# Multimodal deep learning for protein engineering

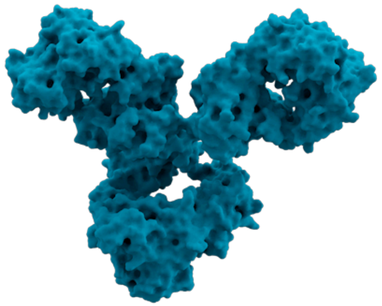
Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

# Proteins are biology’s actuators

- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling

# Proteins are biology’s actuators

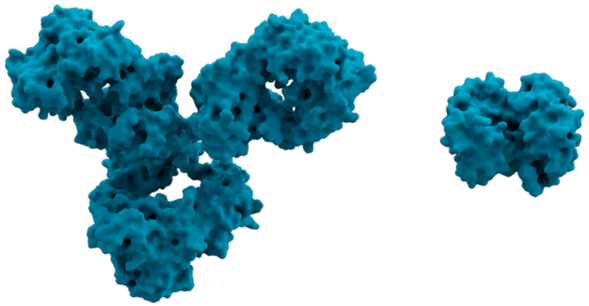
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



IgG

# Proteins are biology’s actuators

- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



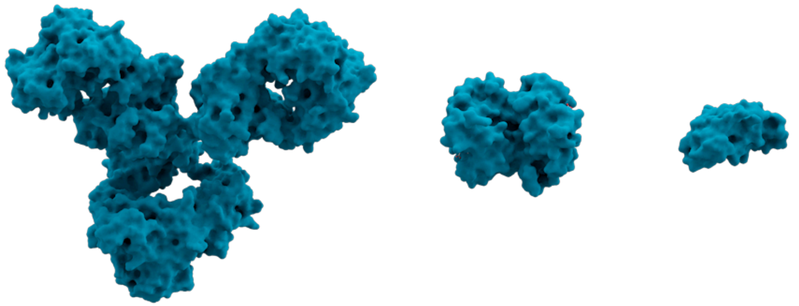
IgG



hemoglobin

# Proteins are biology’s actuators

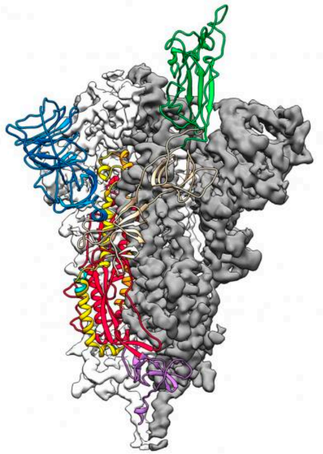
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



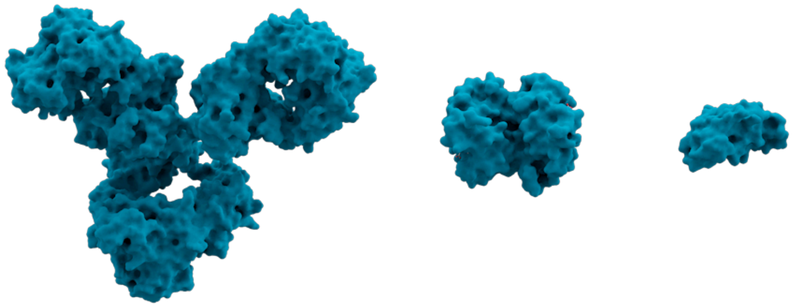
IgG hemoglobin insulin

# Proteins are biology’s actuators

- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



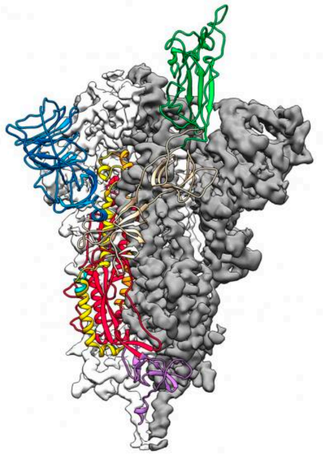
coronavirus spike protein



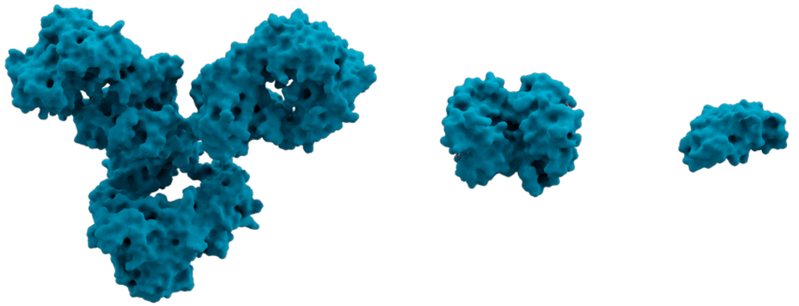
IgG hemoglobin insulin

# Proteins are biology's actuators

- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



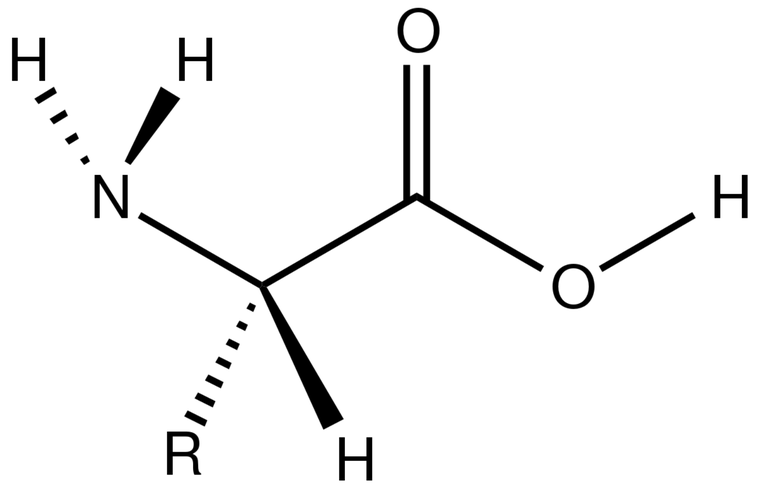
coronavirus spike protein



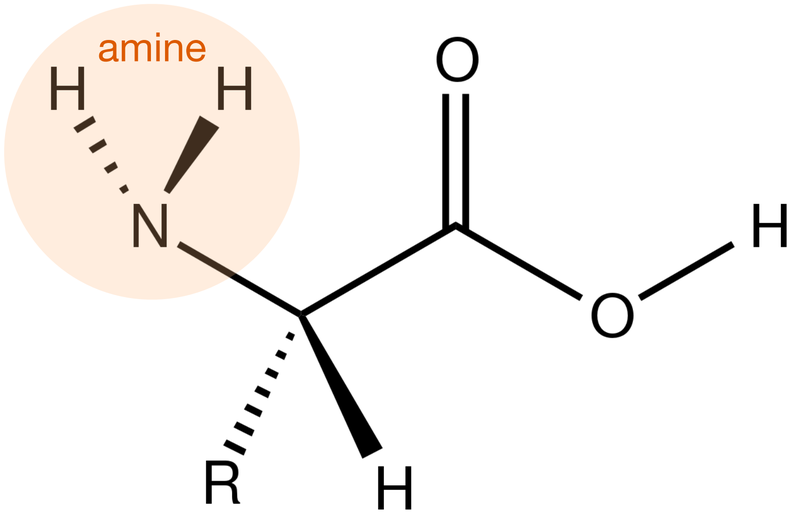
IgG hemoglobin insulin

luciferase

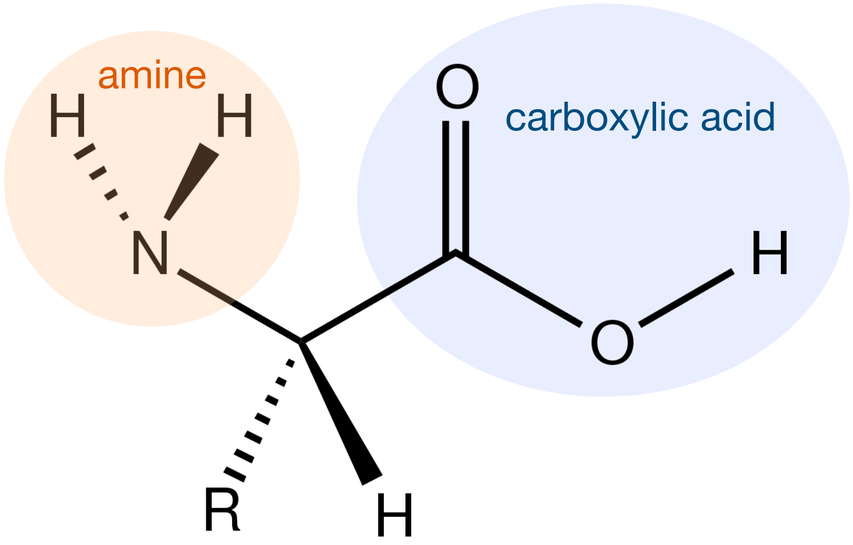
# Diversity arises from 20 building blocks



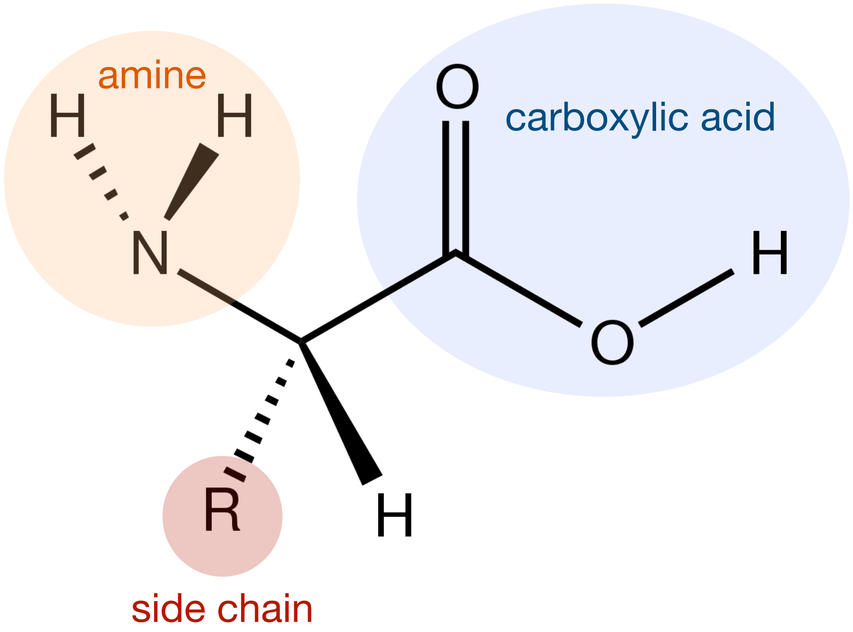
# Diversity arises from 20 building blocks



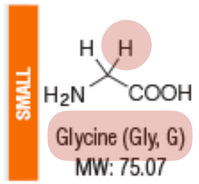
# Diversity arises from 20 building blocks

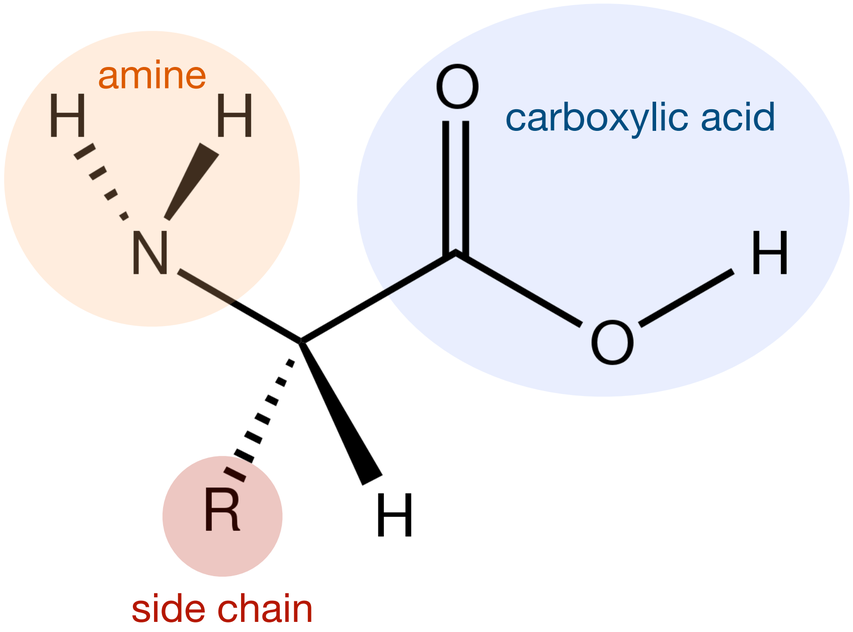


# Diversity arises from 20 building blocks

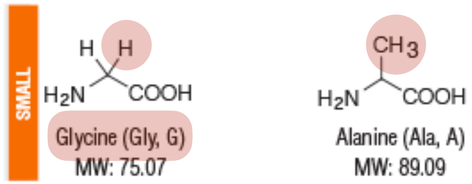


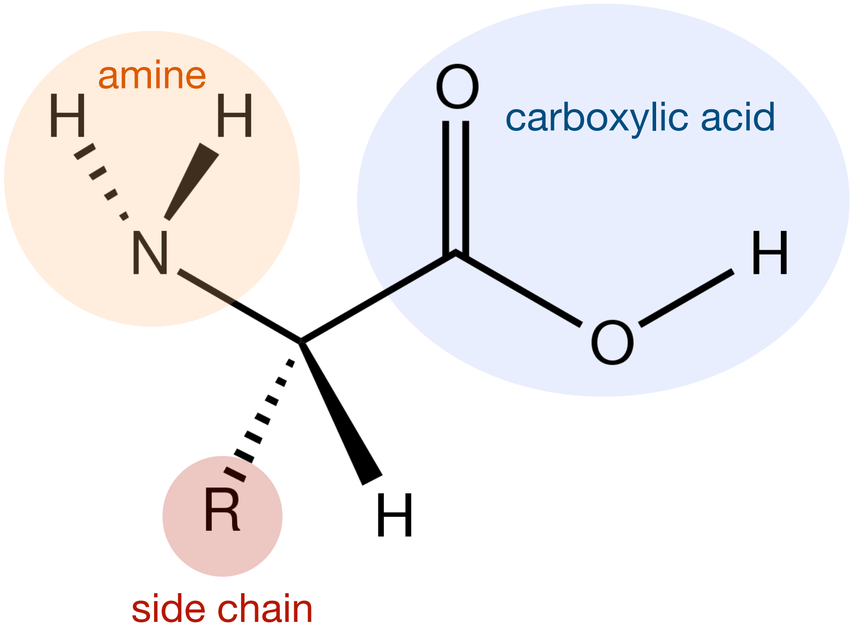
# Diversity arises from 20 building blocks



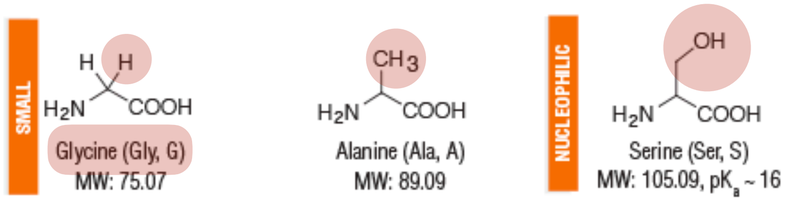



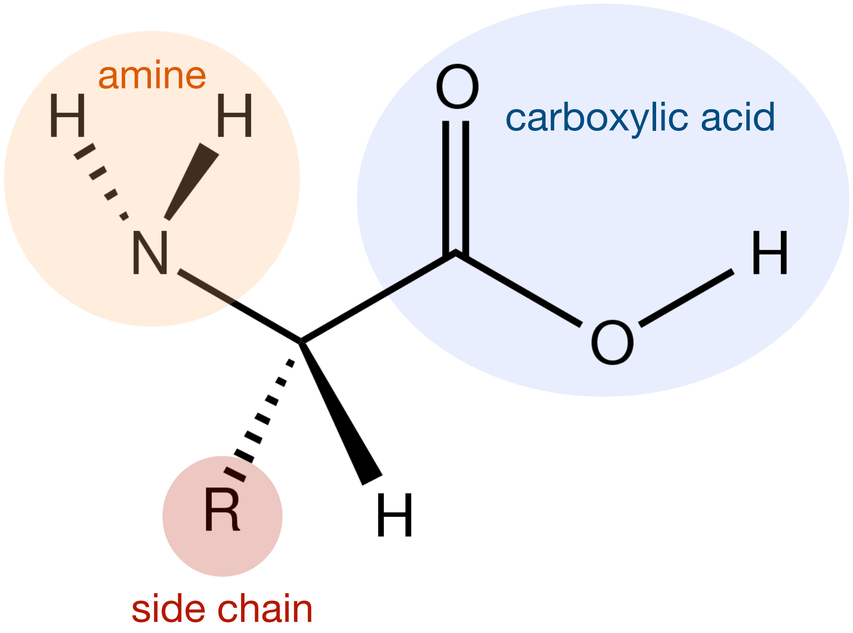
# Diversity arises from 20 building blocks



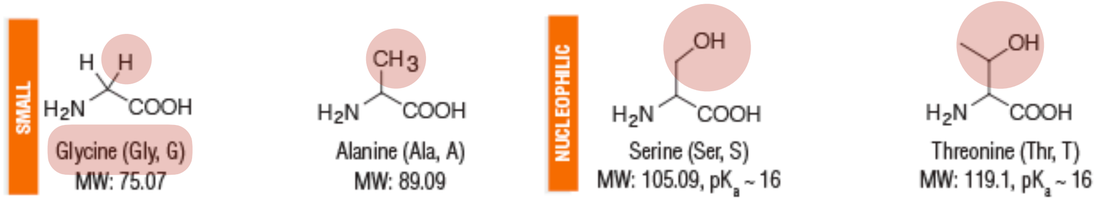



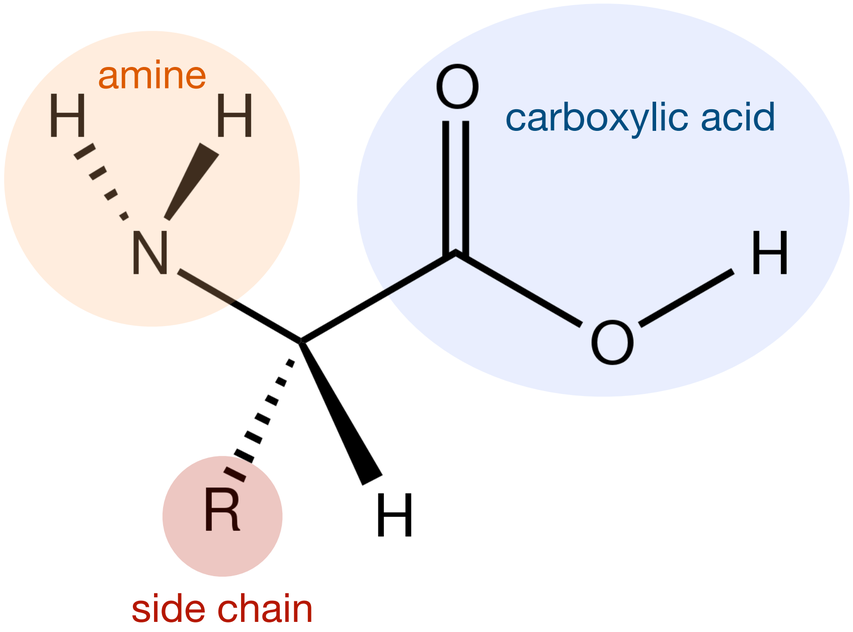
# Diversity arises from 20 building blocks



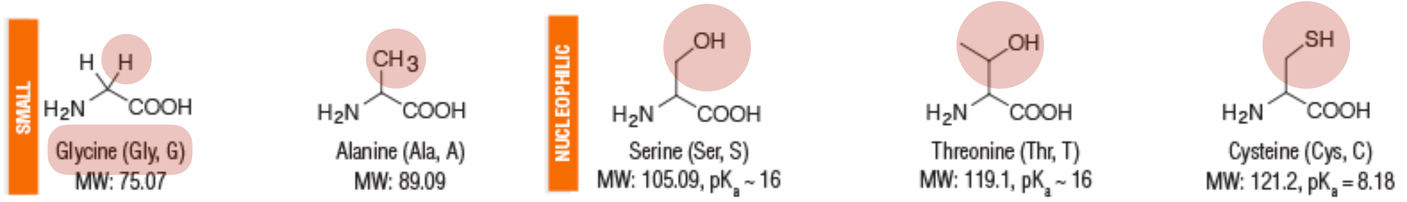



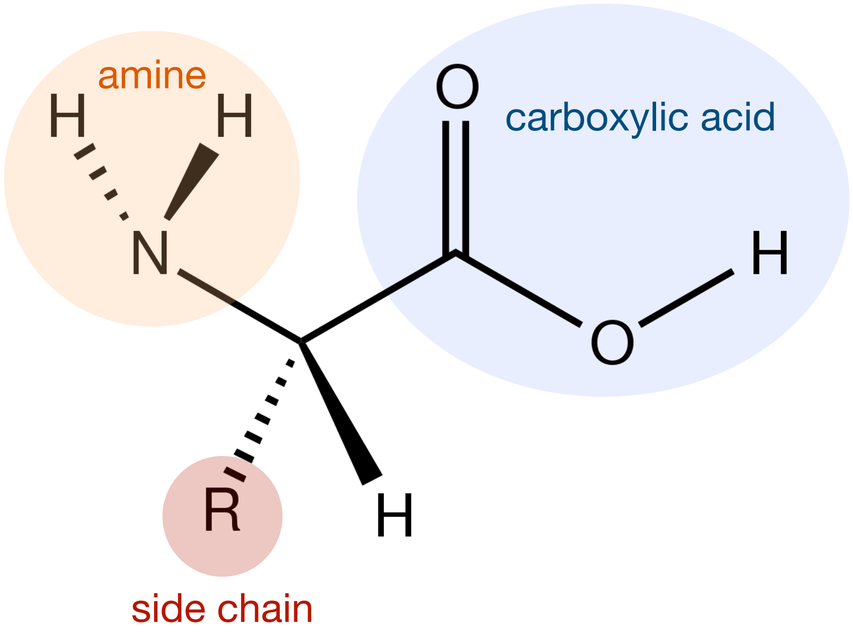
# Diversity arises from 20 building blocks



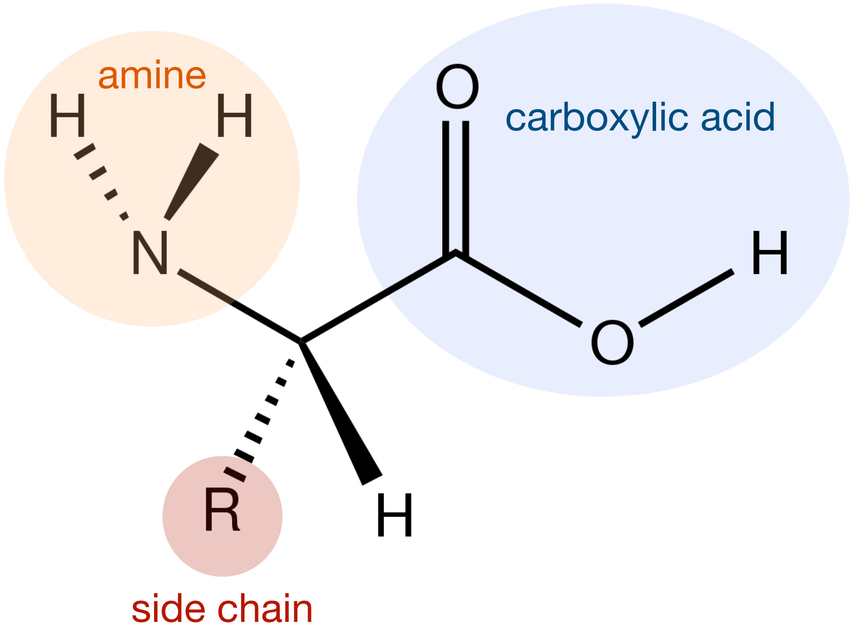



# Diversity arises from 20 building blocks



# Diversity arises from 20 building blocks




SMALL

Glycine (Gly, G)
MW: 75.07

Alanine (Ala, A)
MW: 89.09

NUCLEOPHILIC

Serine (Ser, S)
MW: 105.09, $pK_a \sim 16$

Threonine (Thr, T)
MW: 119.1, $pK_a \sim 16$

Cysteine (Cys, C)
MW: 121.2, $pK_a = 8.18$

HYDROPHOBIC

Valine (Val, V)
MW: 117.1

Leucine (Leu, L)
MW: 131.2

Isoleucine (Ile, I)
MW: 131.2

Methionine (Met, M)
MW: 149.2

Proline (Pro, P)
MW: 115.1

AROMATIC

Phenylalanine (Phe, F)
MW: 165.2

Tyrosine (Tyr, Y)
MW: 181.2, $pK_a = 10.46$

Tryptophan (Trp, W)
MW: 204.2

ACIDIC

Aspartic Acid (Asp, D)
MW: 133.1, $pK_a = 3.9$

Glutamic Acid (Glu, E)
MW: 147.1, $pK_a = 4.07$

AMIDE

Asparagine (Asn, N)
MW: 132.1

Glutamine (Gln, Q)
MW: 146.1

BASIC

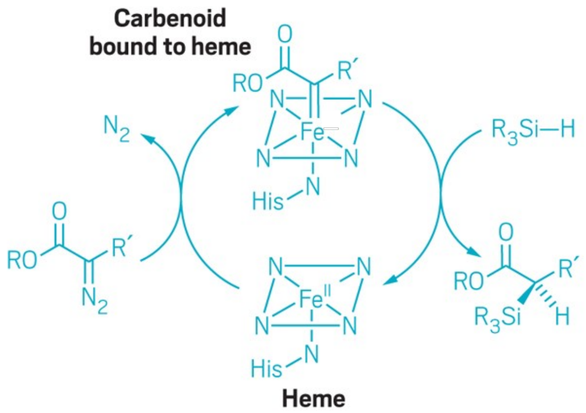
Histidine (His, H)
MW: 155.2, $pK_a = 6.04$

Lysine (Lys, K)
MW: 146.2, $pK_a = 10.79$

Arginine (Arg, R)
MW: 174.2, $pK_a = 12.48$

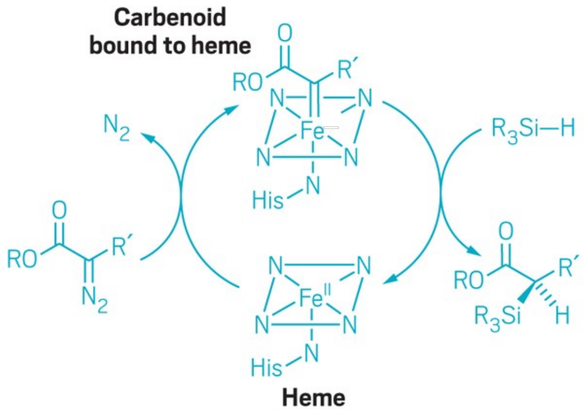
# Why design proteins?

# Why design proteins?

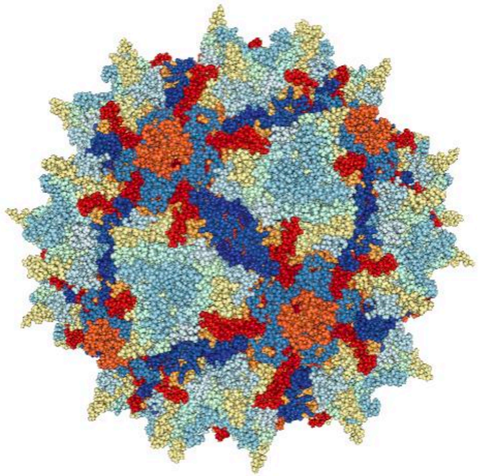



new chemistry

# Why design proteins?

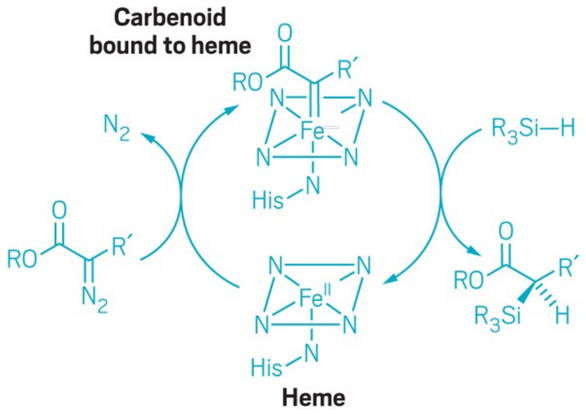



new chemistry

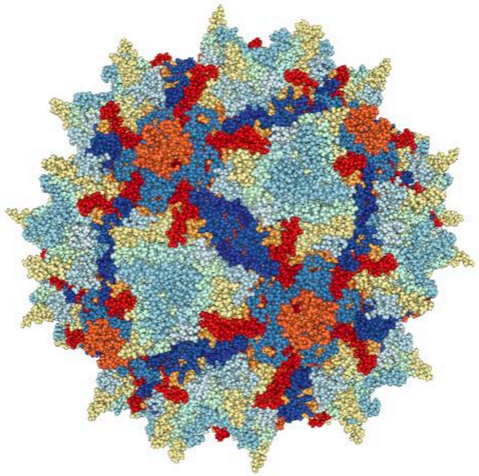


understand function

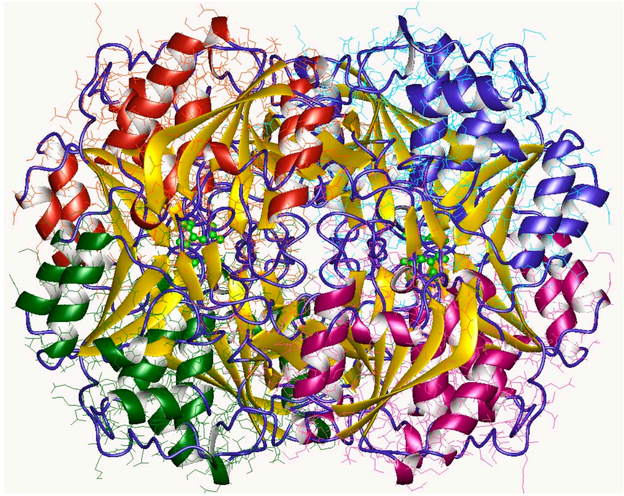
# Why design proteins?




new chemistry

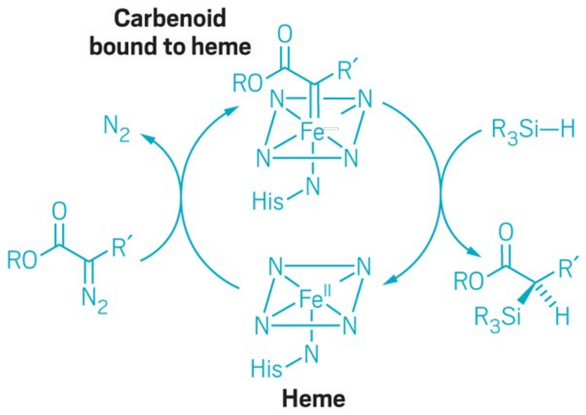


understand function

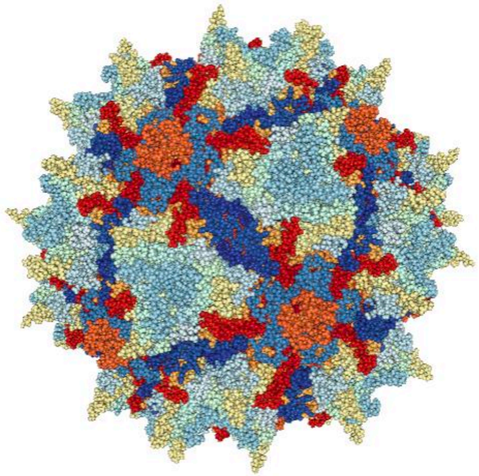


therapeutics

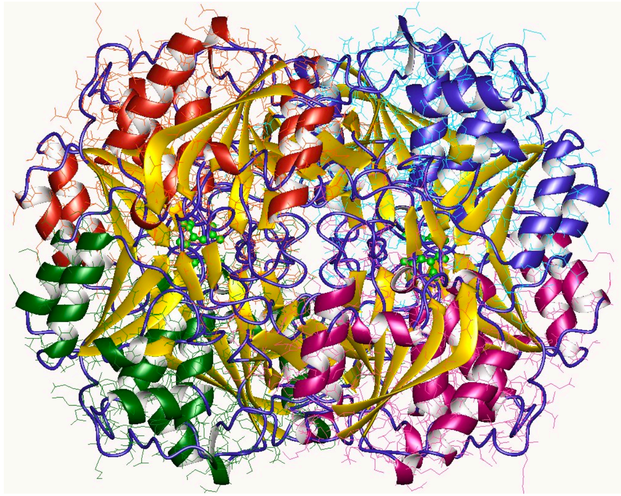
# Why design proteins?



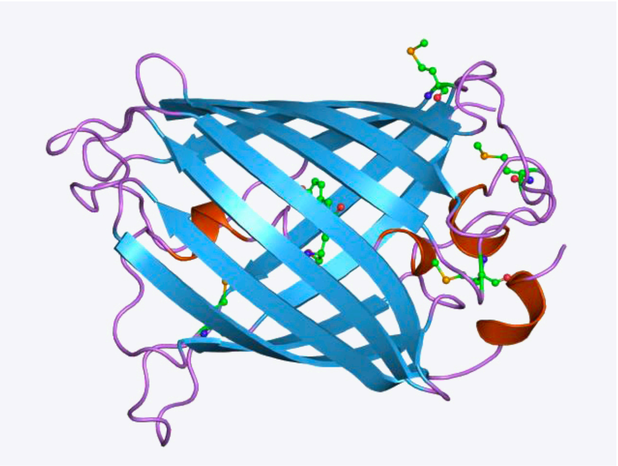

new chemistry



understand function



therapeutics

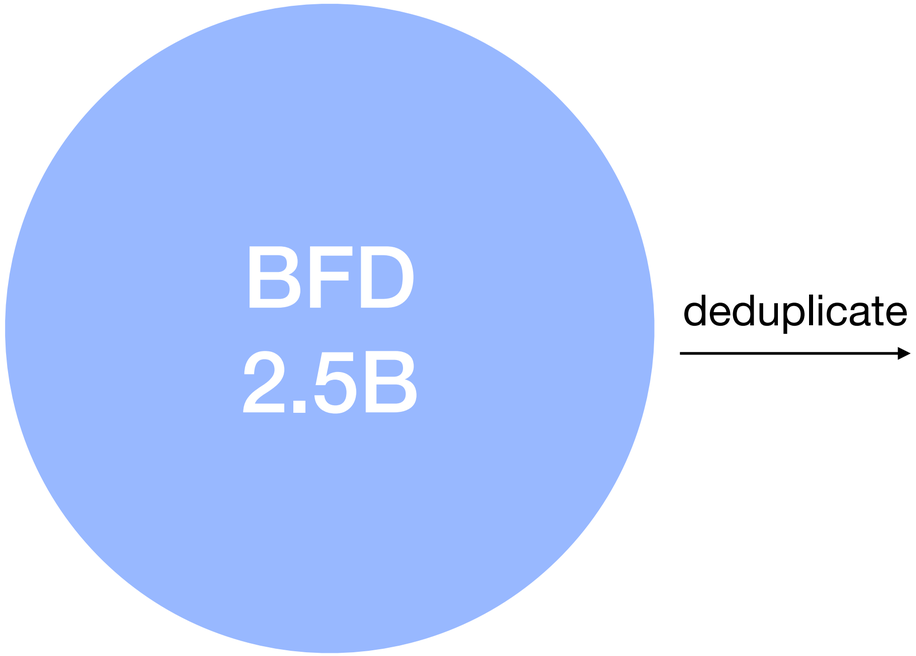


molecular tools

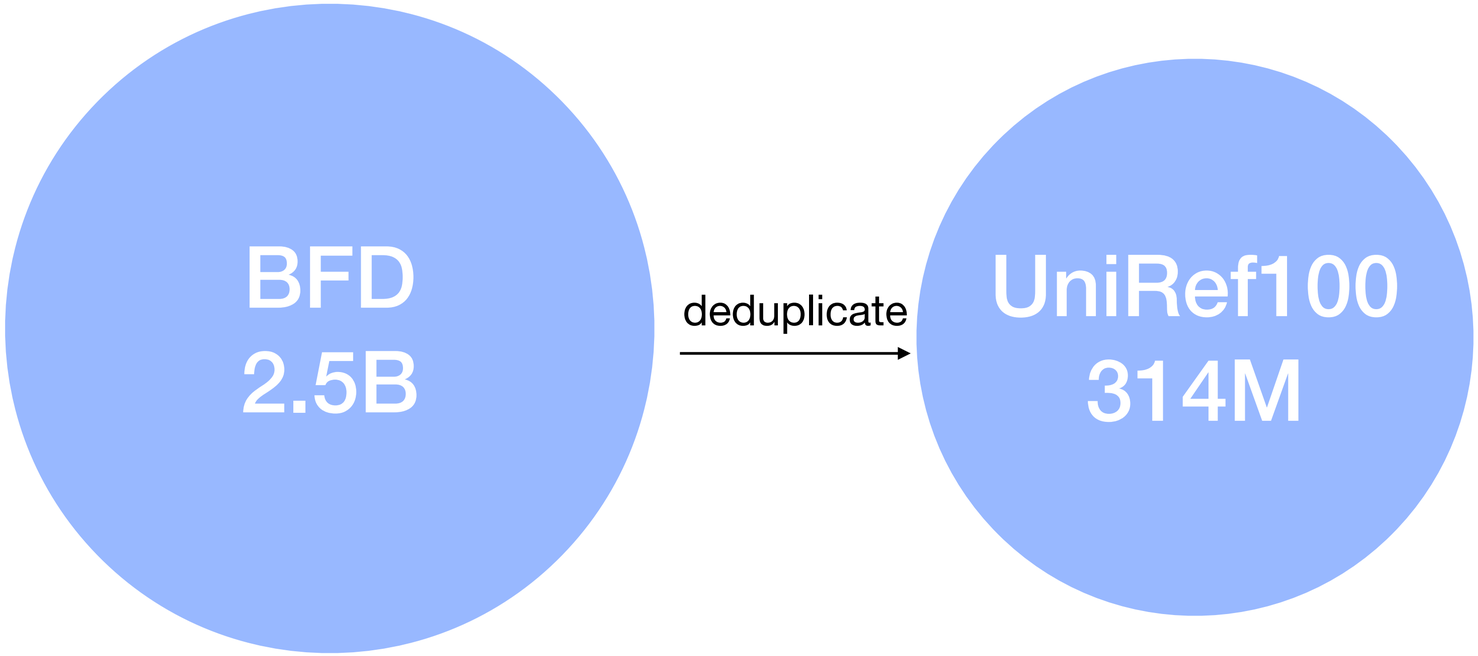
# We have access to large protein databases

# We have access to large protein databases

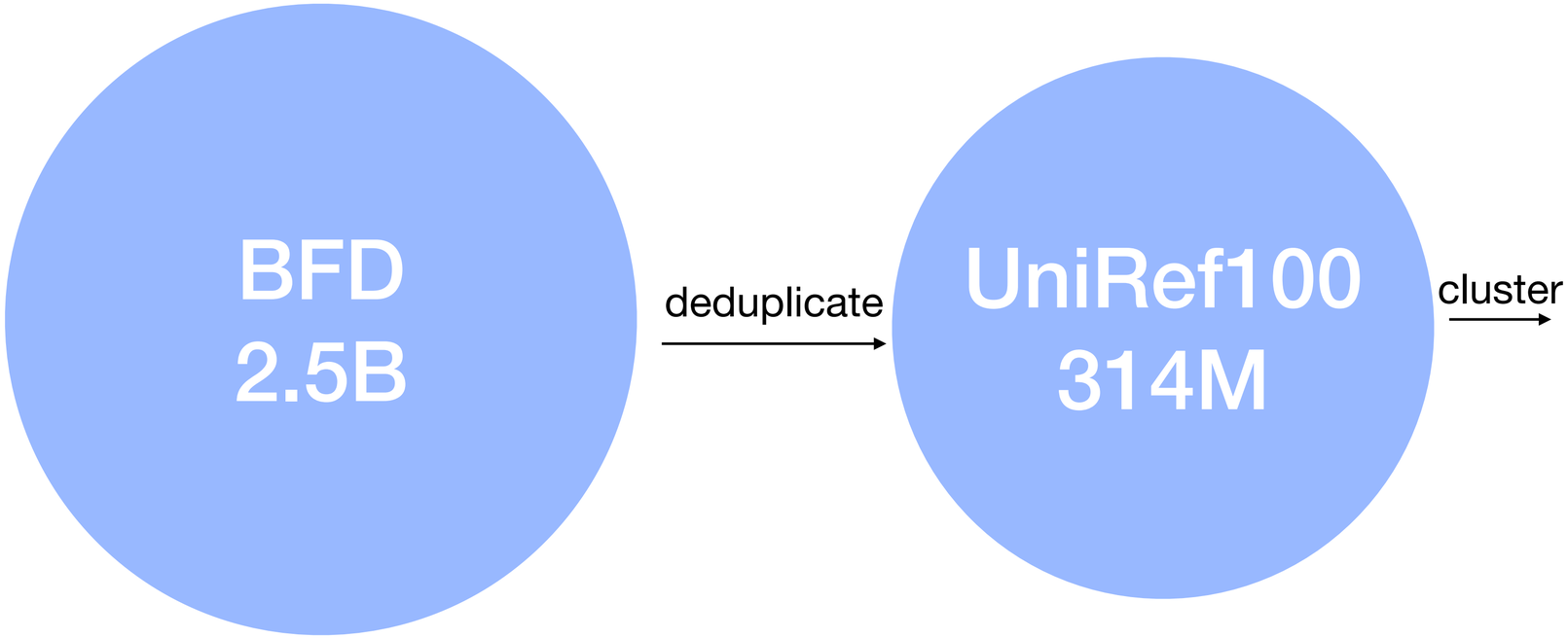
# We have access to large protein databases



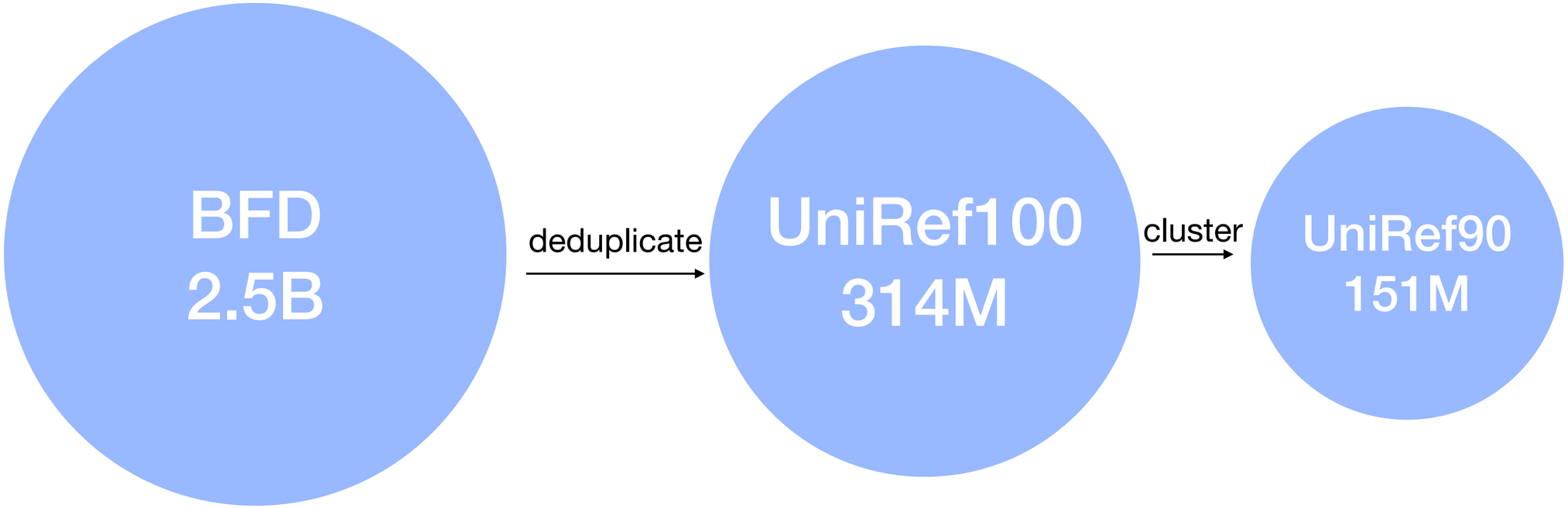
# We have access to large protein databases



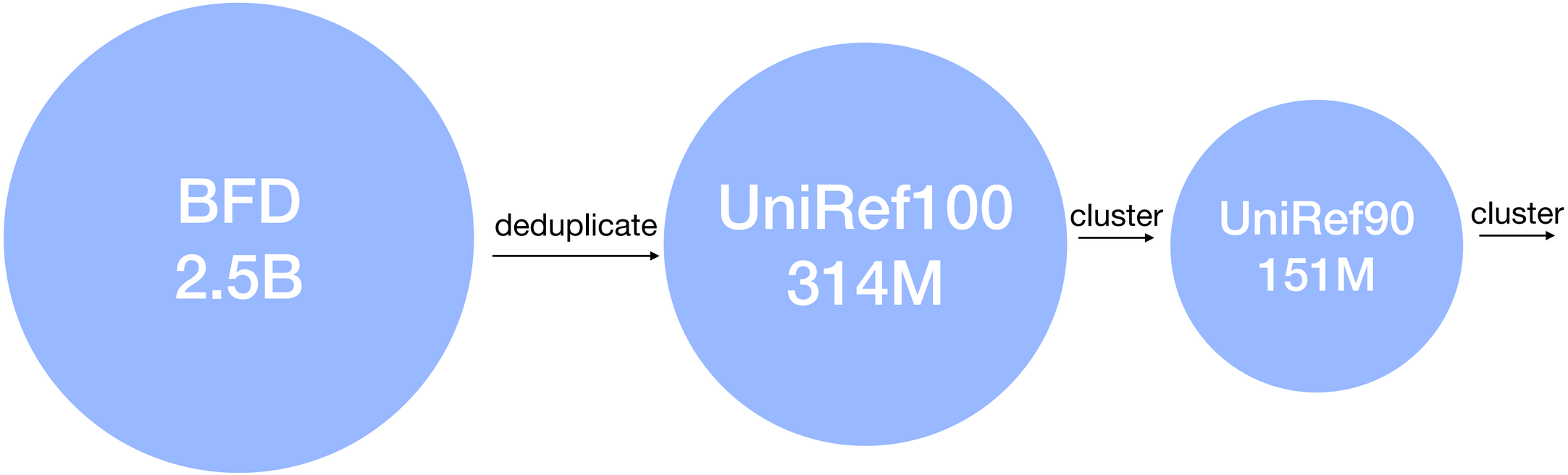
# We have access to large protein databases



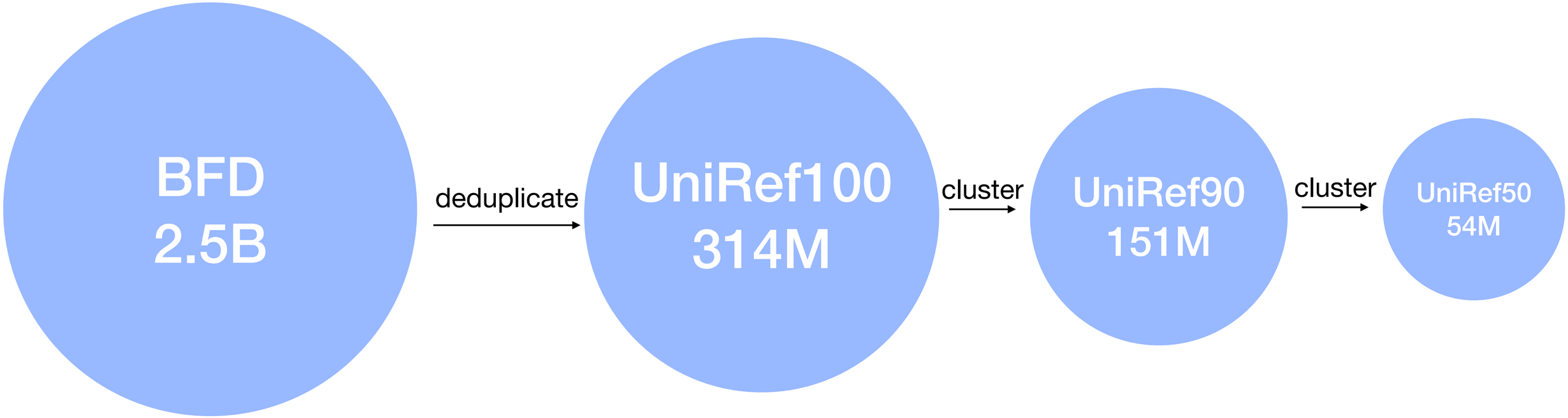
# We have access to large protein databases



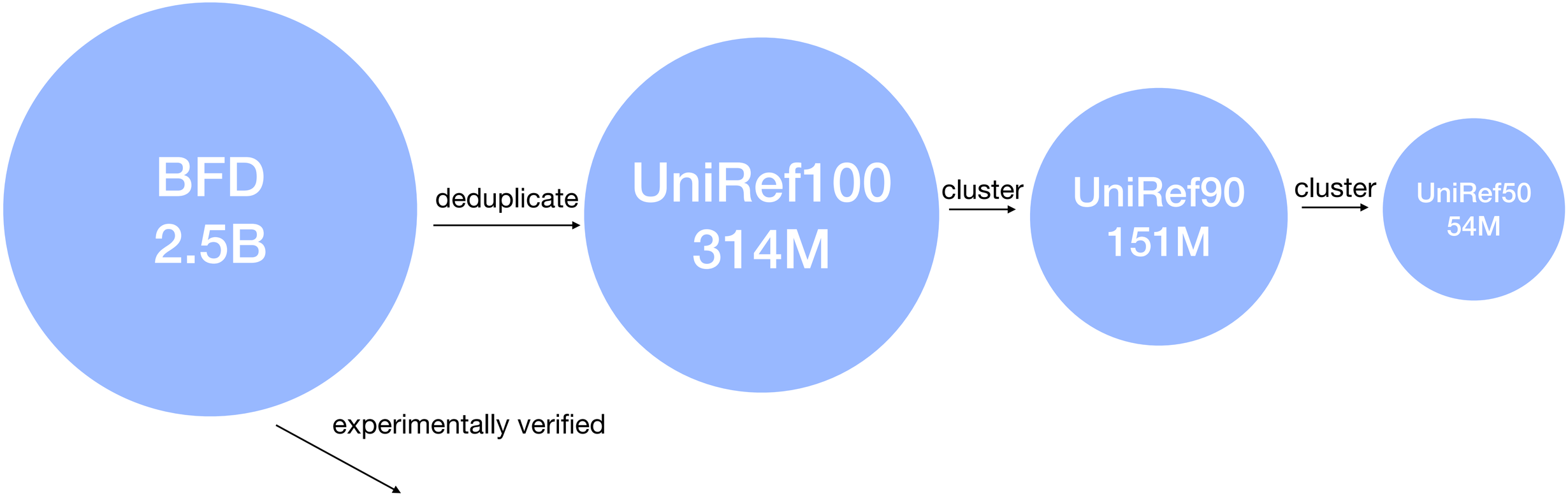
# We have access to large protein databases



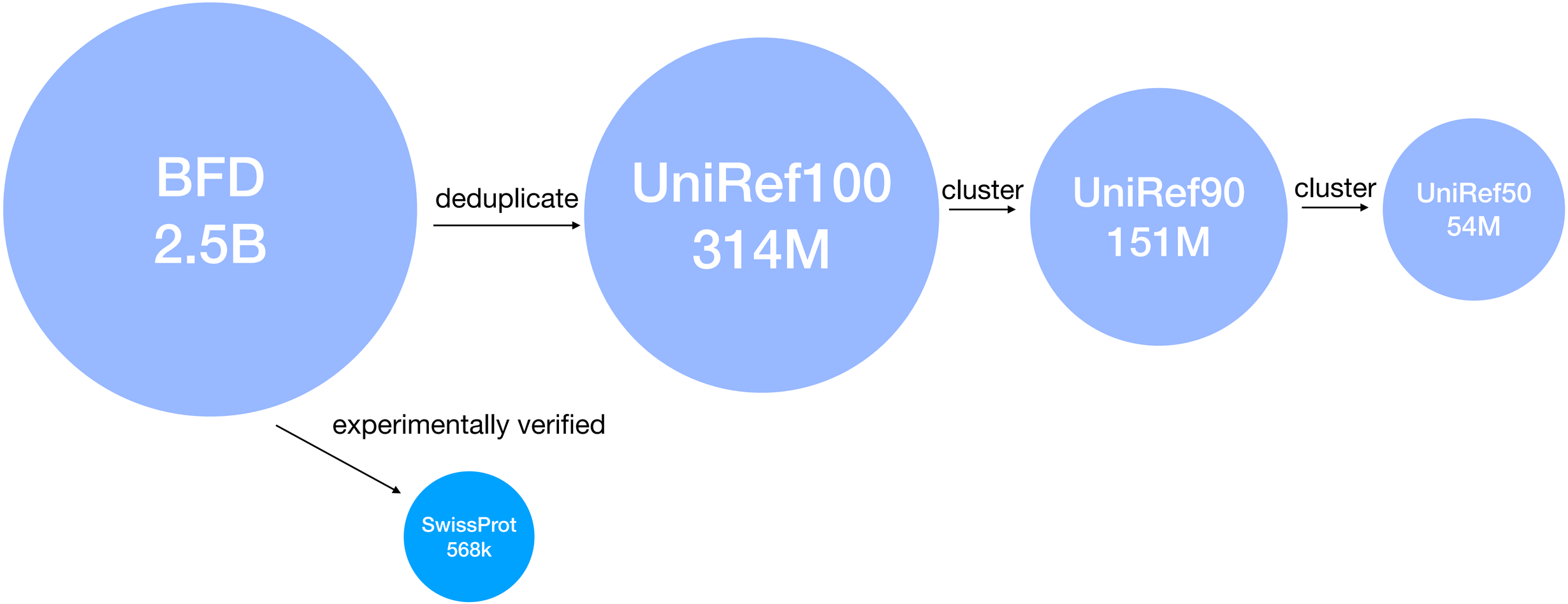
# We have access to large protein databases



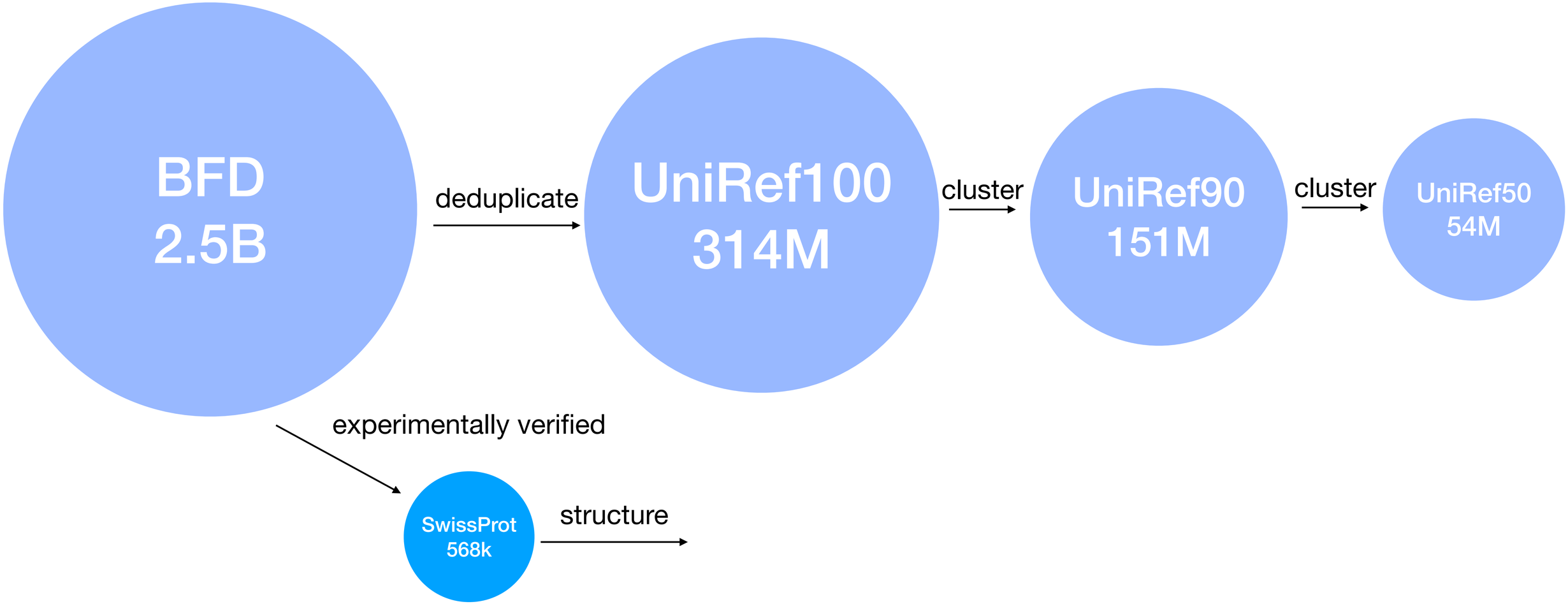
# We have access to large protein databases



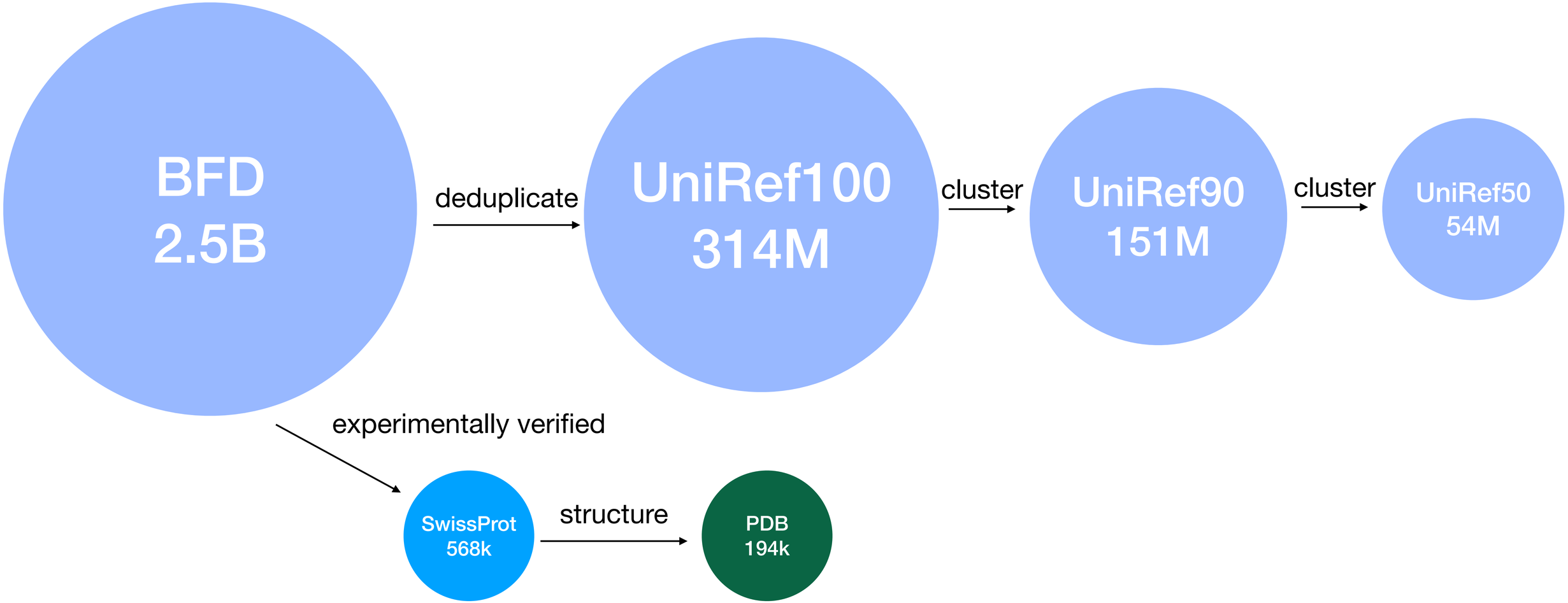
# We have access to large protein databases



# We have access to large protein databases



# We have access to large protein databases



# We have access to large protein databases

BFD
2.5B

deduplicate

UniRef100
314M

cluster

UniRef90
151M

cluster

UniRef50
54M

experimentally verified

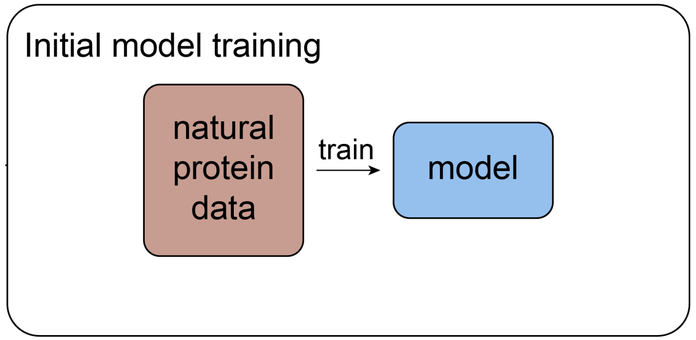
SwissProt
568k

structure

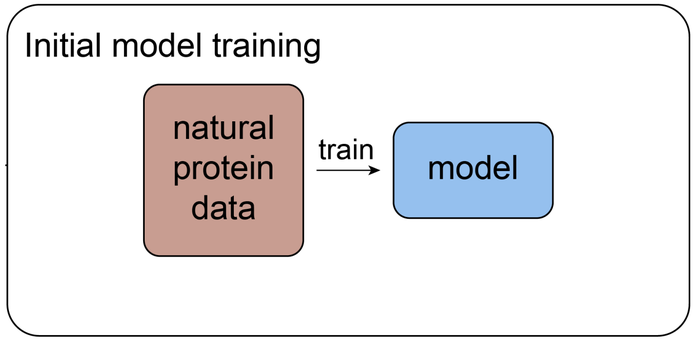
PDB
194k

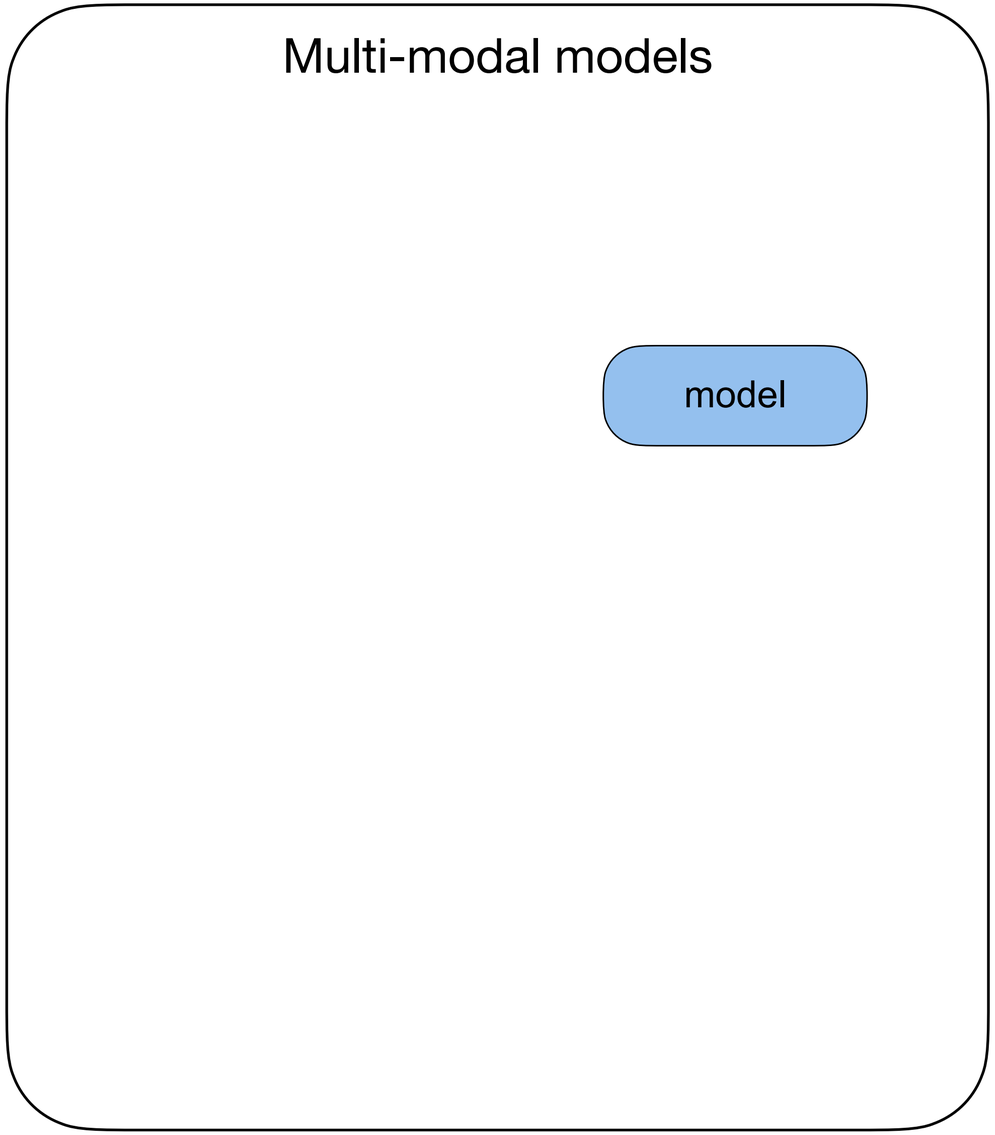
200 measurements

# Use multiple data modalities to discover and design proteins

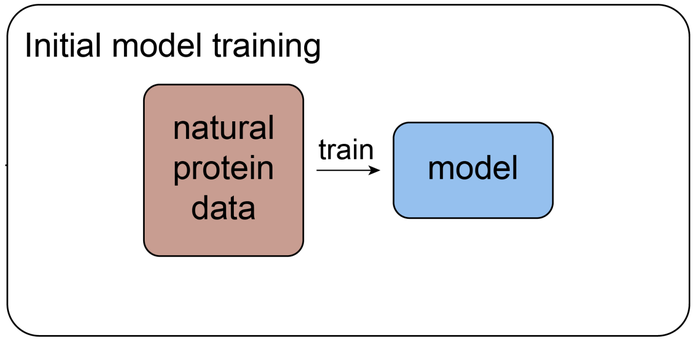


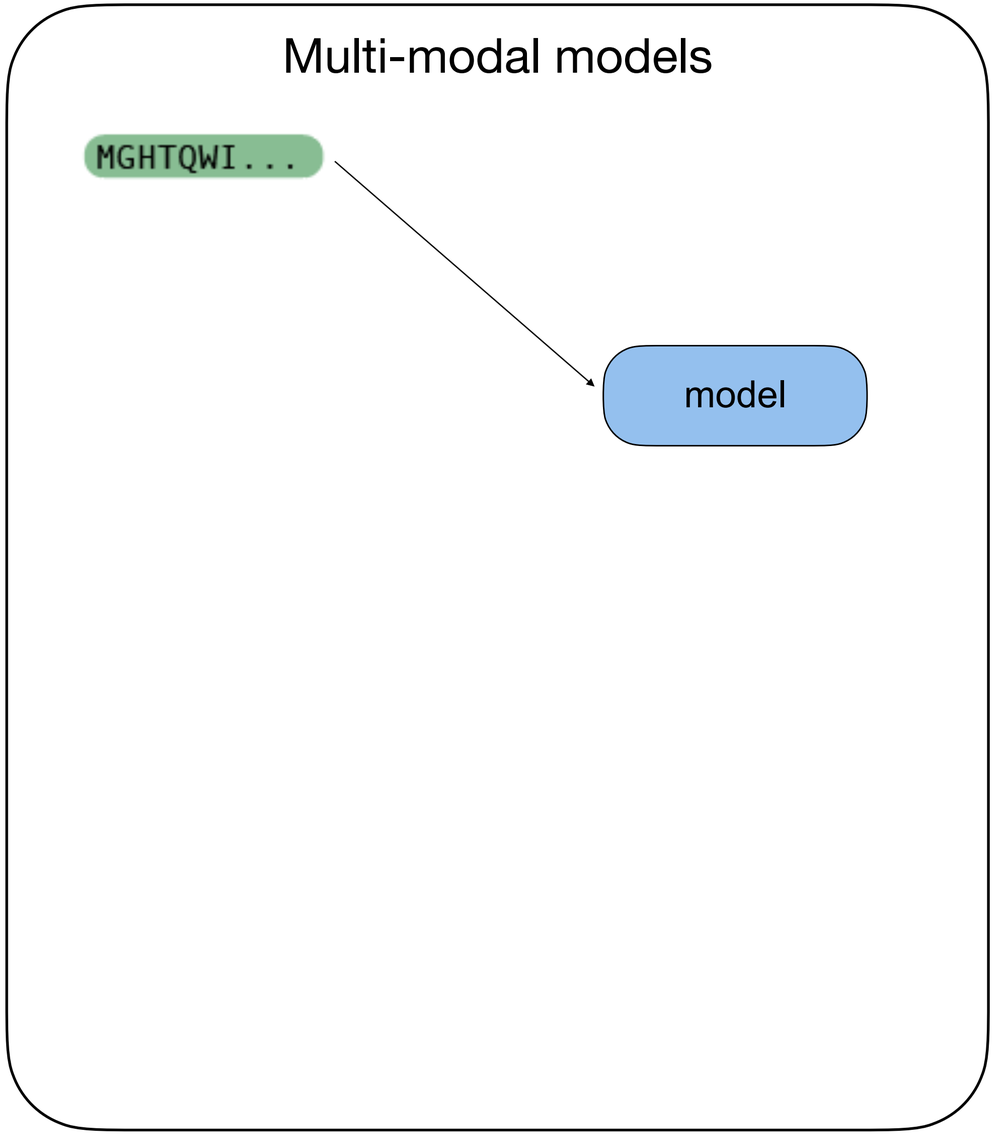
# Use multiple data modalities to discover and design proteins



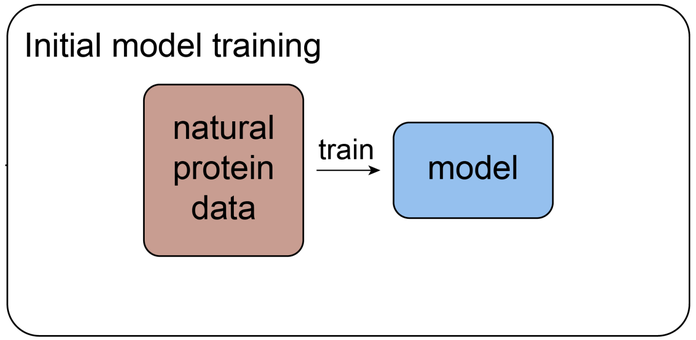



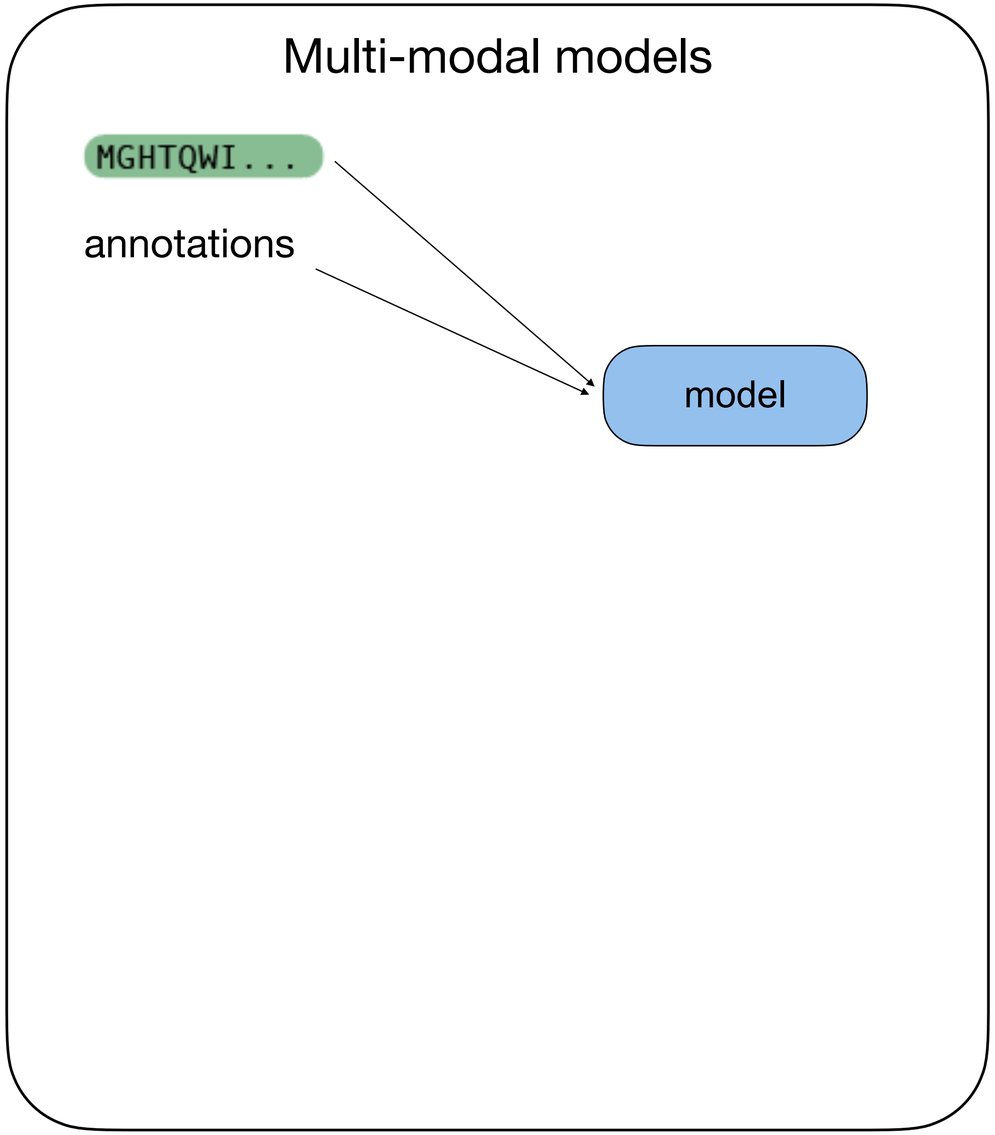
# Use multiple data modalities to discover and design proteins



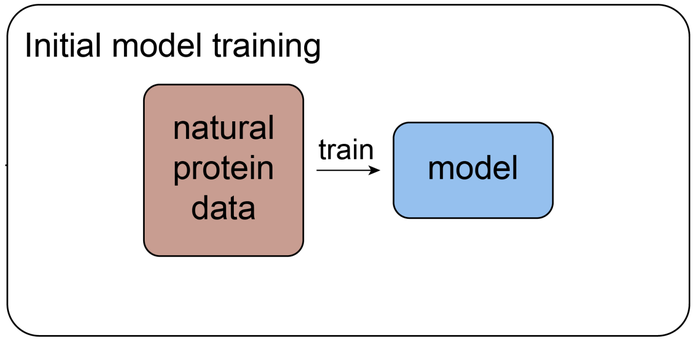



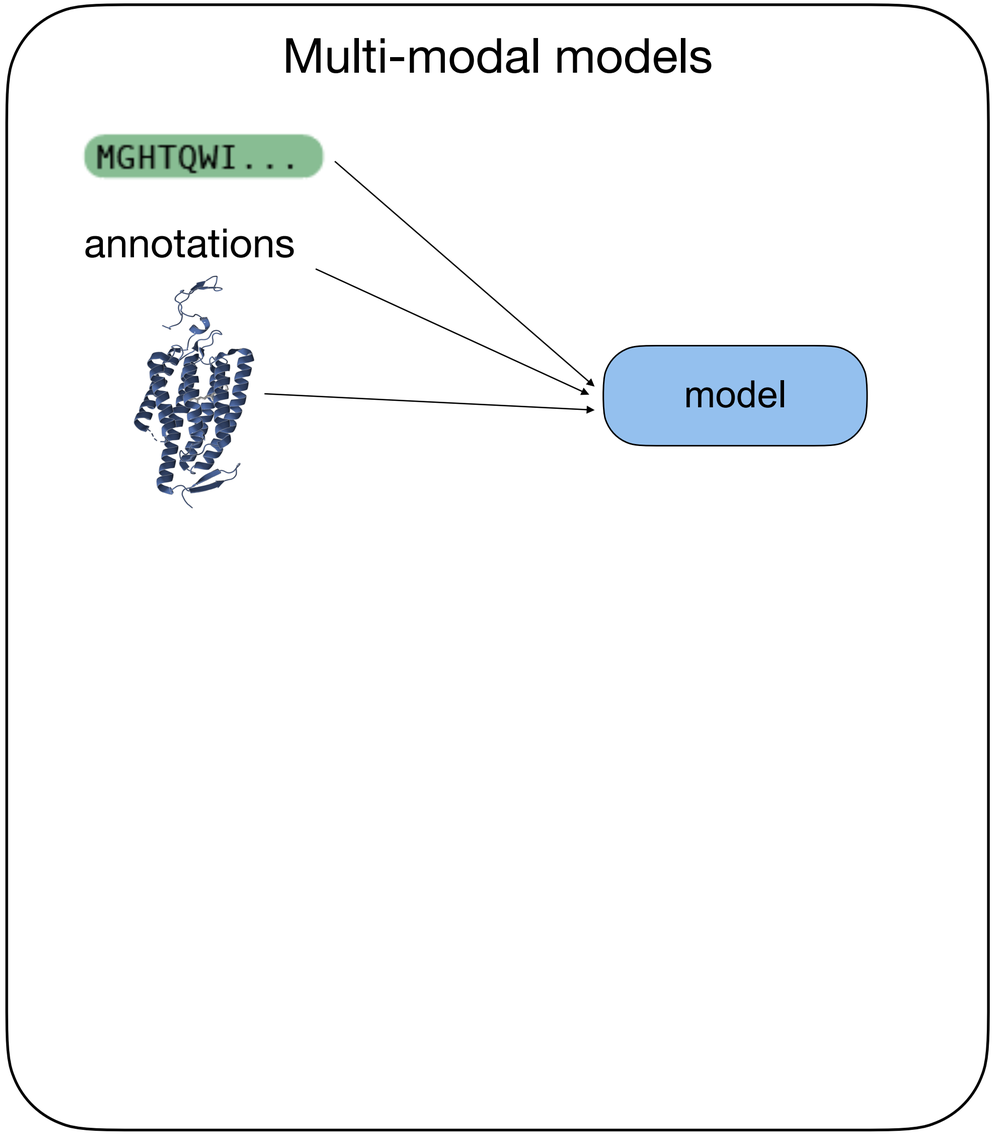
# Use multiple data modalities to discover and design proteins



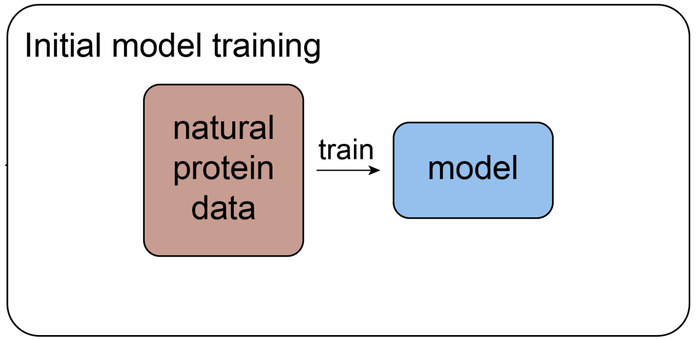



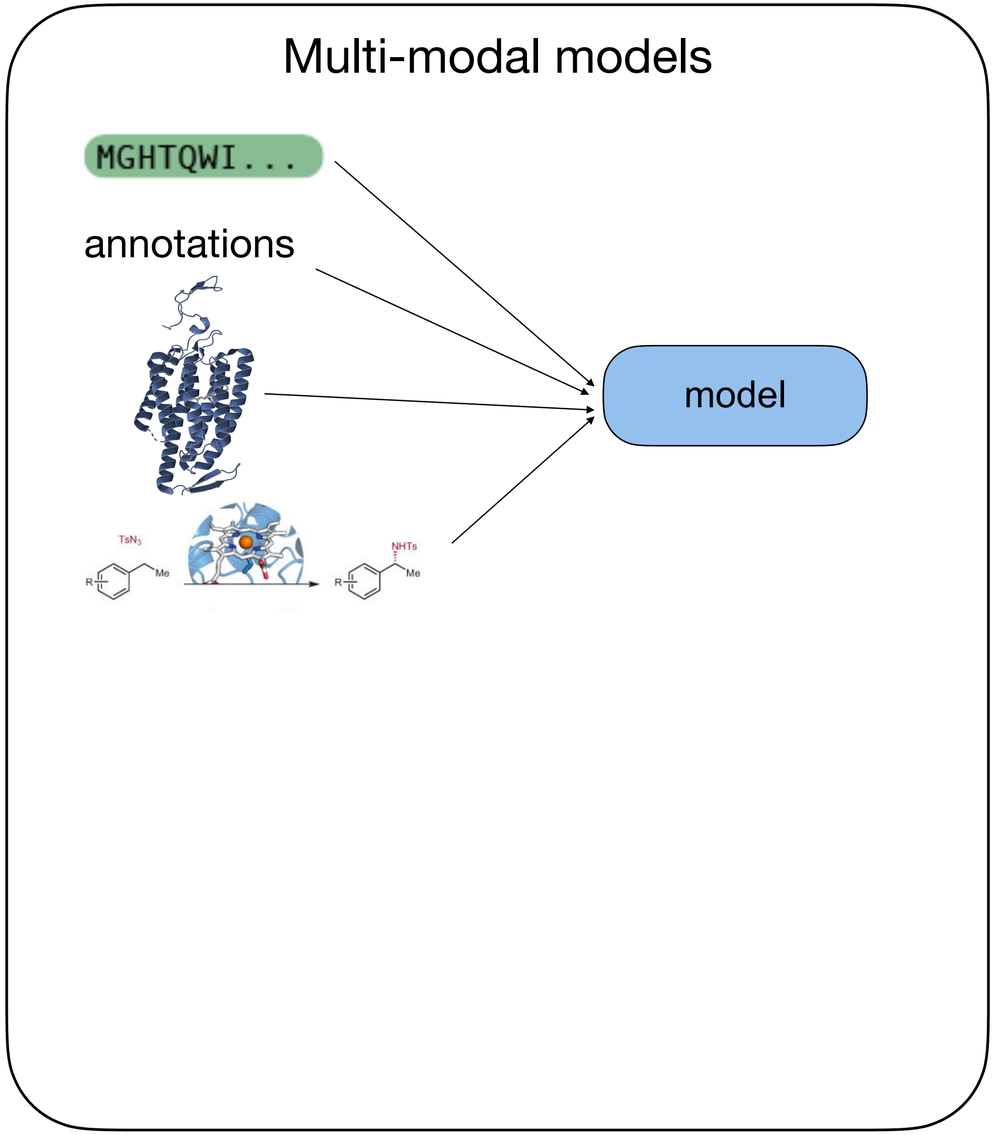
# Use multiple data modalities to discover and design proteins



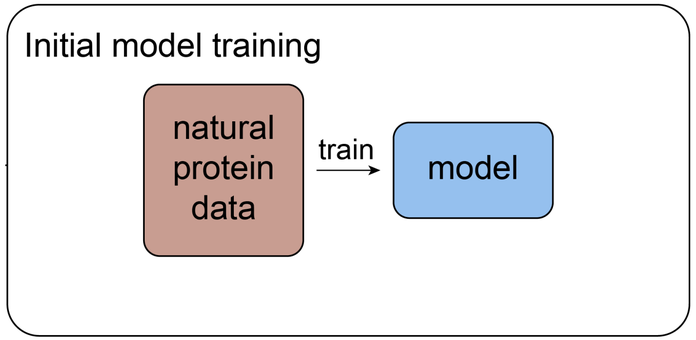



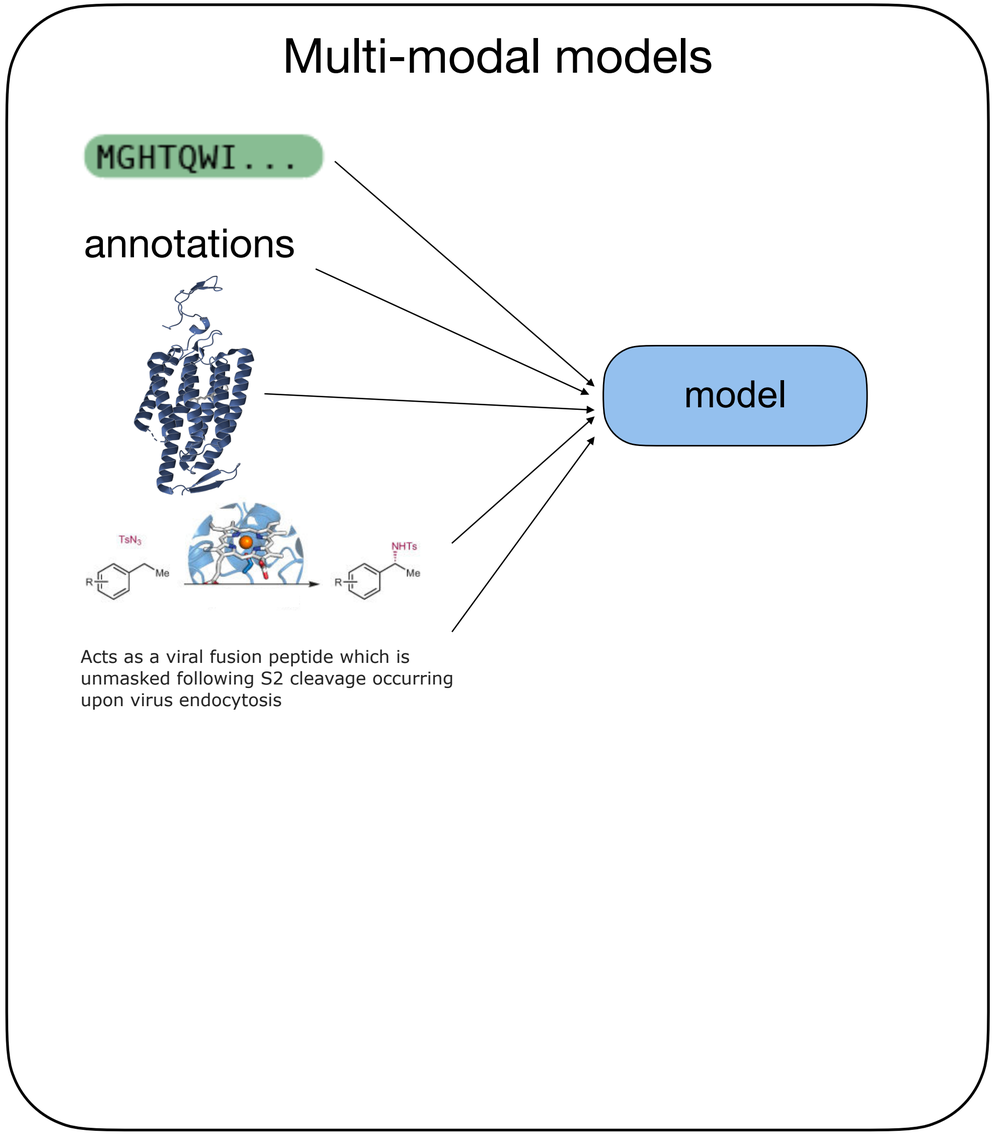
# Use multiple data modalities to discover and design proteins



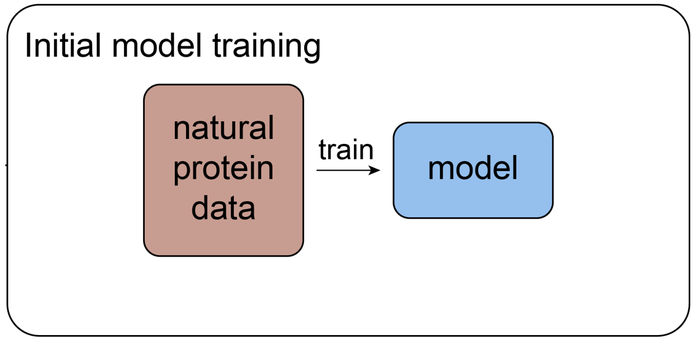



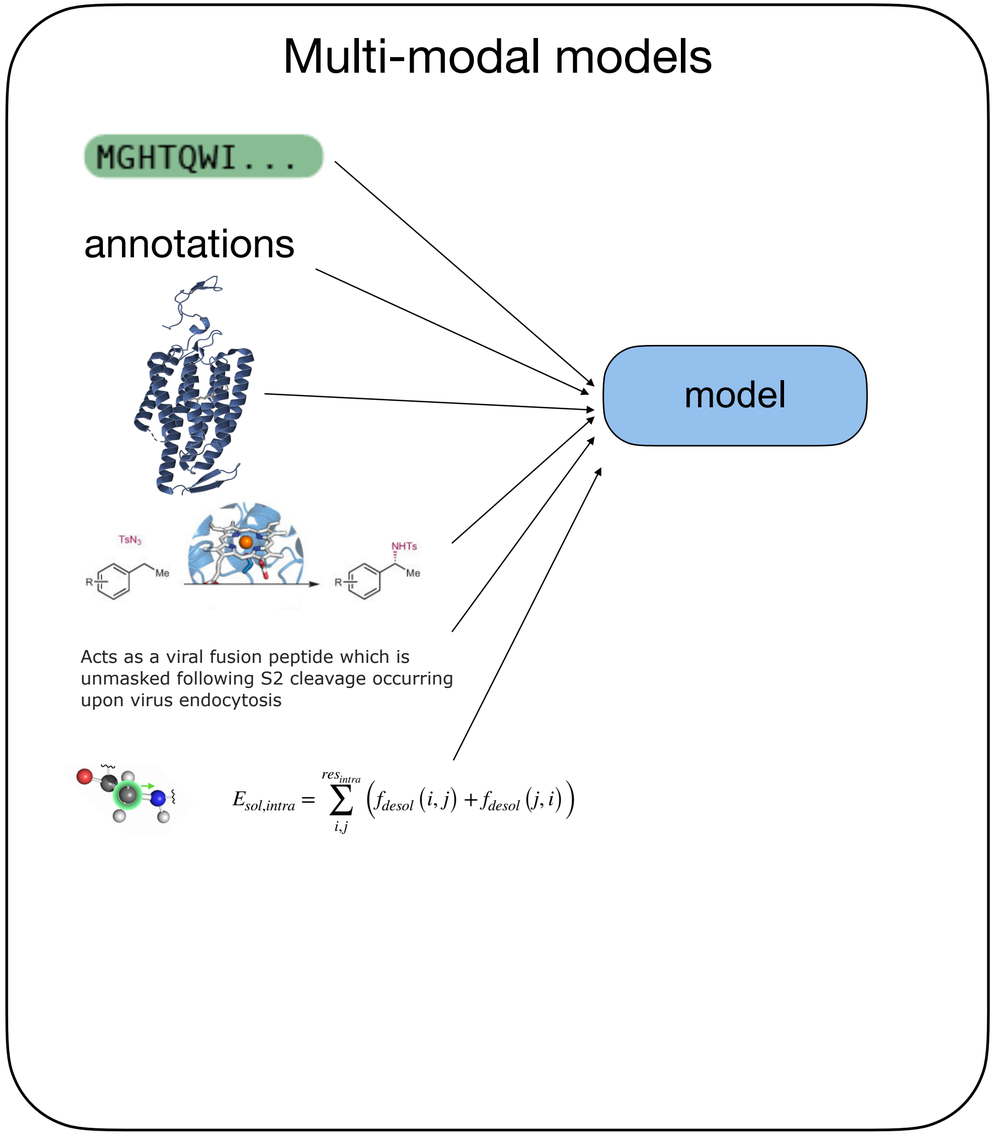
# Use multiple data modalities to discover and design proteins



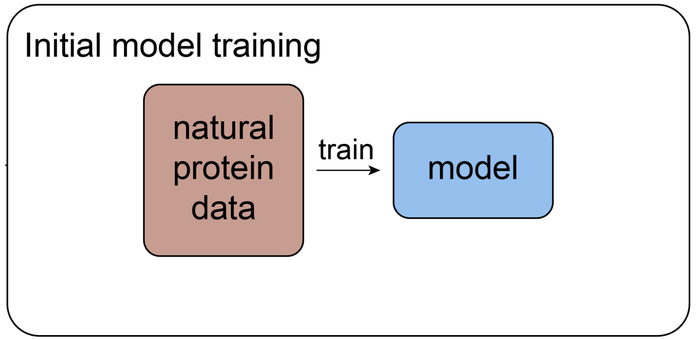



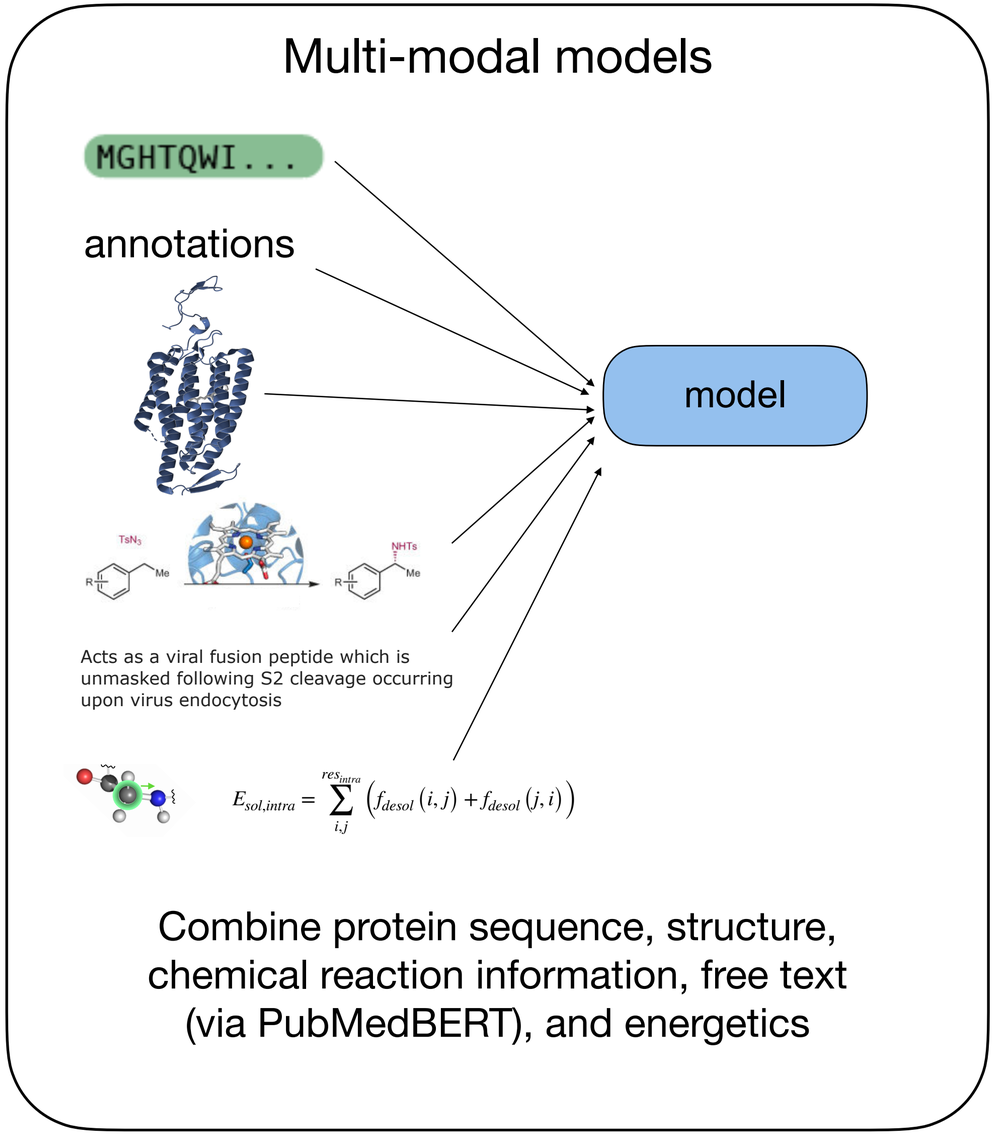
# Use multiple data modalities to discover and design proteins



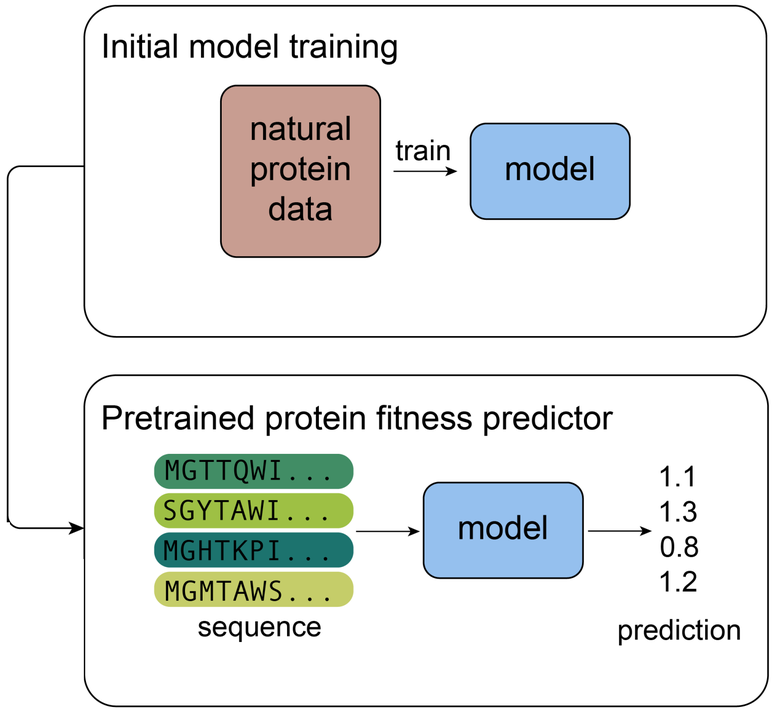



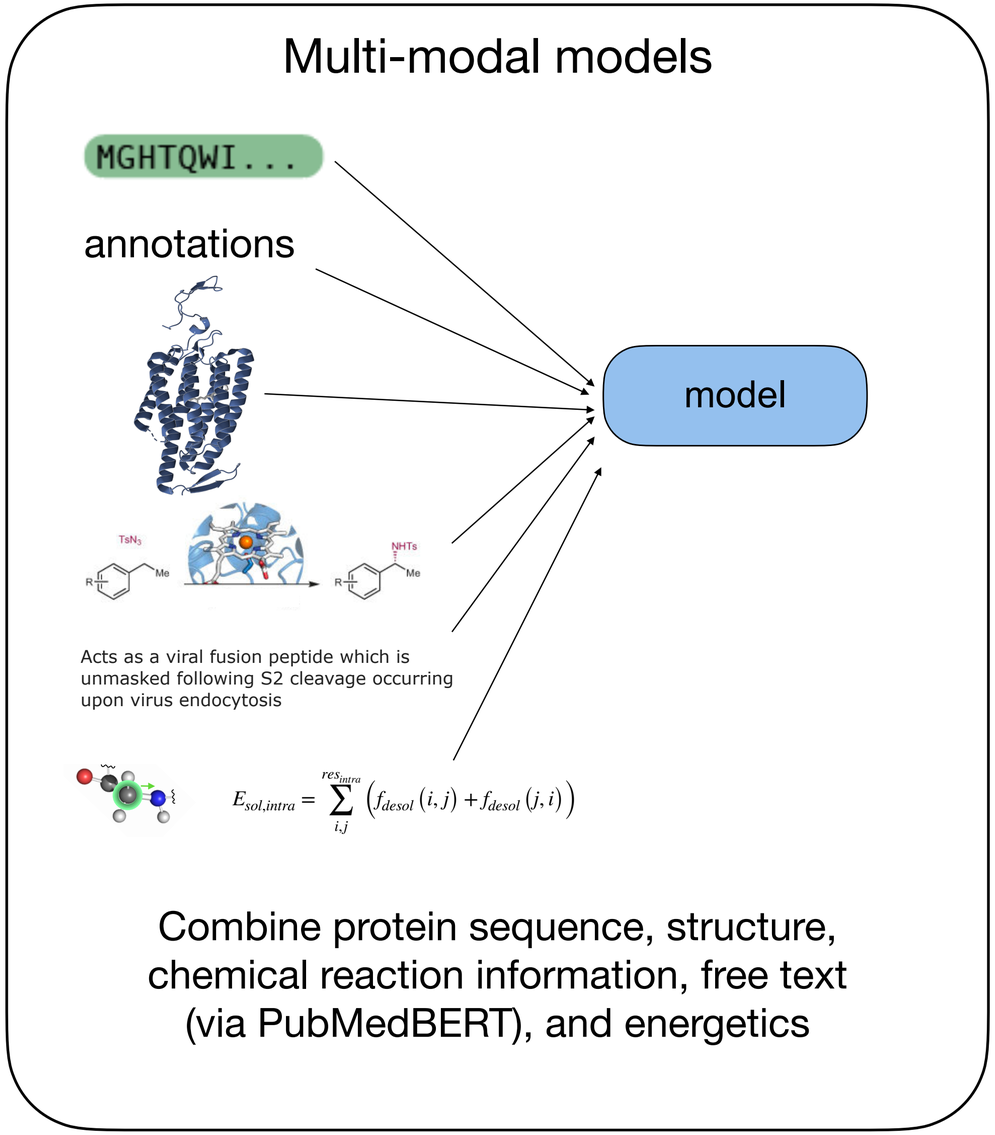
# Use multiple data modalities to discover and design proteins



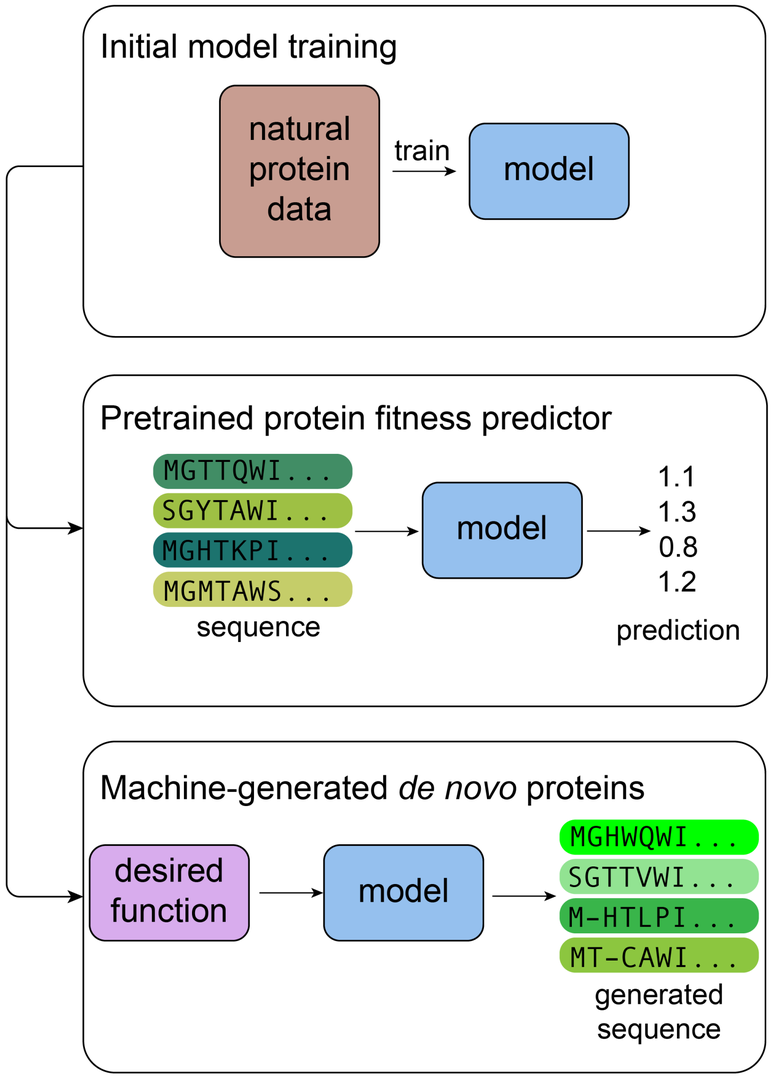



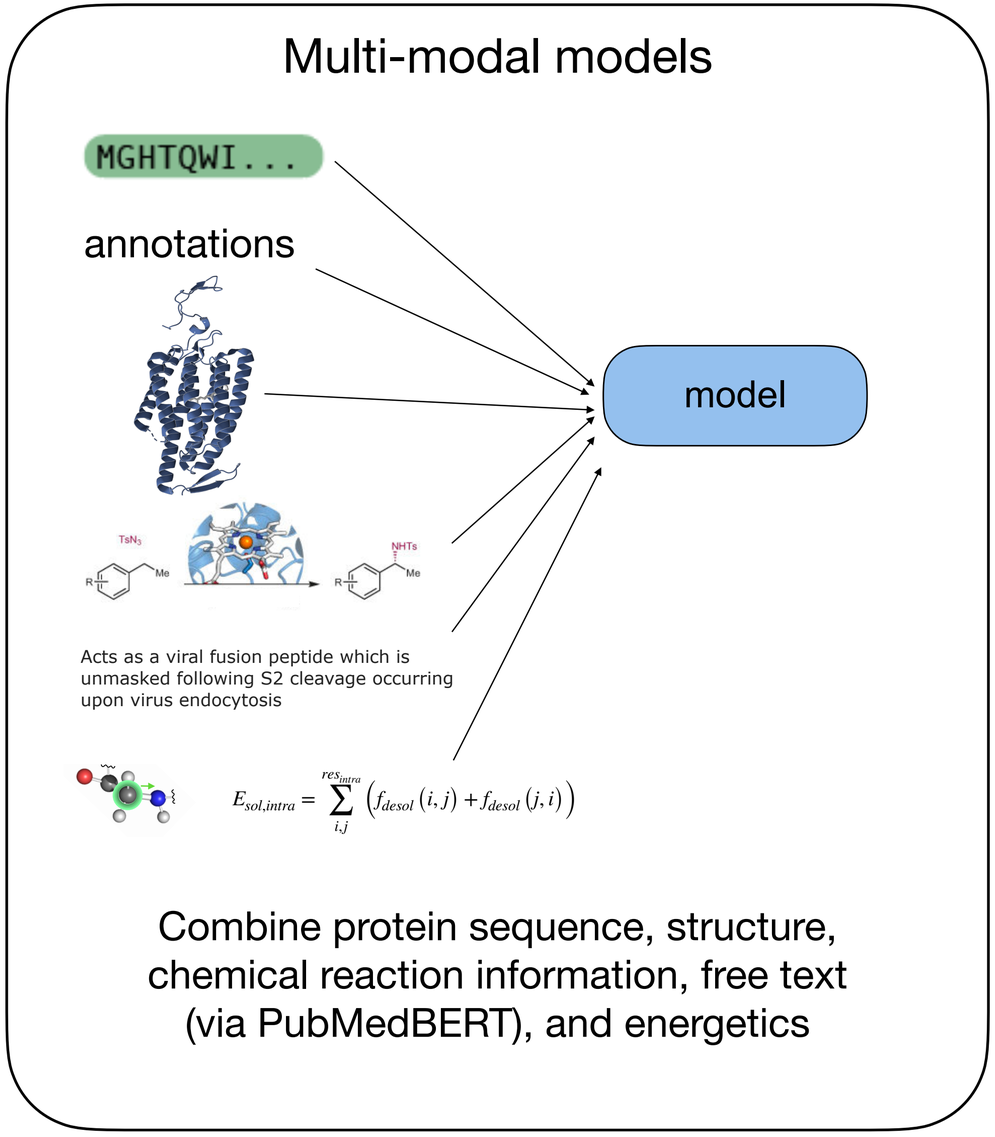
# Use multiple data modalities to discover and design proteins



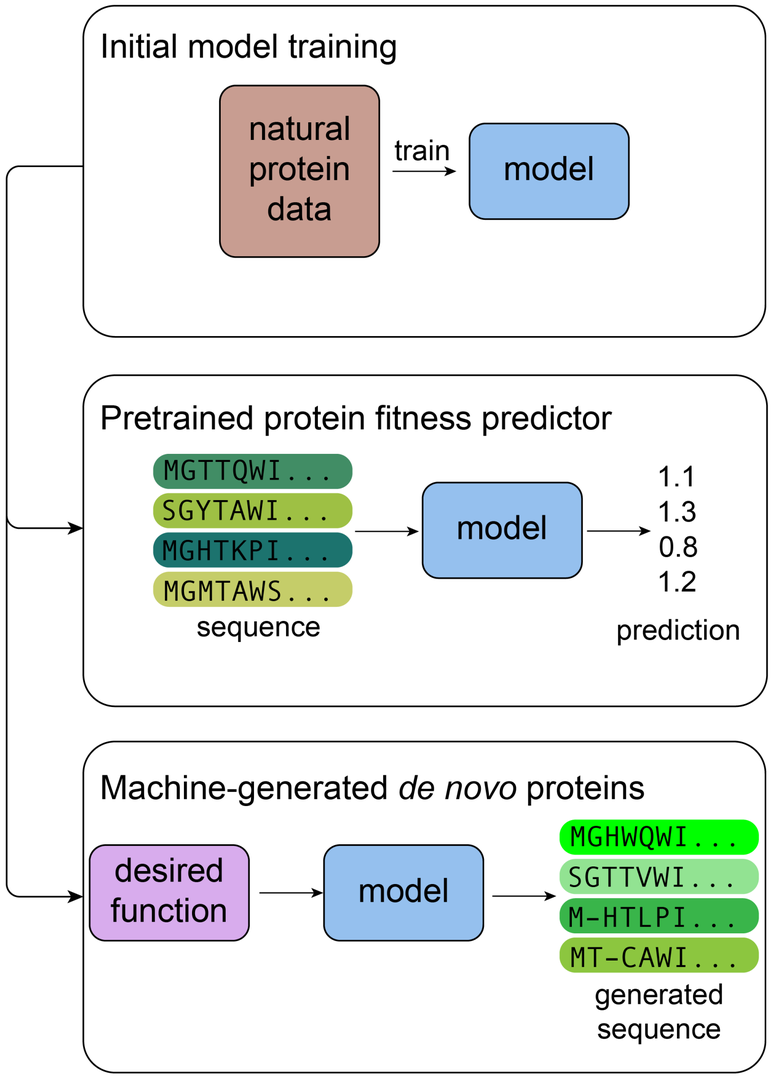



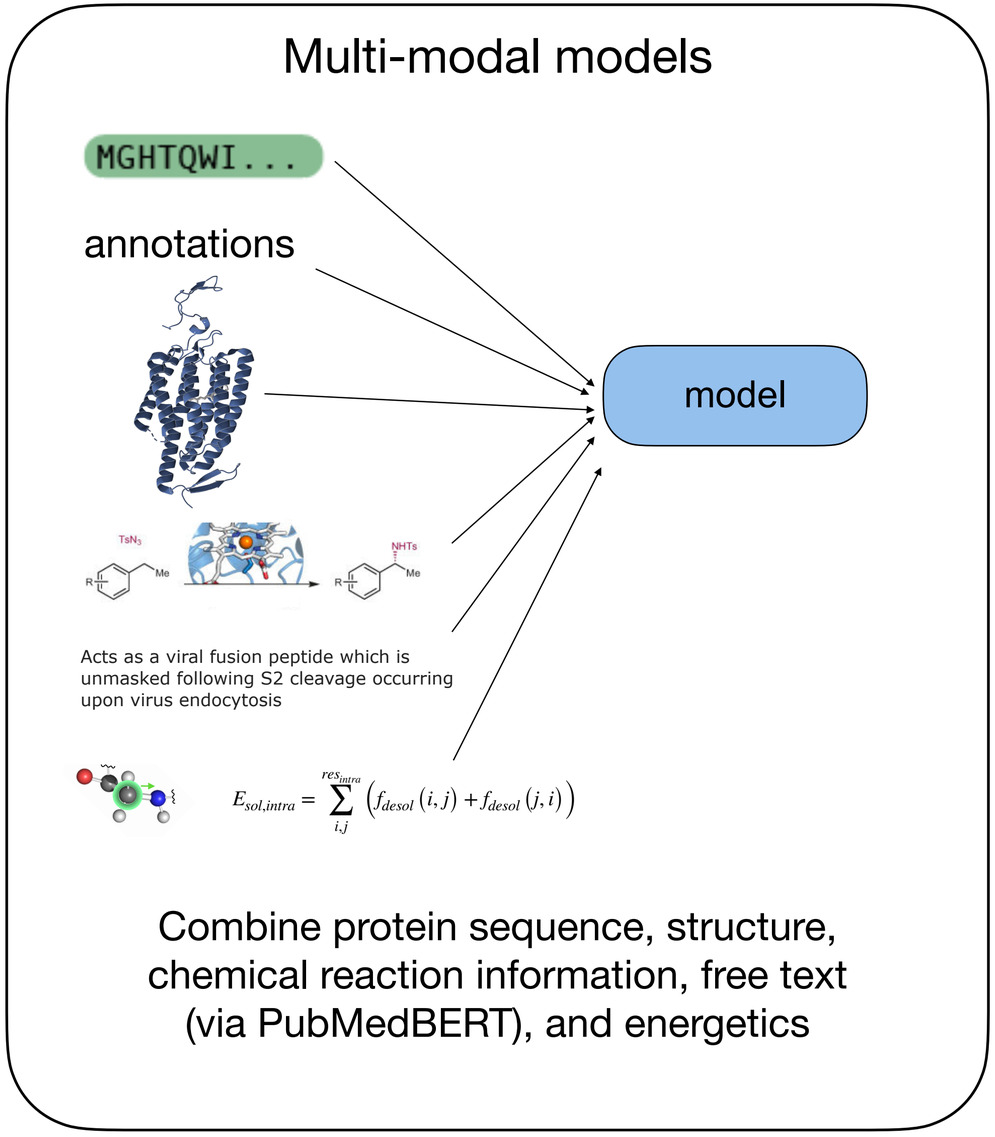
# Use multiple data modalities to discover and design proteins



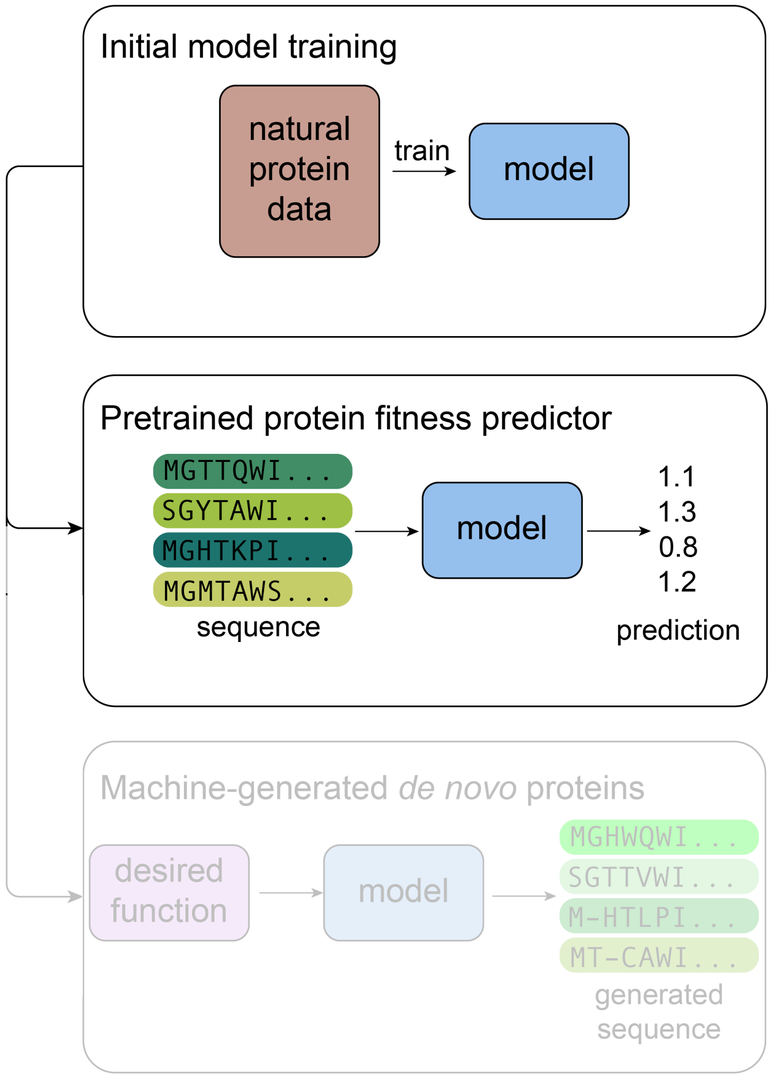



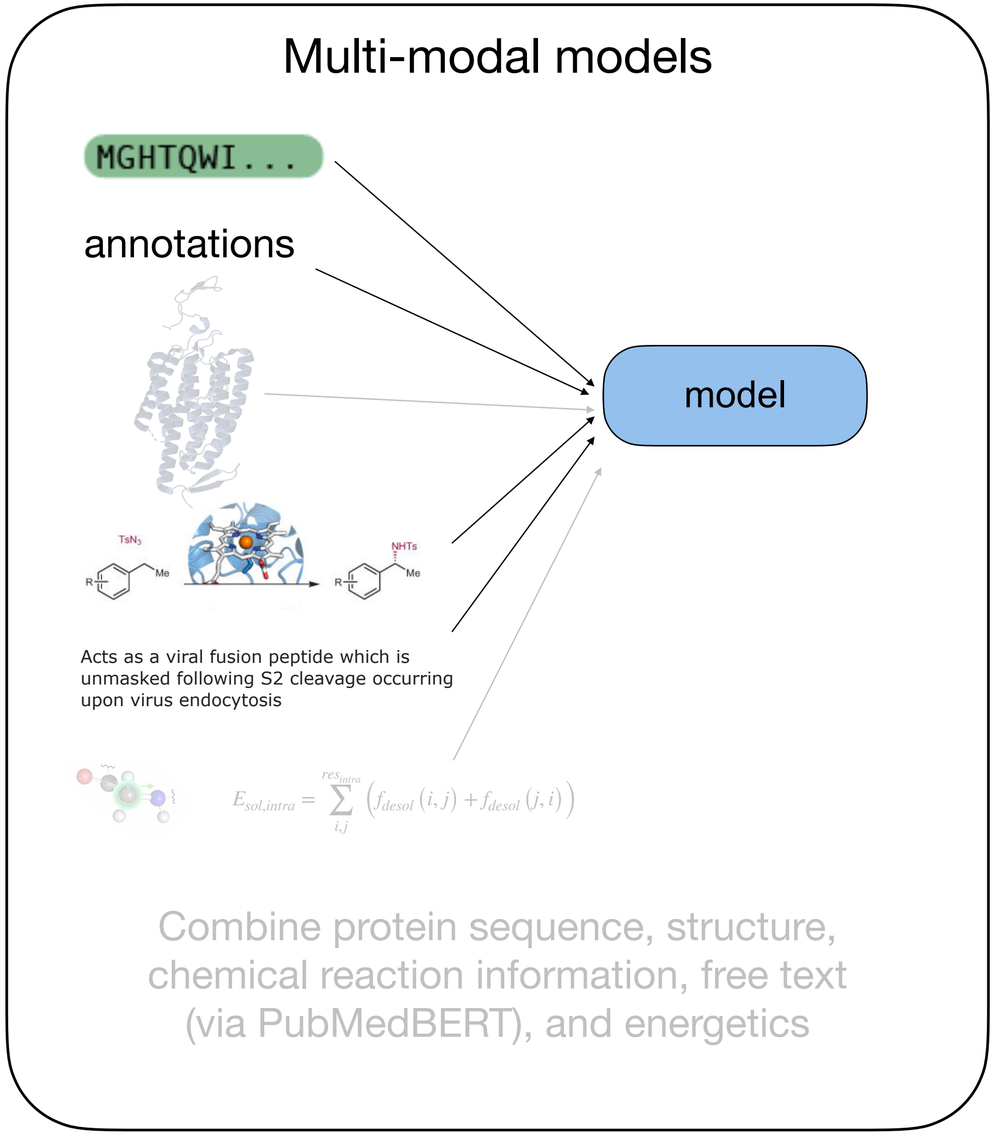
# Use multiple data modalities to discover and design proteins



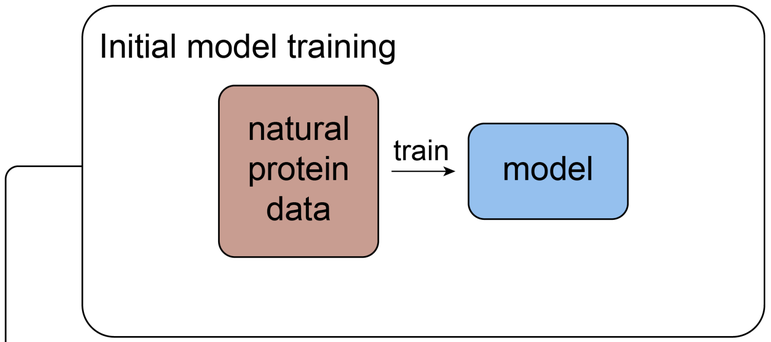
# Use multiple data modalities to discover and design proteins

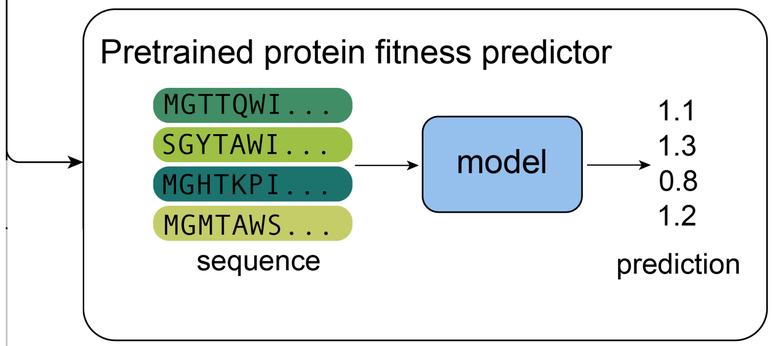


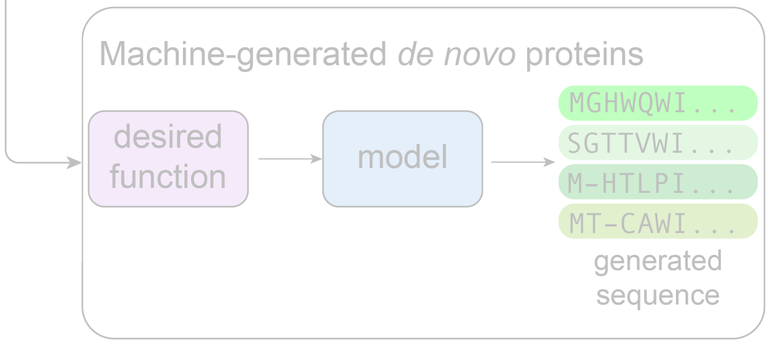



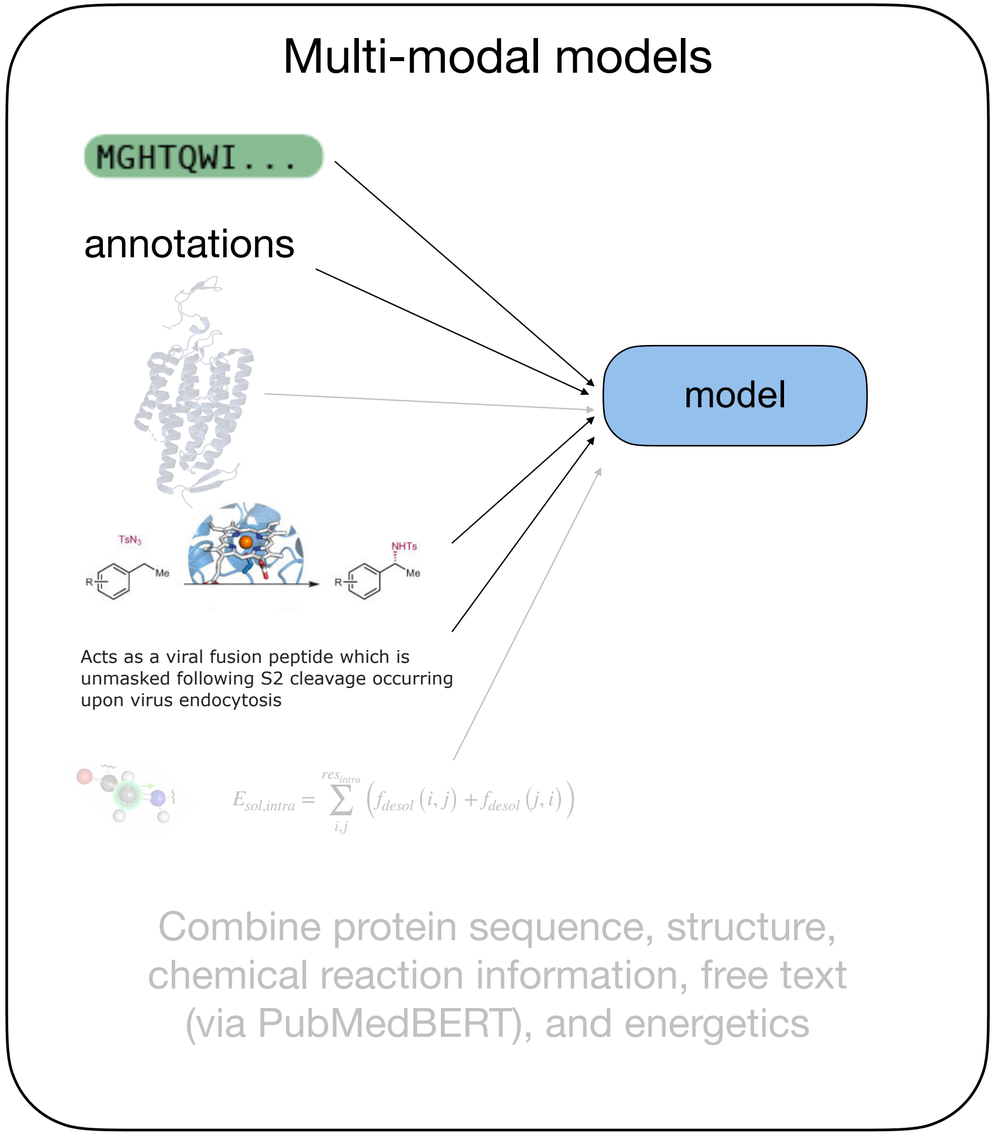
clusters of

# Use multiple data modalities to discover and design proteins

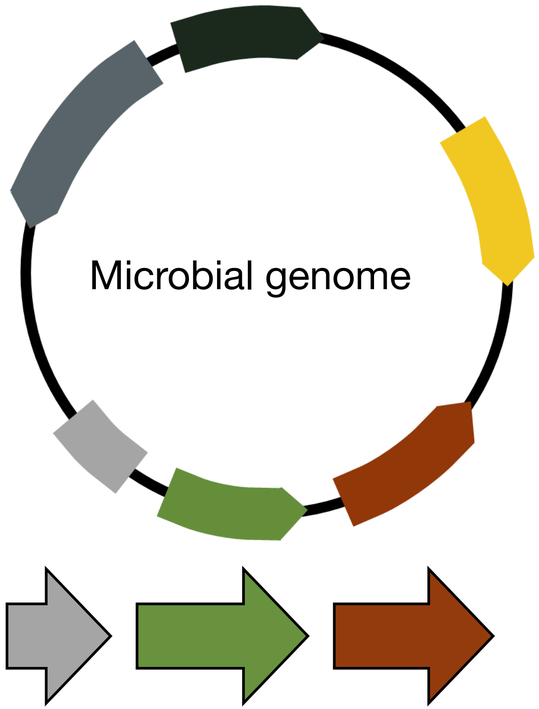


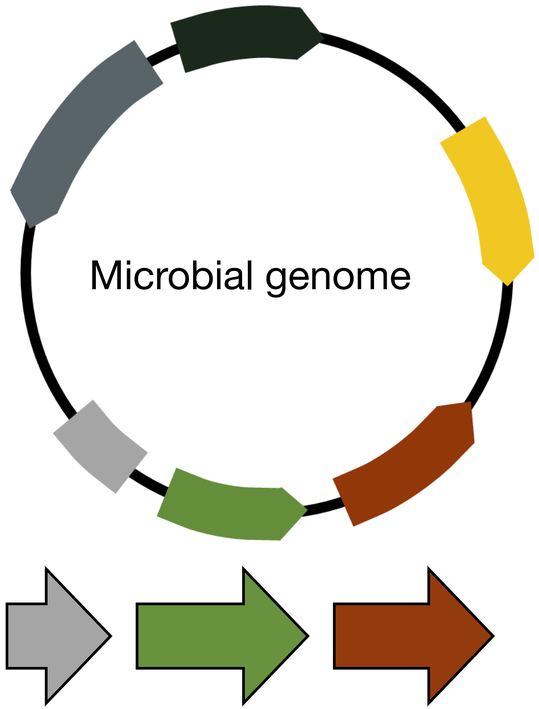



# Biosynthetic gene clusters encode natural products

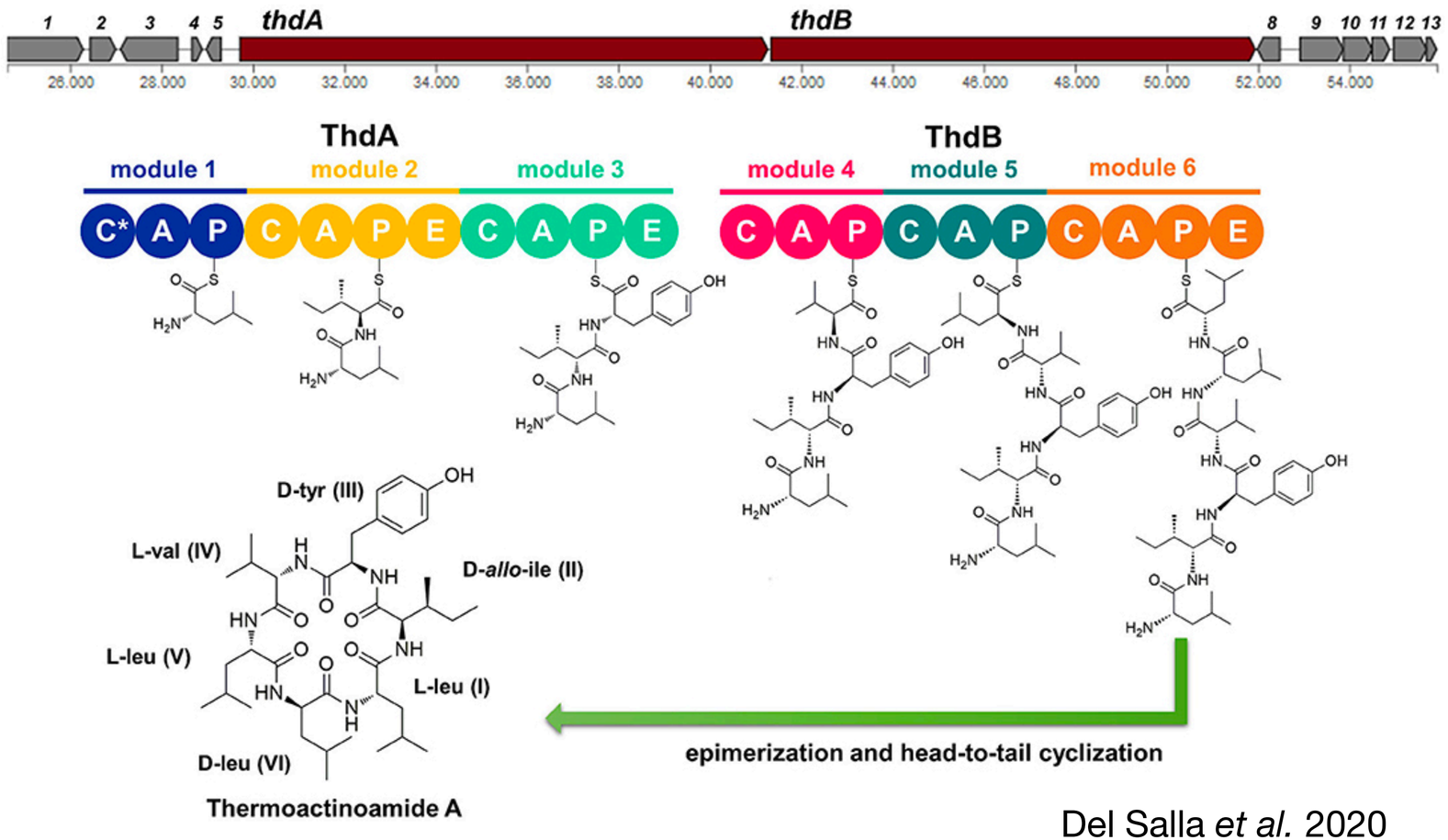



Co-regulated
Physical proximity
Modular
Secondary metabolites

# Biosynthetic gene clusters encode natural products




Co-regulated
Physical proximity
Modular
Secondary metabolites

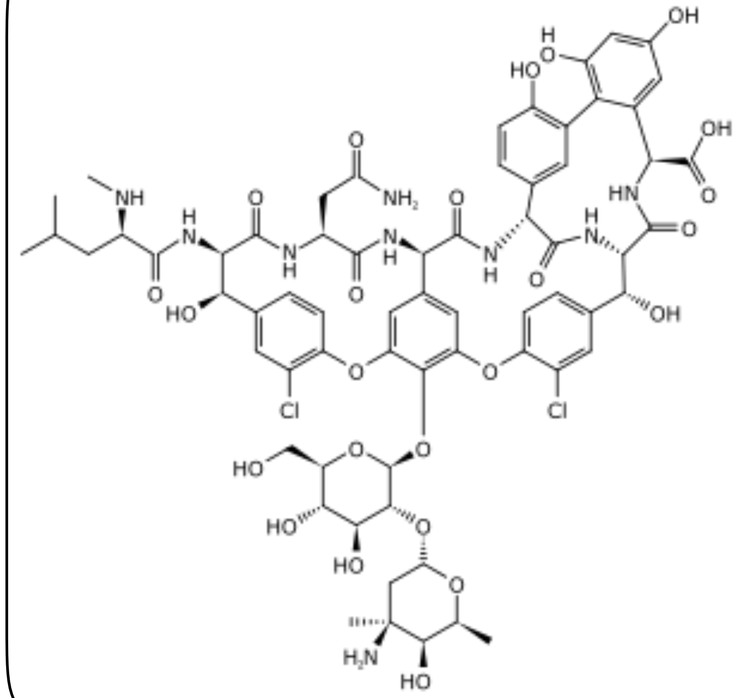



Del Salla *et al.* 2020

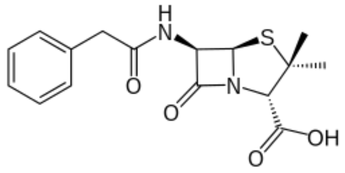
# Natural products are sources of pharmaceuticals

# Natural products are sources of pharmaceuticals

Antibiotics



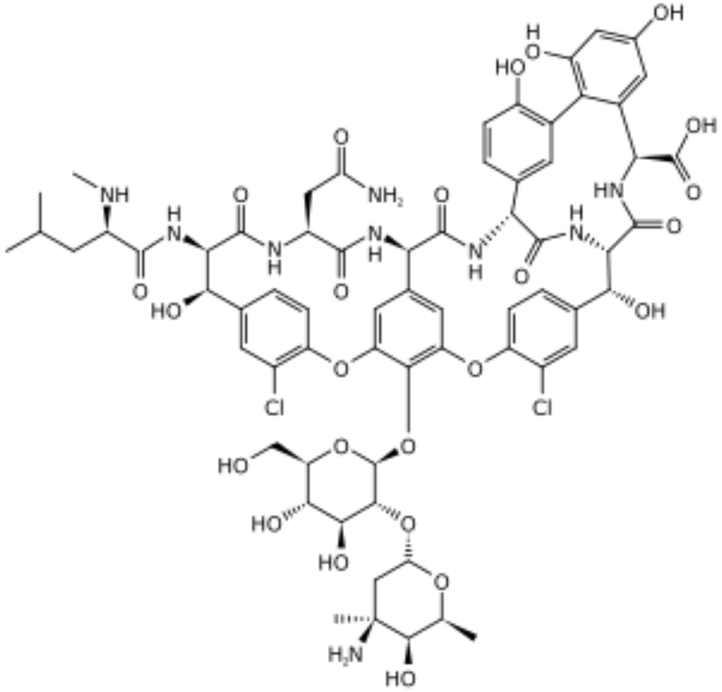
Vancomycin



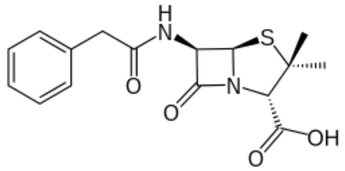
Penicillin

# Natural products are sources of pharmaceuticals

Antibiotics

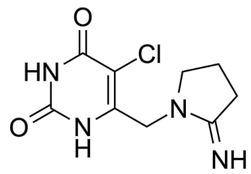


Vancomycin

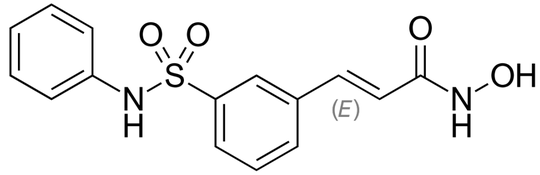


Penicillin

Anti-Cancer



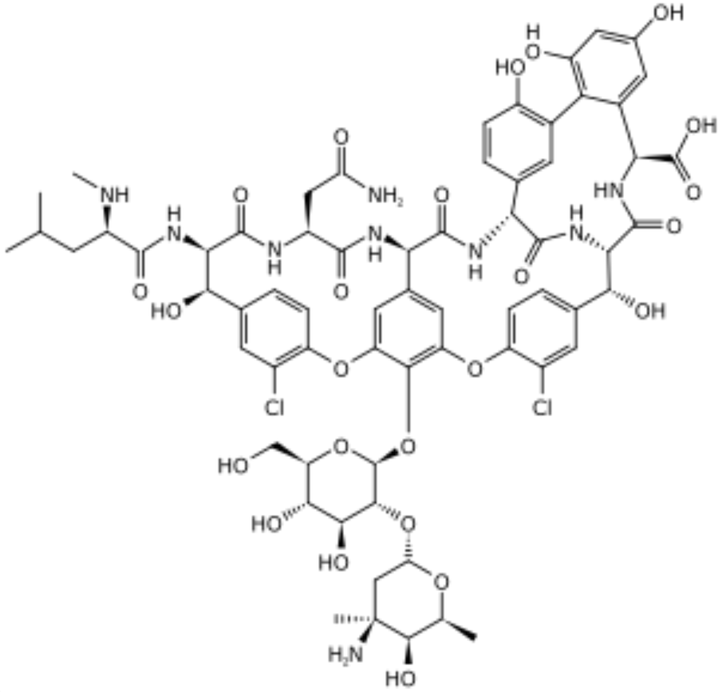
Tipiracil



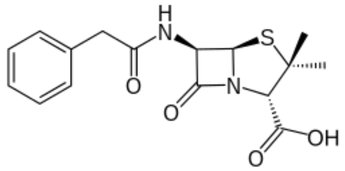
Belinostat

# Natural products are sources of pharmaceuticals

Antibiotics

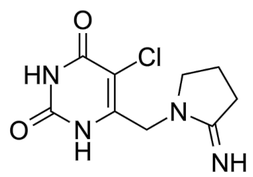


Vancomycin

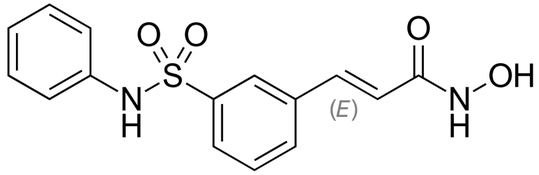


Penicillin

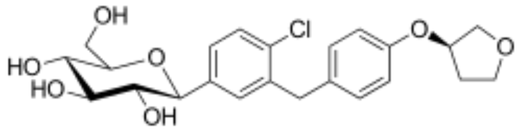
Anti-Cancer



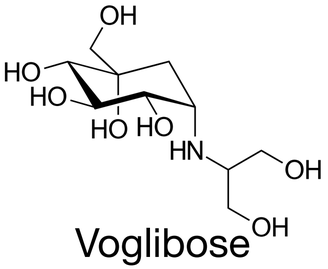
Tipiracil



Belinostat

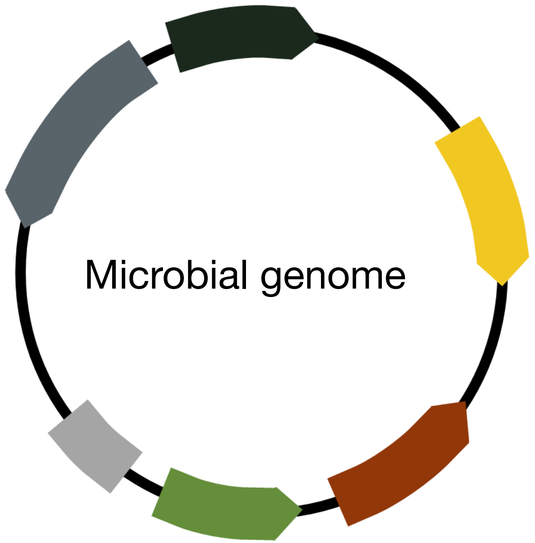


Darunavir



Voglibose

# BGCs are difficult to discover and characterize



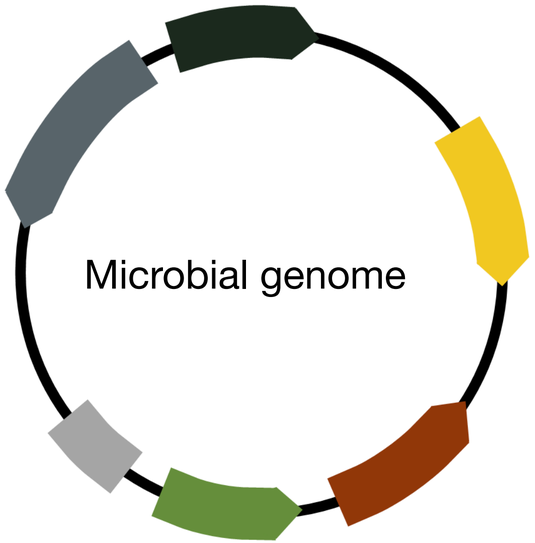

**Questions**

Which genes?
Where are the boundaries?
What does it make?

**Challenges**

Unculturable
Map product to genes?

# BGCs are difficult to discover and characterize




**Questions**

Which genes?
Where are the boundaries?
What does it make?

**Challenges**

Unculturable
Map product to genes?

Only ~1400 “BGCs of known function” in MIBiG

# Many methods pretend proteins are language

MFTGNDAGH

# Many methods pretend proteins are language

MFTGNDAGH

Treat amino acids like tokens

# Many methods pretend proteins are language

MFTGNDAGH

Treat amino acids like tokens

Use tools originally developed for language

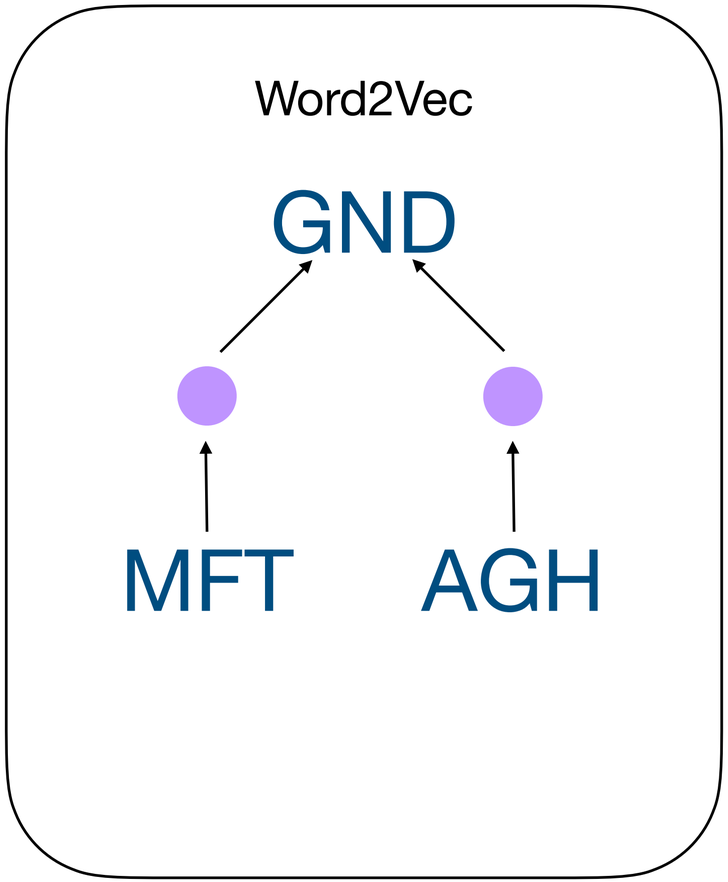
# Many methods pretend proteins are language

Word2Vec

MFT AGH

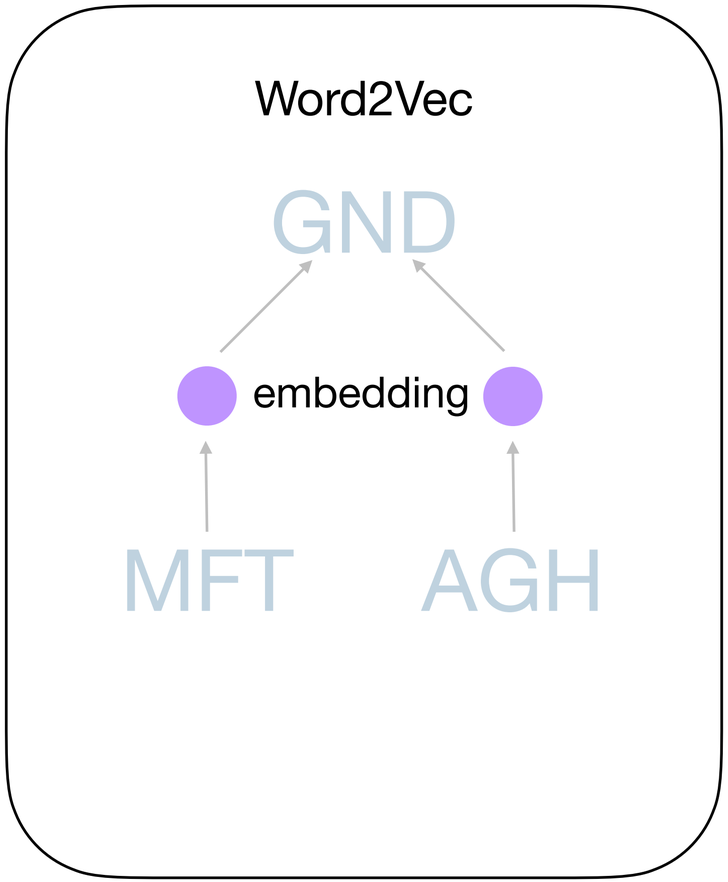
MFTGNDAGH

# Many methods pretend proteins are language



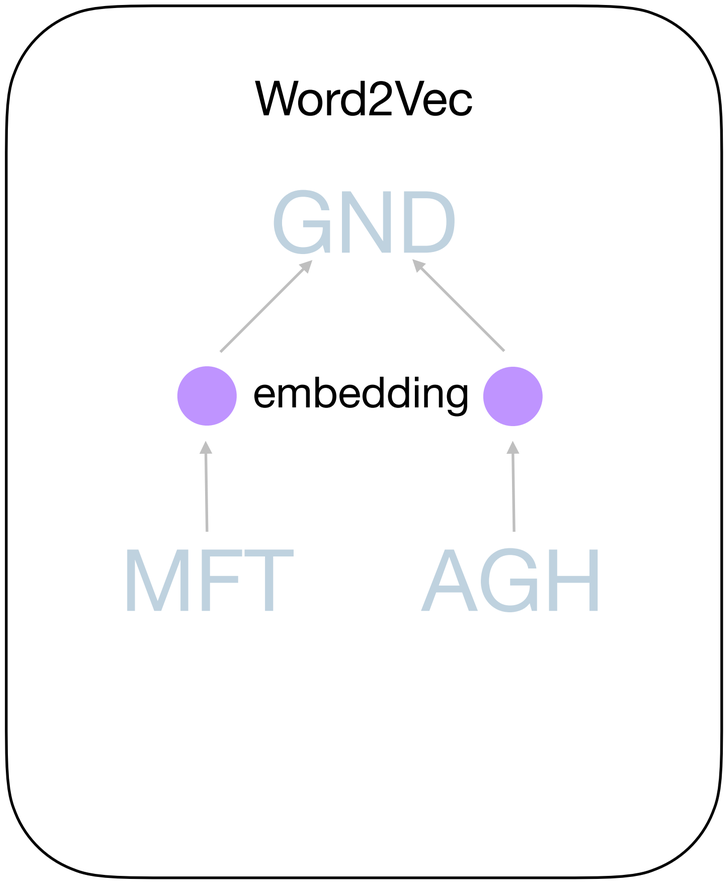

MFTGNDAGH

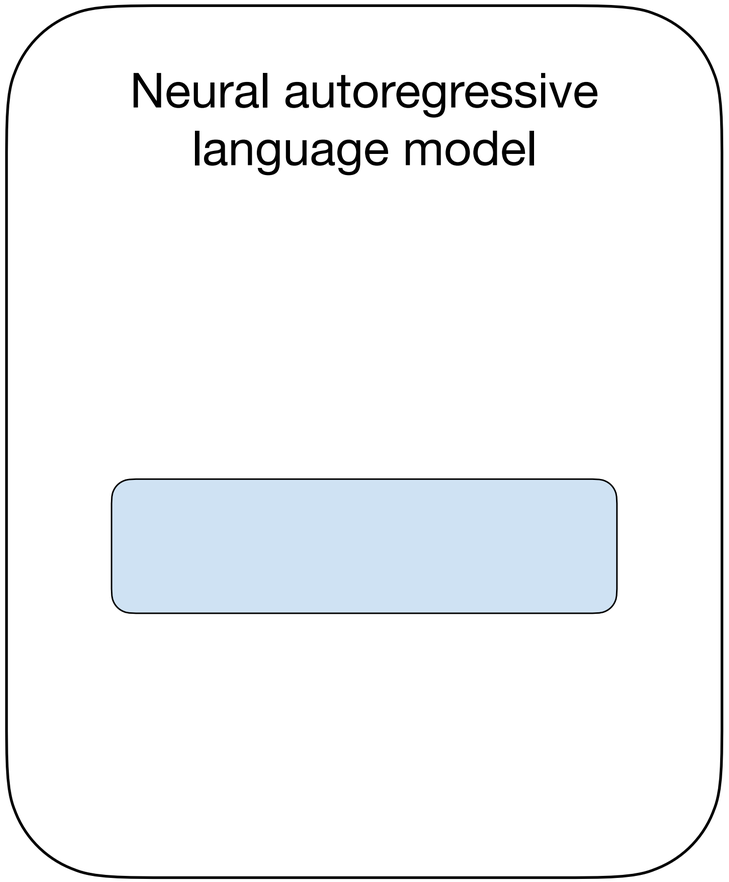
# Many methods pretend proteins are language




MFTGNDAGH

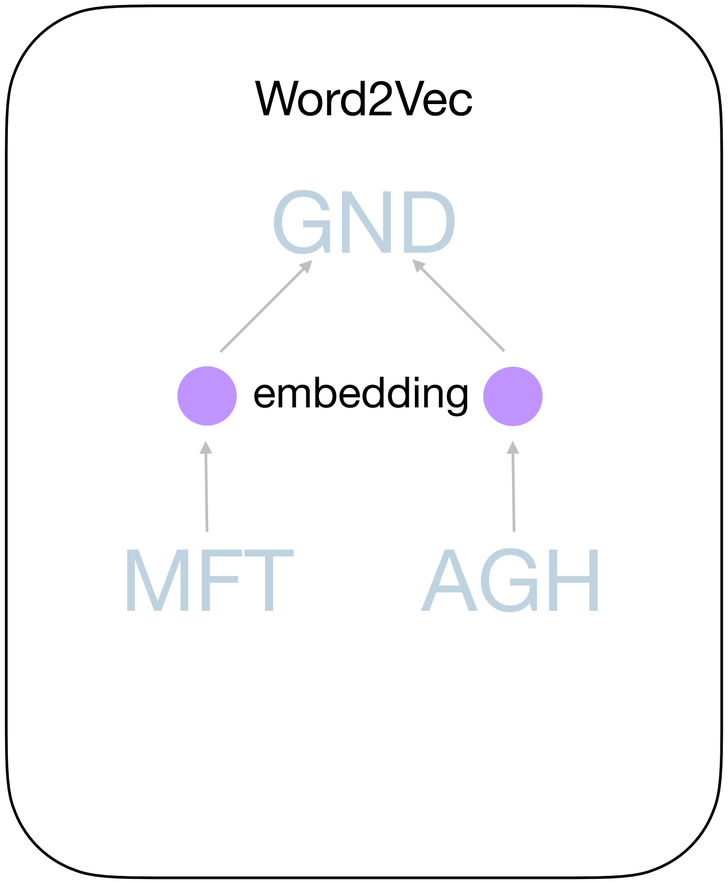
# Many methods pretend proteins are language

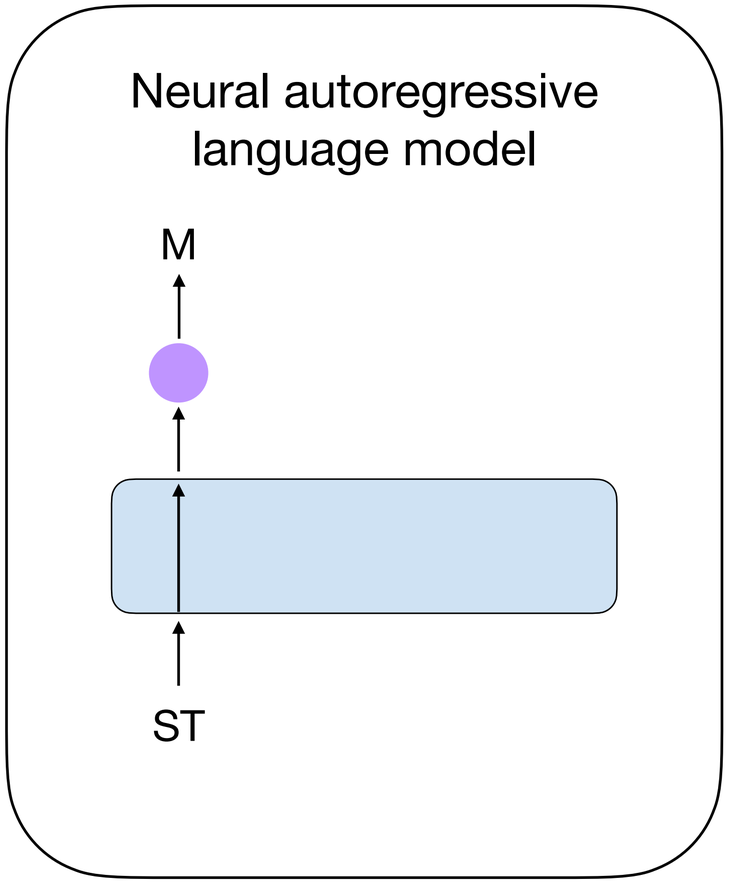



MFTGNDAGH

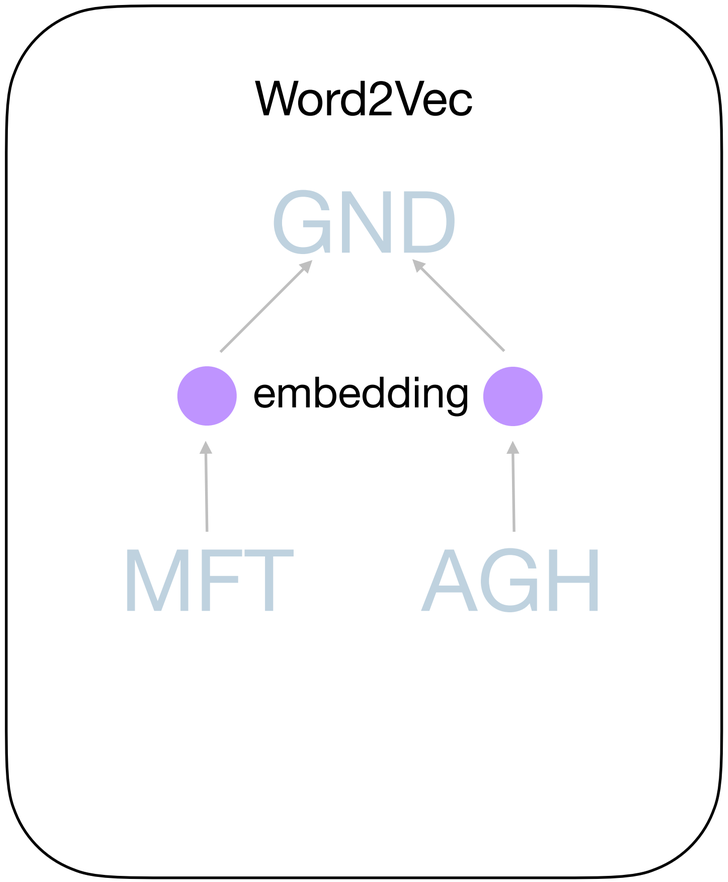
# Many methods pretend proteins are language

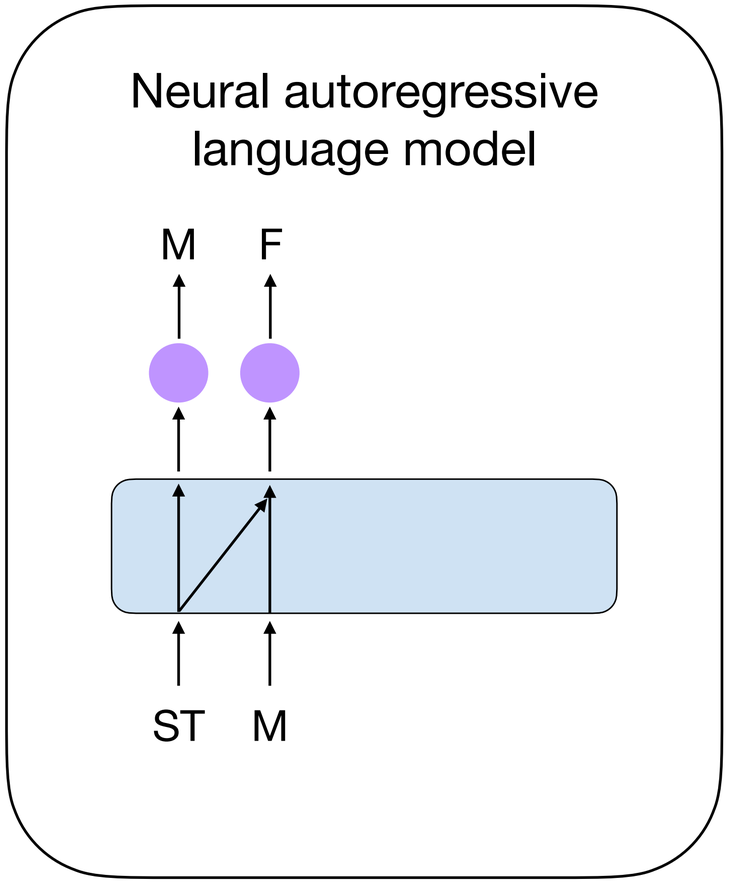



MFTGNDAGH

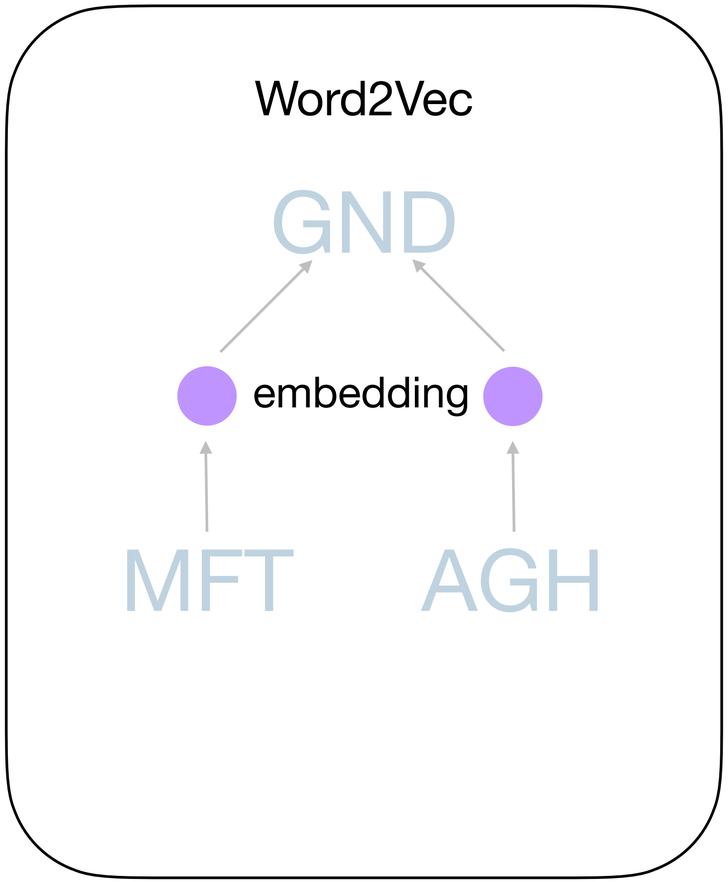
# Many methods pretend proteins are language

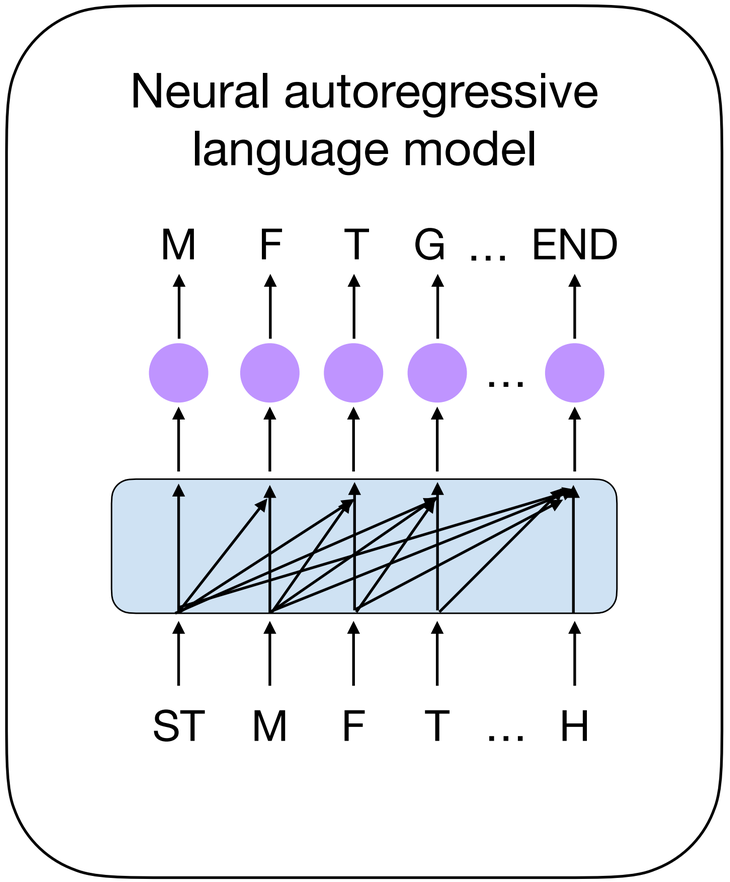



MFTGNDAGH

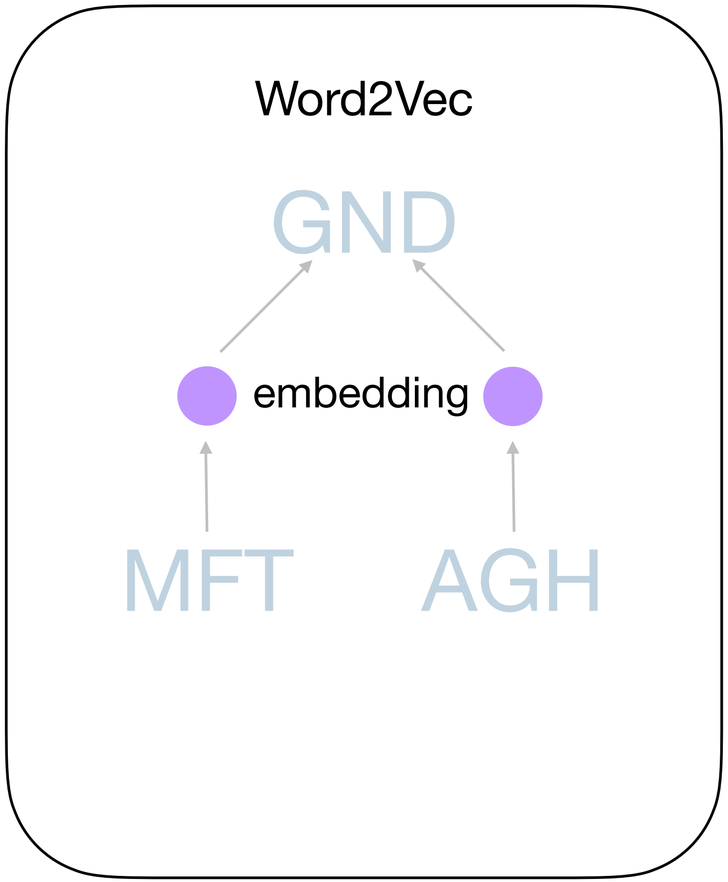
# Many methods pretend proteins are language

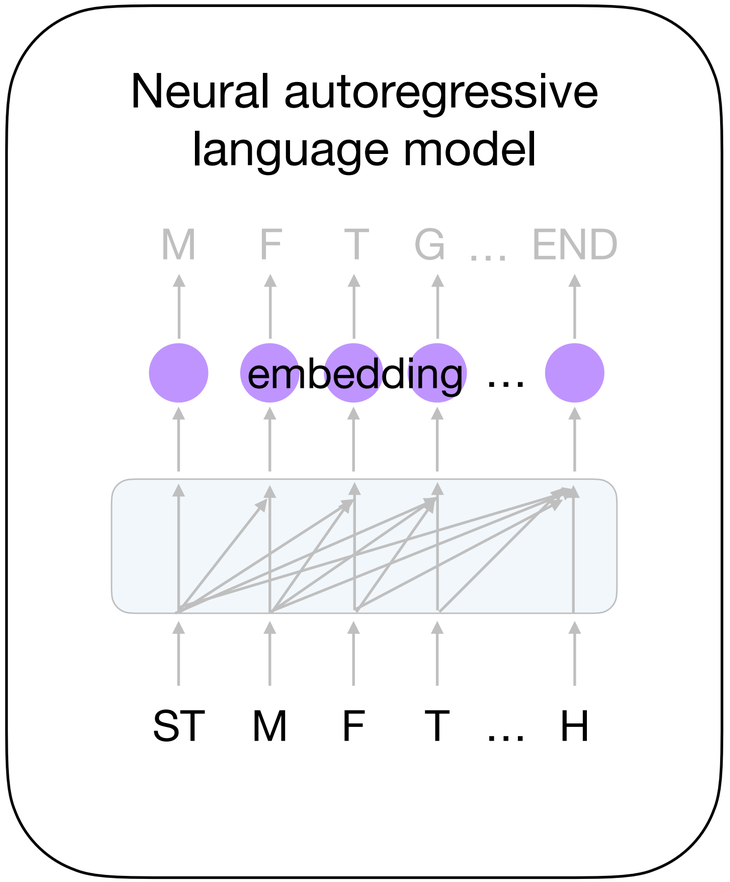



MFTGNDAGH

# Many methods pretend proteins are language

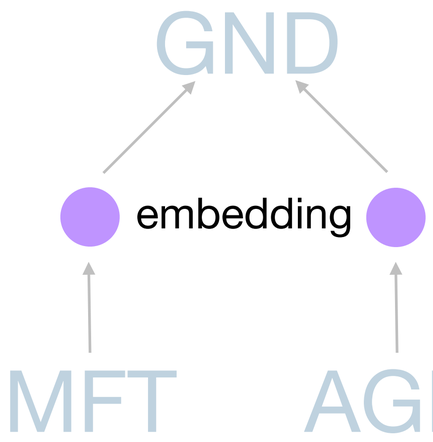


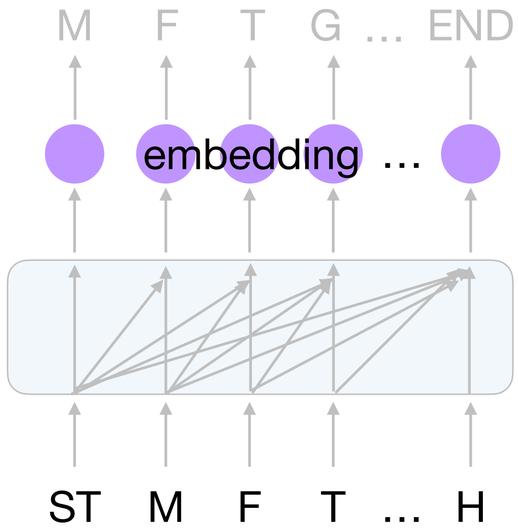

MFTGNDAGH

# Many methods pretend proteins are language

Word2Vec




Neural autoregressive language model

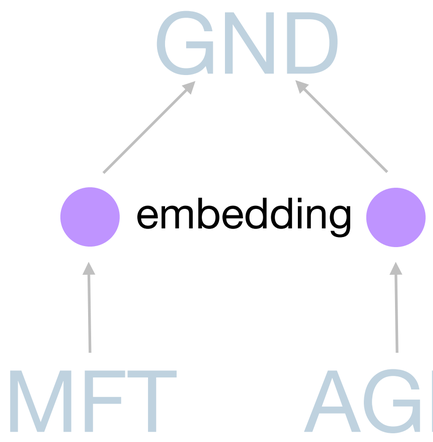



Neural masked language model

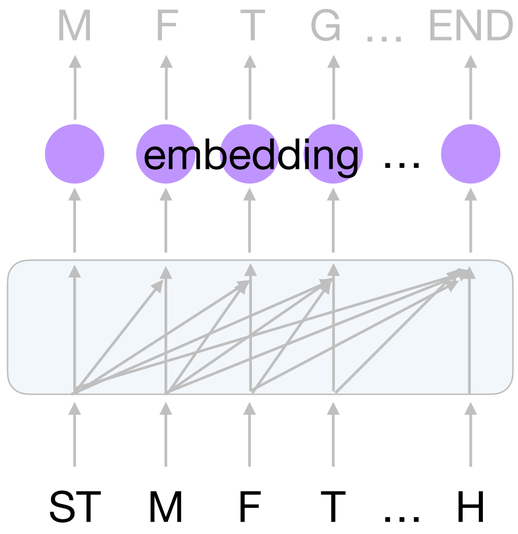
MFTGNDAGH

# Many methods pretend proteins are language

Word2Vec




Neural autoregressive language model



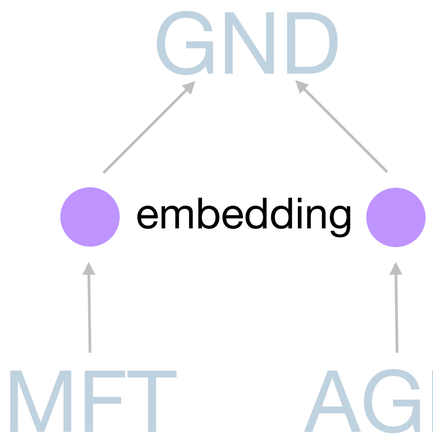

Neural masked language model

M F T G … H

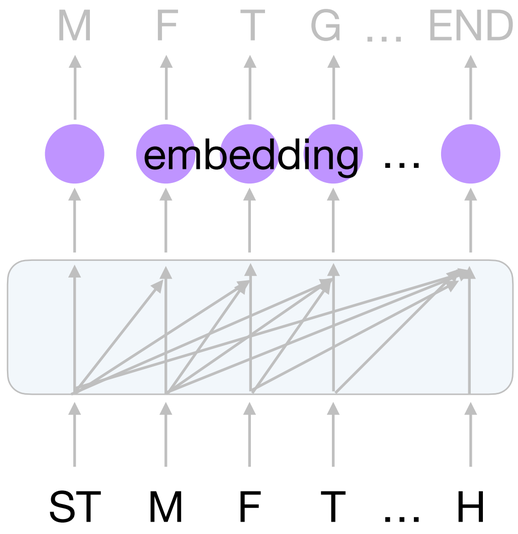
MFTGNDAGH

# Many methods pretend proteins are language

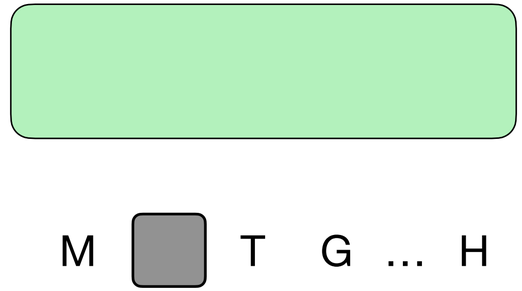
Word2Vec




Neural autoregressive language model



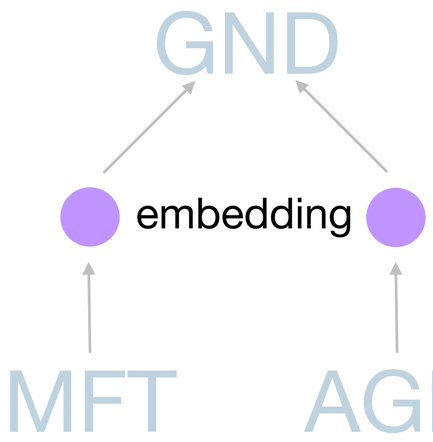

Neural masked language model



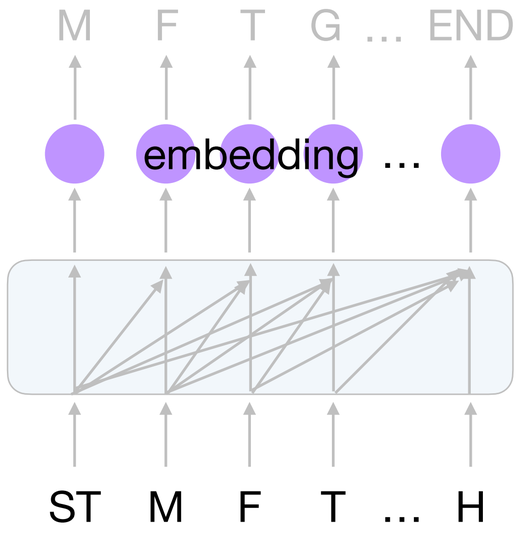

MFTGNDAGH

# Many methods pretend proteins are language

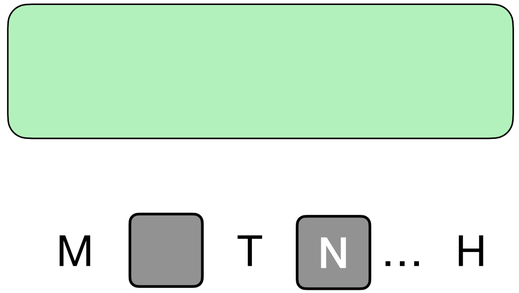
Word2Vec




Neural autoregressive language model



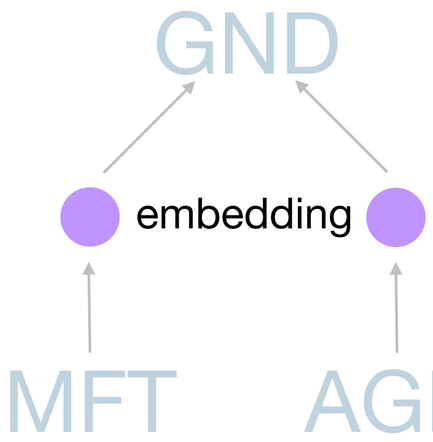

Neural masked language model



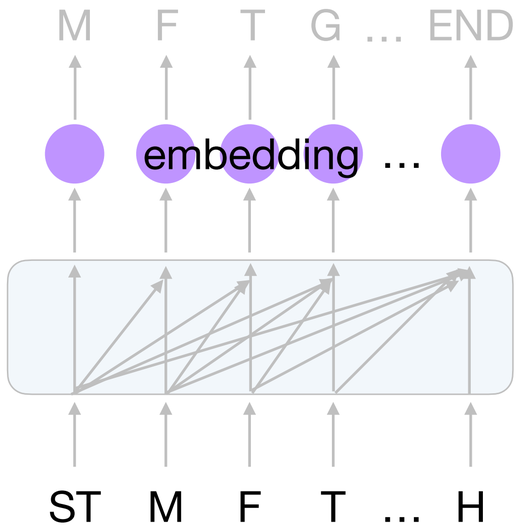

MFTGNDAGH

# Many methods pretend proteins are language

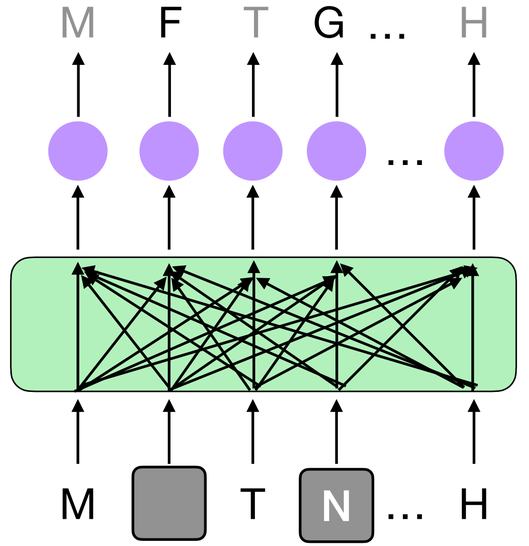
Word2Vec




Neural autoregressive language model



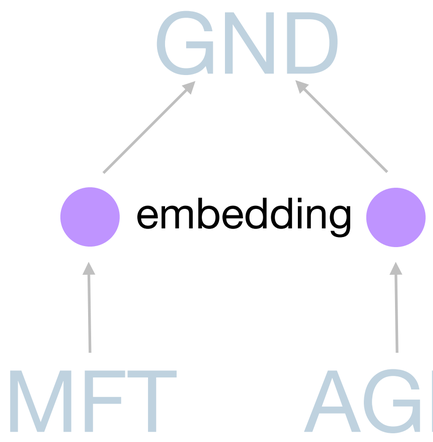

Neural masked language model



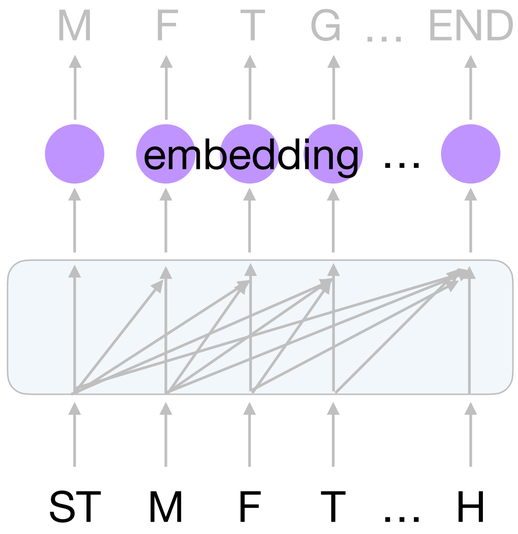

MFTGNDAGH

# Many methods pretend proteins are language

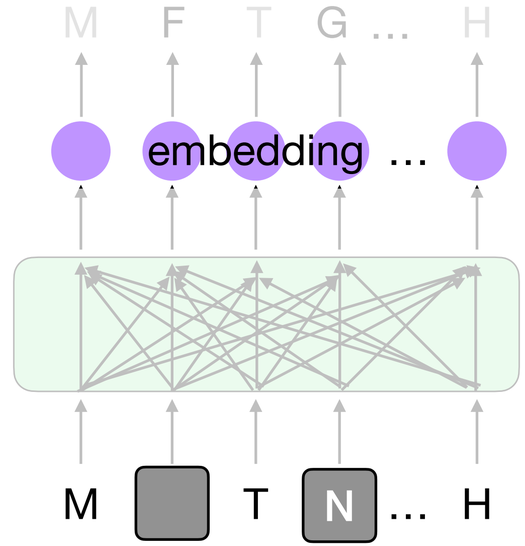
Word2Vec




Neural autoregressive language model




Neural masked language model




MFTGNDAGH

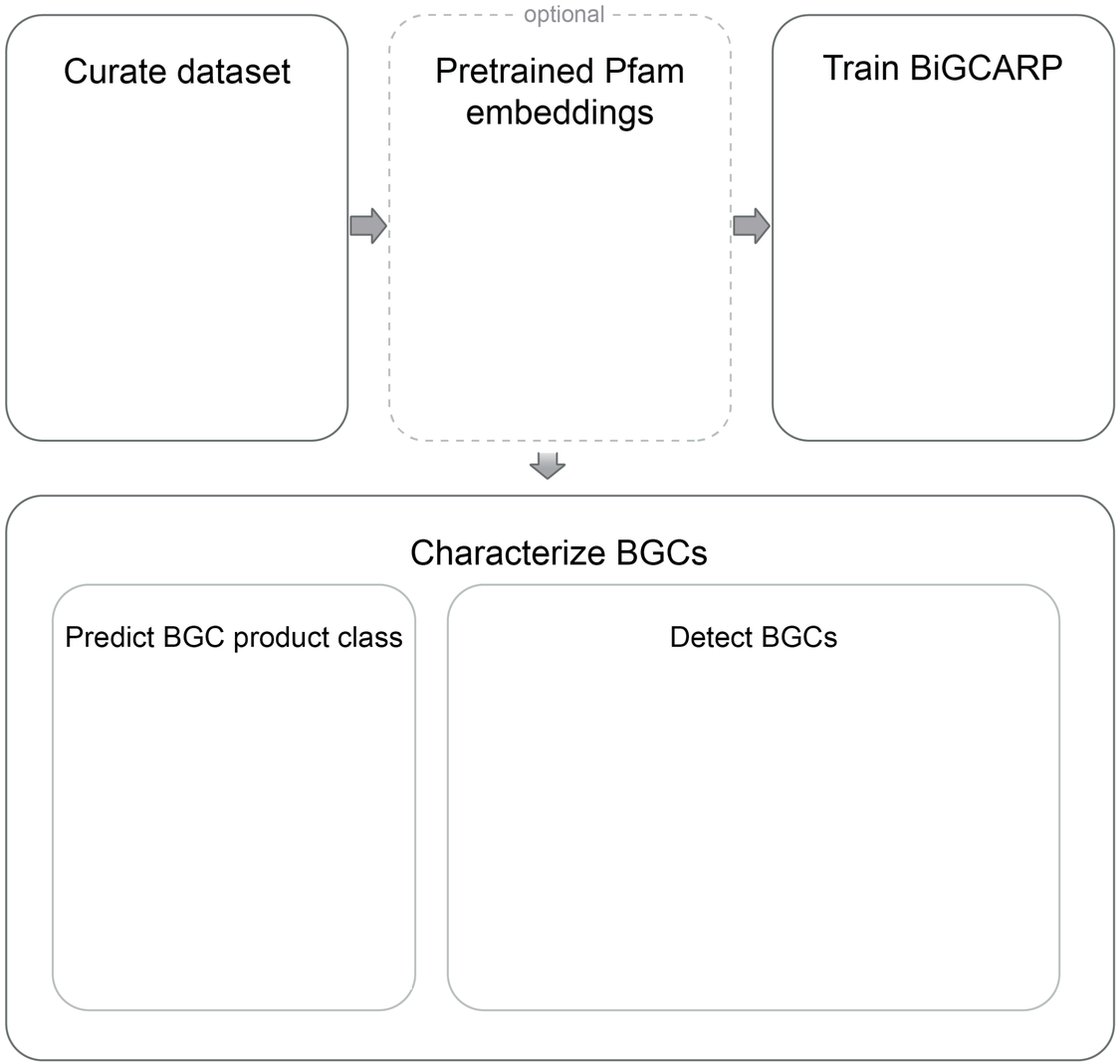
# BiGCARP is a self-supervised BGC model

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins

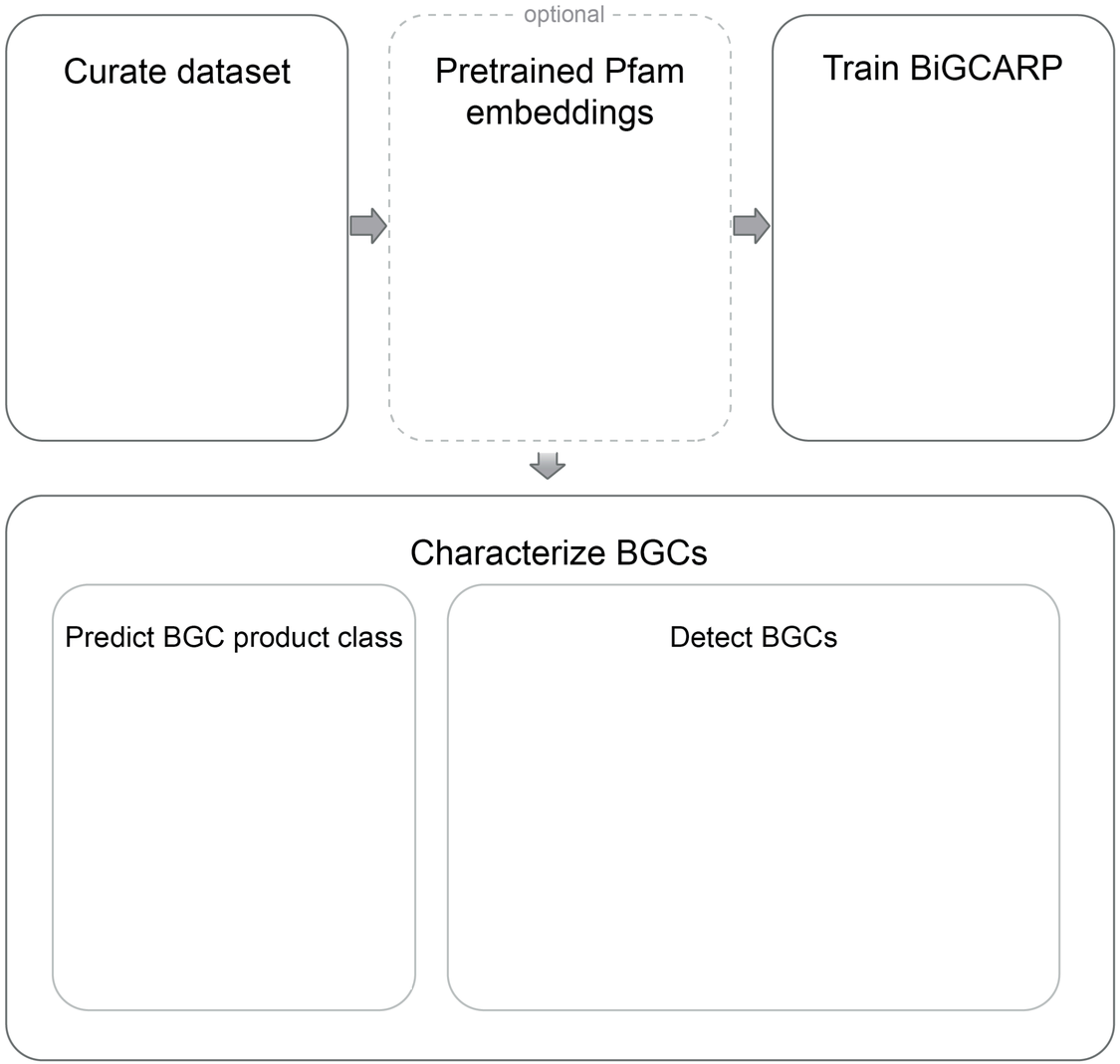
# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins



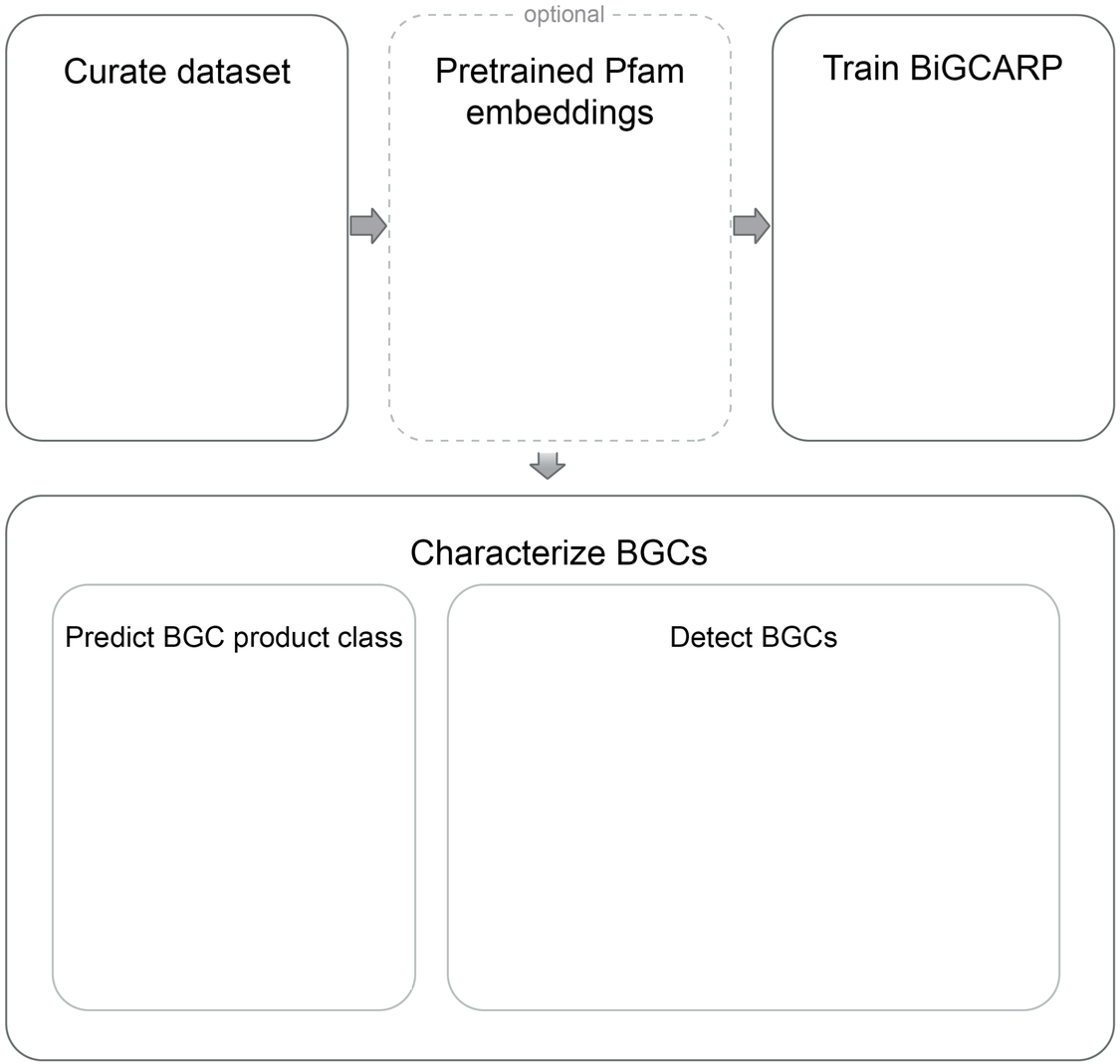
# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins




# BiGCARP is a self-supervised BGC model

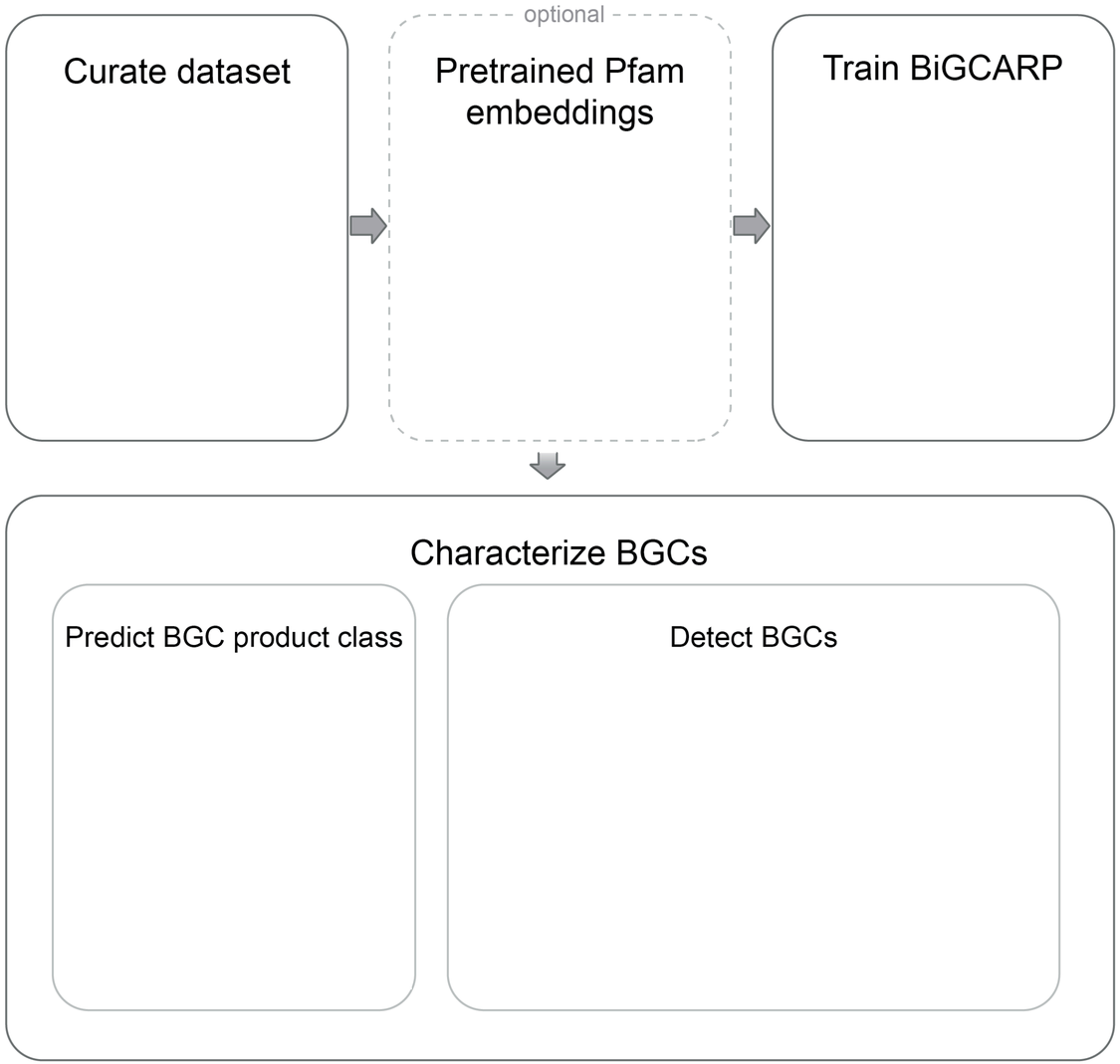
**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins




Previous data-mining efforts find positive examples of BGCs

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins

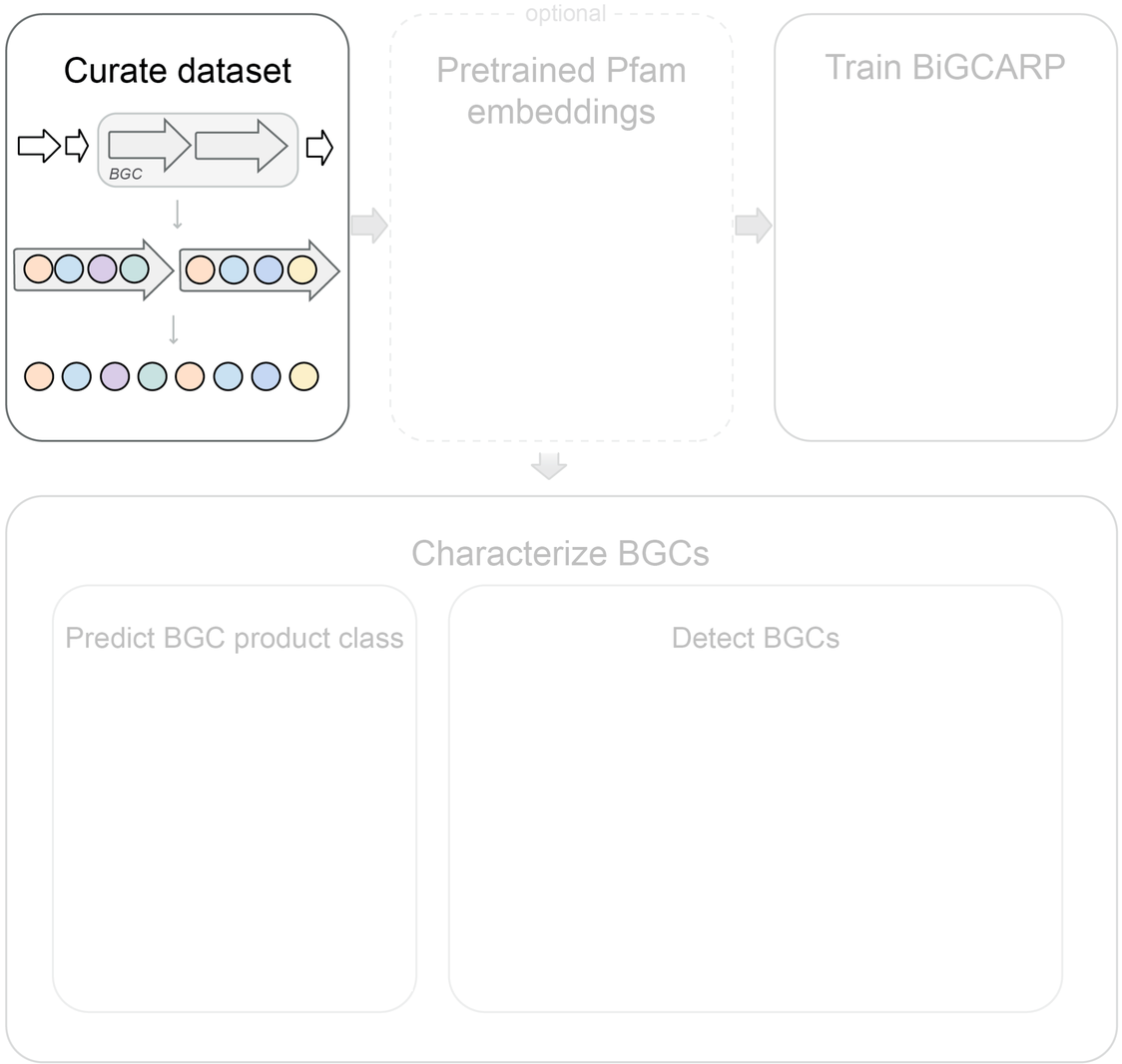



Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins

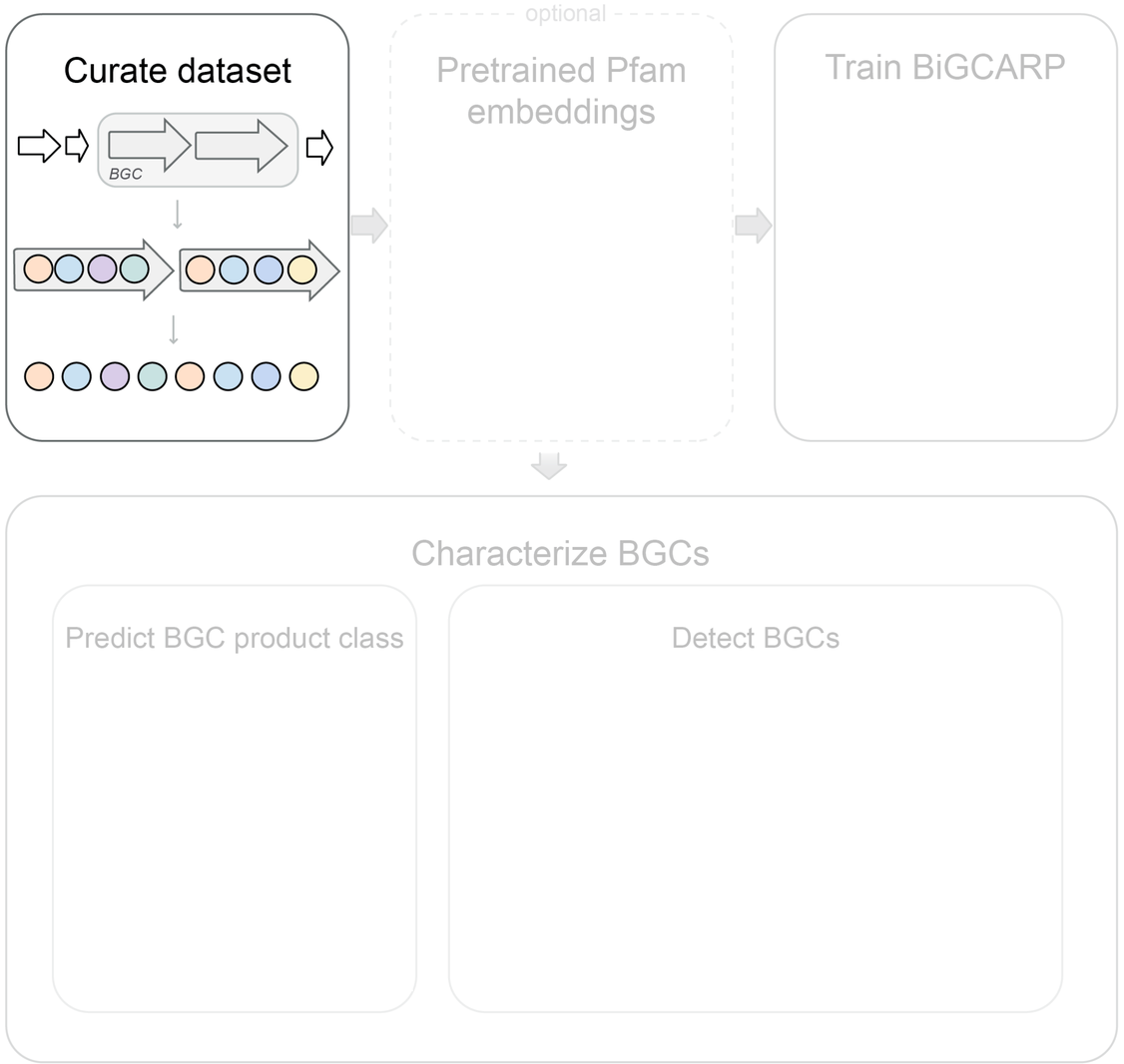



Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins



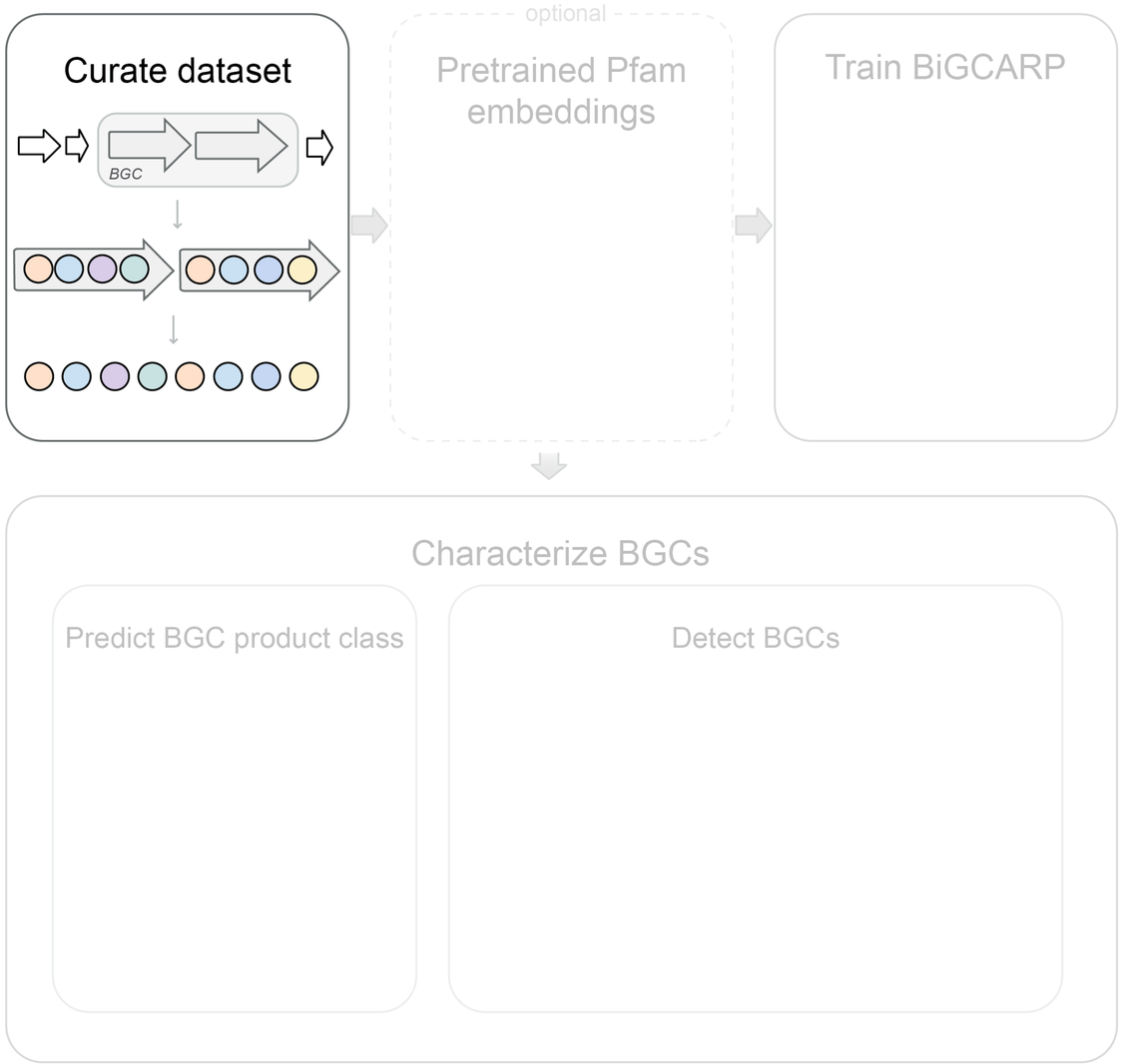

Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

Microbial genome …ACTGCGGTTACG…

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins

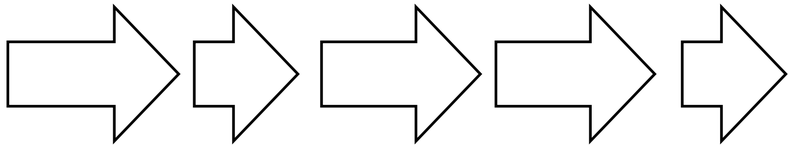



Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

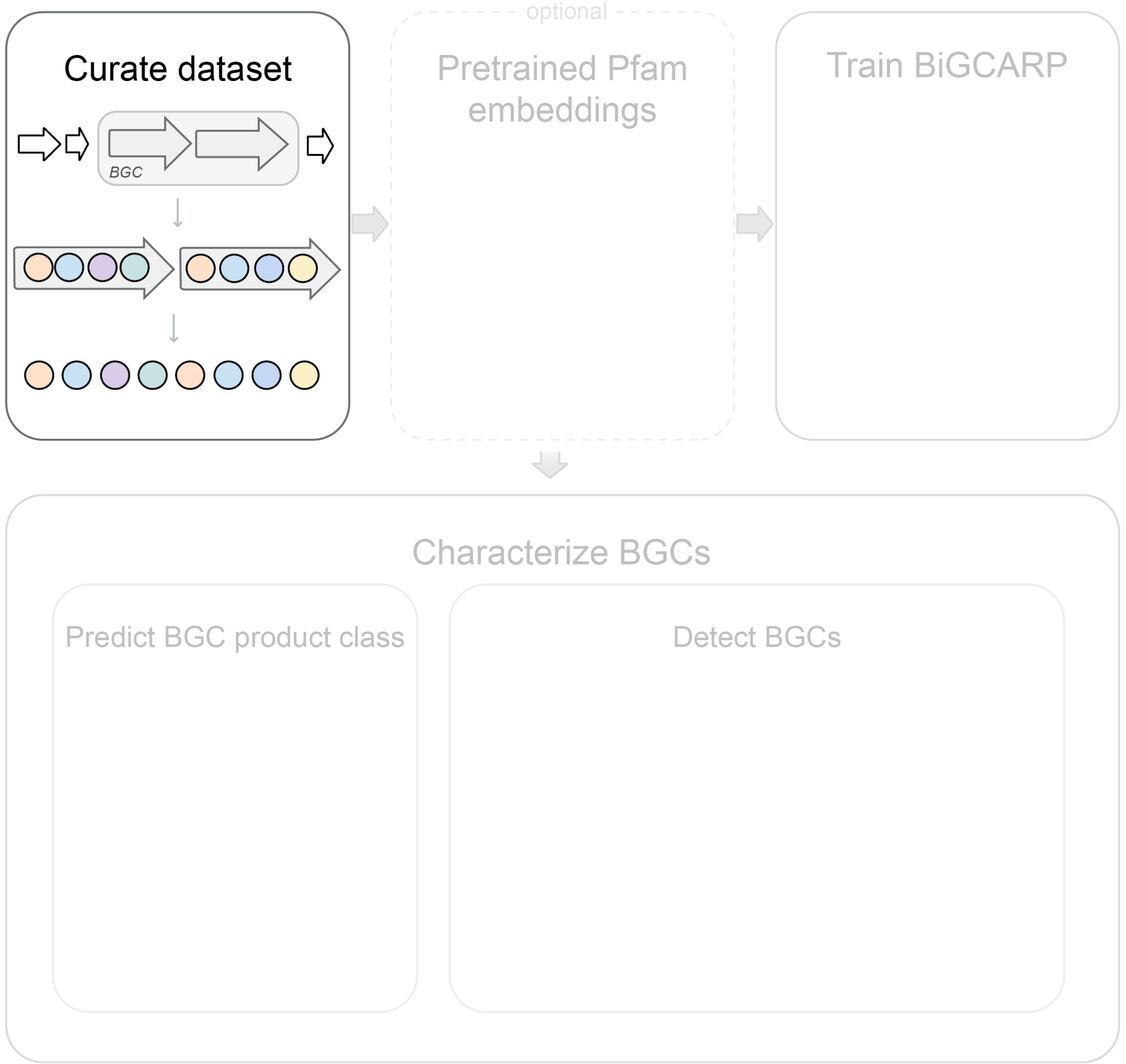
Microbial genome …ACTGCGGTTACG…

Open reading frames



# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins




Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

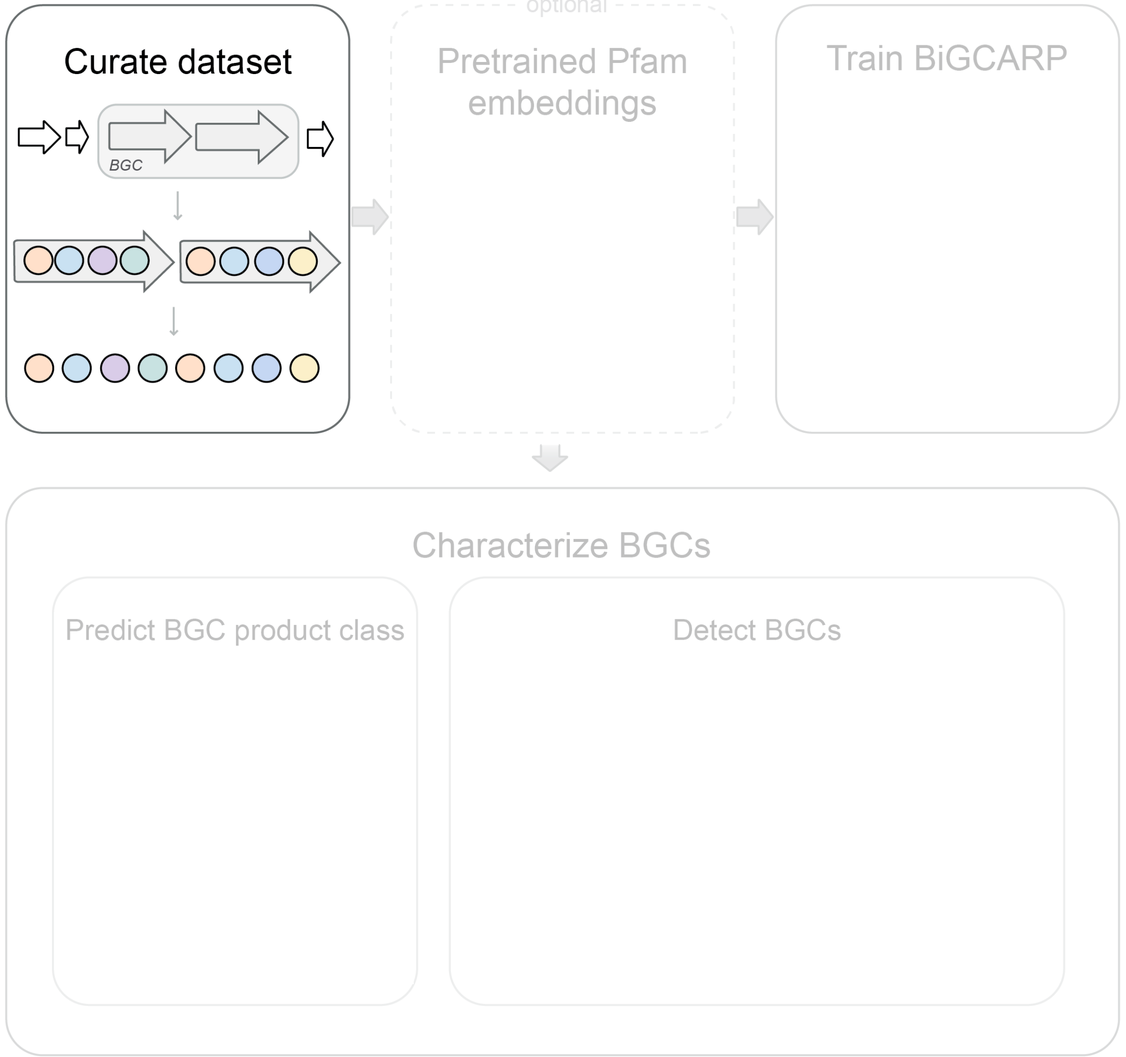
Microbial genome …ACTGCGGTTACG…

Open reading frames

Identify BGCs with antiSMASH

# BiGCARP is a self-supervised BGC model

**Bi**osynthetic **G**ene cluster **C**onvolutional **A**utoencoding **R**epresentations of **P**roteins




Previous data-mining efforts find positive examples of BGCs

True negative examples are difficult to find

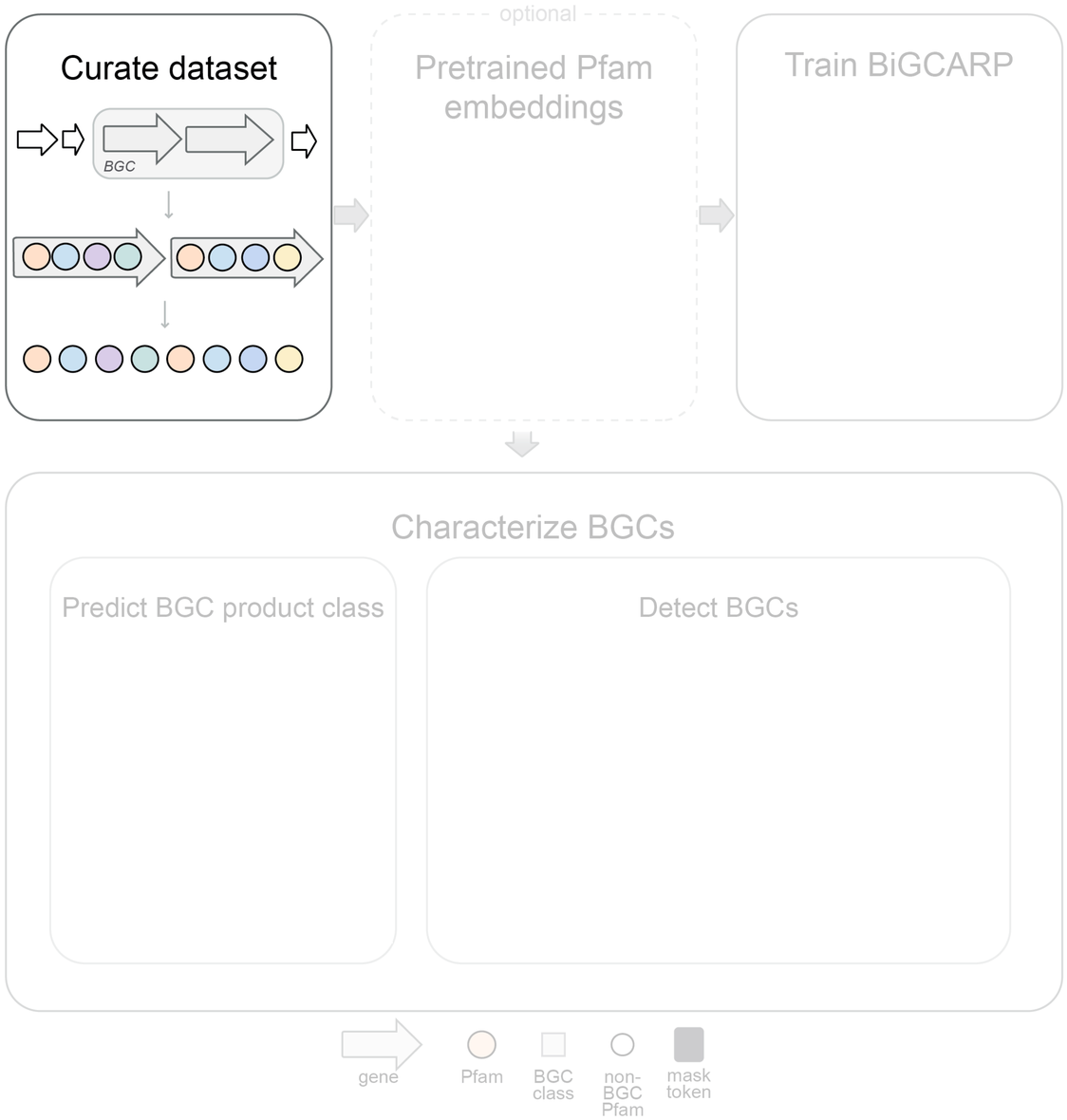
Microbial genome …ACTGCGGTTACG…

Open reading frames

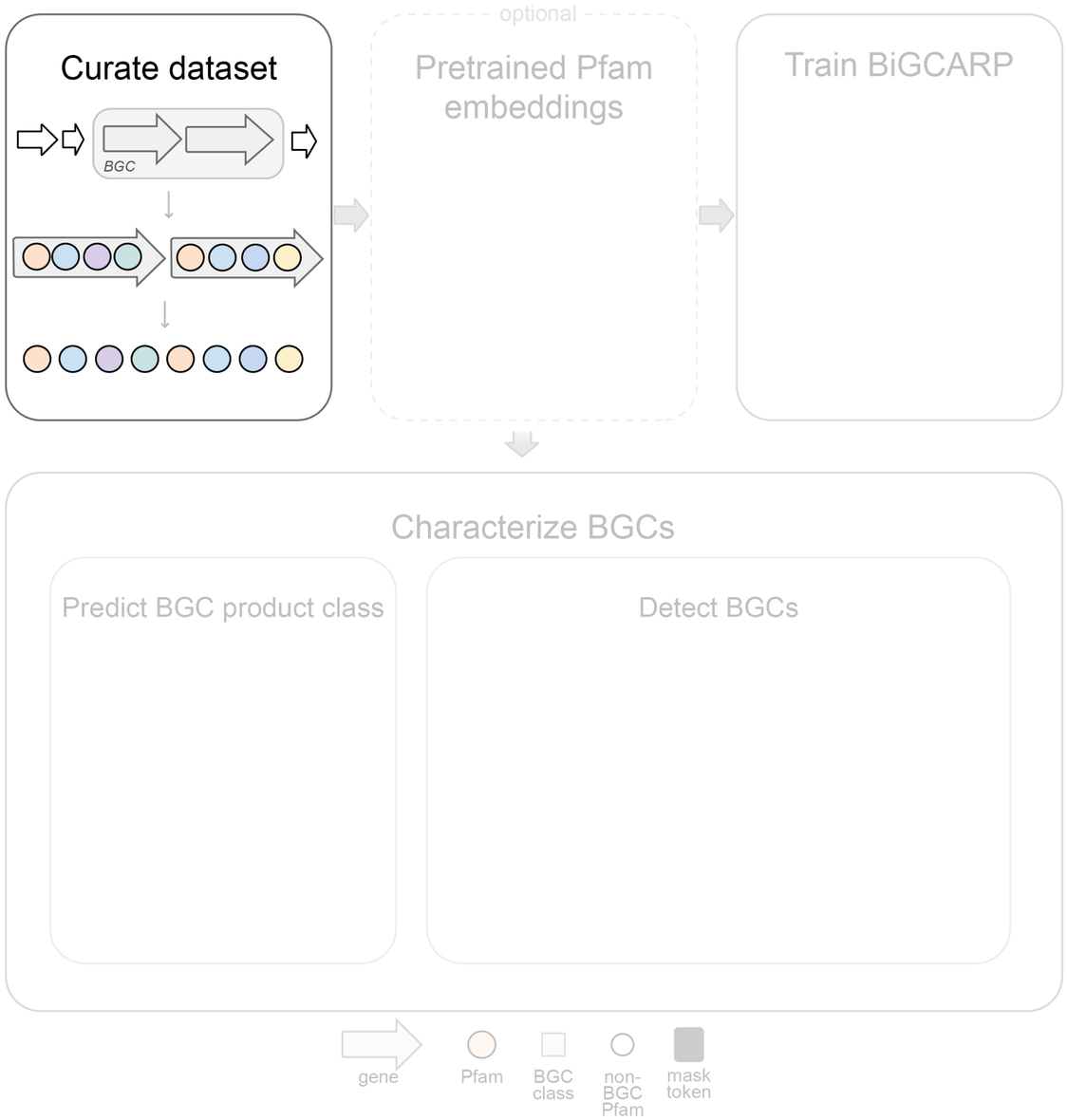
Identify BGCs with antiSMASH

127k BGCs from 55 product classes

# We represent BGCs as sequences of domains

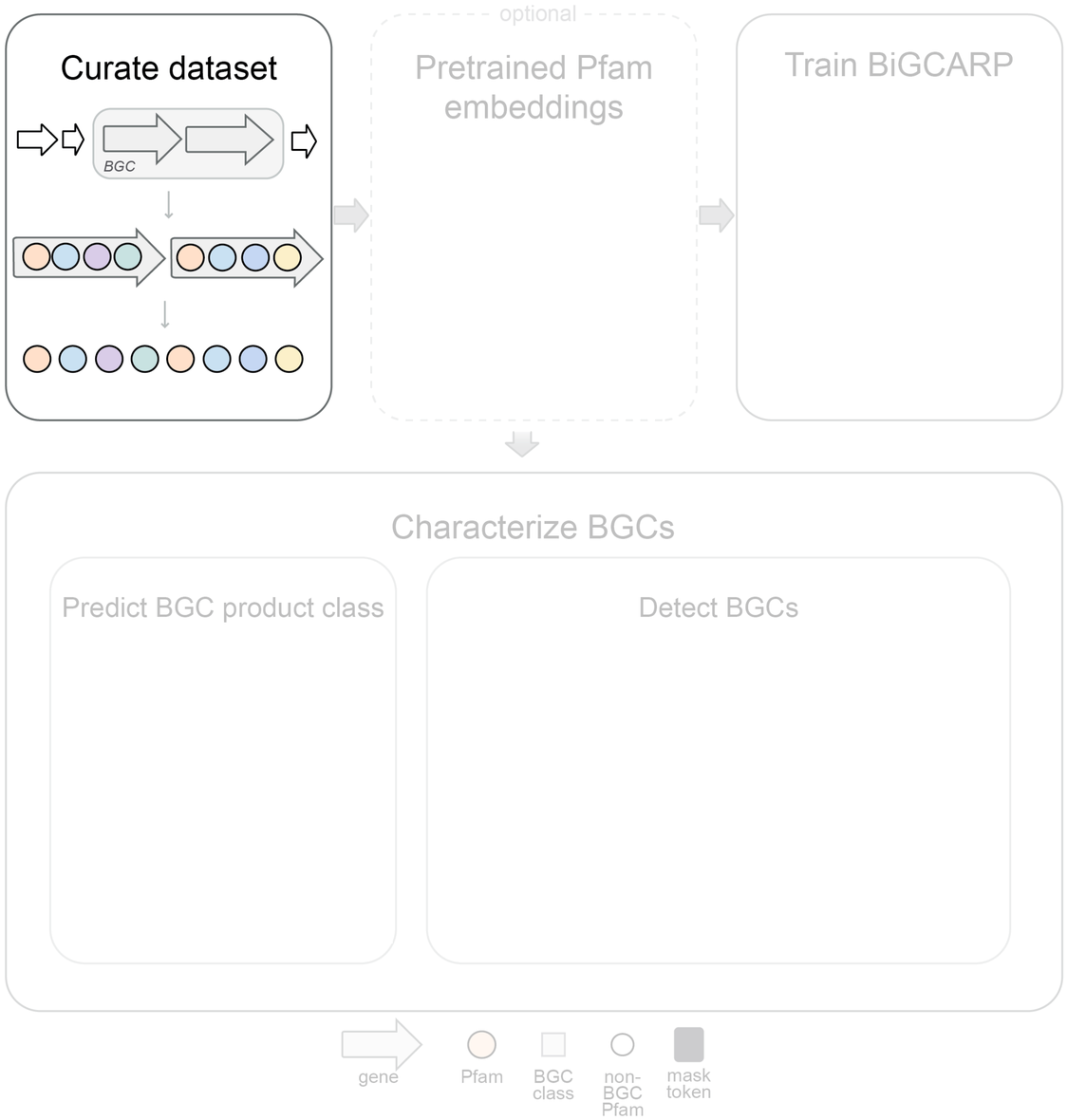


# We represent BGCs as sequences of domains



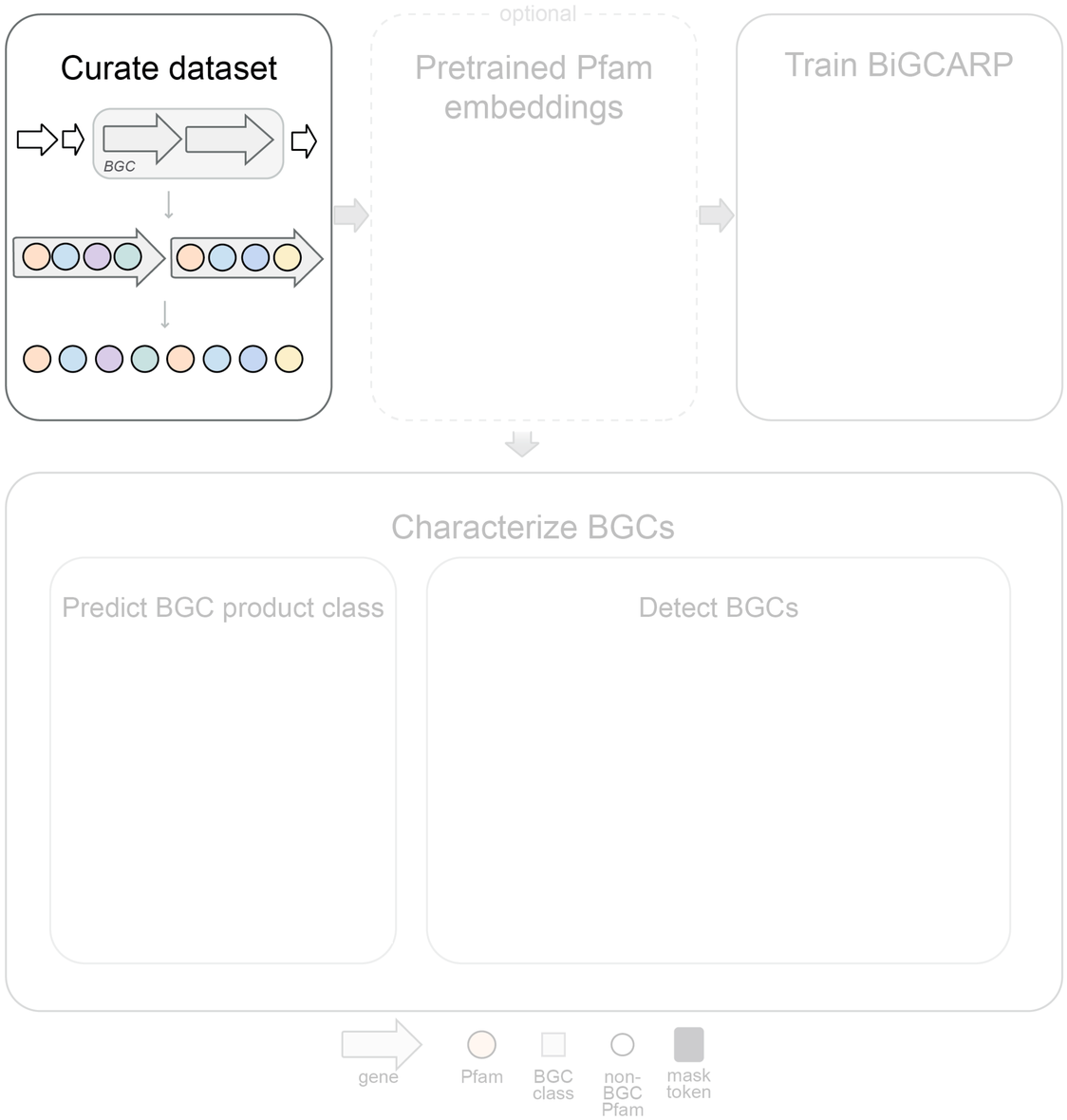

Domain: protein subunit with

# We represent BGCs as sequences of domains



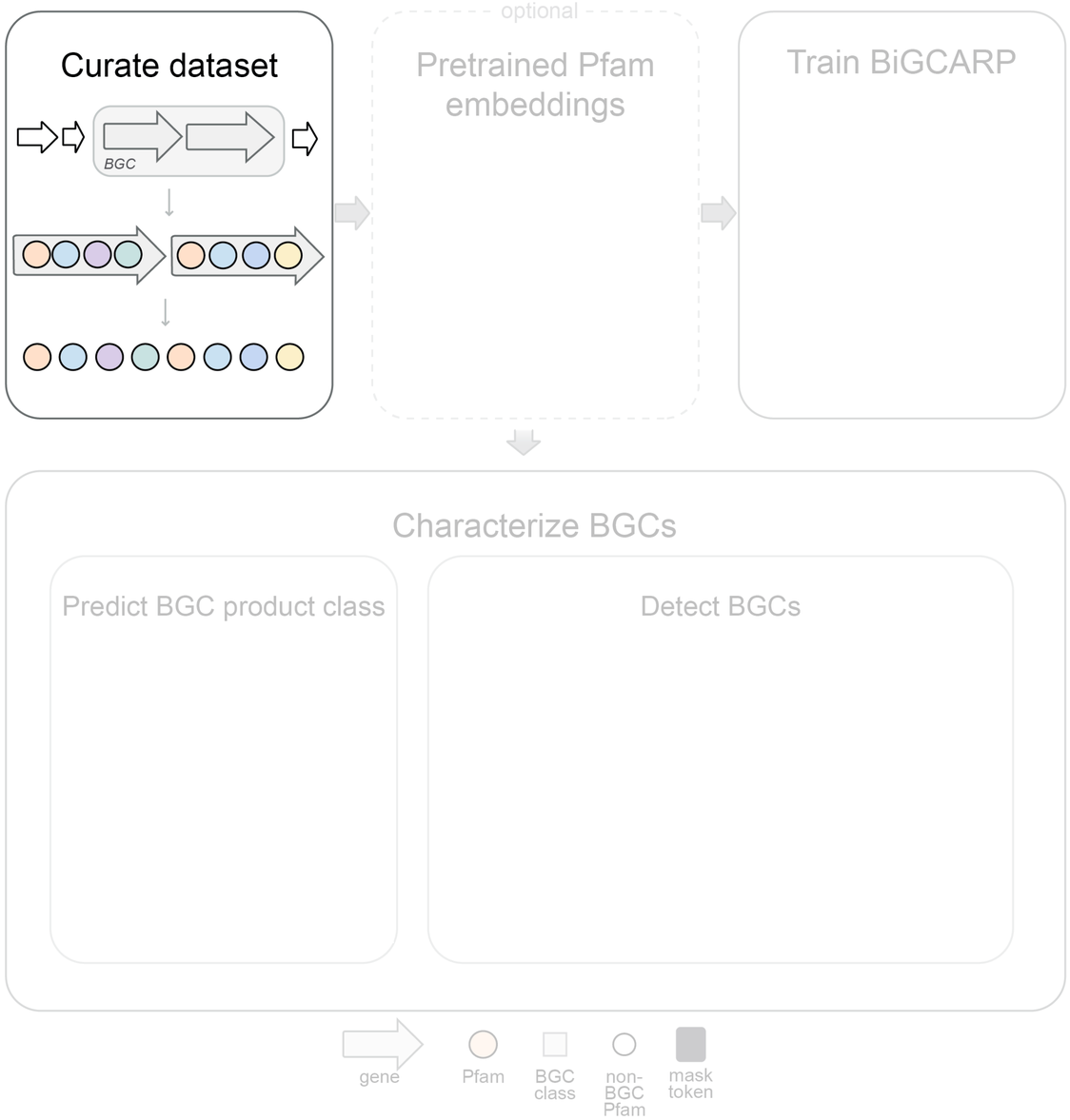

Domain: protein subunit with compact structure

# We represent BGCs as sequences of domains

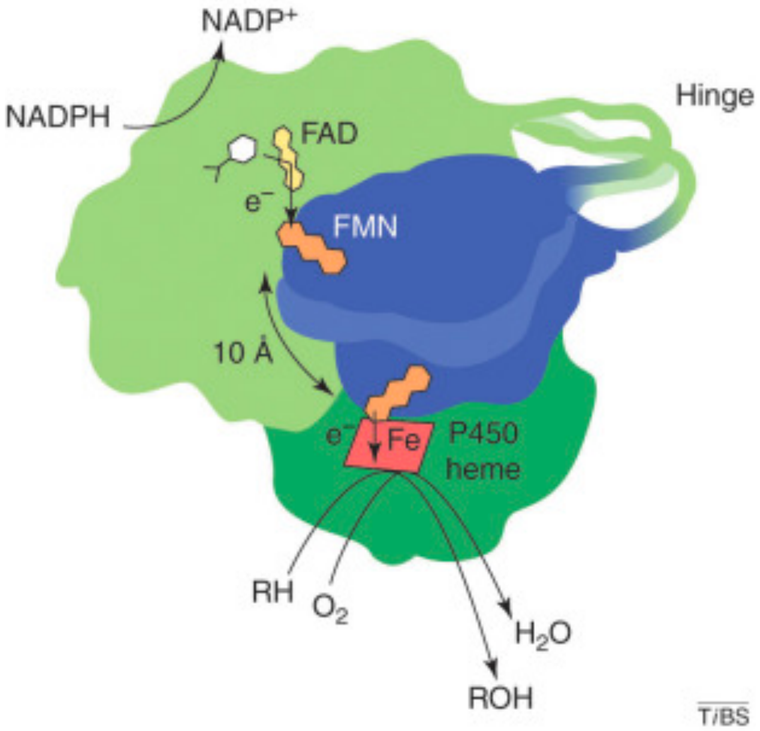



Domain: protein subunit with compact structure
modular function

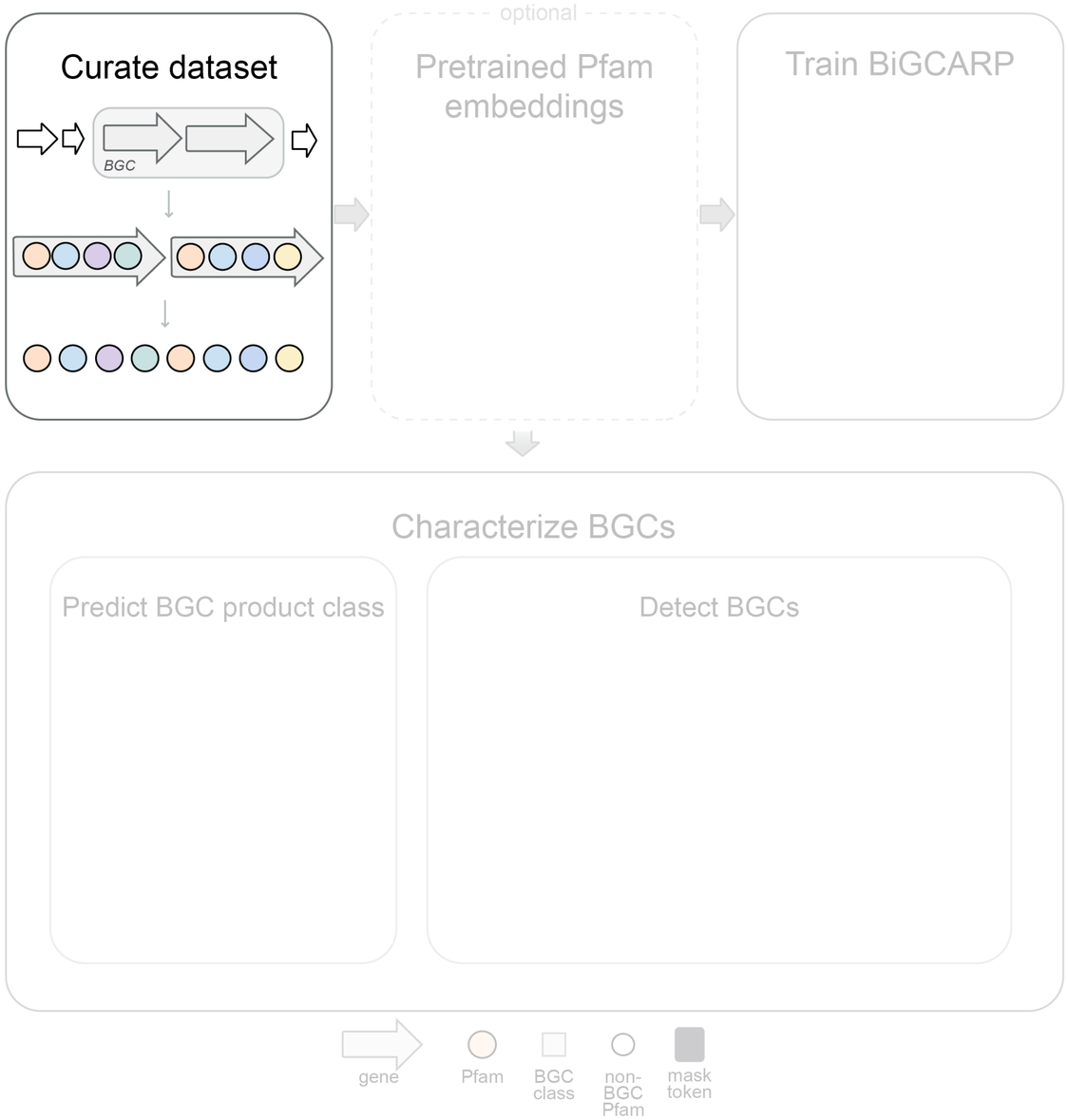
# We represent BGCs as sequences of domains




Domain: protein subunit with compact structure
modular function

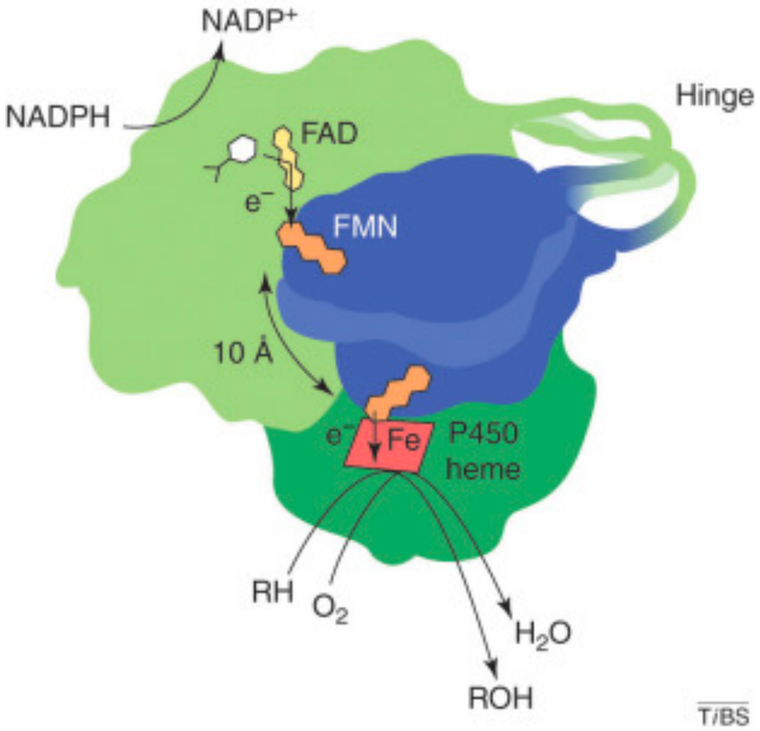


# We represent BGCs as sequences of domains

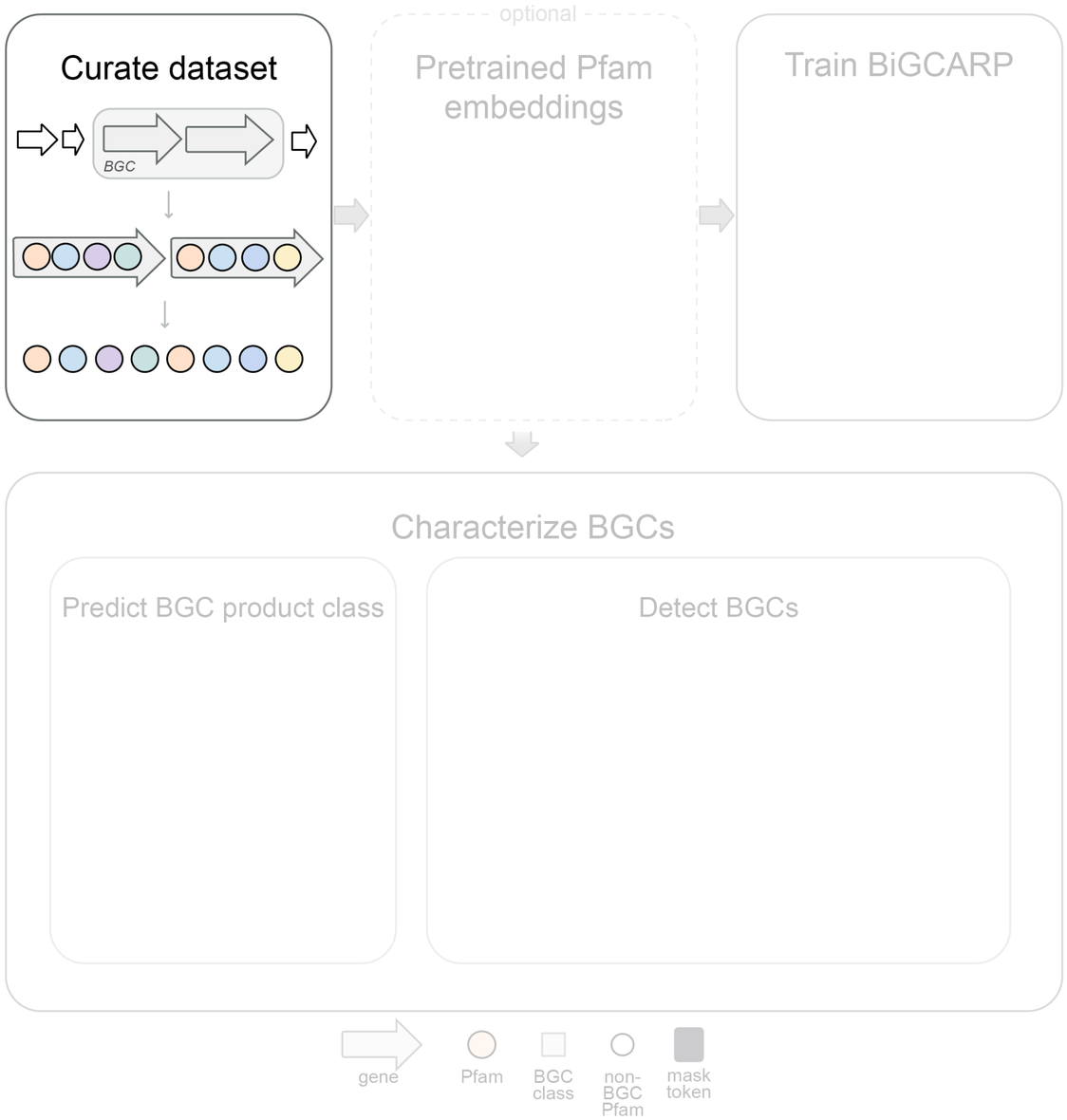



Domain: protein subunit with compact structure
modular function

Compromise between granularity,
vocab size, and sequence length



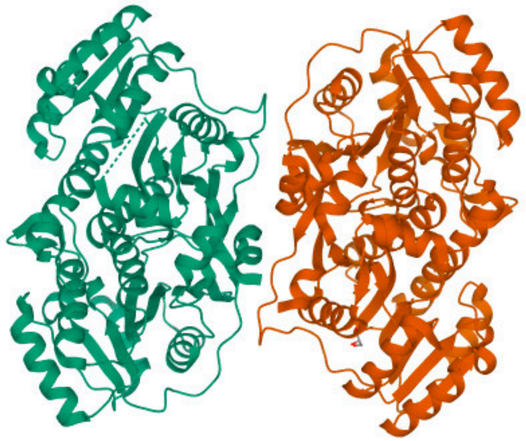
# We represent BGCs as sequences of domains




Domain: protein subunit with compact structure
modular function

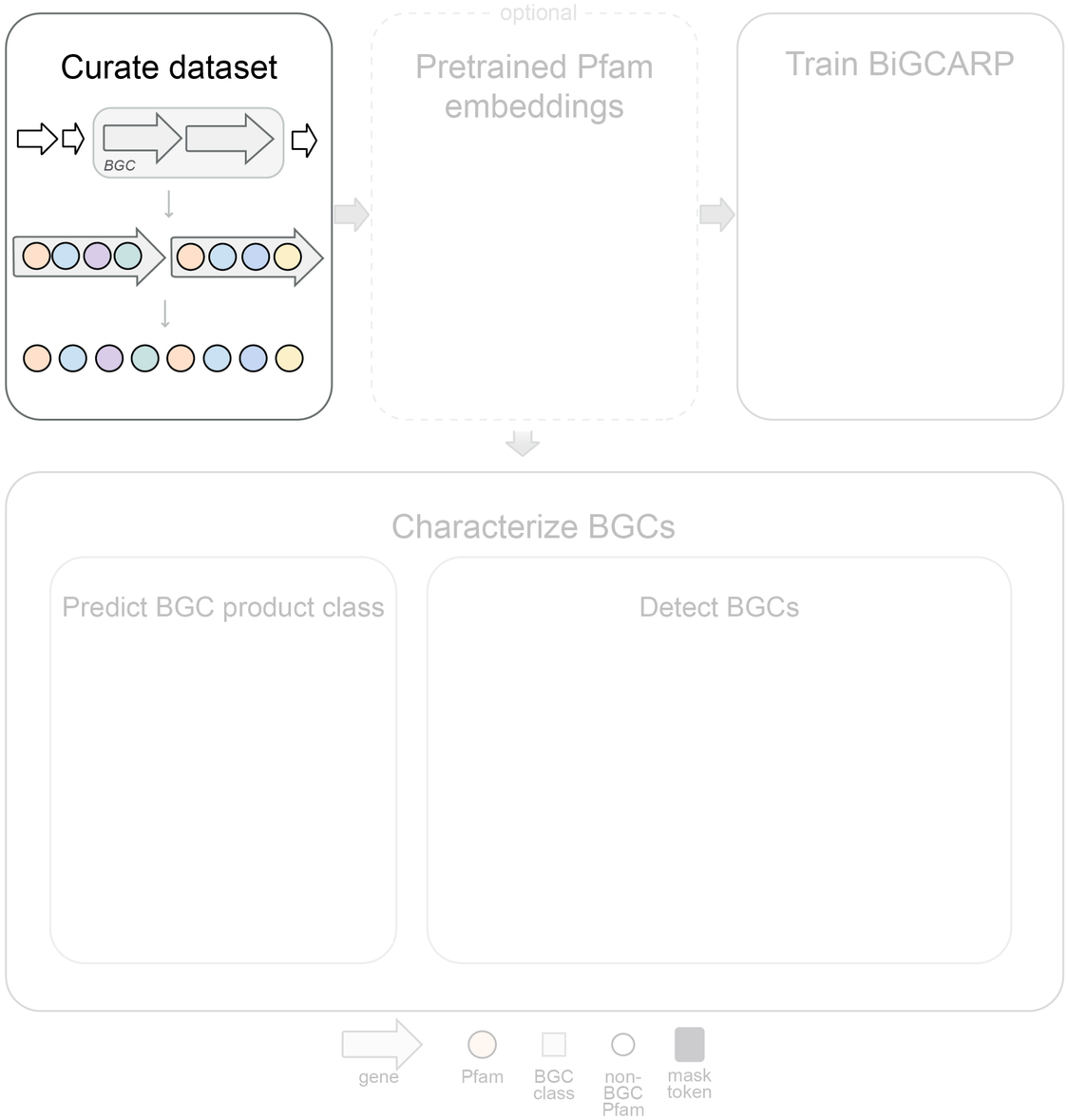
Compromise between granularity,
vocab size, and sequence length

PF02353
Mycolic acid cyclopropane synthetase



PF08659
ketoreductase domain

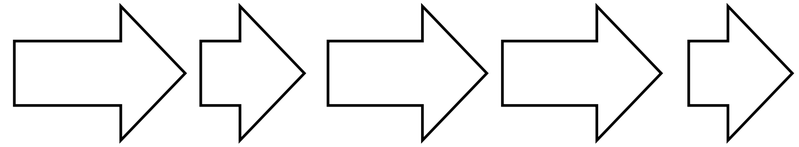
# We represent BGCs as sequences of domains



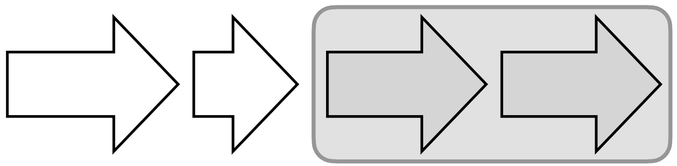

Microbial genome

…ACTGCGGTTACG…

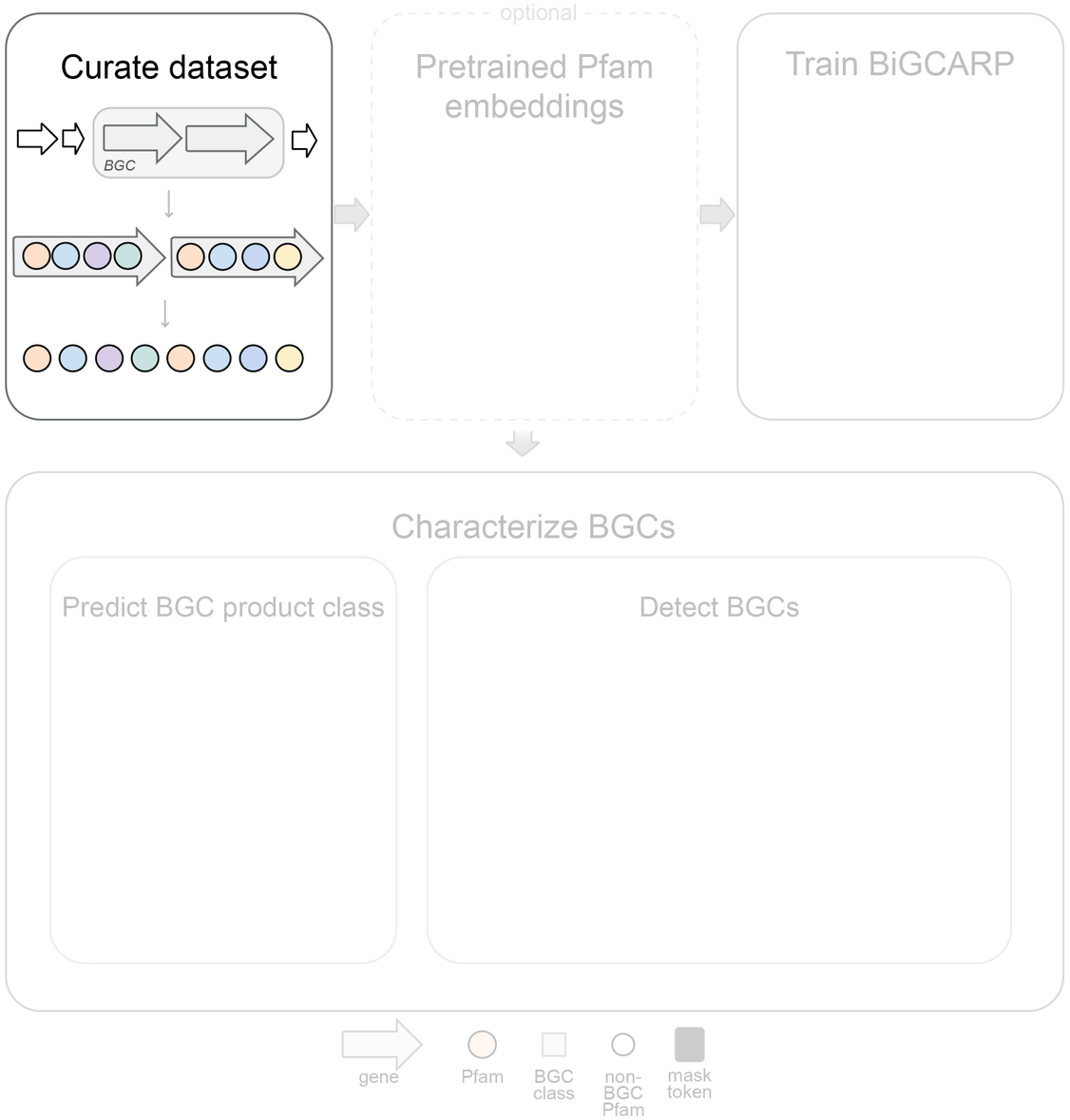
Open reading frames



Identify BGCs with antiSMASH



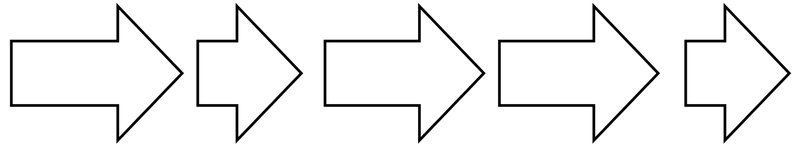
# We represent BGCs as sequences of domains



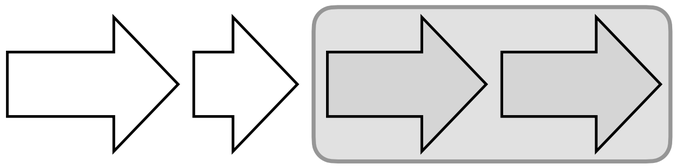

Microbial genome

…ACTGCGGTTACG…

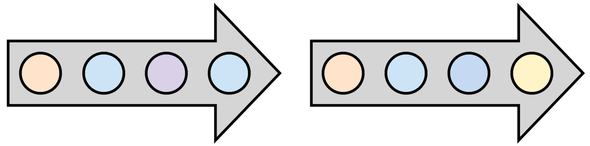
Open reading frames



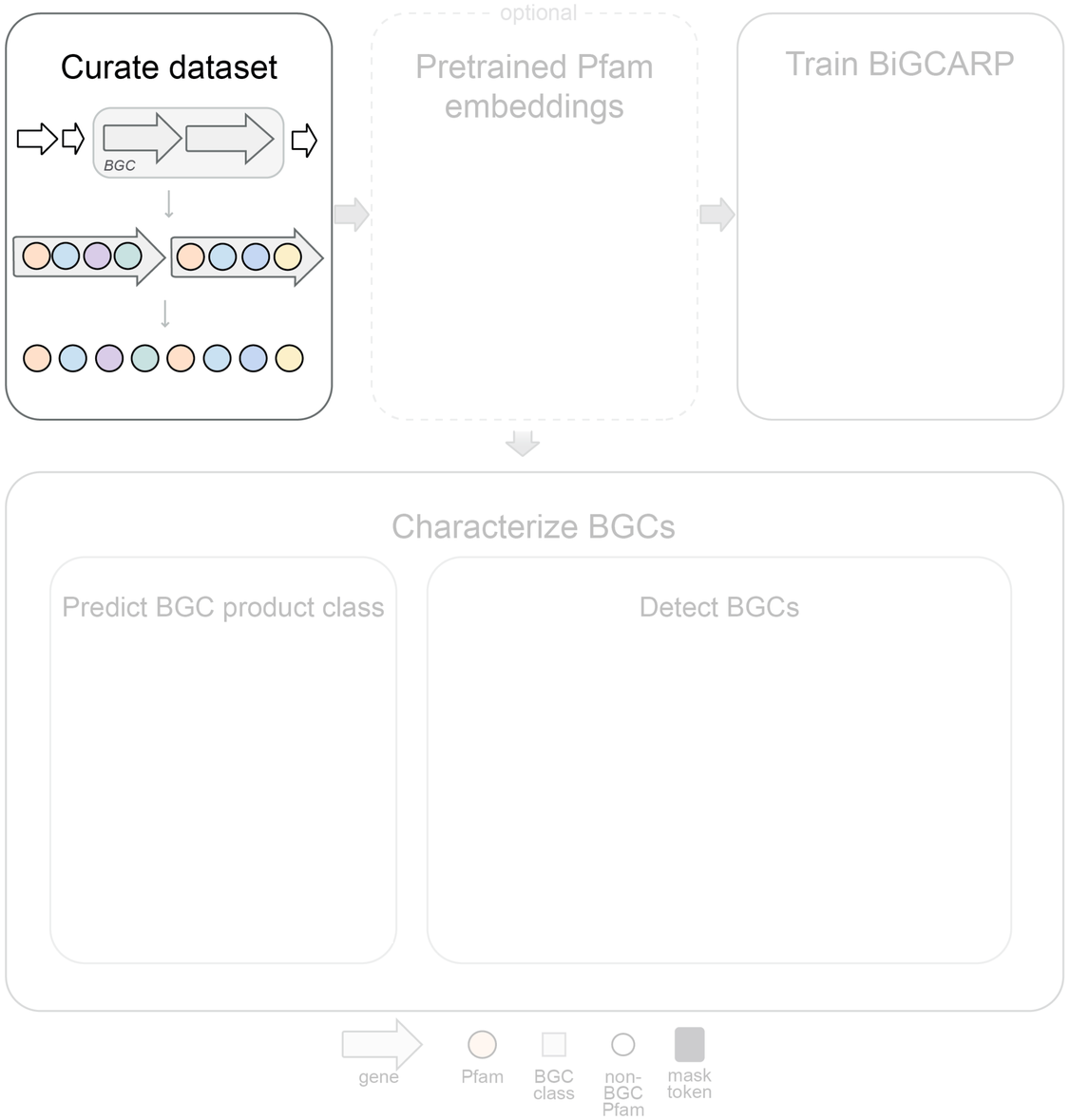
Identify BGCs with antiSMASH



Identify Pfam domains



# We represent BGCs as sequences of domains




Microbial genome

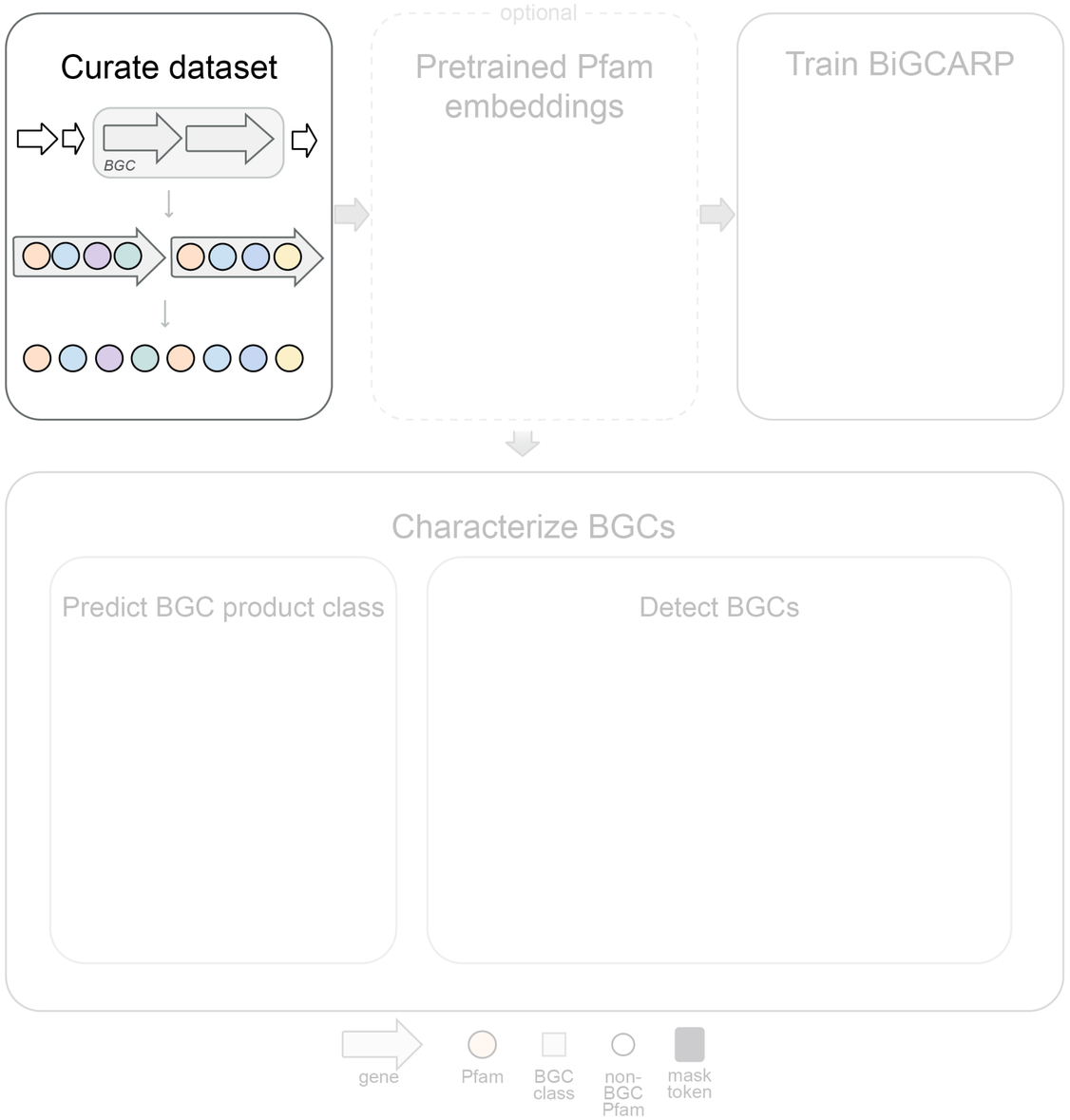
…ACTGCGGTTACG…

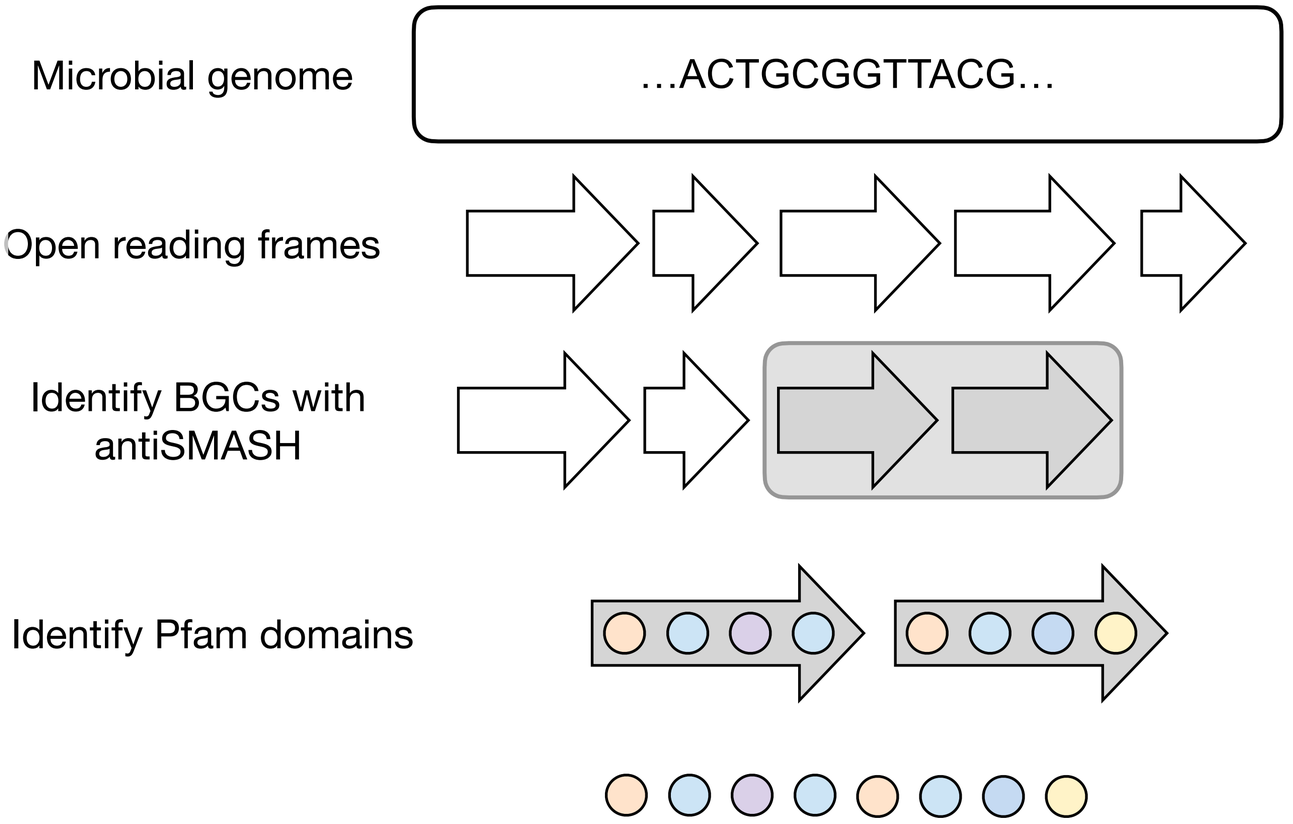
Open reading frames

Identify BGCs with antiSMASH

Identify Pfam domains

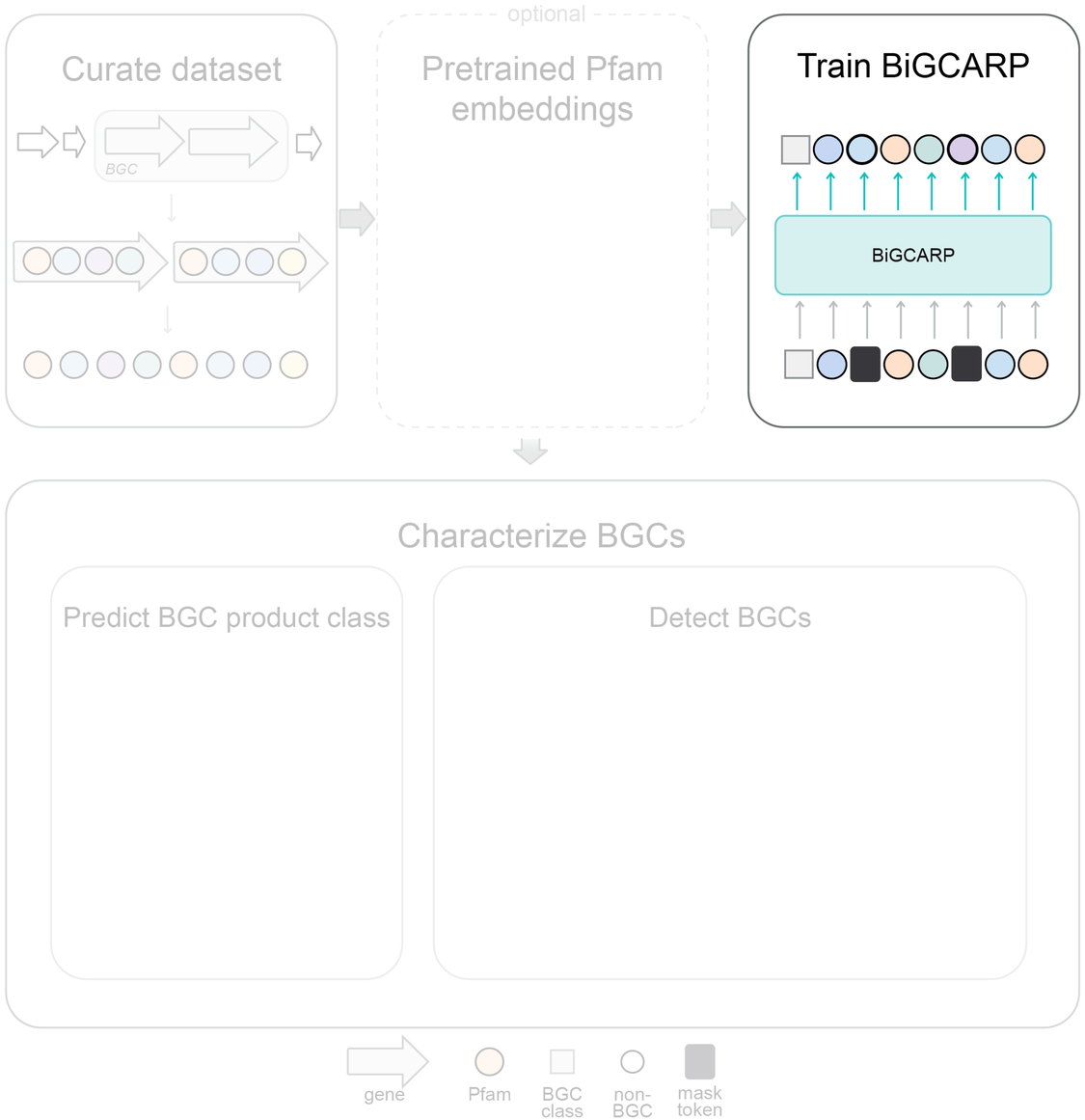
# We represent BGCs as sequences of domains



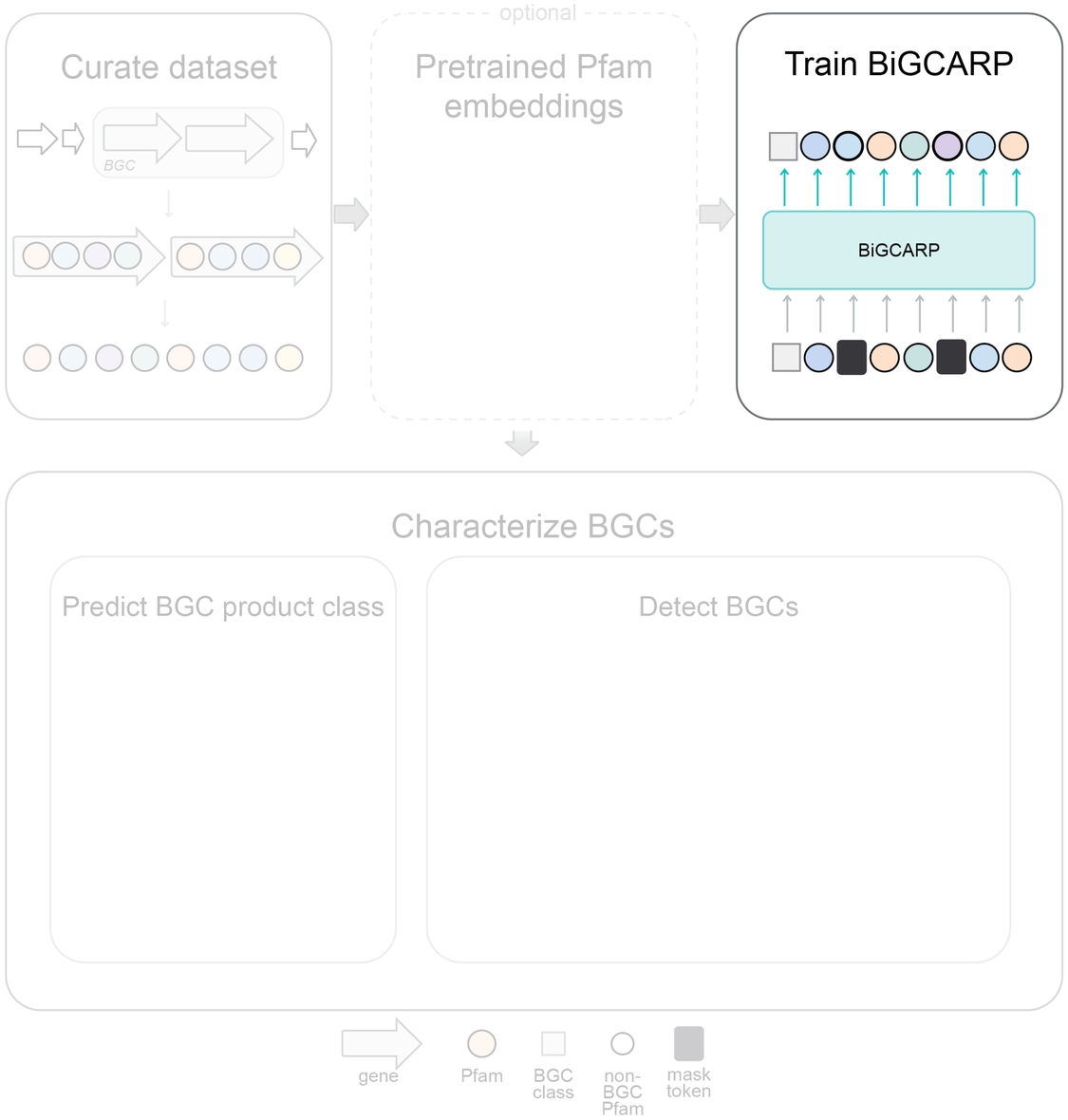




127k BGCs, 19.5k unique domain, 55 product classes

# BiGCARP is trained to reconstruct corrupted BGCs

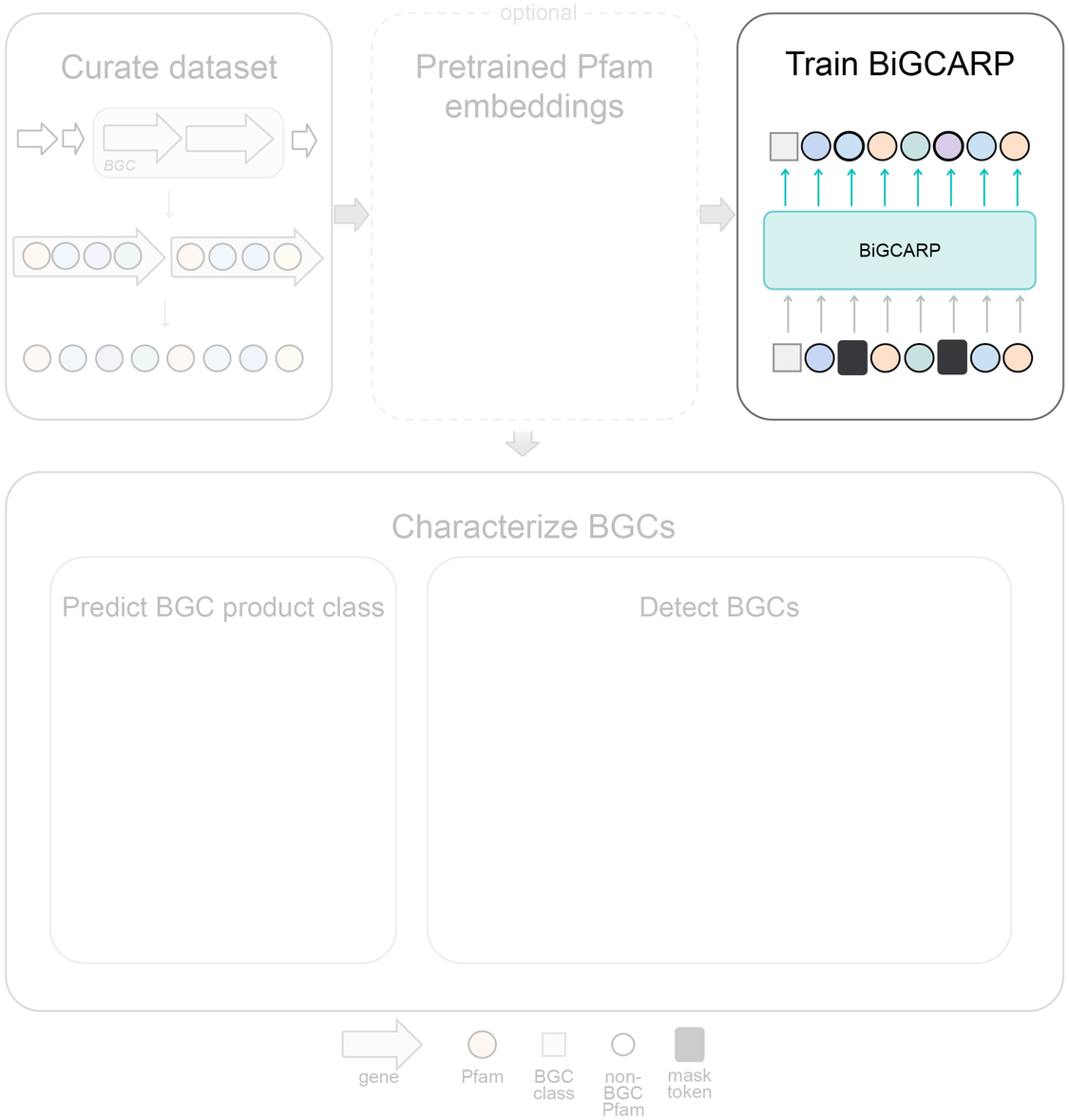


# BiGCARP is trained to reconstruct corrupted BGCs




self-supervised dataset
(127k BGC Pfam sequences)

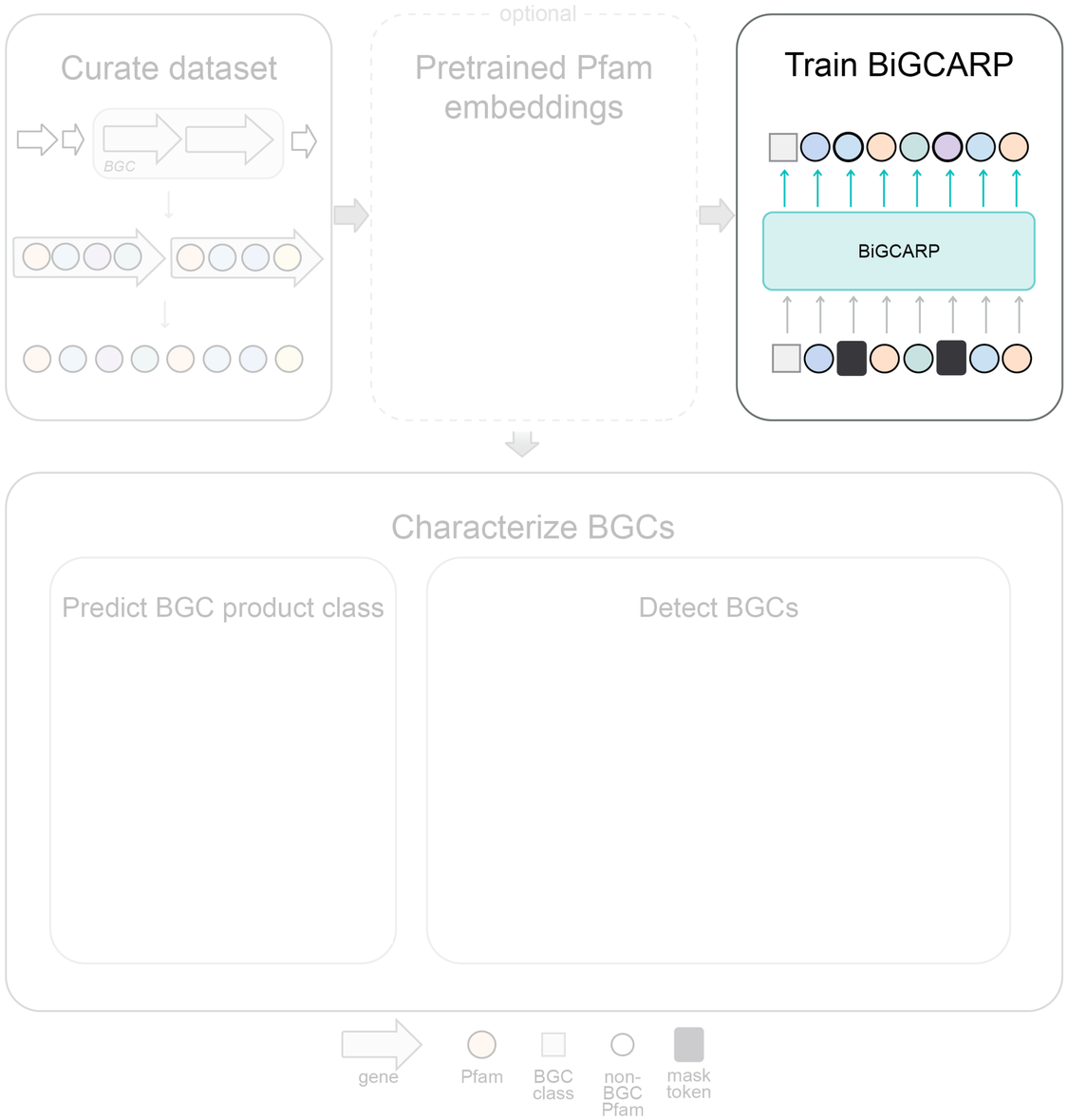
# BiGCARP is trained to reconstruct corrupted BGCs




Prepend
product class token

self-supervised dataset
(127k BGC Pfam sequences)

# BiGCARP is trained to reconstruct corrupted BGCs

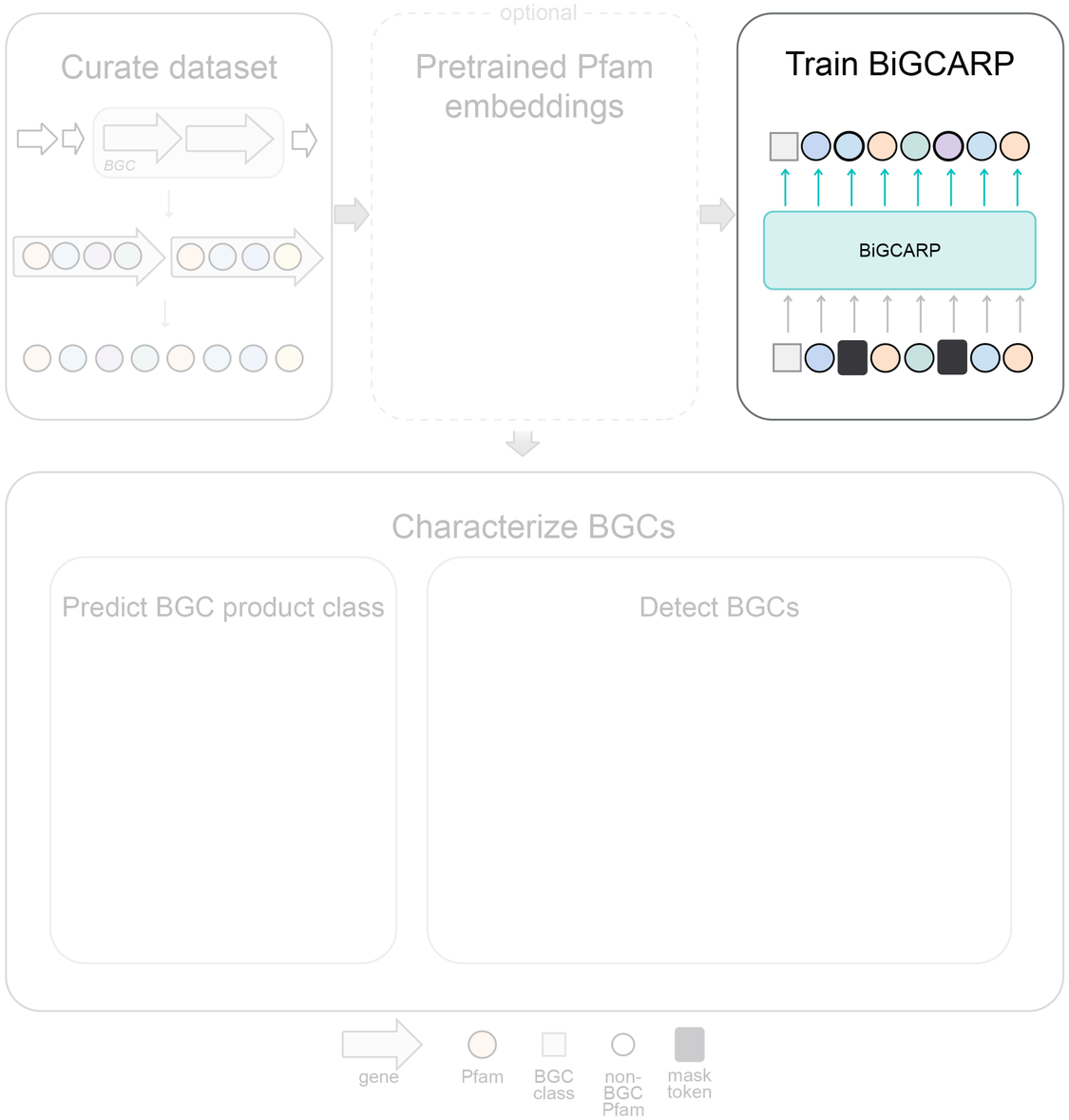


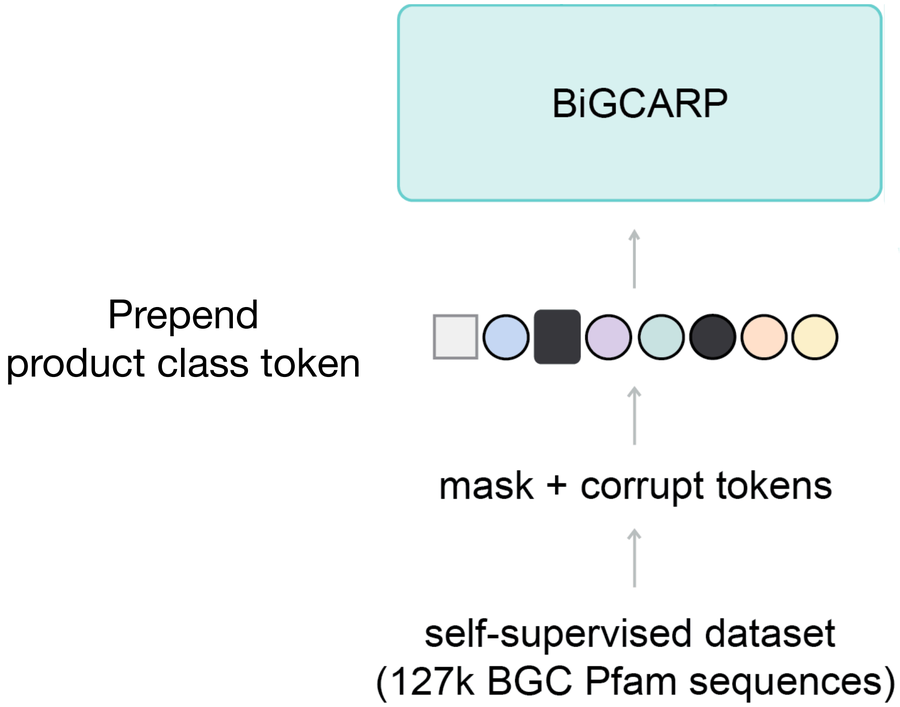

Prepend product class token

mask + corrupt tokens

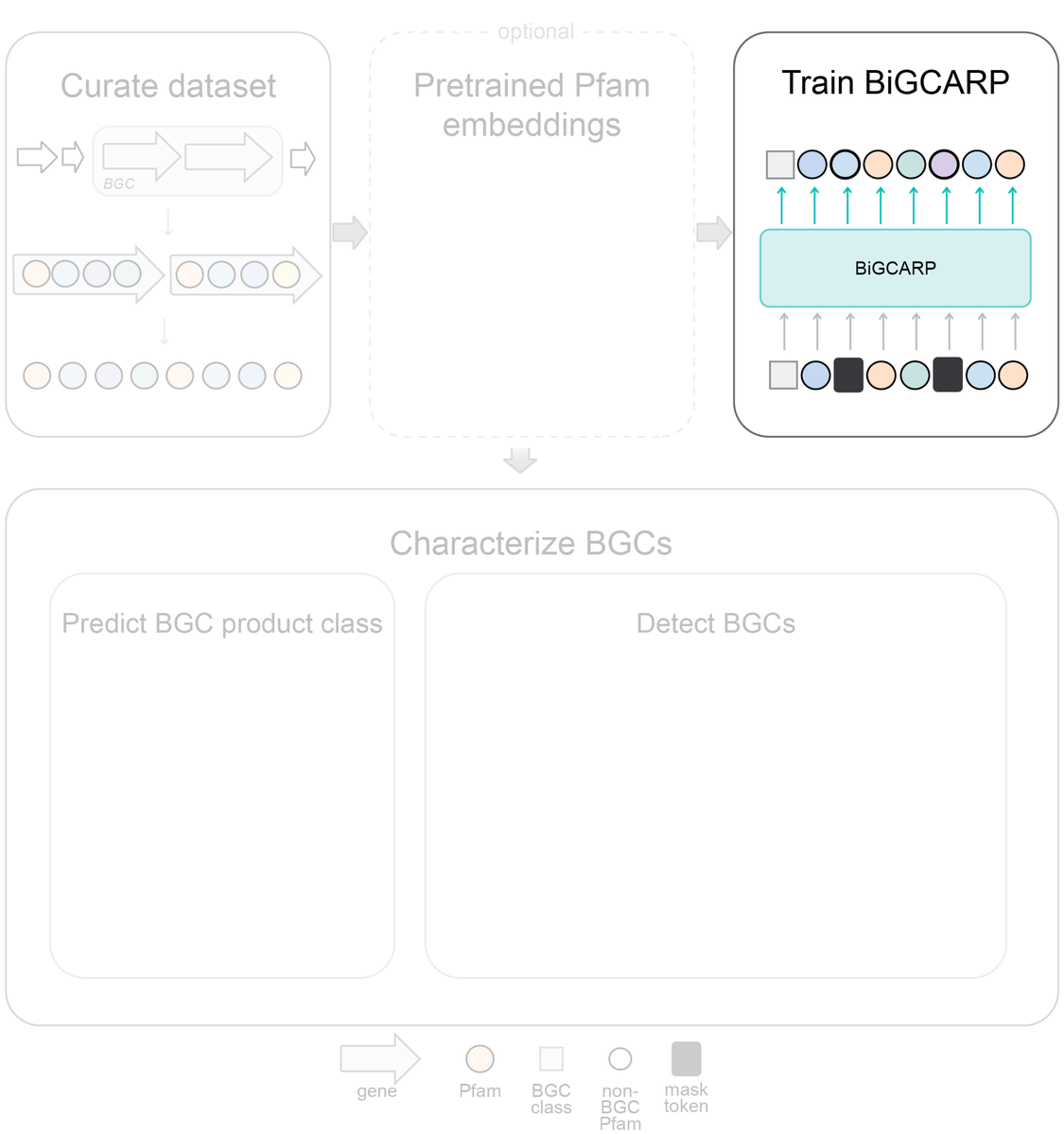
self-supervised dataset
(127k BGC Pfam sequences)

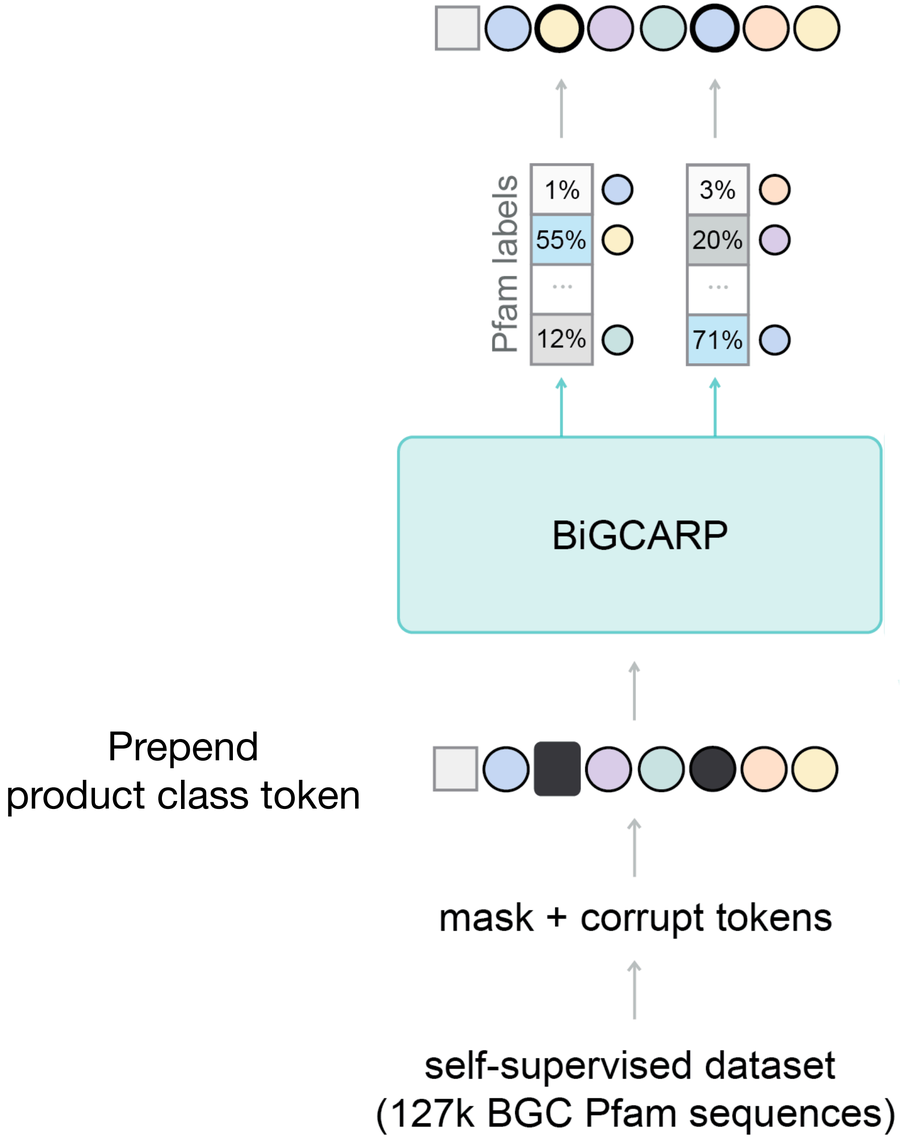
# BiGCARP is trained to reconstruct corrupted BGCs



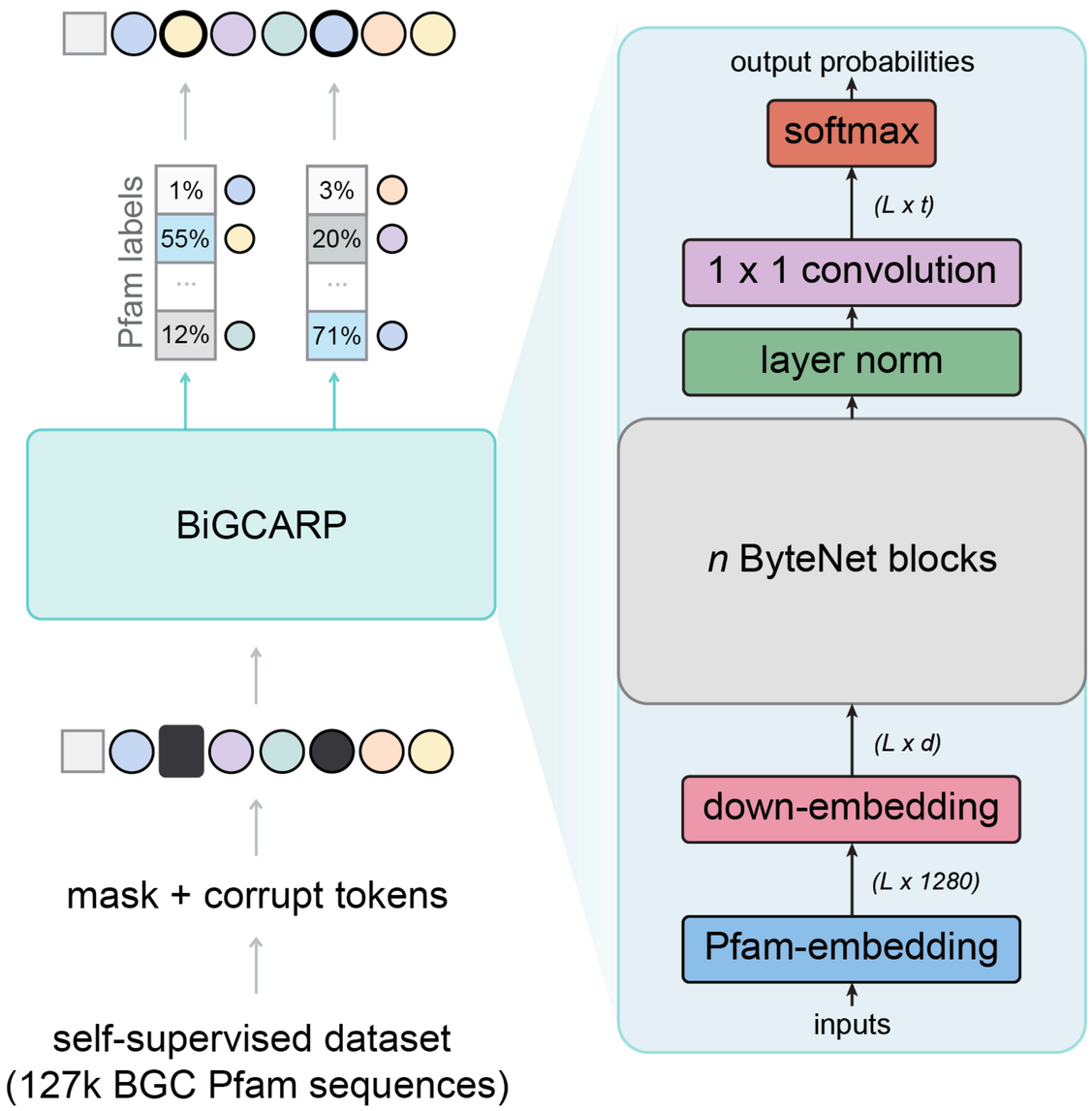



# BiGCARP is trained to reconstruct corrupted BGCs

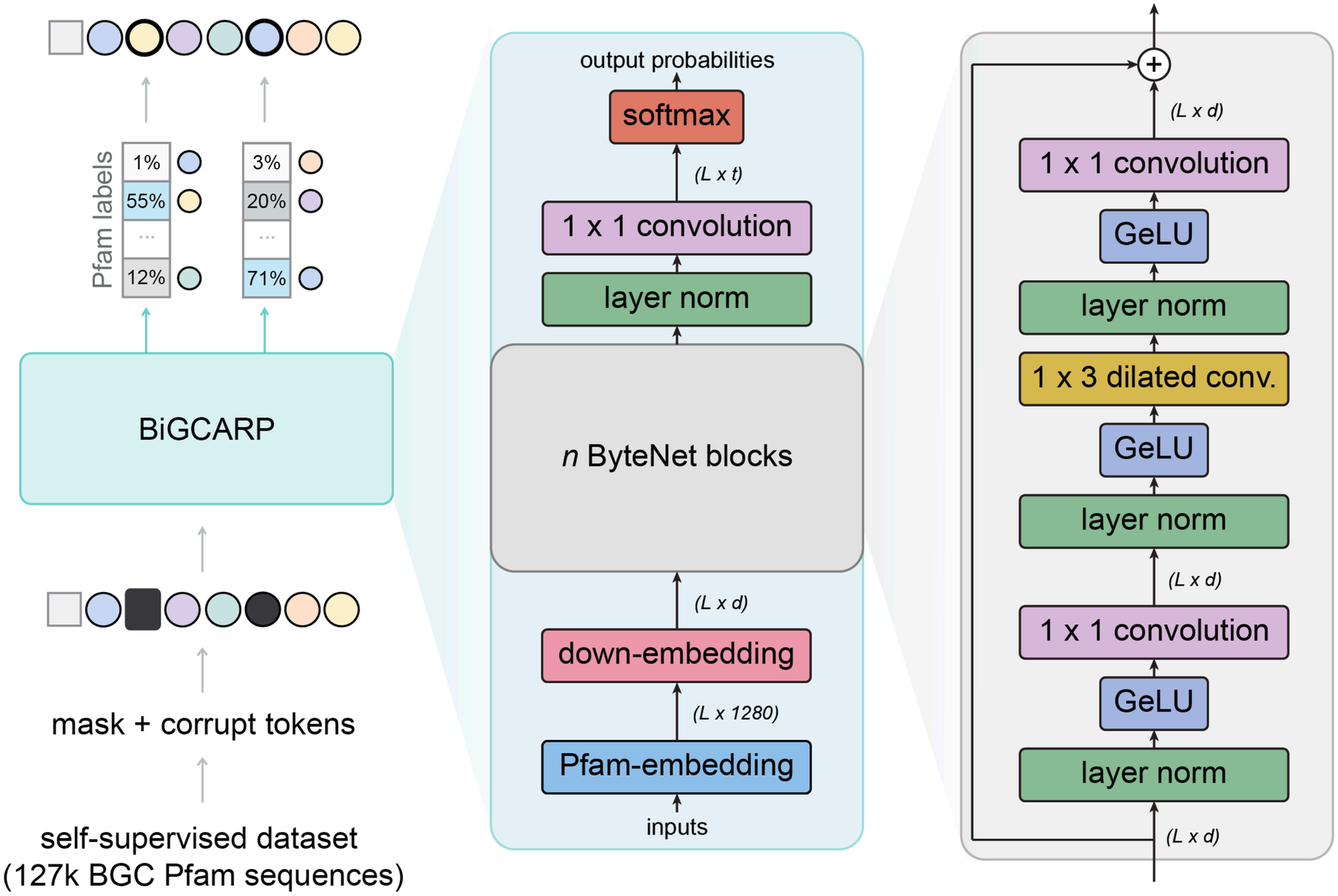


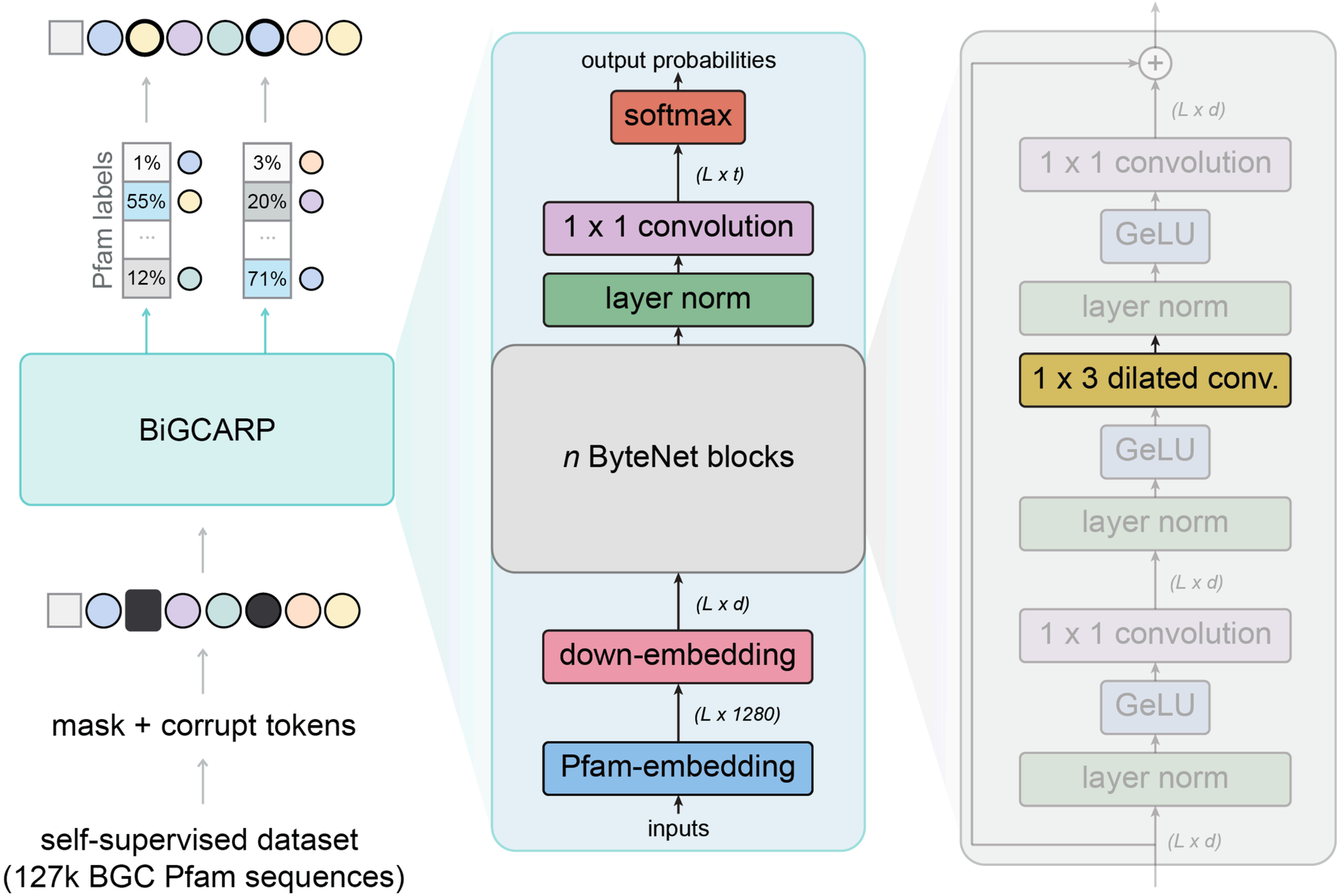
# BiGCARP is trained to reconstruct corrupted BGCs



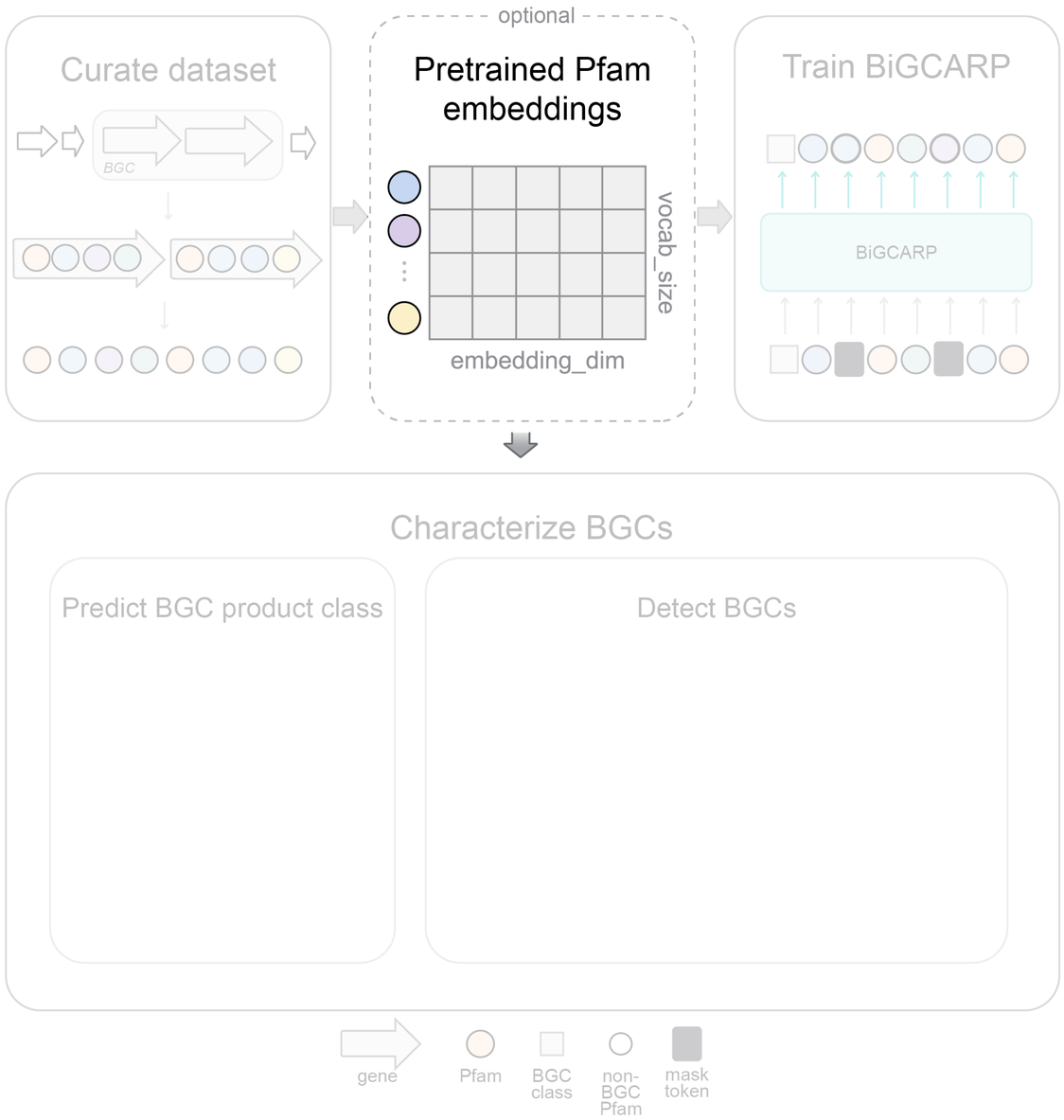
# BiGCARP is trained to reconstruct corrupted BGCs

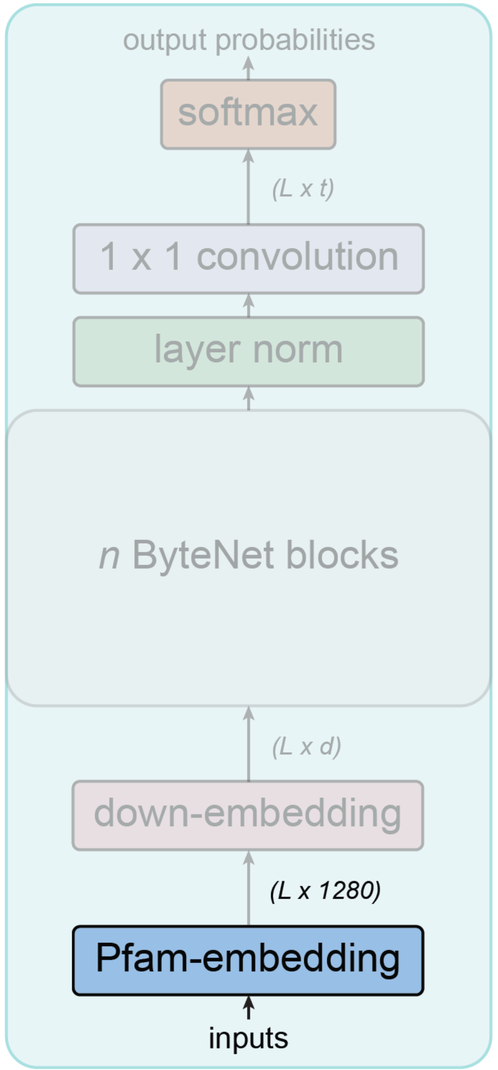


# BiGCARP is trained to reconstruct corrupted BGCs

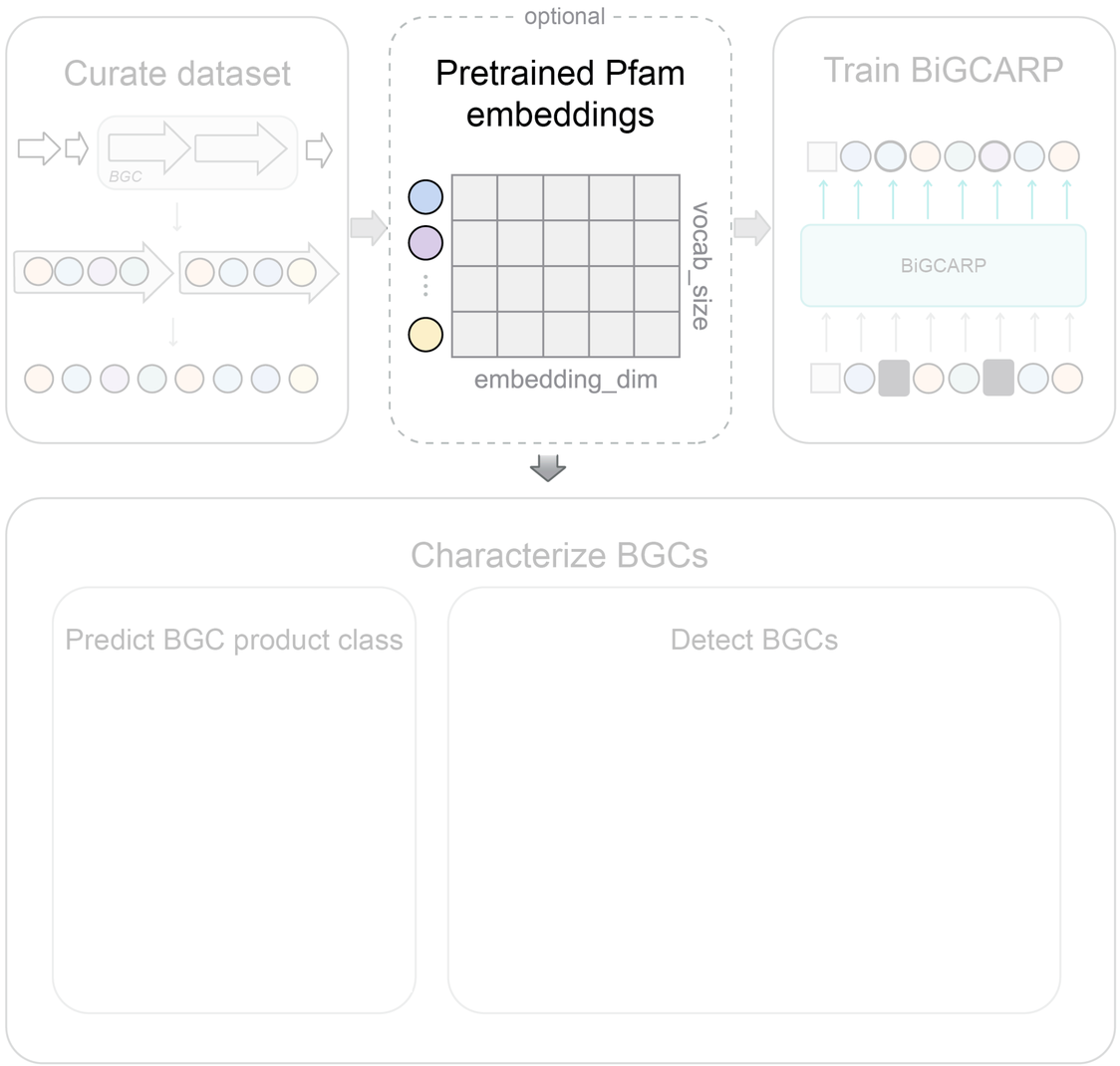


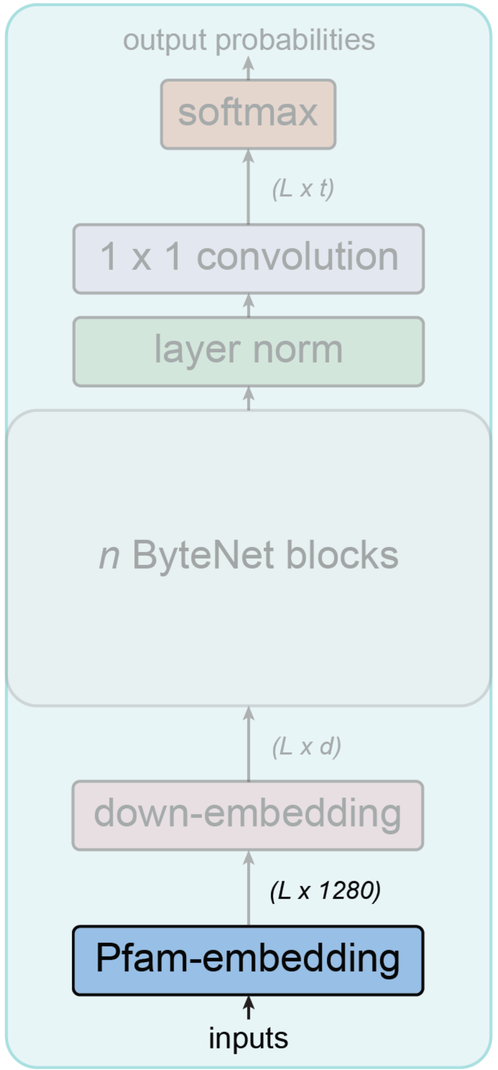
# Try 3 different initial domain embeddings



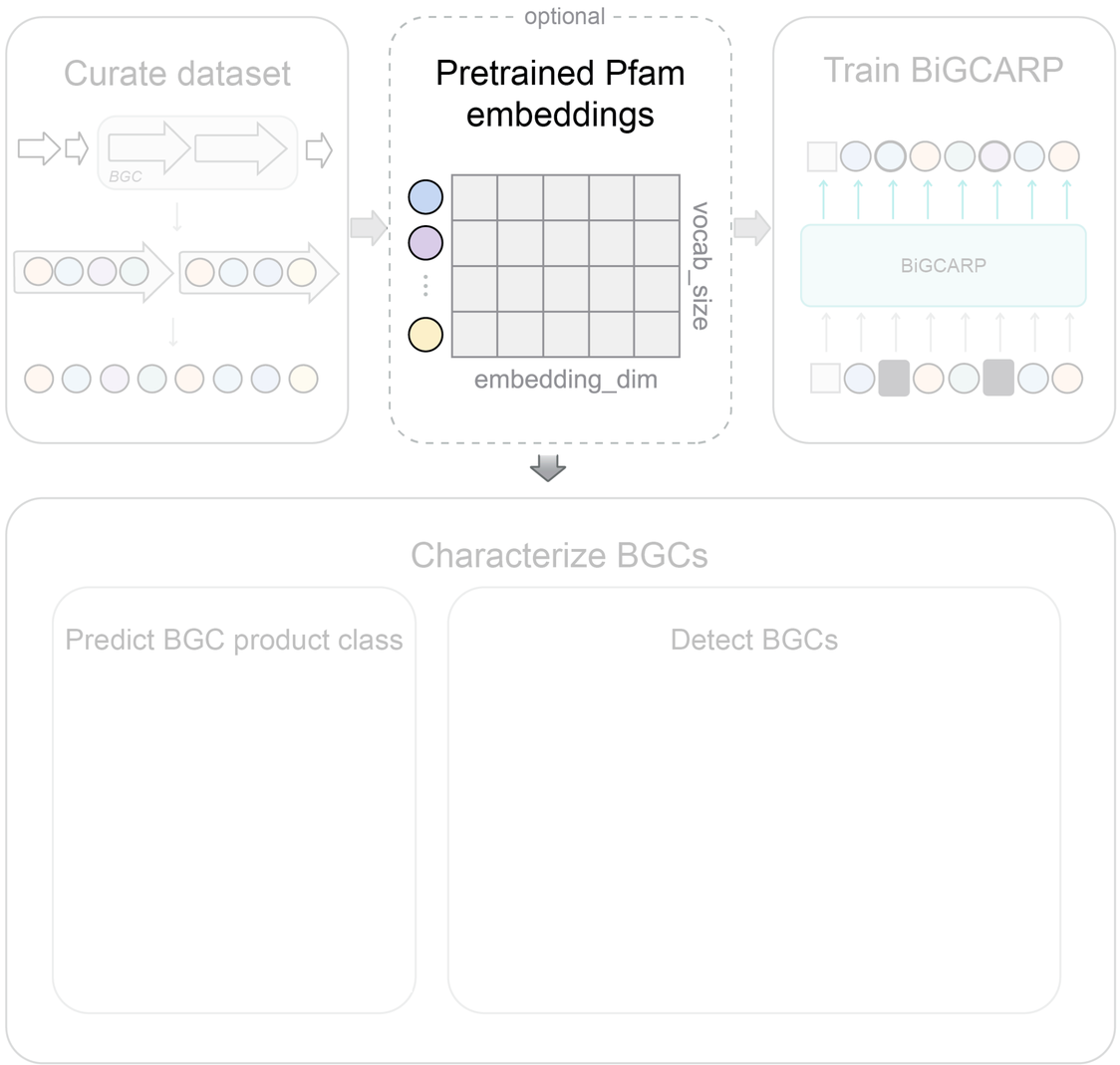



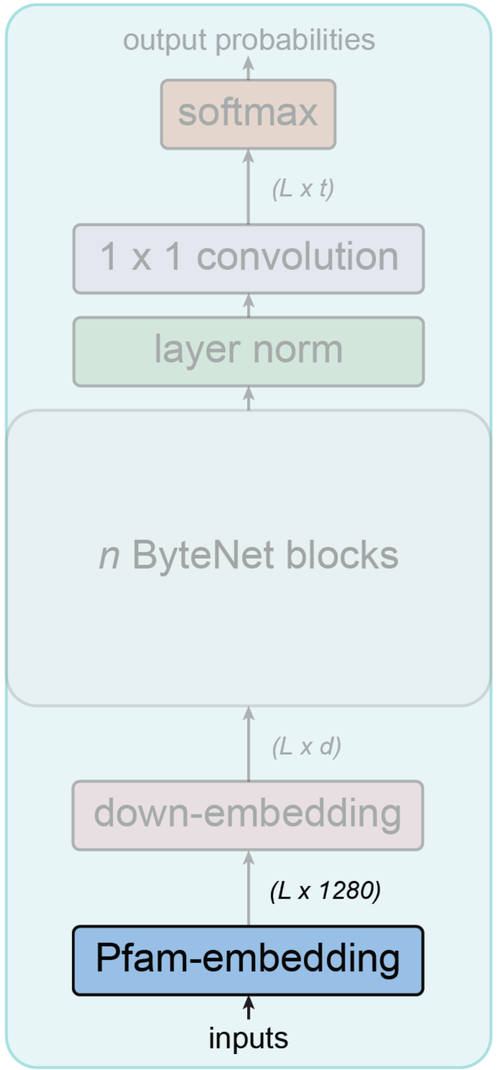
# Try 3 different initial domain embeddings



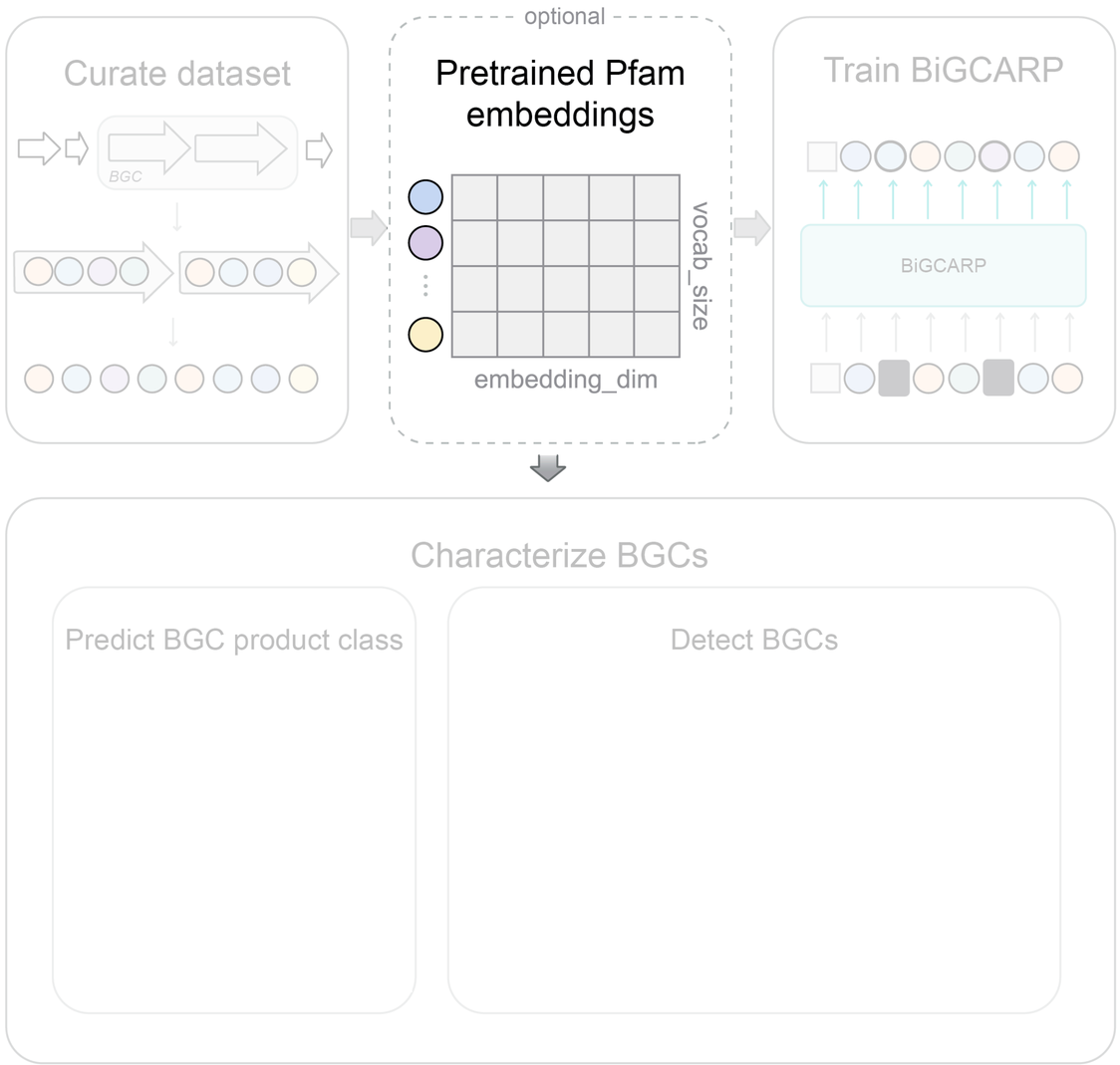

# Try 3 different initial domain embeddings

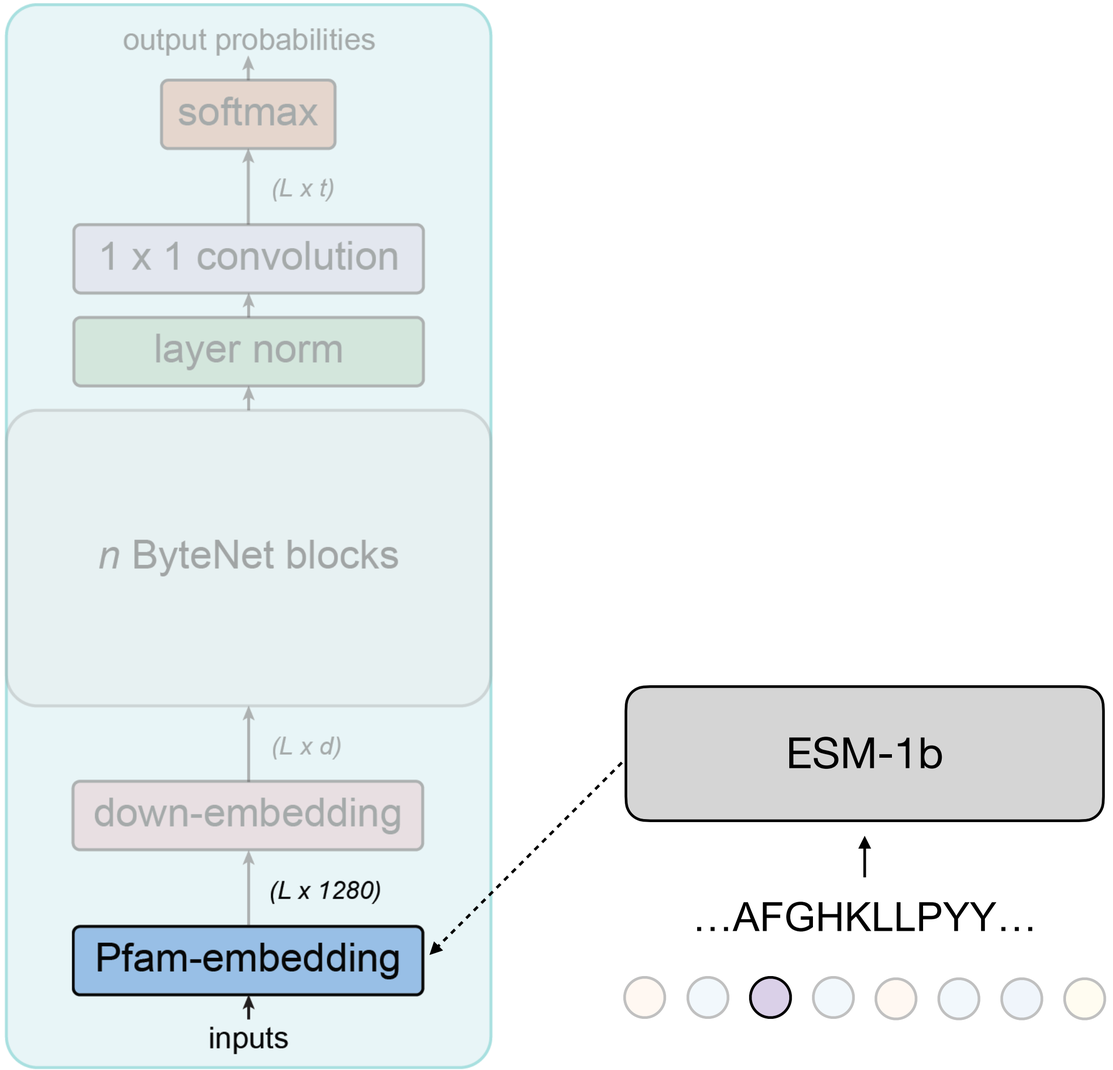


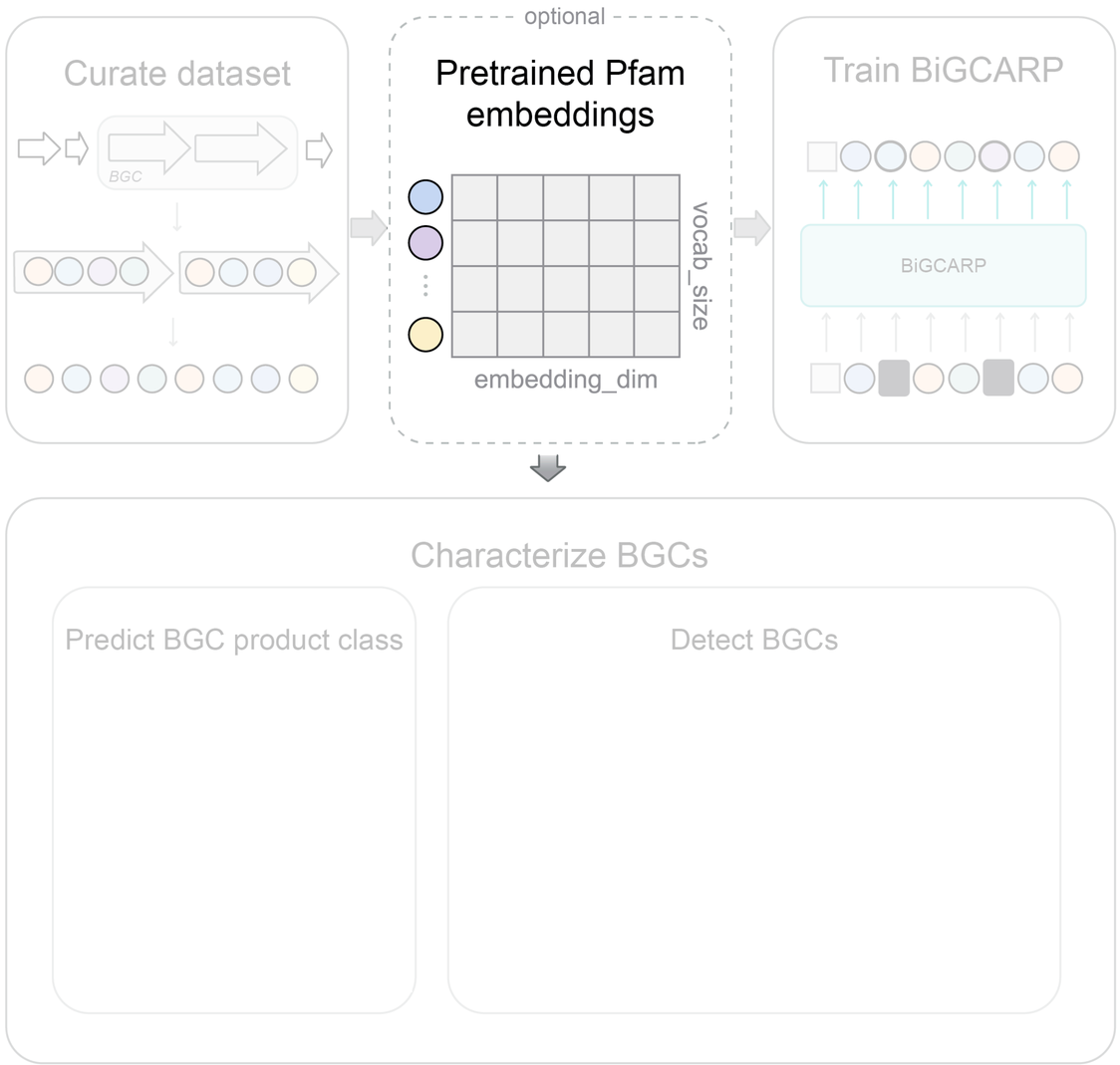

…AFGHKLLPYY…

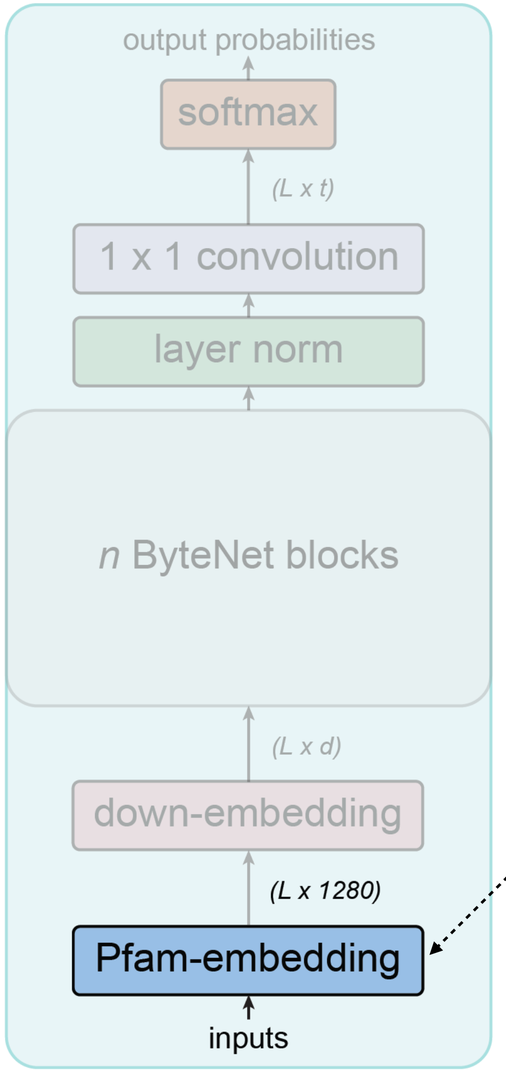
# Try 3 different initial domain embeddings



# Try 3 different initial domain embeddings

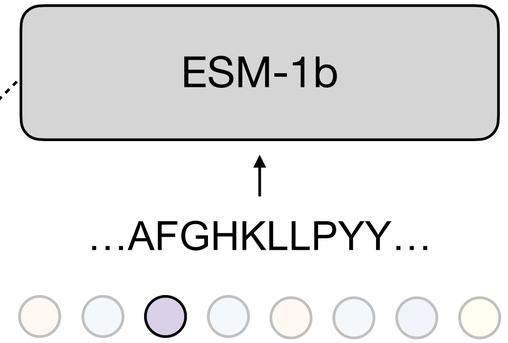


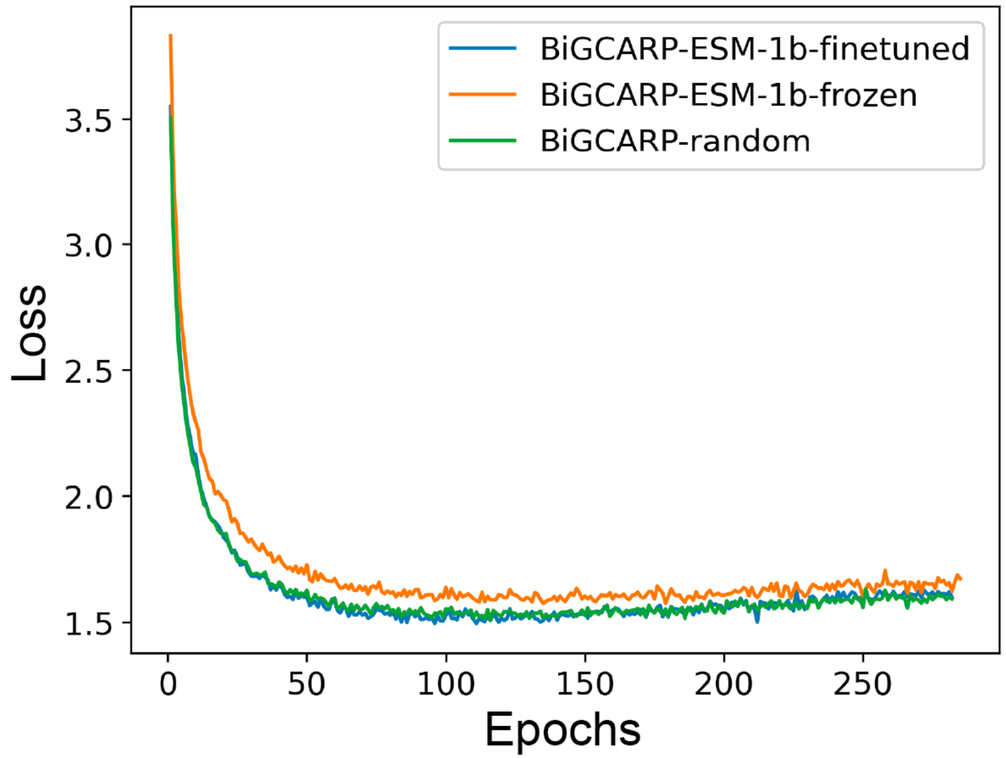

ESM-1b-finetuned

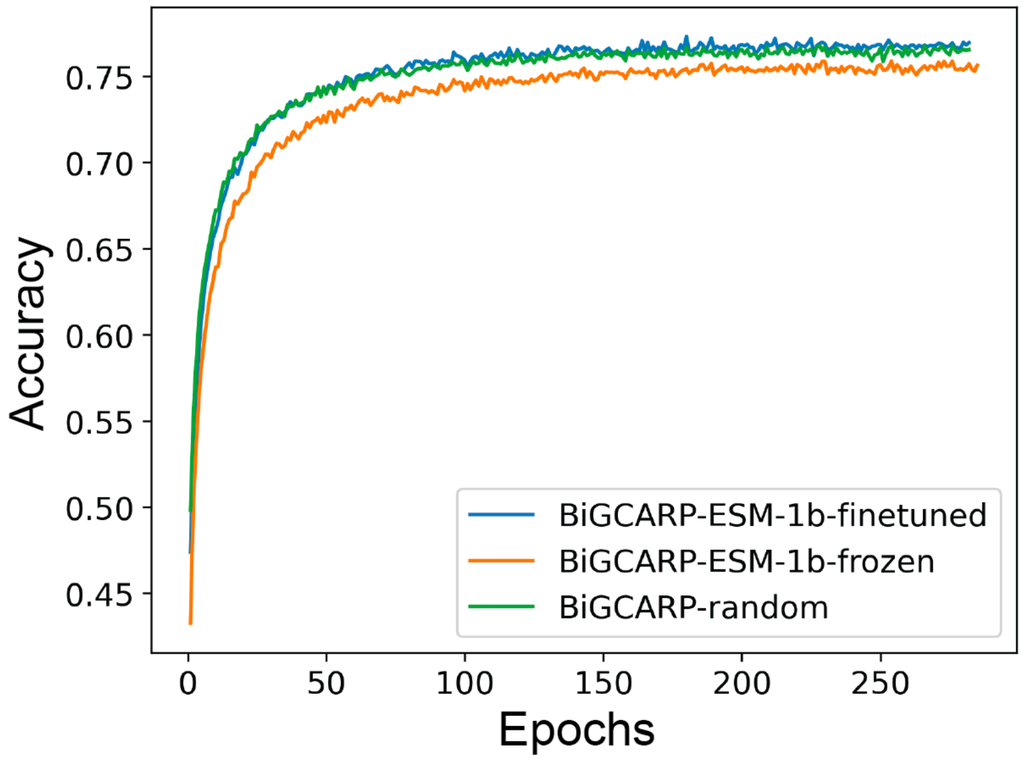
ESM-1b-frozen

Random

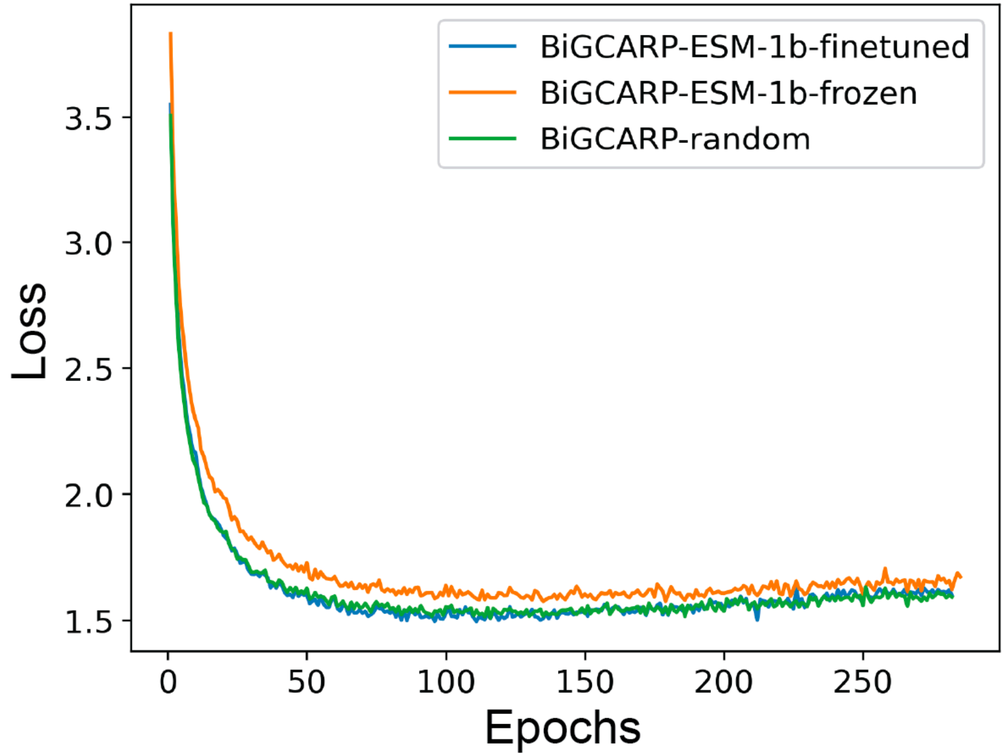


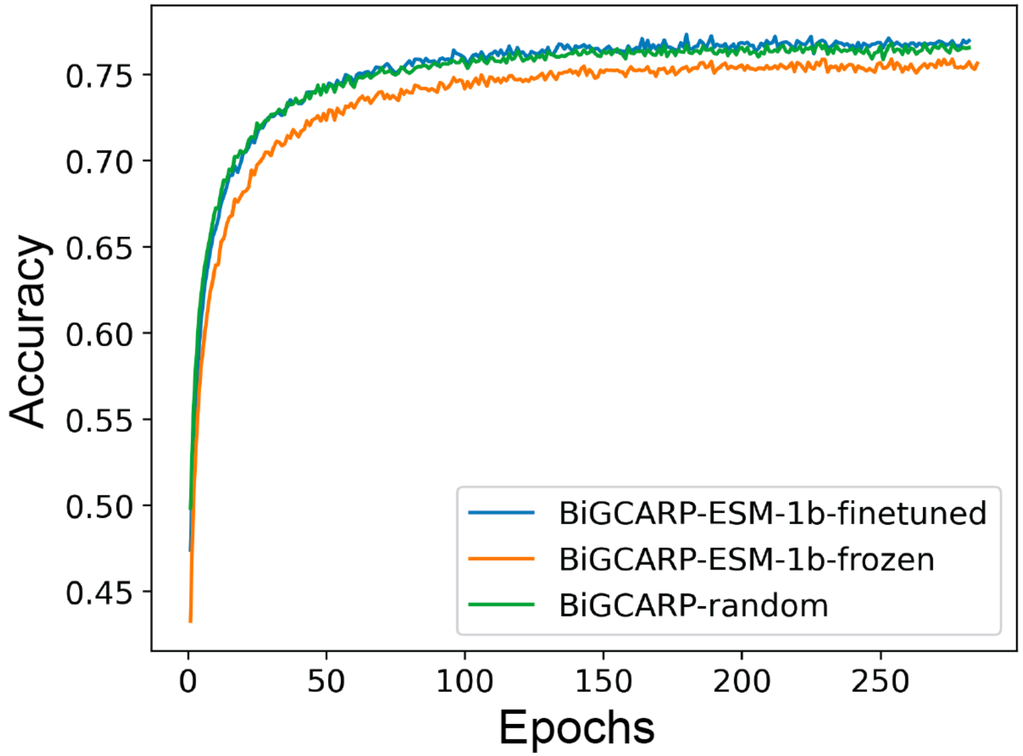
# BiGCARP learns to reconstruct BGC domain sequences



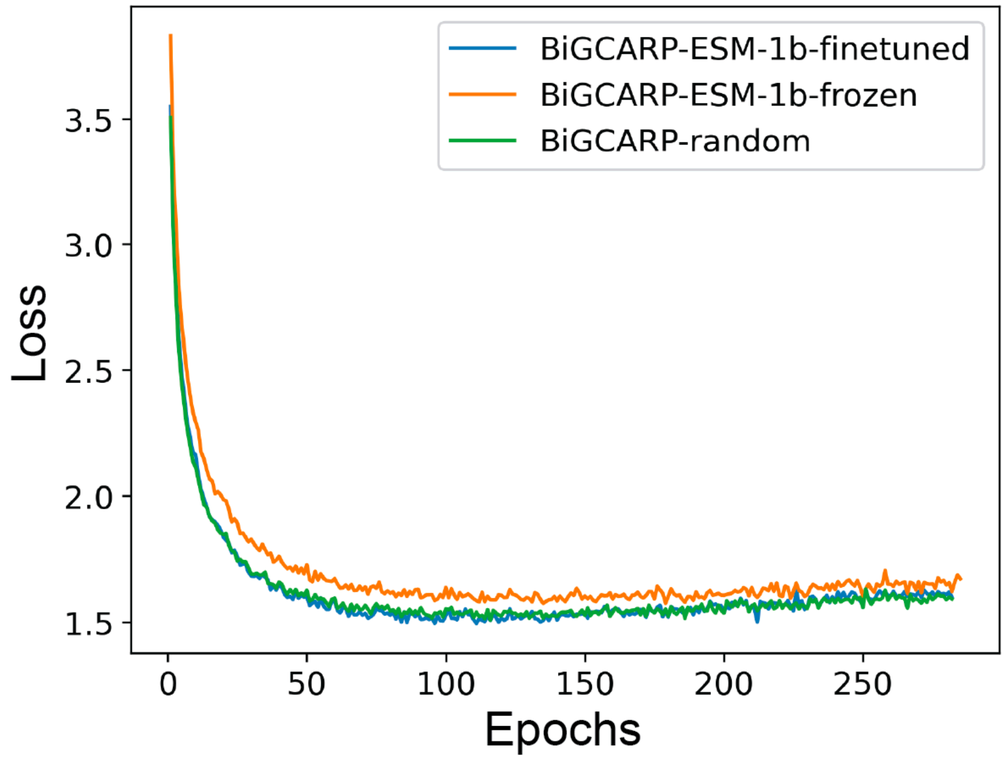
# BiGCARP learns to reconstruct BGC domain sequences

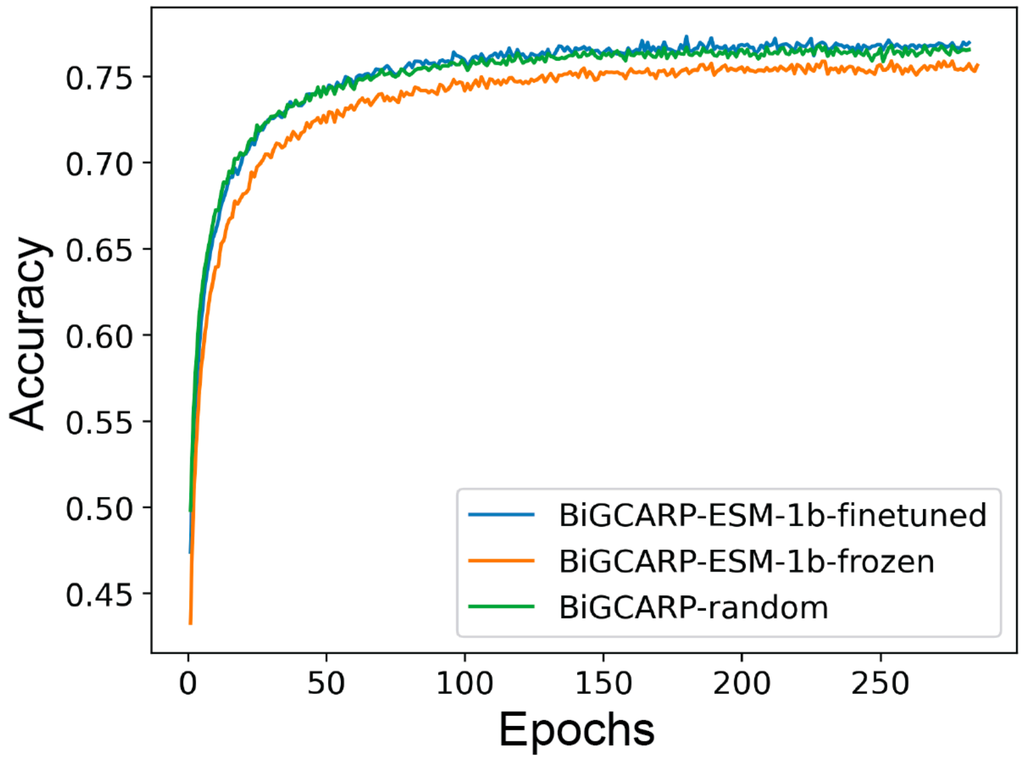



| | BiGCARP | | |
|---|---|---|---|
| | ESM-1b-finetuned | ESM-1b-frozen | random |
| pretrain test set (ECE) | | | |

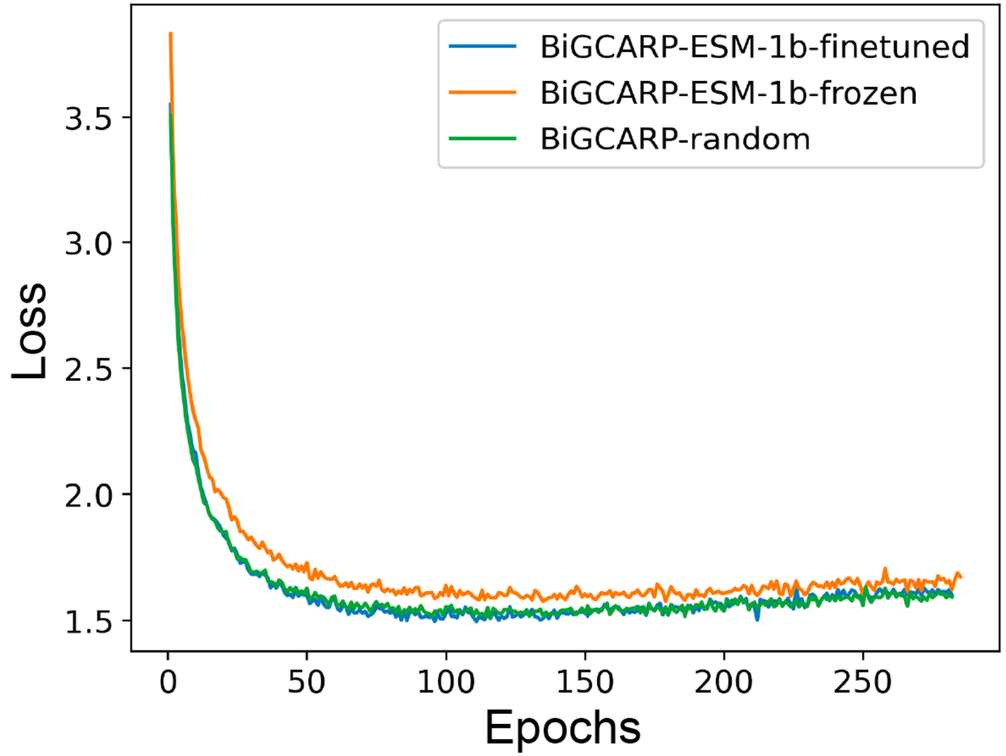
# BiGCARP learns to reconstruct BGC domain sequences

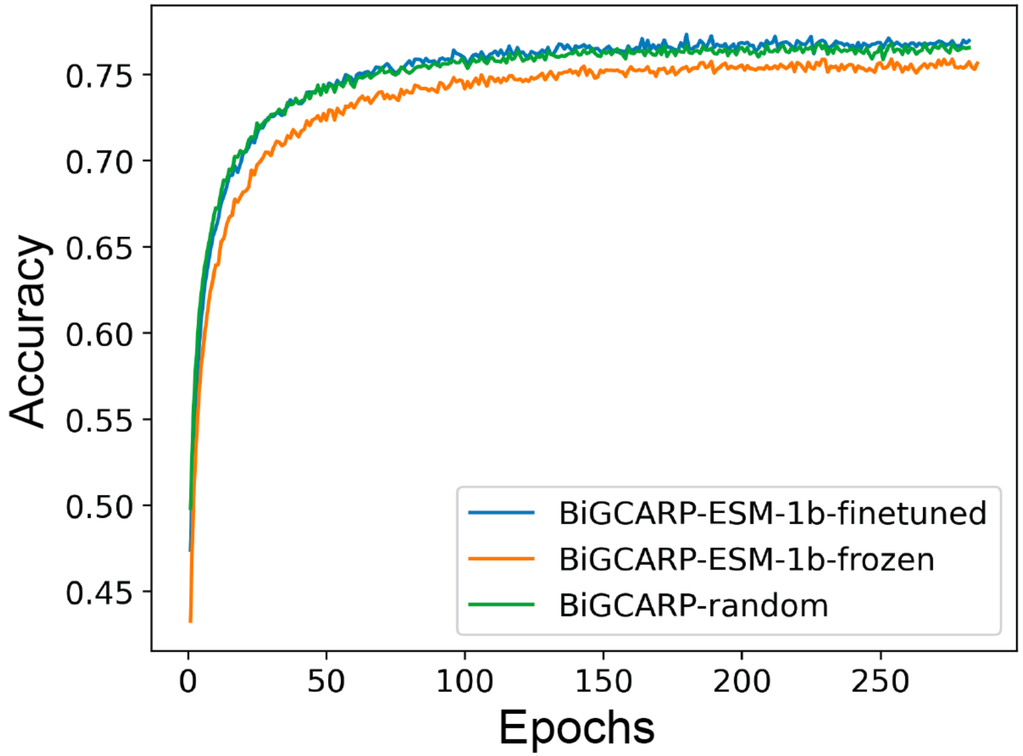



| | | BiGCARP | | |
|---|---|---|---|---|
| | | ESM-1b-finetuned | ESM-1b-frozen | random |
| pretrain test set (ECE) | Pfam domain | **4.64** | 4.99 | 4.67 |

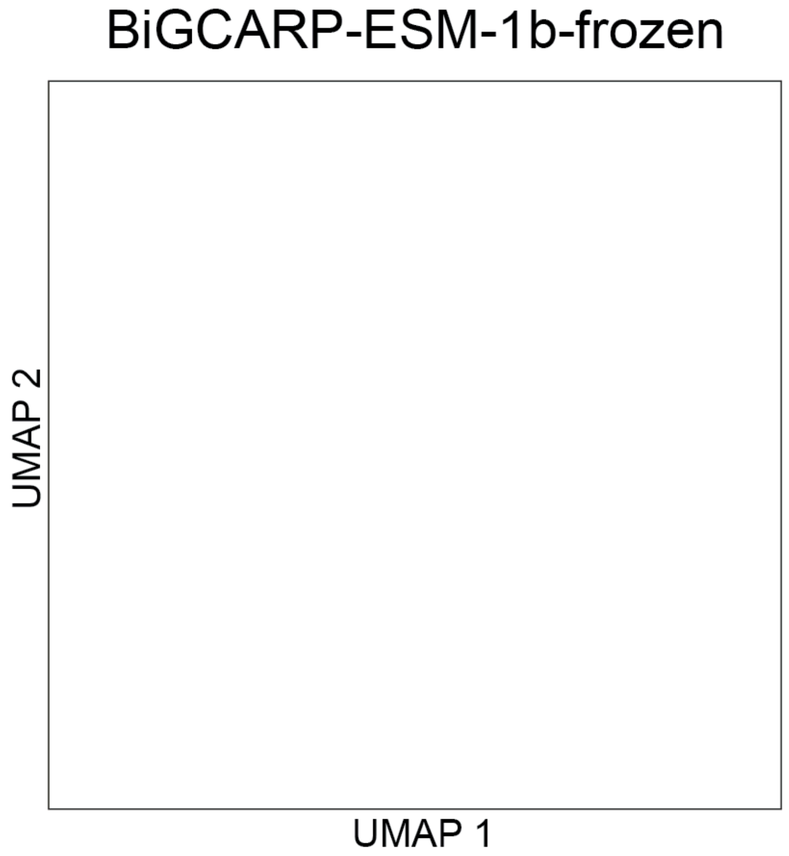
# BiGCARP learns to reconstruct BGC domain sequences



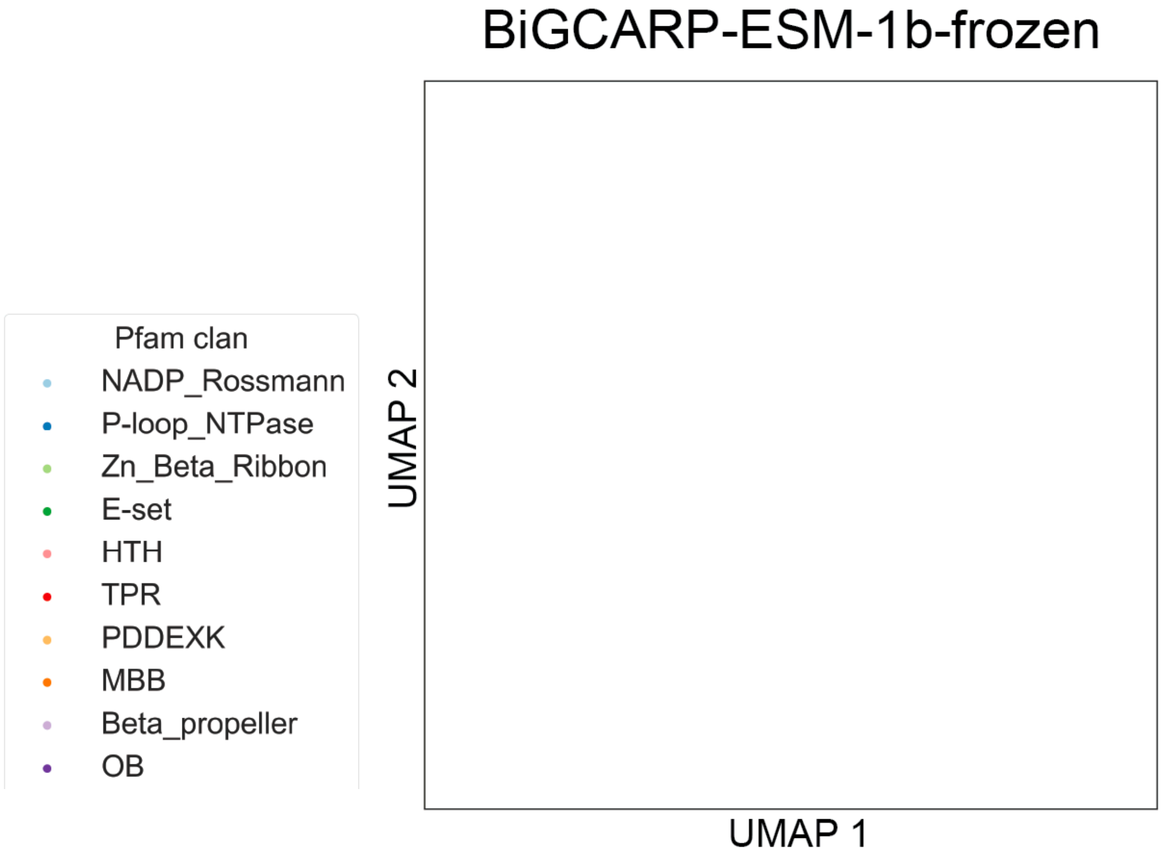




| | | BiGCARP | | |
|---|---|---|---|---|
| | | ESM-1b-finetuned | ESM-1b-frozen | random |
| pretrain test set (ECE) | Pfam domain | **4.64** | 4.99 | 4.67 |
| | product class | 1.50 | **1.46** | 1.50 |

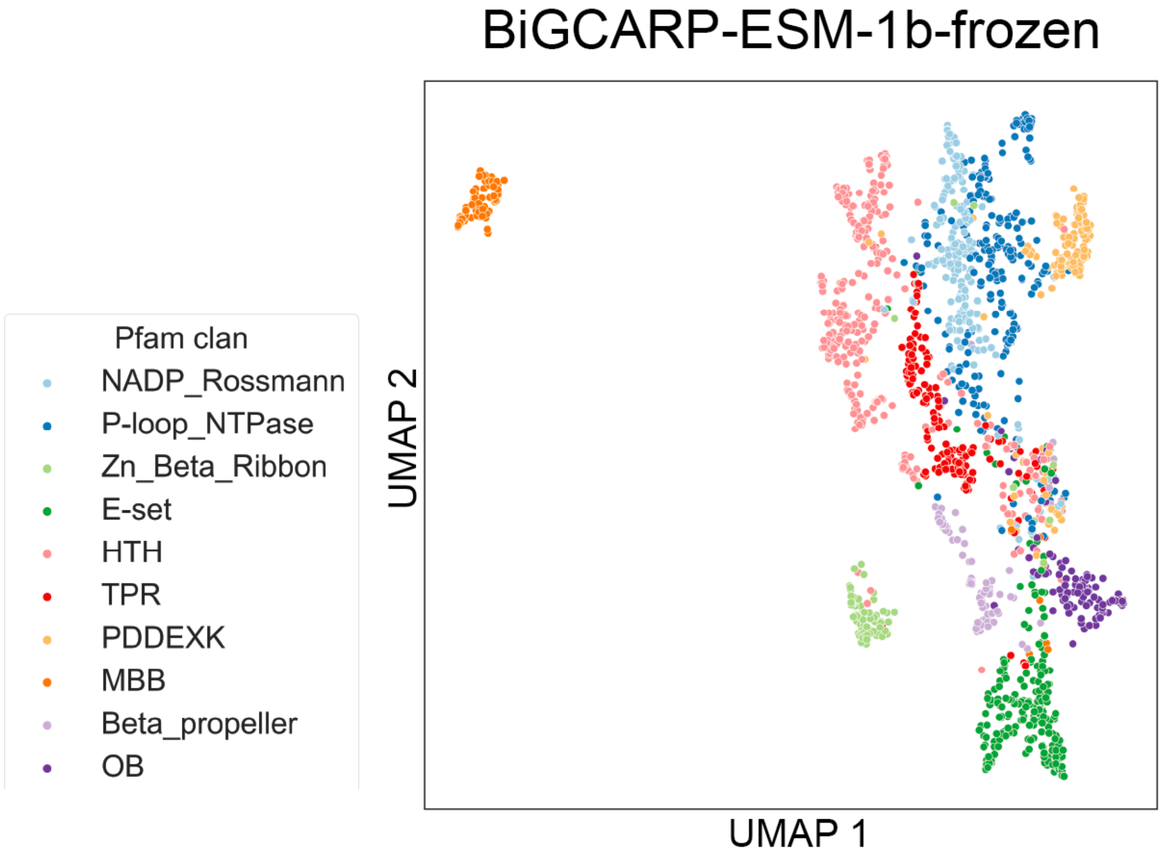
# BiGCARP embeddings encode relevant Pfam information



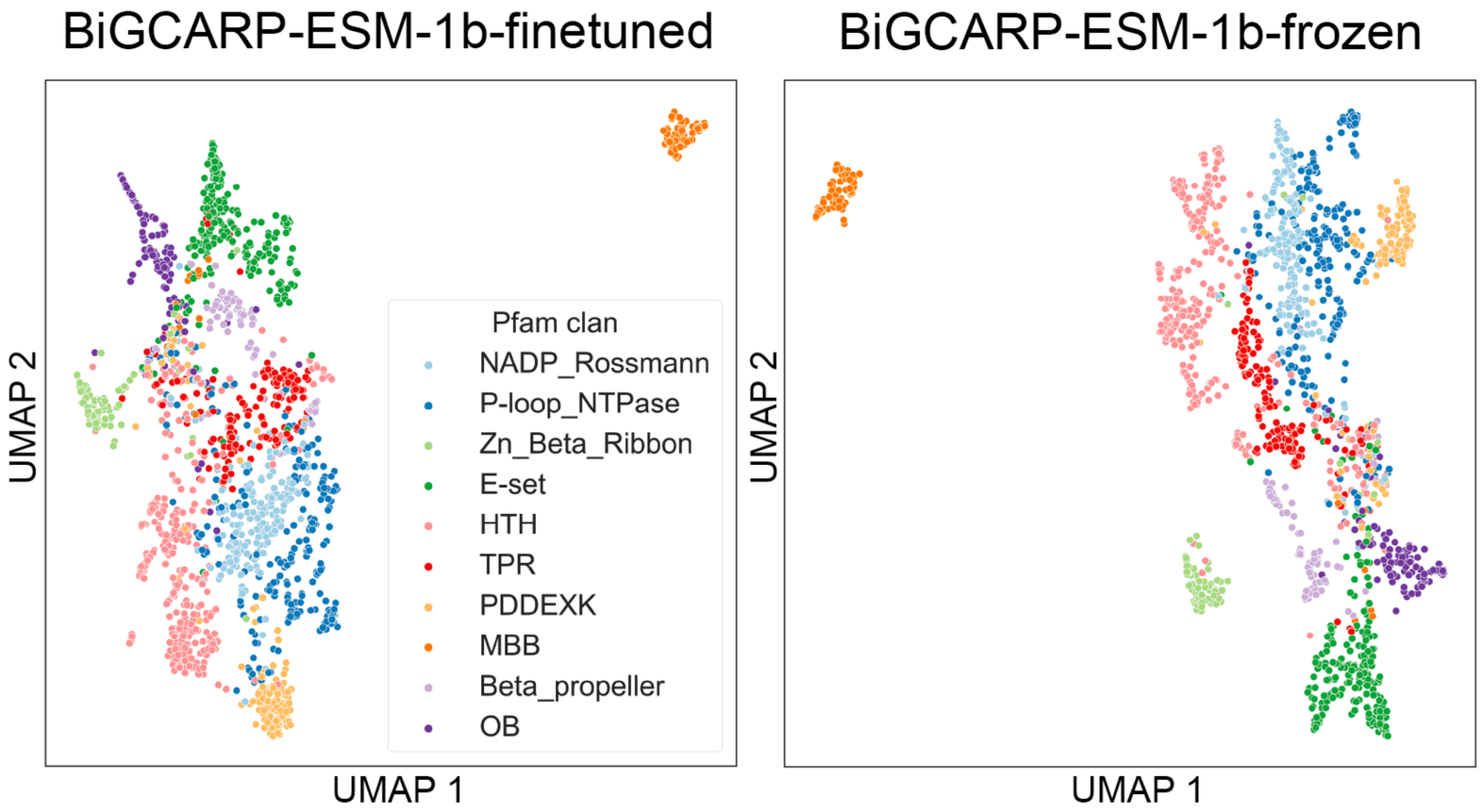
# BiGCARP embeddings encode relevant Pfam information



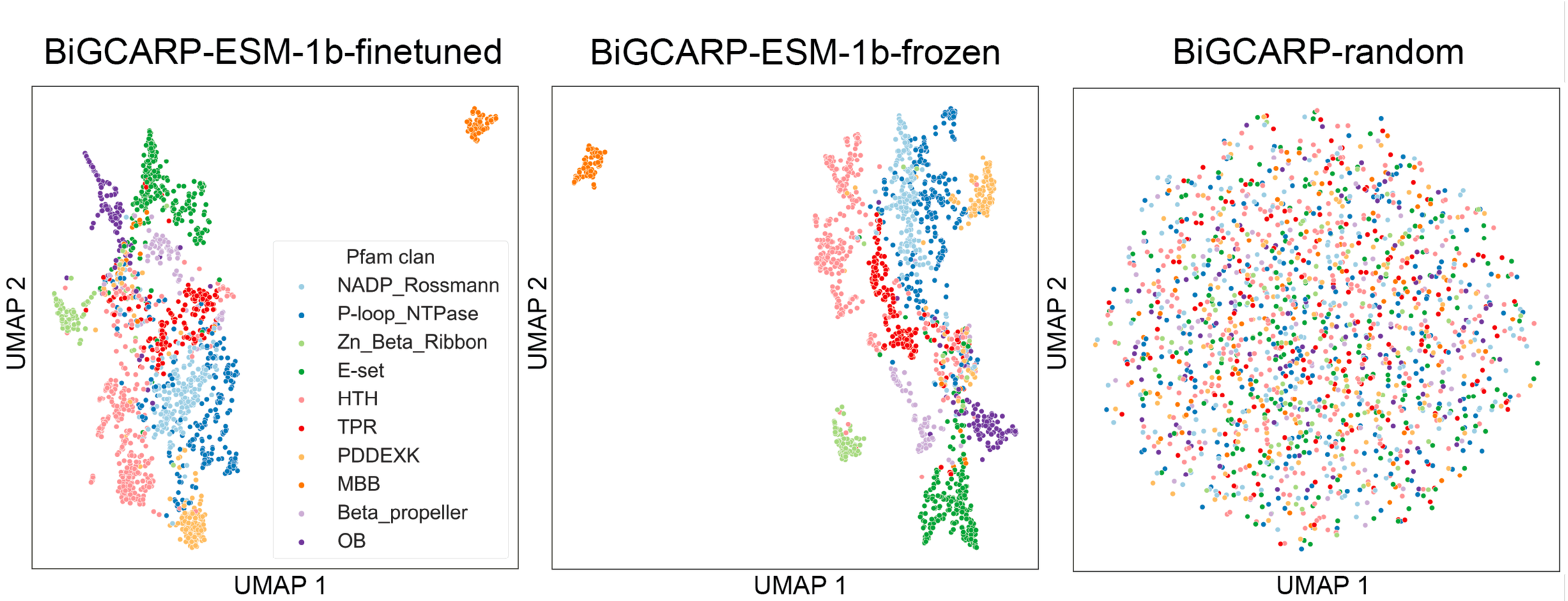
# BiGCARP embeddings encode relevant Pfam information



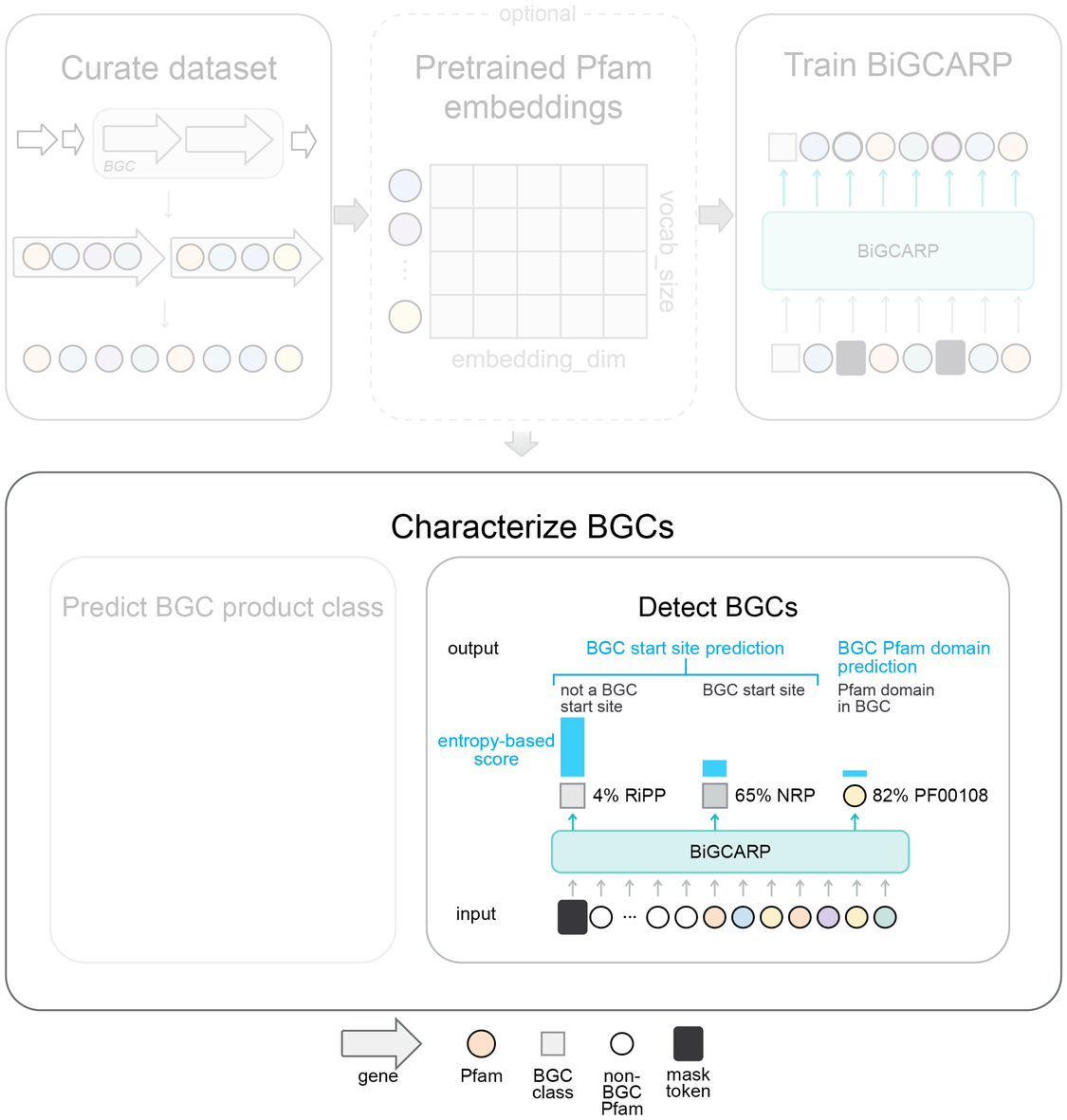
# BiGCARP embeddings encode relevant Pfam information



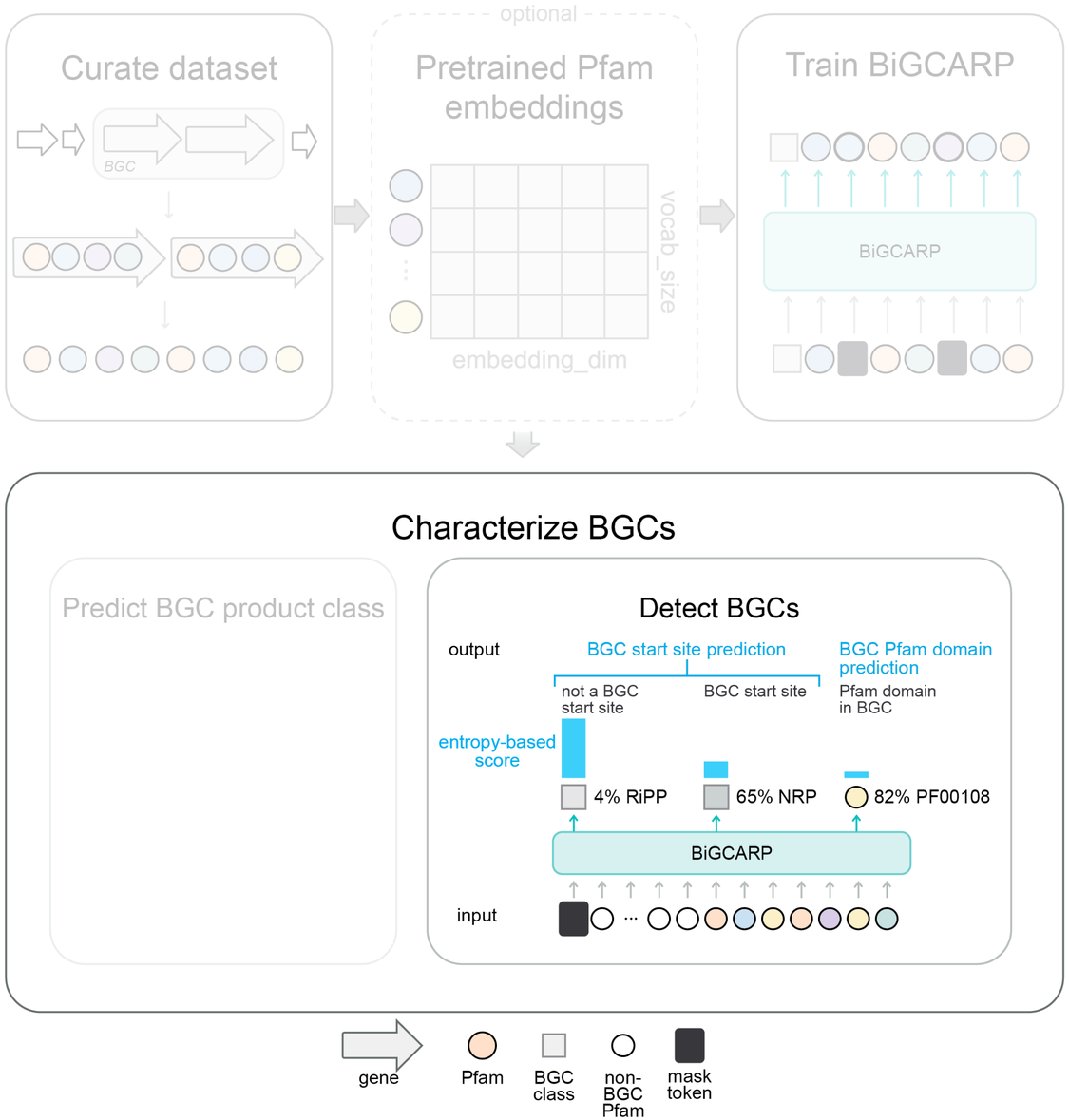
# BiGCARP embeddings encode relevant Pfam information



# BiGCARP is an unsupervised BGC detector



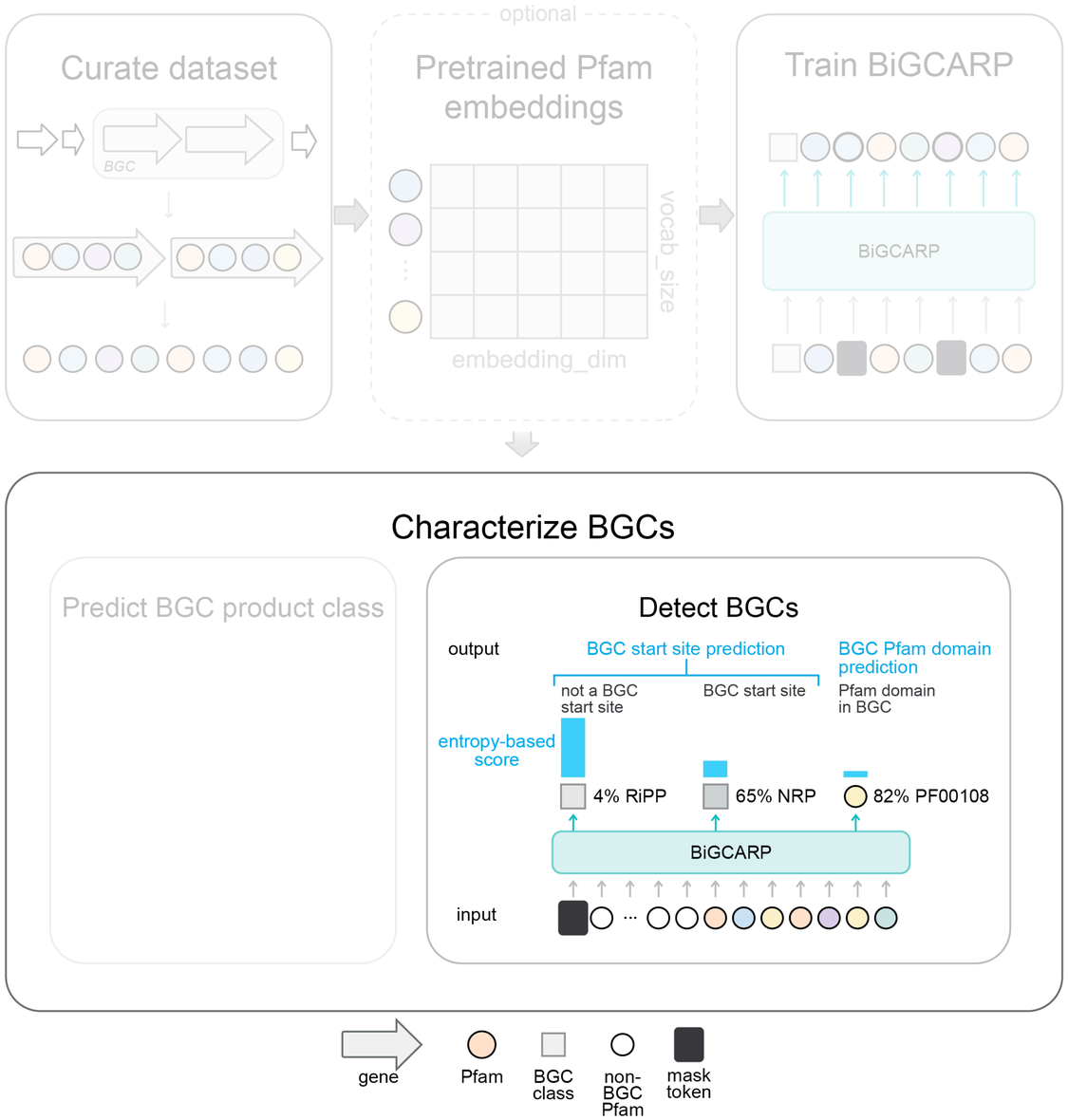
# BiGCARP is an unsupervised BGC detector




9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector

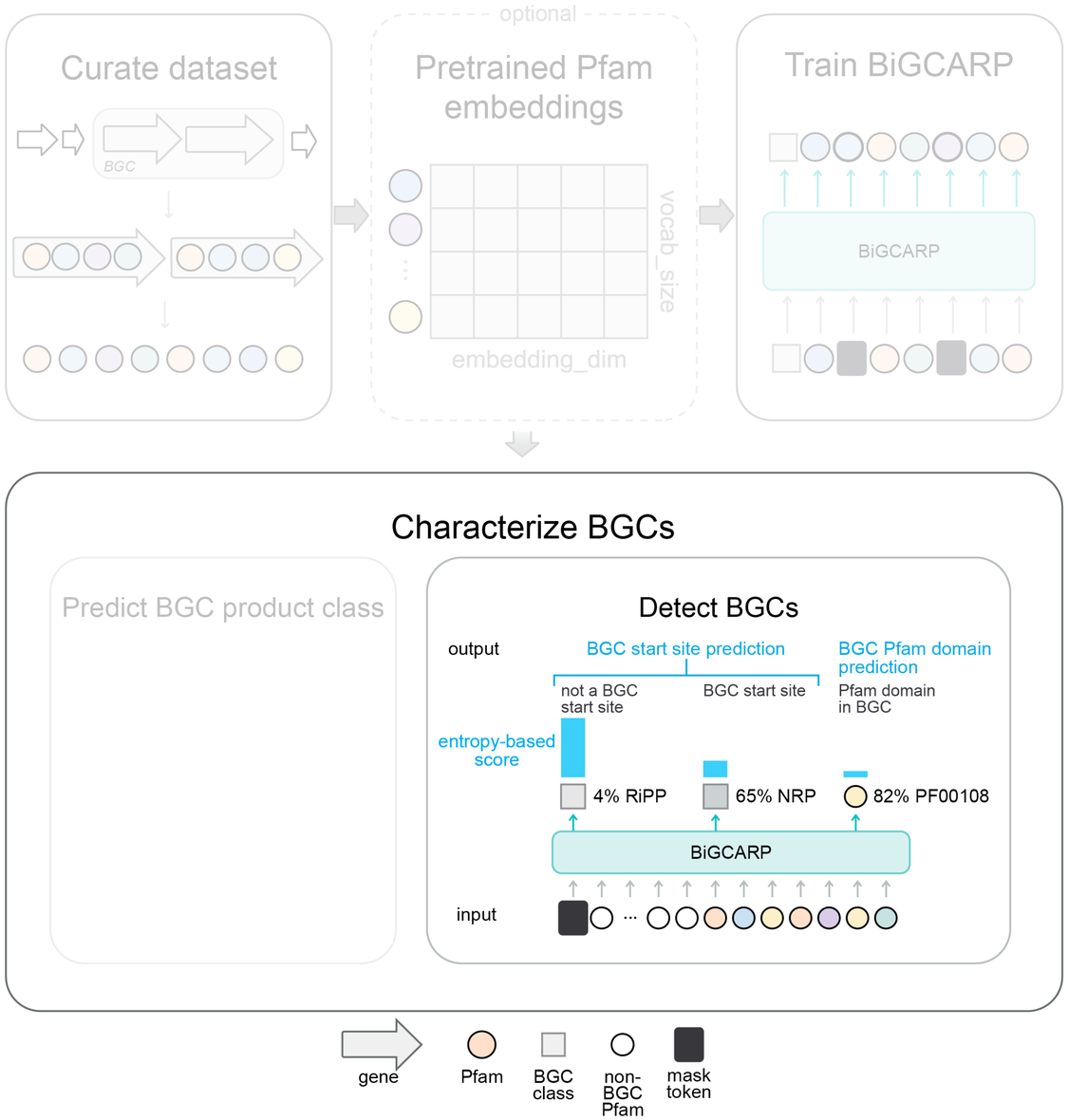



All windows of 64 domains

9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector



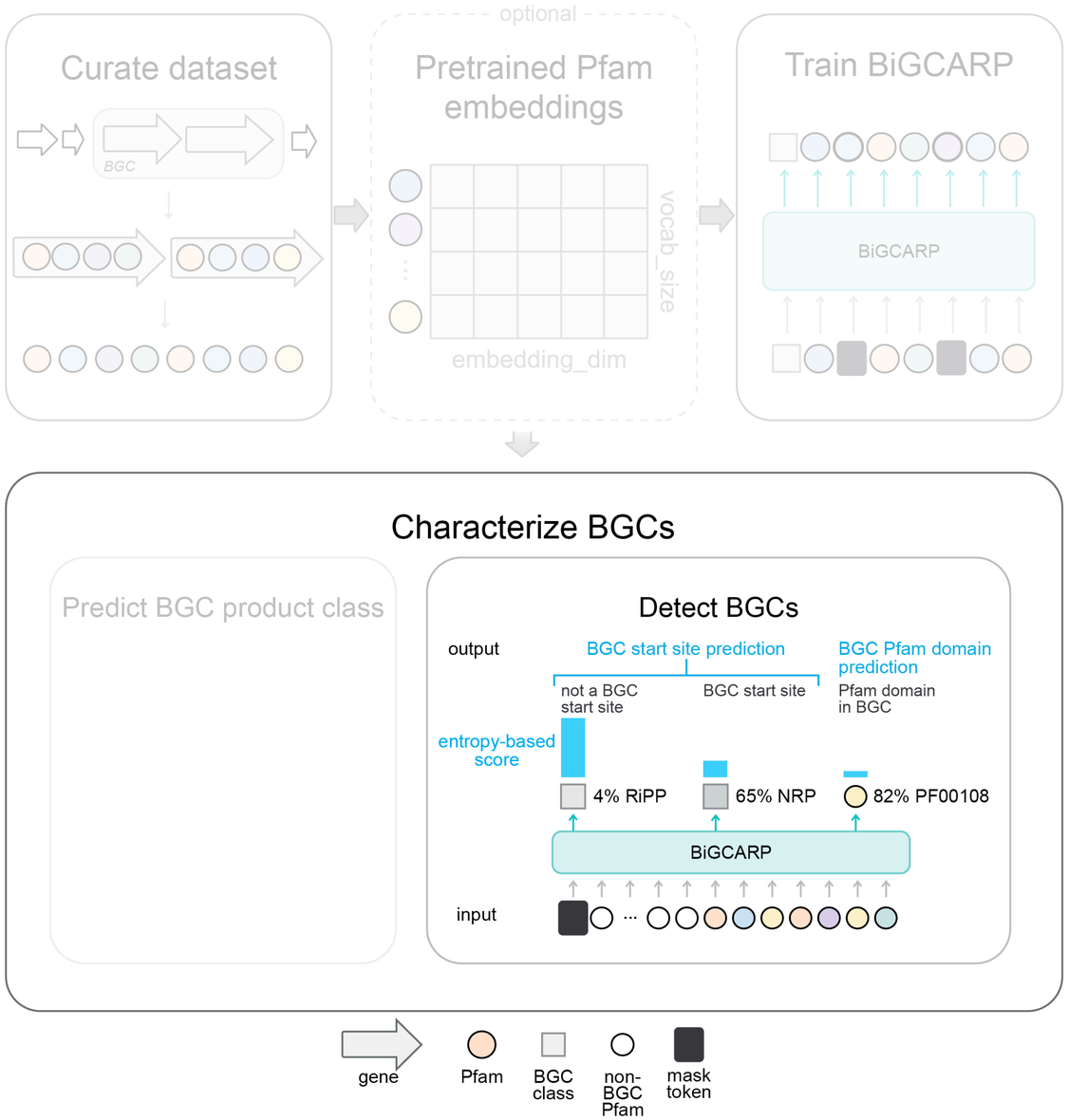

All windows of 64 domains

Append mask

9 bacterial genomes

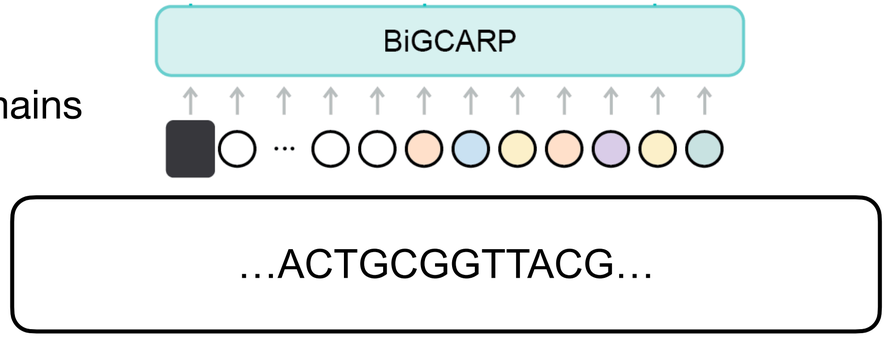
…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector

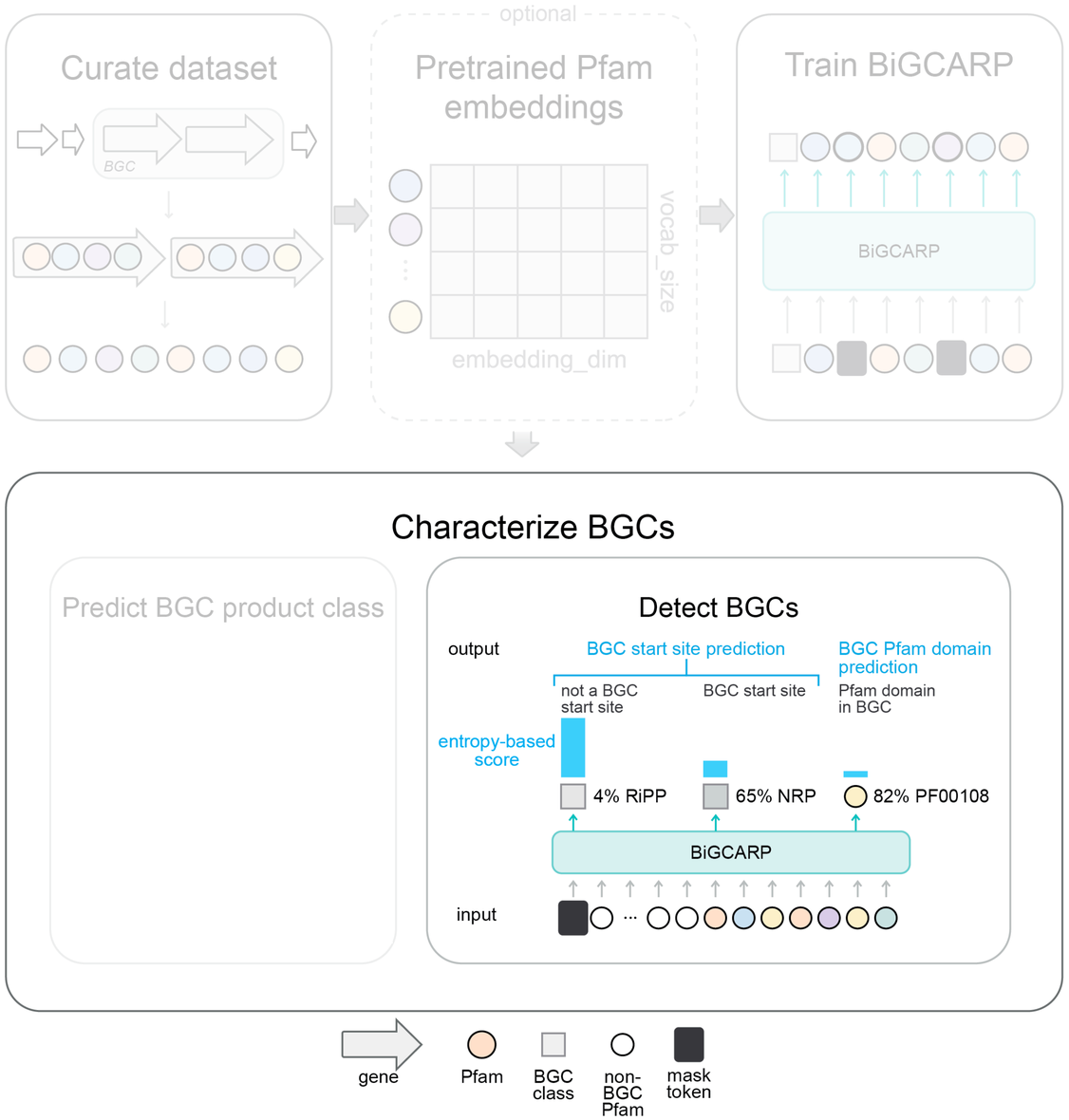


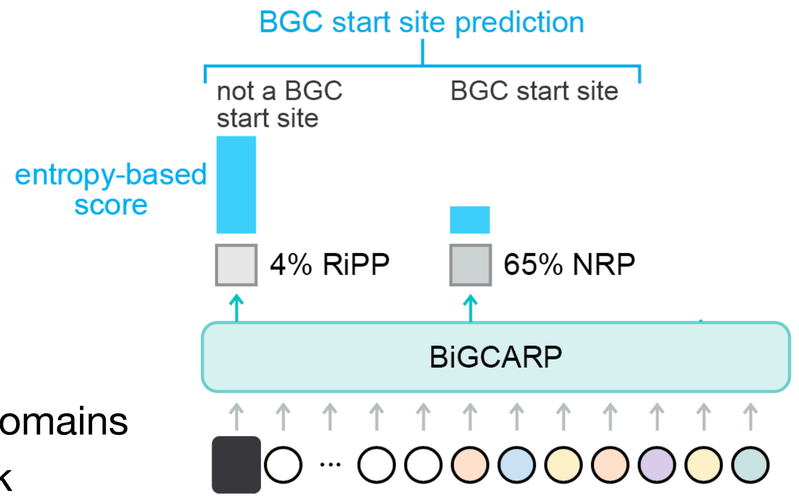

All windows of 64 domains
Append mask

9 bacterial genomes



# BiGCARP is an unsupervised BGC detector



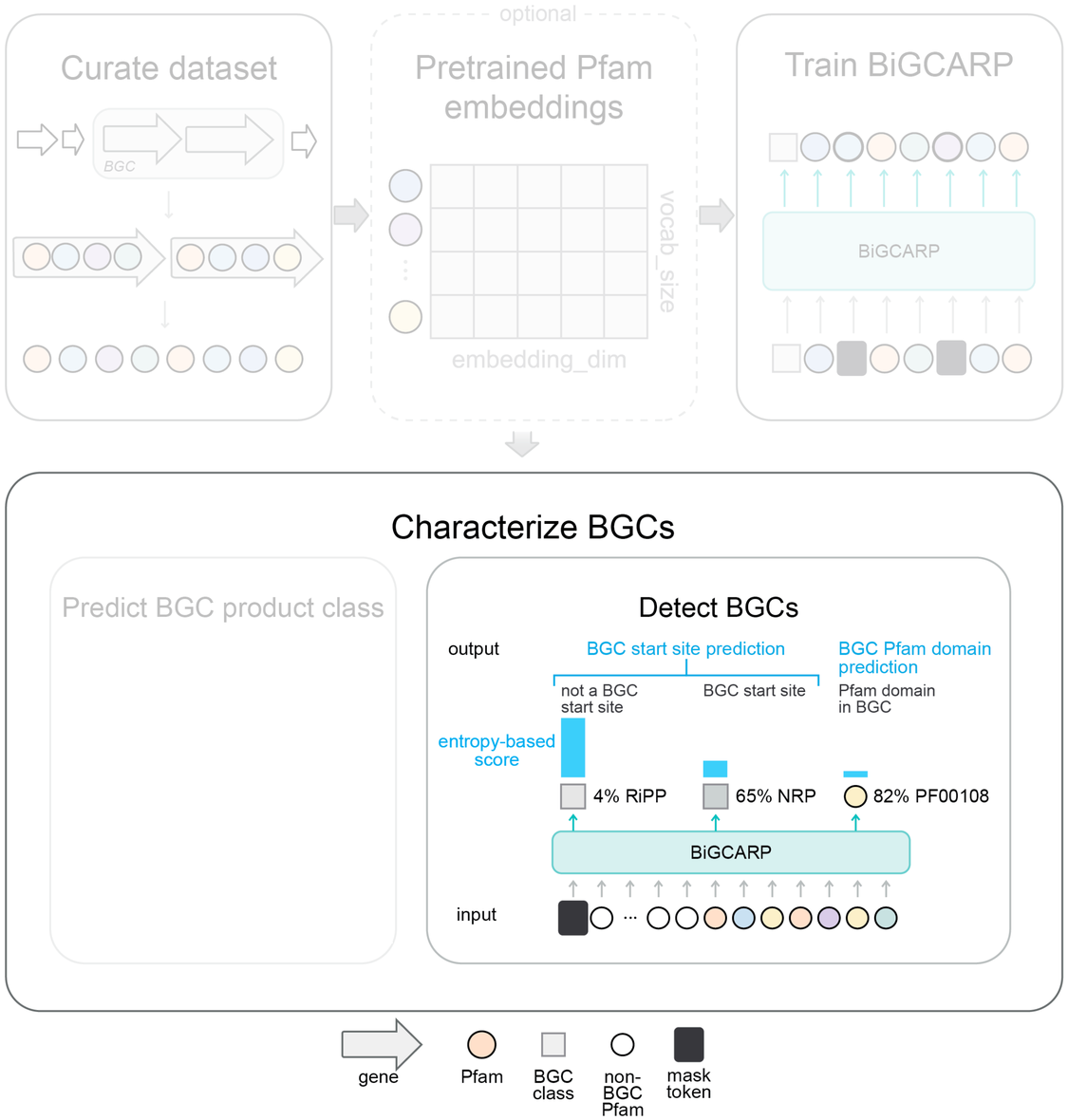

All windows of 64 domains

Append mask

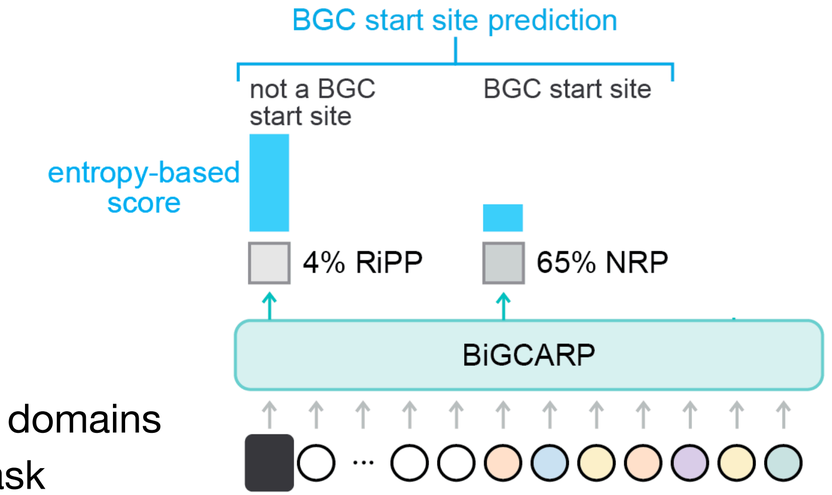
9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector




| | BiGCARP | | |
|---|---|---|---|
| | ESM-1b-finetuned | ESM-1b-frozen | random |
| 9-genomes (AUROC) | | | |



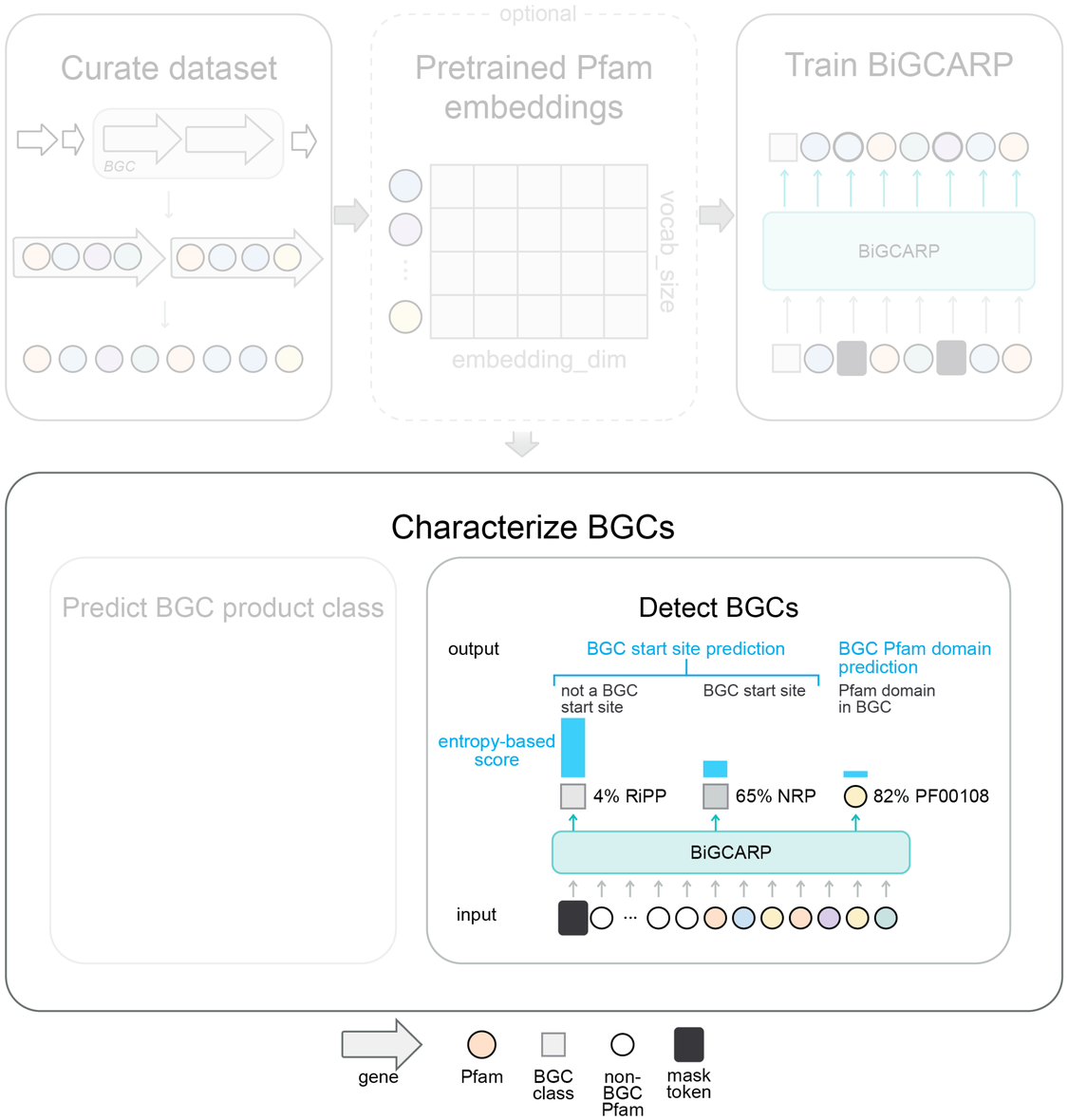

All windows of 64 domains

Append mask

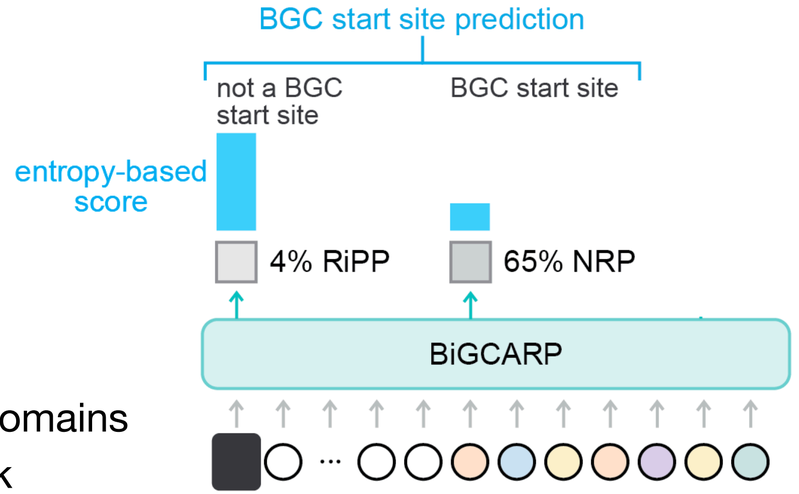
9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector




| | | BiGCARP | | |
|---|---|---|---|---|
| | | ESM-1b-finetuned | ESM-1b-frozen | random |
| 9-genomes (AUROC) | start | 0.720 | 0.701 | **0.723** |



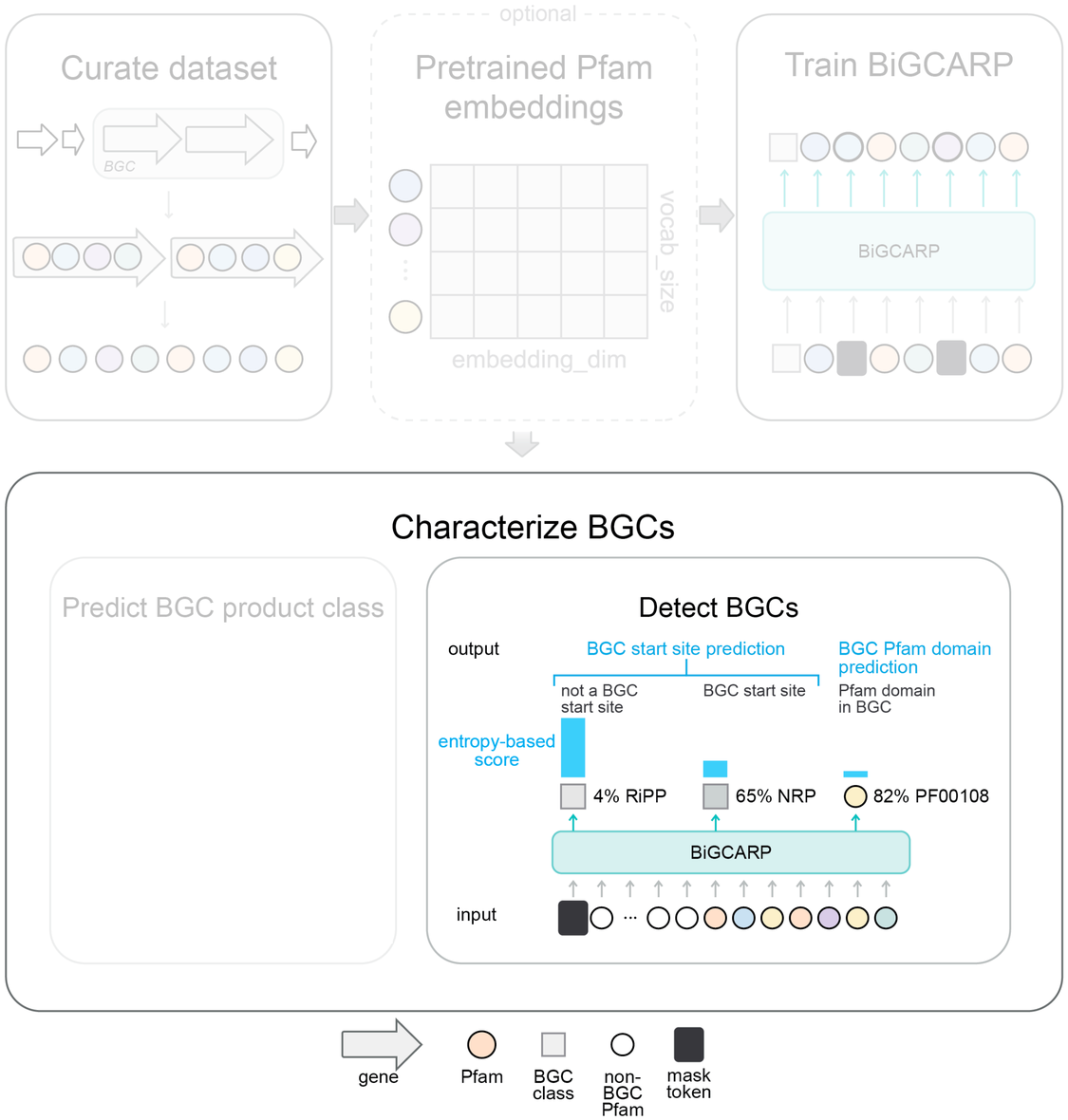

All windows of 64 domains

Append mask

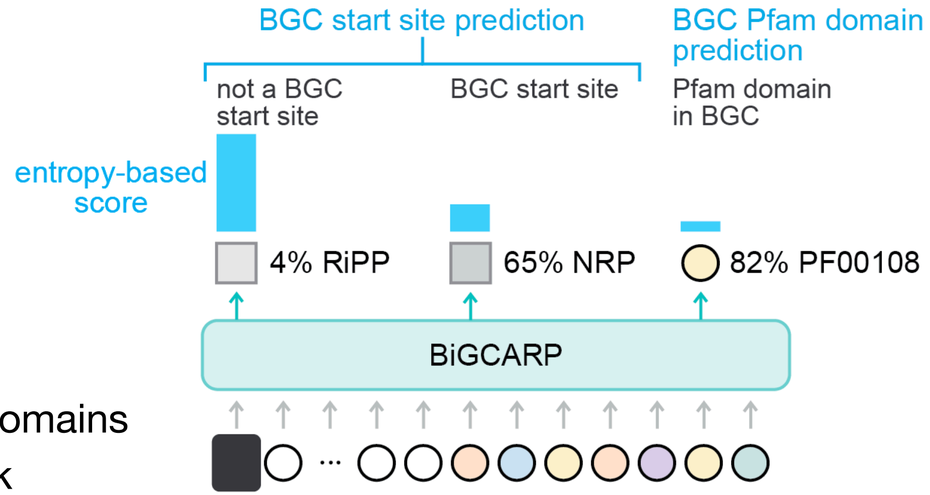
9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector




| | | BiGCARP | | |
|---|---|---|---|---|
| | | ESM-1b-finetuned | ESM-1b-frozen | random |
| 9-genomes (AUROC) | start | 0.720 | 0.701 | **0.723** |



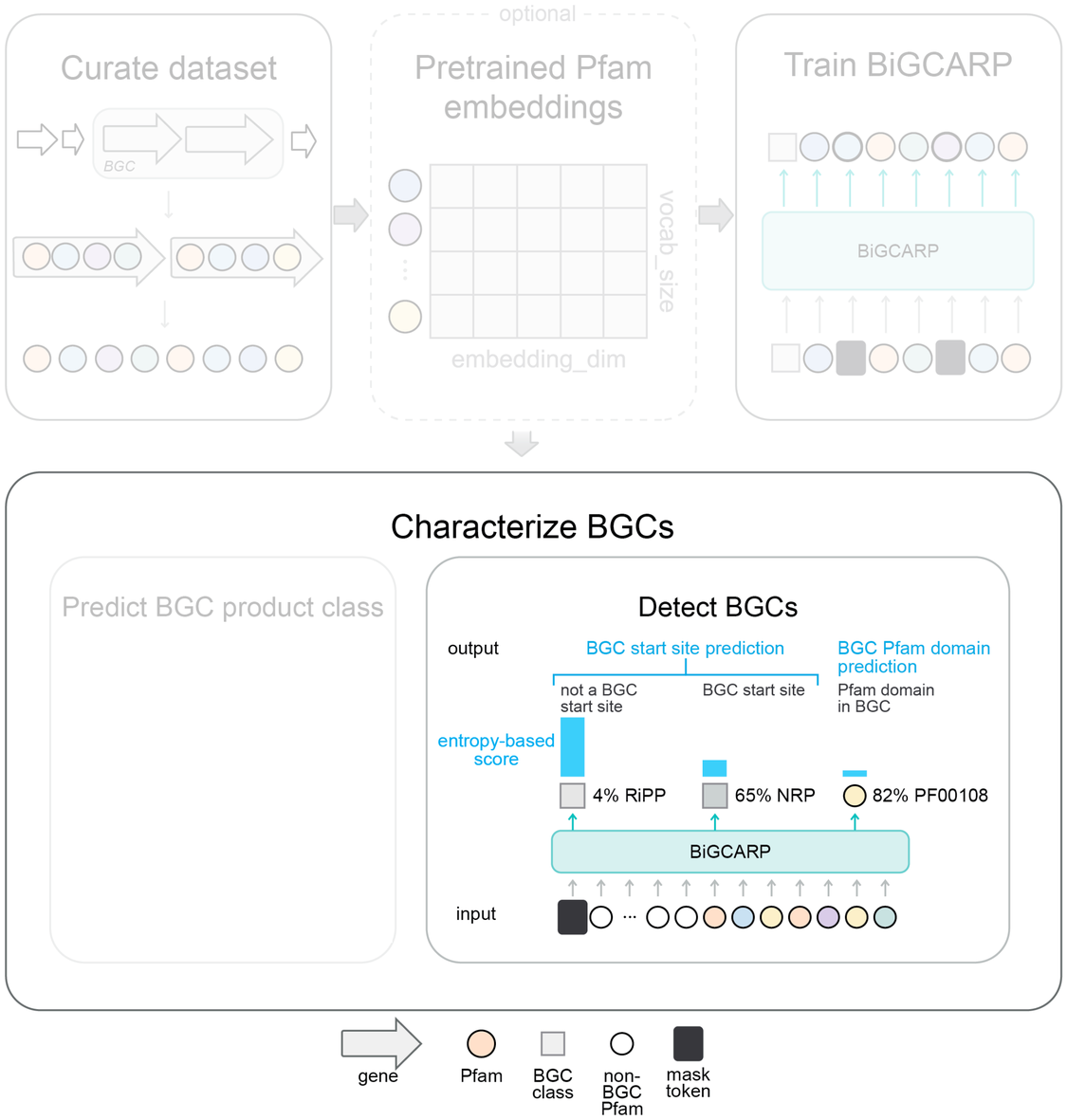

All windows of 64 domains

Append mask

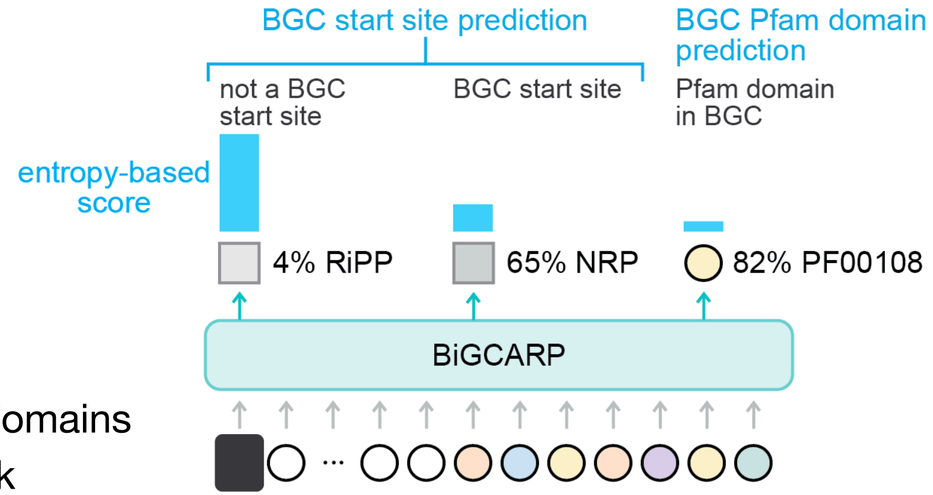
9 bacterial genomes

…ACTGCGGTTACG…

# BiGCARP is an unsupervised BGC detector




| | | BiGCARP | | |
|---|---|---|---|---|
| | | ESM-1b-finetuned | ESM-1b-frozen | random |
| 9-genomes (AUROC) | start | 0.720 | 0.701 | **0.723** |
| | domain | **0.876** | 0.611 | 0.856 |




All windows of 64 domains

Append mask

9 bacterial genomes

…ACTGCGGTTACG…

# Adding negative examples improves detection

# Adding negative examples improves detection

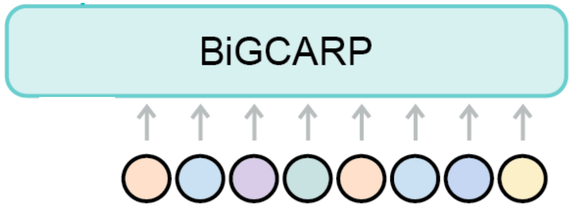
1406 BGCs from MIBiG

+ 10128 negative clusters

# Adding negative examples improves detection

1406 BGCs from MIBiG

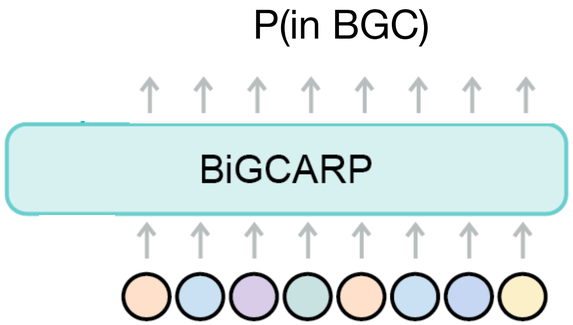
+ 10128 negative clusters



# Adding negative examples improves detection

1406 BGCs from MIBiG

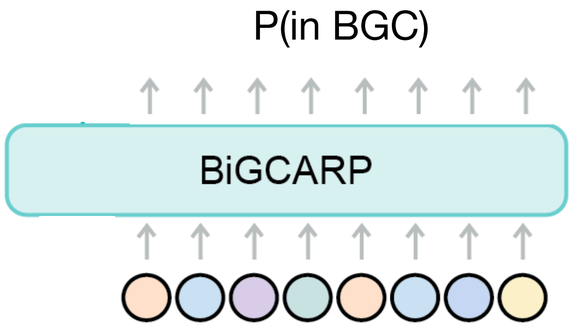
+ 10128 negative clusters



# Adding negative examples improves detection

1406 BGCs from MIBiG

+ 10128 negative clusters

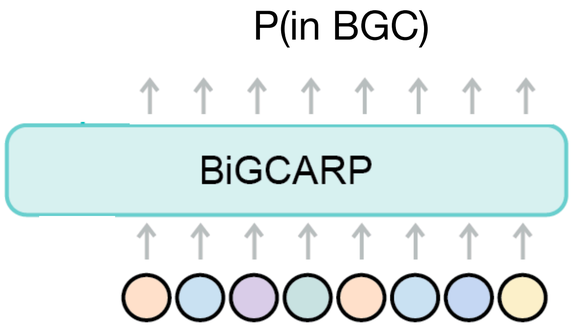



| pretraining |
| --- |
| BiGCARP-ESM-1b-finetuned |
| BiGCARP-ESM-1b-frozen |
| BiGCARP-random |
| none |
| DeepBGC |

# Adding negative examples improves detection

1406 BGCs from MIBiG

+ 10128 negative clusters

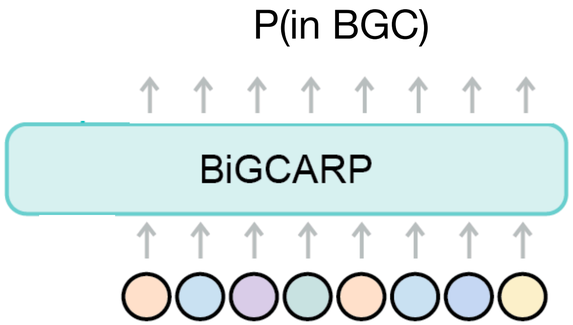



| pretraining |
| --- |
| BiGCARP-ESM-1b-finetuned |
| BiGCARP-ESM-1b-frozen |
| BiGCARP-random |
| none |
| DeepBGC (supervised LSTM) |

# Adding negative examples improves detection

1406 BGCs from MIBiG

+ 10128 negative clusters

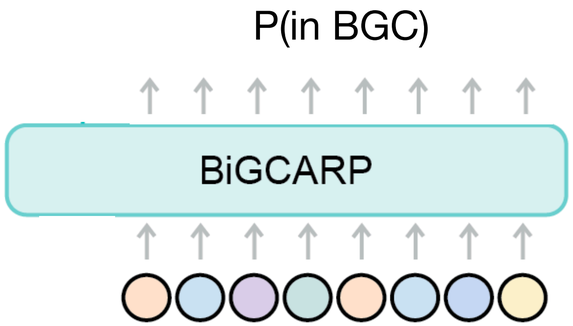



| pretraining | Validation |
|---|---|
| BiGCARP-ESM-1b-finetuned | **0.941** |
| BiGCARP-ESM-1b-frozen | 0.940 |
| BiGCARP-random | 0.936 |
| none | 0.937 |
| DeepBGC (supervised LSTM) | 0.934 |

# Adding negative examples improves detection

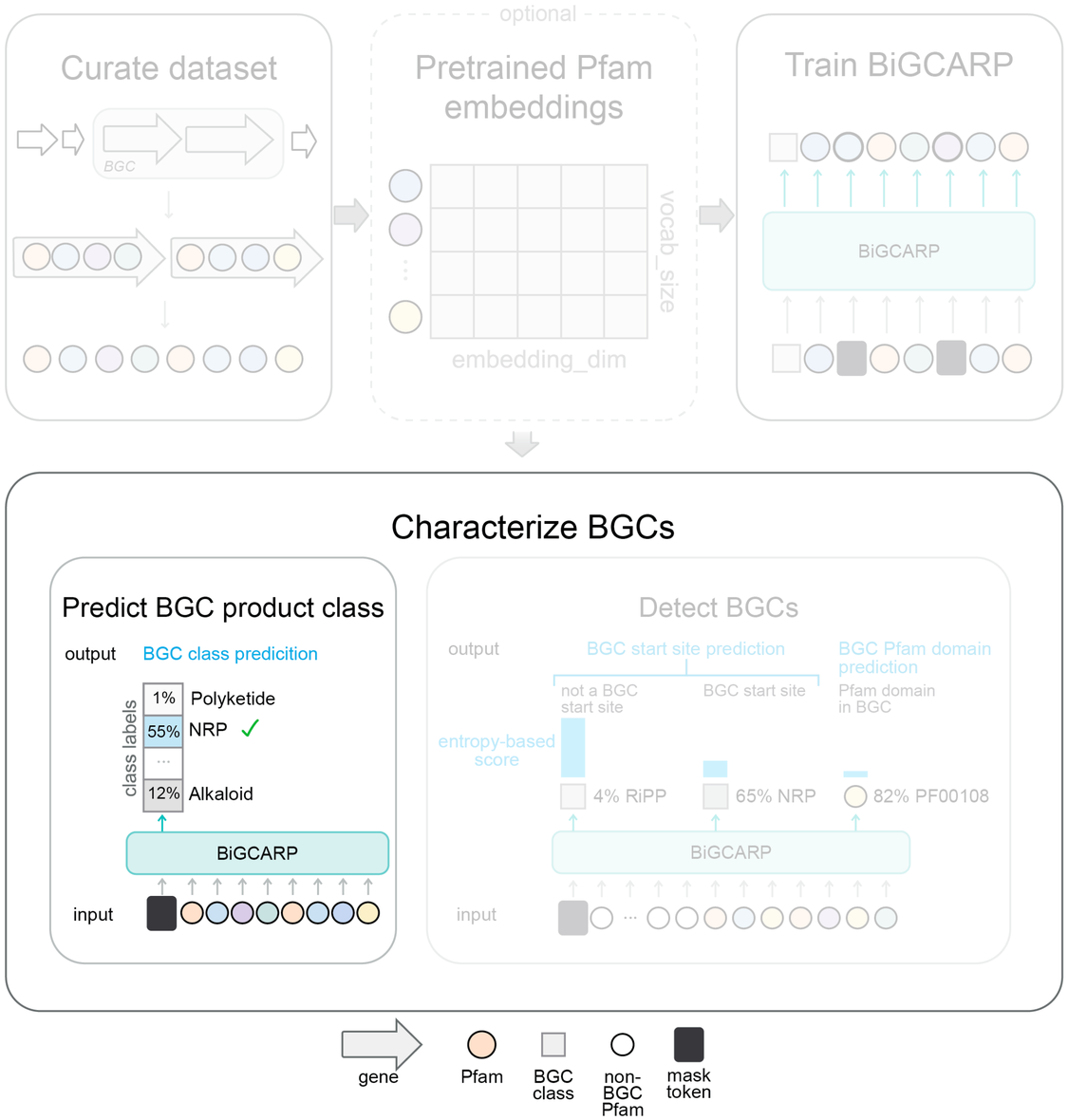
1406 BGCs from MIBiG

+ 10128 negative clusters

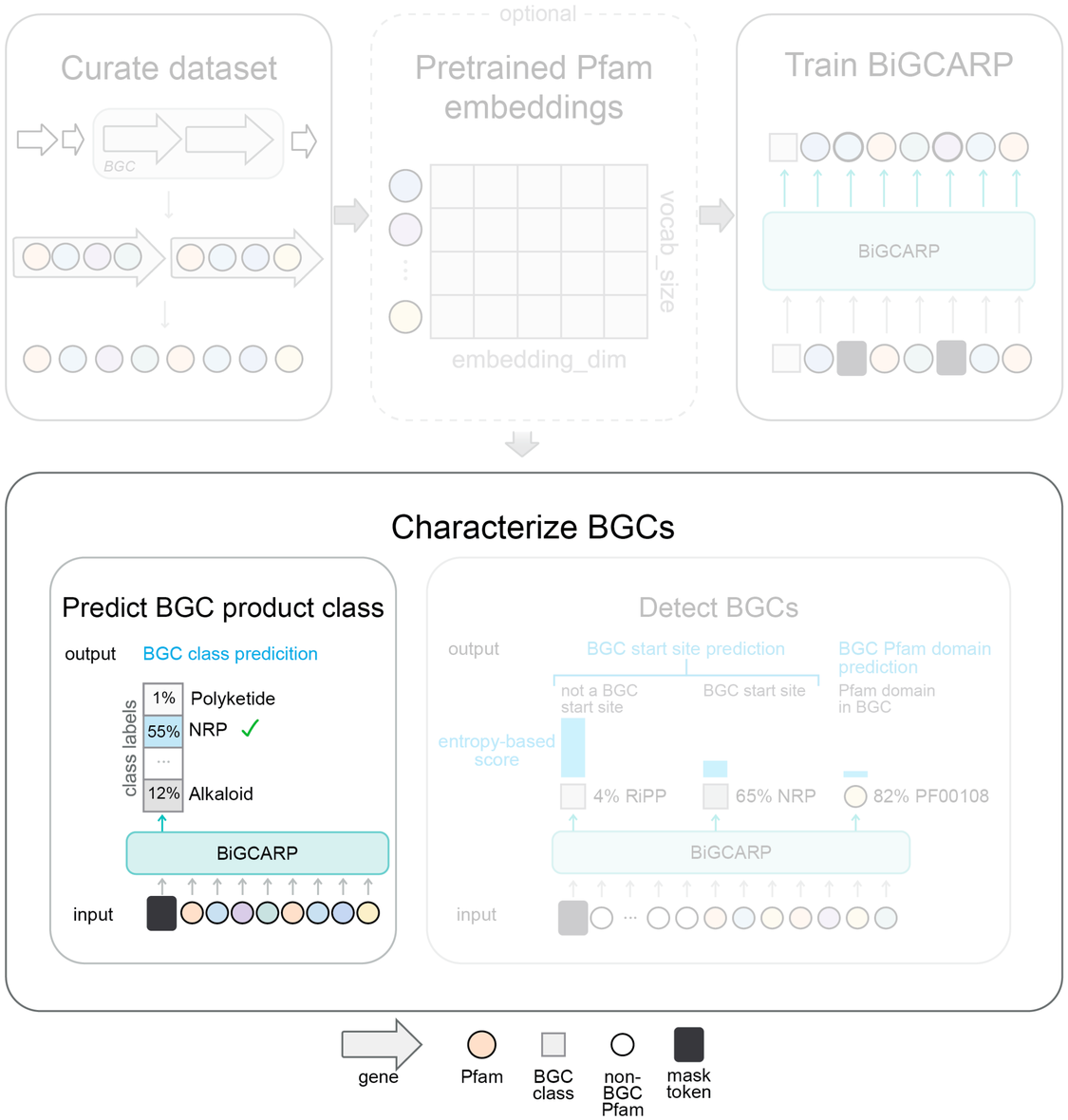



| pretraining | Test | Validation |
|---|---|---|
| BiGCARP-ESM-1b-finetuned | **0.950** | **0.941** |
| BiGCARP-ESM-1b-frozen | 0.946 | 0.940 |
| BiGCARP-random | 0.943 | 0.936 |
| none | **0.950** | 0.937 |
| DeepBGC (supervised LSTM) | 0.921 | 0.934 |

# BiGCARP predicts BGC product classes

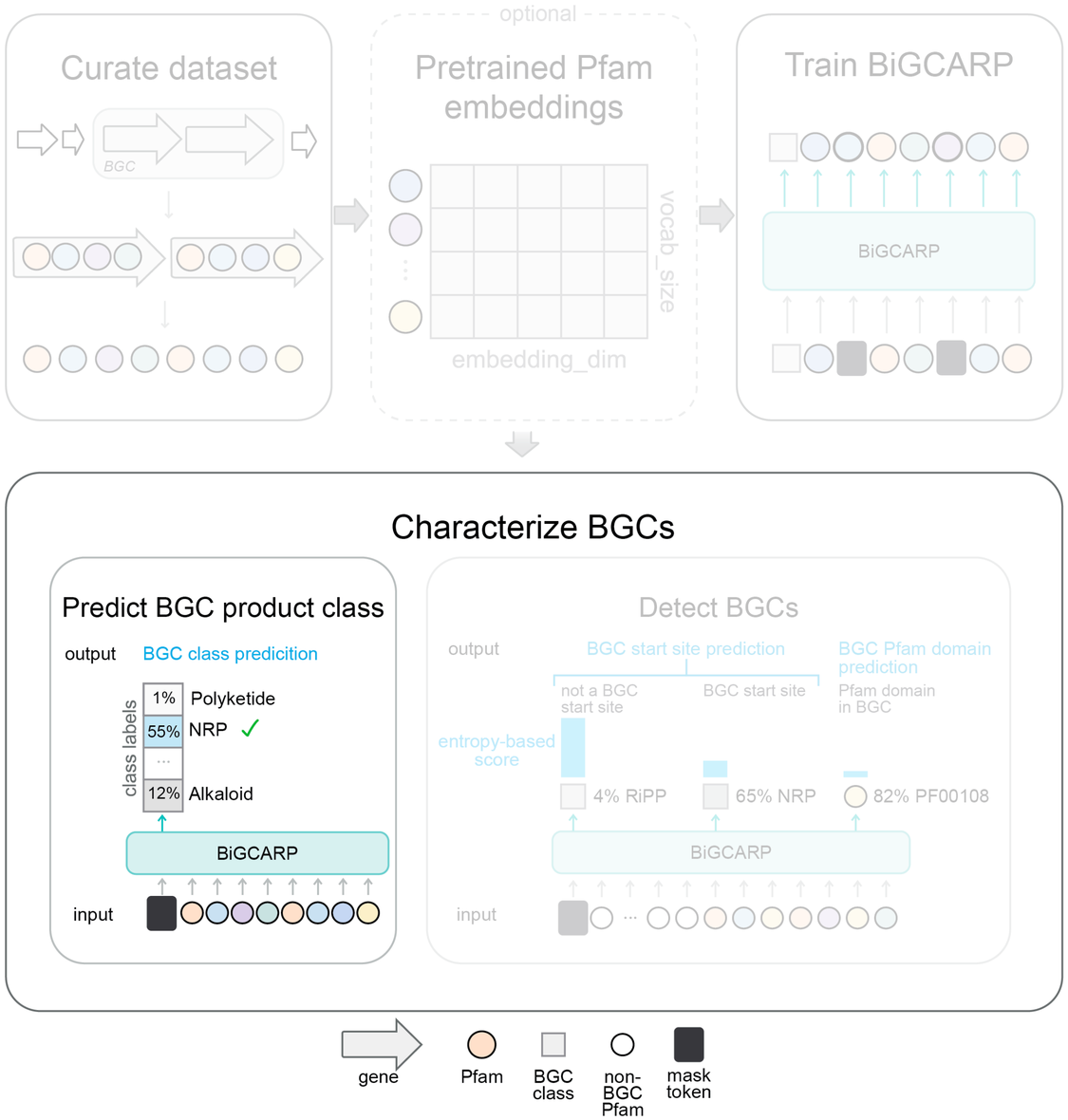


# BiGCARP predicts BGC product classes




1406 BGCs from MIBiG

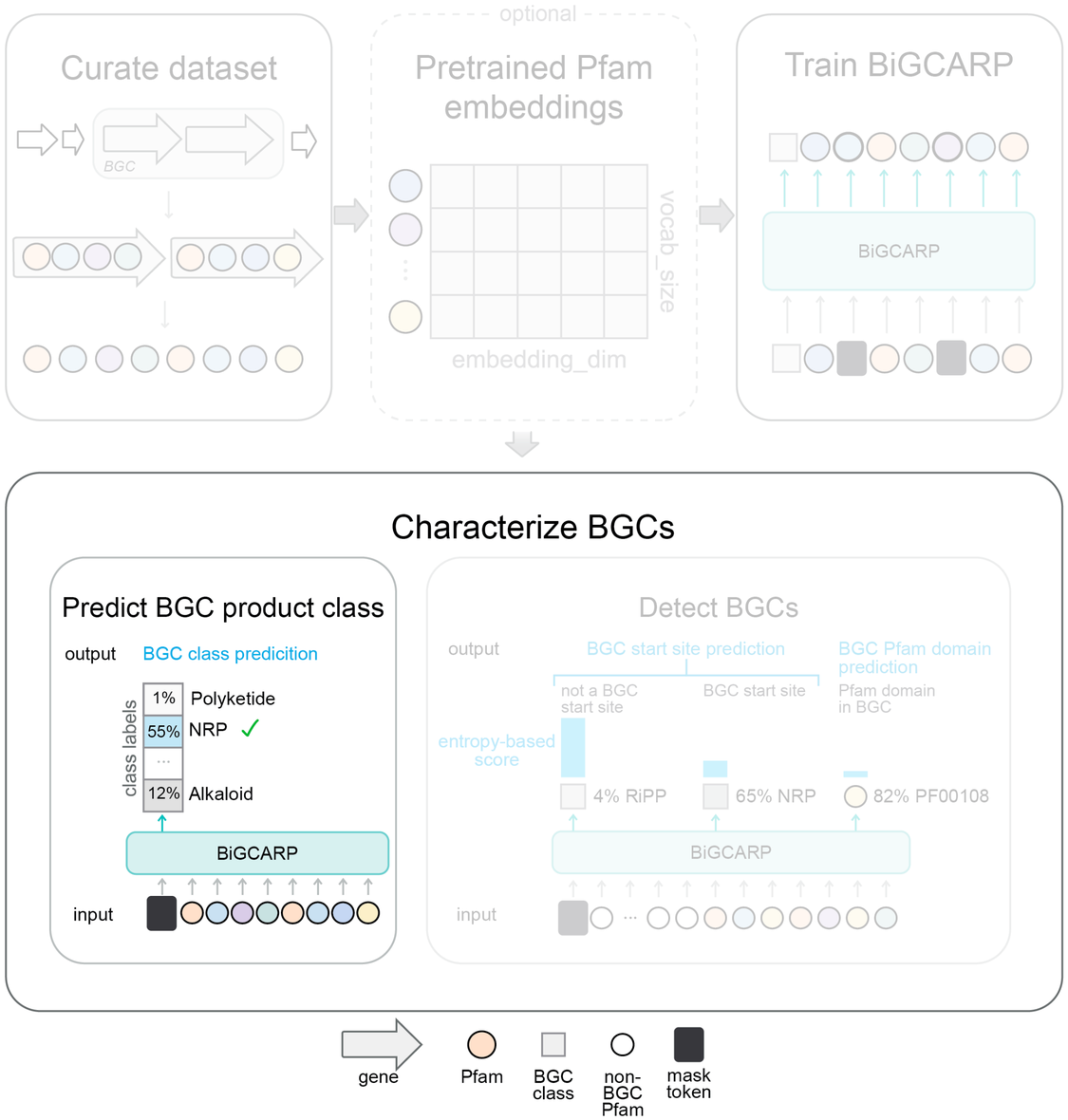
# BiGCARP predicts BGC product classes




1406 BGCs from MIBiG

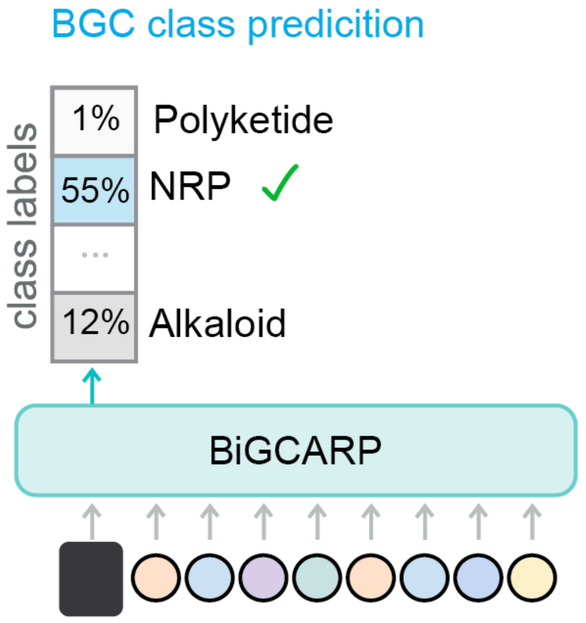
Append mask

# BiGCARP predicts BGC product classes

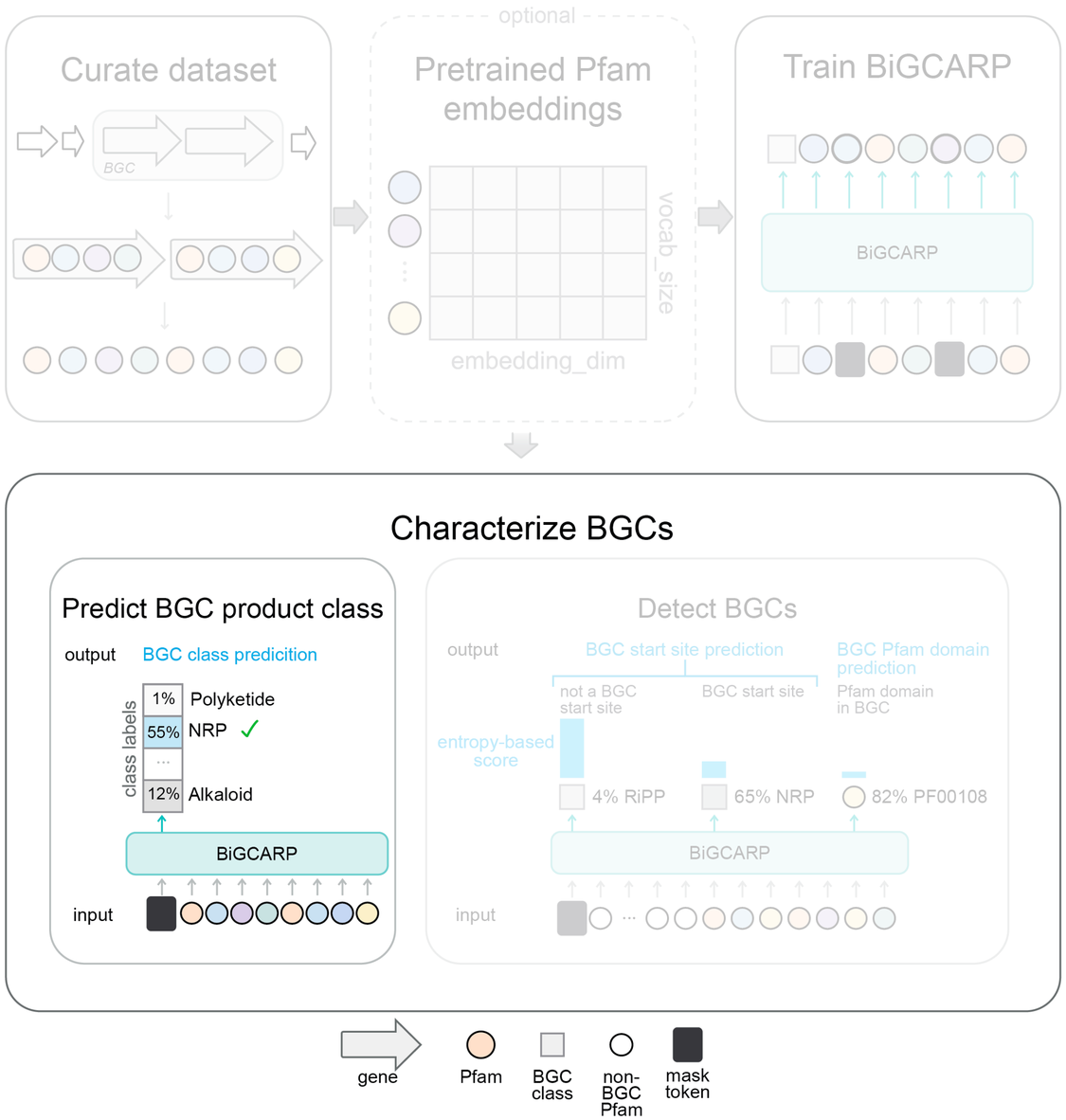



1406 BGCs from MIBiG

Append mask



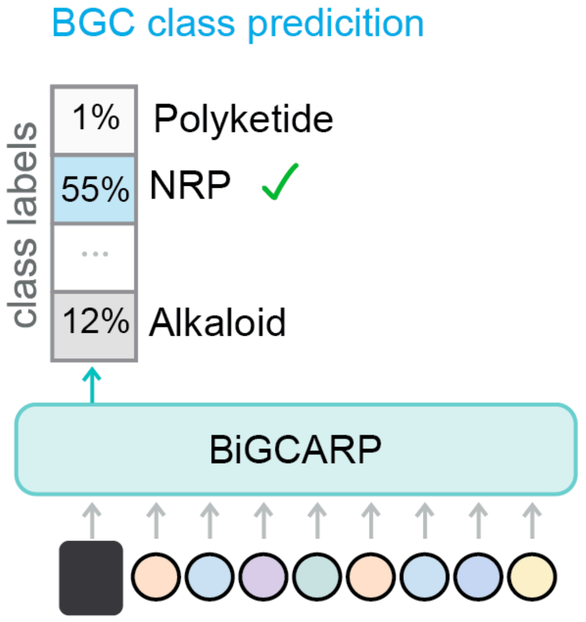
# BiGCARP predicts BGC product classes



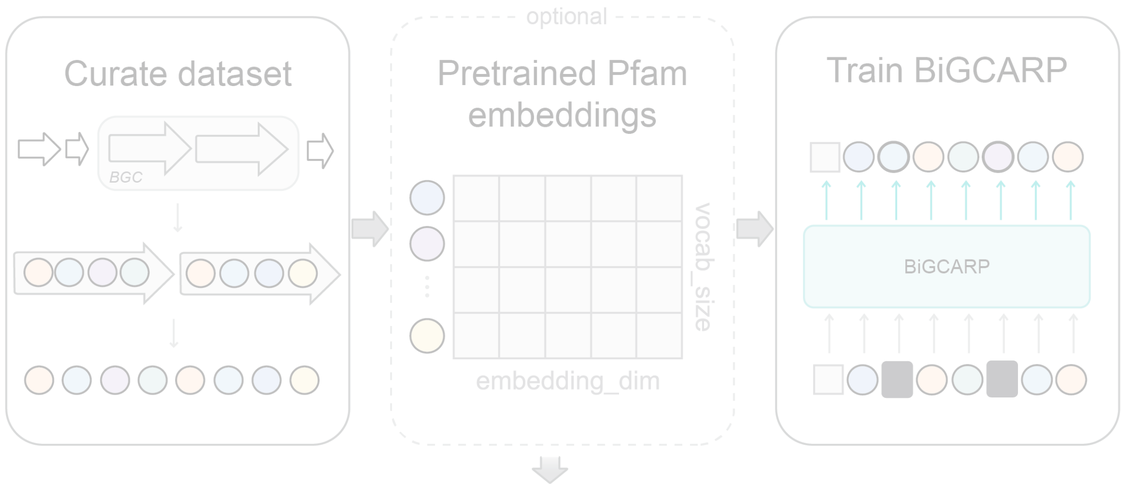

| | BiGCARP-ensemble | DeepBGC |
|---|---|---|
| polyketide | | |
| NRP | | |
| RiPP | | |
| saccharide | | |
| other | | |
| terpene | | |
| alkaloid | | |
| average | | |

1406 BGCs from MIBiG

Append mask

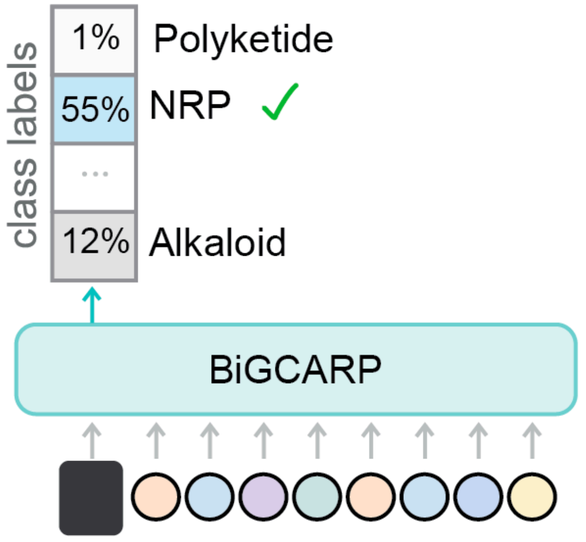


# BiGCARP predicts BGC product classes




| | BiGCARP-ensemble | DeepBGC |
|---|---|---|
| polyketide | | |
| NRP | | |
| RiPP | | |
| saccharide | | |
| other | | |
| terpene | | |
| alkaloid | | |
| average | **0.855** | 0.792 |

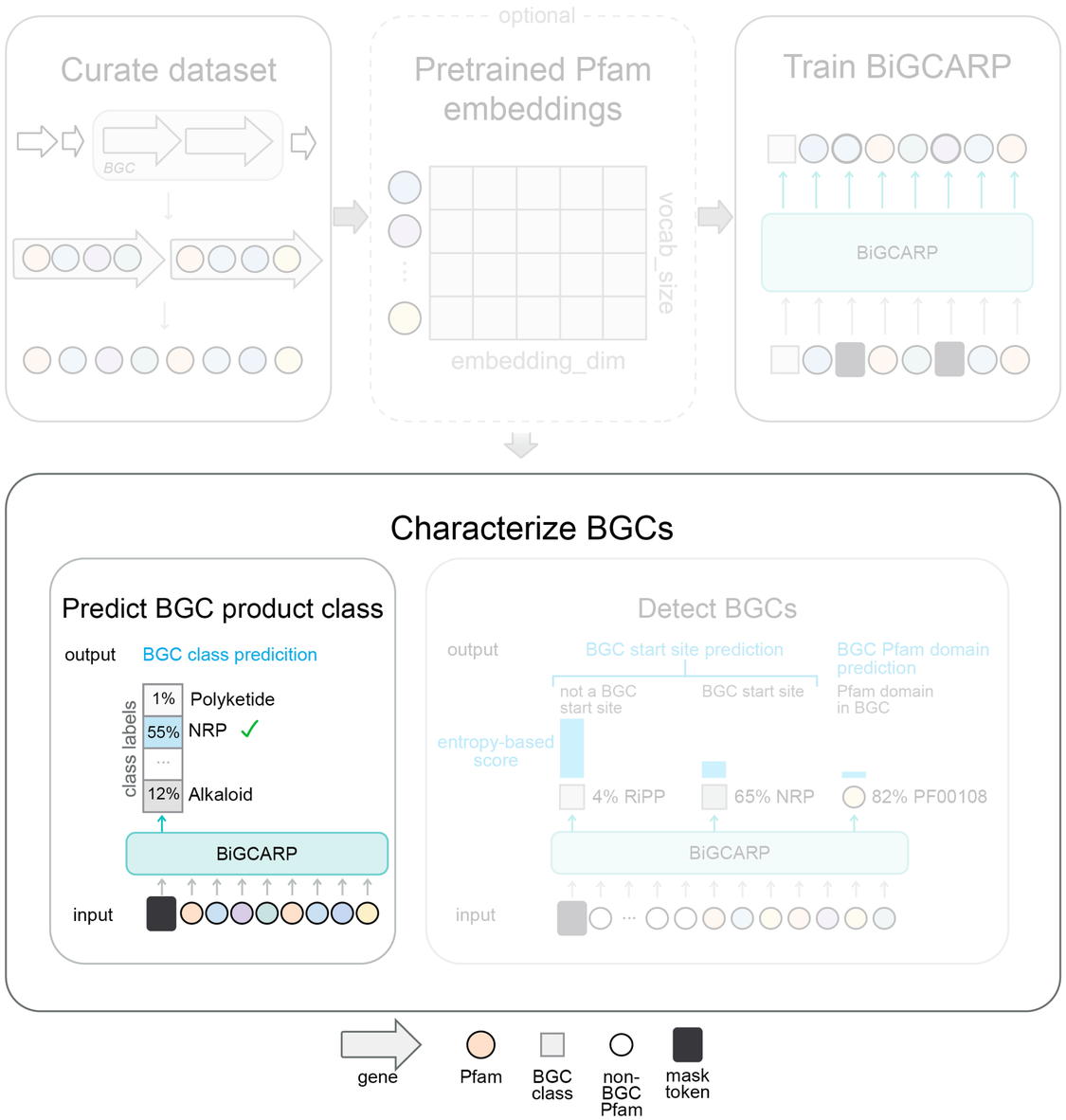
BGC class predicition




1406 BGCs from MIBiG

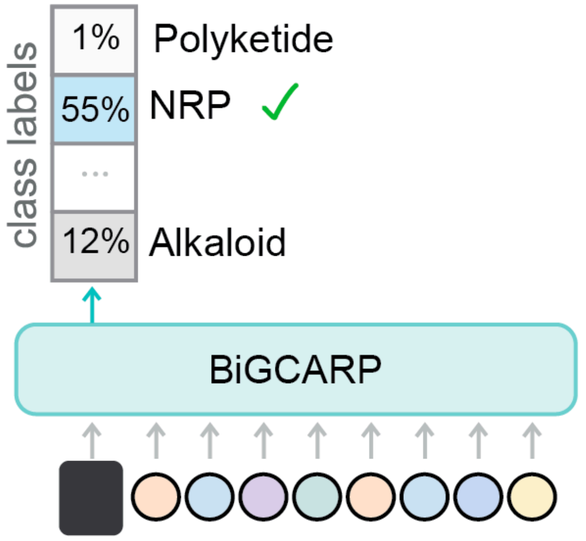
Append mask

# BiGCARP predicts BGC product classes




| | BiGCARP-ensemble | DeepBGC |
|---|---|---|
| polyketide | 0.898 | **0.903** |
| NRP | 0.898 | **0.907** |
| RiPP | **0.963** | 0.907 |
| saccharide | 0.773 | **0.811** |
| other | **0.763** | 0.583 |
| terpene | **0.869** | 0.824 |
| alkaloid | **0.820** | 0.607 |
| average | **0.855** | 0.792 |

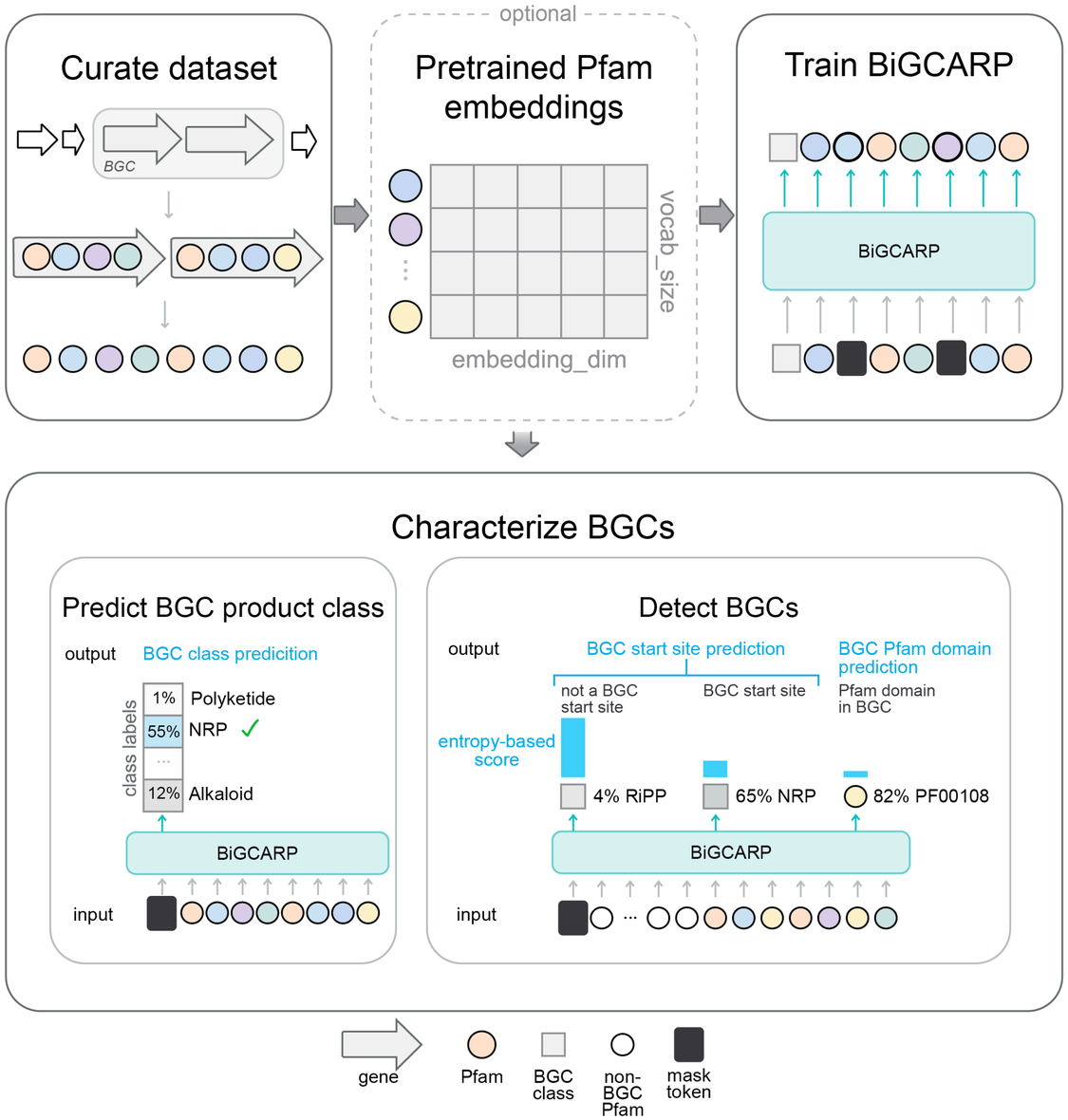
BGC class predicition



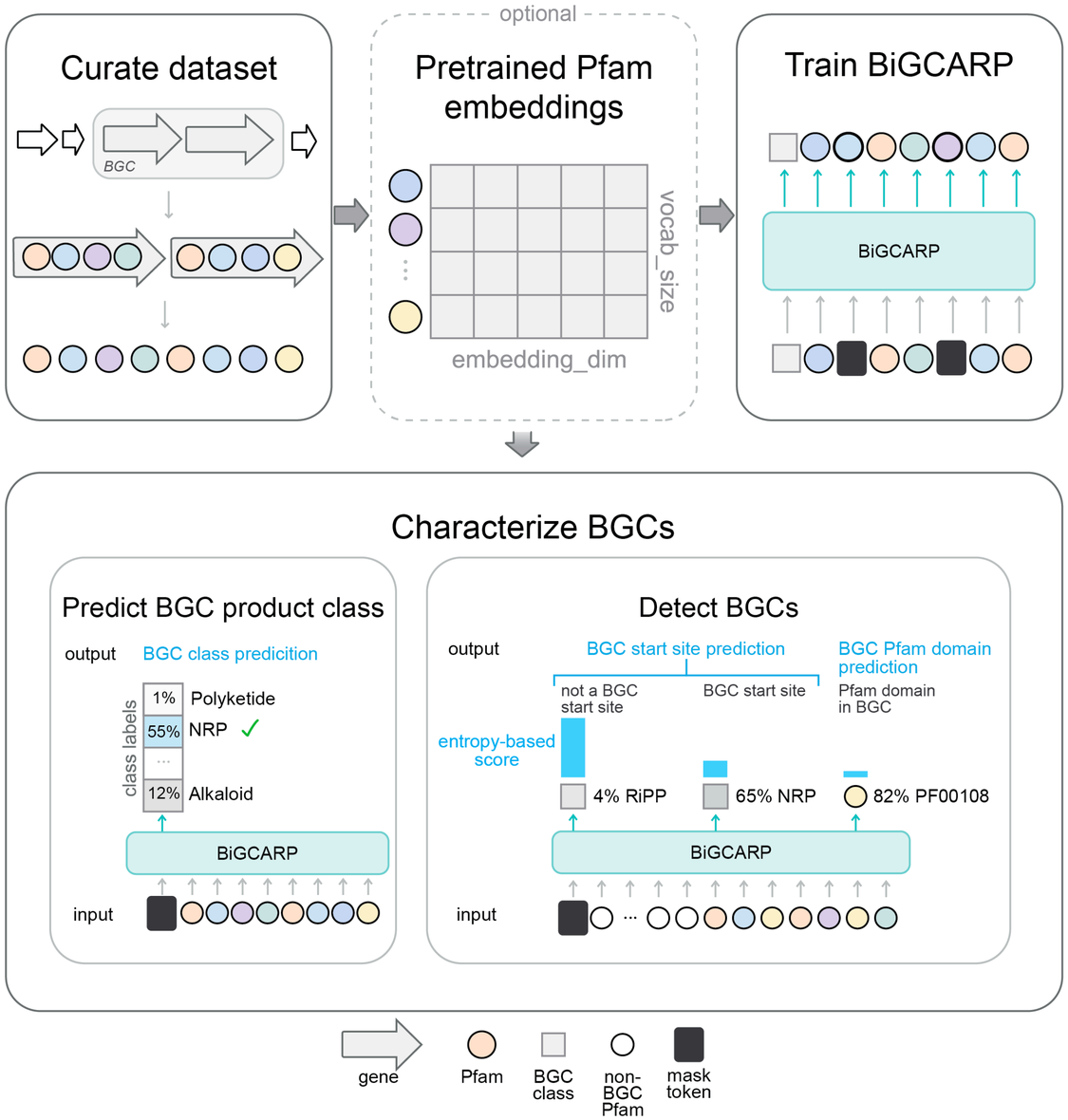

1406 BGCs from MIBiG

Append mask

# BiGCARP detects and classifies BGCs

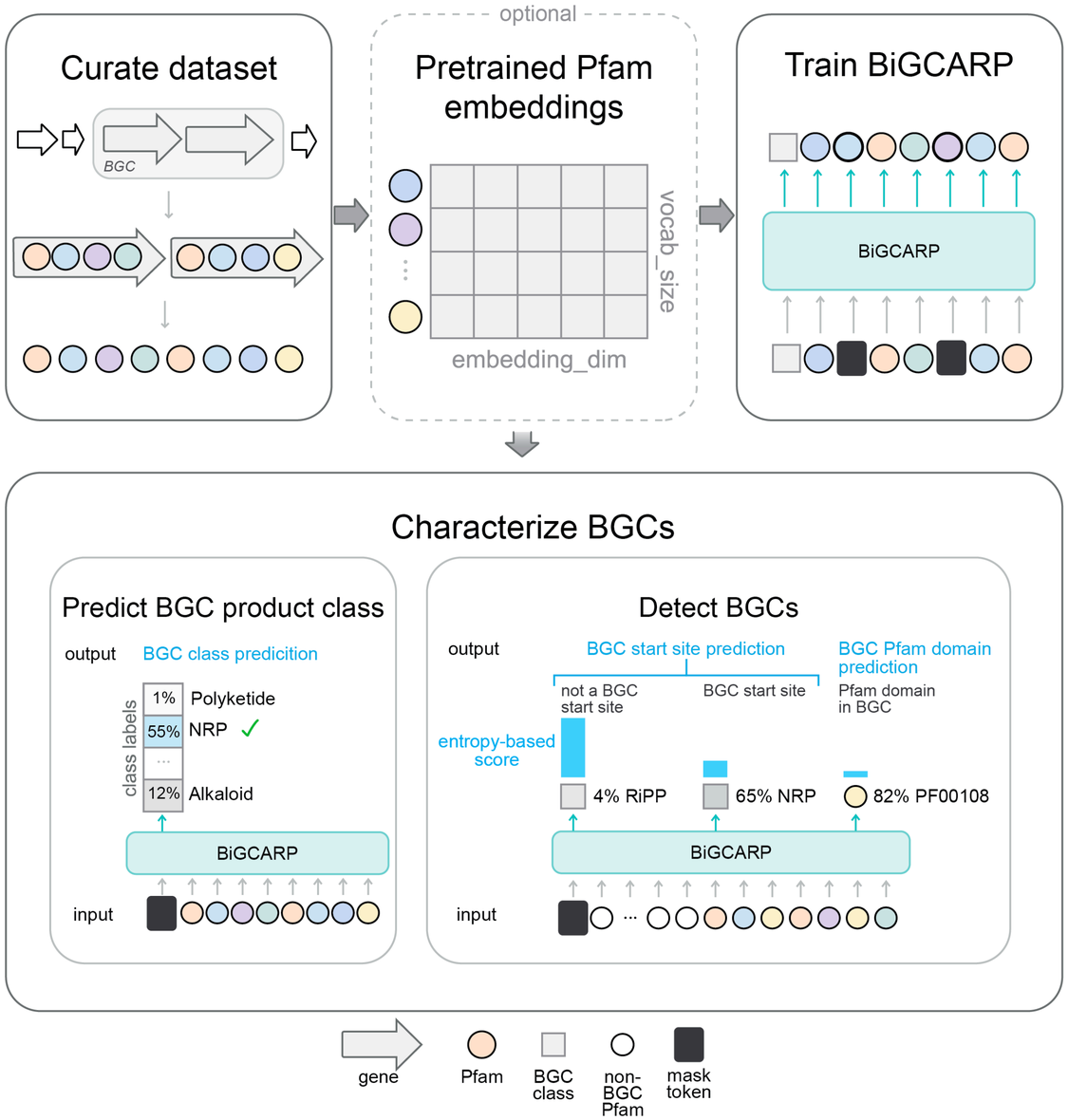


# BiGCARP detects and classifies BGCs



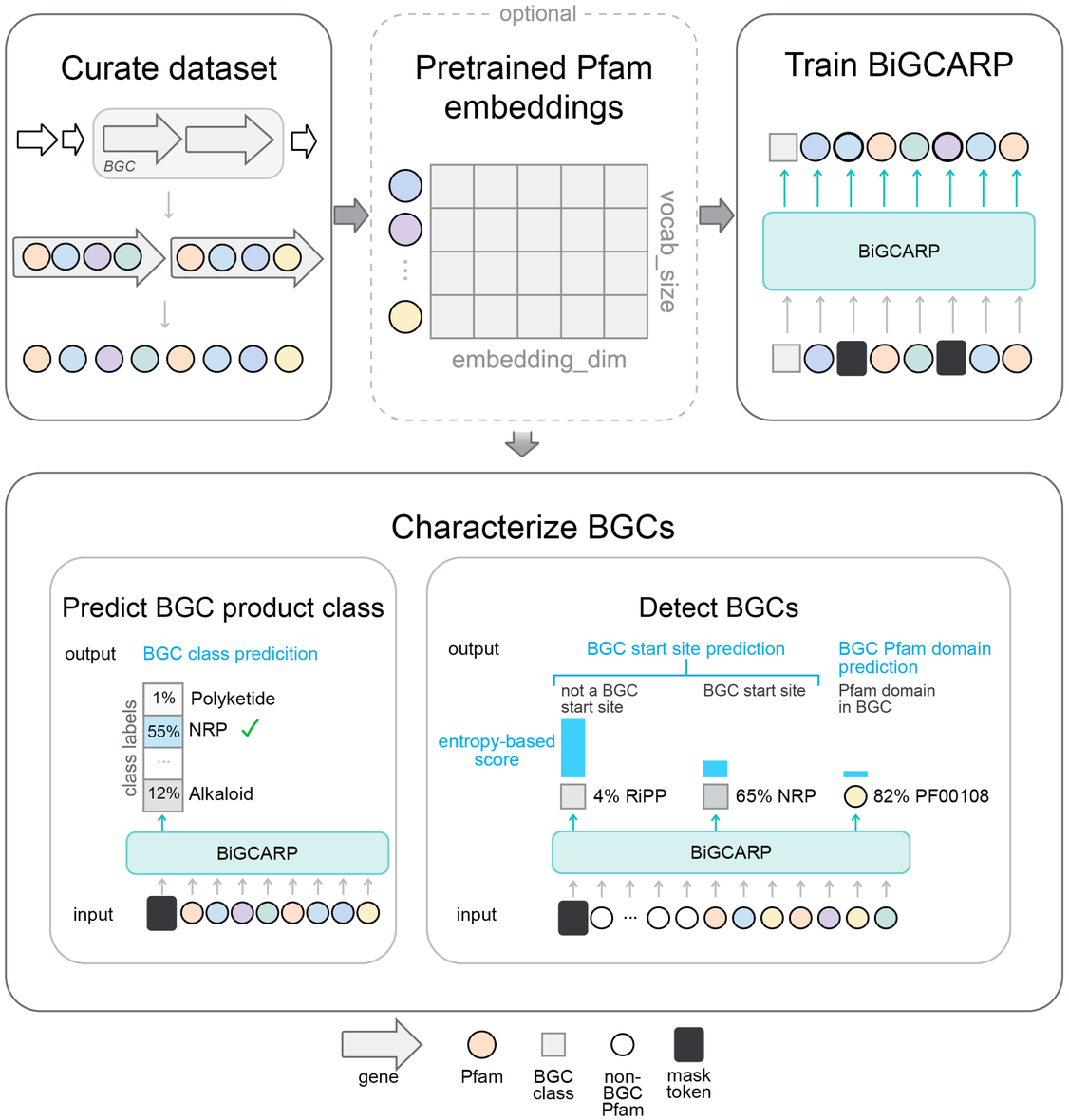

- First masked language model on domains and for BGCs

# BiGCARP detects and classifies BGCs



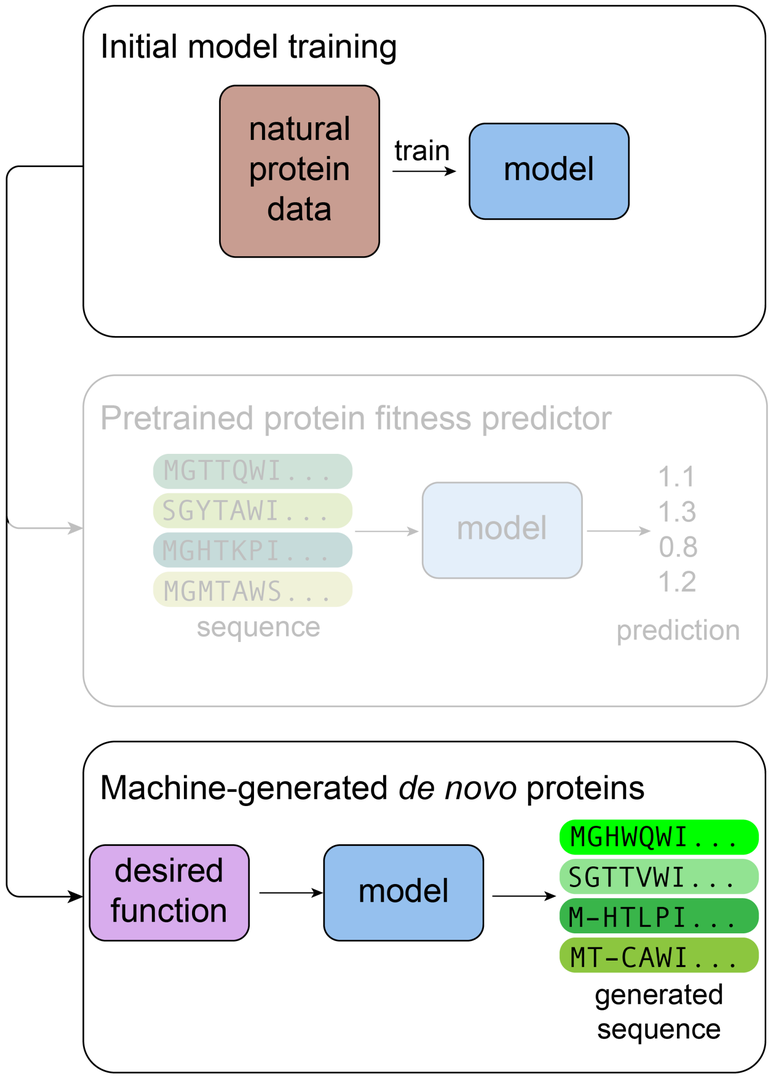

- First masked language model on domains and for BGCs
- Unsupervised BGC detection

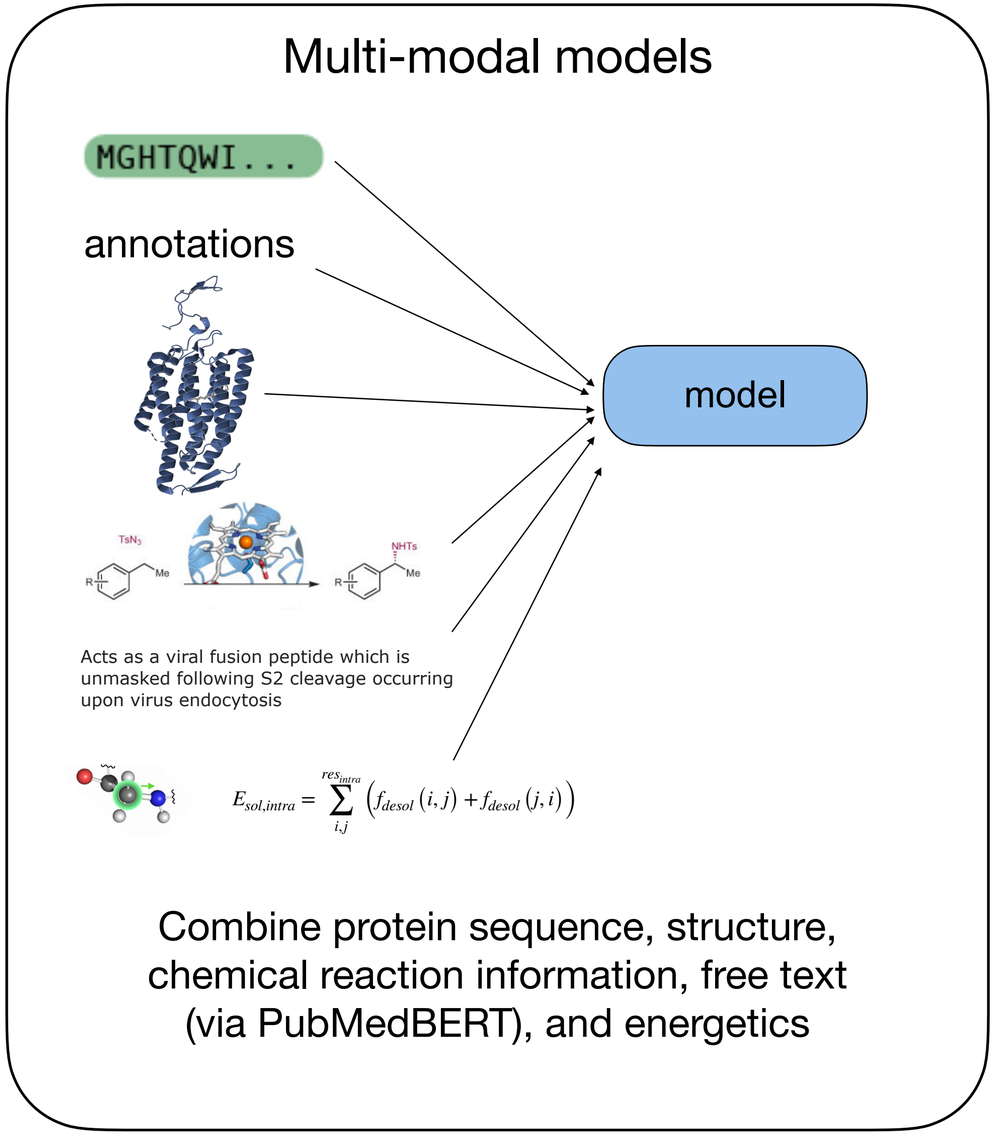
# BiGCARP detects and classifies BGCs



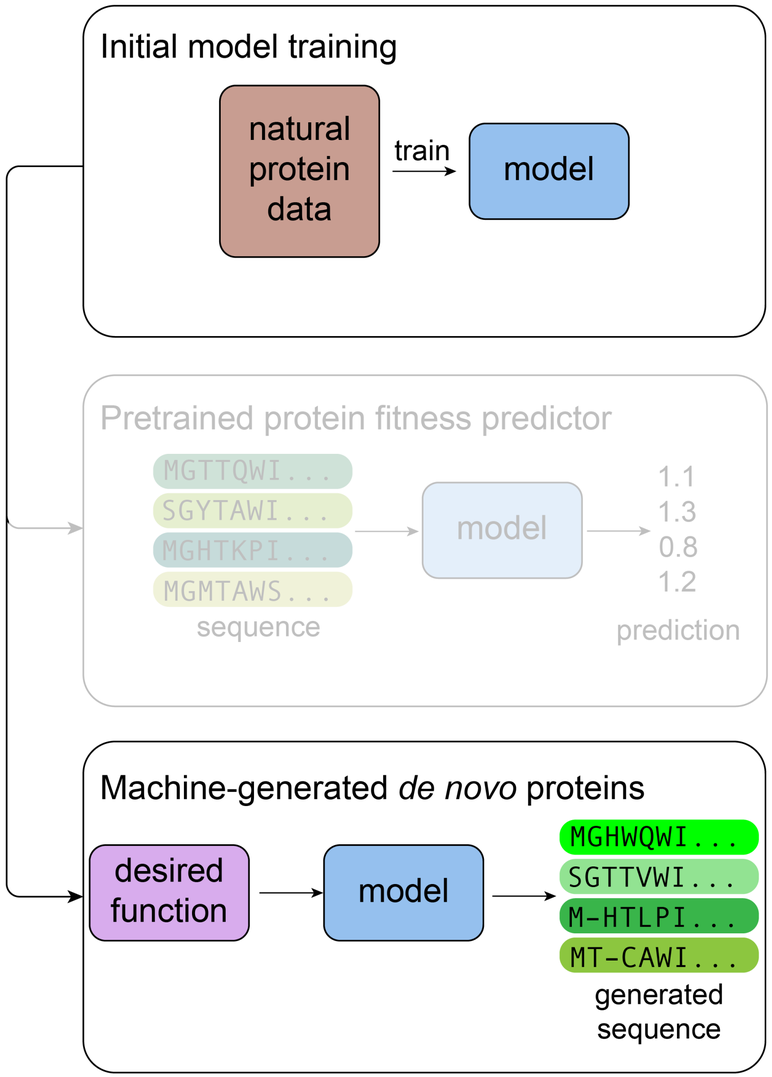

- First masked language model on domains and for BGCs
- Unsupervised BGC detection
- Learn product classes and BGC structure simultaneously

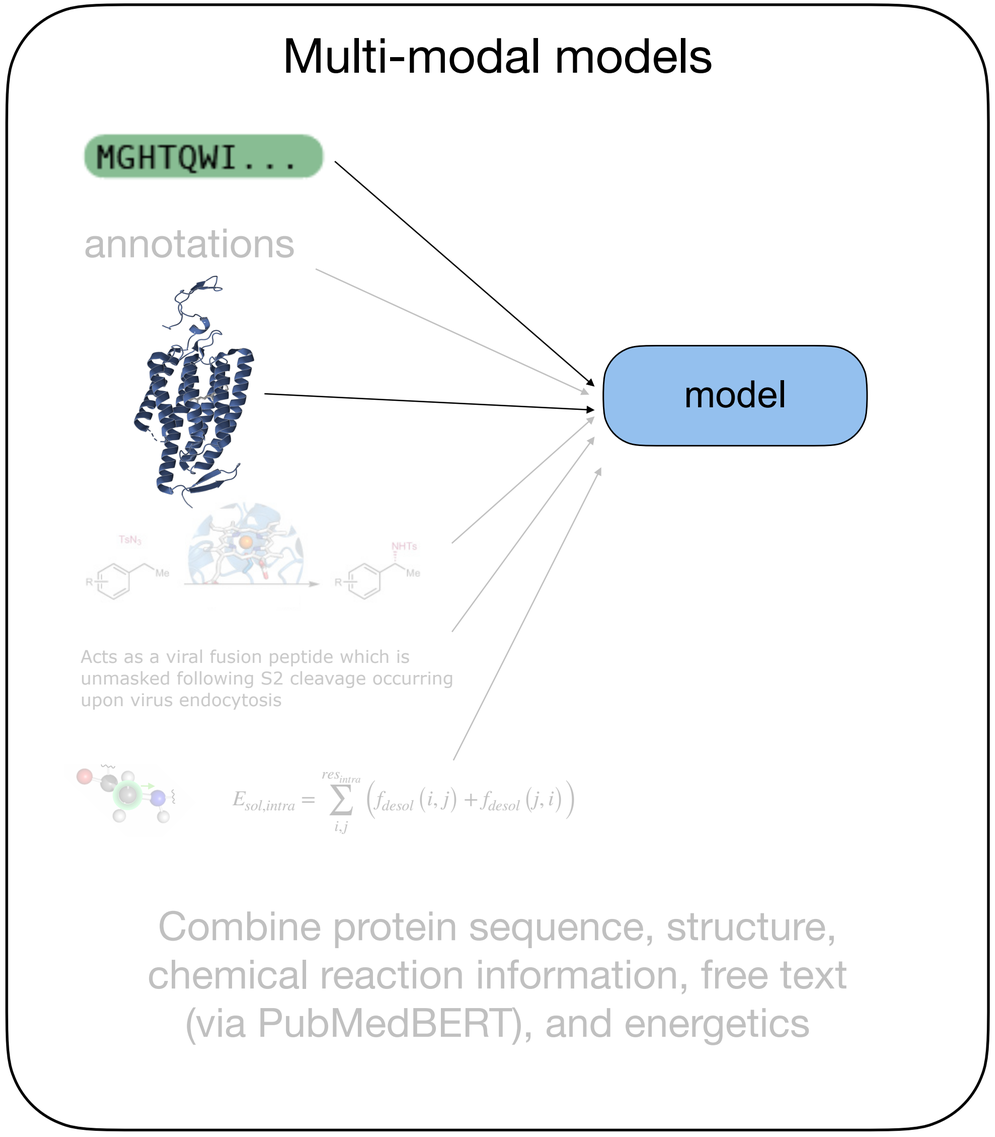
# Use multiple data modalities to discover and design proteins



# Use multiple data modalities to discover and design proteins



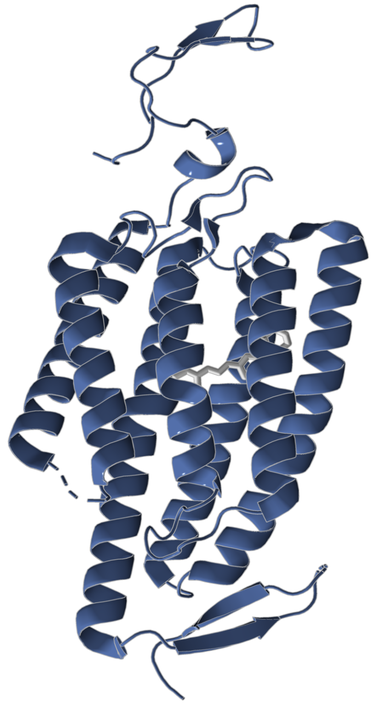
# Sequence determines structure determines function

# Sequence determines structure determines function

MGTGDHDD...

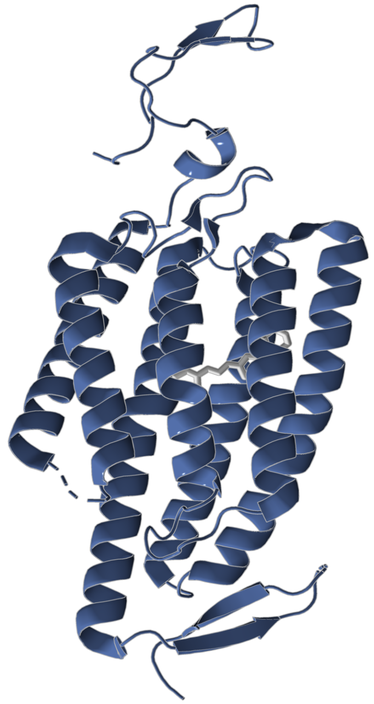
# Sequence determines structure determines function

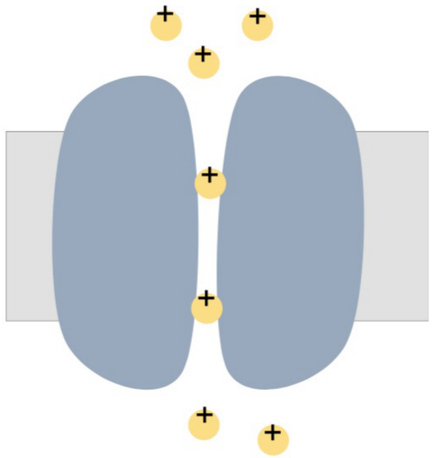
MGTGDHDD...



# Sequence determines structure determines function

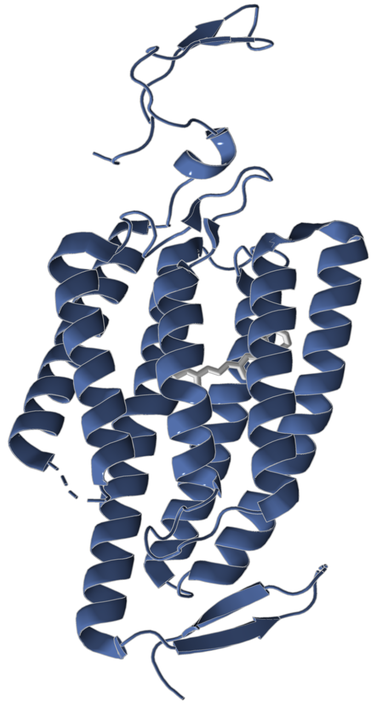
MGTGDHDD...

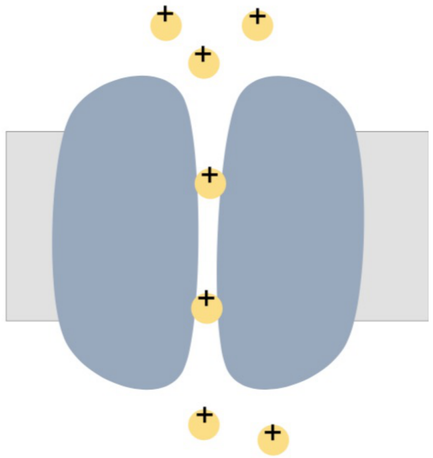




# Generating new, designable structures expands functional space

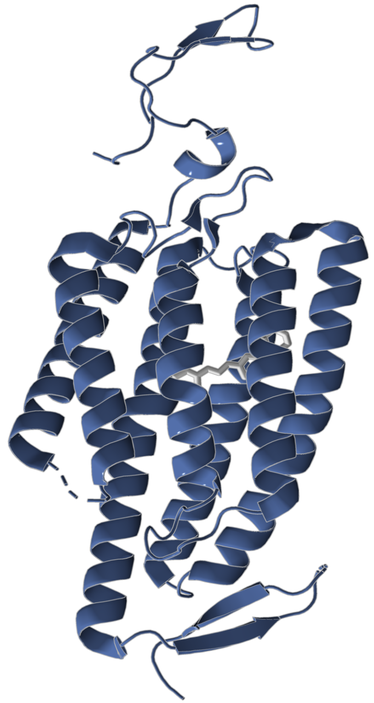
MGTGDHDD...

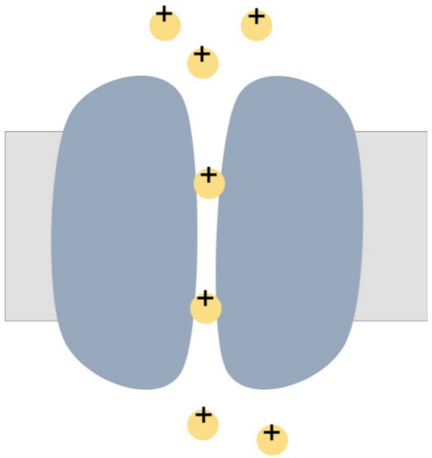




# Generating new, designable structures expands functional space

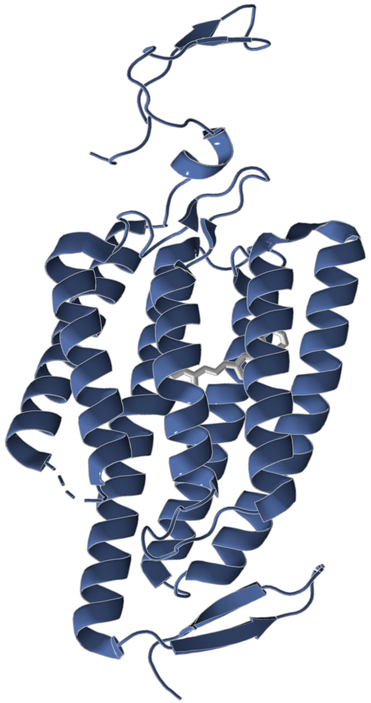
MGTGDHDD...

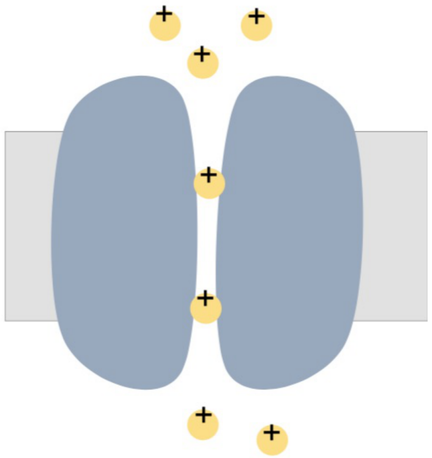




# Generating new, designable structures expands functional space

MGTGDHDD...

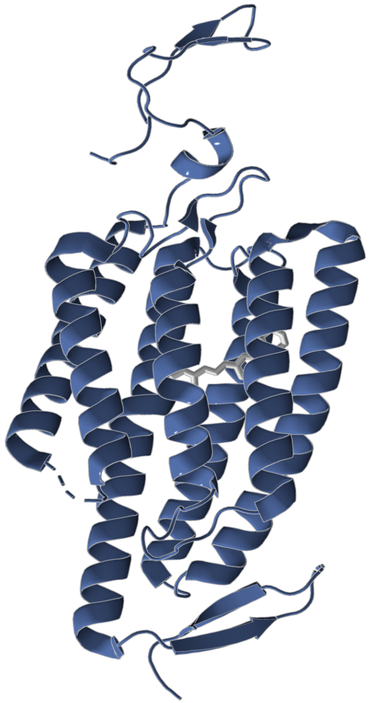


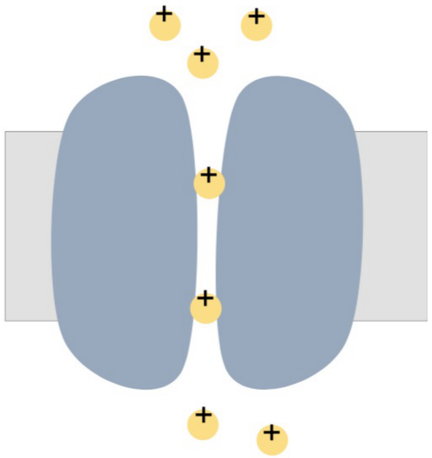


Challenge: Generate diverse and designable structures

# Generating new, designable structures expands functional space

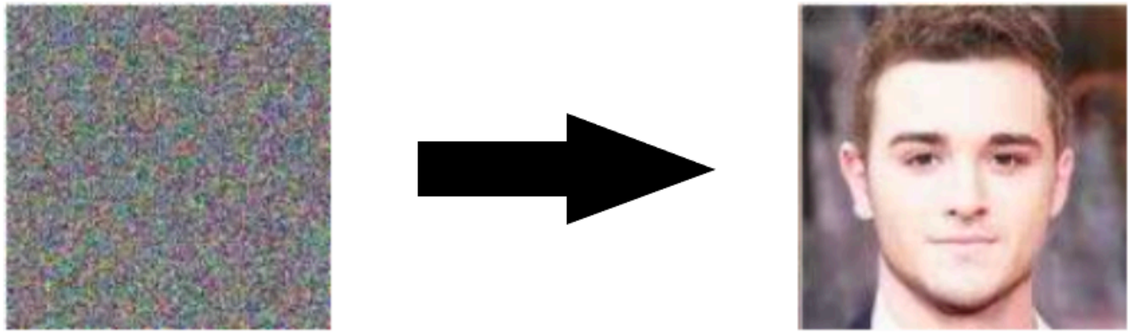
MGTGDHDD...



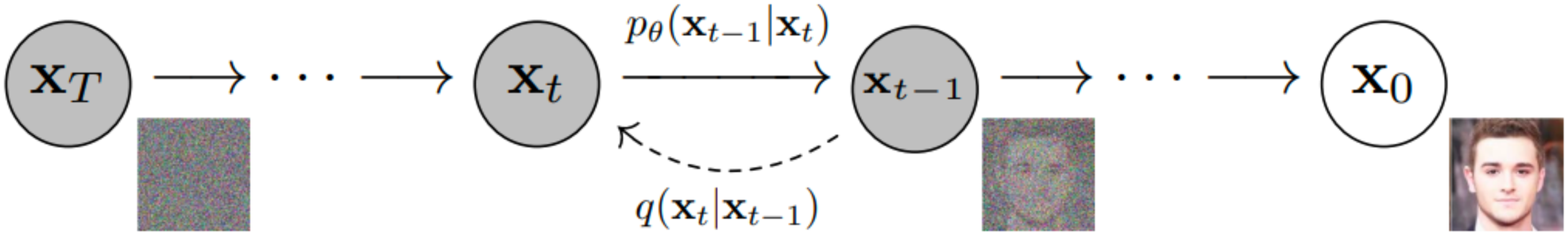


Challenge: Generate diverse and designable structures

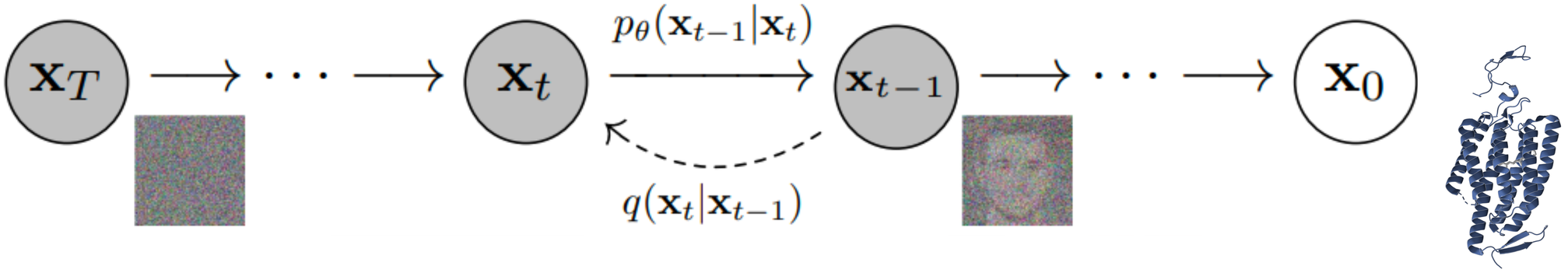
# Diffusion models learn to generate objects from noise



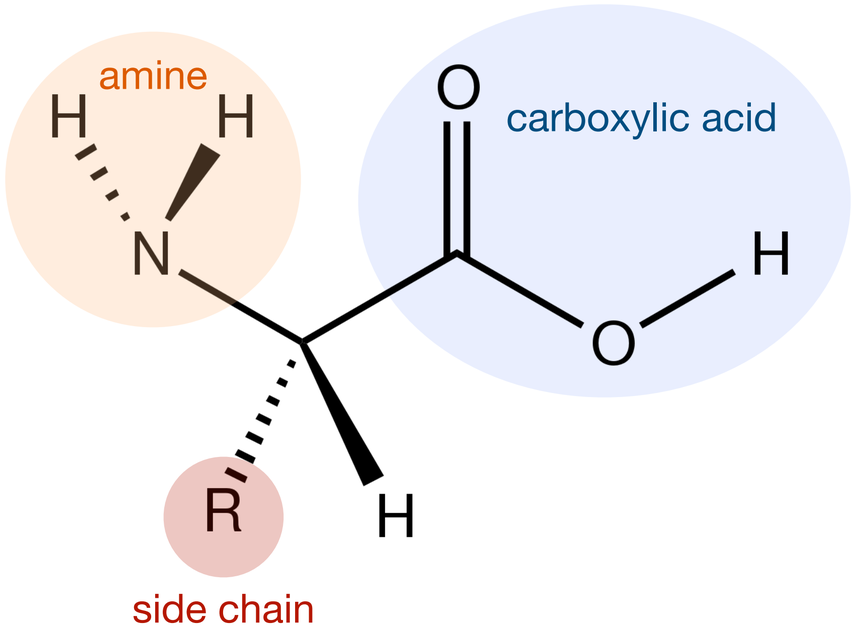
# Diffusion models learn to generate objects from noise



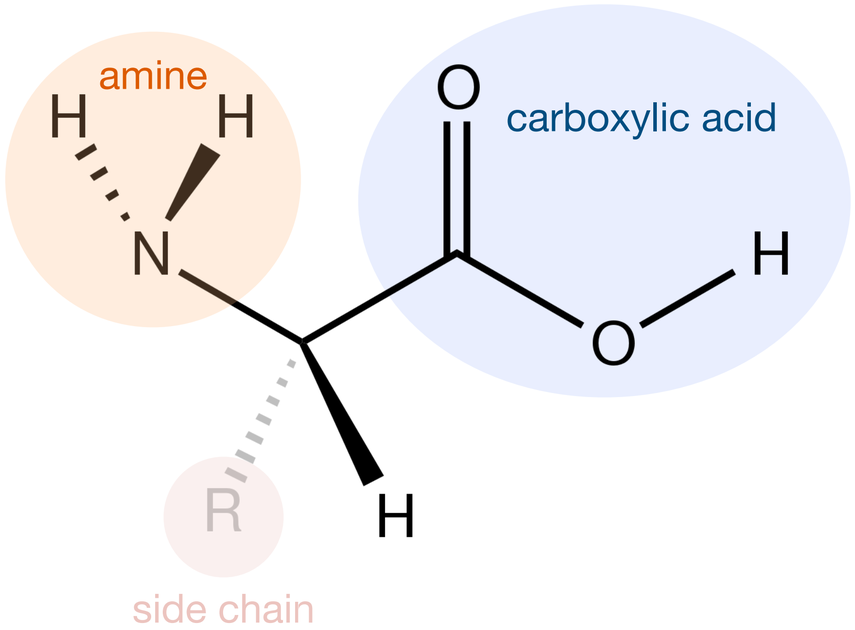
# Diffusion models learn to generate objects from noise



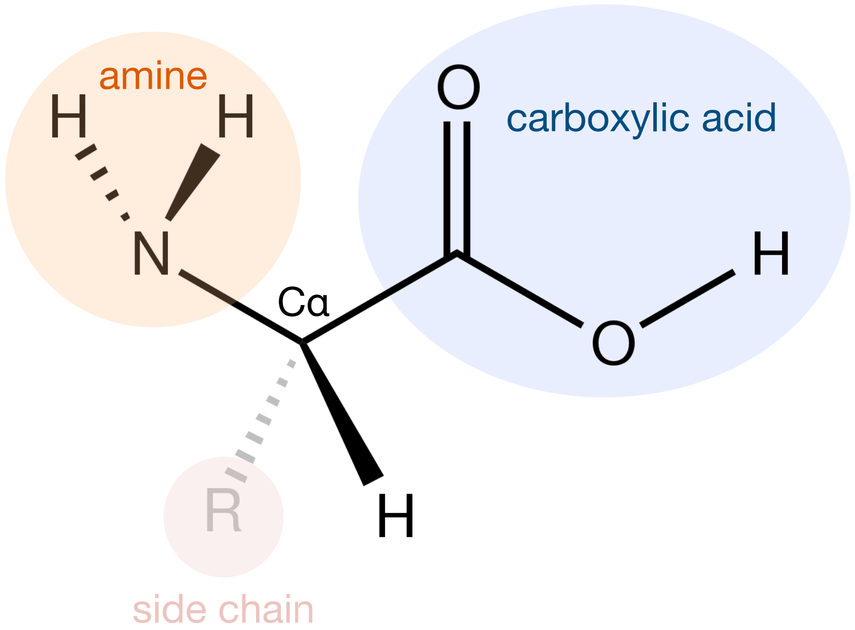
# Generate backbone structures



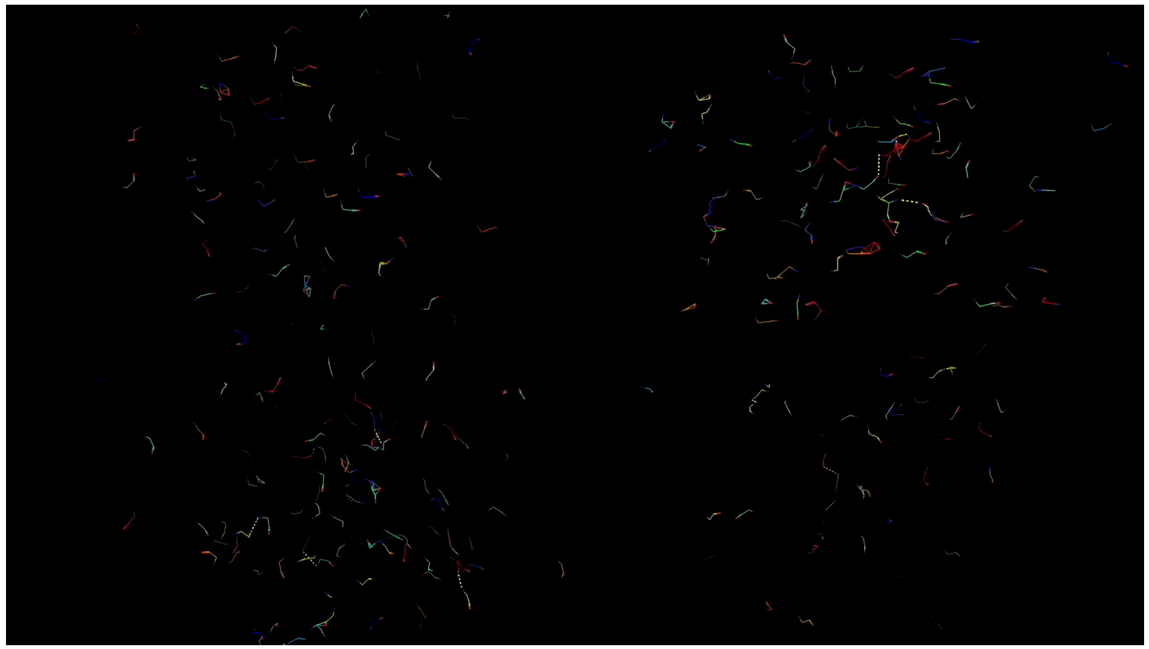
# Generate backbone structures



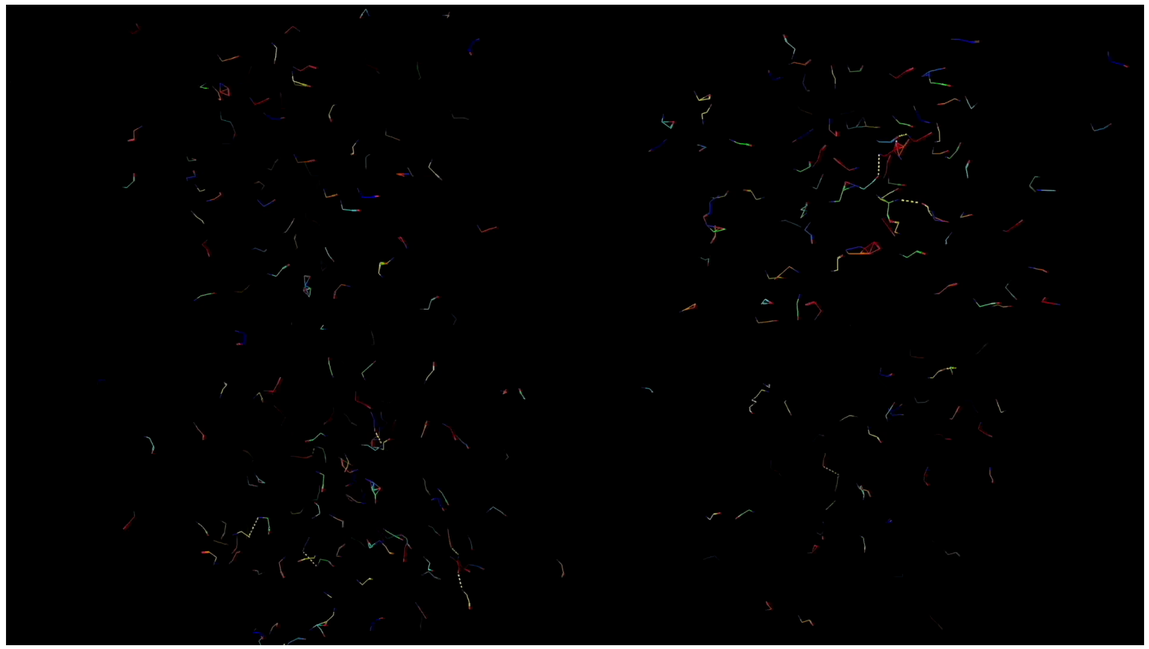
# Generate backbone structures



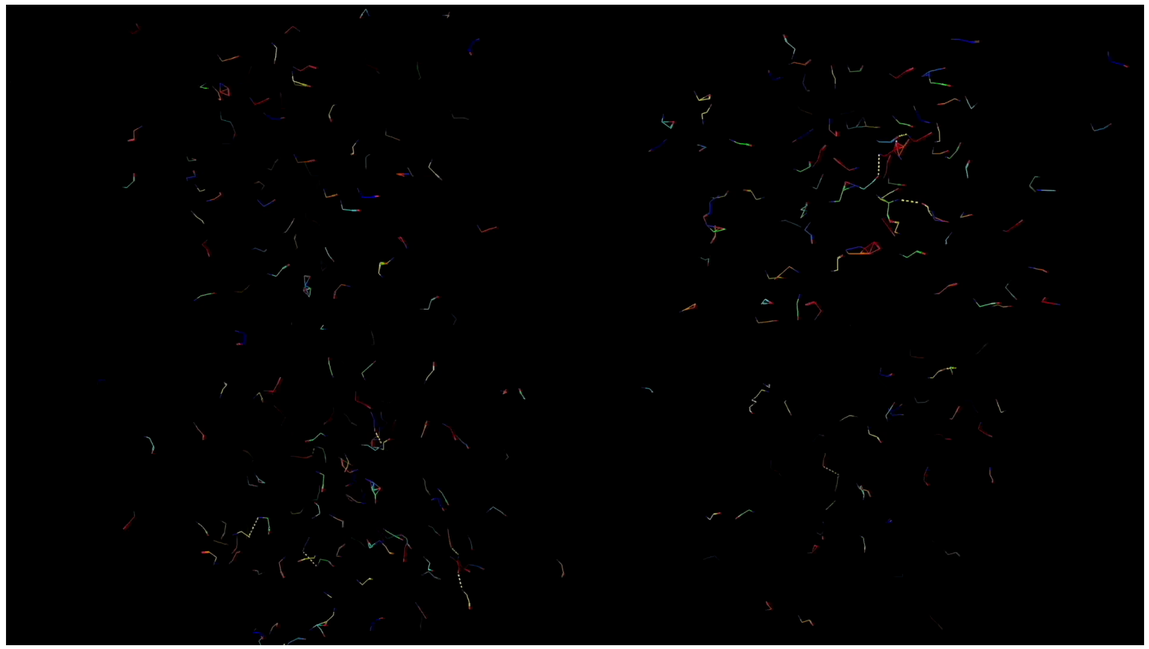
# Diffusion on 3D coordinates requires equivariances

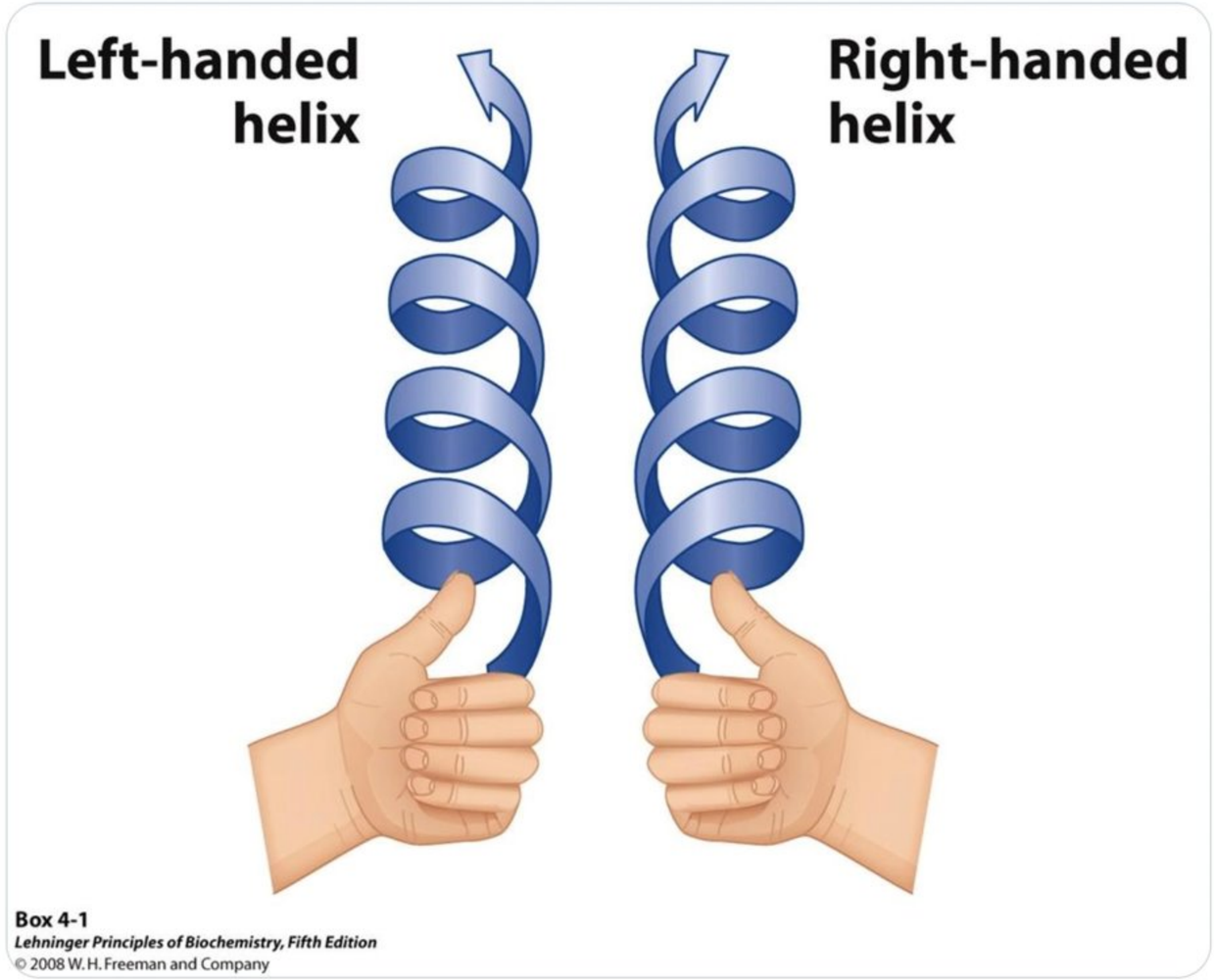


# Diffusion on 3D coordinates requires equivariances



# Diffusion on 3D coordinates requires equivariances



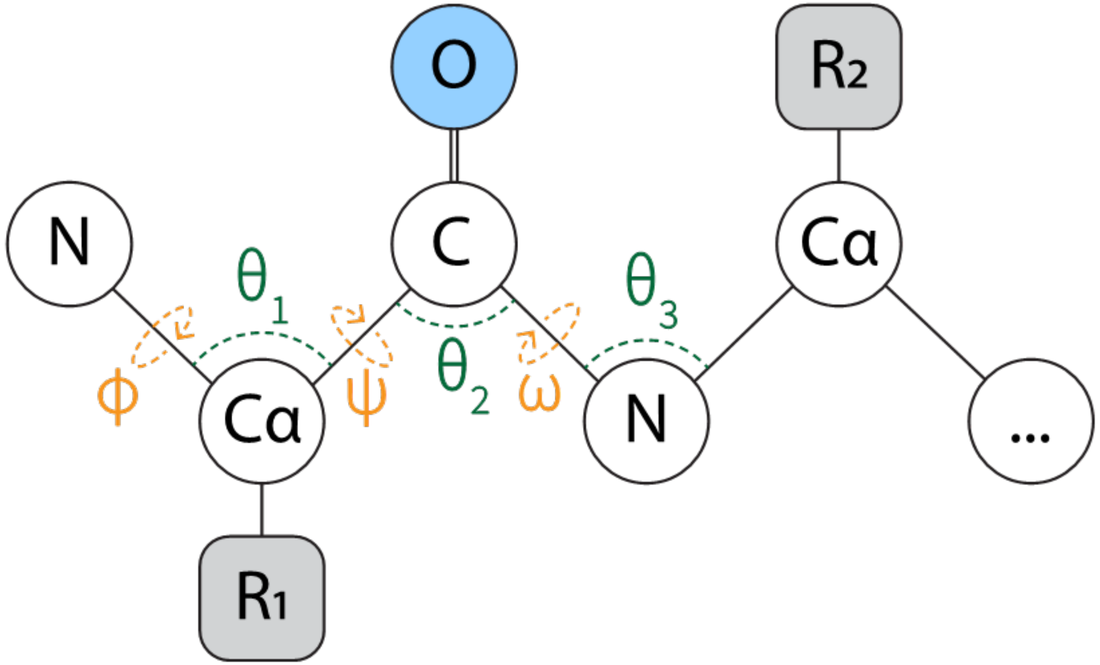



**Box 4-1**
***Lehninger Principles of Biochemistry, Fifth Edition***

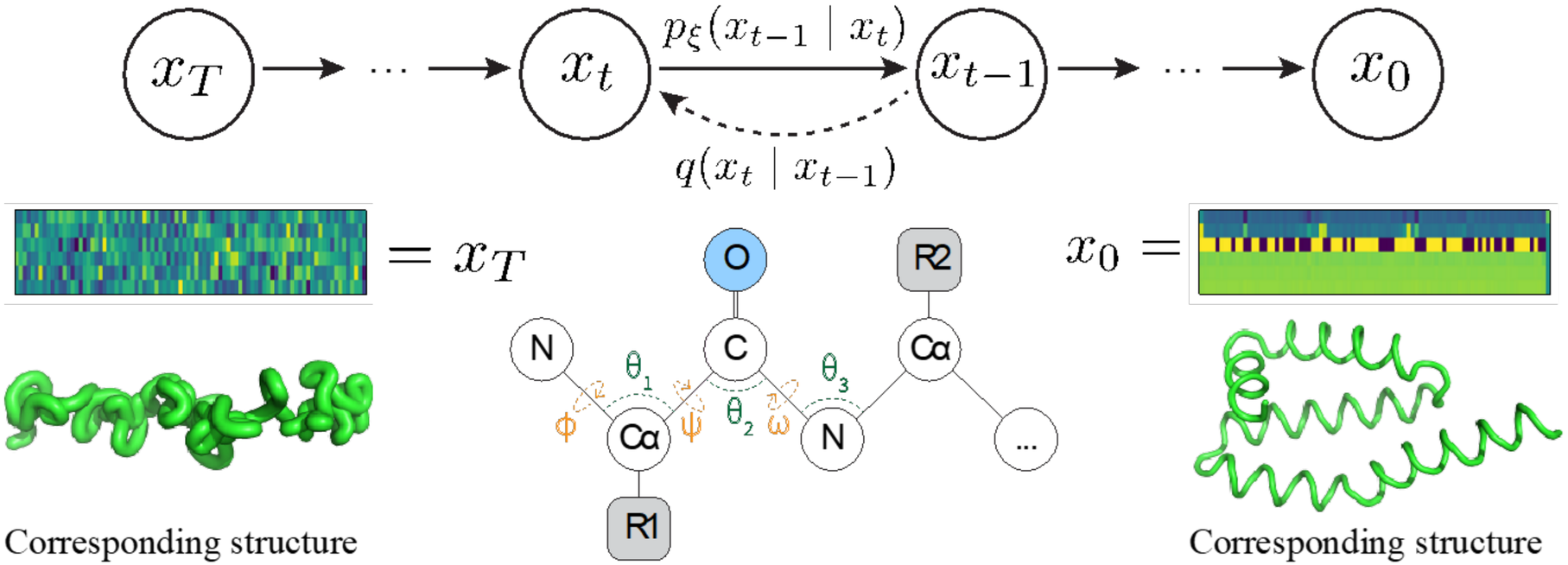
# Folding Diffusion uses coordinates inspired by protein folding

# Folding Diffusion uses coordinates inspired by protein folding

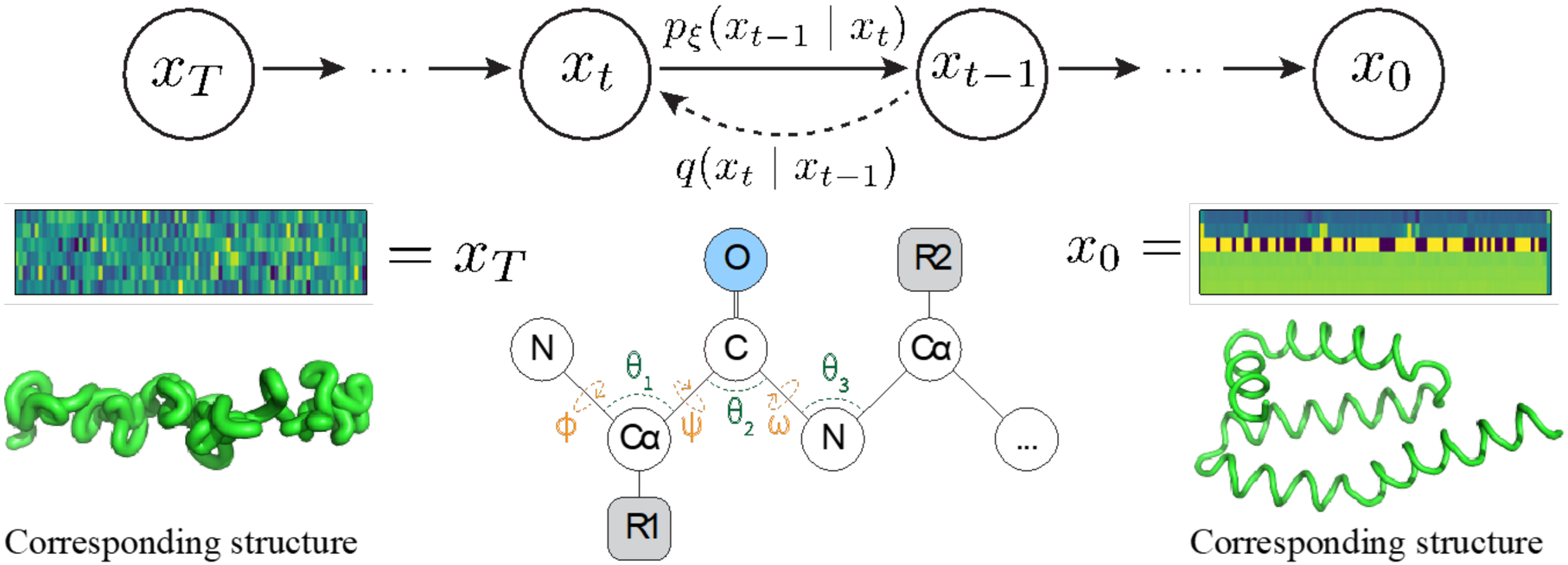
# Folding Diffusion uses coordinates inspired by protein folding



# Folding Diffusion uses coordinates inspired by protein folding

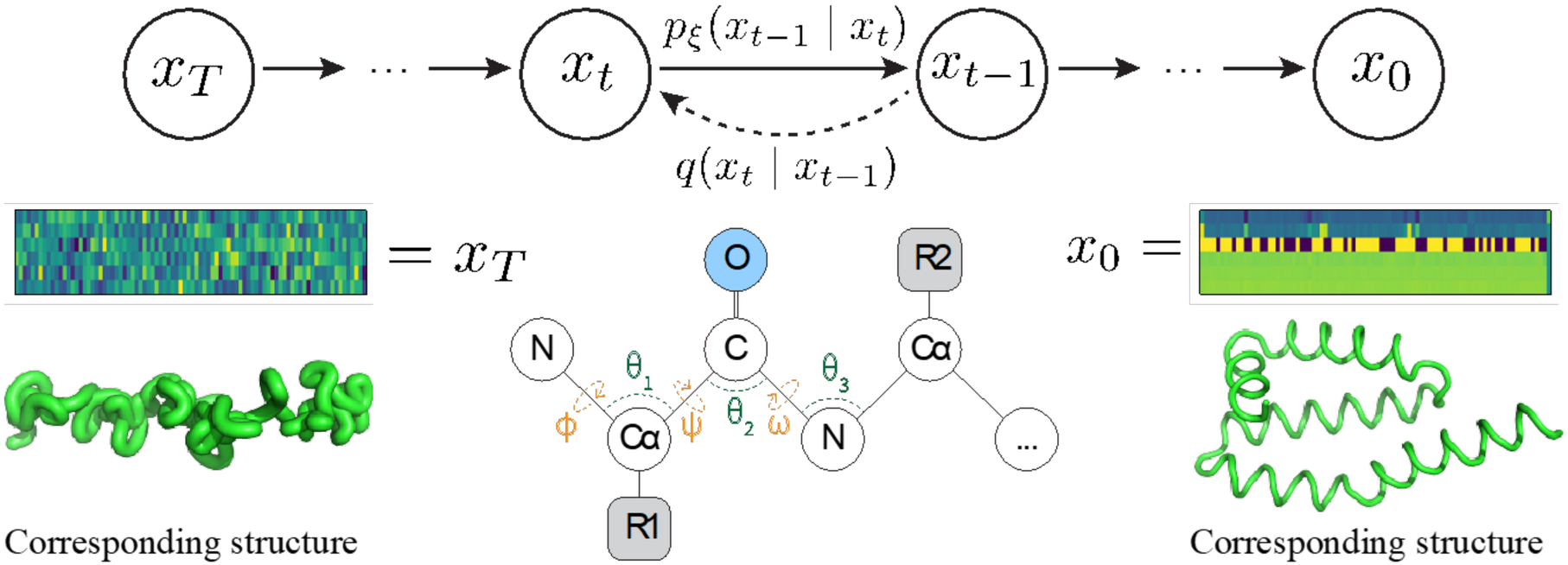


# Folding Diffusion uses coordinates inspired by protein folding




Vanilla transformer

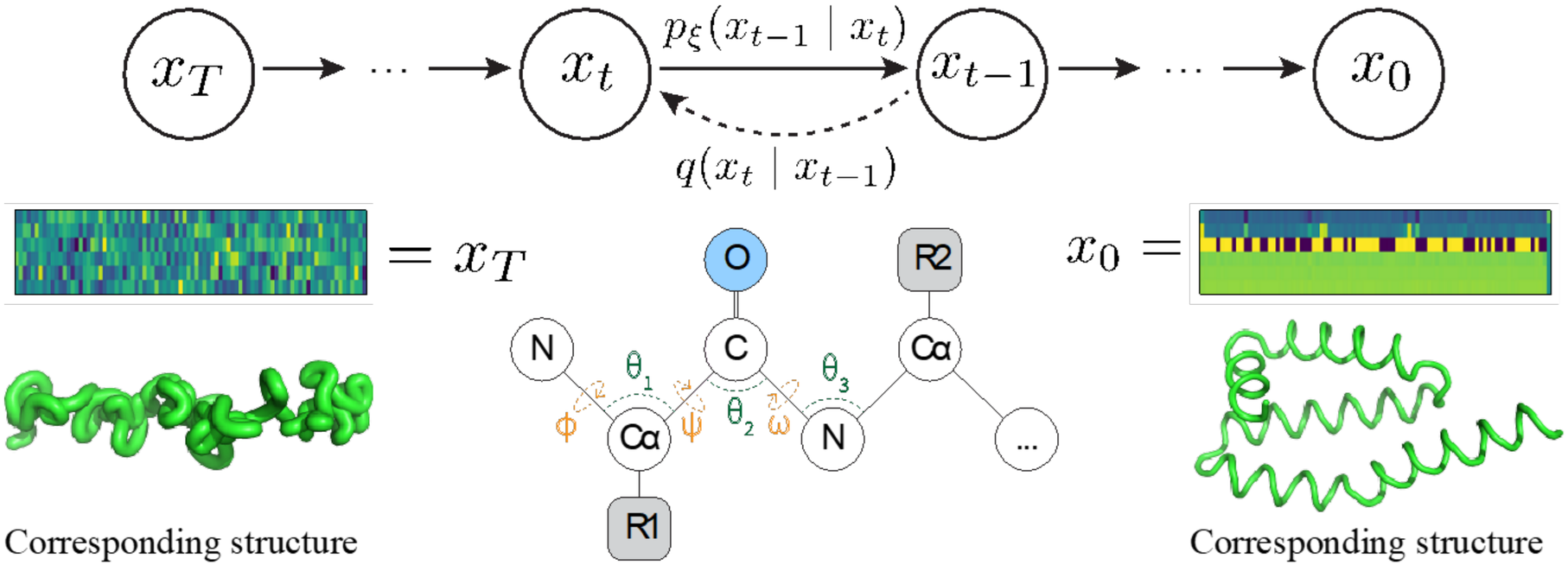
# Folding Diffusion uses coordinates inspired by protein folding




Vanilla transformer

30k structures from CATH

# Folding Diffusion uses coordinates inspired by protein folding

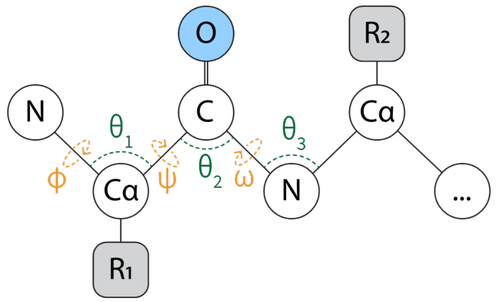


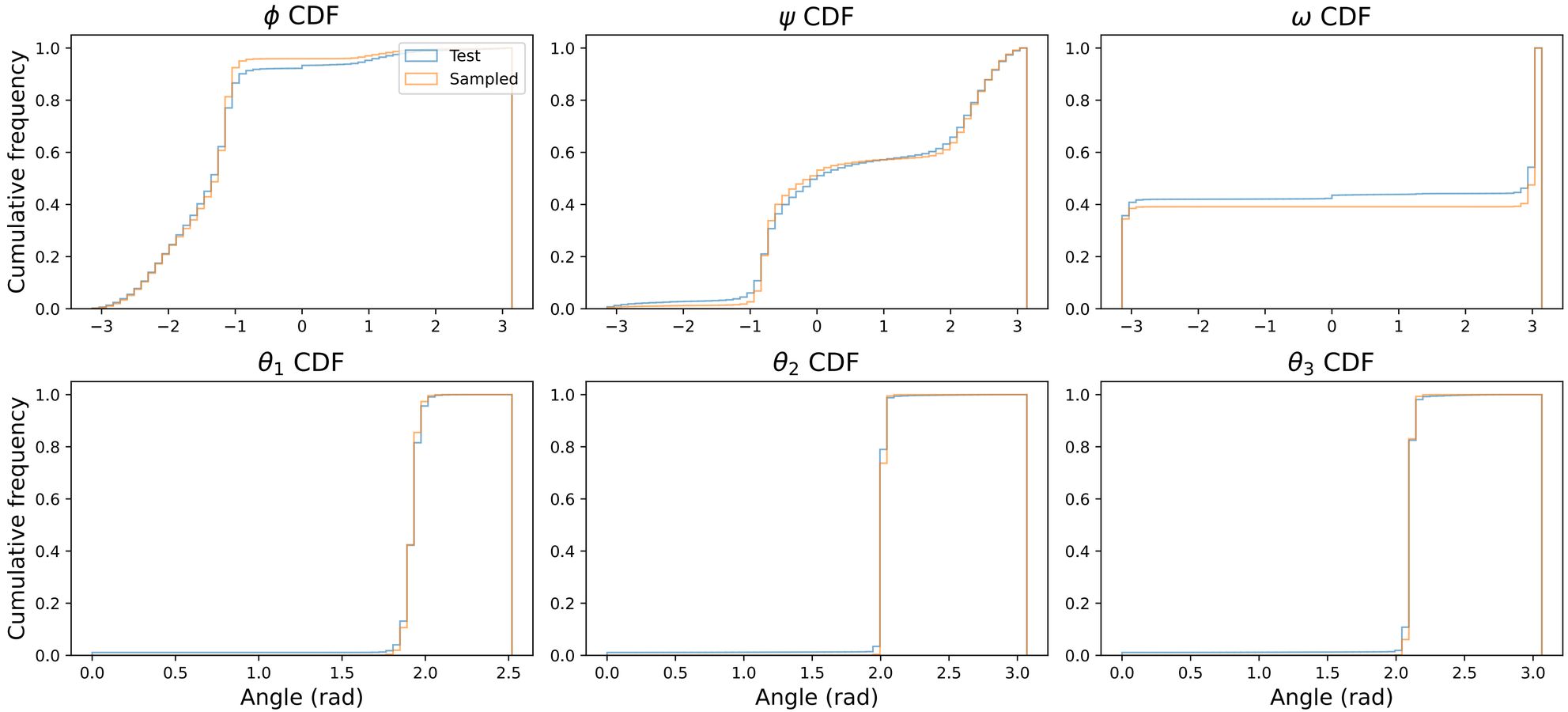

Vanilla transformer

30k structures from CATH

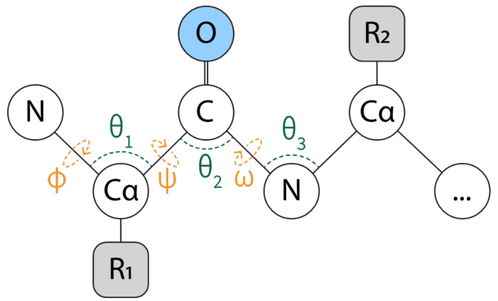
40 - 128 residues

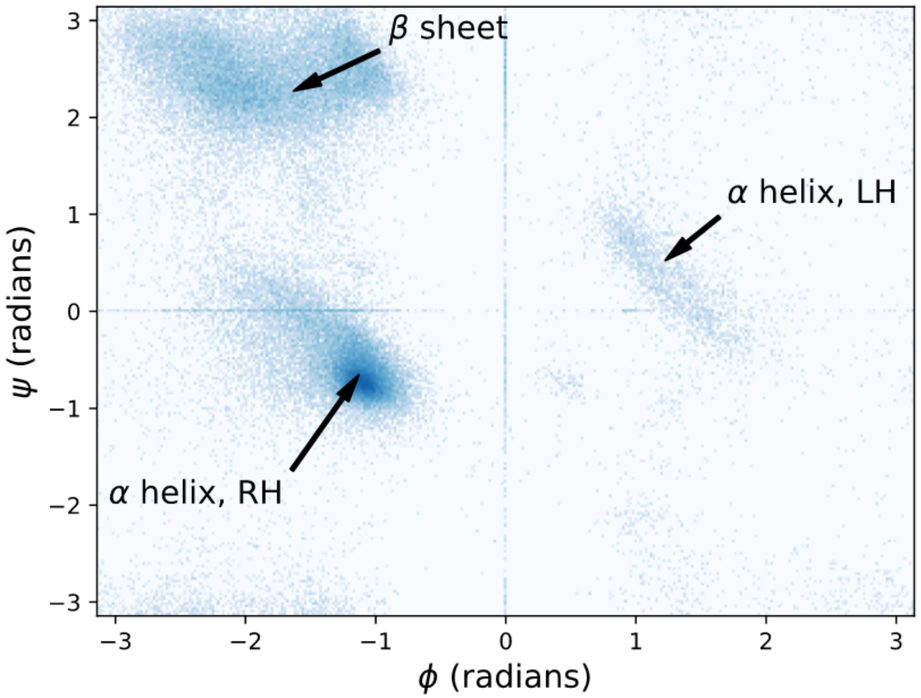
# Generated angles match test distribution



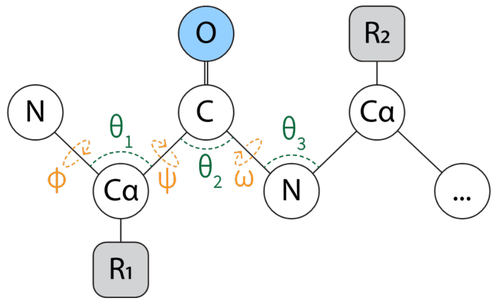



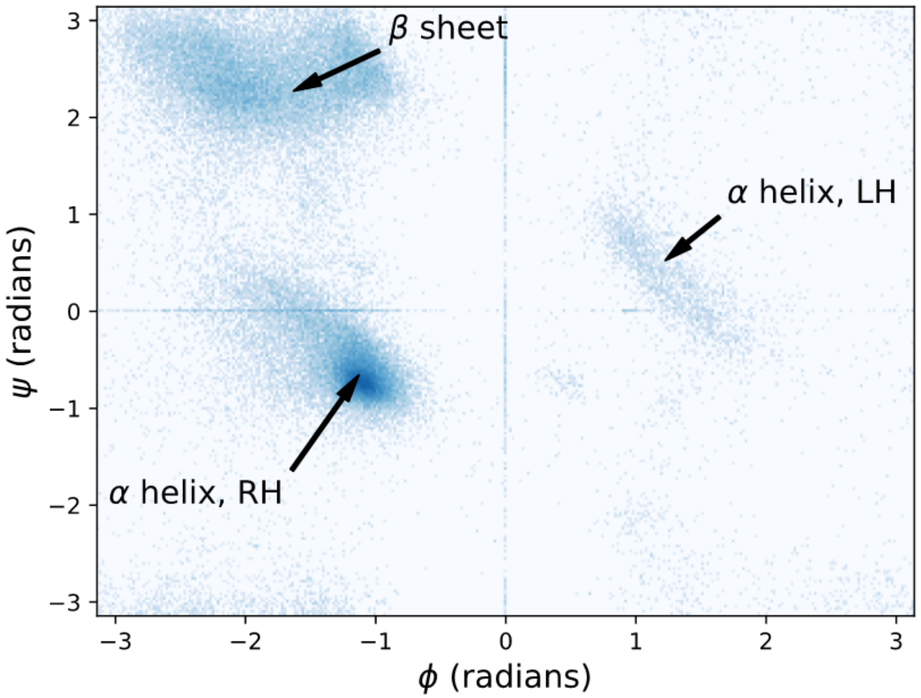
# Generated Ramachandran plots match test structures

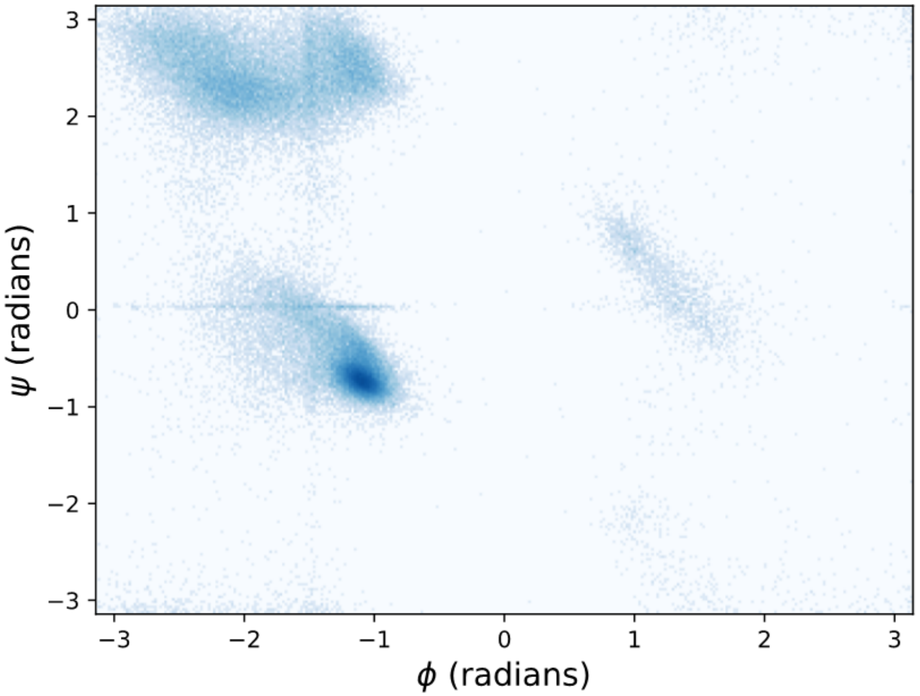


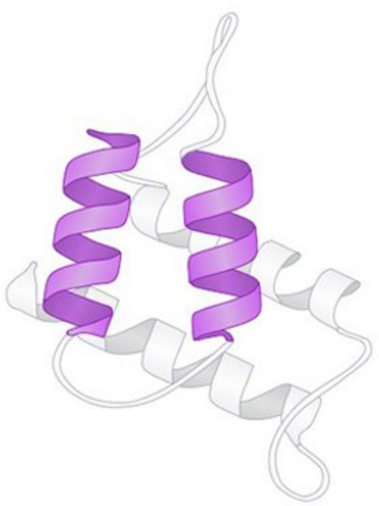
# Generated Ramachandran plots match test structures



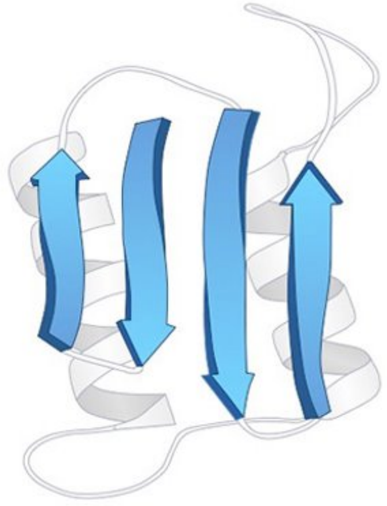



# Generated secondary structures match test structures

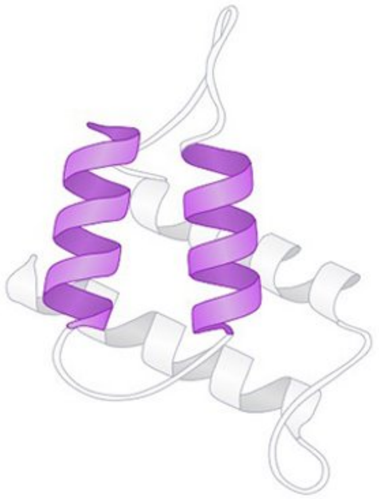


α - helices

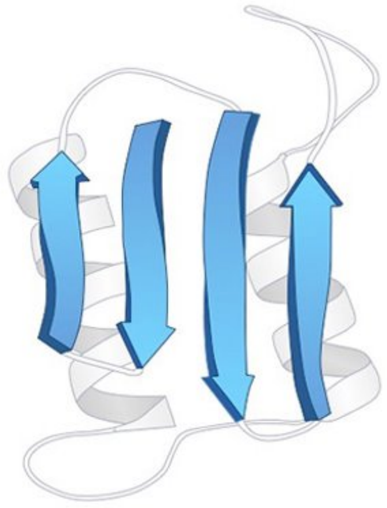


β - pleated sheets

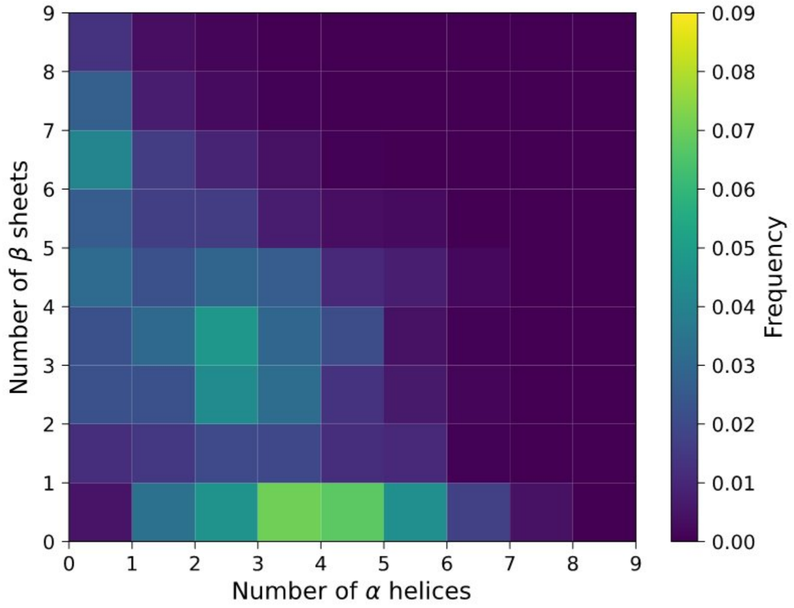
# Generated secondary structures match test structures



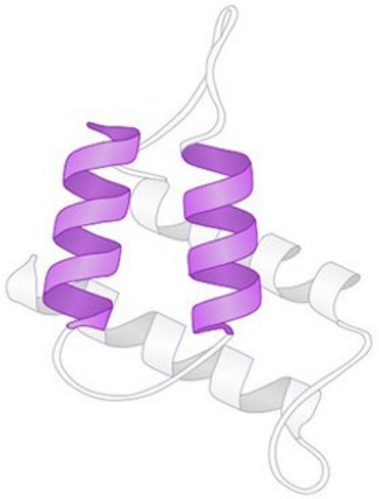
α - helices



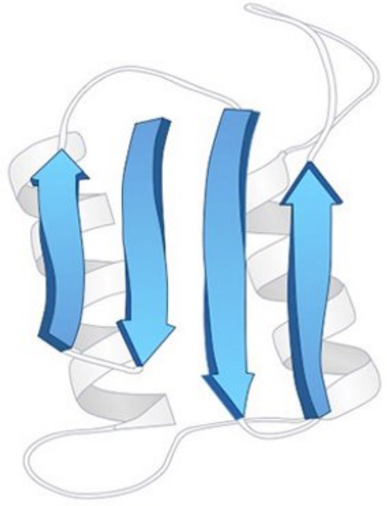
β - pleated sheets



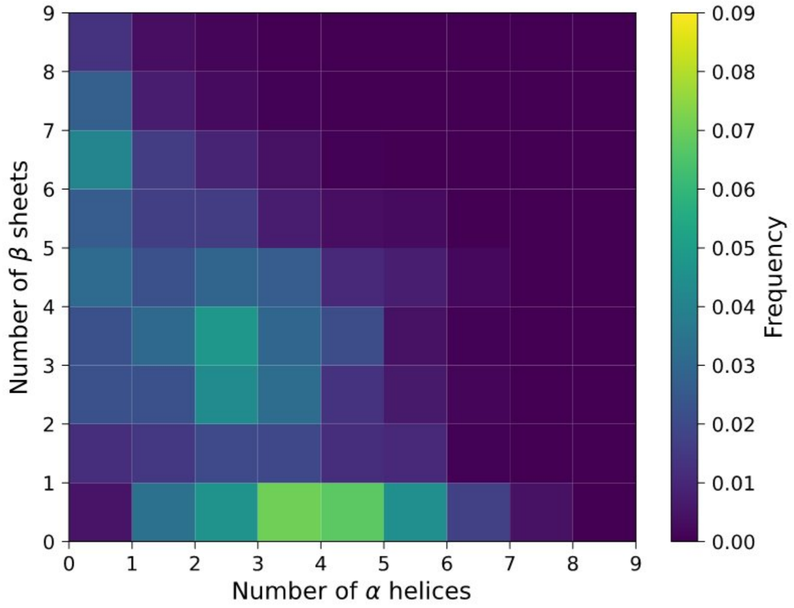
# Generated secondary structures match test structures

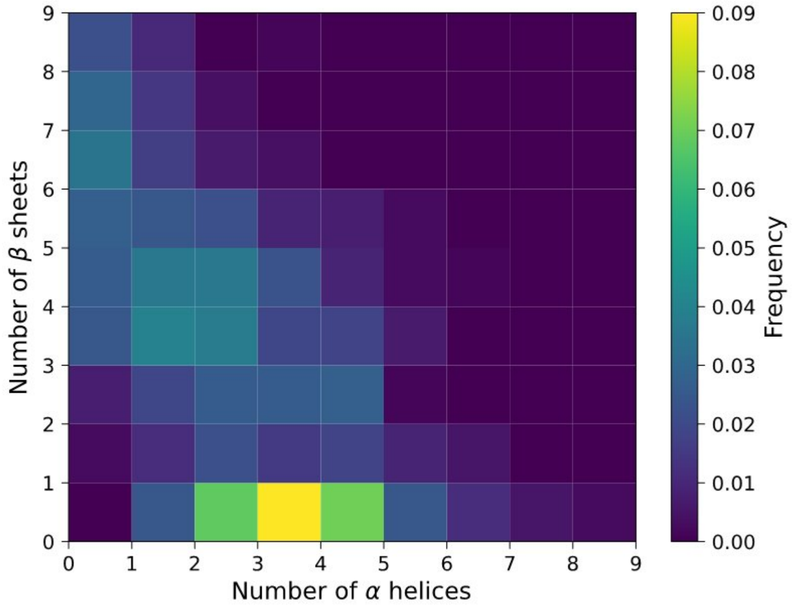


α - helices

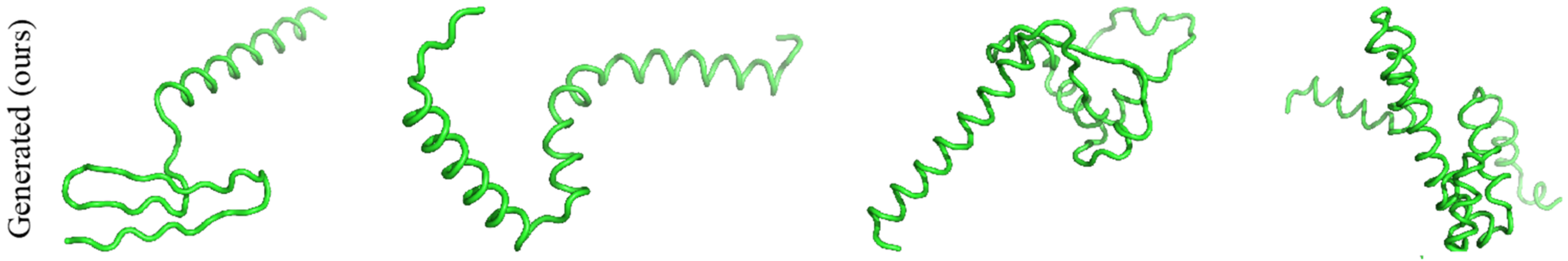


β - pleated sheets



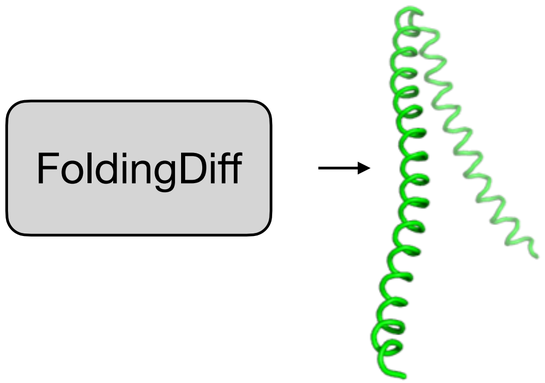



# Generated structures look reasonable

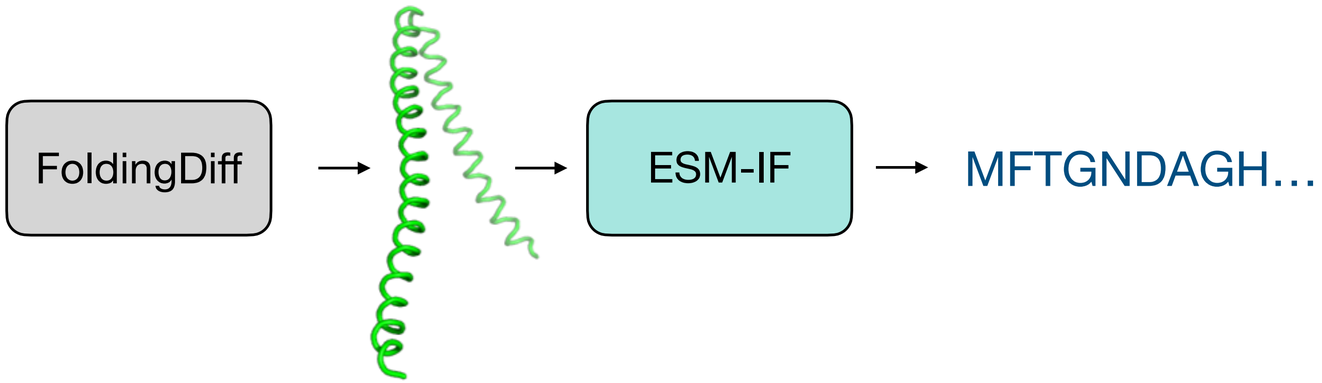


# Measure design ability of structures with self-consistency TMscore

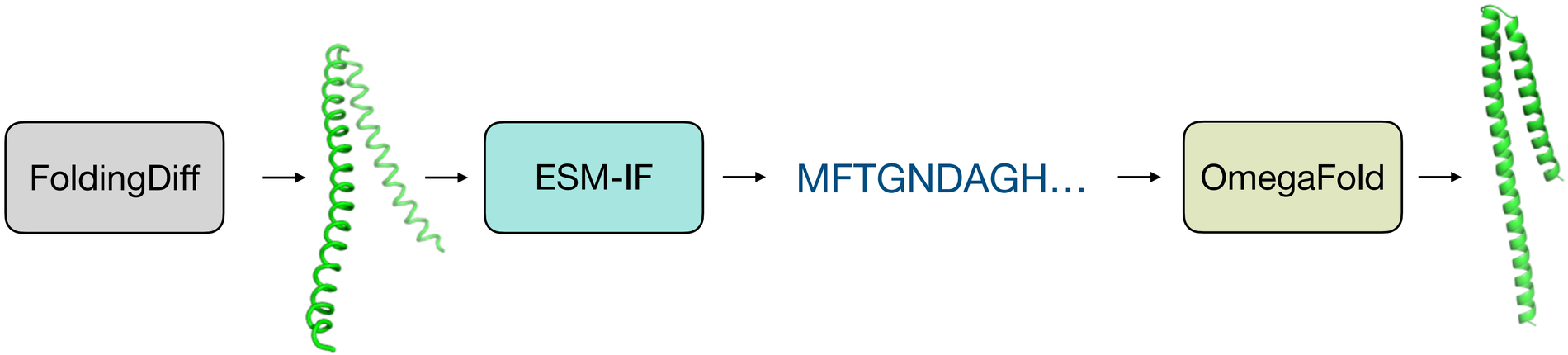
# Measure design ability of structures with self-consistency TMscore



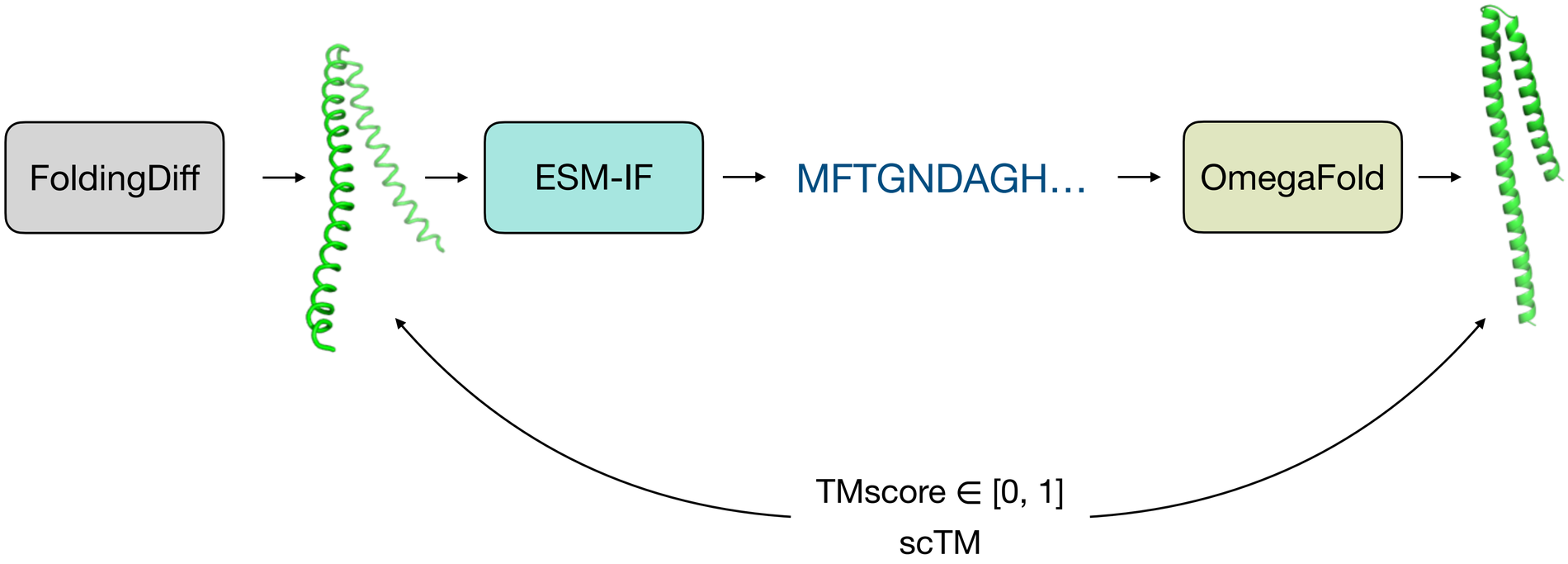
# Measure design ability of structures with self-consistency TMscore



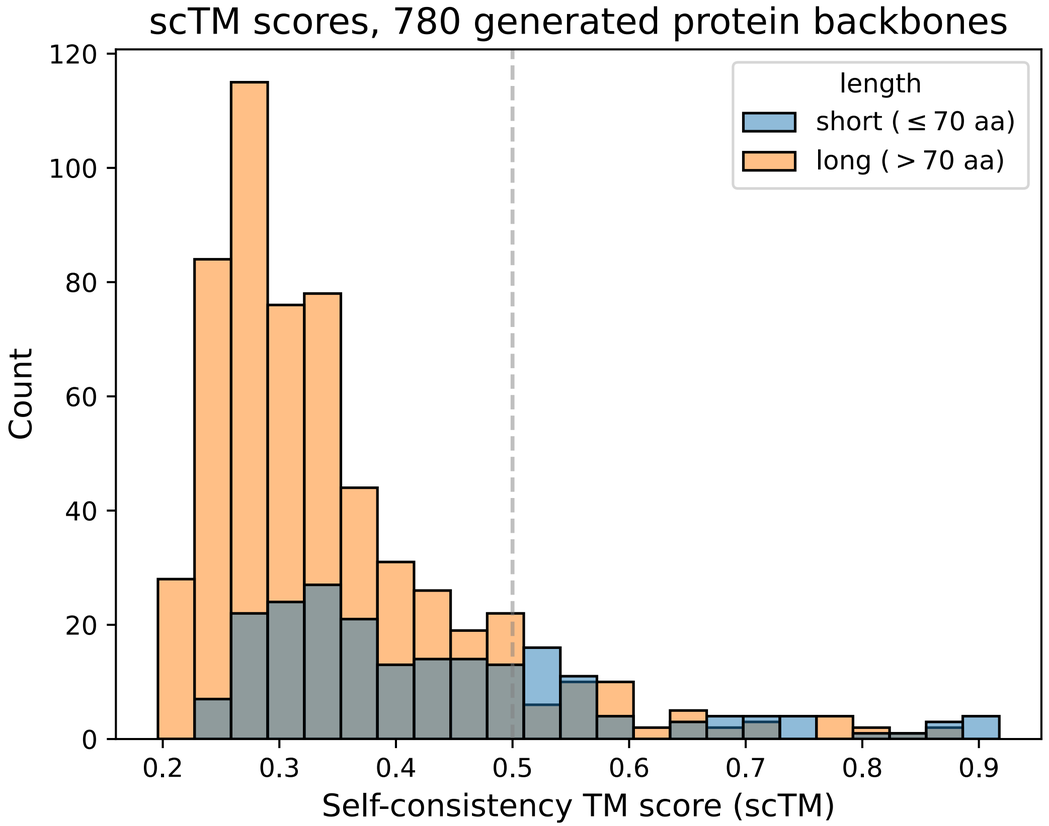
# Measure design ability of structures with self-consistency TMscore



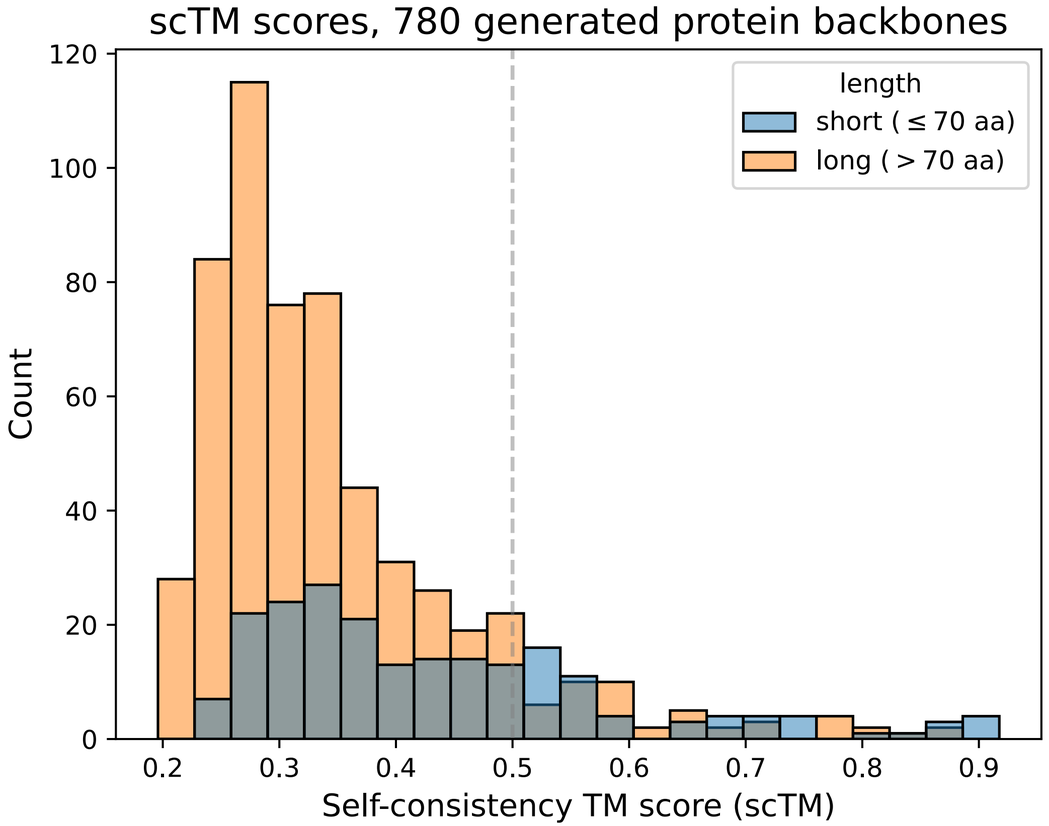
# Measure design ability of structures with self-consistency TMscore



# Many generated structures are designable

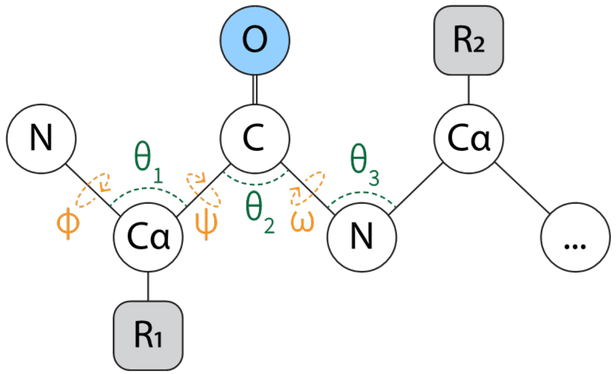


# Many generated structures are designable



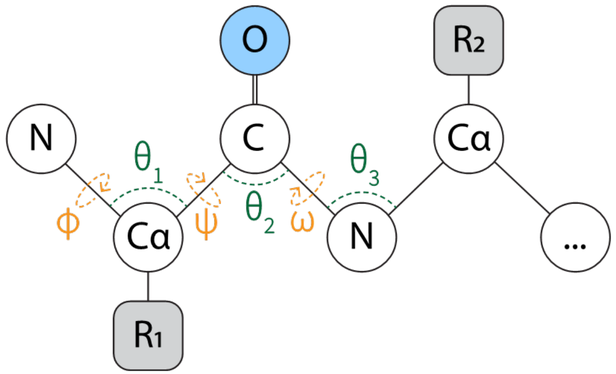

| scTM > 0.5 | ≤ 70 aa | > 70 aa |
|---|---|---|
| FoldingDiff | 57/210 | 54/570 |
| ProtDiff (Trippe *et al.*) | 36/210 | 51/570 |

# FoldingDiff structures are better than random baseline

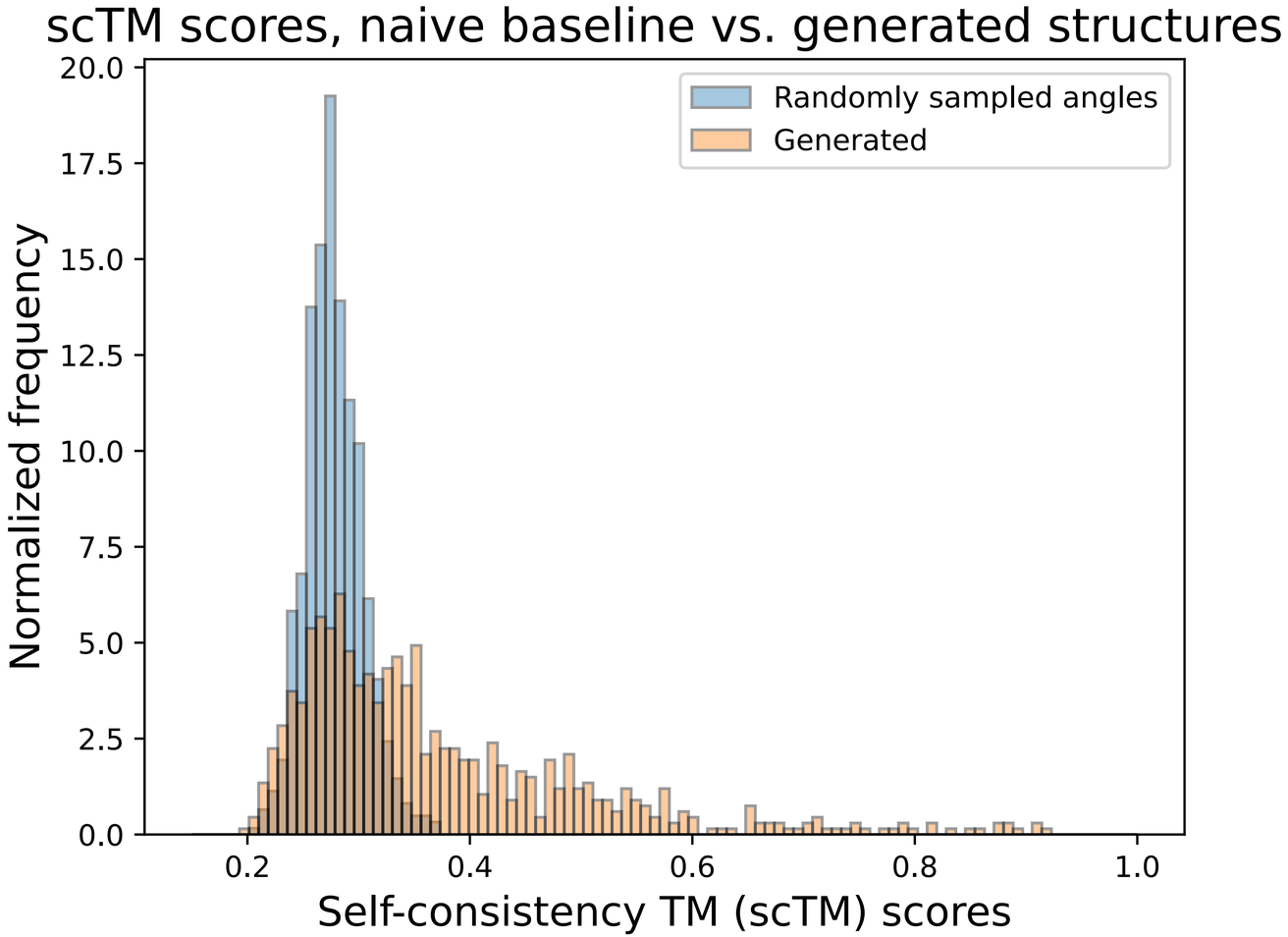



Sample sets of angles
Preserves Ramachandran plot

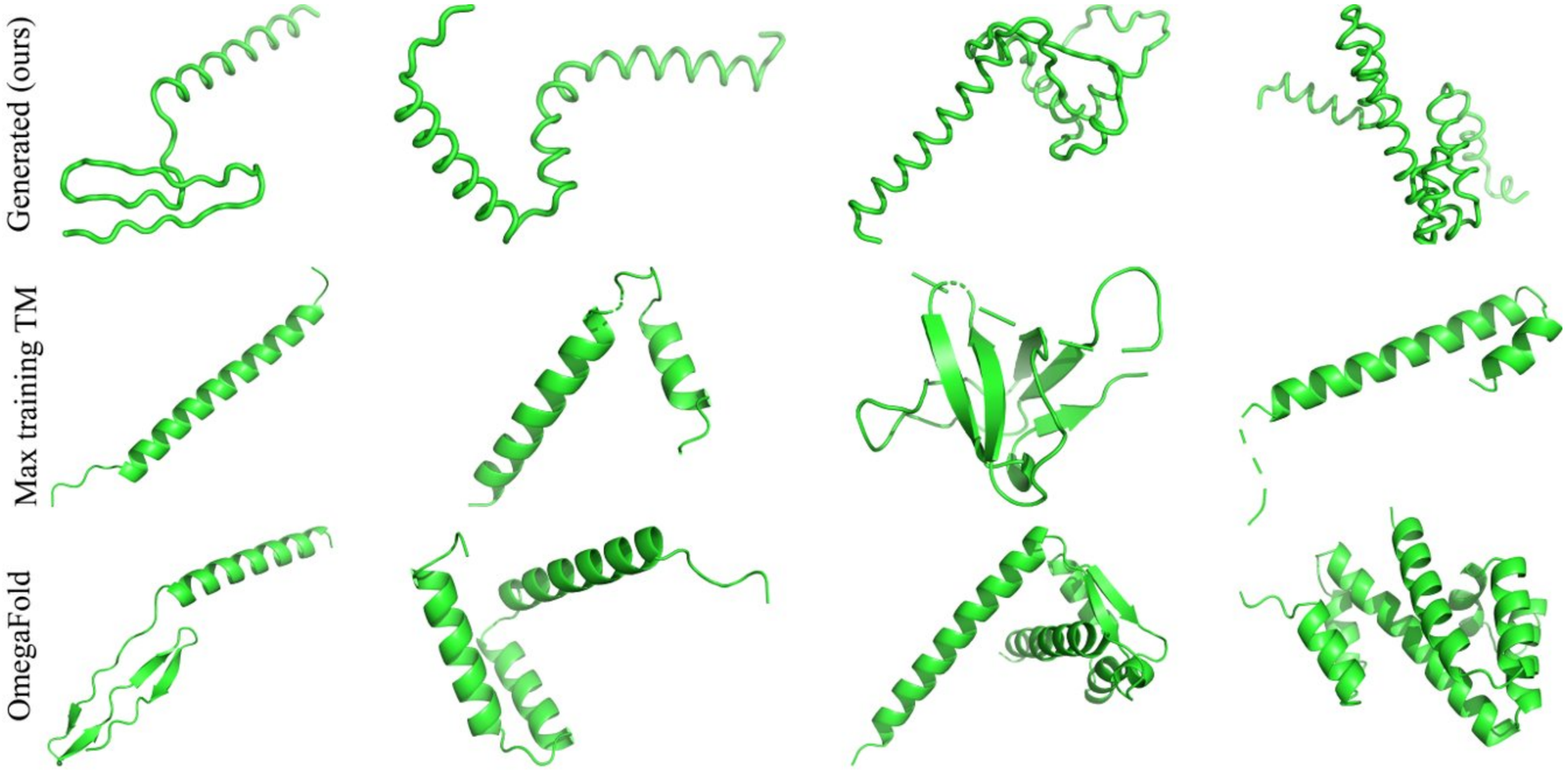
# FoldingDiff structures are better than random baseline




Sample sets of angles
Preserves Ramachandran plot

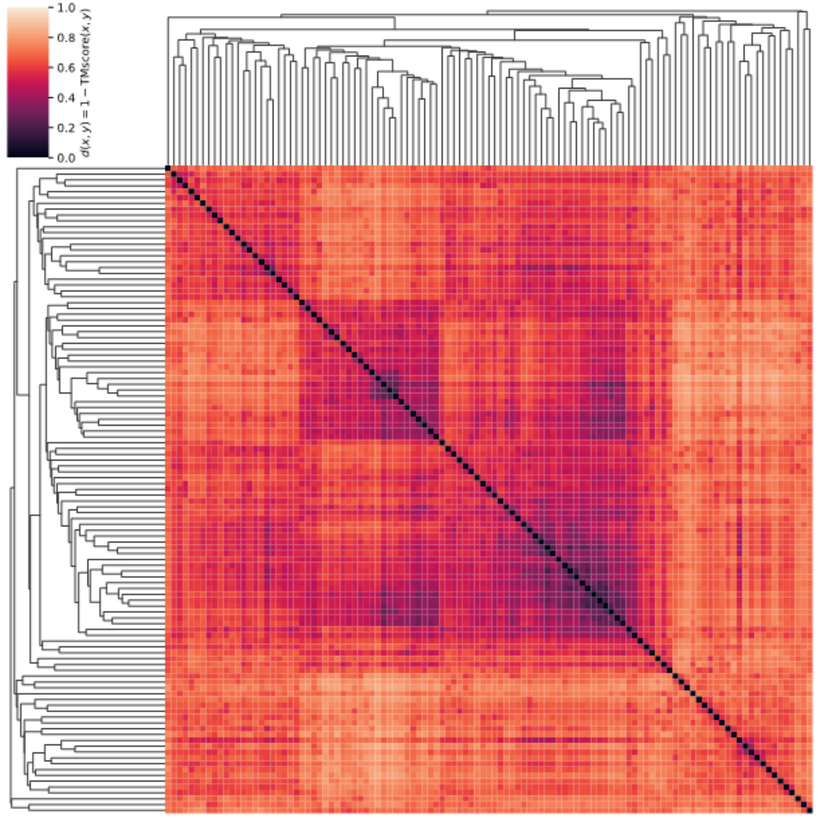


# Generated structures are diverse



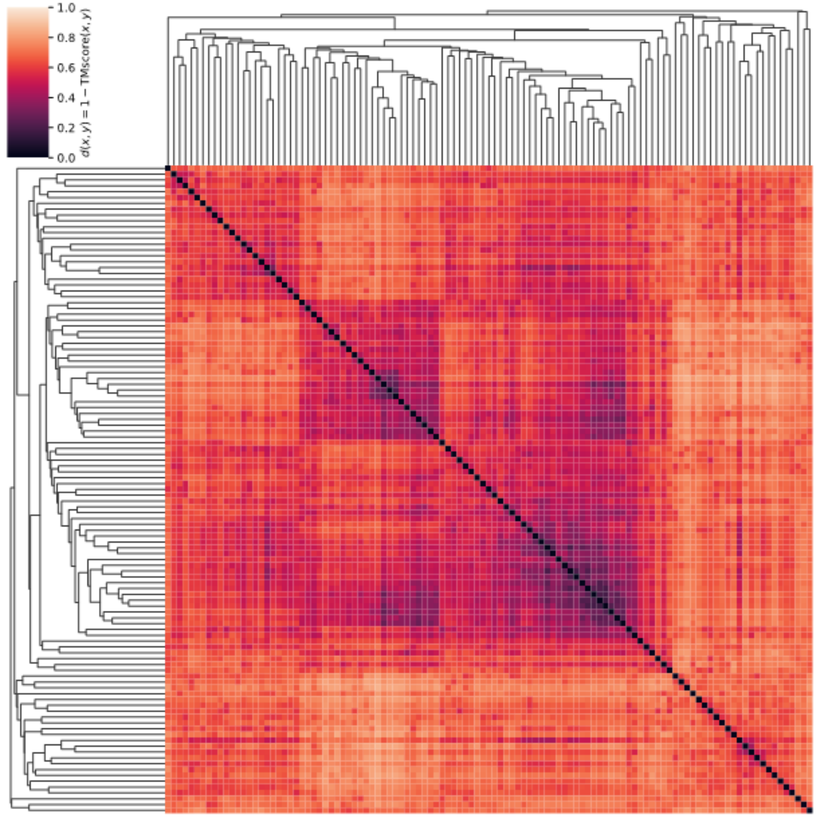
# Generated structures are diverse

FoldingDiff

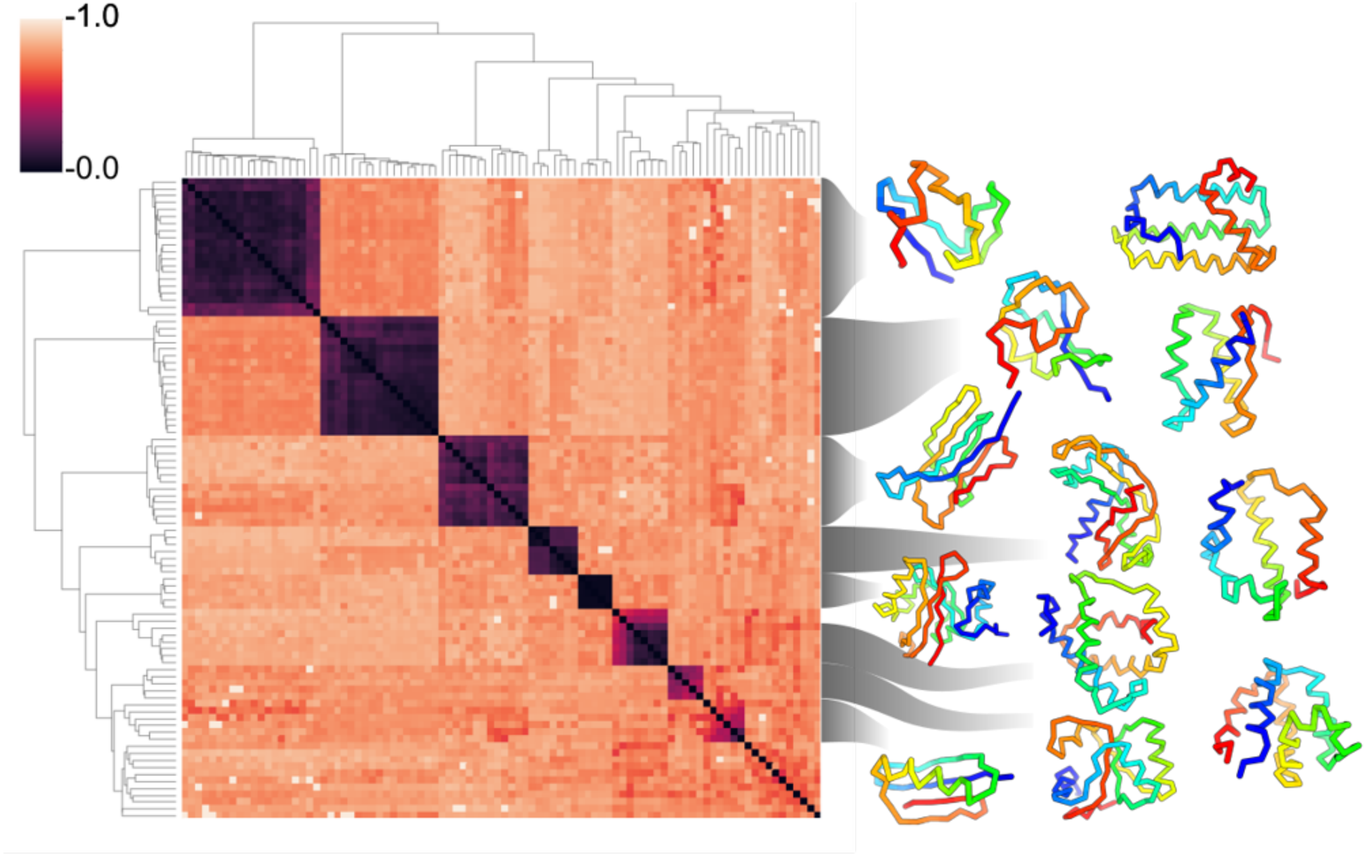


# Generated structures are diverse

FoldingDiff

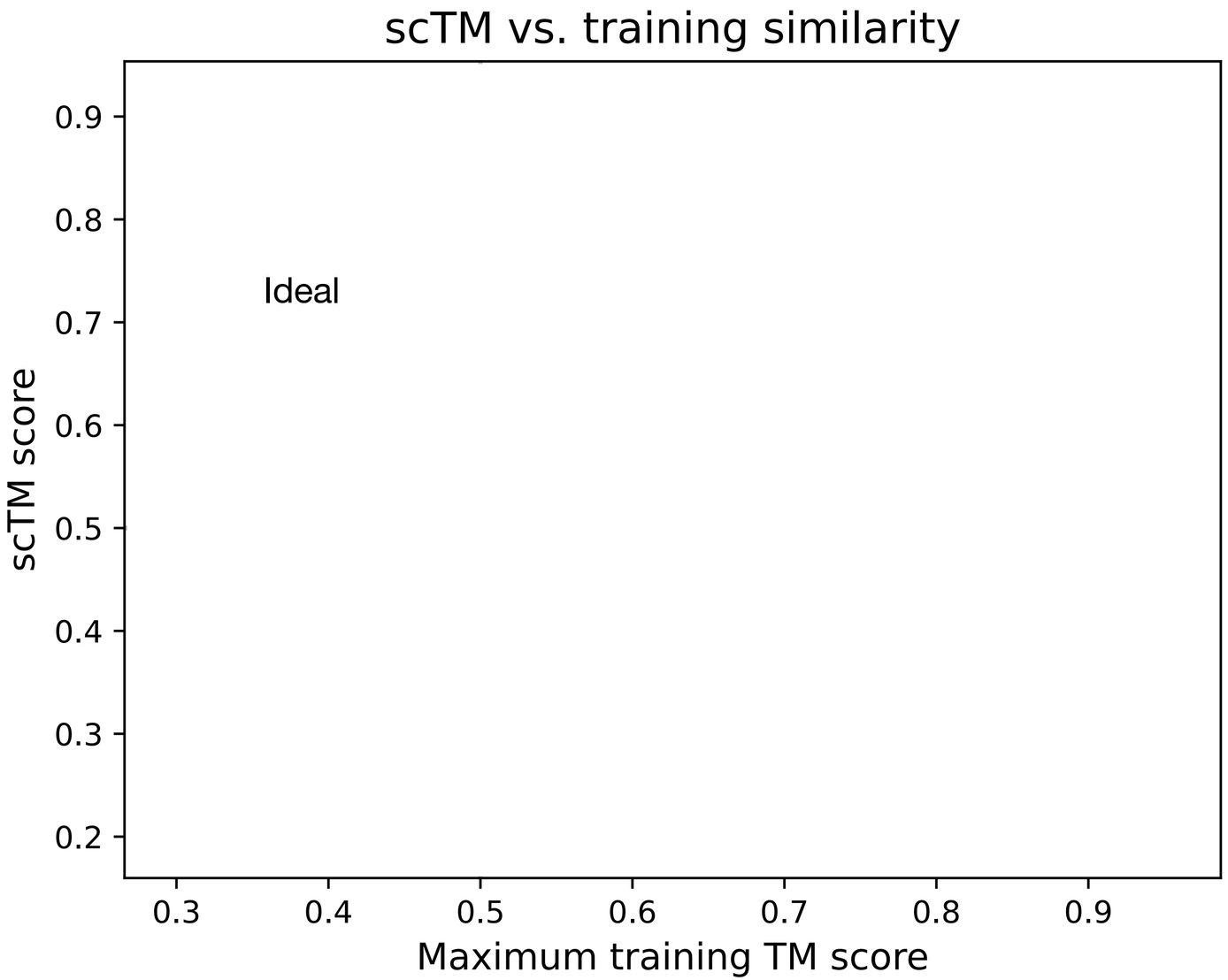



ProtDiff

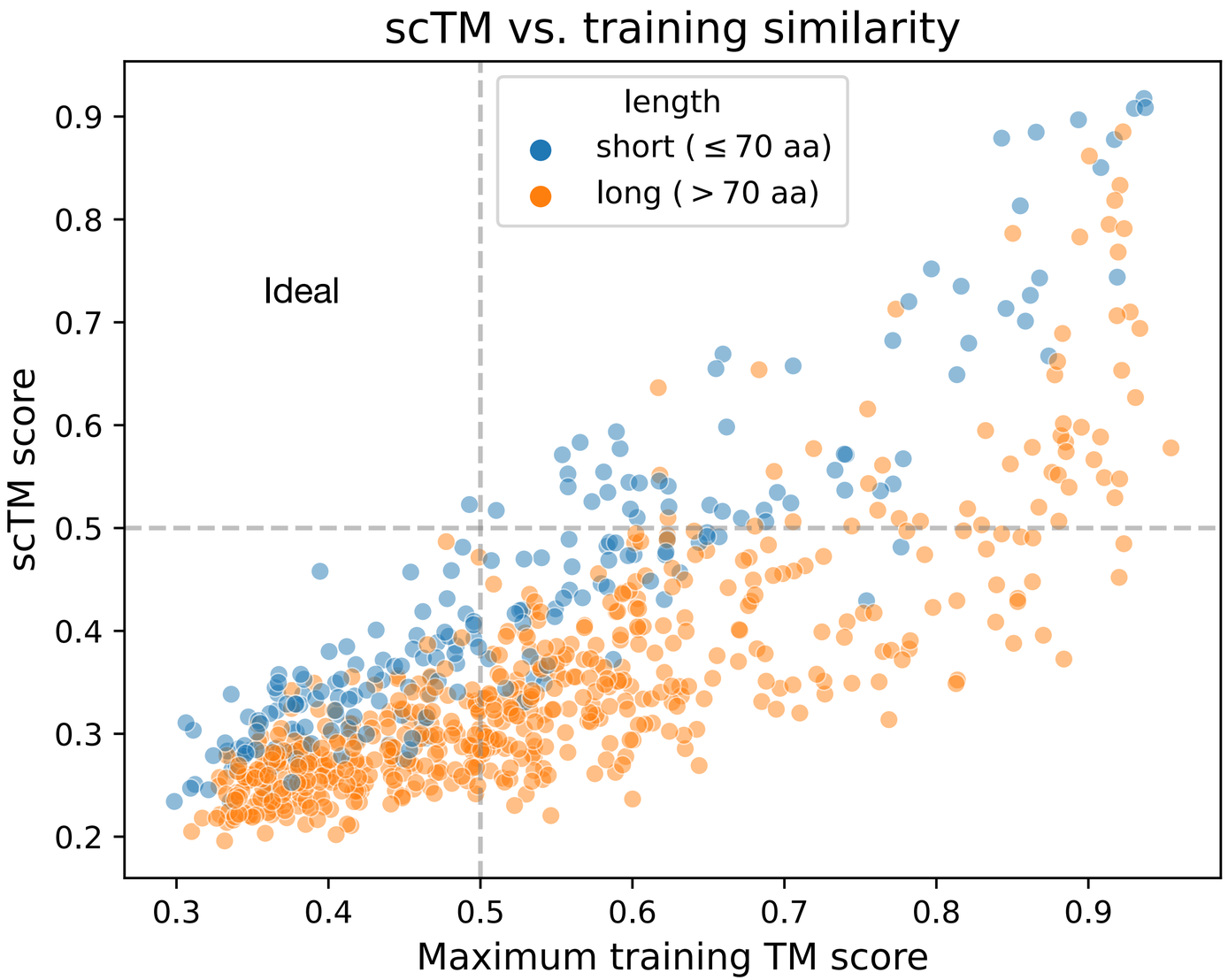


# Generated structures are diverse

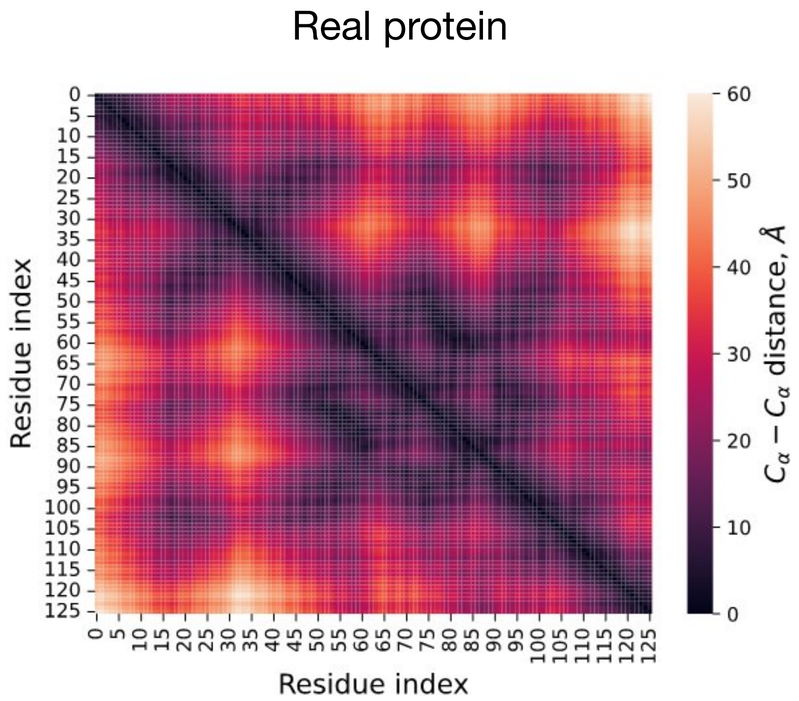
# Generated structures are diverse



# Generated structures are diverse

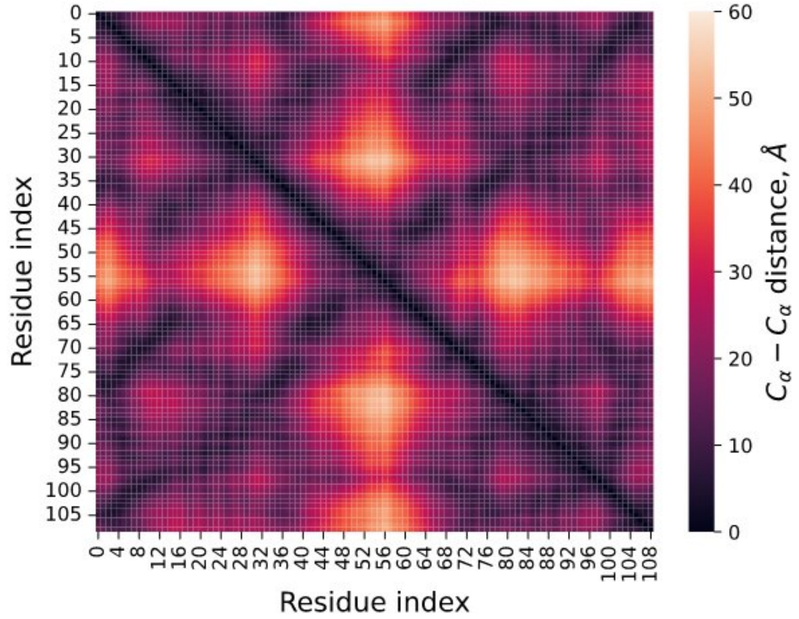


# Angle formulation is enables generation from vanilla transformers

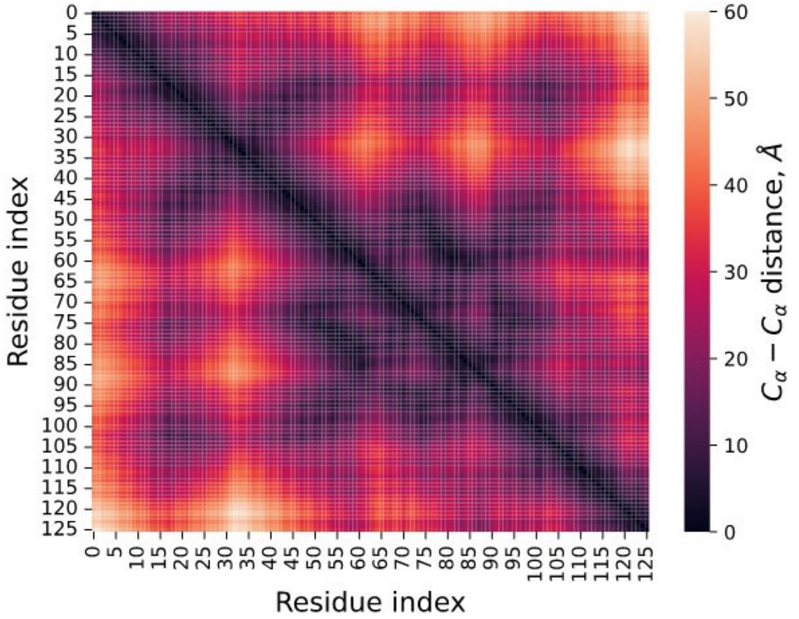


# Angle formulation is enables generation from vanilla transformers

FoldingDiff

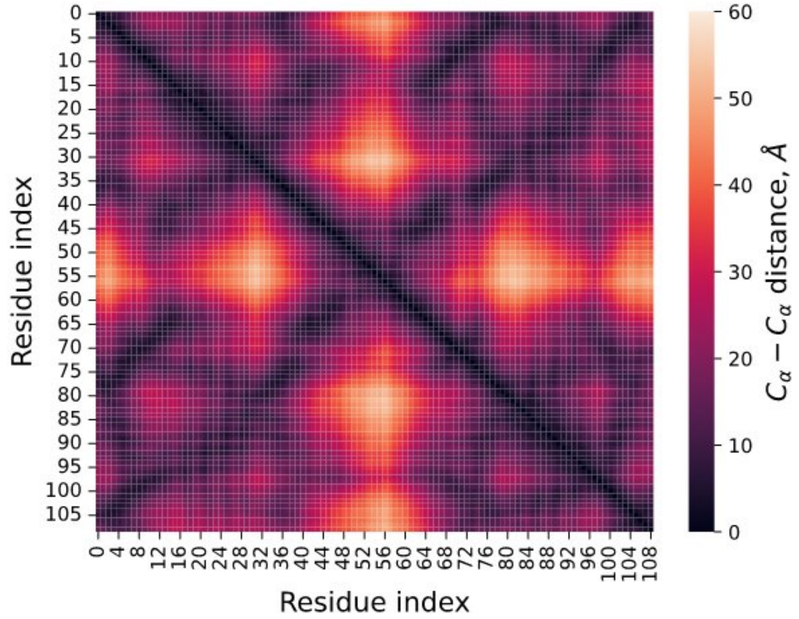



Real protein

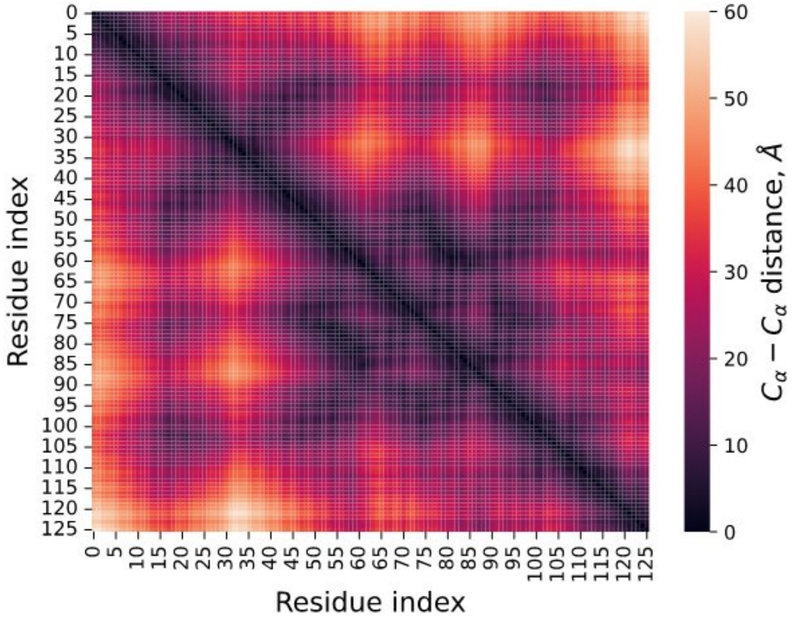


# Angle formulation is enables generation from vanilla transformers

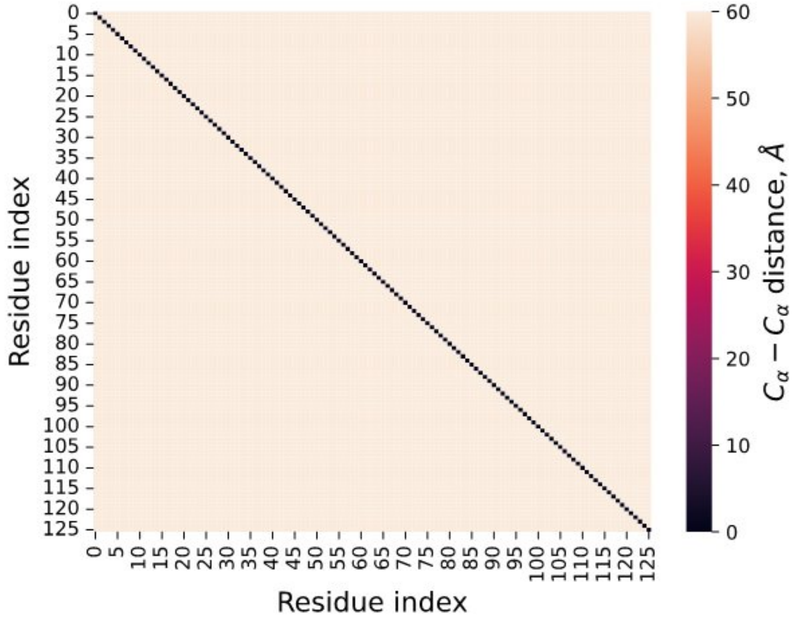
FoldingDiff




Real protein

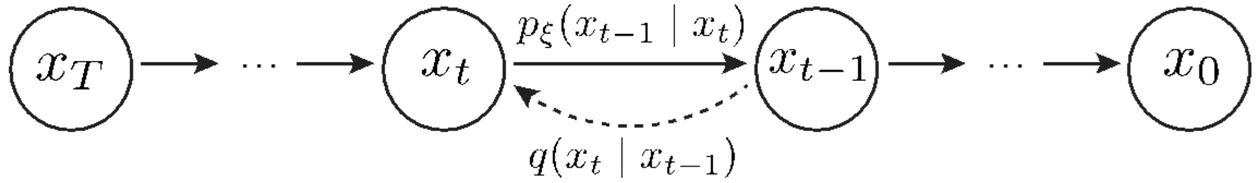


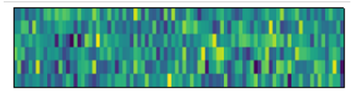

Transformer + point cloud



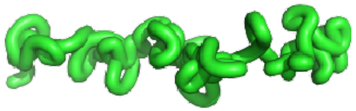
# FoldingDiff is first step towards generating new functions

# FoldingDiff is first step towards generating new functions

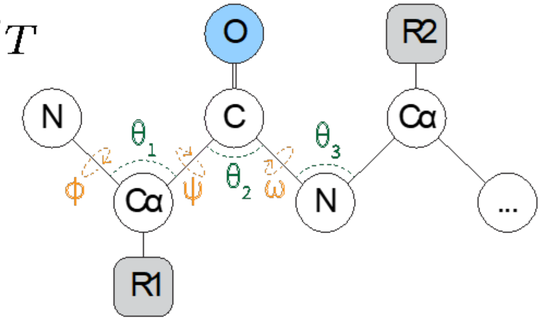


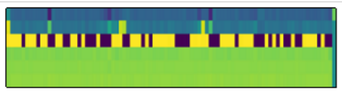
$= x_T$

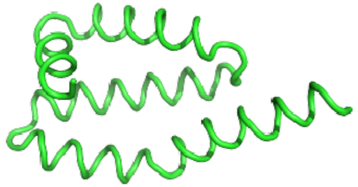


Corresponding structure



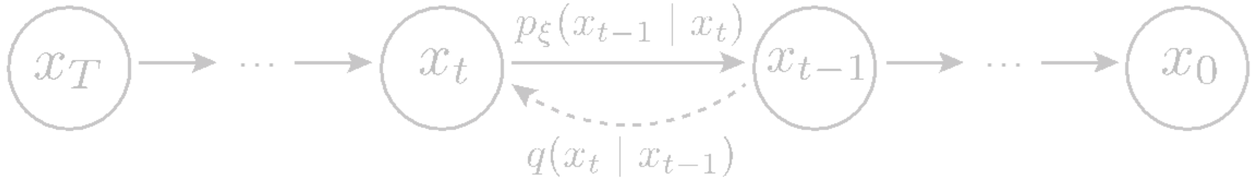

$x_0 =$

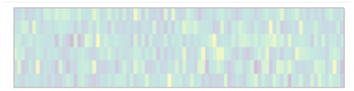




Corresponding structure

# FoldingDiff is first step towards generating new functions



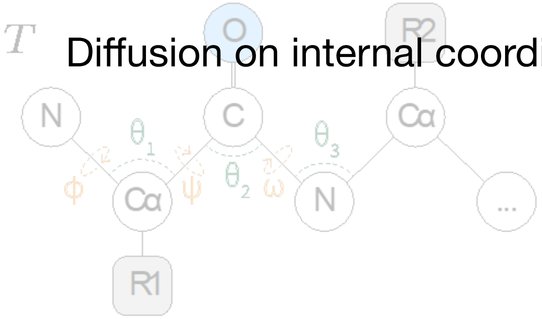



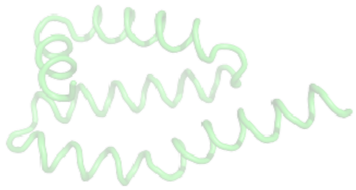
$= x_T$

Diffusion on internal coordinates

$x_0 =$

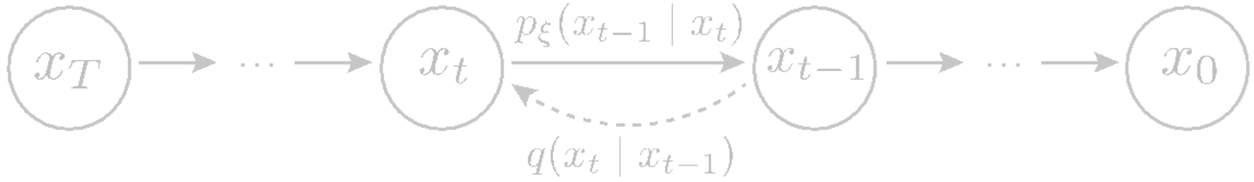
Corresponding structure

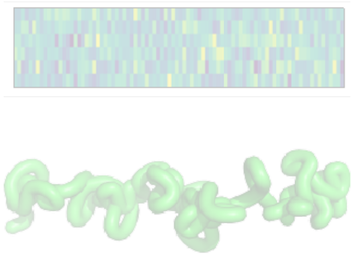


Corresponding structure

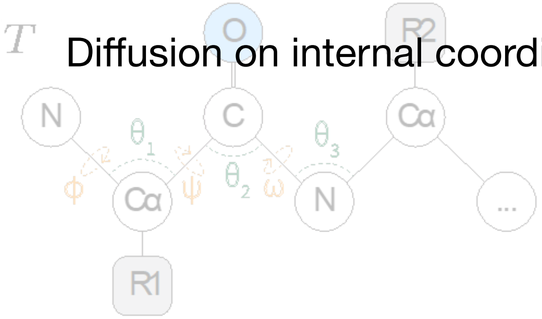
# FoldingDiff is first step towards generating new functions



$= x_T$

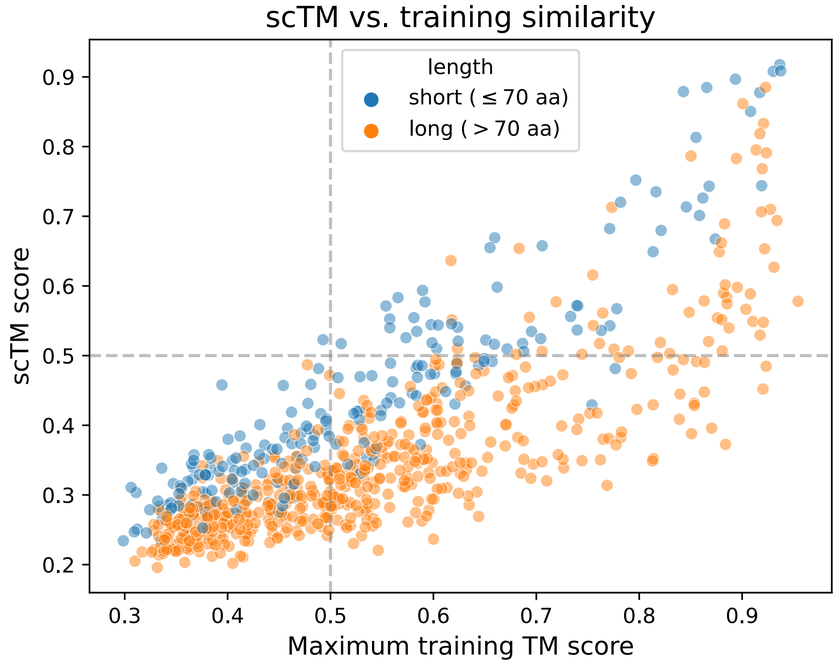
Diffusion on internal coordinates



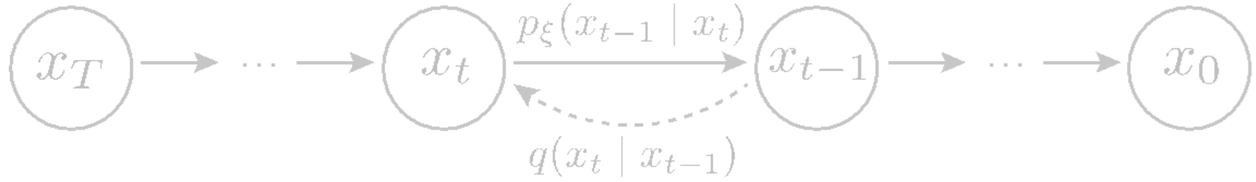

$x_0 =$

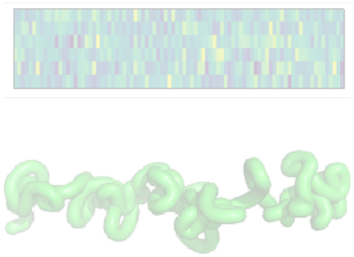
Corresponding structure

Corresponding structure



# FoldingDiff is first step towards generating new functions

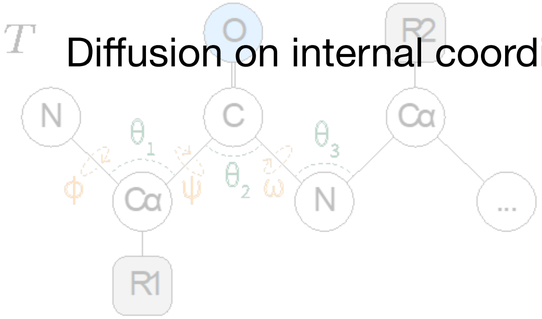


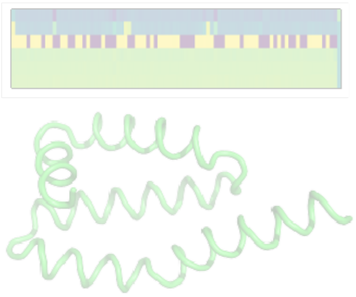
$= x_T$

Corresponding structure

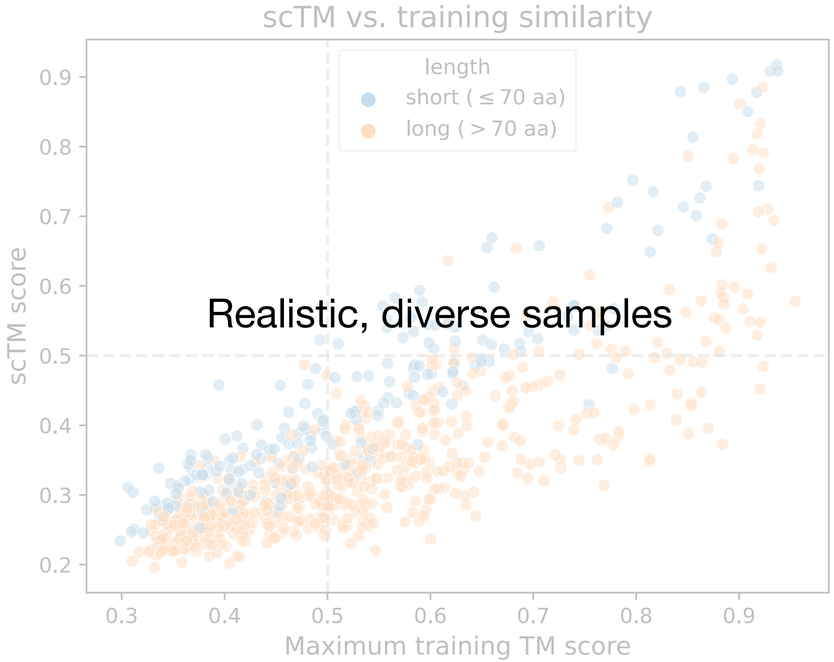
Diffusion on internal coordinates



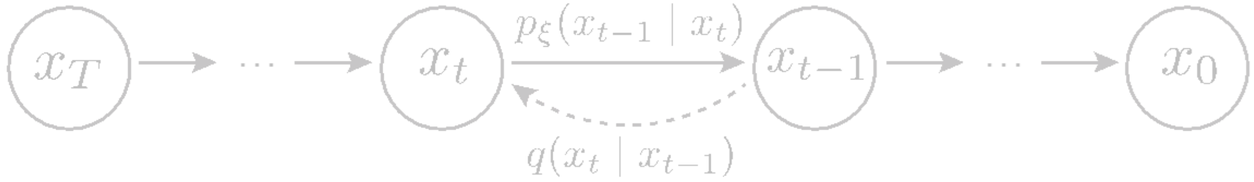

$x_0 =$

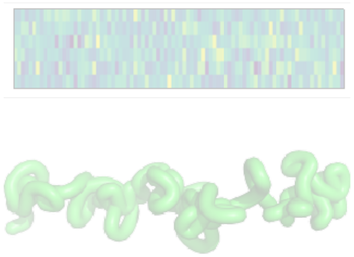


Corresponding structure



# FoldingDiff is first step towards generating new functions

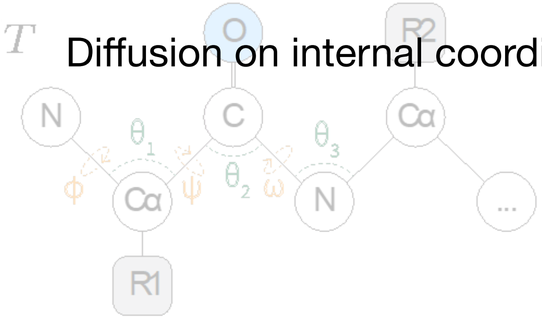


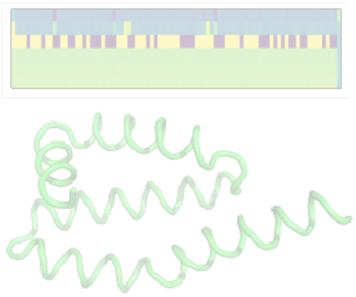
$= x_T$

Corresponding structure

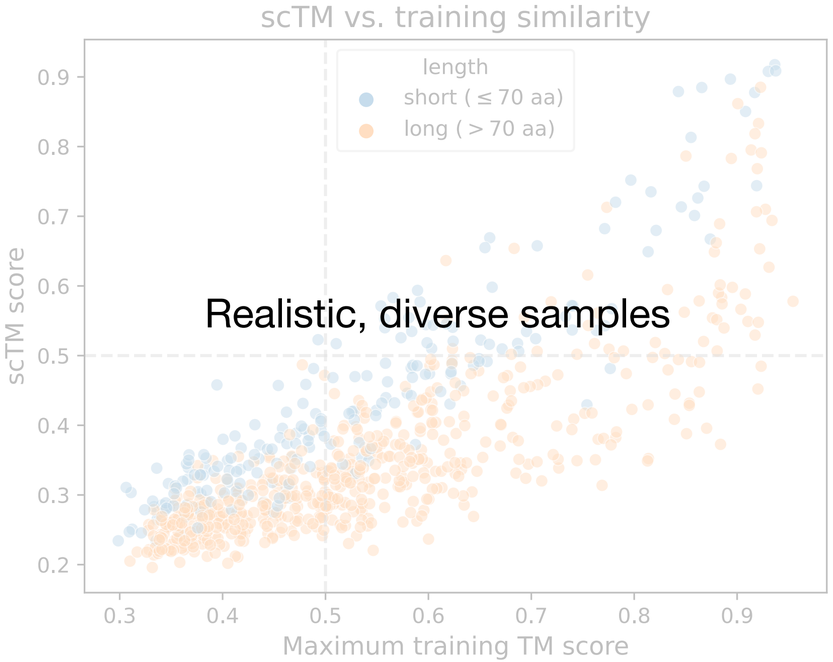
Diffusion on internal coordinates



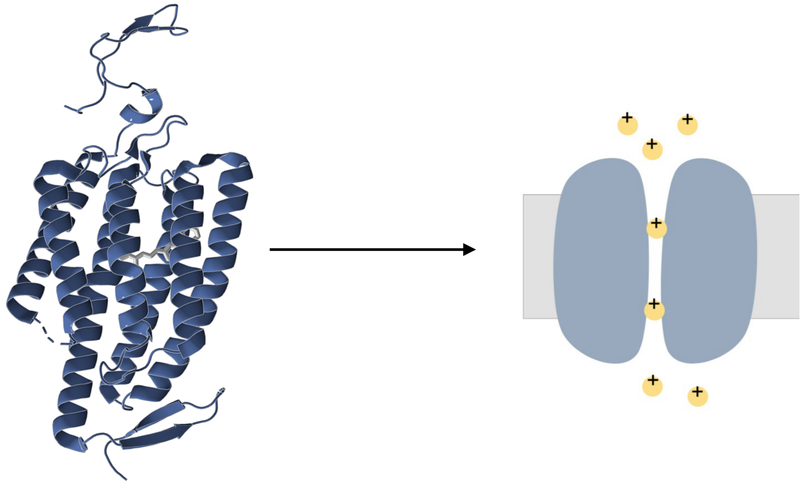

$x_0 =$



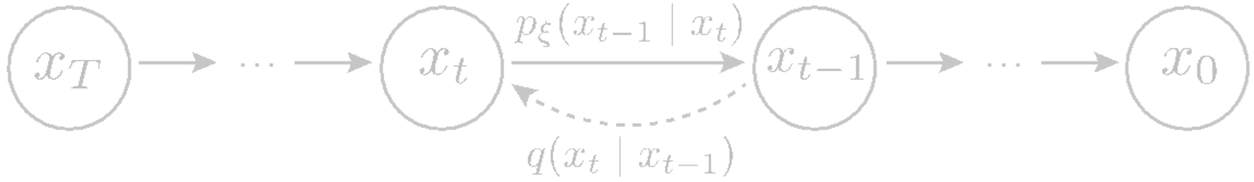
Corresponding structure

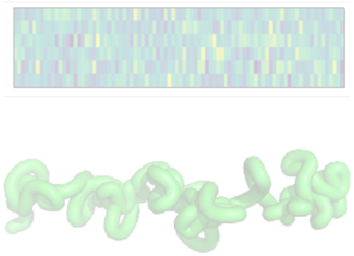



Realistic, diverse samples



# FoldingDiff is first step towards generating new functions

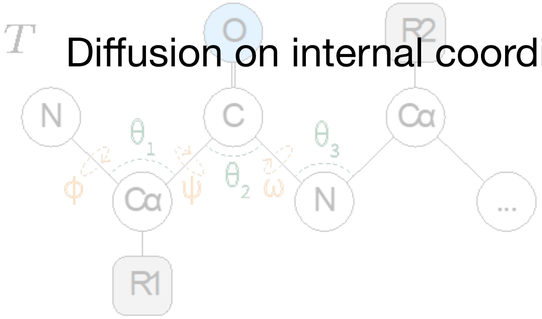


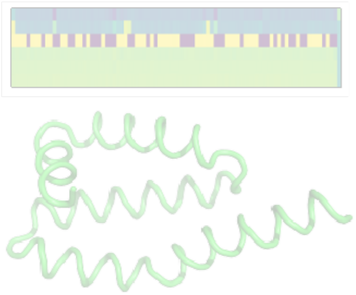
$= x_T$

Corresponding structure

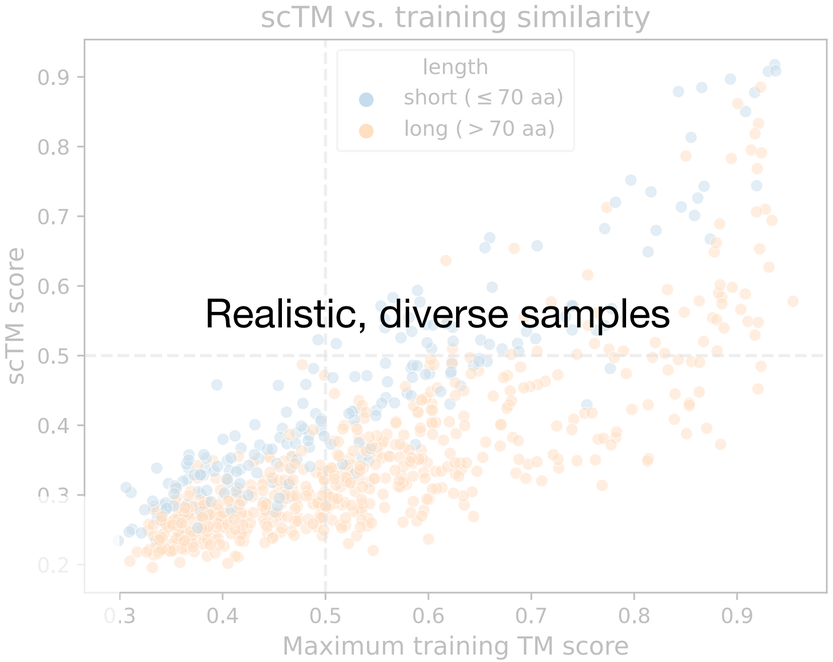
Diffusion on internal coordinates




$x_0 =$



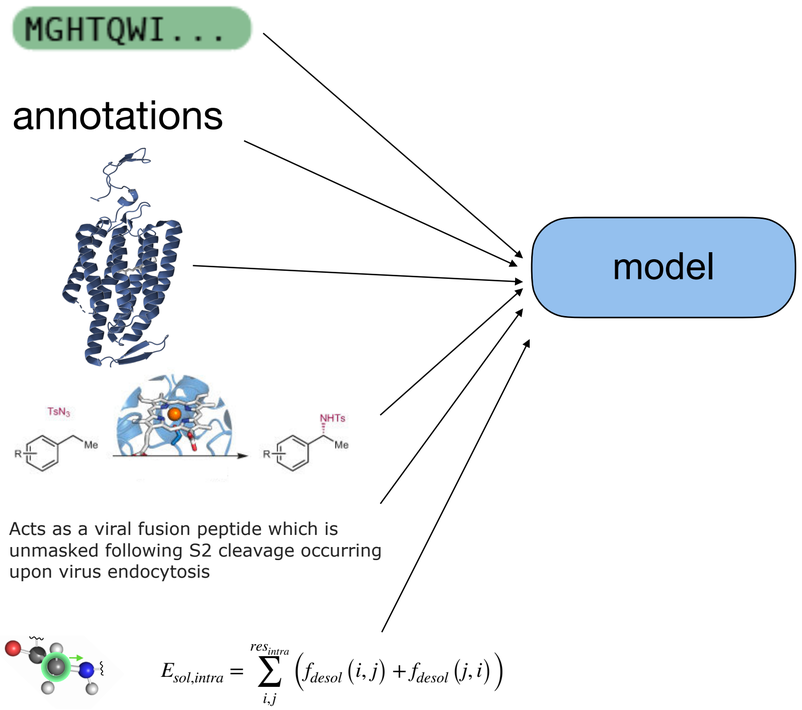
Corresponding structure




Realistic, diverse samples

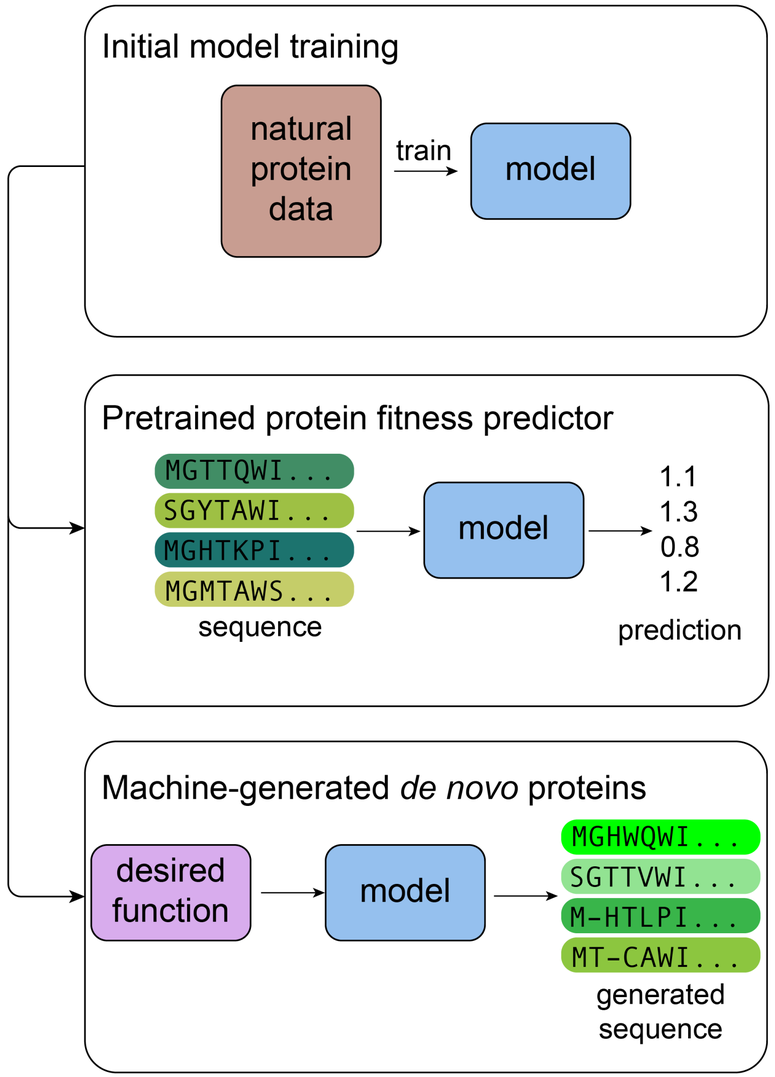
Soon: new function!

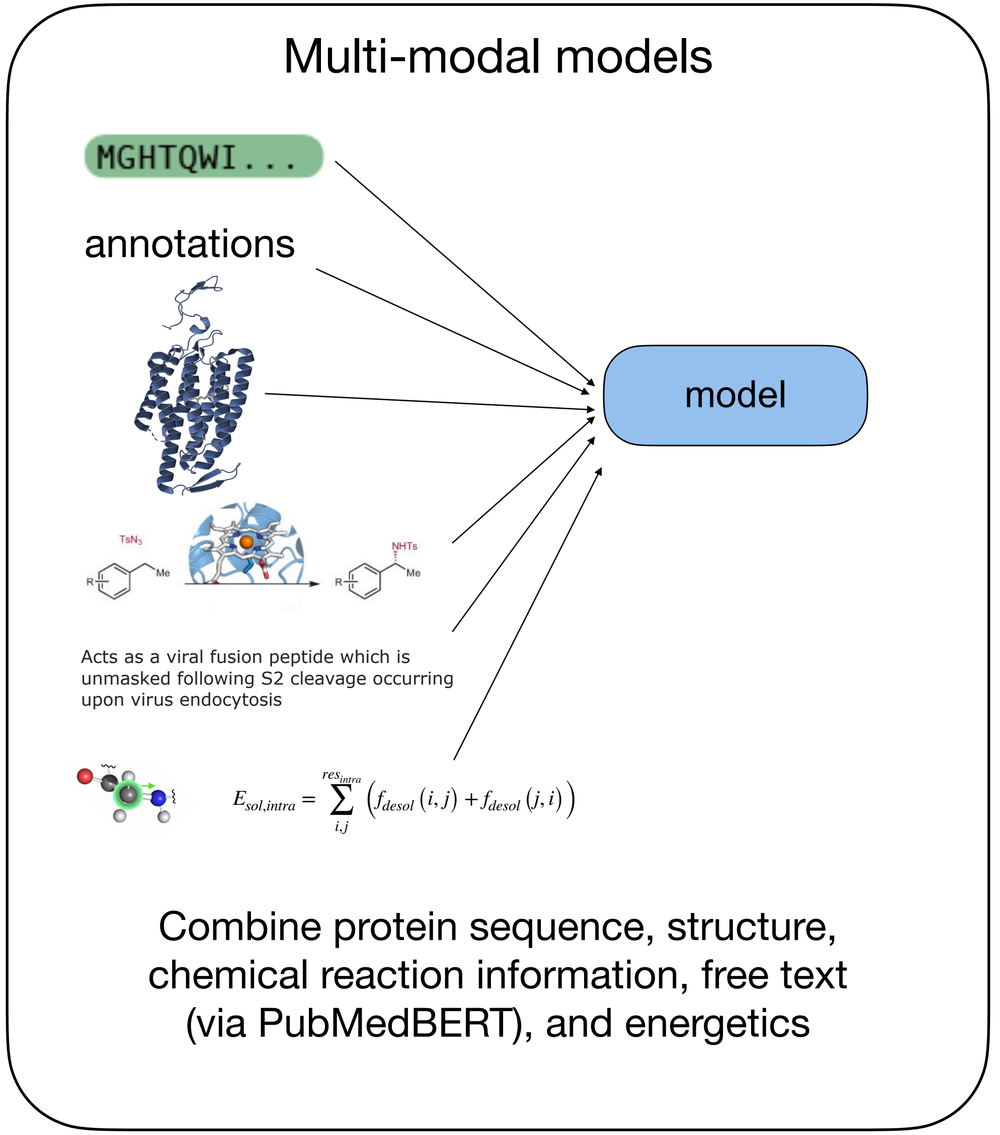
# Multi-modal models




Combine protein sequence, structure, chemical reaction information, free text (via PubMedBERT), and energetics

# Can we predict and generate functional proteins?










BioML at MSR New England

# Acknowledgments

BioML at MSR New England