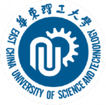
ICLR
International Conference On Learning Representations

Immunogenicity Prediction with Dual Attention Enables Vaccine Target Selection

Song Li *, **Yang Tan** *, **Song Ke**, **Liang Hong**, **Bingxin Zhou** †

SHANGHAI JIAO TONG UNIVERSITY

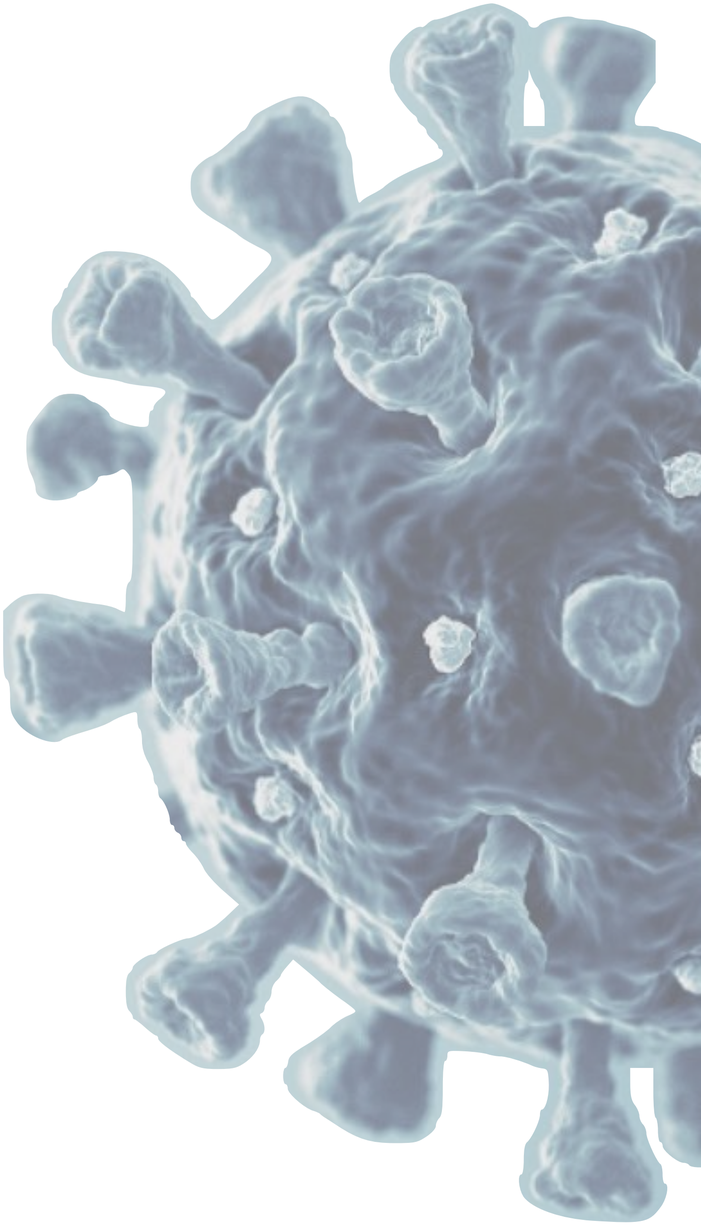


華東理工大學
EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

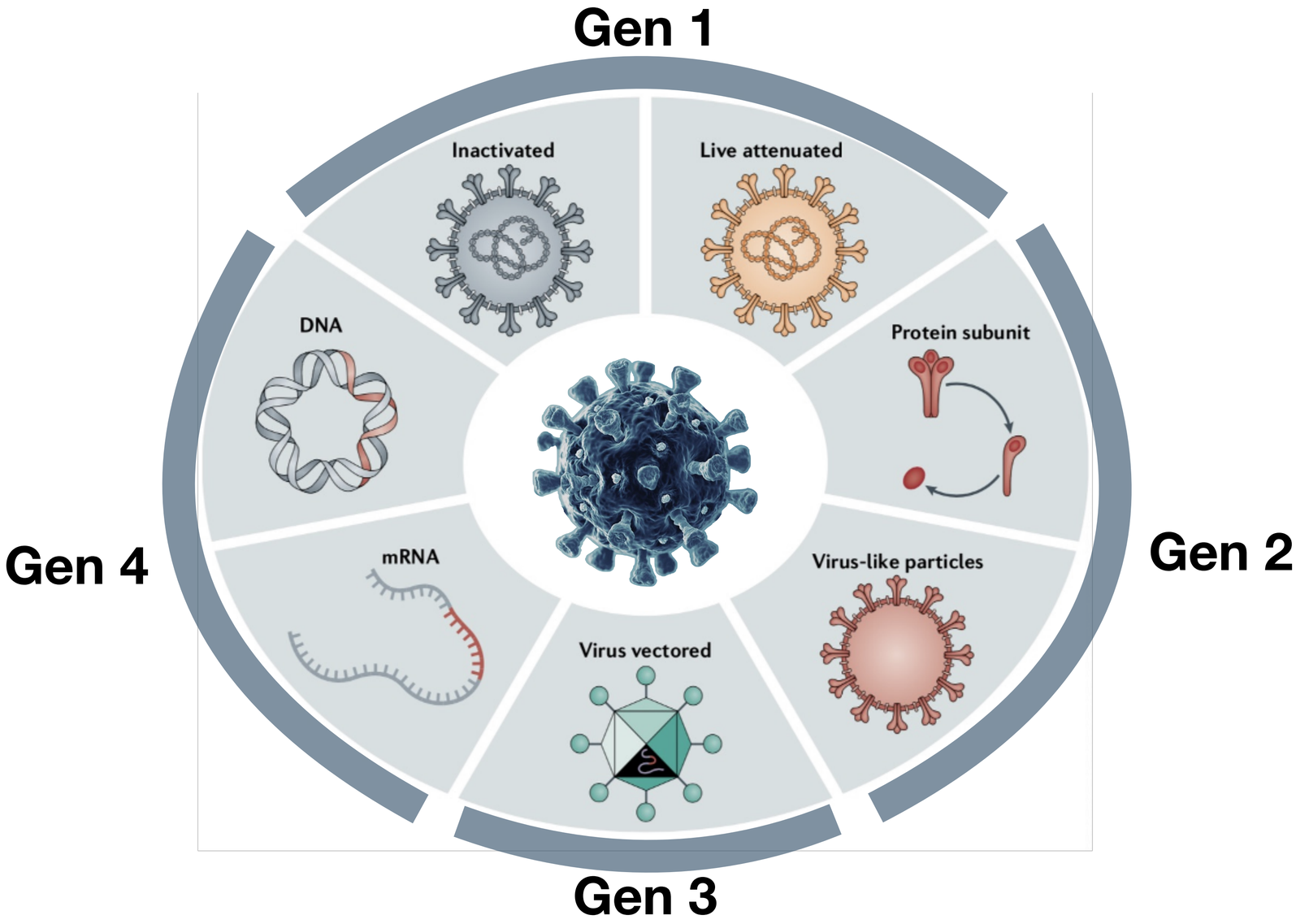
天鹭科技
MATWINGS TECHNOLOGY

ICLR 2025, Singapore

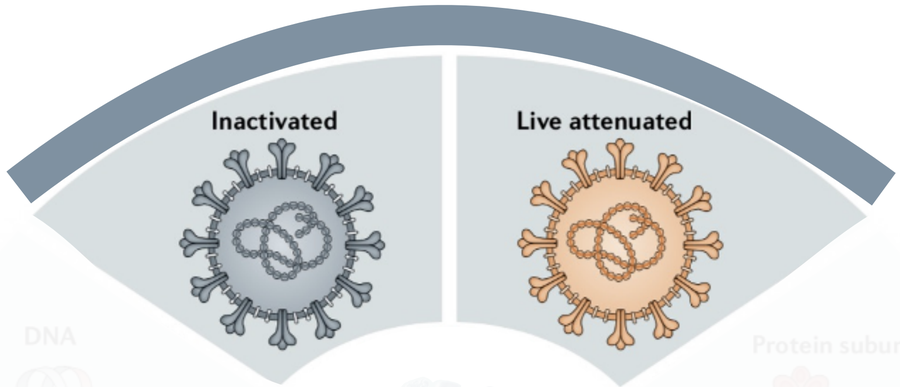
vaccines plays an indispensable role in defending large-scale infectious disease



Four Generations of **Vaccines**



Gen 1



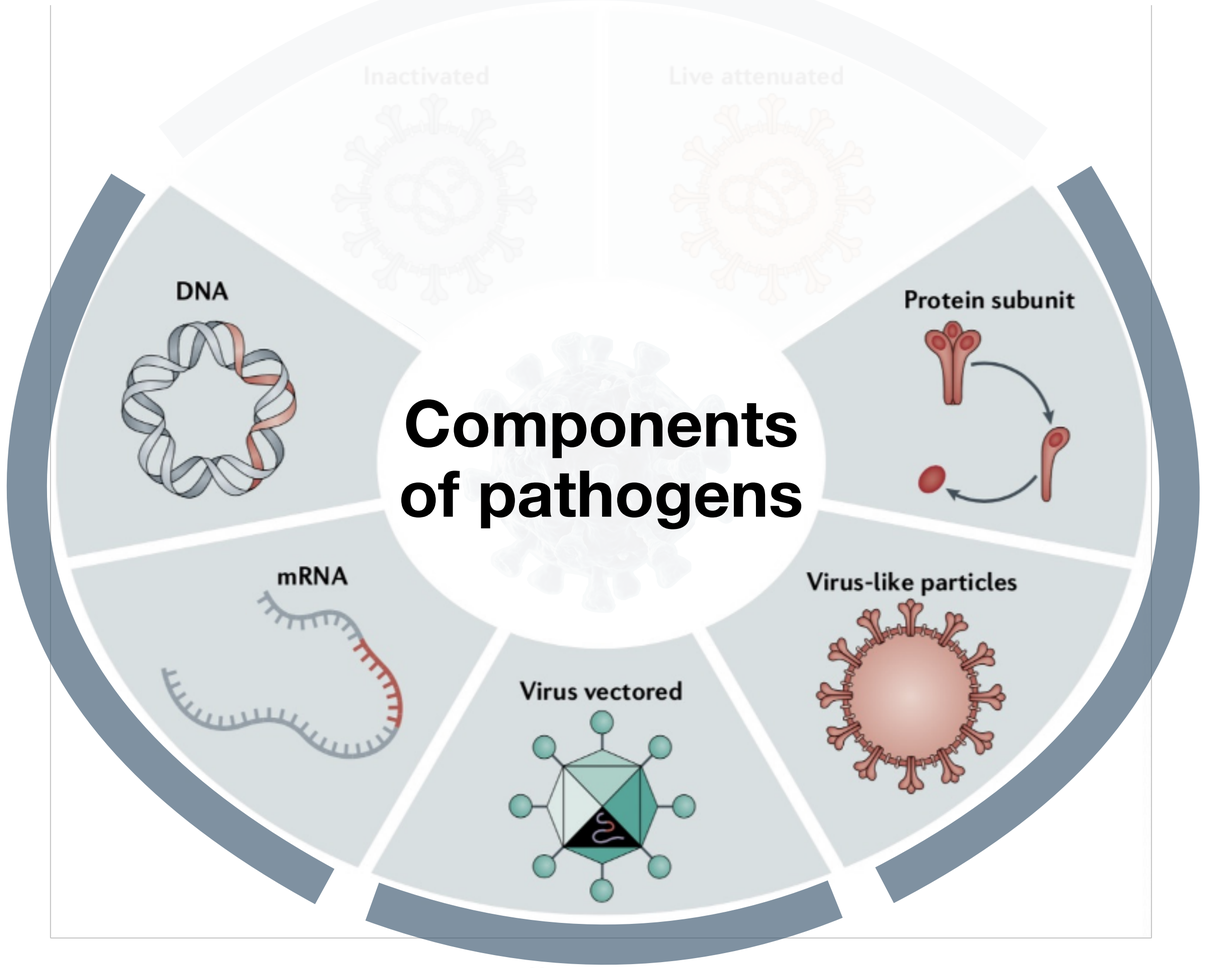
pathogens

Gen 4

Gen 2

Gen 3

Gen 1



Gen 4

Gen 2

Gen 3

Components
of pathogens

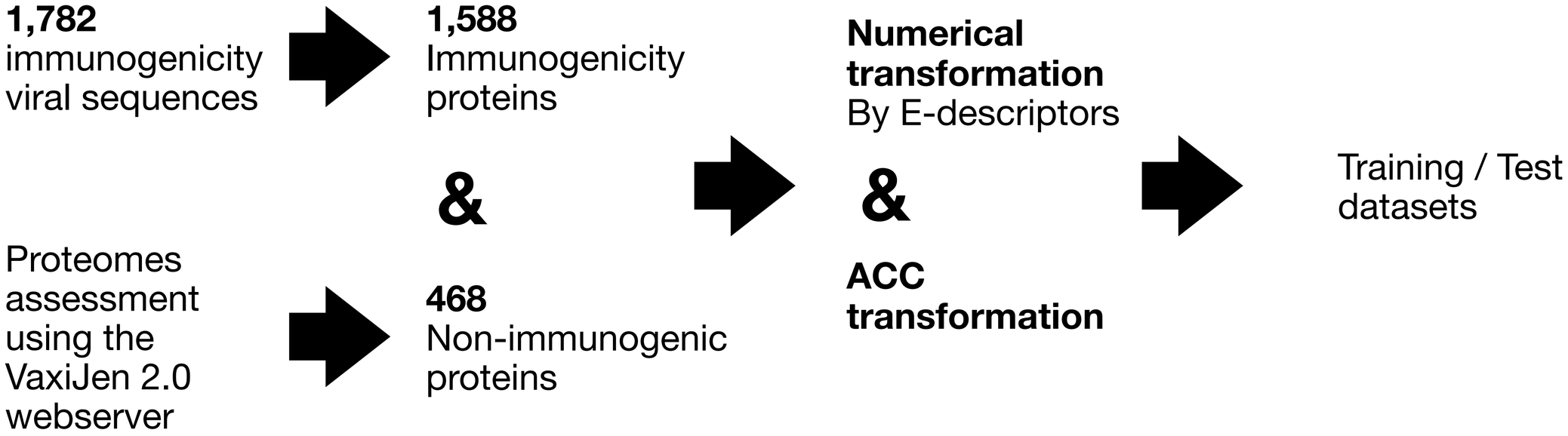
protective antigens

Reverse **Vaccinology**

VaxiJen 3.0

The current SOTA method for immonogenicity prediction

VaxiJen 3.0


1,782
immunogenicity
viral sequences
1,588
Immunogenicity
proteins
&
Proteomes
assessment
using the
VaxiJen 2.0
webserver
468
Non-immunogenic
proteins
Numerical
transformation
By E-descriptors
&
ACC
transformation
Training / Test
datasets

Of course, there are also other machine learning-based methods

In summary:

Method	Model	Feature
	Virus	
VaxiJen3.0	MLP、XGBoost、RandomForest	E-descriptor
VirusImmu	XGBoost、RF、kNN	E-descriptor, Z-descriptor
Vaxi-DL	MLP	biological and physicochemical features
	Bacteria	
VaxiJen3.0	RSM-1NN、XGBoost、RF	E-descriptor
Vaxign-ML	XGBoost	biological and physicochemical features
Vaxi-DL	MLP	biological and physicochemical features
	Tumor	
VaxiJen3.0	SVM, RF, and XGBoost	E-descriptor

Yes, but…

- **Limited Data Volume and Diversity**
- **Manual Feature Extraction with Restricted Information**
- **Insufficient Model Complexity for Capturing Complex Mappings**

And therefore here’s our solution…

- ~~Limited Data Volume and Diversity~~

  Immuno-DB: 3 species, 10,000+ antigen instances

- ~~Manual Feature Extraction with Restricted Information~~

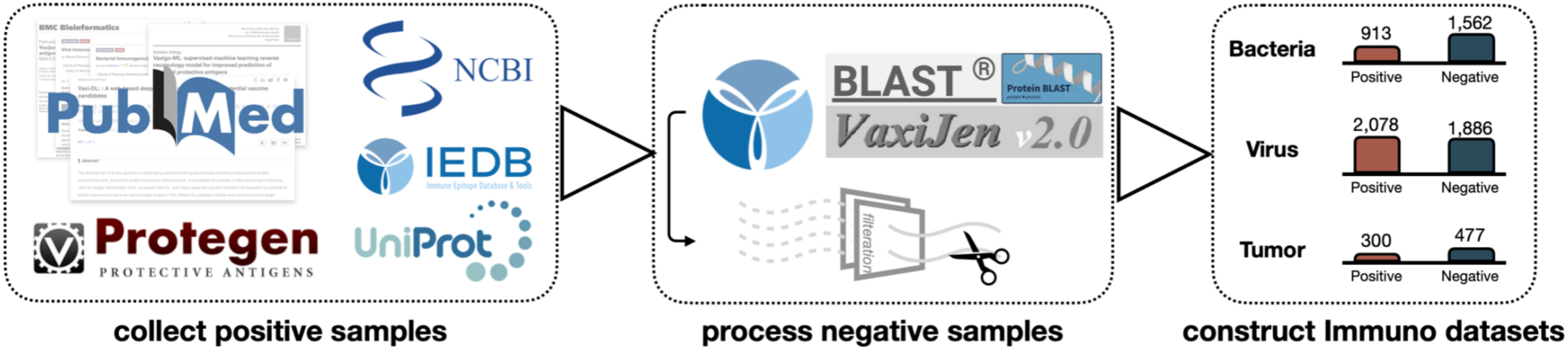
  PLM-based protein representation + expert-guided attributes

- ~~Insufficient Model Complexity for Capturing Complex Mappings~~

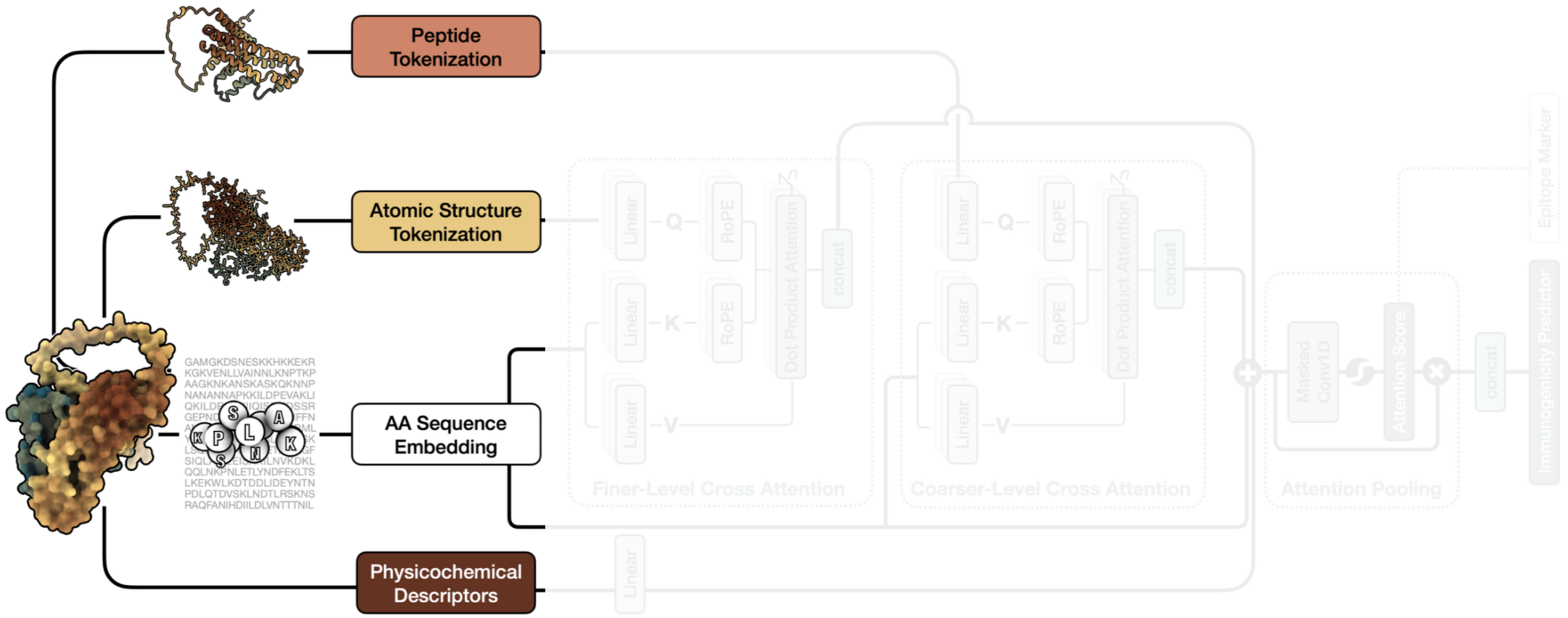
  Dual attention mechanism

- **New benchmarks and evaluation protocols.**

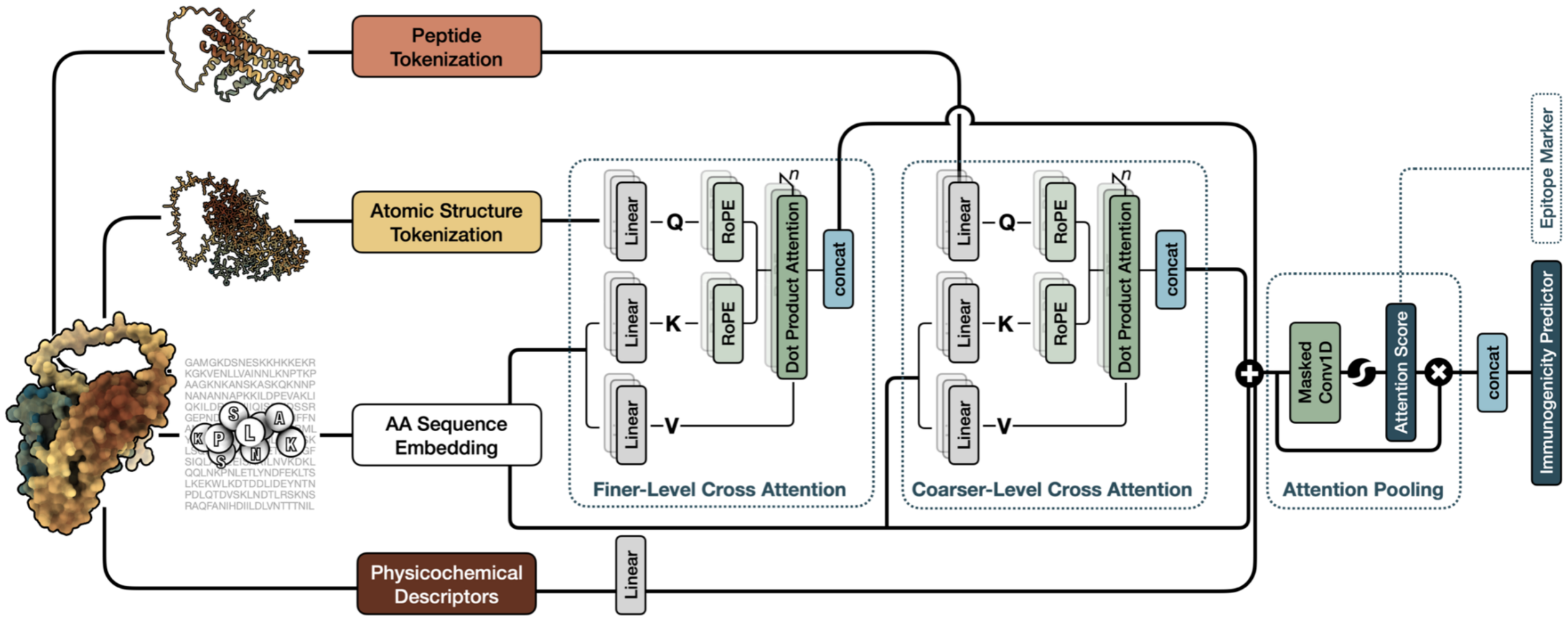
Immuno-DB: 3 species, 10,000+ antigen instances



PLM-based protein representation + expert-guided attributes



Dual attention mechanism



New benchmarks and evaluation protocols

Standard Quantitative Evaluation

Model	Bacteria					Virus					Tumor				
	ACC	T30	MCC	F1	KS	ACC	T30	MCC	F1	KS	ACC	T30	MCC	F1	KS
Random Forest	81.1	**77.7**	57.1	69.7	0.60	88.3	98.7	76.7	88.5	0.80	66.8	**63.5**	27.0	46.4	0.42
Gradient Boosting	80.7	75.9	56.8	71.2	0.62	86.0	97.3	72.3	86.5	0.74	65.4	60.0	26.0	53.1	0.38
XGBoost	80.8	76.1	57.0	71.1	0.61	89.1	**98.8**	78.2	89.2	0.80	69.2	62.2	33.7	56.1	0.40
SGD	78.6	72.4	51.0	65.3	0.56	77.3	88.0	56.0	74.0	0.66	52.6	50.9	29.6	62.2	0.31
Logistic Regression	65.2	67.3	1.8	0.5	0.47	61.4	82.9	35.6	71.9	0.61	60.4	45.2	0.0	0.0	0.25
MLP	78.1	71.6	51.5	68.1	0.57	85.0	90.6	70.5	85.7	0.71	60.4	39.1	0.0	0.0	0.26
SVM	56.7	57.3	18.9	53.0	0.33	61.6	88.7	25.1	67.3	0.53	51.4	61.3	15.2	57.3	0.33
KNN	80.4	73.8	55.3	68.6	0.52	87.4	86.8	74.8	87.3	0.75	57.4	49.6	10.8	45.3	0.12
Vaxi-DL	68.1	55.5	41.7	66.9	0.42	65.3	62.2	33.4	72.9	0.29	54.9	41.3	8.5	47.7	0.10
VaxiJen2.0	75.7	62.2	54.6	72.1	0.57	82.0	74.8	66.6	84.2	0.64	-	-	-	-	-
VaxiJen3.0	**83.3**	58.7	**63.2**	**75.1**	0.63	68.1	62.5	42.3	75.4	0.35	39.0	35.7	0.0	55.9	0.00
VirusImmu	-	-	-	-	-	58.8	59.5	18.1	63.6	0.17	-	-	-	-	-
VENUSVACCINE-Ankh	**82.3**	**78.5**	**60.3**	**73.3**	**0.64**	**92.2**	**99.7**	**84.3**	**92.3**	**0.85**	**76.9**	**73.5**	**55.0**	**73.5**	**0.61**
VENUSVACCINE-ESM2	80.6	77.0	57.0	71.7	**0.66**	**90.3**	**99.5**	**80.6**	**90.6**	**0.82**	**74.0**	61.7	**46.1**	**68.2**	**0.58**
VENUSVACCINE-ProtBert	**84.5**	**84.5**	**65.9**	**77.0**	**0.66**	**91.4**	97.4	**82.8**	**91.2**	**0.84**	**71.5**	**68.7**	**44.7**	**67.6**	**0.54**

ours

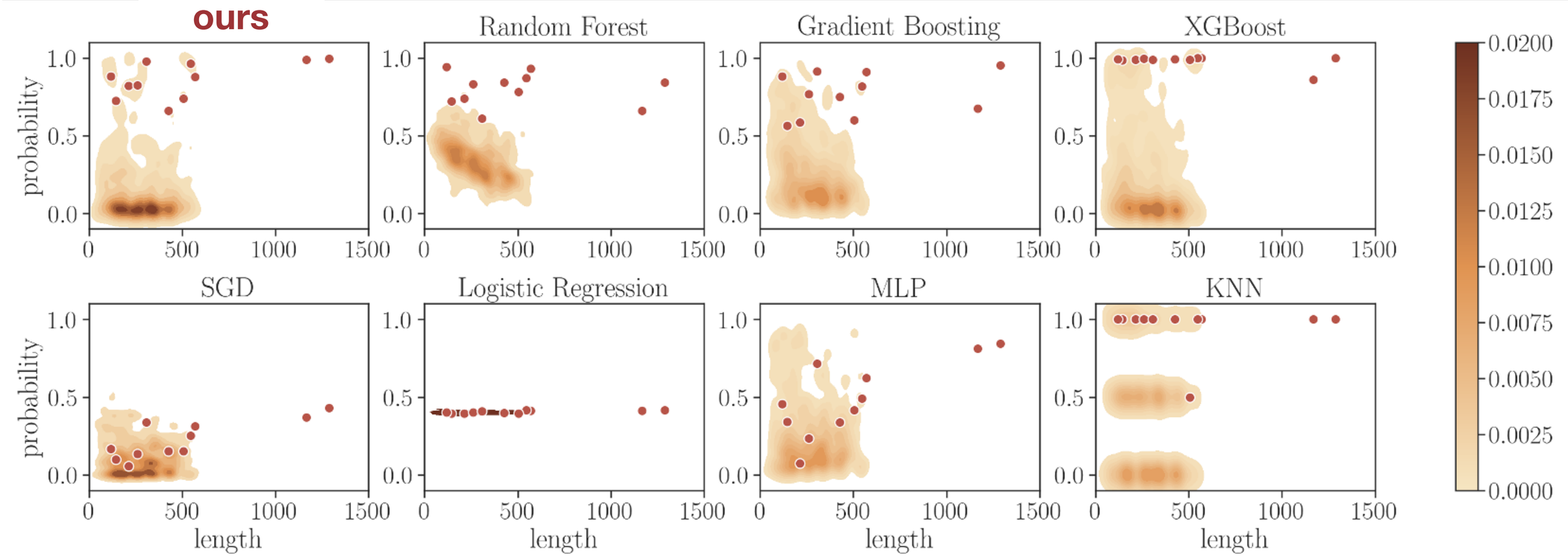
† The top three are highlighted by **First**, **Second**, **Third**.

VenusVaccine achieves SOTA performance on various benchmark datasets with different metrics

(Check more Results in our paper)

New benchmarks and evaluation protocols

Empirical Evaluation 1 - Screening: Helicobacter pylori

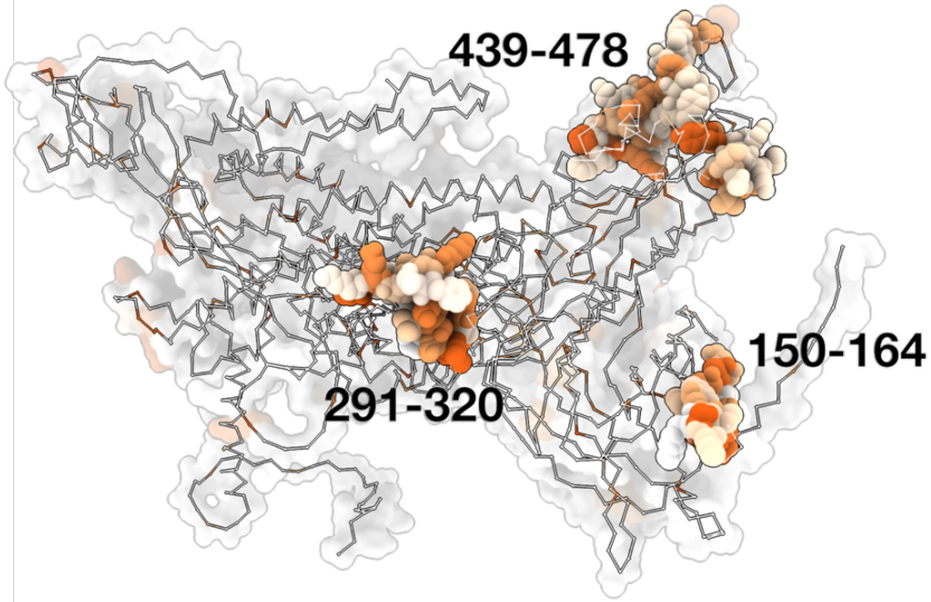


VenusVaccine effectively identifies all 11 PVCs from candidate immunogens

New benchmarks and evaluation protocols

Empirical Evaluation 2 - Prediction: SARS-CoV-2

NCBI ID	Protein Name	predicted label	predicted probability
YP_009724390.1	surface glycoprotein	1	1.000
YP_009742617.1	nsp10	1	1.000
YP_009742616.1	nsp9	1	1.000
YP_009742609.1	nsp2	1	0.999
YP_009742612.1	3C-like proteinase	1	0.998
YP_009725312.1	nsp11	1	0.997
YP_009724396.1	ORF8 protein	1	0.997
YP_009742610.1	nsp3	1	0.997
YP_009724397.2	nucleocapsid phosphoprotein	1	0.995
YP_009742614.1	nsp7	1	0.994
YP_009724395.1	ORF7a protein	1	0.941
YP_009725307.1	RNA-dependent RNA polymerase	1	0.864
YP_009725310.1	endoRNAse	1	0.768
YP_009742611.1	nsp4	0	0.494
YP_009724391.1	ORF3a protein	0	0.463
YP_009742608.1	leader protein	0	0.150
YP_009725308.1	helicase	0	0.140
YP_009742613.1	nsp6	0	0.107
YP_009742615.1	nsp8	0	0.047
YP_009725309.1	3'-to-5' exonuclease	0	0.022
YP_009725311.1	2'-O-ribose methyltransferase	0	0.000
YP_009724393.1	membrane glycoprotein	0	0.000
YP_009724392.1	envelope protein	0	0.000
YP_009724394.1	ORF6 protein	0	0.000
YP_009725318.1	ORF7b	0	0.000
YP_009725255.1	ORF10 protein	0	0.000



VenusVaccine ranks the most important spike protein at the top and marks the vaccine targets by the attention score

Further Details About the Paper

Contact: bingxin.zhou@sjtu.edu.cn

OpenReview

openreview.net/forum?id=hWmwL9gizZ

GitHub

github.com/songleee/VenusVaccine

More Awesome AI4Protein Projects and Tools

github.com/ai4protein/VenusFactory

github.com/tyang816