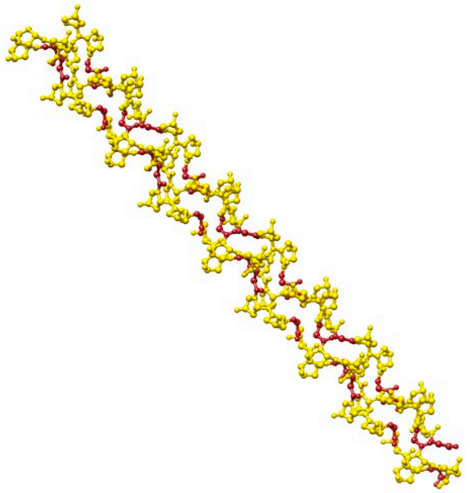
Machine Optimization and Generation of Proteins

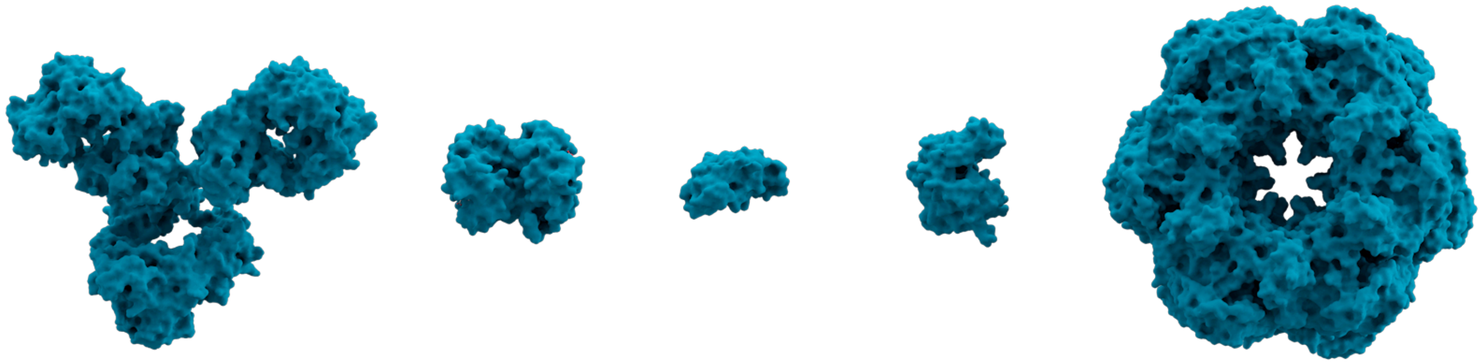
Kevin Kaichuang Yang

Microsoft Research
Caltech

Proteins are biology’s actuators

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

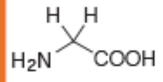




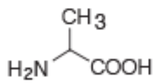
IgG hemoglobin insulin adenylate kinase glutamine synthetase

Diversity arises from 20 building blocks

SMALL

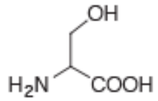


Glycine (Gly, G)
MW: 75.07

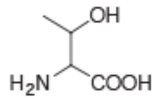


Alanine (Ala, A)
MW: 89.09

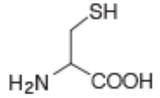
NUCLEOPHILIC



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

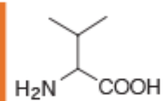


Threonine (Thr, T)
MW: 119.1, $pK_a \sim 16$

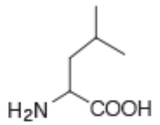


Cysteine (Cys, C)
MW: 121.2, $pK_a = 8.18$

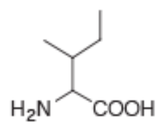
HYDROPHOBIC



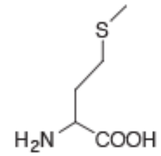
Valine (Val, V)
MW: 117.1



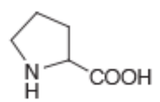
Leucine (Leu, L)
MW: 131.2



Isoleucine (Ile, I)
MW: 131.2

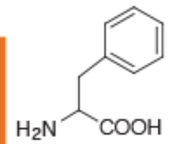


Methionine (Met, M)
MW: 149.2

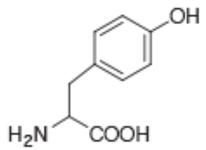


Proline (Pro, P)
MW: 115.1

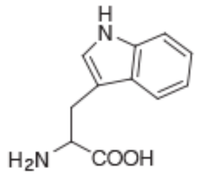
AROMATIC



Phenylalanine (Phe, F)
MW: 165.2

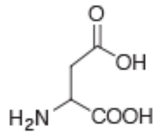


Tyrosine (Tyr, Y)
MW: 181.2, $pK_a = 10.46$

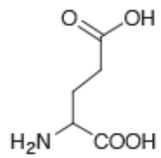


Tryptophan (Trp, W)
MW: 204.2

ACIDIC

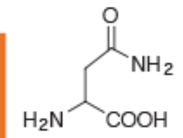


Aspartic Acid (Asp, D)
MW: 133.1, $pK_a = 3.9$

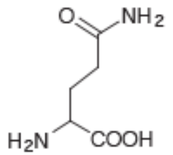


Glutamic Acid (Glu, E)
MW: 147.1, $pK_a = 4.07$

AMIDE

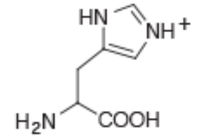


Asparagine (Asn, N)
MW: 132.1

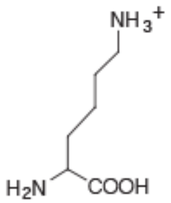


Glutamine (Gln, Q)
MW: 146.1

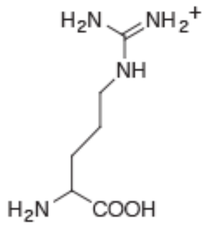
BASIC



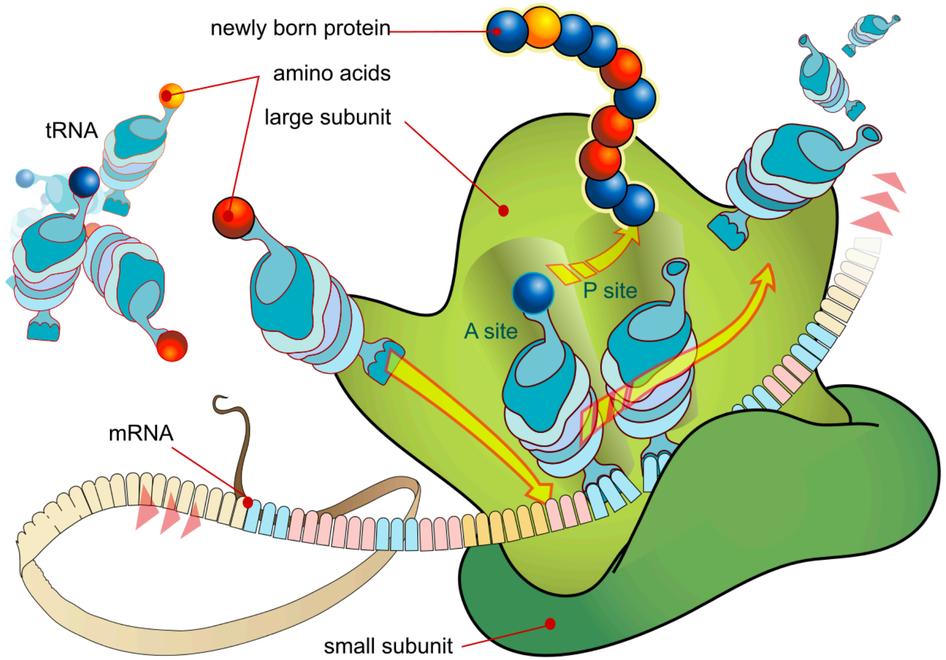
Histidine (His, H)
MW: 155.2, $pK_a = 6.04$



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



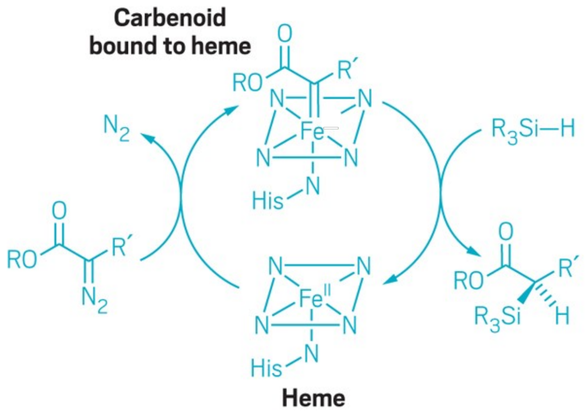
Arginine (Arg, R)
MW: 174.2, $pK_a = 12.48$



Why design proteins?

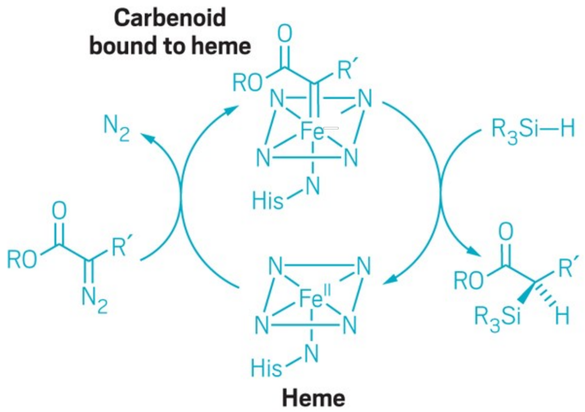
Why design proteins?

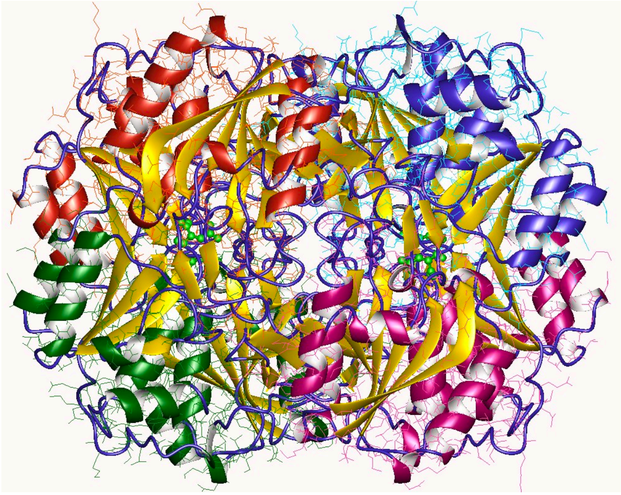
- New chemistry



Why design proteins?

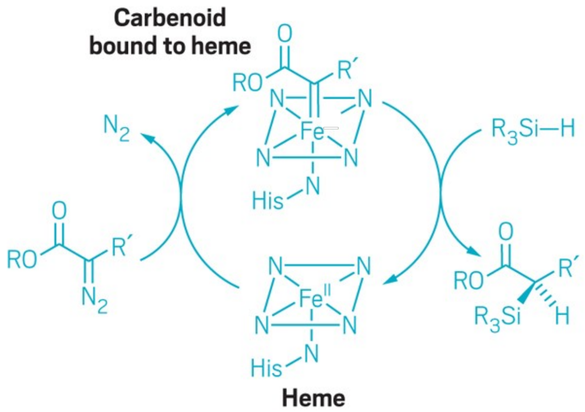
- New chemistry
- Therapeutics

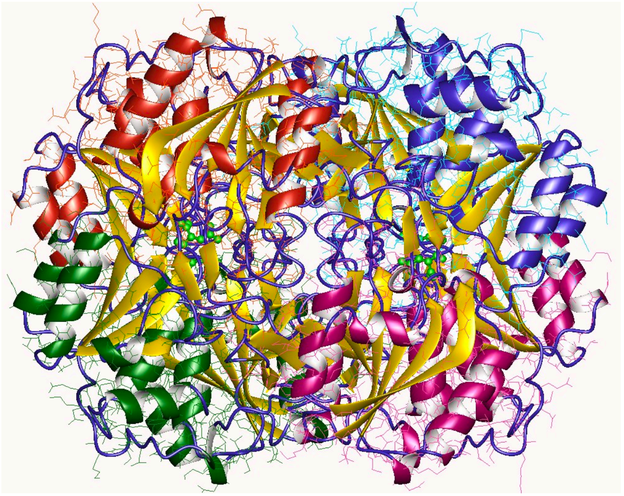


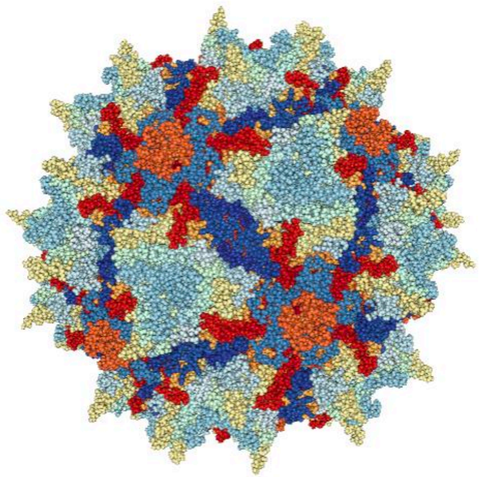


Why design proteins?

- New chemistry
- Therapeutics
- To learn how they work

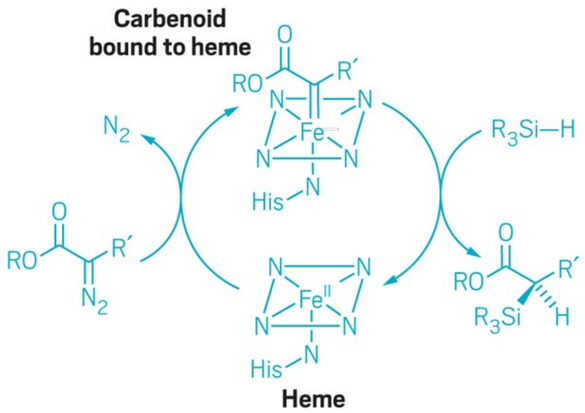


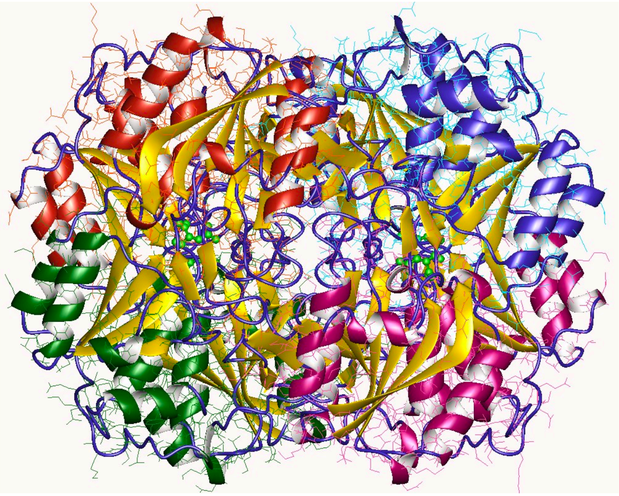


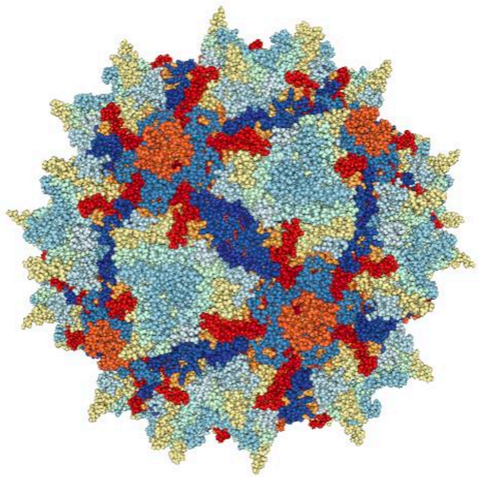


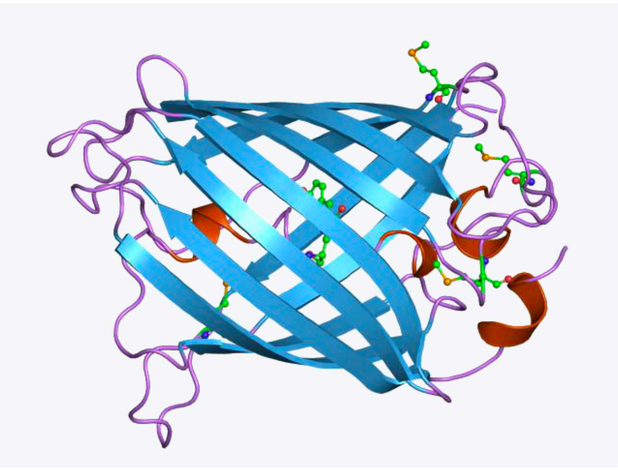
Why design proteins?

- New chemistry
- Therapeutics
- To learn how they work
- Molecular tools









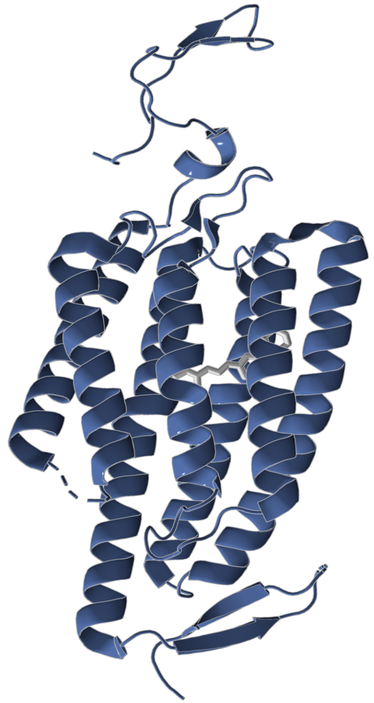
The protein design problem

The protein design problem

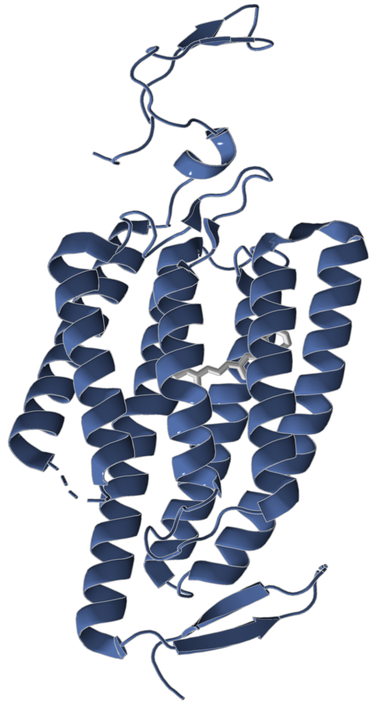
MGTGDHDD...

The protein design problem

MGTGDHDD...



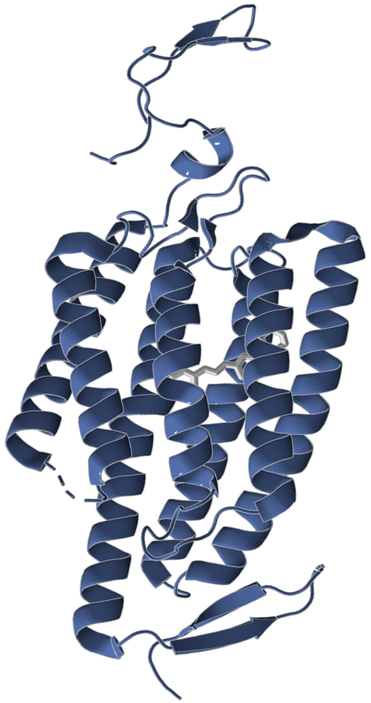
The protein design problem

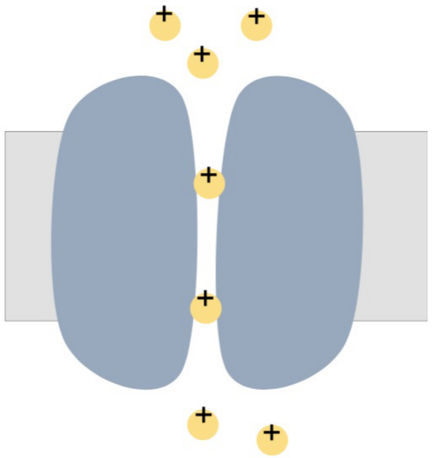
MGTGDHDD... ? 

The protein design problem

MGTGDHDD...

?

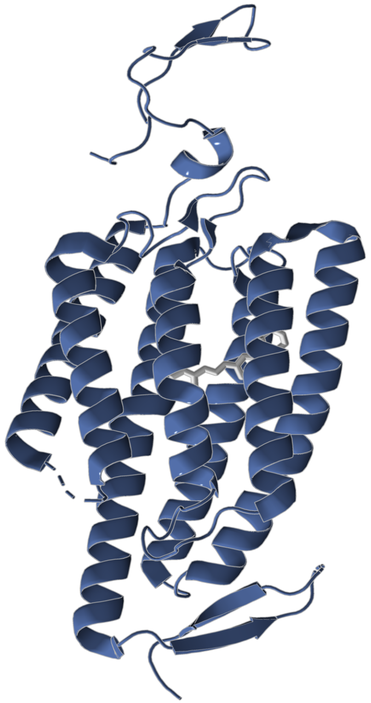




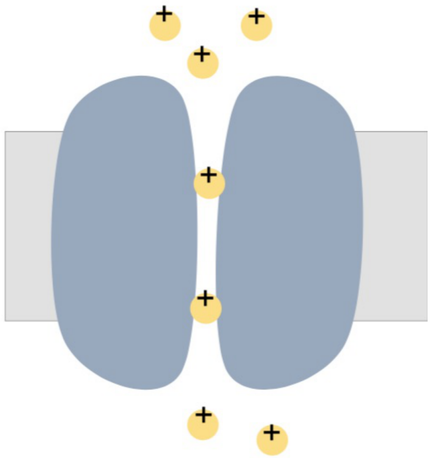
The protein design problem

MGTGDHDD...

?

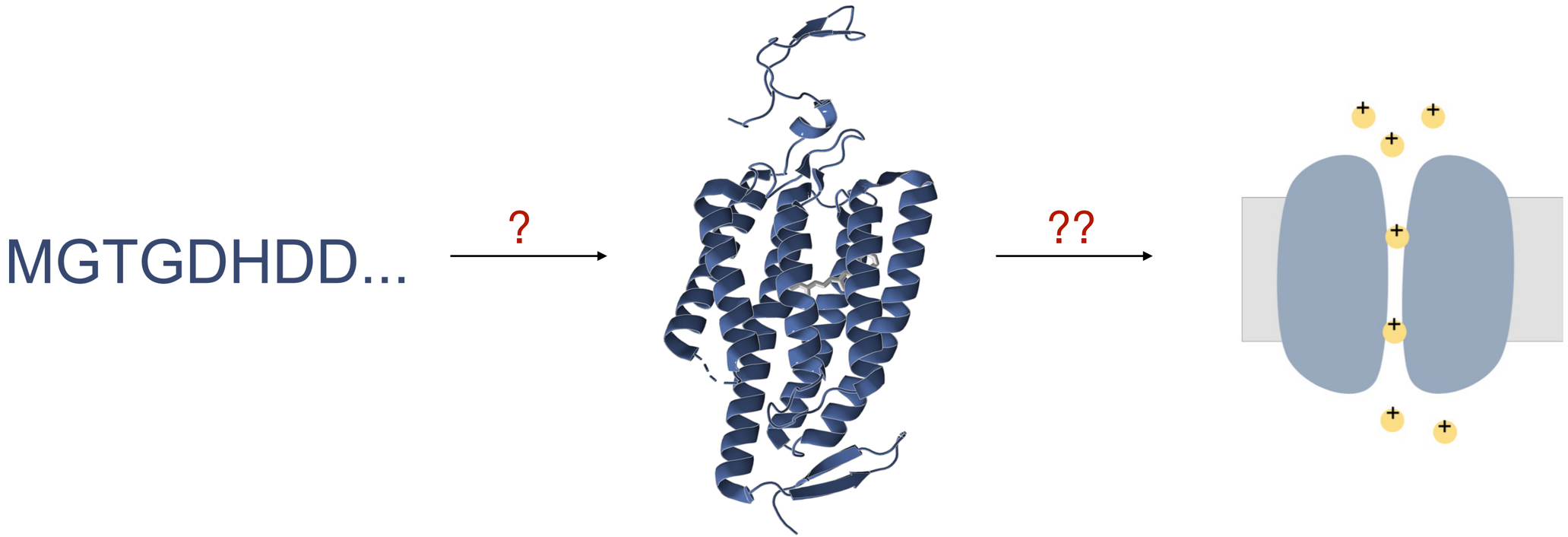


??



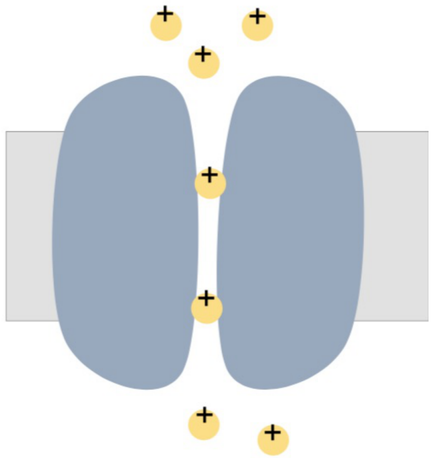
The protein design problem

MGTGDHDD...



What sequence will give the desired function?

The protein design problem



???

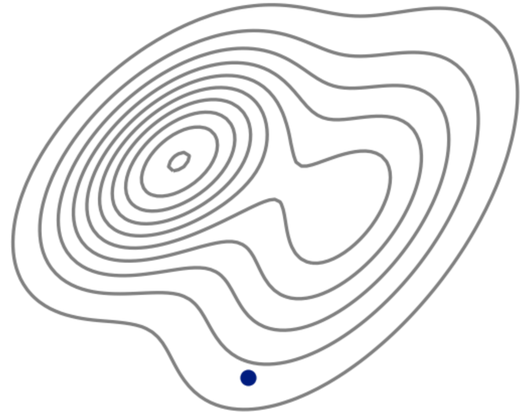
MGTGDHDD...

What sequence will give the desired function?

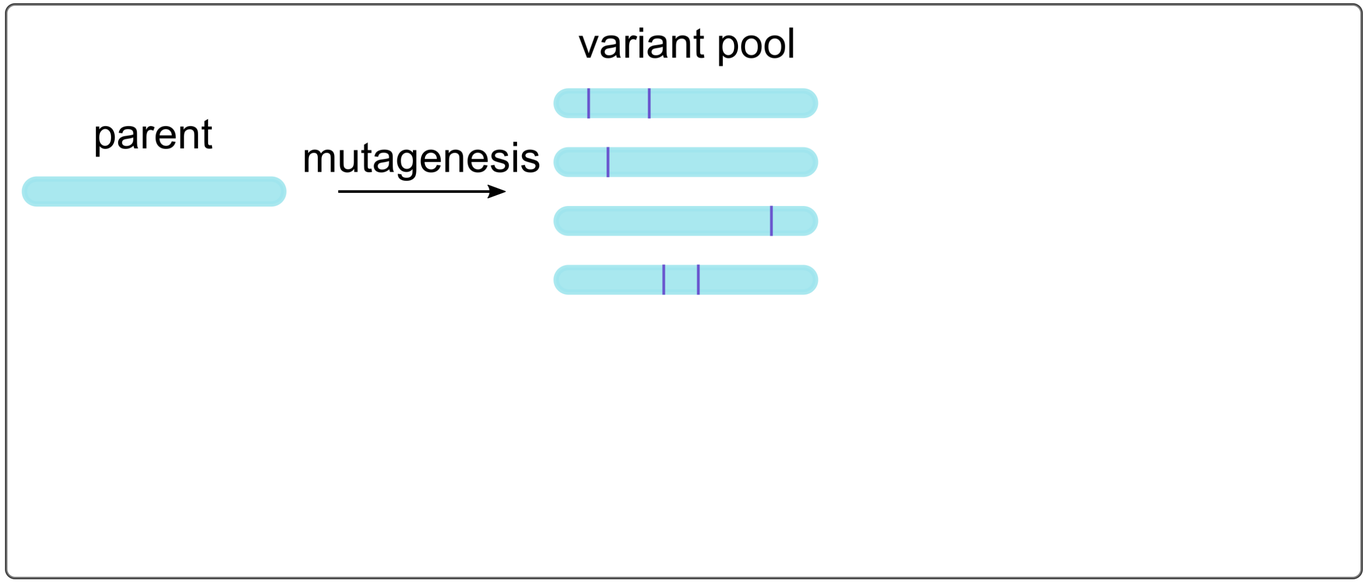
Directed evolution sidesteps the problem

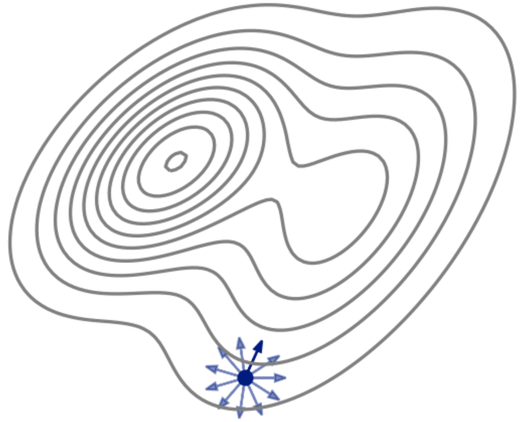
Directed evolution sidesteps the problem

parent

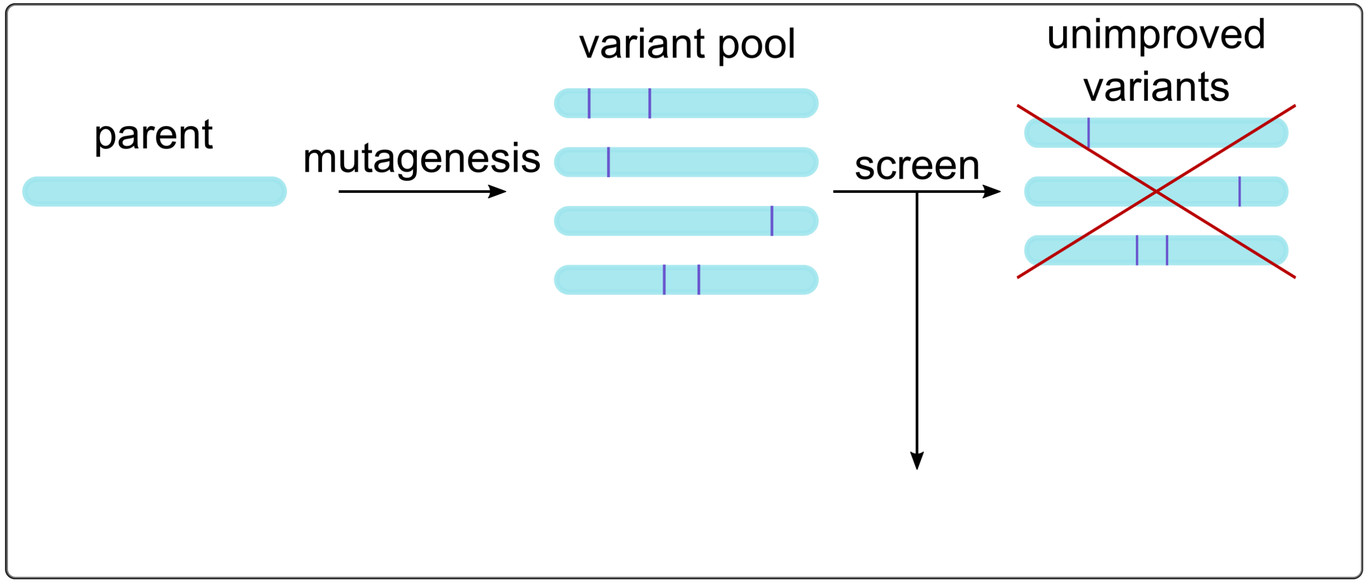


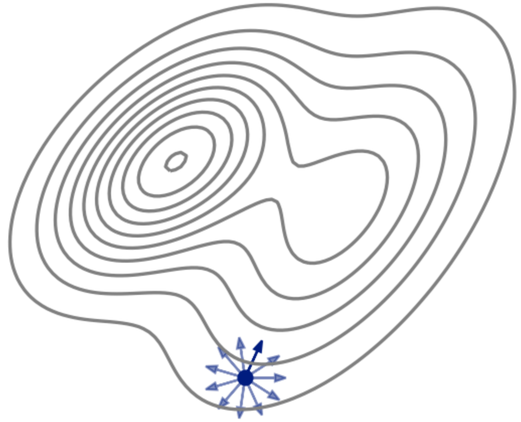
Directed evolution sidesteps the problem



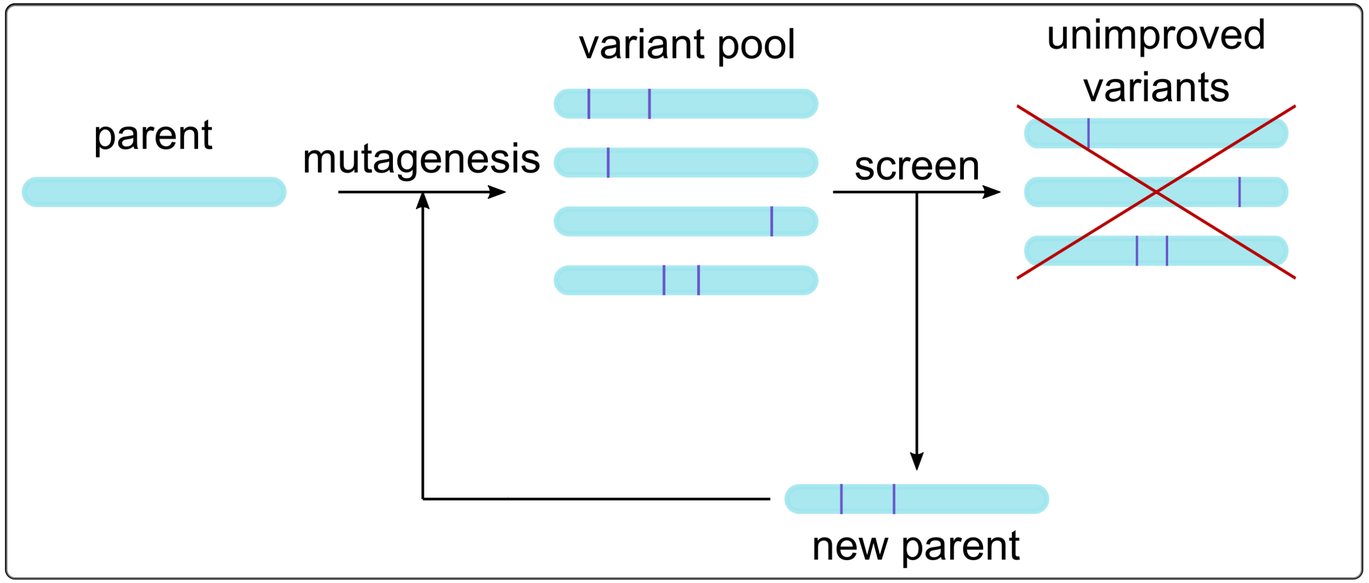


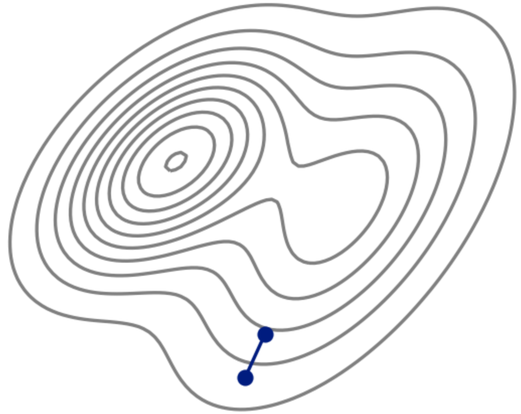
Directed evolution sidesteps the problem



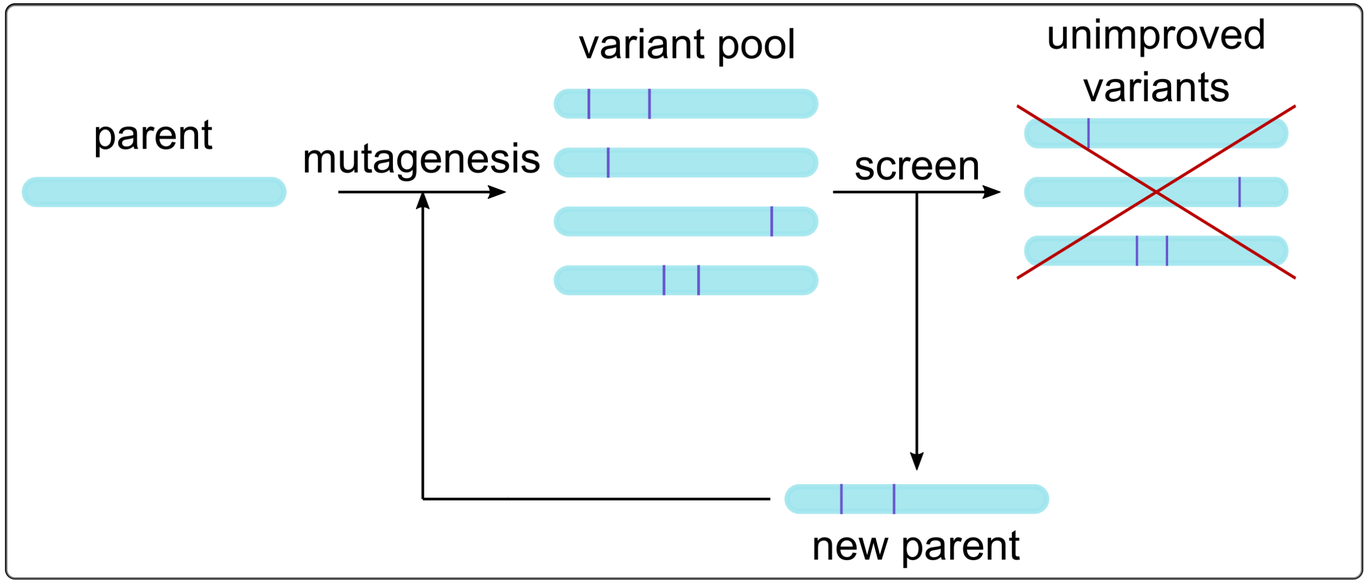


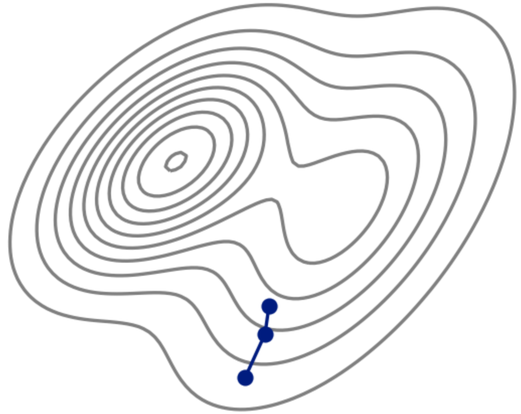
Directed evolution sidesteps the problem



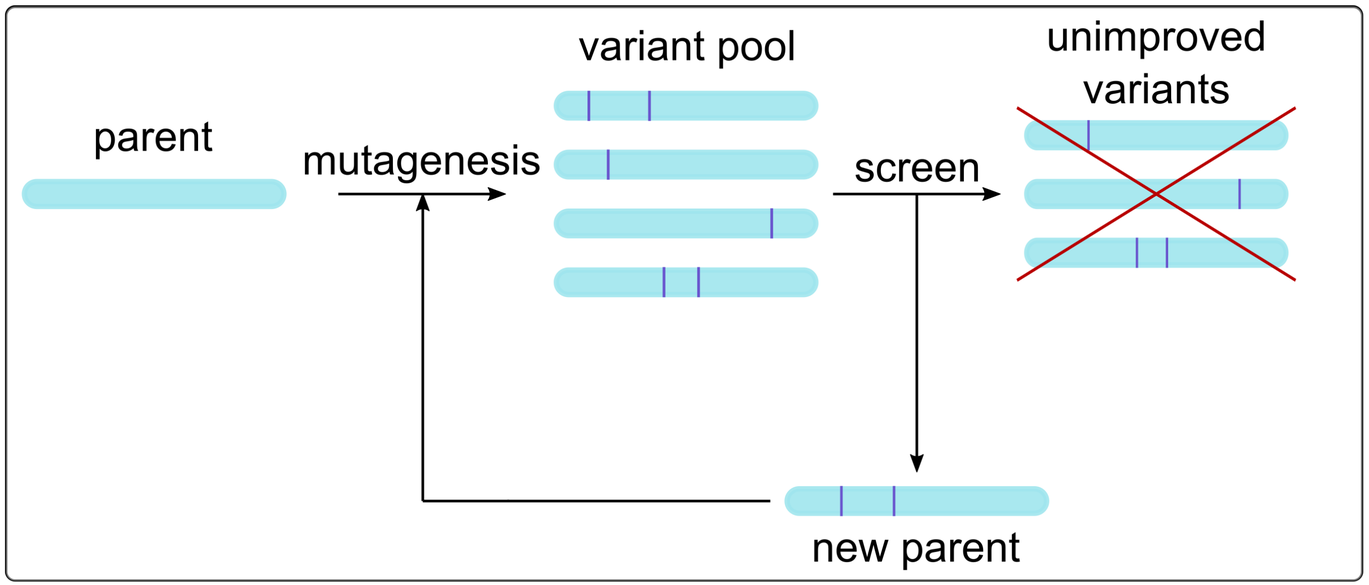


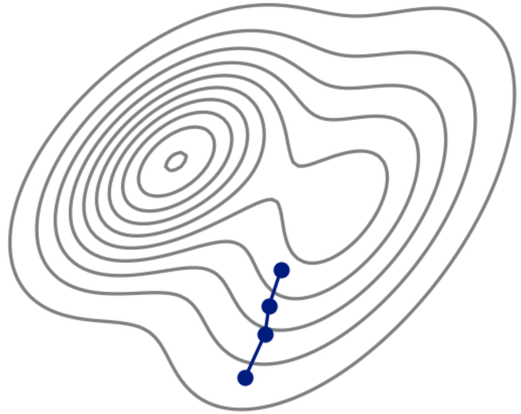
Directed evolution sidesteps the problem



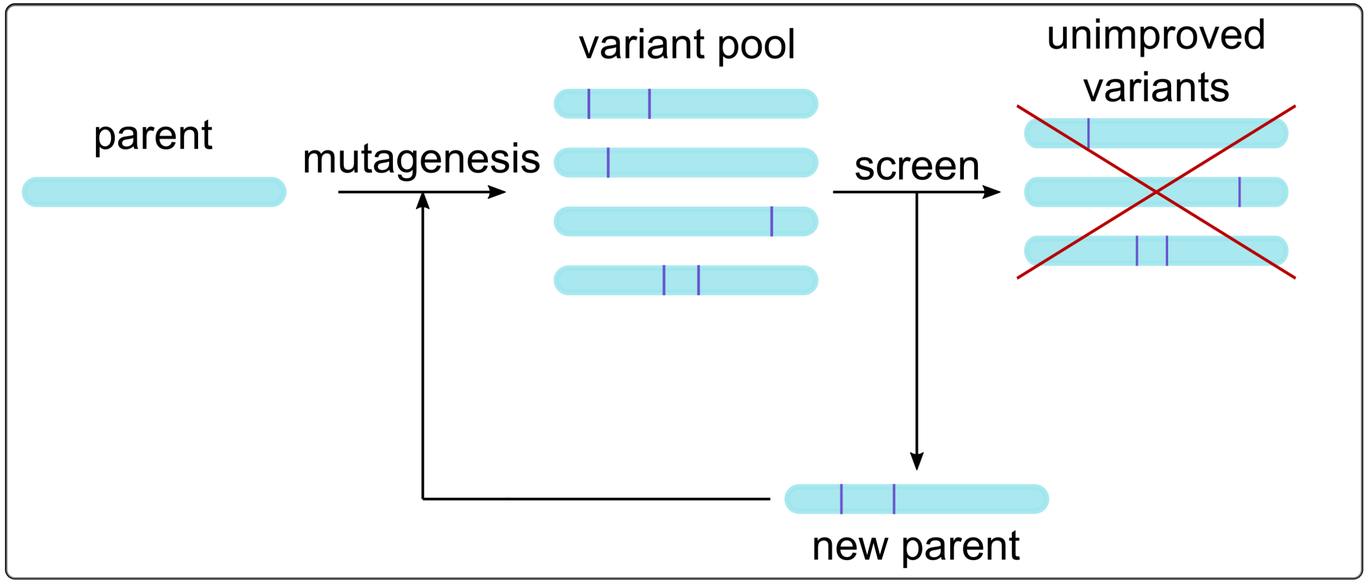


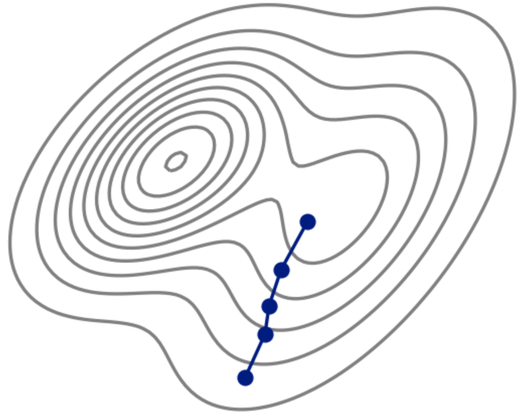
Directed evolution sidesteps the problem



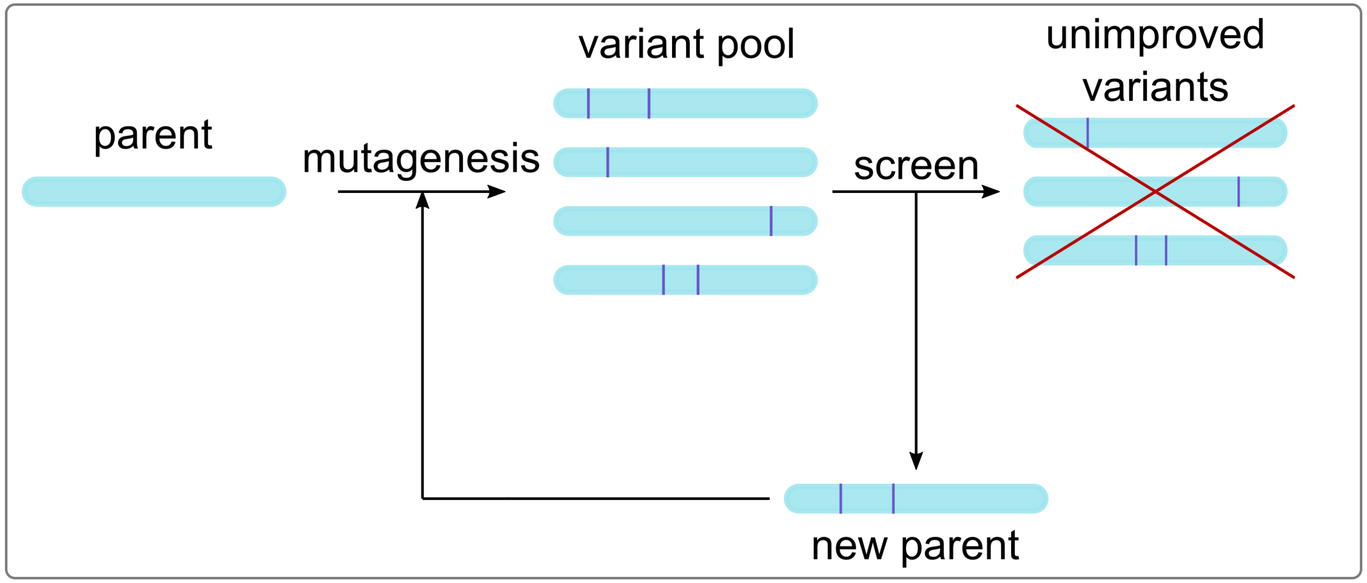


Directed evolution sidesteps the problem

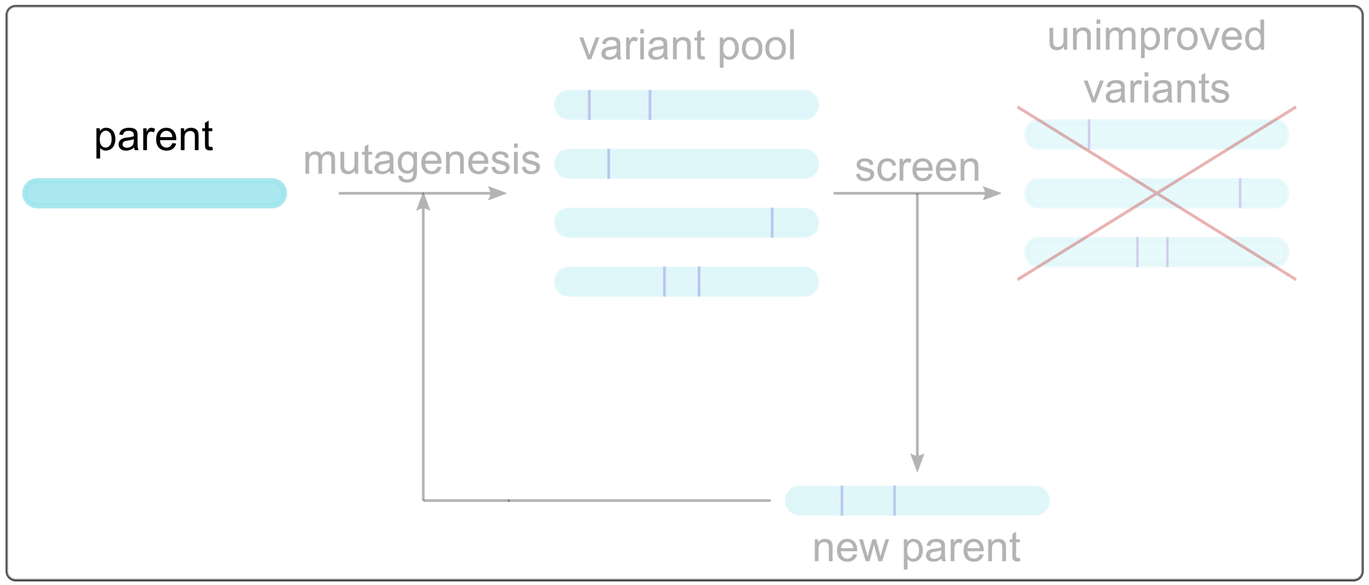




Requirements for directed evolution

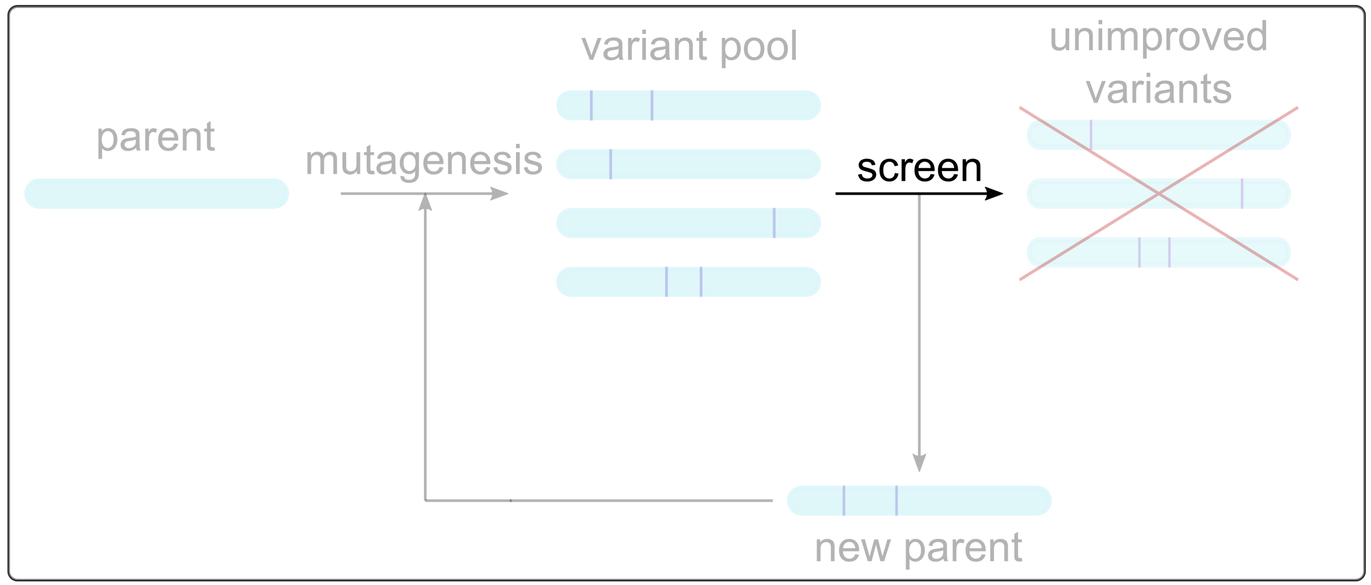


Requirements for directed evolution



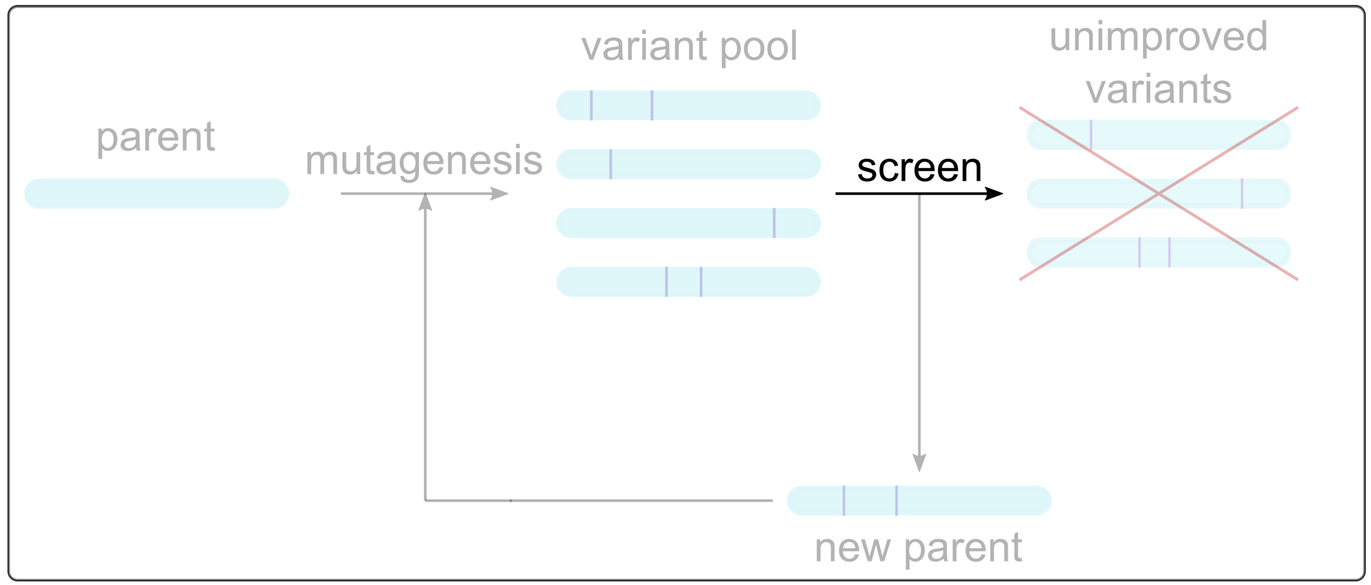
- Parent

Requirements for directed evolution



- Parent
- High-throughput screen (>100 / wk)

Requirements for directed evolution



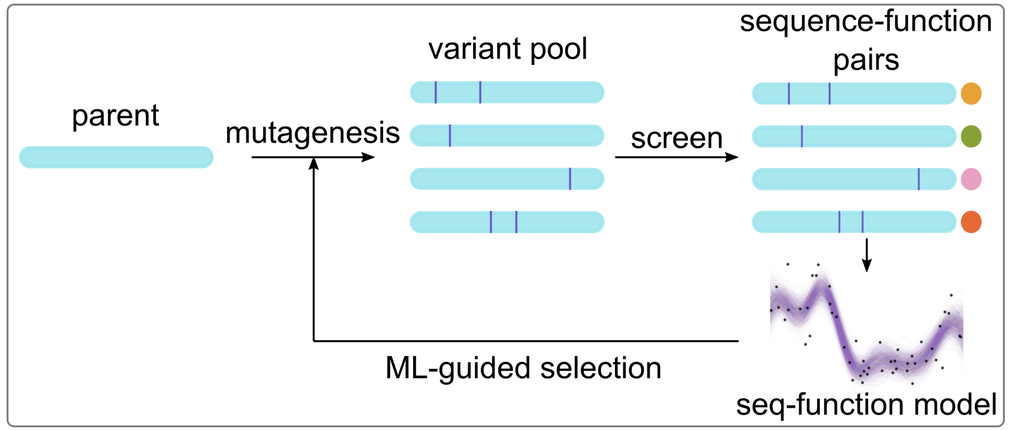
- Parent
- High-throughput screen (>100 / wk)

Sometimes screening is hard
Or we have no parent!

Optimization vs generation

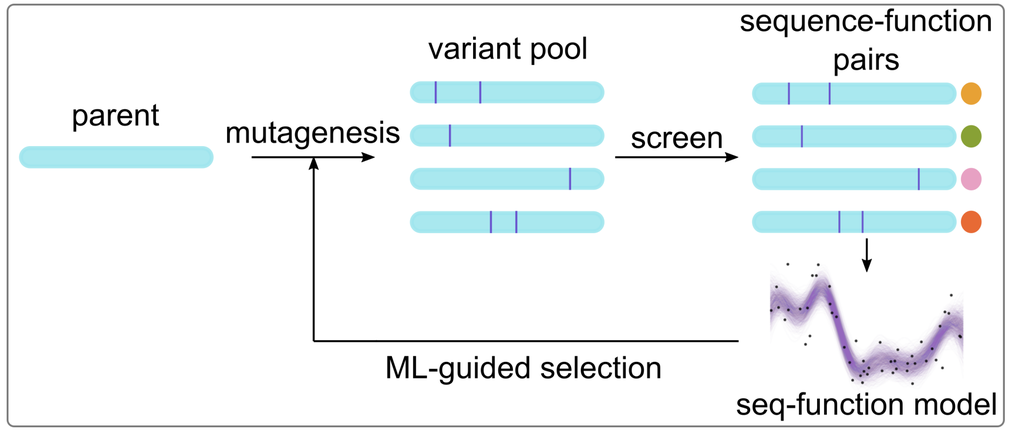
Optimization vs generation

Machine-learning guided directed evolution



Optimization vs generation

Machine-learning guided directed evolution



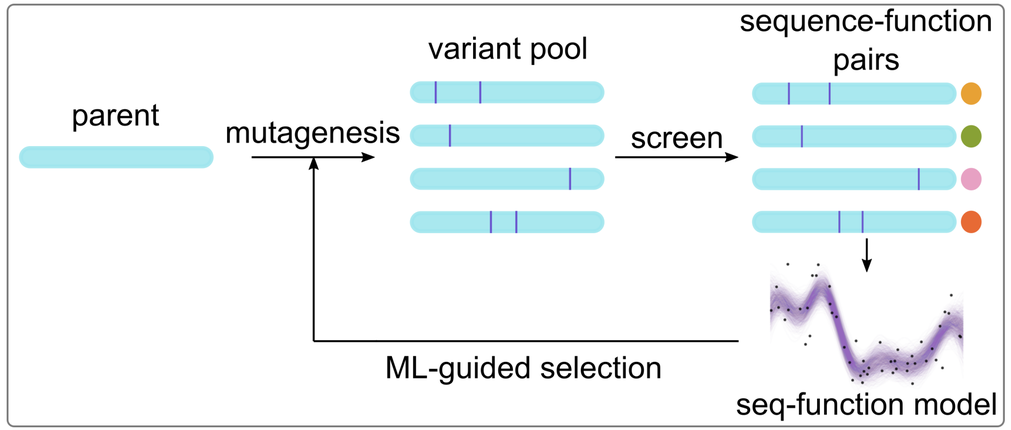
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

Optimization vs generation

Machine-learning guided directed evolution



Designer channelrhodopsins

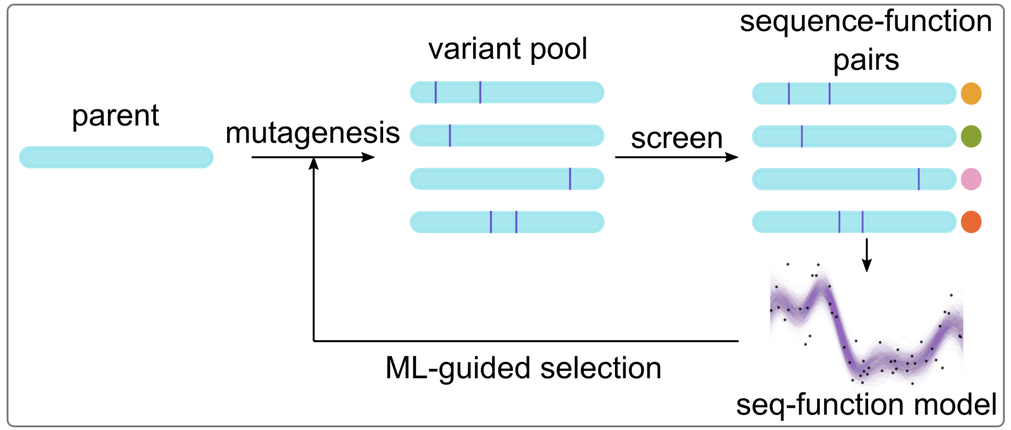
Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

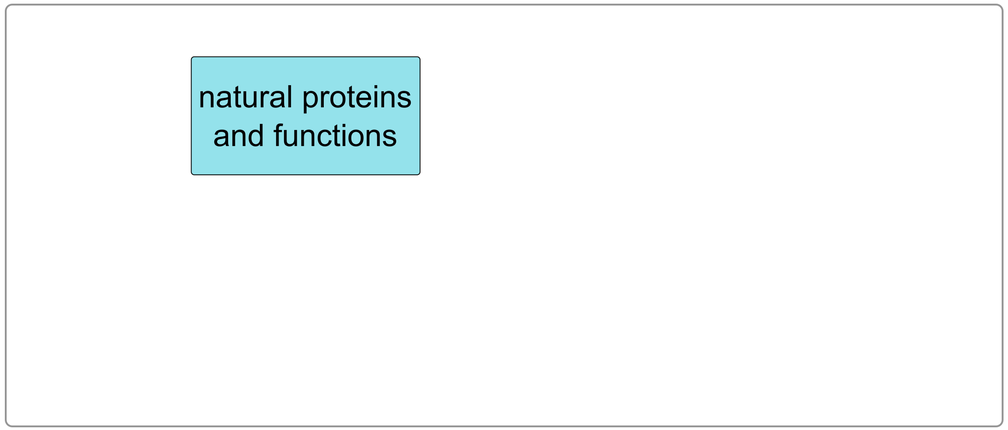
Machine-generated *de novo* proteins

Optimization vs generation

Machine-learning guided directed evolution



Machine-generated *de novo* proteins



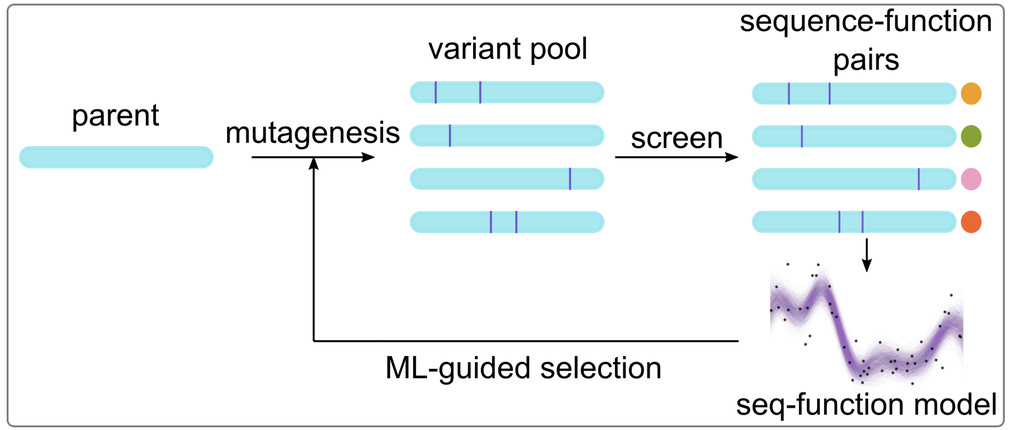
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

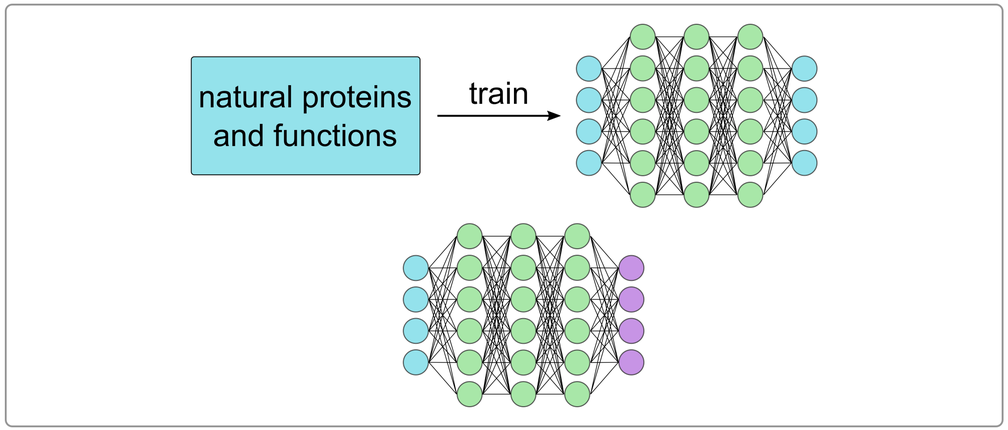
Bedbrook, Yang *et al.* 2019

Optimization vs generation

Machine-learning guided directed evolution



Machine-generated *de novo* proteins



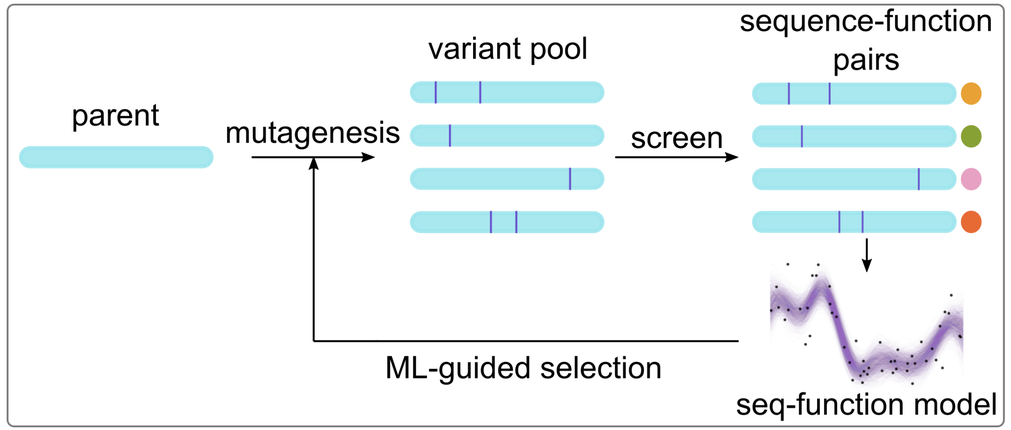
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

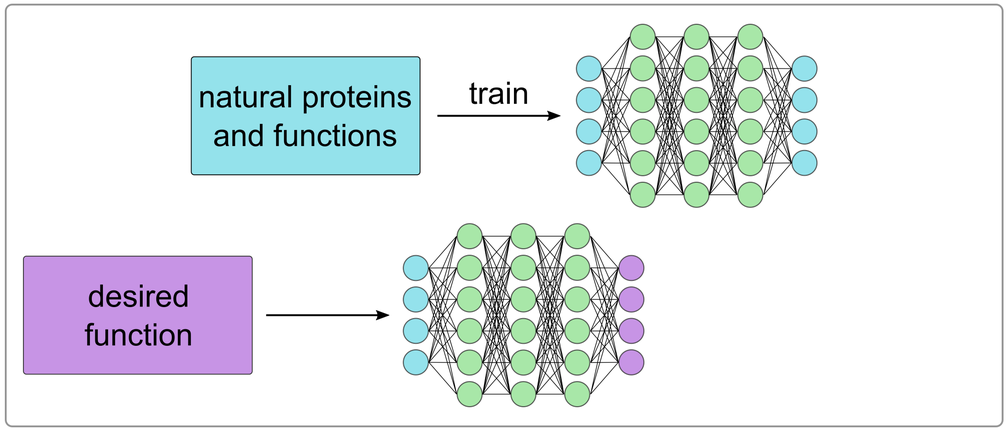
Bedbrook, Yang *et al.* 2019

Optimization vs generation

Machine-learning guided directed evolution



Machine-generated *de novo* proteins



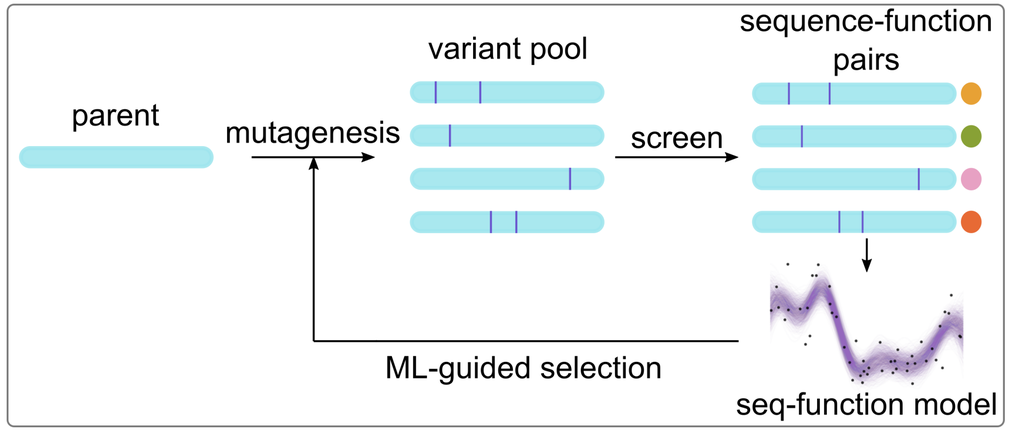
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

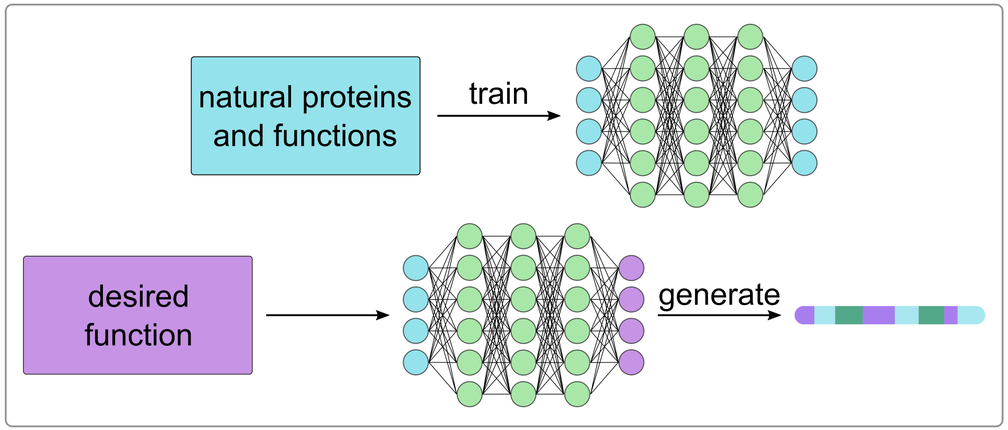
Bedbrook, Yang *et al.* 2019

Optimization vs generation

Machine-learning guided directed evolution



Machine-generated *de novo* proteins



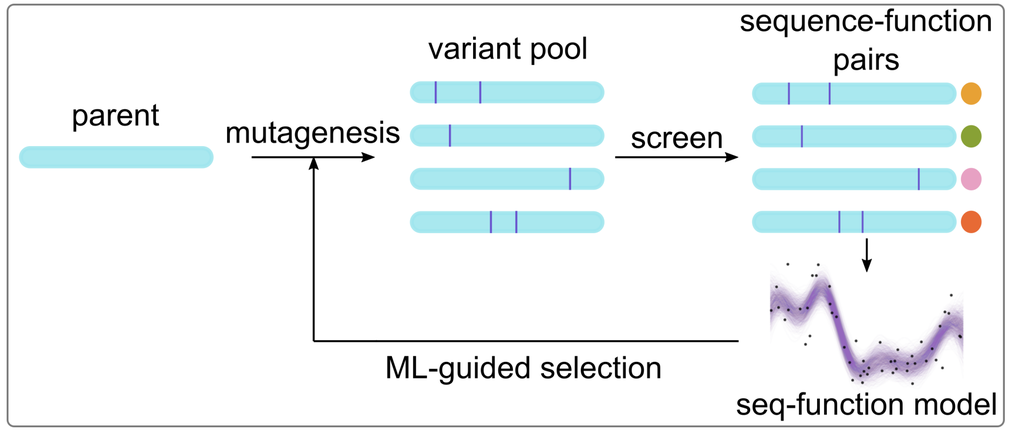
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

Optimization vs generation

Machine-learning guided directed evolution

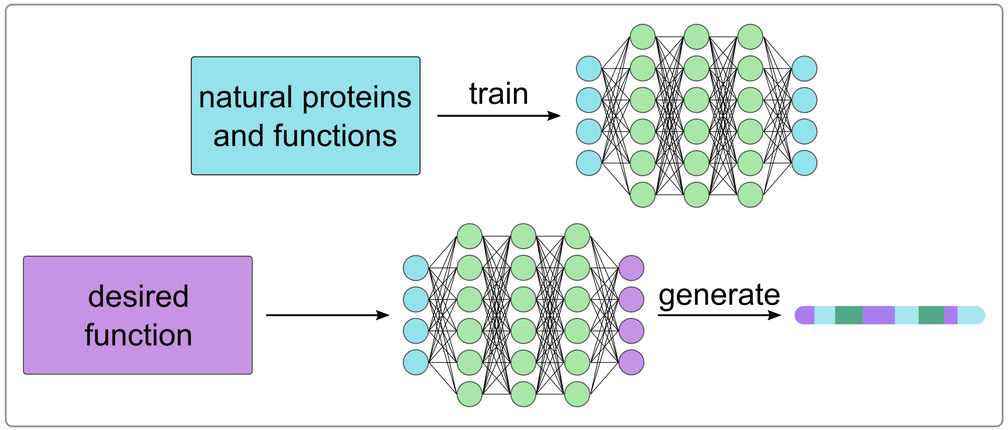


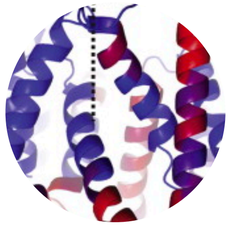
Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

Machine-generated *de novo* proteins

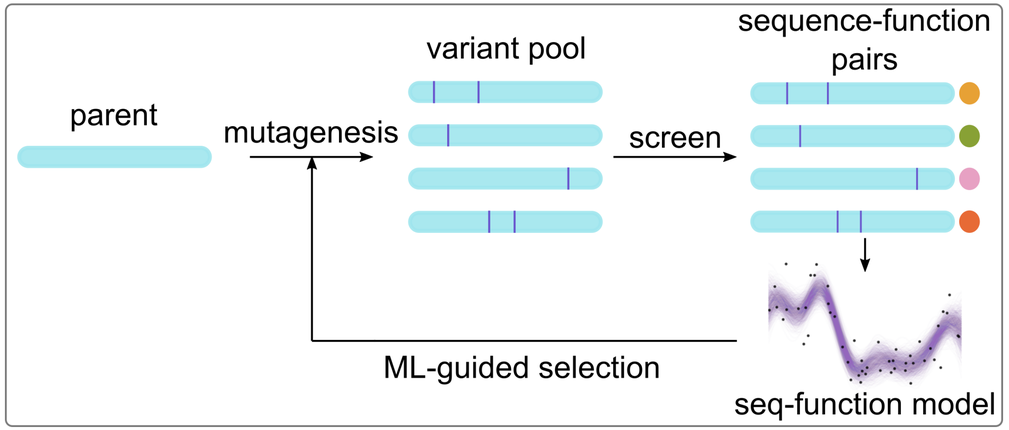




Signal peptide generation

Wu, Yang *et al., 2020*

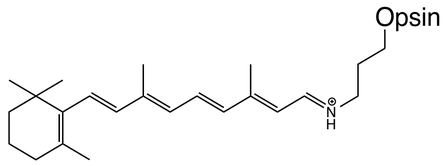
Machine-learning guided directed evolution



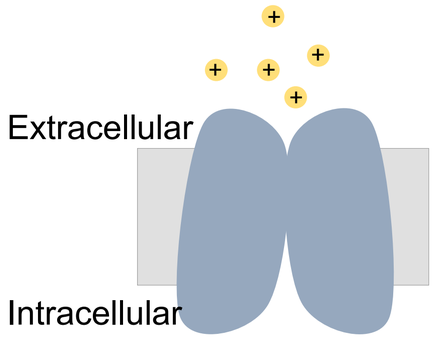
Designer channelrhodopsins

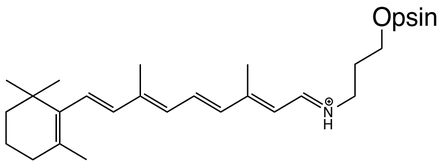
Channelrhodopsins: light-gated ion channels

Channelrhodopsins: light-gated ion channels

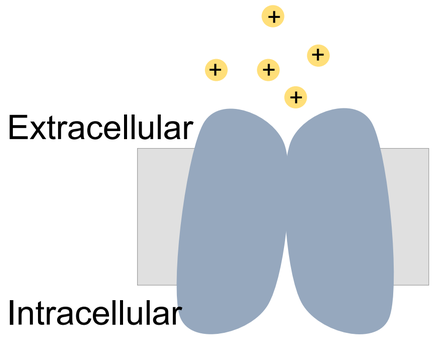


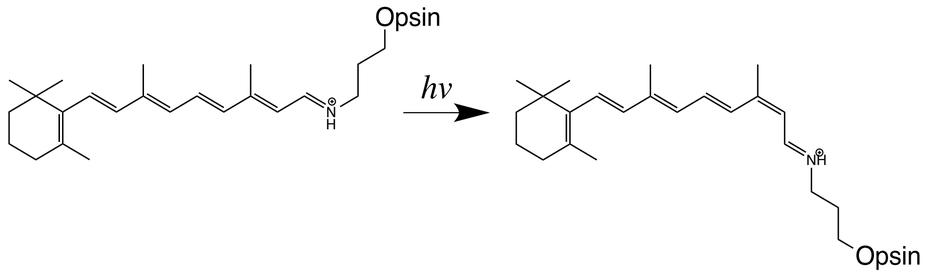
Channelrhodopsins: light-gated ion channels



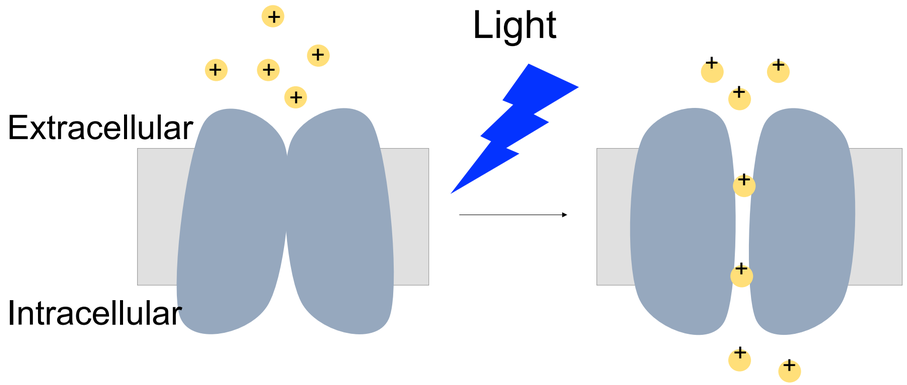


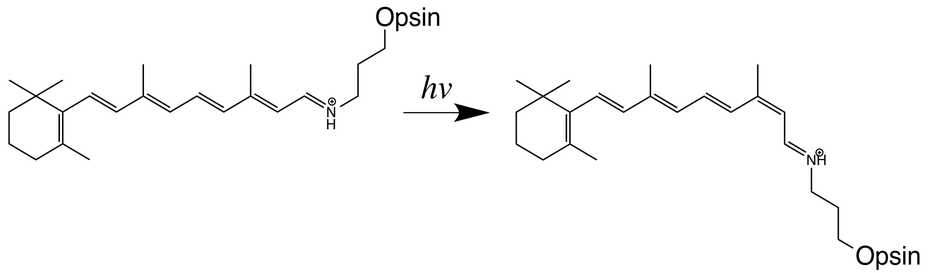
Channelrhodopsins: light-gated ion channels



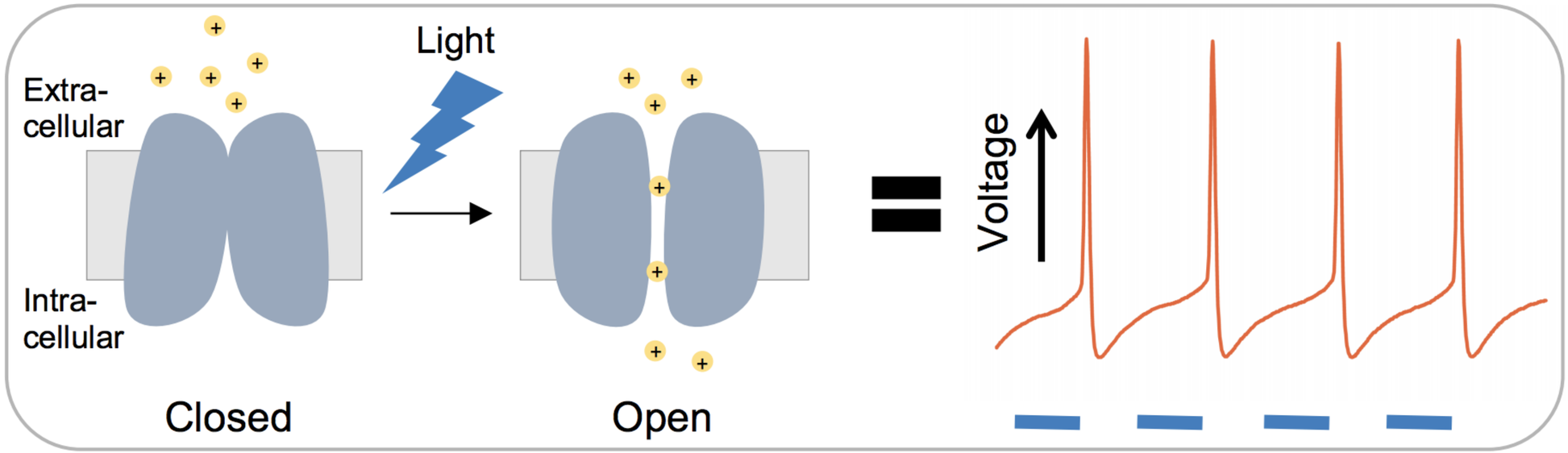


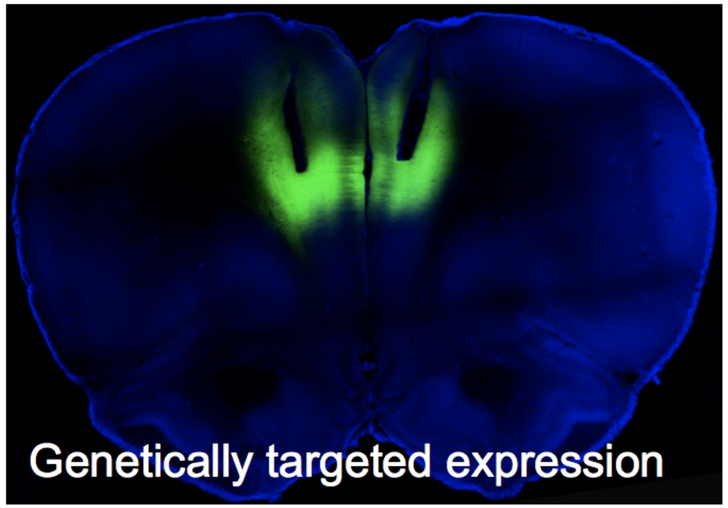
Channelrhodopsins: light-gated ion channels





ChRs are optogenetic tools



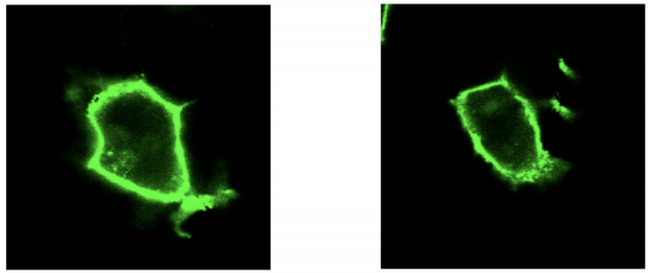


Optogenetics, Nature Method of the year, 2010

Multiple engineering goals

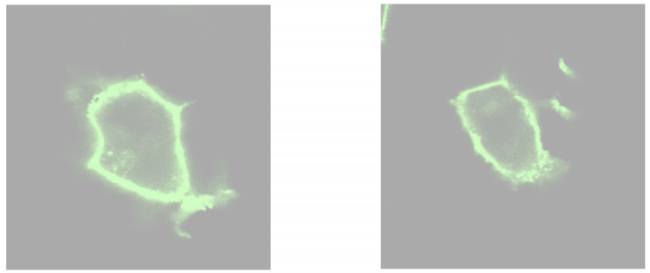
Multiple engineering goals

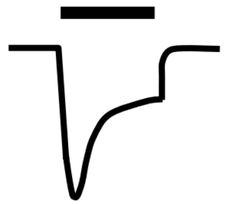
- Heterologous membrane localization



Multiple engineering goals

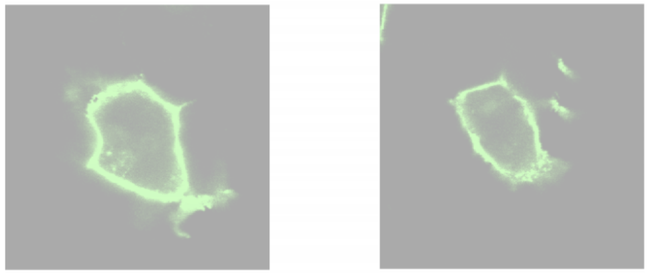
- Heterologous membrane localization
- Increased sensitivity

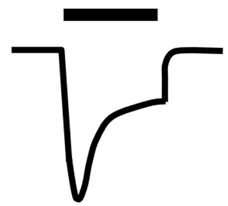




Multiple engineering goals

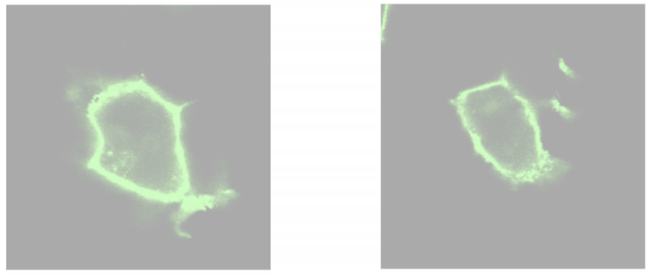
- Heterologous membrane localization
- Increased sensitivity
- Different activation wavelengths



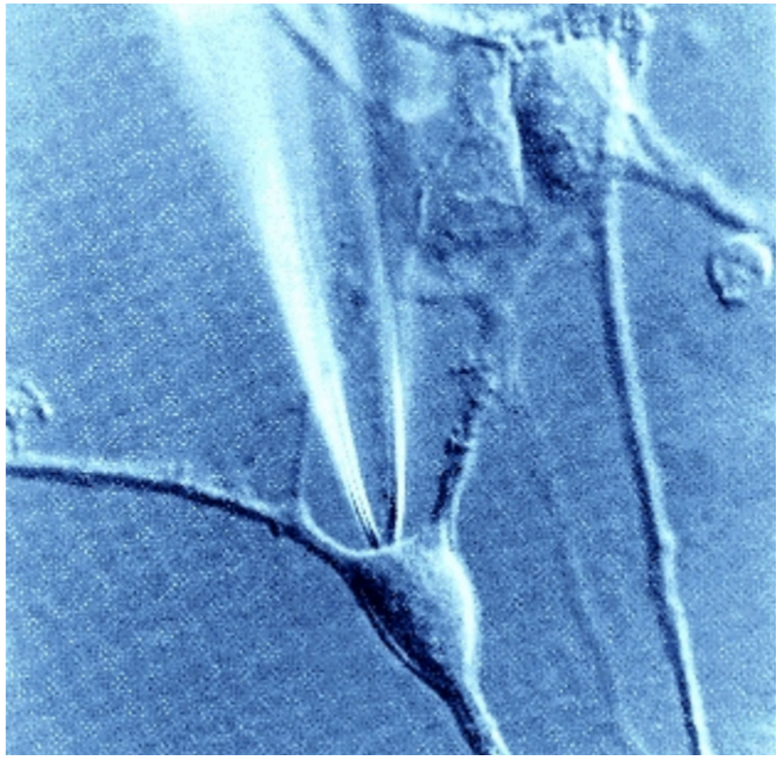


Multiple engineering goals

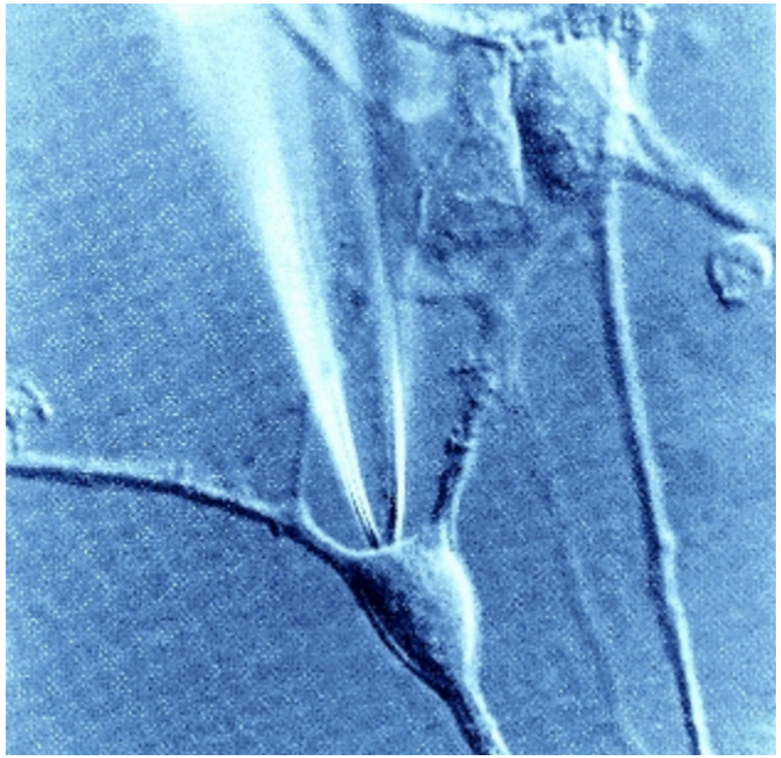
- Heterologous membrane localization
- Increased sensitivity
- Different activation wavelengths
- Different on/off kinetics



No high-throughput screen



No high-throughput screen

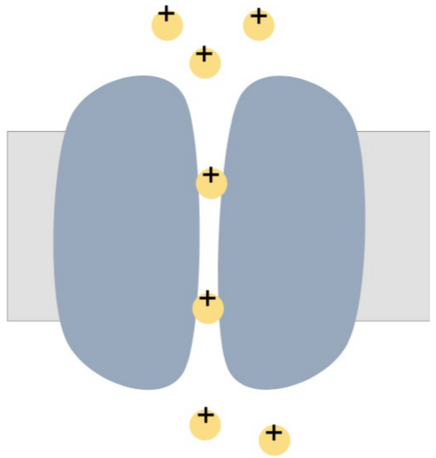


Patch-clamp electrophysiology allows ~2 variants (with replicates) a day

Learn sequence-function relationship

MGTGDHDD...

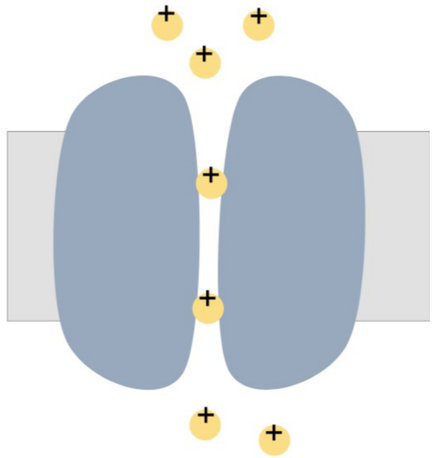
??



Learn sequence-function relationship

MGTGDHDD...

??

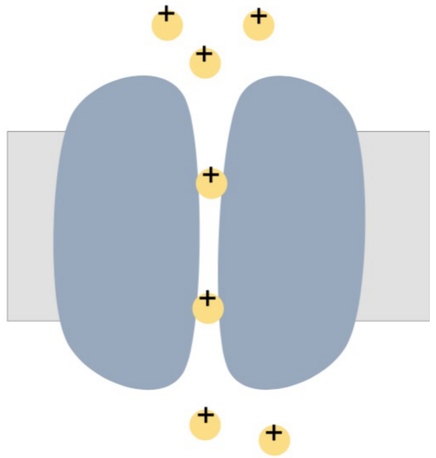


x

Learn sequence-function relationship

MGTGDHDD...

??



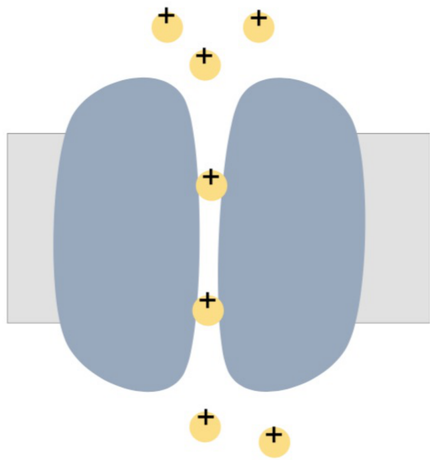
x

f

Learn sequence-function relationship

MGTGDHDD...

??

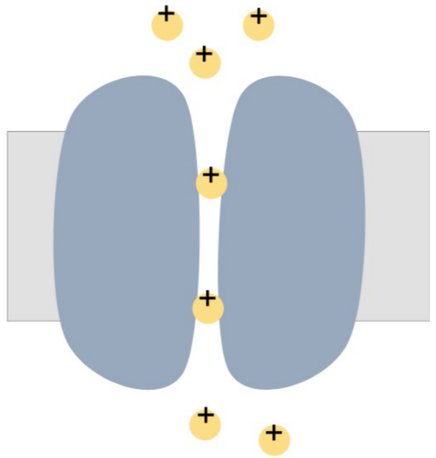


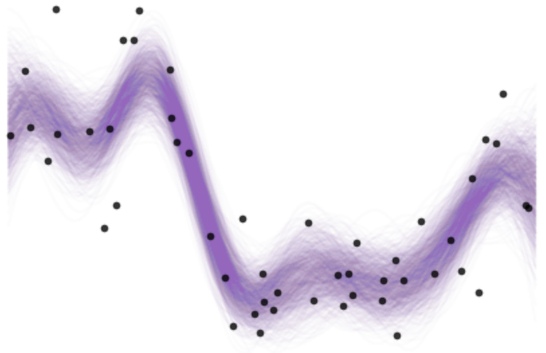
x *f* *f(x)*

Learn sequence-function relationship

MGTGDHDD...

??



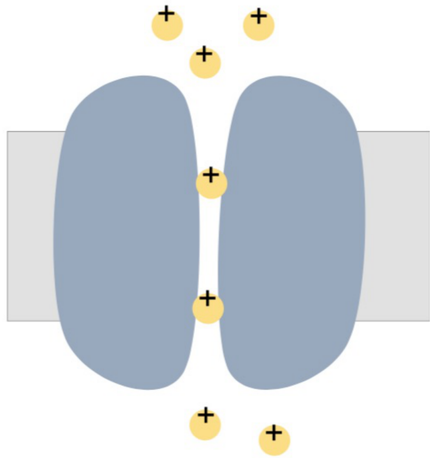


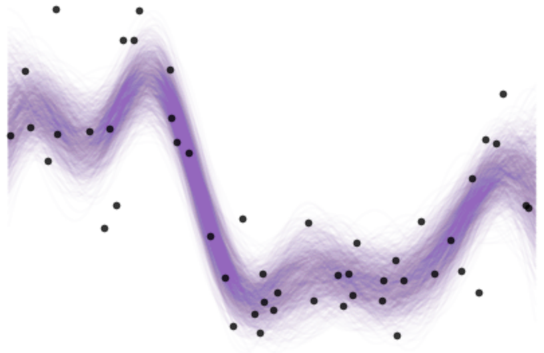
x *f* *f(x)*

Learn sequence-function relationship

MGTGDHDD...

??





x *f* *f(x)*

Approximate *f* from examples of *x and f(x)*

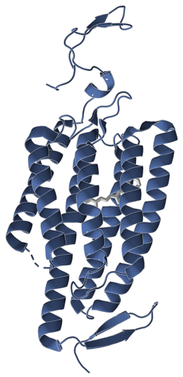
Start from 3 diverse parent ChRs

Start from 3 diverse parent ChRs

C1C2: has structure

Start from 3 diverse parent ChRs

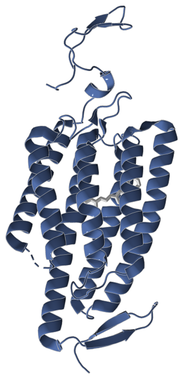
C1C2: has structure



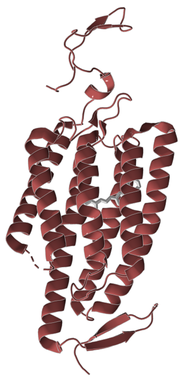
CheRiff: blue-shifted, strong currents

Start from 3 diverse parent ChRs

C1C2: has structure



CheRiff: blue-shifted, strong currents

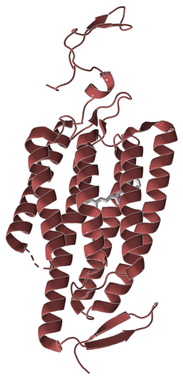


CsChrimsonR: red-shifted

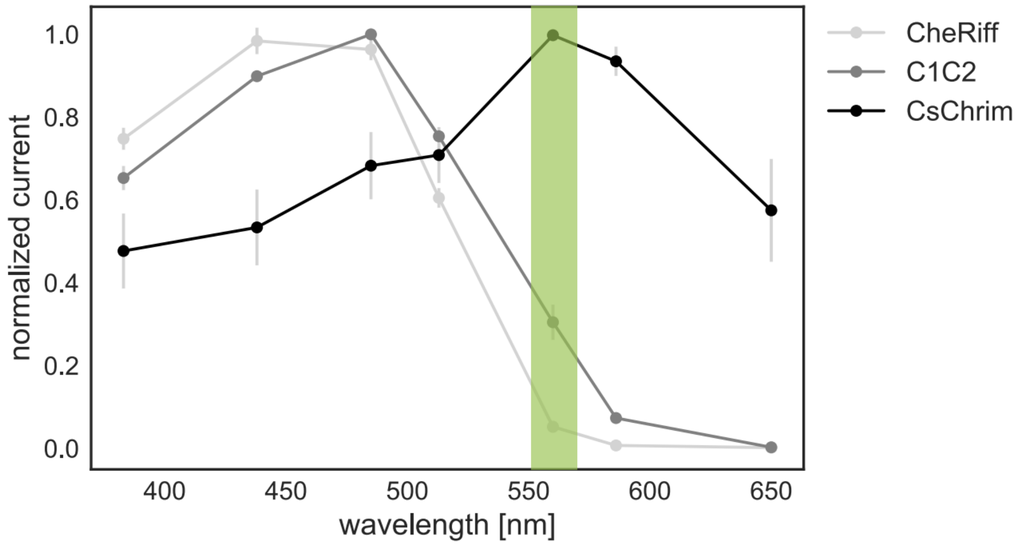
Start from 3 diverse parent ChRs

C1C2: has structure

CheRiff: blue-shifted, strong currents

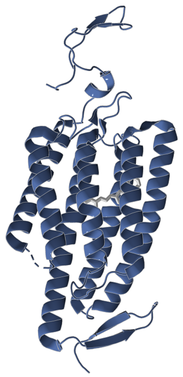


CsChrimsonR: red-shifted

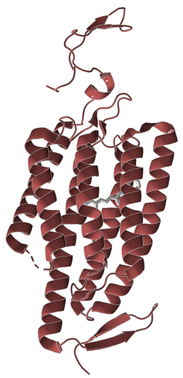


Start from 3 diverse parent ChRs

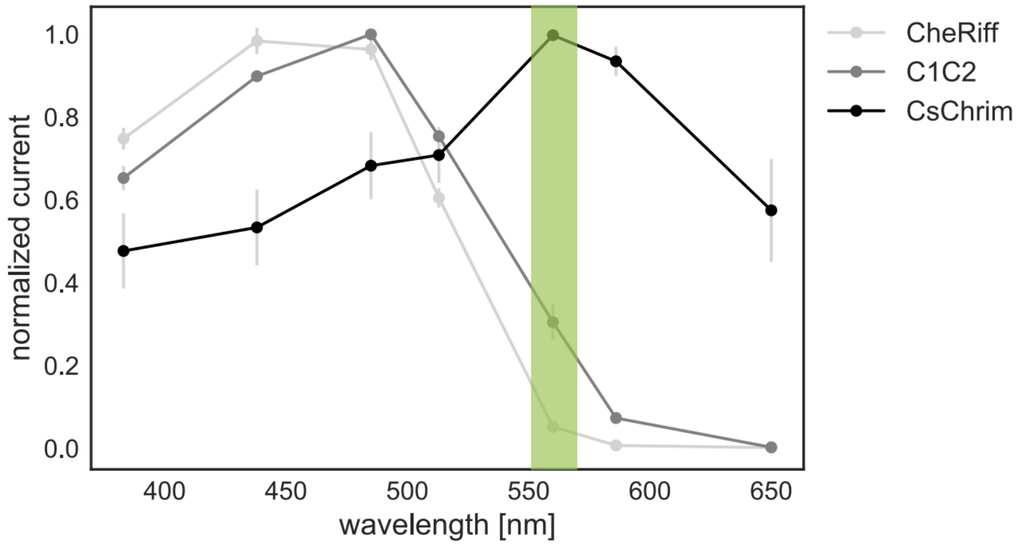
C1C2: has structure

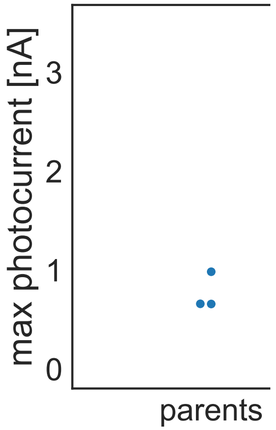


CheRiff: blue-shifted, strong currents



CsChrimsonR: red-shifted

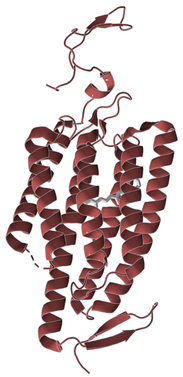




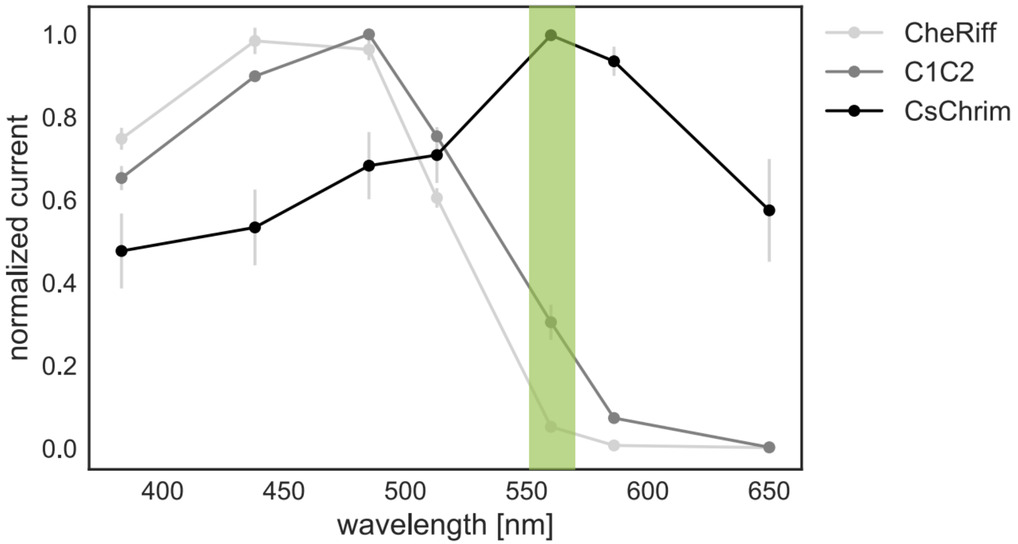
Start from 3 diverse parent ChRs

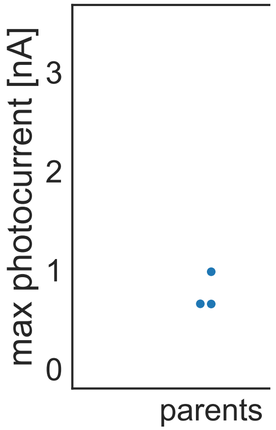
C1C2: has structure

CheRiff: blue-shifted, strong currents



CsChrimsonR: red-shifted



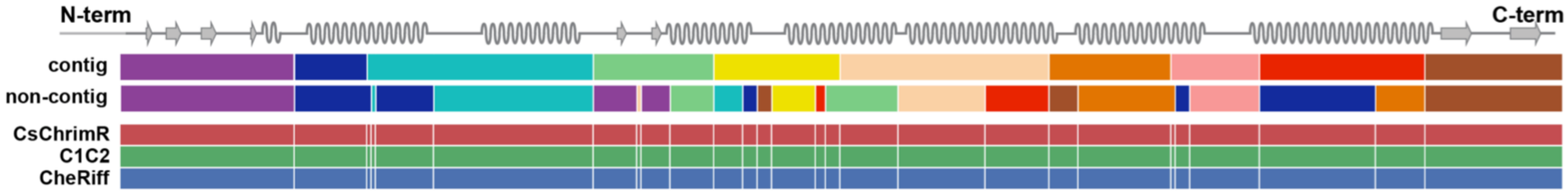


Kato 2012
Hochbaum 2014
Klapoetke 2014

Use recombination to generate diversity

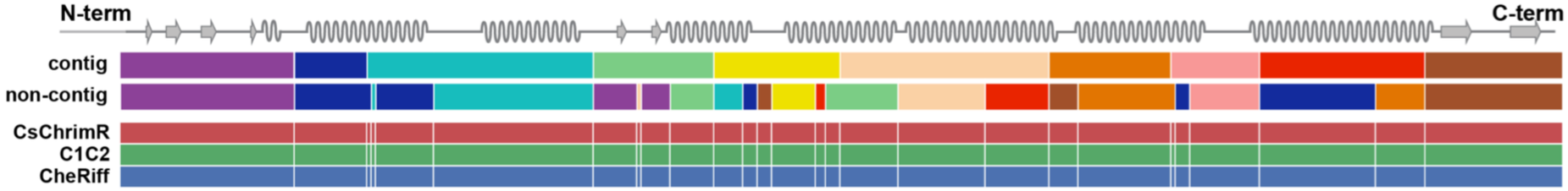
Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



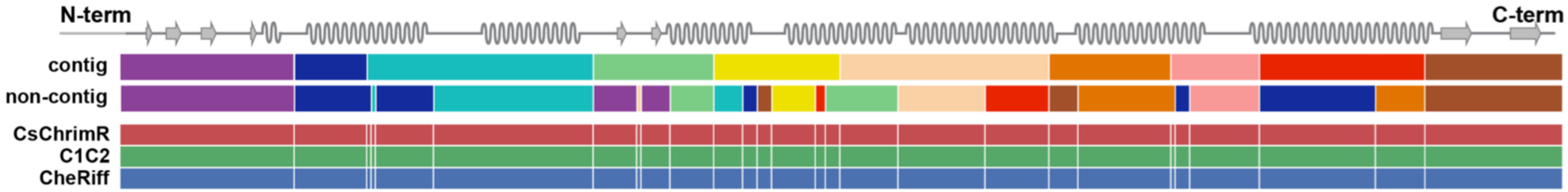
2 libraries

3 parents

10 blocks

Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



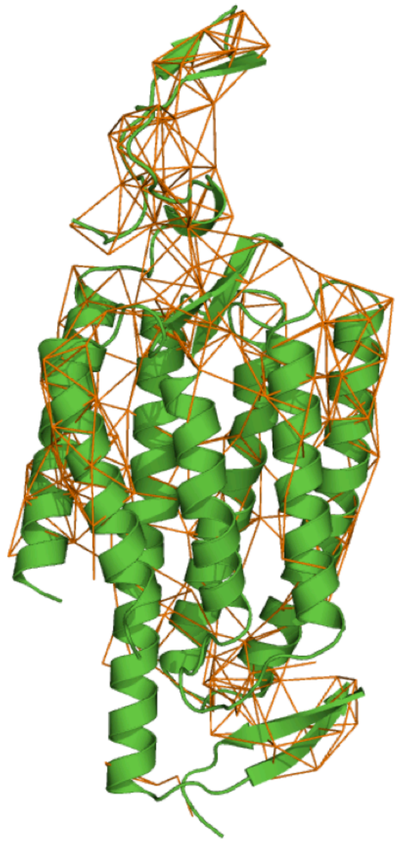
2 libraries

3 parents = 118,098 sequences

10 blocks

Voigt 2002
Otey 2006
Smith 2013

Choose cut sites to minimize broken contacts

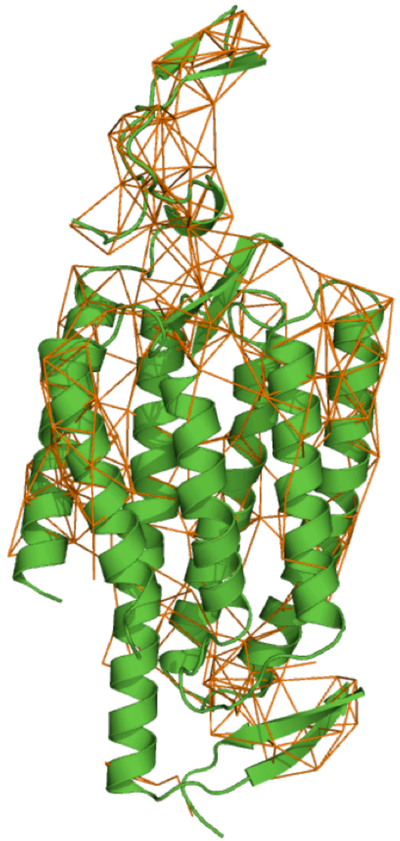


sequence: AGTT

Contacting positions: (0, 2), (0, 3) (from structure)

Voigt 2002
Otey 2006
Smith 2013

Choose cut sites to minimize broken contacts



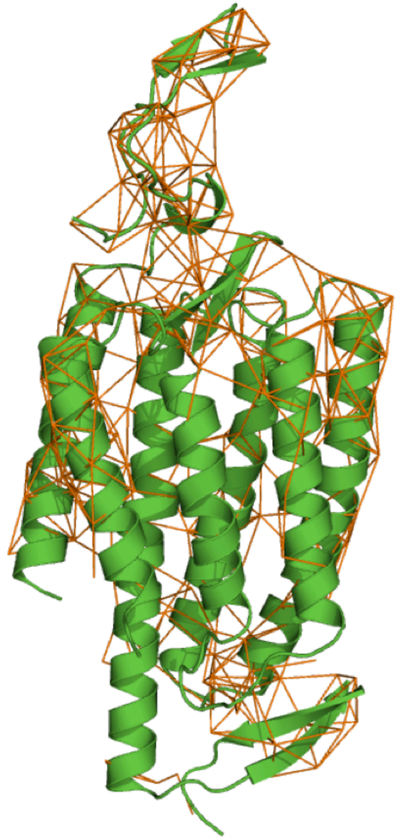
sequence: AGTT

Contacting positions: (0, 2), (0, 3) (from structure)

Contacts: (0A, 2T), (0A, 3T)

Voigt 2002
Otey 2006
Smith 2013

Choose cut sites to minimize broken contacts



sequence: AGTT

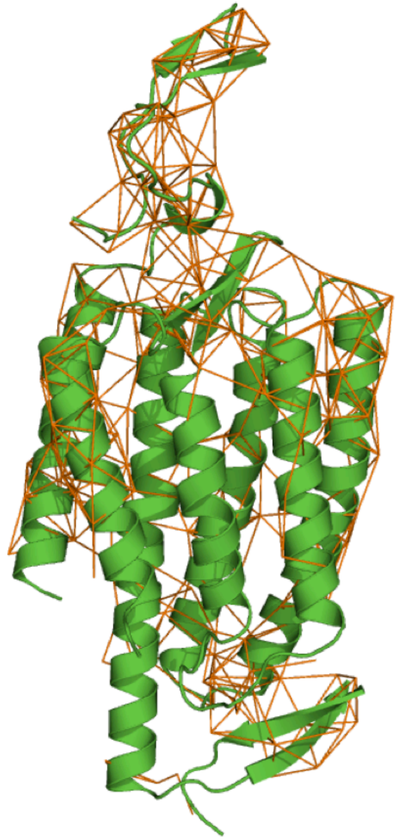
Contacting positions: (0, 2), (0, 3) (from structure)

Contacts: (0A, 2T), (0A, 3T)

Defined by positions *and* identities

Voigt 2002
Otey 2006
Smith 2013

Choose cut sites to minimize broken contacts



sequence: AGTT

Contacting positions: (0, 2), (0, 3) (from structure)

Contacts: (0A, 2T), (0A, 3T)

Defined by positions *and* identities

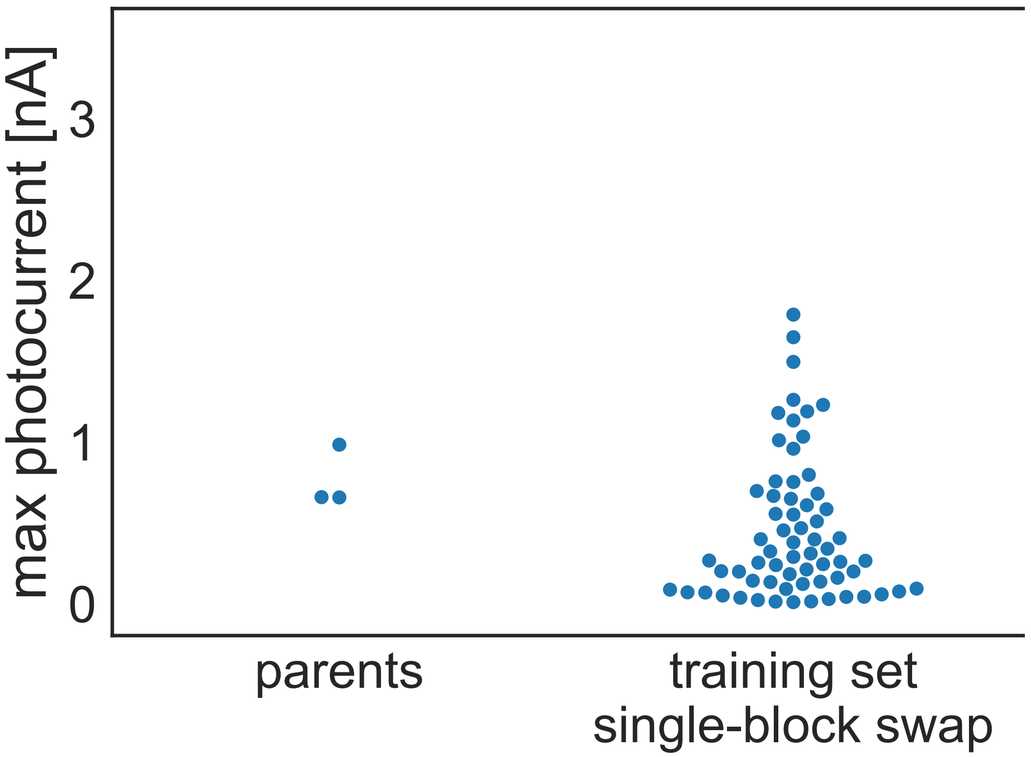
Contacts are “broken” if they are not present in any of the parents

Voigt 2002
Otey 2006
Smith 2013

Chimeras retain photocurrents

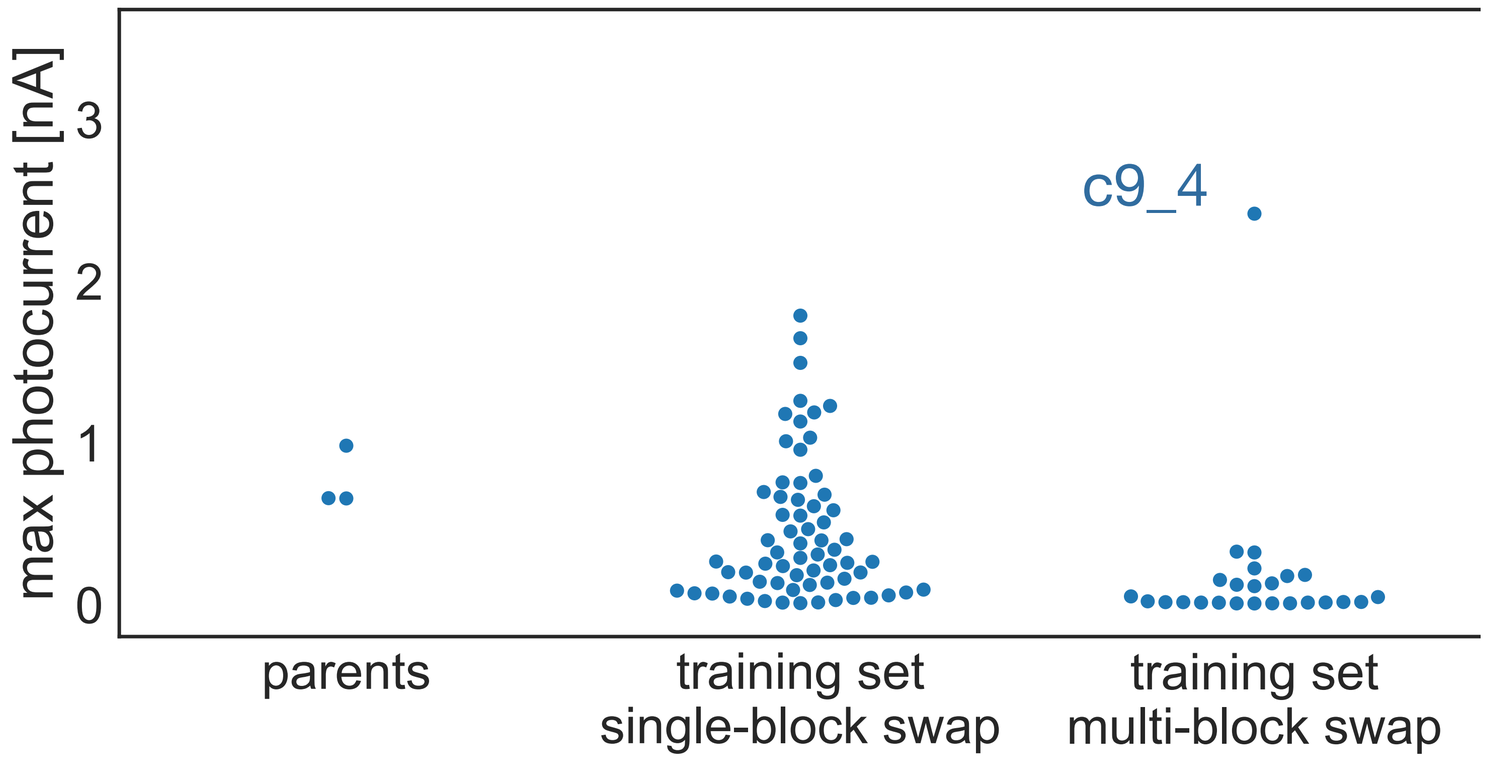
Voigt 2002
Otey 2006
Smith 2013

Chimeras retain photocurrents



Voigt 2002
Otey 2006
Smith 2013

Chimeras retain photocurrents



Voigt 2002
Otey 2006
Smith 2013

Machine learning with Gaussian processes

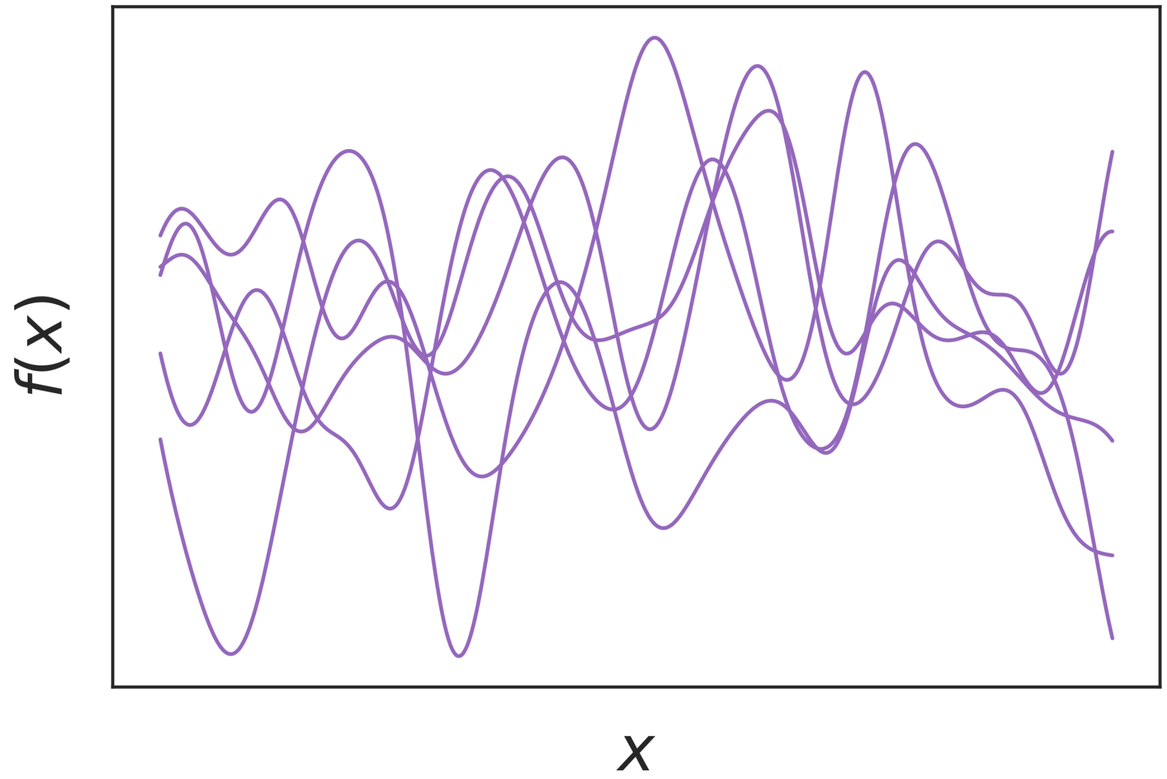
Machine learning with Gaussian processes

Machine learning with Gaussian processes

(Gaussian) processes are distributions over functions

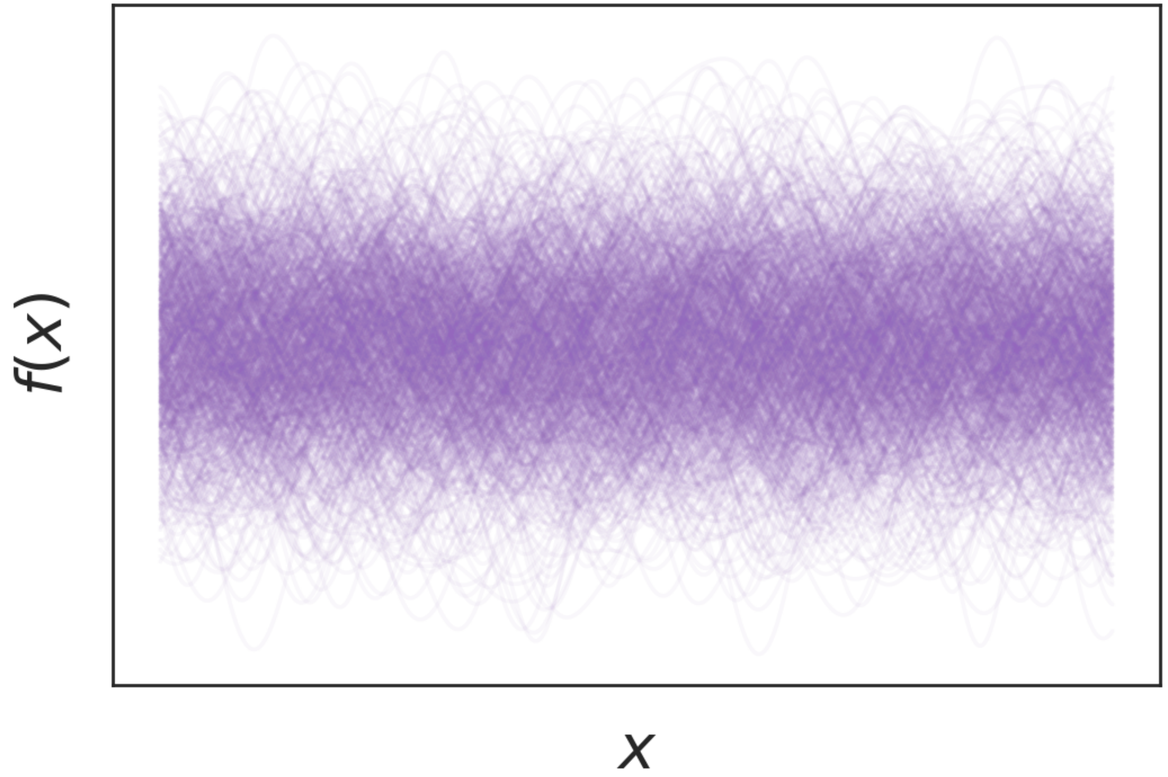
Machine learning with Gaussian processes

(Gaussian) processes are distributions over functions



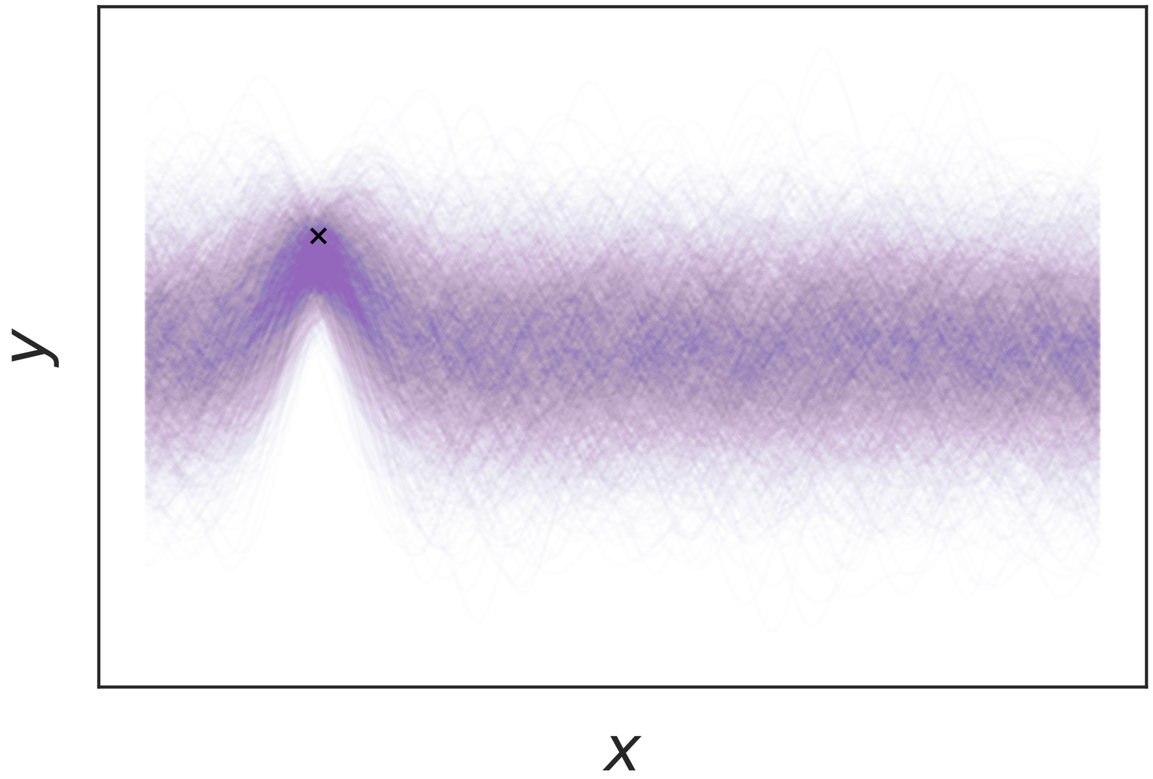
Machine learning with Gaussian processes

(Gaussian) processes are distributions over functions



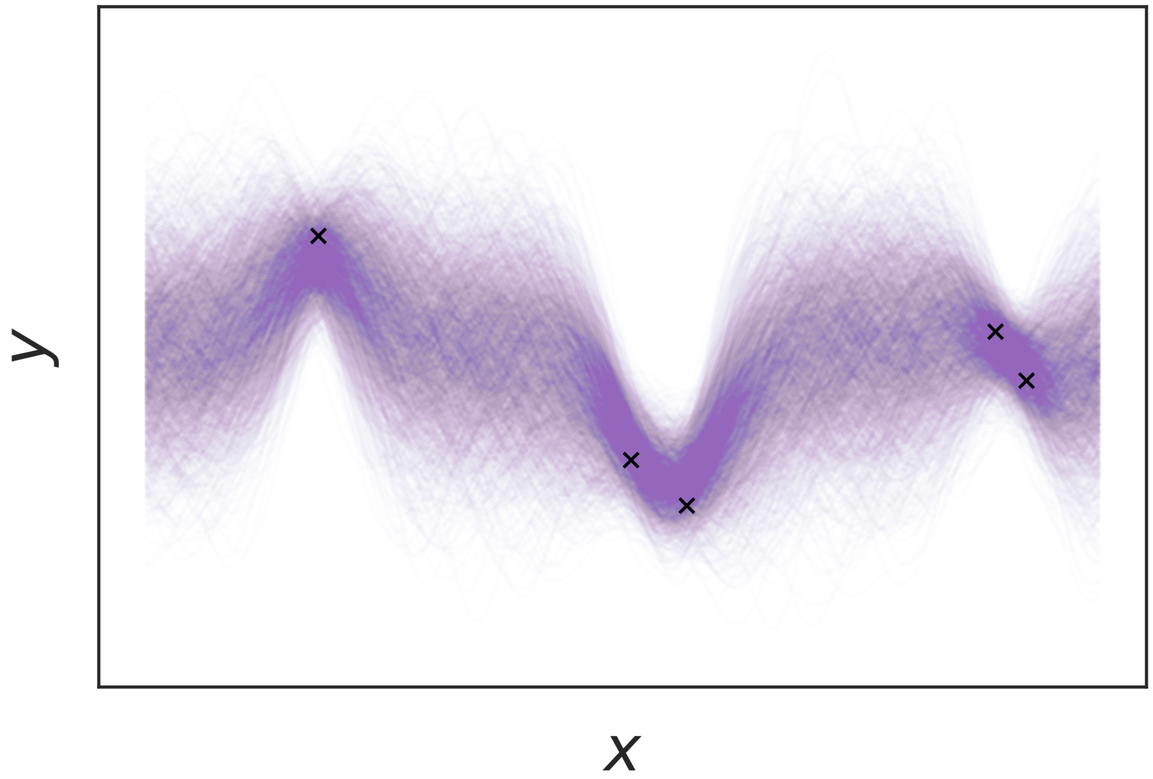
Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



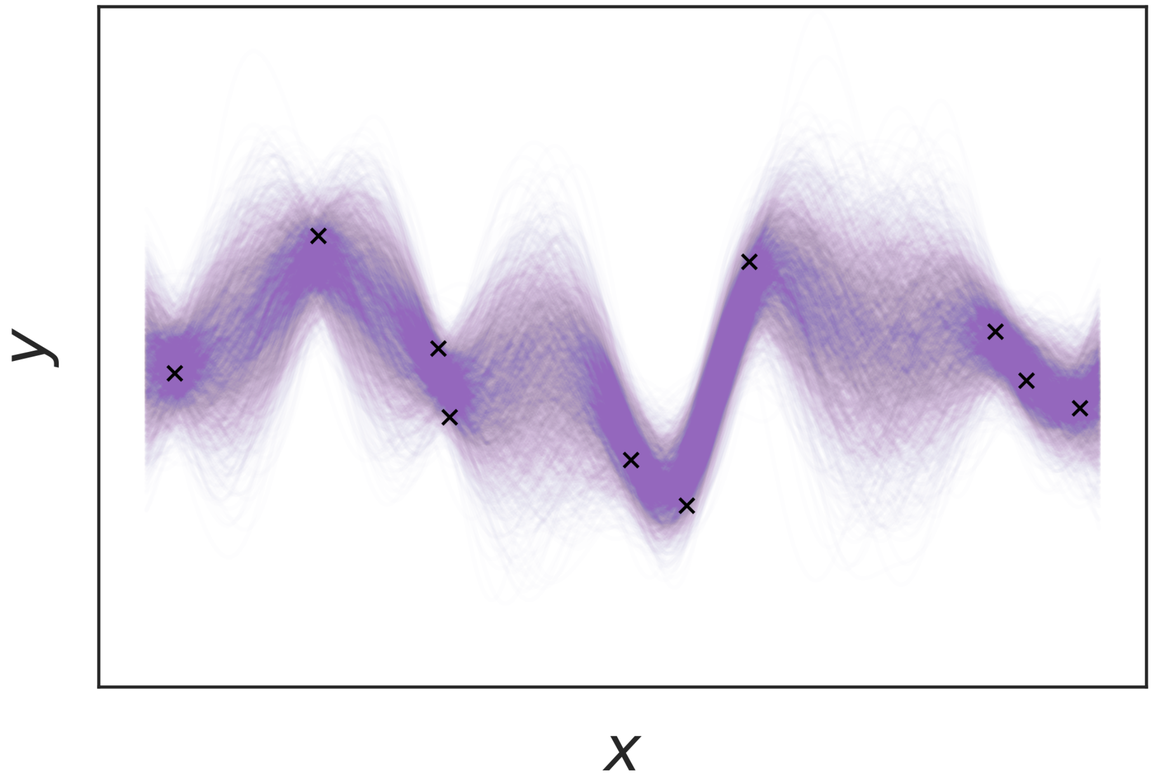
Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



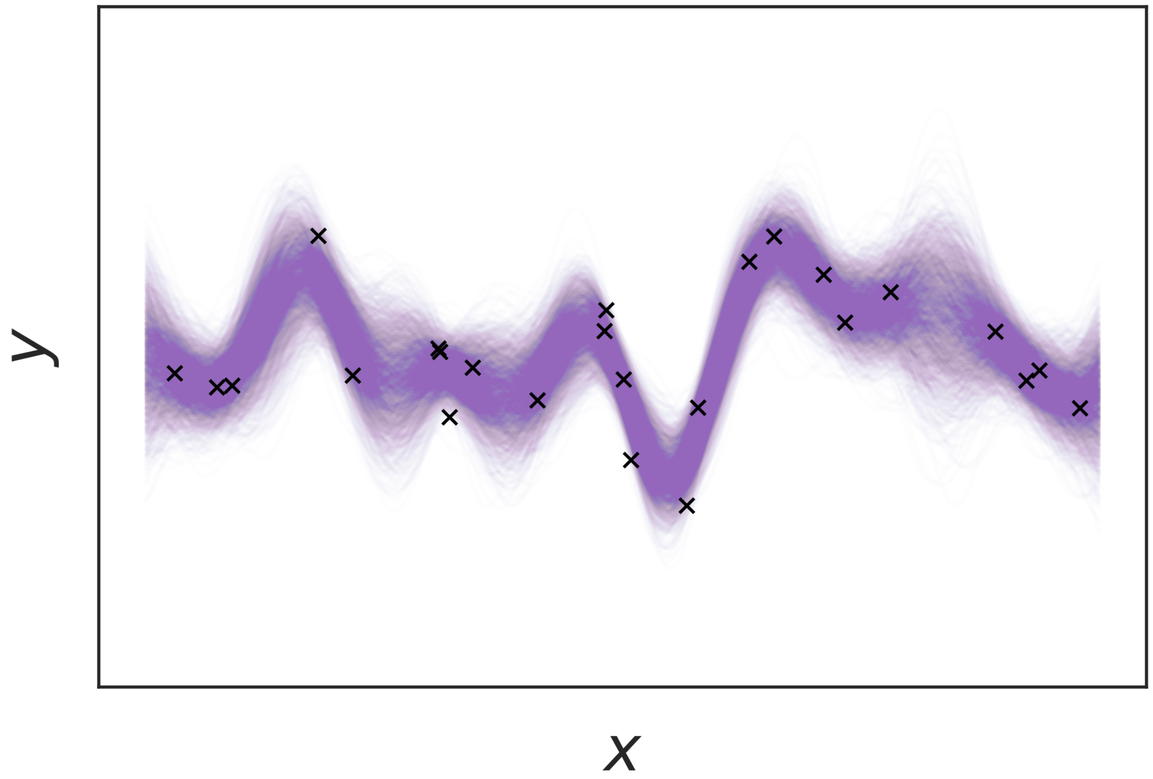
Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



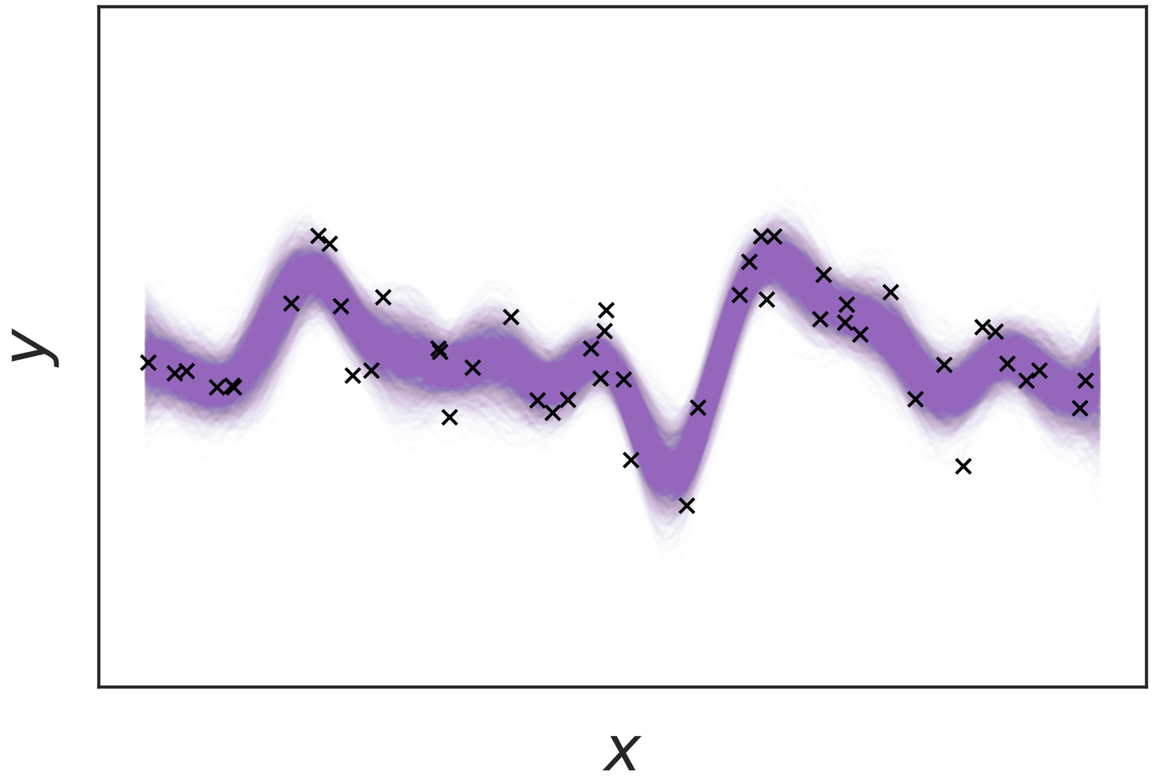
Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



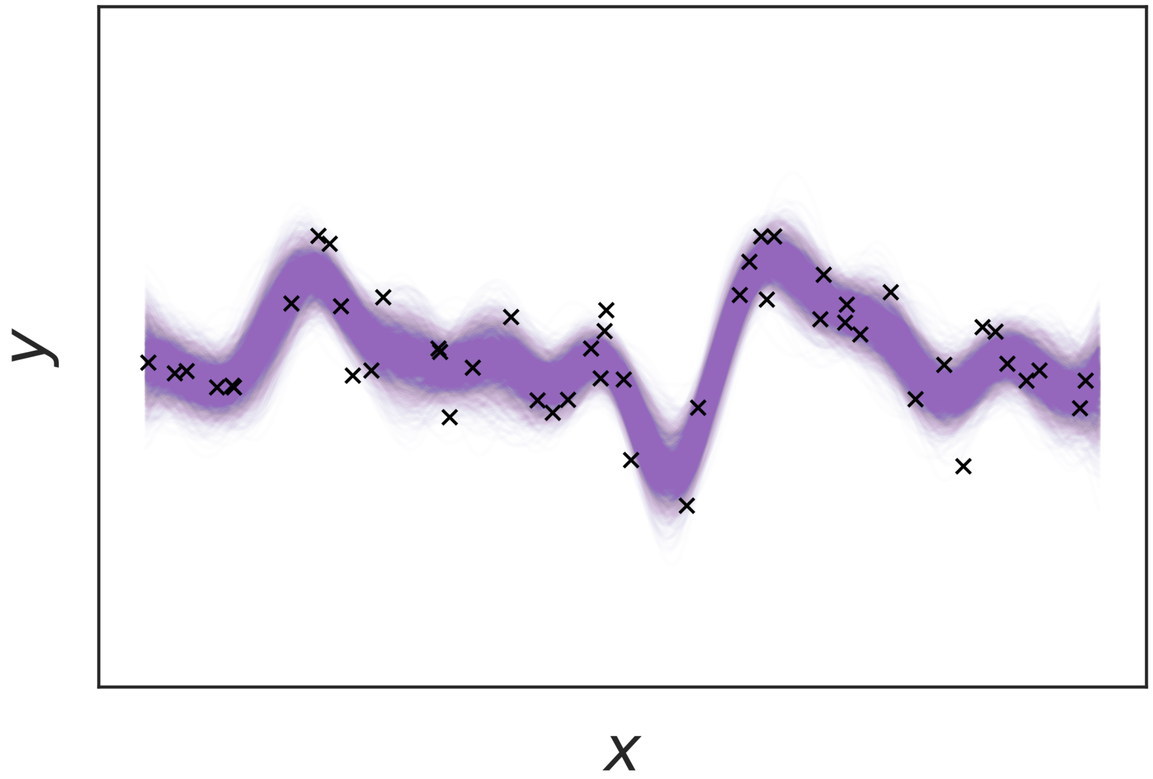
Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



Machine learning with Gaussian processes

As we observe data, we become more certain about the underlying function



Points that are *close* have similar *y*

Uncertainty increases further from observed points

The covariance defines the GP

The covariance defines the GP

- Real-valued *k(x, x')*

The covariance defines the GP

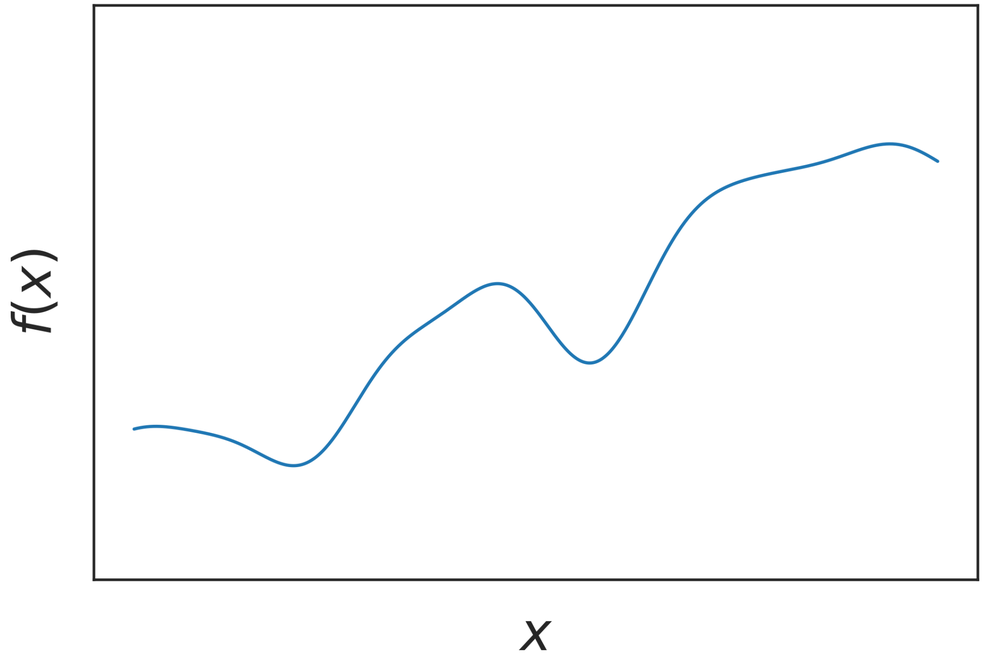
- Real-valued *k(x, x')*
- Defines relatedness between inputs

The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP

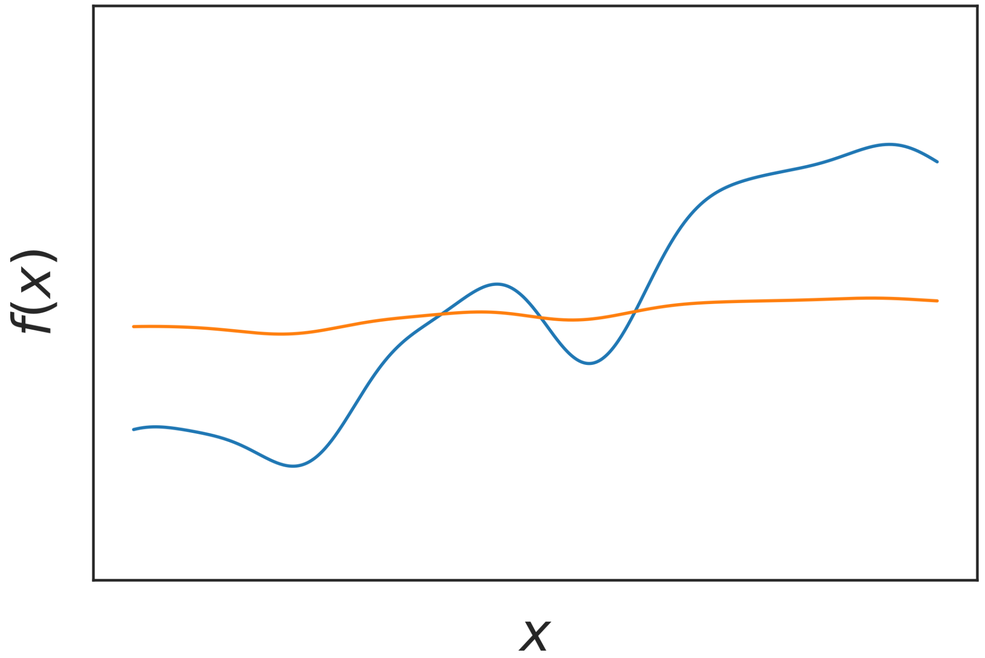
The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP



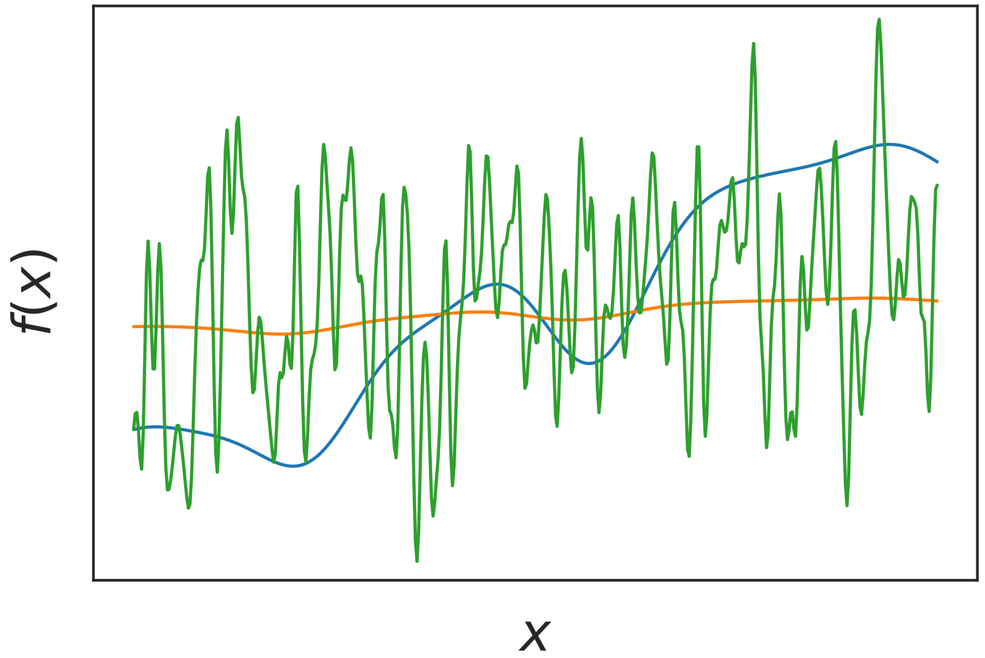
The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP



The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP

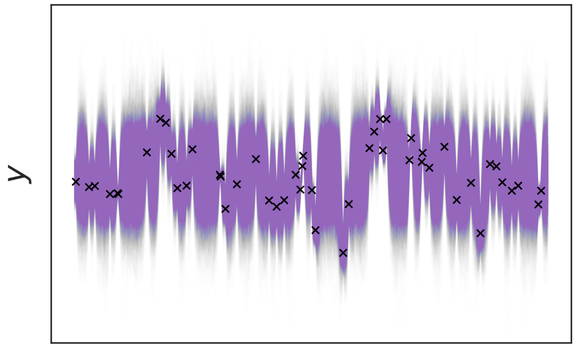


The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP
- The choice of *k* and its *hyperparameters* determines how good the model is

The covariance defines the GP

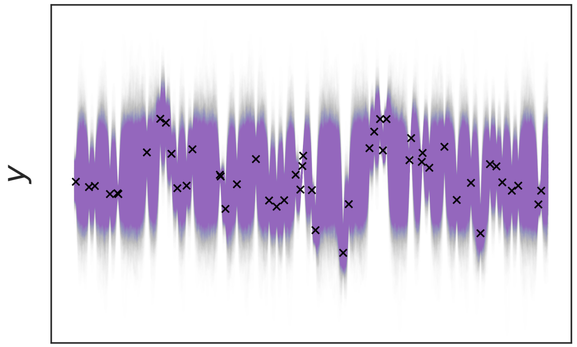
- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP
- The choice of *k* and its *hyperparameters* determines how good the model is



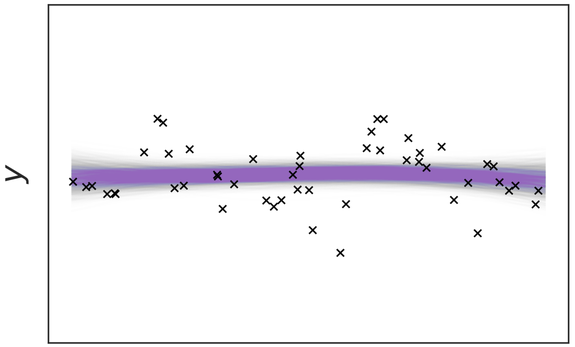
overfit

The covariance defines the GP

- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP
- The choice of *k* and its *hyperparameters* determines how good the model is



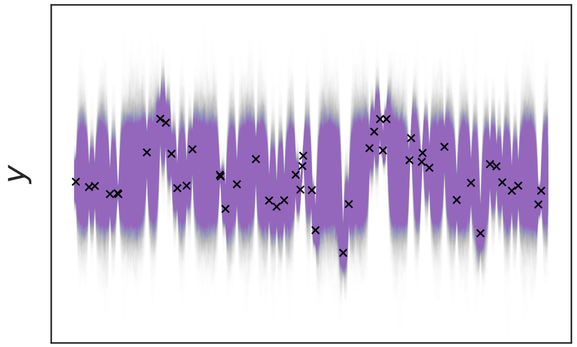
overfit



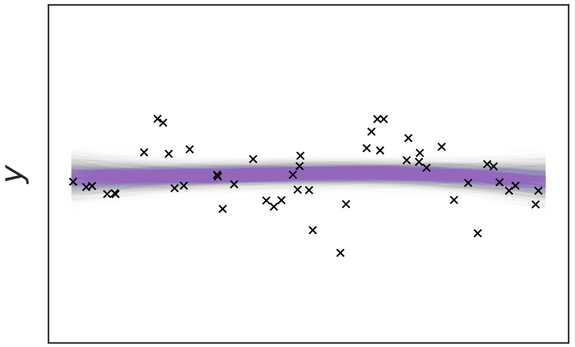
underfit

The covariance defines the GP

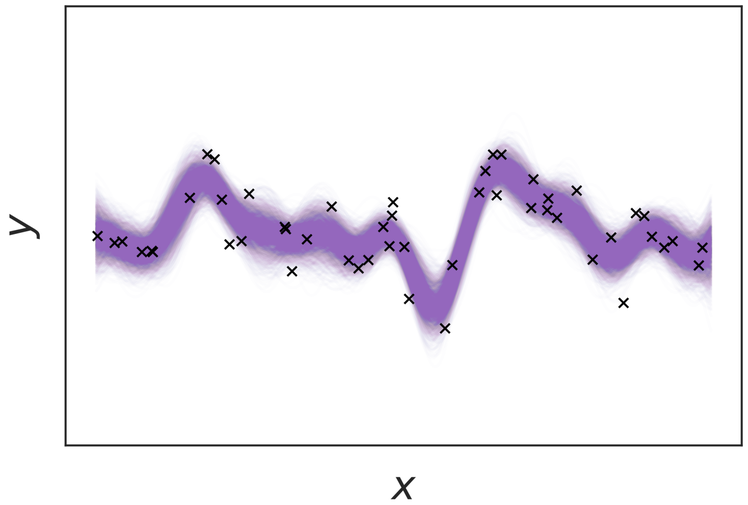
- Real-valued *k(x, x')*
- Defines relatedness between inputs
- Defines the type of functions in GP
- The choice of *k* and its *hyperparameters* determines how good the model is



overfit



underfit



Choosing k and its hyperparameters

polynomial: $k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$

squared exponential (SE): $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$

Matérn: $k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\, \beta^\nu}{\Gamma(\nu)} K_\nu(\beta)$, where $\beta = \left(\frac{2\nu\, \|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$

Choosing k and its hyperparameters

polynomial: $k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$

squared exponential (SE): $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$

Matérn: $k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\, \beta^\nu}{\Gamma(\nu)} K_\nu(\beta)$, where $\beta = \left(\frac{2\nu\, \|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$

Choosing k and its hyperparameters

$$\text{polynomial: } k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$$

$$\text{squared exponential (SE): } k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$$

$$\text{Matérn: } k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\,\beta^\nu}{\Gamma(\nu)} K_\nu(\beta), \text{ where } \beta = \left(\frac{2\nu\,\|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$$

Use the marginal likelihood (differentiable!) to select

Choosing k and its hyperparameters

$$\text{polynomial: } k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$$

$$\text{squared exponential (SE): } k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$$

$$\text{Matérn: } k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\,\beta^\nu}{\Gamma(\nu)} K_\nu(\beta), \text{ where } \beta = \left(\frac{2\nu\,\|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$$

Use the marginal likelihood (differentiable!) to select

$$\log p\left(\mathbf{y} \,|\, \gamma, X\right) \propto -\mathbf{y}^T (K_\gamma + \sigma_n^2 I)^{-1} \mathbf{y} - \log |K_\gamma + \sigma_n^2 I|$$

Choosing k and its hyperparameters

$$\text{polynomial: } k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$$

$$\text{squared exponential (SE): } k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$$

$$\text{Matérn: } k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\,\beta^\nu}{\Gamma(\nu)} K_\nu(\beta), \text{ where } \beta = \left(\frac{2\nu\,\|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$$

Use the marginal likelihood (differentiable!) to select

$$\log p(\mathbf{y} \mid \gamma, X) \propto -\mathbf{y}^T(K_\gamma + \sigma_n^2 I)^{-1}\mathbf{y} - \log|K_\gamma + \sigma_n^2 I|$$

accuracy

Choosing k and its hyperparameters

$$\text{polynomial: } k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \sigma_p^2 \mathbf{x}^\top \mathbf{x}')^d, \quad d \in \{1, 2, 3, \ldots\}$$

$$\text{squared exponential (SE): } k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right]$$

$$\text{Matérn: } k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}\,\beta^\nu}{\Gamma(\nu)} K_\nu(\beta), \text{ where } \beta = \left(\frac{2\nu\,\|\mathbf{x} - \mathbf{x}'\|_2^2}{\ell^2}\right)^{\frac{1}{2}}$$

Use the marginal likelihood (differentiable!) to select

$$\log p(\mathbf{y} \mid \gamma, X) \propto -\mathbf{y}^T (K_\gamma + \sigma_n^2 I)^{-1} \mathbf{y} - \log |K_\gamma + \sigma_n^2 I|$$

accuracy — complexity

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

1-hot encoding for sequence

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

1-hot encoding for sequence

		0	
A	C	G	T
1	0	0	0

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

1-hot encoding for sequence

0				1			
A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

1-hot encoding for sequence

0				1				2			
A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0	0	0	0	1

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

1-hot encoding for sequence

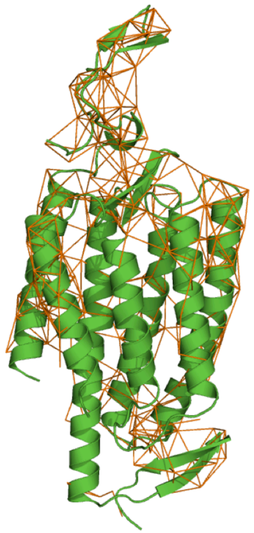
0				1				2				3			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

Contacts: (0, 2), (0, 3)

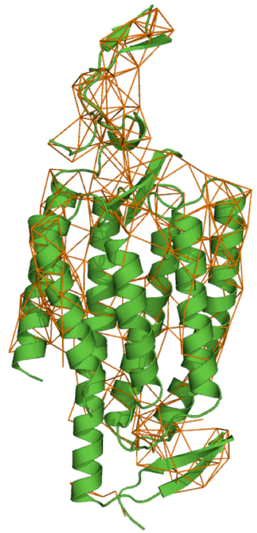


1-hot encoding for sequence

0				1				2				3			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1

1-hot encoding for structure

How do we vectorize a protein?



alphabet: {A, C, G, T}

sequence: AGTT

Contacts: (0, 2), (0, 3)

1-hot encoding for sequence

0				1				2				3			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1

1-hot encoding for structure

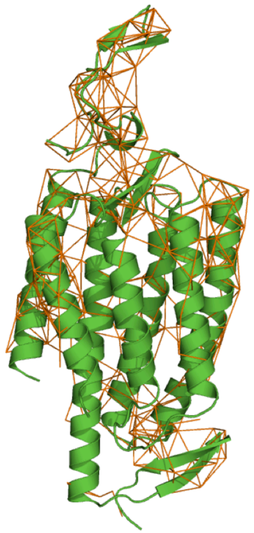
(0,2)											
AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	...
0	0	0	1	0	0	0	0	0	0	0	...

How do we vectorize a protein?

alphabet: {A, C, G, T}

sequence: AGTT

Contacts: (0, 2), (0, 3)



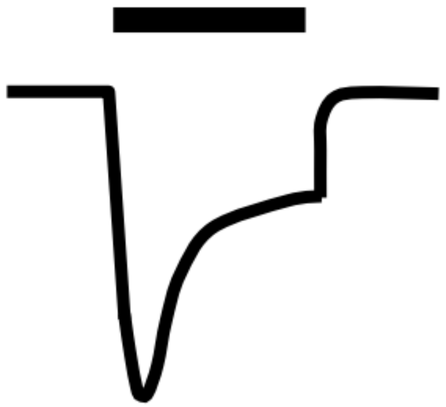
1-hot encoding for sequence

0				1				2				3			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1

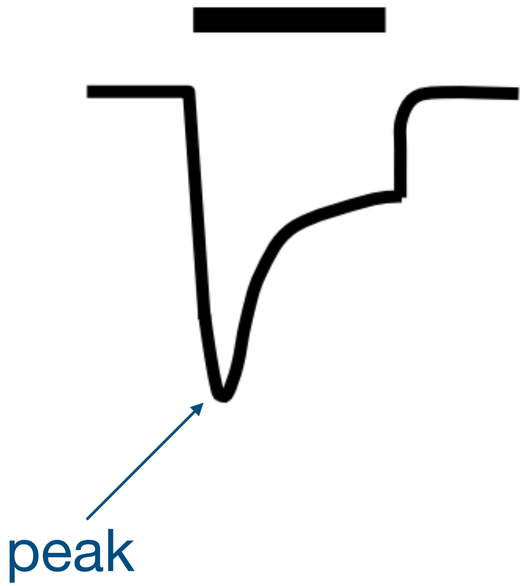
1-hot encoding for structure

			(0,2)							
AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	...
0	0	0	1	0	0	0	0	0	0	0	...

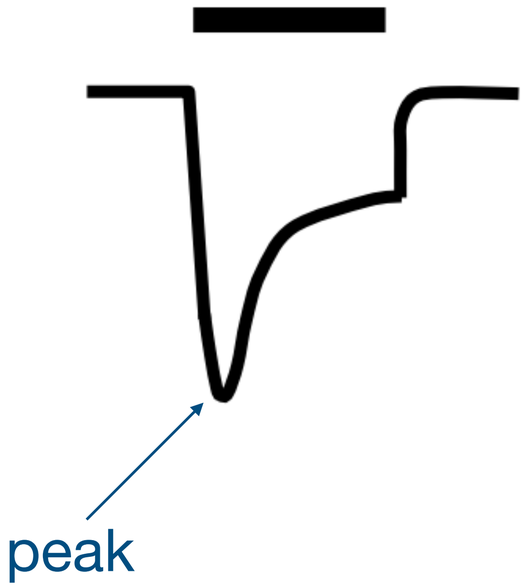
Models for functional properties are accurate

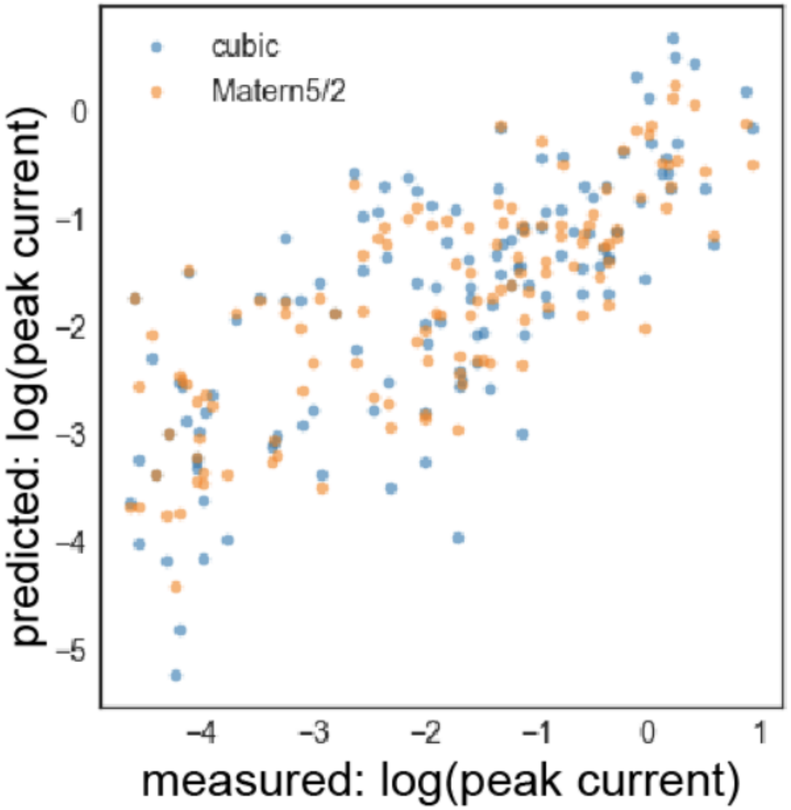


Models for functional properties are accurate

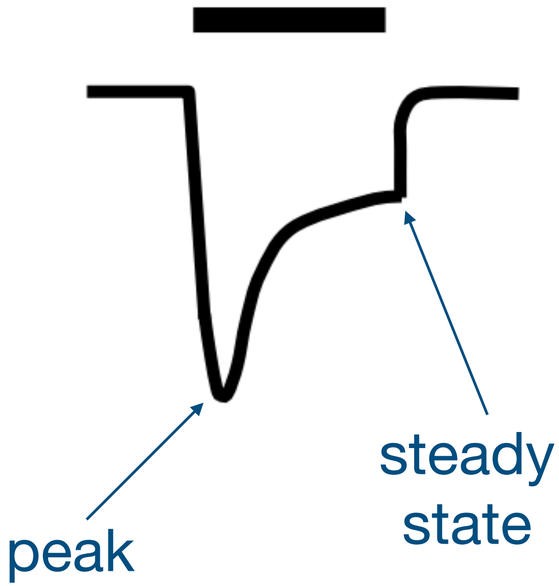


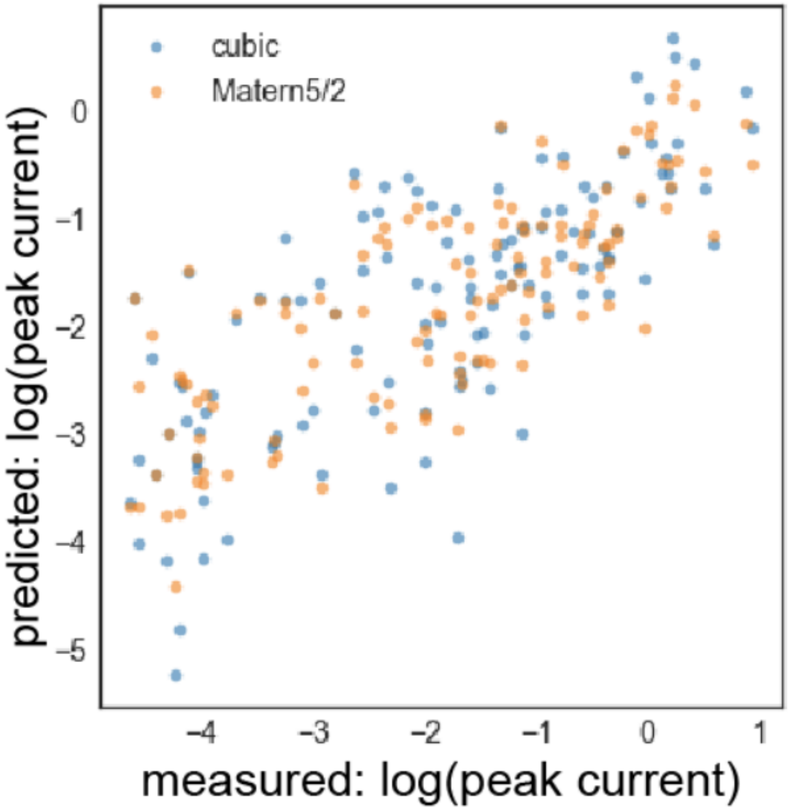
Models for functional properties are accurate



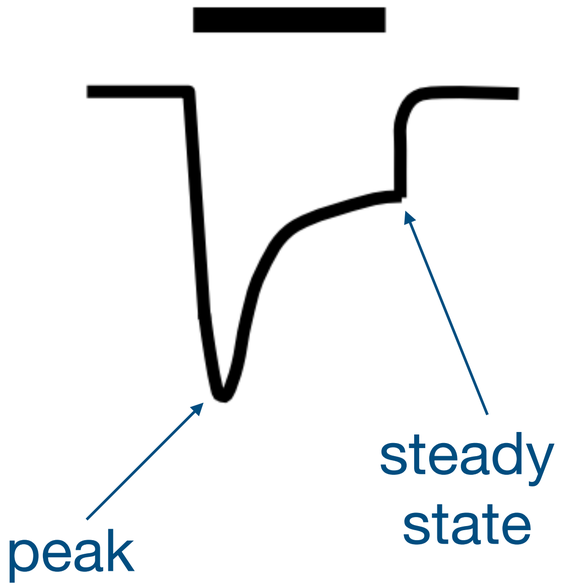


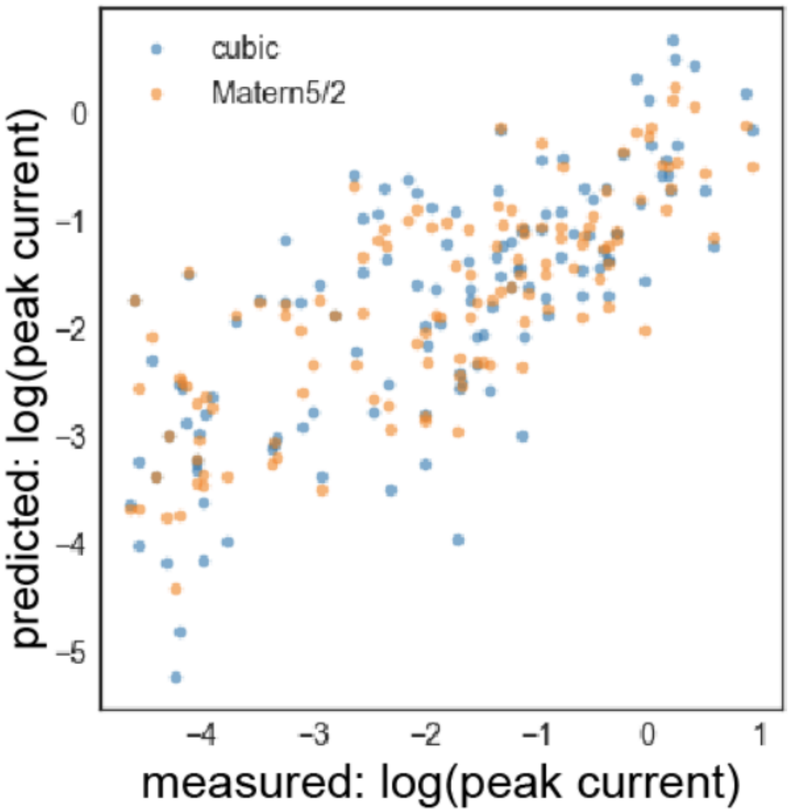
Models for functional properties are accurate

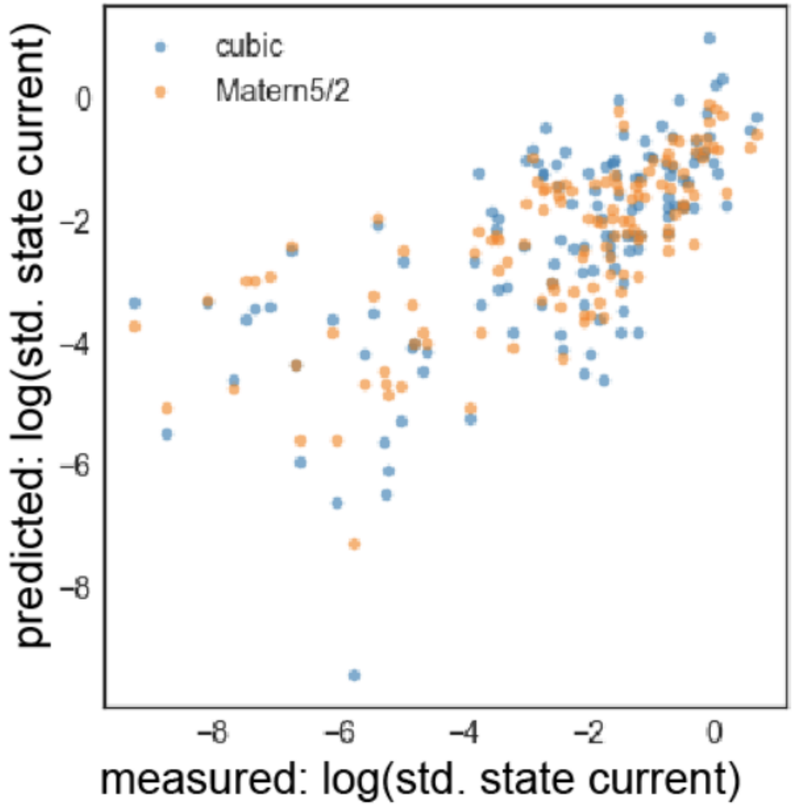




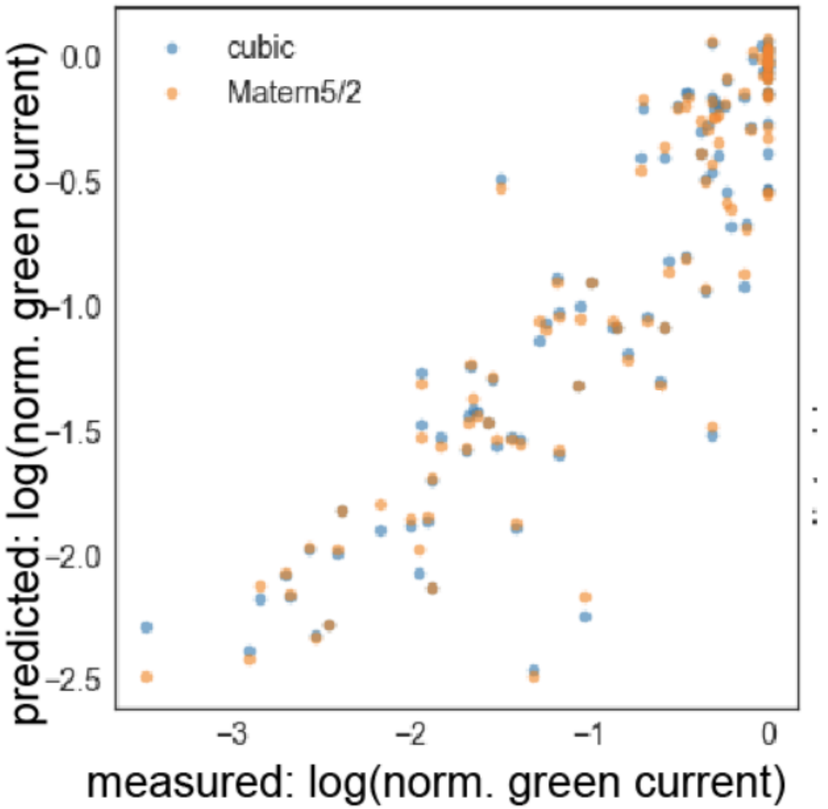
Models for functional properties are accurate

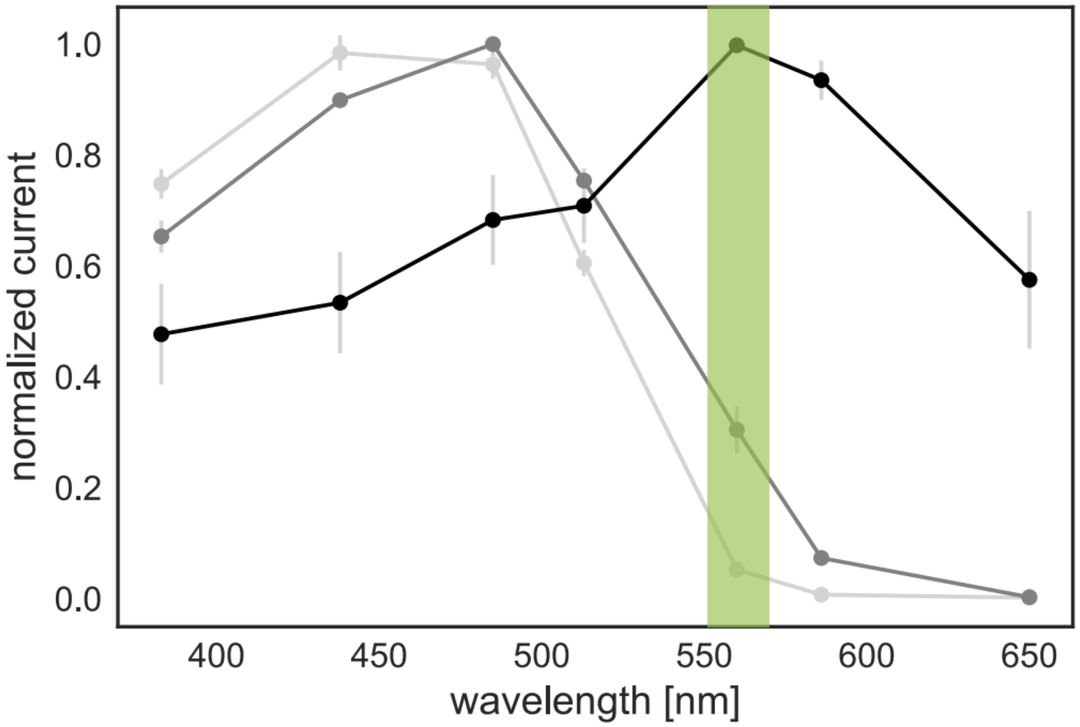




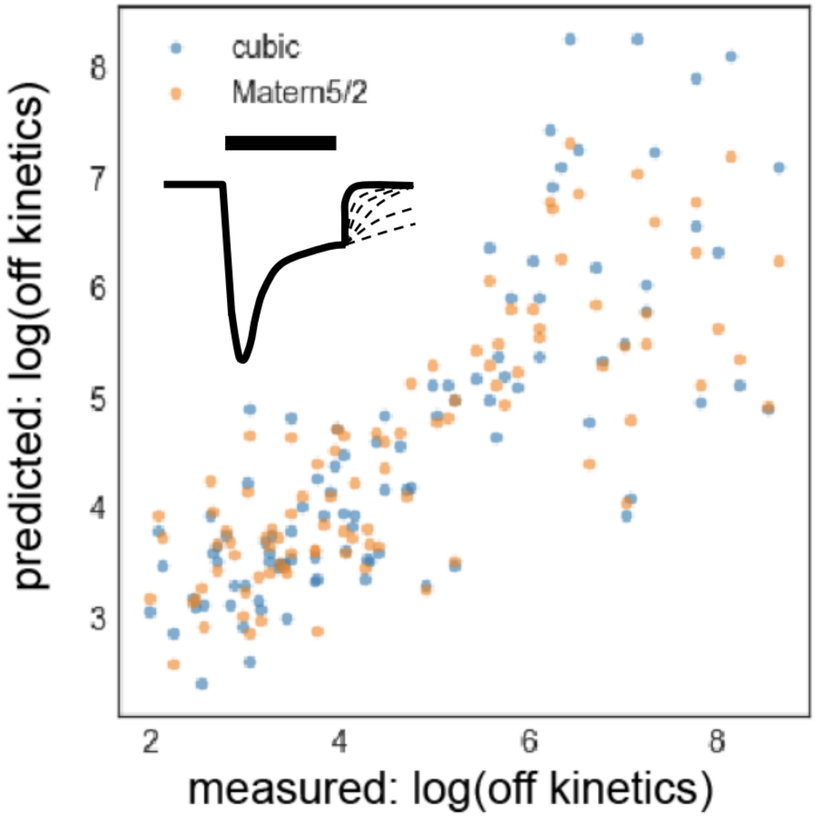


Models for functional properties are accurate

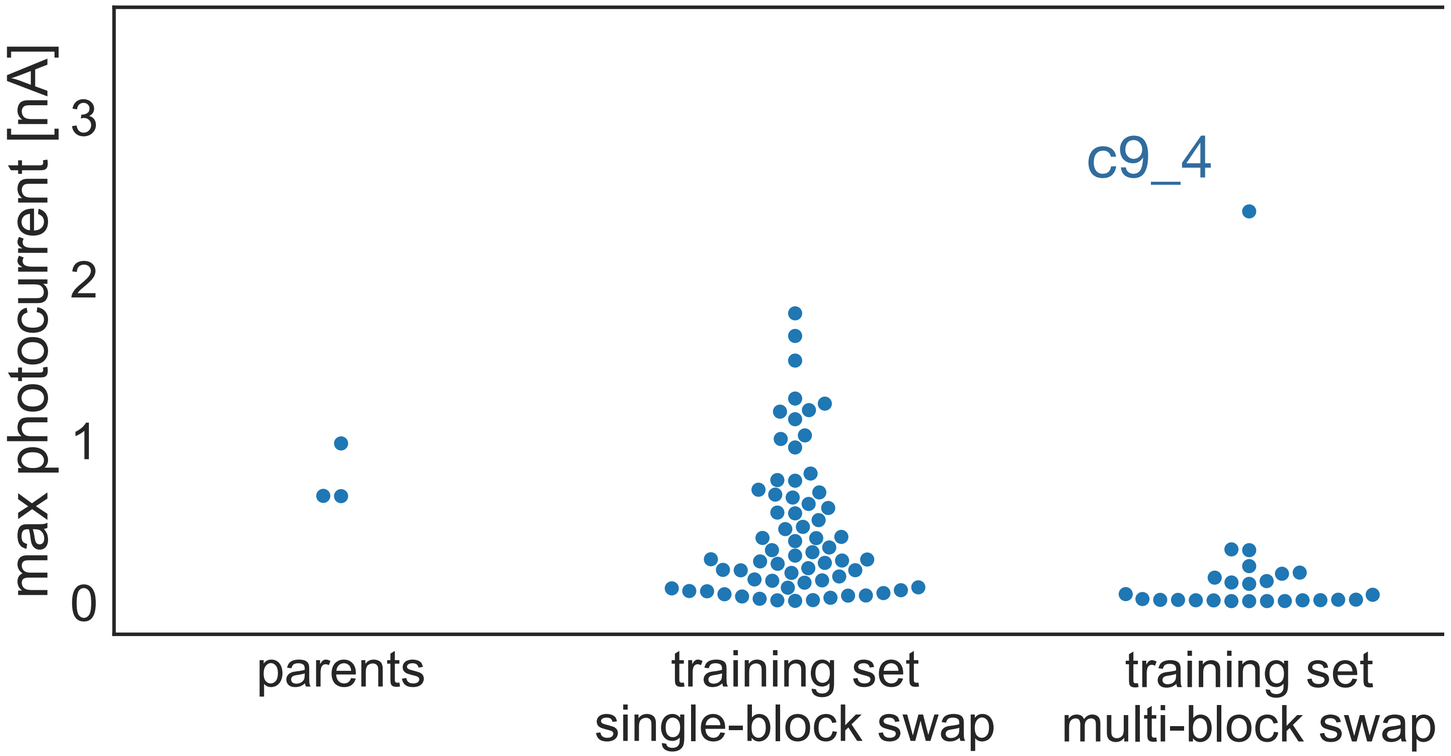




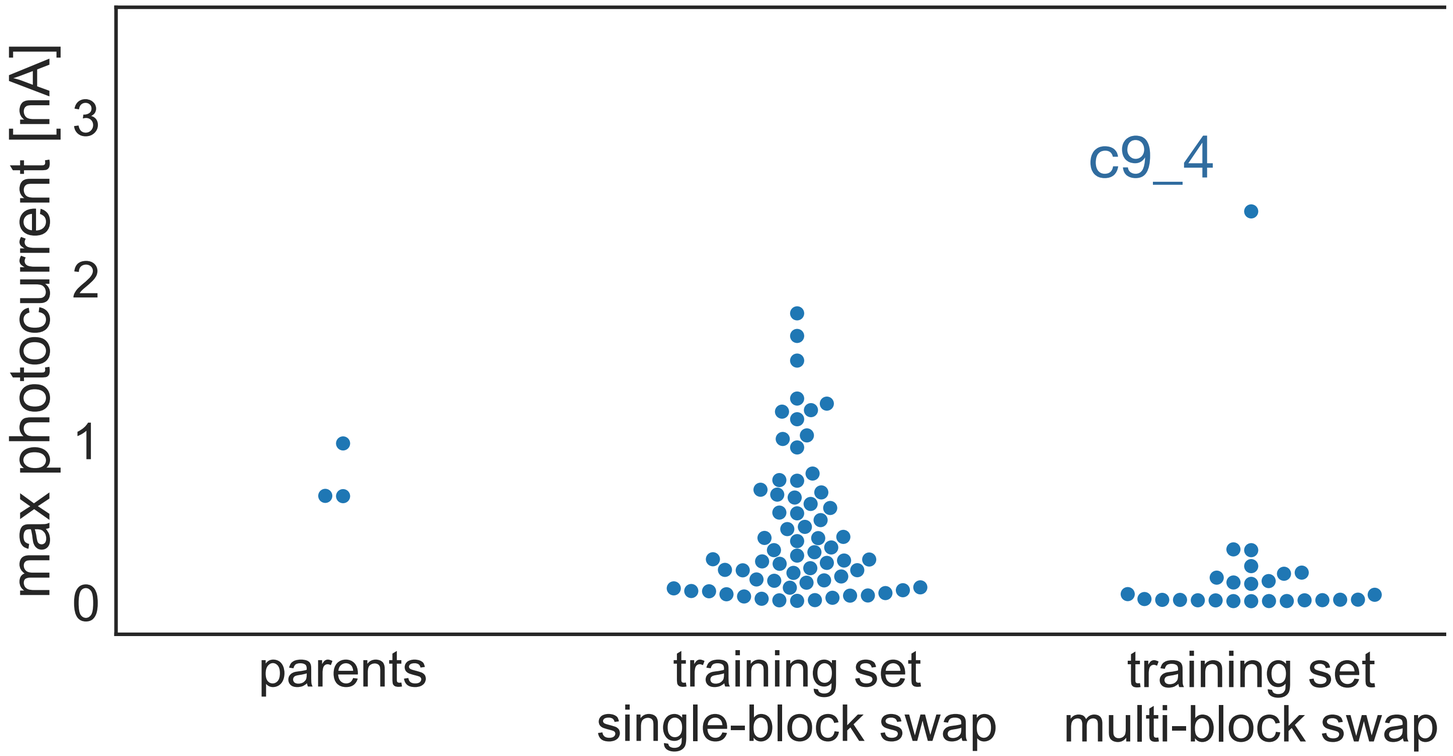
Models for functional properties are accurate



ML finds progressively stronger currents

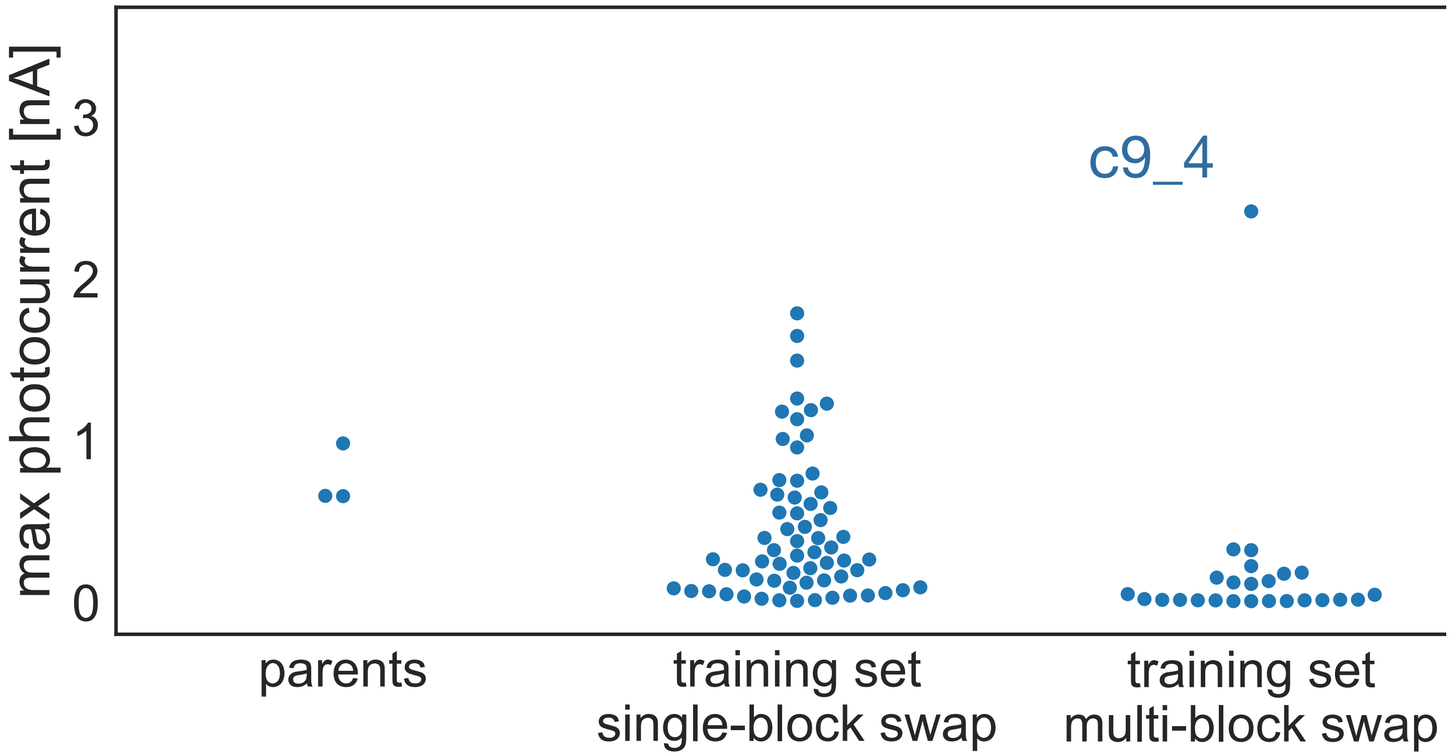


ML finds progressively stronger currents



Localization is a prerequisite to function

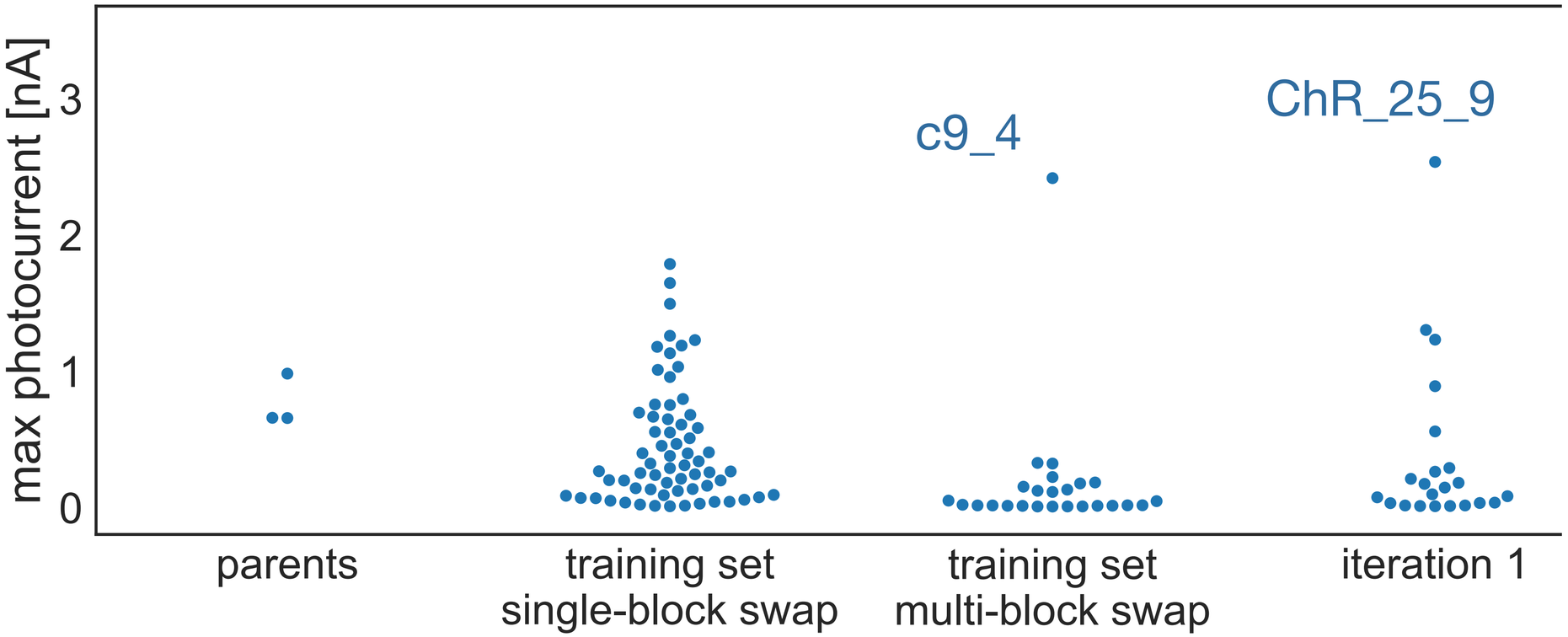
ML finds progressively stronger currents



Localization is a prerequisite to function

Make 30 sequences predicted to localize *and* function

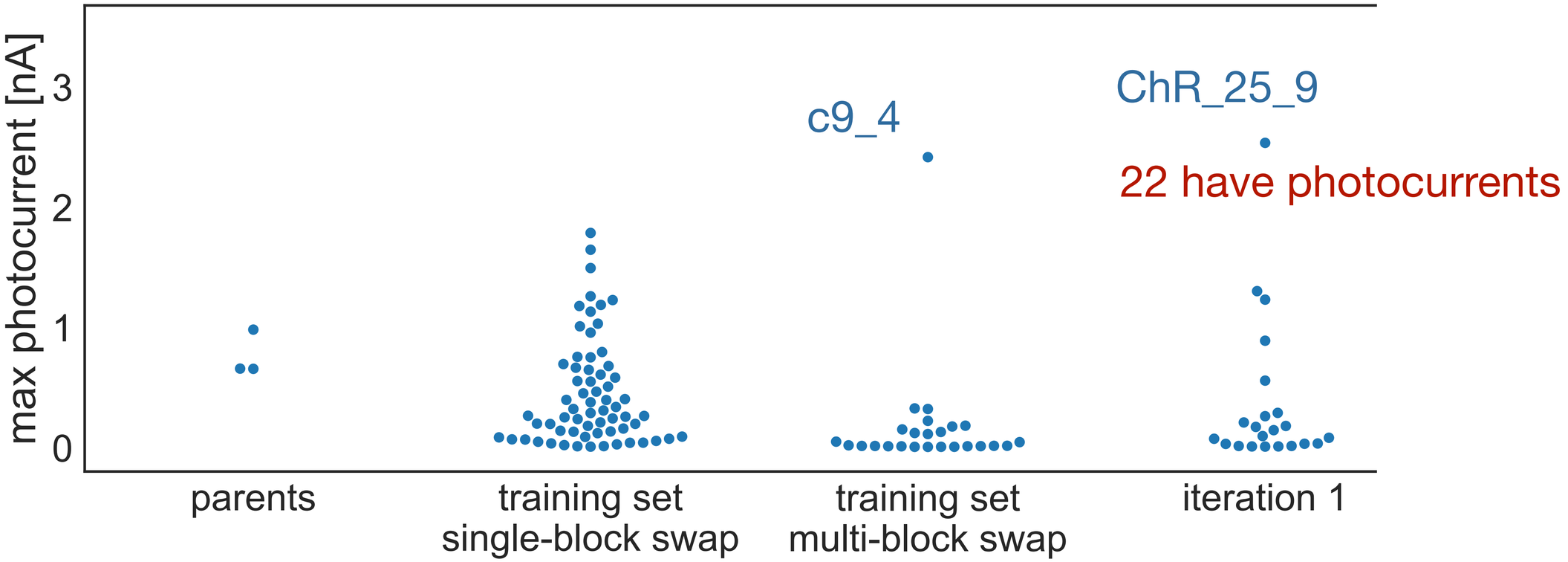
ML finds progressively stronger currents



Localization is a prerequisite to function

Make 30 sequences predicted to localize *and* function

ML finds progressively stronger currents



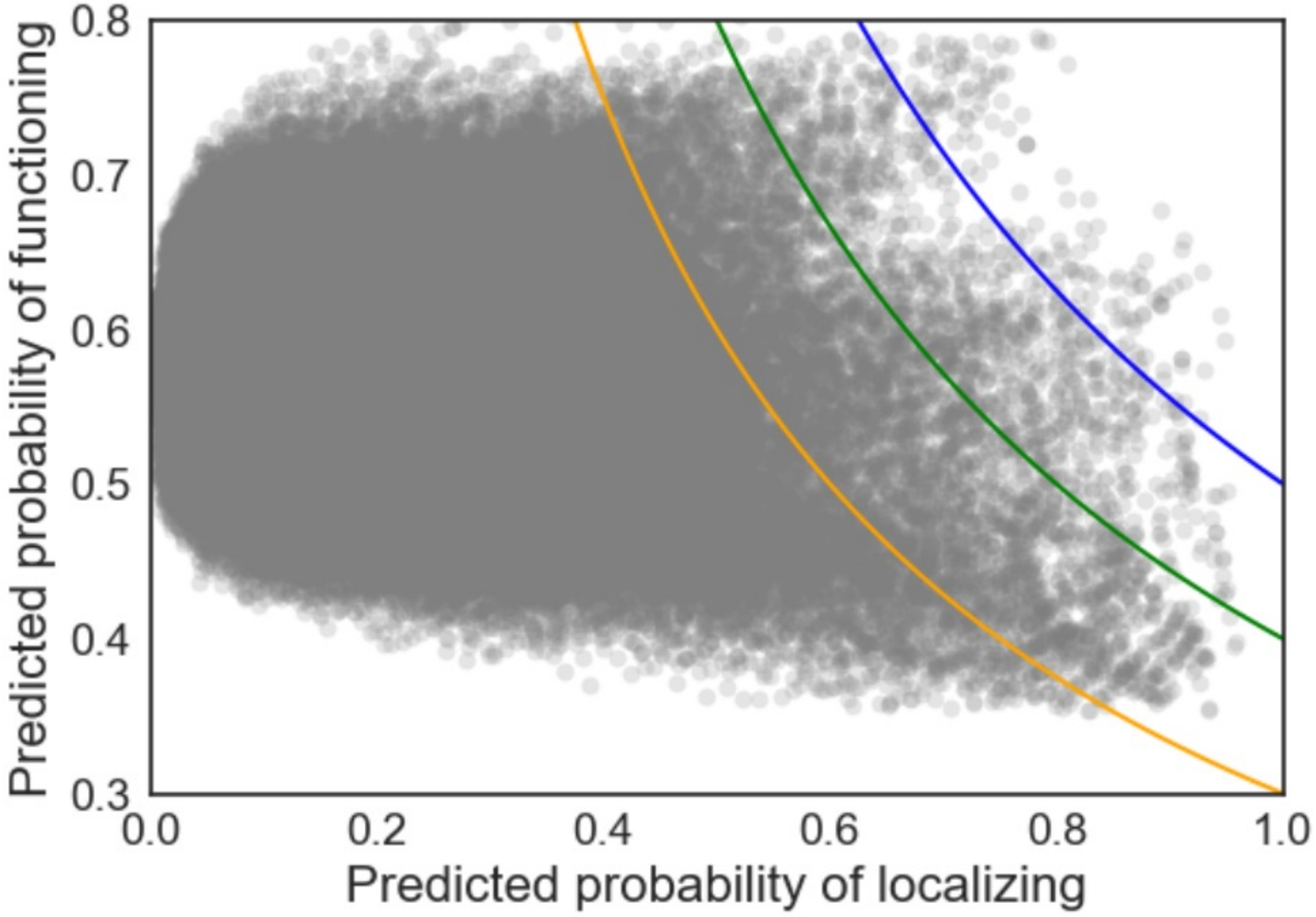
Localization is a prerequisite to function

Make 30 sequences predicted to localize *and* function

Classifiers for localization and currents

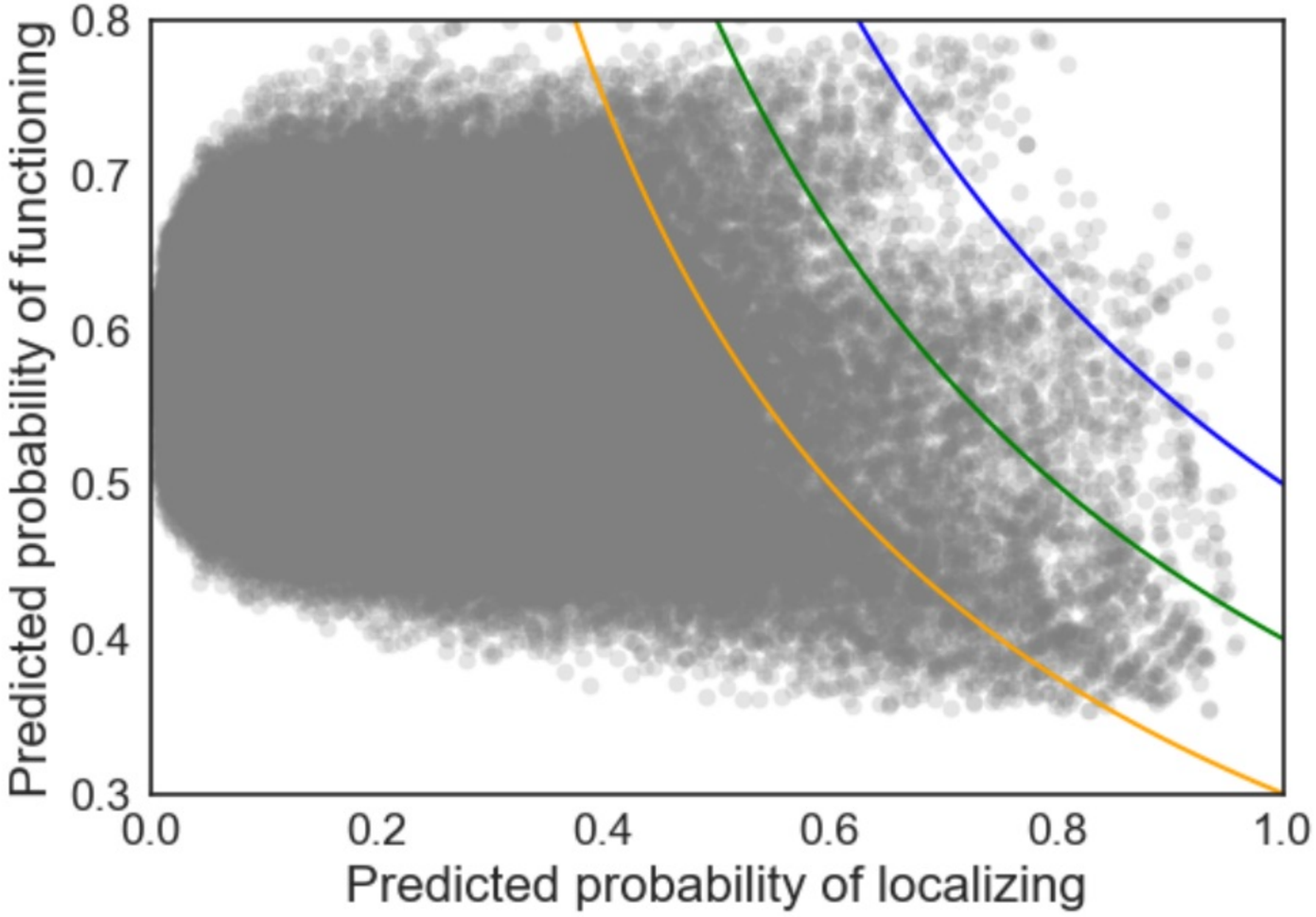
Want variants with interesting properties that also localize and have currents

Classifiers for localization and currents



Want variants with interesting properties that also localize and have currents

Classifiers for localization and currents



Number of chimeras

pp > 0.5: 135

pp > 0.4: 1161

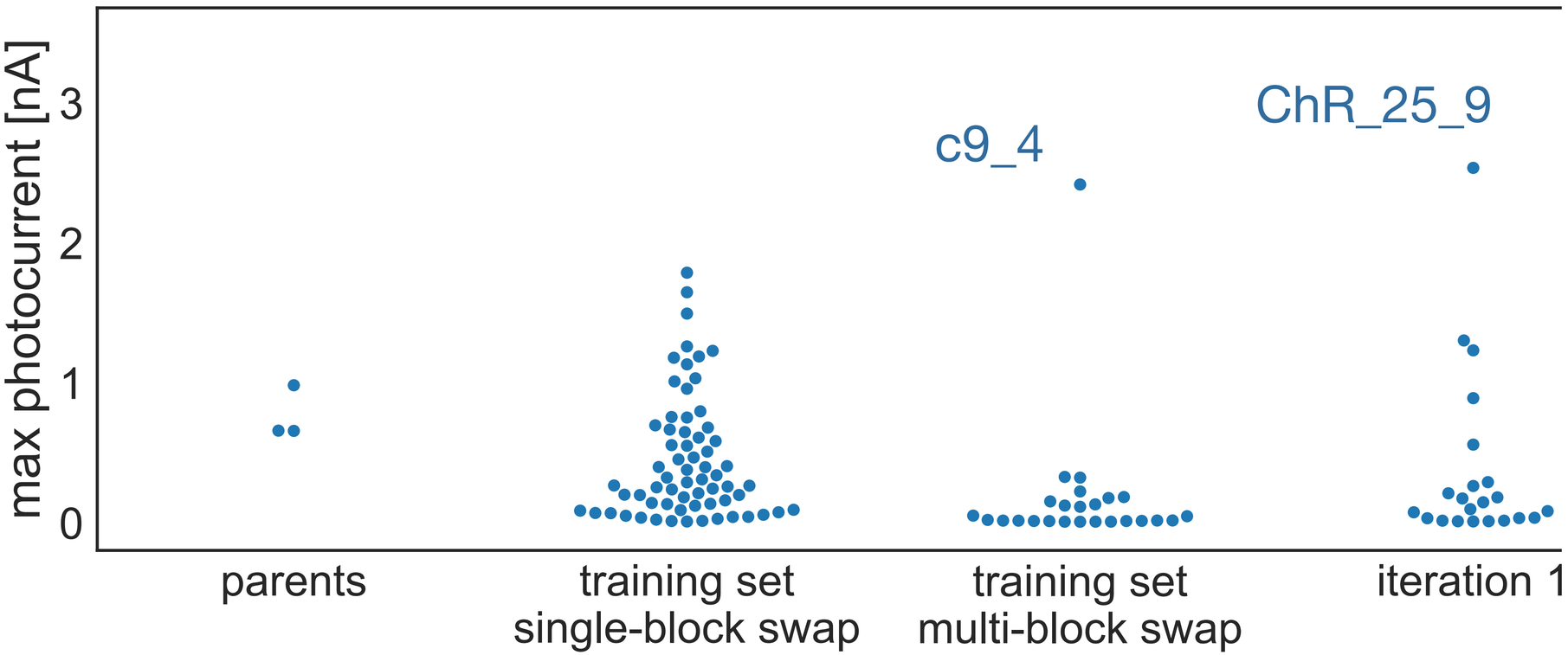
pp > 0.3: 5792

Want variants with interesting properties that also localize and have currents

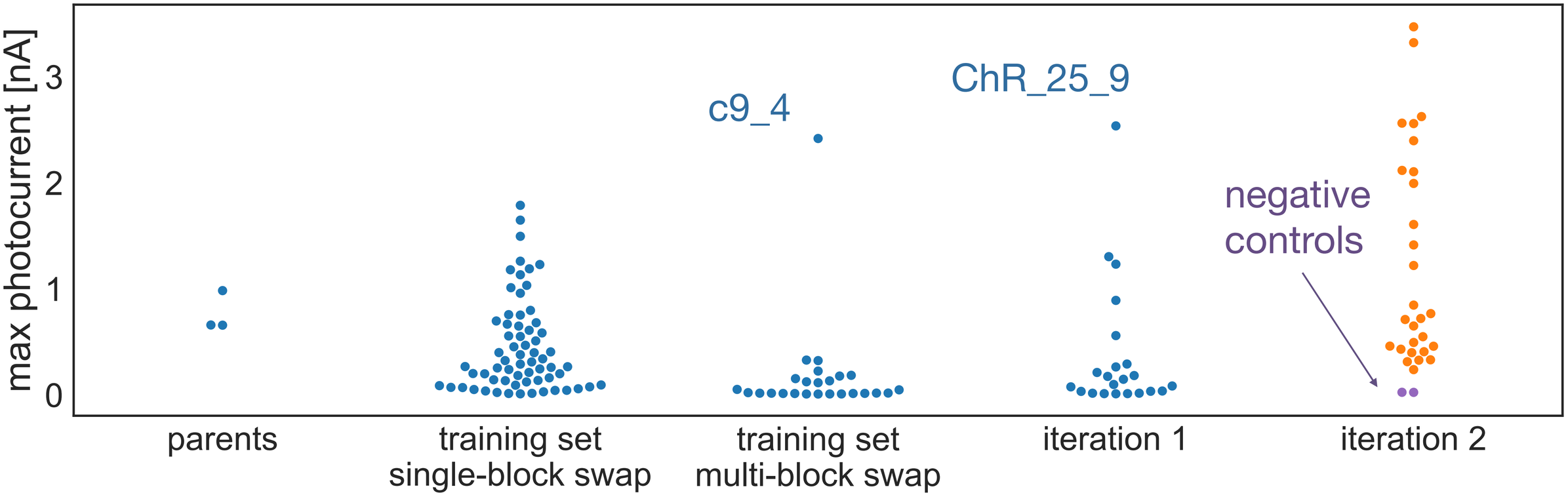
Choose 30 sequences

- 28 predicted to have optimal properties
- 2 predicted to be non-functional (single blocks away from c9_4)

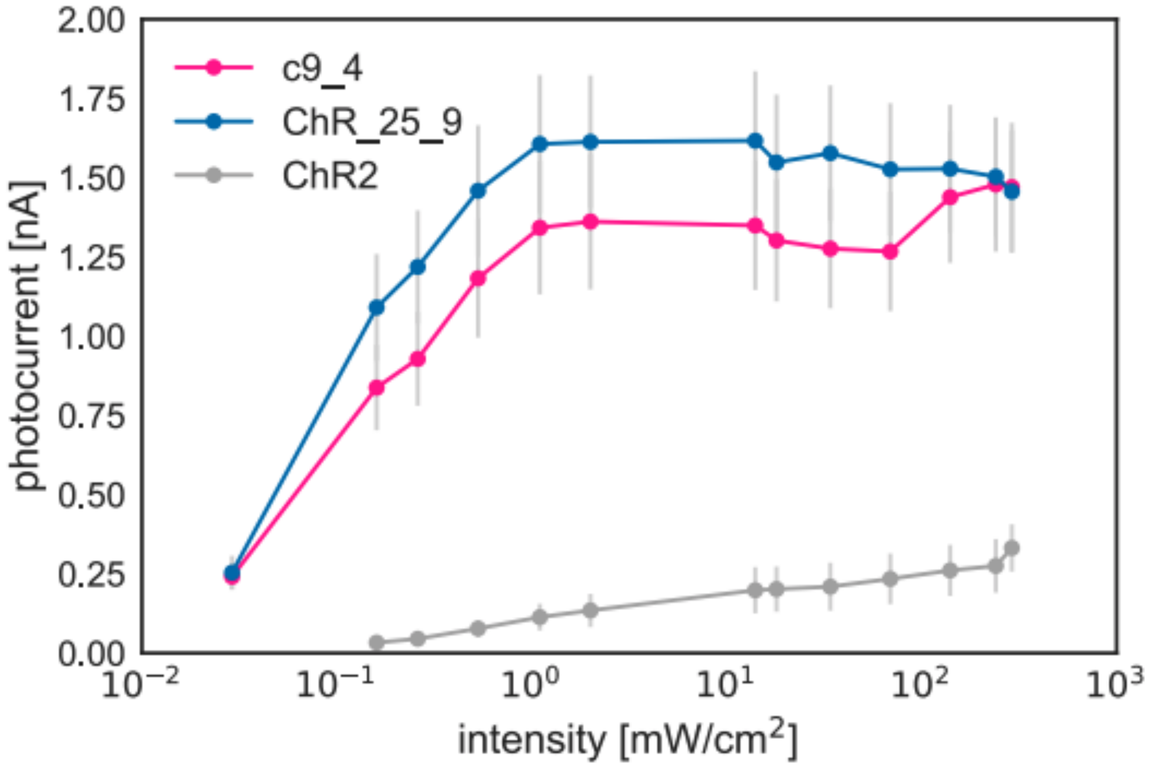
ML finds progressively stronger currents



ML finds progressively stronger currents

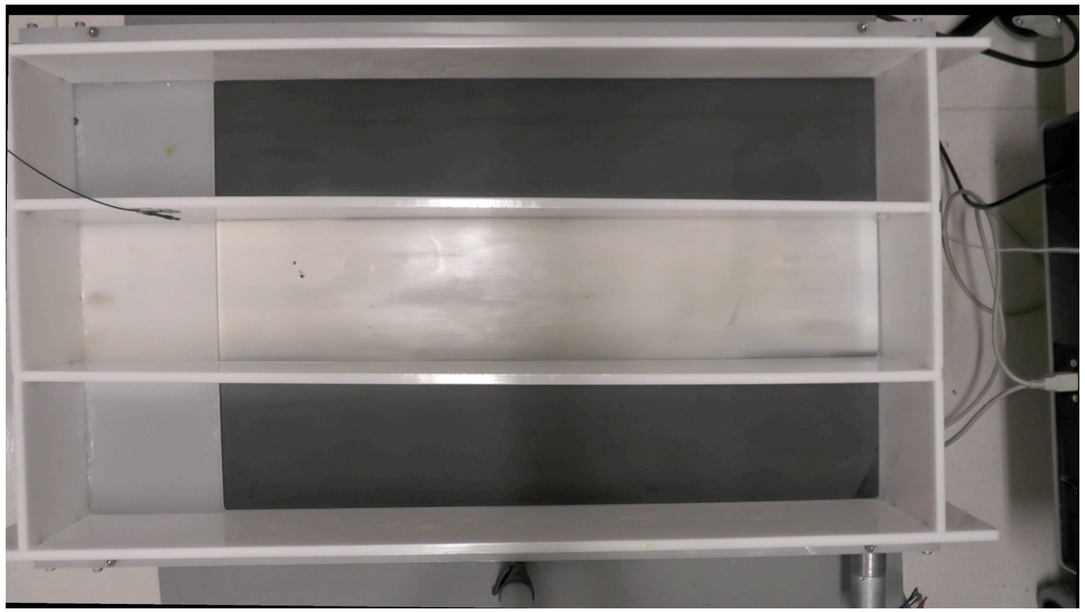


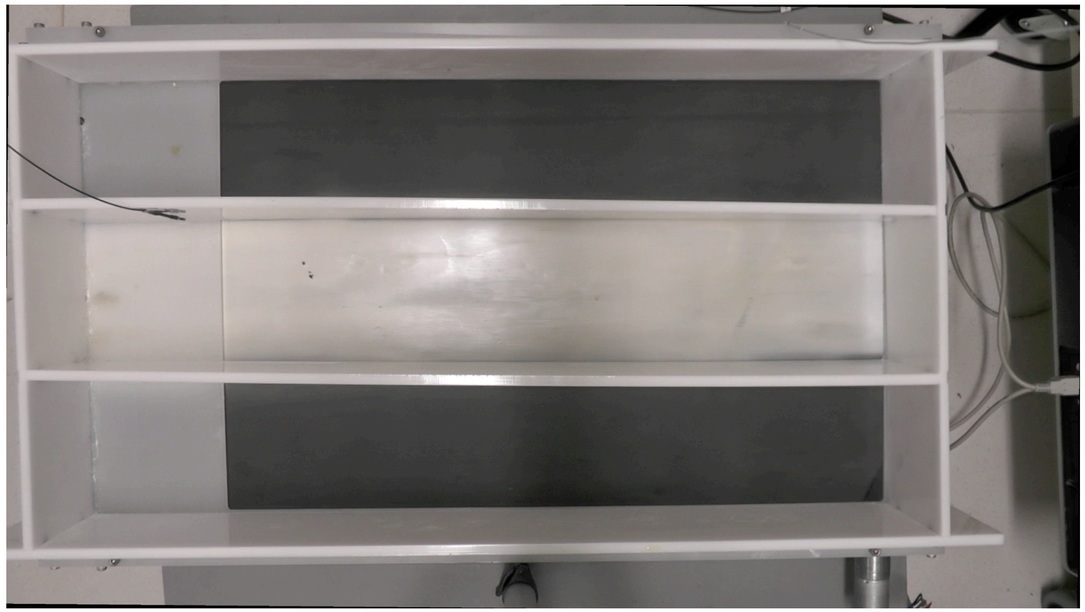
Engineered ChRs are active in vivo



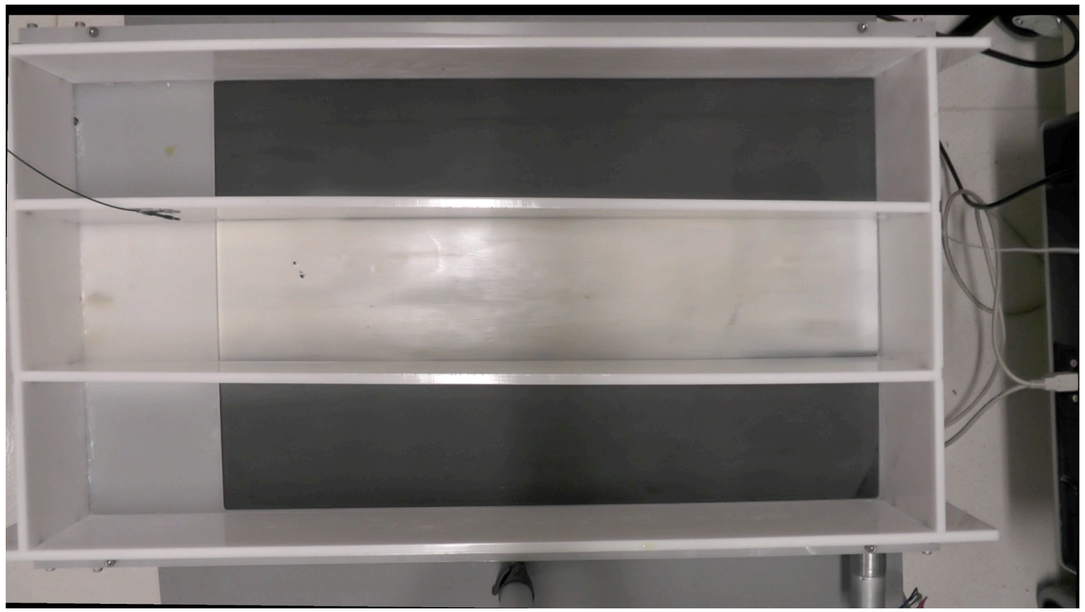
(In mouse neurons, direct intracranial injection)

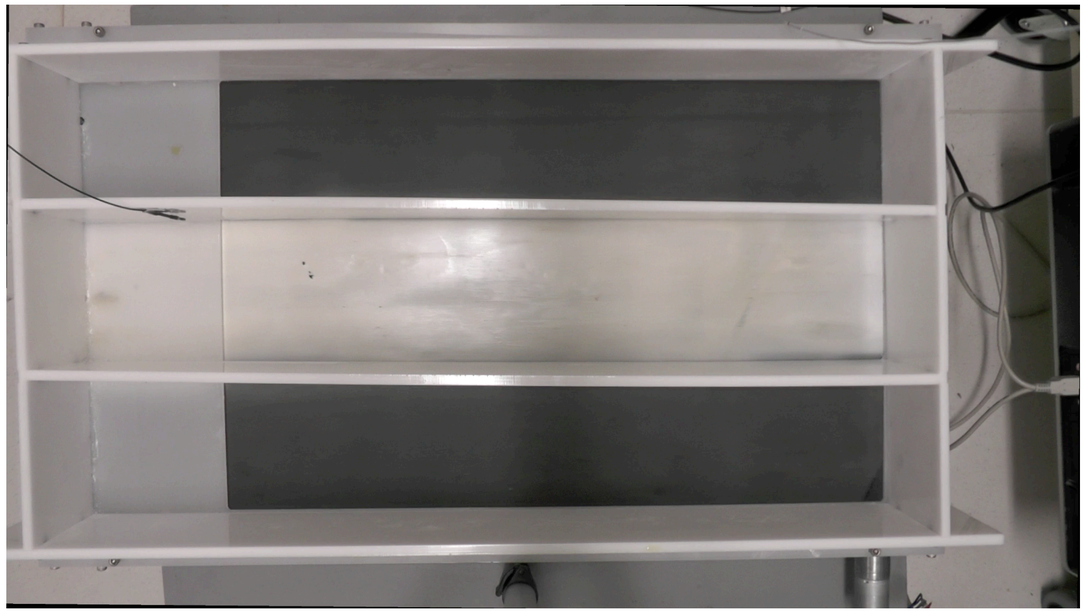
Designed variants modify mouse behavior



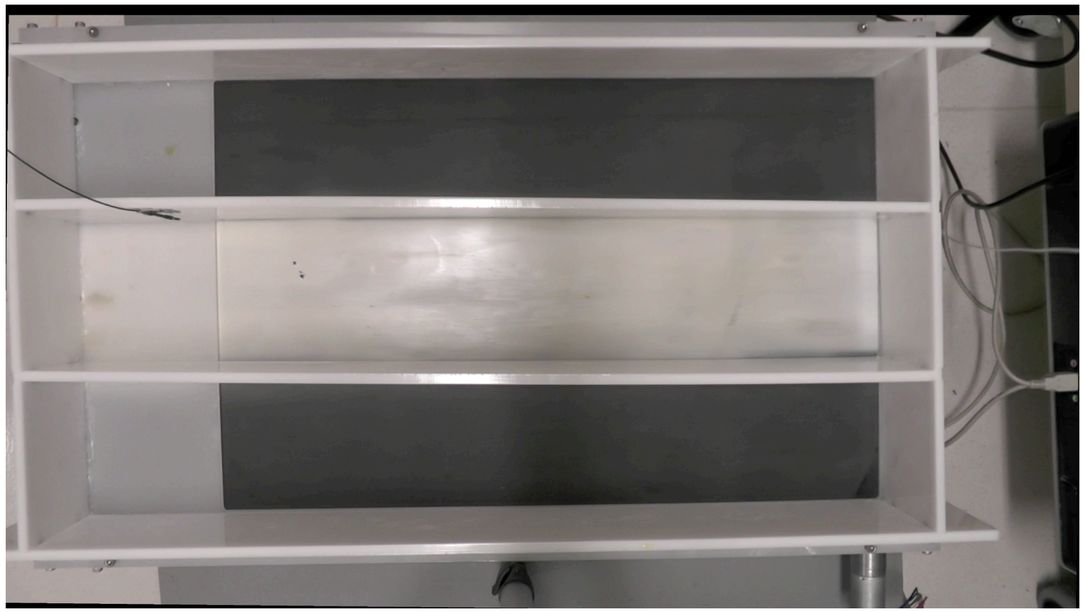


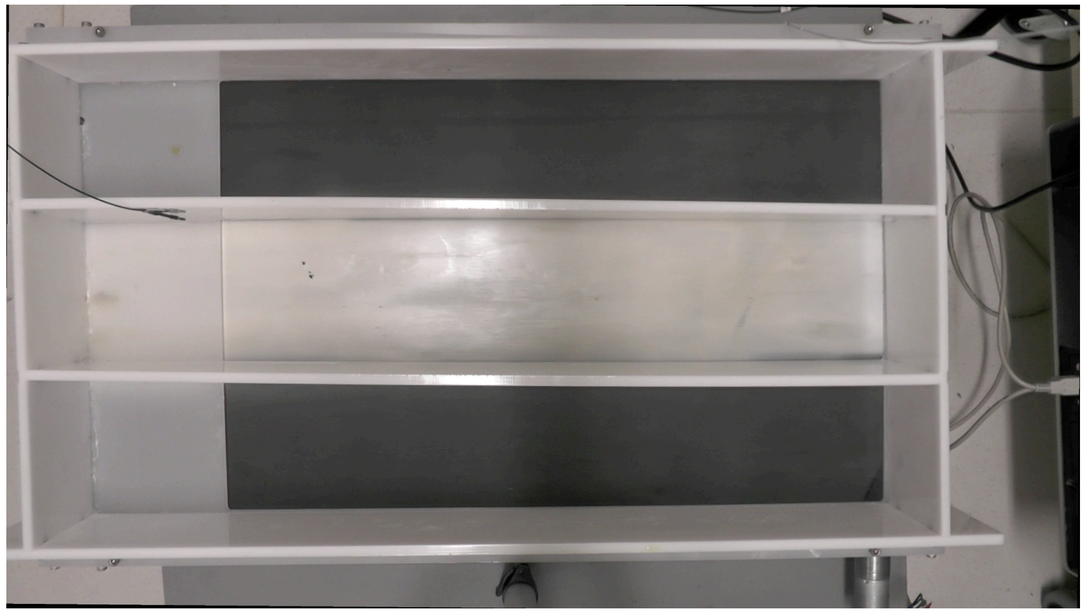
Designed variants modify mouse behavior



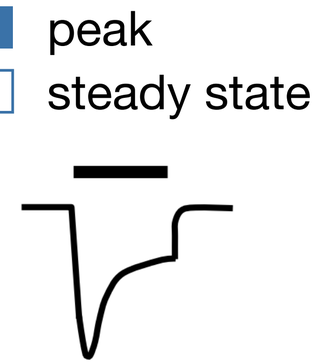


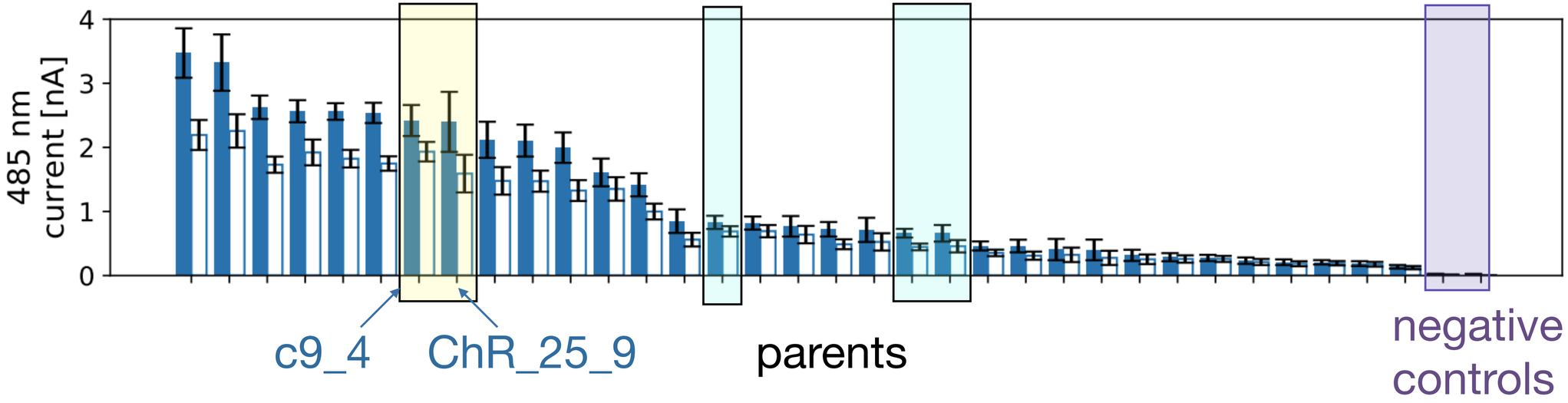
Designed variants modify mouse behavior



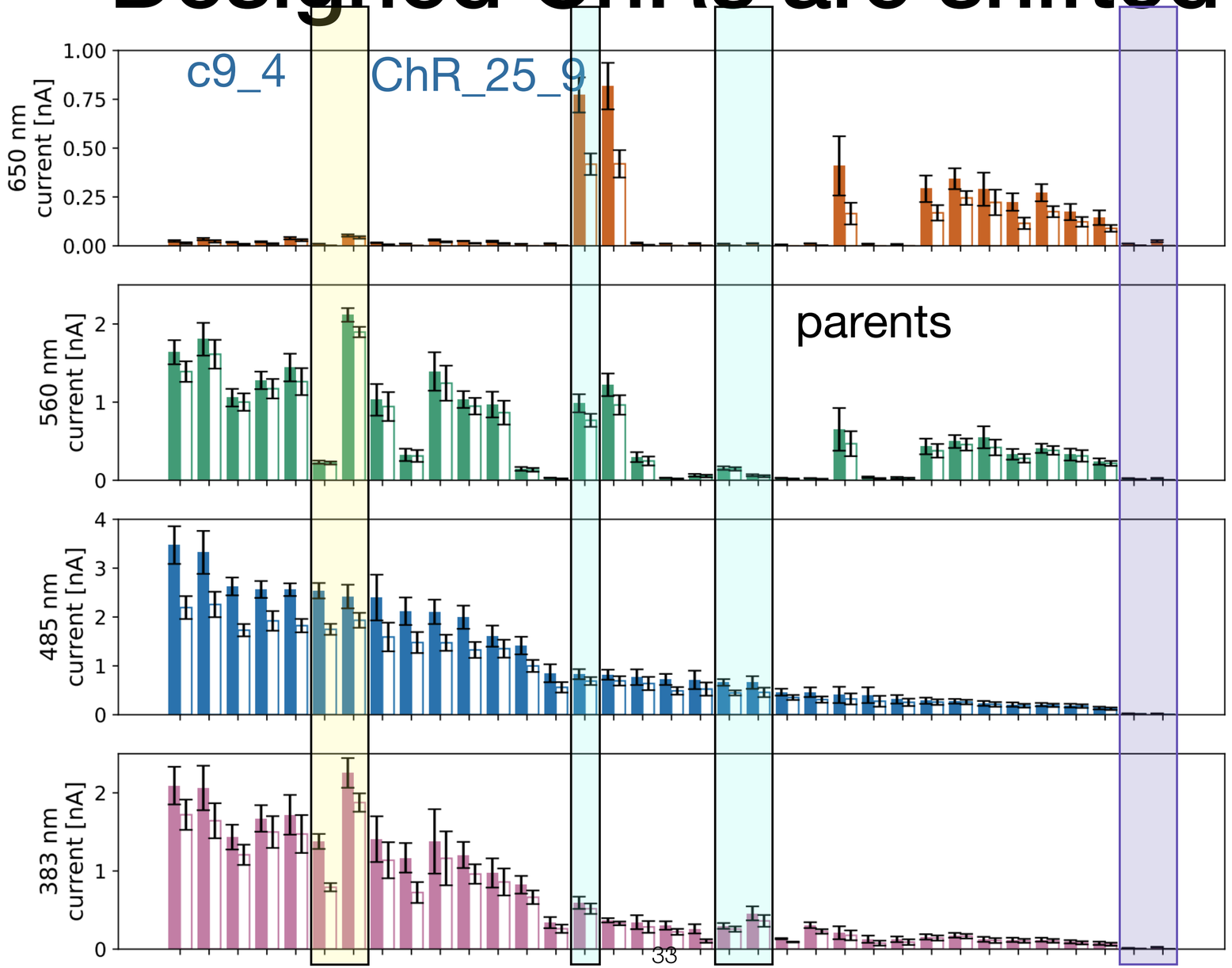


Designed ChRs are shifted

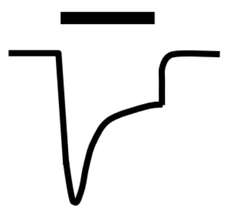




Designed ChRs are shifted



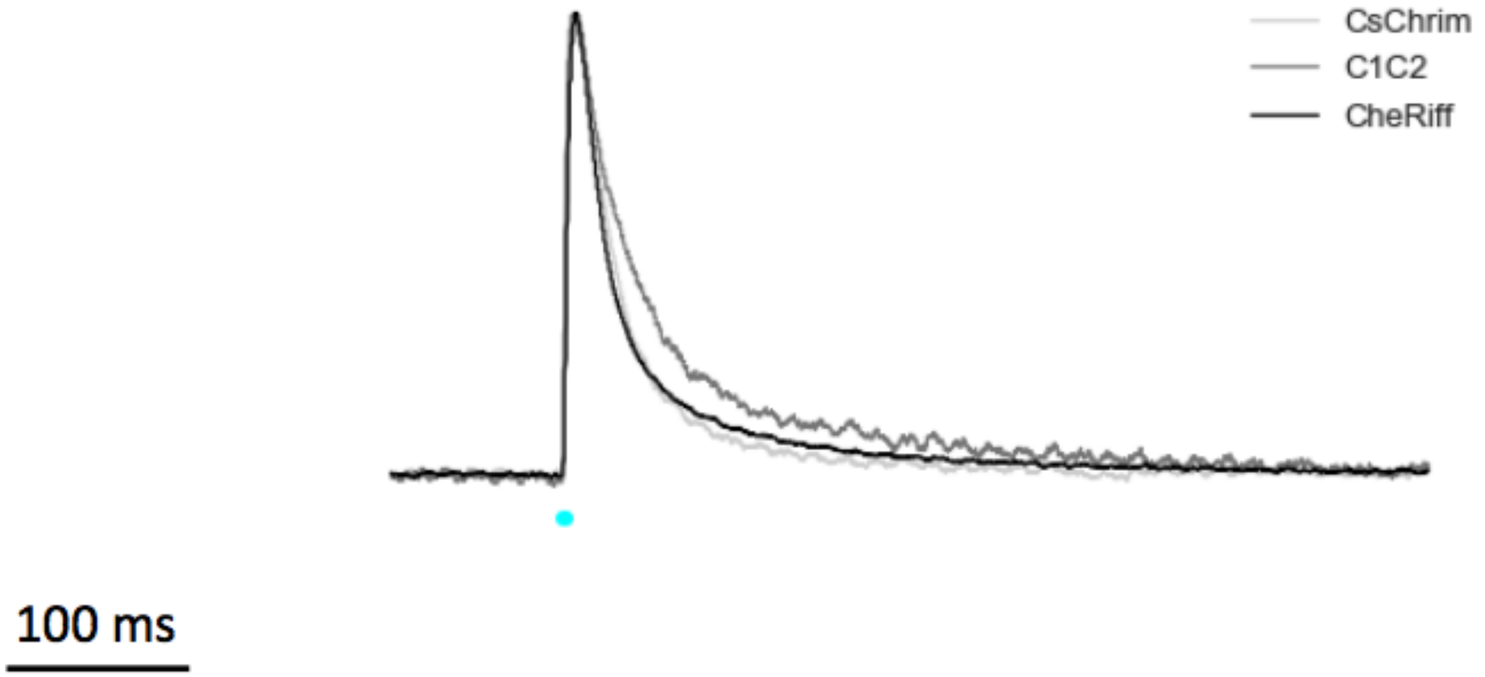
peak
steady state



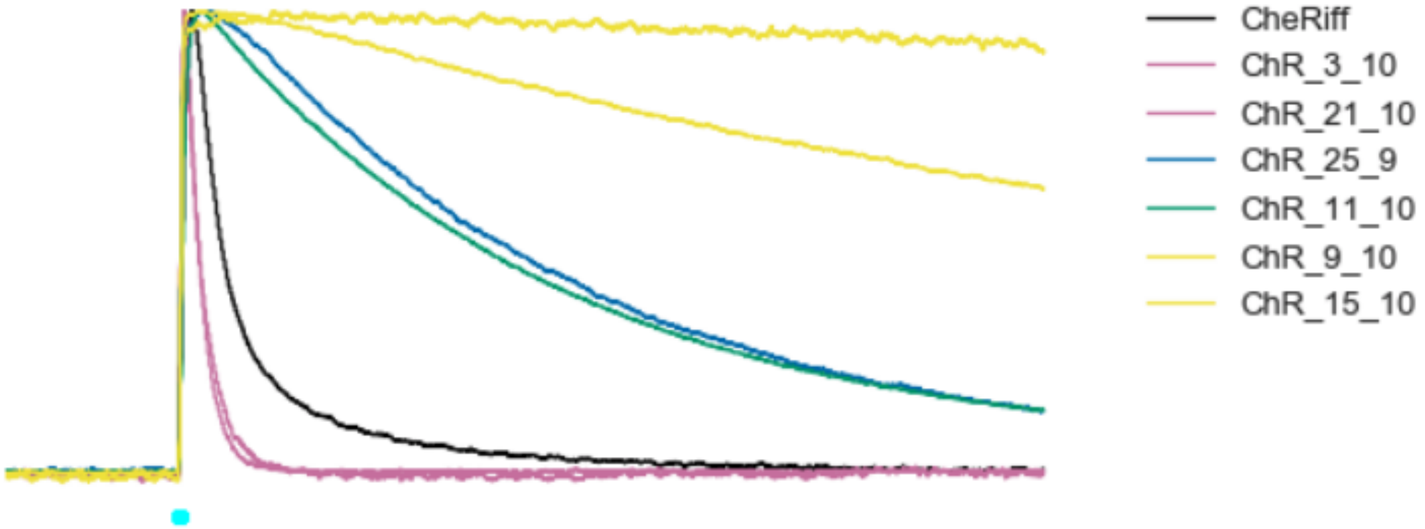
negative controls

Designed ChRs have range of kinetics

Designed ChRs have range of kinetics



Designed ChRs have range of kinetics

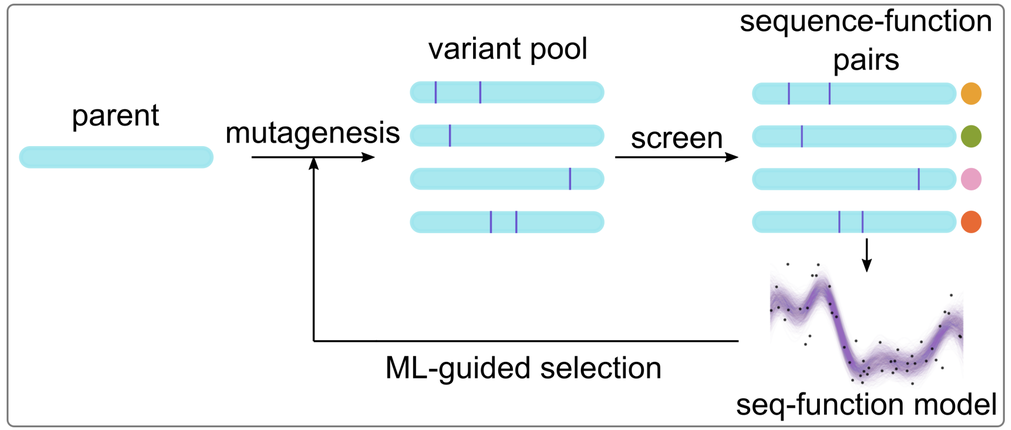


ChR engineering conclusions

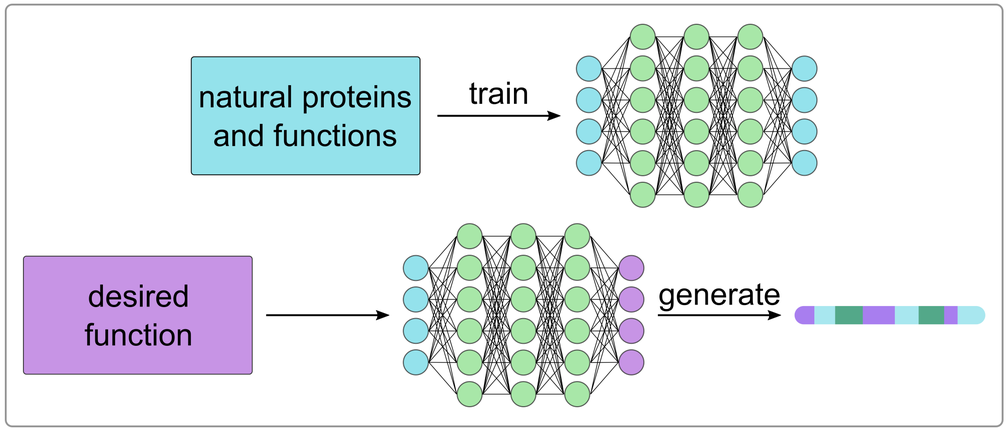
- Engineer multiple properties without a high-throughput screen
- Accurate GP models with only ~120 training points
- Highly-sensitive ChRs with optogenetic applications

Optimization vs generation

Machine-learning guided directed evolution



Machine-generated *de novo* proteins

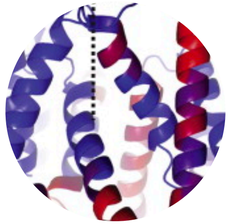


Designer channelrhodopsins

Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

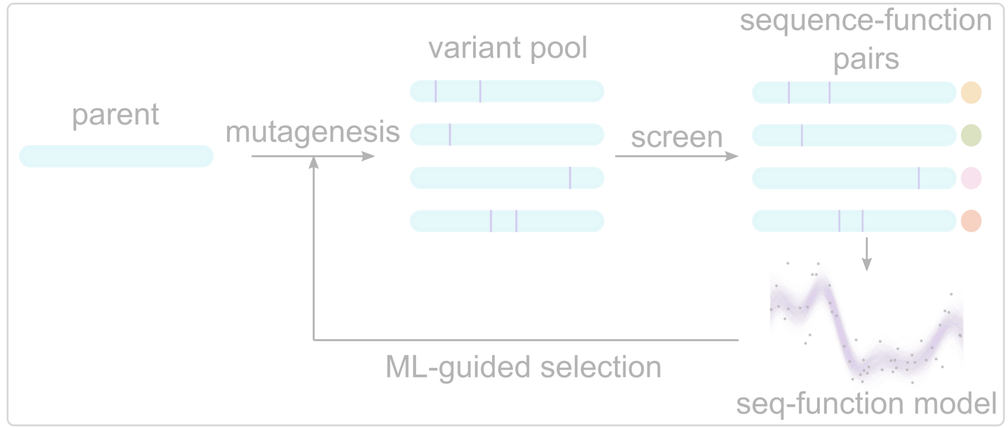
Signal peptide generation



Wu, Yang *et al., 2020*

Optimization vs generation

Machine-learning guided directed evolution

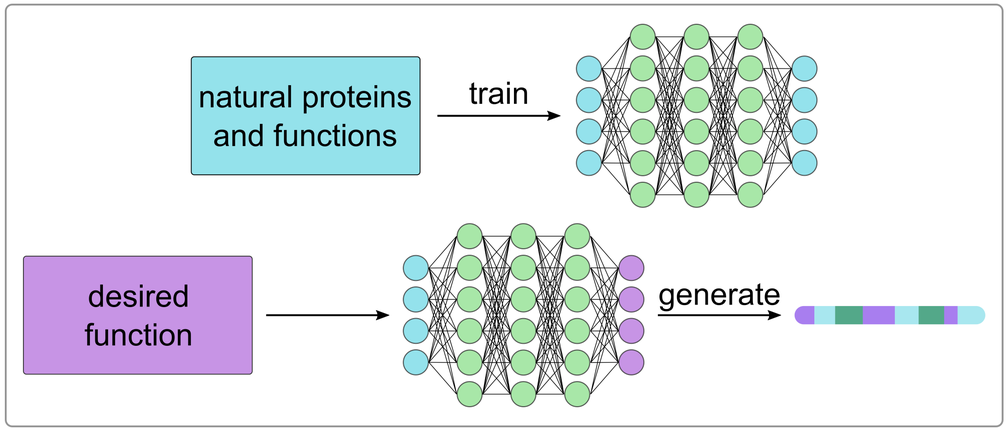


Designer channelrhodopsins

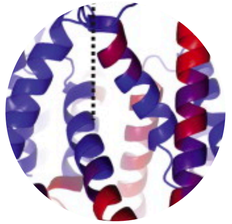
Bedbrook, Yang *et al.* 2017

Bedbrook, Yang *et al.* 2019

Machine-generated *de novo* proteins

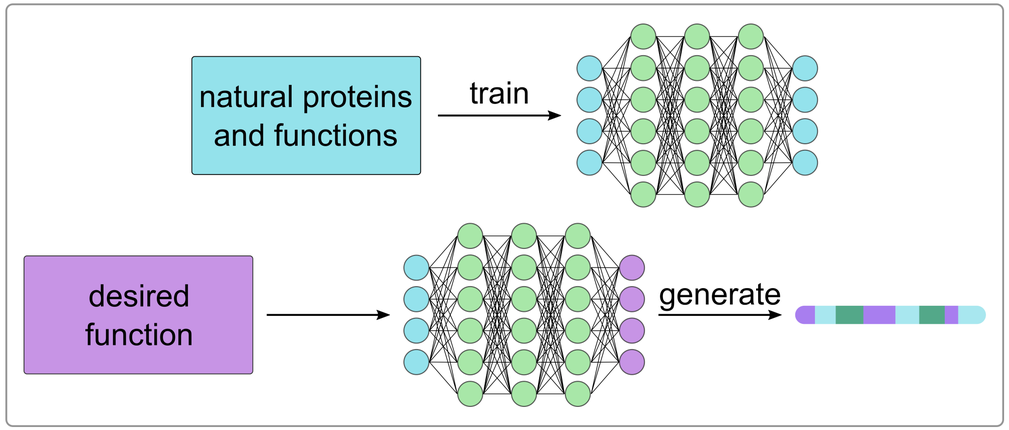


Signal peptide generation

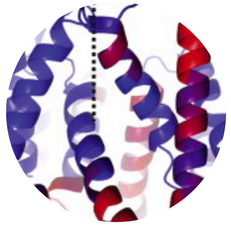


Wu, Yang *et al., 2020*

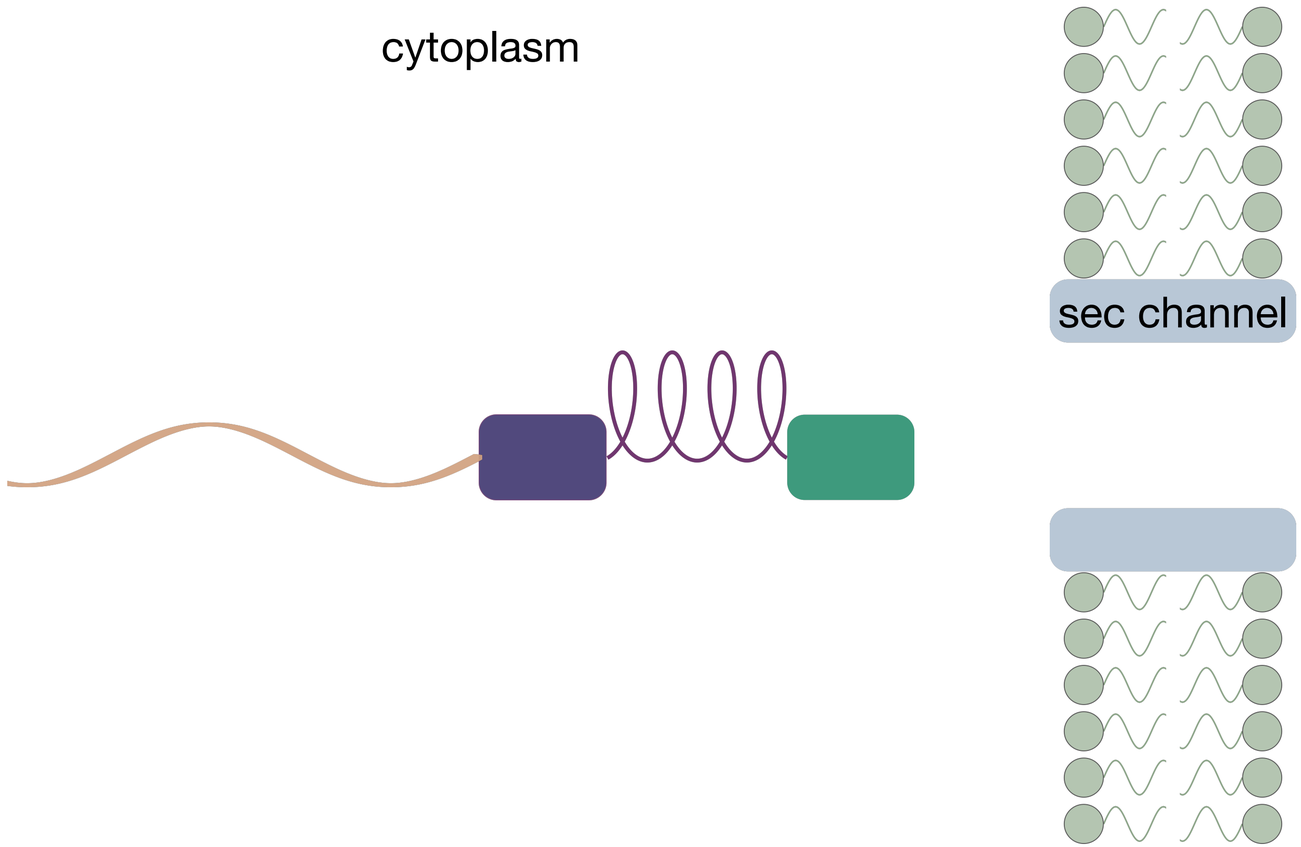
Machine-generated *de novo* proteins



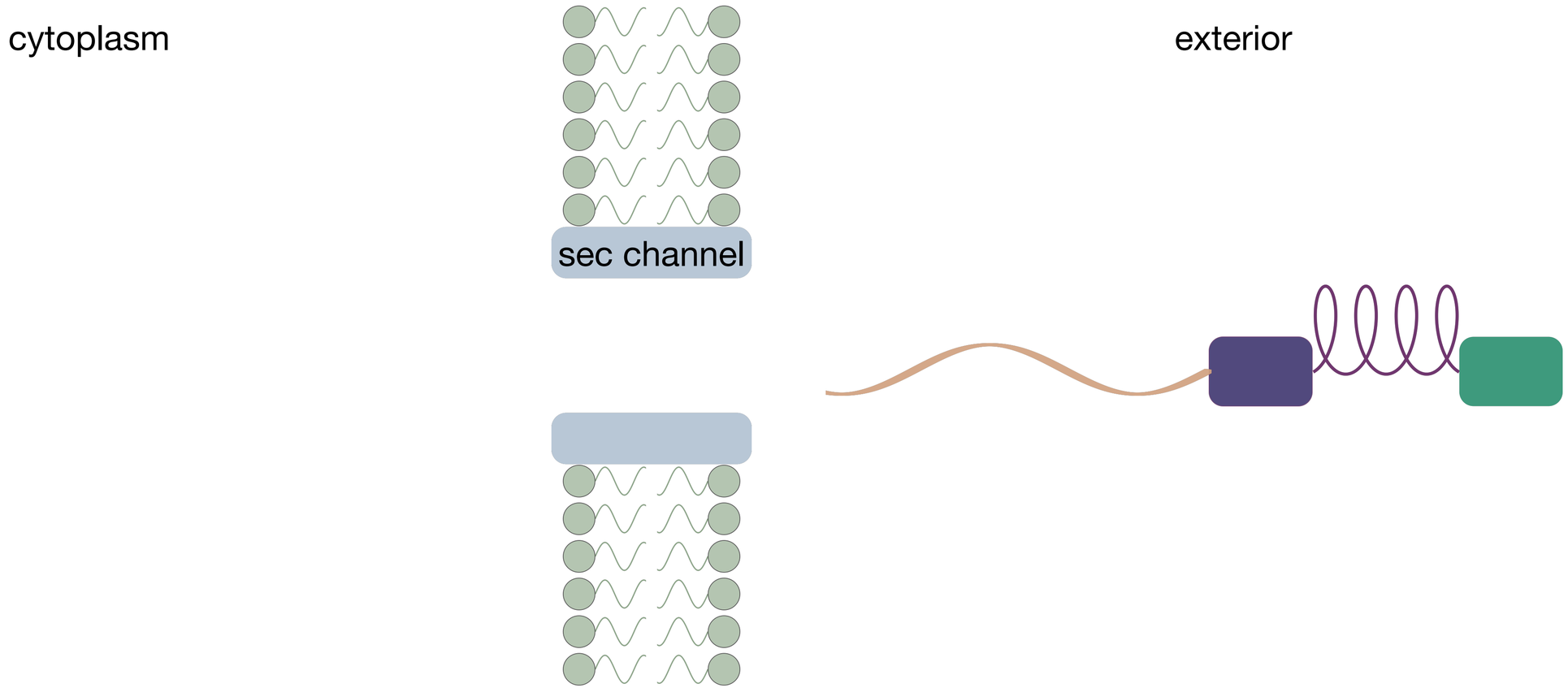
Signal peptide generation



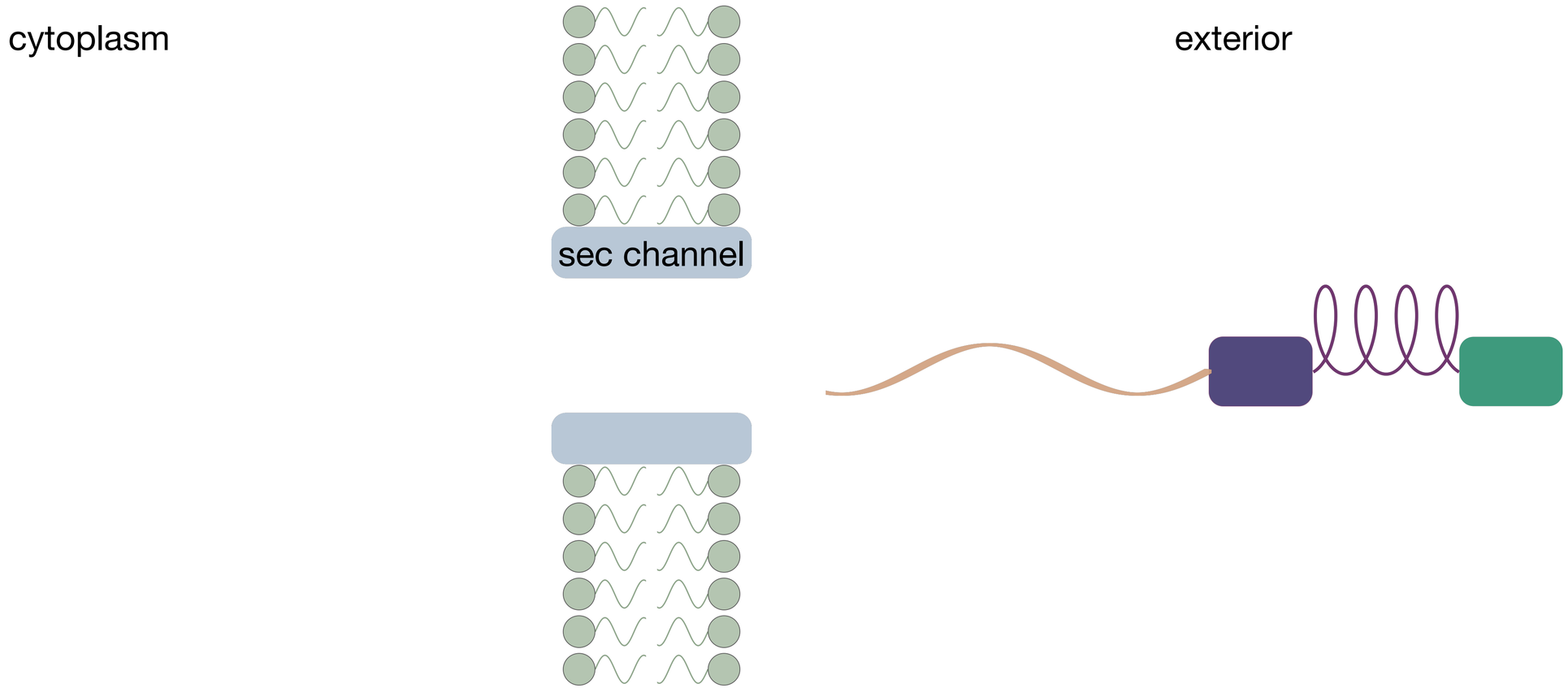
Signal peptides are secretion signals



Signal peptides are secretion signals

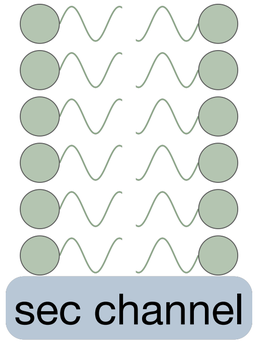


Signal peptides are secretion signals

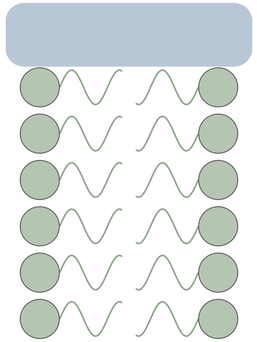


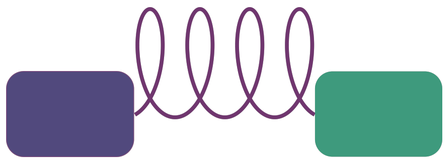
Signal peptides are secretion signals

cytoplasm

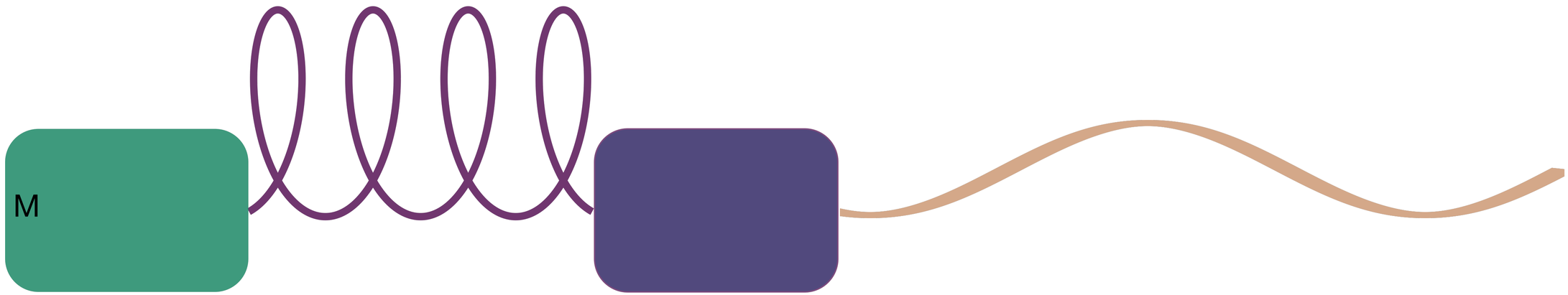


exterior

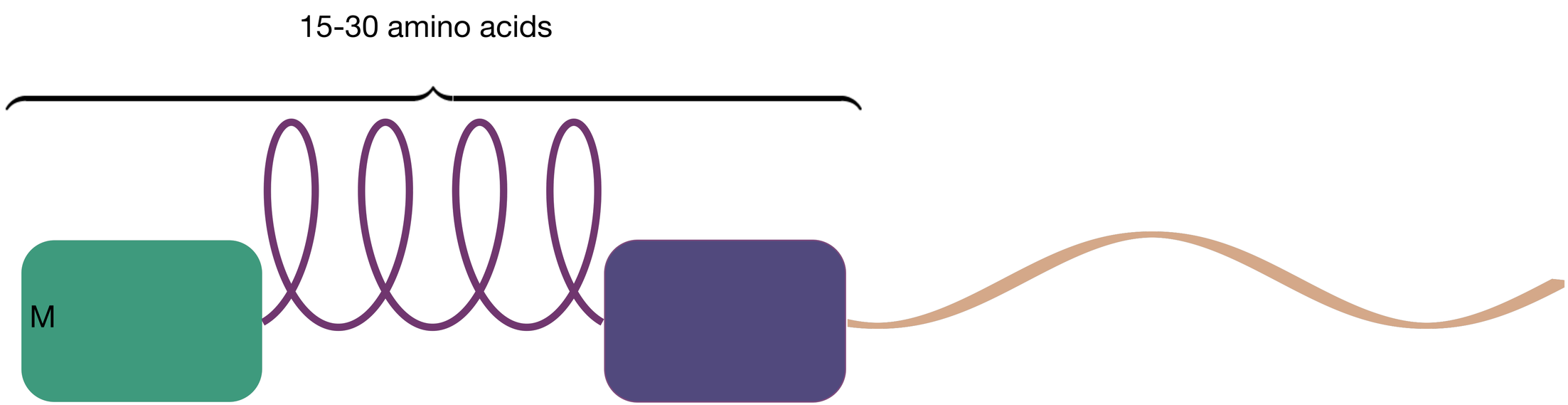




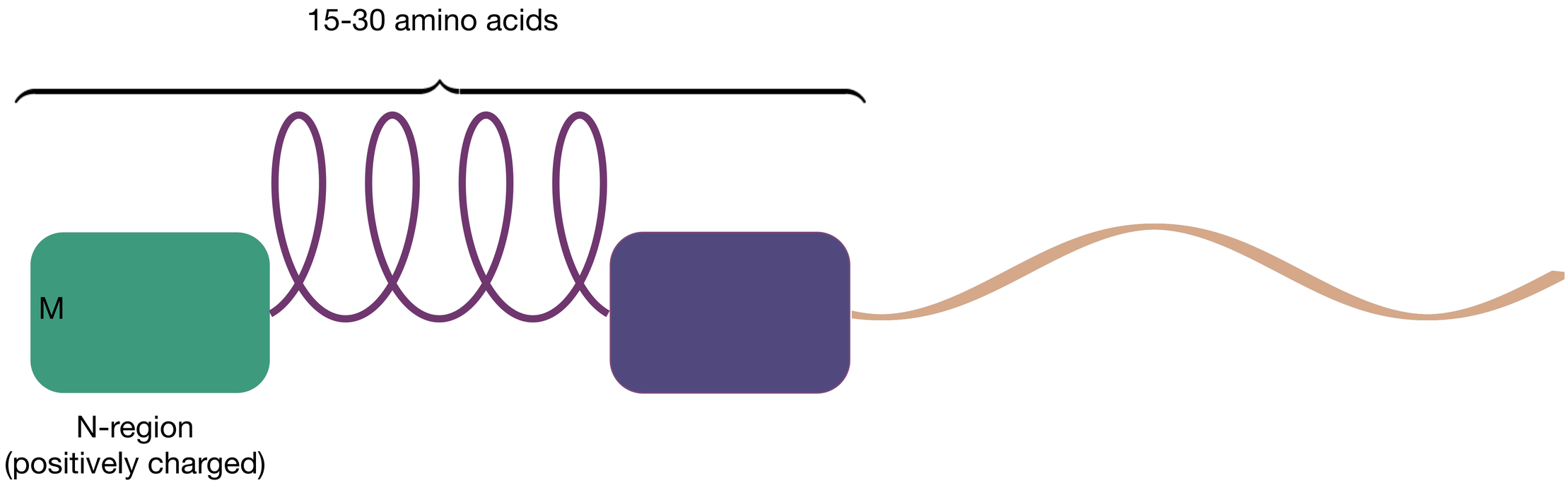
Signal peptides are secretion signals



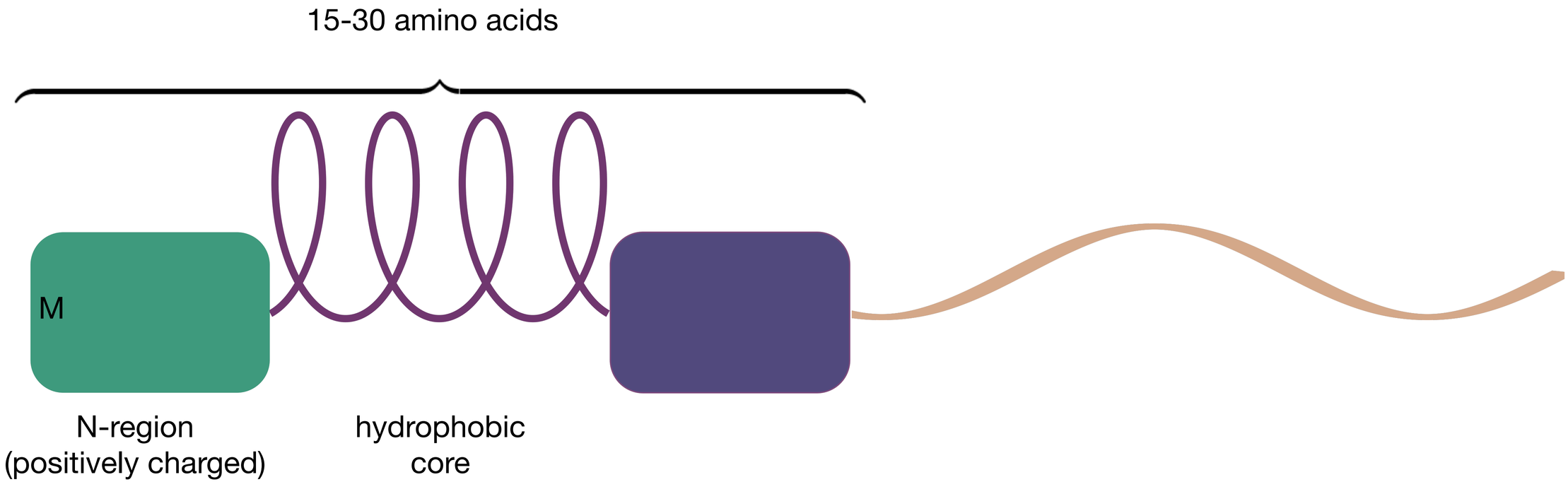
Signal peptides are secretion signals



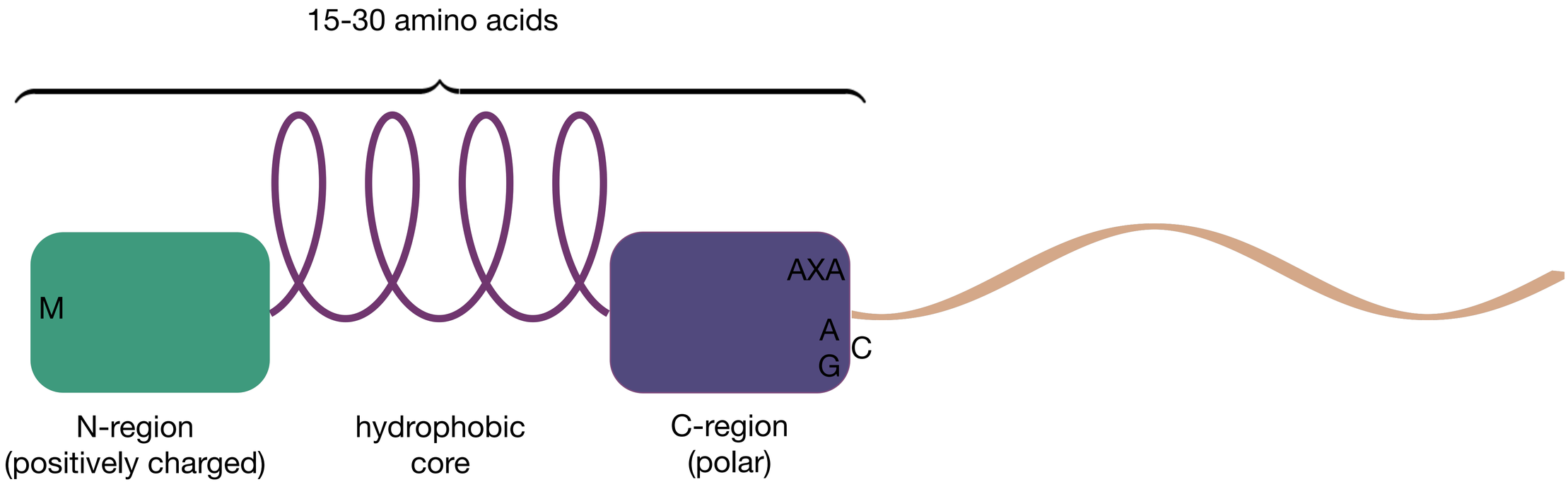
Signal peptides are secretion signals



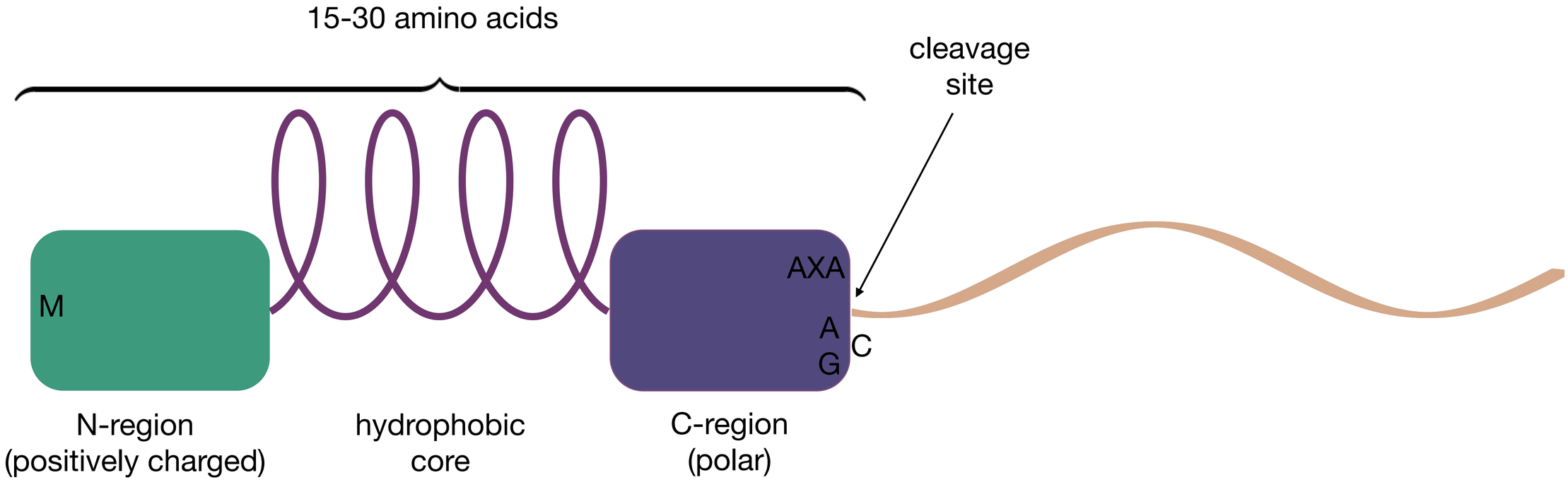
Signal peptides are secretion signals



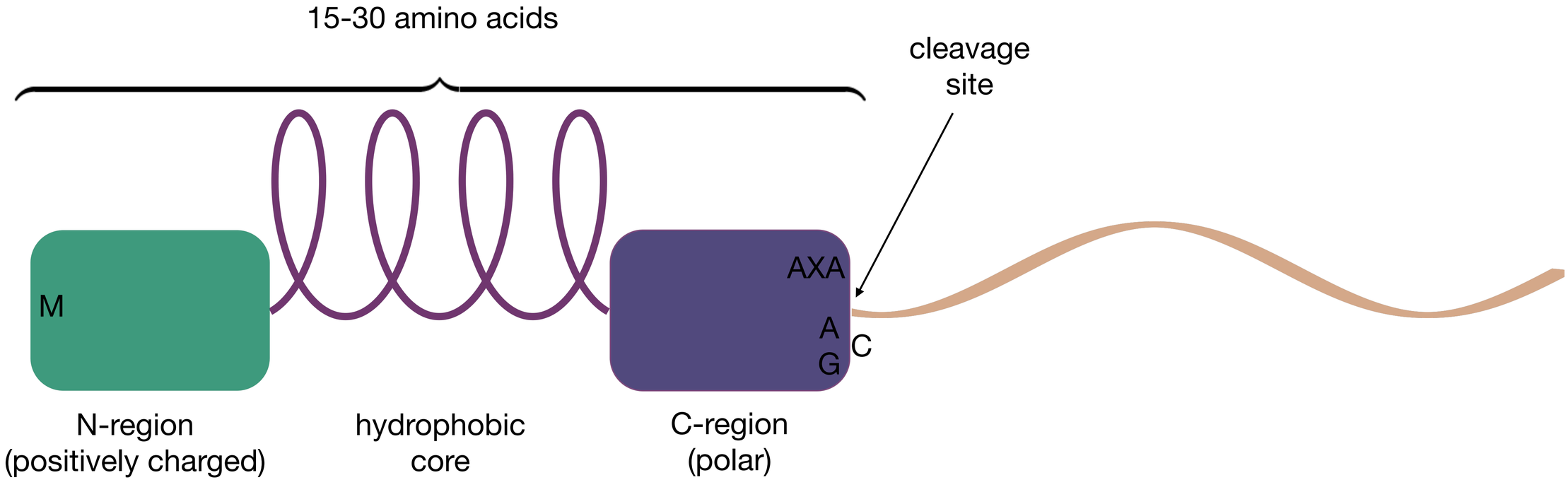
Signal peptides are secretion signals



Signal peptides are secretion signals

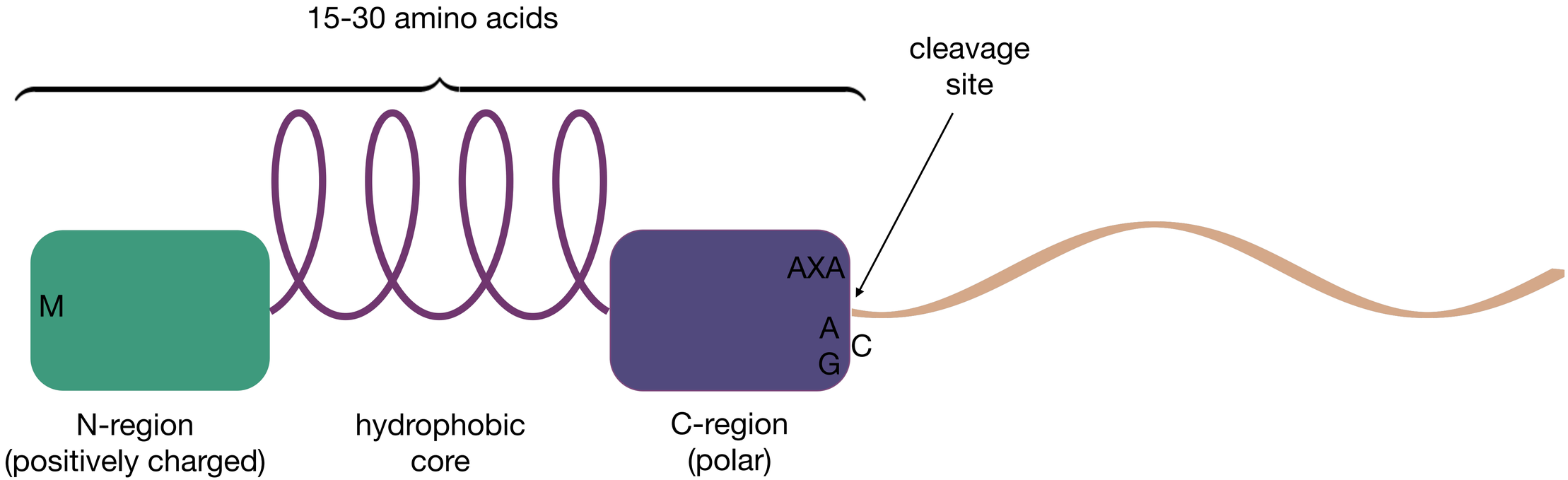


Signal peptides are secretion signals



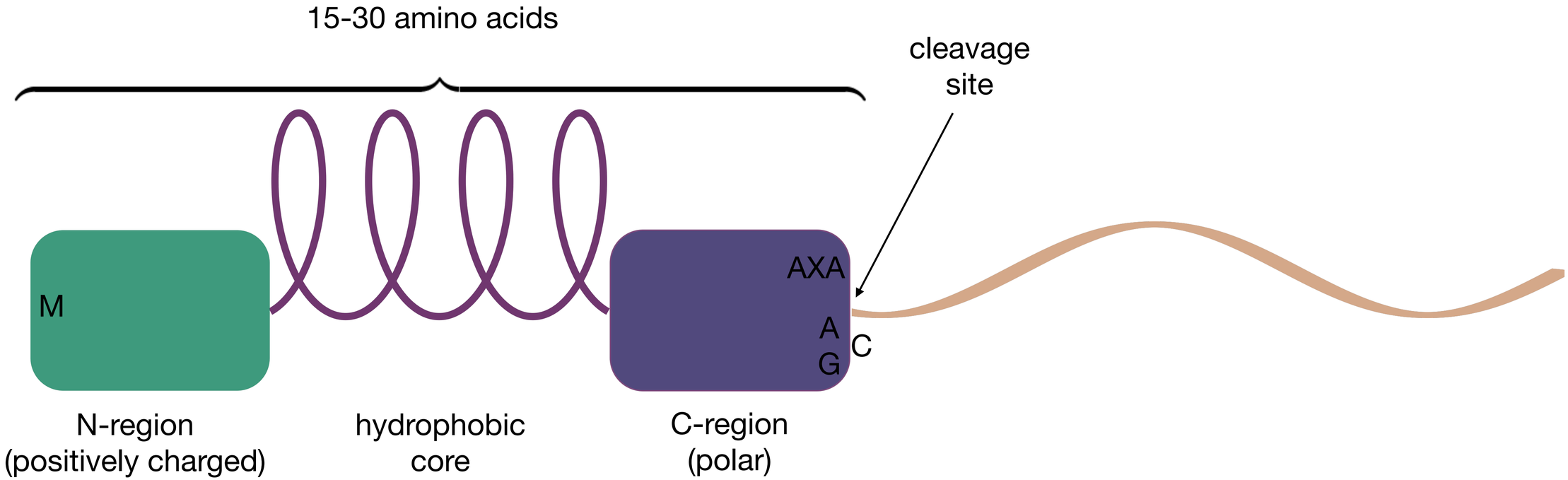
- Extra-cellular secretion in prokaryotes

Signal peptides are secretion signals



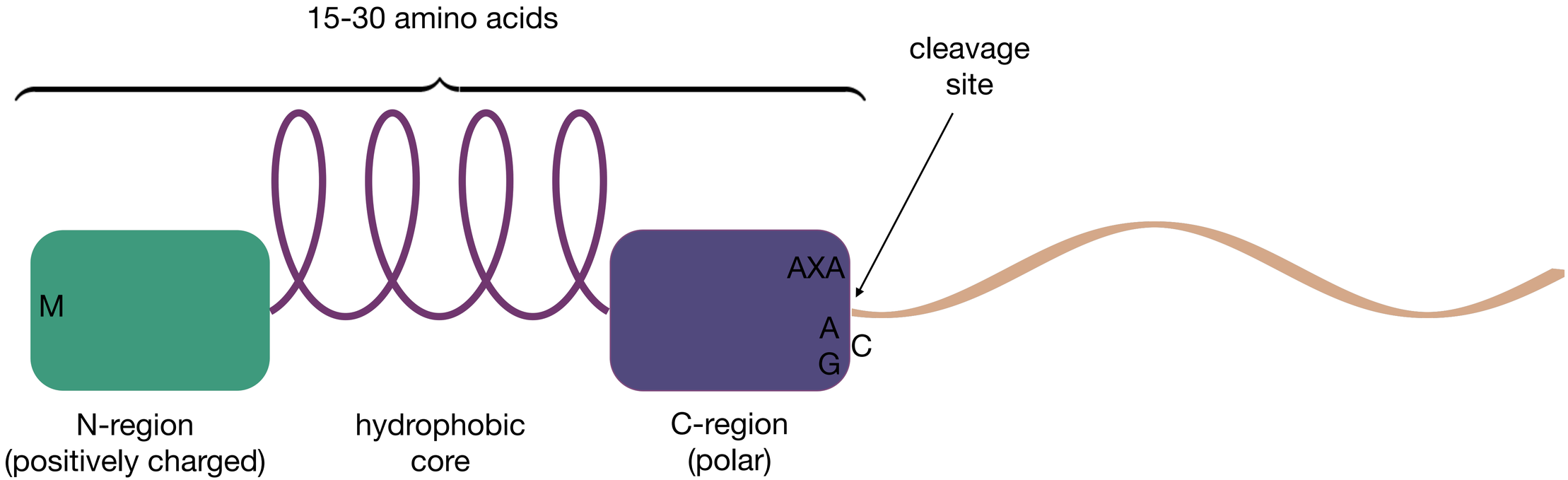
- Extra-cellular secretion in prokaryotes
- Secretion from endoplasmic reticulum in eukaryotes

Signal peptides are secretion signals



- Extra-cellular secretion in prokaryotes
- Secretion from endoplasmic reticulum in eukaryotes
- SP dependent on species and protein

Signal peptides are secretion signals



- Extra-cellular secretion in prokaryotes
- Secretion from endoplasmic reticulum in eukaryotes
- SP dependent on species and protein
- Rules insufficient for generation

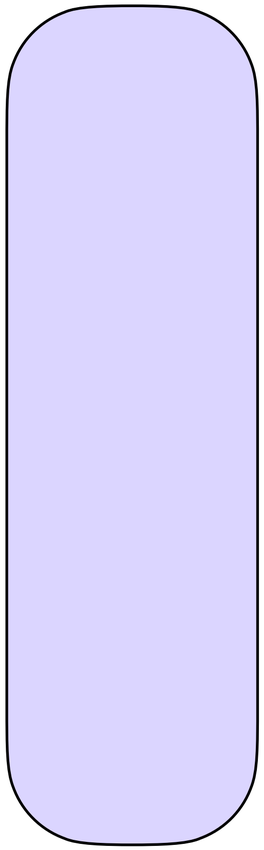
SPs simplify protein production

SPs simplify protein production

- Less burden on cells

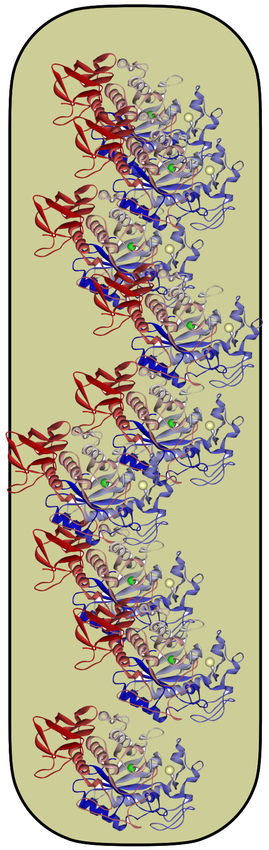
SPs simplify protein production

- Less burden on cells



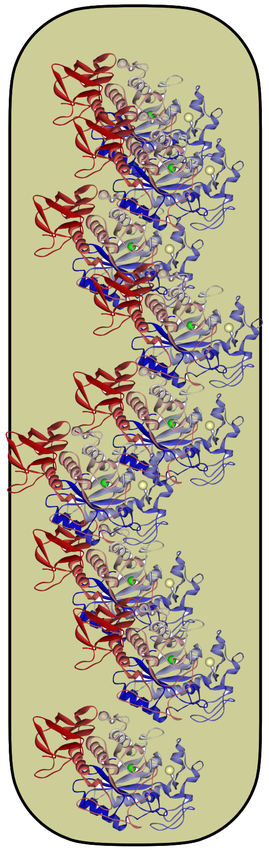
SPs simplify protein production

- Less burden on cells



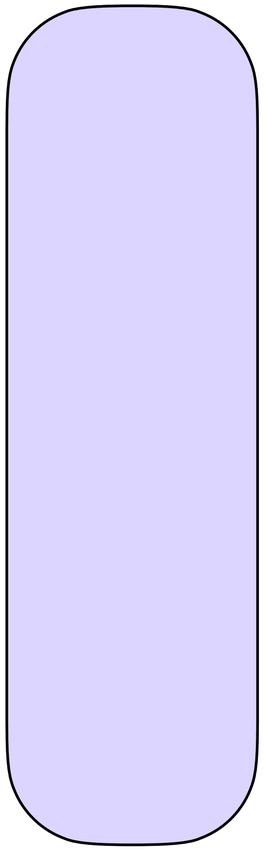
SPs simplify protein production

- Less burden on cells
- Easier separation



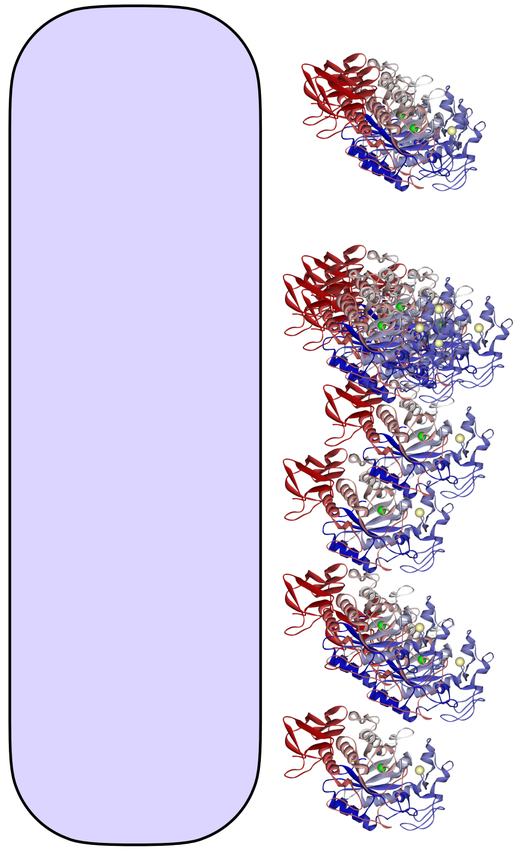
SPs simplify protein production

- Less burden on cells
- Easier separation



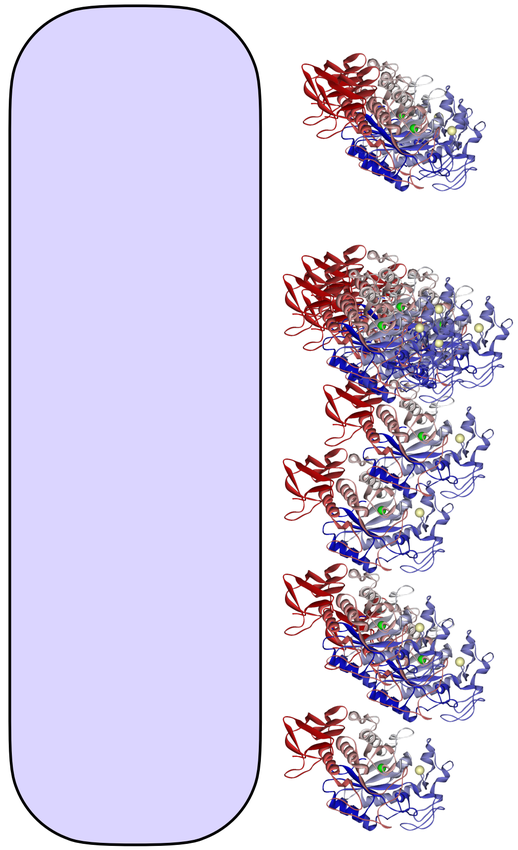
SPs simplify protein production

- Less burden on cells
- Easier separation



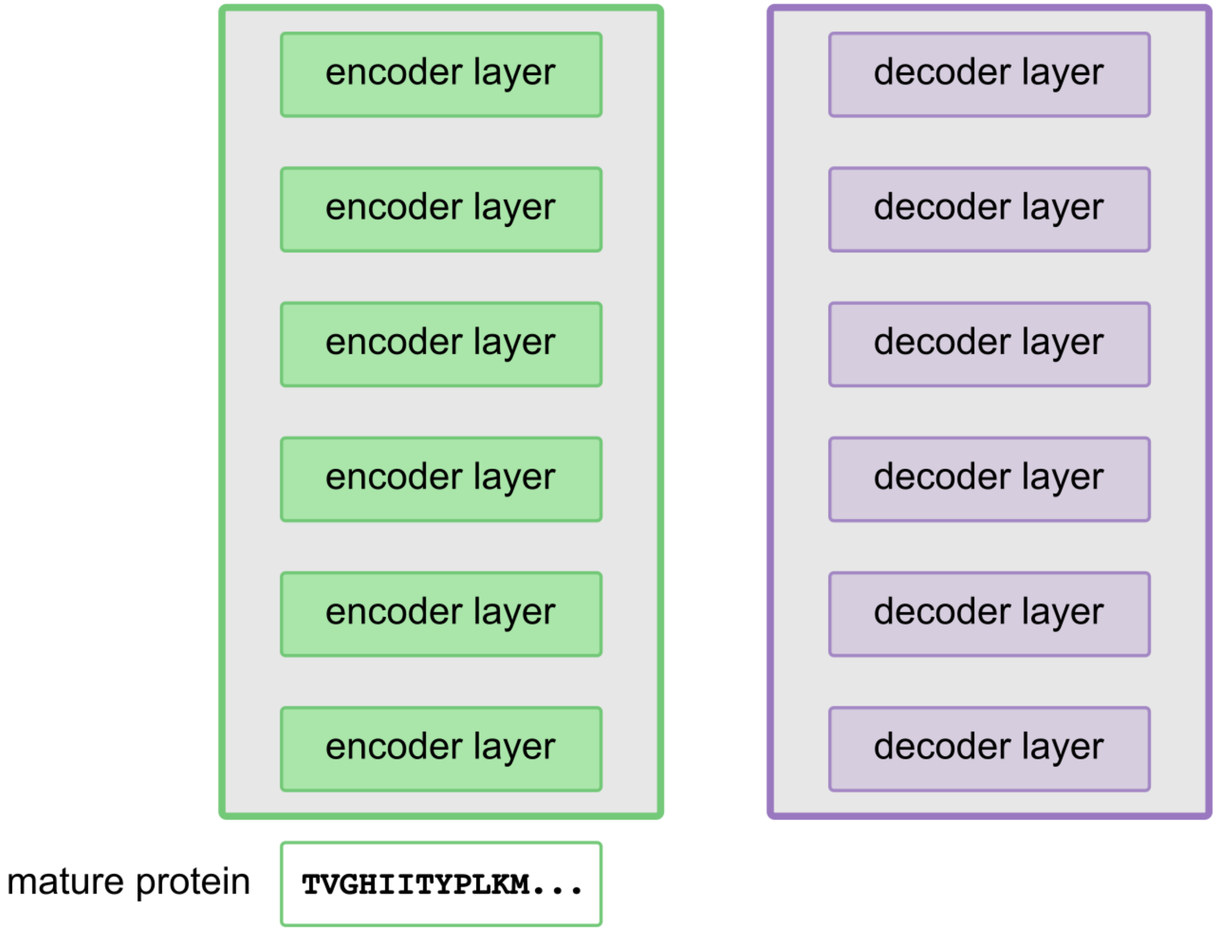
SPs simplify protein production

- Less burden on cells
- Easier separation

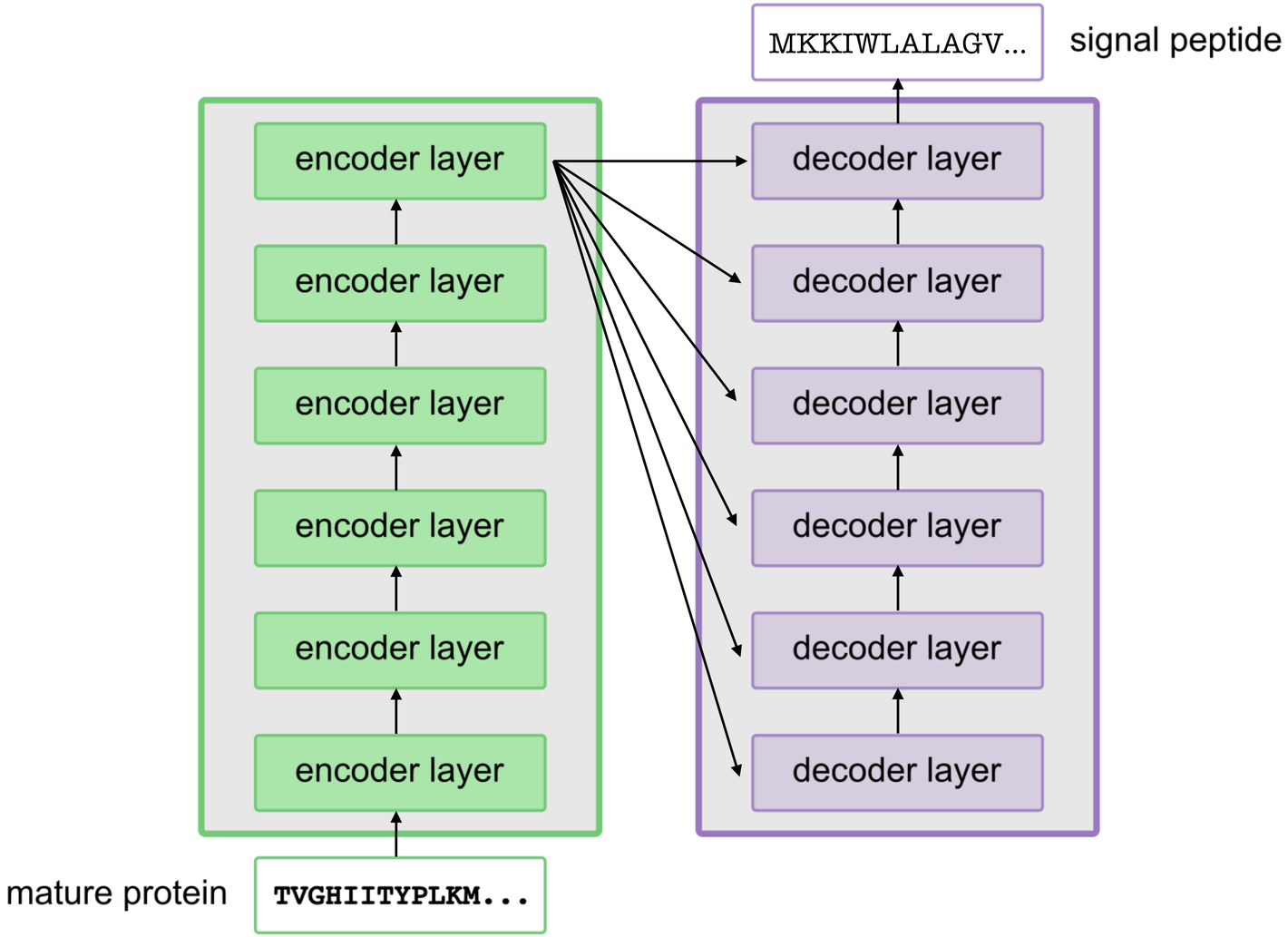


Need: custom SPs for arbitrary proteins

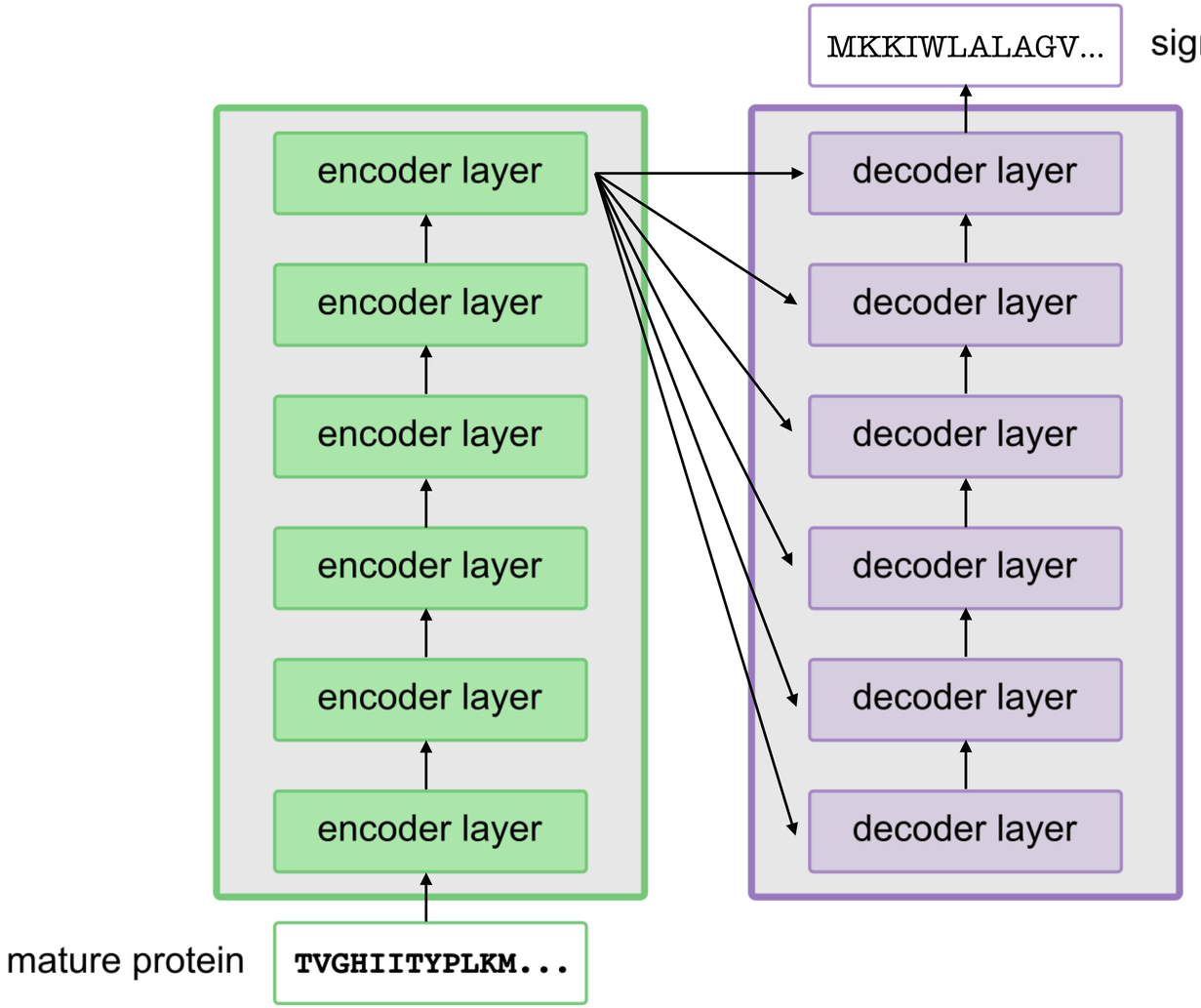
Machine translation for SP generation



Machine translation for SP generation

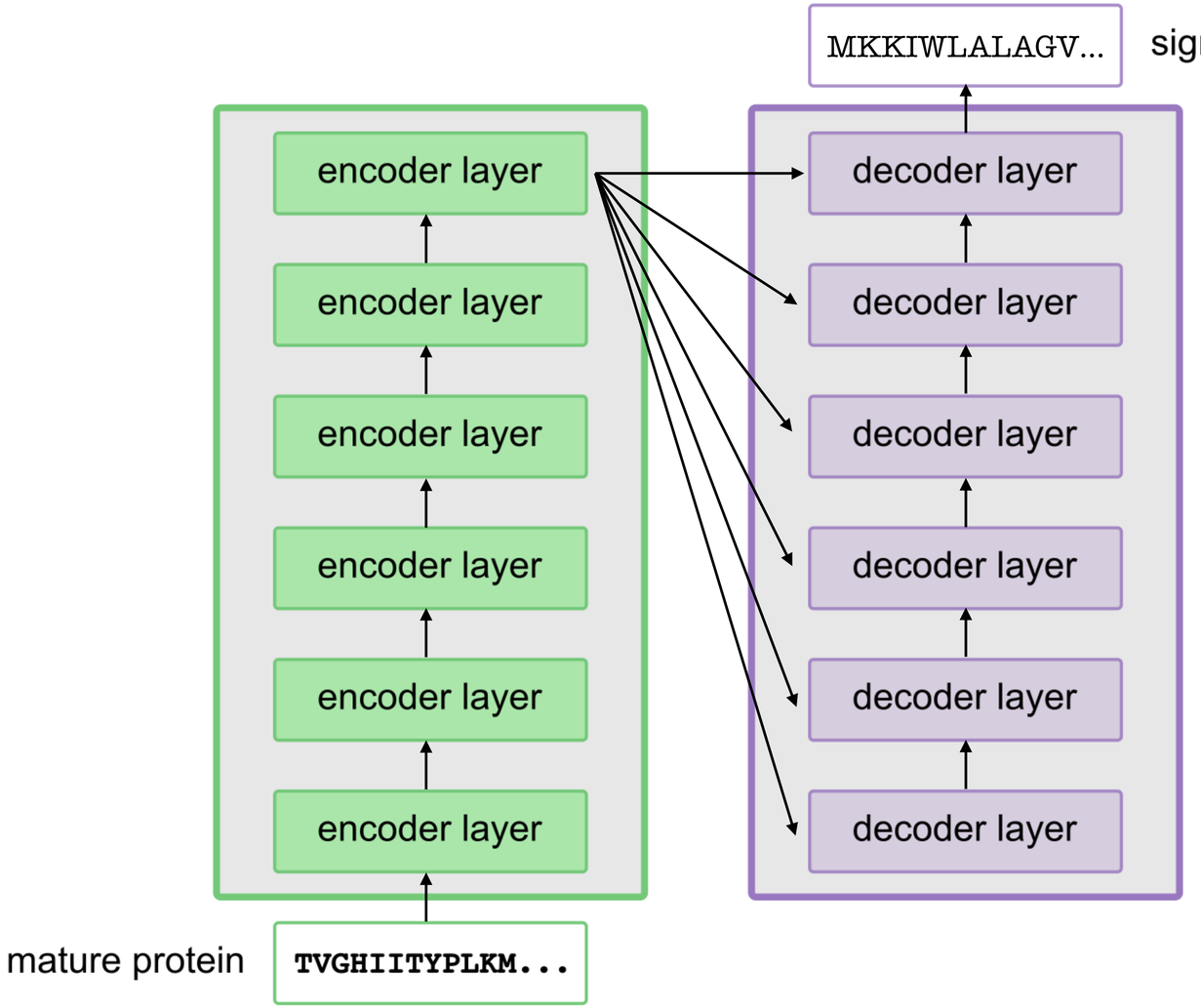


Machine translation for SP generation



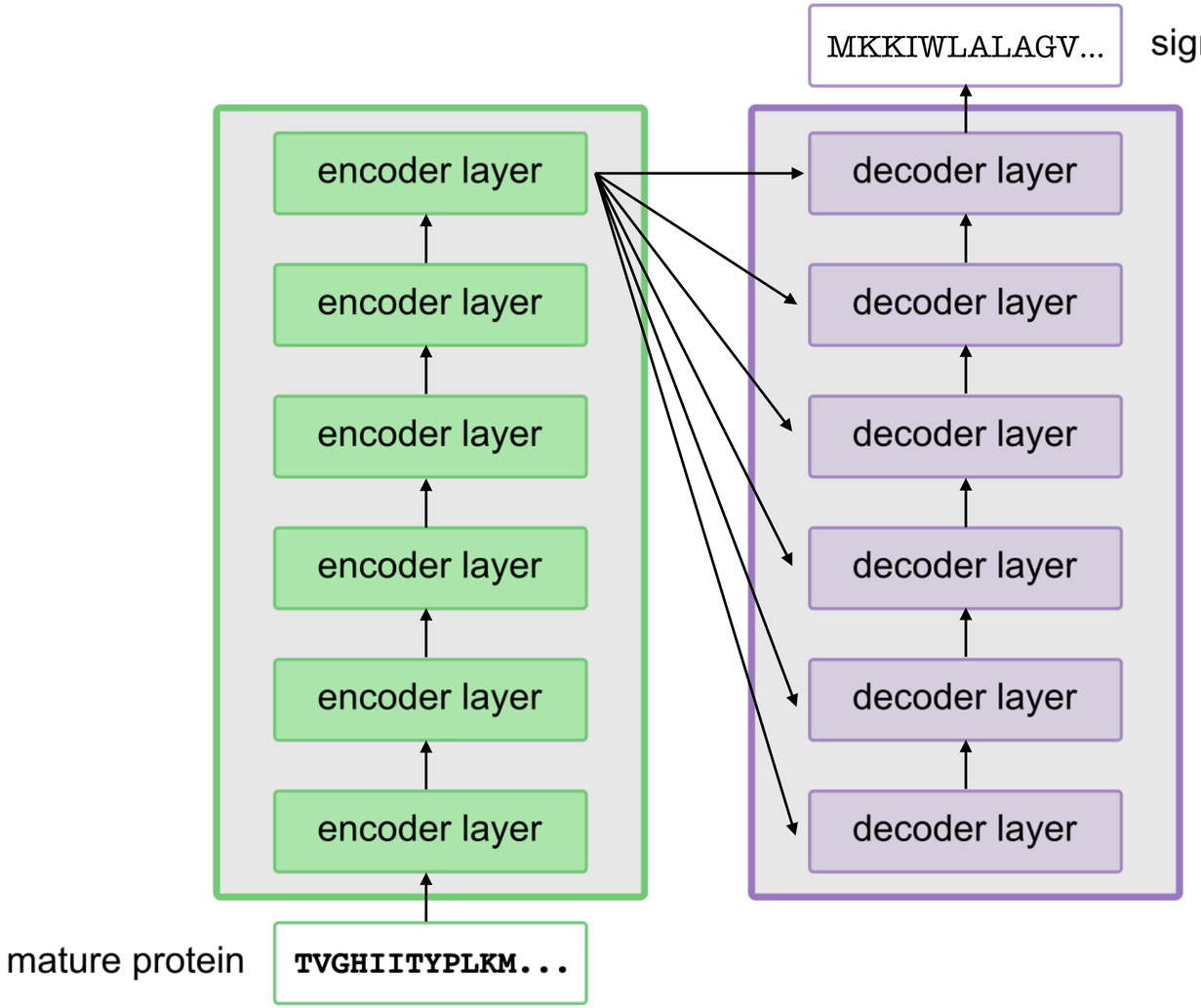
- Transformer architecture

Machine translation for SP generation



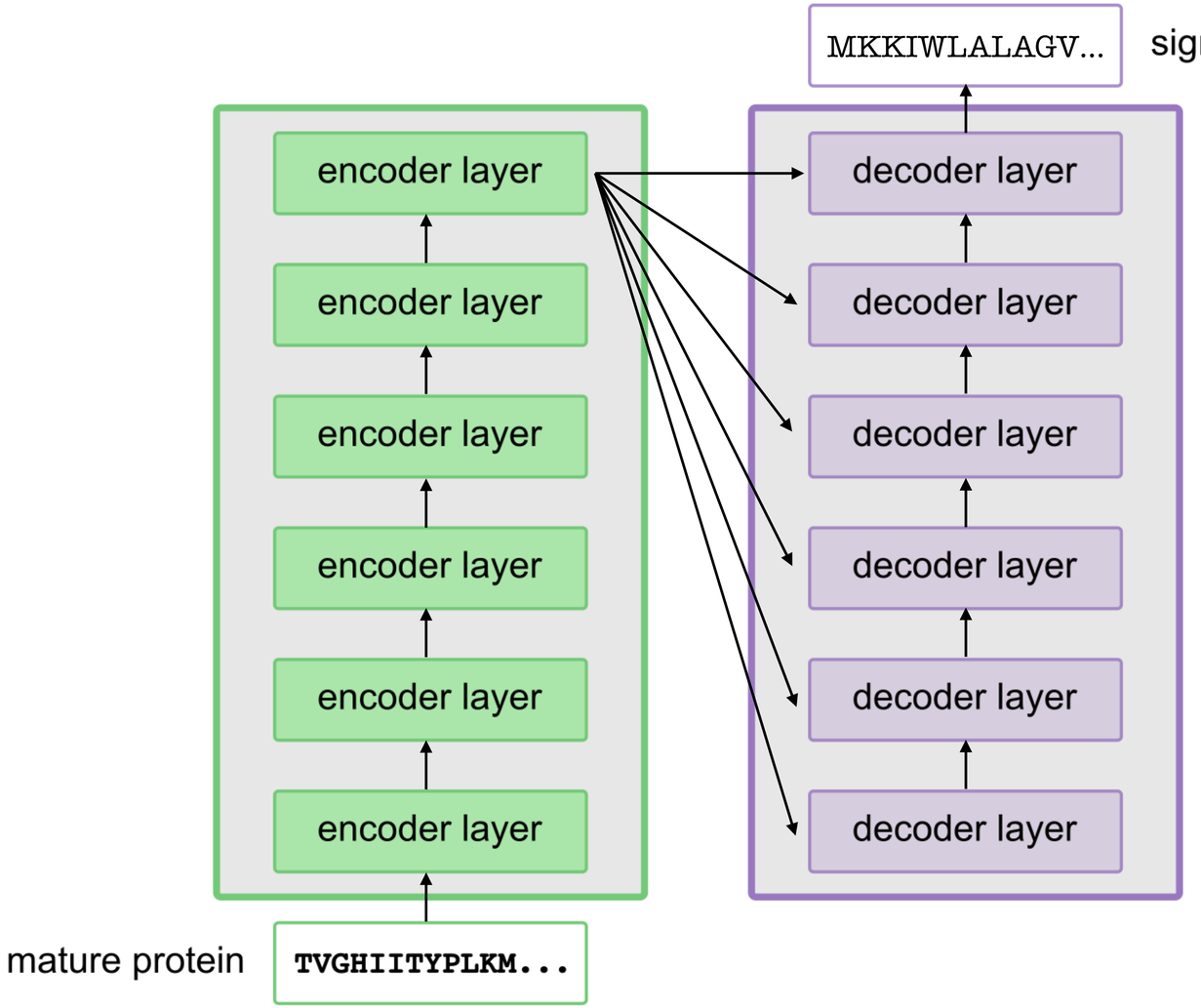
- Transformer architecture
- 40k examples from UniProt

Machine translation for SP generation



- Transformer architecture
- 40k examples from UniProt
- 40% test accuracy

Machine translation for SP generation



- Transformer architecture
- 40k examples from UniProt
- 40% test accuracy

Actual `MNFSKIFIFVVLAVLLLCSQTEA`
Predicted `MNFKLIFLVALVLMAAFLGQTEG`

Actual `MLRKSAVLAGVVLLGASAQA`
Predicted `MRLSTSALVLGAASSAV`

Actual `MASRVECLVVCLLVLRVAQSSG`
Predicted `MNLASVLLLLAACHLSVSVNG`

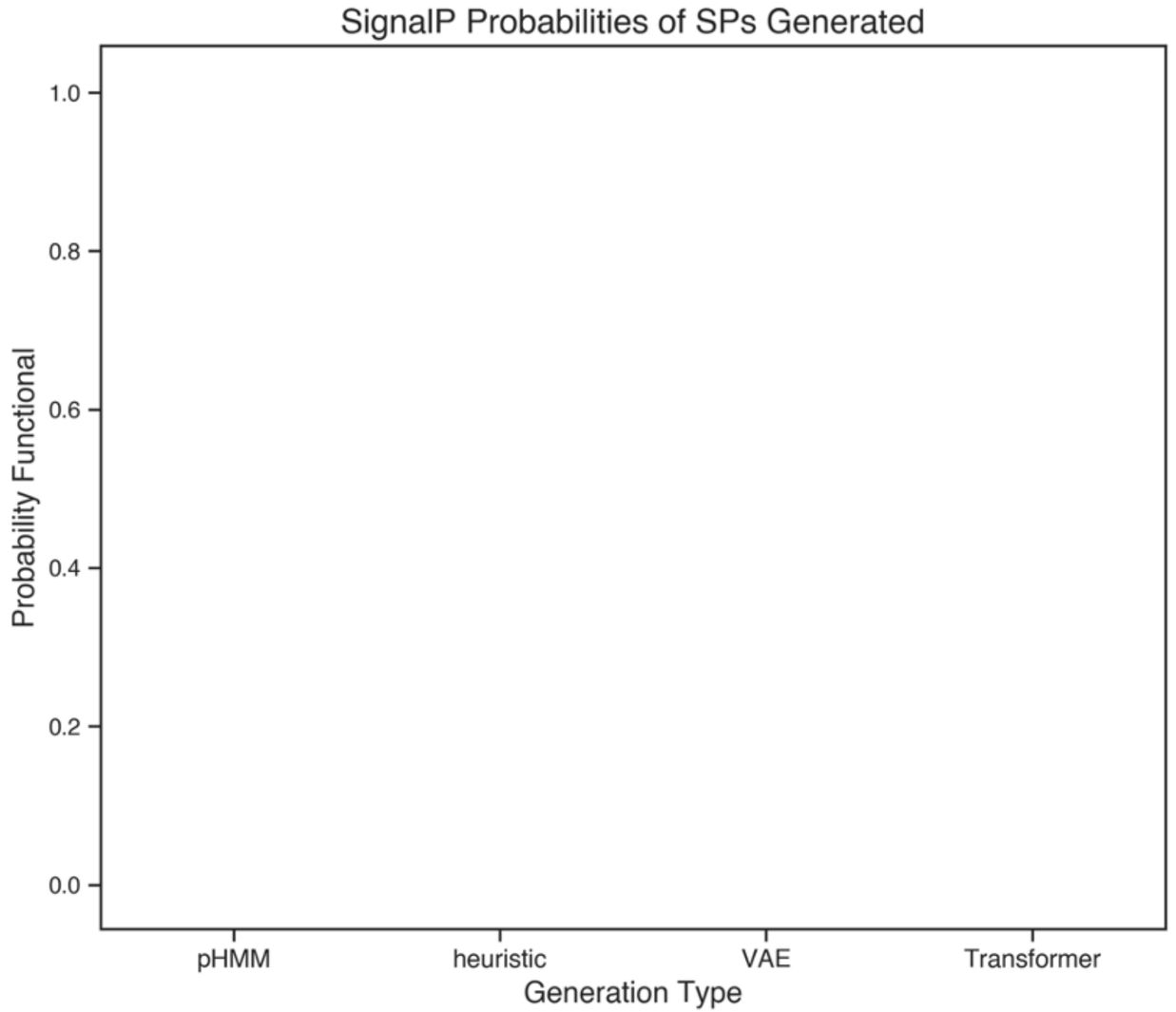
Actual `MLWIRAVLPLVAGSDT`
Predicted `MLGIWTLLPLVLTYVVRLLSKCVNA`

Generated SPs trick SignalP 5.0

Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

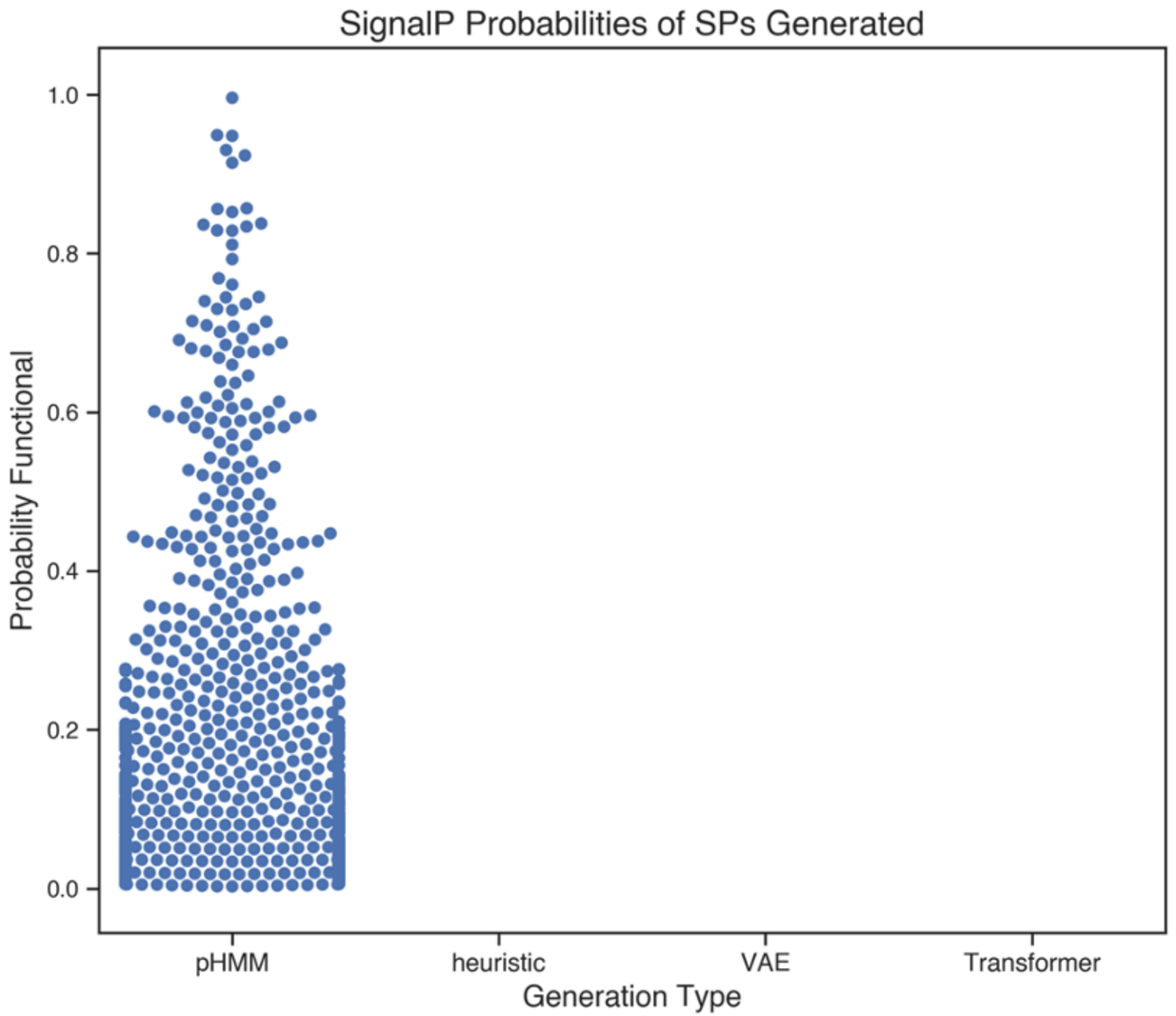
Method	Mean SignalP Probability



Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

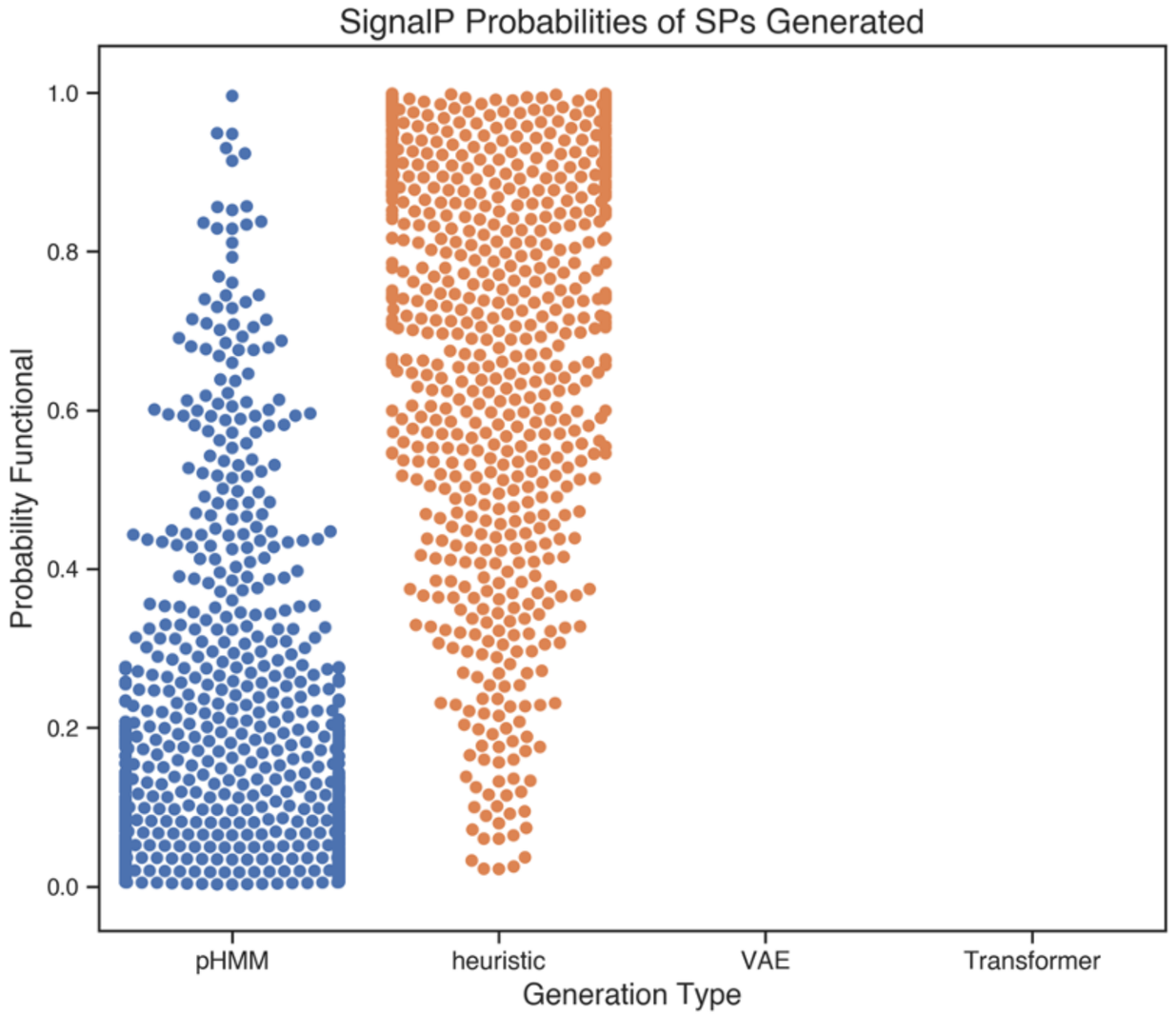
Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%



Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

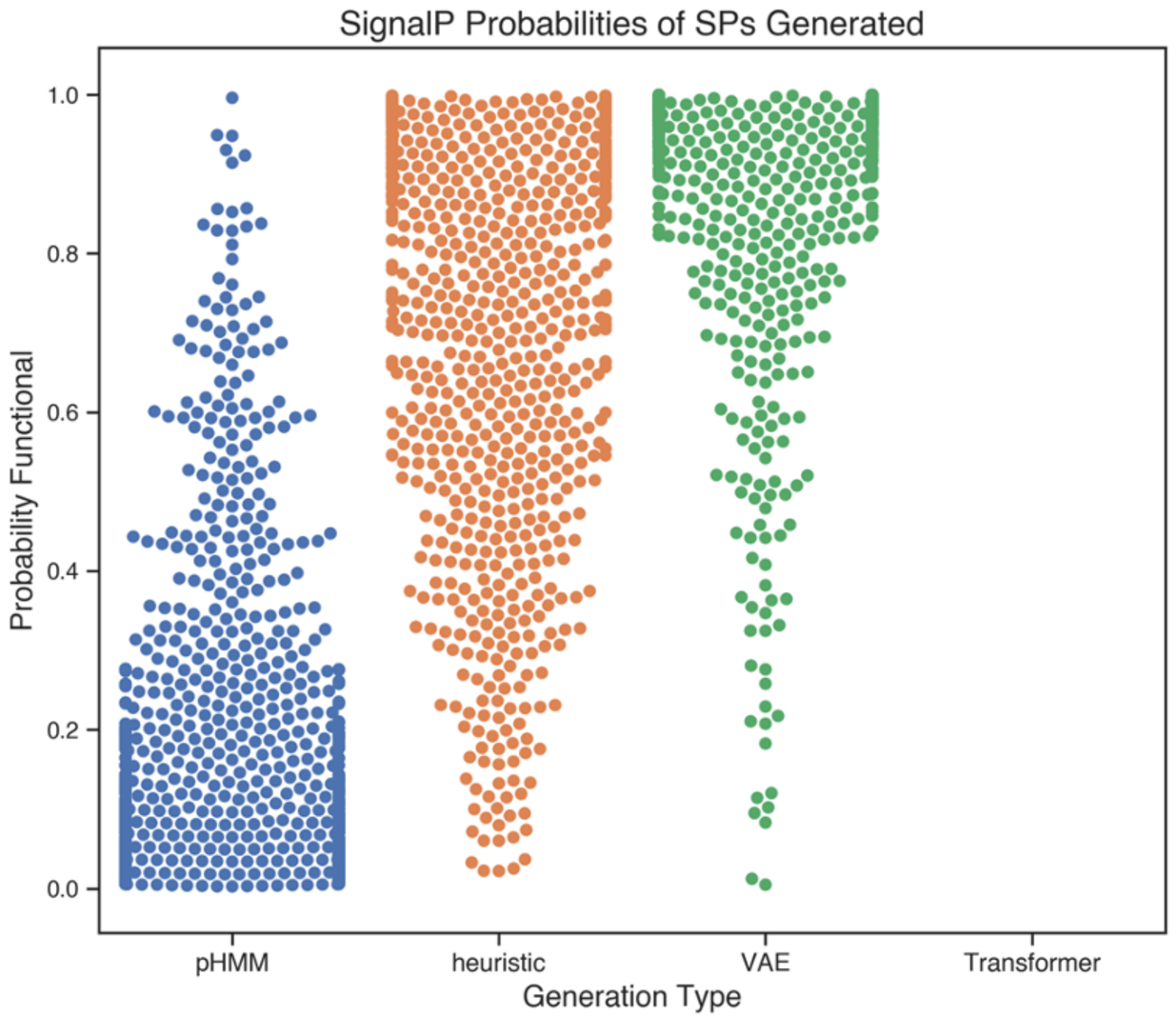
Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%
heuristic	71% ± 25%



Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

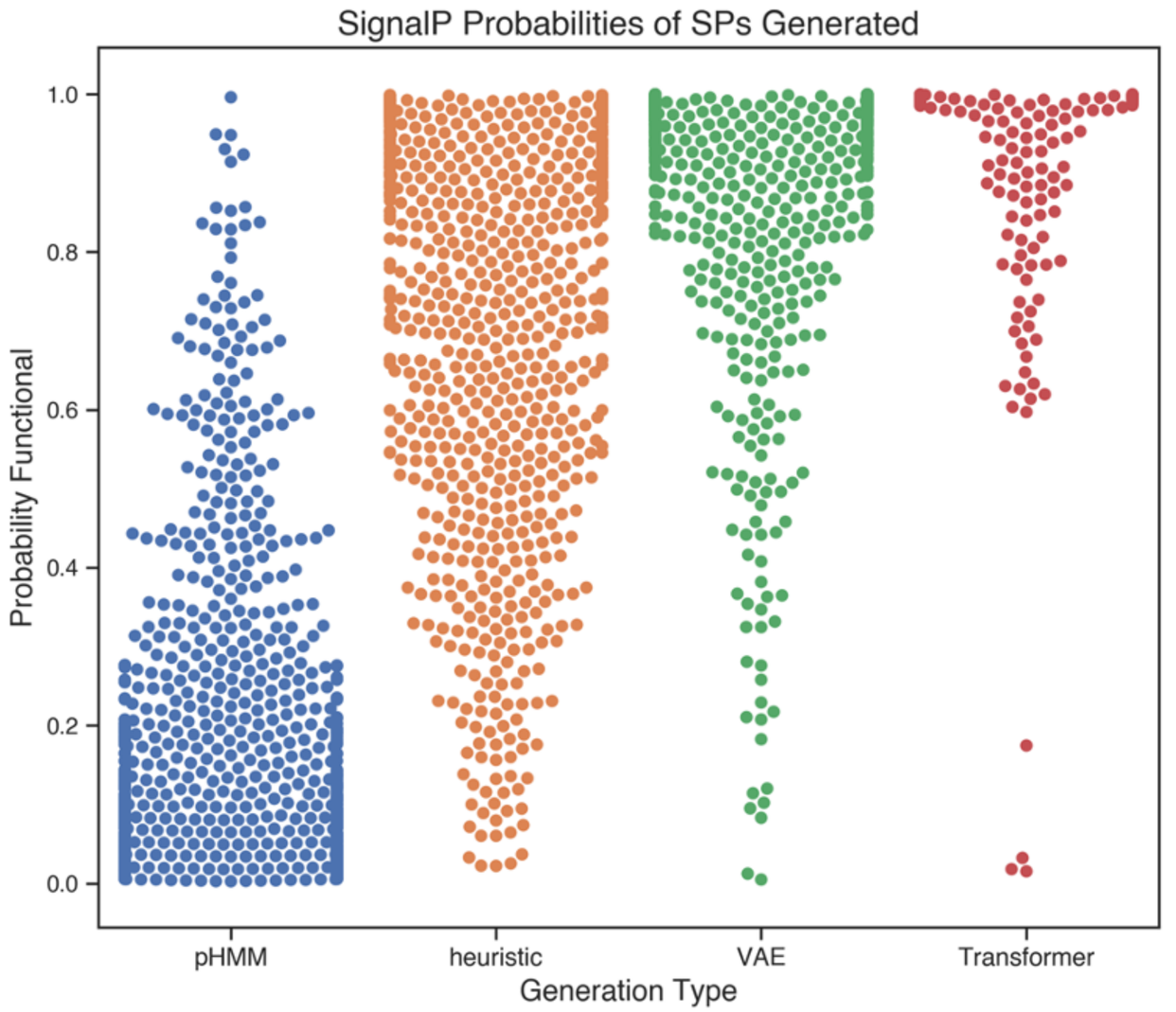
Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%
heuristic	71% ± 25%
Variational Autoencoder (VAE)	**92% ± 15%**



Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%
heuristic	71% ± 25%
Variational Autoencoder (VAE)	**92% ± 15%**
Transformer	90% ± 17%



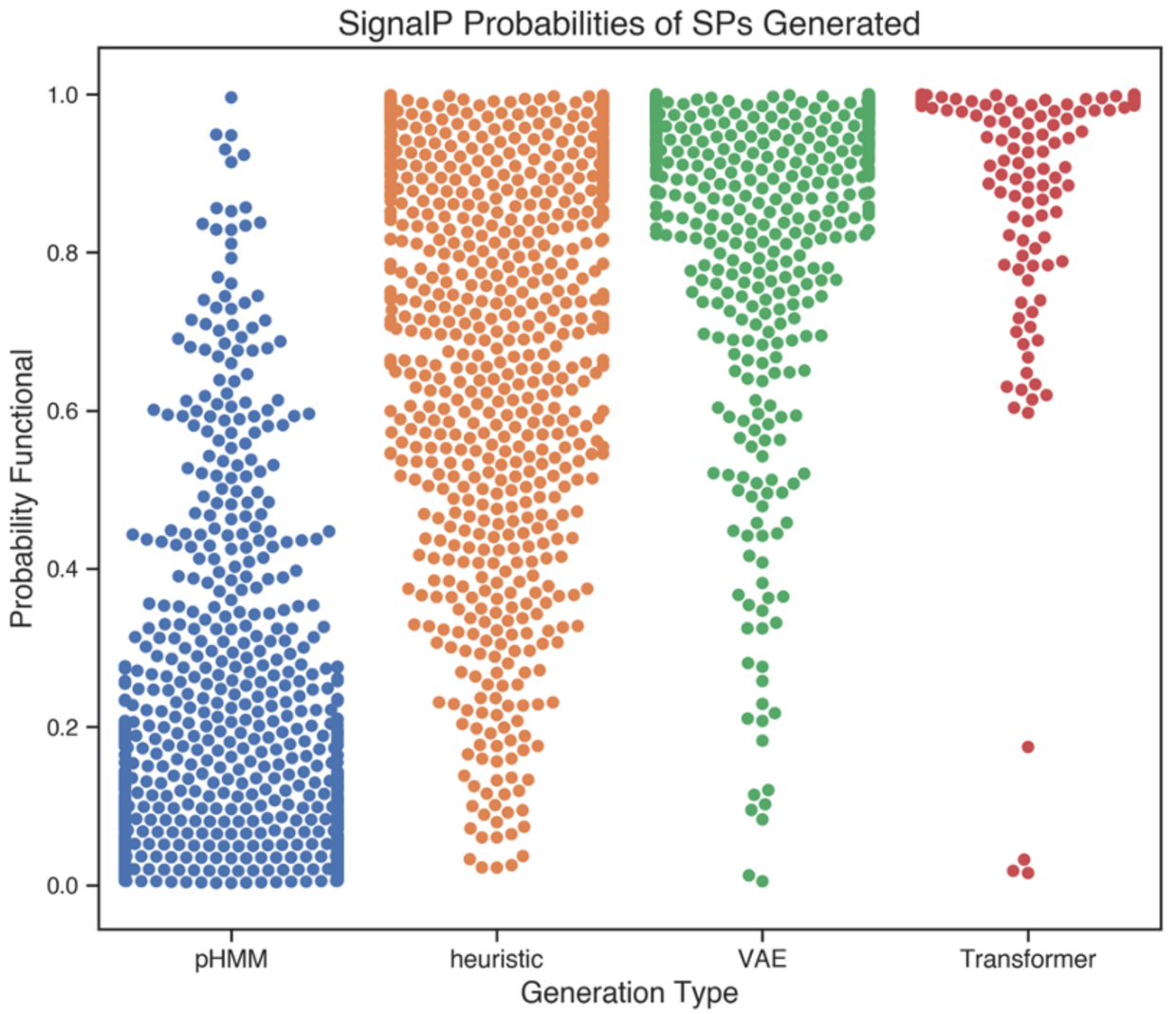
Nielsen *et al.,* 2019

Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%
heuristic	71% ± 25%
Variational Autoencoder (VAE)	**92% ± 15%**
Transformer	90% ± 17%

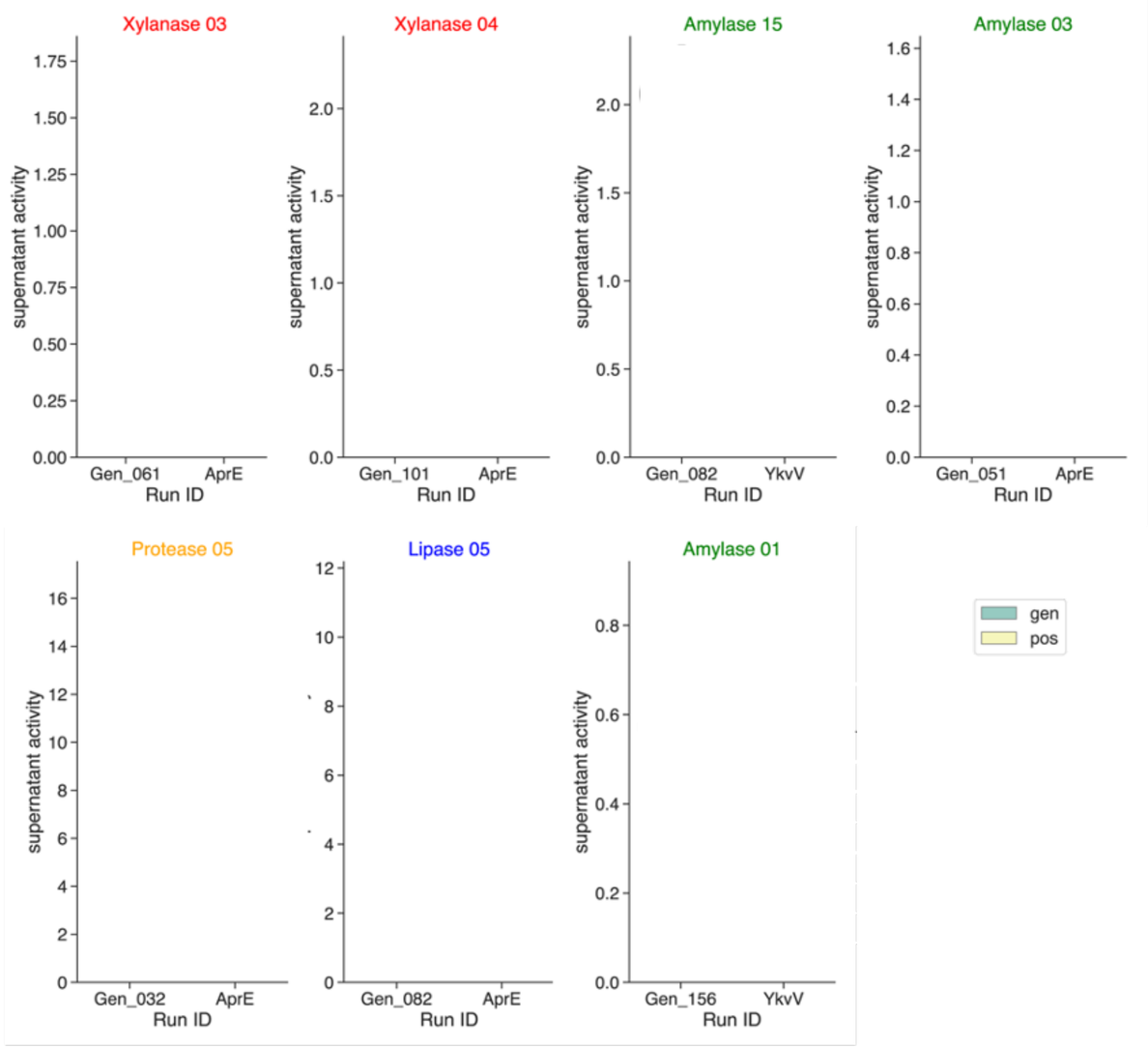
VAE sequences are repetitive

```
MRKRLLALALAALSLLLLLSFGVKALAGSGA
MRLLLLLLLVLLLAAPAPPGLS
MKLLLLLVTLLTSVLALQA
MRLLLLALLAAAAVALASA
MASSSSSLFVVLAVLLLLLLLTLSSA
```



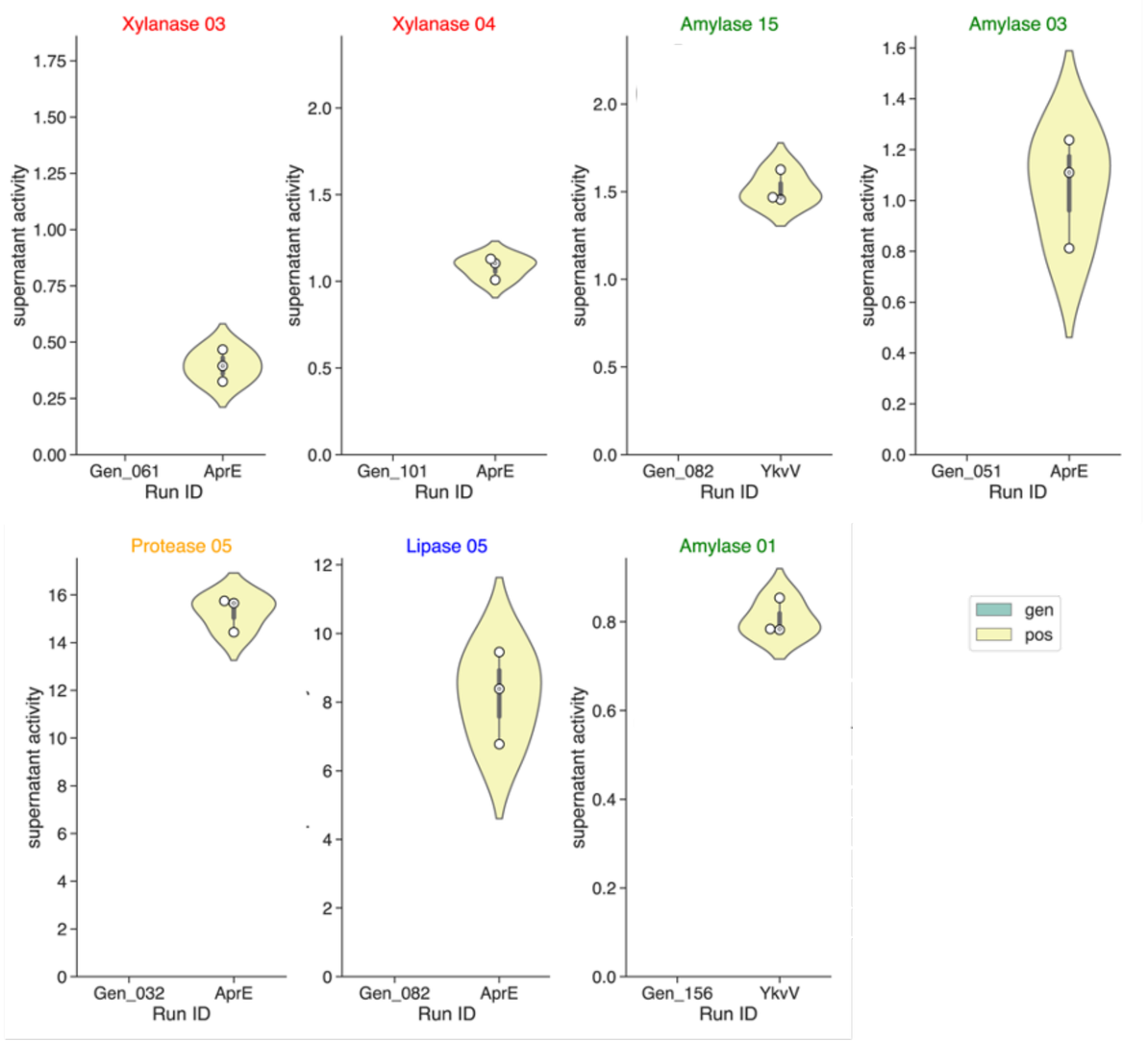
Nielsen *et al.,* 2019

Generated SPs are functional in *Bacillus subtilis*



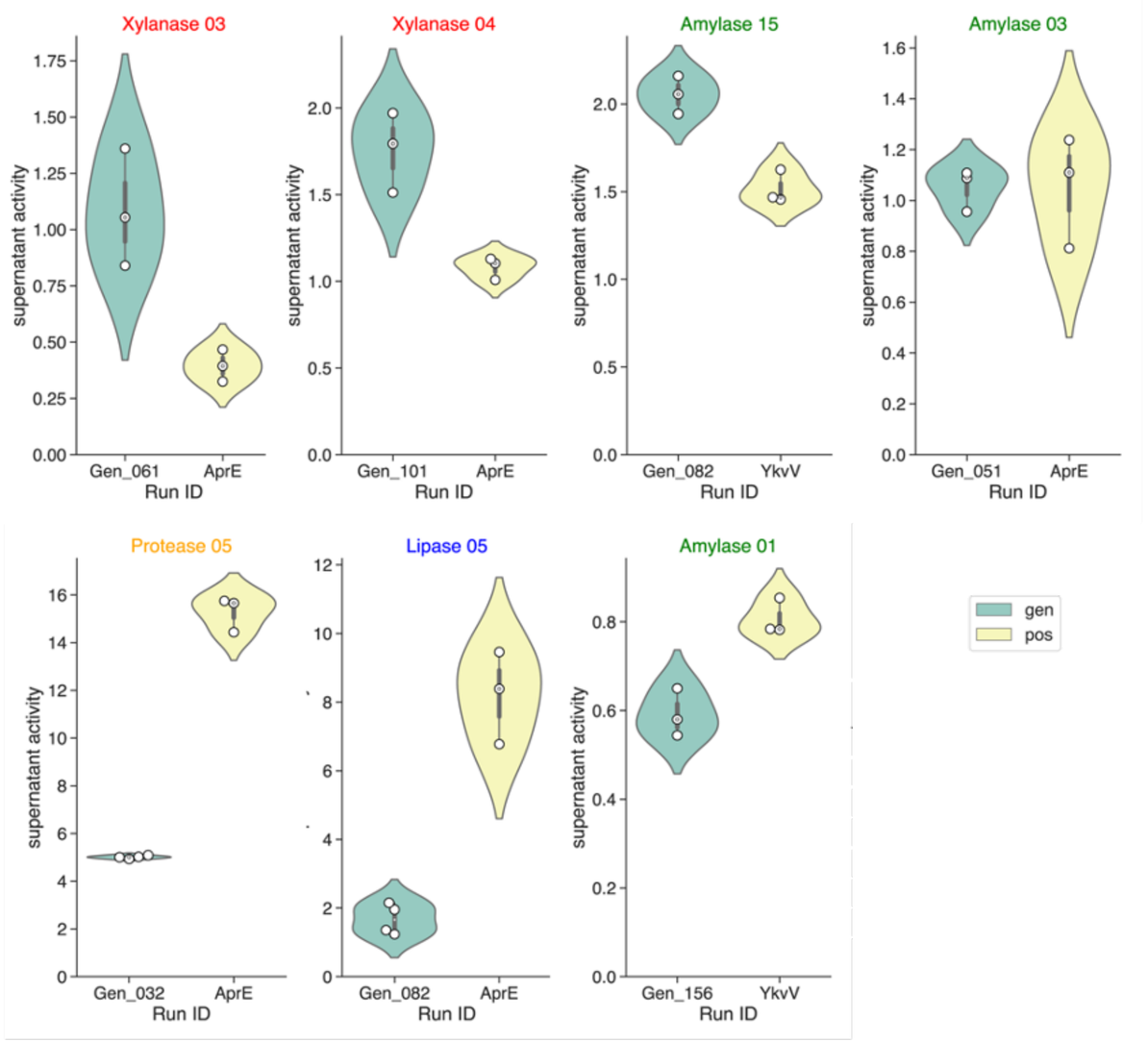
Generated SPs are functional in *Bacillus subtilis*

- 79% of natural SPs are functional



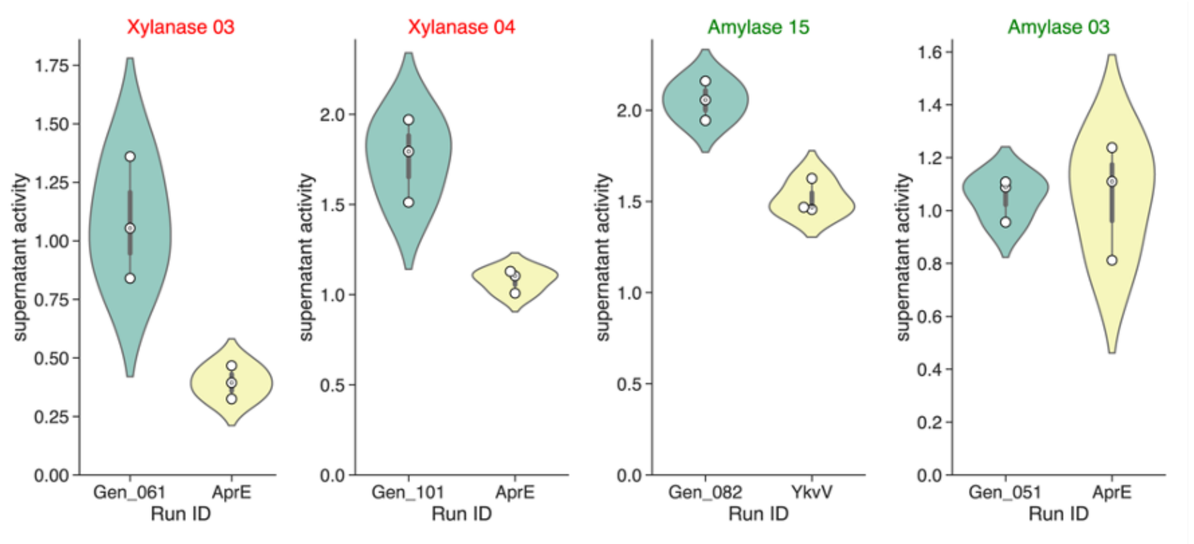
Generated SPs are functional in *Bacillus subtilis*

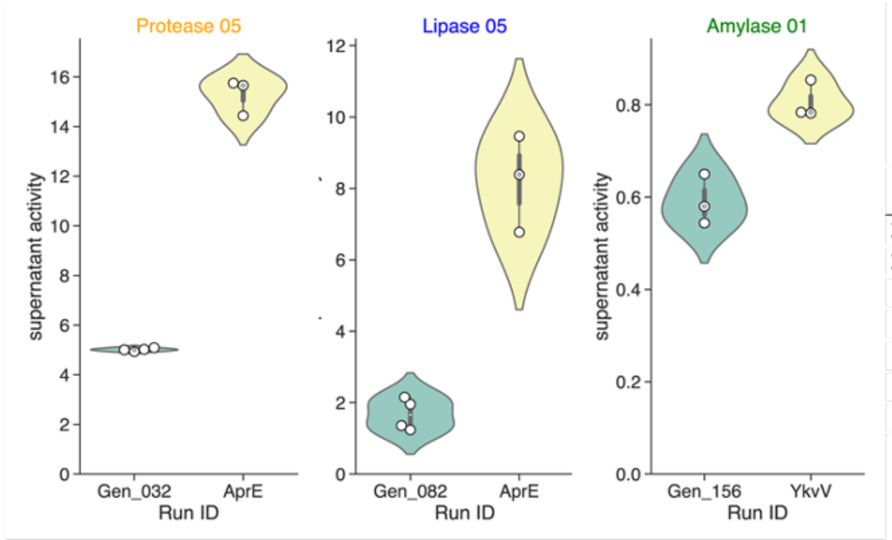
- 79% of natural SPs are functional
- 48% of generated SPs are functional



Generated SPs are functional in *Bacillus subtilis*

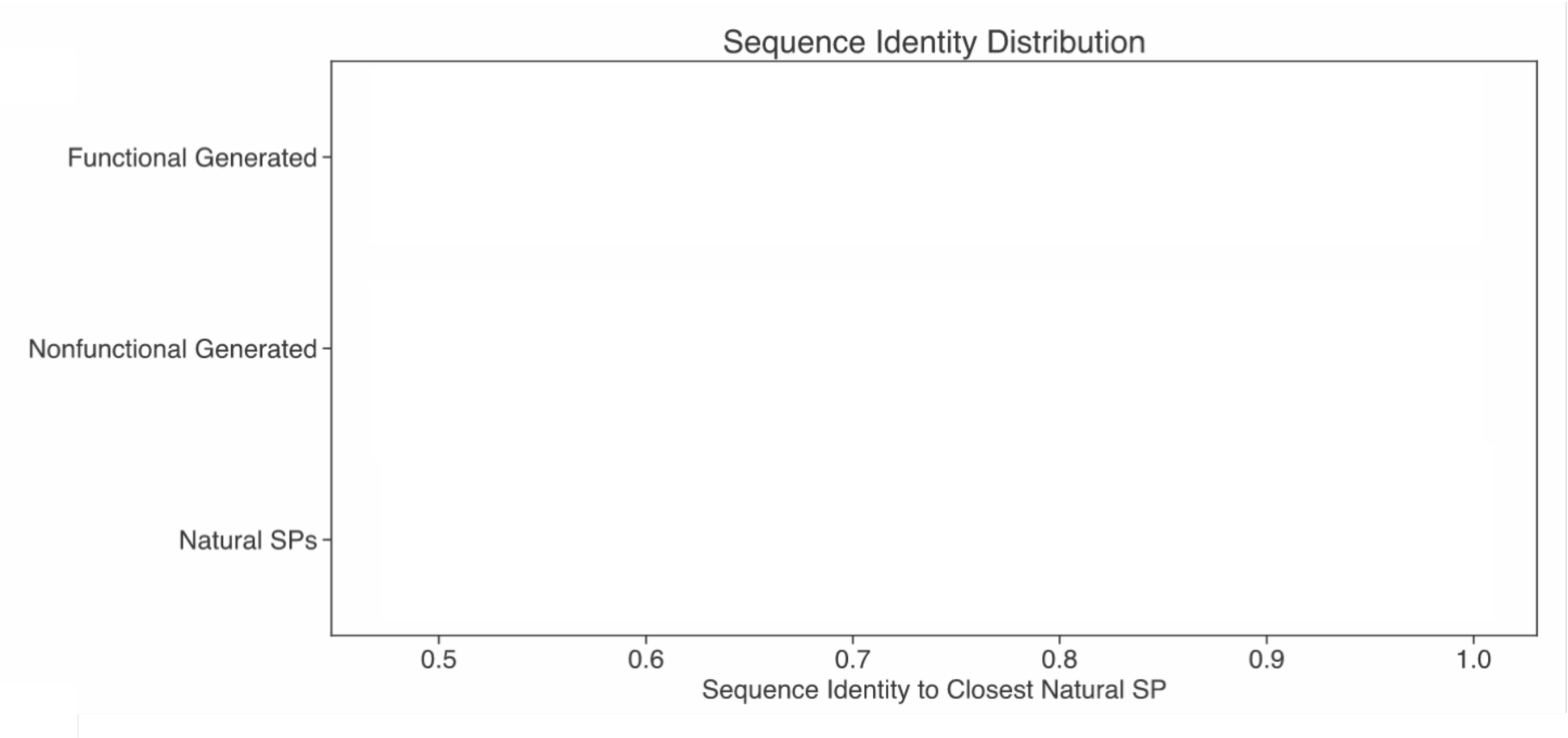
- 79% of natural SPs are functional
- 48% of generated SPs are functional
- Enzymatic activity is comparable



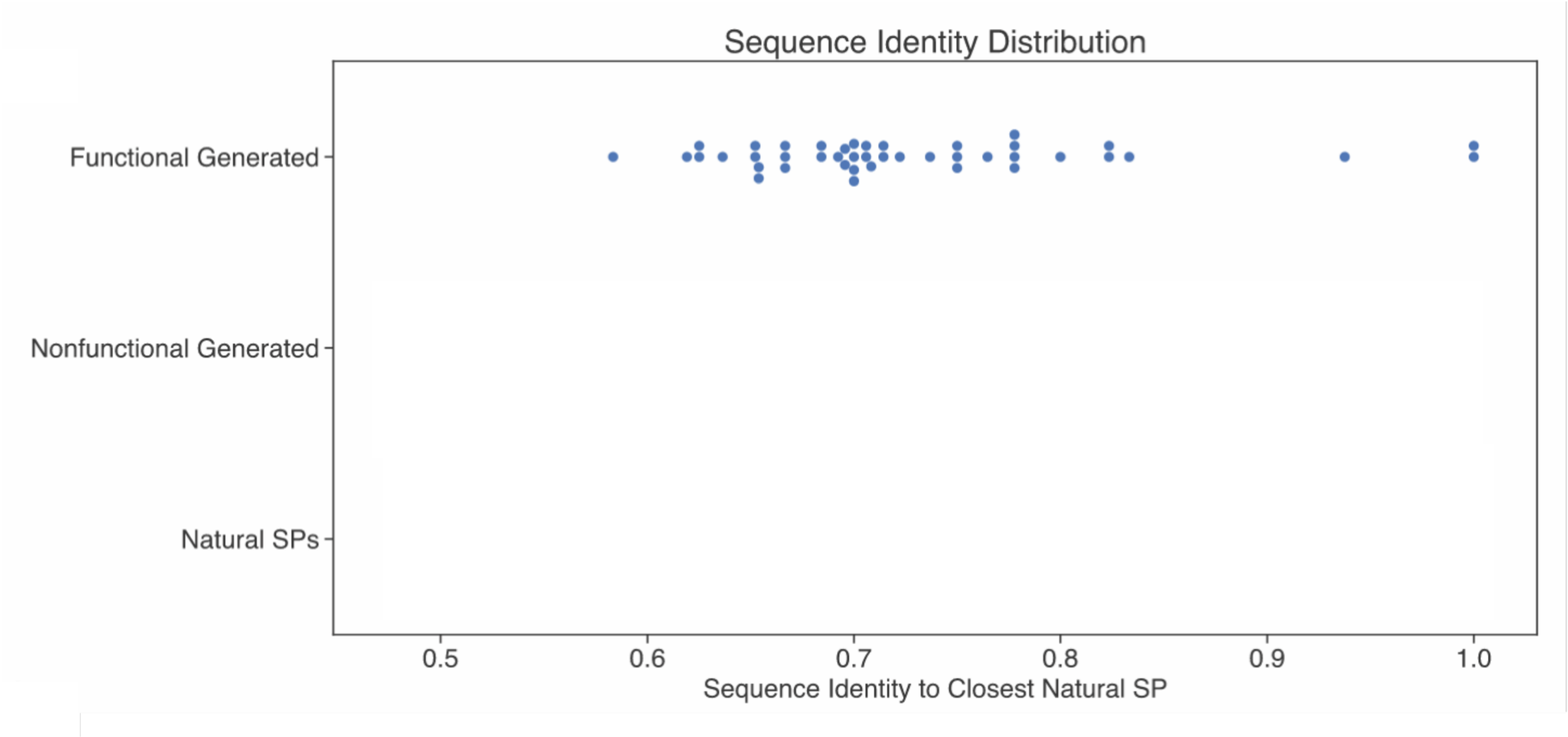


enzyme	p-value	# gen	# pos
Xylanase 03	0.0370	3	3
Xylanase 04	0.0293	3	3
Amylase 15	0.0031	3	3
Amylase 03	0.9857	3	3
Protease 05	0.0016	4	3
Lipase 05	0.0095	4	3
Amylase 01	0.0065	3	3

Generated functional SPs are novel and diverse

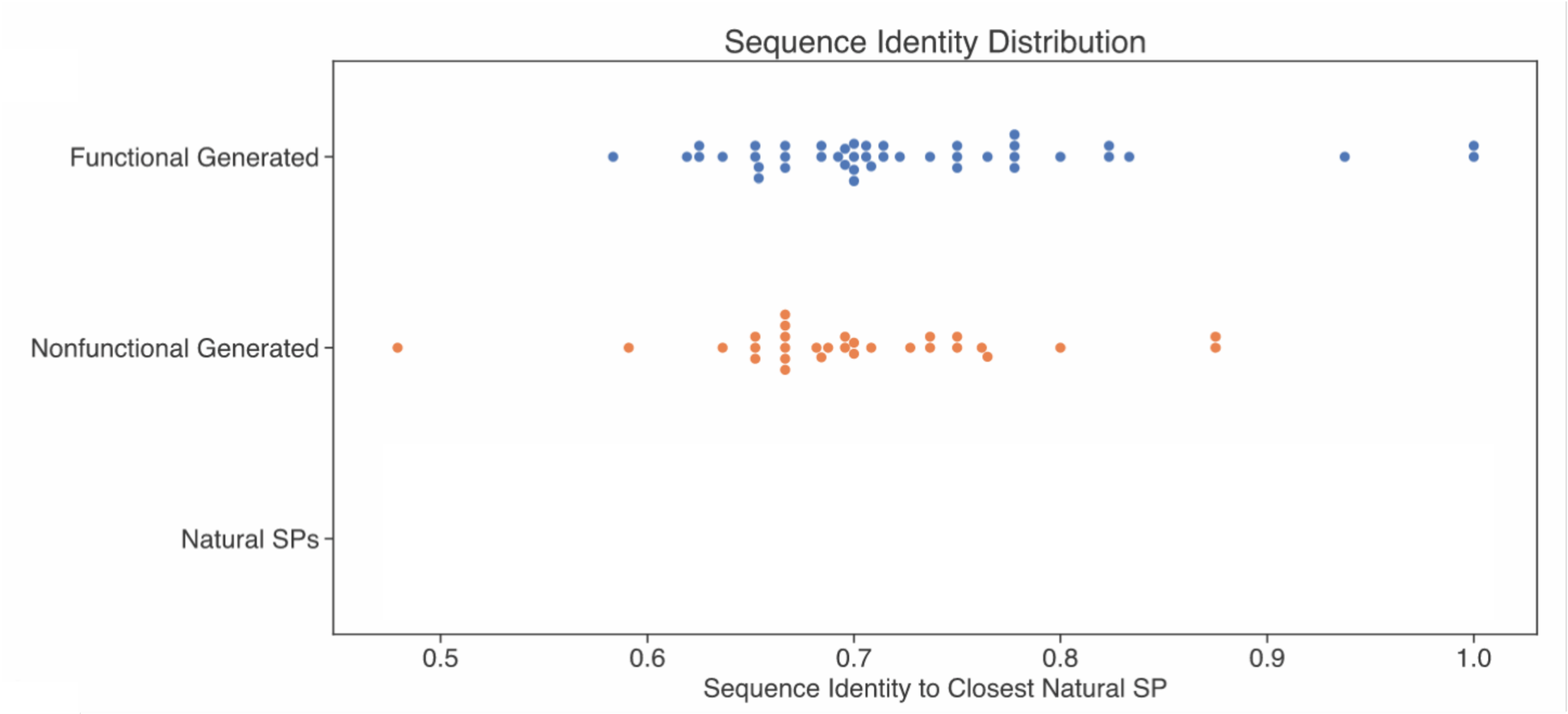


Generated functional SPs are novel and diverse



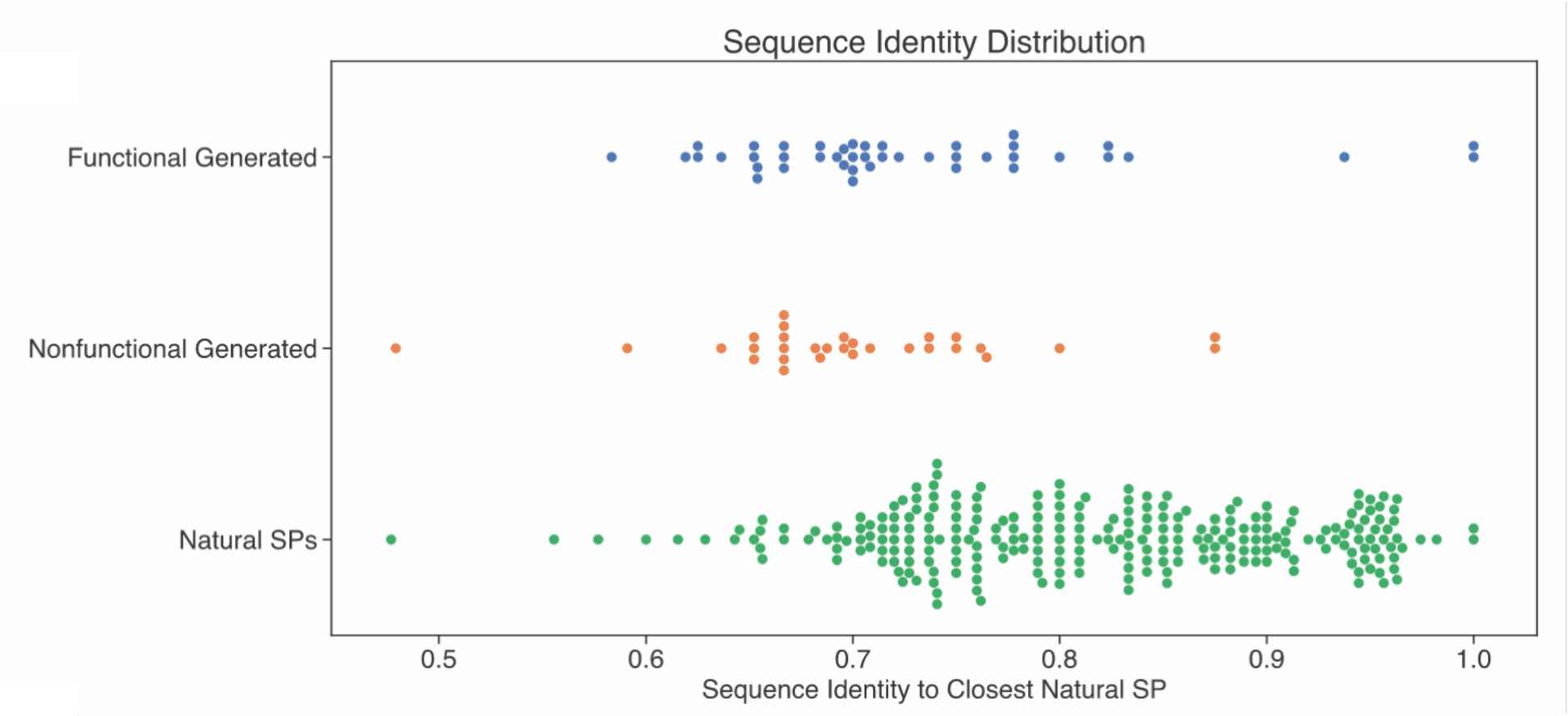
0.75 ± 0.09

Generated functional SPs are novel and diverse



0.75 ± 0.09

Generated functional SPs are novel and diverse



0.75 ± 0.09

0.82 ± 0.1

SP generation conclusions

- Can generate *de novo* signal peptides
- That look like real SPs
- And result in functional secreted enzymes

Acknowledgments

- Frances Arnold
- Yisong Yue
- Zachary Wu, Alycia Lee, Michael J. Liszka
- Claire Bedbrook, Austin Rice

Donna and Benjamin M. Rosen Bioengineering Center

NIH National Institutes of Health

BASF We create chemistry

aws

TWIST BIOSCIENCE

Questions

- yang.kevin@microsoft.com
- @KevinKaichuang
- yangkky.github.io