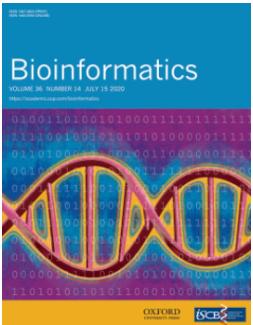


FastSK: Fast Sequence Analysis with Gapped String Kernels



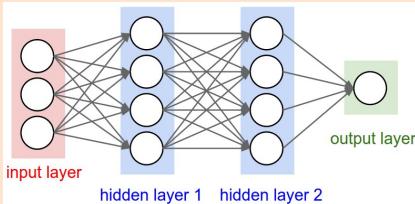
Dr. Yanjun Qi

University of Virginia

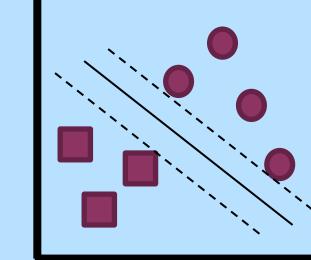
<https://github.com/QData/FastSK>

ECCB Parallel Proceeding Track #03
Monday, Sep 7, 2020 10:45 AM

State-of-the-art String Classification Models

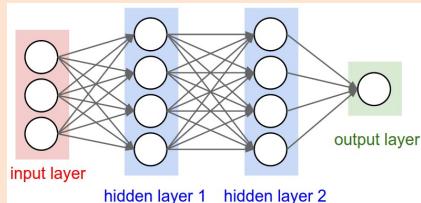


Deep Neural Network
(DNN)

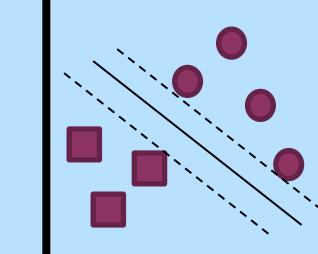
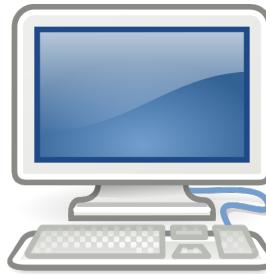


Support Vector Machine
(SVM)

State-of-the-art String Classification Models



Deep Neural Network
(DNN)

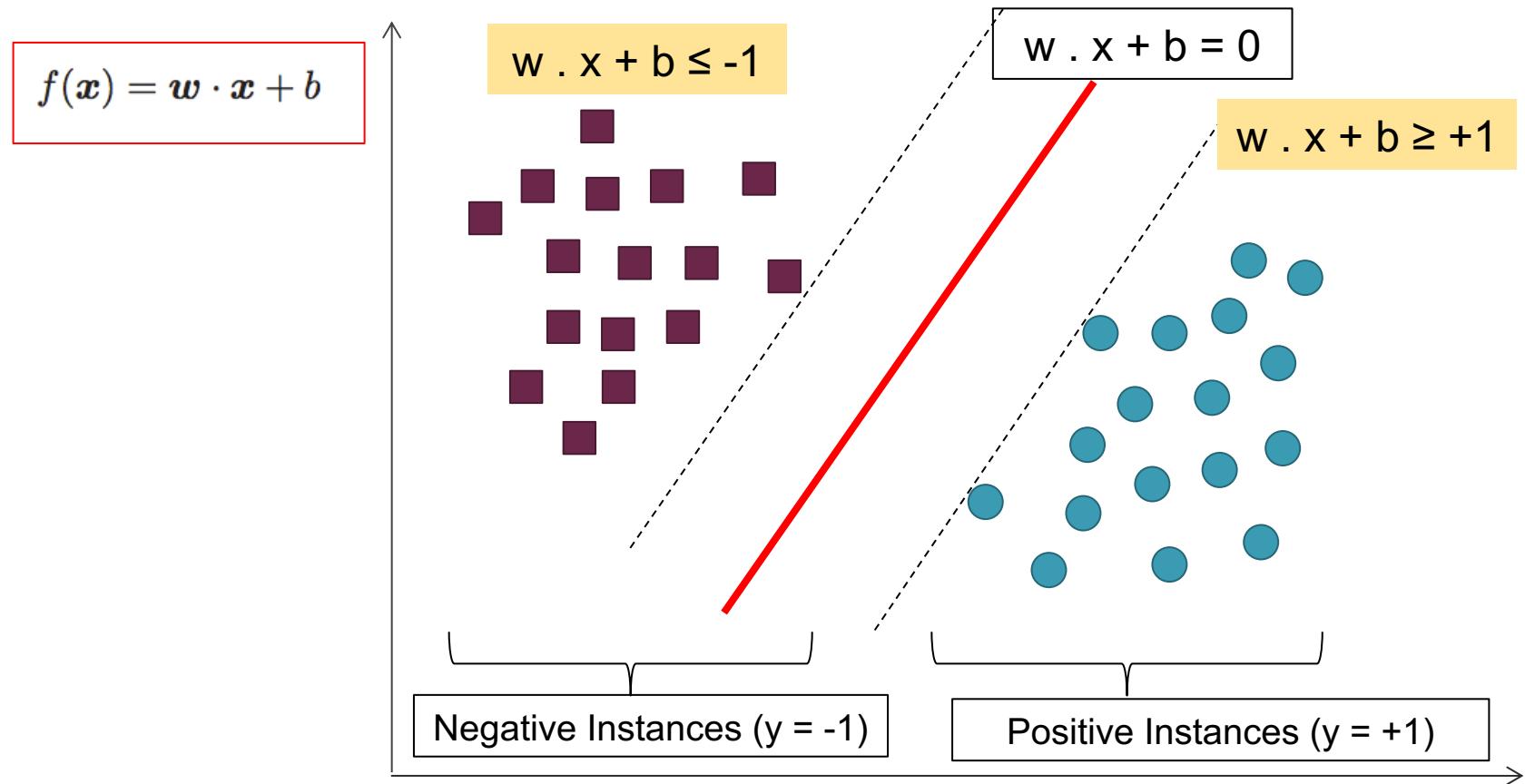


Support Vector Machine
(SVM)



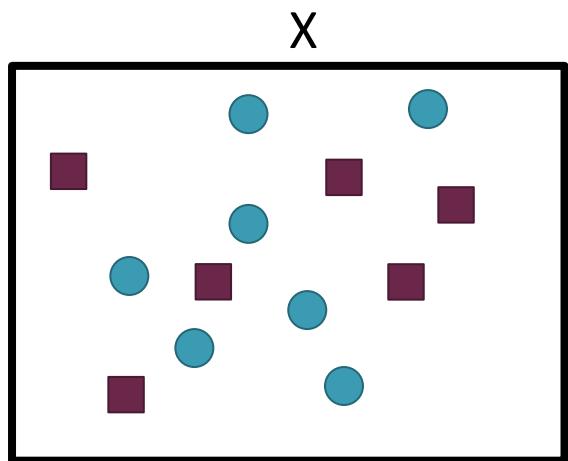
Number of Training Samples

Support Vector Machine (SVM)



Kernel

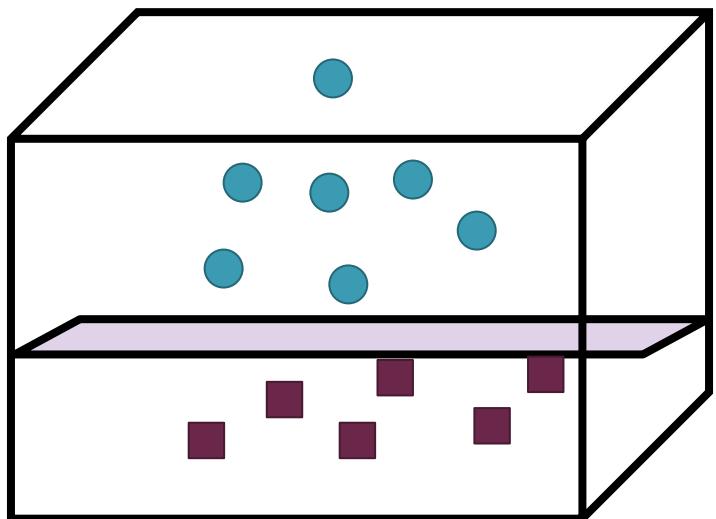
X



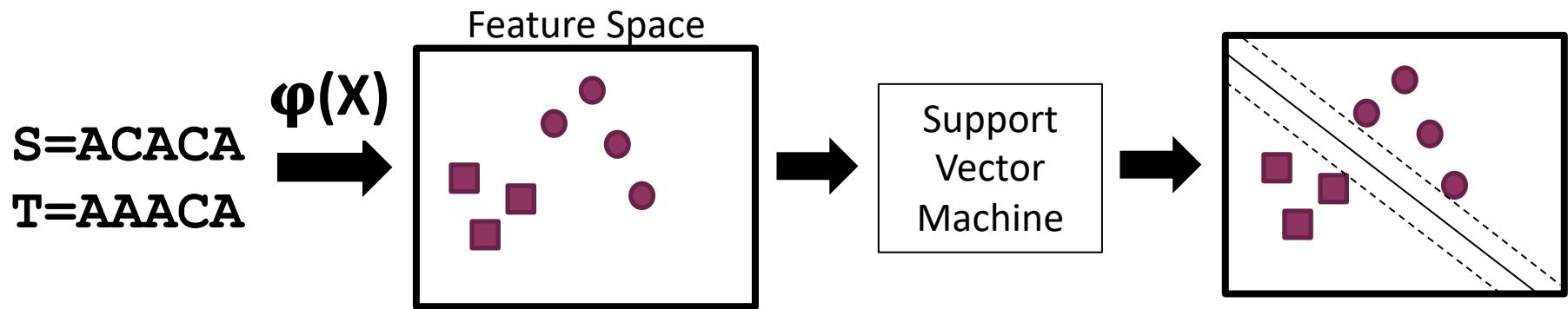
$K(\cdot)$

$$f(x) = w \cdot \Phi(x) + b.$$

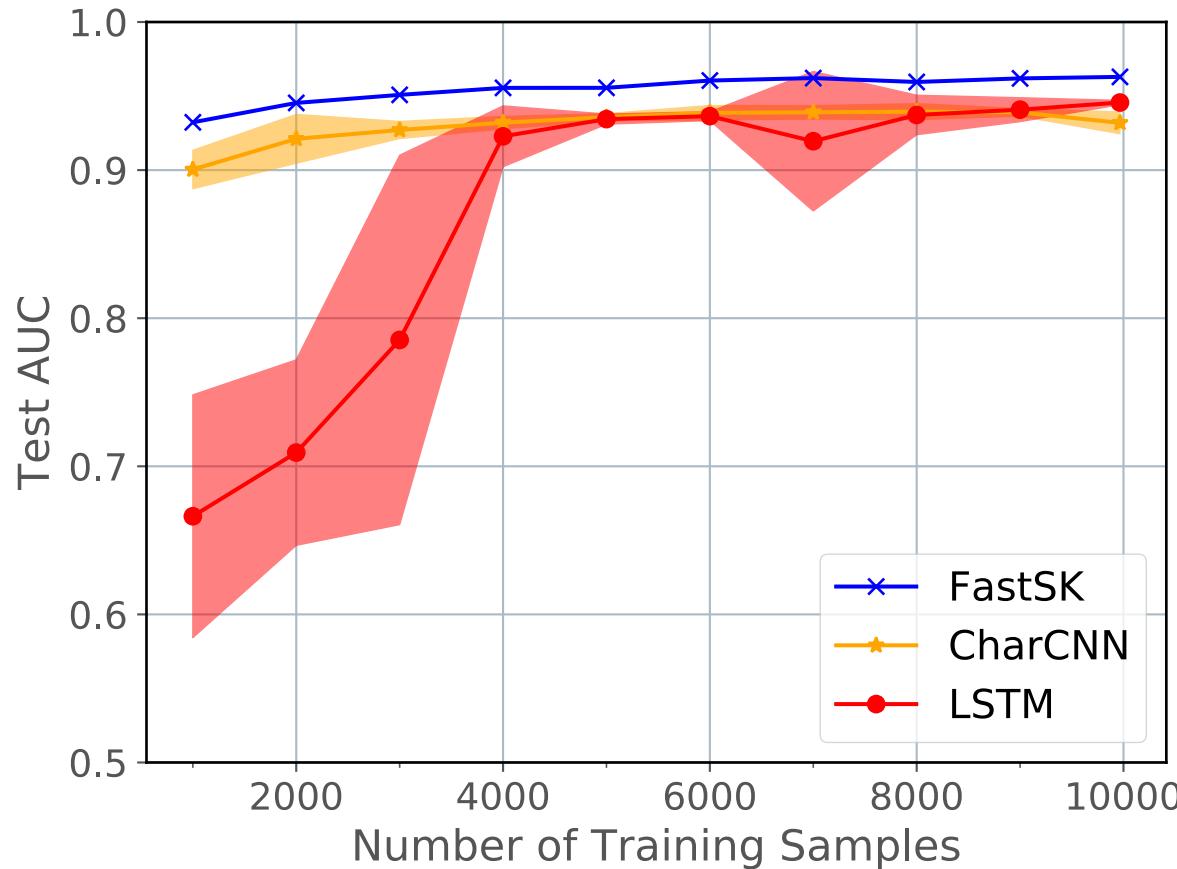
$\varphi(X)$



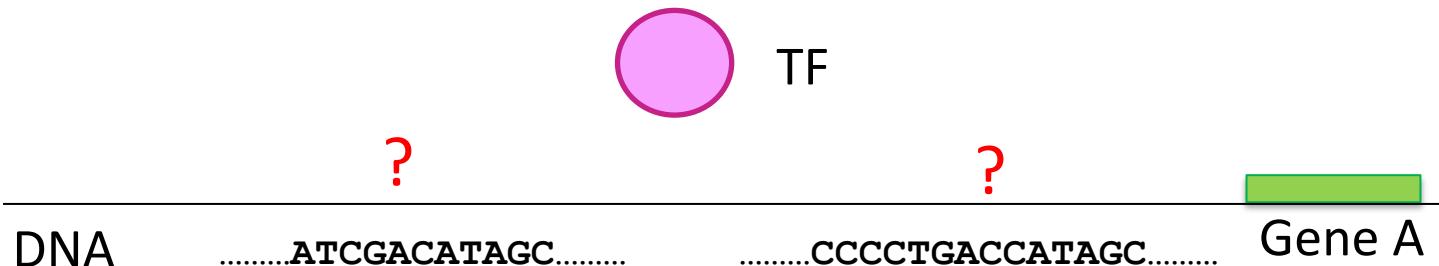
String Kernel + SVM Framework



String Kernel-SVM Predicts Well with Low Variance



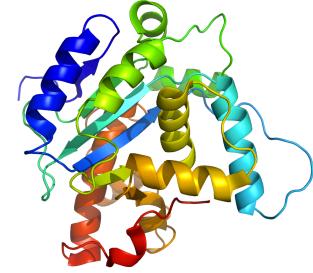
[Application I]: Does TF bind to this DNA sequence?



ATCGAATCCG ✓
CGCTGAATCG ✗
ATCGCTATCG ✓
ATCCCGCTCG ✗

$\Sigma=4$
Minimum Number
of Training Samples
1000

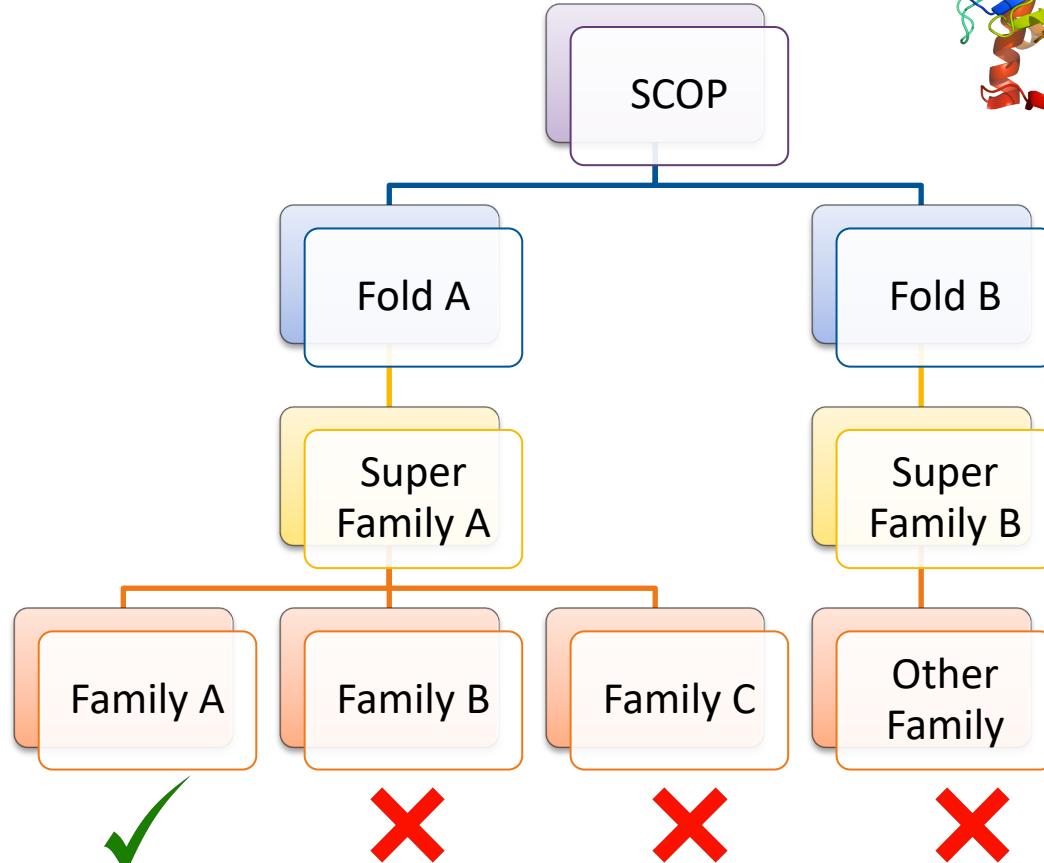
[Application III]: What family does a protein sequence belong to?



vQGGHACCAKKQQQ

$$\Sigma=20$$

Minimum Number
of Training Samples
500



[Application III]: Language Classification:



This Food is not good.

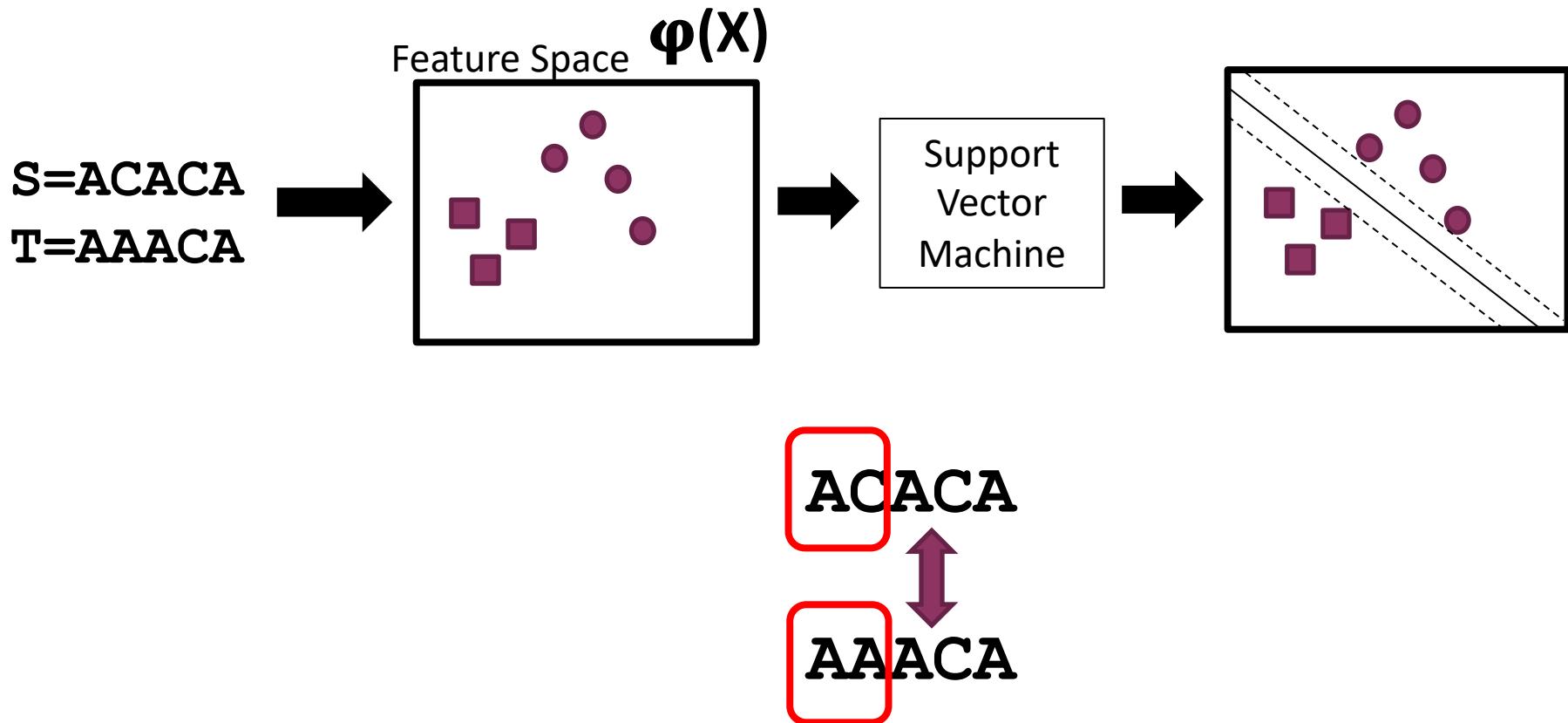


English

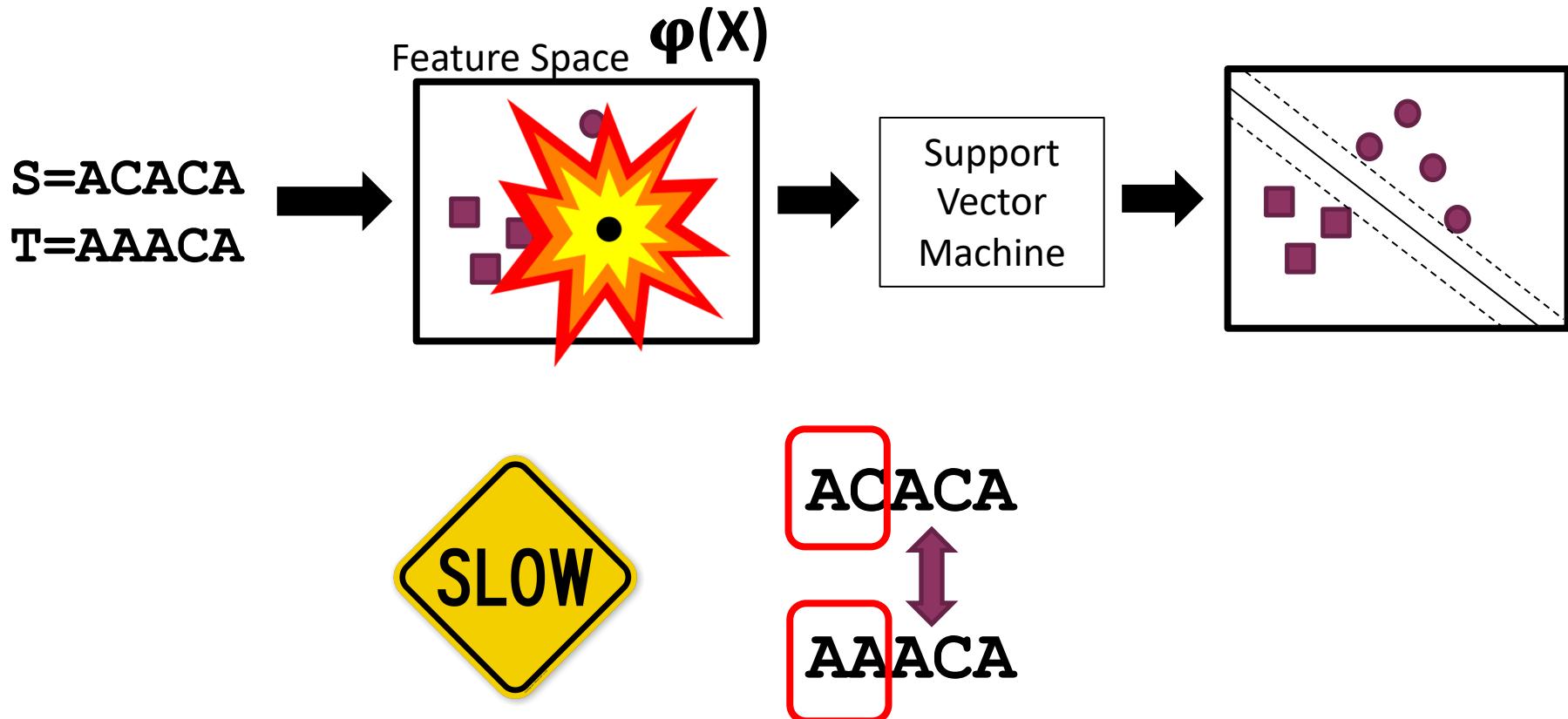
• ام يجي اي شويه يمسا ، هبس انمل اب	CAI
• ام يجي اي شوريه يمسا ، هبس انمل اب	BEI
• ام يجي اي شوريه يمسا ، هركف ىلع	DOH
• ام يجي اي شوريه يتيمس ، هبس انمل اب	RAB
• ام يجي اي شوريه يمسا ، هركف ىلع	TUN
• ام يجي اي شوريه يمسا ، هبس انمل اب	MSA

Can we identify a writer's
region using just text info?

String Kernels



Challenge : SK-SVM methods are slow



String Kernel

$\Sigma=2 \{A,C\}$

$k=3$

$m=1$

Bag of
k-mers

Mismatch Kernel

S:ACACA

T:AAAACA

Feature Space

k-mers

ACA

CAC

ACA

0	AAA	1
0	AAC	1
2	ACA	1
0	CAA	0
1	CAC	0
0	CCC	0

Mismatch Neighborhood
 $\{AAC, ACA, CAA..\}$

k-mers

AAA

AAC

ACA

Computational Challenge

$\Sigma=20$

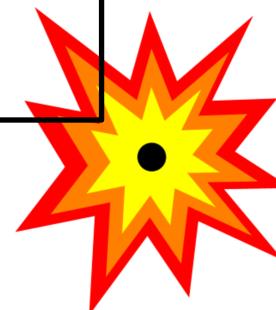
$k=10$

Feature
Space

Σ^k

$$(20)^{10} \sim= 10^{13}$$

Sparse!	AAA	1
	AAC	1
	ACA	0
	CAA	0
	CAC	0
	CCC	0



Related Work: Gapped k-mer Kernel (gkm-SVM)

$$\Sigma=2 \{A,C\}$$

$$g=3$$

$$k=2$$

S:ACACA

g-mers k-mers

ACA {AC,CA,...}

CAC

ACA

T:AAACA

g-mers k-mers

AAA {AA,AA,...}

AAC

ACA

$$\text{Gaps}=g-k=1$$

Feature Space

$$g(|S| + |T|) \ll \Sigma^k$$

Related Work: Gapped k-mer Kernel (gkm-SVM)

(1)

$$K(x, x') = \sum_{\gamma \in \Theta_g} c_x(\gamma) \cdot c_{x'}(\gamma)$$

(2)

$$K(x, x') = \sum_{i=0}^{l_1} \sum_{j=0}^{l_2} h_{gk}(g_i^x, g_j^{x'})$$

(3)

$$K(x, x') = \sum_{m=0}^{g-k} N_m(x, x') h_m$$

Related Work: Gapped k-mer Kernel

$$\sum = 2 \{A, C\}$$

$$g=3$$

$$k=2$$

S:ACACA

g-mers

ACA

CAC

ACA

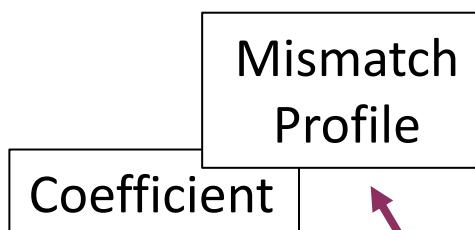
T:AAACA

g-mers

AAA

AAC

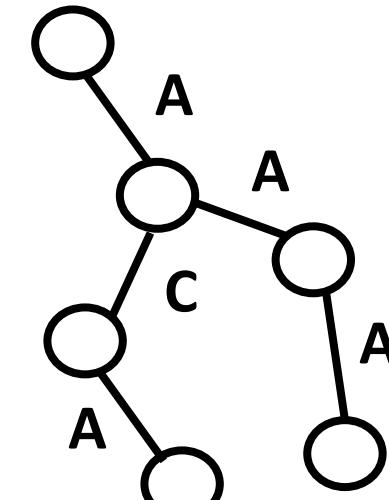
ACA



$$\text{Gaps}=g-k=1$$

gkm-SVM

Trie



$$m=0$$

$$N_m = 2$$

$$m=1$$

$$N_m = 3$$

Related Work: Gapped k-mer Kernel

$$\Sigma=2 \{A,C\}$$

$$g=3$$

S:ACACA

T:AAACA

Time taken to compute Kernel for Protein dataset > 5 HOURS

ACA
CAC
ACA

Mismatch Profile

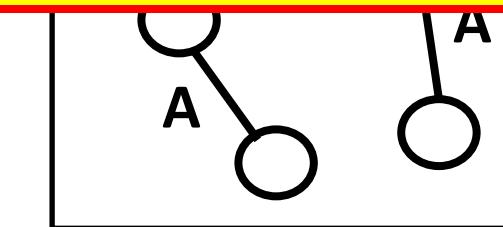
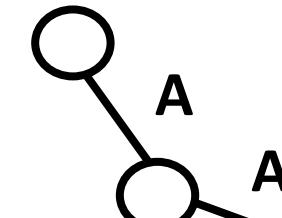
Coefficient

$$K(S, T) = \sum_{m=0}^{g-k} h_m N_m$$

$$\text{Gaps}=g-k=1$$

gkm-SVM

Trie



$$m=0$$

$$N_m = 2$$

$$m=1$$

$$N_m = 3$$

Proposed: Three Key Strategies in FastSK

1. Reduced feature space

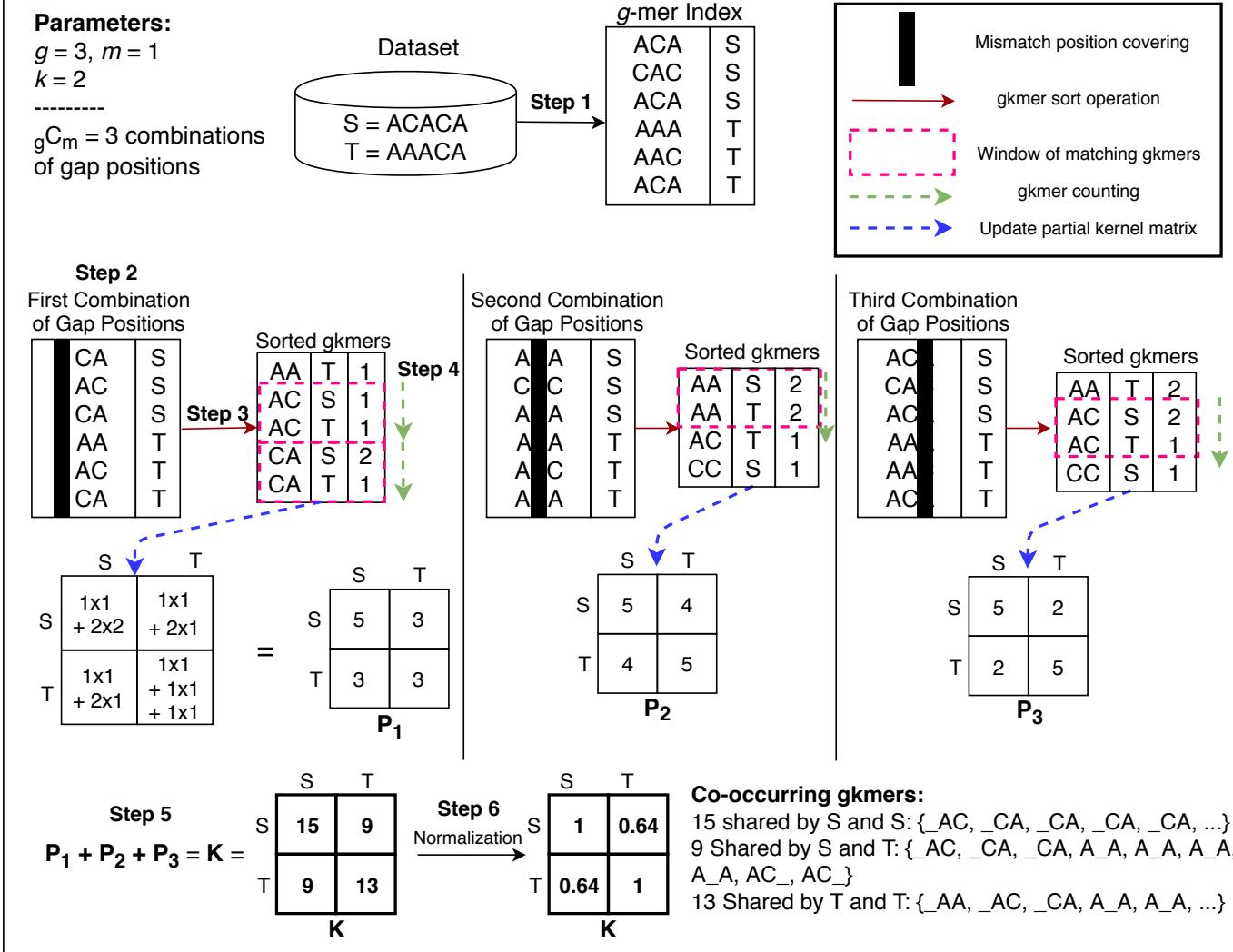
2. Direct counting algorithm

3. Approximation algorithm

$$K_{FSK} = \sum_{i=1}^{gCm} P_i$$

$$P_i(x, y) = \sum_{\gamma \in \Theta_i} c_x(\gamma) c_y(\gamma)$$

FastSK -Exact



Theorem 1. *The FastSK kernel function is equivalent to the gkmSVM kernel function [6, 7].*

Theorem 1

Proof. Because $h_{gk}(d) = 0$ for $d > m = g - k$, equation 15 is equal to

$$K_{gkm}(x, y) = \sum_{d=0}^{m=g-k} N_d(x, y) h_{gk}(d) = \sum_{d=0}^{m=g-k} \binom{g-d}{k} N_d(x, y) \quad (17)$$

To *directly* count the gkmers, we replace $\binom{g-d}{k} N_d(x, y), d \in 0, 1, 2, \dots, m$ with $\binom{g}{m}$, the number of possible mismatch positions. Therefore, at the i th mismatch position, we simply count the gapped k -mers in the set Θ_i . Therefore, we have

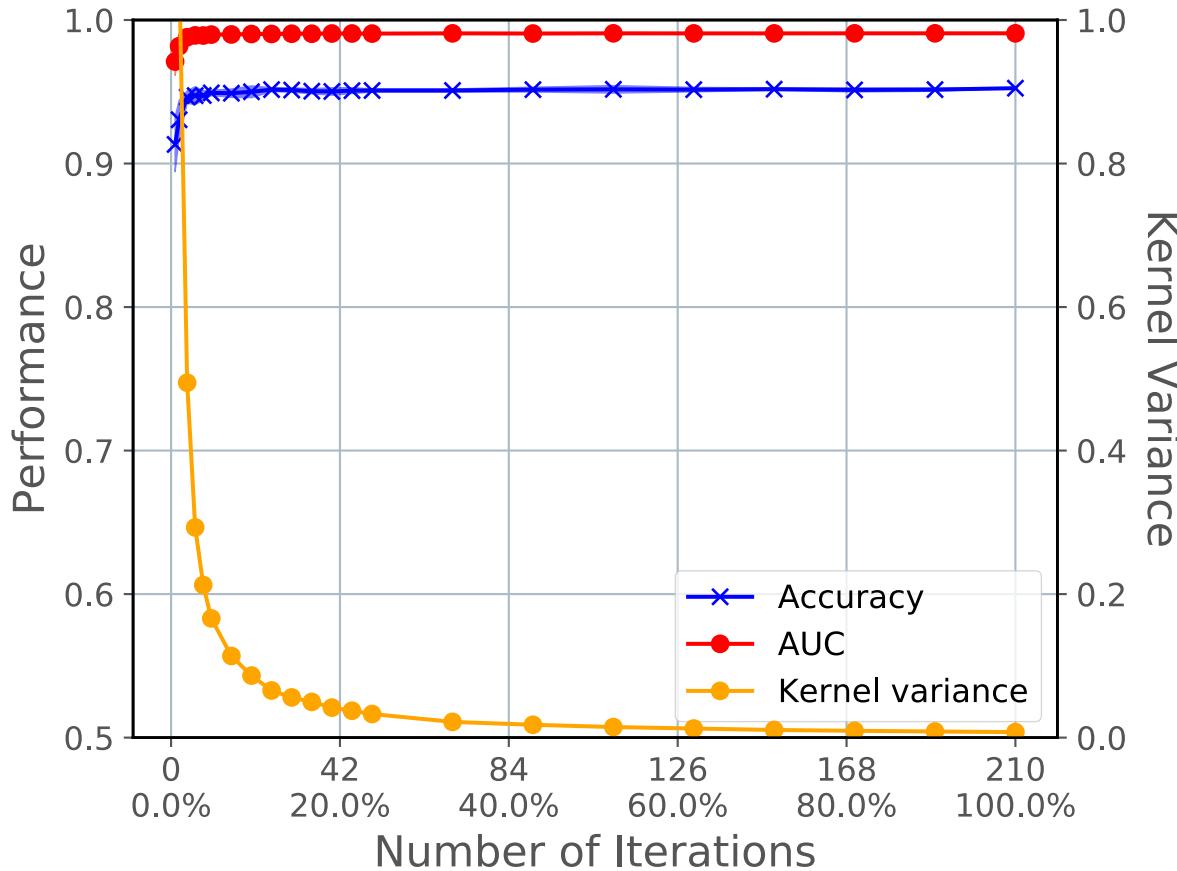
$$K_{gkm}(x, y) = \sum_{d=0}^{m=g-k} \binom{g-d}{k} N_d(x, y) = \sum_{i=1}^{gCm} \sum_{\gamma \in \Theta_i} c_x(\gamma) c_y(\gamma) \quad (18)$$

as desired. Our method is both faster and simpler because we cut out the middleman: we directly count the gkmers. It also allows us to create a better approximation algorithm than what is possible using equation 15, because we can sample mismatch positions to estimate the kernel function.

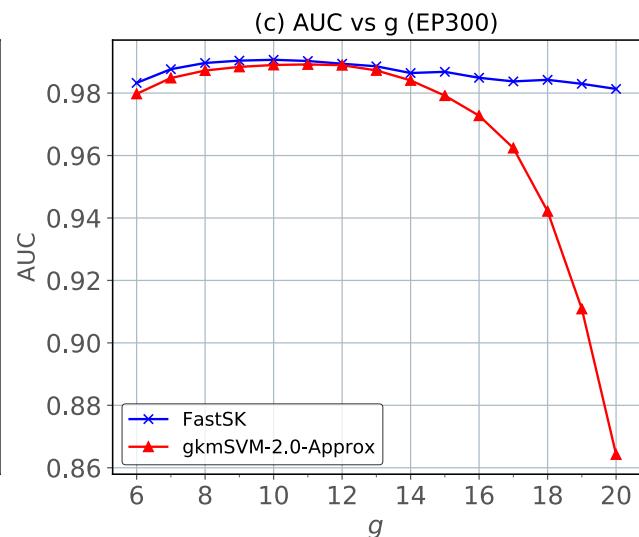
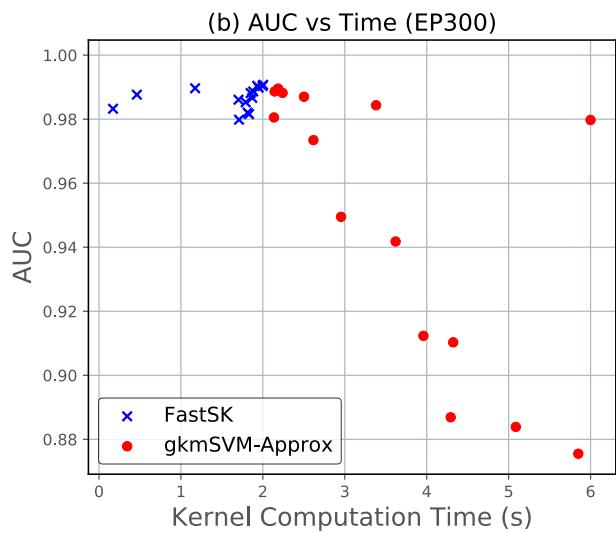
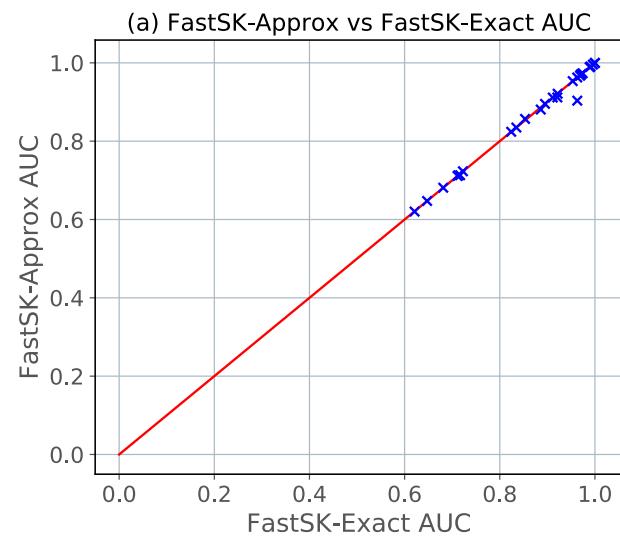
FastSK -Approx

Fig. 2: Only a small proportion of the possible mismatch combinations are needed for strong performance. We use the EP300 TFBS (DNA) dataset with $g = 10$ and $m = 4$, hence 210 possible mismatch combinations. On the x-axis, we show the number of mismatch positions used and the percentage of the total possible below each one.

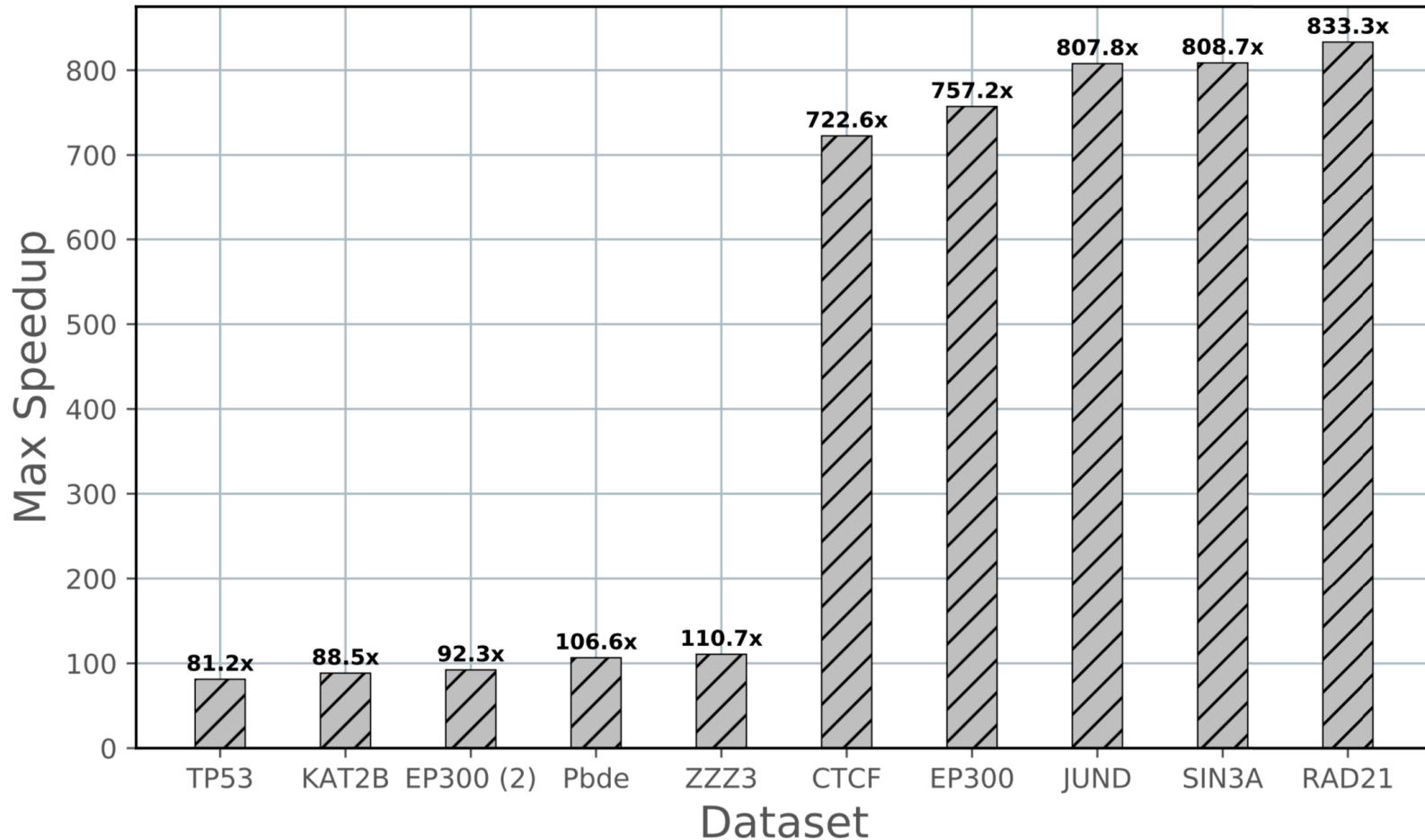
To compute FastSK -Approx, we sample possible mismatch combinations for up to $I_{max} \leq \binom{g}{m}$ iterations. That is, at iteration



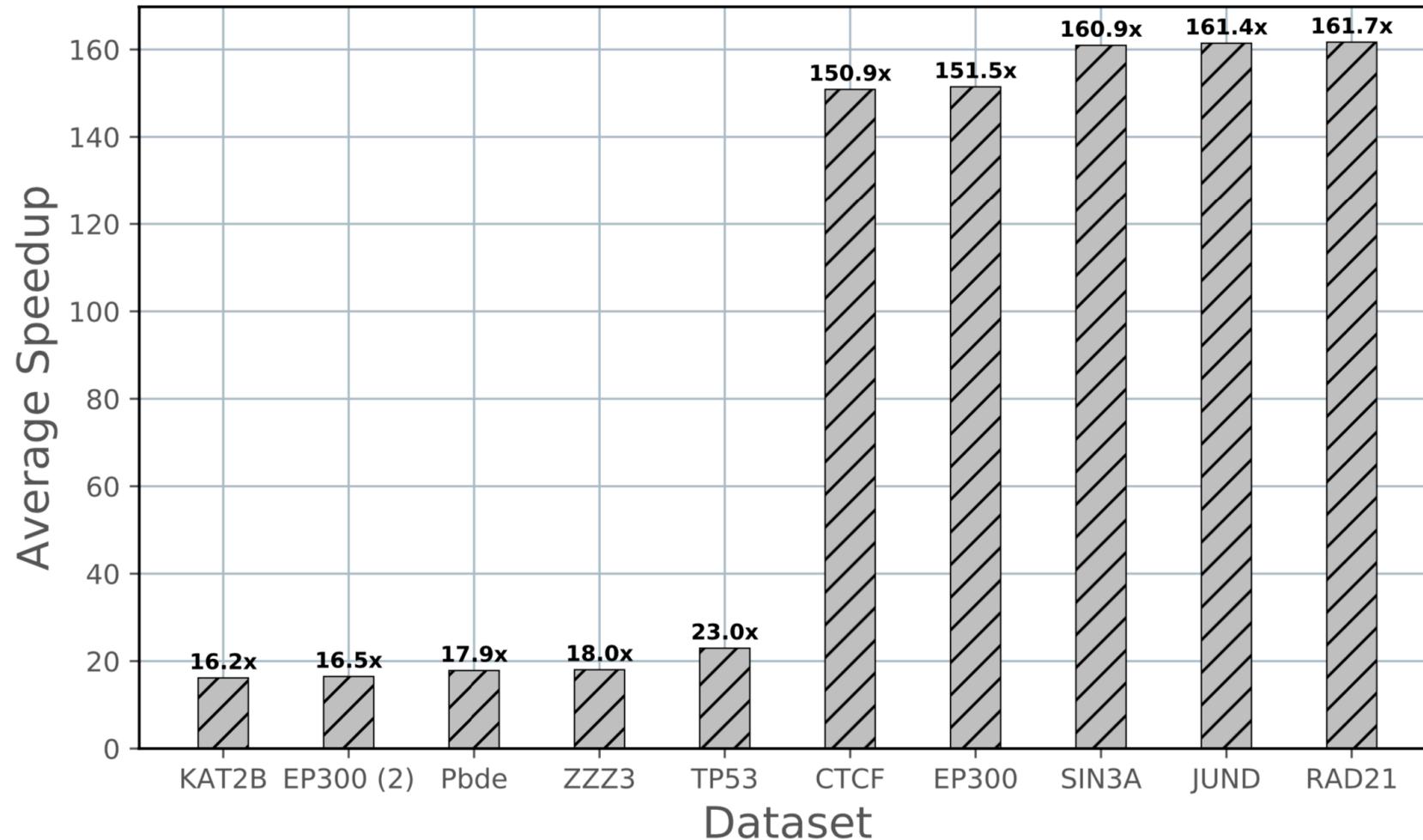
FastSK-Exact vs. FastSK-Approx vs. gkm-svm-Approx



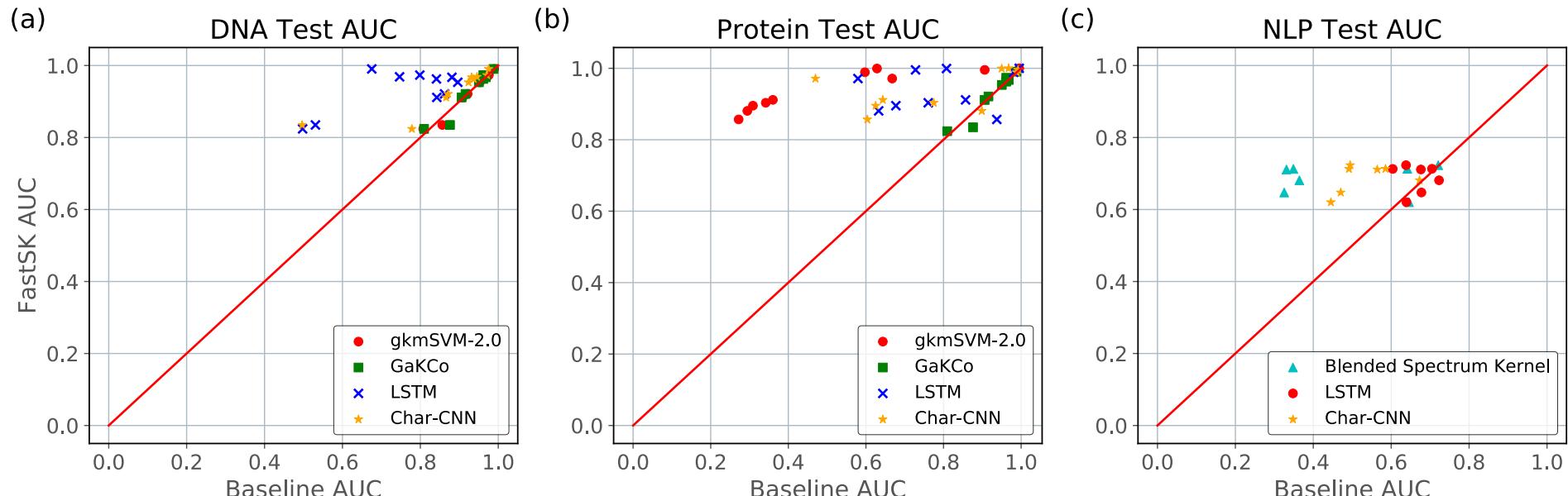
FastSK is faster!



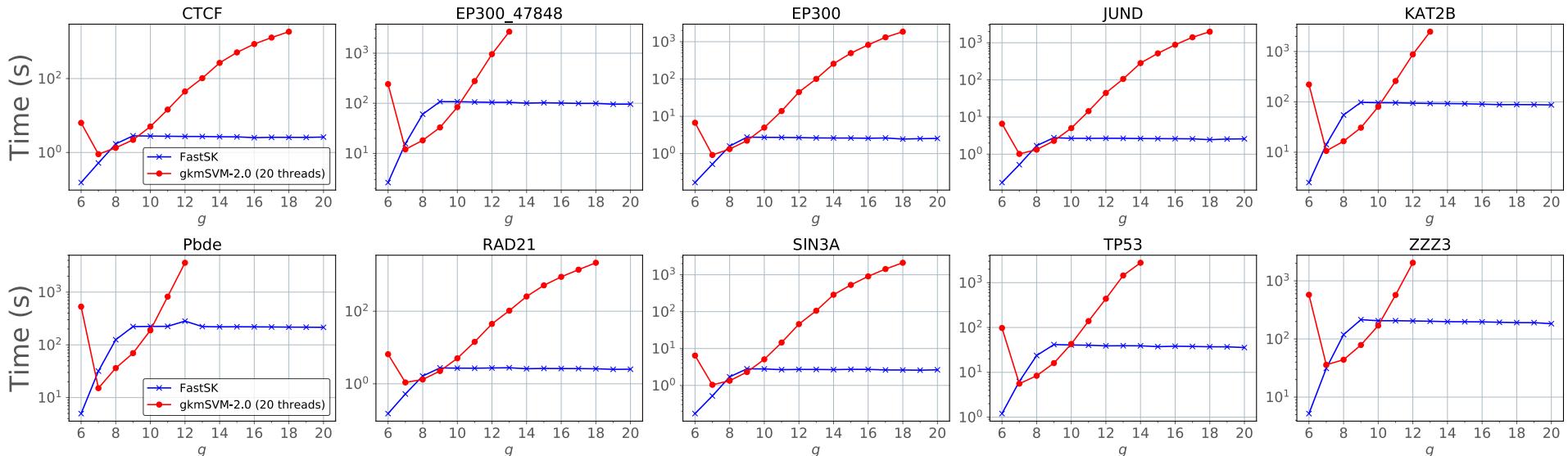
FastSK is faster!



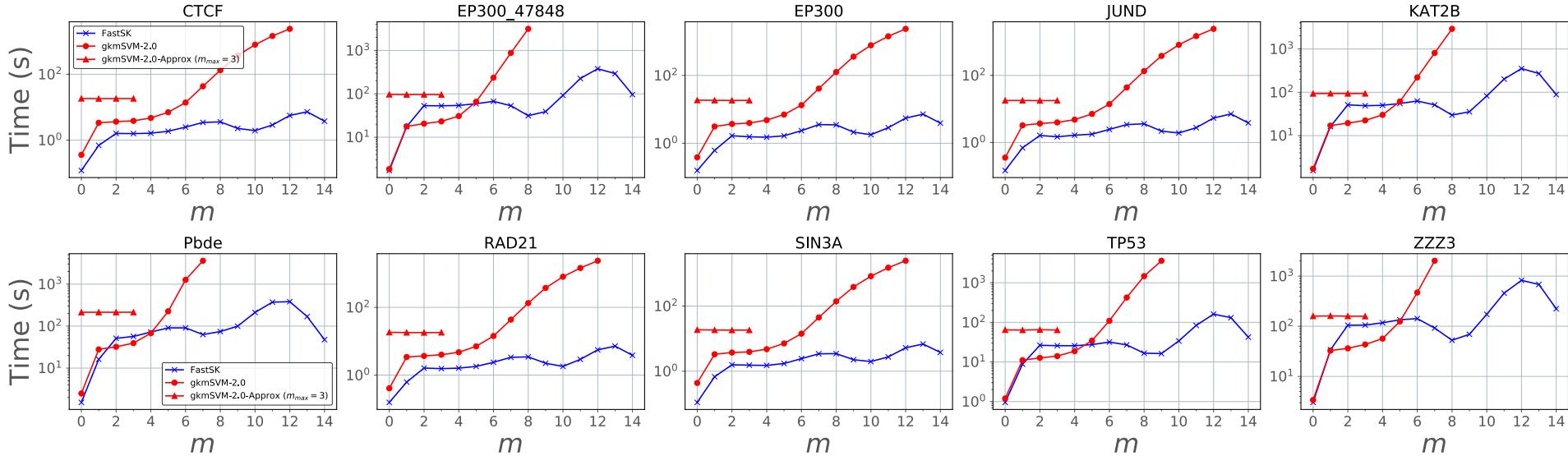
FastSK Predicts Well!



FastSK Scales well with increasing g



FastSK Scales well with increasing m



Summary:

- (A) Strong prediction performance
- (B) Scale with Σ —allow any sequence task
- (C) Scale with g —allow longer features
- (D) Scale with m —allow more mismatch versatility
- (E) Modern software architecture + usable toolbox
- Bonus property: conceptually simple kernel algorithm
- Code and datasets available at:
<https://github.com/QData/FastSK>
- bioArXiv:
<https://www.biorxiv.org/content/10.1101/2020.04.21.053975v1>



Thank you

FastSK vs. gkm-SVM

$$K_{FSK}(x, y) = \sum_{i=1}^{gCm} \sum_{\gamma \in \Theta_i} c_x(\gamma) c_y(\gamma)$$

$$K_{GSK}(x, y) = \sum_{\gamma \in \Theta_{g,m}} c_x(\gamma) c_y(\gamma)$$