写作技巧分享

Center for Informational Biology, UESTC, China, Chengdu

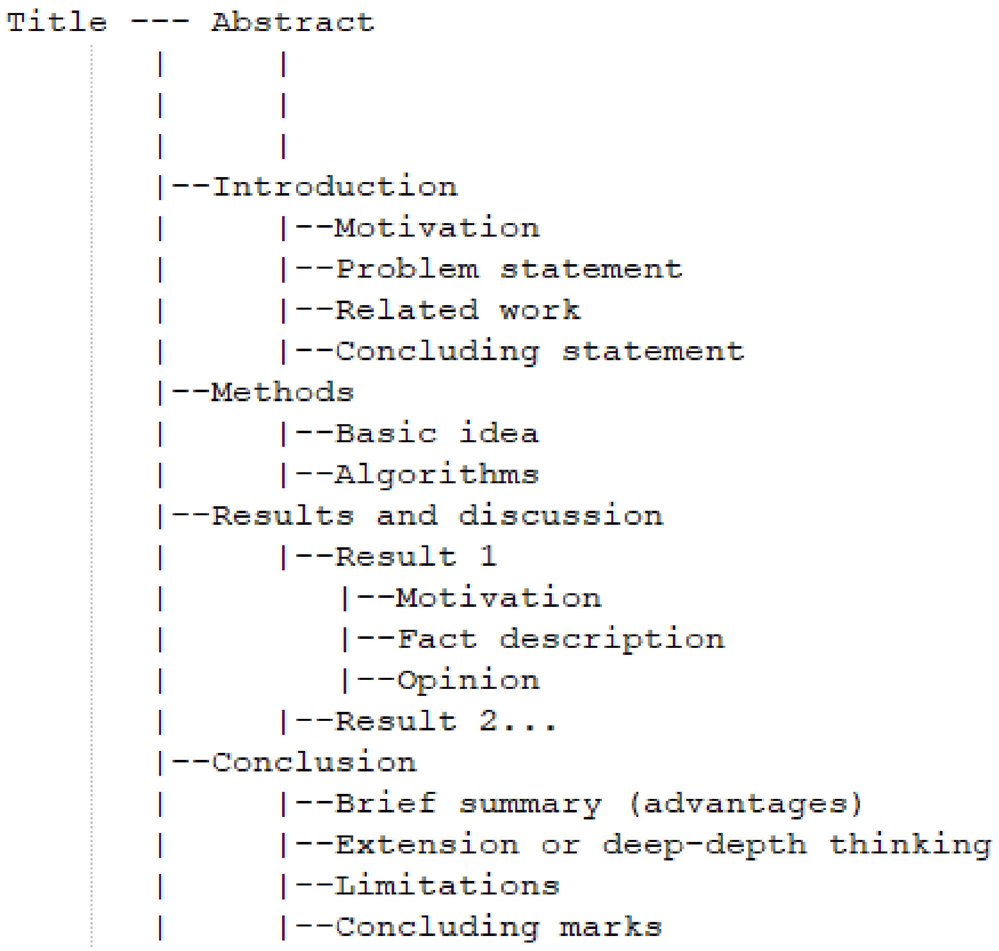
吕 昊

2021-07-18

-- Research article

-- Review

组织全文结构



```
Title --- Abstract
       |     |
       |     |
       |     |
       |--Introduction
       |     |--Motivation
       |     |--Problem statement
       |     |--Related work
       |     |--Concluding statement
       |--Methods
       |     |--Basic idea
       |     |--Algorithms
       |--Results and discussion
       |     |--Result 1
       |        |--Motivation
       |        |--Fact description
       |        |--Opinion
       |     |--Result 2...
       |--Conclusion
       |     |--Brief summary (advantages)
       |     |--Extension or deep-depth thinking
       |     |--Limitations
       |     |--Concluding marks
```

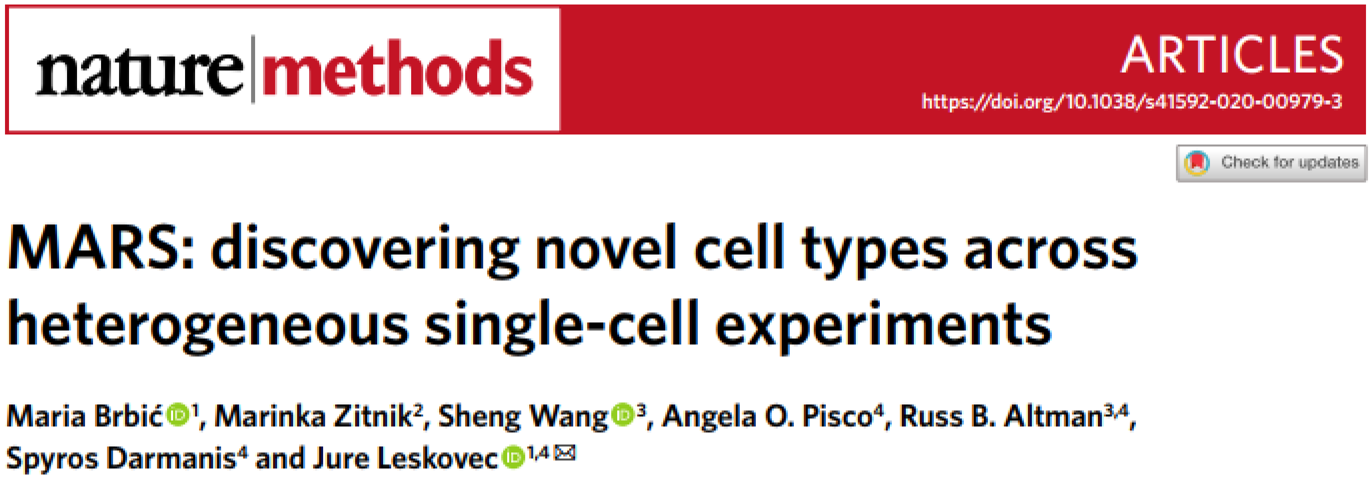
写之前要有整体把握，按照树形结构，由根 (title) 到枝 (paragraphs)，由枝到叶 (sentences)

Abstract 结构

- 问题是什么?
- 为什么选择要做?（重要性)
- 怎么做?
- 结果表现如何?

- ~7句话
- ~150词

举例



nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-00979-3

Check for updates

MARS: discovering novel cell types across heterogeneous single-cell experiments

Maria Brbić[1], Marinka Zitnik[2], Sheng Wang[3], Angela O. Pisco[4], Russ B. Altman[3,4], Spyros Darmanis[4] and Jure Leskovec[1,4]

- **问题是什么？为什么选择要做？(重要性) 1-2句**

-- Although tremendous effort has been put into cell-type annotation (问题), identification of previously uncharacterized cell types in heterogeneous single-cell RNA-seq data remains a challenge (重要性).

- 怎么做? ~3句

-- Here we present MARS, a meta-learning approach（方法）for identifying and annotating known as well as new cell types.

-- MARS overcomes the heterogeneity of cell types by transferring latent cell representations across multiple datasets. (方法亮点部分)

-- MARS uses deep learning to learn a cell embedding function as well as a set of landmarks in the cell embedding space. (方法亮点部分)

- 结果表现如何? ~3句

-- The method has a unique ability to discover cell types that have never been seen before and annotate experiments that are as yet unannotated.

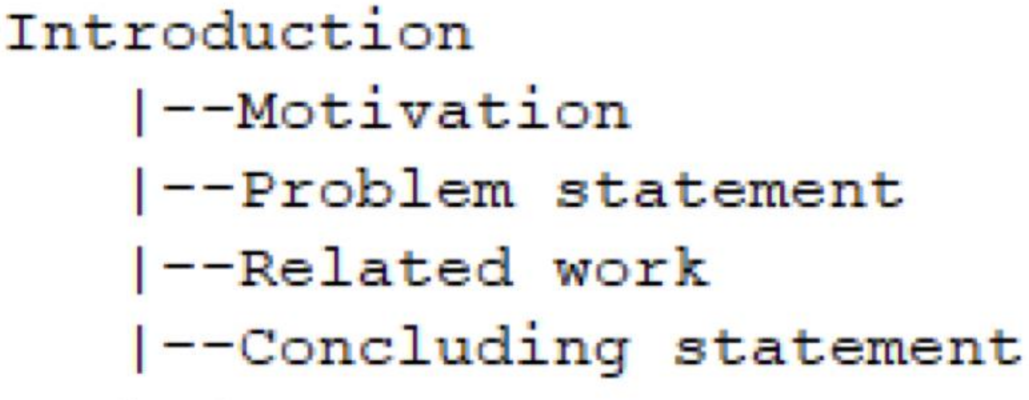
-- We apply MARS to a large mouse cell atlas and show its ability to accurately identify cell types, even when it has never seen them before.

-- Further, MARS automatically generates interpretable names for new cell types by probabilistically defining a cell type in the embedding space. (新的发现)

最后，如果有网站或代码链接，可以再加一句话，eg:

Based on the proposed model, a webserver called xxxxx was established and is freely accessible at http://lin-group.cn/server/xxxxx.

Introduction 结构 3-5段



```
Introduction
     |--Motivation
     |--Problem statement
     |--Related work
     |--Concluding statement
```

第一段 – motivation(动机是什么)

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly transmissible and pathogenic coronavirus that emerged in late 2019 and has caused a pandemic of acute respiratory disease, named ‘coronavirus disease 2019’ (COVID-19), which presents a massive health and socioeconomic crisis. To devise therapeutic strategies to conquer SARS-CoV-2 infection and the associated COVID-19 pathology, it is urgent to develop new drugs and repurpose existing ones to dampen the disease course and reduce the burden of medical institutions. As of 2 October 2020, there were about 405 therapeutic drugs in development for COVID-19, but mostly remain computational without tests in infection models. Comprehensive understanding of the molecular mechanisms of SARS-CoV-2 infection and the changes within the host cell pathways is essential to rationally repurpose drugs.

逻辑性很重要!

第二段 – Problem statement(具体问题是什么)

Proteomics approaches are powerful tools to elucidate mechanisms of pathogenesis by quantifying changes in protein abundance and phosphorylation [6]. For instance, Stukalov et al. characterized interactome, proteome and signaling process in a systems wide manner to study the relationship of SARS-CoV-2 and host cells [7]. Bouhaddou et al. presented a quantitative mass spectrometry-based phosphoproteomics survey of SARS-CoV-2 infection in Vero E6 cells to reveal dramatic rewiring of phosphorylation on host and viral proteins [8]. Klann et al. used a SARS-CoV-2 infection system in Caco-2 human cells to study signaling changes by phosphoproteomics [5]. Hekman et al. performed a quantitative phosphoproteomics survey of SARS-CoV-2 infection in iAT2 cells to exploit the mechanisms driving infection and pathology [9]. The high-throughput Mass Spectrometry techniques used in the above studies can annotate phosphorylation sites accurately, therefore accumulating a large number of phosphorylation examples. However, traditional experimental methods are labor-intensive and time-consuming especially applied in verifying huge amounts of candidate phosphorylation sites. Alternately, as a complementary technique to traditional experimental strategies, the computational approach is a better choice.

第三段 – Related work(总结前人在该问题上的研究)

To date, a considerable number of predictors for identifying phosphorylation sites have been proposed. Most of them show a common strategy that can be summarized as two steps: (i) to encode original sequence based on artificially designed feature extraction method; (ii) to choose an optimized machine learning algorithm for classification and prediction. For example, PhosPred-RF used information theory feature, overlapping property feature, twenty-bit features, twenty-one-bit features, and Skip-n-gram features, trained by random forest-based algorithm for phosphorylation sites prediction [10]. Quokka applied a variety of sequence scoring functions combined with an optimized logistic regression algorithm for the prediction of phosphorylation sites [11]. GPS 5.0 utilized two novel methods named position weight determination and scoring matrix optimization followed by logistic regression algorithm to identify phosphorylation sites [12]. Although features involved in these methods achieved good performance phosphorylation sites predictions, there is limitation of ‘feature engineering’, which requires artificially design that may result in biased features [13].

总结完前人工作后，要提出它们的不足，这个“不足”也是自己工作开展的意义所在

第四段 – Concluding statement(总结性的描述本研究的内容)

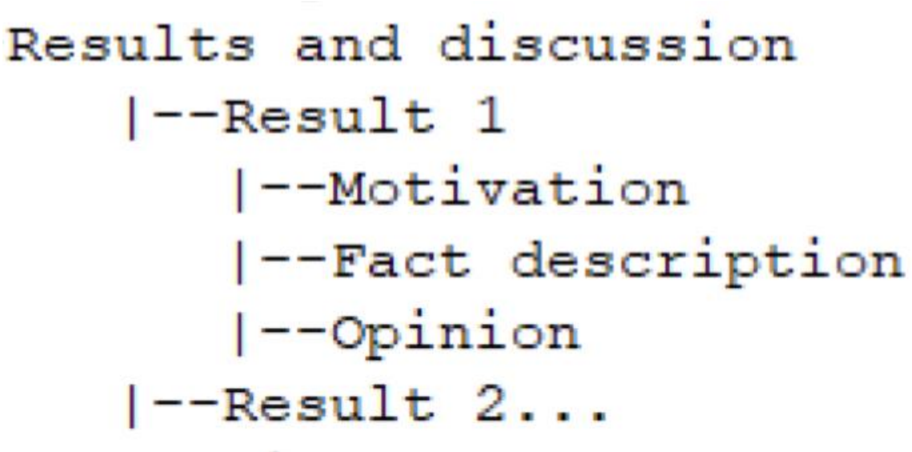
Here, we present a novel CNN-LSTM architecture, DeepIPs, to accurately predict phosphorylation sites in host cells infected with SARS-CoV-2 (**做了什么**) (Figure 1). Different from aforementioned deep-learning methods, DeepIPs uses word embedding approaches in natural language processing to obtain protein sequence representation, which avoids the limitation of 'feature engineering' and effectively improves the performance of the model(**基于什么方法**). To evaluate the performance of DeepIPs, we built different independent datasets to assess the model. The evaluation results reveal that the robust representations generated by word embedding and CNN-LSTM architecture have a strong discriminant power in recognizing general phosphorylation sites (**结果是什么**). We believe that the proposed architecture can also address other bioinformatics problems better than previous methods. In addition, our study provides an early example use-case of popular word embedding methods in biological sequence analysis, and may shed light on other biological prediction problems. (**结论是什么**)

做了什么，基于什么方法，结果是什么，结论是什么

Methods 结构

1. Benchmark dataset construction
2. Basic idea
3. Algorithm statement
4. Architecture and hyperparameters
5. Evaluation metrics

Results and discussion 看图说话



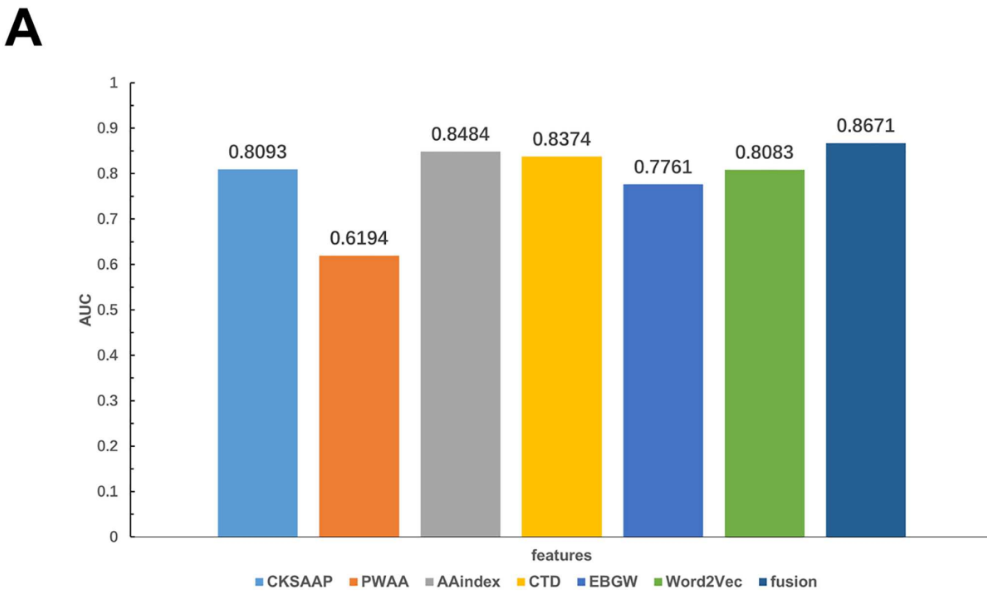
```
Results and discussion
    |--Result 1
        |--Motivation
        |--Fact description
        |--Opinion
    |--Result 2...
```

Motivation： why draw this figure?

Fact description: As shown in figure 1…

Opinion: these results indicate that…

举例

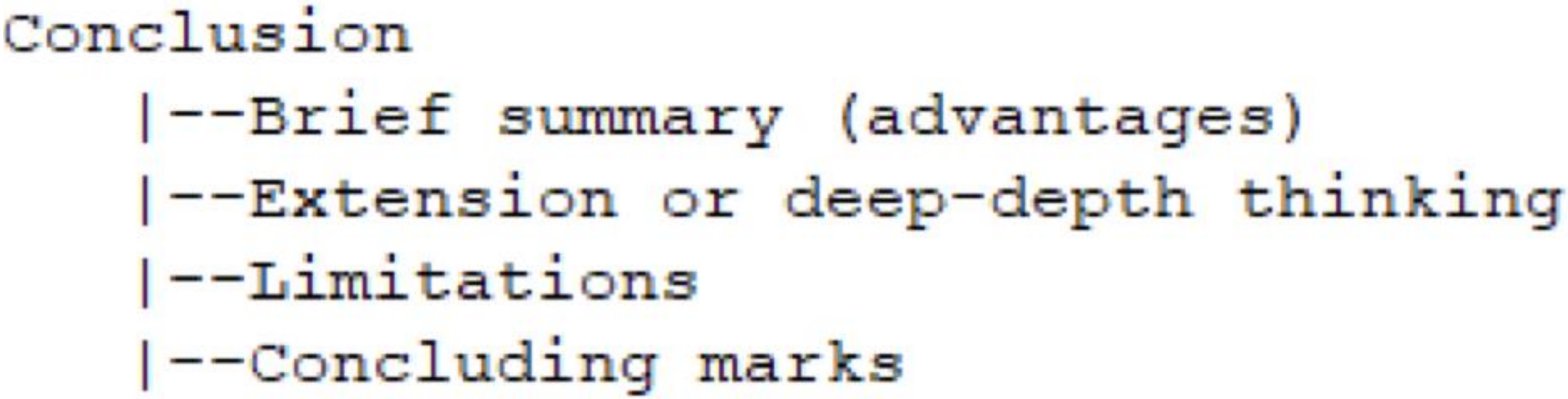


Motivation：why draw this figure? --- To further improve the predictive performance of the models, the extracted features were combined to form a 3184-dimensional feature set.

Fact description: As shown in figure 1… --- As shown in Figure A, the model trained on the fusion feature set achieved better performance (AUC = 0.8671) compared with the models trained on the original feature sets (all AUCs are below 0.8484).

Opinion: these results indicate that… --- indicating that the feature fusion strategy is effective in the prediction of Kcr sites to produce significant performance improvement.

Conclusion



```
Conclusion
    |--Brief summary (advantages)
    |--Extension or deep-depth thinking
    |--Limitations
    |--Concluding marks
```

举例 – brief summary

Phosphorylation is of significance in biological process, which relates to the occurrence of SARS-CoV-2 infection. Due to the limitations of experimental verifying sites that cost time and money, it is very urgent to develop effective computational methods for phosphorylation identification in SARS-CoV-2 infection. Hence, in this study, we propose DeepIPs, which consists of the most popular word embedding methods and CNN-LSTM architecture, to predict phosphorylation sites. The independent test demonstrates that DeepIPs has a better performance than existing phosphorylation sites predictors. Furthermore, a freely accessible web-server called DeepIPs was established.

从动机→方法的提出→方法的结果表现

例如：挖掘磷酸化修饰的信息对阐明新冠发生机制很重要 (动机)，因此需要提出有效的方法；本文提出了DeepIPs (方法的提出)；多组结果显示DeepIPs模型表现很好 (方法的结果表现)

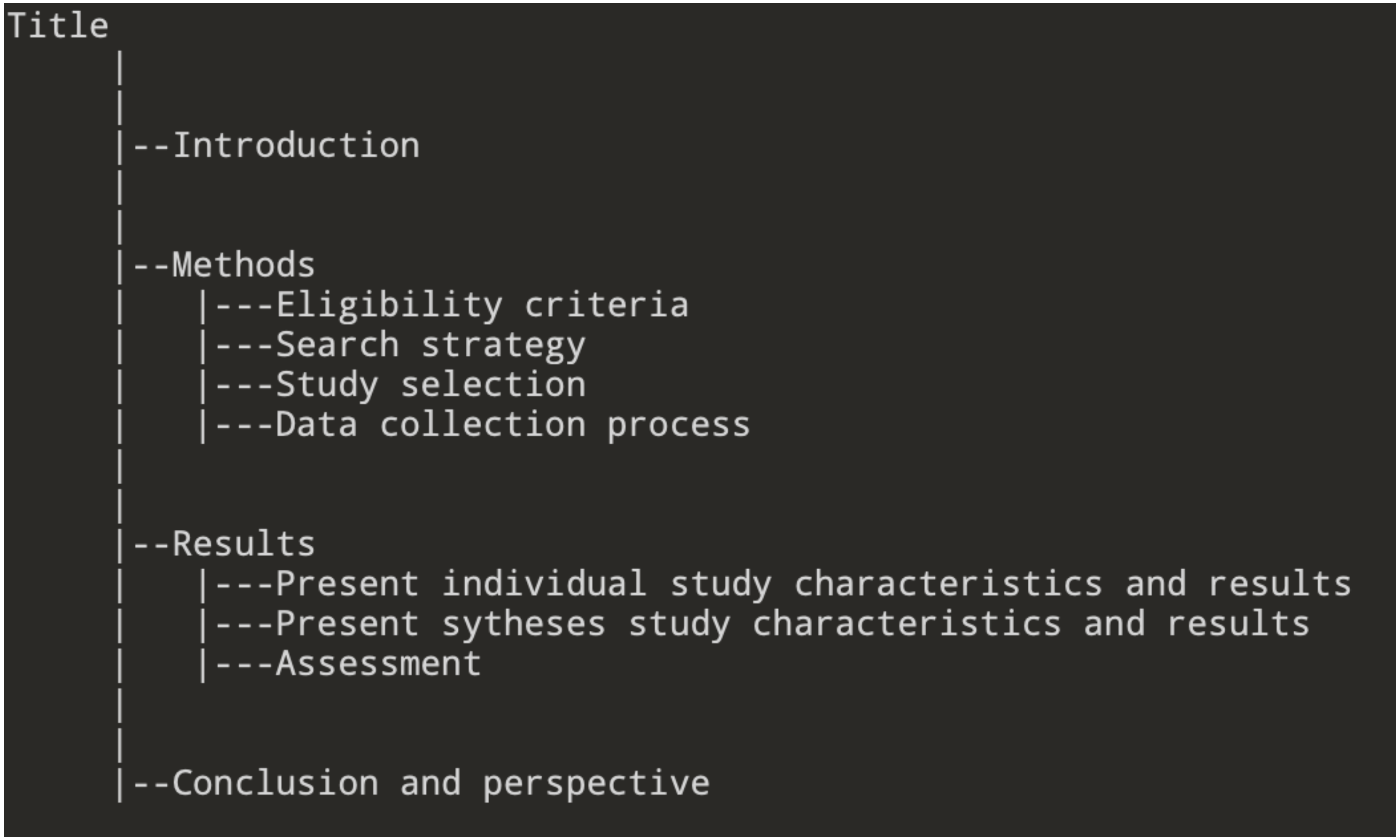
举例 – extension or deep depth thinking

In addition, the following aspects can be further improved in the future. Firstly, the word embedding methods such as SEL, Word2Vec, fastText, and GloVe used in this study are all based on fixed representations of word vectors, which cannot represent the different meanings of word in different contexts. The dynamic word representation methods such as ELMo, GPT, and BERT can extract contextual semantic information based on the words in the context, thus have stronger word representation capabilities. Secondly, the CNN-LSTM architecture we designed cannot explain meaningful biological process well due to ‘black box’ property. Therefore, we will use some interpretable deep-learning algorithms in future works, such as generating adversarial network.

-- Research article

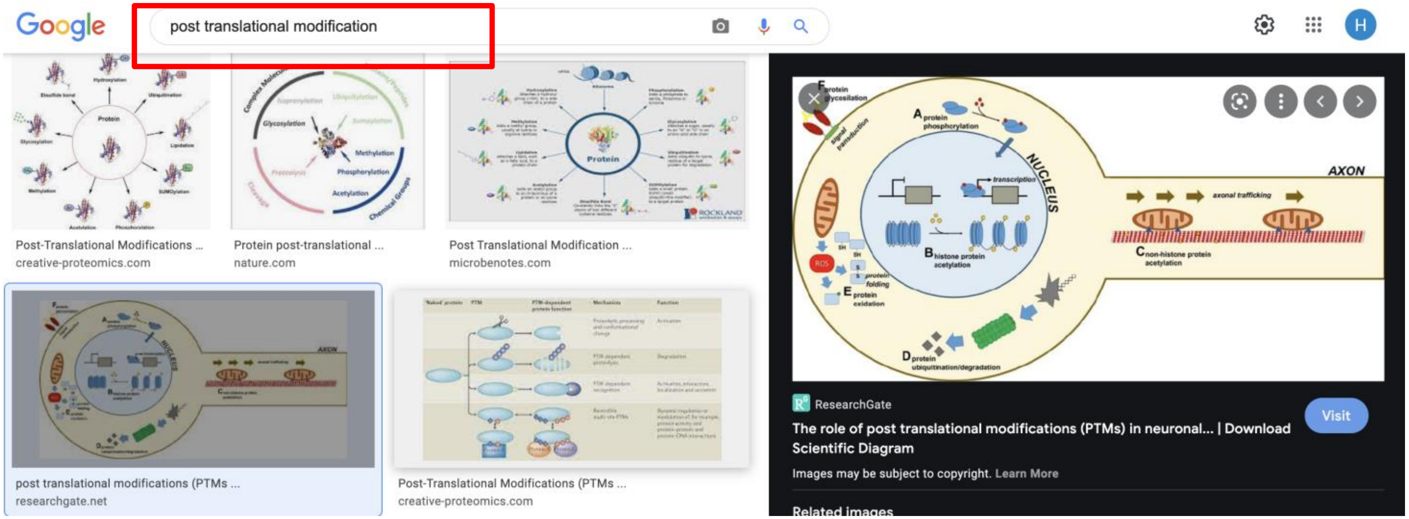
-- Review

结构



```
Title
     |
     |
     |--Introduction
     |
     |
     |--Methods
     |   |---Eligibility criteria
     |   |---Search strategy
     |   |---Study selection
     |   |---Data collection process
     |
     |
     |--Results
     |   |---Present individual study characteristics and results
     |   |---Present sytheses study characteristics and results
     |   |---Assessment
     |
     |
     |--Conclusion and perspective
```

图



post translational modification

Post-Translational Modifications ...
creative-proteomics.com

Protein post-translational ...
nature.com

Post Translational Modification ...
microbenotes.com

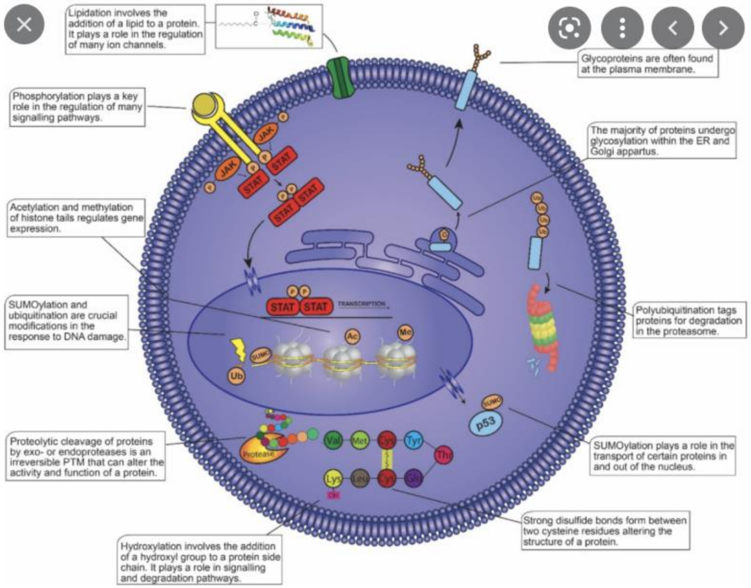
post translational modifications (PTMs ...
researchgate.net

Post-Translational Modifications (PTMs ...
creative-proteomics.com

ResearchGate

The role of post translational modifications (PTMs) in neuronal... | Download Scientific Diagram

Images may be subject to copyright. Learn More



Lipidation involves the addition of a lipid to a protein. It plays a role in the regulation of many ion channels.

Glycoproteins are often found at the plasma membrane.

Phosphorylation plays a key role in the regulation of many signalling pathways.

The majority of proteins undergo glycosylation within the ER and Golgi appartus.

Acetylation and methylation of histone tails regulates gene expression.

SUMOylation and ubiquitination are crucial modifications in the response to DNA damage.

Polyubiquitination tags proteins for degradation in the proteasome.

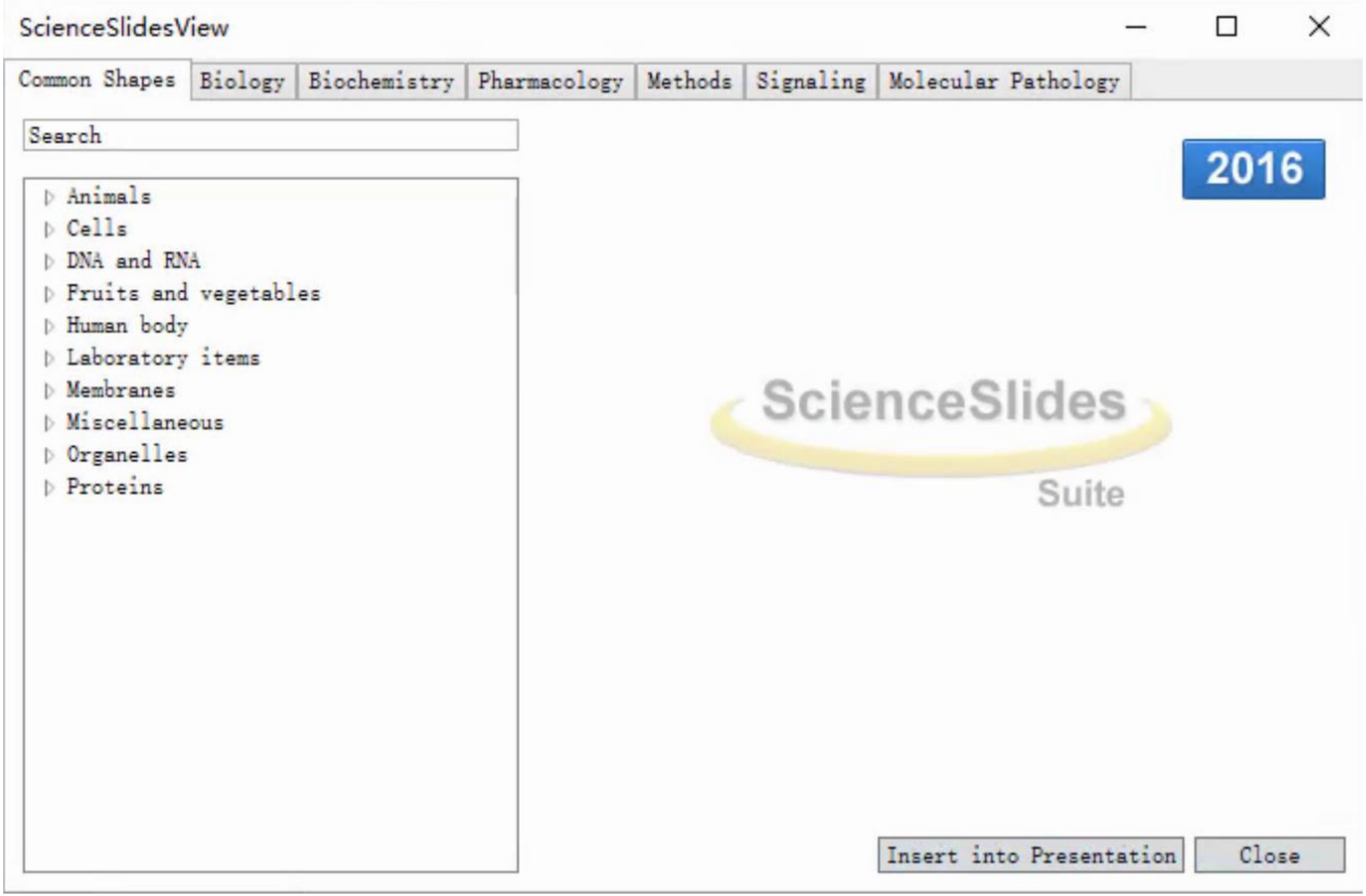
Proteolytic cleavage of proteins by exo- or endoproteases is an irreversible PTM that can alter the activity and function of a protein.

SUMOylation plays a role in the transport of certain proteins in and out of the nucleus.

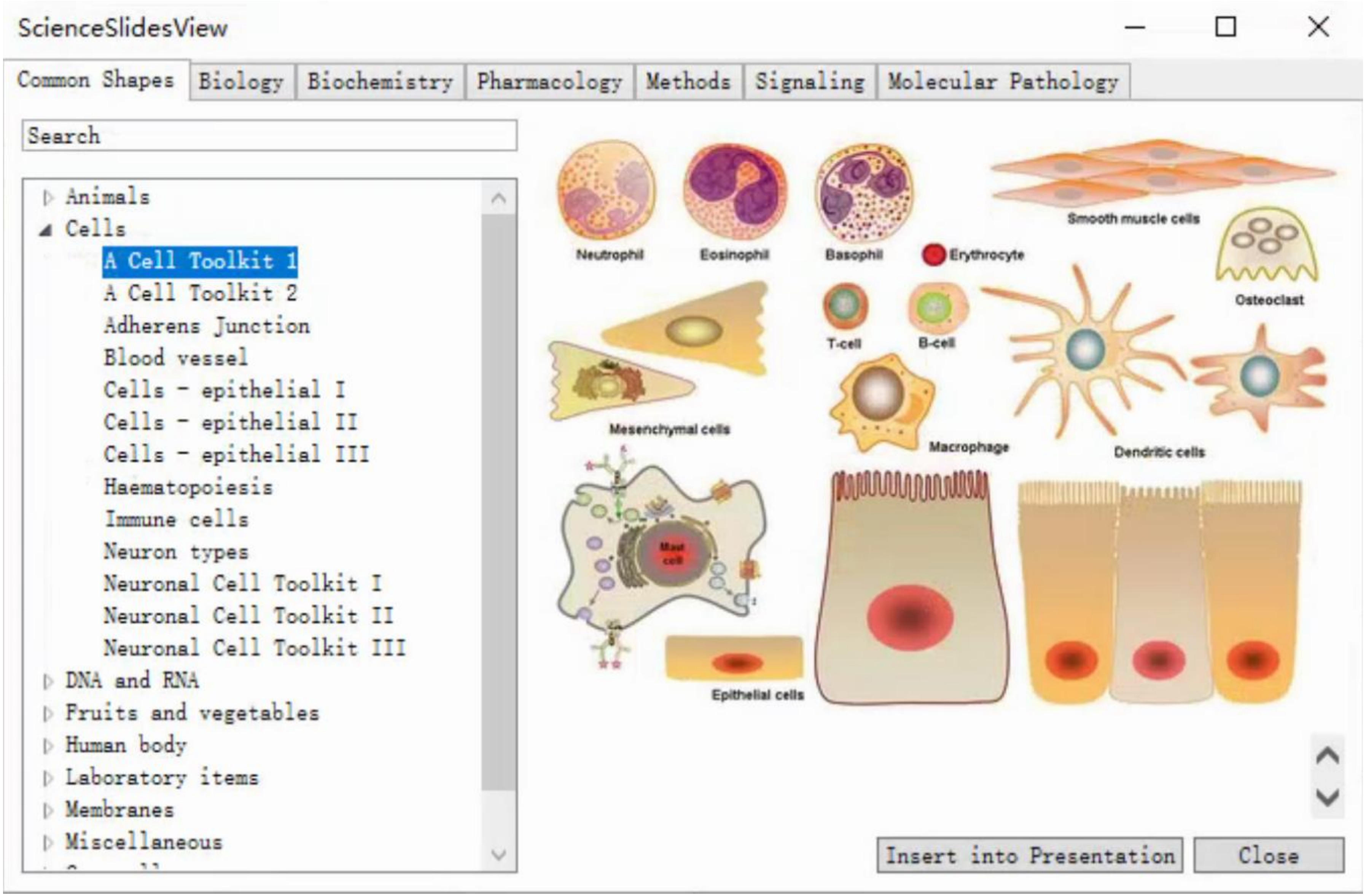
Strong disulfide bonds form between two cysteine residues altering the structure of a protein.

Hydroxylation involves the addition of a hydroxyl group to a protein side chain. It plays a role in signalling and degradation pathways.

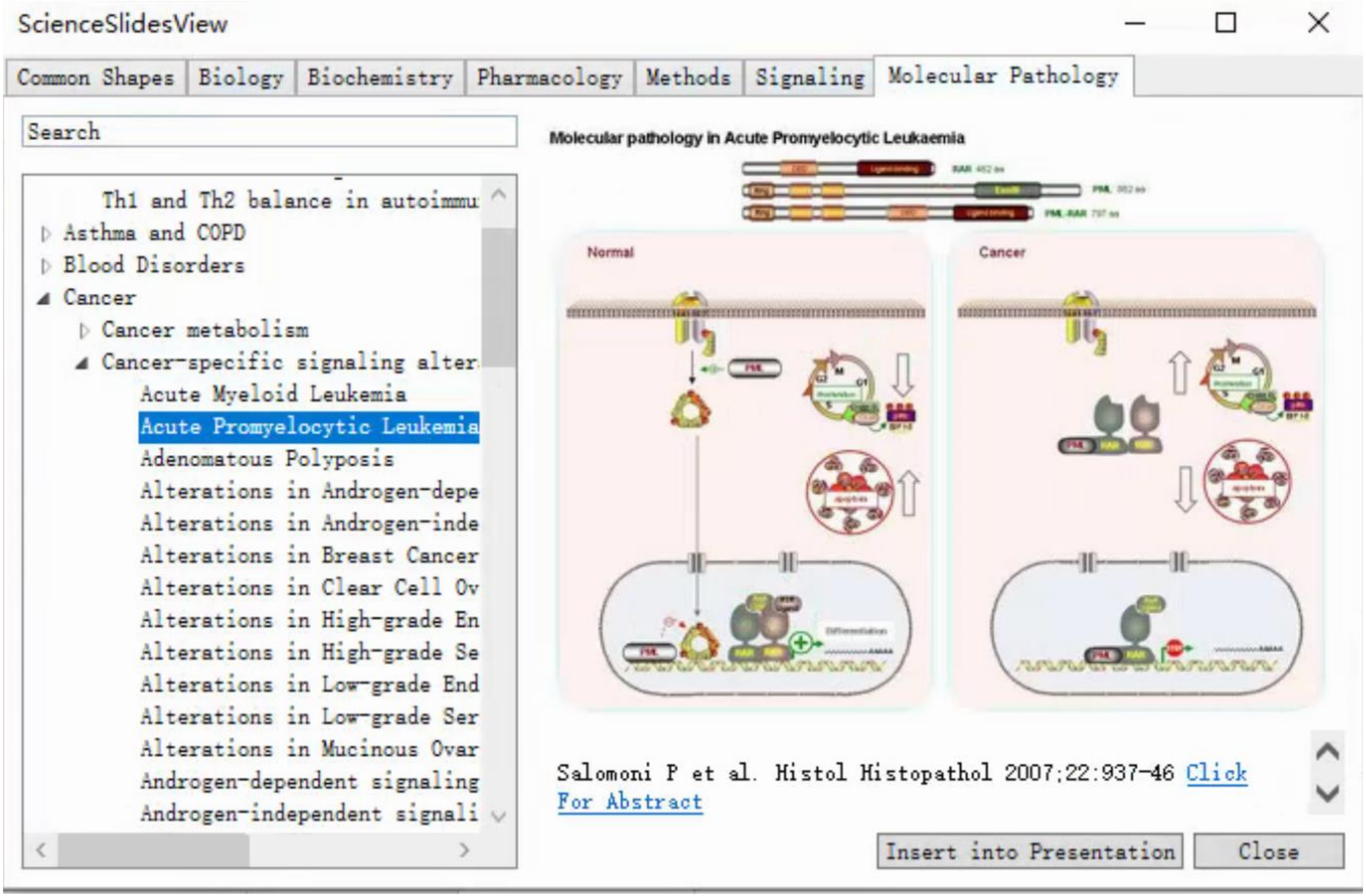
图


ScienceSlidesView
Common Shapes
Biology
Biochemistry
Pharmacology
Methods
Signaling
Molecular Pathology
Search
Animals
Cells
DNA and RNA
Fruits and vegetables
Human body
Laboratory items
Membranes
Miscellaneous
Organelles
Proteins
2016
ScienceSlides
Suite
Insert into Presentation
Close

图


ScienceSlidesView
Common Shapes
Biology
Biochemistry
Pharmacology
Methods
Signaling
Molecular Pathology
Search
Animals
Cells
A Cell Toolkit 1
A Cell Toolkit 2
Adherens Junction
Blood vessel
Cells - epithelial I
Cells - epithelial II
Cells - epithelial III
Haematopoiesis
Immune cells
Neuron types
Neuronal Cell Toolkit I
Neuronal Cell Toolkit II
Neuronal Cell Toolkit III
DNA and RNA
Fruits and vegetables
Human body
Laboratory items
Membranes
Miscellaneous
Neutrophil
Eosinophil
Basophil
Erythrocyte
Smooth muscle cells
Osteoclast
T-cell
B-cell
Mesenchymal cells
Macrophage
Dendritic cells
Mast cell
Epithelial cells
Insert into Presentation
Close

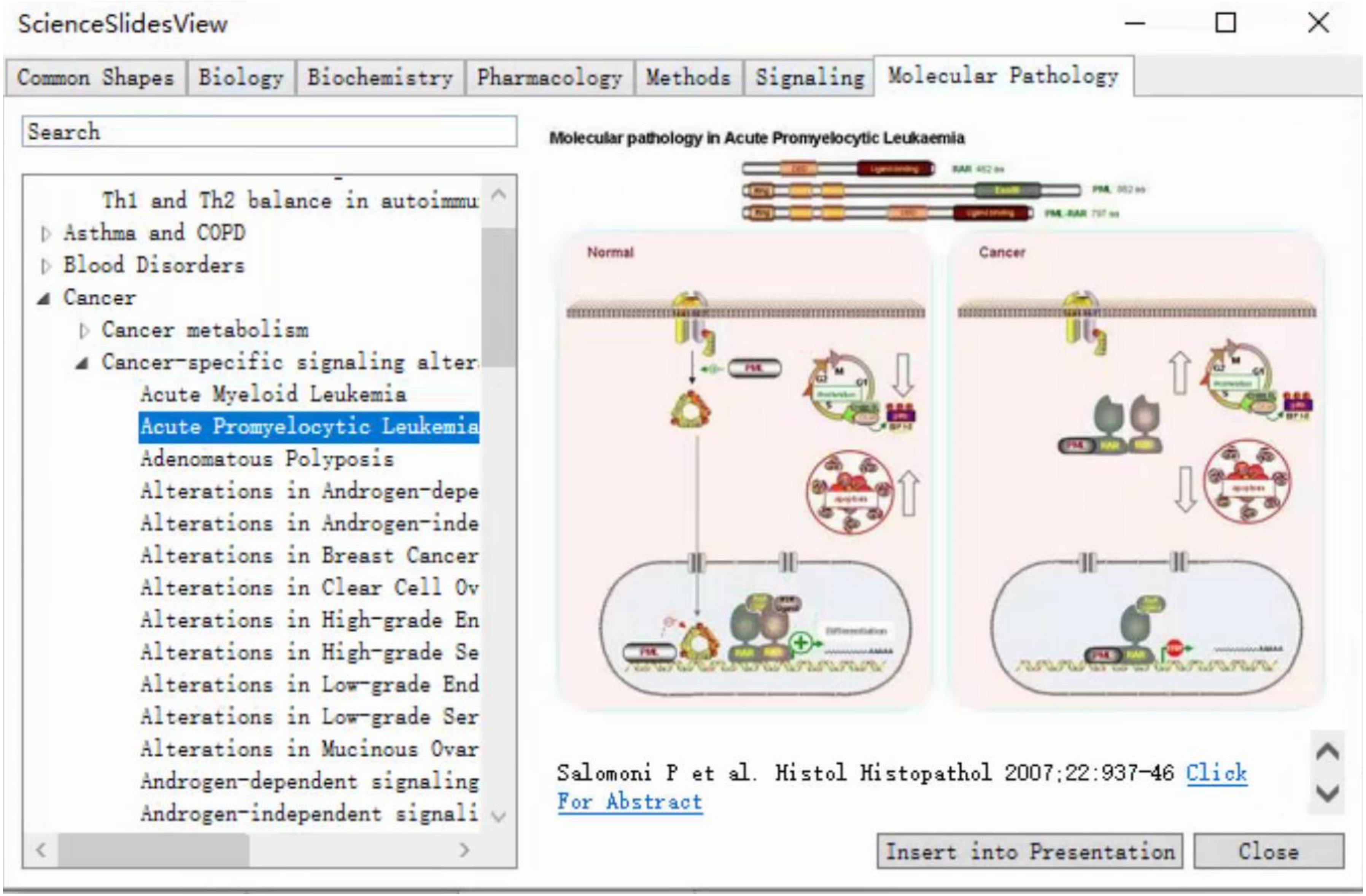
图


ScienceSlidesView
Common Shapes
Biology
Biochemistry
Pharmacology
Methods
Signaling
Molecular Pathology
Search
Th1 and Th2 balance in autoimmu
Asthma and COPD
Blood Disorders
Cancer
Cancer metabolism
Cancer-specific signaling alter
Acute Myeloid Leukemia
Acute Promyelocytic Leukemia
Adenomatous Polyposis
Alterations in Androgen-depe
Alterations in Androgen-inde
Alterations in Breast Cancer
Alterations in Clear Cell Ov
Alterations in High-grade En
Alterations in High-grade Se
Alterations in Low-grade End
Alterations in Low-grade Ser
Alterations in Mucinous Ovar
Androgen-dependent signaling
Androgen-independent signali
RAR
PML
PML-RAR
Normal
Cancer
M
G2
G1
S
PML
RAR
Insert into Presentation
Close

Molecular pathology in Acute Promyelocytic Leukaemia

Salomoni P et al. Histol Histopathol 2007;22:937-46 Click For Abstract

图



ScienceSlidesView

Common Shapes | Biology | Biochemistry | Pharmacology | Methods | Signaling | Molecular Pathology

Search

Th1 and Th2 balance in autoimmu
- Asthma and COPD
- Blood Disorders
- Cancer
  - Cancer metabolism
  - Cancer-specific signaling alter
    - Acute Myeloid Leukemia
    - Acute Promyelocytic Leukemia
    - Adenomatous Polyposis
    - Alterations in Androgen-depe
    - Alterations in Androgen-inde
    - Alterations in Breast Cancer
    - Alterations in Clear Cell Ov
    - Alterations in High-grade En
    - Alterations in High-grade Se
    - Alterations in Low-grade End
    - Alterations in Low-grade Ser
    - Alterations in Mucinous Ovar
    - Androgen-dependent signaling
    - Androgen-independent signali

Molecular pathology in Acute Promyelocytic Leukaemia

Salomoni P et al. Histol Histopathol 2007;22:937-46 Click For Abstract

Insert into Presentation | Close

图

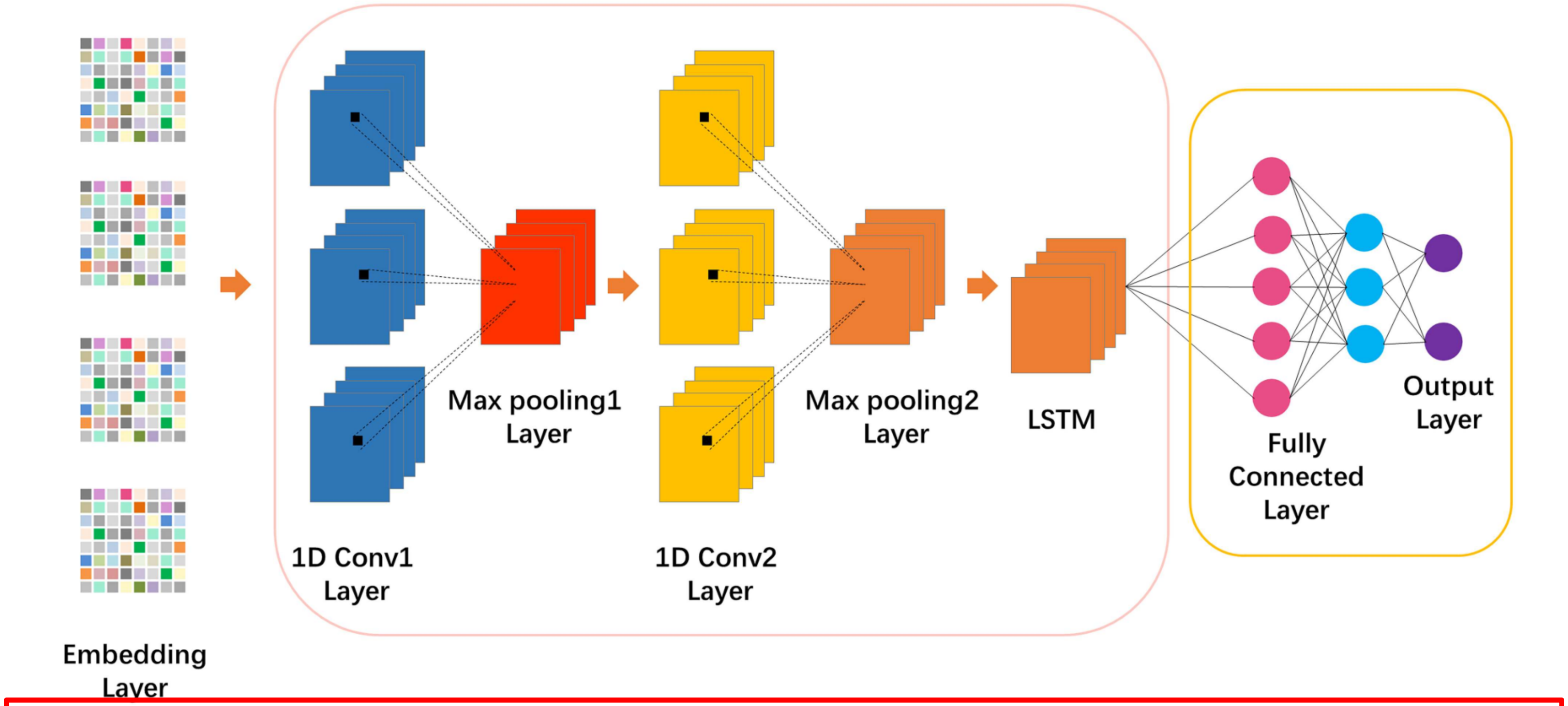


Figure 2. Visualization of the detailed architecture of DeepIPs. The input of DeepIPs is four different word embedding methods. The protein sequences are encoded as vectors that are fed into CNN-LSTM block. The convolution block was used for initial feature extraction and LSTM block was used to further capture the features from convolutional layer. Finally, the output of CNN-LSTM is fed into an additional fully connected layer and a Softmax layer to produce the final output.

表

不同的期刊有不同的要求

三线表

调整至窗口大小

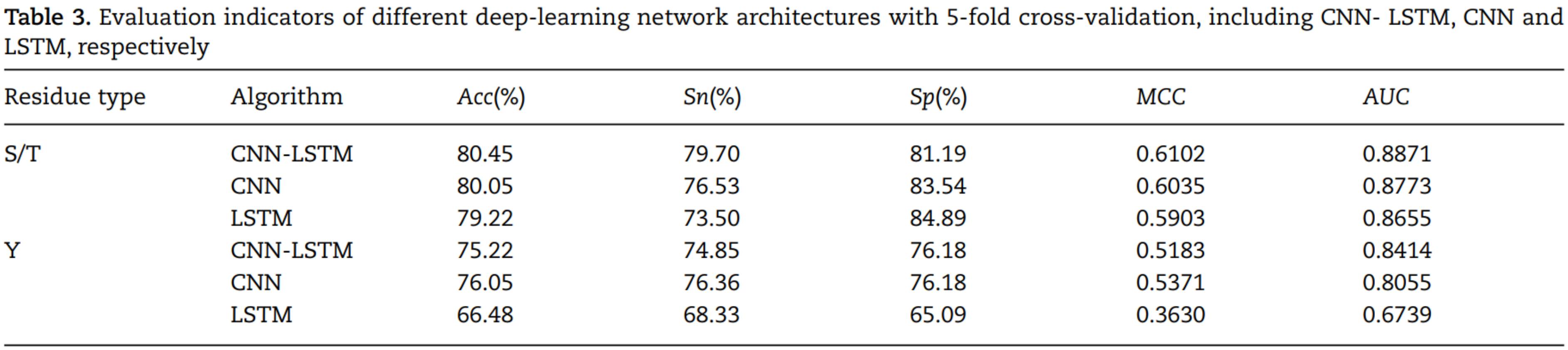
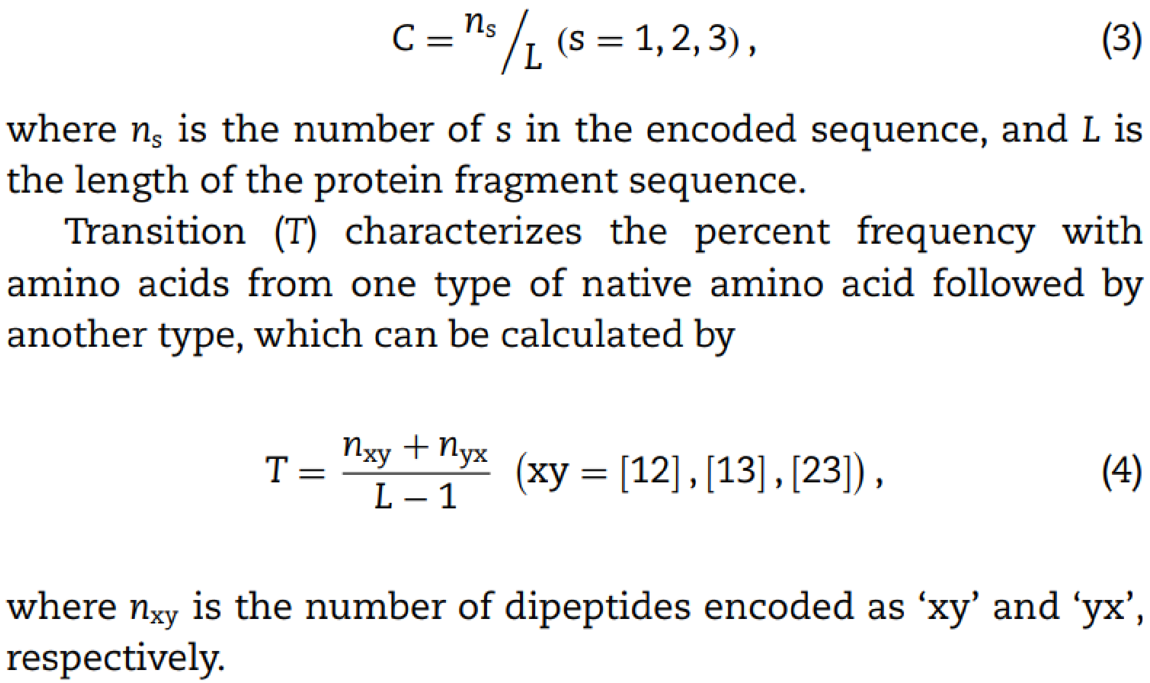


Table 3. Evaluation indicators of different deep-learning network architectures with 5-fold cross-validation, including CNN- LSTM, CNN and LSTM, respectively

Residue type	Algorithm	*Acc*(%)	*Sn*(%)	*Sp*(%)	*MCC*	*AUC*
S/T	CNN-LSTM	80.45	79.70	81.19	0.6102	0.8871
	CNN	80.05	76.53	83.54	0.6035	0.8773
	LSTM	79.22	73.50	84.89	0.5903	0.8655
Y	CNN-LSTM	75.22	74.85	76.18	0.5183	0.8414
	CNN	76.05	76.36	76.18	0.5371	0.8055
	LSTM	66.48	68.33	65.09	0.3630	0.6739

Tips

1. 公式



$$C = {}^{n_s}\!/_{L} \quad (s = 1, 2, 3)\,, \tag{3}$$

where n_s is the number of s in the encoded sequence, and L is the length of the protein fragment sequence.

Transition (T) characterizes the percent frequency with amino acids from one type of native amino acid followed by another type, which can be calculated by

$$T = \frac{n_{xy} + n_{yx}}{L - 1} \quad (xy = [12]\,, [13]\,, [23])\,, \tag{4}$$

where n_{xy} is the number of dipeptides encoded as ‘xy’ and ‘yx’, respectively.

变量斜体

公式标号右对齐

解释公式中符号含义的句子 (where开头的句子)不缩进，且where小写

直到这一part公式写完前，后面都需要加逗号

Tips

2. 正文字体字号：arial/times new roman 10号

3. Figure分辨率：大于300dpi

4. Reference需要包括的内容：作者-题目-期刊名-年份-卷号-页码；如果是preprint或Epub，没有卷号和页码，则加doi号

5. Data availability, supplementary data, author contributions, acknowledgements, fundings

谢谢

THANK YOU