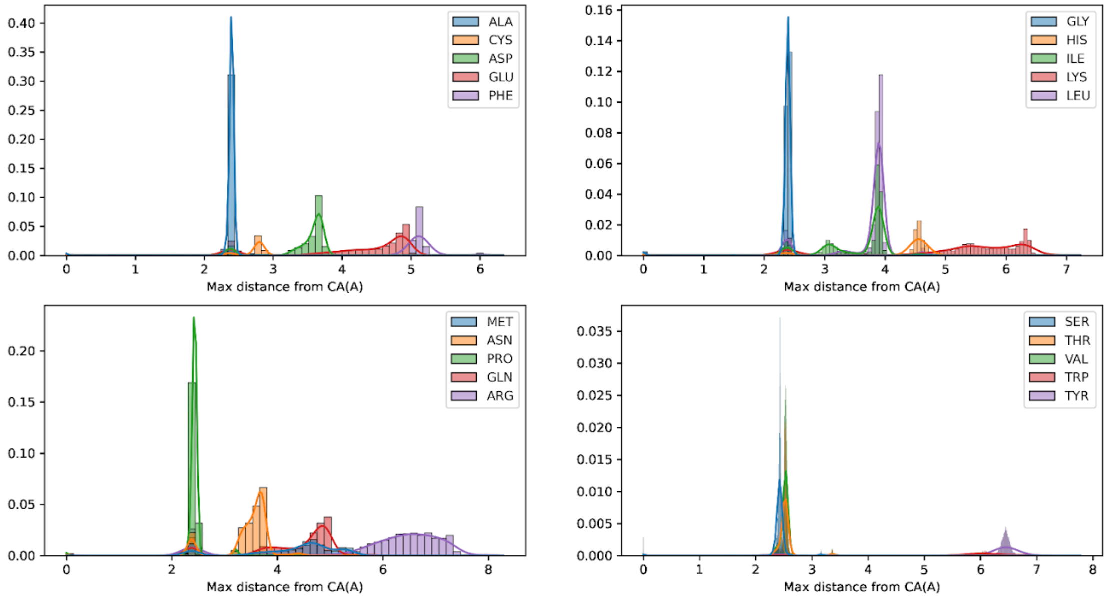
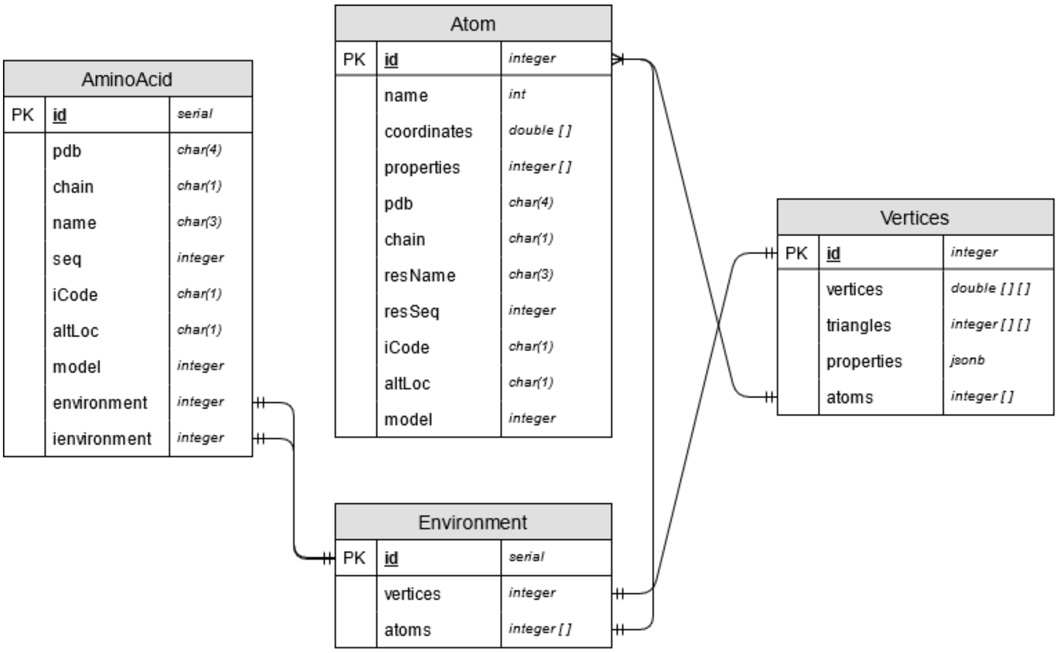
Supplementary material

Supplementary Table 1. Ten folds split of the available environments for arginine used in the training with augmentation

Fold	+ Environments	+ Proteins	- Environments	Proteins
1	1,867	187	18,670	224
2	1,978	154	18,780	191
3	1,881	160	18,810	197
4	1,934	200	19,340	237
5	1,957	164	19,570	201
6	1,869	186	18,690	223
7	1,870	182	18,700	219
8	1,886	162	18,860	219
9	1,877	167	18,770	204
10	1,658	126	16,580	163



Supplementary Figure 1: Distribution of the distances from its CA to its most distant atom in all of the amino acids in the first 1000 PDB structures.



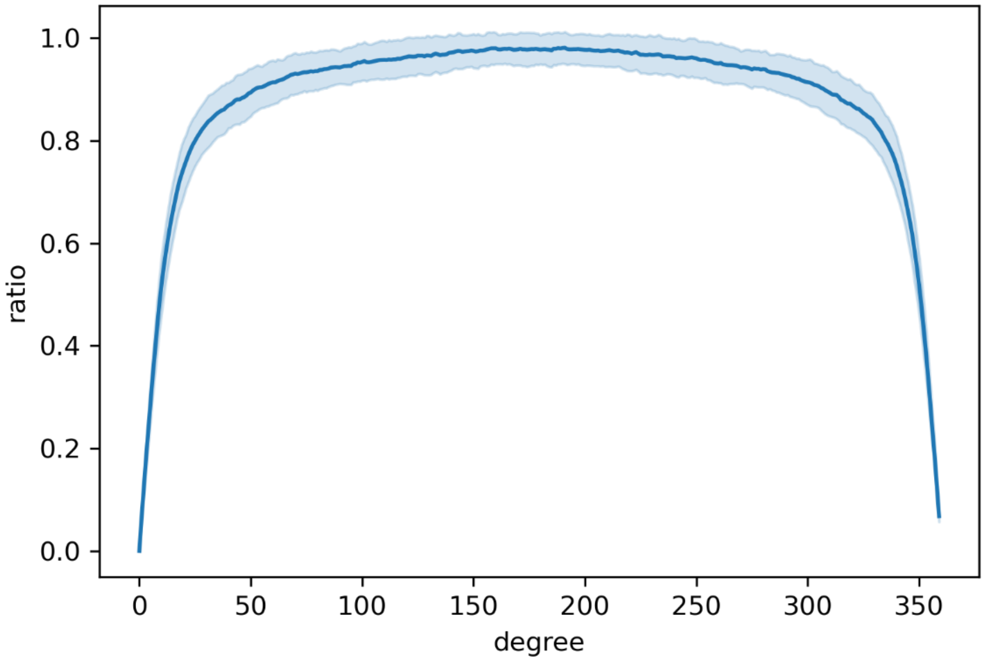
AminoAcid		
PK	id	serial
	pdb	char(4)
	chain	char(1)
	name	char(3)
	seq	integer
	iCode	char(1)
	altLoc	char(1)
	model	integer
	environment	integer
	ienvironment	integer

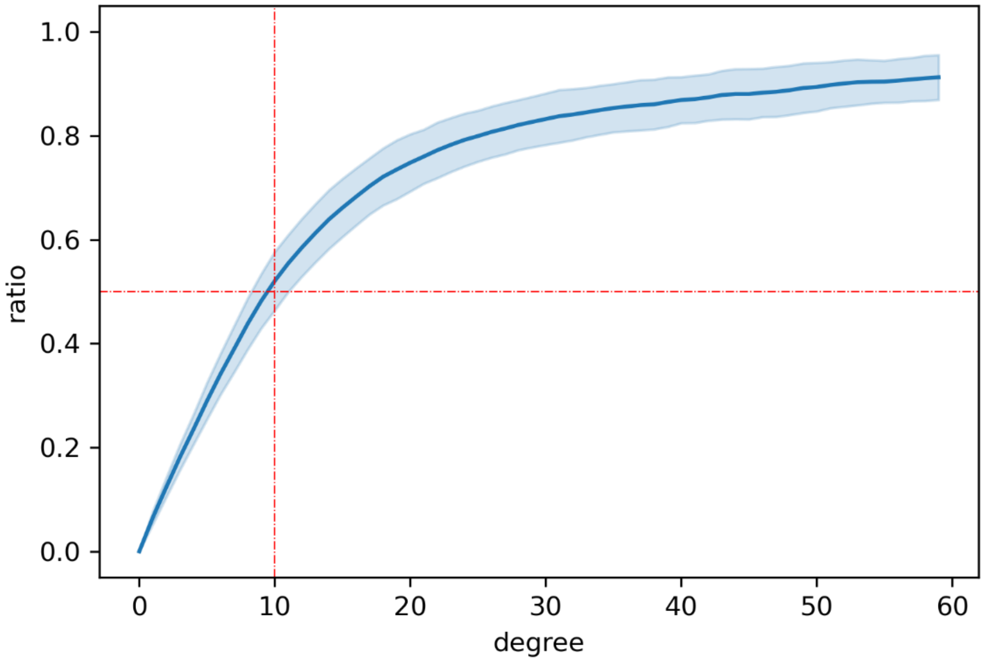
Atom		
PK	id	integer
	name	int
	coordinates	double [ ]
	properties	integer [ ]
	pdb	char(4)
	chain	char(1)
	resName	char(3)
	resSeq	integer
	iCode	char(1)
	altLoc	char(1)
	model	integer

Vertices		
PK	id	integer
	vertices	double [ ] [ ]
	triangles	integer [ ] [ ]
	properties	jsonb
	atoms	integer [ ]

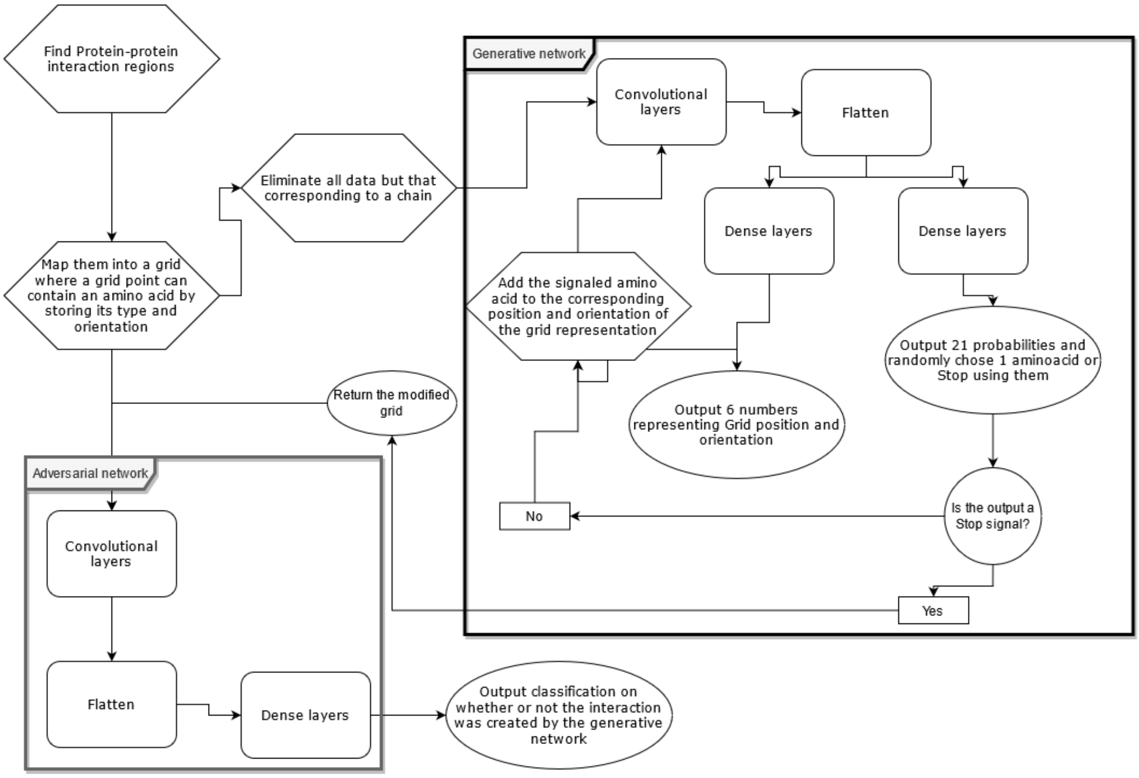
Environment		
PK	id	serial
	vertices	integer
	atoms	integer [ ]

Supplementary Figure 2: The database schema for the environments and related information storage. [ ] Represents one dimensional arrays while [ ] [ ] represents two-dimensional arrays.





Supplementary Figure 3: Portion of initially true points of the grid that changed after a turn of x degrees. The shadow shows the standard deviation obtained after computing the same ratio for fifty environments.

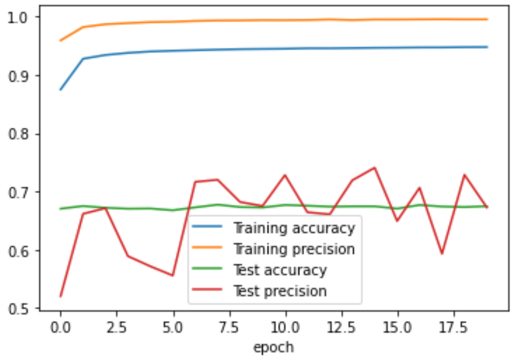

Find Protein-protein interaction regions
Map them into a grid where a grid point can contain an amino acid by storing its type and orientation
Eliminate all data but that corresponding to a chain
Return the modified grid
Generative network
Convolutional layers
Flatten
Dense layers
Dense layers
Add the signaled amino acid to the corresponding position and orientation of the grid representation
Output 6 numbers representing Grid position and orientation
Output 21 probabilities and randomly chose 1 aminoacid or Stop using them
Is the output a Stop signal?
No
Yes
Adversarial network
Convolutional layers
Flatten
Dense layers
Output classification on whether or not the interaction was created by the generative network

Supplementary Figure 4: Alternative flowchart that was considered using a generative adversarial network approach. The idea was for the generative network to output the whole interaction when given one Protein-Protein interaction region with all chains erased but one. The grids used would not be the same as in the main flowchart and instead, contain only amino acids position and orientation because the generative network adds amino acids to the input grid. The needed code was written but it was not trained because of time constraints as the main workflow was already difficult to train and the generative network is computationally expensive. Moreover, this flowchart does not quite fit the project objectives as even if it perfectly worked it would output protein-protein-like interactions, which might not directly help with *de novo* peptide design. Additionally, it losses the surface representation to a more minimalistic one, most likely generating an important information loss.

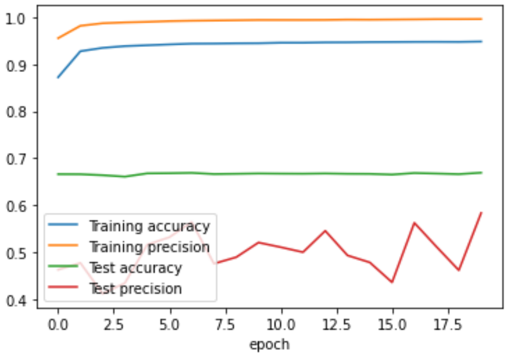
MODELS PERFORMANCE

Arginine binary model with **data augmentation** (original + 9 rotations of 90 degrees) (data augmentation was only performed at positive cases, more negative ones were sampled so that training was balanced)

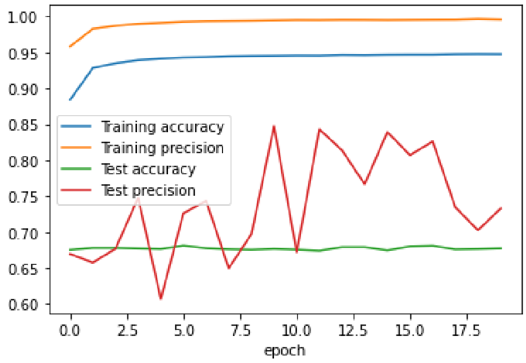
Crossvalidation 1



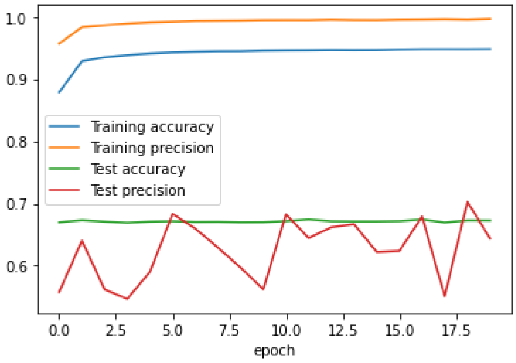
Crossvalidation 2

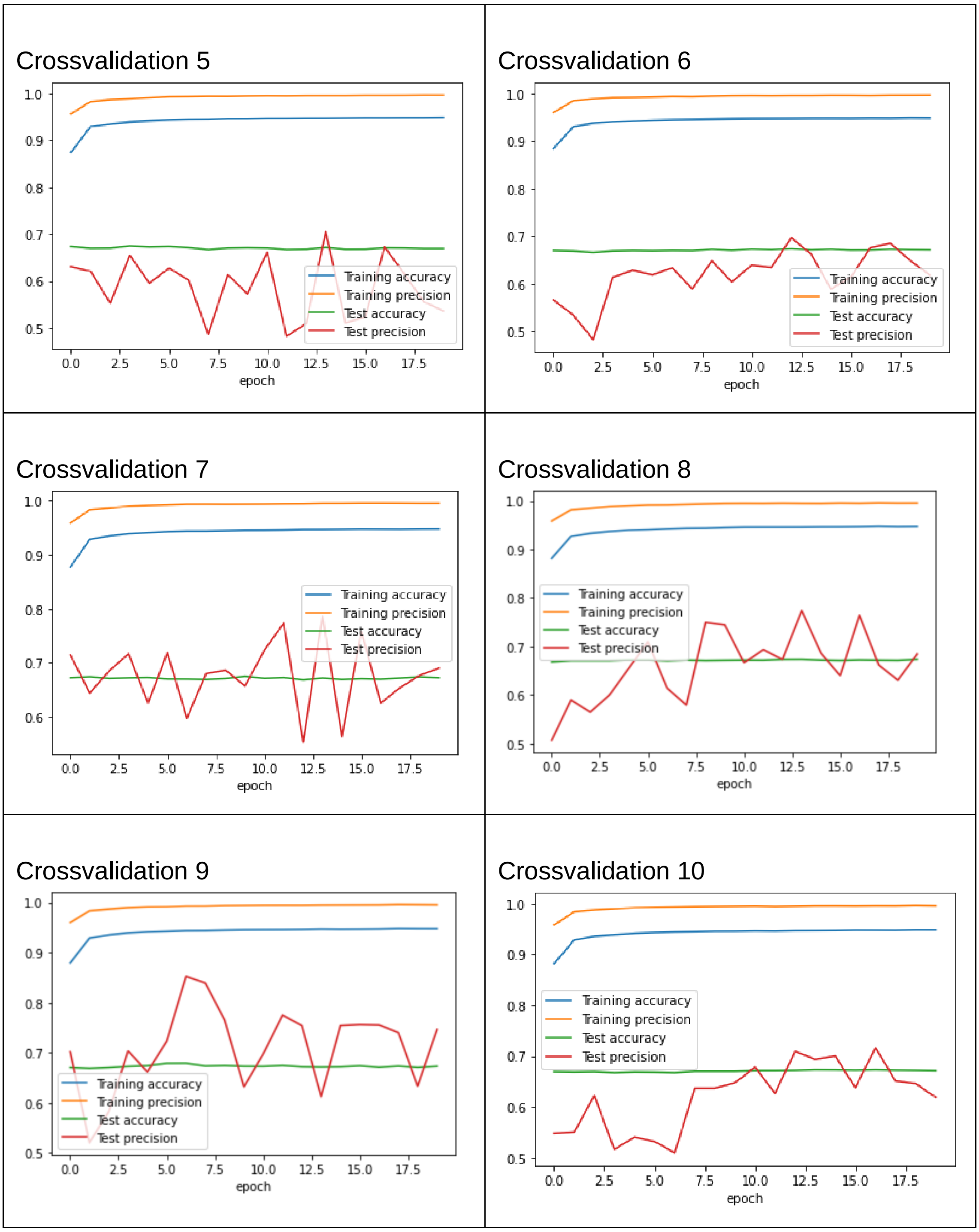


Crossvalidation 3



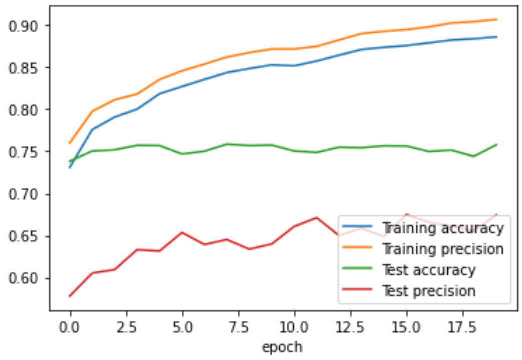
Crossvalidation 4



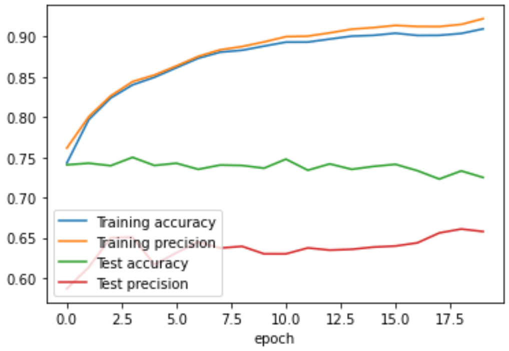


Proline binary model environments **aligned** and **cropped** and positive environments **oversampled**, this model was trained to try to determine the influence of different factors on other models.

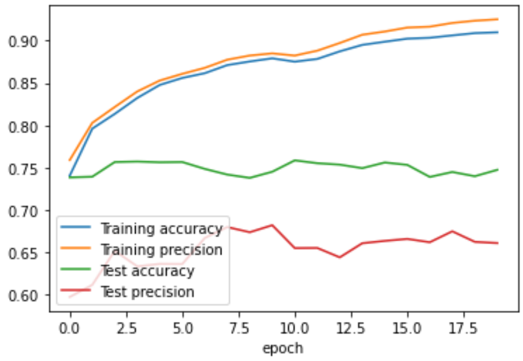
Crossvalidation 1



Crossvalidation 2

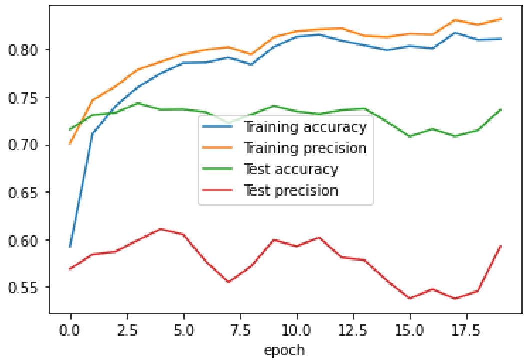


Crossvalidation 3

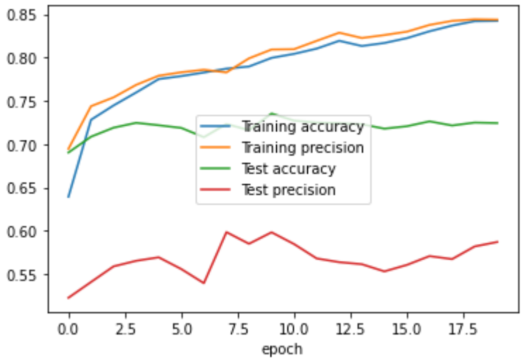


Proline binary model environments **aligned** and **cropped**, this model was trained to try to determine the influence of different factors on other models.

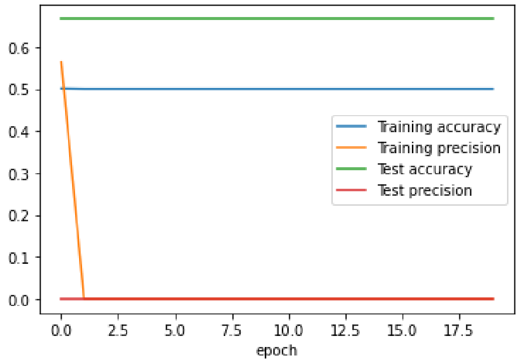
Crossvalidation 1



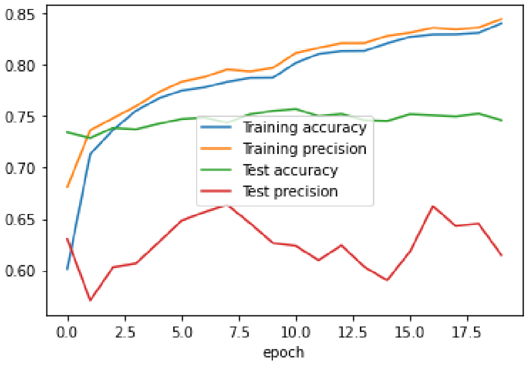
Crossvalidation 2



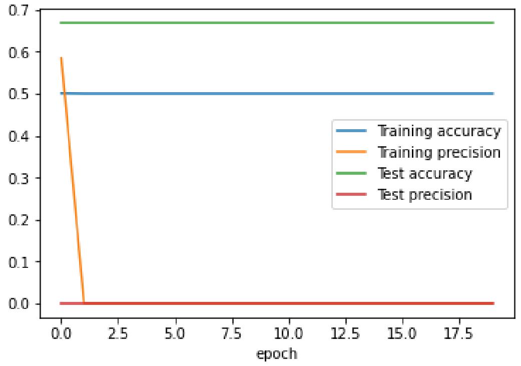
Crossvalidation 3



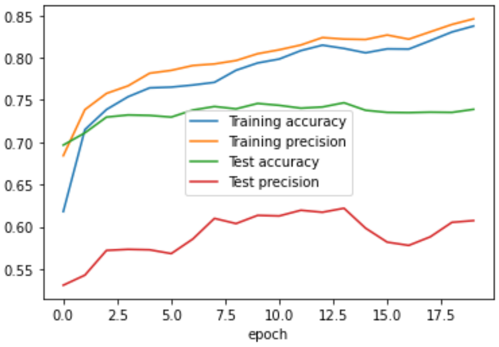
Crossvalidation 4



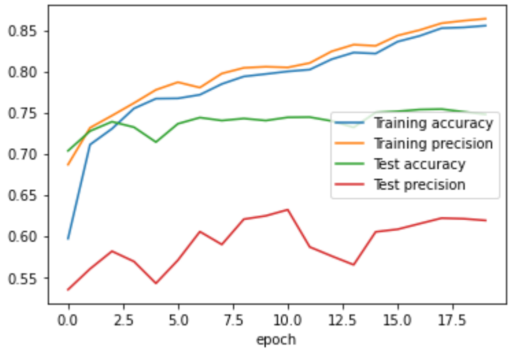
Crossvalidation 5



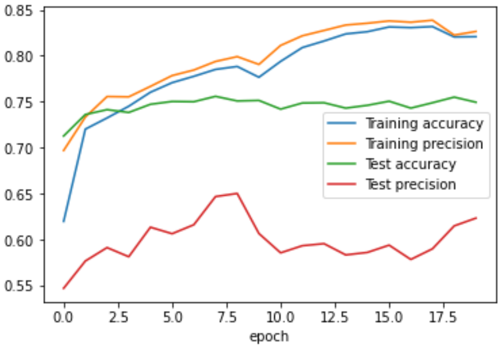
Crossvalidation 6



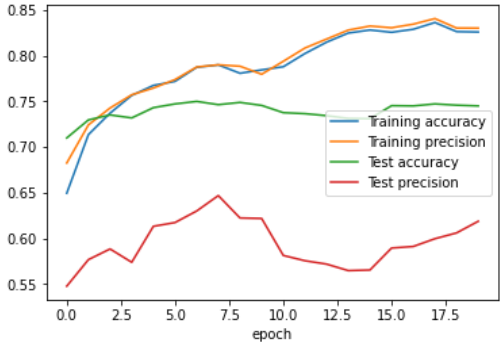
Crossvalidation 7



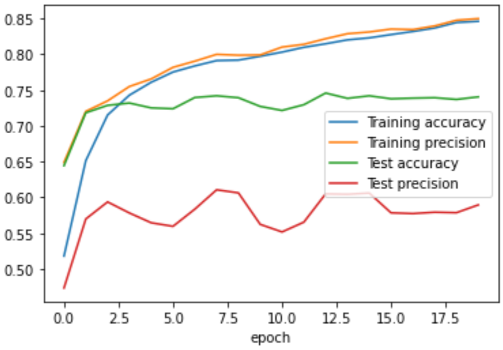
Crossvalidation 8



Crossvalidation 9

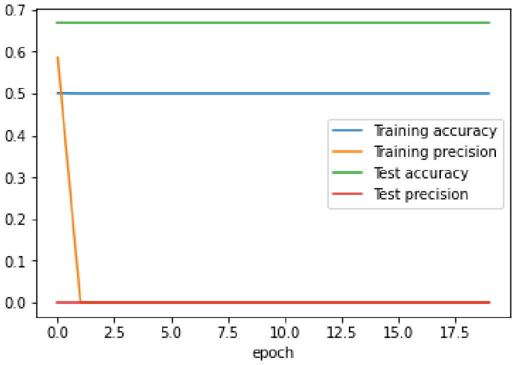


Crossvalidation 10



Proline binary model environments **aligned**

The training of this model quickly started to get 0 training precision for all folds:



Arginine binary model

Multiclass model. For the multiclass model the 10 folds approach was not yet chosen and training was more informal and experimental, some information of those first attempts was eventually lost. The maximum accuracy reached on the training set was 0.5, that same model obtained 0.2 precision in a randomly chosen test set (independence constrains throw UniProt codes was not applied yet)