Class core values

1. Be **respect**ful to yourself and others
2. Be **confident** and believe in yourself
3. Always do your **best**
4. Be **cooperative**
5. Be **creative**
6. Have **fun**
7. Be **patient** with yourself while you learn
8. Don’t be shy to **ask “stupid” questions**

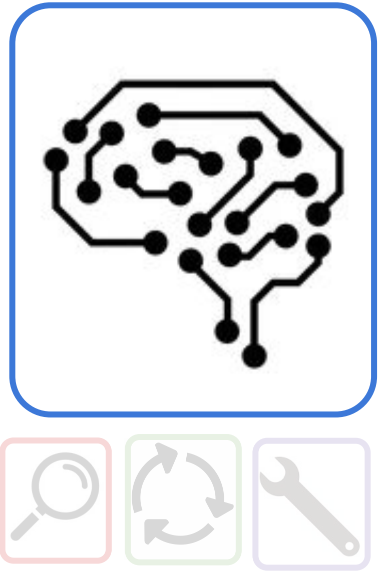
Week 8, Lecture 1

Machine learning in protein engineering

Learning Objectives

1. Describe applications of machine learning in protein engineering
2. Identify challenges in the application of neural nets
3. Evaluate learning literature based on performance and application

Machine learning



Machine learning

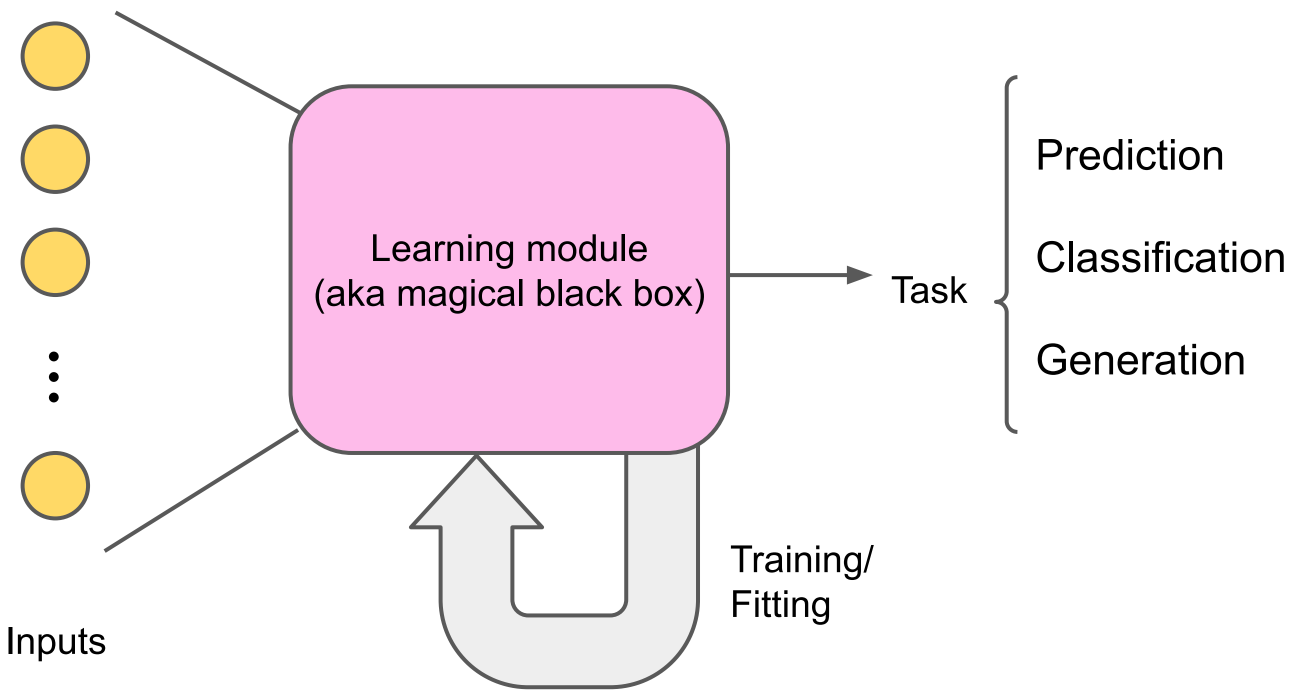
learns from large databases for prediction, classification, generation

Machine learning

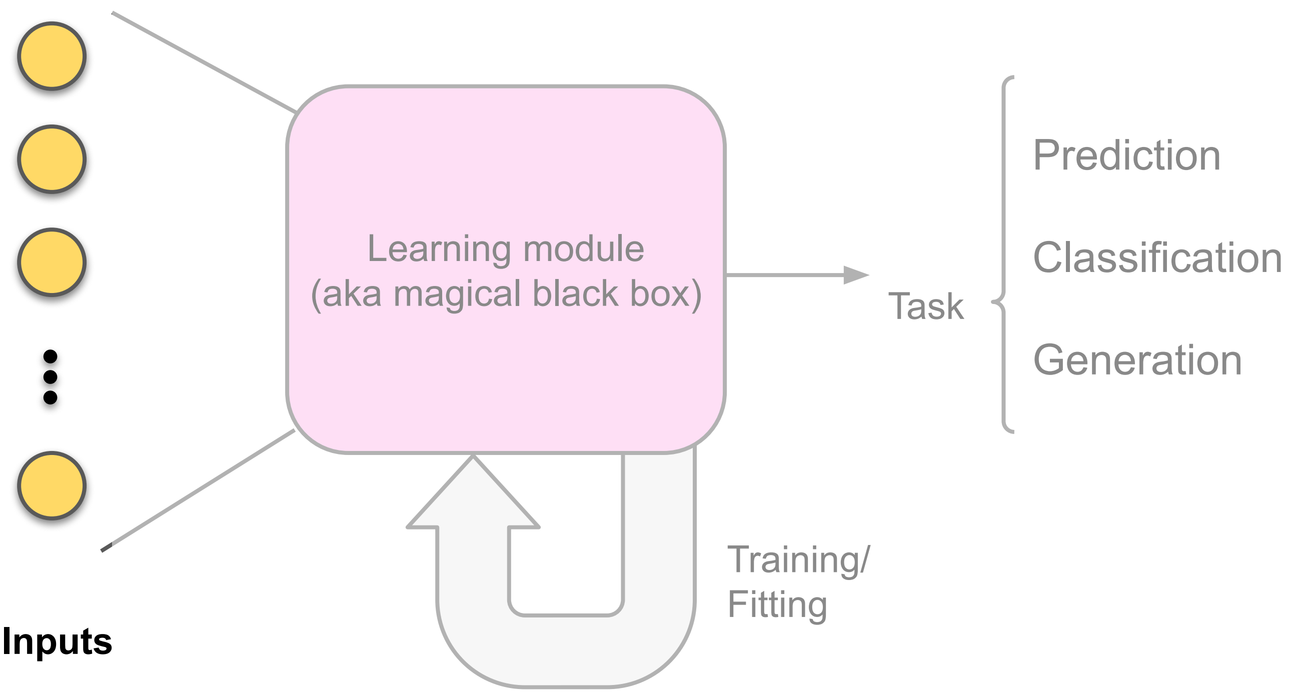
learns from large databases for prediction, classification, generation

Advantage	Disadvantage
The computational power is getting better	Needs large amount of data
It can uncover new patterns and rules that we can’t	Can’t be specific/target-based (great for general cases)
Fully reproducible	Can sample new sequences but in the neighborhood of what exist
Can take in multiple considerations into account	

Basic components of a learning module



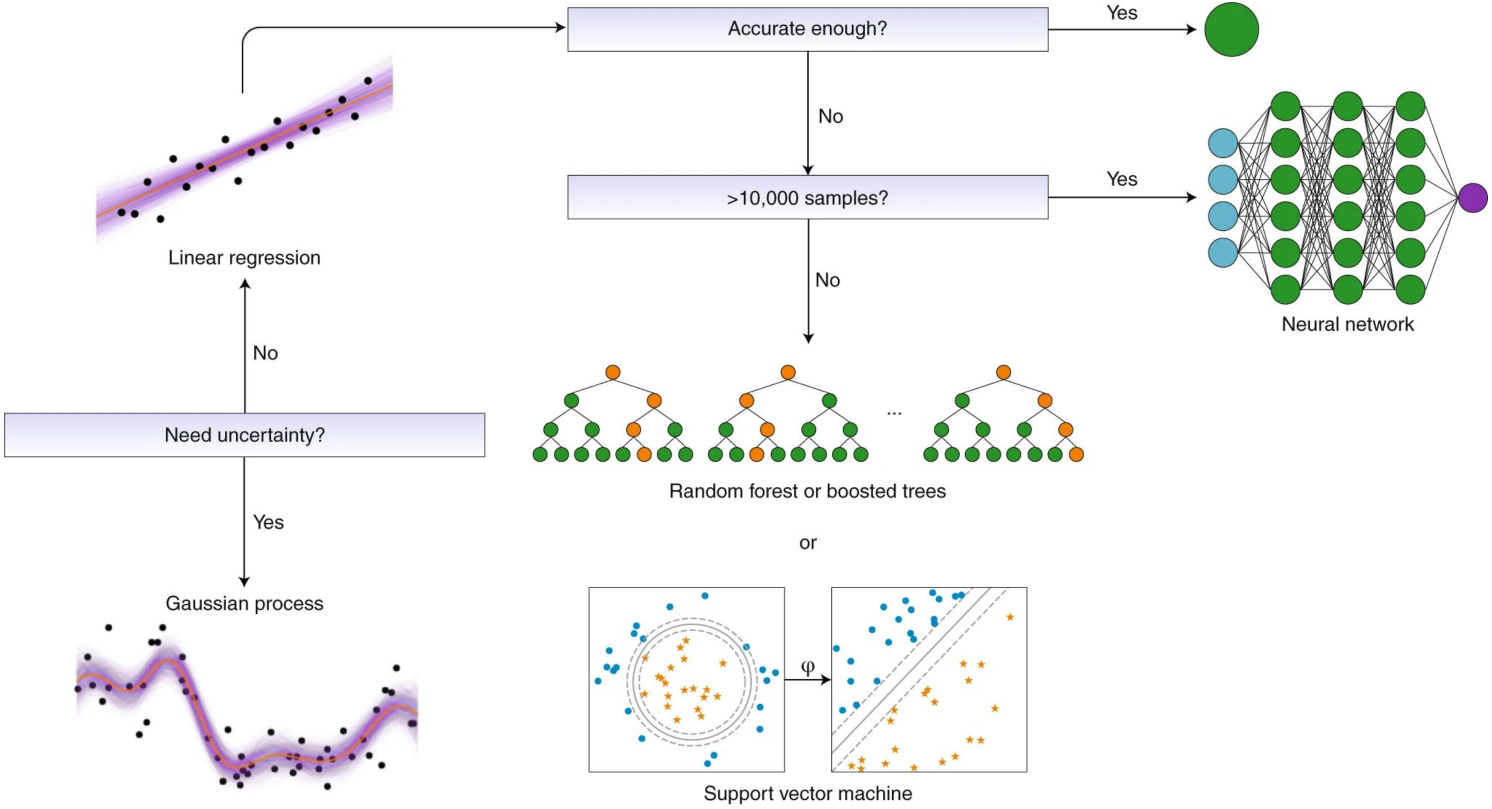
The inputs define the model to use



What to check for in inputs

1. How much data do I have?

The number of inputs defines types of ML model



Nature Methods 16, 687–694 (**2019**)

What to check for in inputs

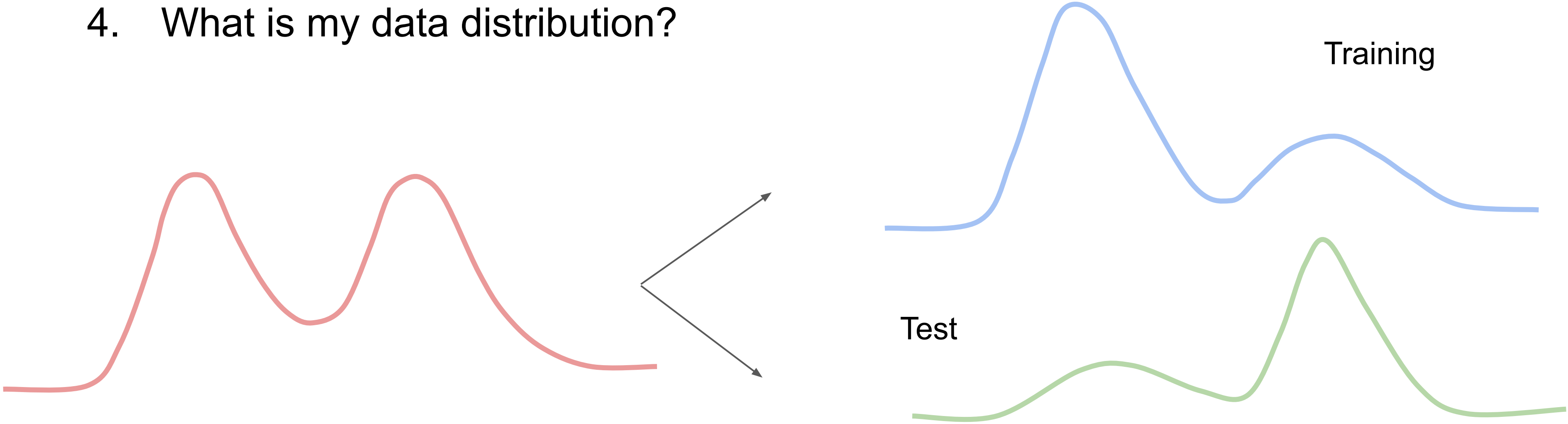
1. How much data do I have?
2. What is my data type?
   a. Sequence
   b. Structure
   c. Image
   d. …

What to check for in inputs

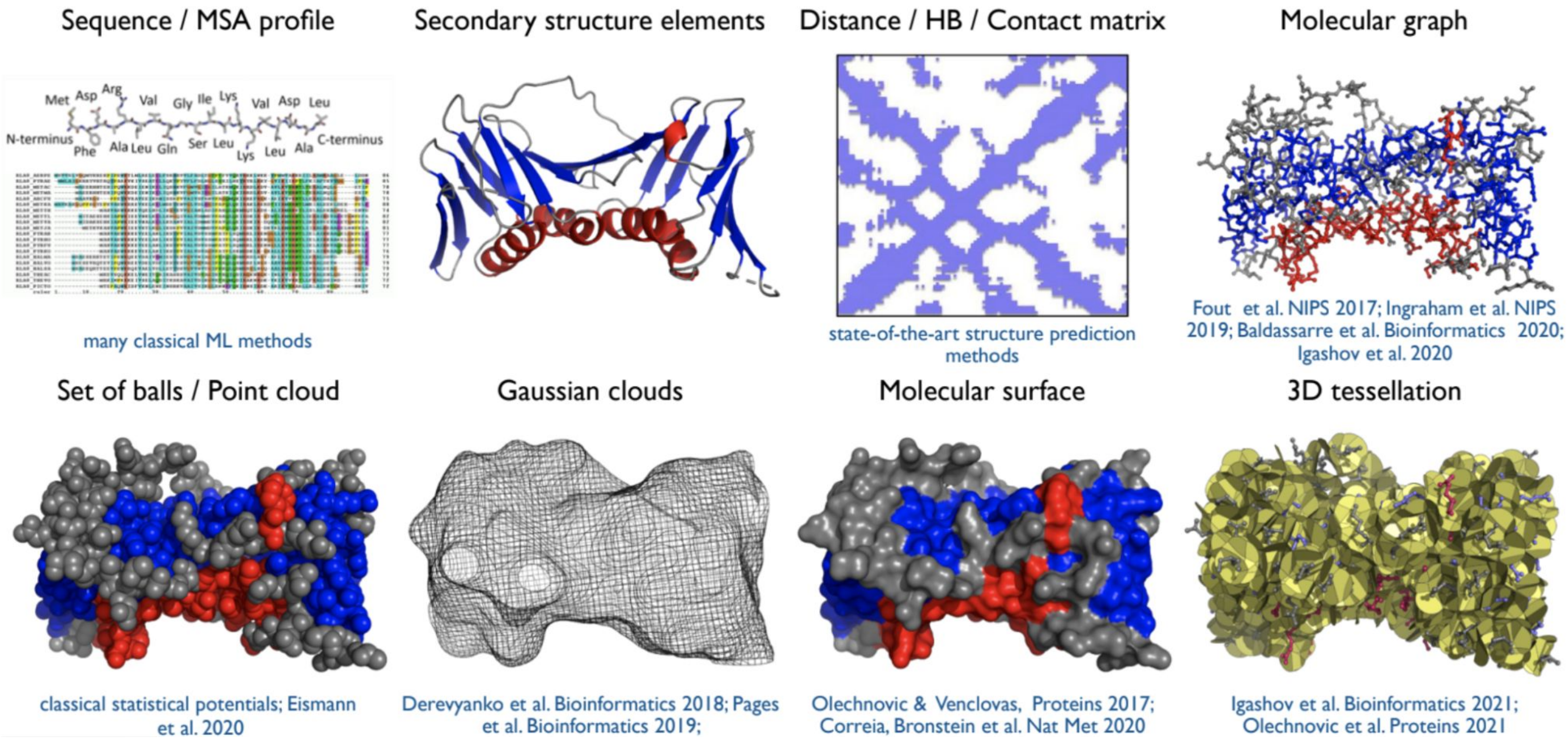
1. How much data do I have?
2. What is my data type?
3. How much noise do I have in my data?
   a. How can I clean it?

What to check for in inputs

1. How much data do I have?
2. What is my data type?
3. How much noise do I have in my data?
4. What is my data distribution?



Proteins can be represented in many different ways



Sequence / MSA profile

many classical ML methods

Secondary structure elements

Distance / HB / Contact matrix

state-of-the-art structure prediction methods

Molecular graph

Fout et al. NIPS 2017; Ingraham et al. NIPS 2019; Baldassarre et al. Bioinformatics 2020; Igashov et al. 2020

Set of balls / Point cloud

classical statistical potentials; Eismann et al. 2020

Gaussian clouds

Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;

Molecular surface

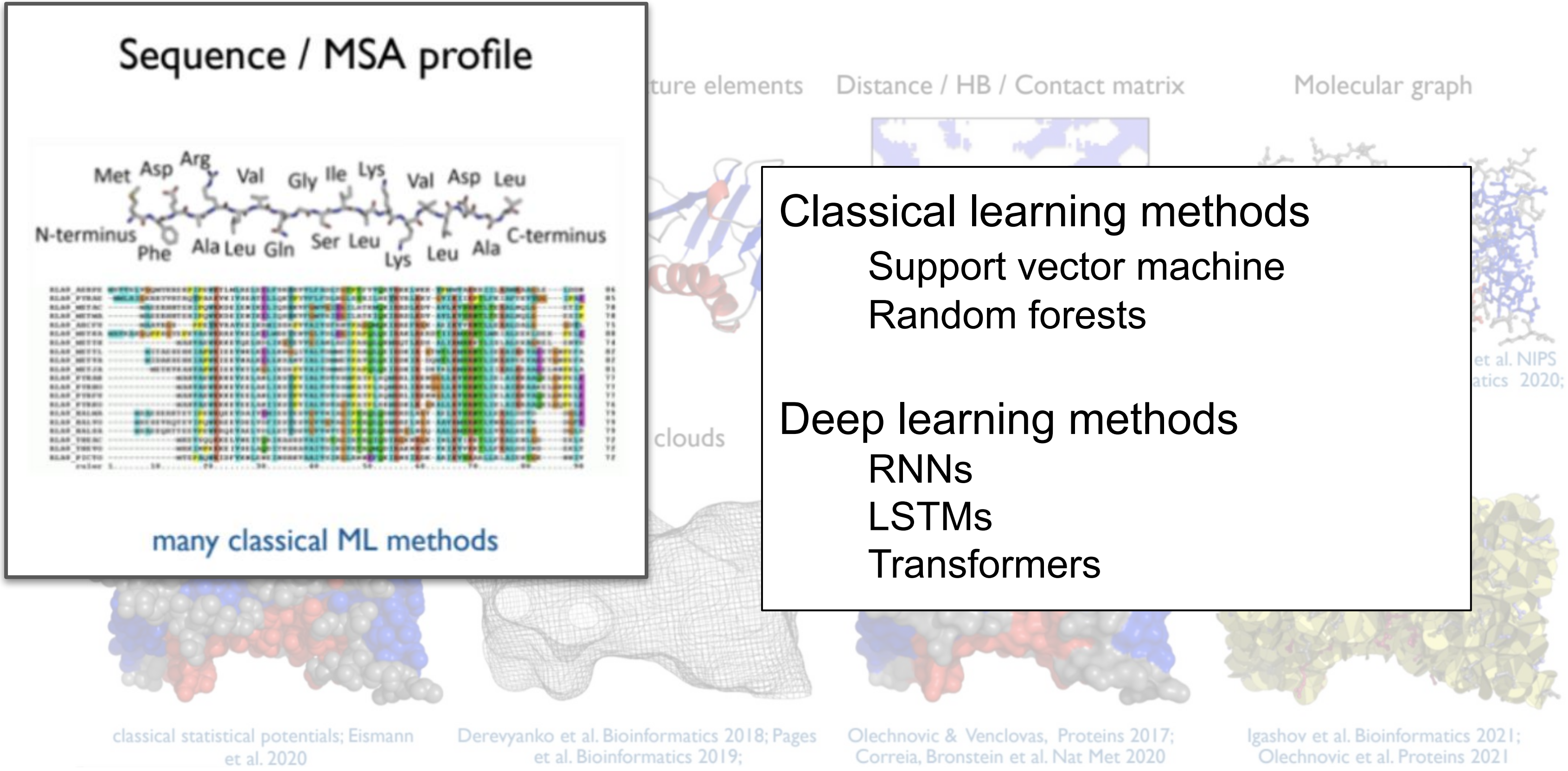
Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020

3D tessellation

Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

Slide from Sergei Grudinin, AlgoSB 2021

Each representation fits a different model better



classical statistical potentials; Eismann et al. 2020

Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;

Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020

Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

Classical learning methods

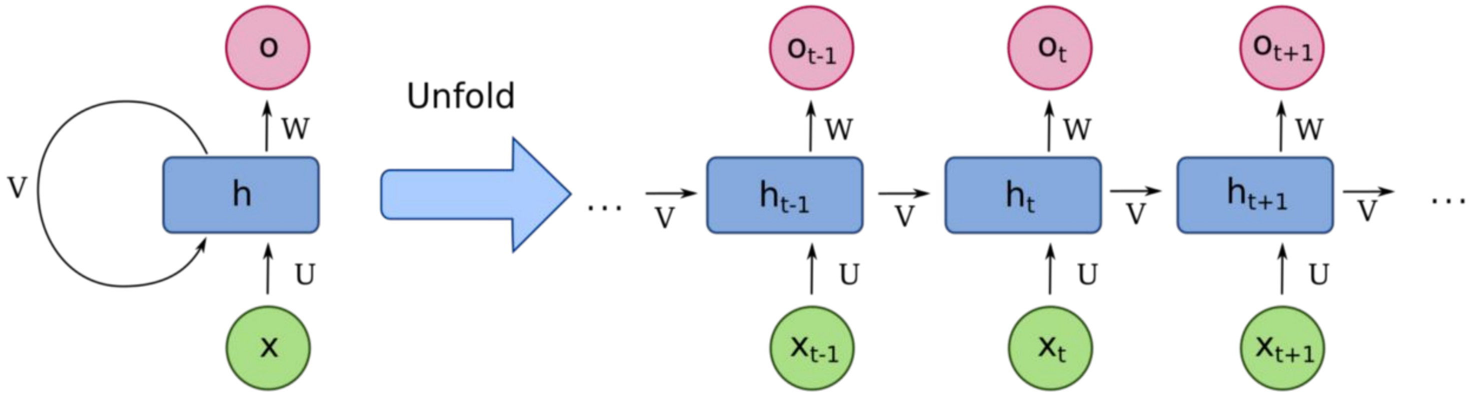
- Support vector machine
- Random forests

Deep learning methods

- RNNs
- LSTMs
- Transformers

Slide from Sergei Grudinin, AlgoSB 2021

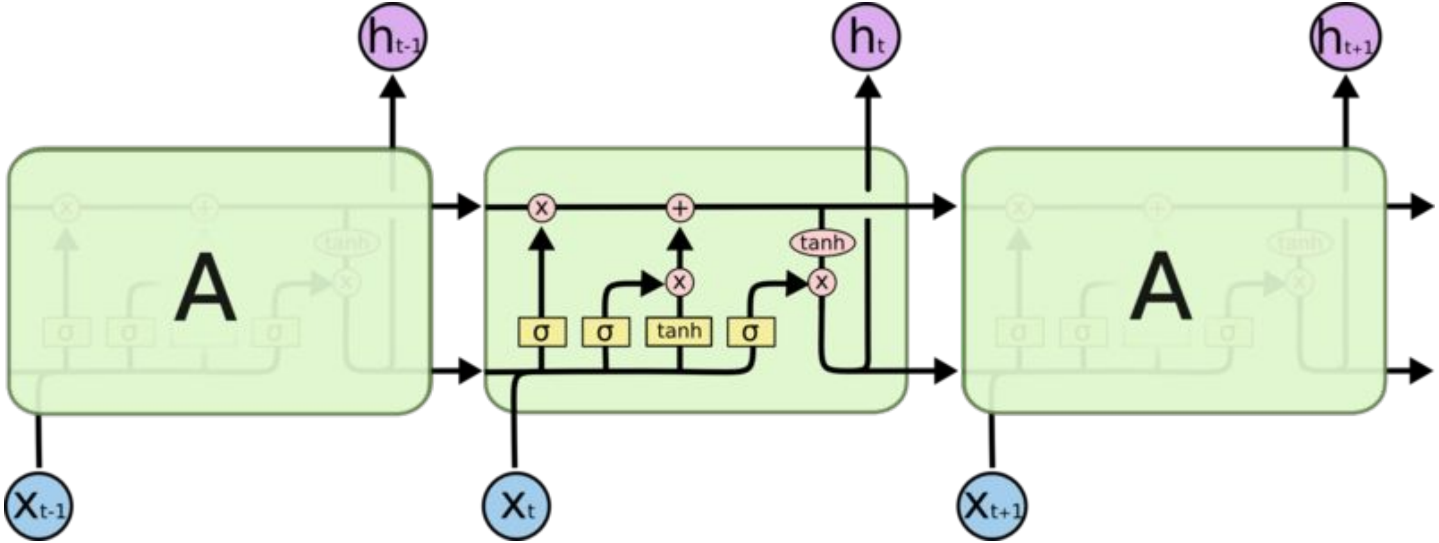
Sequences are natural inputs for NLP (natural language processing) models



Recurrent **N**eural **N**ets (RNNs)

Images from medium post on sequence-based models

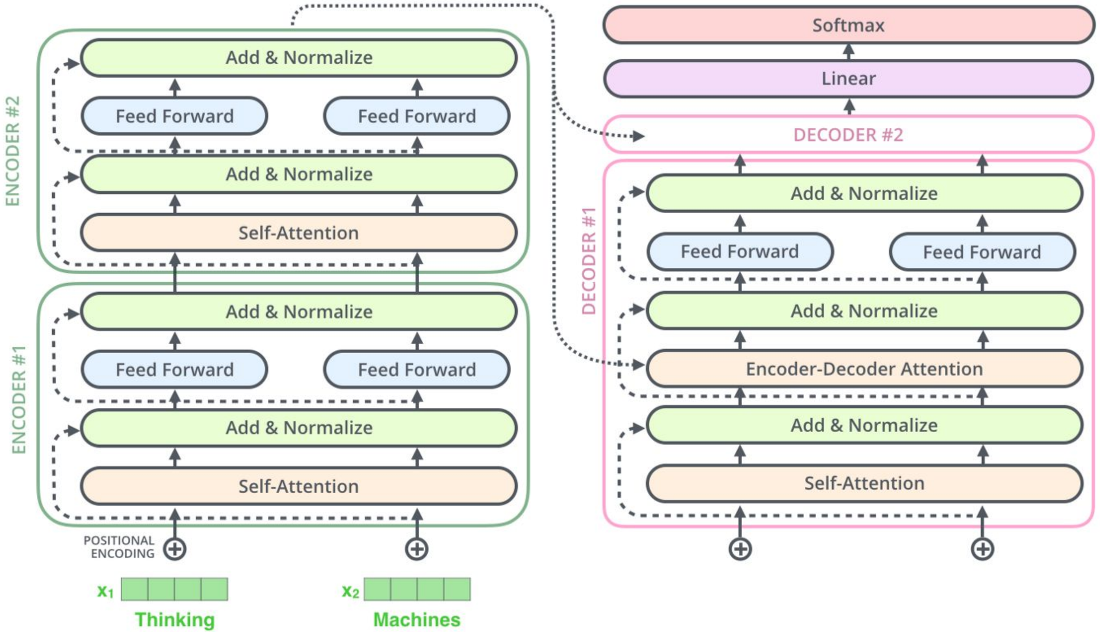
Sequences are natural inputs for NLP (natural language processing) models



Long **S**hort **T**erm **M**emory (LSTM)

Images from medium post on sequence-based models

Sequences are natural inputs for NLP (natural language processing) models

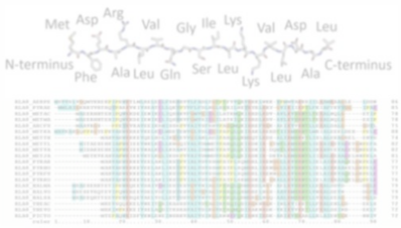


Transformer

Images from medium post on sequence-based models

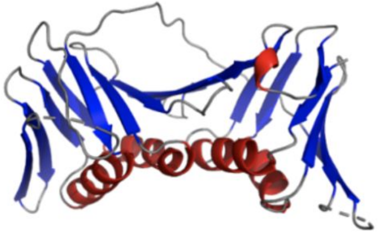
Each representation fits a different model better

Sequence / MSA profile

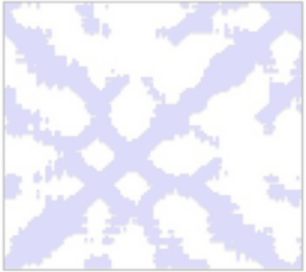


many classical ML methods

Secondary structure elements

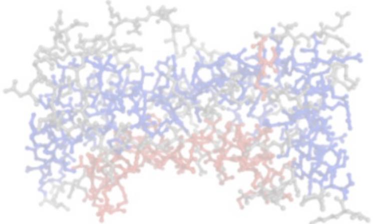


Distance / HB / Contact matrix



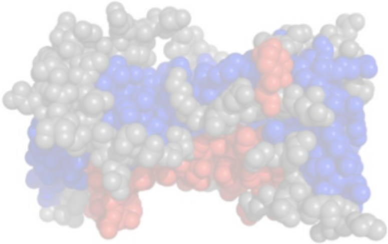
state-of-the-art structure prediction methods

Molecular graph



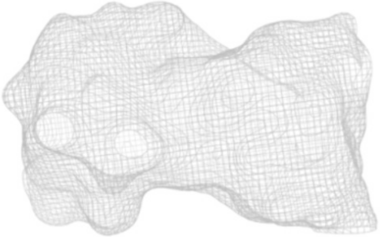
Fout et al. NIPS 2017; Ingraham et al. NIPS 2019; Baldassarre et al. Bioinformatics 2020; Igashov et al. 2020

Set of balls / Point cloud



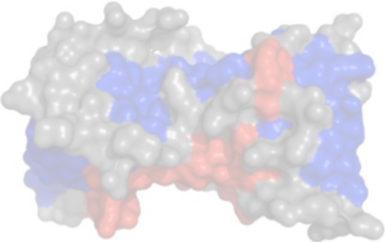
classical statistical potentials; Eismann et al. 2020

Gaussian clouds



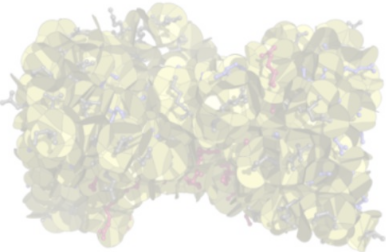
Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;

Molecular surface



Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020

3D tessellation



Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

Slide from Sergei Grudinin, AlgoSB 2021

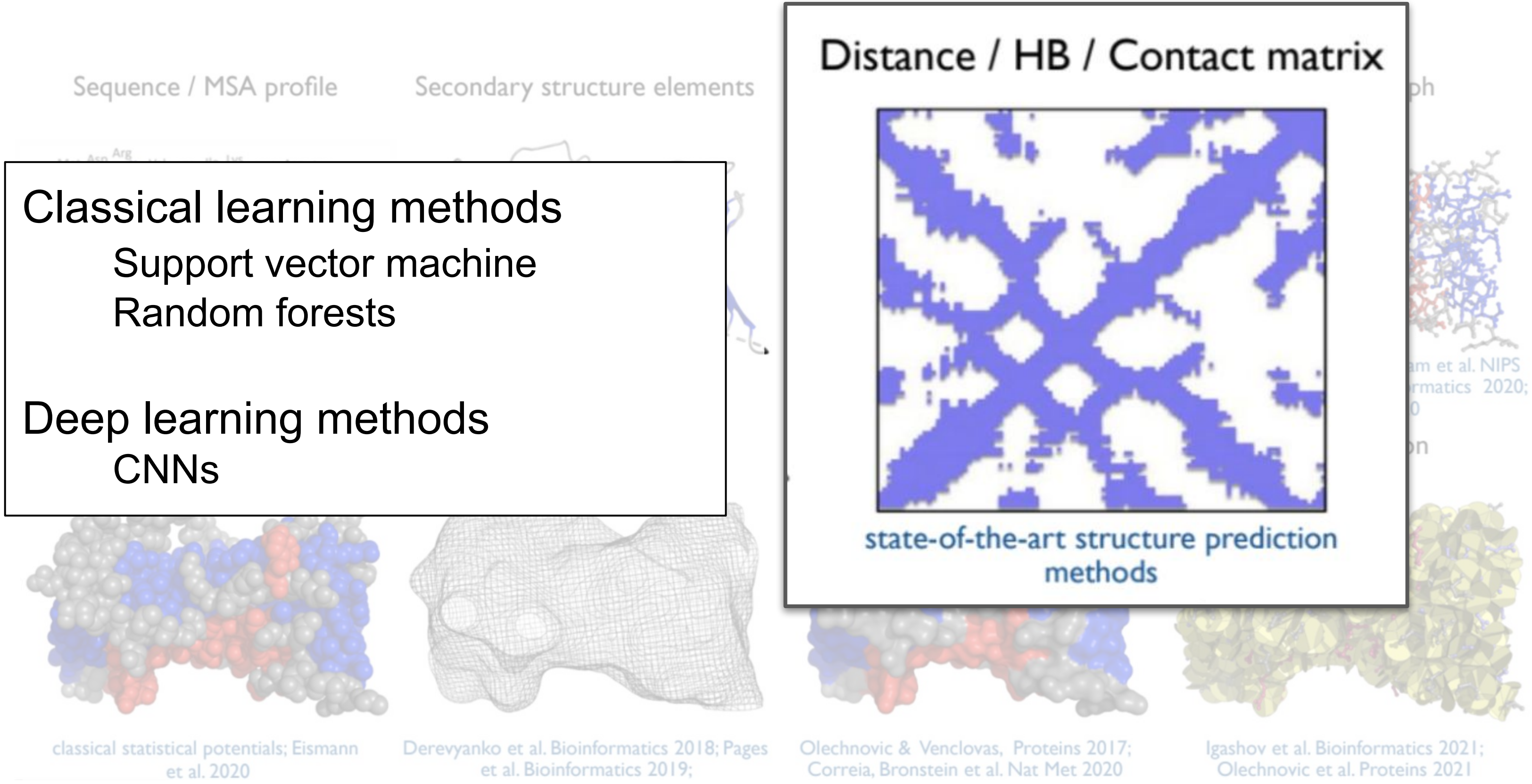
Each representation fits a different model better

Classical learning methods

- Support vector machine
- Random forests

Deep learning methods

- CNNs



Sequence / MSA profile

Secondary structure elements

Distance / HB / Contact matrix

state-of-the-art structure prediction methods

ph

am et al. NIPS
rmatics 2020;
0

on

classical statistical potentials; Eismann et al. 2020

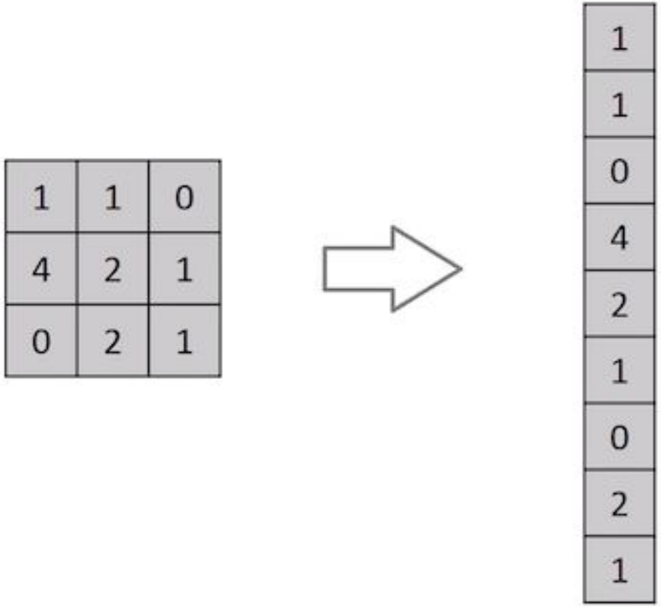
Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;

Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020

Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

Slide from Sergei Grudinin, AlgoSB 2021

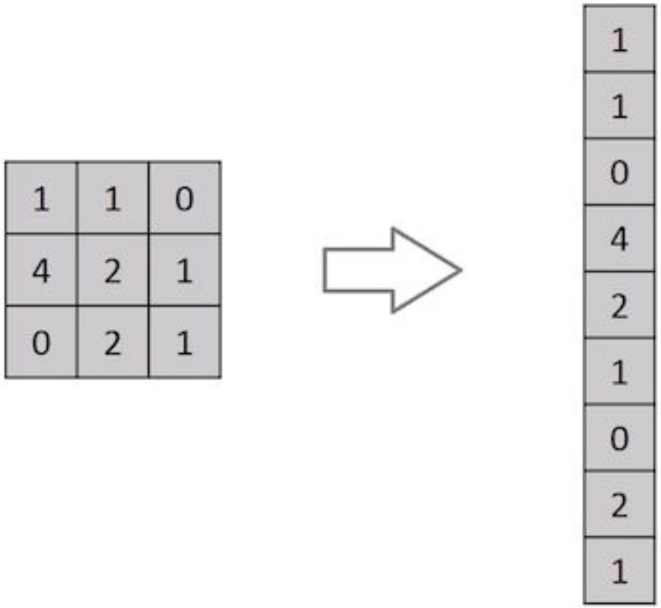
Convolutional neural nets (CNNs) are the preferred methods for images



$$\begin{bmatrix}1 & 1 & 0\\ 4 & 2 & 1\\ 0 & 2 & 1\end{bmatrix} \Rightarrow \begin{bmatrix}1\\ 1\\ 0\\ 4\\ 2\\ 1\\ 0\\ 2\\ 1\end{bmatrix}$$

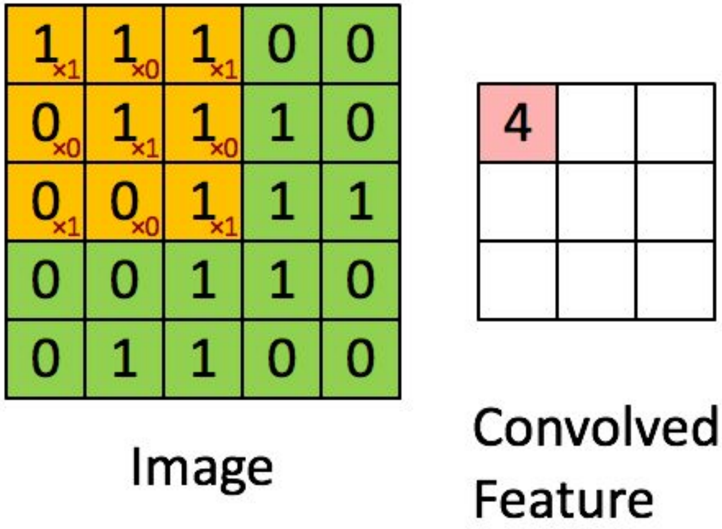
Images from towards data science, a comprehensive guide to CNNs

Convolutional neural nets (CNNs) are the preferred methods for images



1	1	0
4	2	1
0	2	1

1
1
0
4
2
1
0
2
1



$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature

Images from towards data science, a comprehensive guide to CNNs

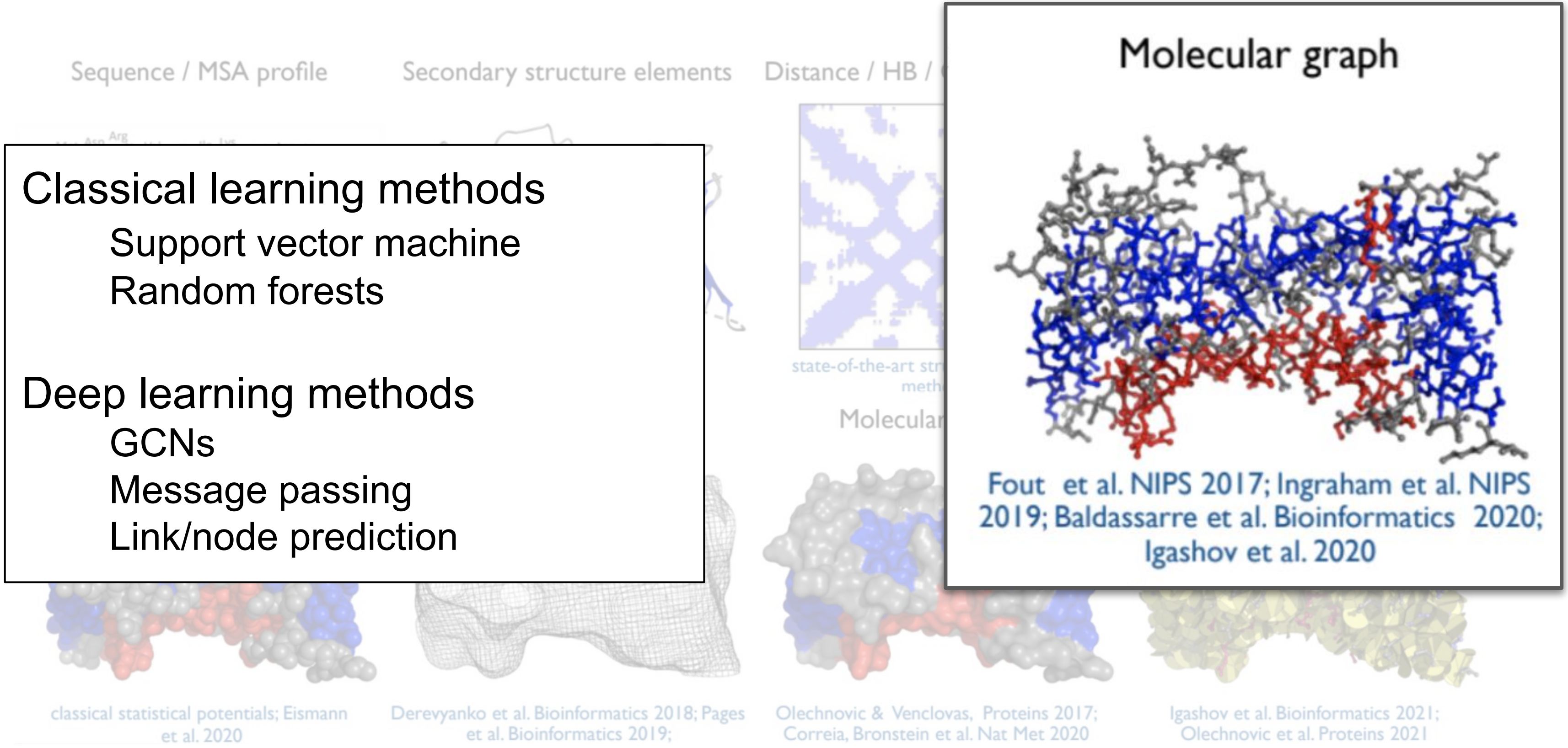
Each representation fits a different model better

Classical learning methods

- Support vector machine
- Random forests

Deep learning methods

- GCNs
- Message passing
- Link/node prediction



Sequence / MSA profile

Secondary structure elements

Distance / HB /

state-of-the-art str
meth

Molecular

Molecular graph

Fout et al. NIPS 2017; Ingraham et al. NIPS 2019; Baldassarre et al. Bioinformatics 2020; Igashov et al. 2020

classical statistical potentials; Eismann et al. 2020

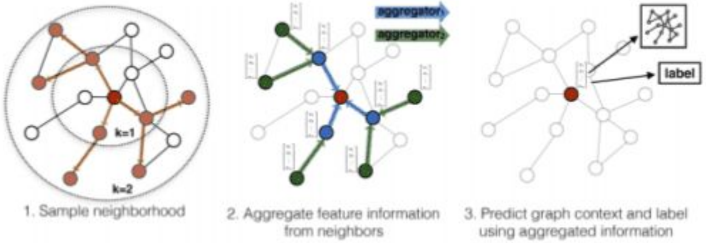
Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;

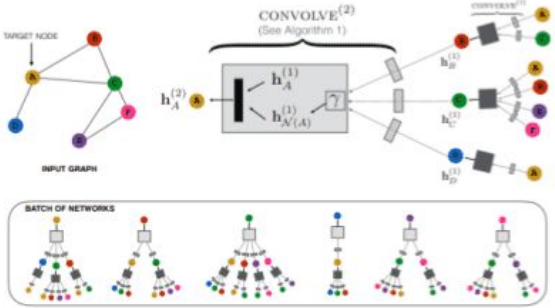
Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020

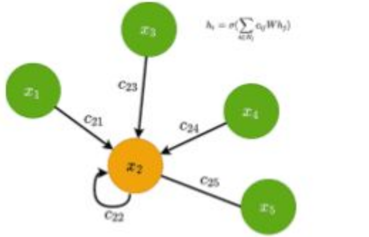
Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

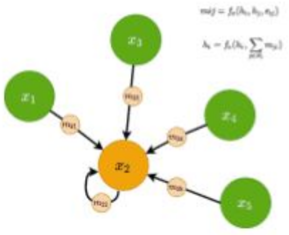
Slide from Sergei Grudinin, AlgoSB 2021

The underlying idea behind convolution on graphs is message passing


k=1
k=2
aggregator1
aggregator2
label
1. Sample neighborhood
2. Aggregate feature information from neighbors
3. Predict graph context and label using aggregated information






$mij = f_e(h_i, h_j, e_{ij})$
$h_i = f_v(h_i, \sum_{j \in N_i} m_{ji})$
x_1
x_2
x_3
x_4
x_5

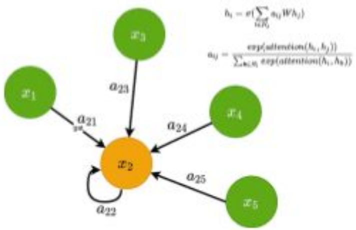

$h_i = \sigma(\sum_{j \in N_i} a_{ij} W h_j)$
$a_{ij} = \frac{exp(attention(h_i, h_j))}{\sum_{k \in N_i} exp(attention(h_i, h_k))}$
x_1
x_2
x_3
x_4
x_5
a_{21}
a_{22}
a_{23}
a_{24}
a_{25}

Image from AI summer

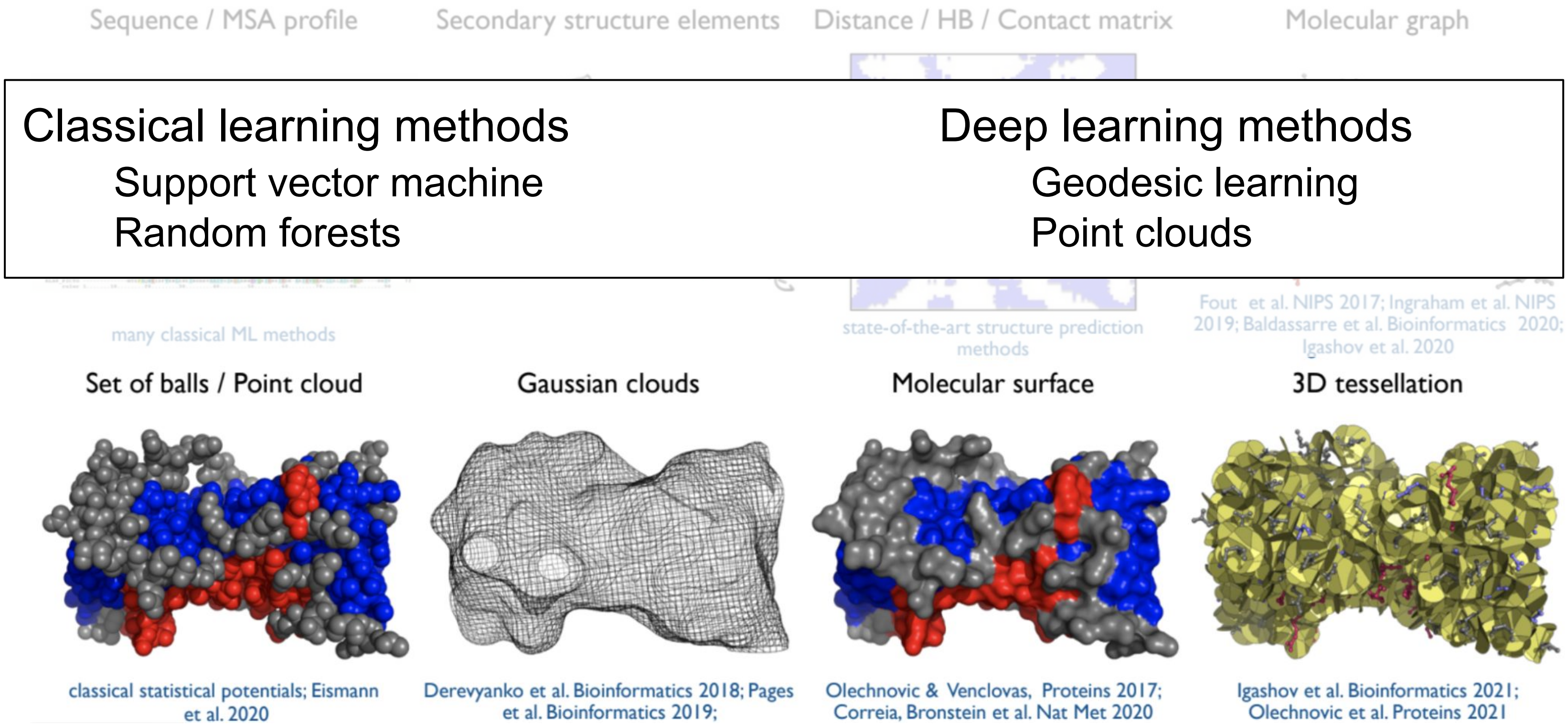
Each representation fits a different model better

Classical learning methods

- Support vector machine
- Random forests

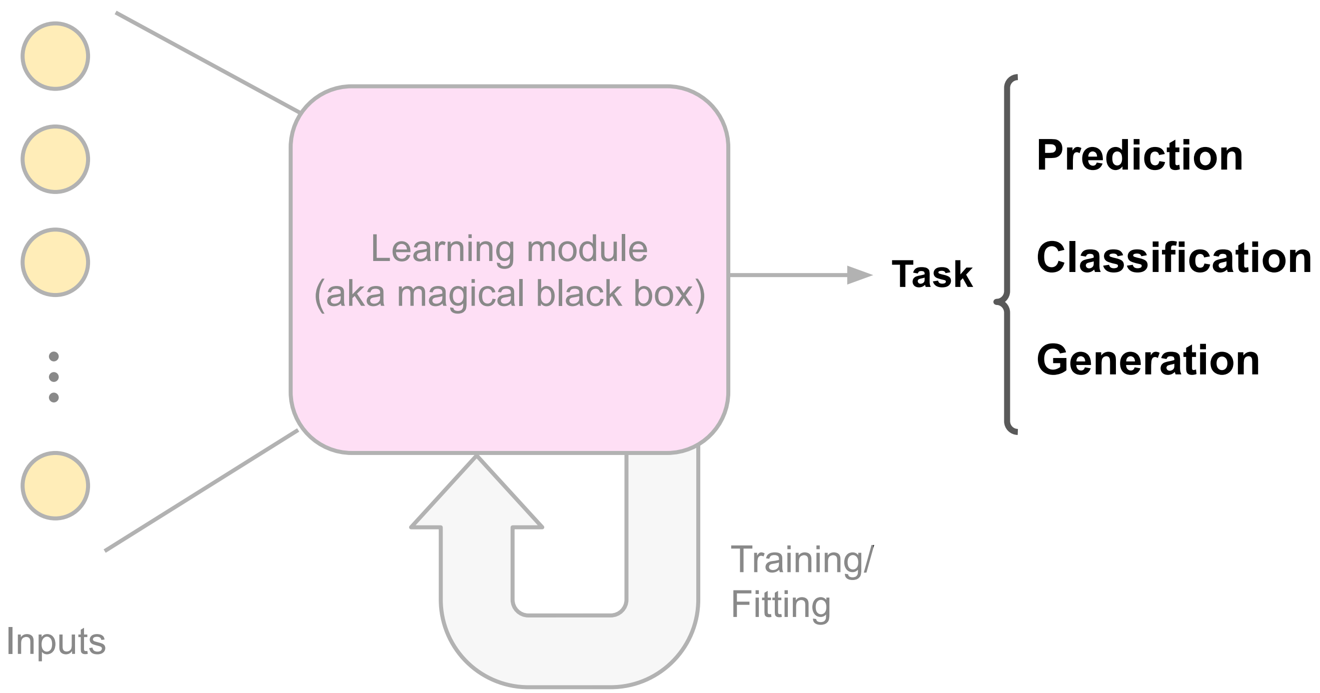
Deep learning methods

- Geodesic learning
- Point clouds


Sequence / MSA profile
Secondary structure elements
Distance / HB / Contact matrix
Molecular graph
many classical ML methods
state-of-the-art structure prediction methods
Fout et al. NIPS 2017; Ingraham et al. NIPS 2019; Baldassarre et al. Bioinformatics 2020; Igashov et al. 2020
Set of balls / Point cloud
Gaussian clouds
Molecular surface
3D tessellation
classical statistical potentials; Eismann et al. 2020
Derevyanko et al. Bioinformatics 2018; Pages et al. Bioinformatics 2019;
Olechnovic & Venclovas, Proteins 2017; Correia, Bronstein et al. Nat Met 2020
Igashov et al. Bioinformatics 2021; Olechnovic et al. Proteins 2021

Slide from Sergei Grudinin, AlgoSB 2021

Basic components of a learning module



ML applications in protein engineering

- Machine learning guided directed evolution
- Representation learning
- Unsupervised variant prediction
- Generative models
- Predicting stability
- Predicting structure from sequence
- Predicting interactions with other molecules
- Classification and annotation
- Predicting sequence from structure

Classification adopted from @yangkky github

ML applications in protein engineering

Machine learning guided directed evolution

Representation learning

Unsupervised variant prediction

Generative models

Predicting stability

Predicting structure from sequence

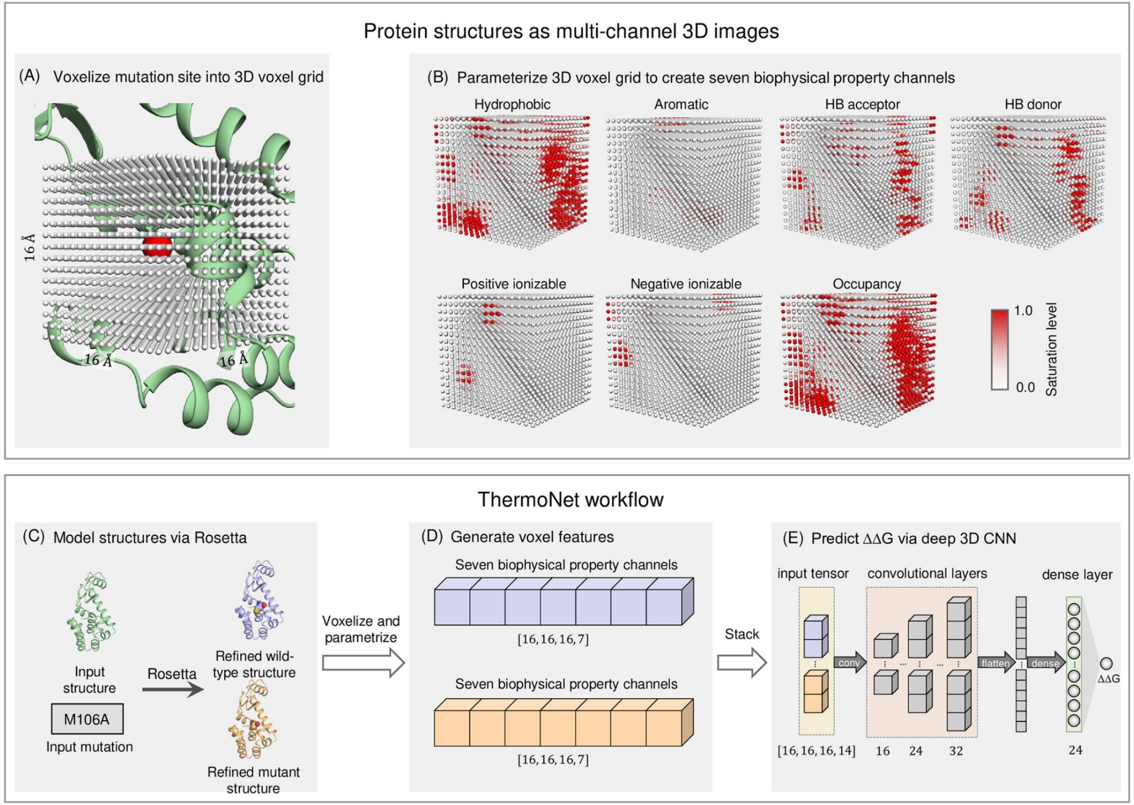
Predicting interactions with other molecules

Classification and annotation

Predicting sequence from structure

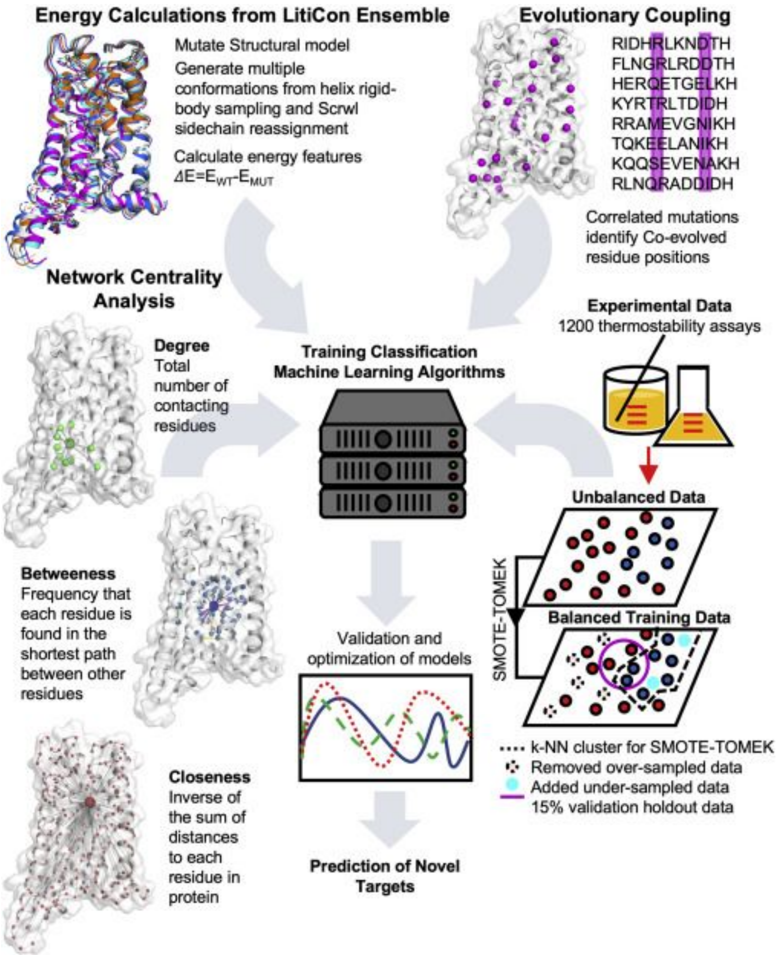
Classification adopted from @yangkky github

3D CNN for prediction of stability


Protein structures as multi-channel 3D images
(A) Voxelize mutation site into 3D voxel grid
16 Å
16 Å
16 Å
(B) Parameterize 3D voxel grid to create seven biophysical property channels
Hydrophobic
Aromatic
HB acceptor
HB donor
Positive ionizable
Negative ionizable
Occupancy
1.0
0.0
Saturation level
ThermoNet workflow
(C) Model structures via Rosetta
Input structure
Rosetta
Refined wild-type structure
M106A
Input mutation
Refined mutant structure
Voxelize and parametrize
(D) Generate voxel features
Seven biophysical property channels
[16, 16, 16, 7]
Seven biophysical property channels
[16, 16, 16, 7]
Stack
(E) Predict ΔΔG via deep 3D CNN
input tensor
convolutional layers
dense layer
conv
flatten
dense
ΔΔG
[16, 16, 16, 14]
16
24
32
24

Plos Comp Biol November 30, **2020**

ML-based models for the specific case of GPCRs


Energy Calculations from LitiCon Ensemble
Mutate Structural model
Generate multiple conformations from helix rigid-body sampling and Scrwl sidechain reassignment
Calculate energy features
$\Delta E = E_{WT} - E_{MUT}$
Evolutionary Coupling
RIDHRLKNDTH
FLNGRLRDDTH
HERQETGELKH
KYRTRLTDIDH
RRAMEVGNIKH
TQKEELANIKH
KQQSEVENAKH
RLNQRADDIDH
Correlated mutations identify Co-evolved residue positions
Network Centrality Analysis
Degree
Total number of contacting residues
Training Classification Machine Learning Algorithms
Experimental Data
1200 thermostability assays
Unbalanced Data
SMOTE-TOMEK
Balanced Training Data
Betweeness
Frequency that each residue is found in the shortest path between other residues
Validation and optimization of models
k-NN cluster for SMOTE-TOMEK
Removed over-sampled data
Added under-sampled data
15% validation holdout data
Closeness
Inverse of the sum of distances to each residue in protein
Prediction of Novel Targets

Biophysical journal, volume 117, issue 11, P2228-2239, December 03, **2019**

ML applications in protein engineering

Machine learning guided directed evolution

Representation learning

Unsupervised variant prediction

Generative models

Predicting stability

Predicting structure from sequence

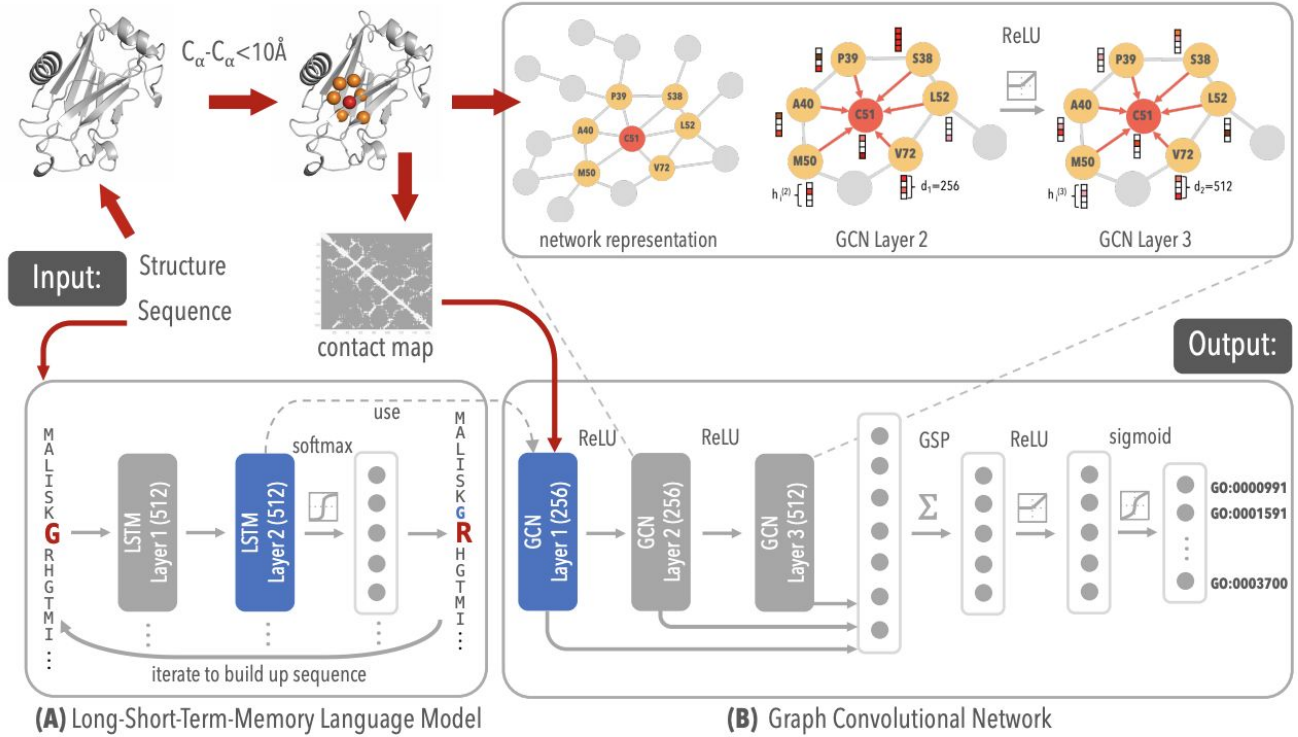
Predicting interactions with other molecules

Classification and annotation

Predicting sequence from structure

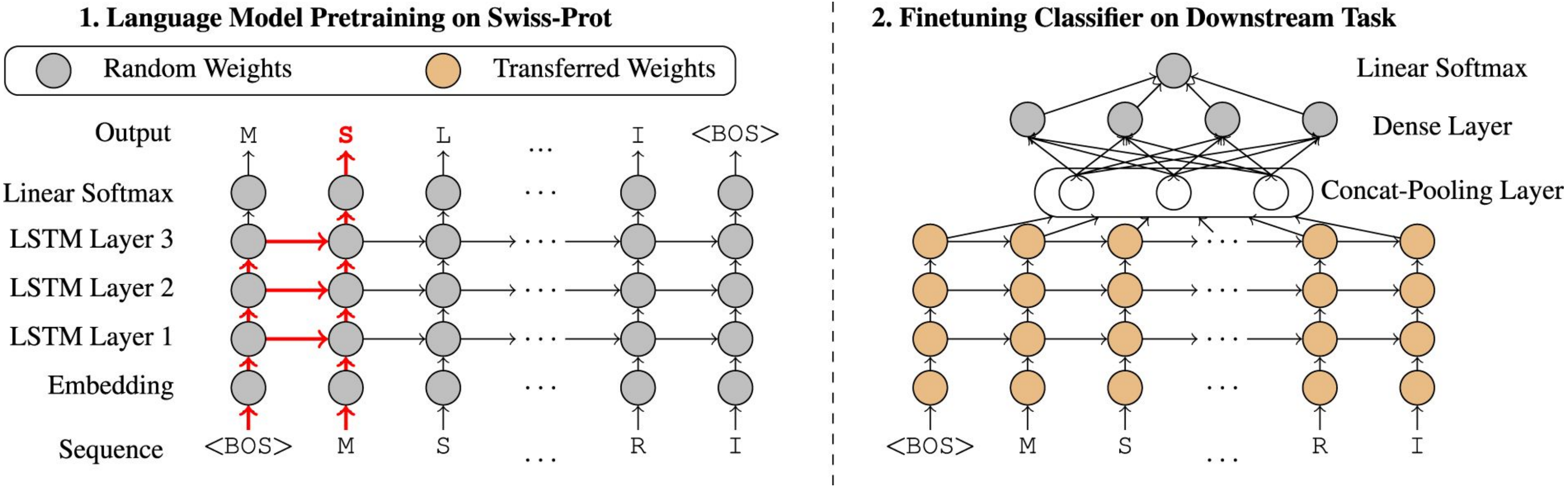
Classification adopted from @yangkky github

Using GCNs for function prediction



Nature Communications **2021**, 12

Deep sequence model for protein classification



Bioinformatics, Volume 36, Issue 8, 15 April **2020**, Pages 2401–2409

ML applications in protein engineering

Machine learning guided directed evolution

Generative models

Predicting interactions with other molecules

Representation learning

Predicting stability

Classification and annotation

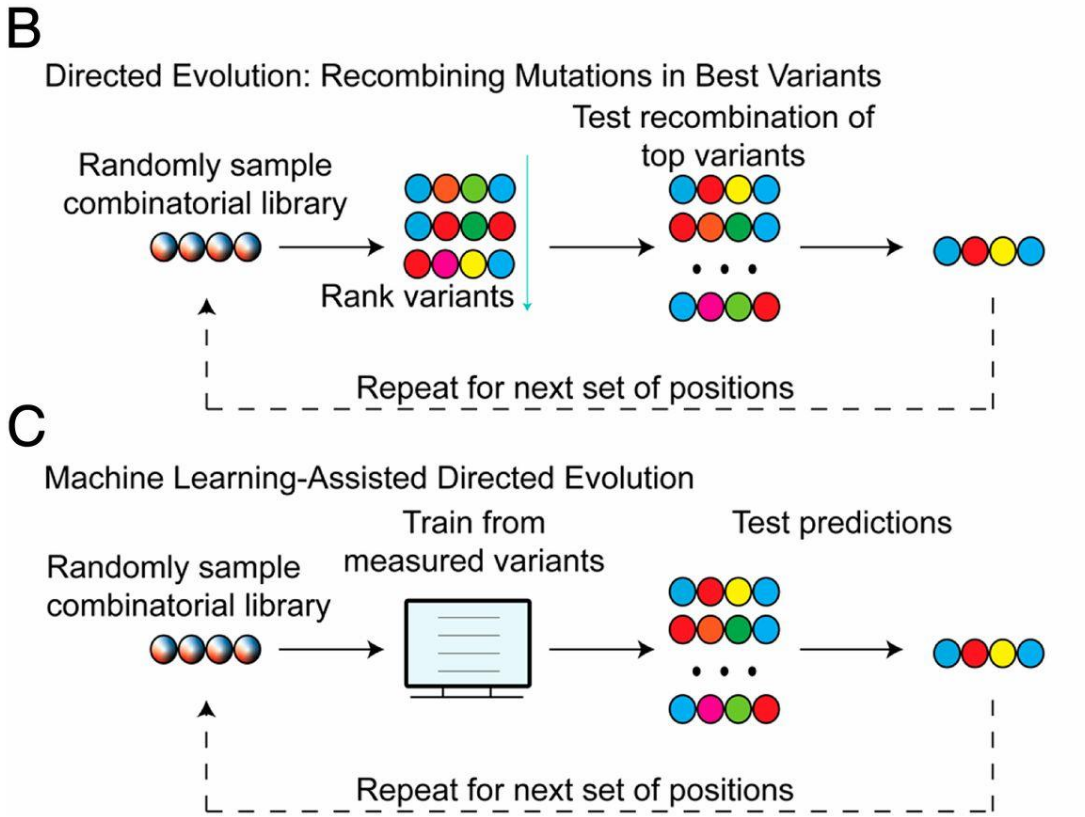
Unsupervised variant prediction

Predicting structure from sequence

Predicting sequence from structure

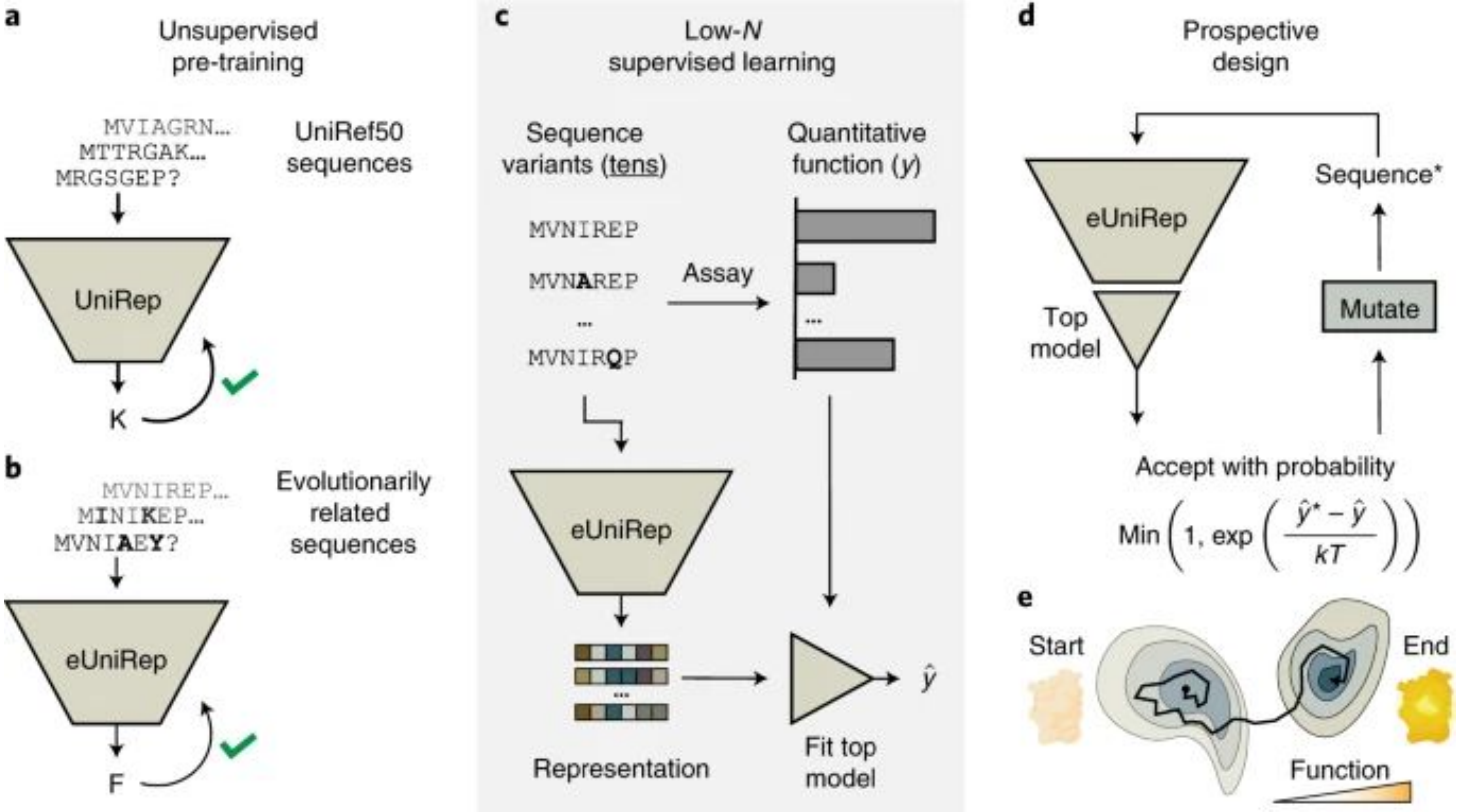
Classification adopted from @yangkky github

Machine learning assisted directed evolution


B
Directed Evolution: Recombining Mutations in Best Variants
Randomly sample combinatorial library
Rank variants
Test recombination of top variants
Repeat for next set of positions
C
Machine Learning-Assisted Directed Evolution
Randomly sample combinatorial library
Train from measured variants
Test predictions
Repeat for next set of positions

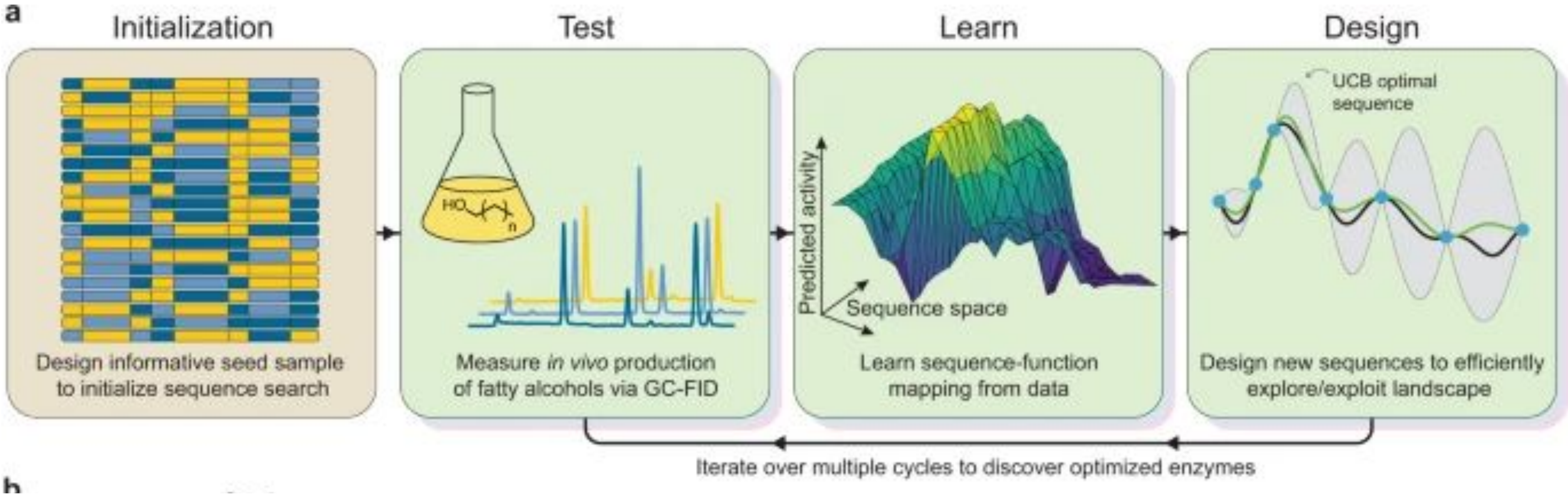
PNAS April 30, **2019** 116 (18) 8852-8858

low-N protein engineering



Nature Methods **2021** 18, 386-396

ML-accelerated protein sequence optimization



Initialization

Design informative seed sample to initialize sequence search

Test

Measure *in vivo* production of fatty alcohols via GC-FID

Learn

Learn sequence-function mapping from data

Design

Design new sequences to efficiently explore/exploit landscape

Iterate over multiple cycles to discover optimized enzymes

Nature Communications **2021** 12, 5825

ML applications in protein engineering

Machine learning guided directed evolution

Generative models

Predicting interactions with other molecules

Representation learning

Predicting stability

Classification and annotation

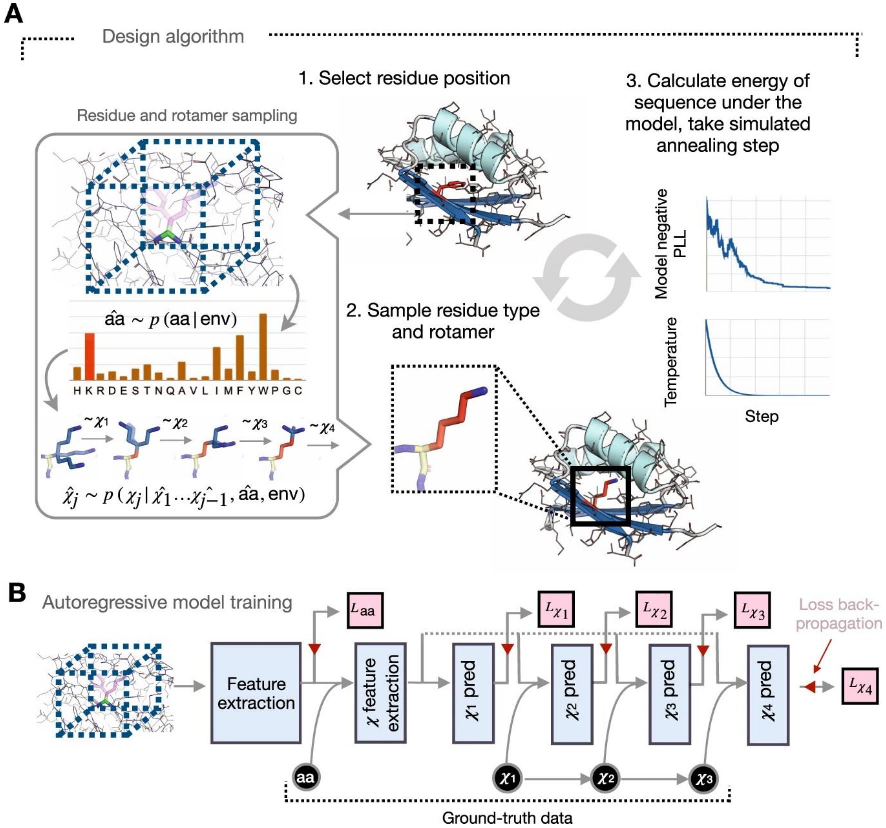
Unsupervised variant prediction

Predicting structure from sequence

Predicting sequence from structure

Classification adopted from @yangkky github

Sequence design with a learned potential


A
Design algorithm
1. Select residue position
Residue and rotamer sampling
3. Calculate energy of sequence under the model, take simulated annealing step
$\hat{aa} \sim p(aa|env)$
H K R D E S T N Q A V L I M F Y W P G C
$\sim\chi_1$
$\sim\chi_2$
$\sim\chi_3$
$\sim\chi_4$
$\hat{\chi}_j \sim p(\chi_j|\hat{\chi}_1 \cdots \hat{\chi}_{j-1}, \hat{aa}, env)$
2. Sample residue type and rotamer
Model negative PLL
Temperature
Step
B
Autoregressive model training
Feature extraction
χ feature extraction
χ_1 pred
χ_2 pred
χ_3 pred
χ_4 pred
L_{aa}
$L_{\chi 1}$
$L_{\chi 2}$
$L_{\chi 3}$
$L_{\chi 4}$
Loss back-propagation
aa
χ_1
χ_2
χ_3
Ground-truth data

Nature Communications **2022** 13, 746

Autoregressive models

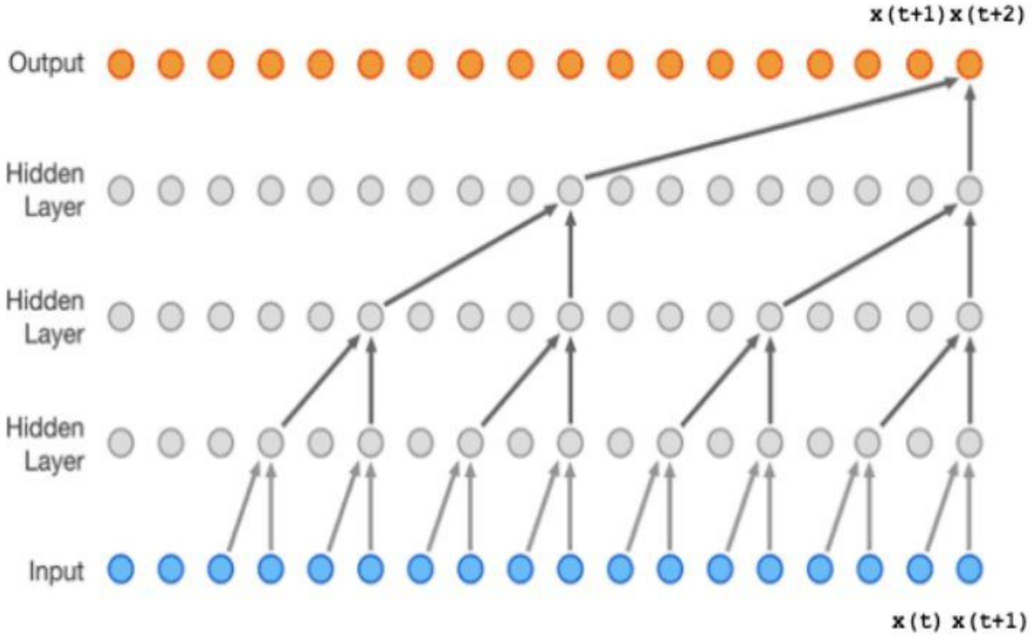
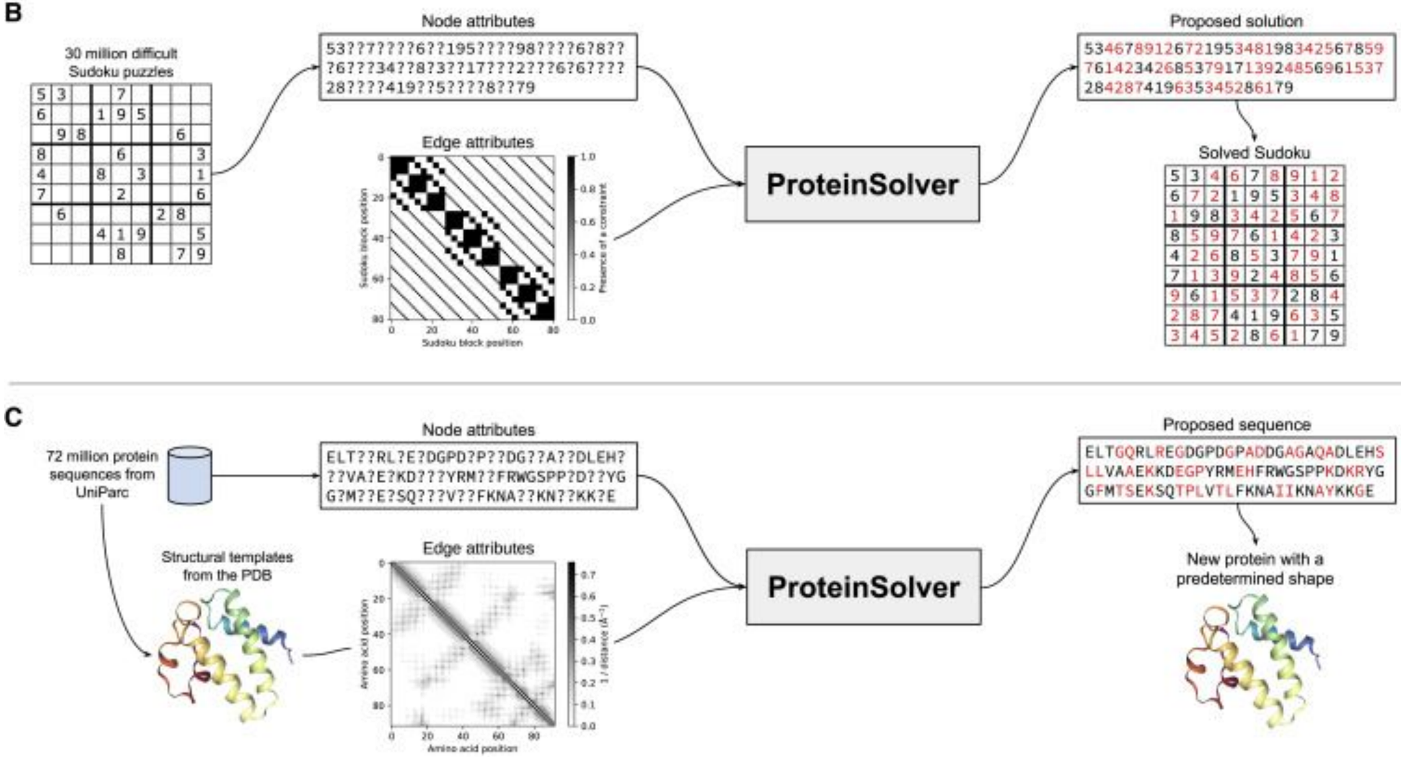


Image from DeepMind, Google's WaveNet

Deep GCNs to design protein sequences



5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

```
53??7????6??195????98????6?8??
?6???34??8?3??17???2???6?6????
28????419??5????8??79
```

```
534678912672195348198342567859
761423426853791713924856961537
284287419635345286179
```

5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

```
ELT??RL?E?DGPD?P??DG??A??DLEH?
??VA?E?KD???YRM??FRWGSPP?D??YG
G?M??E?SQ???V??FKNA??KN??KK?E
```

```
ELTGQRLREGDGPDGPADDGAGAQADLEHS
LLVAAEKKDEGPYRMEHFRWGSPPKDKRYG
GFMTSEKSQTPLVTLFKNAIIKNAYKKGE
```

Cell systems **2020**, 11(4)

ML applications in protein engineering

Machine learning guided directed evolution

Generative models

Predicting interactions with other molecules

Representation learning

Predicting stability

Classification and annotation

Unsupervised variant prediction

Predicting structure from sequence

Predicting sequence from structure

Classification adopted from @yangkky github

Variational Autoencoders (VAEs)

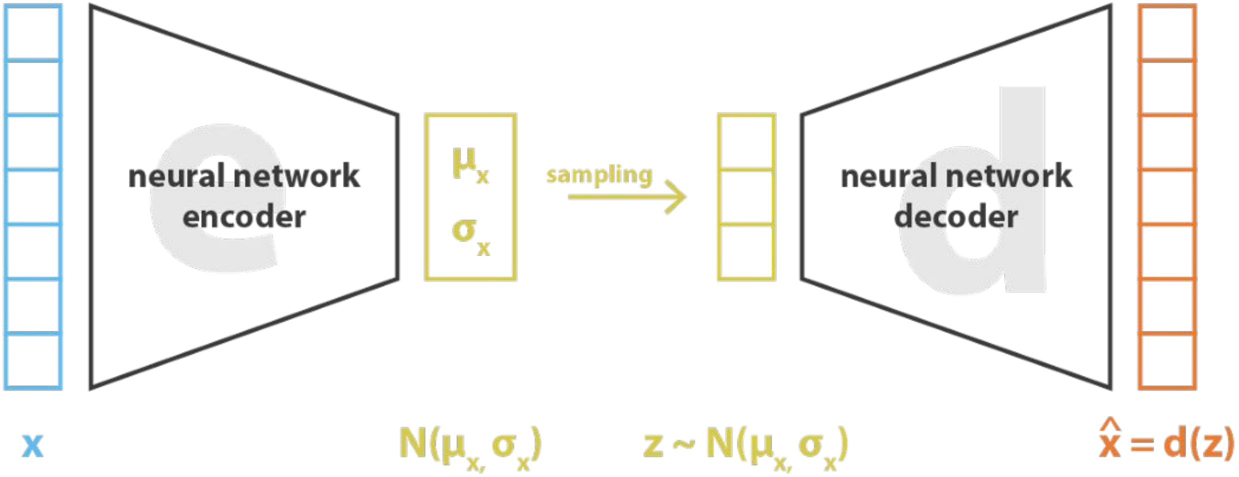
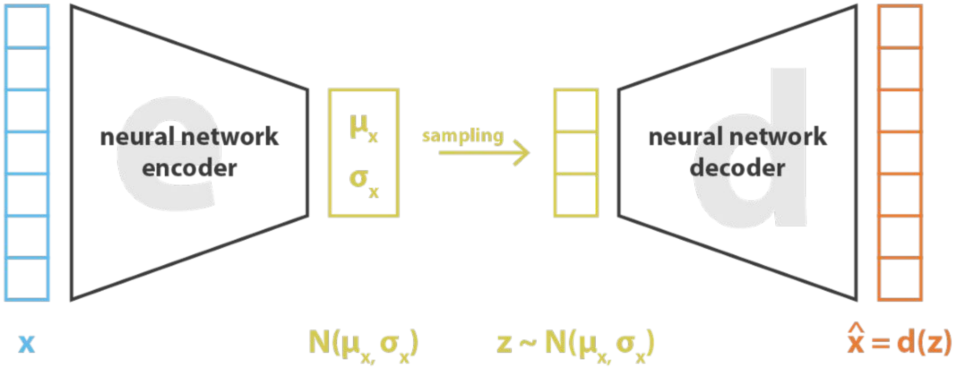


Image from towards data science, understanding VAEs

Variational Autoencoders (VAEs)



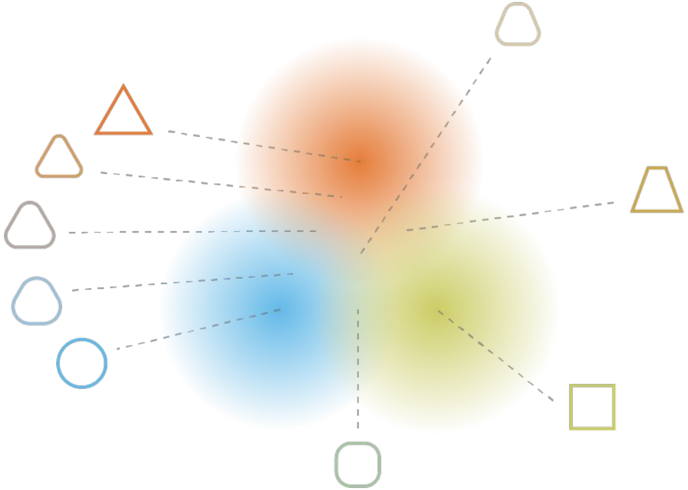
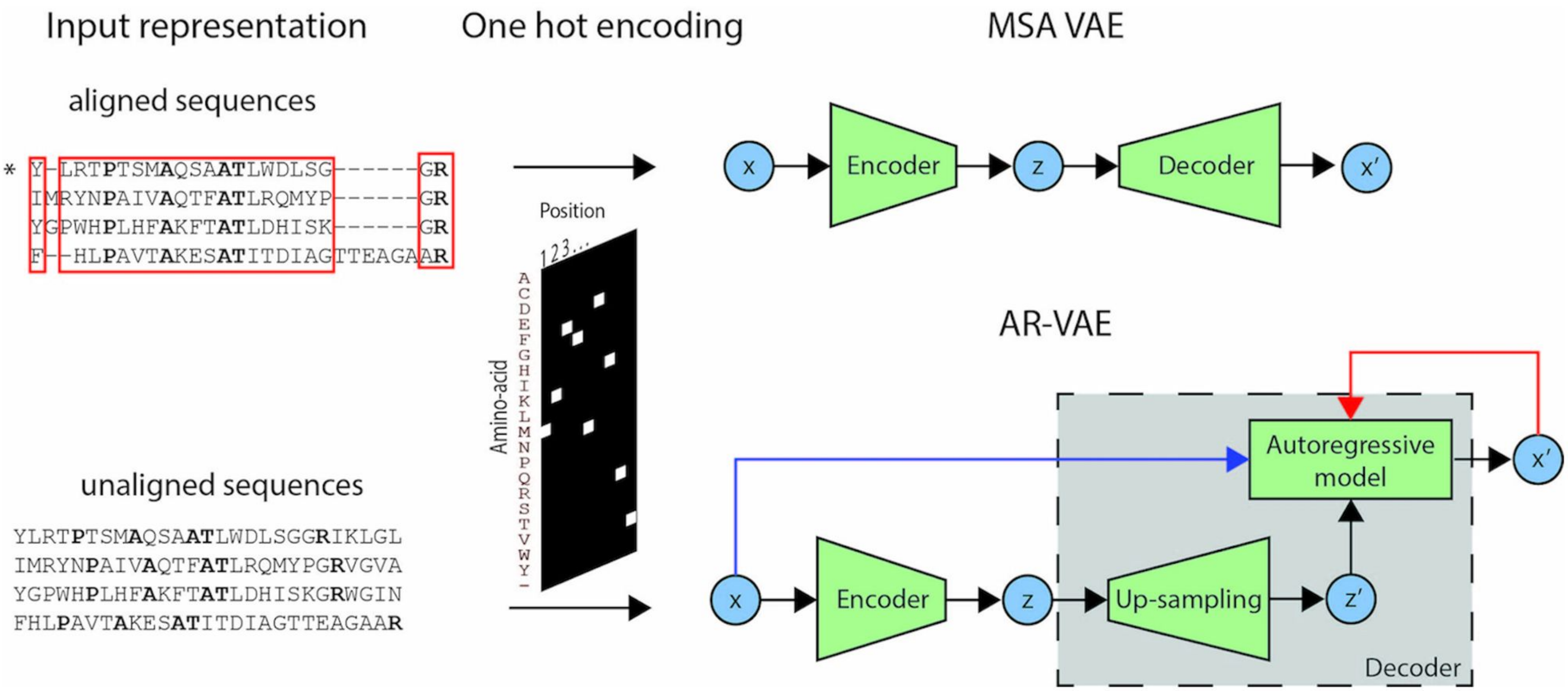


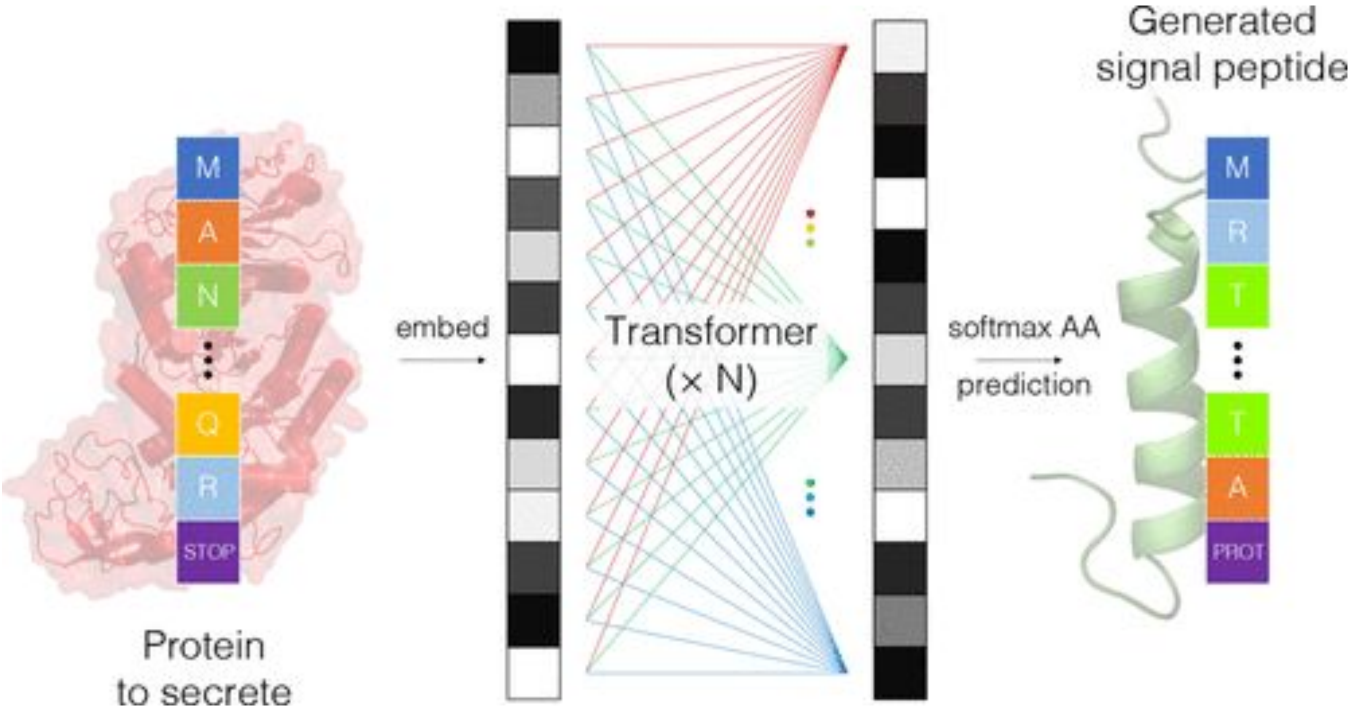
Image from towards data science, understanding VAEs

VAEs for protein design


Input representation
One hot encoding
MSA VAE
aligned sequences
* Y-LRTPTSMAQSAATLWDLSG------GR
IMRYNPAIVAQTFATLRQMYP------GR
YGPWHPLHFAKFTATLDHISK------GR
F--HLPAVTAKESATITDIAGTTEAGAAR
Position
1 2 3...
Amino-acid
A C D E F G H I K L M N P Q R S T V W Y -
x
Encoder
z
Decoder
x′
AR-VAE
unaligned sequences
YLRTPTSMAQSAATLWDLSGGRIKLGL
IMRYNPAIVAQTFATLRQMYPGRVGVA
YGPWHPLHFAKFTATLDHISKGRWGIN
FHLPAVTAKESATITDIAGTTEAGAAR
Autoregressive model
x′
x
Encoder
z
Up-sampling
z′
Decoder

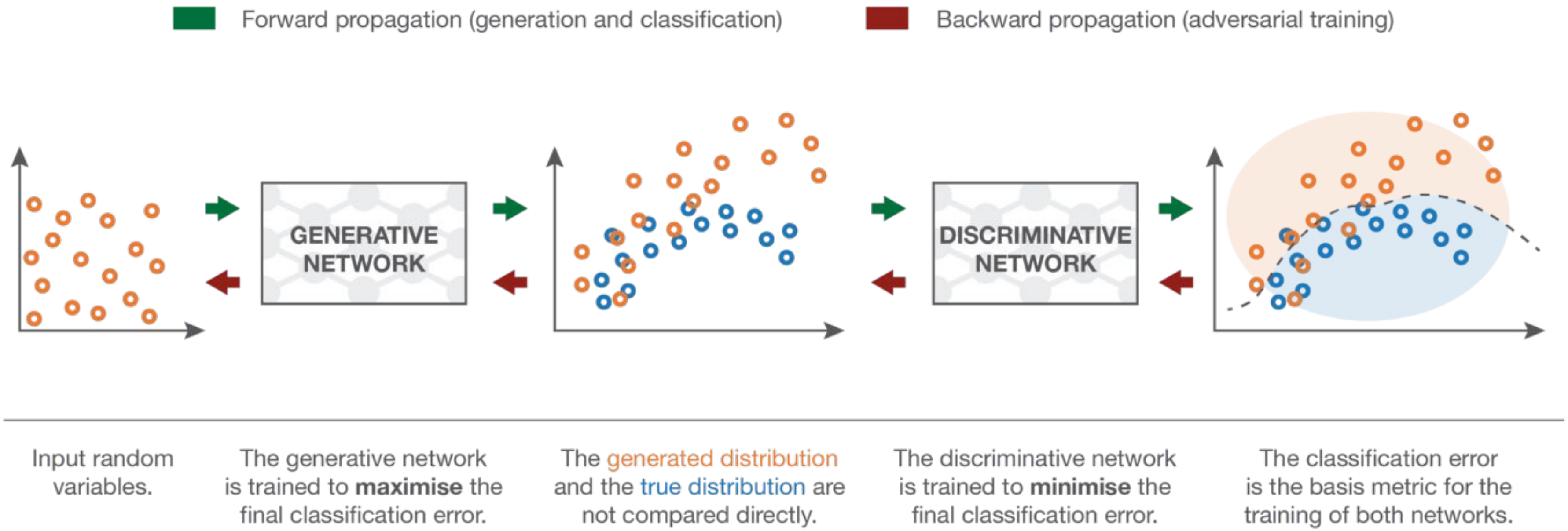
Plos Comp Biol **2021**

Transformers for translating from protein to sequence space



ACS Synth Biol **2020** 9(8), 2154-2161

Generative Adversarial Networks (GANs)



🟩 Forward propagation (generation and classification) 🟥 Backward propagation (adversarial training)

GENERATIVE NETWORK

DISCRIMINATIVE NETWORK

Input random variables.

The generative network is trained to **maximise** the final classification error.

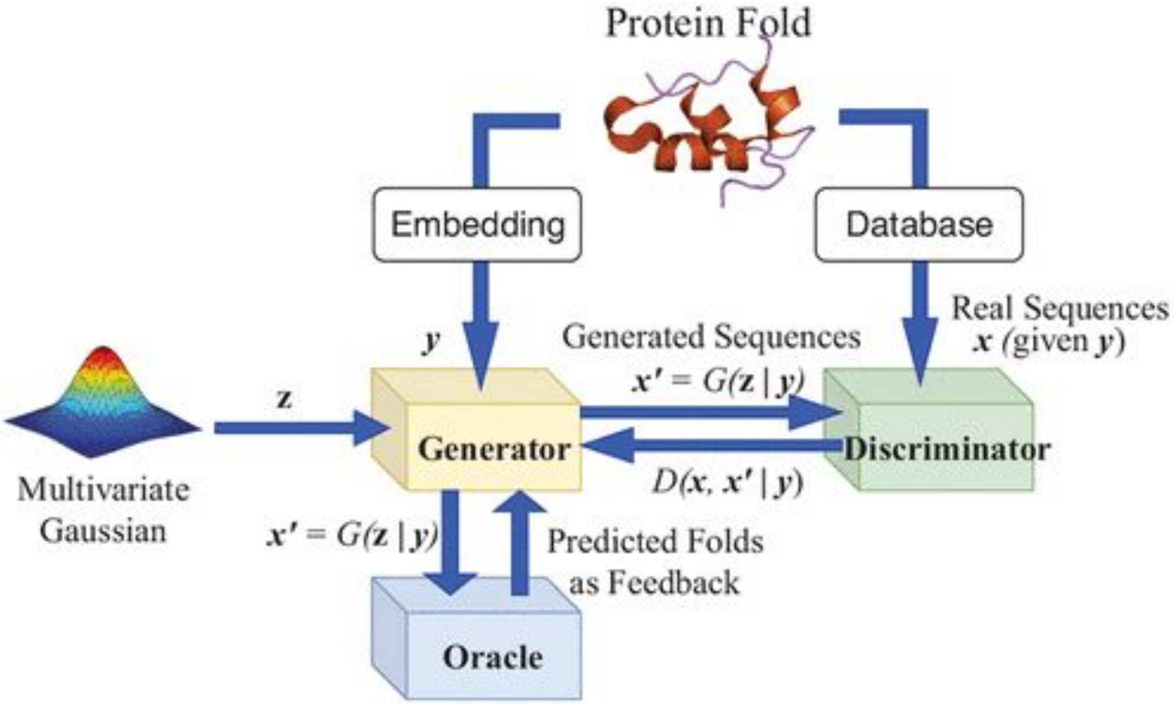
The generated distribution and the true distribution are not compared directly.

The discriminative network is trained to **minimise** the final classification error.

The classification error is the basis metric for the training of both networks.

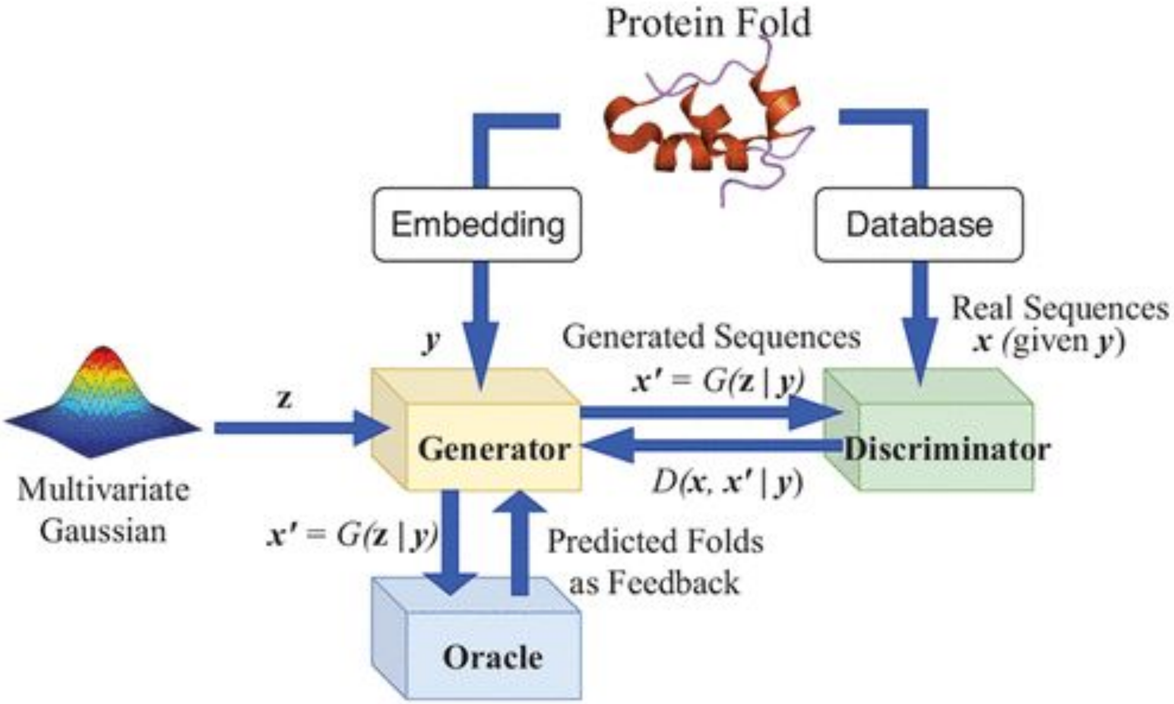
Image from towards data science,, Understanding GANs

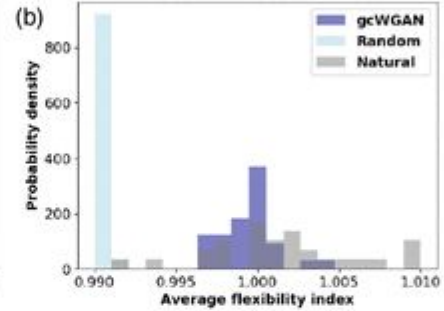
GANs for protein design



J Chem Info Med **2020**, 60,12, 5667-5681

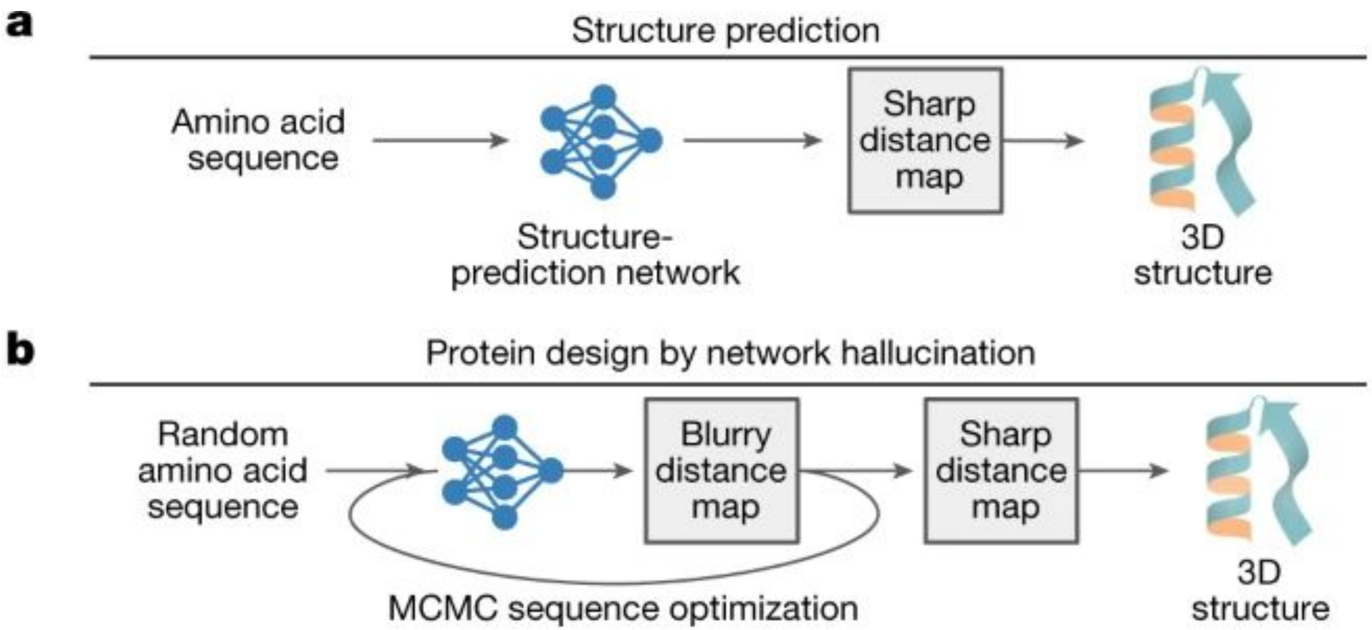
GANs for protein design





J Chem Info Med **2020**, 60,12, 5667-5681

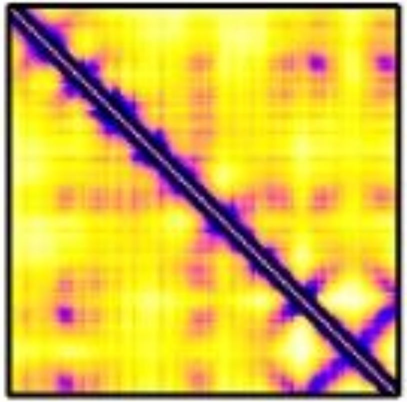
Protein design by reversing protein structure prediction

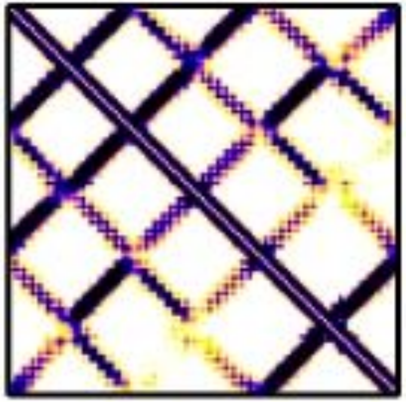


Nature **2021** 600, 547-552

In class activity:

Design new proteins using deep hallucination





Nature **2021** 600, 547-552

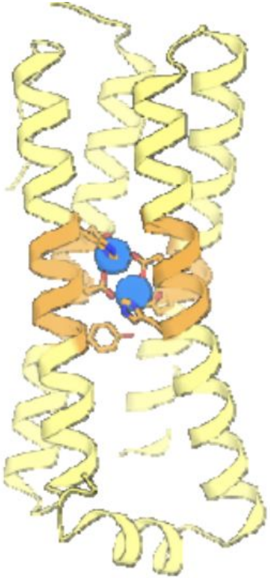
For the next lecture:

1. Read journal for the next lecture
    a. Moderated by **group IV**

2. W7L2 assignment due next lecture

Next lecture:

Design of proteins with functional sites



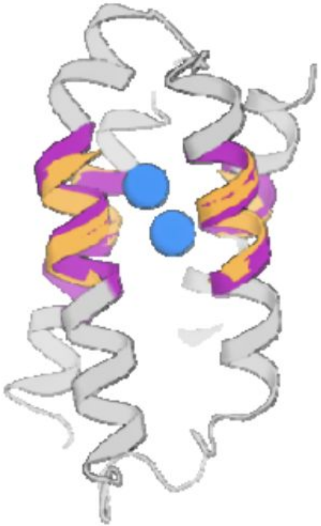


Image from Wang et al, 2021 bioRxiv