ITERATIVE REFINEMENT GRAPH NEURAL NETWORK FOR ANTIBODY SEQUENCE-STRUCTURE CO-DESIGN

2022-06-07 / JiWung Han

Department of Artificial Intelligence
Korea University

MAI Lab
EDICAL

Abstract

◆ **Abstract**

- Antibody → bind to pathogens like viruses and stimulate the adaptive immune system
- The specificity of antibody binding → complementarity-determining regions (CDRs) at the tips of these Y-shaped proteins.
- In this paper,
  1. Propose a generative model to **automatically design the CDRs of antibodies** with enhanced binding specificity or neutralization capabilities
  2. Propose to co-design the sequence and 3D structure of CDRs as graphs

Abstract

◆ Abstract

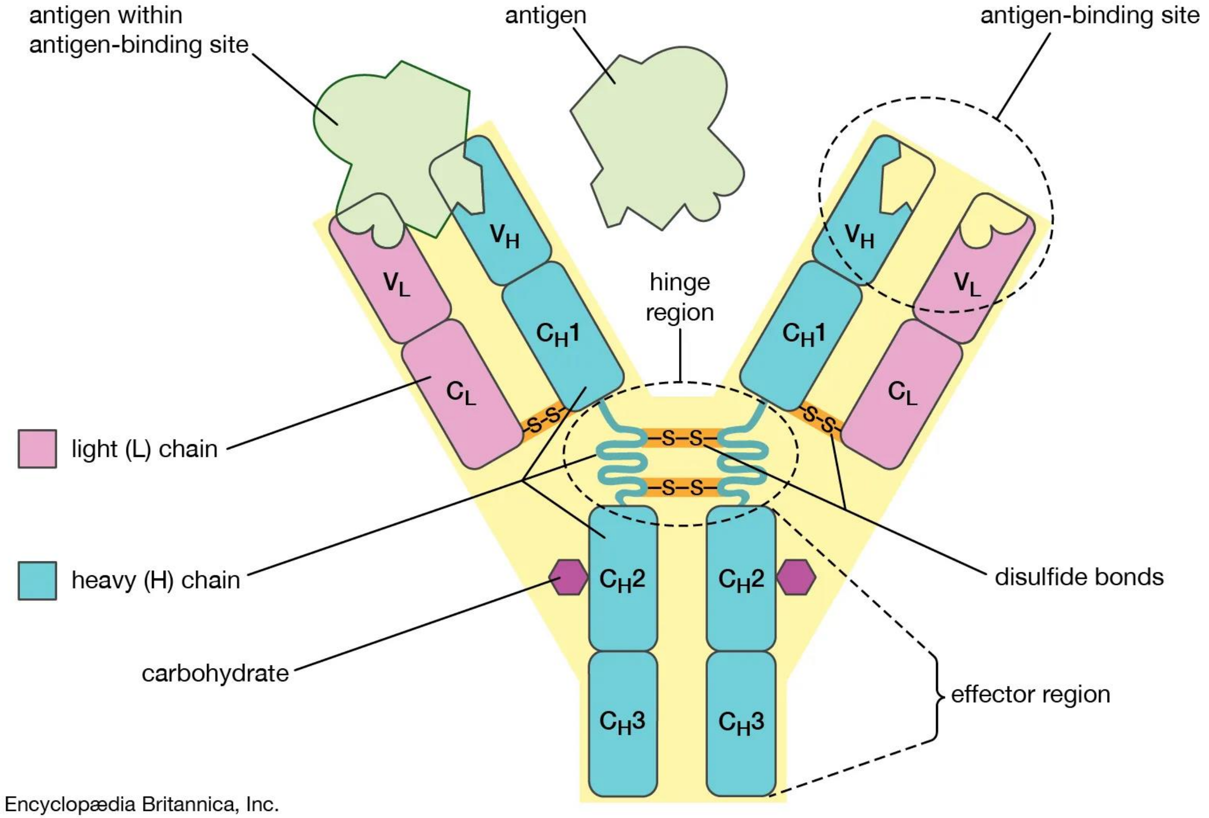
- The model unravels a sequence autoregressively while iteratively refining its predicted global structure (Structure 를 계속해서 수정해 나간다는 뜻)
- The inferred structure in turn guides subsequent residue choices
- We model the conditional dependence between residues inside and outside of a CDR in a coarse-grained manner (뭉뚱그려서 한다는 뜻, Figure 2 참조)
- Our method achieves superior log-likelihood on the test set
- Outperforms previous baselines in designing antibodies capable of neutralizing the SARS-CoV-2 virus1

Introduction

- **Antibodies**
  - Monoclonal antibodies are increasingly adopted as therapeutics targeting a wide range of pathogens such as SARS-CoV-2
  - Binding specificity

    → Determined by their complementarity-determining regions (CDRs)
  - Main Goal

    → To automate the creation of CDR subsequences **with desired properties**

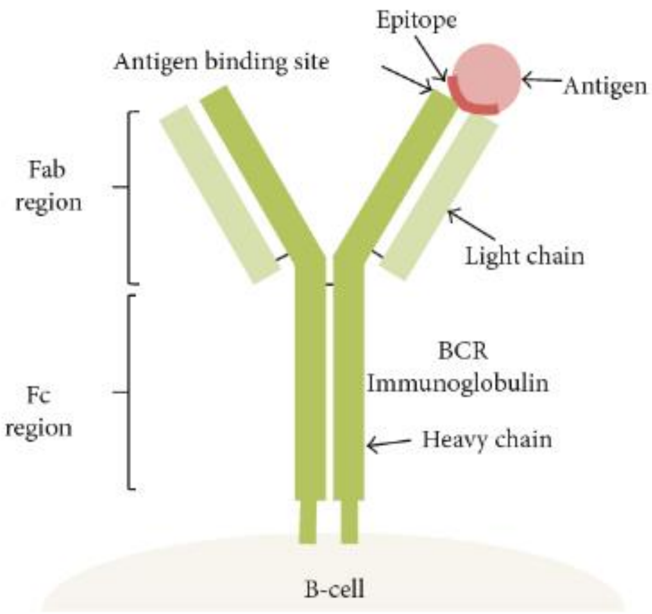


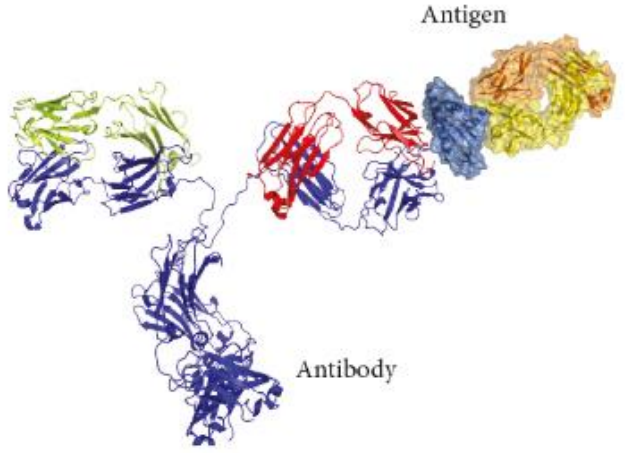
© Encyclopædia Britannica, Inc.

Introduction

◆ CDR (Antibody) Epitope (Antigen)

1. Complementarity-determining region
    - '항체' 에 존재하는 것
    - '항원' 과 상보적인 결합을 하는 부위
    - Hypervariable region (HV) 라고도 부름
    - 아미노산들의 서열 변화가 집중되어 있음
2. Epitope
    - '항원' 에 존재하는 것
    - CDR 혹은 HV와 결합하는 부위





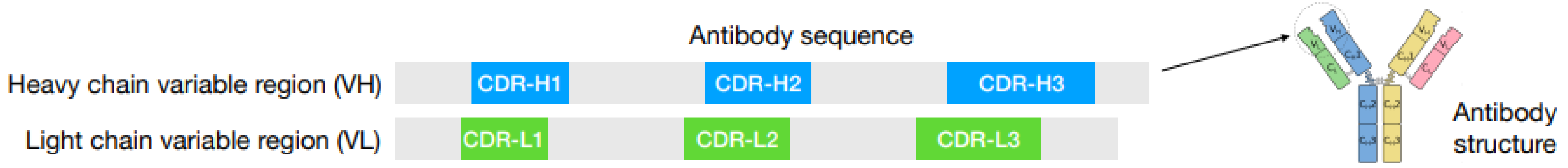
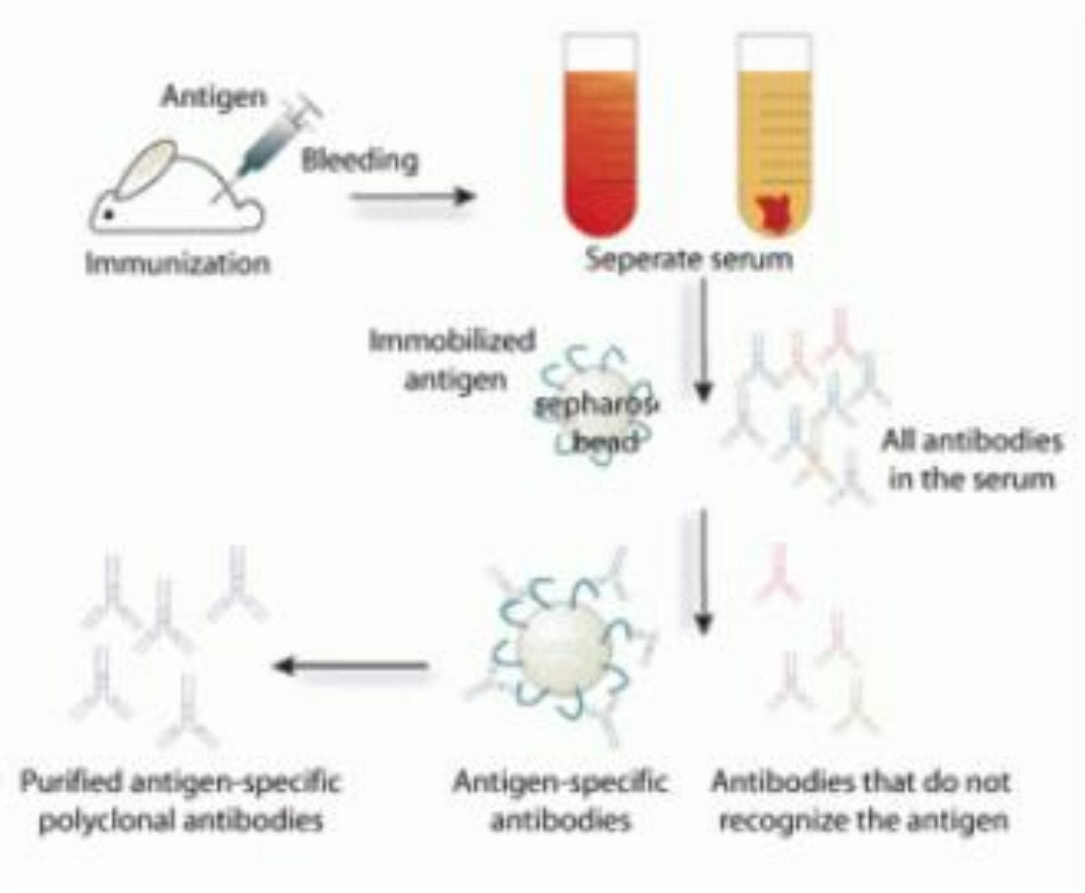


Figure 1: Schematic structure of an antibody (figure modified from Wikipedia).

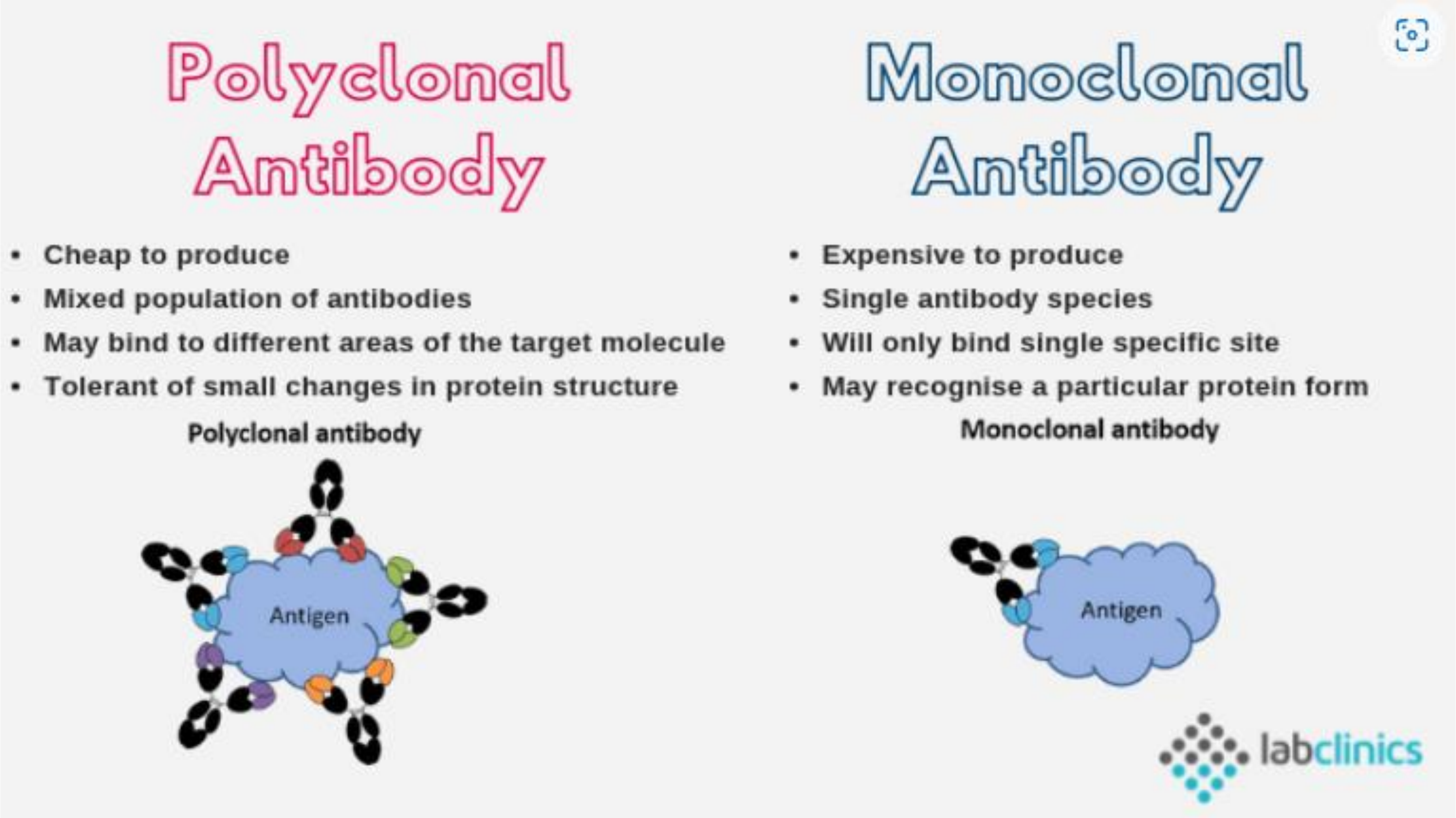
Introduction

- **Monoclonal Antibody**
  - Monoclonal antibodies (mAbs) are generated by identical B cells which are clones from a single parent cell
  - This means that the monoclonal antibodies have monovalent affinity and only recognize the same epitope of an antigen.


Antigen
Bleeding
Immunization
Seperate serum
Immobilized antigen
sepharose bead
All antibodies in the serum
Purified antigen-specific polyclonal antibodies
Antigen-specific antibodies
Antibodies that do not recognize the antigen

Introduction

- Monoclonal vs Polyclonal



Polyclonal Antibody

- Cheap to produce
- Mixed population of antibodies
- May bind to different areas of the target molecule
- Tolerant of small changes in protein structure

Monoclonal Antibody

- Expensive to produce
- Single antibody species
- Will only bind single specific site
- May recognise a particular protein form

labclinics

Introduction

◆ **Three Key modeling questions**

1. How to model the relation between a sequence and its underlying 3D structure
   - Structure 을 고려 안하면? → lead to sub-optimal performance (안 좋은 것)
   - Predefined 된 3D structure 를 고려하면 되지 않나? → 이상적으로 알려진 priori 는 지극히 적다
2. How to model the conditional distribution of CDRs given the remainder of a sequence (context)
   - 여기서 context 란 우리가 고려하는 CDR 의 sequence 를 말하는 듯
   - 이 sequence 를 이용해 CDR 의 조건부 분포를 만드는 것이 중요하다는 의미
   - Attention-based 방법은 sequence 단에서만 conditional dependence 를 고려하지만 Context 와 CDR region 의 구조적인 관계는 generation 에 있어서 매우 중요
3. Model's ability to optimize for various properties
   - 전통적인 물리적 방법은 binding E 를 최소화 시키는 것에만 집중했으나
   - 저자는 binding E 보다 더 중요한, 또 다른 objective 에 집중했음

Introduction

- **In this paper,**
  - Represent a sequence-structure pair as a graph
  - Formulate the co-design task as a graph generation problem

  - CDR 과 그에 해당하는 context 사이의 조건부 의존성을 Sequence 와 Structure level 둘 다 에서 고려하여 모델링 함
    - Sequence level – Residue amino acid
    - 3D structure – Pairwise residue distance

Introduction

◆ **Solve,**

- Antibody graph generation poses unique challenges because **the global structure is expected to change when new nodes are inserted**
- Previous autoregressive cannot modify a generated structure because they are trained under teacher forcing

  → 이미 이 다음 structure 가 주어진 상태에서 그 structure 에 맞춰 훈련을 하는 방식이기 때문에 robust 한 model 이 될 수가 없음

  → Cascade of errors 를 가져오는 문제가 생김

Proposed Model

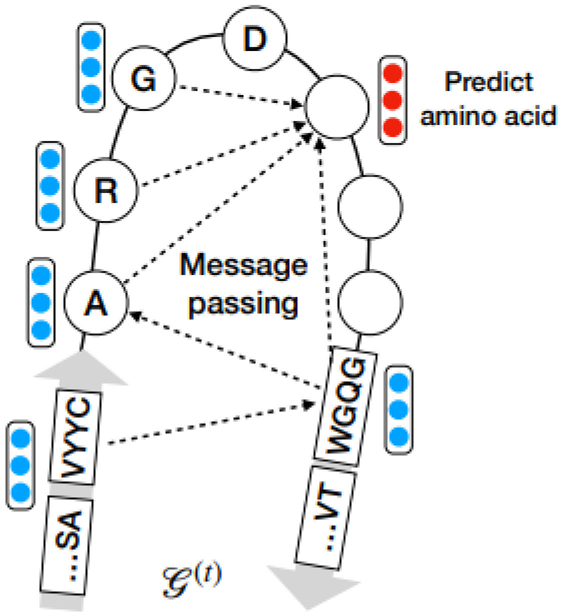
◆ **Solve,**

- To address these problems, we propose a novel architecture

  → **Interleaves the generation of amino acid nodes with the prediction of 3D structures**

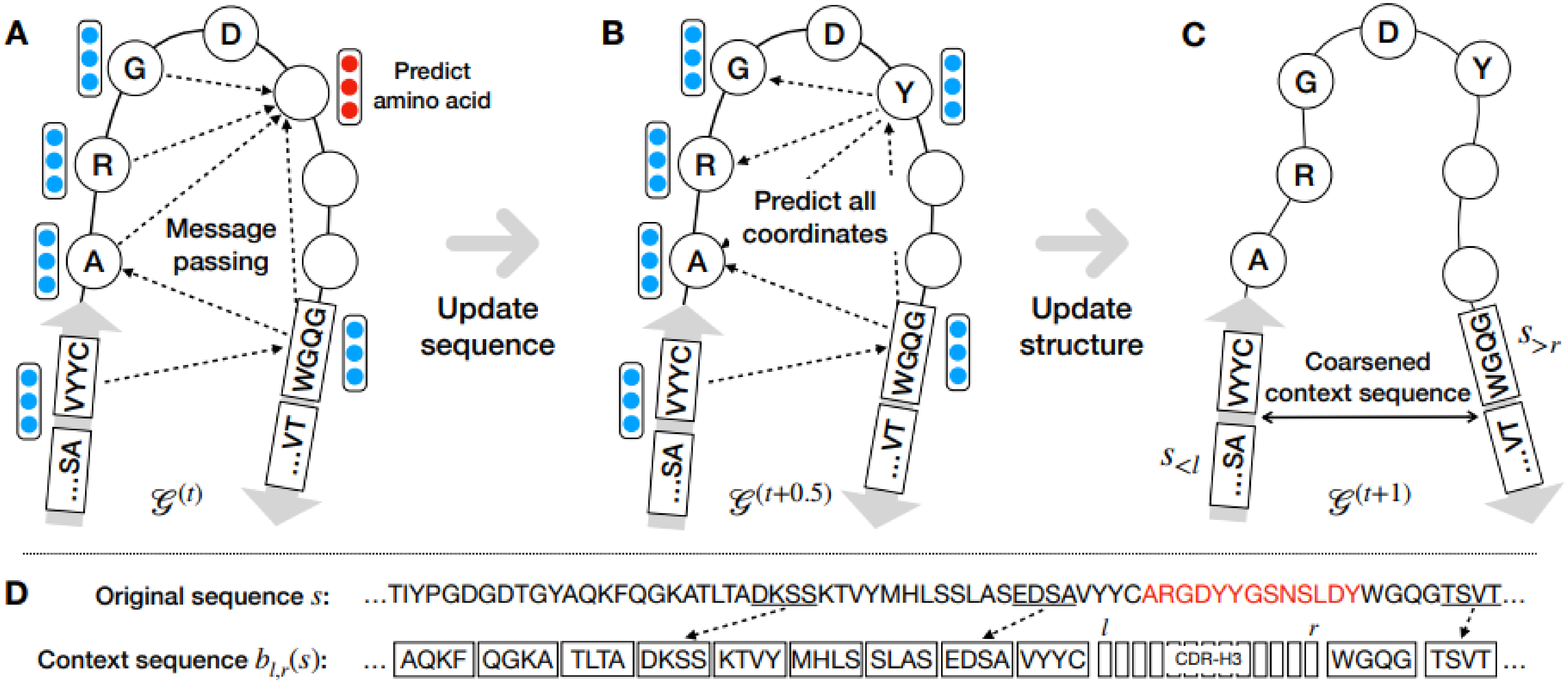
  → Structure generation is based on an **iterative refinement of a global graph** rather than a sequential expansion of a partial graph with teacher forcing (단순히 sequence 를 늘려가는 방식이 아닌, global graph 자체를 계속해서 다듬는 과정을 말함)

- Since the context sequence is long, we further introduce a coarsened graph representation by grouping **nodes into blocks**



Antibody Sequence & Structure Co-design

- **Overview,**



Related Work

◆ **Current Methods,**

- Computational antibody design roughly fall into two categories
  - Based on energy function optimization
  - Based on generative models

1. Based on energy function optimization
   - Use Monte Carlo simulation to iteratively modify a sequence & its structure
     → Until reaching a local energy minimum
   - Similar approaches are used in protein design
   - Weak point
     → computationally expensive (Ingraham et al., 2019) and
     → Our desired objective can be much more complicated than low binding energy
     (웬공진씨가 만든 모델은 이것 보다 더 복잡하다는 뜻 → Resource 의 한계가 뚜렷하다)

Related Work

◆ **Current Methods,**

- Computational antibody design roughly fall into two categories
  - Based on energy function optimization
  - Based on generative models

2. Based on generative models
   - Mostly sequence-based
   - Developed models conditioned on a backbone structure or protein fold
   - Weak point

     → Not consider both structure and sequence

     → Our model also seeks to incorporate 3D structure information for antibody generation

     → Since the best CDR structures are often unknown for new pathogens, we **co-design sequences and structures for specific properties**

Related Work

◆ **Generative models for graphs**

- Very related to autoregressive models for graph generation
  - Weak point

    → Generate edges sequentially and cannot modify a previously generated subgraph when new nodes arrive

    → (Even iterative model) Assumes all the node labels are given and **predicts edges only**

  - Our work combines **autoregressive models** with iterative refinement to generate a full graph with **node and edge labels**, including node labels and coordinates

Related Work

◆ **3D structure prediction**

- Closely related to protein folding
  - Weak point

    → AlphaFold : Require a complete protein sequence, its multi-sequence alignment (MSA), and its template features

    → (Even iterative model) Assumes all the node labels are given and **predicts edges only**

  - Our work : Models are **not directly applicable** because **we need to predict the structure of an incomplete sequence** and the MSA is not specified in advance

Related Work

◆ **3D structure prediction**

- Our iterative refinement model is also related to score matching methods for molecular conformation prediction and diffusion-based methods for point clouds
  - Iteratively refine a predicted 3D structure
  - Weak point
    → Only for a complete molecule or point cloud (완전한 분자에만 적용)
  - Our work : Learns to predict the 3D structure for incomplete graphs and interleaves 3D structure refinement with graph generation (Incomplete graph 에도 적용이 가능하다)

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ 3D structure prediction

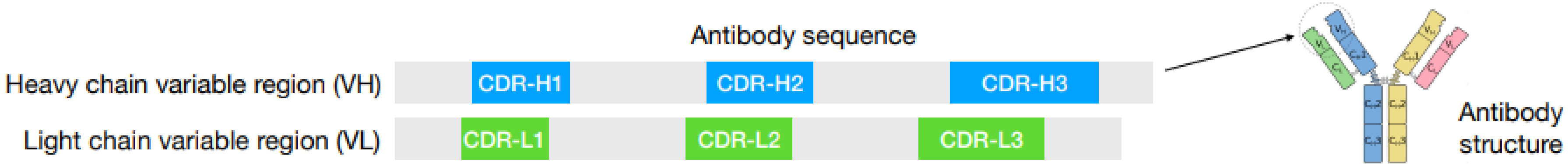


Figure 1: Schematic structure of an antibody (figure modified from Wikipedia).

- Antibody 의 구조 → Variable region 이 있는데 VR 은 2가지 부분으로 나뉨
  - Framework region
  - Three complementarity determining regions (CDRs)
- This work

  → formulate antibody design as a CDR generation task, **conditioned on** the framework region

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **3D structure prediction**

1. Represent an antibody as a graph, which encodes both its sequence and 3D structure
2. Propose a new graph generation approach called RefineGNN and extend it to handle **conditional generation given a fixed framework region**
3. Describe how to apply RefineGNN to property-guided optimization to design new antibodies with better neutralization properties

   → For simplicity, we **focus on the generation of heavy chain CDRs**, though our method can be easily extended to model light chains CDRs

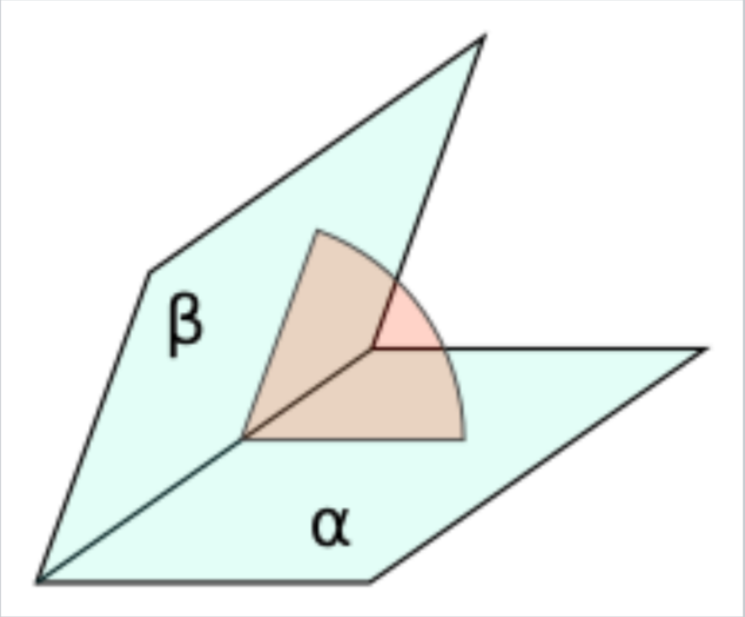
ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ GRAPH REPRESENTATION

- Each node (Residue) → Has three dihedral angle (이면각? 이거 나만 몰랐나?)

$$(\phi_i, \psi_i, \omega_i)$$

- Related to three backbone coordinates of residue i



Angle between two half-planes (α, β, pale blue) in a third plane (red) which cuts the line of intersection at right angles

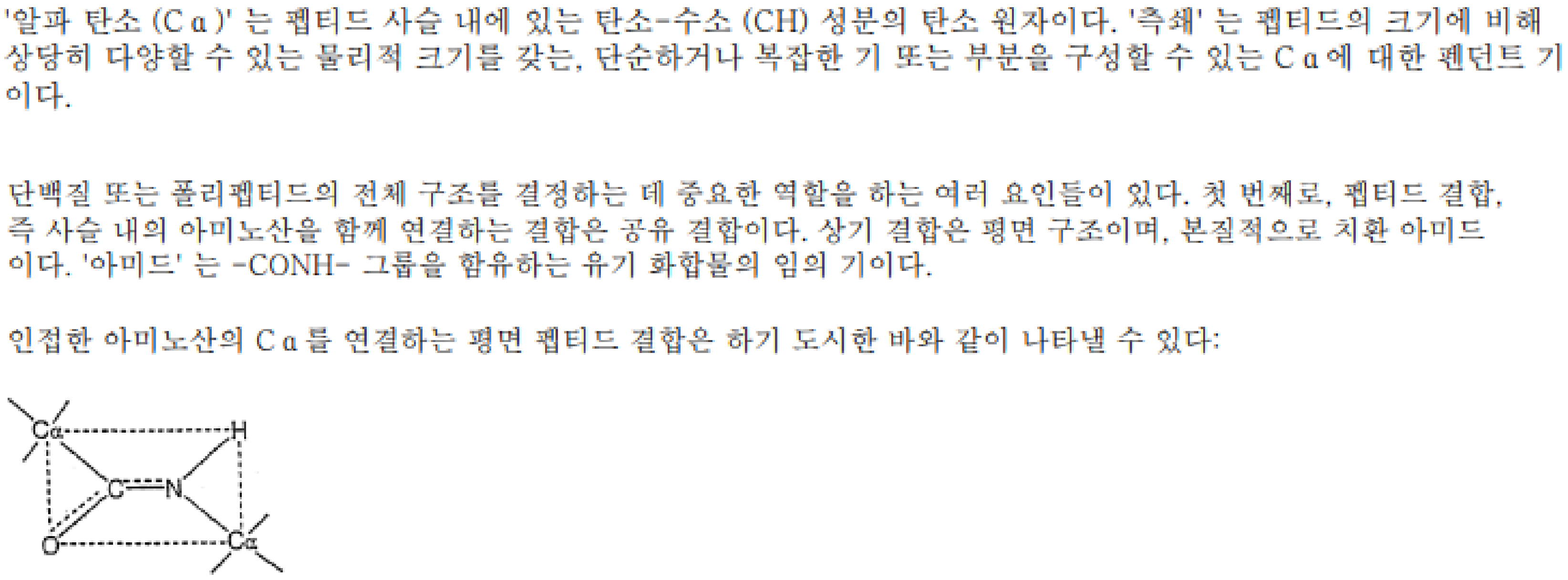
ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ GRAPH REPRESENTATION

- For each residue, → The coordinate of alpha-carbon

  → The other side chain atoms → Nitrogen, Carbon

- (참고 - 감소된 면역원성을 갖는 개질 에리트로포이에틴)



'알파 탄소 (C α)' 는 펩티드 사슬 내에 있는 탄소-수소 (CH) 성분의 탄소 원자이다. '측쇄' 는 펩티드의 크기에 비해 상당히 다양할 수 있는 물리적 크기를 갖는, 단순하거나 복잡한 기 또는 부분을 구성할 수 있는 C α에 대한 펜던트 기이다.

단백질 또는 폴리펩티드의 전체 구조를 결정하는 데 중요한 역할을 하는 여러 요인들이 있다. 첫 번째로, 펩티드 결합, 즉 사슬 내의 아미노산을 함께 연결하는 결합은 공유 결합이다. 상기 결합은 평면 구조이며, 본질적으로 치환 아미드이다. '아미드' 는 -CONH- 그룹을 함유하는 유기 화합물의 임의 기이다.

인접한 아미노산의 C α 를 연결하는 평면 펩티드 결합은 하기 도시한 바와 같이 나타낼 수 있다:

→ Compute orientation matrix O_i (Local coordinate frame 을 나타냄)

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ GRAPH REPRESENTATION

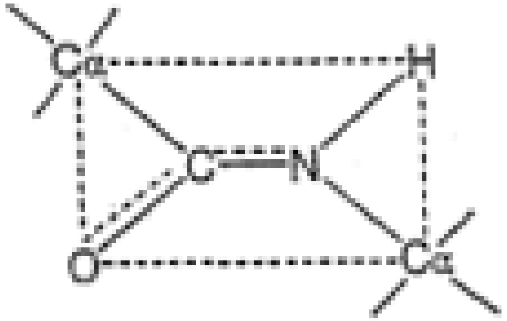
- Edge Feature : Contains four parts

$$\boldsymbol{e}_{ij} = \Big(E_{\text{pos}}(i-j),\quad \text{RBF}(\|\boldsymbol{x}_{i,\alpha}-\boldsymbol{x}_{j,\alpha}\|),\quad \boldsymbol{O}_i^\top \frac{\boldsymbol{x}_{j,\alpha}-\boldsymbol{x}_{i,\alpha}}{\|\boldsymbol{x}_{i,,\alpha}-\boldsymbol{x}_{j,,\alpha}\|},\quad \boldsymbol{q}(\boldsymbol{O}_i^\top \boldsymbol{O}_j)\Big)$$

- e_{ij} = Distance of sequence level
- RBF = Distance encoding lifted into radial basis (방사형, 아래 그림)

  = Distance between the alpha carbon of two residues

  = 1. Take the distance, 2. Lift it to the radial basis form (RBF kernel 이라 생각)



ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ GRAPH REPRESENTATION

- Edge Feature : Contains four parts

$$\boldsymbol{e}_{ij} = \Big(E_{\text{pos}}(i-j), \quad \text{RBF}(\|\boldsymbol{x}_{i,\alpha} - \boldsymbol{x}_{j,\alpha}\|), \quad \boldsymbol{O}_i^\top \frac{\boldsymbol{x}_{j,\alpha} - \boldsymbol{x}_{i,\alpha}}{\|\boldsymbol{x}_{i,,\alpha} - \boldsymbol{x}_{j,,\alpha}\|}, \quad \boldsymbol{q}(\boldsymbol{O}_i^\top \boldsymbol{O}_j)\Big)$$

- $\boldsymbol{O}_i^\top \frac{\boldsymbol{x}_{j,\alpha} - \boldsymbol{x}_{i,\alpha}}{\|\boldsymbol{x}_{i,,\alpha} - \boldsymbol{x}_{j,,\alpha}\|}$ = Orientation matrix of between residue i and j

  = Local coordinate frame

- $\boldsymbol{q}(\boldsymbol{O}_i^\top \boldsymbol{O}_j)$ = Orientation encoding of the quaternion representation of the spatial rotation matrix $(\boldsymbol{O}_i^\top \boldsymbol{O}_j)$

- These four parts → **Input** to the graph neural network

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)

- $\mathcal{G}^{(0)}$: initial guess of the true antibody graph
- Each residue → initialized as a special token <MASK>
- Each edge (i, j) → initialized to be of distance $3|i - j|$ (Consecutive residues → three?)
- Direction and orientation features are set to 0

- Each generation step
  - Model learns to revise a current antibody graph (그래프를 수정하는 법을 배움)
  - Predict the label of the next residue $t + 1$ **(노드 하나를 추가 하는게 하나의 step 임!!! 중요)**

$$\{\boldsymbol{h}_1^{(t)}, \cdots, \boldsymbol{h}_n^{(t)}\} = \mathrm{MPN}_\theta(\mathcal{G}^{(t)})$$

  - Given the current graph structure → Use message passing, encode to get the hidden state of each residue

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)

$$\{\boldsymbol{h}_1^{(t)}, \cdots, \boldsymbol{h}_n^{(t)}\} = \mathrm{MPN}_\theta(\mathcal{G}^{(t)})$$

- Any message passing network is adopted

$$\boldsymbol{h}_i^{(t,l+1)} = \mathrm{LayerNorm}\Big(\sum_j \mathrm{FFN}\big(\boldsymbol{h}_i^{(t,l)}, \boldsymbol{h}_j^{(t,l)}, E(\boldsymbol{s}_j), \boldsymbol{e}_{i,j}\big)\Big), \quad 0 \le l \le L-1$$

- 여기서는 단순하게 Feedforward network 를 사용
  - Residue i 의 learned representation
  - Residue j 의 learned representation
  - Learned embedding of amino acid type S_j
  - Residue i 와 Residue j 의 distance Residue i

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)

$$\{\boldsymbol{h}_1^{(t)}, \cdots, \boldsymbol{h}_n^{(t)}\} = \mathrm{MPN}_\theta(\mathcal{G}^{(t)})$$

- Any message passing network is adopted

$$\boldsymbol{h}_i^{(t,l+1)} = \mathrm{LayerNorm}\left(\sum_j \mathrm{FFN}\left(\boldsymbol{h}_i^{(t,l)}, \boldsymbol{h}_j^{(t,l)}, E(\boldsymbol{s}_j), \boldsymbol{e}_{i,j}\right)\right), \quad 0 \le l \le L-1$$

- Based on the learned residue representations, we predict the **amino acid type** of the next residue $t+1$ = $h_i^{(t,l+1)}$ 을 이용해서 그 다음 residue 를 예측

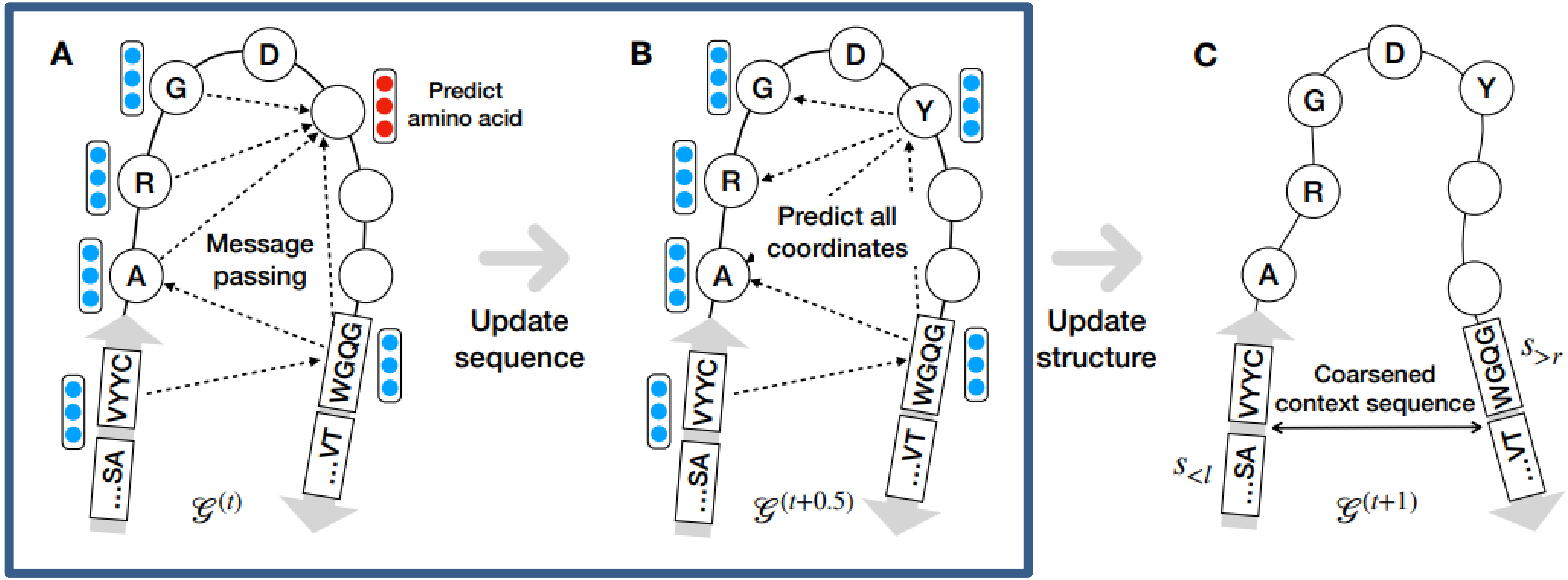
$$\boldsymbol{p}_{t+1} = \mathrm{softmax}(\boldsymbol{W}_a \boldsymbol{h}_{t+1}^{(t)})$$

→ Classification task

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)

- Based on the learned residue representations, we predict the amino acid type of the next residue $t+1$ = $h_i^{(t,l+1)}$ 을 이용해서 그 다음 residue 를 예측
- 아래 그림을 보면 그 다음 residue 인 Y 를 예측함



ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)

- Next, we need to update the structure to accommodate the new residue t+ 1
- Encode graph $\mathcal{G}^{(t+0.5)}$ by another MPN with a different parameter $\tilde{\theta}$, predict the coordinate of all residues

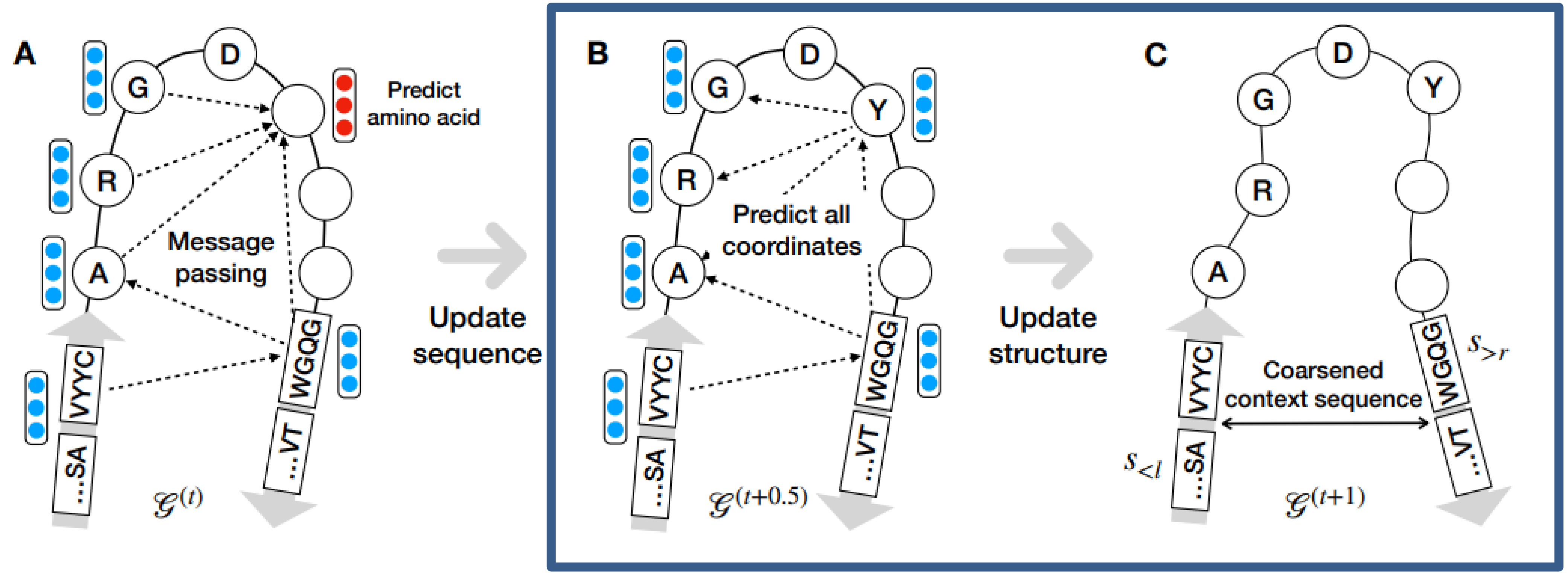
$$\begin{aligned}\{\boldsymbol{h}_1^{(t+0.5)}, \cdots, \boldsymbol{h}_n^{(t+0.5)}\} &= \mathrm{MPN}_{\tilde{\theta}}(\mathcal{G}^{(t+0.5)}) \\ \boldsymbol{x}_{i,e}^{(t+1)} &= \boldsymbol{W}_x^e \boldsymbol{h}_i^{(t+0.5)}, \qquad 1 \leq i \leq n, e \in \{\alpha, c, n\}\end{aligned}$$

- $x_{i,e}^{(t+1)}$ = new coordinate of each residue

→ $\mathcal{G}^{(t+1)}$ 을 예측

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

◆ ITERATIVE REFINEMENT GRAPH NEURAL NETWORK (REFINEGNN)



- The structure prediction (coordinates x_i) and sequence prediction (amino acid types p_{t+1}) are carried out by two different MPNs, namely the structure network $\tilde{\theta}$ and sequence network θ
- This disentanglement allows the two networks to focus on two distinct tasks

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **Training**

1. Apply teacher forcing to the discrete amino acid type prediction
2. In each generation step t, residues 1 to t are set to their ground truth amino acid types $s_1, \dots, s_t$ while all future residues $t+1, \dots, n$ are set to a padding token
3. In contrast, the continuous structure prediction is carried out without teacher forcing
4. In each iteration, the model refines the entire structure predicted in the previous step and constructs a new K-nearest neighbors graph $\mathcal{G}^{(t+1)}$ of all residues based on the predicted coordinates $\left\{x_{i,e}^{(t+1)} \mid 1 \leq i \leq n, e \in \{\alpha, c, n\}\right\}$

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **Loss function**
  - **Loss function = Structure prediction loss + Sequence prediction loss**
  - The loss function for antibody structure prediction consists of three parts
    1. Distance Loss
    2. Dihedral Loss
    3. C_α angle loss (Backbone angle)
  - The loss function for sequence prediction is Cross-entropy loss
    1. Between predicted and true residue

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **Loss function**
  - **Loss function = Structure prediction loss + Sequence prediction loss**
  - The loss function for antibody structure prediction consists of three parts

    1. Distance Loss (Pairwise distances)

$$\mathcal{L}_d^{(t)} = \sum\nolimits_{i,j} \ell_{\text{huber}}(\|\boldsymbol{x}_{i,\alpha}^{(t)} - \boldsymbol{x}_{j,\alpha}^{(t)}\|^2, \|\boldsymbol{x}_{i,\alpha} - \boldsymbol{x}_{j,\alpha}\|^2)$$

- $\|\boldsymbol{x}_{i,\alpha}^{(t)} - \boldsymbol{x}_{j,\alpha}^{(t)}\|^2$ = Predicted
- $\|\boldsymbol{x}_{i,\alpha} - \boldsymbol{x}_{j,\alpha}\|^2$ = Ground truth

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **Loss function**
  - **Loss function = Structure prediction loss + Sequence prediction loss**
  - The loss function for antibody structure prediction consists of three parts

    2. Dihedral Loss (Angle)

$$\mathcal{L}_a^{(t)} = \sum_i \sum_{\boxed{a\in\{\phi,\psi,\omega\}}} (\cos a_i^{(t)} - \cos a_i)^2 + (\sin a_i^{(t)} - \sin a_i)^2$$

- $\boxed{\left(\phi_i^{(t)}, \psi_i^{(t)}, \omega_i^{(t)}\right)}$ = Dihedral angle based on predicted atom coordinates

$\boldsymbol{x}_{i,\alpha}^{(t)}, \boldsymbol{x}_{i,c}^{(t)}, \boldsymbol{x}_{i,n}^{(t)}$ and $\boldsymbol{x}_{i+1,\alpha}^{(t)}, \boldsymbol{x}_{i+1,c}^{(t)}, \boldsymbol{x}_{i+1,n}^{(t)}$

ANTIBODY SEQUENCE AND STRUCTURE CO-DESIGN

- **Loss function**
  - **Loss function = Structure prediction loss + Sequence prediction loss**
  - The loss function for antibody structure prediction consists of three parts

    3. C_α angle loss : Calculate angle

$$\mathcal{L}_c^{(t)} = \sum\nolimits_i (\cos\gamma_i^{(t)} - \cos\gamma_i)^2 + (\cos\beta_i^{(t)} - \cos\beta_i)^2$$

- Calculate angles $\gamma_i^{(t)}$ between two vectors $\boldsymbol{x}_{i-1,\alpha}^{(t)} - \boldsymbol{x}_{i,\alpha}^{(t)}$ and $\boldsymbol{x}_{i,\alpha}^{(t)} - \boldsymbol{x}_{i+1,\alpha}^{(t)}$
- Calculate angles $\beta_i^{(t)}$ between two planes $\boldsymbol{x}_{i-2,\alpha}^{(t)}, \boldsymbol{x}_{i-1,\alpha}^{(t)}, \boldsymbol{x}_{i,\alpha}^{(t)}, \boldsymbol{x}_{i+1,\alpha}^{(t)}$

- Structure Loss = $\mathcal{L}_{\text{struct}} = \sum\nolimits_t \mathcal{L}_d^{(t)} + \mathcal{L}_a^{(t)} + \mathcal{L}_c^{(t)}$
- Sequence Loss = $\mathcal{L}_{\text{seq}} = \sum\nolimits_t \mathcal{L}_{ce}(\boldsymbol{p}_t, \boldsymbol{s}_t)$
- **Total Loss** = $\mathcal{L} = \mathcal{L}_{\text{seq}} + \mathcal{L}_{\text{struct}}$

Conditional Generation Given the Framework Region

◆ Given framework condition

- The model architecture described so far is designed for unconditional generation
- It generates an entire antibody graph without any constraints

$$\boldsymbol{s}_{<l} = \boldsymbol{s}_1 \cdots \boldsymbol{s}_{l-1}$$

- However, in practice,
  - Usually fix the framework region of an antibody and
  - Design the CDR sequence only
- Therefore,
  - We need to extend **the model architecture to learn the conditional distribution** $\boxed{P(\boldsymbol{s}'|\boldsymbol{s}_{<l}, \boldsymbol{s}_{>r})}$ where $\boldsymbol{s}_{<l} = \boldsymbol{s}_1 \cdots \boldsymbol{s}_{l-1}$ and $\boldsymbol{s}_{>r} = \boldsymbol{s}_{r+1} \cdots \boldsymbol{s}_n$ are residues outside of the CDR $s_l, \ldots, s_r$

Conditional Generation Given the Framework Region

- **Conditioning via attention**
  - A simple extension of RefineGNN is to encode the non-CDR sequence using a recurrent neural network and propagate information to the CDR through an attention layer.
  - Leverage the information from the framework residues
  - Apply attention over all the framework residues

$$\begin{aligned}
\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_n\} &= \boldsymbol{c}_{1:n} = \mathrm{GRU}(\tilde{\boldsymbol{s}}) \\
\boldsymbol{p}_{t+1} &= \mathrm{softmax}\big(\boldsymbol{W}_a \boldsymbol{h}_{t+1}^{(t)} + \boldsymbol{U}_a^\top \mathrm{attention}(\boldsymbol{c}_{1:n}, \boldsymbol{h}_{t+1}^{(t)})\big) \\
\boldsymbol{x}_{i,e}^{(t+1)} &= \boldsymbol{W}_x^e \boldsymbol{h}_i^{(t+0.5)} + \boldsymbol{U}_x^{e\top} \mathrm{attention}(\boldsymbol{c}_{1:n}, \boldsymbol{h}_i^{(t+0.5)})
\end{aligned}$$

  - 'Weak Point!'

    → Only modeling the structure of CDR (not entire antibody)

Conditional Generation Given the Framework Region

- **Multi-resolution modeling**
  - The attention-based approach alone is not sufficient
    - Because it **does not model the structure of the context sequence**, thus ignoring how its residues **structurally interact with the CDR's**
    - 앞서 말한 것과 같이 CDR 부분의 structure 만 신경씀, CDR 의 앞뒤로 있는 **framework 부분과의 interaction (structure) 는 신경쓰지 않음**

Conditional Generation Given the Framework Region

◆ **Multi-resolution modeling**

- So, during training step, we use the known structure to predict the interaction
- However,
  - Computationally expensive because we need to recompute the MPN (message passing network) encoding for all residues in each generation step
  - Cannot predict the context residue coordinates at the outset and fix them
    → Not adjusted accordingly when the coordinates of CDR residues are updated in each generation step

Conditional Generation Given the Framework Region

◆ Multi-resolution modeling

- Solution
  - Propose a coarse-grained model that reduces the context sequence length **by clustering it into residue blocks** (Coarse-grained → 거칠고 큼직큼직 하나는 뜻 → Context sequence 를 만들어 줌)
  - 각 residue 의 coordinate 의 mean 값을 residue block 의 coordinate 로 사용

$$E(\boldsymbol{b}_i) = \sum\nolimits_{\boldsymbol{s}_j \in \boldsymbol{b}_i} E(\boldsymbol{s}_j)/b, \qquad \boldsymbol{x}_{\boldsymbol{b}_i,e} = \sum\nolimits_{\boldsymbol{s}_j \in \boldsymbol{b}_i} \boldsymbol{x}_{j,e}/b, \qquad e \in \{\alpha, c, n\}$$

Property-guided sequence optimization

◆ Ultimate goal

- Generate new antibodies with desired properties such as neutralizing a particular virus

  → Cen be formulated as an optimization problem

  → Conditional generative model $P_{\Theta}(s'|b_{l,r}(s))$

  → Maximizes the probability of neutralization for a training set of antibodies $\mathcal{D}$

$$\sum_{\boldsymbol{s}\in\mathcal{D}} \log P(Y=1|\boldsymbol{b}_{l,r}(\boldsymbol{s})) = \sum_{\boldsymbol{s}\in\mathcal{D}} \log \sum_{\boldsymbol{s}'} \boxed{\boxed{f(\boldsymbol{s}')}\boxed{P_{\Theta}(\boldsymbol{s}'|\boldsymbol{b}_{l,r}(\boldsymbol{s}))}}$$

$f(s')$ →Predictor for $P_{\Theta}(s'|b_{l,r}(s))$

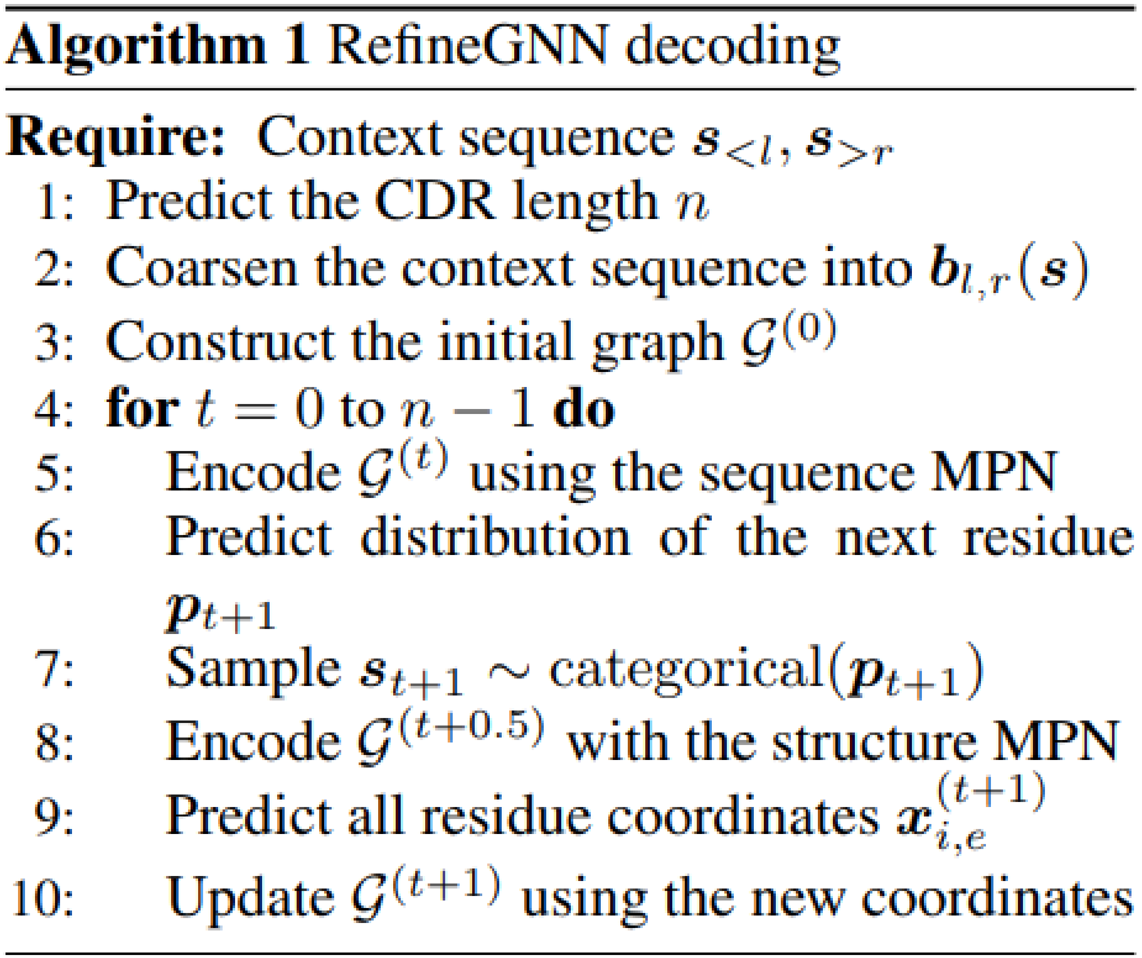
- Context sequence 가 주어졌을 때 desire property 를 가질 수 있도록 모든 amino acid (s') 에 대해 loglikelihood 값을 최대화 시켜 줌

  → f 가 주어졌을 때, 위의 식은 iterative target augmentation 으로 풀이 될 수 있다 (오 이거 내가 발표했던 논문임!!!)

  → 반복적으로 추가적인 target molecule 을 붙여나갈 수 있게 작업

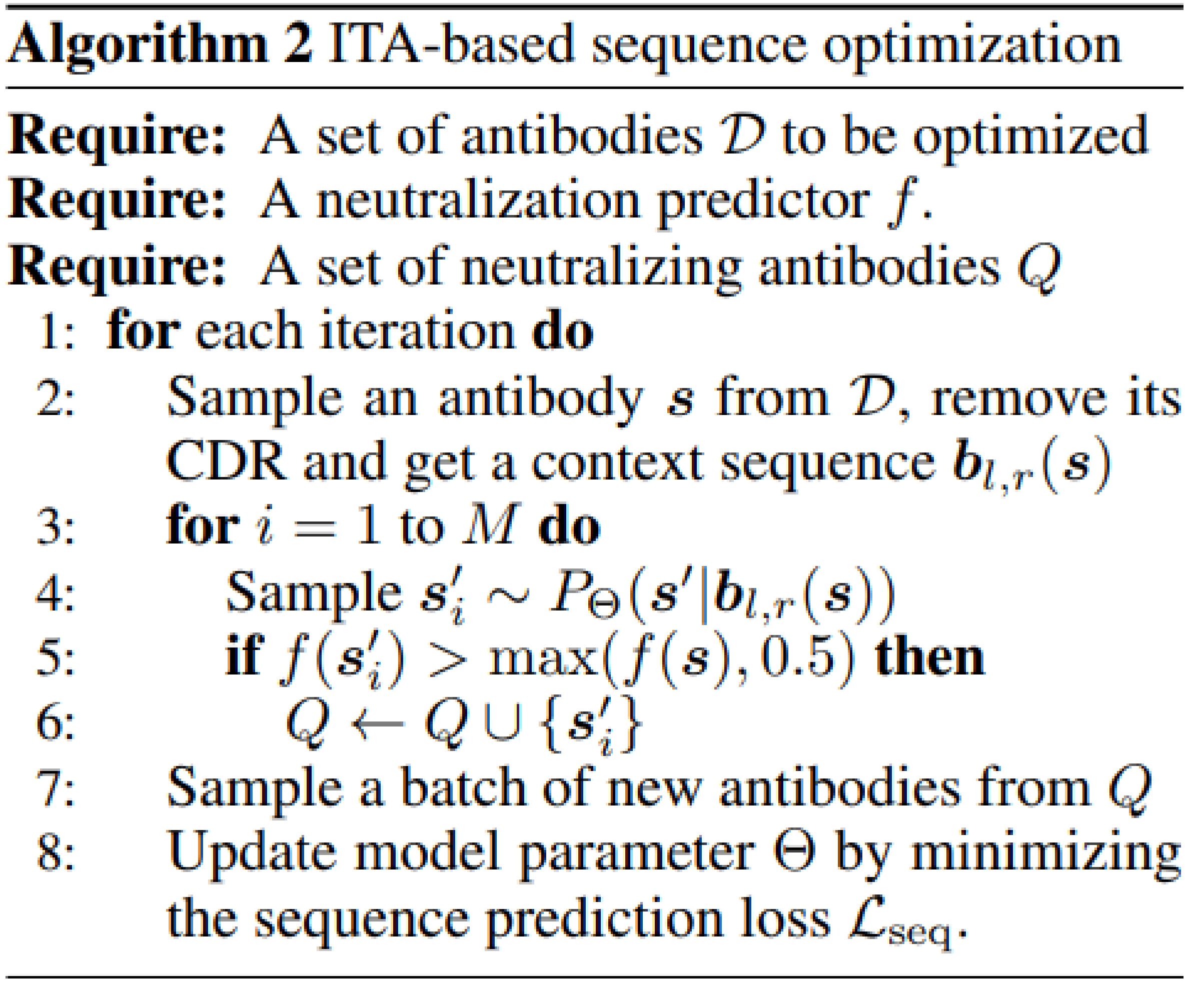
Property-guided sequence optimization

- **Ultimate goal**



Algorithm 1 RefineGNN decoding

```
Require: Context sequence $\boldsymbol{s}_{<l}, \boldsymbol{s}_{>r}$
Predict the CDR length $n$
Coarsen the context sequence into $\boldsymbol{b}_{l,r}(\boldsymbol{s})$
Construct the initial graph $\mathcal{G}^{(0)}$
for $t = 0$ to $n - 1$ do
    Encode $\mathcal{G}^{(t)}$ using the sequence MPN
    Predict distribution of the next residue $\boldsymbol{p}_{t+1}$
    Sample $\boldsymbol{s}_{t+1} \sim \text{categorical}(\boldsymbol{p}_{t+1})$
    Encode $\mathcal{G}^{(t+0.5)}$ with the structure MPN
    Predict all residue coordinates $\boldsymbol{x}_{i,e}^{(t+1)}$
    Update $\mathcal{G}^{(t+1)}$ using the new coordinates
```



Algorithm 2 ITA-based sequence optimization

```
Require: A set of antibodies $\mathcal{D}$ to be optimized
Require: A neutralization predictor $f$.
Require: A set of neutralizing antibodies $Q$
for each iteration do
    Sample an antibody $\boldsymbol{s}$ from $\mathcal{D}$, remove its CDR and get a context sequence $\boldsymbol{b}_{l,r}(\boldsymbol{s})$
    for $i = 1$ to $M$ do
        Sample $\boldsymbol{s}_i' \sim P_\Theta(\boldsymbol{s}'|\boldsymbol{b}_{l,r}(\boldsymbol{s}))$
        if $f(\boldsymbol{s}_i') > \max(f(\boldsymbol{s}), 0.5)$ then
            $Q \leftarrow Q \cup \{\boldsymbol{s}_i'\}$
    Sample a batch of new antibodies from $Q$
    Update model parameter $\Theta$ by minimizing the sequence prediction loss $\mathcal{L}_{\text{seq}}$.
```

Q & A

Thank You!