efsa

European Food Safety Authority

preDQ – a tool for peptide binding prediction to HLA-DQ2 and HLA-DQ8

https://r4eu.efsa.europa.eu/app/predq

Questions & Answers

Irini Doytchinova, Ivan Dimitrov, Mariyana Atanasova, Antonio F. Dumont, Frits Koning, Javier F. Moreno, Estefania N. Fernandez

User Experience

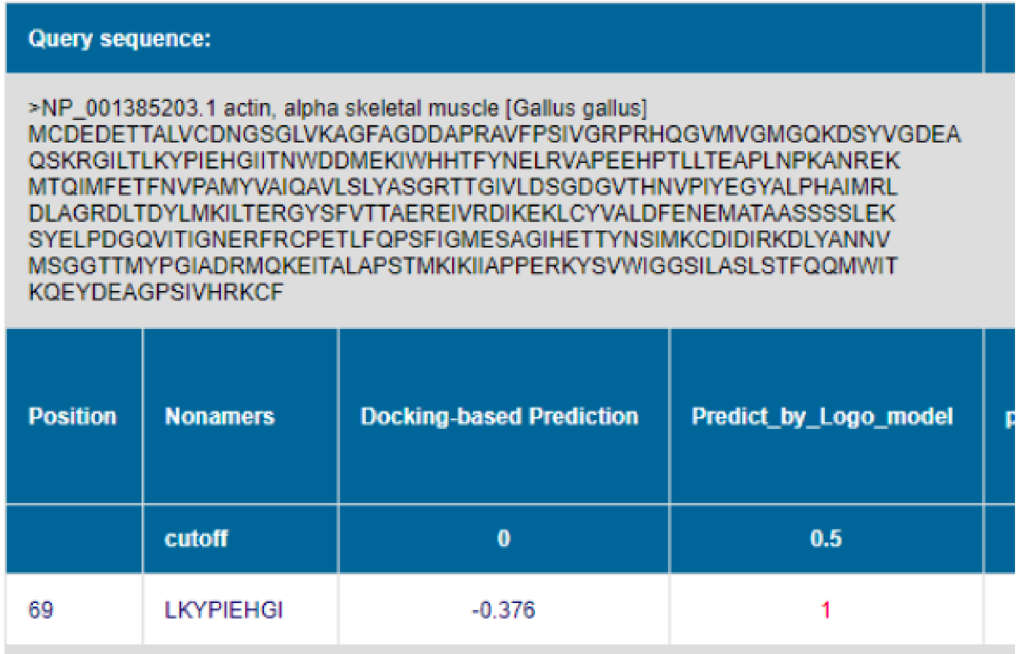
Question:

- The text field does not correctly accept a sequence in FASTA format, but rather only accepts a "bare sequence". This is unusual because putting a sequence in FASTA format into a text field is a feature common to most Internet-based bioinformatics search tools. If a sequence in FASTA format is placed in the text field, uppercase letters in the header are treated as amino acids and appended to the sequence before searching. Conversely, lowercase letters in the header and amino acid sequence remain unrecognized.

Question:

- The text box only accepts one sequence at a time, the ability to process sequences in bulk is needed as it would facilitate more rigorous testing and be more consistent with real-life use cases.

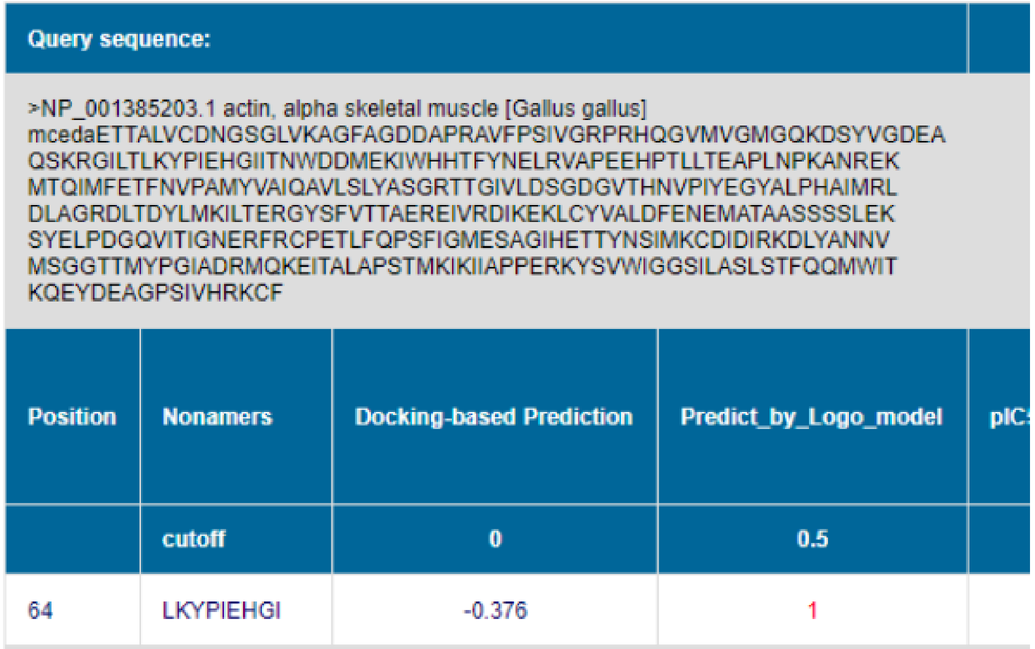
Amino Acids in Lower Case Not Recognized



Query sequence:

>NP_001385203.1 actin, alpha skeletal muscle [Gallus gallus]
MCDEDETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA
QSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREK
MTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRL
DLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEK
SYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV
MSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIT
KQEYDEAGPSIVHRKCF

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	p
	cutoff	0	0.5	
69	LKYPIEHGI	-0.376	1	



Query sequence:

>NP_001385203.1 actin, alpha skeletal muscle [Gallus gallus]
mcedaETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA
QSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREK
MTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRL
DLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEK
SYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV
MSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIT
KQEYDEAGPSIVHRKCF

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	pIC
	cutoff	0	0.5	
64	LKYPIEHGI	-0.376	1	

Uppercase amino acids, hit at position 69

First 5 amino acids lowercase, hit now at position 64

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. Biochem J. 1984;219(2):345-373. doi:10.1042/bj2190345

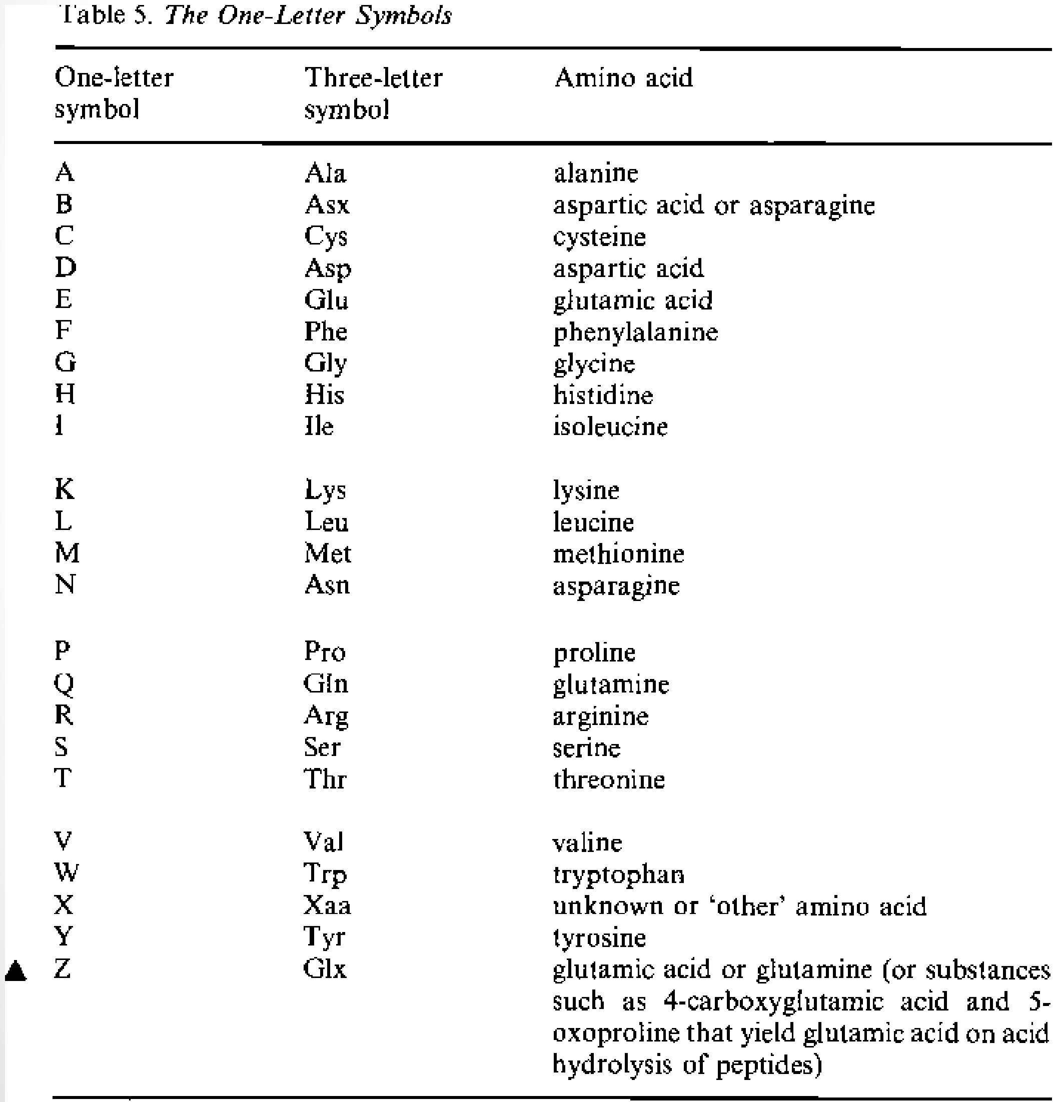


Table 5. *The One-Letter Symbols*

One-letter symbol	Three-letter symbol	Amino acid
A	Ala	alanine
B	Asx	aspartic acid or asparagine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xaa	unknown or 'other' amino acid
Y	Tyr	tyrosine
Z	Glx	glutamic acid or glutamine (or substances such as 4-carboxyglutamic acid and 5-oxoproline that yield glutamic acid on acid hydrolysis of peptides)

Amino acids are coded
with **one-CAPITAL-letter**
or with three letters:
the first is CAPITAL,
the rest are small.
In a protein sequence,
the aa residues are coded
with one-CAPITAL-letter.

User Experience

Answer:

- The tool accepts plain format in the text box and fasta format as an uploaded file. The user can copy/paste a single protein or a peptide in the text box for a prompt test or upload a file with many proteins in fasta format. In the latter case, predictions take longer.
- For user convenience, an example in plain format is added in the text box.

Search Result Observations

Question:

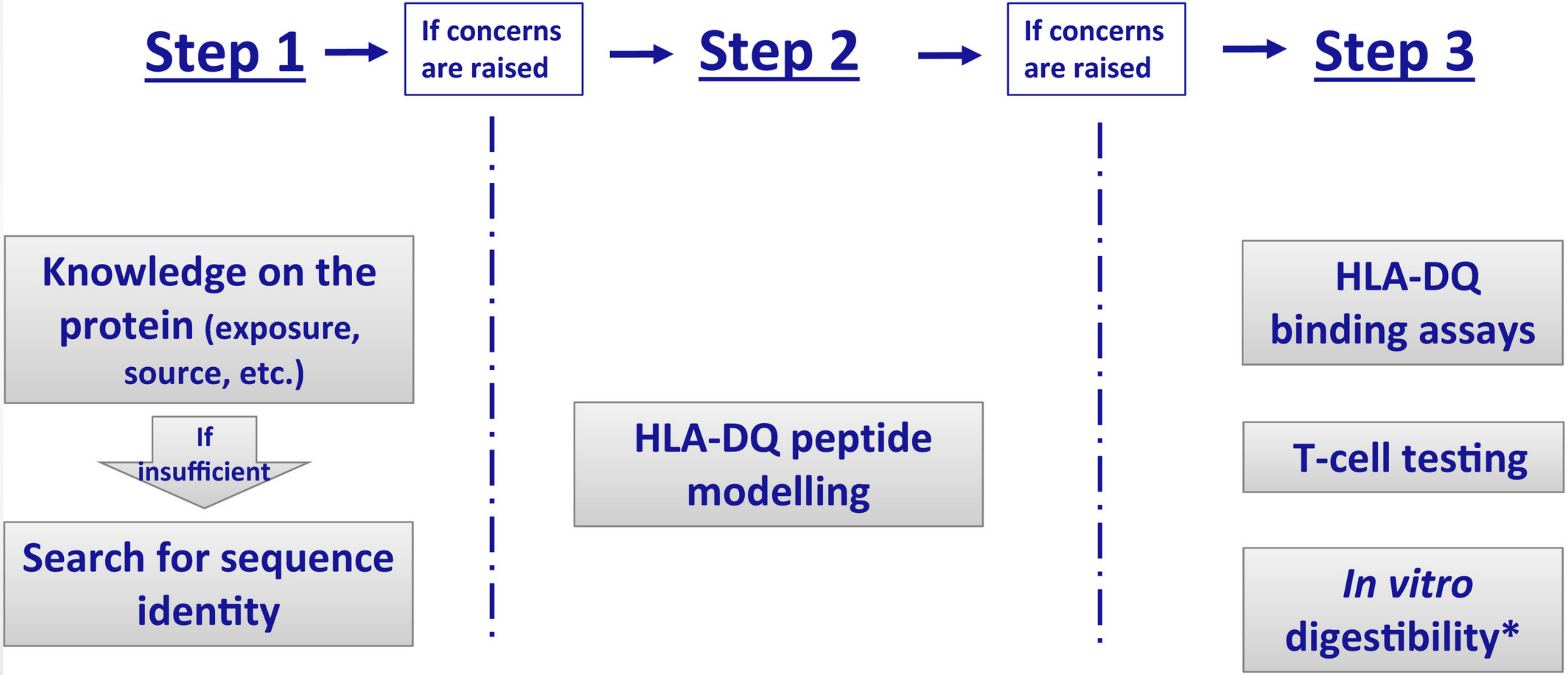
- When using the tool, proteins that are not expected to trigger a celiac response appear to have nonamer binders. Proteins that have been tested include animal muscle proteins such as actin and myosin, and plant storage and metabolic proteins. Despite having no association with celiac disease and having annual consumption levels of kilograms per year, chicken muscle myosin heavy chain (NP_001107181.2) returns 73 potential DQ8.1 and 22 potential DQ2.5 binding sites. In the absence of a detailed documentation describing how the tool operates or the data used to train the tool, it is impossible to draw a conclusion regarding such results.

Search Result Observations

Answer:

- People with celiac disease carry one or both of the HLA-DQ2 and DQ8 genes, but so does up to 25-30% of the general population. **Carrying HLA-DQ2 and/or DQ8 is not a diagnosis of celiac disease**, nor does it mean the carrier will ever develop celiac disease. However, if you carry HLA-DQ2 and/or DQ8, your risk of developing celiac disease is 3% instead of the general population risk of 1%.

Stepwise approach for risk assessment



Step 1 → If concerns are raised → Step 2 → If concerns are raised → Step 3

Step 1:

- Knowledge on the protein (exposure, source, etc.)
- If insufficient
- Search for sequence identity

Step 2:

- HLA-DQ peptide modelling

Step 3:

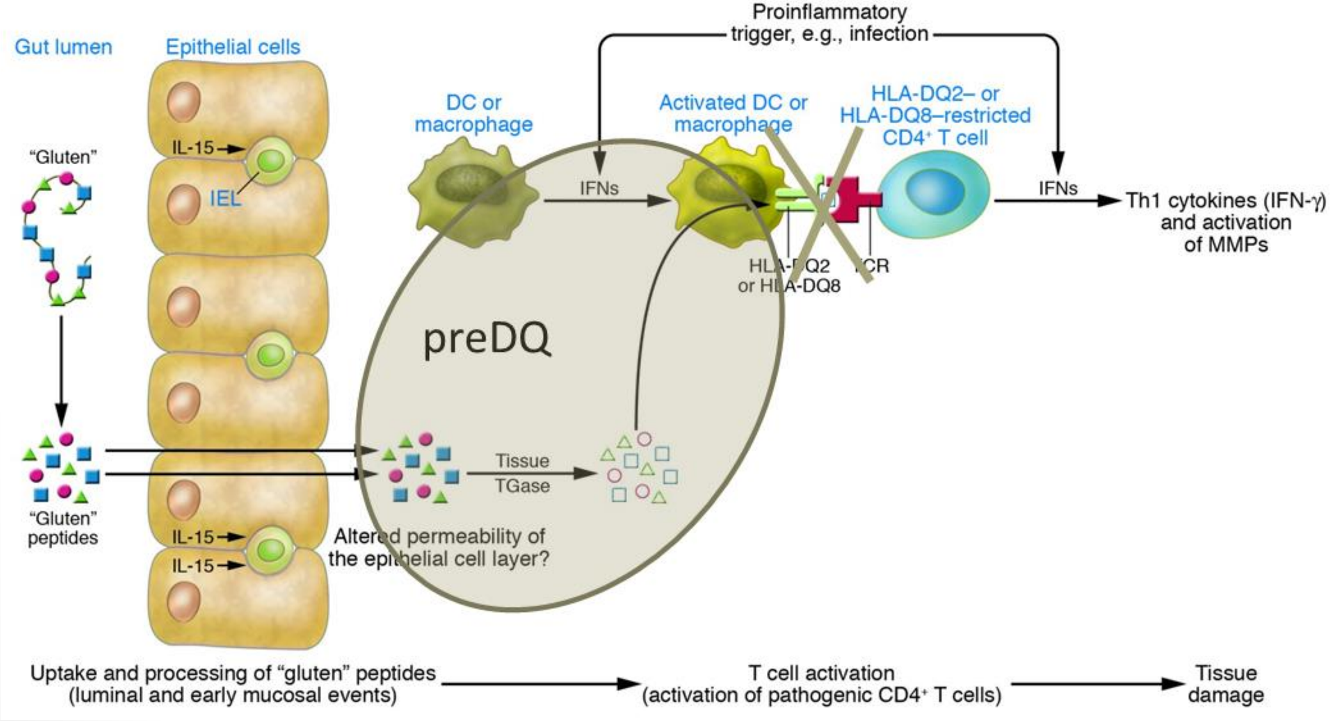
- HLA-DQ binding assays
- T-cell testing
- *In vitro* digestibility*

EFSA Panel on Genetically Modified Organisms (GMO), Naegeli H, Birch AN, et al. Guidance on allergenicity assessment of genetically modified plants. *EFSA J*. 2017;15(6):e04862.

Search Result Observations

Answer:

- preDQ mimics two steps in the antigen processing:
  1. Deamidation of peptides: Gln to Glu at certain positions
  2. Binding to HLA-DQ2 or HLA-DQ8

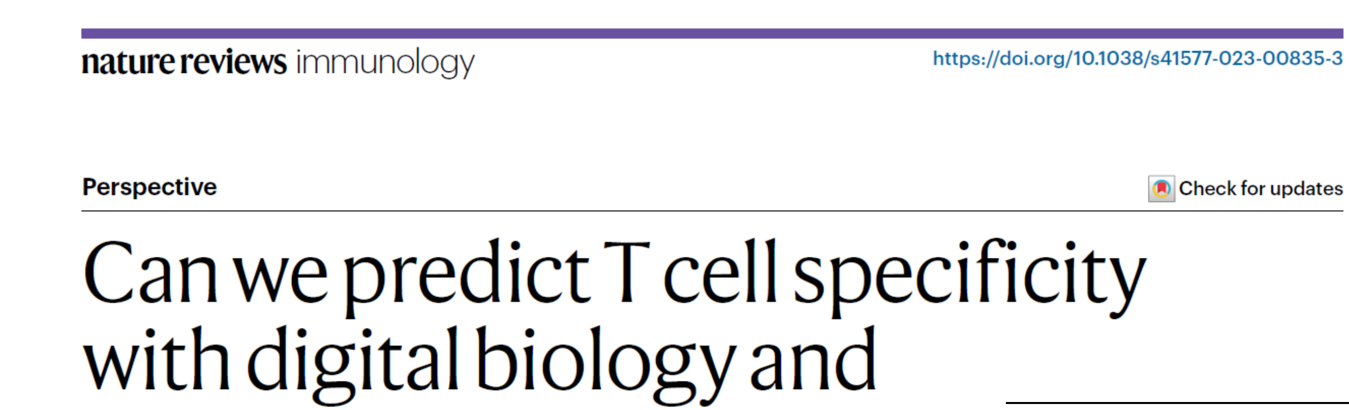

Proinflammatory trigger, e.g., infection
Gut lumen
Epithelial cells
DC or macrophage
Activated DC or macrophage
HLA-DQ2– or HLA-DQ8–restricted CD4+ T cell
"Gluten"
IL-15
IEL
IFNs
IFNs
Th1 cytokines (IFN-γ) and activation of MMPs
HLA-DQ2 or HLA-DQ8
TCR
preDQ
Tissue TGase
"Gluten" peptides
IL-15
IL-15
Altered permeability of the epithelial cell layer?
Uptake and processing of "gluten" peptides (luminal and early mucosal events)
T cell activation (activation of pathogenic CD4+ T cells)
Tissue damage

Kagnoff M. J Clin Invest. 2007;117(1):41-49.

Search Result Observations

Answer:

- The pHLA-TCR interaction is not modelled here because the TCR repertoire is a diverse set of TCRs (approx. 100 million) unique for each individual developed throughout life under the pressure of environmental factors (exposure to different antigens), age (thymus gradually losses function), immunological history (previous infections and vaccinations), diseases (chronic inflammation, cancers, immunosuppressive therapy). Even more, the TCR repertoire changes throughout an individual's life, influenced again by aging, infections and vaccination.
- preDQ is a tool for peptide binding prediction to HLA-DQ2 and HLA-DQ8. preDQ is not a tool for T-cell epitope prediction.



nature reviews immunology https://doi.org/10.1038/s41577-023-00835-3

Perspective

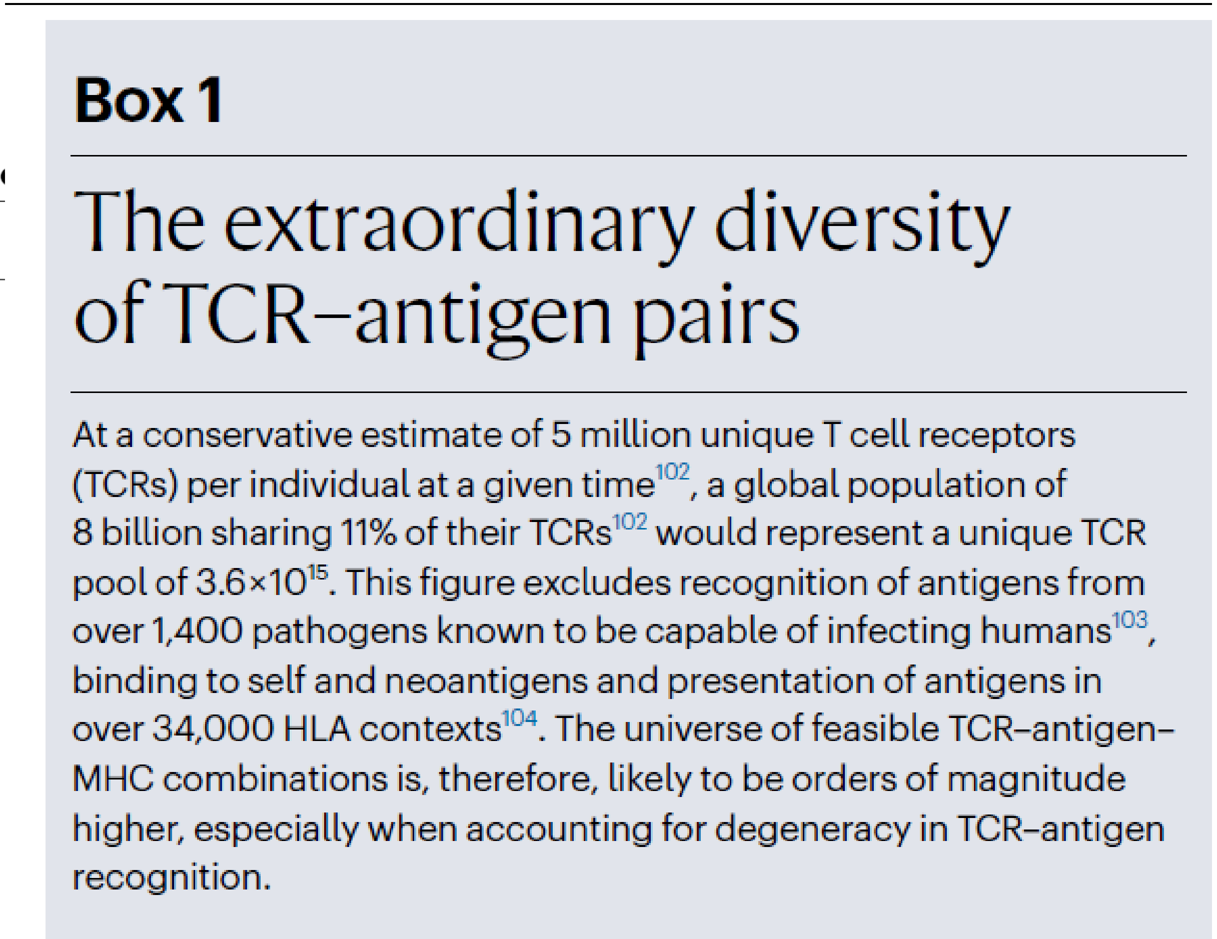
Check for updates

Can we predict T cell specificity with digital biology and machine learning?

Dan Hudson[1,2], Ricardo A. Fernandes[3], Mark Basham[2], Graham Ogg[1,3] & Hashem Koohy

Abstract

Recent advances in machine learning and experimental biology have offered breakthrough solutions to problems such as protein structure



Box 1

The extraordinary diversity of TCR–antigen pairs

At a conservative estimate of 5 million unique T cell receptors (TCRs) per individual at a given time[102], a global population of 8 billion sharing 11% of their TCRs[102] would represent a unique TCR pool of 3.6×10^{15}. This figure excludes recognition of antigens from over 1,400 pathogens known to be capable of infecting humans[103], binding to self and neoantigens and presentation of antigens in over 34,000 HLA contexts[104]. The universe of feasible TCR–antigen–MHC combinations is, therefore, likely to be orders of magnitude higher, especially when accounting for degeneracy in TCR–antigen recognition.

Search Result Observations

Question:

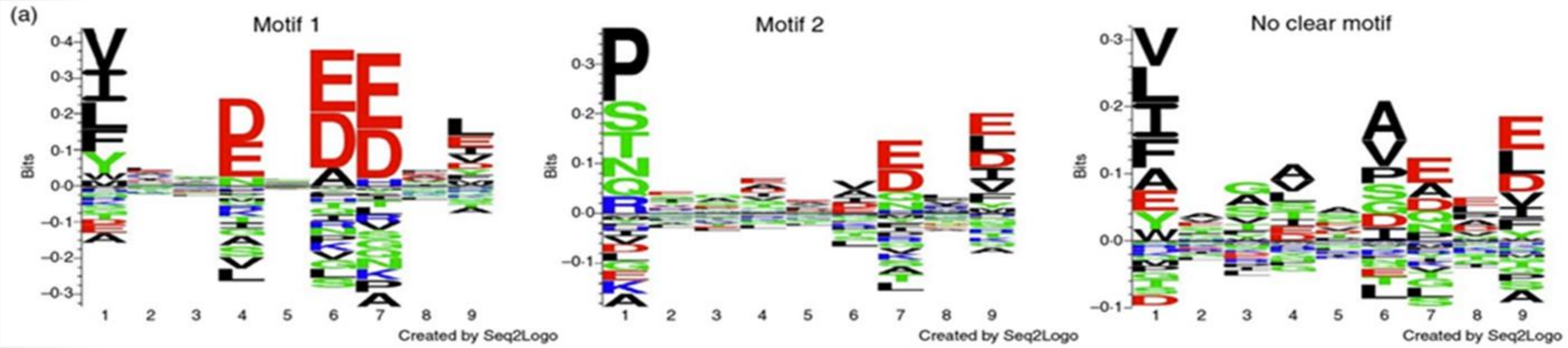
- While the tool returned a celiac peptide, PMPMPELPY, spiked into a protein sequence, it was unable to identify PMPMP**D**LPY or PM**G**MPELPY as being potential celiac peptide candidates despite having 8 out of 9 matching residues and a single conservative substitution.

Search Result Observations

Answer:

- Revisiting the negative set.

HLA-DQ2.5



Koşaloğlu-Yalçın et al. Immunology 2021;162:235-247.

1: ['A', 'E', 'D', 'T', 'G', 'S', 'K', 'R', 'N'],
2: ['L', 'A'],
3: ['L', 'E', 'I', 'R', 'D', 'K'],
4: ['L', 'V', 'S', 'A', 'G', 'I', 'T', 'K', 'R', 'Q', 'M', 'F', 'W'],
5: ['L', 'T', 'K',' R’],

6: ['K', 'L', 'T', 'R', 'S', 'G', 'V', 'P', 'N', 'Q’],
7: ['K', 'A', 'P', 'N', 'Q', 'G', 'S', 'V', 'R', 'L', 'T’],
8: ['L', 'A', 'I', 'Q', 'T'],
9: ['K', 'A', 'S', 'P', 'Q', 'T', 'S', 'M', ‘R’]

27 799 200 non-binding nonamers

Search Result Observations

Answer:

- Revisiting the negative set.

0 – non-preferred aas; 1 – preferred or neutral aas

aa	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
p1	**0**	1	**0**	**0**	1	**0**	1	1	**0**	1	1	**0**	1	1	**0**	**0**	**0**	1	1	1
p2	**0**	1	1	1	1	1	1	1	1	**0**	1	1	1	1	1	1	1	1	1	1
p3	1	1	**0**	**0**	1	1	1	**0**	**0**	**0**	1	1	1	1	0	1	1	1	1	1
p4	**0**	1	1	1	**0**	**0**	1	**0**	**0**	**0**	**0**	1	1	**0**	**0**	**0**	**0**	**0**	**0**	1
p5	1	1	1	1	1	1	1	1	**0**	**0**	1	1	1	1	0	1	0	1	1	1
p6	1	1	1	1	1	**0**	1	1	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1
p7	**0**	1	1	1	1	**0**	1	1	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1
p8	**0**	1	1	1	1	1	1	**0**	1	**0**	1	1	1	**0**	1	1	**0**	1	1	1
p9	**0**	1	1	1	1	1	1	1	**0**	1	**0**	1	**0**	**0**	**0**	**0**	**0**	1	1	1

NB score = p1 + p2 + p3 + p4 + p5 + p6 + p7 + p8 + p9

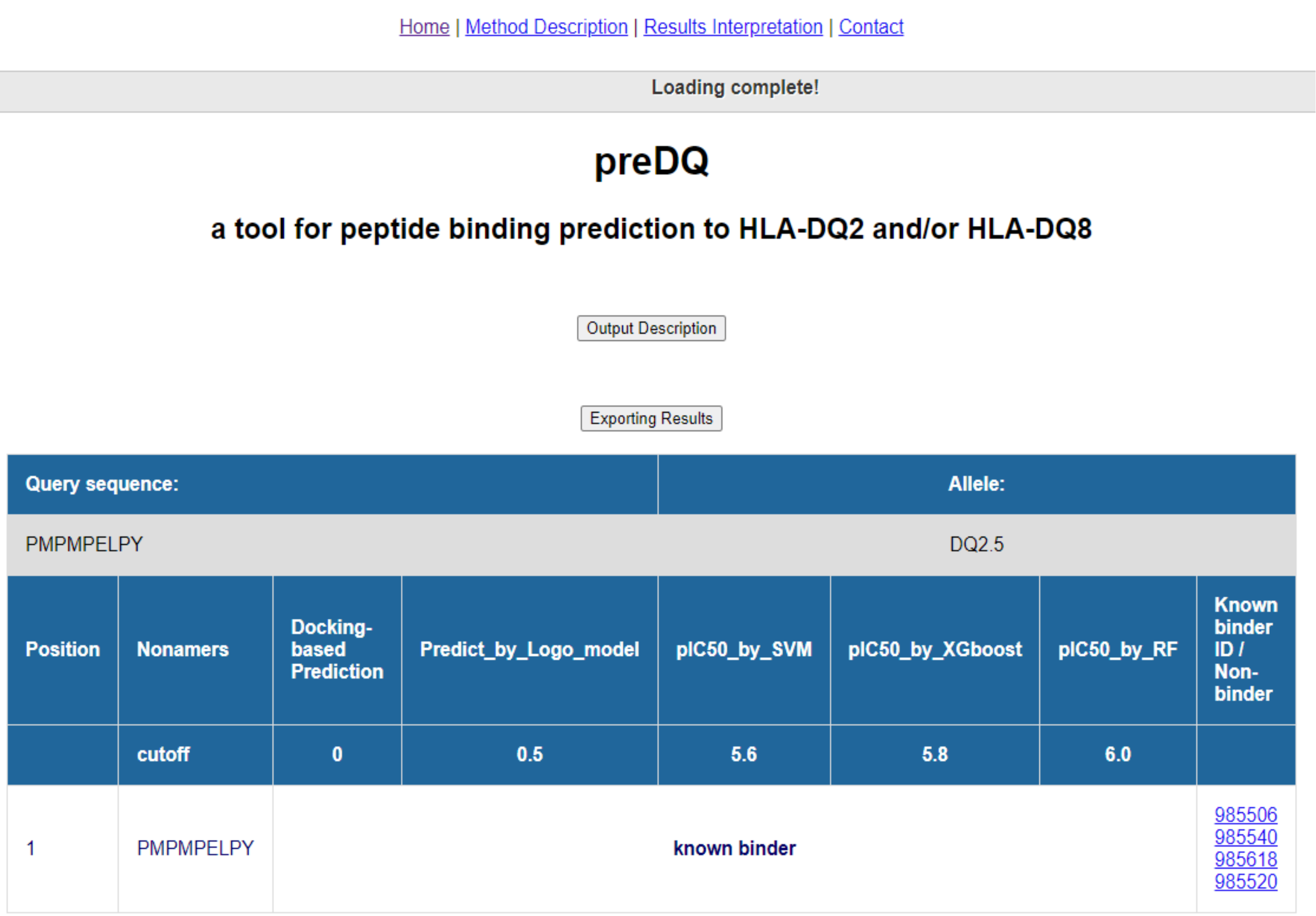
If NBS = 0 → The peptide is classified as a non-binder.

If NBS > 0 → The peptide binding affinity is calculated.

Search Result Observations

Answer:

- PMPMPELPY is a known binder to HLA-DQ2.5 in IEDB.


Home | Method Description | Results Interpretation | Contact
Output Description
Exporting Results

Loading complete!

preDQ

a tool for peptide binding prediction to HLA-DQ2 and/or HLA-DQ8

Query sequence:	Allele:
PMPMPELPY	DQ2.5

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	pIC50_by_SVM	pIC50_by_XGboost	pIC50_by_RF	Known binder ID / Non-binder
	cutoff	0	0.5	5.6	5.8	6.0	
1	PMPMPELPY			known binder			985506 985540 985618 985520

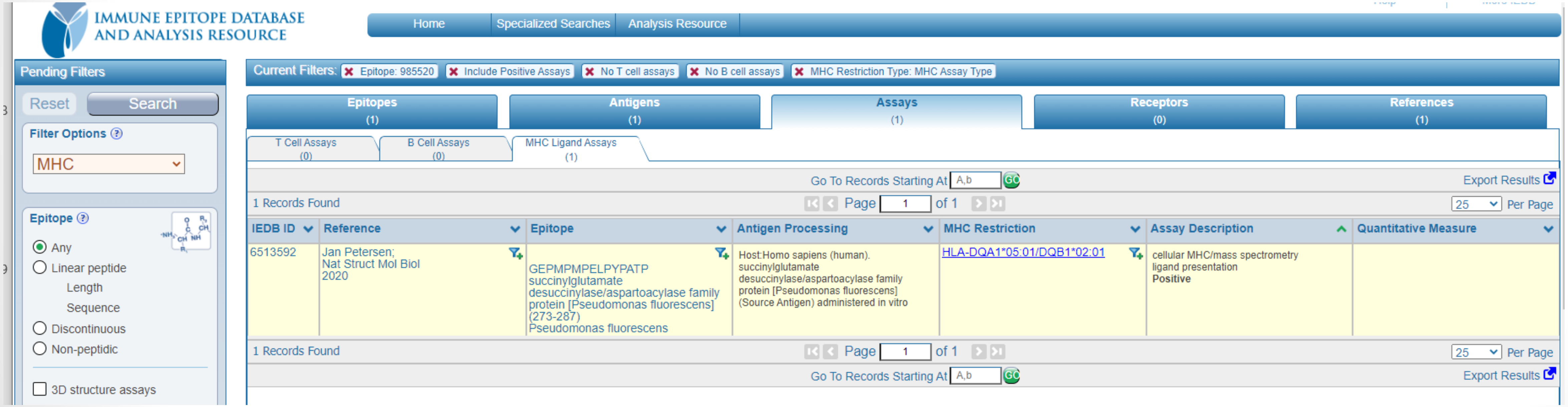
3/5 predict: binder;

majority voting: probable binder

Search Result Observations

Answer:

- PMPMPELPY is a known binder to HLA-DQ2.5 in IEDB.



Current Filters: Epitope: 985520 | Include Positive Assays | No T cell assays | No B cell assays | MHC Restriction Type: MHC Assay Type

1 Records Found

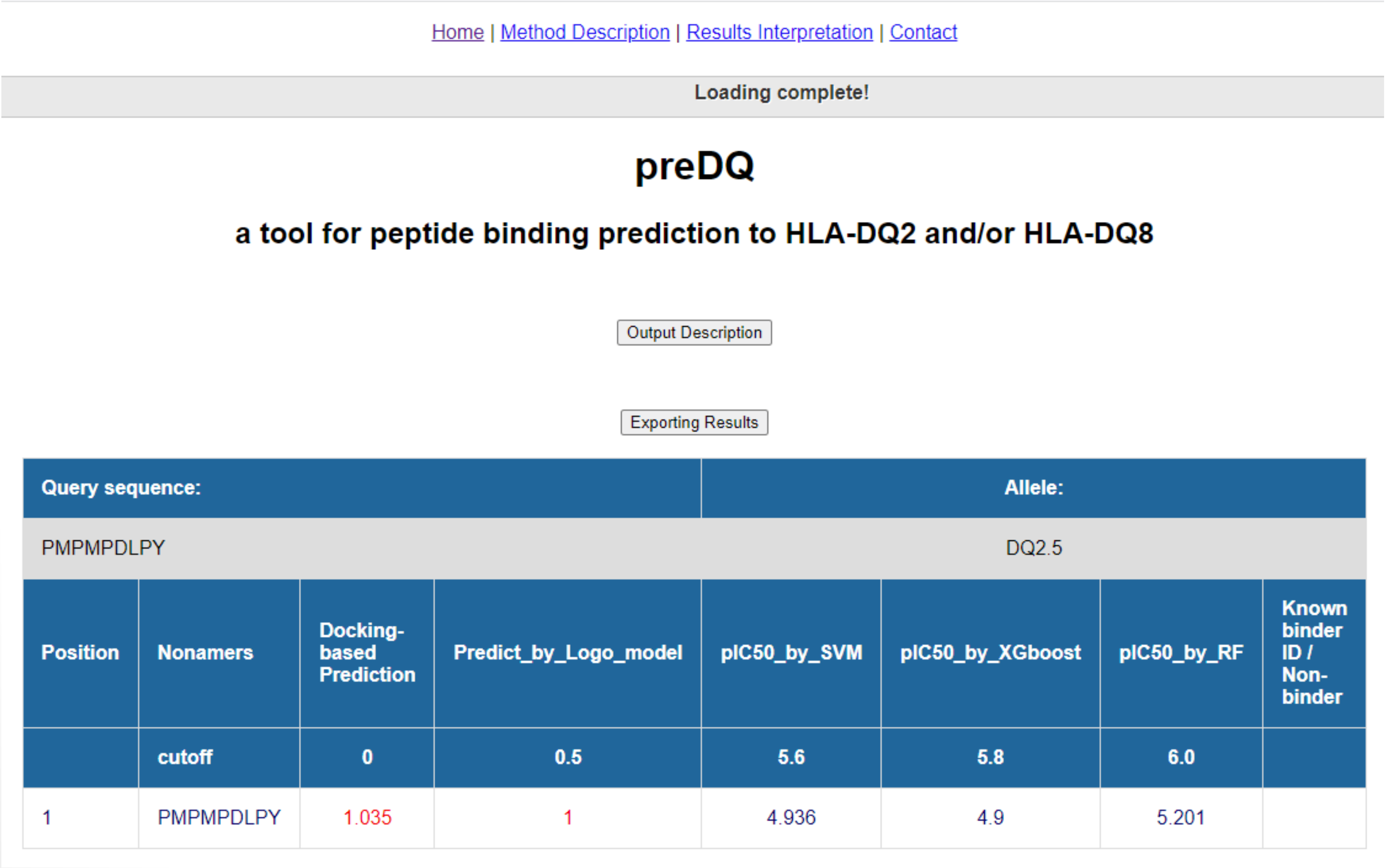
IEDB ID	Reference	Epitope	Antigen Processing	MHC Restriction	Assay Description	Quantitative Measure
6513592	Jan Petersen; Nat Struct Mol Biol 2020	GEPMPMPELPYPATP succinylglutamate desuccinylase/aspartoacylase family protein [Pseudomonas fluorescens] (273-287) Pseudomonas fluorescens	Host:Homo sapiens (human). succinylglutamate desuccinylase/aspartoacylase family protein [Pseudomonas fluorescens] (Source Antigen) administered in vitro	HLA-DQA1*05:01/DQB1*02:01	cellular MHC/mass spectrometry ligand presentation **Positive**	

1 Records Found

Search Result Observations

Answer:

- PMPMP**D**LPY – 2/5 predict: binder; majority voting: probable non-binder


Home | Method Description | Results Interpretation | Contact
Output Description
Exporting Results

Loading complete!

preDQ

a tool for peptide binding prediction to HLA-DQ2 and/or HLA-DQ8

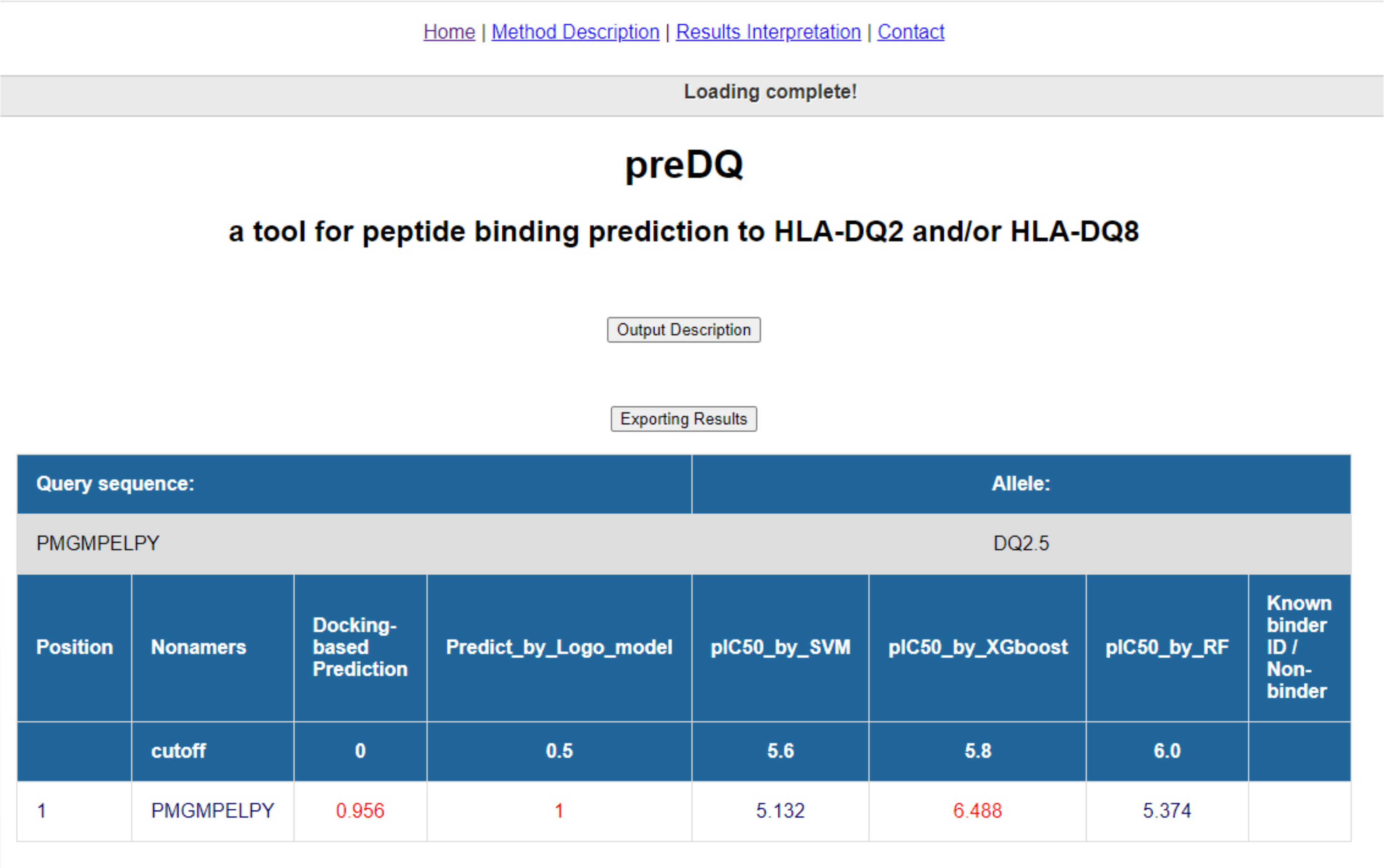
Query sequence:	Allele:
PMPMPDLPY	DQ2.5

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	pIC50_by_SVM	pIC50_by_XGboost	pIC50_by_RF	Known binder ID / Non-binder
	cutoff	0	0.5	5.6	5.8	6.0	
1	PMPMPDLPY	1.035	1	4.936	4.9	5.201	

Search Result Observations

Answer:

- PM**G**MPELPY – 3/5 predict: binder; majority voting: probable binder


Home | Method Description | Results Interpretation | Contact
Output Description
Exporting Results

Loading complete!

preDQ

a tool for peptide binding prediction to HLA-DQ2 and/or HLA-DQ8

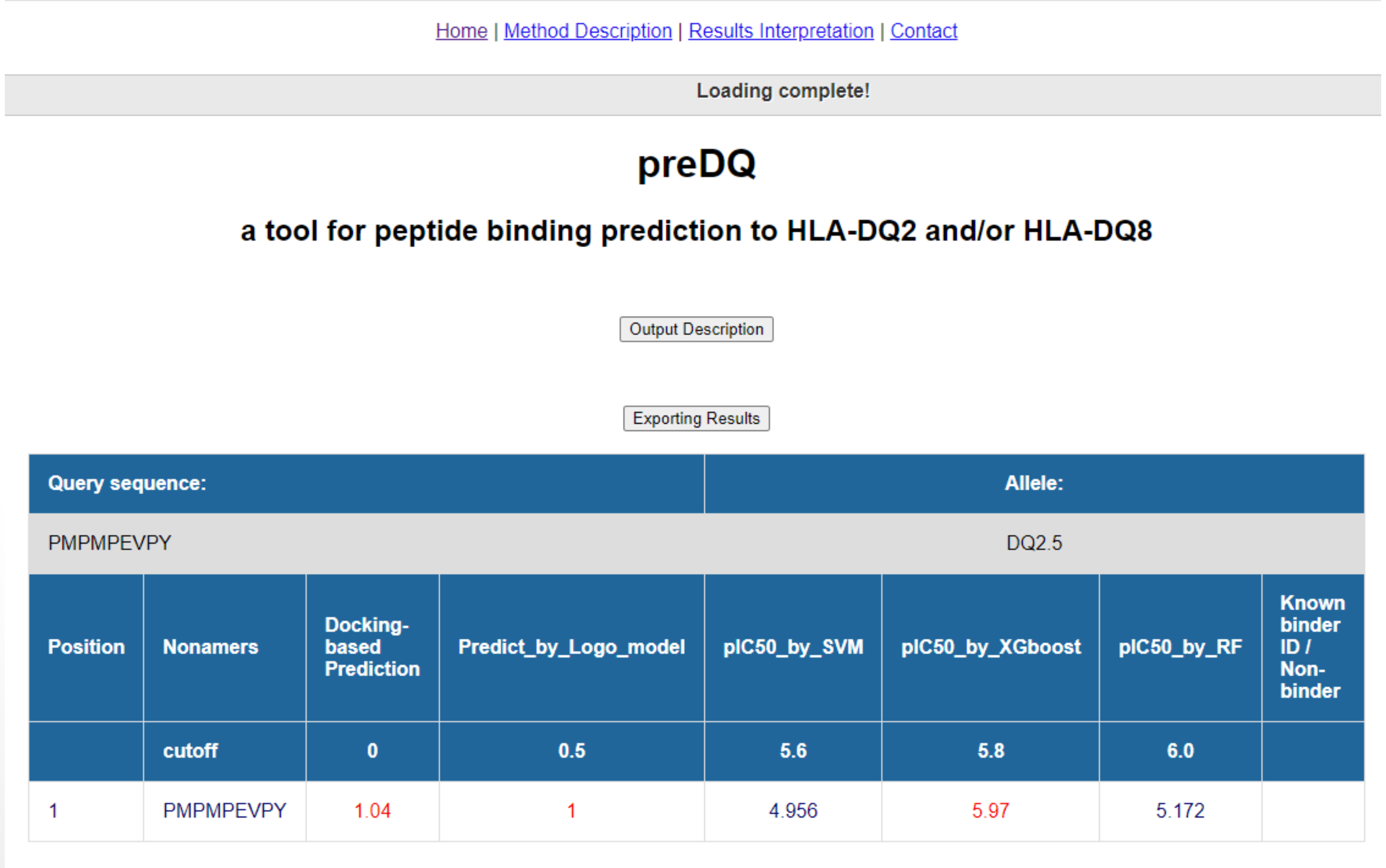
Query sequence:	Allele:
PMGMPELPY	DQ2.5

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	pIC50_by_SVM	pIC50_by_XGboost	pIC50_by_RF	Known binder ID / Non-binder
	cutoff	0	0.5	5.6	5.8	6.0	
1	PMGMPELPY	0.956	1	5.132	6.488	5.374	

Search Result Observations

Answer:

- PMPMPE**V**PY – 3/5 predict: binder; majority voting: probable binder



Home | Method Description | Results Interpretation | Contact
Loading complete!

preDQ

a tool for peptide binding prediction to HLA-DQ2 and/or HLA-DQ8

Output Description
Exporting Results

Query sequence:	Allele:
PMPMPEVPY	DQ2.5

Position	Nonamers	Docking-based Prediction	Predict_by_Logo_model	pIC50_by_SVM	pIC50_by_XGboost	pIC50_by_RF	Known binder ID / Non-binder
	cutoff	0	0.5	5.6	5.8	6.0	
1	PMPMPEVPY	1.04	1	4.956	5.97	5.172	

Structure-based models

Model	Cutoff	Sensitivity % true positives/ all positives	Specificity % true negatives/ all negatives	Accuracy % (true positives + true negatives)/all
DQ2.5 11mers – neg p3	0	92	77	84
DQ8.1 11mers – neg p6	0	94	94	94

Ligand-based models for HLA-DQ2.5

Model	pIC_{50} Cutoff	Sensitivity % true positives/ all positives	Specificity % true negatives/ all negatives	Accuracy % (true positives + true negatives)/all
logo	BS > NBS	90	100	95
Random forest	6.0	80	83	81
SVM	5.6	84	77	80
xgboost	5.8	80	80	80

Ligand-based models for HLA-DQ8.1

Model	pIC$_{50}$ Cutoff	Sensitivity % true positives/ all positives	Specificity % true negatives/ all negatives	Accuracy % (true positives + true negatives)/all
logo	BS > NBS	98	100	99
Random forest	5.6	95	98	96
SVM	5.6	93	97	95
xgboost	5.5	92	94	93

Technical Considerations

Question:

- The model is based on binding affinity calculations made from various models that calculate lipophilicity, steric and electronic properties, etc. and takes into consideration methods like random forest regression, nearest neighbour regression, etc. These models don’t have a clear established role in celiac peptide binding.

Answer:

- Machine learning methods are widely used for peptide-protein and ligand-protein binding predictions in bioinformatics (Wikberg JES. Introduction to Pharmaceutical Bioinformatics. Oakleaf Academic, Sweden, 2020).
- Even more, there is a branch in bioinformatics, named immunoinformatics dealing with peptide binding predictions to HLA proteins.

Technical Considerations

Question:

- It is difficult to understand how the results of the preDQ tool should be interpreted. For example, there is no reference to in vitro or in vivo validation if a peptide is predicted as binding to the HLA. Furthermore, the model predicts HLA binding, but does not take into consideration the role of the T-cell receptor (TCR), that binds to the peptide/HLA complex to trigger celiac disease.

Answer:

- In the Results table, references for known binders and non-binders are included (if available) in the last column.
- A new tab “Results Interpretation” is added on the Results page and it is explained with examples how to interpret the obtained results.

Technical Considerations

Question:

- There is minimal documentation regarding the tool. A full description of machine learning (ML), the docking methodology, details on data used for ML training such as a database version, sources of LC mass data from literature etc. preferably in the form of a paper published in a peer-reviewed journal publication are needed. This publication will support transparency and will facilitate understanding of the results returned by the tool.

Answer:

- A set of several papers on the models implemented in the tool will be published soon in journals with an open access. The pdf files will be uploaded to the tool.

Technical Considerations

Question:

- CropLife Europe wishes to stress the importance of validating bioinformatics tools, especially those used for regulatory purposes. We look forward to meeting with EFSA experts to discuss the preDQ tool in detail.

Answer:

- The tool is a work in progress. During the next years, it will be updated and upgraded at least once a year. Recommendations are welcome. They will be carefully considered and included in the next version of the tool.

User Experience

Question:

- Access to the tool has been intermittent as users were unable to login into the tool.

Answer:

- The tool is hosted by EFSA servers. Please provide us with more details about what happened and when.

User Experience

Question:

- It is unclear what, if any, measures have been implemented to support user/data security.

Answer:

- For each e-service, a controller determines the purposes and means of the personal data handling, if any, and ensures the conformity with Regulation (EU) 2018/1725. https://www.efsa.europa.eu/en/personal-data-protection
- As Docker container technology is used, an isolated environment is created for the user when the application is launched, then when the session is finished, the isolated environment is destroyed.

Technical Considerations

Question:

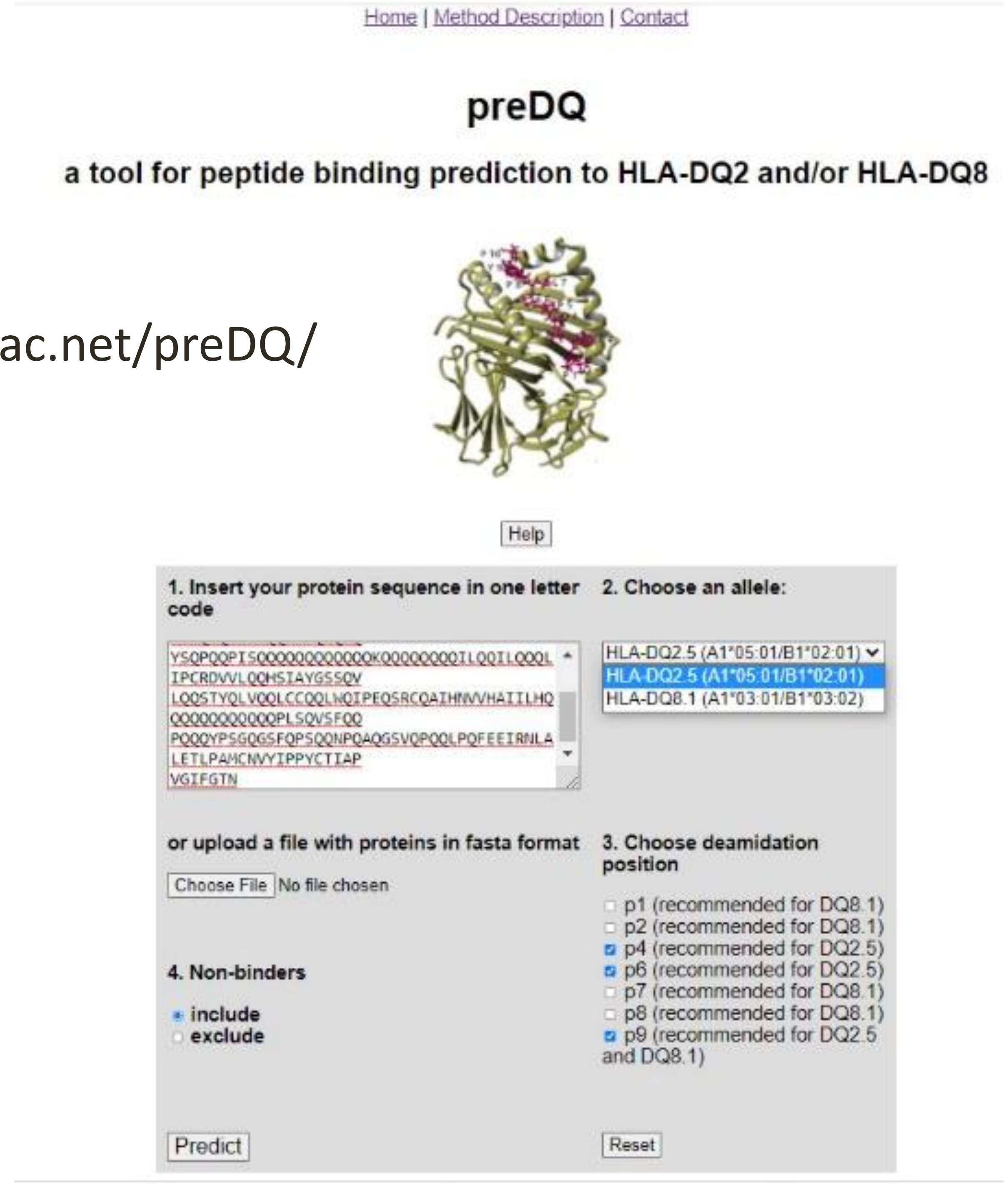
- As mentioned previously, the tool in its current form does not allow high throughput testing of sequences. Programmatic access to the tool is needed to screen internal data for throughput but more importantly to support data privacy. A command line tool available through, e.g., Github, that is fully accessible to the data science community and that can be installed and is pipelined on internal computing resources is needed.

Answer:

- The IP of the tool is held by EFSA. The tool will be distributed free of charge for offline usage under a license agreement.

preDQ

https://ddg-pharmfac.net/preDQ/



preDQ

a tool for peptide binding prediction to HLA-DQ2 and/or HLA-DQ8

1. Insert your protein sequence in one letter code

```
YSQPQQPISQQQQQQQQQQQQKQQQQQQQQILQQILQQQL
IPCRDVVLQQHSIAYGSSQV
LQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILHQ
QQQQQQQQQQQPLSQVSFQQ
PQQQYPSGQGSFQPSQQNPQAQGSVQPQQLPQFEEIRNLA
LETLPAMCNVYIPPYCTIAP
VGIFGTN
```

or upload a file with proteins in fasta format

Choose File No file chosen

2. Choose an allele:

HLA-DQ2.5 (A1*05:01/B1*02:01)

HLA-DQ2.5 (A1*05:01/B1*02:01)

HLA-DQ8.1 (A1*03:01/B1*03:02)

3. Choose deamidation position

- □ p1 (recommended for DQ8.1)
- □ p2 (recommended for DQ8.1)
- ☑ p4 (recommended for DQ2.5)
- ☑ p6 (recommended for DQ2.5)
- □ p7 (recommended for DQ8.1)
- □ p8 (recommended for DQ8.1)
- ☑ p9 (recommended for DQ2.5 and DQ8.1)

4. Non-binders

- ● include
- ○ exclude