TooT-T: Discrimination of transport proteins from non-transport proteins

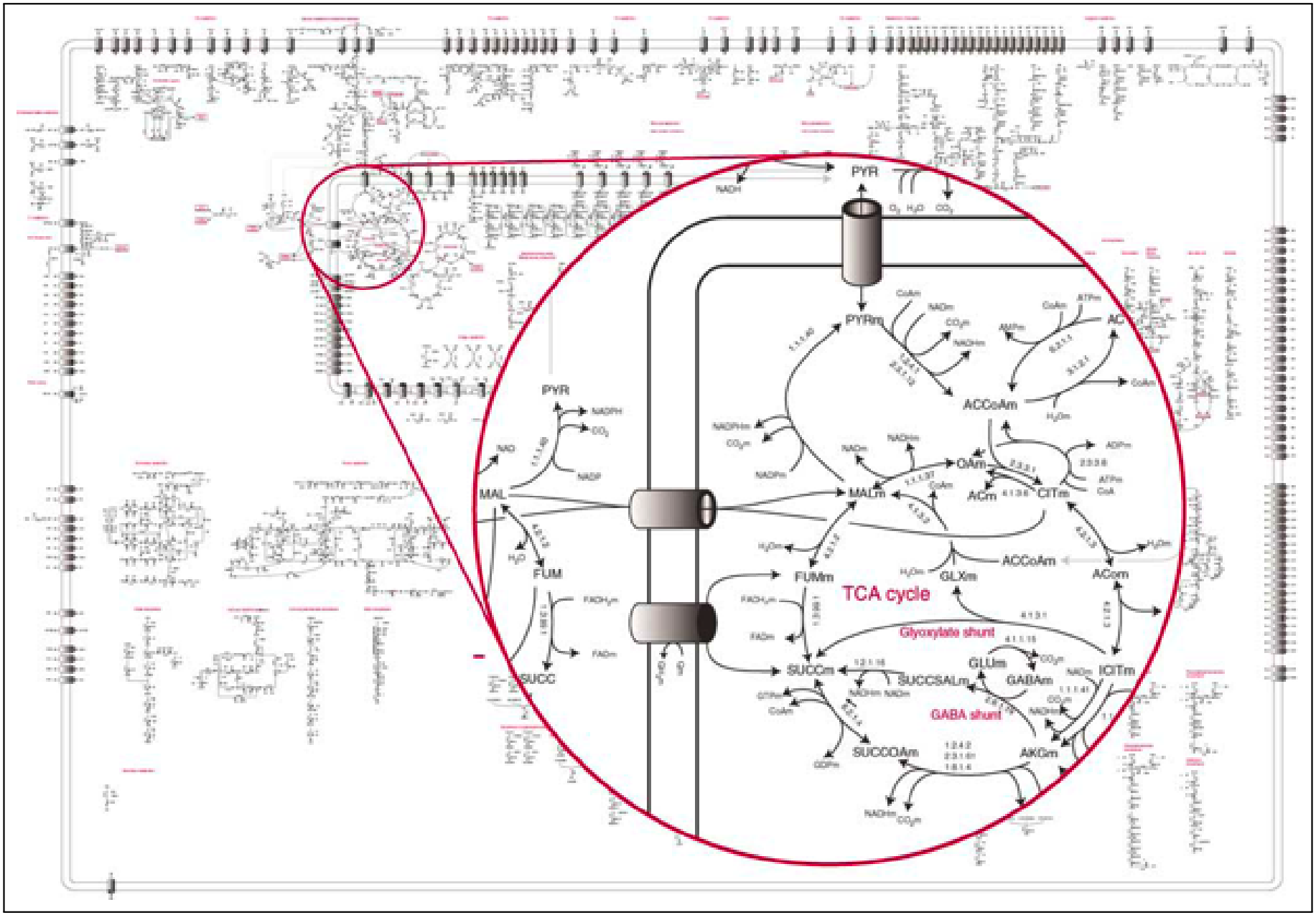
Munira Alballa and Gregory Butler

Department of Computer Science & Software Engineering
Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

December 2019 — GIW/ABACBS 2019 Sydney

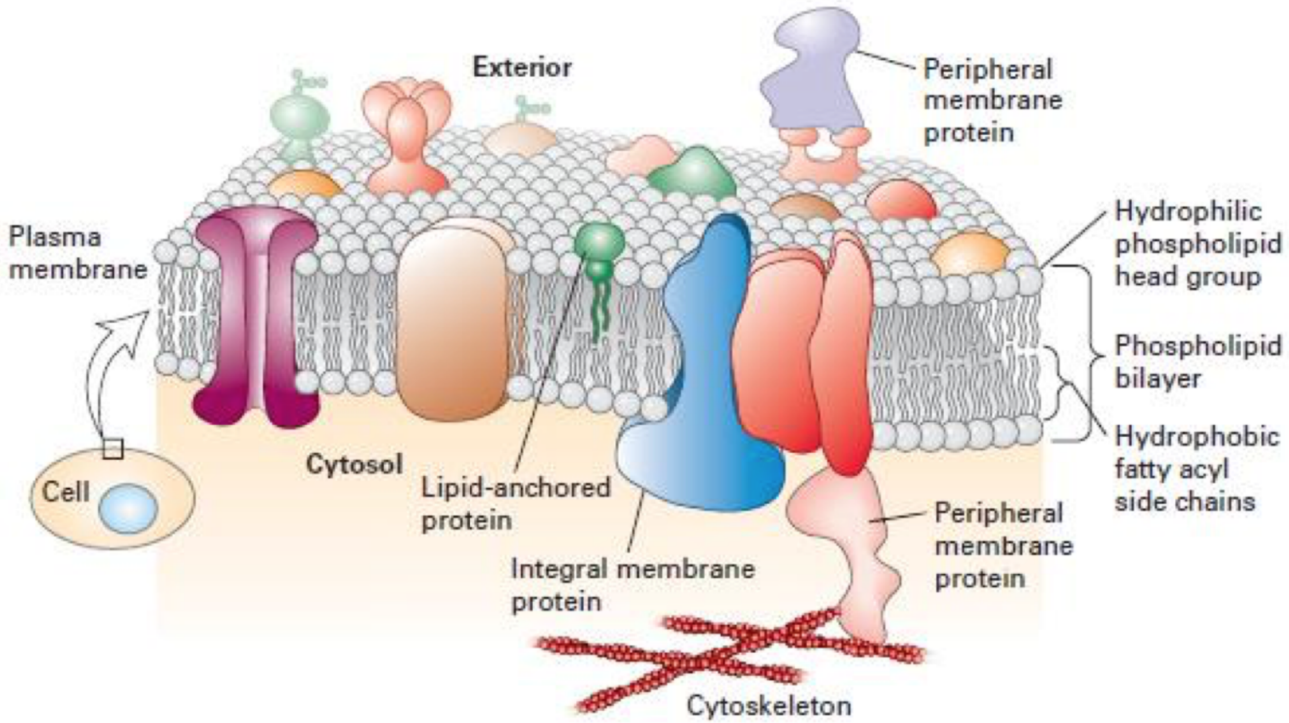
Outline

Example of Metabolism and Transport

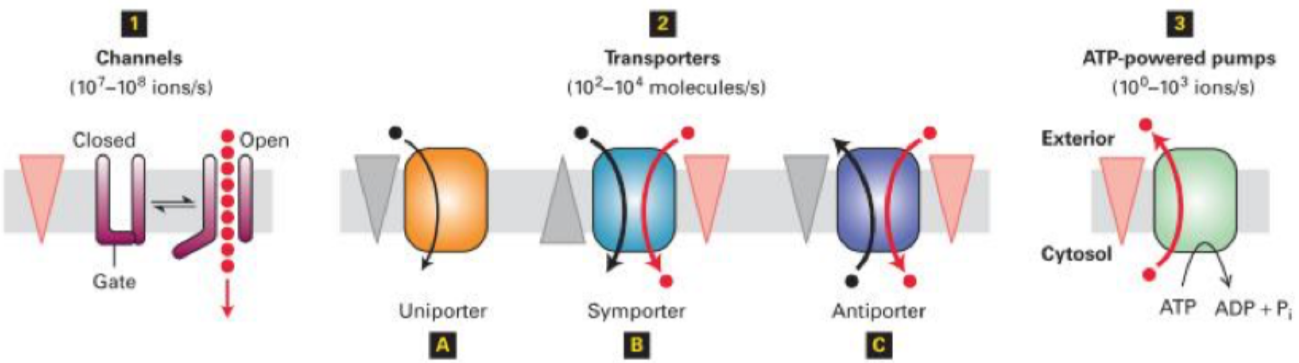


Transport Proteins

Biomembrane



Transmembrane Transport Proteins



Previous Work on Transport Prediction

TrSSP — Mishra *et al*, PLoS ONE, 2014

SVM with AAIndex, PSSM; *MCC 0.57*

SCMMTP — Liou *et al*, BMC Bioinformatics, 2015

Scoring card method using dipeptide freq. (PAAC); *MCC 0.47*

Ou'2019 — Ho *et al*, Analytical Biochem., 2019

word embeddings (from NLP); *MCC 0.73*

Li'2019 — Li *et al*, Trans. Comp. Bio & Bioinf., 2019

SVM with PSSM, PseAAC, and **GO terms**; *MCC 0.91*

The Toot Suite Project

Genome Canada BCB 2017 Competition

TooT Suite: Predication and classification of membrane transport proteins, Gregory Butler and Tristan Glatard, 2018–2021

Bioinformatics and Machine Learning

Develop predictors for transporter proteins and membrane proteins

Open Science

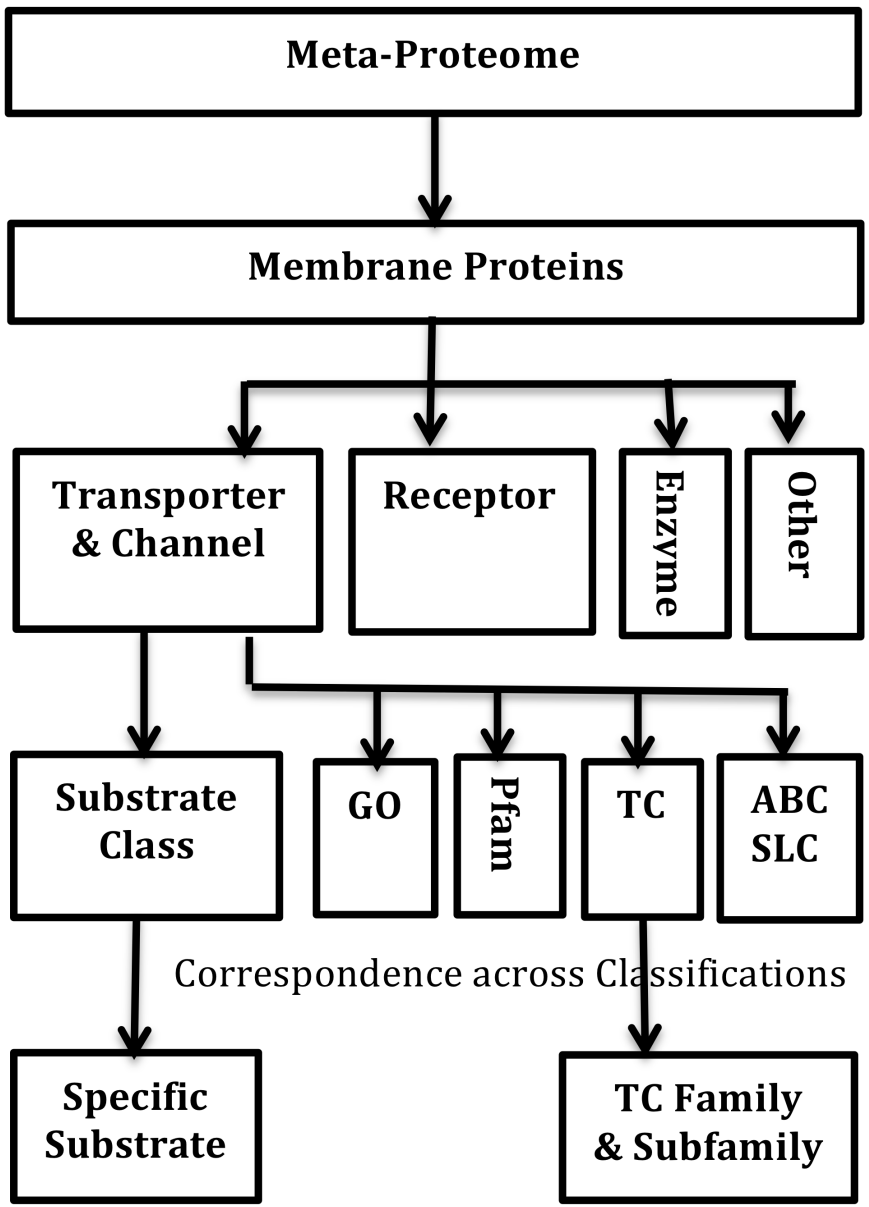
tools — open source
platform for experiments — Boutiques + bfx tools + ML tools
reproducible experiments

Scale to microbiomes

Motivation

Improve agricultural productivity
provide tools to help understand microbiome-host interaction

Toot Suite — Prediction Overview



Predictors

TooT-T — transporter?
TooT-M — membrane type

TooT-SC — substrate
TooT-TC — TC info
TooT-All — all classifications

TooT-Proteome predict classification for membrane protein in a proteome, or meta-proteome

TooT-SS specific substrate for transport protein

Experimental Platform

Experiments

EPRCS Methodology for Protein Sequence Analysis

Evolution **[E]**

Classical `blastp`, `PSI-blast`
MSA, TMS-aware MSA

Position **[P]**

Focus on important sites
classical PSSM

Region **[R]**

Split sequence into regions
eg C-terminus, Rest, N-terminus
eg TMS and non-TMS

Composition **[C]**

Classical amino acid composition
AAC, PAAC, PseAAC (Chou), split

Sequence **[S]**

HMM capture patterns along sequence

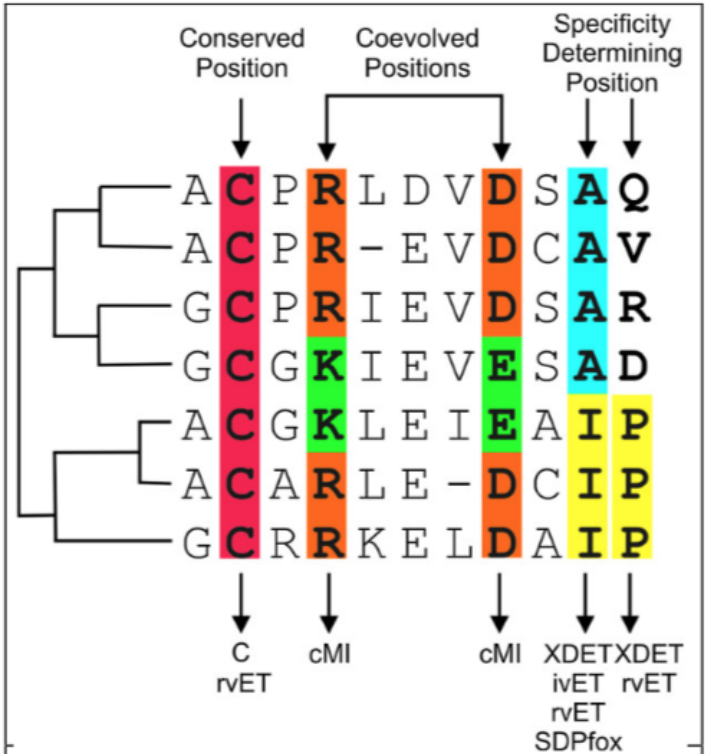


Figure 4: Types of Positional Information [Teppa'12]

TooT-T Overview

Dataset — Mishra *et al*, 2014

Class	Training dataset	Testing dataset
Transporter	780	120
Non-Transporter	600	60
Total	1380	180

Novel psi-Composition Introduced

Ensemble of Six Classifiers

- Similarity-based ($\times 3$)
- SVM and psi-composition based ($\times 3$)

Evaluation

10-fold cross validation

independent test set

Novel psi-Composition

Idea (avoid costly MSA)

1) Run PSI-BLAST against Swiss-Prot

2) Trimd lignmentsto original sequence

3) Combine with amino acid composition techniques

Comparison

Table 3 Average performance of different models

	name	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
SVM	psiPAAC*	86.73 ± 0.29	87.99 ± 0.54	87.29 ± 0.11	0.7448 ± 0.0027
	blast-PAAC	87.03 ± 0.37	86.08 ± 0.24	86.62 ± 0.22	0.7299 ± 0.0045
	psiAAC*	82.69 ± 0.21	90.64 ± 0.41	86.13 ± 0.15	0.7278 ± 0.0036
	psiPseAAC*	80.18 ± 0.58	91.51 ± 0.45	85.13 ± 0.40	0.7125 ± 0.0075
	blast-AAC	84.97 ± 0.35	84.14 ± 0.52	84.61 ± 0.22	0.6897 ± 0.0050
	PSSM	83.83 ± 0.59	82.03 ± 0.59	83.06 ± 0.21	0.6579 ± 0.0038
	blast-PseAAC	84.59 ± 0.53	78.19 ± 0.82	81.81 ± 0.35	0.6306 ± 0.0077
	PseAAC	80.45 ± 0.42	70.62 ± 0.70	76.19 ± 0.44	0.5149 ± 0.0098
	AAC	79.73 ± 0.50	70.66 ± 0.89	75.79 ± 0.51	0.5069 ± 0.0101
	PAAC	77.93 ± 0.31	72.14 ± 0.56	75.41 ± 0.31	0.5014 ± 0.0062

(PSI-BLAST 3 iterations, e-value 0.001; blast e-value 0.001)

TooT-T — Use Similarity

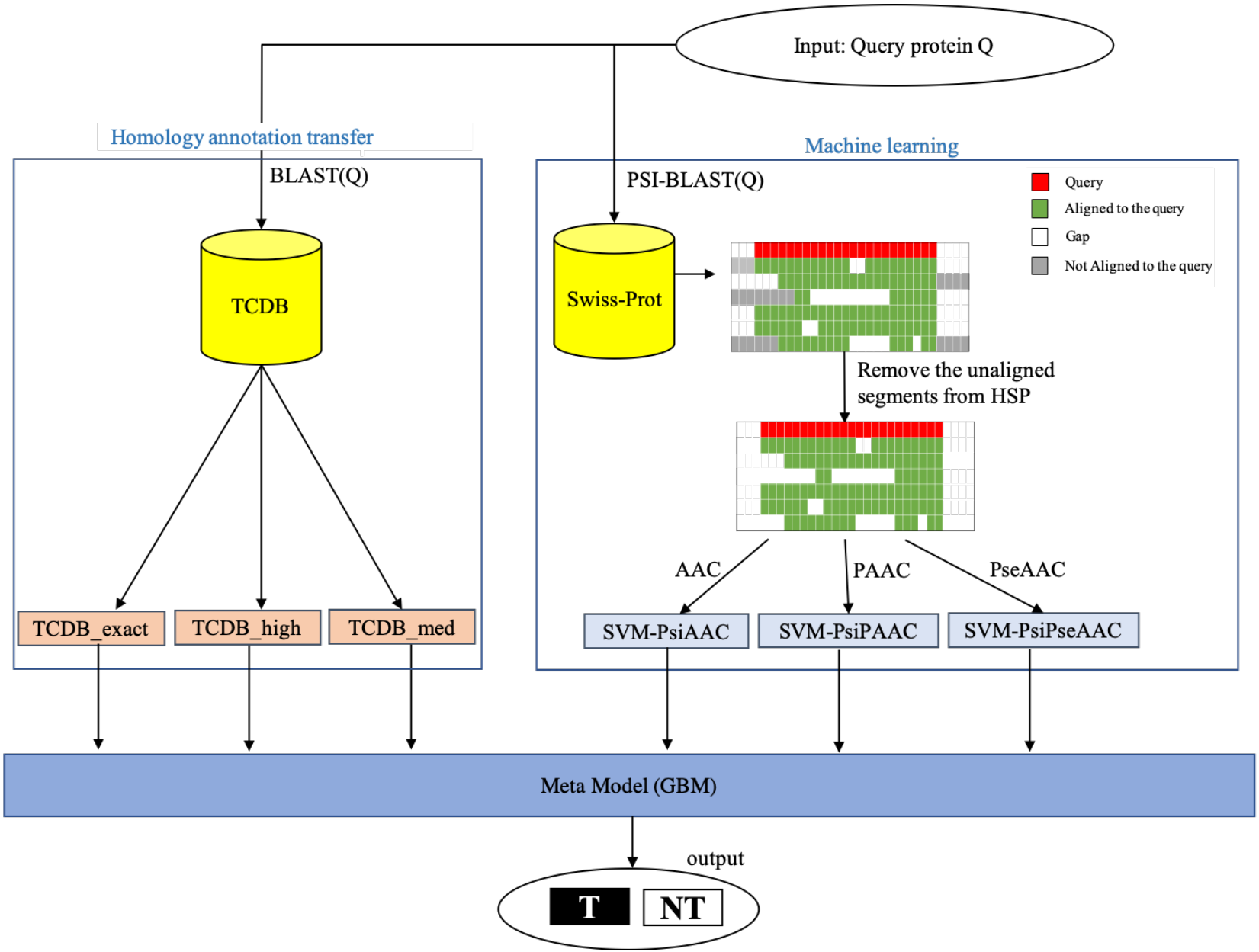
Table 2 Different Blast thresholds on TCDB

Name	BLAST Threshold	Motivation
TCDB_exact	e-value=0; percent identity 100%	exact match
TCDB_high	e-value 1e–20; percent identity 40%; query coverage 70%; subject coverage 70%; and difference in length of 10%	thresholds recommended by Butler *et al.* [3] for TCDB Blast
TCDB_med	e-value 1e–8%	threshold recommended by Barghash *et al.* [4] as an acceptable normalized BLAST threshold when dealing with a TC system

Table 5 Performance of annotation transfer by homology

	name	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
ATH	TCDB_exact	56.92	95.17	73.55	0.5440
	TCDB_high	85.90	85.50	85.72	0.7112
	TCDB_med	90.38	64.17	78.98	0.5737

TooT-T Ensemble



TooT-T Performance

Table 6 Cross-validation performance of the proposed model

	name	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
SVM	psiAAC	82.69 ± 00.21	90.64 ± 00.41	86.13 ± 00.15	0.7278 ± 0.0036
	psiPAAC	86.73 ± 00.29	87.99 ± 00.54	87.29 ± 00.11	0.7448 ± 0.0027
	psiPseAAC	80.43 ± 00.43	91.47 ± 00.46	85.23 ± 00.34	0.7142 ± 0.0069
ATH	TCDB_exact	56.92	95.17	73.55	0.5440
	TCDB_high	85.90	85.50	85.72	0.7112
	TCDB_med	90.38	64.17	78.98	0.5737
Proposed_Ensemble*		**90.15 ± 00.24**	**89.97± 00.34**	**90.07 ±00.07**	**0.7995 ±0.001**

Table 7 Independent testing performance of the proposed model

	name	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
SVM	psiAAC	83.33	95.00	87.22	0.75
	psiPAAC	89.17	88.33	88.89	0.76
	psiPseAAC	80.00	96.67	85.56	0.73
ATH	TCDB_exact	56.67	91.67	68.33	0.46
	TCDB_high	86.67	80.00	84.44	0.66
	TCDB_med	92.5	58.33	81.11	0.56
Proposed_Ensemble*		**94.17**	**88.33**	**92.22**	**0.82**

TooT-T — Comparison with Previous Work

Table 9 Comparison with other published work

Tool	Sensitivity(%)		Specificity (%)		Accuracy (%)		MCC	
	Ind.	CV	Ind.	CV	Ind.	CV	Ind.	CV
SCMMTP [7]	80.00	83.76	68.33	77.68	76.11	81.12	0.47	0.62
TrSSP [6]	76.67	76.67	81.67	78.46	80.00	78.99	0.57	0.58
Ou et al. [9]	100.00	83.14	77.50	84.48	85.00	83.94	0.73	0.68
Proposed model	**94.17**	**90.15**	**88.33**	**89.97**	**92.22**	**90.07**	**0.82**	**0.80**
Li et al. [8]	96.67	99.50	95.83	97.44	96.11	98.33	0.91	0.97

Note

Li *et al*, 2019 use **GO terms** as features

when building classifier

and for protein sequences being classified

These are available for dataset from Swiss-Prot, but not in general!

Conclusion

TooT-T is the State-of-the-Art

TooT-T outperforms all methods relying only on protein sequence!

On independent test set, TooT-T achieves

- accuracy of 92.22%
- MCC of 0.82

Thank You!

Questions, Please?