KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(오프라인)

Deep Learning in Bioinformatics

노미나 _ 한양대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 (2.6 월)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	개회사/공지사항전달			
09:30-10:50 (80)	**Best practice for single-cell data analysis**	**박종은 교수**	**Introduction to ML & DNN (이론)**	**이상근 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Practice1: Scanpy basic workflow**	**김우석 김성룡 조교**	**CNN (이론)**	**이상근 교수**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**Public data, batch correction, cell annotation**	**박종은 교수**	**RNN, GAN, XAI (이론)**	**이상근 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**Practice2: Advanced single-cell analysis**	**김우석 김성룡 조교**	**AI 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)**	**이정현 한성민 조교**

DAY2 (2.7 화)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**Introduction to protein structure prediction - Homology modeling - Coevolution-guided modeling Early AI-based approaches**	**백민경 교수**	**Pre-trained Models for Transfer Learning (이론)**	**전민지 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**단백질 구조 예측 실습 - MSA generation, template search - homology modeling contact prediction & modeling**	**백민경 교수**	**Pre-trained Models for Transfer Learning (실습)**	**정민수 조교**
12:10-13:40 (90)	점심			
13:40-15:10 (90)	**AI-based protein structure prediction - AlphaFold/RoseTTAFold Applications to PPI prediction & protein design**	**백민경 교수**	**Deep learning in Bioinformatics**	**노미나 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**단백질 구조 예측 실습 II AlphaFold, RoseTTAFold 실습 및 응용**	**백민경 교수**	**Deep learning model을 이용한 실습**	**곽호진 박예솔 조교**

DAY3 (2.8 수)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**화학정보학 기초(Cheminformatics) 약물특성 및 약물다움(druglikeness) Molecular Notations & Descriptors AI 신약개발을 위한 Databases AI 신약개발을 위한 Programming 기초**	**김동섭 교수**	**마이크로바이옴 기본 이론**	**이선재 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Google Colab에 RDKit 설치 화합물 정보 읽기 실습 Bioactivity database 검색 및 정보 읽기 실습 Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습**	**문채영 나민주 조교**	**16S rRNA amplicon seq. - DADA2**	**서영창 조준우 조교**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**AI 신약개발을 위한 기계학습법 기초 QSAR 모델링 기초 AI 신약개발을 위한 딥러닝 모델 Virtual screening (ligand-based, structure-based) 및 de novo design**	**김동섭 교수**	**최신 메타지놈 분석 기법의 현황**	**이선재 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**QSAR modeling 전체 과정 실습 화합물의 Bioactivity 예측 모델 개발 Virtual screening 과정을 통한 신약후보물질 발굴 실습**	**문채영 나민주 조교**	**Shotgun metagenome 분석 (Linux)**	**서영창 조준우 조교**

강의개요

Deep Learning in Bioinformatics

최근 다양한 생물 종, 인체 부위 및 질환 관련 오믹스 데이터는 점점 더 많이 생산되고 있으며, 중요한 의생명과학의 문제에 대한 해답을 구하고자 활용되고 있다. 신약 개발을 포함한 많은 문제에서 대규모 분자구조를 이용한 기능, 반응성 예측도 매우 중요한 문제이다. 딥러닝은 비선형 변환의 조합을 이용하여 데이터의 높은 수준의 추상화를 통해 특성을 추출함으로써 예측이나 분류와 같은 문제를 해결하고자 한다. 따라서 딥러닝은 의생명과학 분야에서도 좋은 성과를 보이고 있다.

본 강의에서는 대용량 유전체, 분자구조, 텍스트 데이터와 다양한 딥러닝 모델을 이용하여 유전인자나 그들의 기능을 예측하는 방법들을 소개한다.

강의는 다음의 내용을 포함한다:

- 딥러닝 모델 개요
- 시퀀스 데이터를 이용한 유전인자 예측
- 분자구조 데이터를 이용한 기능 예측
- 텍스트 데이터를 이용한 정보 추출

* 참고강의교재:

  Deep Learning (Aaron Courville, Ian Goodfellow, and Yoshua Bengio, 2015)

  Deep Learning for the Life Sciences (Peter Eastman, Patrick Walters, Bharath Ramsundar, Vijay S. Pande, 2019)

* 교육생준비물:

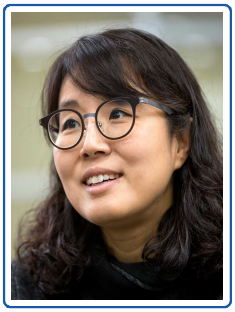
  노트북 (Google Colab 이용)

* 강의 난이도: 중급

* 강의: 노미나교수 (한양대학교 컴퓨터소프트웨어학부)

Curriculum Vitae

Speaker Name: Mina Rho, Ph.D.



▶ **Personal Info**

Name	Mina Rho
Title	Associate Professor
Affiliation	Hanyang University

▶ **Contact Information**

Address	222, Wangsimni-ro, Sungdong-Gu, Seoul, 04763
Email	minarho@hanyang.ac.kr
Phone Number	010-3460-9257

Research Interest

Translational bioinformatics, machine learning, and (meta)genomics

Educational Experience

1998	B.S. in Computer Science, Ewha Womans University, Korea
2001	M.S. in Computer Engineering, Boston University, USA
2009	Ph.D. in Computer Science, Indiana University, USA

Professional Experience

2009-2012	Postdoctoral Associate, Dept. of Computer Science, Indiana University, Seoul, Korea
2012-2013	Assistant Professor, Dept. of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, USA
2013-2017	Assistant Professor, Dept. of Computer Science, Hanyang University, Seoul, Korea
2017-Current	Associate Professor, Dept. of Computer Science, Hanyang University, Seoul, Korea

Selected Publications (5 maximum)

1. HJ Gwak, M Rho (2022) "ViBE: a deep learning model to classify viruses using metagenome sequencing data", Briefings in Bioinformatics: bbac204
2. J Jeon, J Lee, SM Jung, JH Shin, WJ Song, M Rho (2021) "Genomic Determinants Encode for the Reactivity and Regioselectivity of Flavin-Dependent Halogenases in Bacterial Genomes and Metagenomes", mSystems: 6(3)
3. Y Park, J Lee, H Moon, Y Choi, M Rho (2021) "Discovering microbe-disease associations from the literature using a hierarchical long short-term memory network and an ensemble parser model", Scientific Reports: 11(5874)
4. HJ Gwak, M Rho (2020) "Data-driven modeling for species-level taxonomic assignment from 16S rRNA: Application to human microbiomes", Frontiers in microbiology :11
5. SK Lim, D Kim, DC Moon, Y Cho, M Rho (2020) "Antibiotic resistomes discovered in the gut microbiomes of swine and cattle", Giga Science :9(5)

KSBi-BIML 2023

Deep Learning in Bioinformatics

Overview on Machine Learning

Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed

- Arthur Samuel (1959)

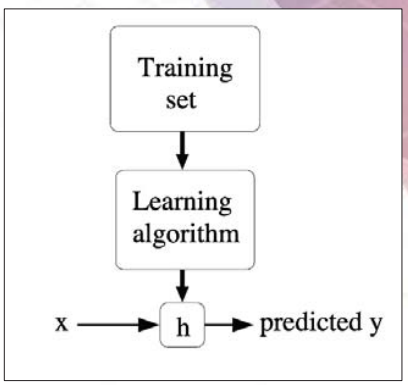
A computer program is said to learn from experience E with respect to some tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E

- Tom Mitchell

2

Overview on Machine Learning

Function approximation



Problem setting:

Set of instances (examples) $X = \{ x^1, \ldots, x^n \}$

Unknown target function $f: X \rightarrow Y$

Set of function hypothesis $H = \{h \mid h : X \rightarrow Y\}$

Input:

Training examples $\{ (x^i, y^i) \}$ of unknown target function f

Output:

Hypothesis $h \in H$ that best approximates target function f, $h \approx f$

3

Overview on Deep Learning

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

[1]Facebook AI Research, 770 Broadway, New York, New York 10003 USA. [2]New York University, 715 Broadway, New York, New York 10003, USA. [3]Department of Computer Science and Operations Research Université de Montréal, Pavillon André-Aisenstadt, PO Box 6128 Centre-Ville STN Montréal, Quebec H3C 3J7, Canada. [4]Google, 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. [5]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3G4, Canada.

4

What is Representation?

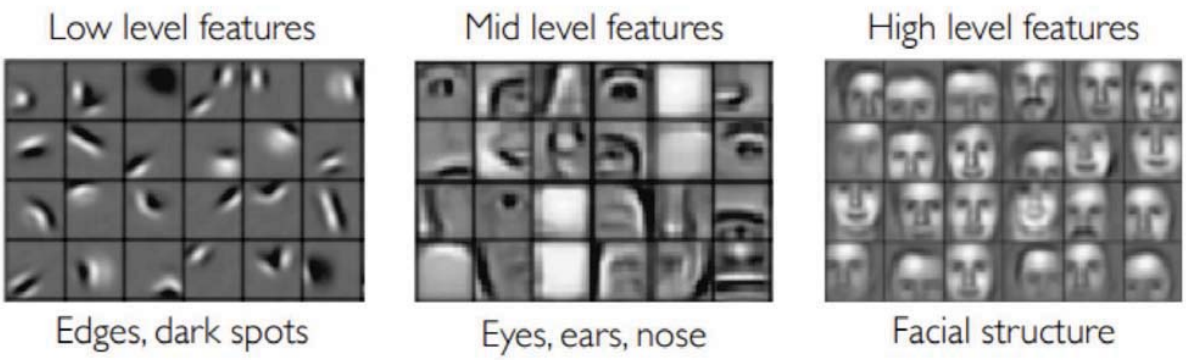
	# of words	# of attached files	# of links	# of malicious words	.....	spam
mail #1	256	0	3	7	.....	1 (Yes)
mail #2	56	1	0	3	.....	0 (No)
mail #3	24	1	0	1	.....	0
mail #4	672	0	0	0	.....	0
mail #5	67	2	4	3	.....	1
mail #6	48	0	2	6	.....	0
mail #7	79	1	3	8	.....	1

$$X^1 = (x_1^1, x_2^1, ..., x_n^1)$$

5

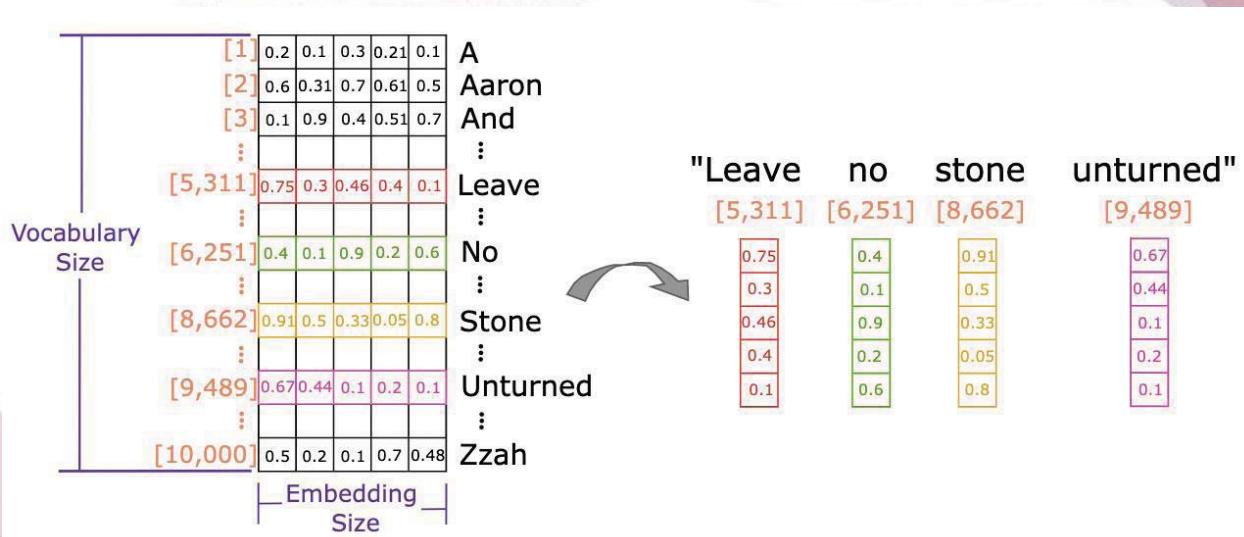
Learning Representations of data with multiple levels of representation

- Hand-engineered features are time-consuming and not scalable in practice.
- Can we learn the underlying features directly from the data?



6

Learning Representations of data with multiple levels of representation



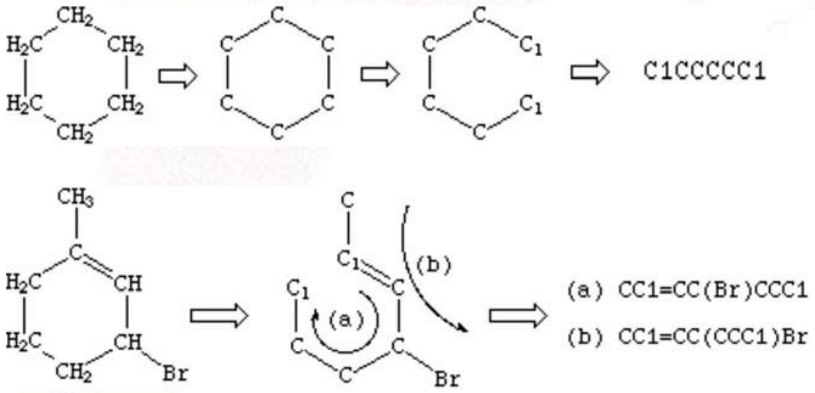
Learning Representations of data with multiple levels of representation

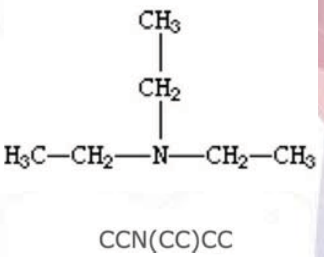
ACGCGCTGATGCCCGACACTGACTGACGCG

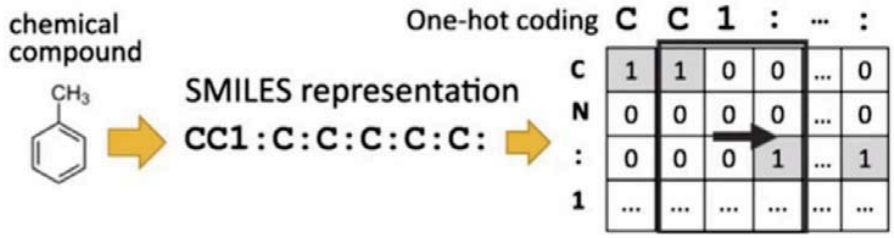
$X^1 = (x_1^1, x_2^1, \ldots)$ → (A, C, G, C, …)

(ACG, CGC, GCG, …)

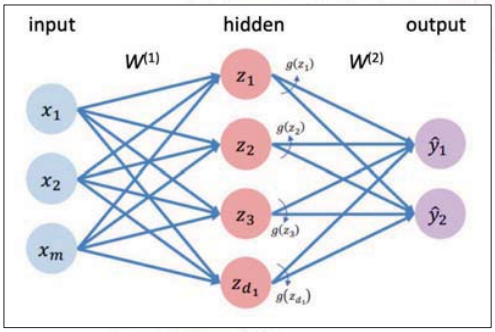
Learning Representations of data with multiple levels of representation

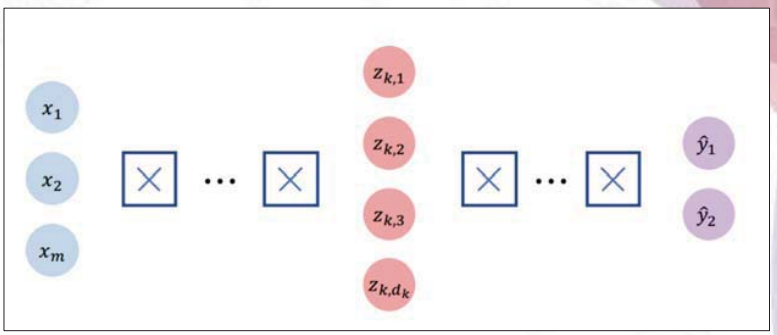






Models that are composed of multiple processing layers





$$z_i = w_{0,i}^{(1)} + \sum_{j=1}^{m} x_j \, w_{j,i}^{(1)}$$

$$z_{k,i} = w_{0,i}^{(k)} + \sum_{j=1}^{n_{k-1}} g(z_{k-1,j}) w_{j,i}^{(k)}$$

$$\hat{y}_i = g(w_{0,i}^{(2)} + \sum_{j=1}^{d_1} g(z_j) w_{j,i}^{(2)})$$

Loss optimization

- The loss of the network measures the cost incurred from incorrect predictions
- Cross entropy loss can be used with models that output a probability between 0 and 1

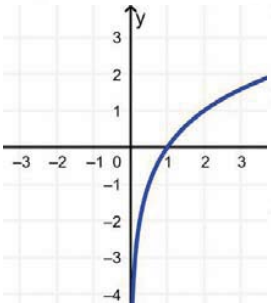
$$L(\hat{y}, y) = -(y log \hat{y} + (1-y) log(1-\hat{y}))$$

$$y = 1 : L(\hat{y}, y) = -log \hat{y}$$

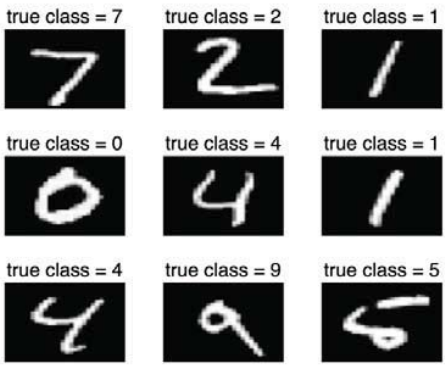
$$y = 0 : L(\hat{y}, y) = -log(1-\hat{y})$$

$$J(W) = \frac{1}{N} \sum_{i=1}^{N} L(f(x; W), y)$$

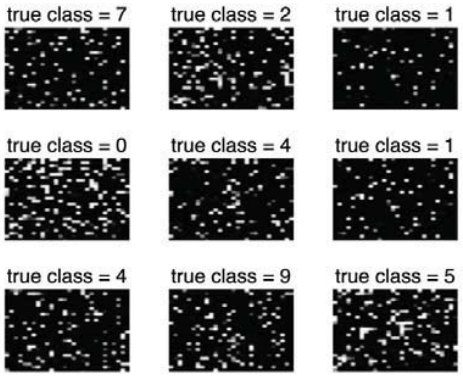
$$W^* = \underset{W}{argmin} J(W)$$



Neural Network



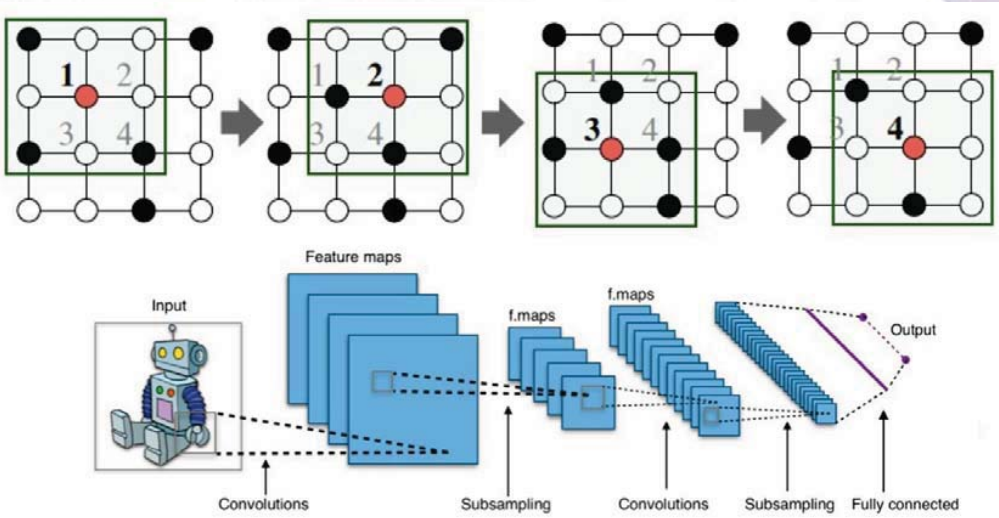
original image 28 x 28 grayscale 0:255



features permuted randomly

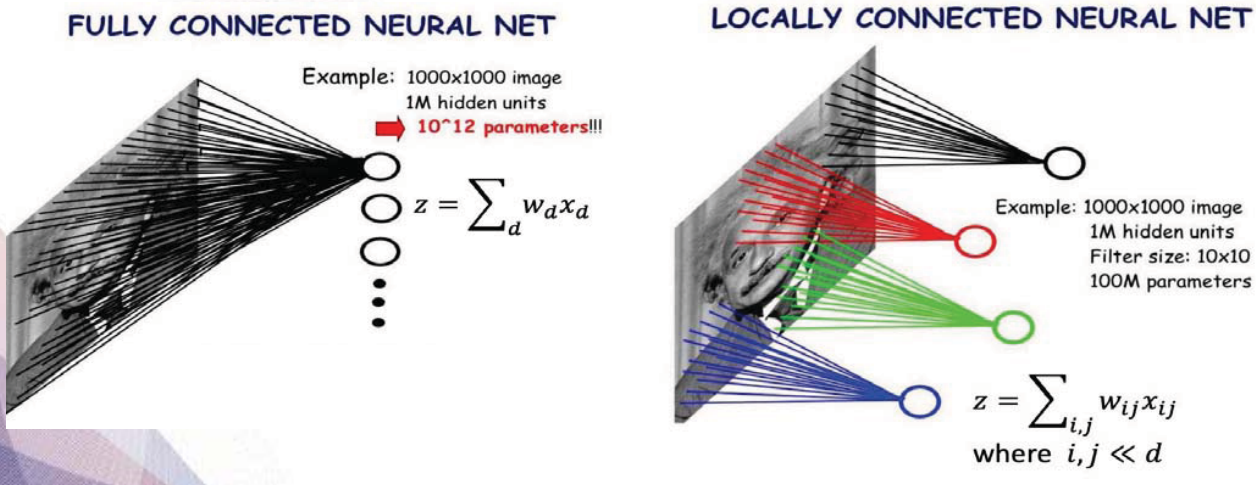
- images are 28 x 28 pixels
- represent input image as a vector $x \in \mathbb{R}^{784}$
- Learn a classifier $f(x)$ such that
  $f: x \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

Convolution neural network

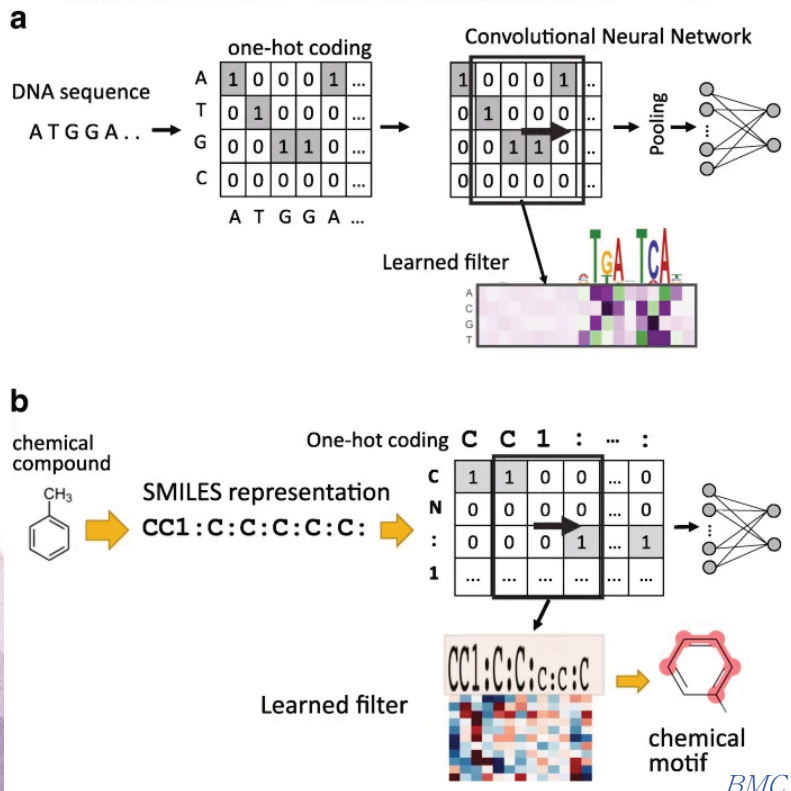


Convolution neural network

- Each convolution can be seen as a **locally-connected neuron** sliding on the entire input features

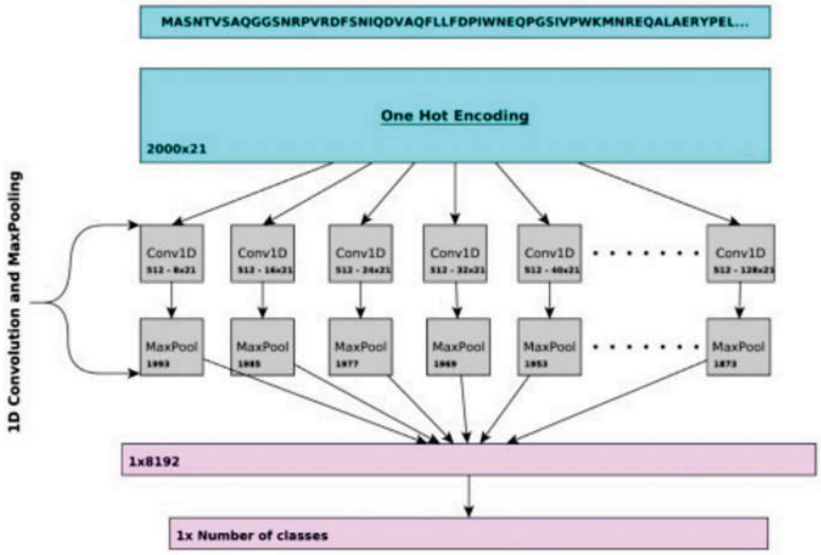


Learning representation with CNN



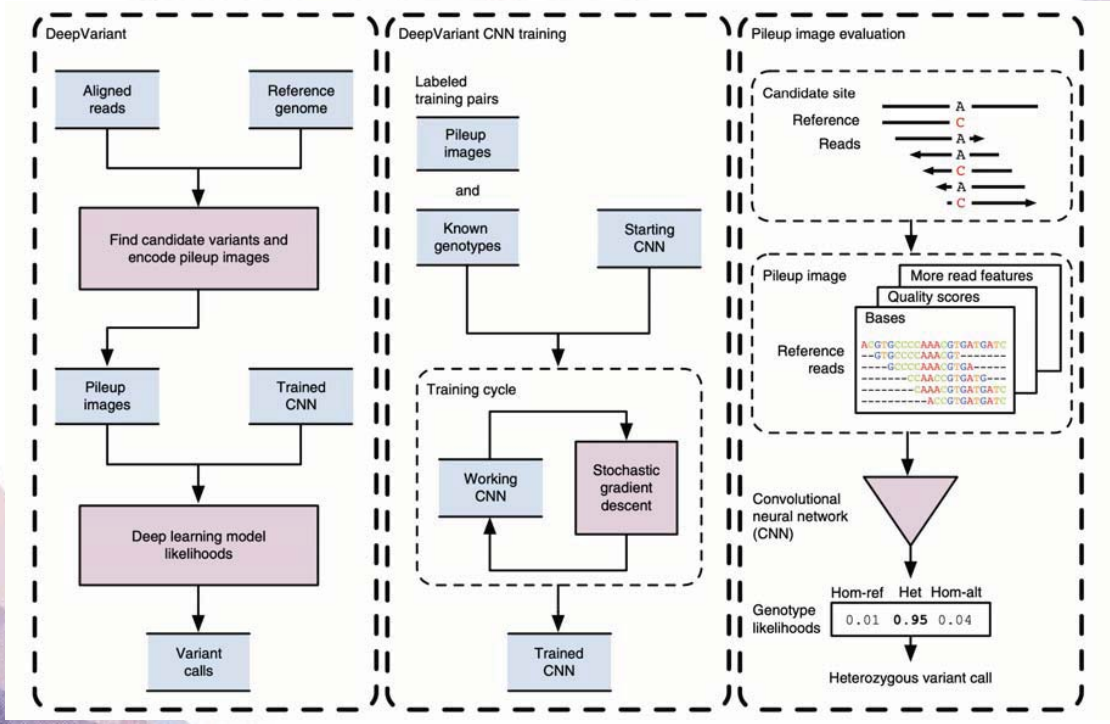
[BMC Bioinformatics](#) 19, 526 (2018)

Protein Function Prediction with CNN



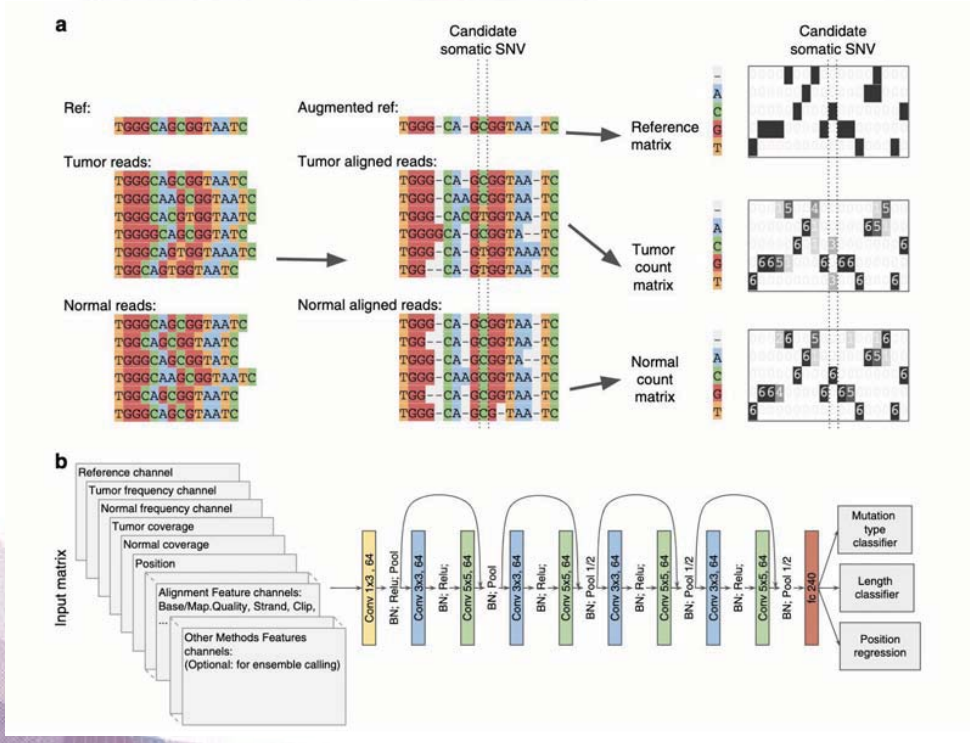
Bioinformatics, 36(2), 2020, 422–429 16

Learning representation with CNN: DeepVariant



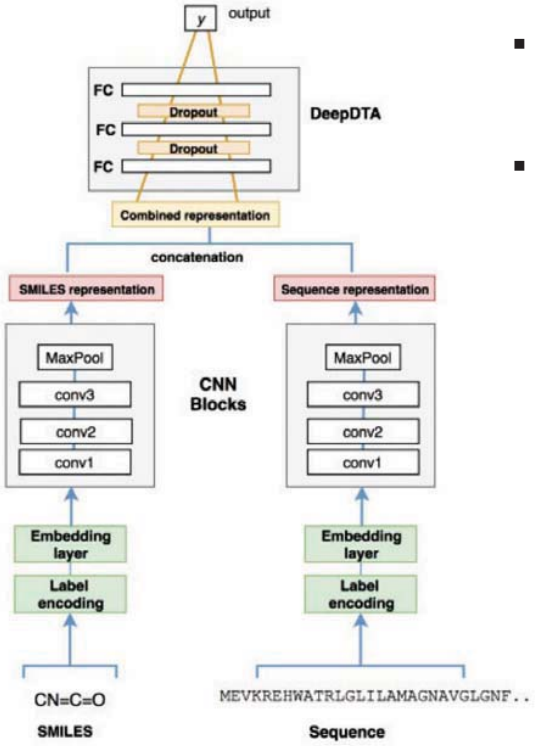
Nature Biotechnology 36, 983–987 (2018)17

Learning representation with CNN: NeuSomatic



NATURE COMMUNICATIONS | (2019)10:1041 18

Prediction of drug-target interaction using CNN



- The identification of novel drug–target (DT) interactions is a substantial part of the drug discovery process
- A deep-learning based model that uses only sequence information of both targets and drugs to predict DT interaction binding affinities

Table 6. The average r_m^2 and AUPR scores of the test set trained on five different training sets for the KIBA dataset

	Proteins	Compounds	r_m^2 (std)	AUPR (std)
KronRLS (Pahikkala *et al.*, 2014)	S–W	Pubchem Sim	0.342 (0.001)	0.635 (0.004)
SimBoost (He *et al.*, 2017)	S–W	Pubchem Sim	0.629 (0.007)	0.760 (0.003)
DeepDTA	CNN	CNN	0.673 (0.009)	0.788 (0.004)

Bioinformatics, 34, 2018, i821–i829 19

Representation

output distribution

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|}$$

hidden layer

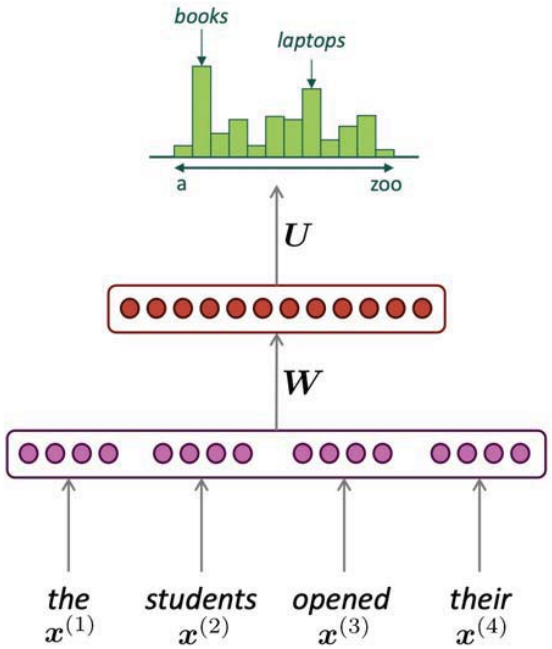
$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{e} + \boldsymbol{b}_1)$$

concatenated word embeddings

$$\boldsymbol{e} = [\boldsymbol{e}^{(1)}; \boldsymbol{e}^{(2)}; \boldsymbol{e}^{(3)}; \boldsymbol{e}^{(4)}]$$

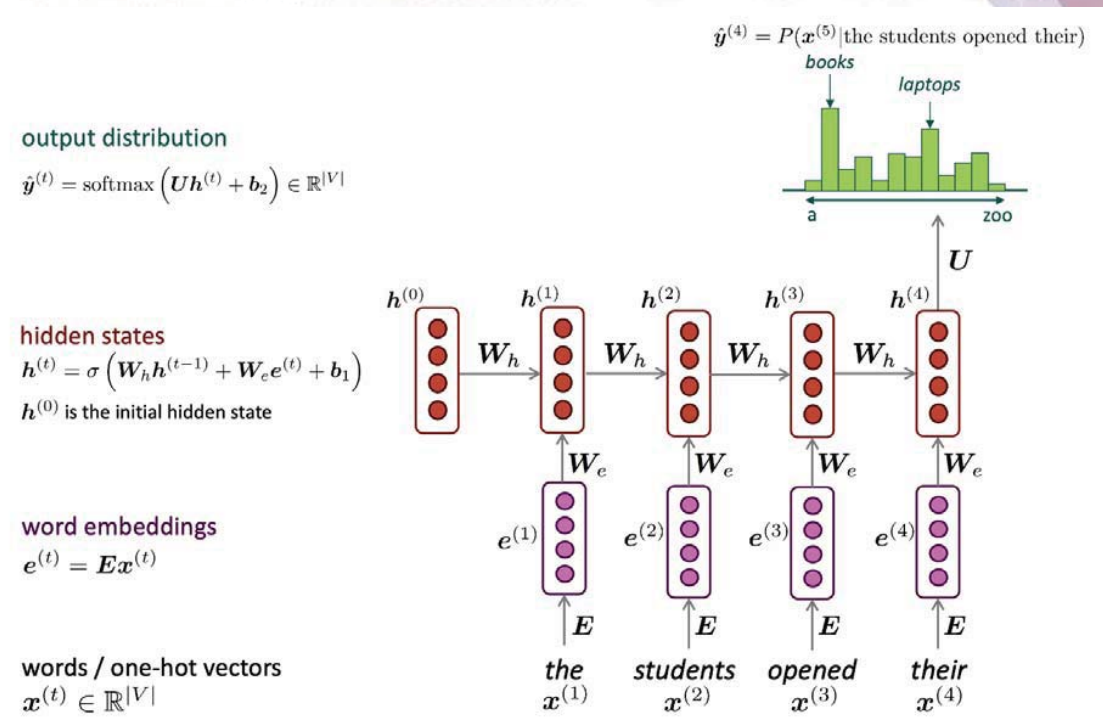
words / one-hot vectors

$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$$

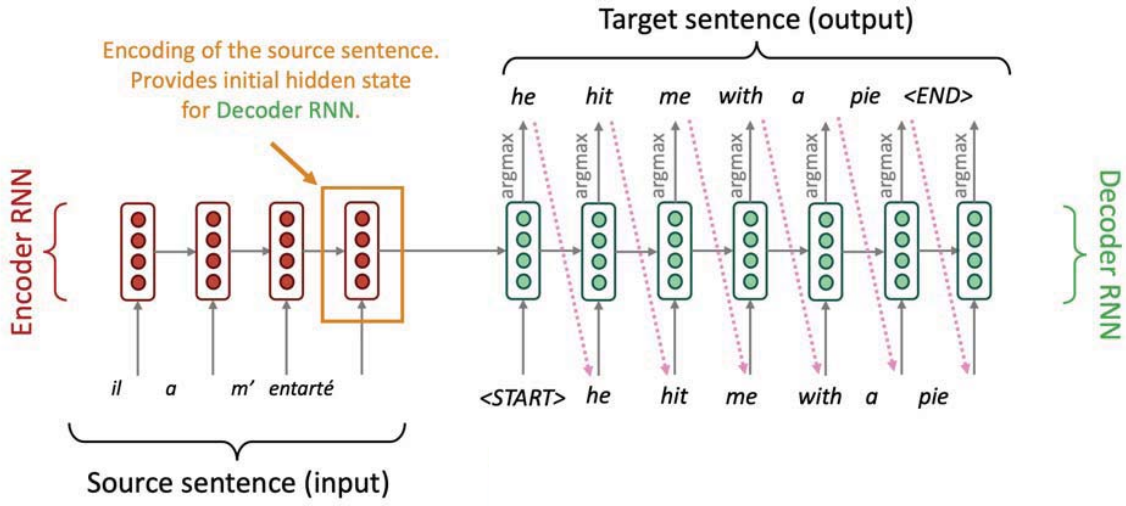


20

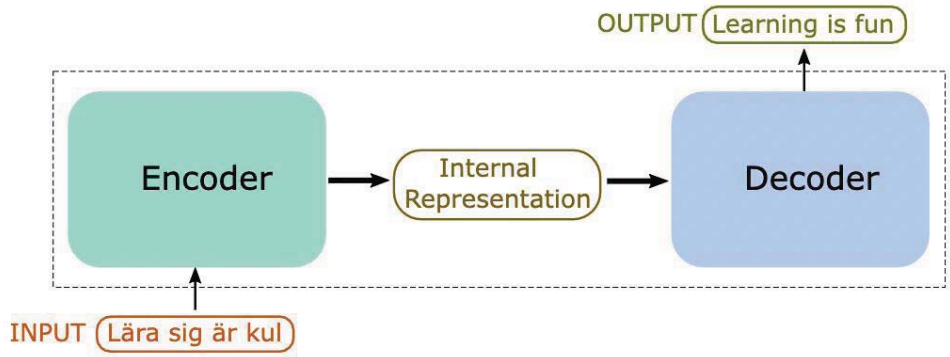
Embedding in RNN

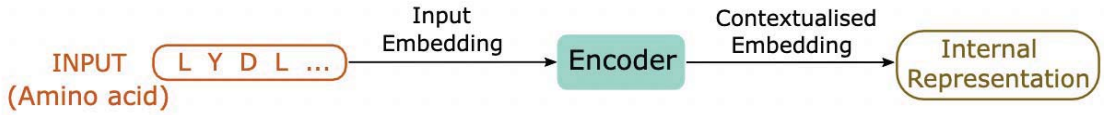


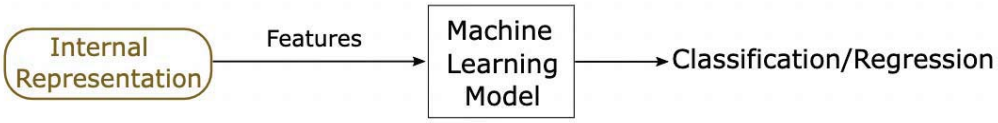
Encoder-decoder model



Encoder-decoder model

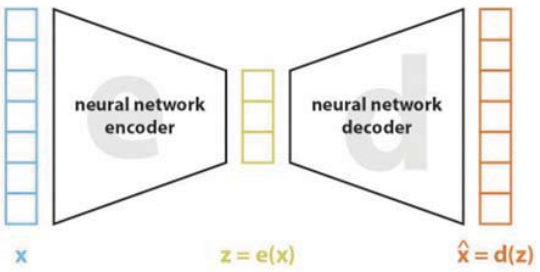






Encoder-decoder model

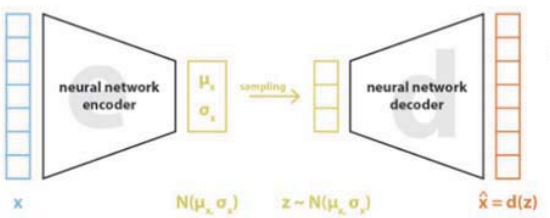
Use encoder-decoder model to generate efficient representations

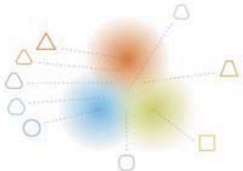


$$L = ||x - \hat{x}||^2$$
$$= ||x - d(e(x))||^2$$

Instead of encoding an input as a single point, encode it as a distribution over the latent space.

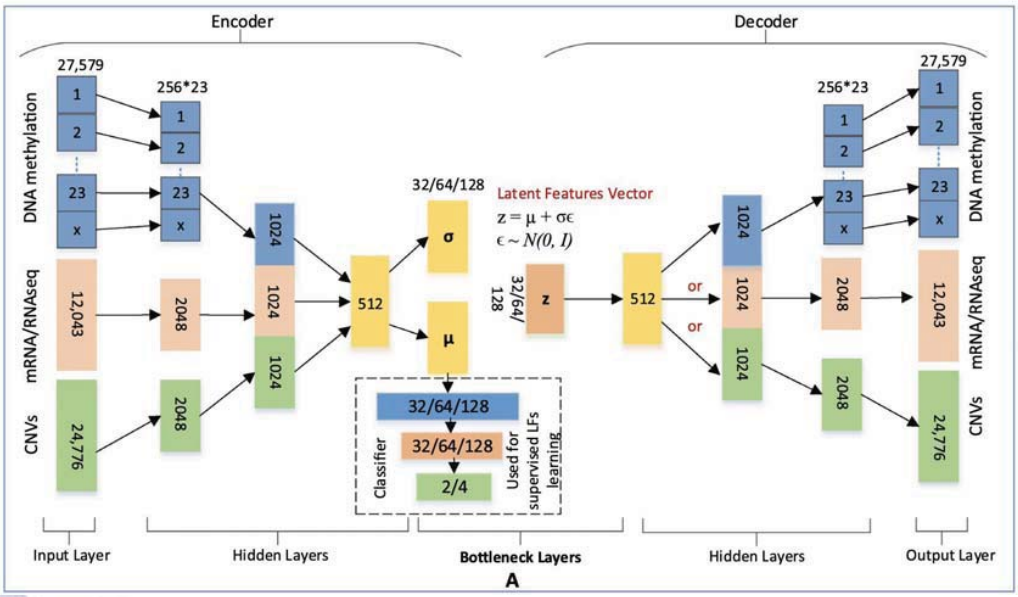
$$L = \underbrace{||x - \hat{x}||^2}_{\text{reconstruction}} + \underbrace{KL[N(\mu_x, \sigma_x), N(0,1)]}_{\text{regularization}}$$





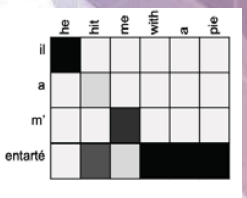
Encoder-decoder model

- Integrated multi-omics analysis of ovarian cancer using variational autoencoders



Scientific Reports 11: 6265 (2021)

Attention model



$$a_t = \sum_{i=1}^{N} \alpha_i^t h_i \in \mathbb{R}^h$$

Weighted sum of encoder hidden states based on the attention distribution

Softmax

$$\alpha^t = \text{softmax}(e^t) \in \mathbb{R}^N$$

Dot product

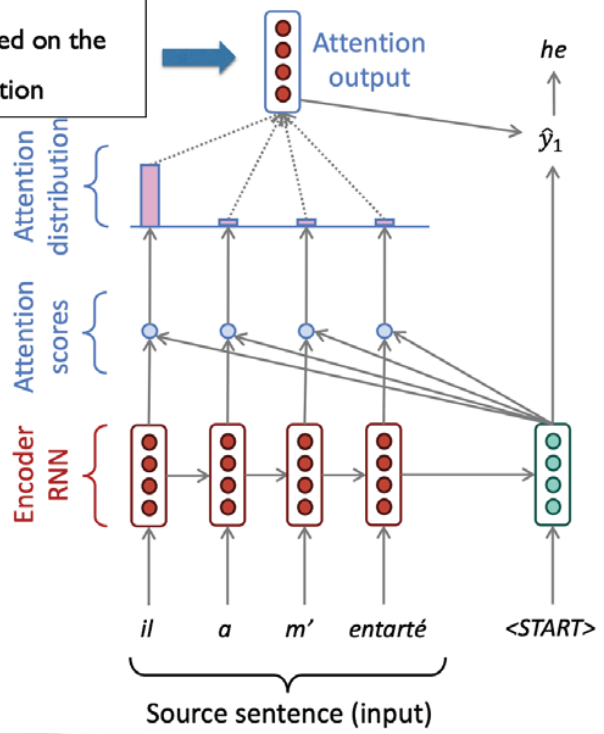
$$e^t = [s_t^T h_1, \ldots, s_t^T h_N] \in \mathbb{R}^N$$

Encoder hidden states

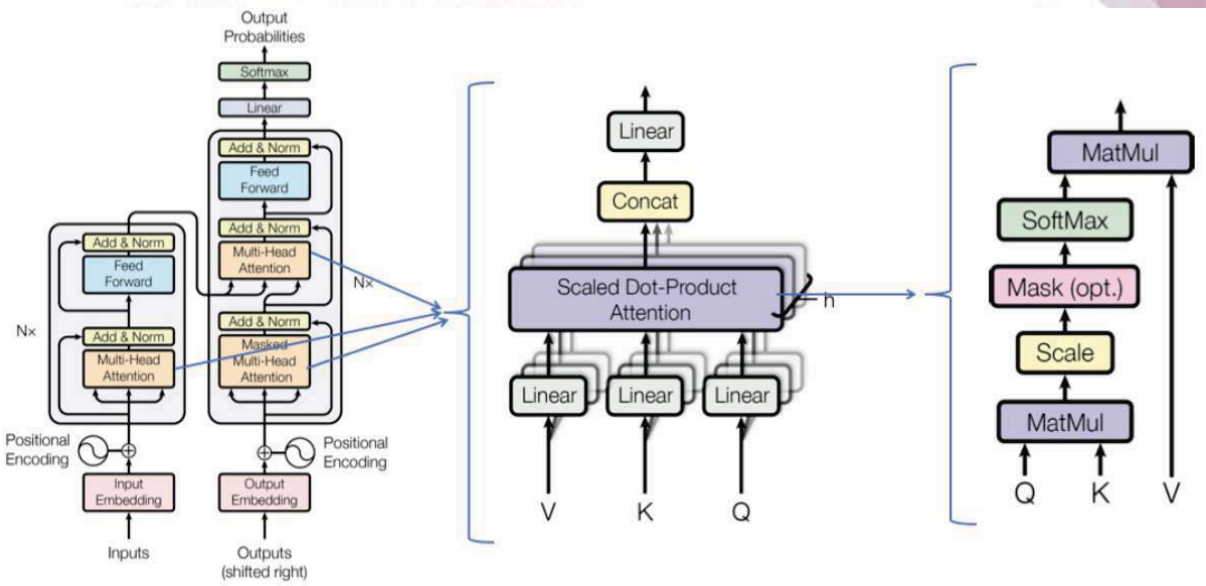
$$h_1, \ldots, h_N \in \mathbb{R}^h$$

Decoder hidden states

$$s_t \in \mathbb{R}^h$$



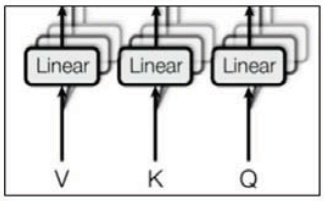
Self- attention model

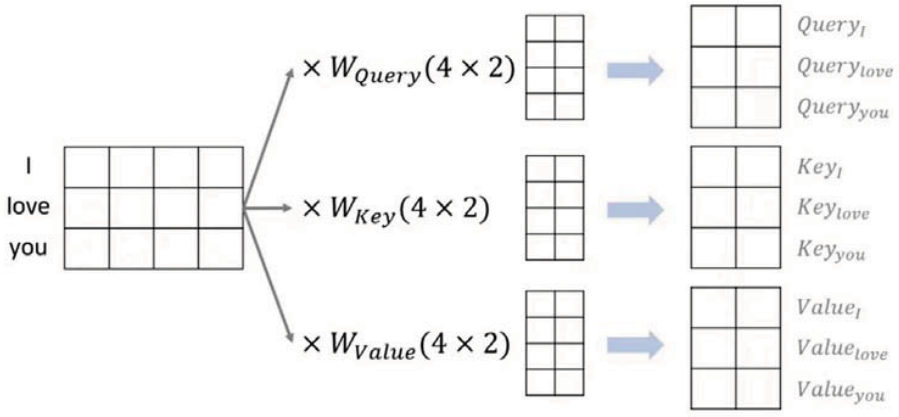


NIPS2017; https://arxiv.org/abs/1706.03762 27

Self- attention model

- Transform the token embedding to the embedding for query, key, and value

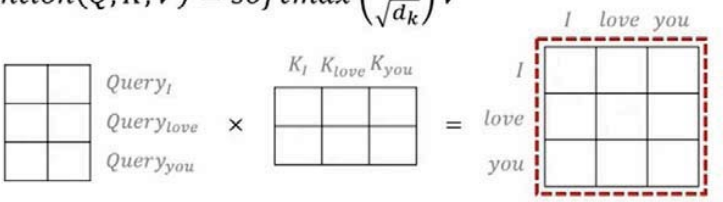


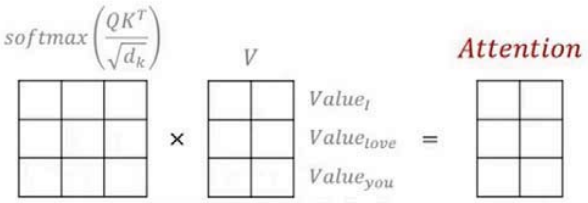


NIPS2017; https://arxiv.org/abs/1706.03762 28

Self- attention model

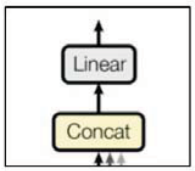
- $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$





NIPS2017; https://arxiv.org/abs/1706.03762 29

Self- attention model

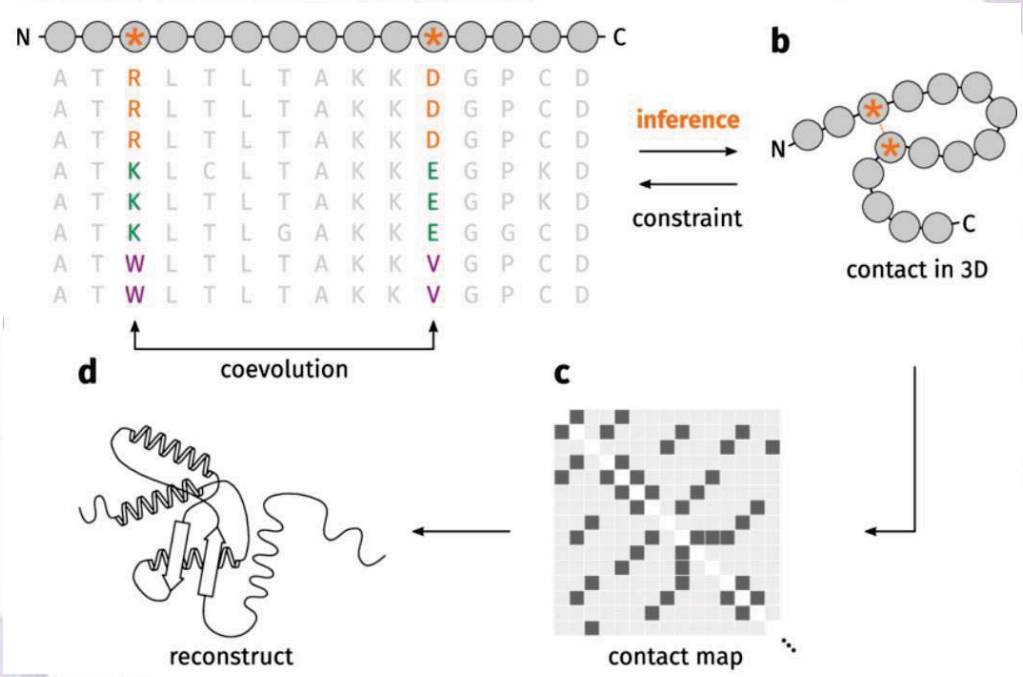


$$Concat(head_1, \dots, head_h) = \underbrace{head_1 \quad head_2 \quad head_3 \quad \dots \quad head_h}_{d_{model} = d_v \times h}$$

$$MultiHead(Q, K, V) = \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{d_{model} = d_v \times h} seq\_len \times \underbrace{\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}}_{d_{model}} d_{model}$$

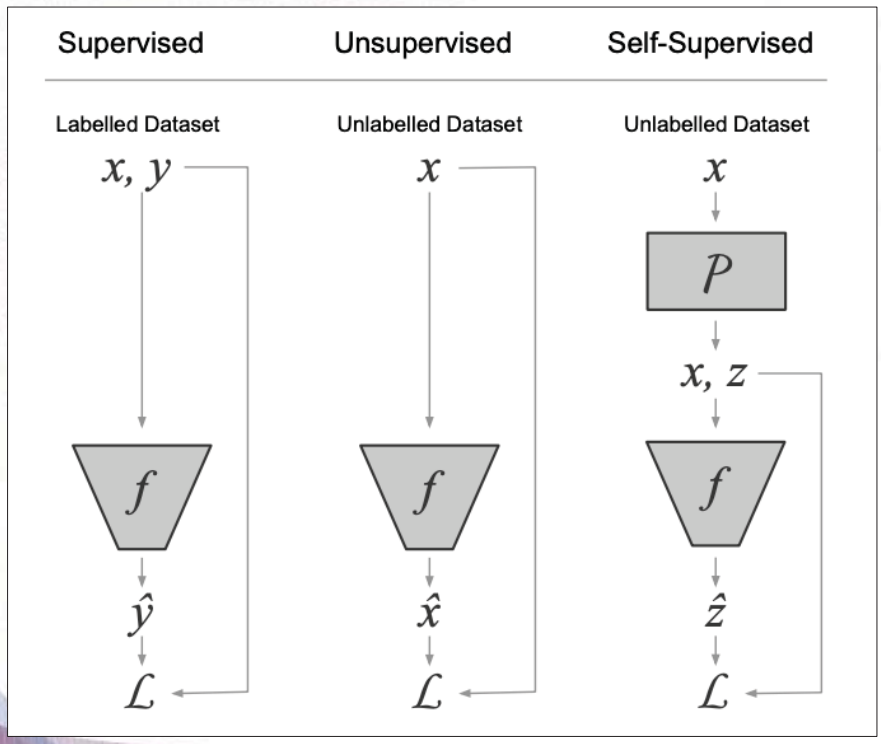
NIPS2017; https://arxiv.org/abs/1706.03762 30

Self- attention model



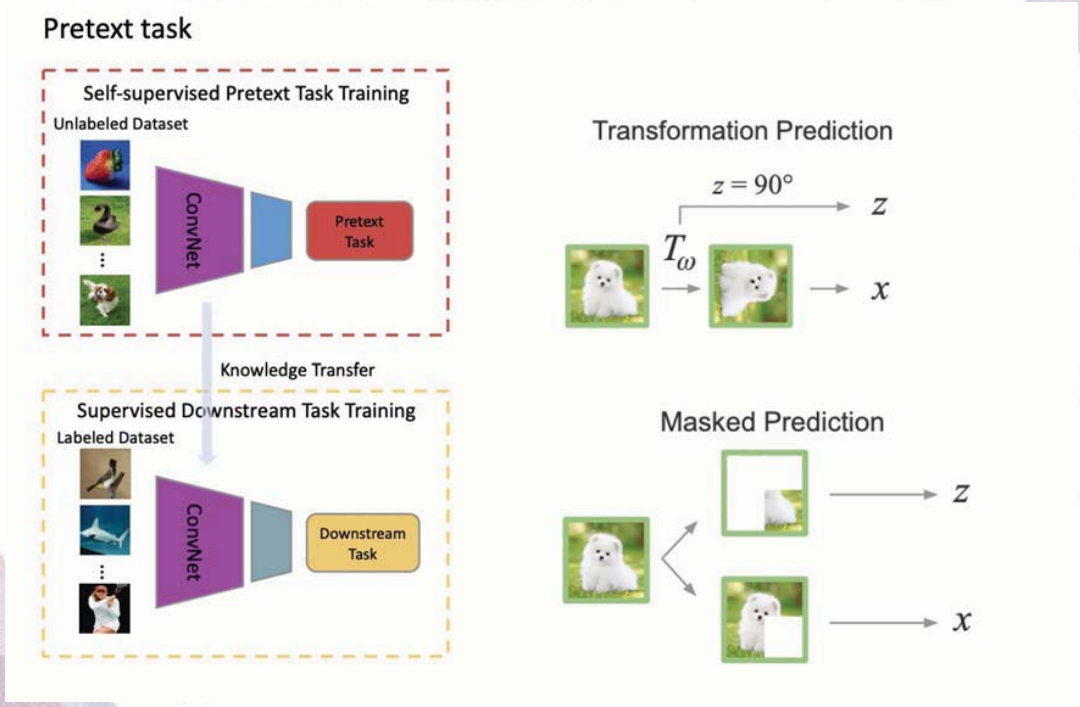
https://www.nature.com/articles/s41586-021-03819-2 31

Self-Supervised Learning



32

Self-Supervised Learning



33

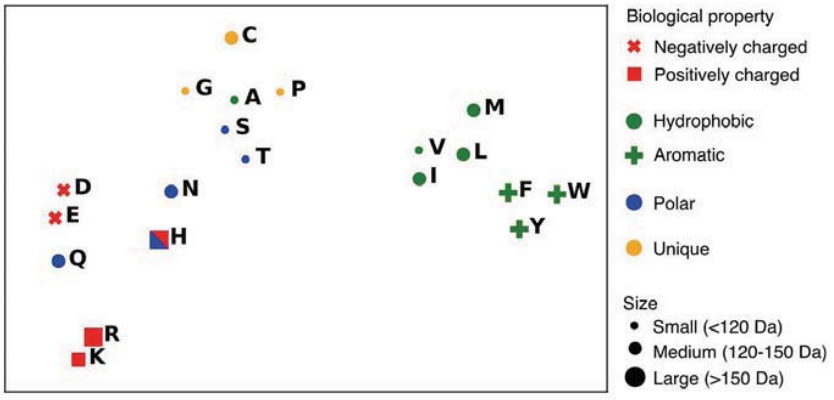
Self-Supervised Learning with protein sequences

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives[a,b,1,2], Joshua Meier[a,1], Tom Sercu[a,1], Siddharth Goyal[a,1], Zeming Lin[b], Jason Liu[a], Demi Guo[c,3], Myle Ott[a], C. Lawrence Zitnick[a], Jerry Ma[d,e,3], and Rob Fergus[b]

[a]Facebook AI Research, New York, NY 10003; [b]Department of Computer Science, New York University, New York, NY 10012; [c]Harvard University, Cambridge, MA 02138; [d]Booth School of Business, University of Chicago, Chicago, IL 60637; and [e]Yale Law School, New Haven, CT 06511

Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)



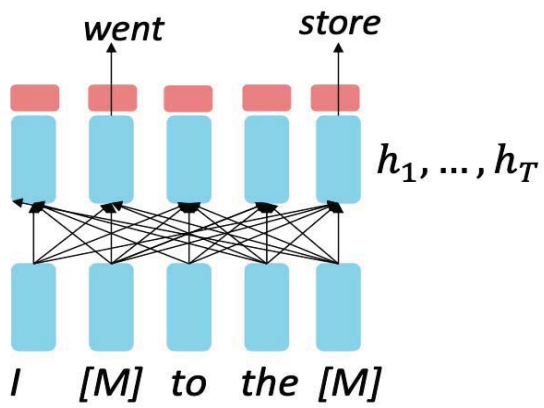
trained a deep contextual language model on 86 billion amino acids across 250 million protein sequence

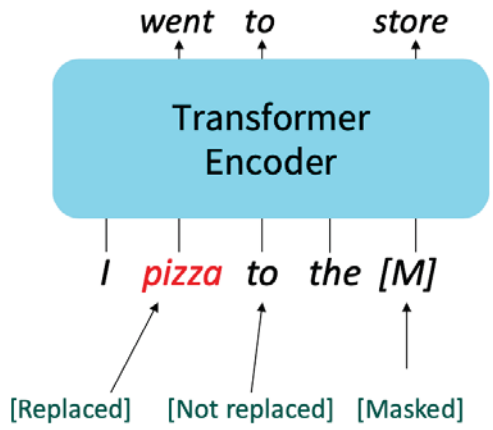
PNAS 2021 Vol. 118 No. 15 34

Masked Language Model

$\tilde{x}$ is the masked version of x, we're learning $p(x|\tilde{x})$

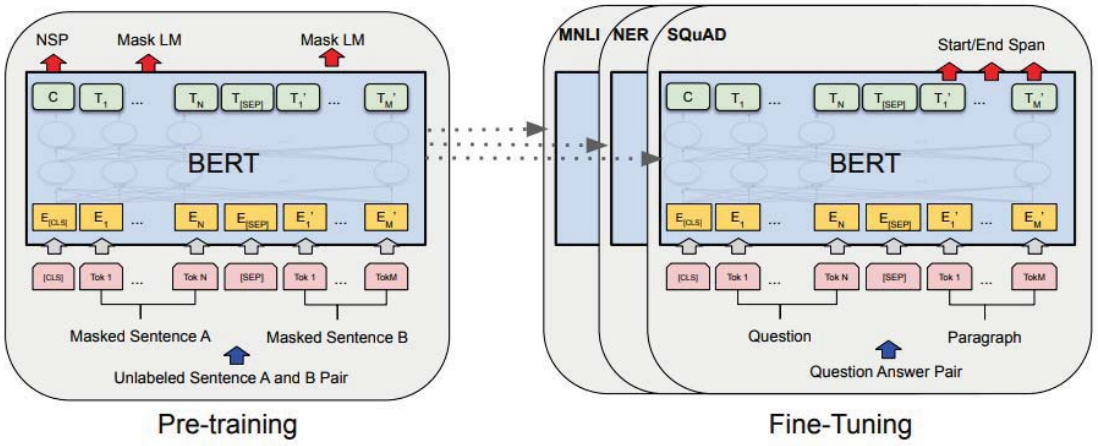
- replace some fraction of words in the input with a special [MASK] token
- predict these words.





35

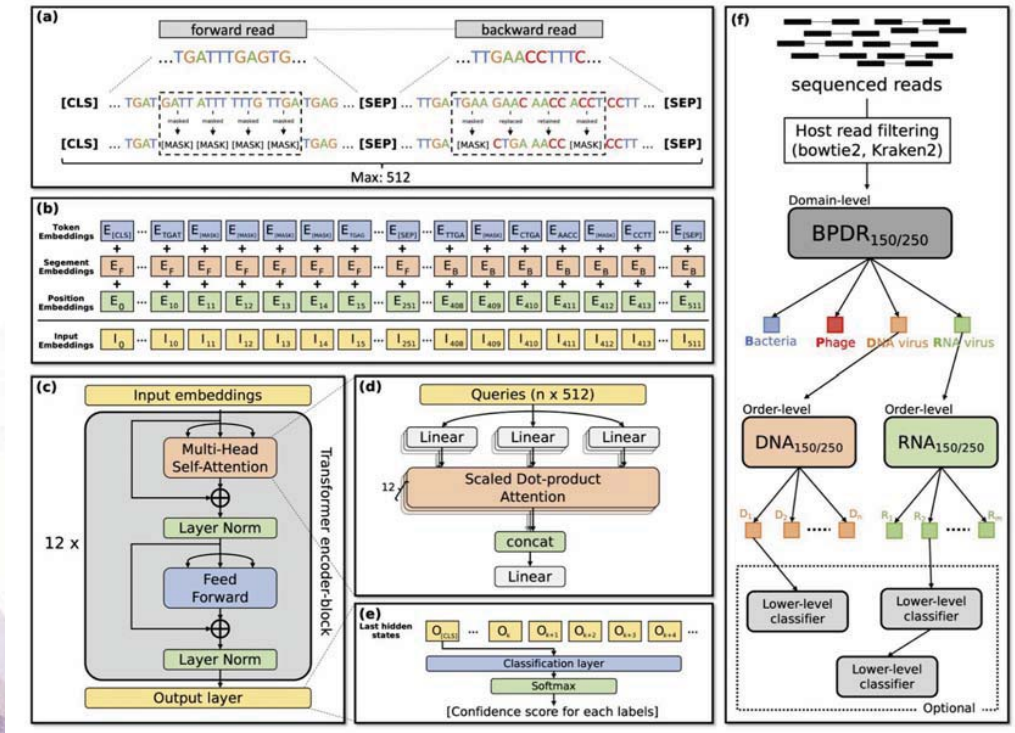
BERT



- Apart from output layers, the same architectures are used in both pre-training and fine-tuning
- The same pre-trained model parameters are used to initialize models for different down-stream tasks
- During fine-tuning, all parameters are fine-tuned.

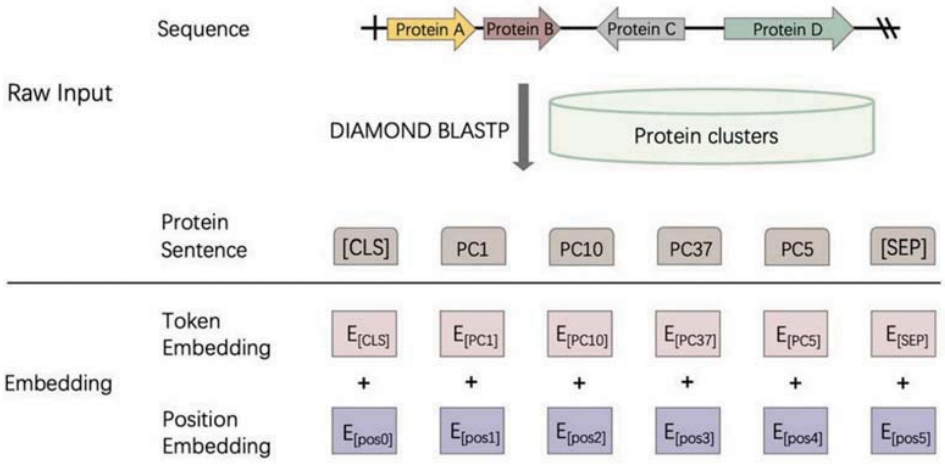
36

Identifying eukaryotic viruses using BERT and metagenome sequencing



Briefings in Bioinformatics, Volume 23, Issue 4 37

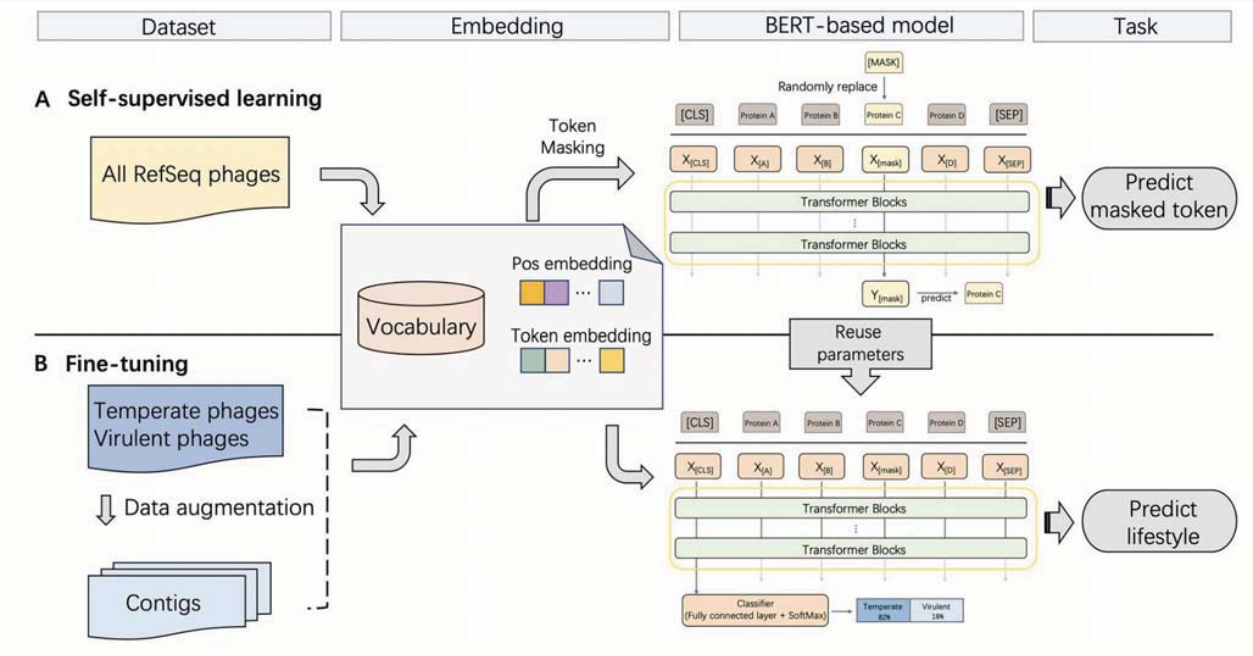
Predicting the lifestyle for bacteriophages using BERT



- Used all the phages proteins from the RefSeq database
- Clustered to generate

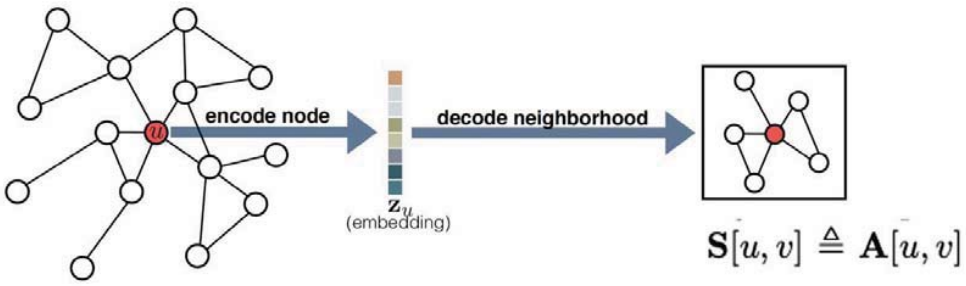
Briefings in Bioinformatics, 2023, 24(1), 1–11 38

Predicting the lifestyle for bacteriophages using BERT



Briefings in Bioinformatics, 2023, 24(1), 1–11 39

Graph Neural Network



$$\mathrm{ENC} : \mathcal{V} \rightarrow \mathbb{R}^d.$$

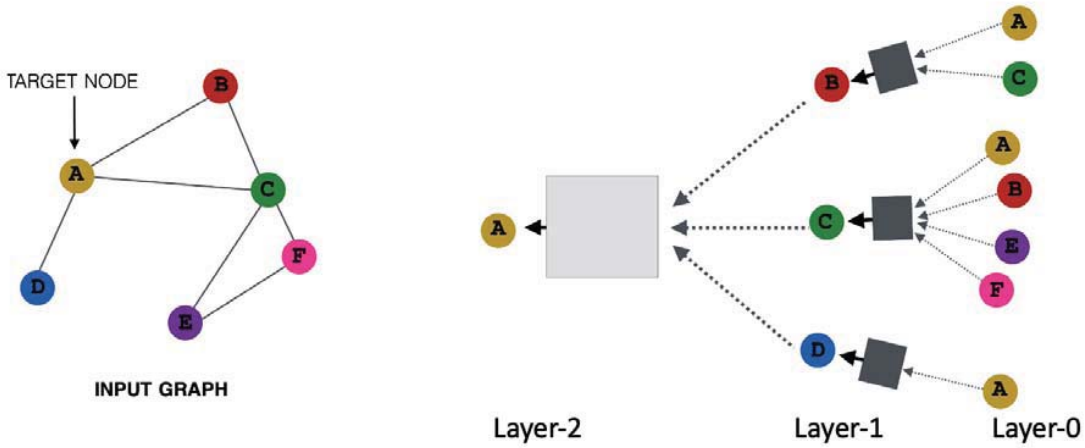
$$\mathrm{DEC} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$$

$$\mathrm{DEC}(\mathrm{ENC}(u), \mathrm{ENC}(v)) = \mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \approx \mathbf{z}_v^{\mathrm{T}} \mathbf{z}_u$$

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \ell\left(\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v), \mathbf{S}[u,v]\right) \approx \sum_{(u,v)\in\mathcal{D}} \mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \cdot \mathbf{S}[u,v].$$

40

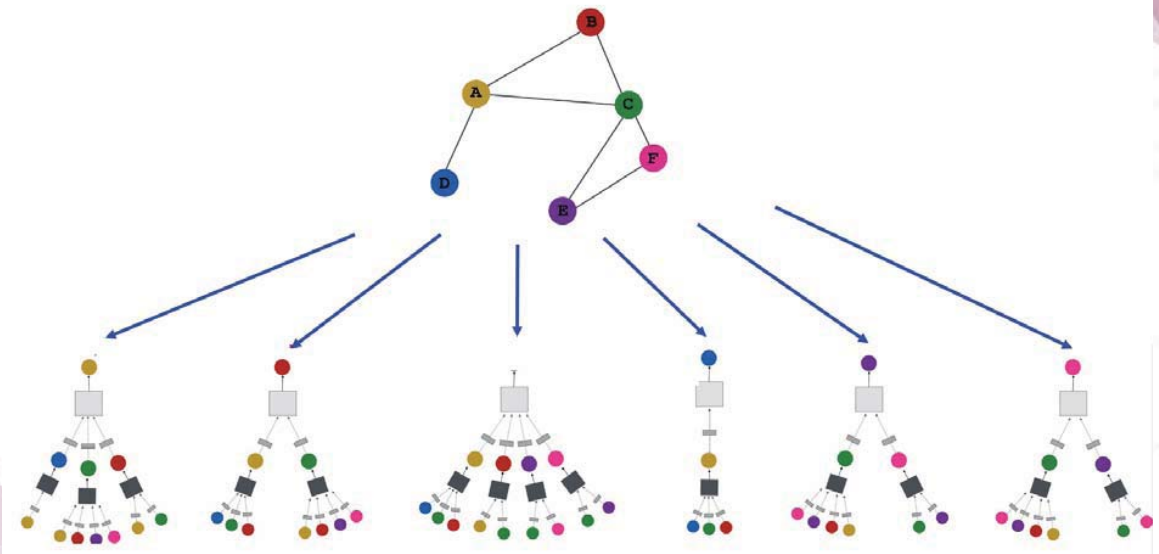
Graph Neural Network



- Nodes have embeddings at each layer
- Layer-0 embedding of node *v* is its input feature x_v
- Layer-k embedding gets information from nodes that are k hops away

41

Graph Neural Network



42

Graph Neural Network

$$\mathbf{h}_u^{(k+1)} = \text{UPDATE}^{(k)}\left(\mathbf{h}_u^{(k)}, \text{AGGREGATE}^{(k)}(\{\mathbf{h}_v^{(k)}, \forall v \in \mathcal{N}(u)\})\right)$$
$$= \text{UPDATE}^{(k)}\left(\mathbf{h}_u^{(k)}, \mathbf{m}_{\mathcal{N}(u)}^{(k)}\right)$$

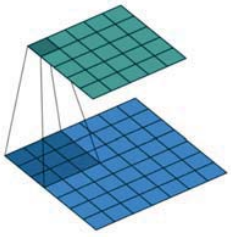
$$\mathbf{m}_{\mathcal{N}(u)} = \sum_{v \in \mathcal{N}(u)} \mathbf{h}_v$$
$$\text{UPDATE}(\mathbf{h}_u, \mathbf{m}_{\mathcal{N}(u)}) = \sigma\left(\mathbf{W}_{\text{self}}\mathbf{h}_u + \mathbf{W}_{\text{neigh}}\mathbf{m}_{\mathcal{N}(u)}\right)$$

$$\mathbf{h}_u^{(k)} = \sigma\left(\mathbf{W}_{\text{self}}^{(k)}\mathbf{h}_u^{(k-1)} + \mathbf{W}_{\text{neigh}}^{(k)} \sum_{v \in \mathcal{N}(u)} \mathbf{h}_v^{(k-1)} + \mathbf{b}^{(k)}\right)$$

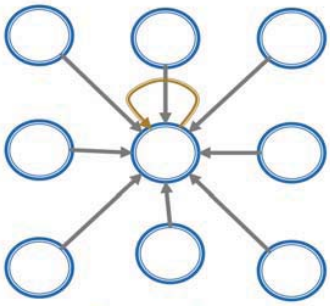
$$\mathbf{h}_u^{(0)} = \mathbf{x}_u, \forall u \in \mathcal{V}. \quad \mathbf{z}_u = \mathbf{h}_u^{(K)}, \forall u \in \mathcal{V}$$

43

Graph convolution network



Image



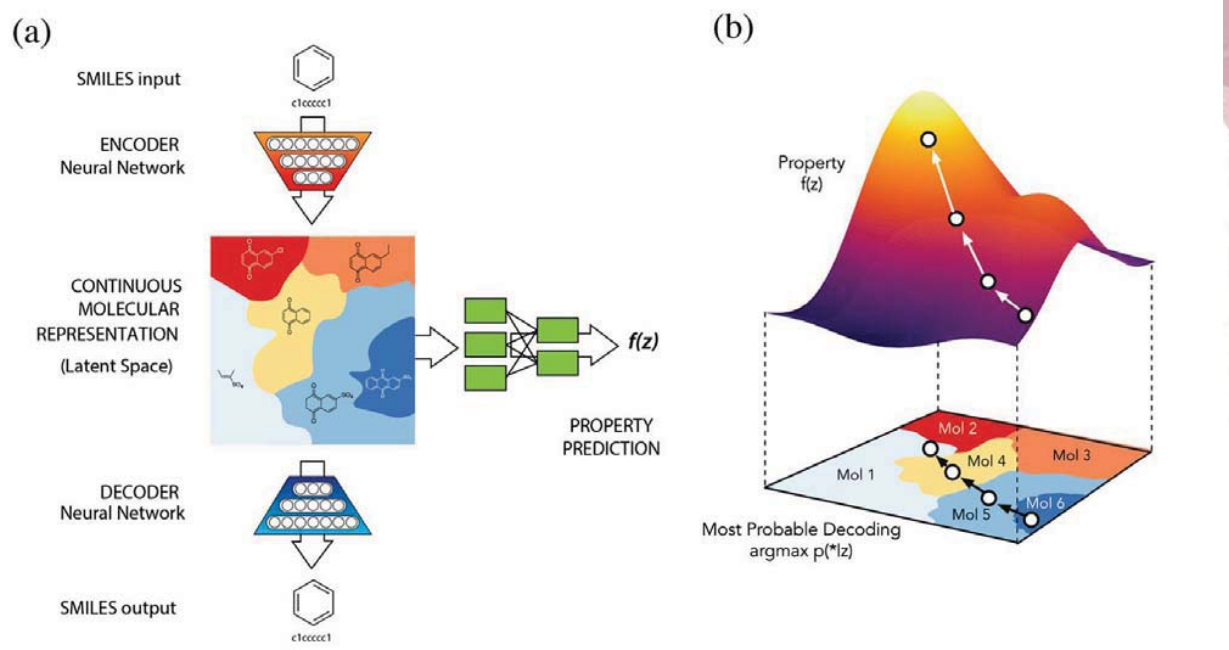
Graph

GNN $$\mathrm{h}_v^{(l+1)} = \sigma(\mathbf{W}_l \textstyle\sum_{u \in \mathrm{N}(v)} \frac{\mathrm{h}_u^{(l)}}{|\mathrm{N}(v)|} + \mathrm{B}_l \mathrm{h}_v^{(l)}), \forall l \in \{0, \ldots, L-1\}$$

CNN $$\mathrm{h}_v^{(l+1)} = \sigma(\textstyle\sum_{u \in \mathrm{N}(v) \cup \{v\}} \mathrm{W}_l^u \mathrm{h}_u^{(l)}), \forall l \in \{0, \ldots, L-1\}$$
$$\mathrm{h}_v^{(l+1)} = \sigma(\textstyle\sum_{u \in \mathrm{N}(v)} \mathbf{W}_l^u \mathrm{h}_u^{(l)} + \mathrm{B}_l \mathrm{h}_v^{(l)}), \forall l \in \{0, \ldots, L-1\}$$

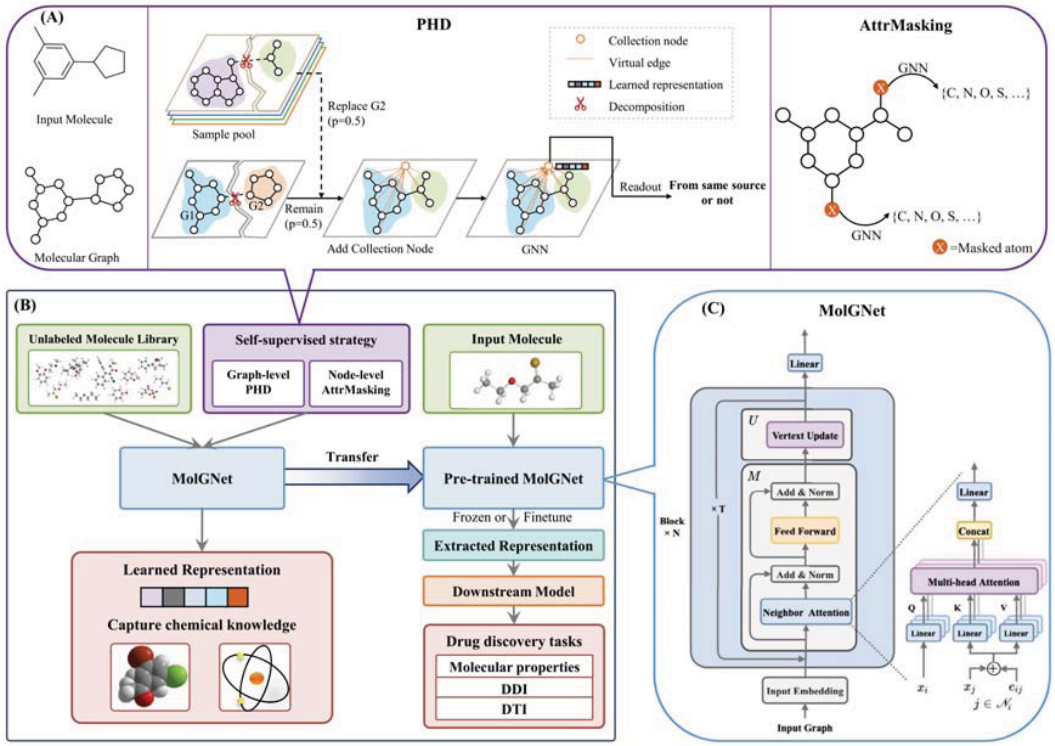
44

Graph Neural Network



ACS Cent. Sci. 2018, 4, 268−276

Learning molecular representation based on self-supervised approach



Briefings in Bioinformatics, Volume 22, Issue 6,

47