Illuminating the twilight zone of protein function prediction and deep learning model assessment

Wong Limsoon

This is work of my student, Neamul Kabir

Protein function assignment

A protein is a large complex molecule made up of one or more chains of amino acids

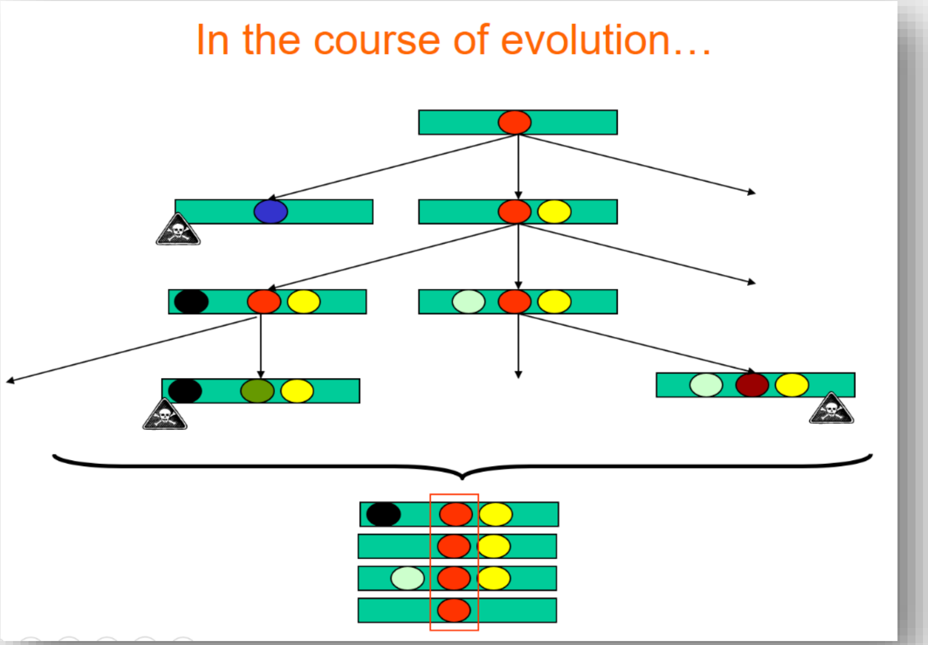
Usually, only the sequence of amino acid is known

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

Proteins perform a wide variety of activities in the cell

How do we predict the function of a protein?

A standard postulate based on evolution



Two proteins (not) inheriting their function from a common ancestor (do not) have similar amino acid sequences

Guilt by association

Compare *T* with seqs of known function in a db

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                            60        70        80        90        100
Amicyanin          MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                              :..:  . ::. ::
Ascorbate Oxidase  ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                       70        80        90        100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|   unknown protein [Mesorhizobium loti]
         Length = 105

Score =  105 bits (262), Expect = 1e-22
Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
            MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

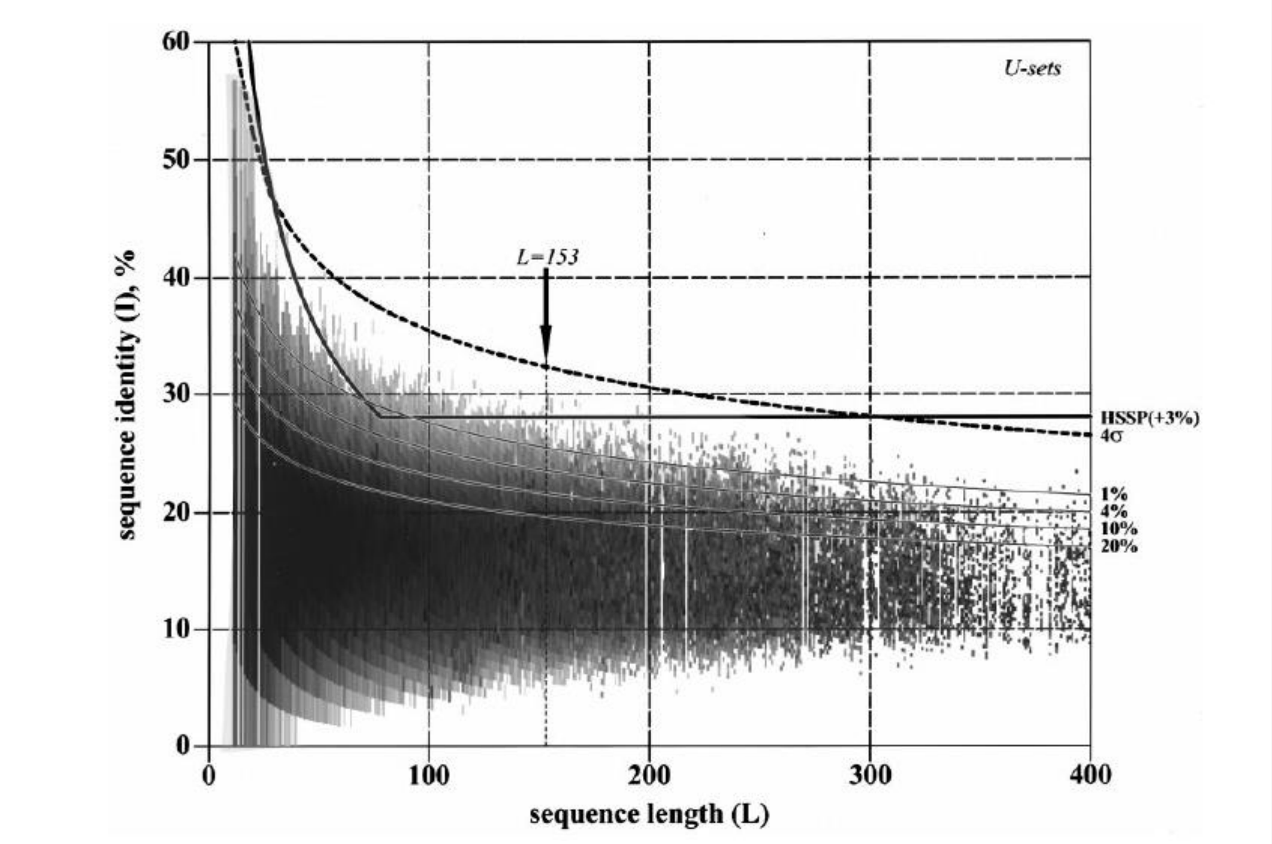
good match between Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Twilight zone: Limit of sequence similarity-based protein function assignment

So, need clever methods for the twilight zone



Similarity to ref proteins high ⇒ easy

Similarity is low (~30%) ⇒ error prone

Similarity is very low ⇒ really hard

Abagyan RA, Batalov S. *J Mol Biol.*, 273(1):355-68, 1997

Many deep learning models to the rescue?

DeepFam (2018, CNN)

DeepGO (2018, CNN + DNN)

DeepPred (2019, hierarchical DNN)

DeepGoPlus (2019, CNN + DNN)

UDSMProt (2020, LSTM)

MultiPredGO (2020, multimodal DL)

TALE+ (2021, transformer)

DeepGraphGO (2021, CNN + DNN, multimodal)

DeepGoZero (2022, zero-shot learning), …

DeepFam, deep learning for protein family prediction

This looks good

Really?

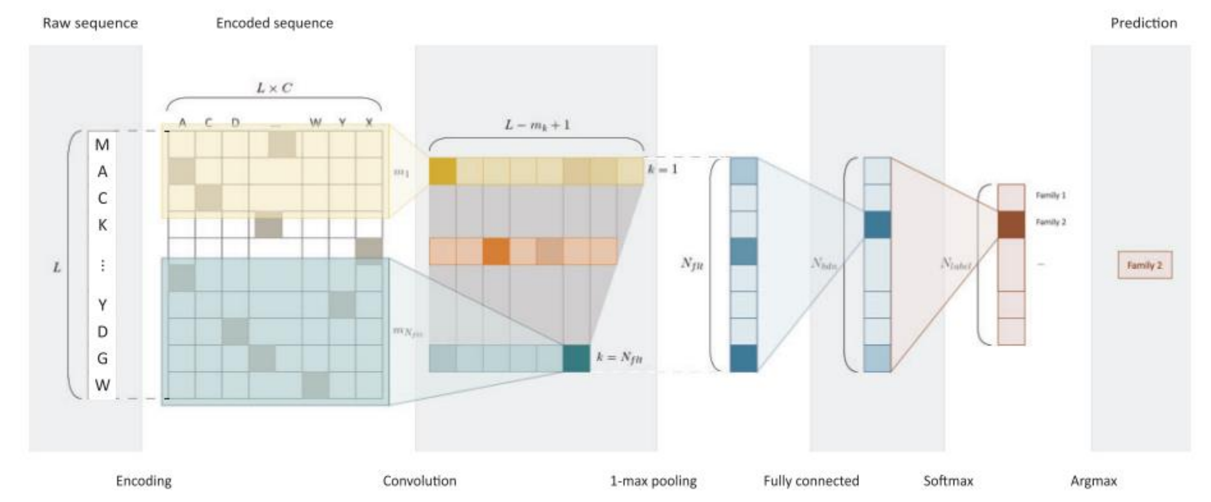


Fig. 1. The overview of DeepFam model. It is a feedforward convolution neural network whose last layer represents the probabilities of each family. convolution layer and 1-max pooling layer calculate a score (activation) of the existence of a conserved regions. The next layer is fully-connected neural network which can detect longer or complex sites. In order to infer the probability of each family, the last layer is designed as softmax layer (multinomial logistic regression), generally used for multi-class classification

Table 2. Prediction accuracy (%) comparison of COG dataset

Dataset	COG-500-1074	COG-250-1796	COG-100-2892
DeepFam	**95.40**	**94.08**	91.40
pHMM	91.75	91.78	**91.67**
3-mer LR	85.59	81.15	75.44
Protvec LR	47.34	41.76	37.05

Bold indicates the best performance for each dataset.

Seo et al., *Bioinformatics*, 34(13):i254-i262, 2018

Illuminating the twilight zone of deep-learning model assessment

Image credit, https://www.earth.com/earthpedia-articles/secrets-of-the-oceans-mysterious-mesopelagic-zone/

How do you set final exam of your course?

This is how I set exams

Some easy questions

Enough hard questions

And often some surprising questions

Students don't get "A" answering easy questions

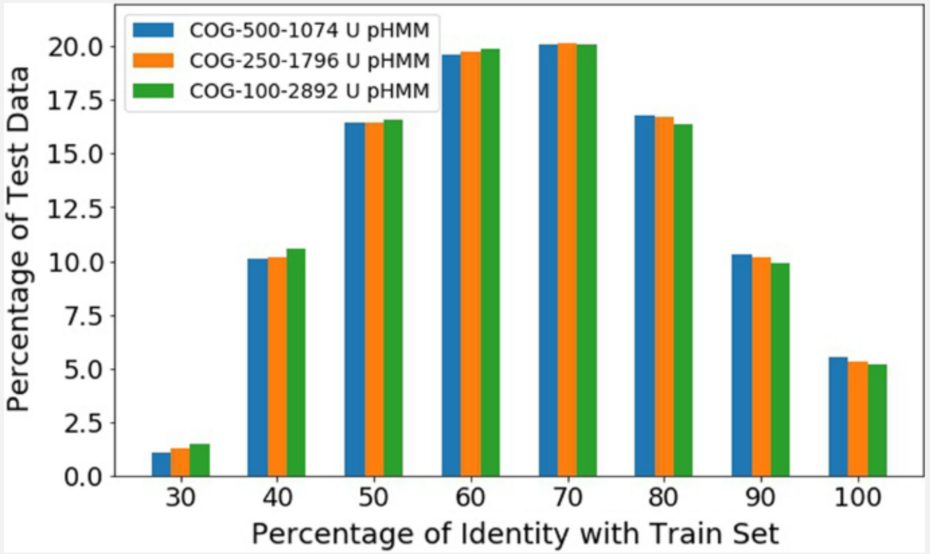
Table 2. Prediction accuracy (%) comparison of COG dataset

Dataset	COG-500-1074	COG-250-1796	COG-100-2892
DeepFam	**95.40**	**94.08**	91.40
pHMM	91.75	91.78	**91.67**
3-mer LR	85.59	81.15	75.44
Protvec LR	47.34	41.76	37.05

Bold indicates the best performance for each dataset.

Is this a good assessment of DeepFam? Do the test sets have enough hard questions and surprising questions?

DeepFam's good accuracy is largely due to "easy" proteins



Dataset	Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
Identity: $0 < x \leq 30$							
COG-500-1074	EnsembleFam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**
	pHMM	69.54	73.75	55.51	70.62	70.85	73.55
	DeepFam	57.14	54.52	49.90	46.92	43.64	35.94
COG-250-1796	EnsembleFam	72.84	**77.07**	**81.02**	**82.14**	**84.66**	**86.45**
	pHMM	**75.39**	73.82	73.84	71.02	67.44	72.43
	DeepFam	32.44	32.54	30.24	29.53	30.02	28.68
COG-100-2892	EnsembleFam	**75.24**	**79.55**	**81.21**	**80.63**	**82.05**	**88.95**
	pHMM	63.44	59.69	53.45	48.16	47.42	57.57
	DeepFam	27.30	26.13	25.54	27.62	24.83	25.36

Neamul Kabir & Wong, *BMC Bioinformatics,* 23:90, 2022

If there are few twilight zone proteins in real life, maybe DeepFam's poor twilight zone performance is ok?

The reference database comprises proteins with known function

If no function is predicted for a protein, or a wrong function is predicted, there won't be any validated result for the protein

∴ Few twilight zone proteins can get into the reference database

A self-fulfilling prophecy!

How often do you encounter twilight-zone proteins

Here is a typical distribution of protein similarity of a new (fungal) genome to large reference protein databases

Identity region	Percentage of proteins from genome
Zero identity	54.29%
$0 < \text{identity} \leq 30$	35.23%
$30 < \text{identity} \leq 40$	3.81%
$\text{identity} > 40$	6.67%

New genomes are dominated by twilight-zone proteins

Don’t be fooled by high accuracy on test sets with too many easy examples

Need to stratify wrt easy and hard test instances

Does the test sets contain surprising questions?

How do DeepFam perform on these?

The test sets don’t have surprises (proteins from novel classes)

If you give DeepFam a protein from a novel class it was not trained on, it will always (wrongly) assign it to one of the classes it was trained on

There are thousands of function classes DeepFam cannot be trained on due to too few samples…

Don’t be fooled by high accuracy on test sets without surprises

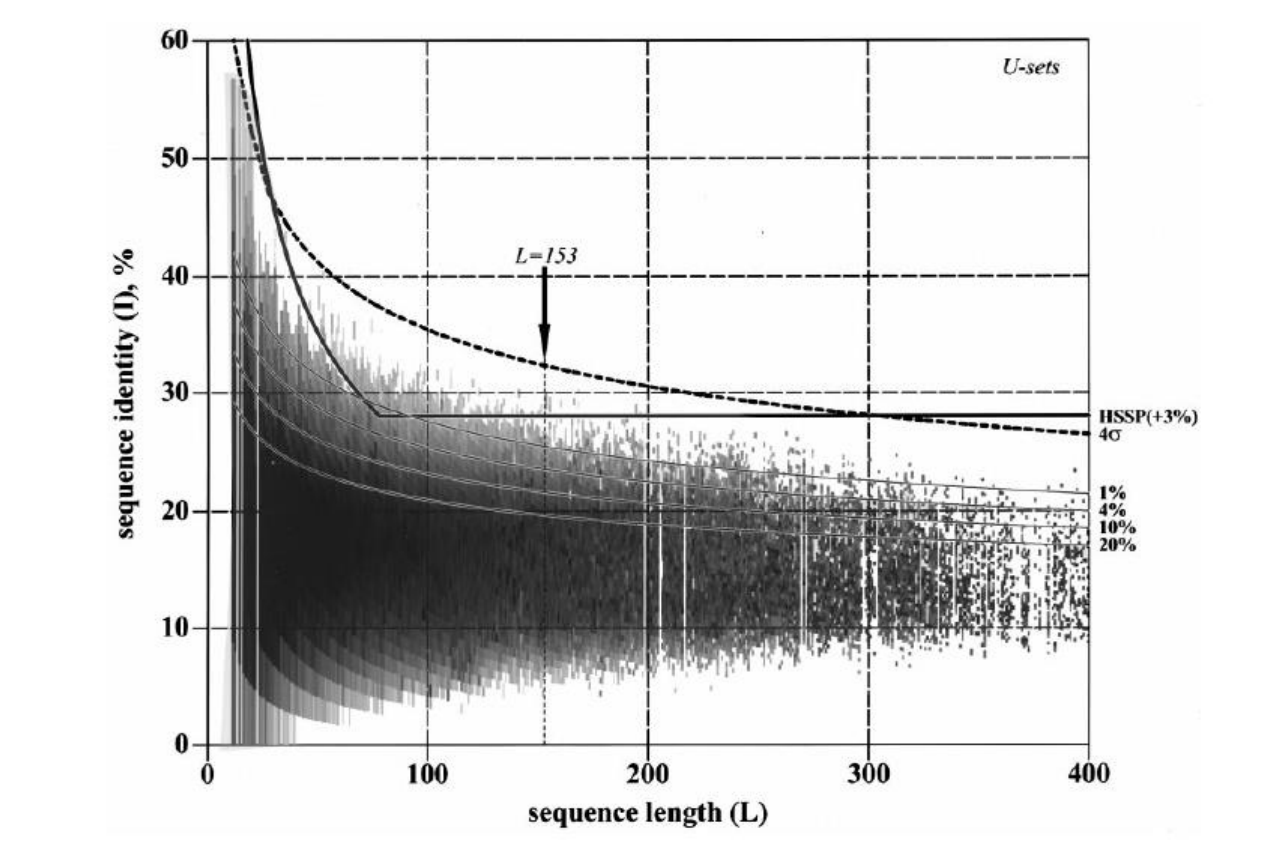
Real world is full of them

Illuminating the twilight zone of protein function prediction

Image credit, https://phys.org/news/2014-03-ninety-five-cent-world-fish-mesopelagic.html

Inferring protein function from low-similarity reference proteins is harmful

Really?



Similarity to ref proteins high ⇒ easy

Similarity is low (~30%) ⇒ error prone

Similarity is very low ⇒ really hard

Abagyan RA, Batalov S. *J Mol Biol.*, 273(1):355-68, 1997

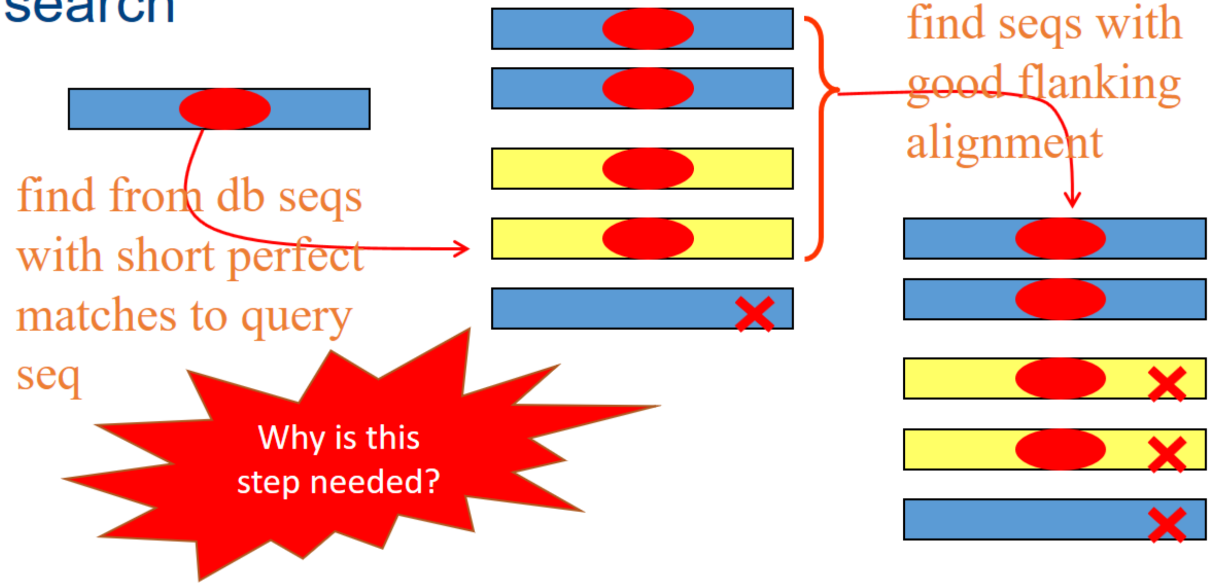
Many homology search tools are optimized to skip comparing & retrieving low-similarity proteins

A twilight-zone protein is low similarity ⇒ get nothing back!

BLAST: How it works

Altschul et al., *JMB*, 215:403-410, 1990

BLAST is one of the most popular tool for doing fast "guilt-by-association" sequence homology search



Inspiration

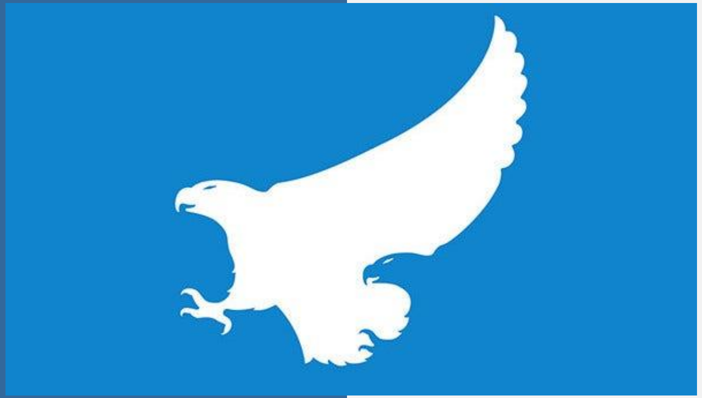


Image credit: ArtStation, Peter Hermes Furian, & TEMPLATE.NET

Similarities of dissimilarities

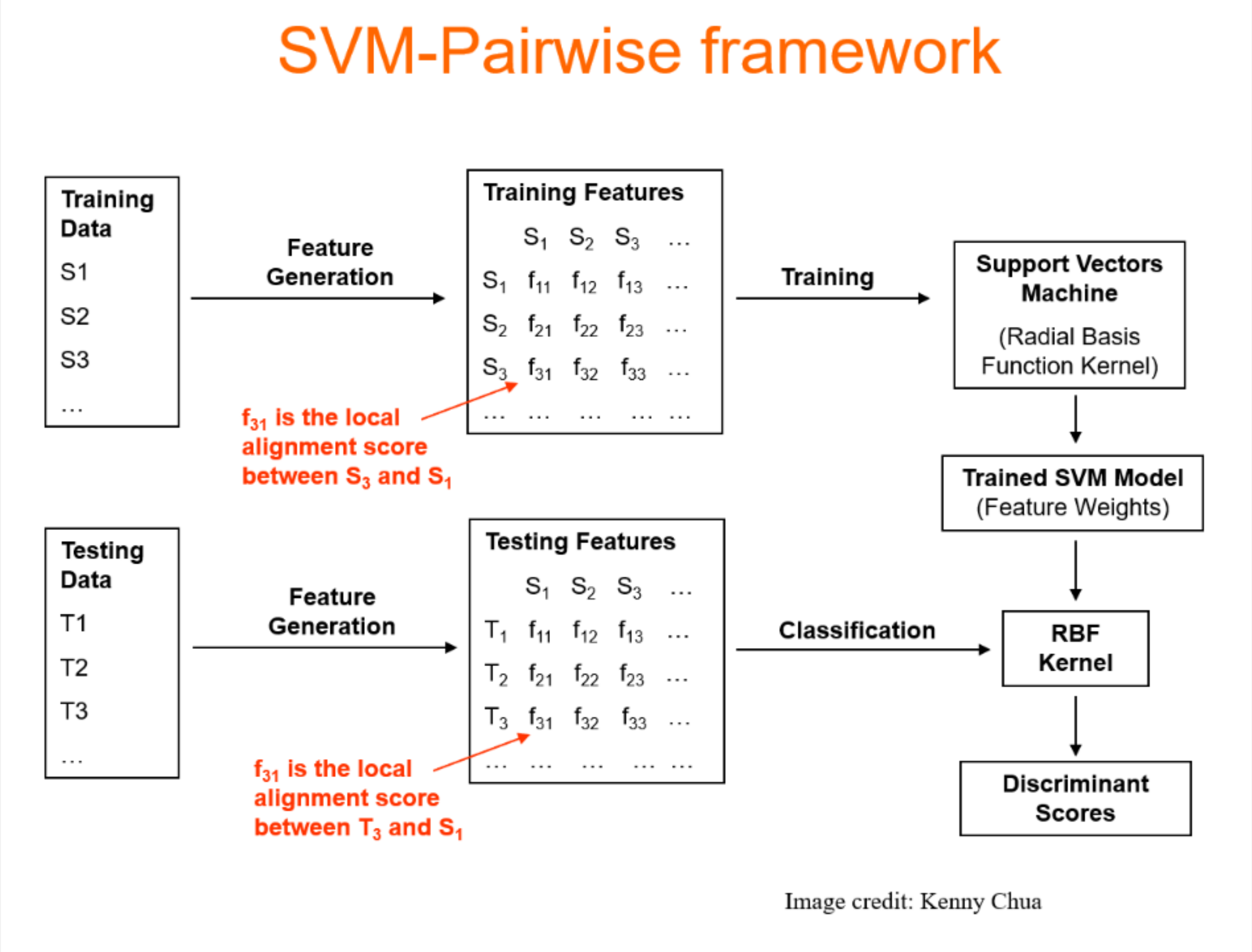
The diff betw any apple to orange / banana / mango / etc. are mostly same as the diff betw any other apple with that orange / banana / mango / etc.

The diff betw a mysterious fruit X to orange / banana / mango / etc. are mostly same as the diff betw an apple to orange / banana /mango / etc.

$\Rightarrow$The fruit X is likely an apple

EnsembleFam uses low-/dis-similarity information discarded by other methods!

Inspired by SVM-pairwise



Neamul Kabir & Wong, *BMC Bioinformatics,* 23:90, 2022

Design of EnsembleFam

One ensemble per protein family

Each ensemble has 3 base SVMs

Base SVMs use diff combinations of similarity & dissimilarity features

Ensemble decides (in vs not-in target family) by a majority vote

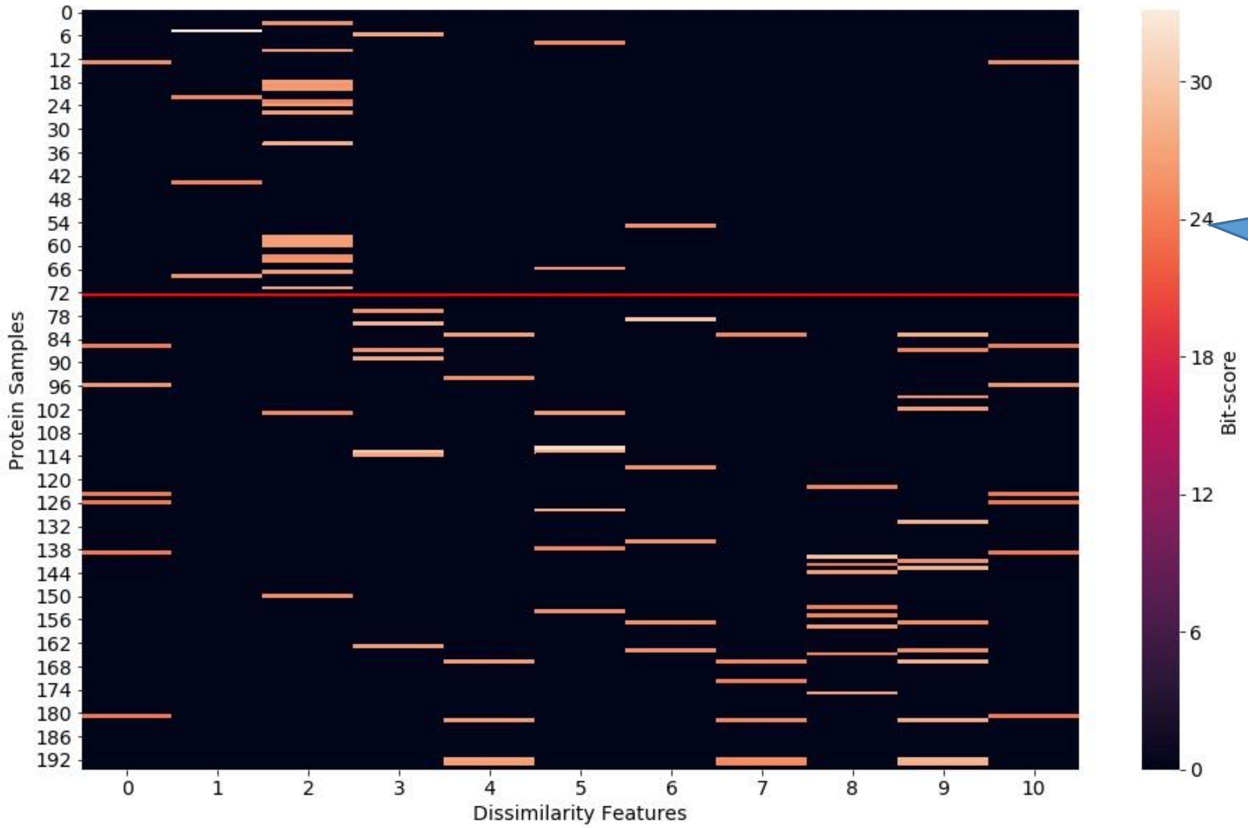
If no ensemble predicts "yes", the protein is from a novel class

Feature group #1 (Dissimilarities)
Best BLAST scores from each non-target class (only 10 ref proteins used per class)

Feature group #2 (Dissimilarities)
pHMM scores from each pFam family

Feature #3 (Similarity)
Best BLAST score from target class

HeatMap of dissimilarity features
Extracted from EnsembleFam



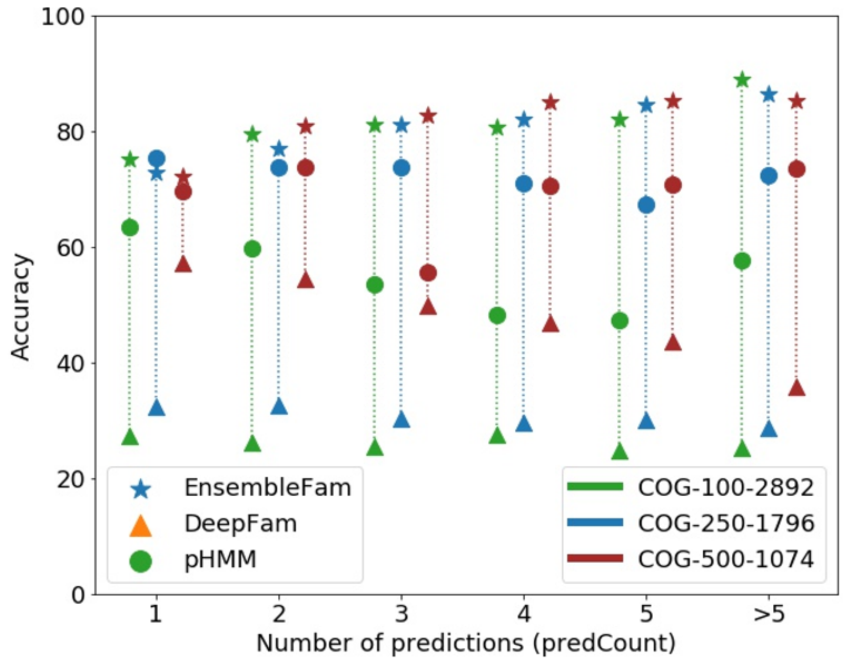
Note: Bitscore of proteins in the same family is typically ~200

There is consistency in the way two proteins of the same family differ from the other families

EnsembleFam performance on the whole COG test set

Dataset	Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
COG-500-1074	EnsembleFam	**98.14**	**98.59**	**98.73**	**98.87**	**98.87**	**99.13**
	pHMM	96.46	97.22	97.06	96.63	95.74	95.44
	DeepFam	84.88	82.49	80.86	79.46	78.18	77.25
COG-250-1796	EnsembleFam	**97.74**	**98.41**	**98.54**	**98.81**	**98.80**	**99.17**
	pHMM	96.44	97.08	96.89	96.13	94.88	95.04
	DeepFam	72.29	71.60	71.22	71.28	71.20	70.48
COG-100-2892	EnsembleFam	**98.01**	**98.44**	**98.71**	**98.82**	**98.96**	**99.37**
	pHMM	96.59	96.75	95.83	94.36	95.70	95.84
	DeepFam	61.51	62.59	64.87	67.41	68.12	67.89

EnsembleFam performance in the twilight zone



0 < identity <=30

0 < identity <= 30							
Dataset	**Method**	**Pred Count=1**	**Pred Count=2**	**Pred Count=3**	**Pred Count=4**	**Pred Count=5**	**Pred Count > 5**
COG-500-1074	Ensemble Fam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**
	pHMM	69.54	73.75	55.51	70.62	70.85	73.55
	DeepFam	57.14	54.52	49.90	46.92	43.64	35.94
COG-250-1796	Ensemble Fam	72.84	**77.07**	**81.02**	**82.14**	**84.66**	**86.45**
	pHMM	**75.39**	73.82	73.84	71.02	67.44	72.43
	DeepFam	32.44	32.54	30.24	29.53	30.02	28.68
COG-100-2892	Ensemble Fam	**75.24**	**79.55**	**81.21**	**80.63**	**82.05**	**88.95**
	pHMM	63.44	59.69	53.45	48.16	47.42	57.57
	DeepFam	27.30	26.13	25.54	27.62	24.83	25.36

Contribution of dissimilarities

Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
Identity: $0 \leq x \leq 30$						
SVM Model 1	59.82	66.96	67.60	67.96	67.32	74.67
SVM Model 2	57.11	65.09	65.01	65.86	64.55	71.80
SVM Model 3	57.34	65.34	64.29	65.02	63.36	70.13
EnsembleFam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**

Bases SVM have similar performance, Not much better than e.g. DeepFam

Where does performance increment of the ensemble come from?

Intersection of the two **dissimilarity-based SVMs** is ~50% of the extra predictions!

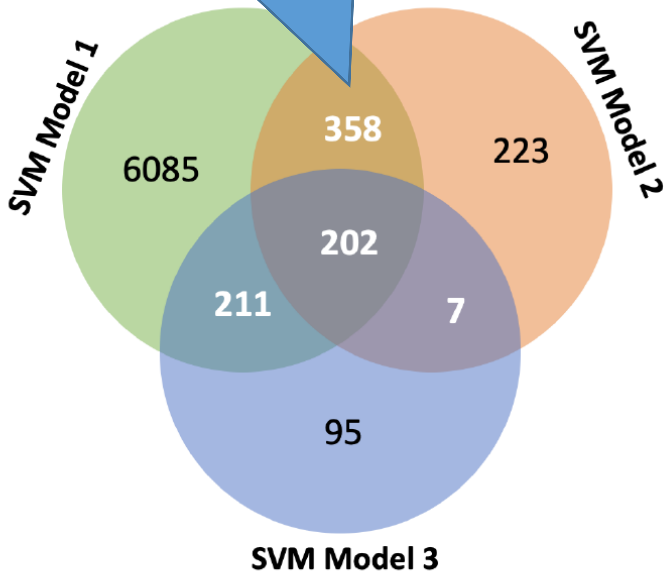


Figure 3.7: Prediction overlap of the three base classifier on the twilight zone proteins in $0 \leq$ identity ≤ 30 region. The Venn diagram is drawn based on the prediction made on twilight zone proteins of the testset of COG-500-1074 dataset. The number of predictions made by each base classifier is indicated in the figure. Numbers highlighted in white indicate overlap between at least two methods, hence predicted by EnsembleFam.

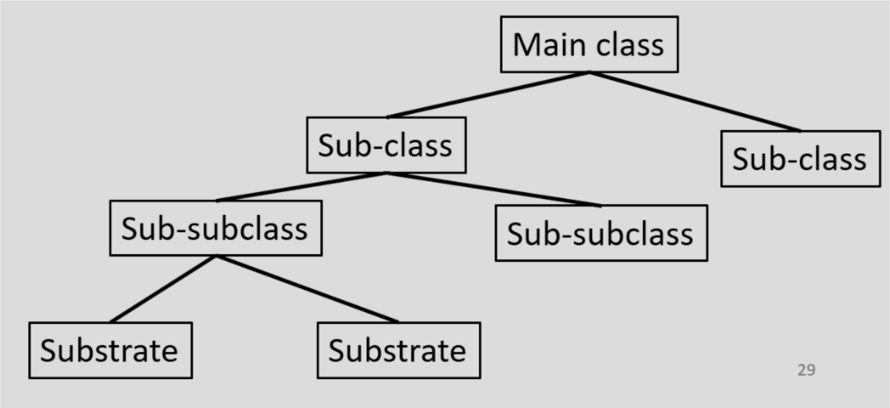
Enzyme Commission (EC) # prediction

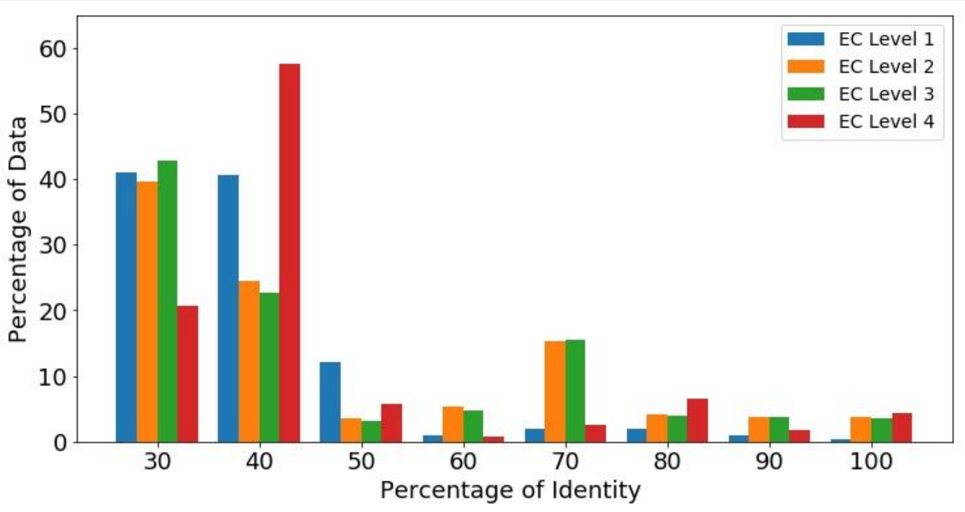
Enzymes are more heterogeneous due to hierarchical nature

This heterogeneity adds more difficulty in annotation

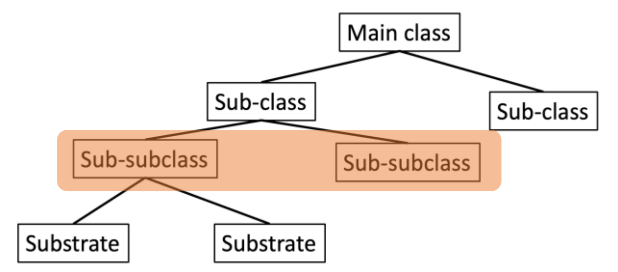
There also exists twilight zone enzymes

Current method do not provide good support for twilight zone





EC # - level 3



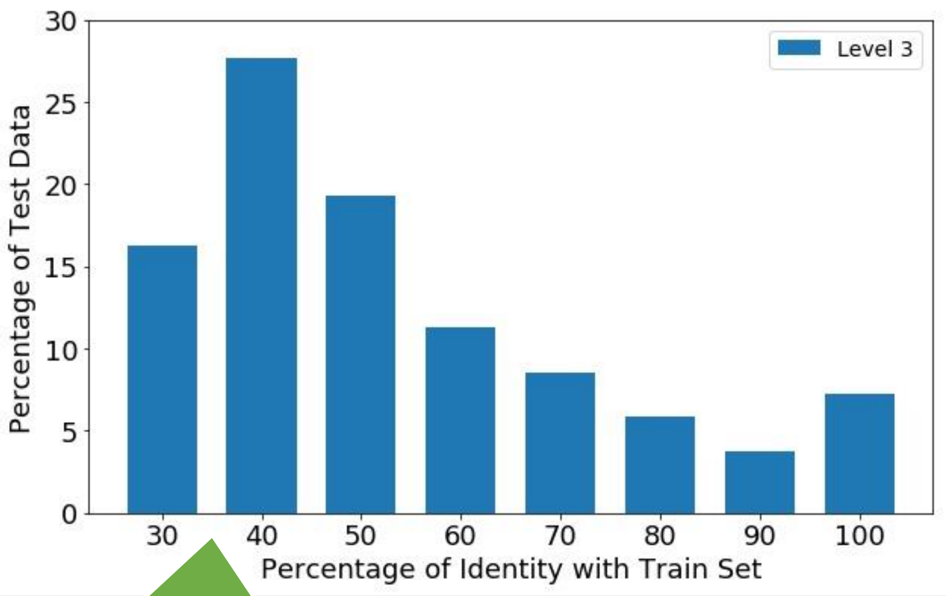
Dataset used for expt: Swiss-Prot

Total # of families: 185

Min # of annotated member: 30

Total	Training Data	Blind Test Set*
SWISS-PROT	Annotated before 2019	Annotated in 2019 or later
254,176 seq	251,817 seq	2,359 seq

*As all methods, compared in our experiment, are trained using Swiss-Prot annotations of 2018 or earlier, we used the rest, annotated in 2019 or later, as our blind test set.



45% test data lies in or near twilight zone

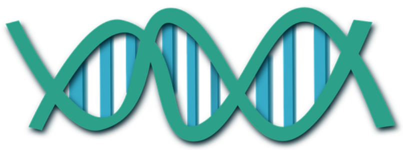
EnsembleFam's performance on EC-level 3 predictions in and near the twilight zone in Swiss-Prot

0 <= identity <= 30	TP	FP	Specificity	Sensitivity	Precision	F1-score	Geo Mean
e-EnsembleFam	**109**	290	98.54	**17.69**	27.32	**21.48**	**41.75**
DEEPre	82	231	98.71	13.31	26.20	17.65	36.25
DeepEC	46	29	**99.98**	7.47	**61.33**	13.31	27.33
EFICAz2.5	81	75	99.92	13.15	51.92	20.98	36.25
CatFam	33	32	99.96	5.36	50.77	9.69	23.14
ECPred	17	**25**	**99.98**	2.76	40.48	5.17	16.61

30 < identity <= 40	TP	FP	Specificity	Sensitivity	Precision	F1-score	Geo Mean
e-EnsembleFam	**398**	121	98.89	**57.85**	76.69	**65.95**	**75.64**
DEEPre	211	196	96.64	30.67	51.84	38.54	54.44
DeepEC	77	47	99.96	11.20	62.10	18.97	33.46
EFICAz2.5	357	78	99.92	51.89	92.07	63.58	72.01
CatFam	109	**14**	**99.99**	15.84	**88.62**	26.88	39.80
ECPred	34	27	99.98	4.94	55.73	9.08	22.22

Identifying enzymes in new genomes

- **Input**

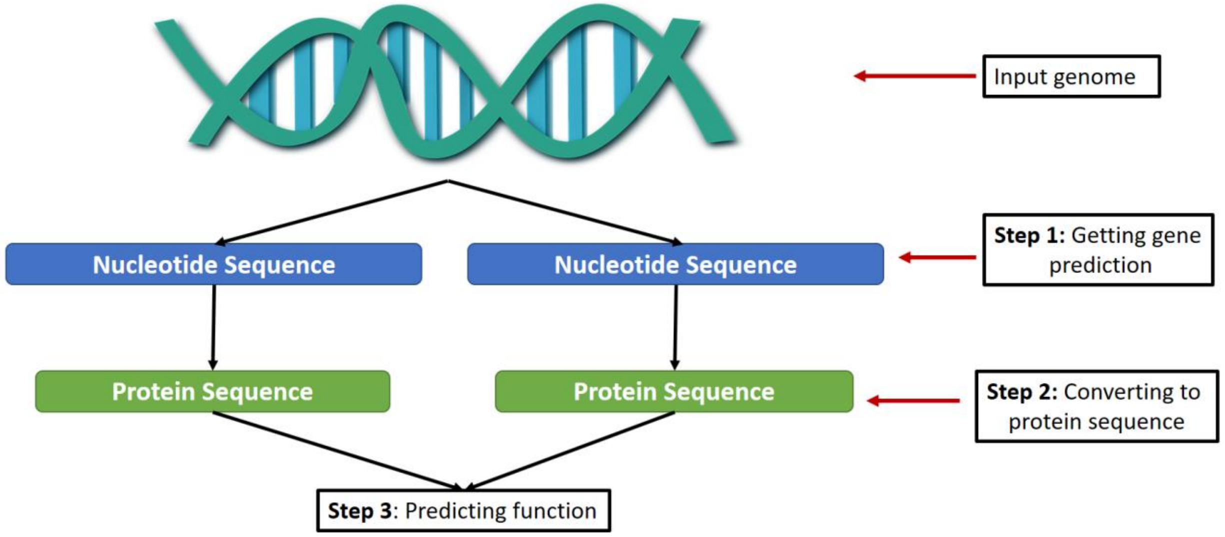


Genome

Enzyme Name	EC Number
PETase (PET hydrolase)	3.1.1.101
MHETase (MHET hydrolase)	3.1.1.102
Terpene (MLH)	3.1.1.83

Enzymes of interest

- **Output**
  - Enzymes of interest exist in genome or not



Enzyme hunting in new fungal genomes

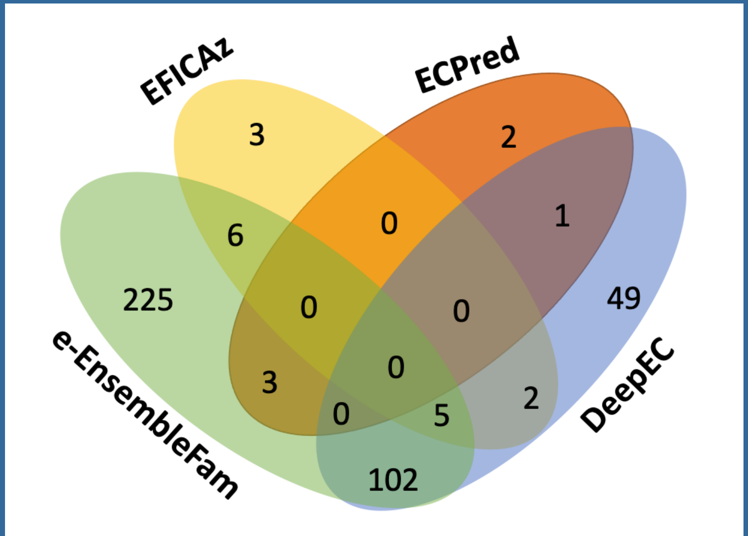
EC level 3 prediction of diff methods. EnsembleFam provides more predictions than competing methods

Genome 1: # predicted genes = 3302		
	Enzyme	Non-enzyme
EnsembleFam	635 (19%)	2667 (81%)
DeepEC	56 (2%)	3246 (98%)

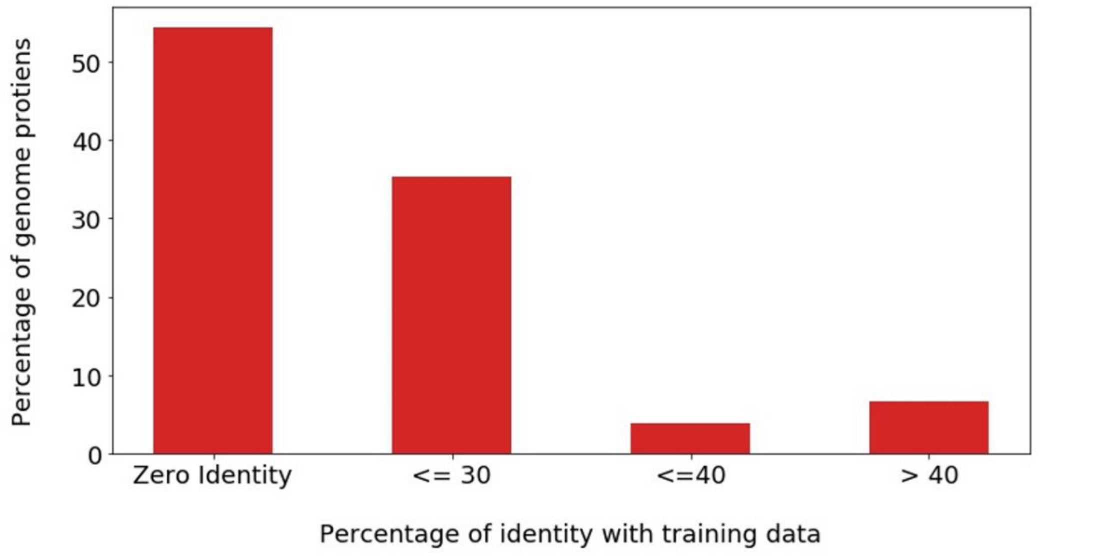
Genome 2: # predicted genes = 3864		
	Enzyme	Non-enzyme
EnsembleFam	725 (19%)	3139 (81%)
DeepEC	54 (1%)	3810 (99%)

Genome 3: # predicted genes = 3599		
	Enzyme	Non-enzyme
EnsembleFam	684 (19%)	2915 (81%)
DeepEC	68 (2%)	3531 (98%)

EC level-3 prediction for chr1 of a new fungal genome



Genome 1 – chr1: Total predicted gene = 504		
	Predicted Enzyme	Non-Enzyme
e-EnsembleFam	**341 *(67.66%)***	163 *(32.34%%)*
DeepEC	159 *(31.35%)*	345 *(68.65%)*
EFICAz2.5	16 *(3.18%)*	488 *(96.82%)*
ECPred	6 *(1.20%)*	498 *(98.80%)*



Useful info is overlooked

There are similarities in dissimilarities

Two take-home messages

Prediction model assessment needs careful thought

Easy questions, hard questions, surprise questions, & many more nuances

A lot of useful information get overlooked

Similarity of dissimilarities

Neamul Kabir & Limsoon Wong, “EmsembleFam: Towards more accurate protein family prediction in the twilight zone,” *BMC Bioinformatics*, 23:90, 2022