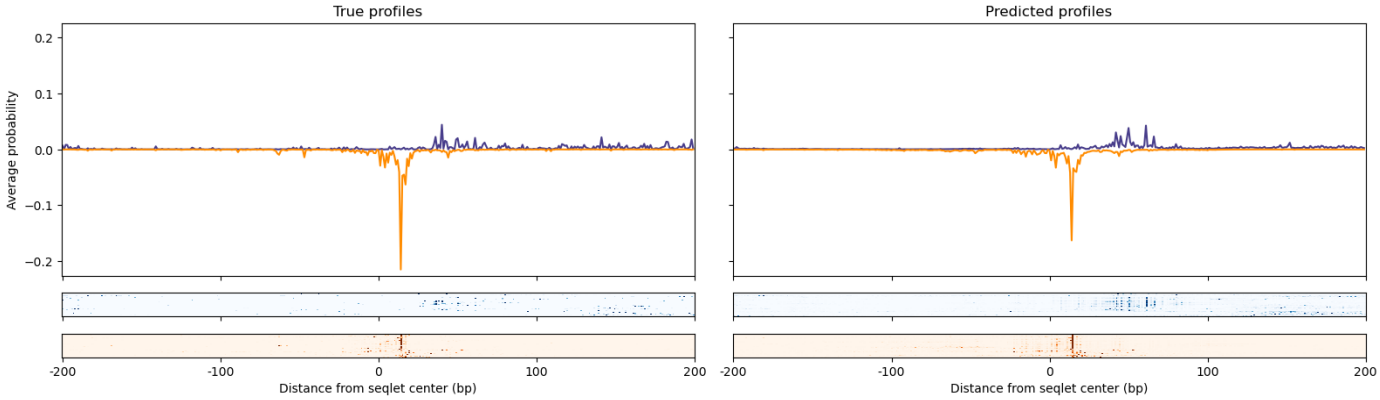
Pattern 28/39

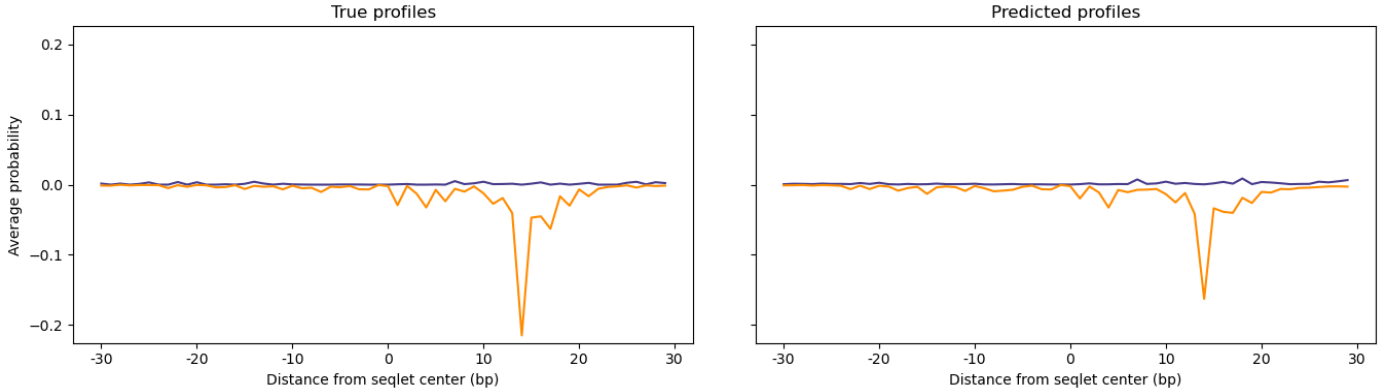
105 seqlets

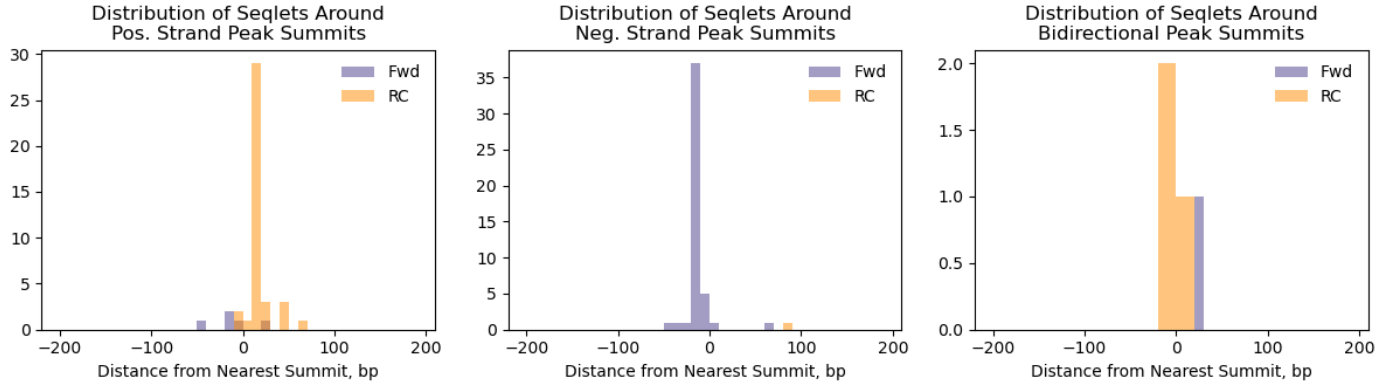
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







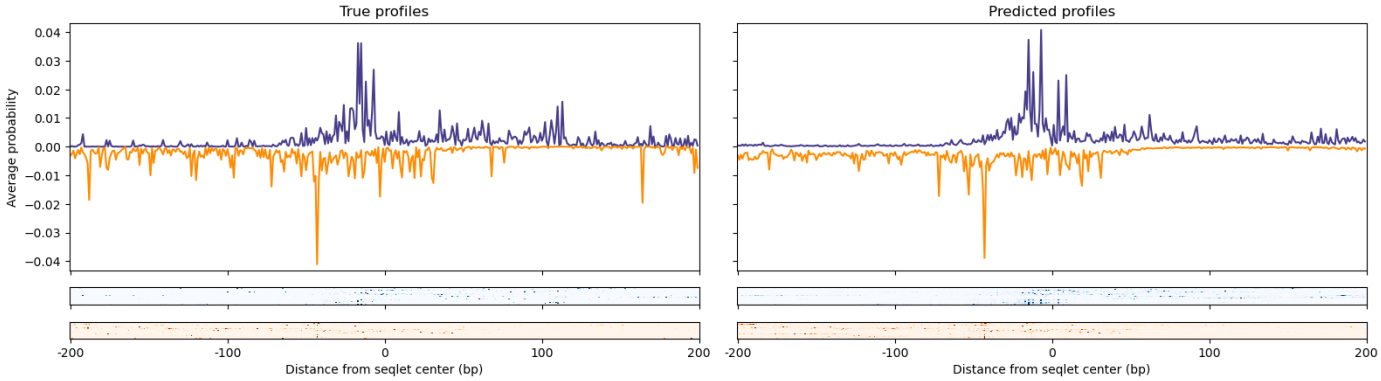
Pattern 29/39

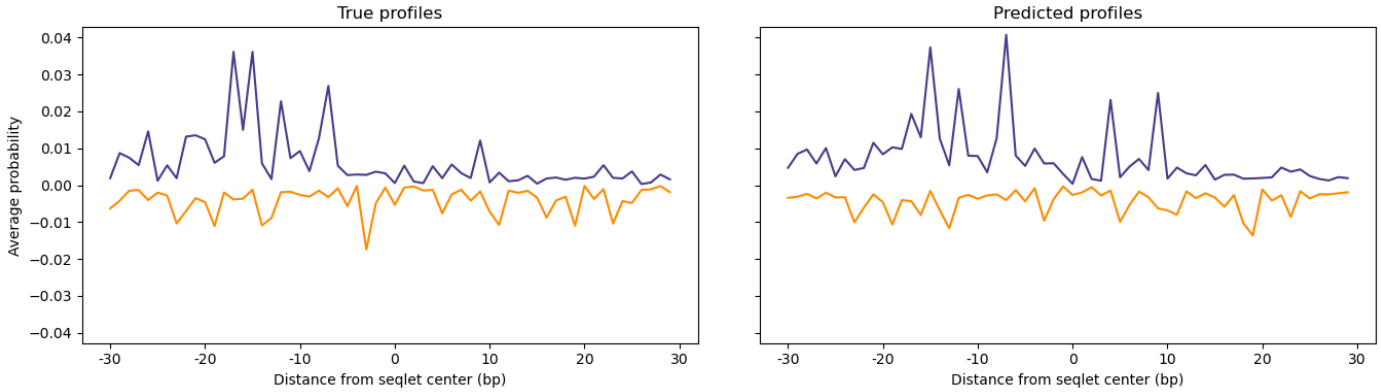
82 seqlets

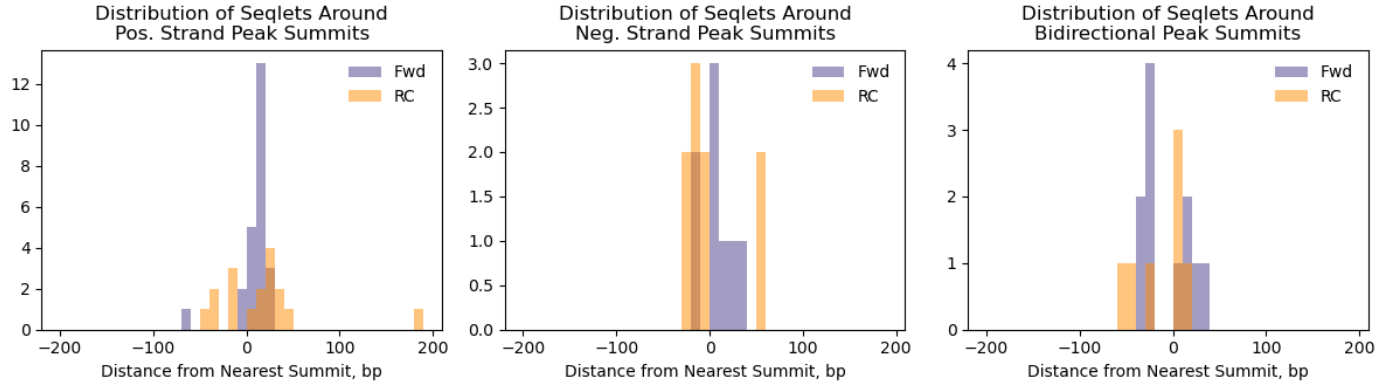
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





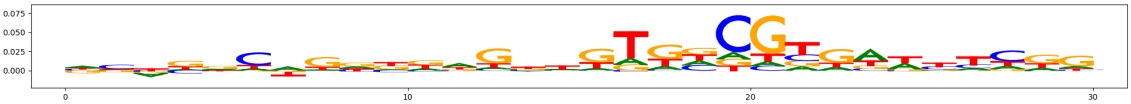


Pattern 30/39

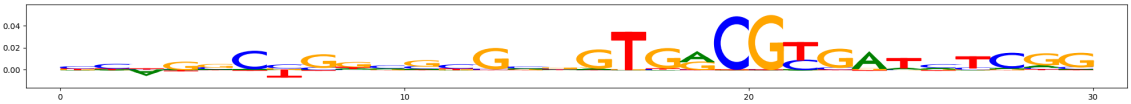
67 seqlets

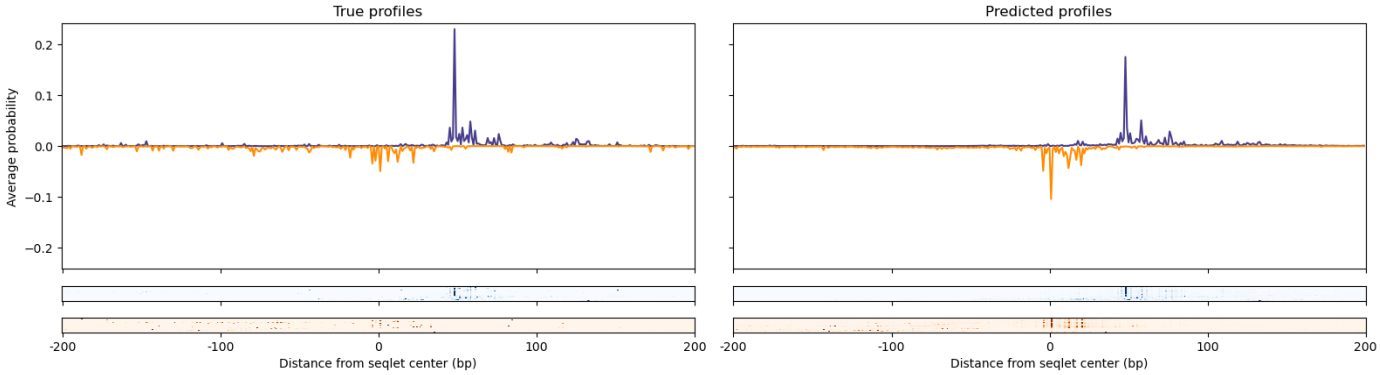
Sequence (PFM)

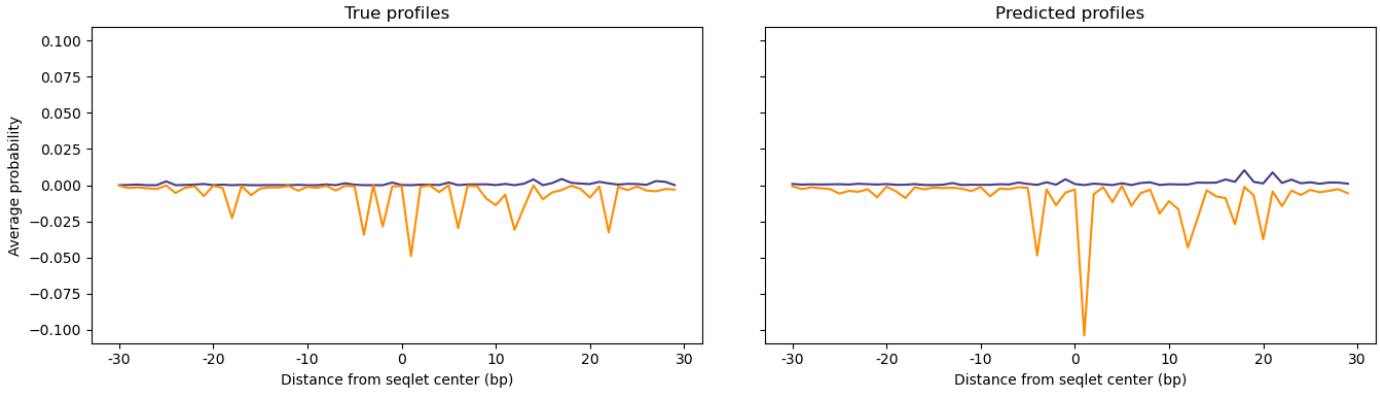
Hypothetical contributions (hCWM)

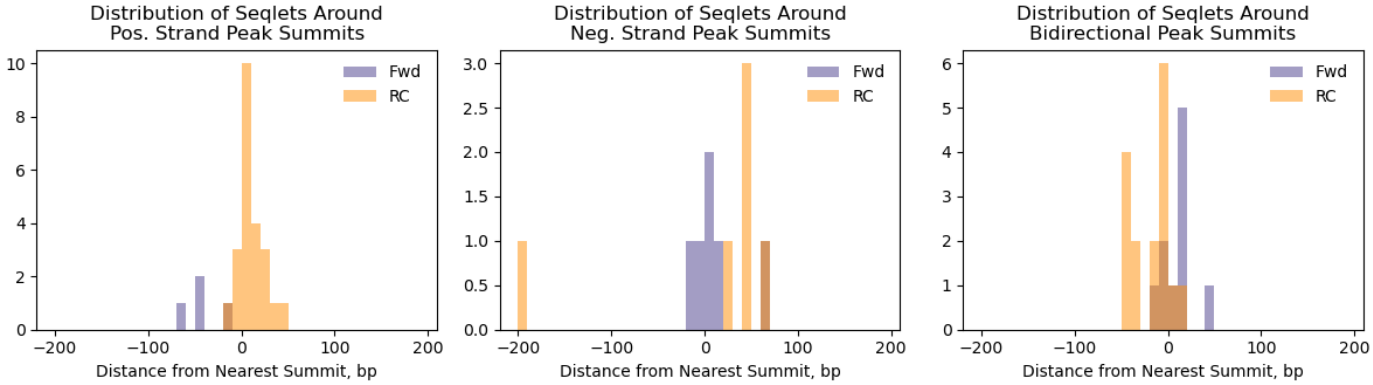


Actual contributions (CWM)



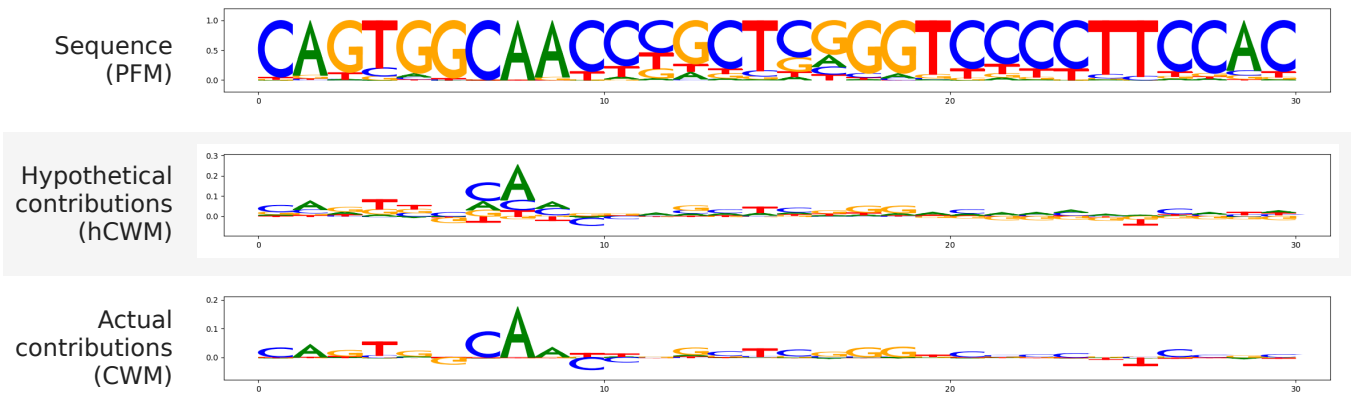


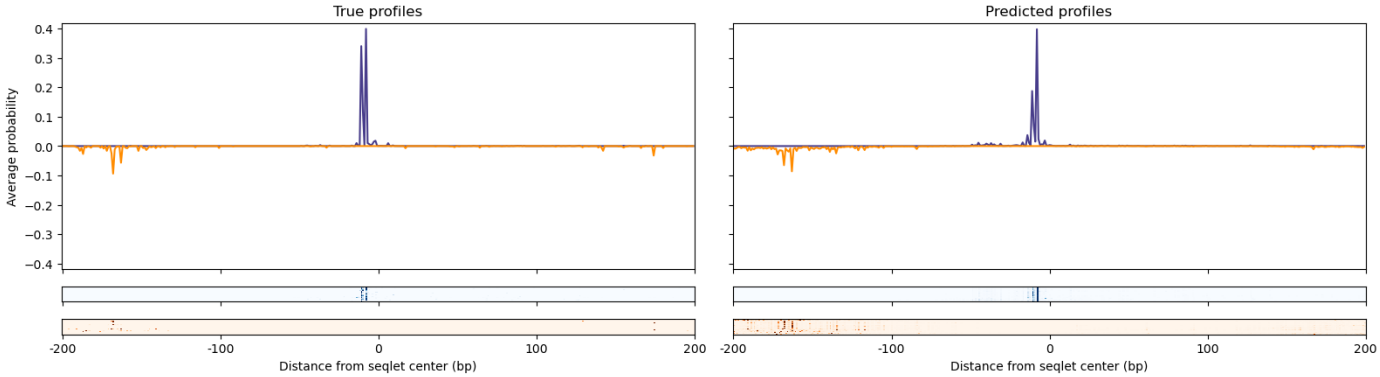


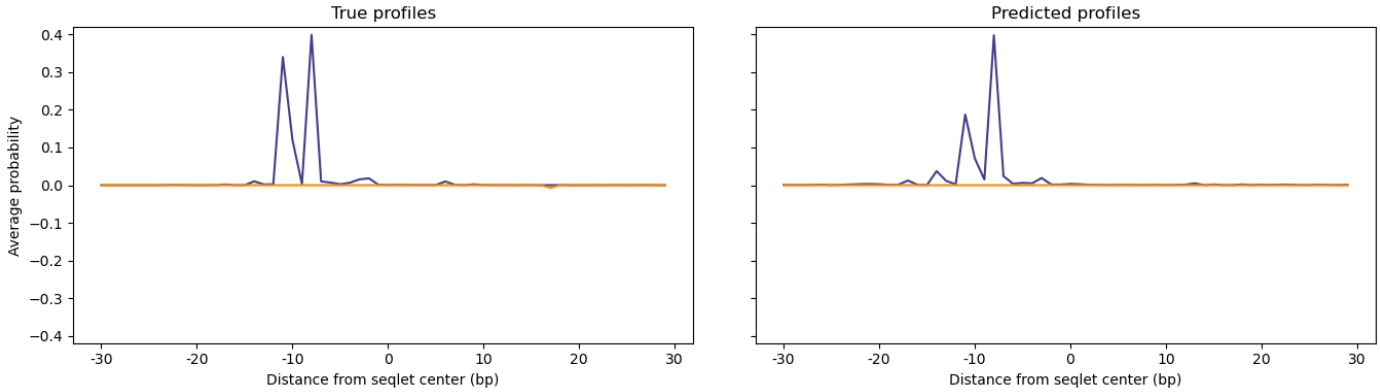


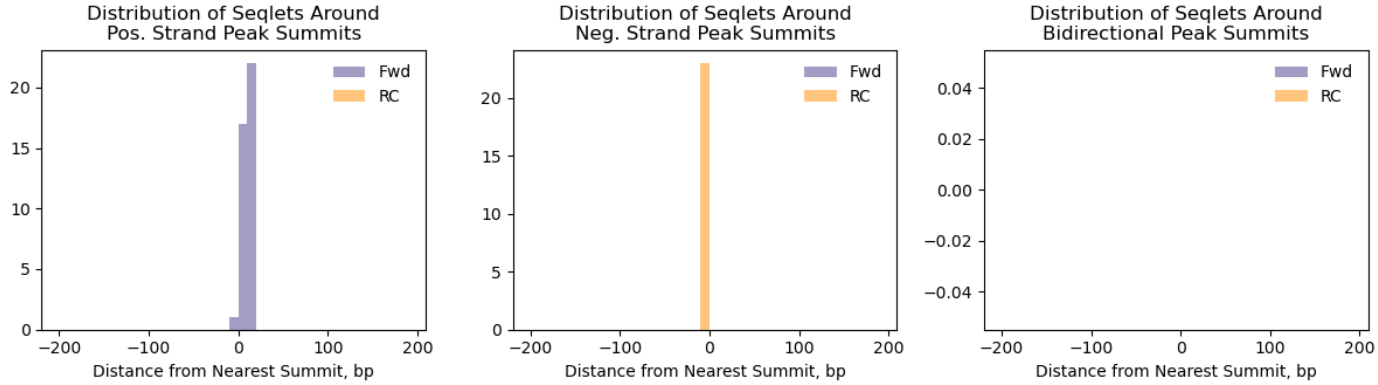
Pattern 31/39

63 seqlets









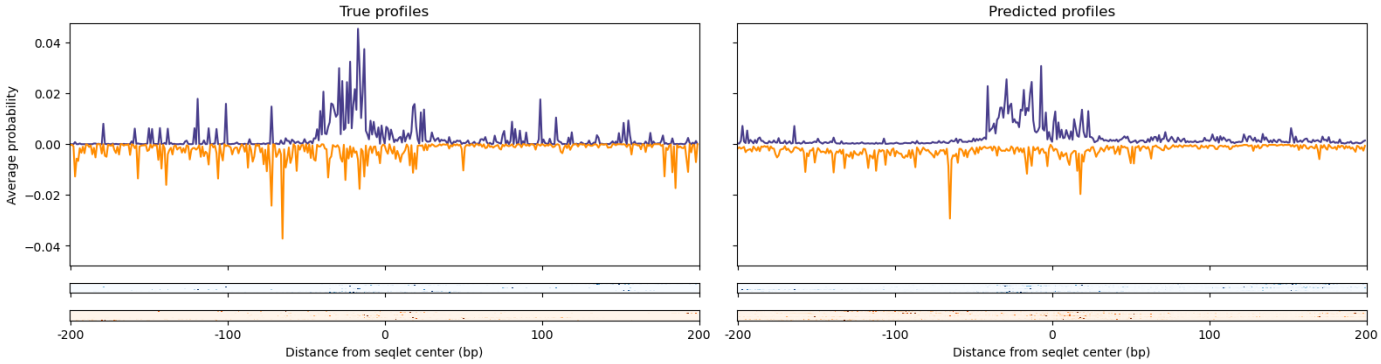
Pattern 32/39

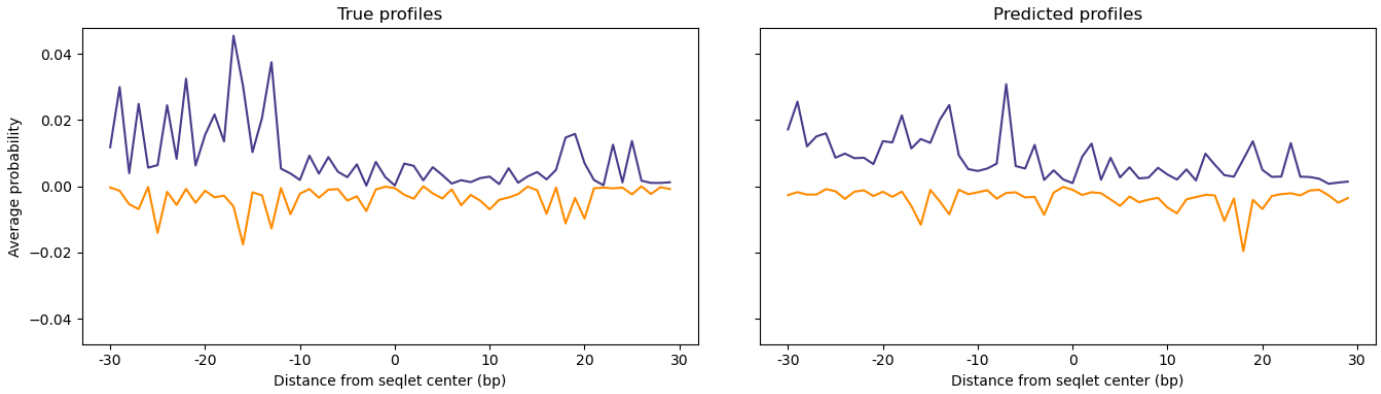
54 seqlets

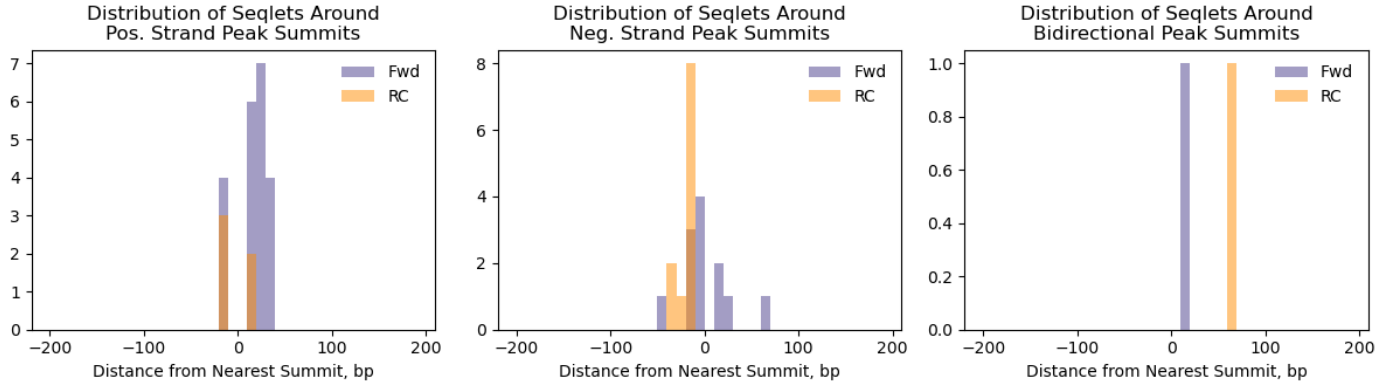
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





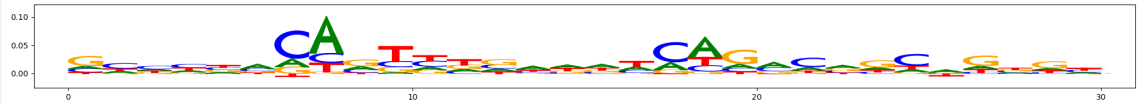


Pattern 33/39

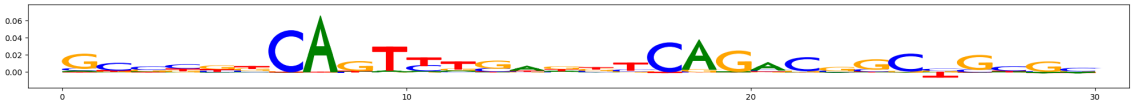
49 seqlets

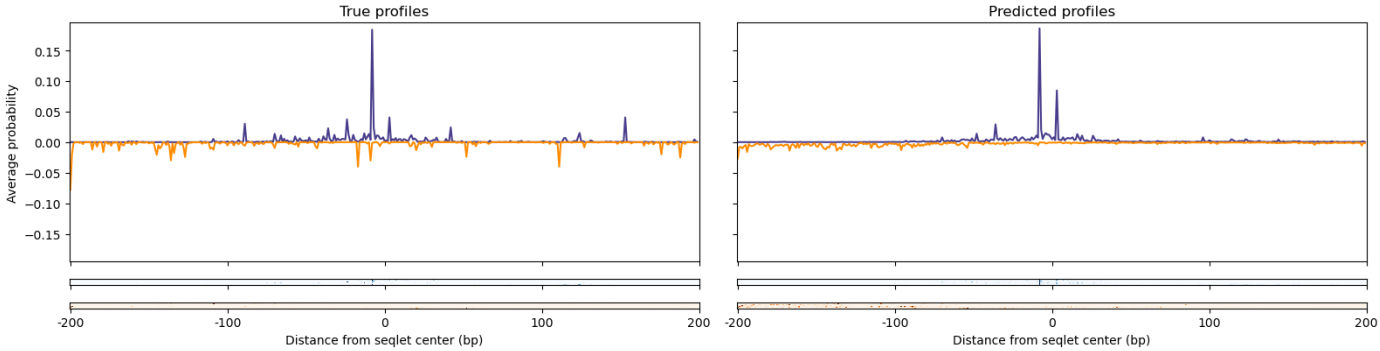
Sequence (PFM)

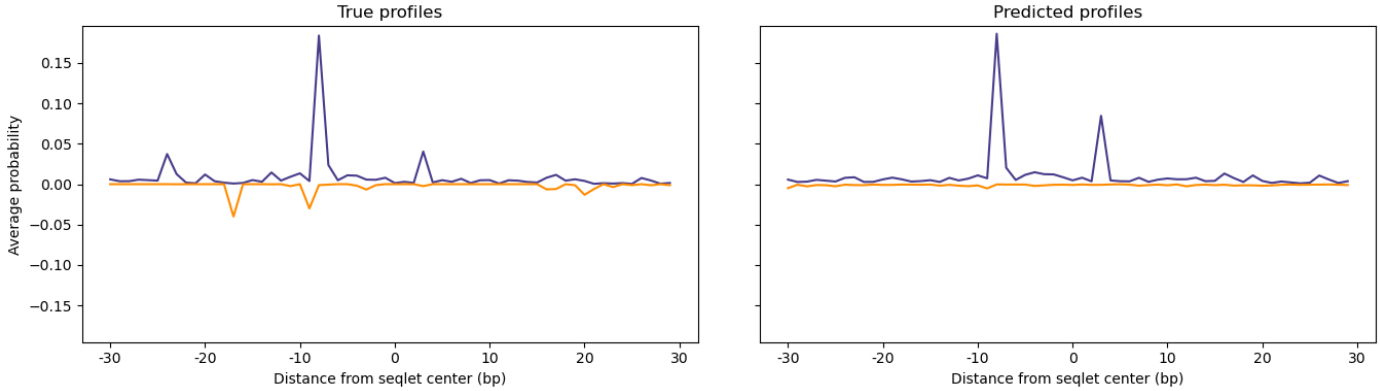
Hypothetical contributions (hCWM)

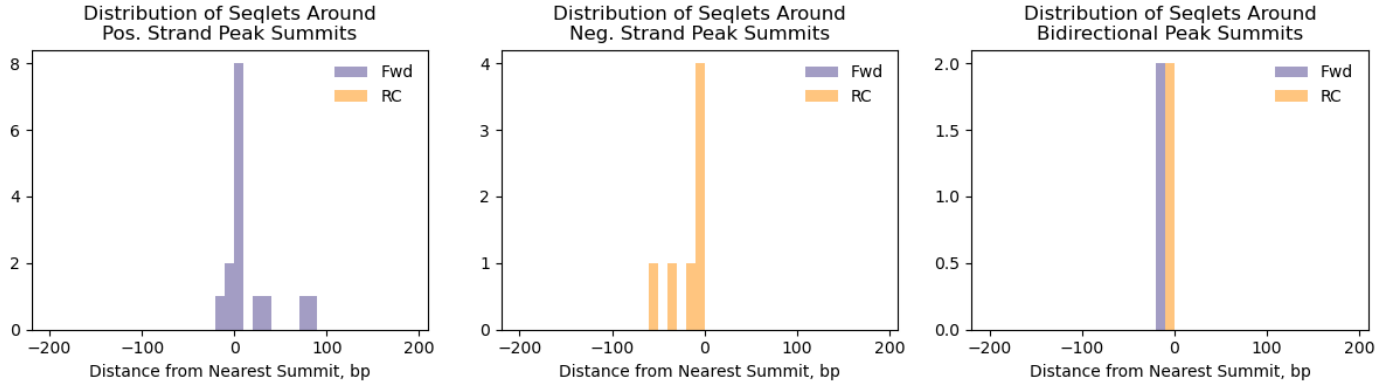


Actual contributions (CWM)









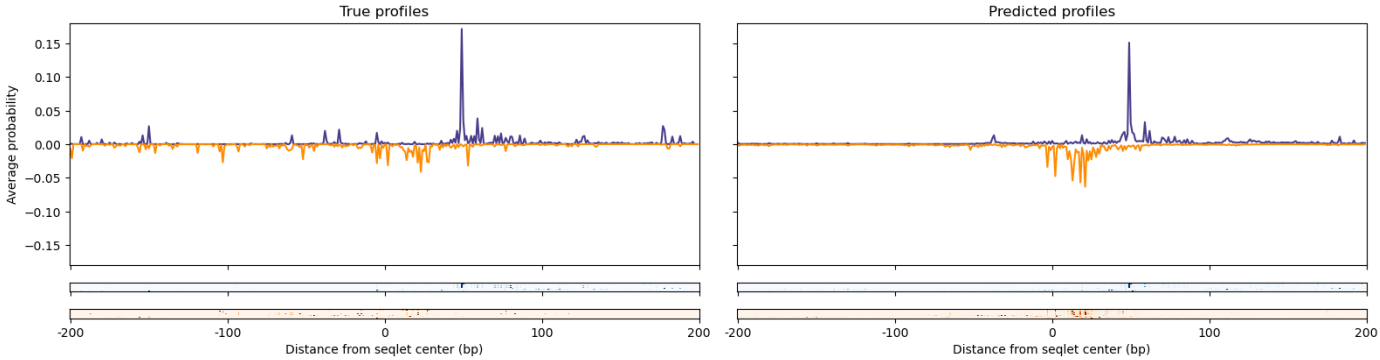
Pattern 34/39

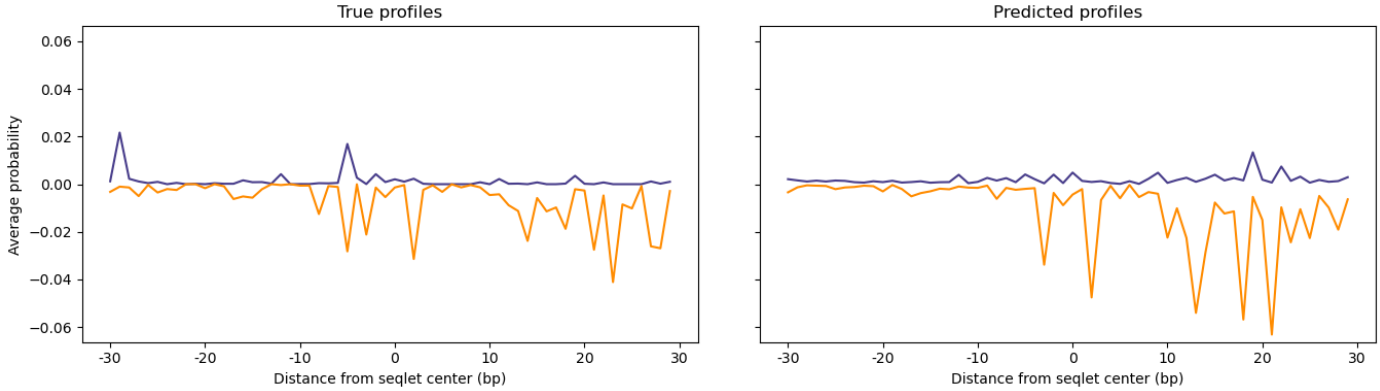
41 seqlets

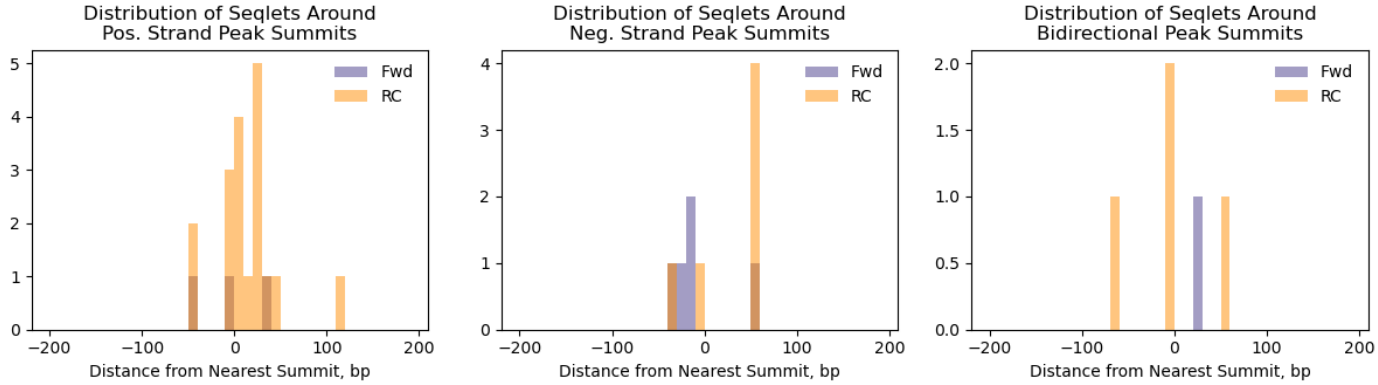
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)







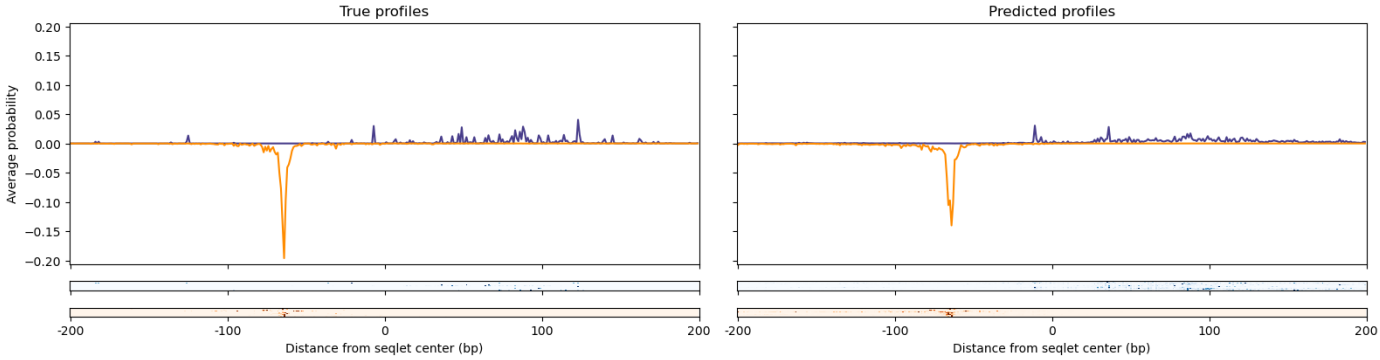
Pattern 35/39

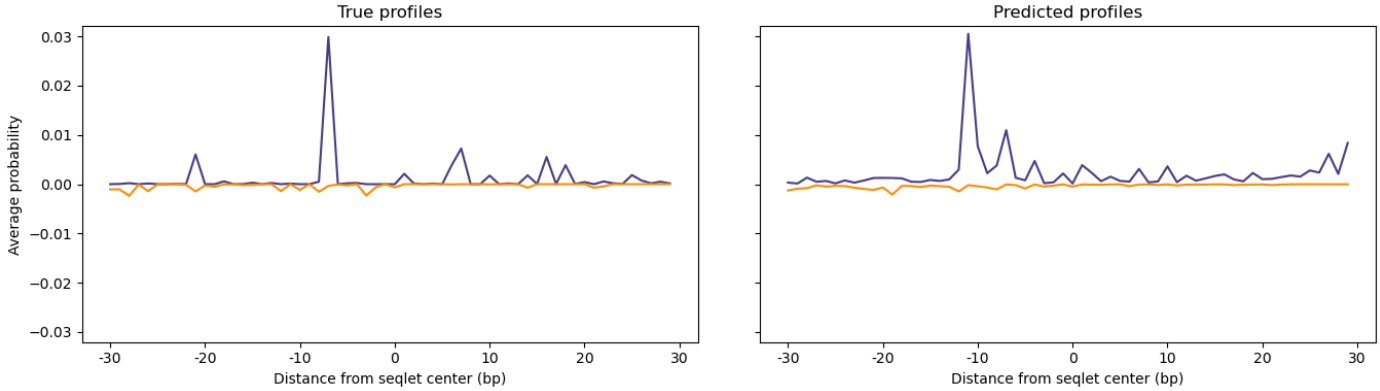
40 seqlets

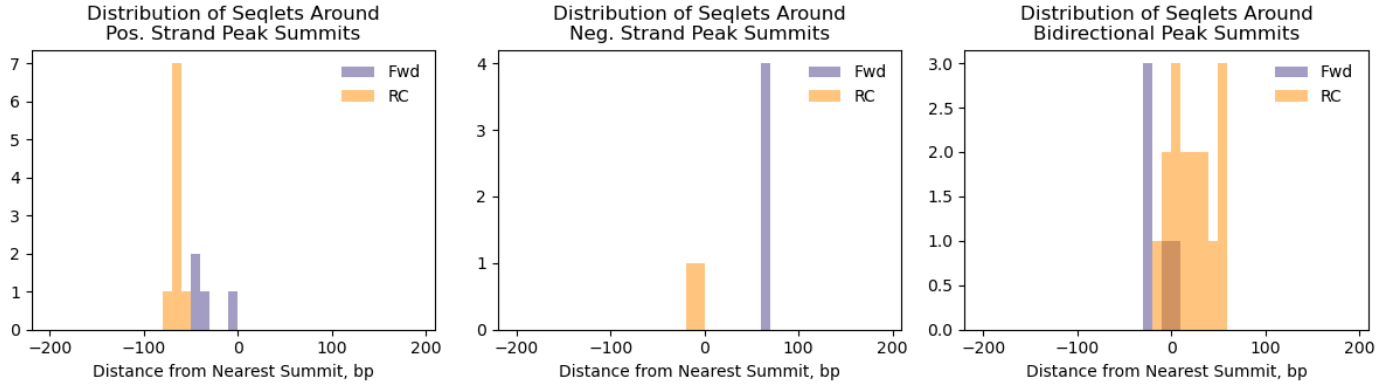
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)





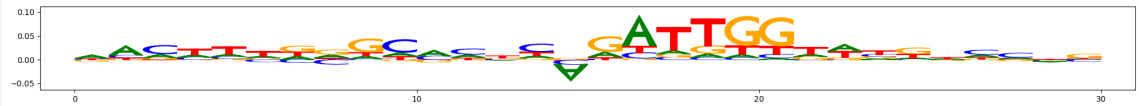


Pattern 36/39

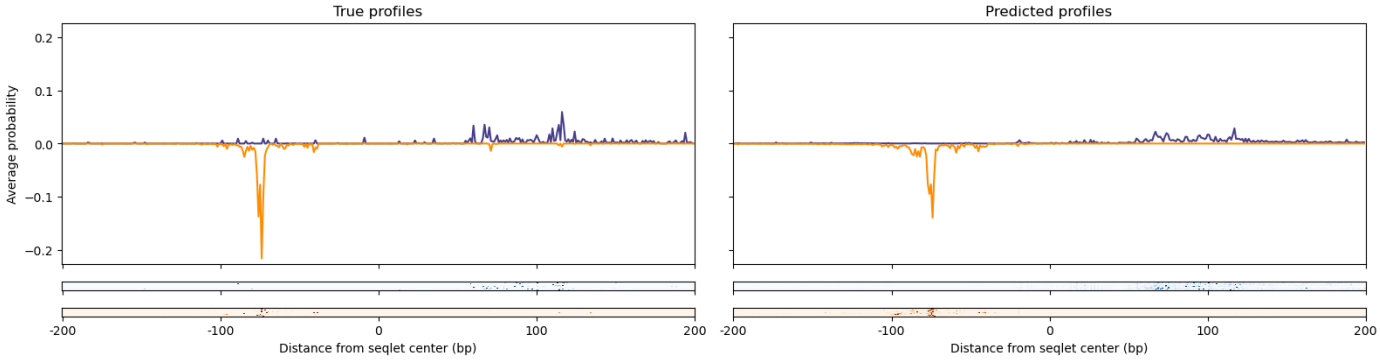
39 seqlets

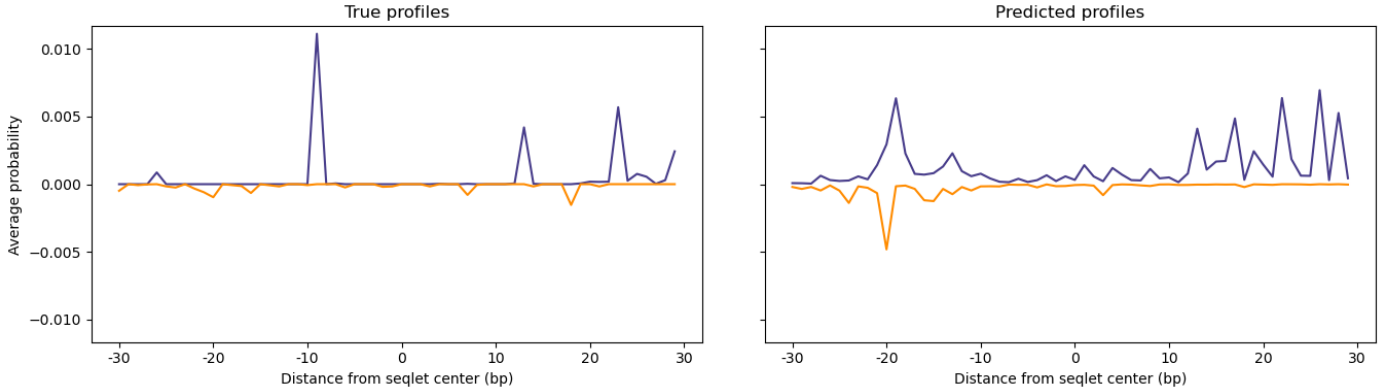
Sequence (PFM)

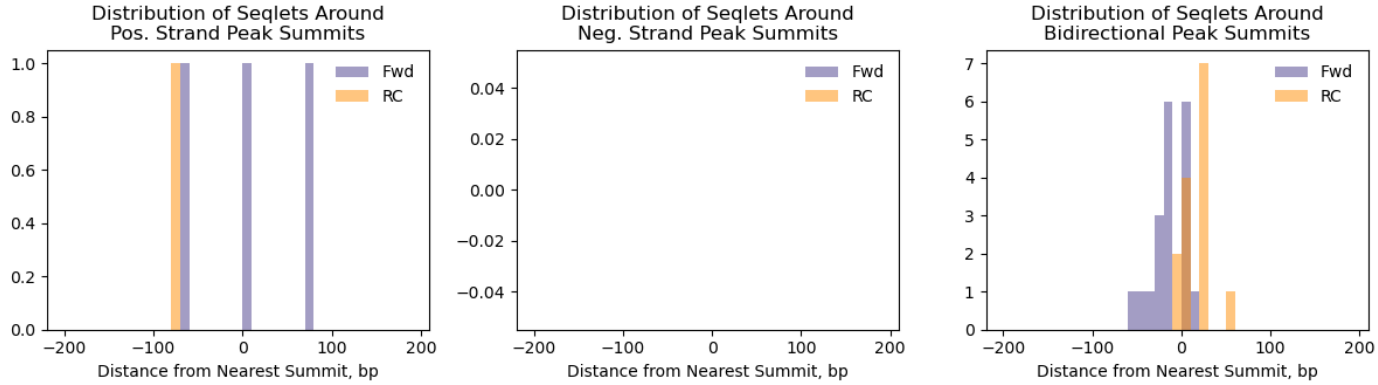
Hypothetical contributions (hCWM)



Actual contributions (CWM)





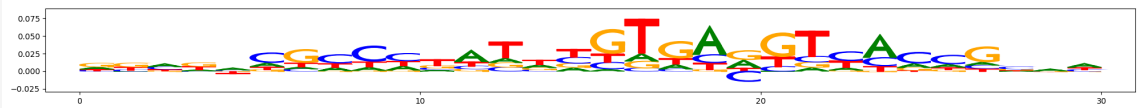


Pattern 37/39

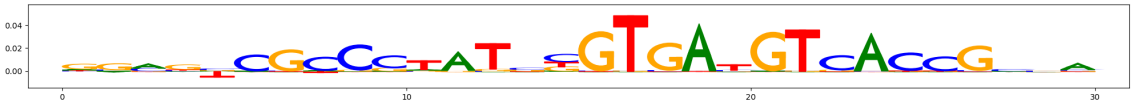
34 seqlets

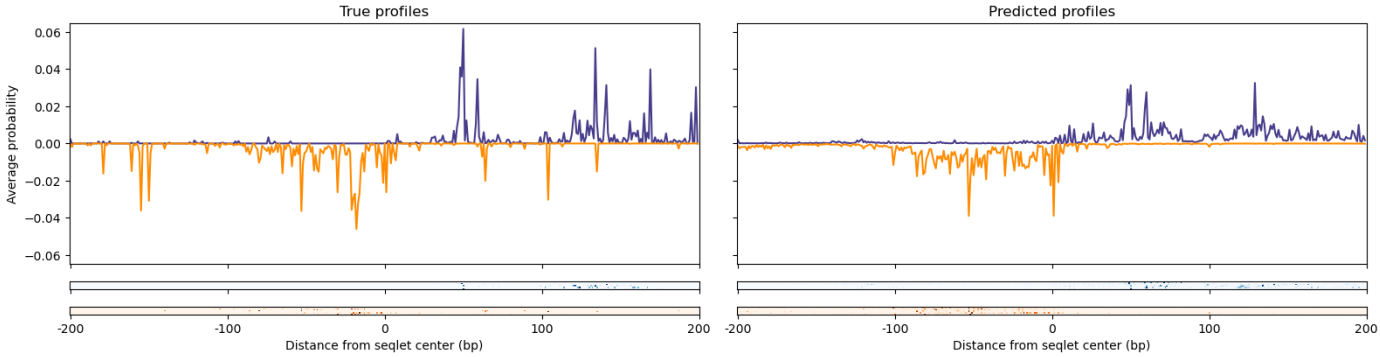
Sequence (PFM)

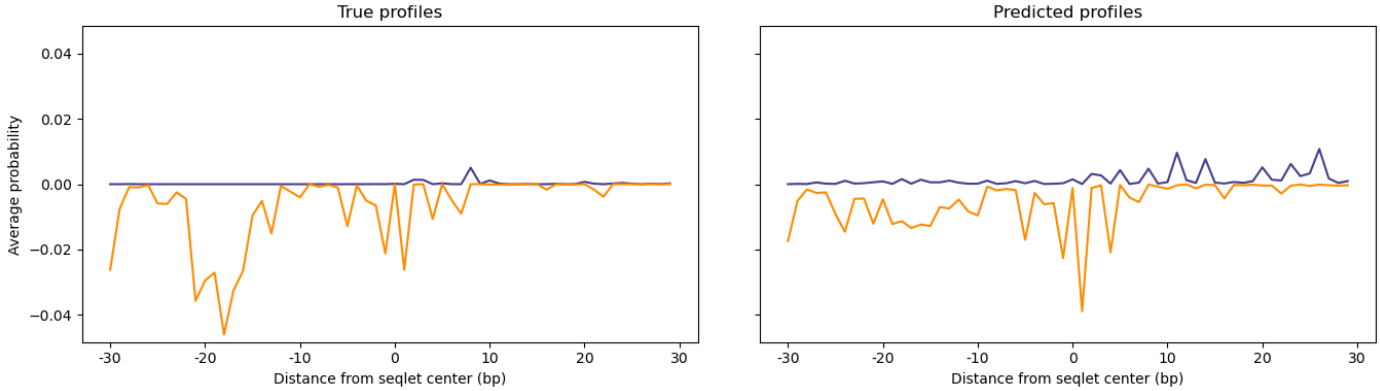
Hypothetical contributions (hCWM)

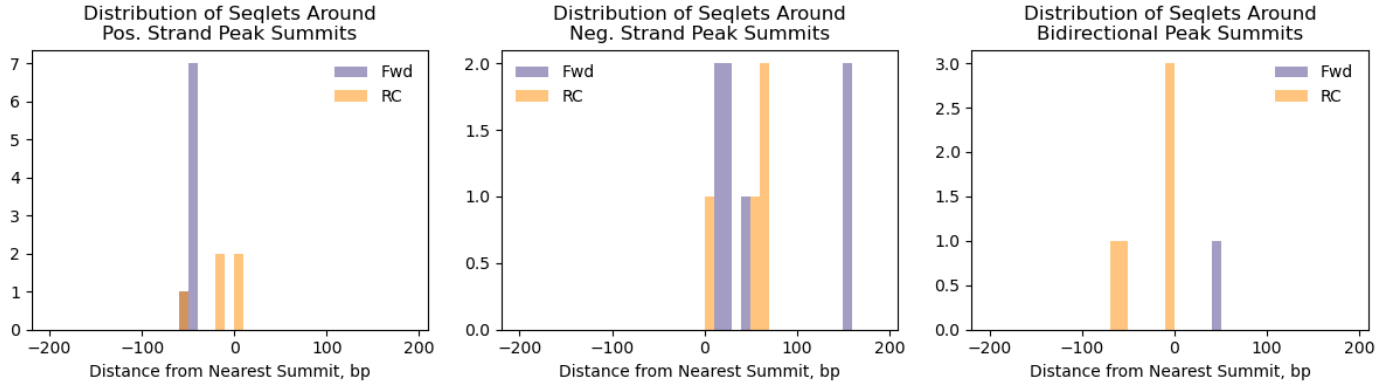


Actual contributions (CWM)







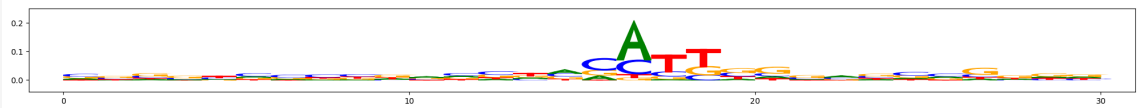


Pattern 38/39

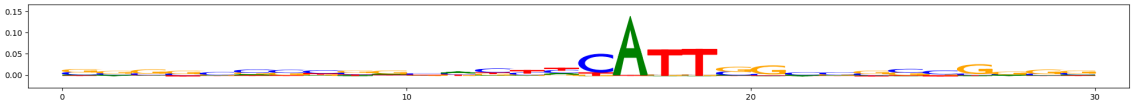
24 seqlets

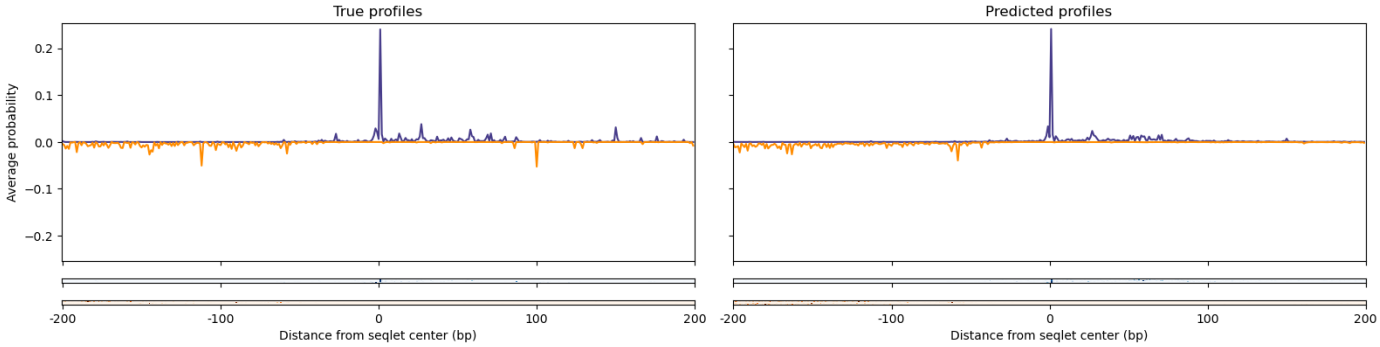
Sequence (PFM)

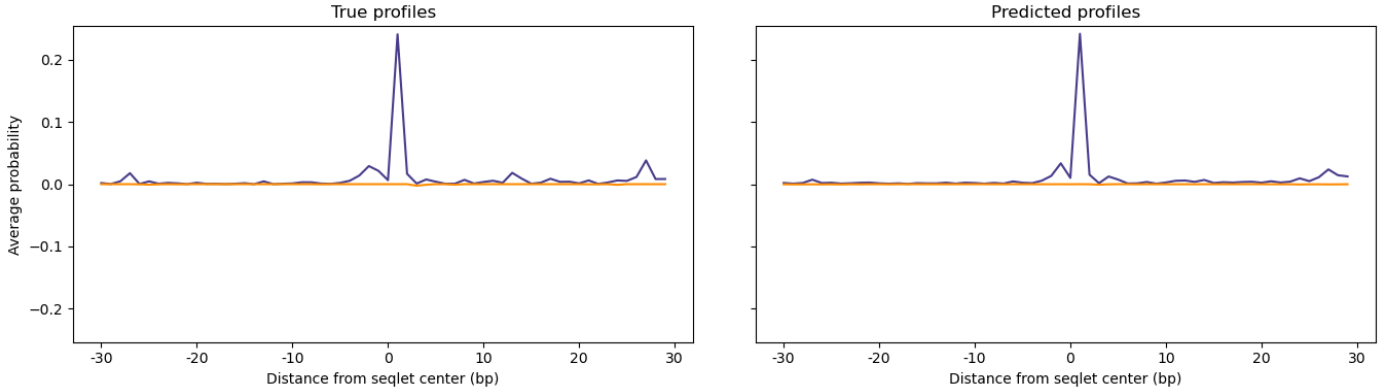
Hypothetical contributions (hCWM)

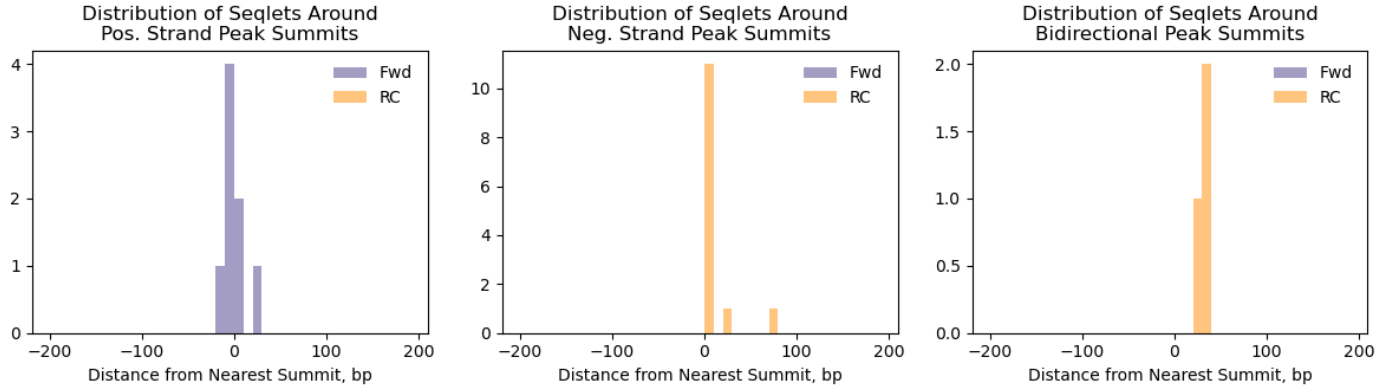


Actual contributions (CWM)







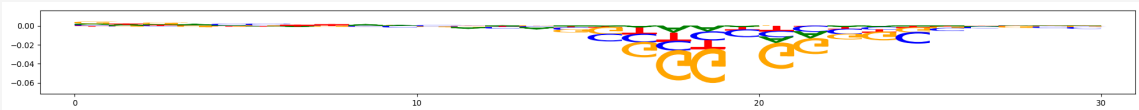


Pattern 0/4

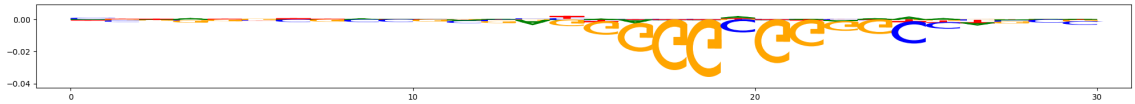
87 seqlets

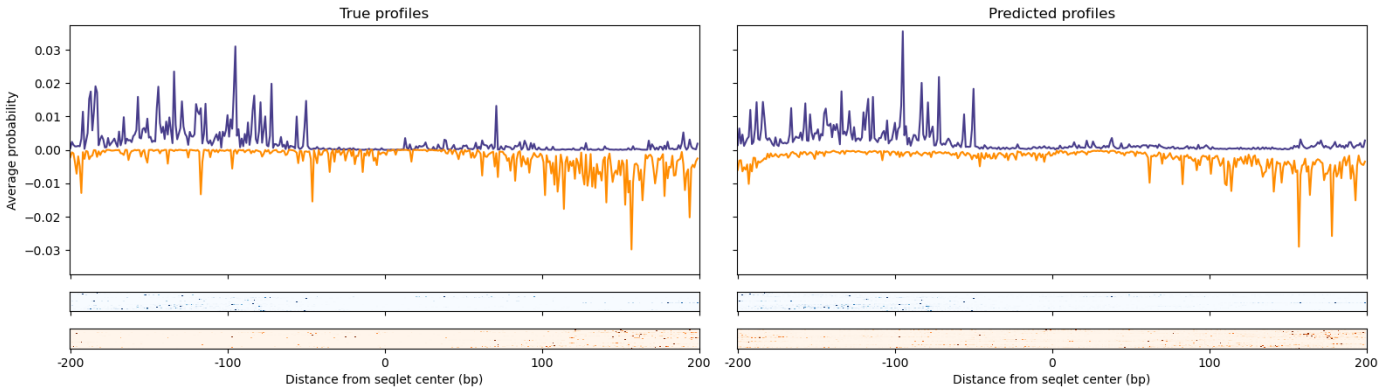
Sequence (PFM)

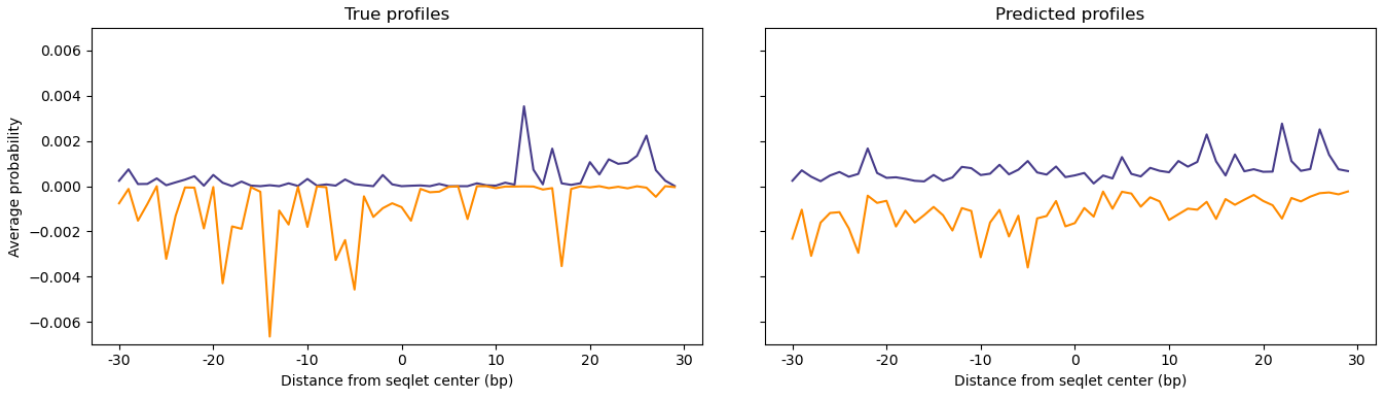
Hypothetical contributions (hCWM)

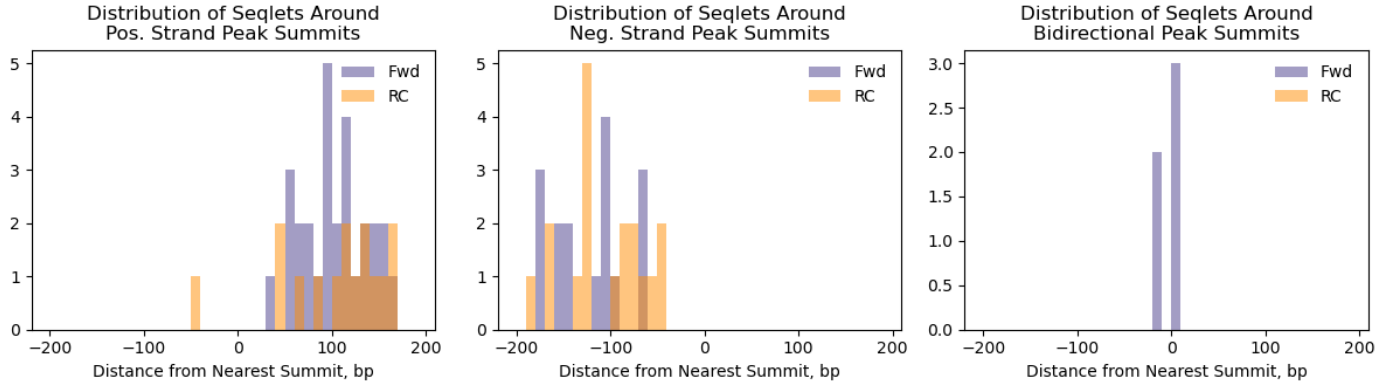


Actual contributions (CWM)







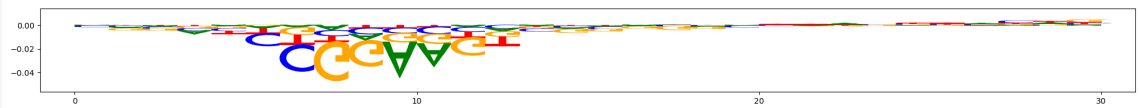


Pattern 1/4

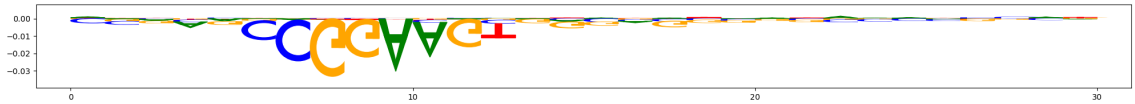
61 seqlets

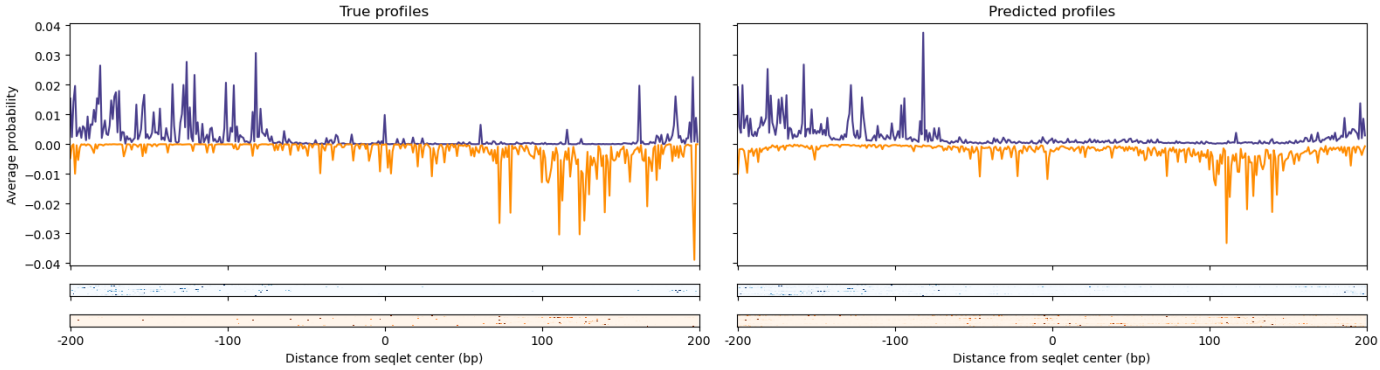
Sequence (PFM)

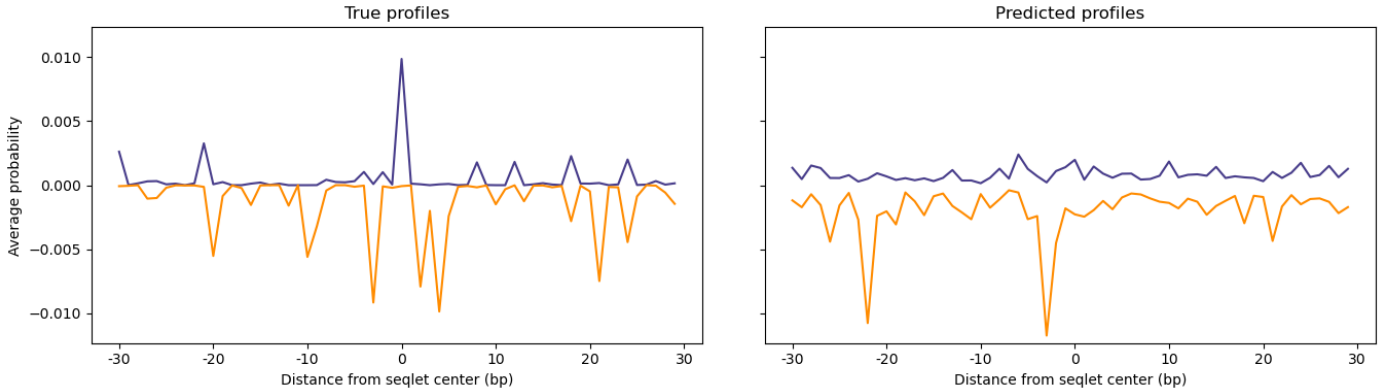
Hypothetical contributions (hCWM)

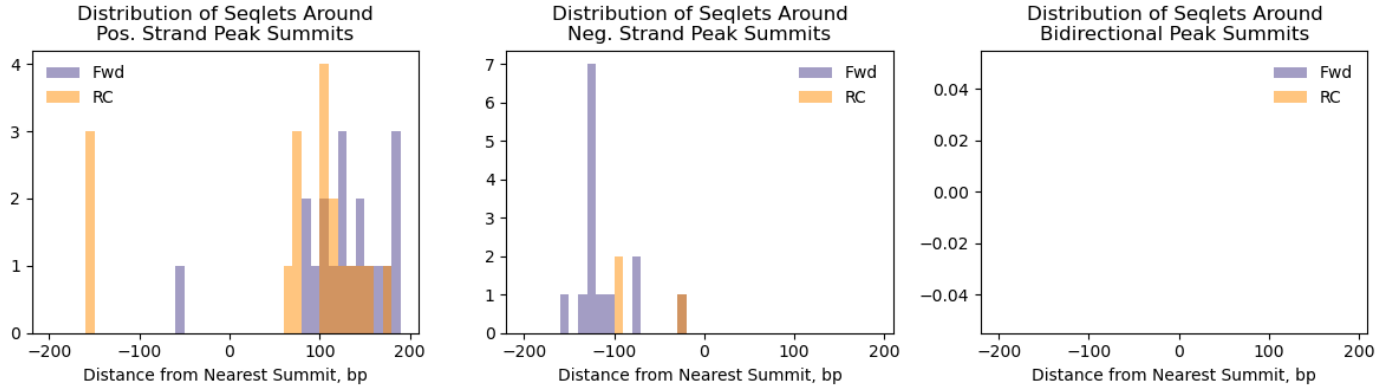


Actual contributions (CWM)







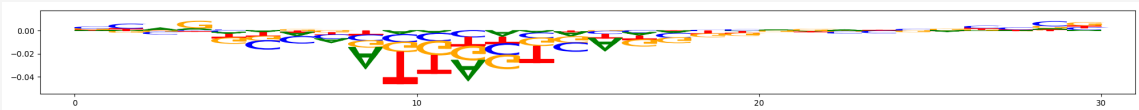


Pattern 2/4

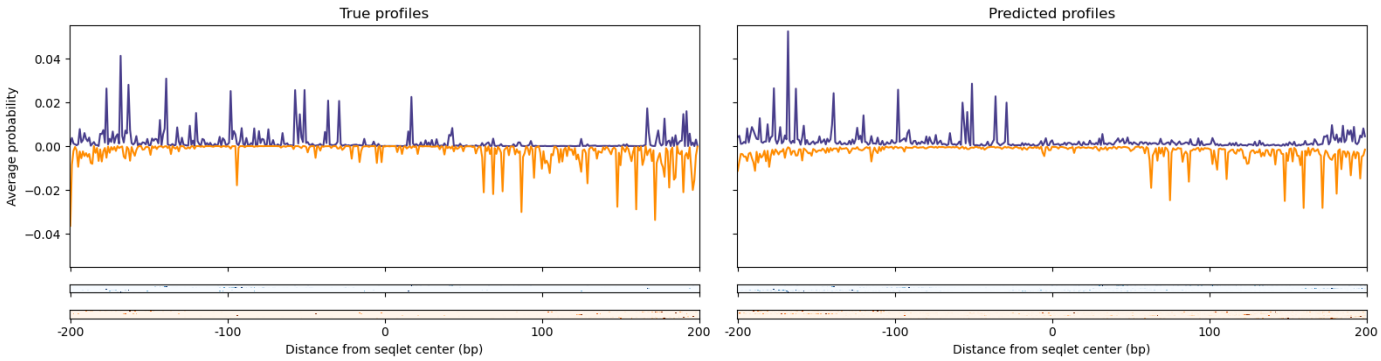
42 seqlets

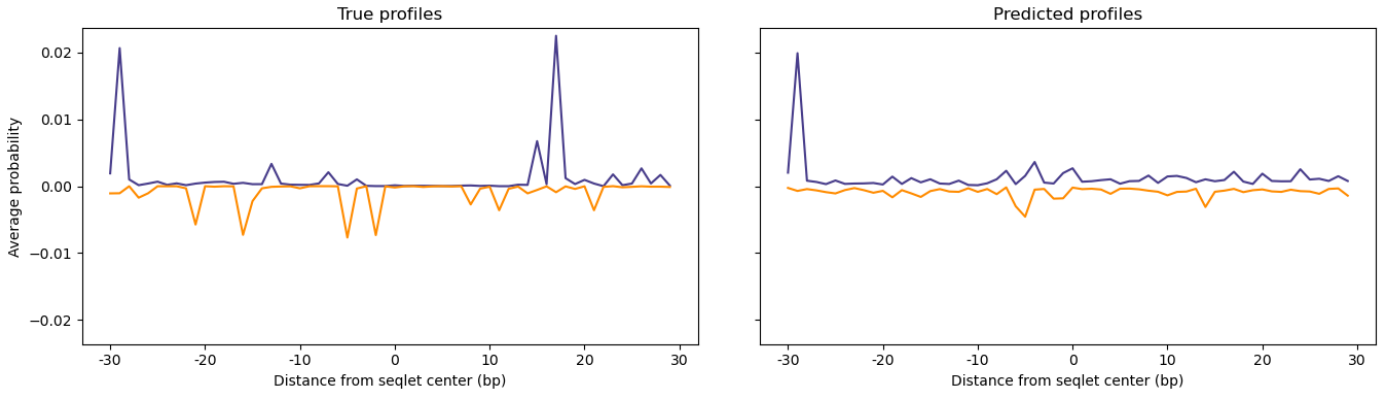
Sequence (PFM)

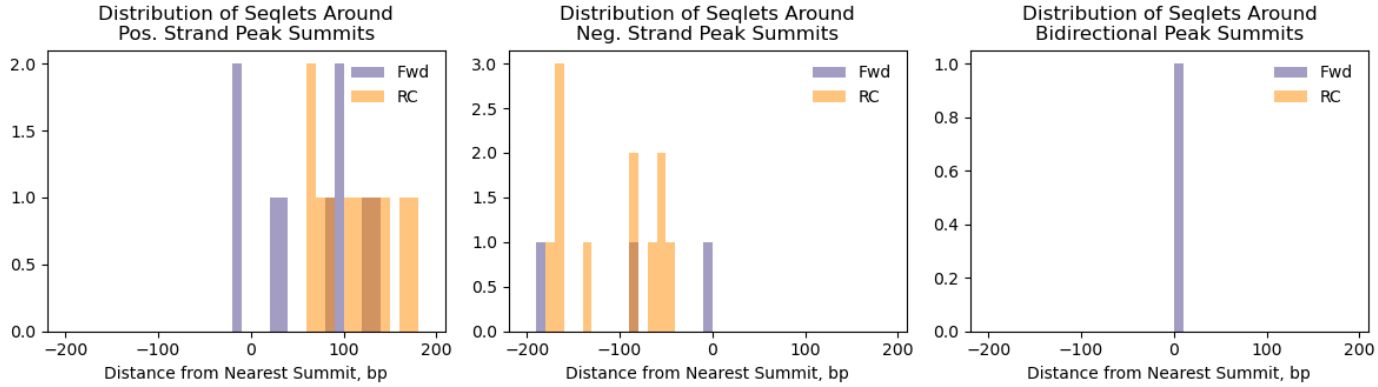
Hypothetical contributions (hCWM)



Actual contributions (CWM)







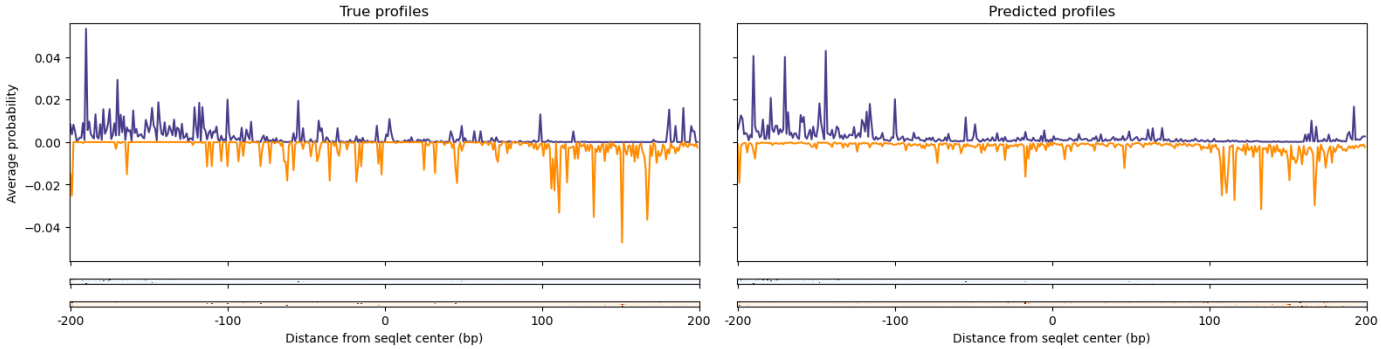
Pattern 3/4

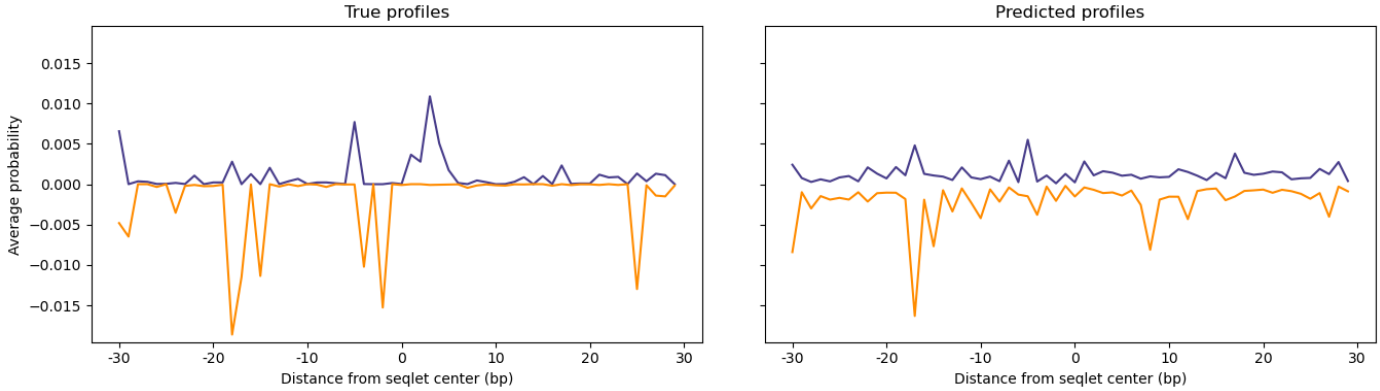
25 seqlets

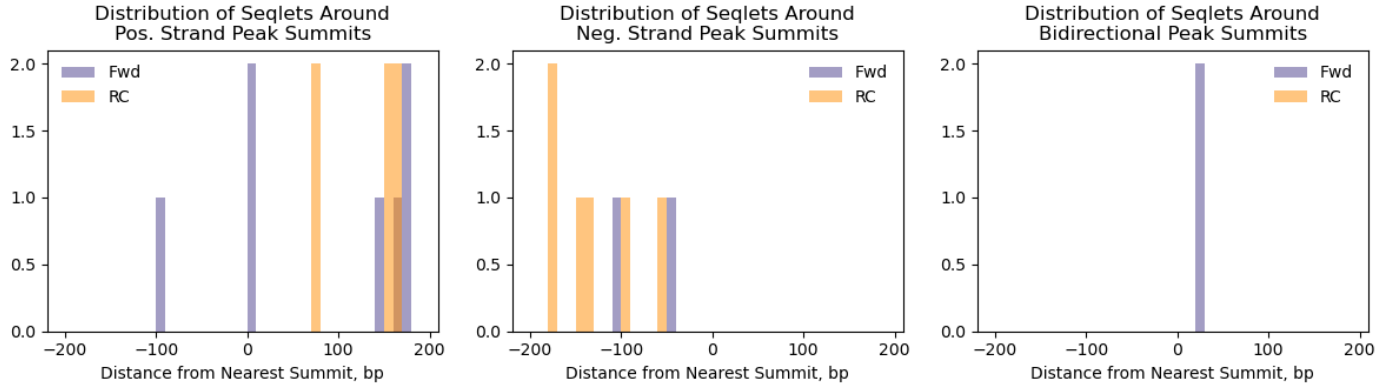
Sequence (PFM)

Hypothetical contributions (hCWM)

Actual contributions (CWM)






Distribution of Seqlets Around
Pos. Strand Peak Summits
Fwd
RC
Distance from Nearest Summit, bp
Distribution of Seqlets Around
Neg. Strand Peak Summits
Fwd
RC
Distance from Nearest Summit, bp
Distribution of Seqlets Around
Bidirectional Peak Summits
Fwd
RC
Distance from Nearest Summit, bp

In [ ]: