Reasoning about biology with data-driven approaches

Howes Scholar Presentation

Adam Riesselman

Reasoning about biology with data-driven approaches

Howes Scholar Presentation

Adam Riesselman

DOE CSGF Fellow: Harvard University, 2014-2018

Practicum: Lawrence Berkeley National Lab
(Joint Genome Institute), 2016

Understanding how **genotype** affects **phenotype** is a critical **challenge**.

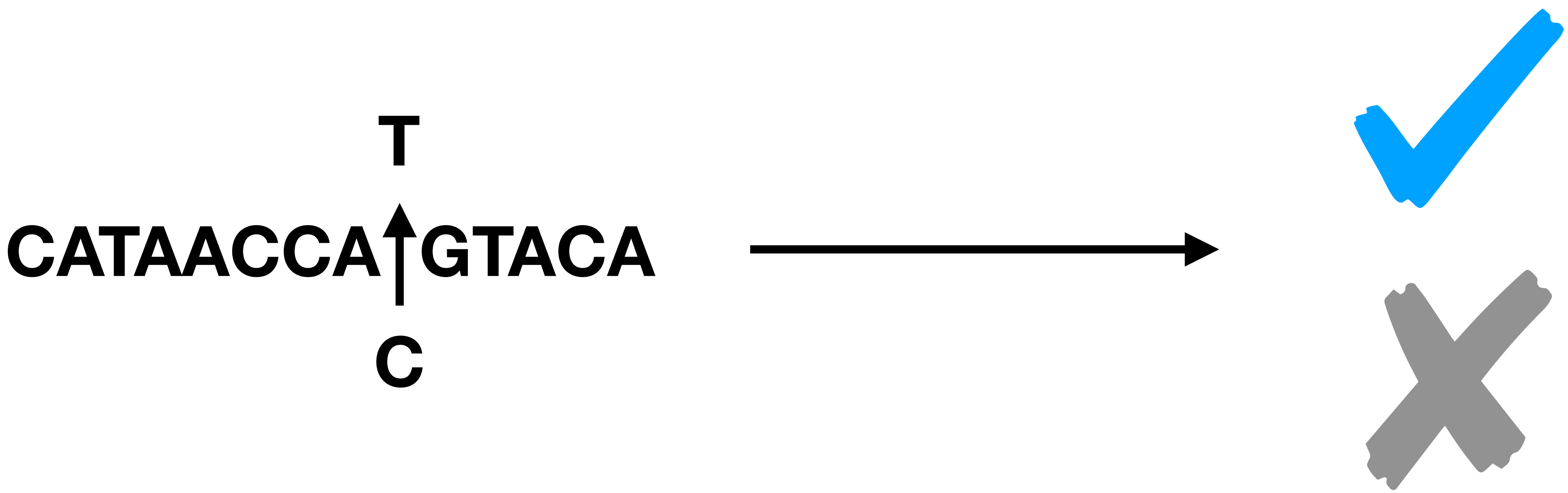
T

CATAACCA GTACA

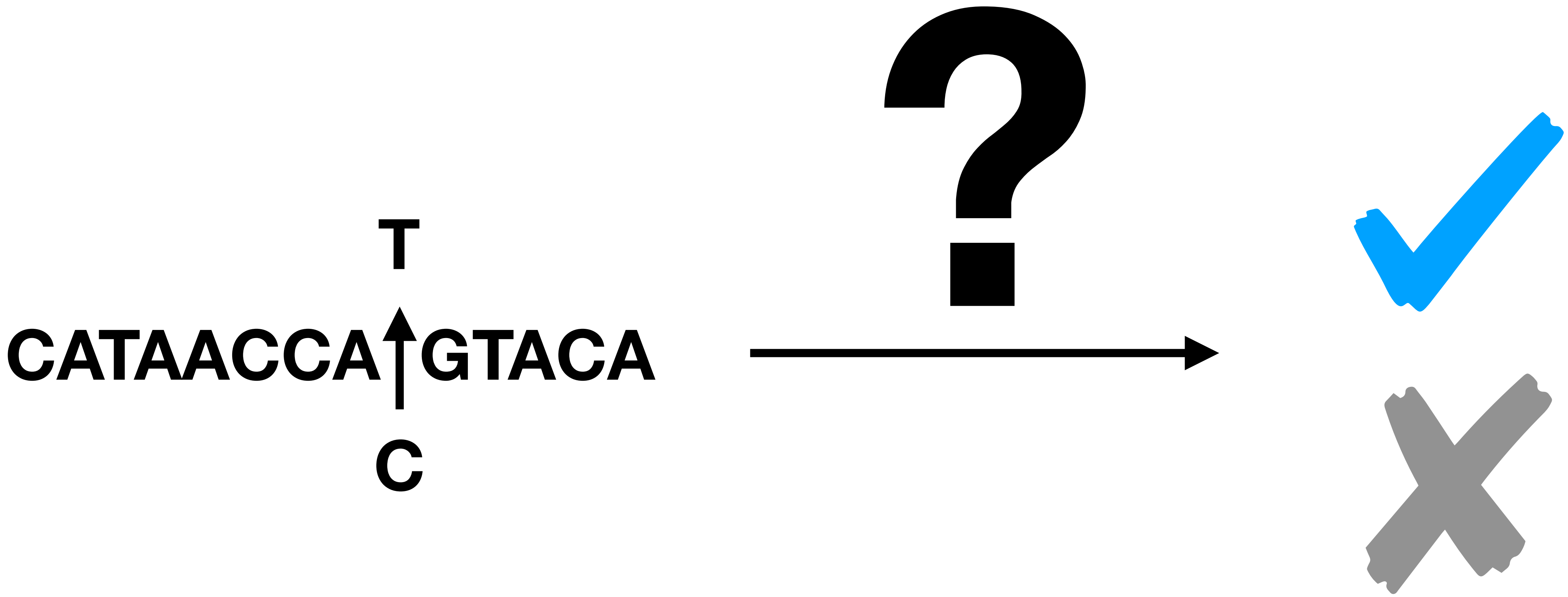
C

Understanding how **genotype** affects **phenotype** is a critical **challenge**.

Understanding how **genotype** affects **phenotype** is a critical **challenge**.

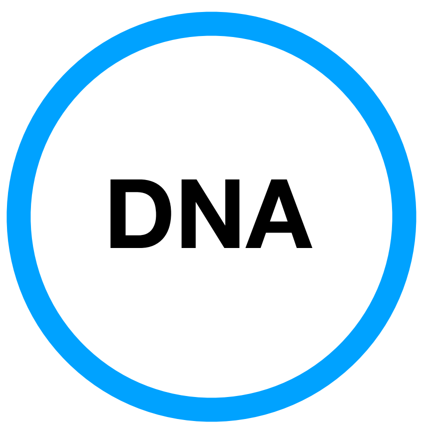


Understanding how **genotype** affects **phenotype** is a critical **challenge**.



The **Central Dogma** of **Biology**

The **Central Dogma** of **Biology**



The **Central Dogma** of **Biology**

DNA

```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

4 different bases

The **Central Dogma** of **Biology**

DNA → **RNA**

```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

4 different bases

The **Central Dogma** of **Biology**

DNA → **RNA**

DNA	RNA
ATGAGTATTCAACATTTCCGTGT	AUGAGUAUUCAACAUUUCCGUGU
CGCCCTTATTCCCTTTTTTGCGG	CGCCCUUAUUCCCUUUUUUGCGG
CATTTTGCCTTCCTGTTTTTGCT	CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCTGGTGAAAGT	CACCCAGAAACGCUGGUGAAAGU
AAAAGATGCTGAAGATCAGTTGG	AAAAGAUGCUGAAGAUCAGUUGG
GTGCACGAGTGGGTTACATCGAA	GUGCACGAGUGGGUUACAUCGAA
CTGGATCTCAACAGCGGTAAGAT	CUGGAUCUCAACAGCGGUAAGAU
CCTTGAGAGTTTTCGCCCCGAAG	CCUUGAGAGUUUUCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT	AACGUUUUCCAAUGAUGAGCACU
TTTAAAGTTCTGCTATGTGGCGC	UUUAAAGUUCUGCUAUGUGGCGC
GGTATTATCCCGTGTTGACGCCG	GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACTCGGTCGCCGC	GGCAAGAGCAACUCGGUCGCCGC
ATACACTATTCTCAGAATGACTT	AUACACUAUUCUCAGAAUGACUU
GGTTGAGTACTCACCAGTCACAG	GGUUGAGUACUCACCAGUCACAG
AAAAGCATCTTACGGATGGCATG	AAAAGCAUCUUACGGAUGGCAUG
ACAGTAAGAGAATTATGCAGTGC	ACAGUAAGAGAAUUAUGCAGUGC
TGCCATAACCATGAGTGATAACA	UGCCAUAACCAUGAGUGAUAACA
CTGCGGCCAACTTACTTCTGACA	CUGCGGCCAACUUACUUCUGACA
ACGATCGGAGGACCGAAGGAGCT	ACGAUCGGAGGACCGAAGGAGCU
AACCGCTTTTTTGCACAACATGG	AACCGCUUUUUUGCACAACAUGG
GGGATCATGTAACTCGCCTTGAT	GGGAUCAUGUAACUCGCCUUGAU
CGTTGGGAACCGGAGCTGAATGA	CGUUGGGAACCGGAGCUGAAUGA
AGCCATACCAAACGACGAG...	AGCCAUACCAAACGACGAG...

4 different bases

The **Central Dogma** of **Biology**

DNA → **RNA** → **Protein**

ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...

AUGAGUAUUCAACAUUUCCGUGU
CGCCCUUAUUCCCUUUUUUGCGG
CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCUGGUGAAAGU
AAAAGAUGCUGAAGAUCAGUUGG
GUGCACGAGUGGGUUACAUCGAA
CUGGAUCUCAACAGCGGUAAGAU
CCUUGAGAGUUUUCGCCCCGAAG
AACGUUUUCCAAUGAUGAGCACU
UUUAAAGUUCUGCUAUGUGGCGC
GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACUCGGUCGCCGC
AUACACUAUUCUCAGAAUGACUU
GGUUGAGUACUCACCAGUCACAG
AAAAGCAUCUUACGGAUGGCAUG
ACAGUAAGAGAAUUAUGCAGUGC
UGCCAUAACCAUGAGUGAUAACA
CUGCGGCCAACUUACUUCUGACA
ACGAUCGGAGGACCGAAGGAGCU
AACCGCUUUUUUGCACAACAUGG
GGGAUCAUGUAACUCGCCUUGAU
CGUUGGGAACCGGAGCUGAAUGA
AGCCAUACCAAACGACGAG...

4 different bases

The **Central Dogma** of **Biology**

DNA → **RNA** → **Protein**

DNA

```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

RNA

```
AUGAGUAUUCAACAUUUCCGUGU
CGCCCUUAUUCCCUUUUUUGCGG
CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCUGGUGAAAGU
AAAAGAUGCUGAAGAUCAGUUGG
GUGCACGAGUGGGUUACAUCGAA
CUGGAUCUCAACAGCGGUAAGAU
CCUUGAGAGUUUUCGCCCCGAAG
AACGUUUUCCAAUGAUGAGCACU
UUUAAAGUUCUGCUAUGUGGCGC
GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACUCGGUCGCCGC
AUACACUAUUCUCAGAAUGACUU
GGUUGAGUACUCACCAGUCACAG
AAAAGCAUCUUACGGAUGGCAUG
ACAGUAAGAGAAUUAUGCAGUGC
UGCCAUAACCAUGAGUGAUAACA
CUGCGGCCAACUUACUUCUGACA
ACGAUCGGAGGACCGAAGGAGCU
AACCGCUUUUUUGCACAACAUGG
GGGAUCAUGUAACUCGCCUUGAU
CGUUGGGAACCGGAGCUGAAUGA
AGCCAUACCAAACGACGAG...
```

Protein

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

20 different amino acids

The **Central Dogma** of **Biology**

DNA → **RNA** → **Protein**

DNA

```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

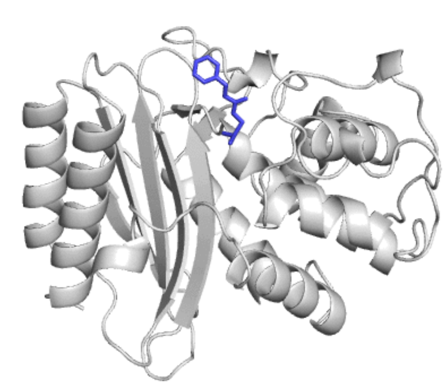
RNA

```
AUGAGUAUUCAACAUUUCCGUGU
CGCCCUUAUUCCCUUUUUUGCGG
CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCUGGUGAAAGU
AAAAGAUGCUGAAGAUCAGUUGG
GUGCACGAGUGGGUUACAUCGAA
CUGGAUCUCAACAGCGGUAAGAU
CCUUGAGAGUUUUCGCCCCGAAG
AACGUUUUCCAAUGAUGAGCACU
UUUAAAGUUCUGCUAUGUGGCGC
GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACUCGGUCGCCGC
AUACACUAUUCUCAGAAUGACUU
GGUUGAGUACUCACCAGUCACAG
AAAAGCAUCUUACGGAUGGCAUG
ACAGUAAGAGAAUUAUGCAGUGC
UGCCAUAACCAUGAGUGAUAACA
CUGCGGCCAACUUACUUCUGACA
ACGAUCGGAGGACCGAAGGAGCU
AACCGCUUUUUUGCACAACAUGG
GGGAUCAUGUAACUCGCCUUGAU
CGUUGGGAACCGGAGCUGAAUGA
AGCCAUACCAAACGACGAG...
```

Protein

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

20 different amino acids



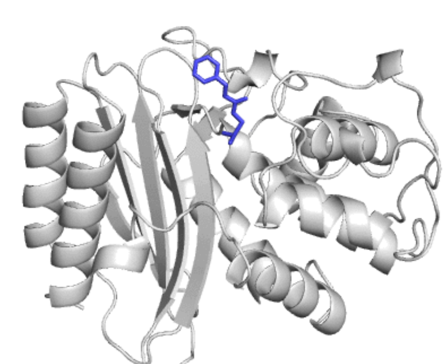
The **Central Dogma** of **Biology**

DNA → **RNA** → **Protein** → **Function**

```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

```
AUGAGUAUUCAACAUUUCCGUGU
CGCCCUUAUUCCCUUUUUUGCGG
CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCUGGUGAAAGU
AAAAGAUGCUGAAGAUCAGUUGG
GUGCACGAGUGGGUUACAUCGAA
CUGGAUCUCAACAGCGGUAAGAU
CCUUGAGAGUUUUCGCCCCGAAG
AACGUUUUCCAAUGAUGAGCACU
UUUAAAGUUCUGCUAUGUGGCGC
GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACUCGGUCGCCGC
AUACACUAUUCUCAGAAUGACUU
GGUUGAGUACUCACCAGUCACAG
AAAAGCAUCUUACGGAUGGCAUG
ACAGUAAGAGAAUUAUGCAGUGC
UGCCAUAACCAUGAGUGAUAACA
CUGCGGCCAACUUACUUCUGACA
ACGAUCGGAGGACCGAAGGAGCU
AACCGCUUUUUUGCACAACAUGG
GGGAUCAUGUAACUCGCCUUGAU
CGUUGGGAACCGGAGCUGAAUGA
AGCCAUACCAAACGACGAG...
```

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```



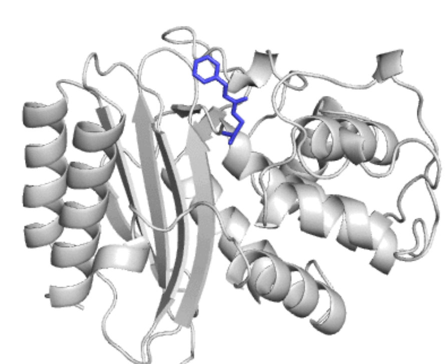
The **Central Dogma** of **Biology**

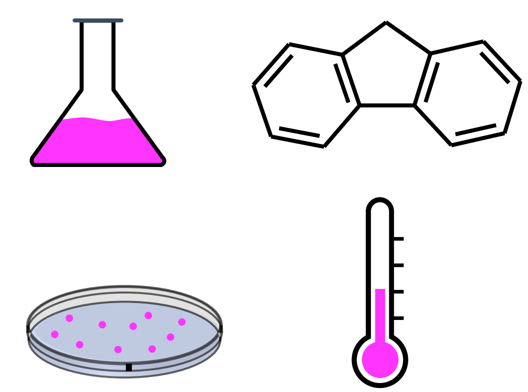
DNA → **RNA** → **Protein** → **Function**

ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...

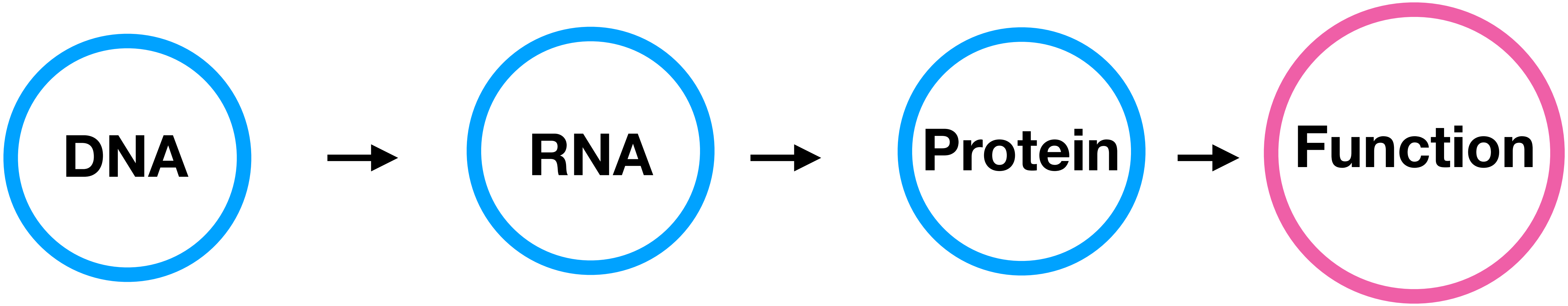
AUGAGUAUUCAACAUUUCCGUGU
CGCCCUUAUUCCCUUUUUUGCGG
CAUUUUGCCUUCCUGUUUUUGCU
CACCCAGAAACGCUGGUGAAAGU
AAAAGAUGCUGAAGAUCAGUUGG
GUGCACGAGUGGGUUACAUCGAA
CUGGAUCUCAACAGCGGUAAGAU
CCUUGAGAGUUUUCGCCCCGAAG
AACGUUUUCCAAUGAUGAGCACU
UUUAAAGUUCUGCUAUGUGGCGC
GGUAUUAUCCCGUGUUGACGCCG
GGCAAGAGCAACUCGGUCGCCGC
AUACACUAUUCUCAGAAUGACUU
GGUUGAGUACUCACCAGUCACAG
AAAAGCAUCUUACGGAUGGCAUG
ACAGUAAGAGAAUUAUGCAGUGC
UGCCAUAACCAUGAGUGAUAACA
CUGCGGCCAACUUACUUCUGACA
ACGAUCGGAGGACCGAAGGAGCU
AACCGCUUUUUUGCACAACAUGG
GGGAUCAUGUAACUCGCCUUGAU
CGUUGGGAACCGGAGCUGAAUGA
AGCCAUACCAAACGACGAG...

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

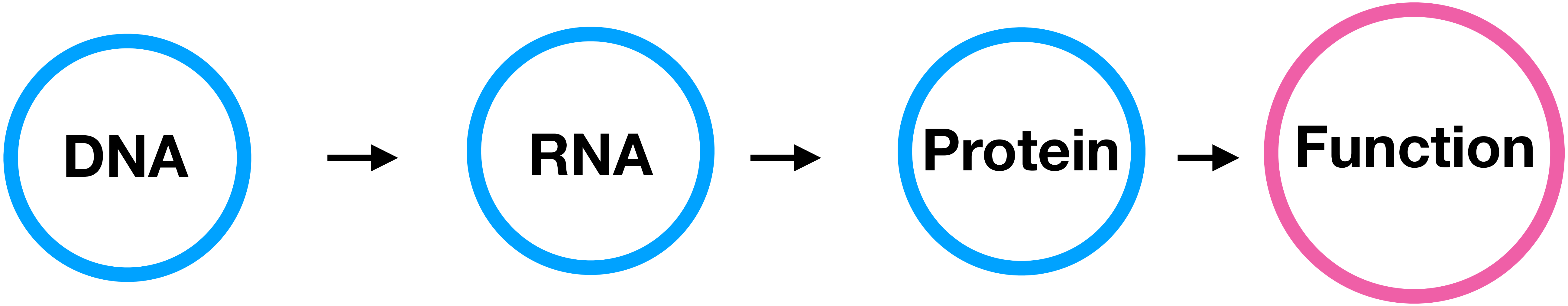




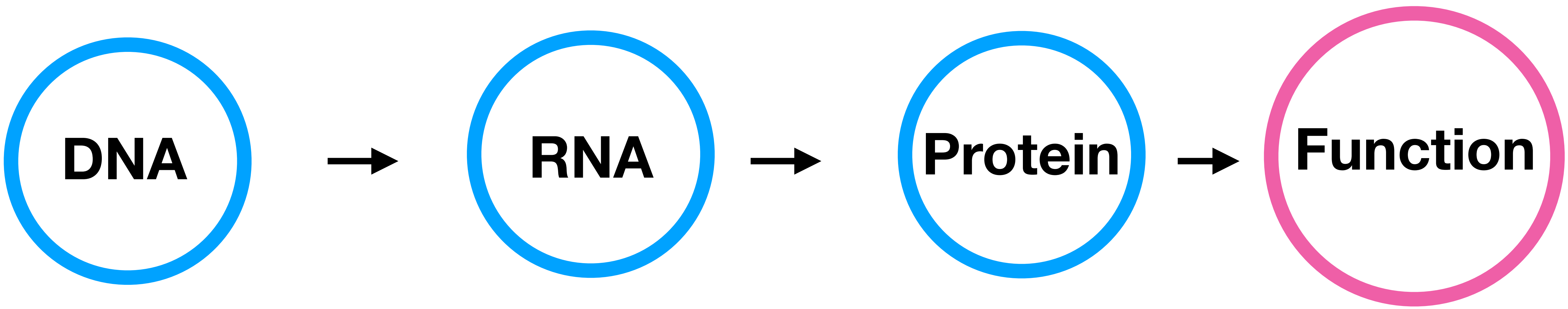
The **Central Dogma** of **Biology**



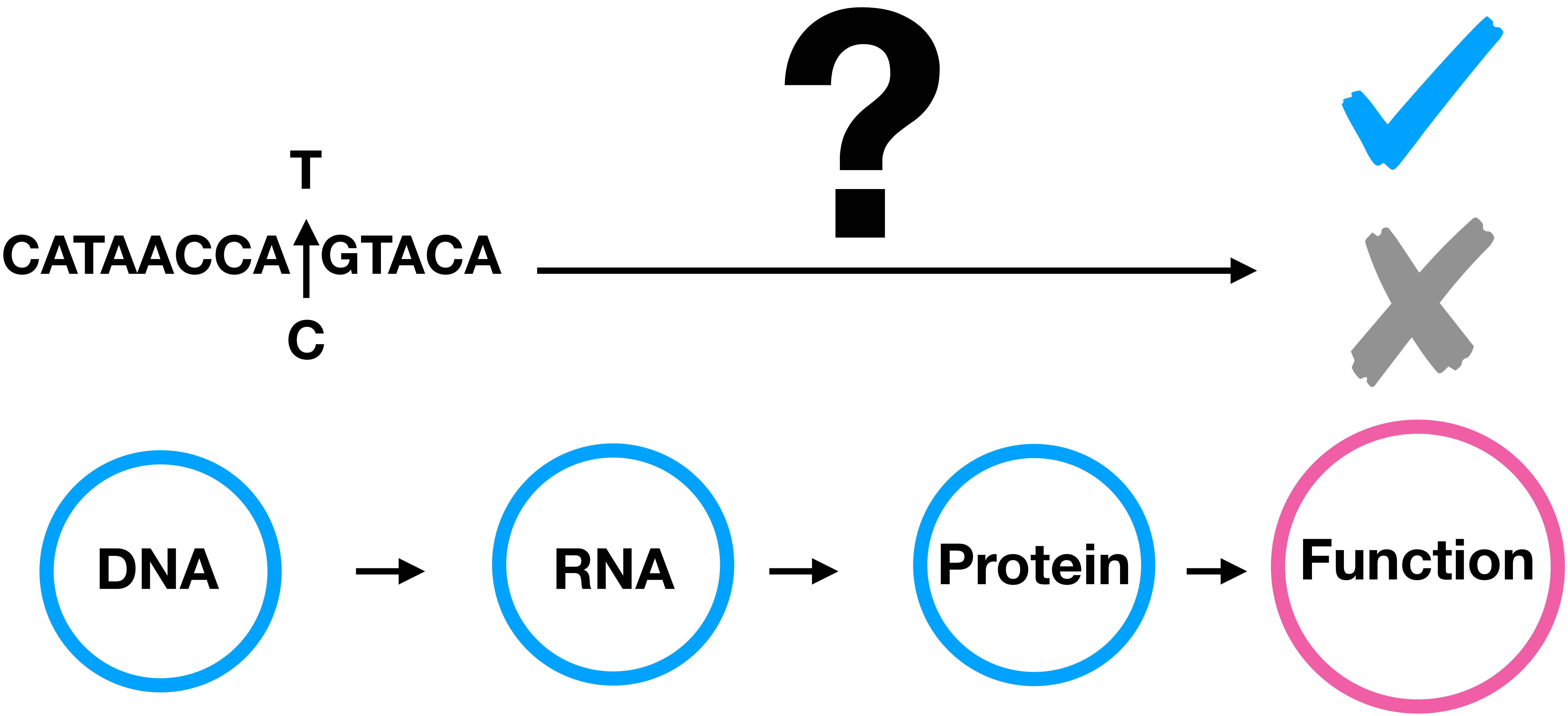
Mutations alter biological **function**



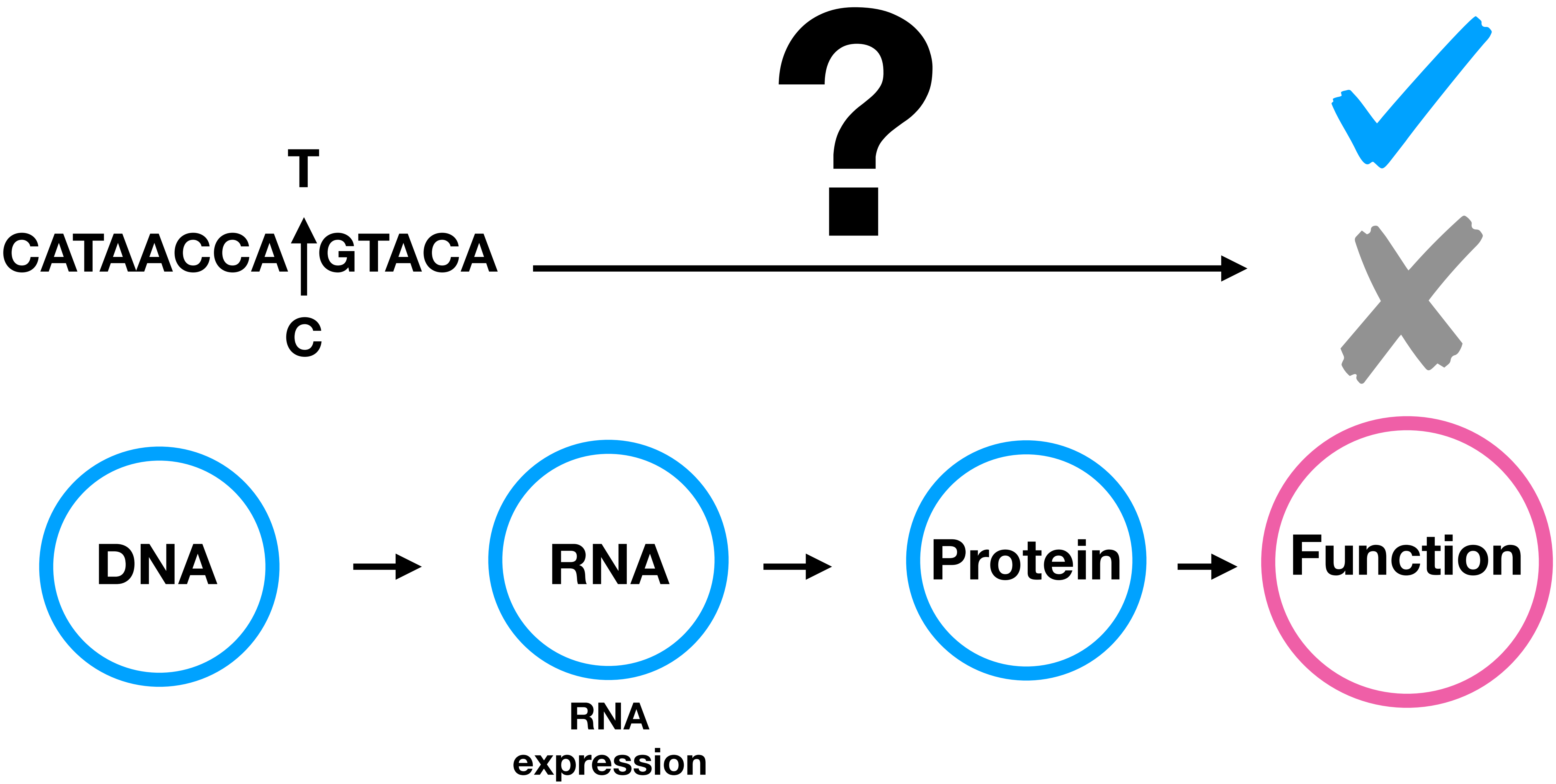
Mutations alter biological **function**



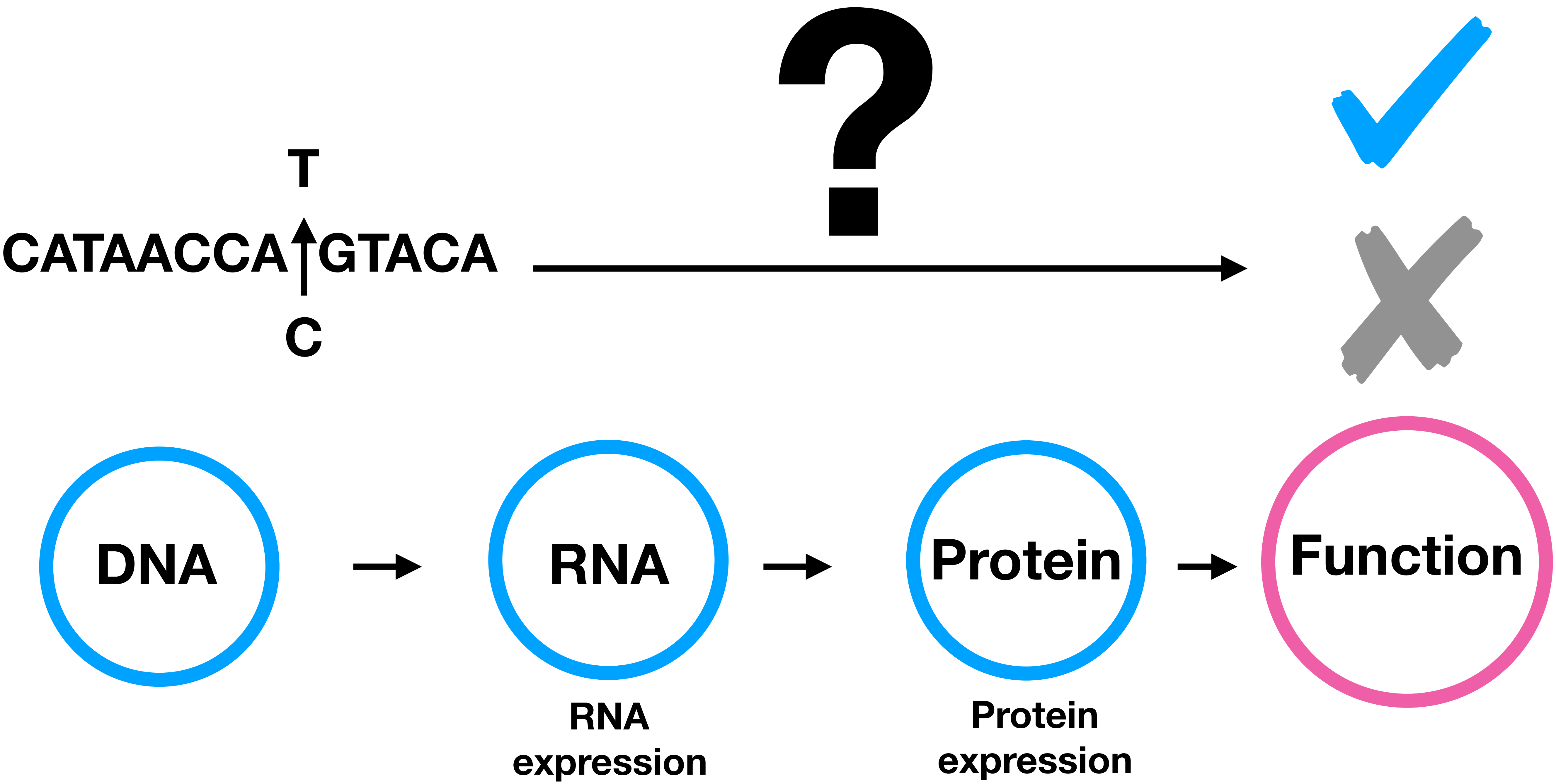
Mutations alter biological **function**



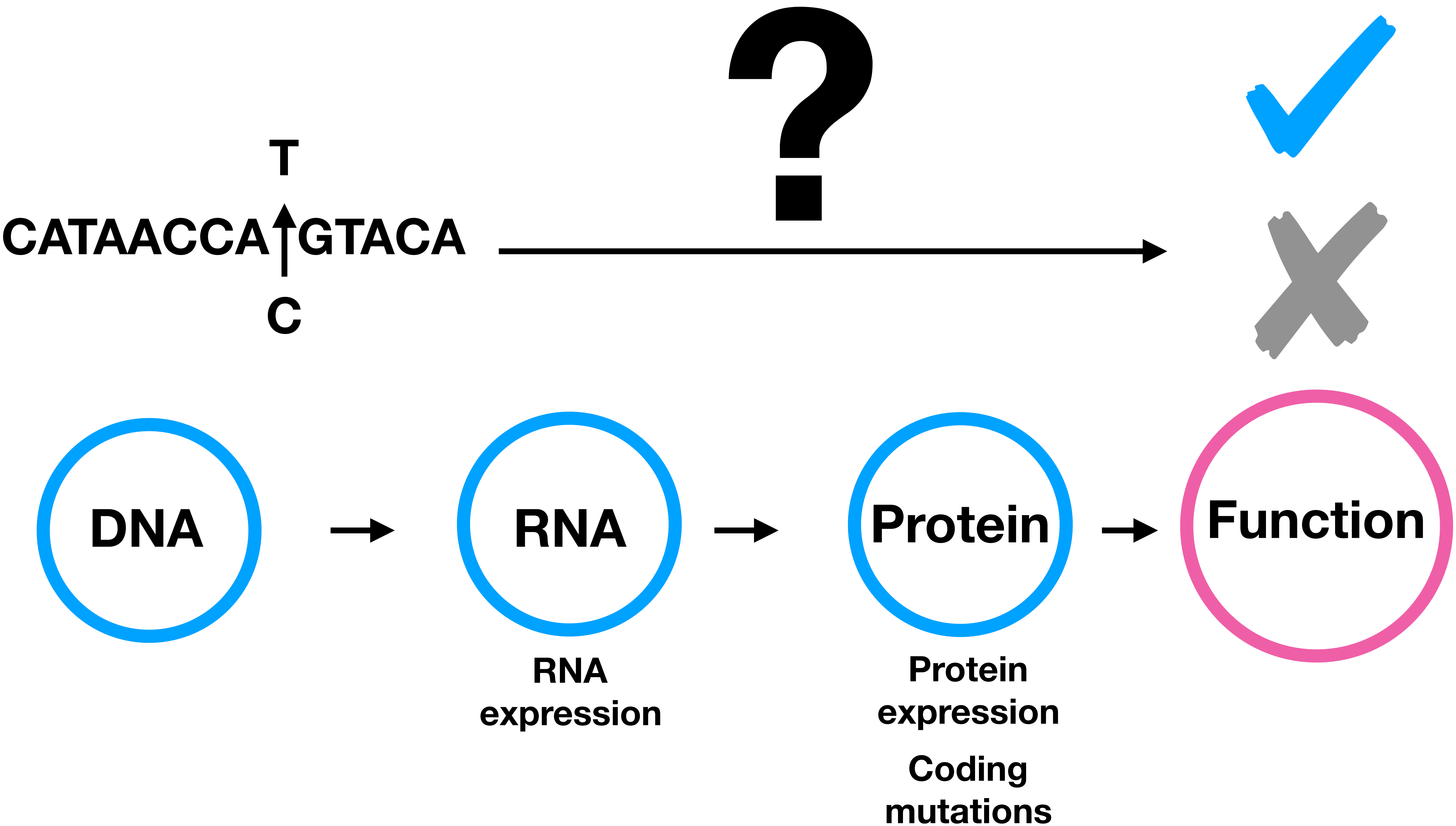
Mutations alter biological **function**



Mutations alter biological **function**



Mutations alter biological **function**



Mutation effect prediction is **important**

Mutation effect prediction is **important**

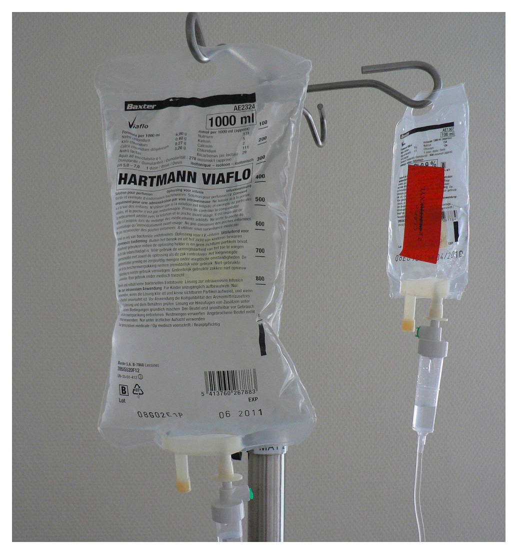
Understanding disease

“Does this mutation cause cancer?”

Mutation effect prediction is **important**

Understanding disease

Biomedicine



"Does this mutation cause cancer?"

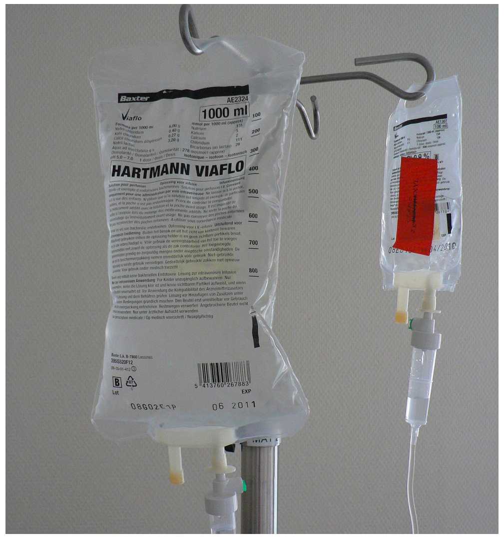
"Is this antibody stable in a patient?"

Mutation effect prediction is important

Understanding disease

Biomedicine

Bioengineering

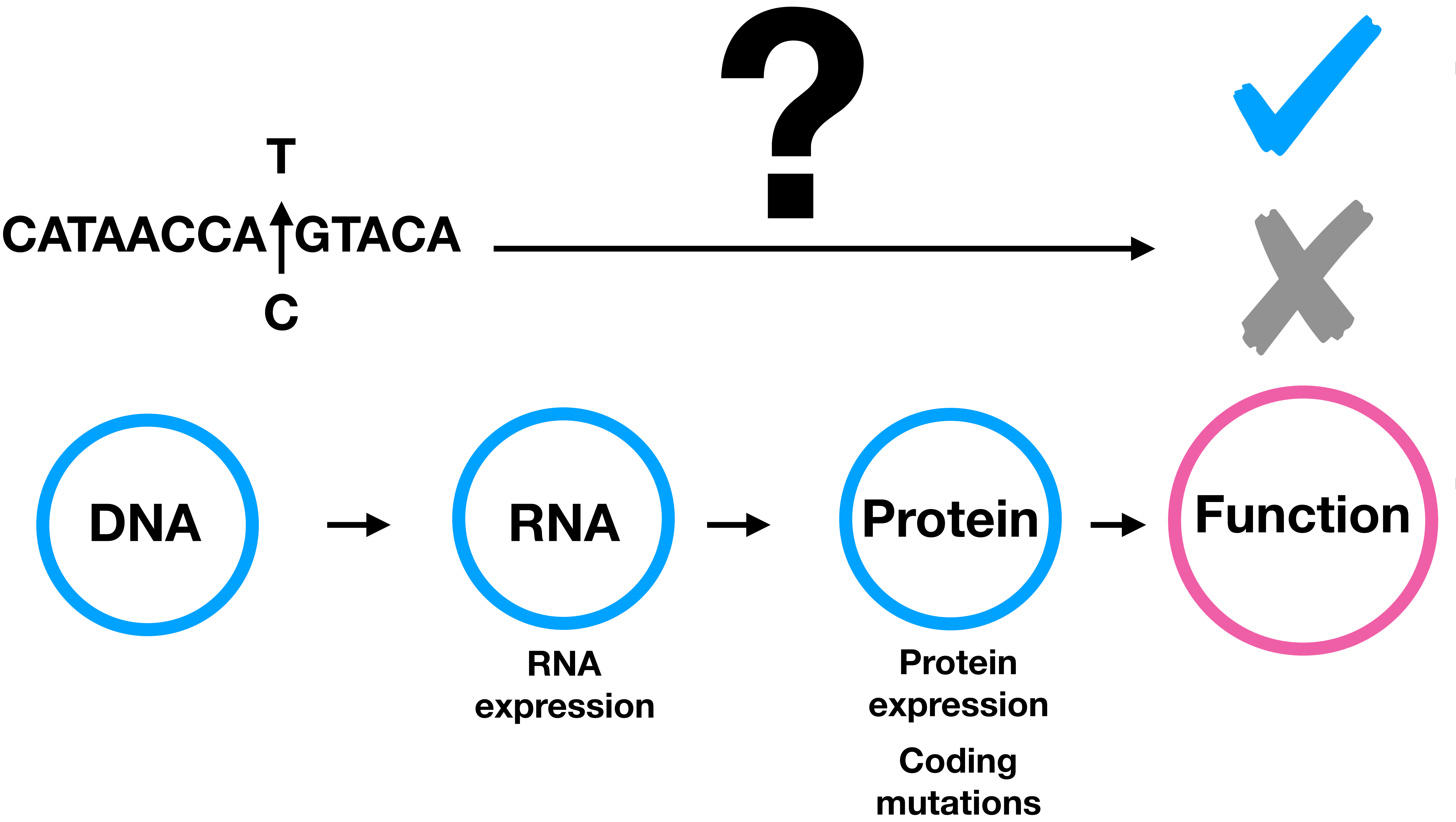


"Does this mutation cause cancer?"

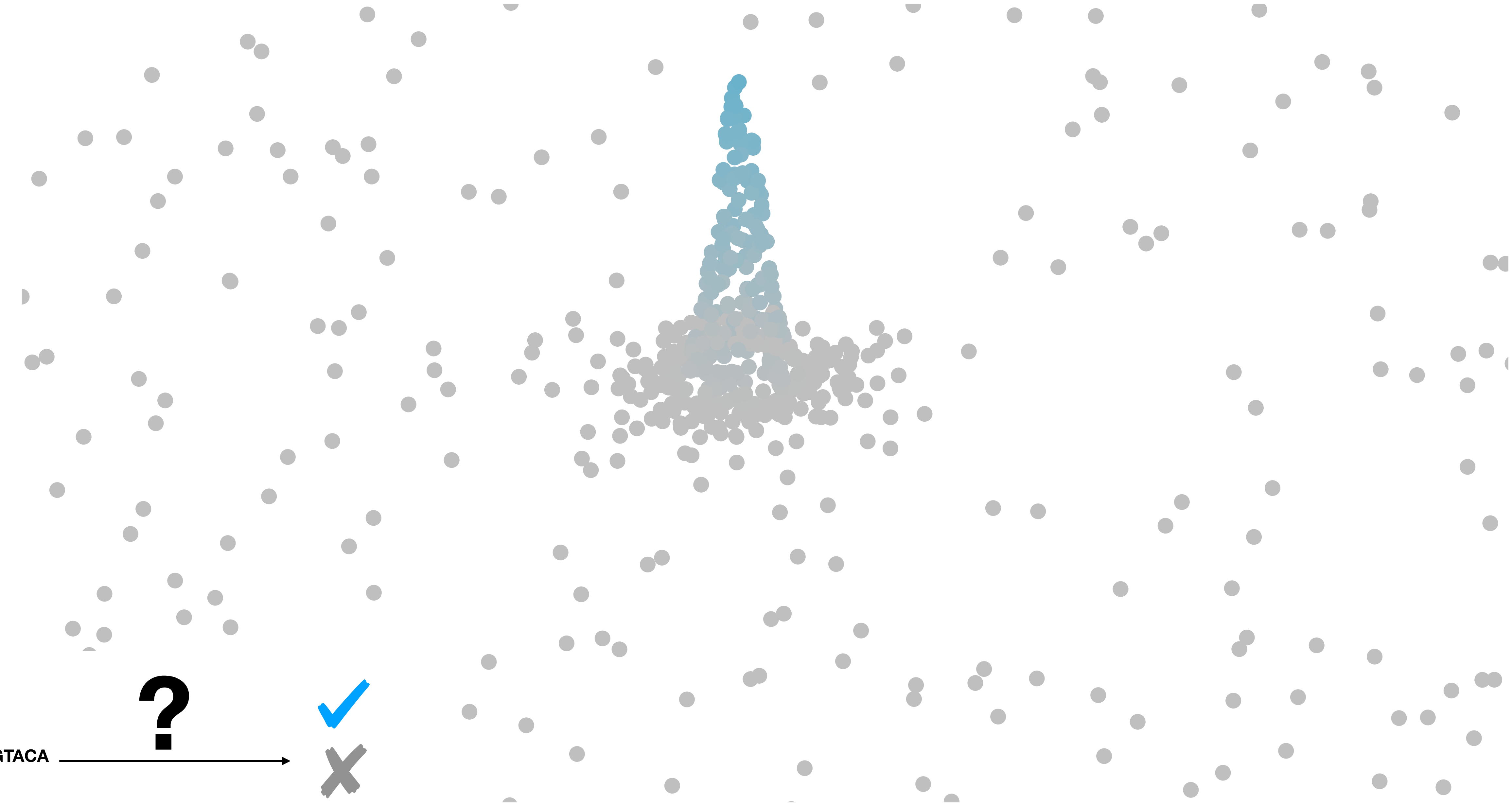
"Is this antibody stable in a patient?"

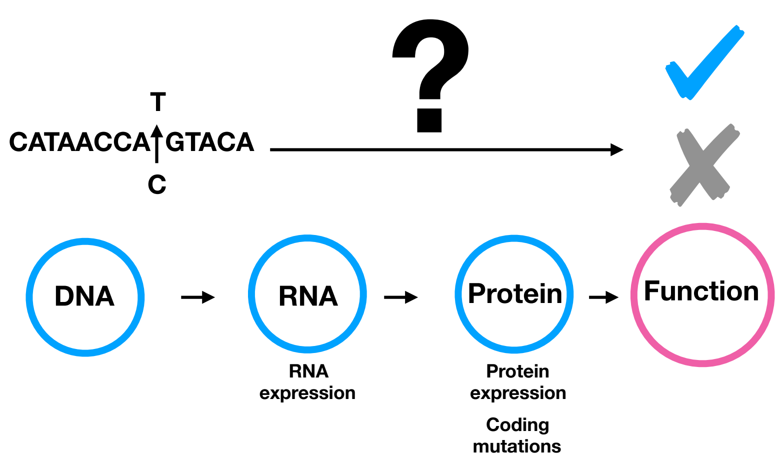
"Can this microbe be used to create vitamins?"

Mutations alter biological **function**

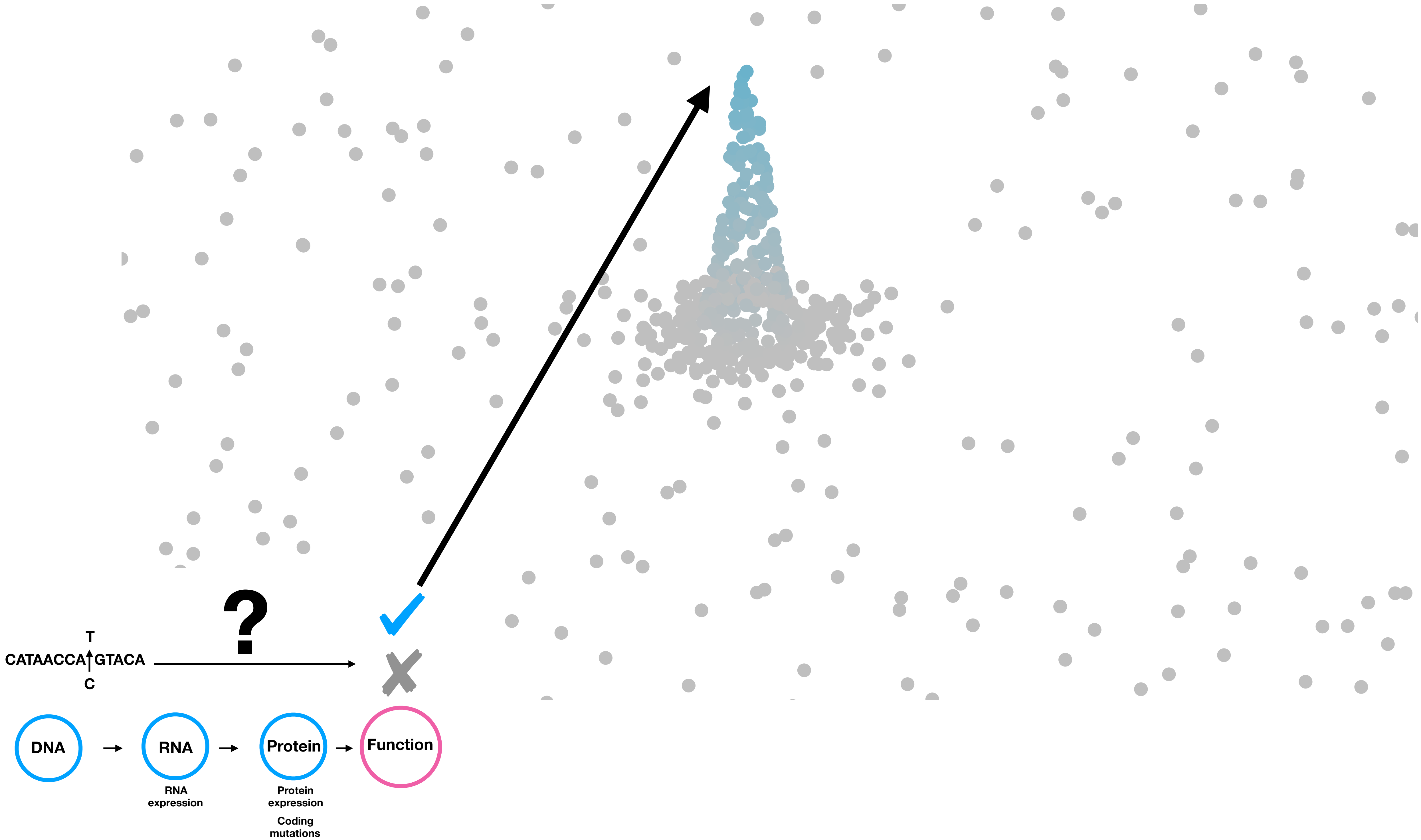


Mutations alter biological **function**

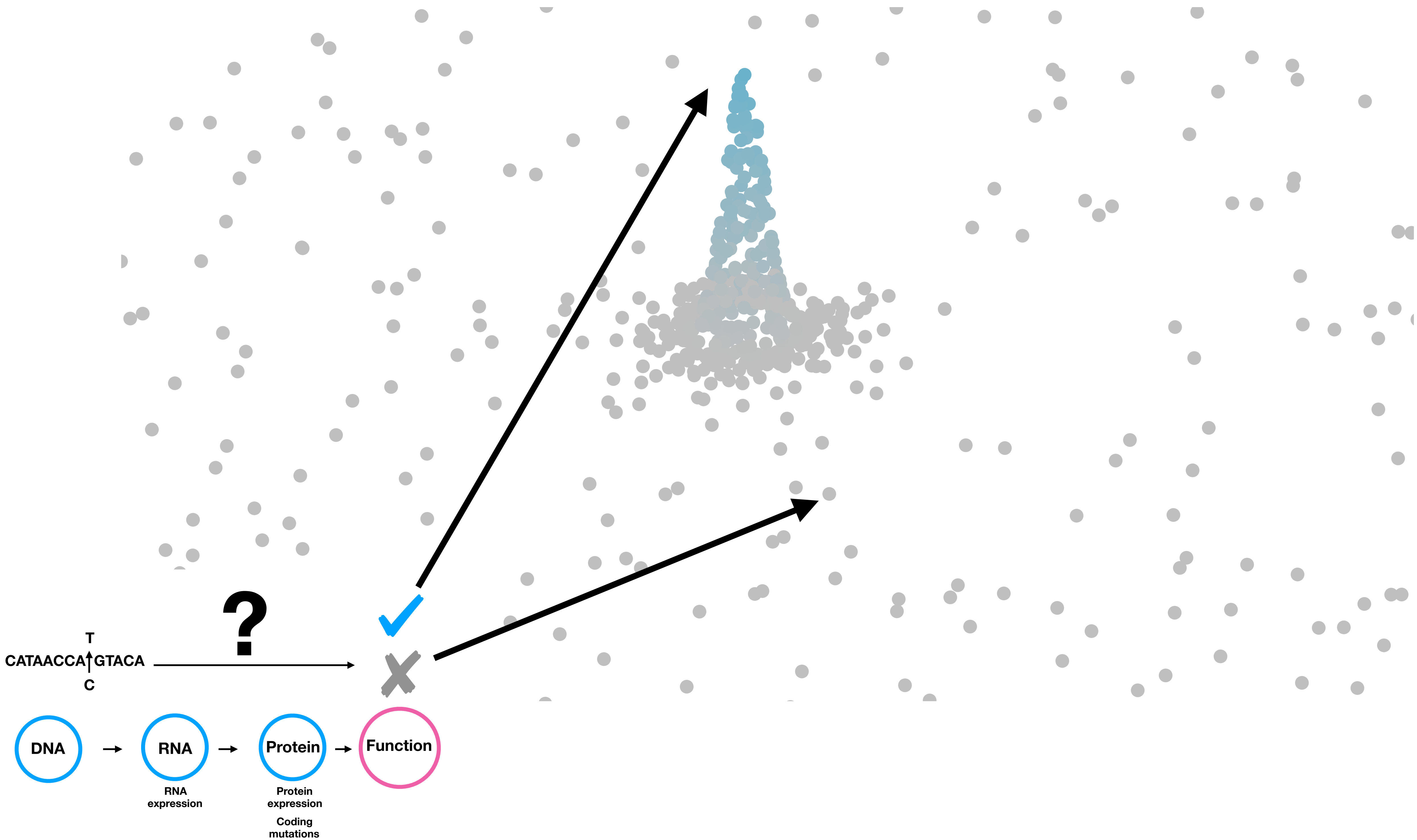




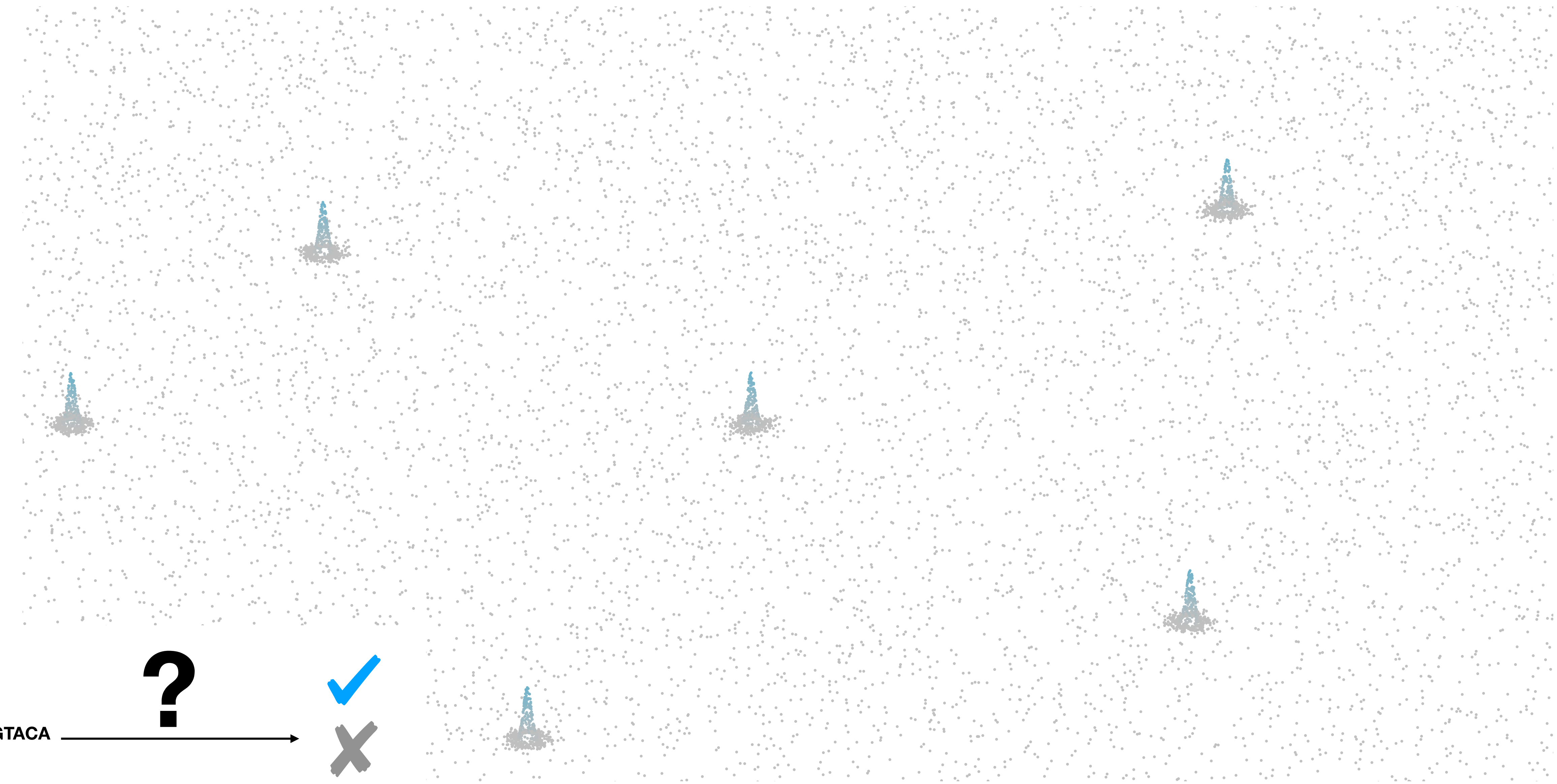
Mutations alter biological **function**

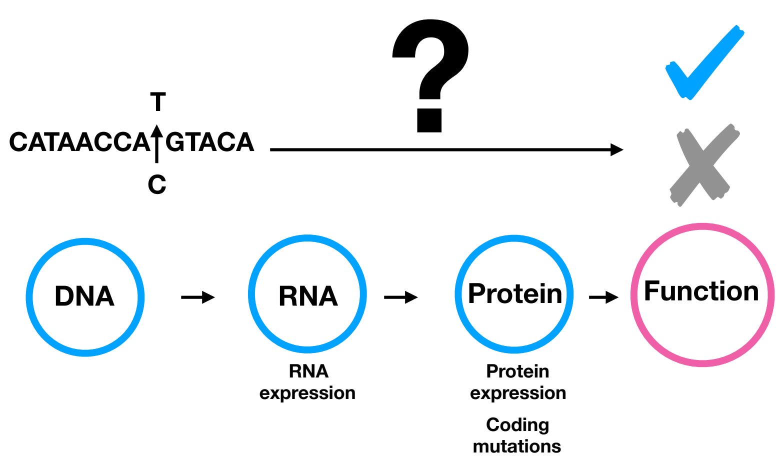


Mutations alter biological **function**

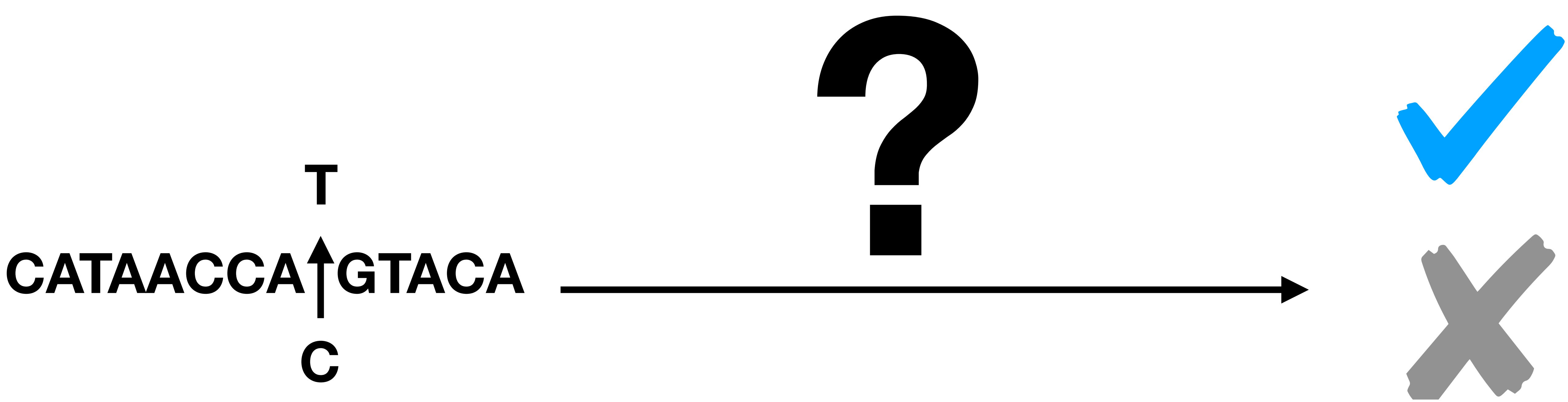


Mutations alter biological **function**



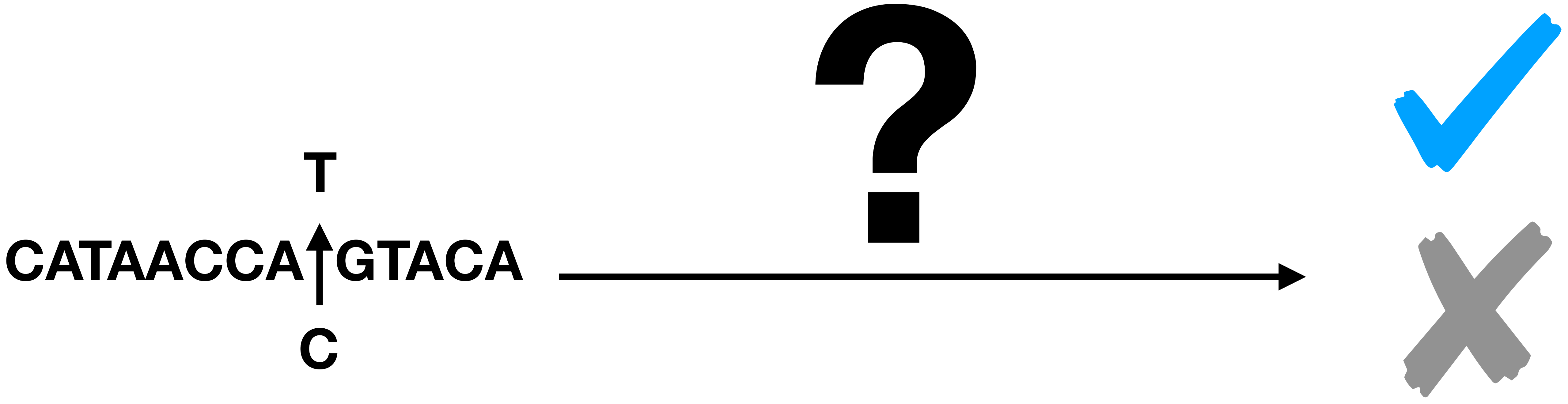


Mutations alter biological **function**



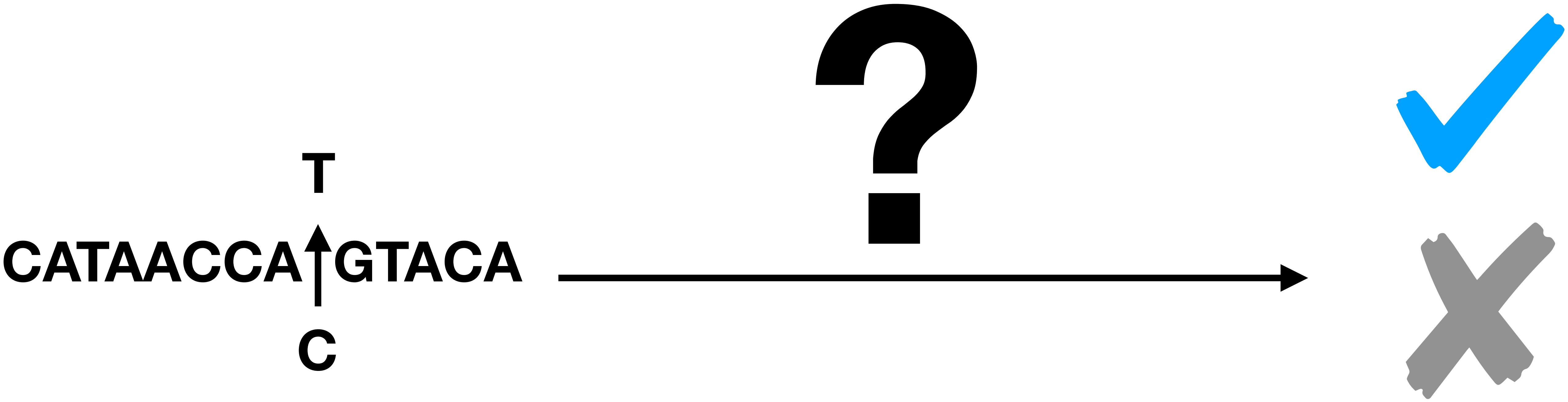
Mutation effect prediction

is **hard**



Mutation effect prediction
is **hard**

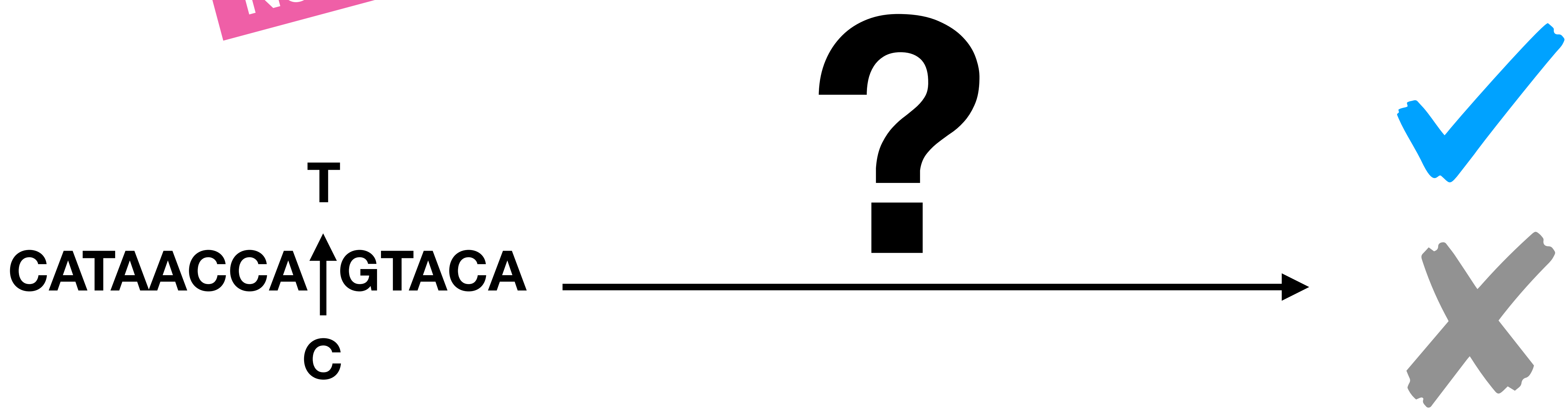
Sparsely sampled



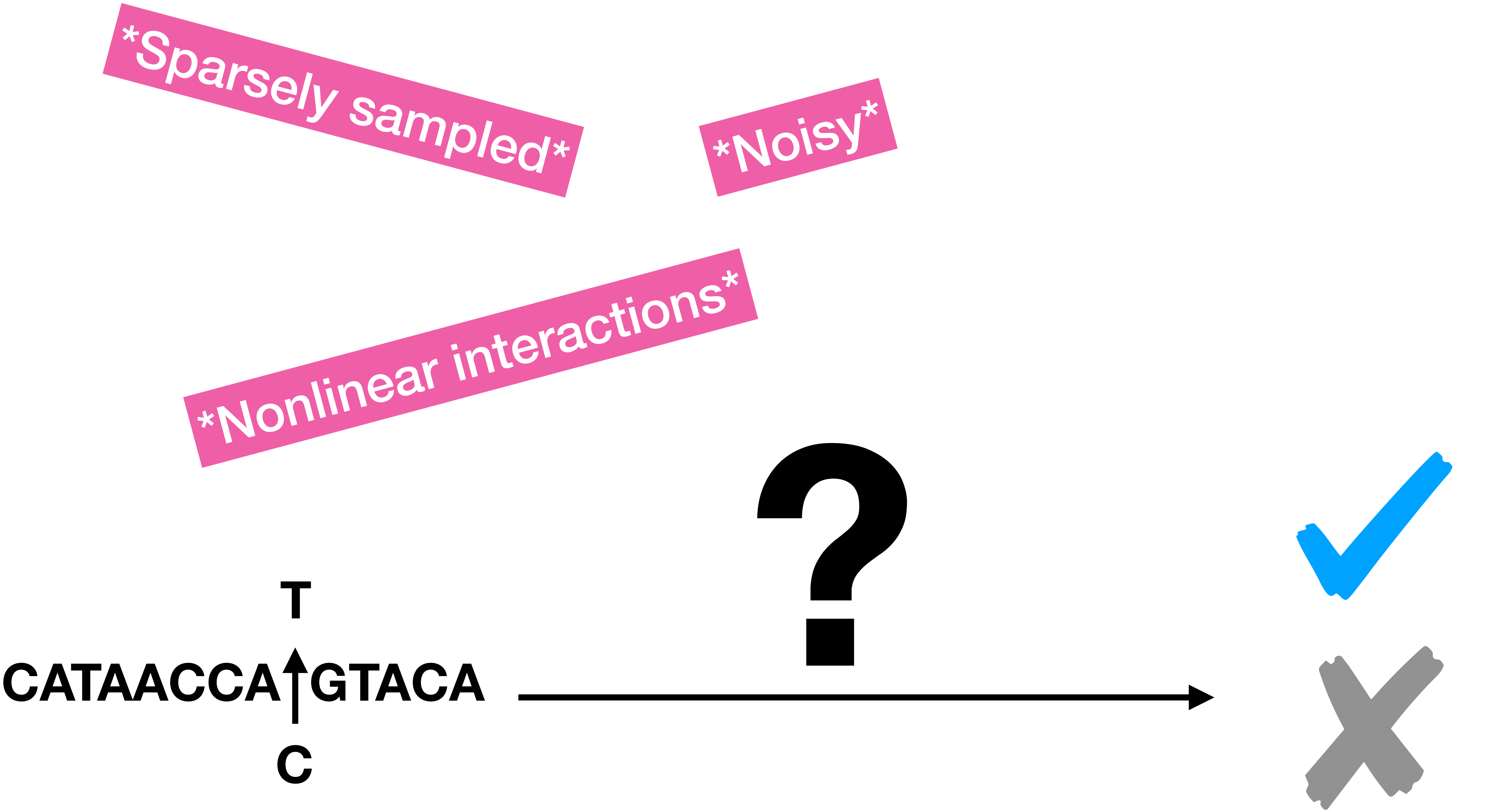
Mutation effect prediction is **hard**

Sparsely sampled

Nonlinear interactions



Mutation effect prediction is **hard**



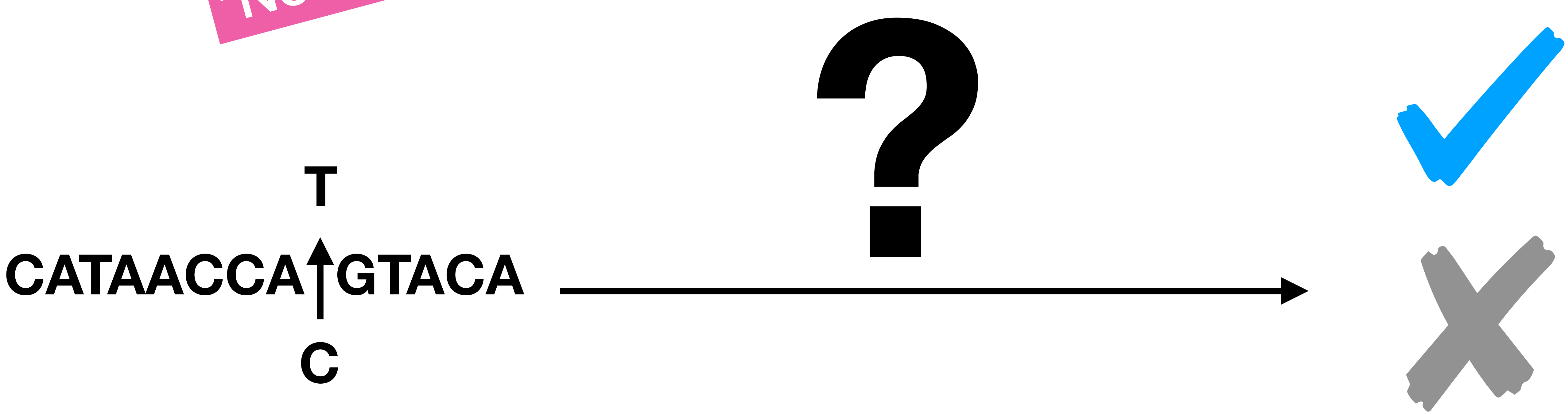
Mutation effect prediction is **hard**

Sparsely sampled

Noisy

Nonlinear interactions

Confounders



Mutation effect prediction is **hard**

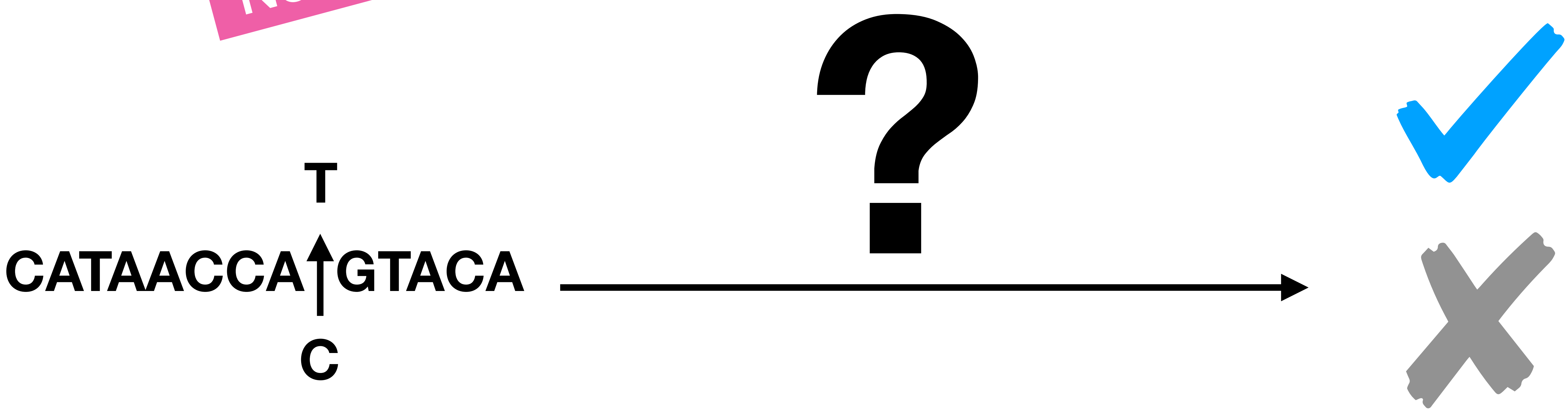
Sparsely sampled

Noisy

Effect not measured

Nonlinear interactions

Confounders



Part I:

Genotype -> Phenotype

in **proteins**

Part I:

Genotype -> Phenotype in **proteins**

DNA

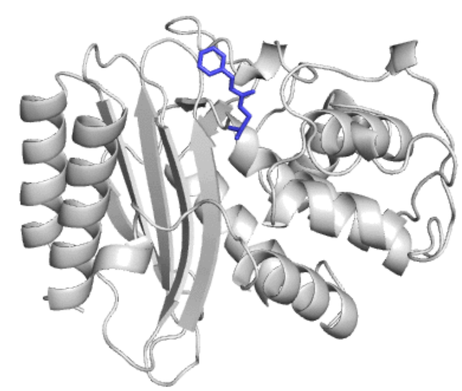
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...

RNA

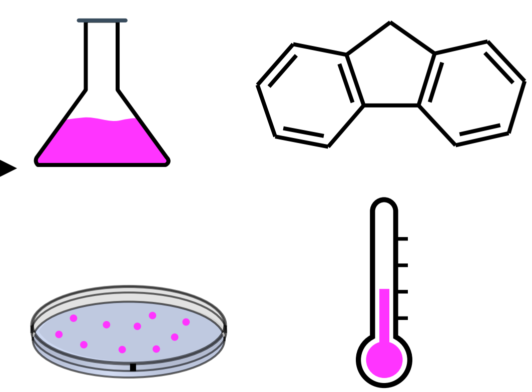
Protein

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

Structure



Function



Part I:

Genotype -> Phenotype
in **proteins**

DNA

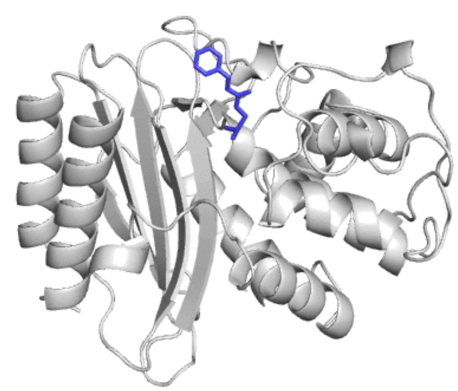
```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

RNA

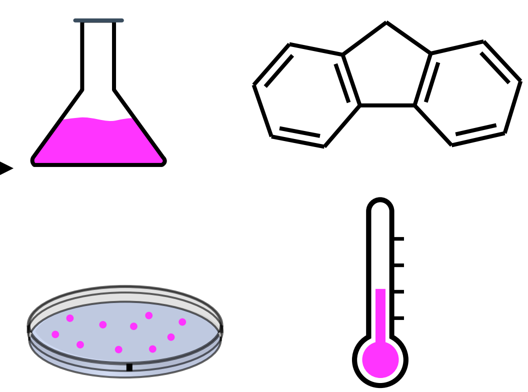
Protein

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

Structure



Function



Mutations impact protein **function**

Mutations impact protein **function**

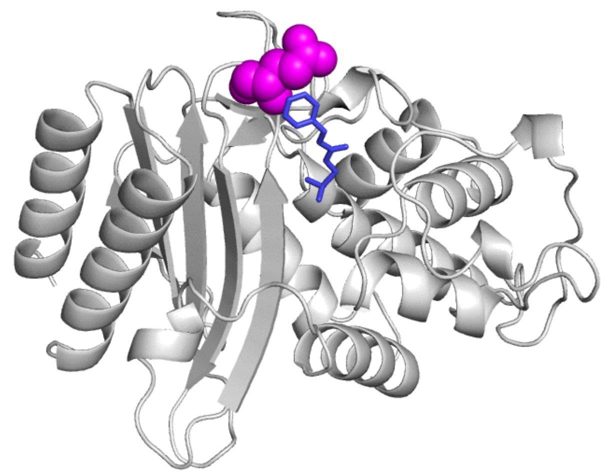
```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

Sequence

Mutations impact protein **function**

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→L

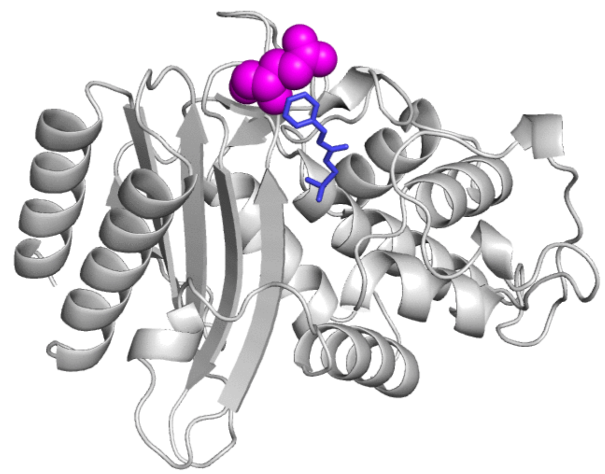


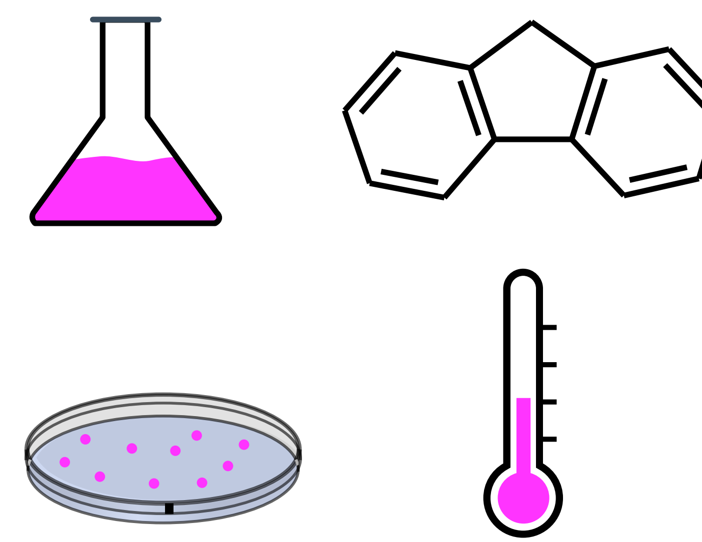
Sequence **Structure**

Mutations impact protein **function**

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→L



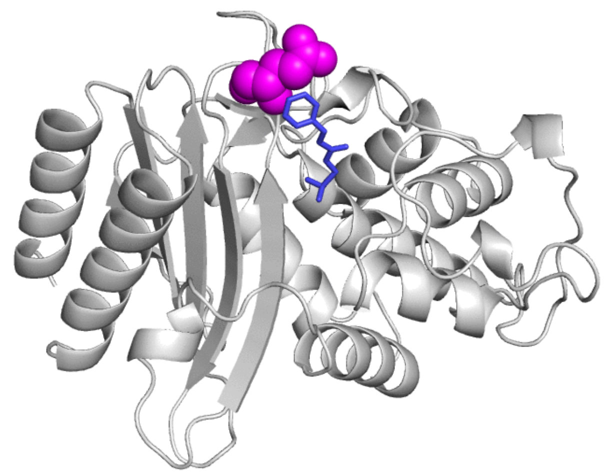


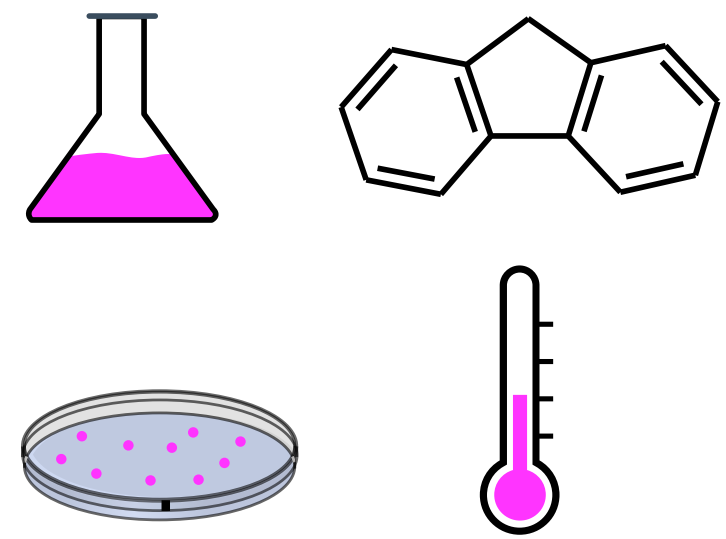
Sequence **Structure** **Function**

Mutations impact protein **function**

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→L





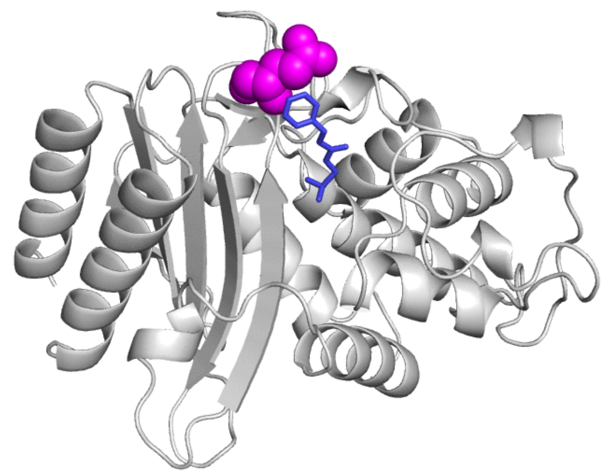
```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

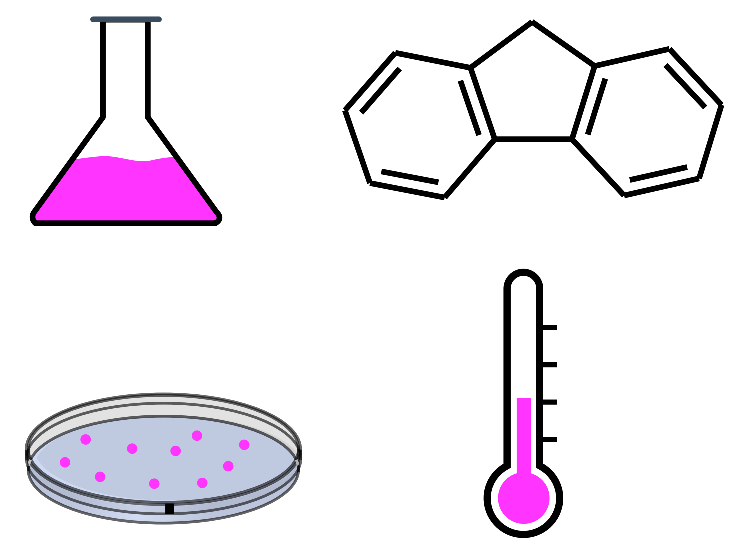
Sequence **Structure** **Function**

Mutations impact protein **function**

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

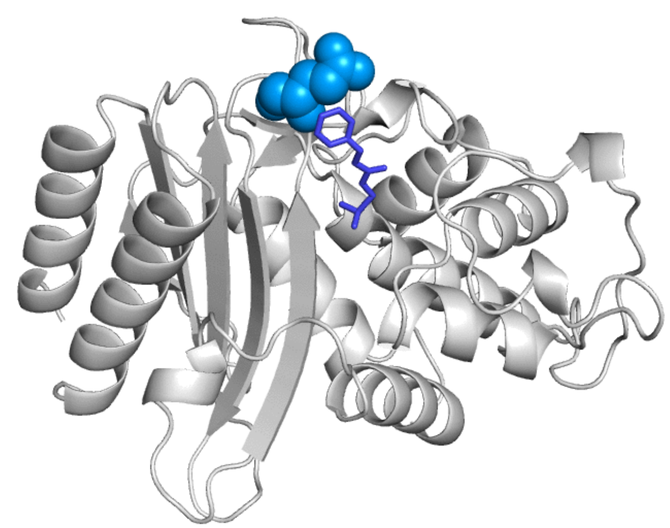
E→L





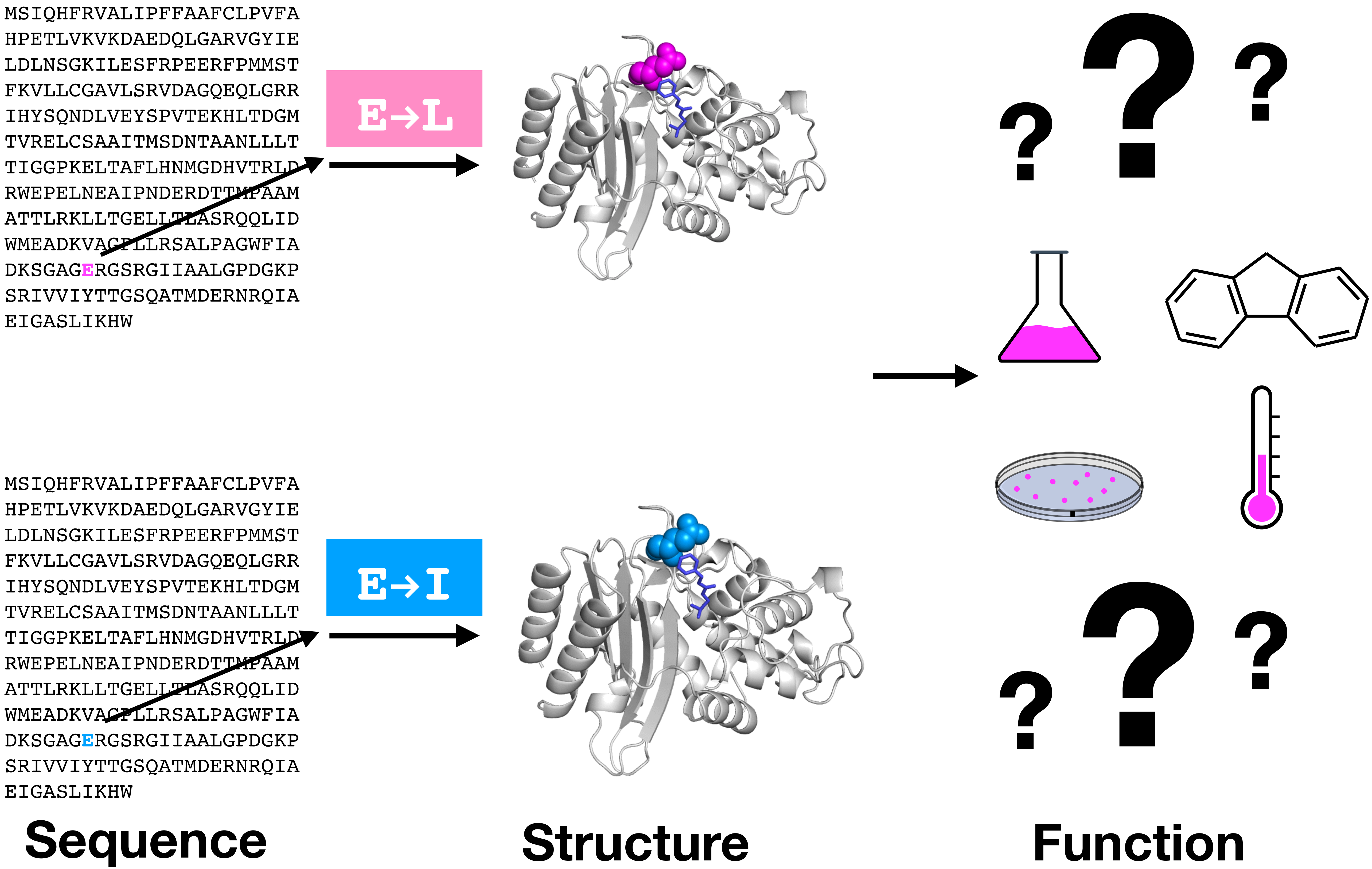
```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→I

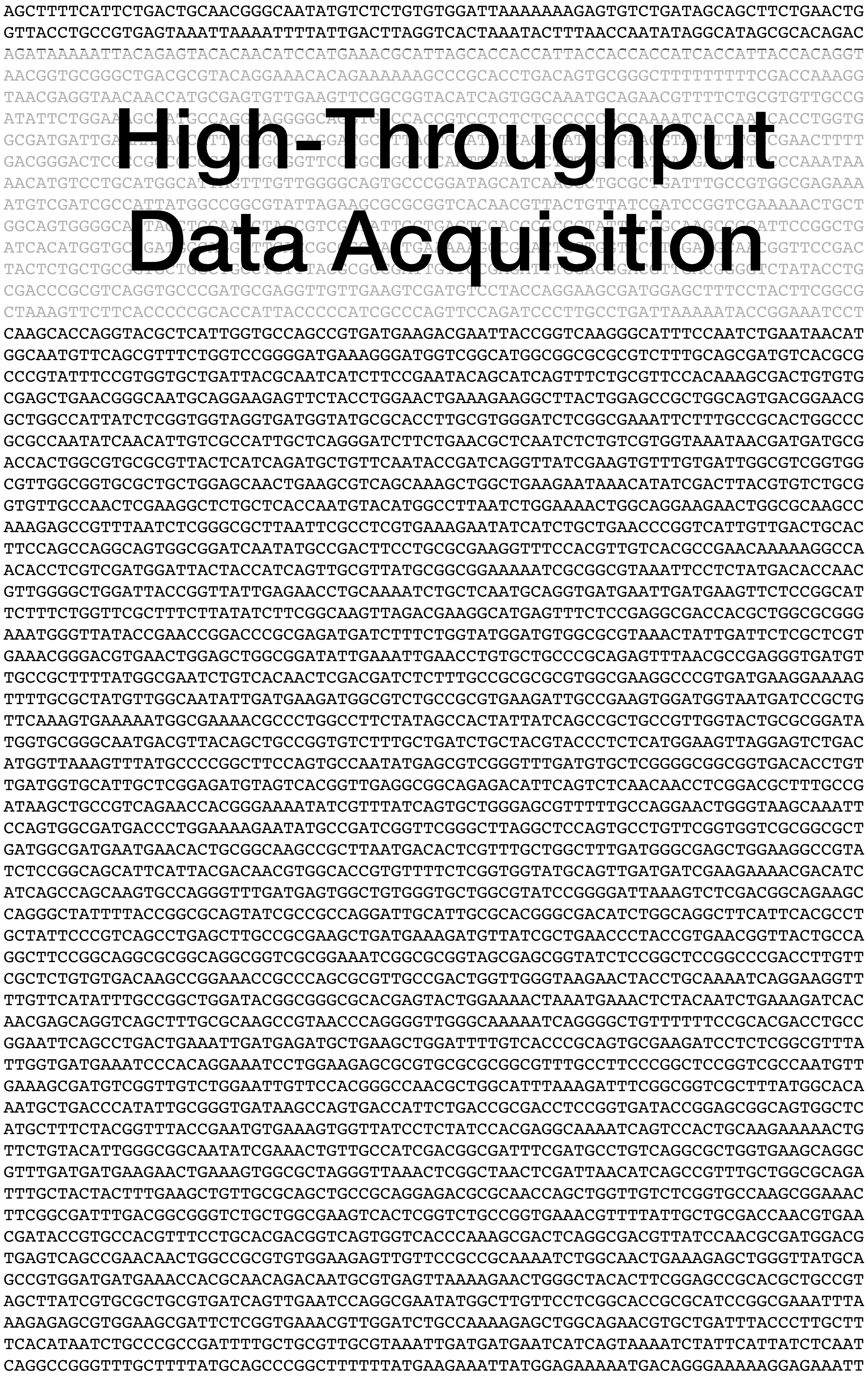


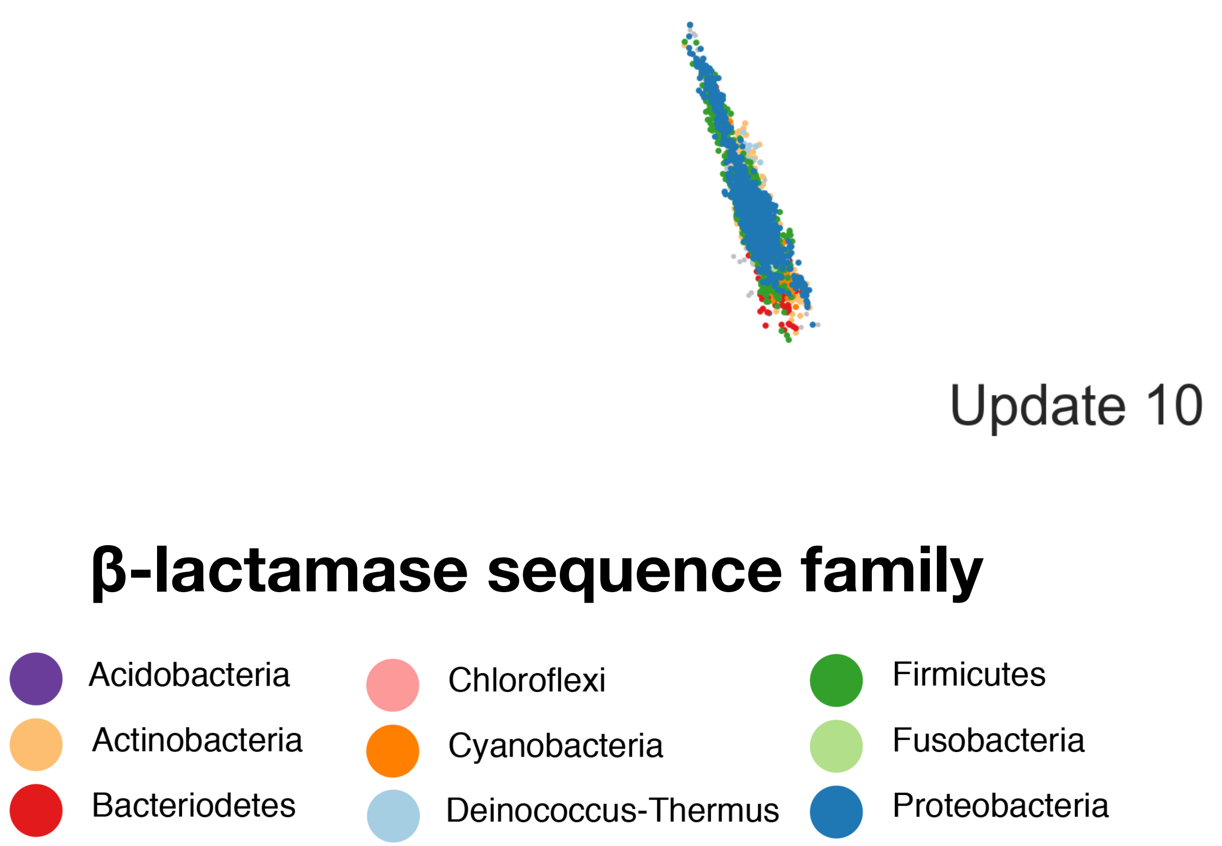
Sequence **Structure** **Function**

Mutations impact protein **function**



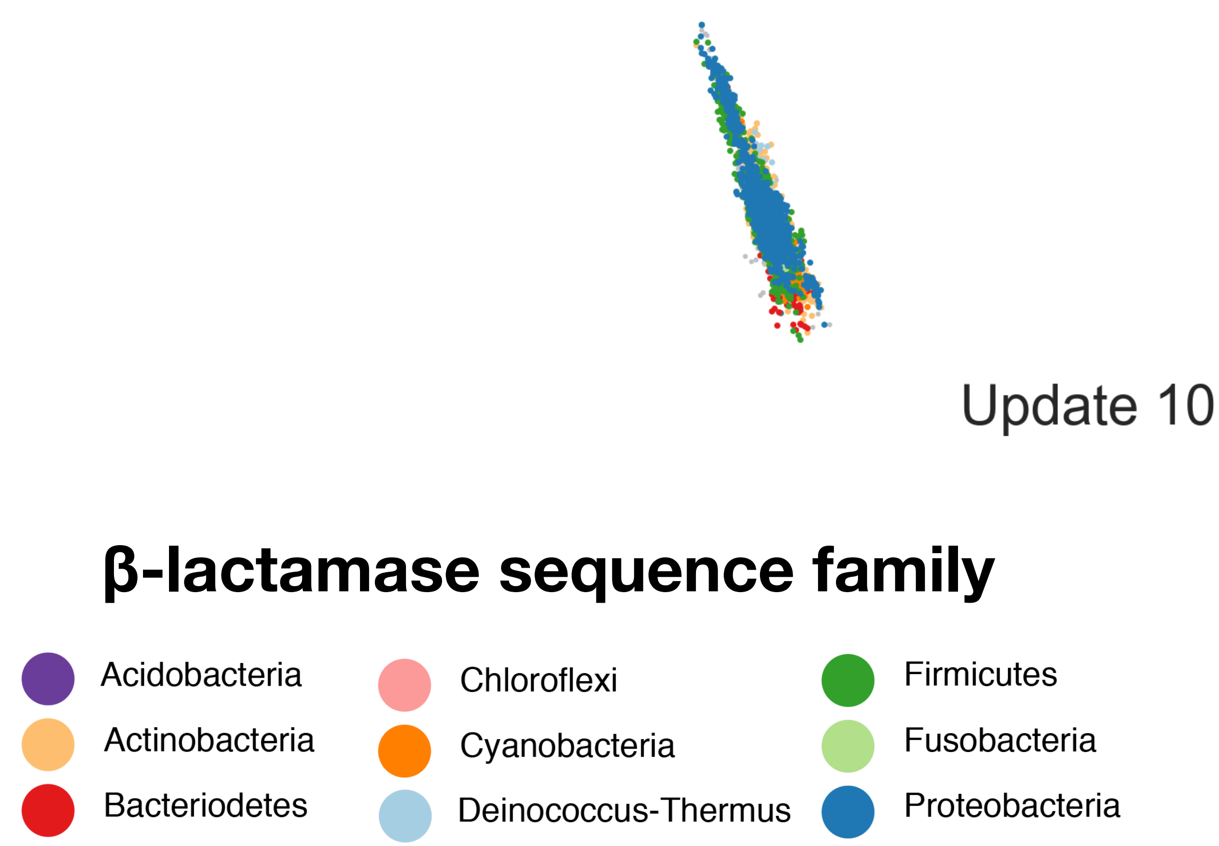
High-Throughput Data Acquisition + Machine Learning





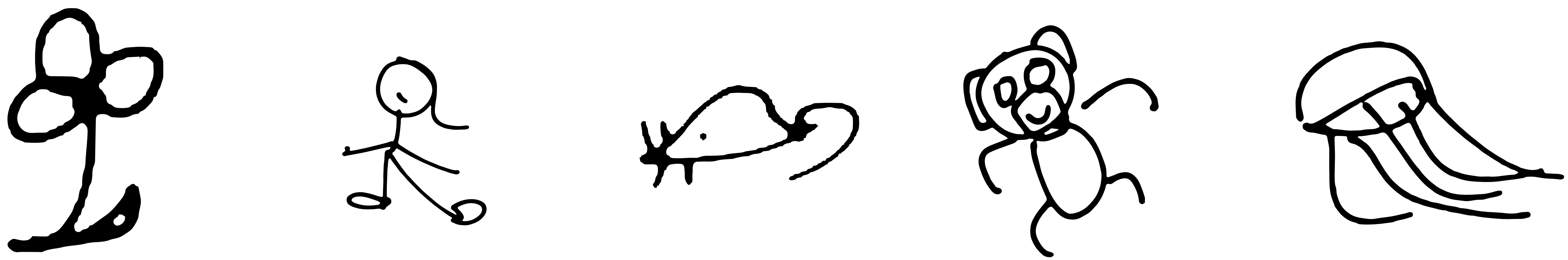
Riesselman, A.J.*, Ingraham, J.B.* and Marks, D.S., 2018. *Nat. Methods*, *15*, pp.816-822.

High-Throughput Data Acquisition + Machine Learning



Riesselman, A.J.*, Ingraham, J.B.* and Marks, D.S., 2018. *Nat. Methods*, *15*, pp.816-822.

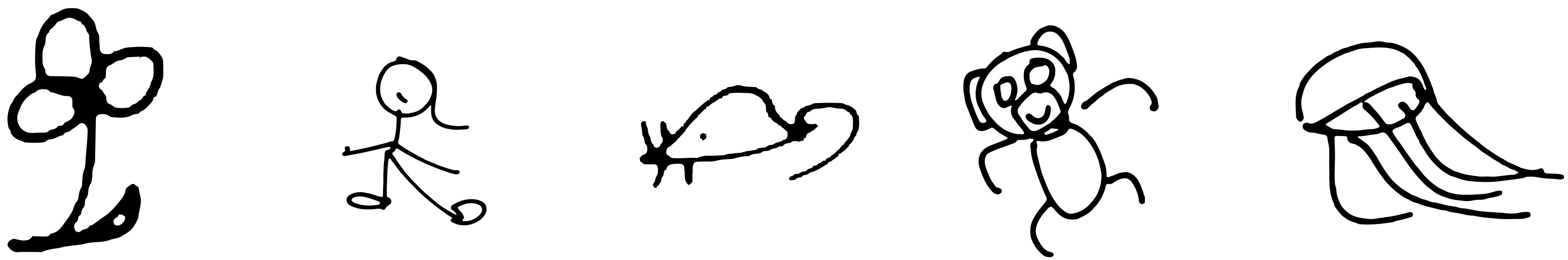
DNA sequencing is becoming very **cheap** and **easy**



https://www.uniprot.org/statistics/TrEMBL

https://www.youtube.com/watch?v=GUb1TZvMWsw

DNA sequencing is becoming very **cheap** and **easy**

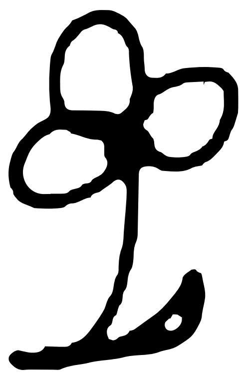
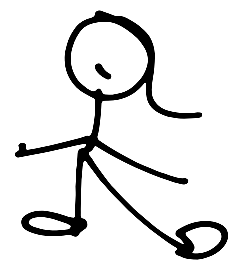
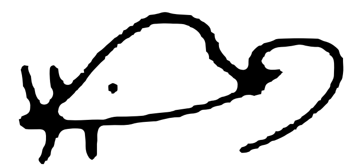


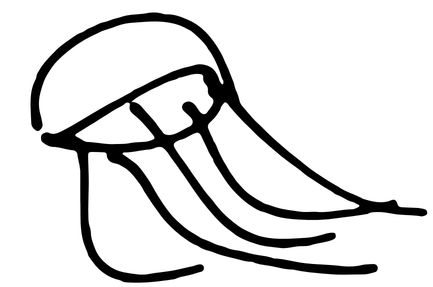
Isolate DNA

https://www.uniprot.org/statistics/TrEMBL

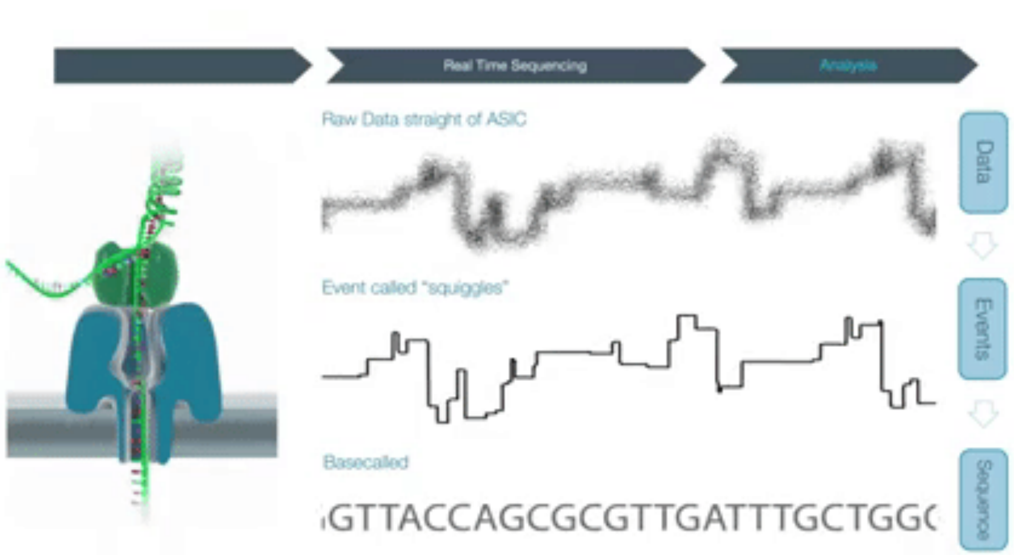
https://www.youtube.com/watch?v=GUb1TZvMWsw

DNA sequencing is becoming very **cheap** and **easy**



Isolate DNA

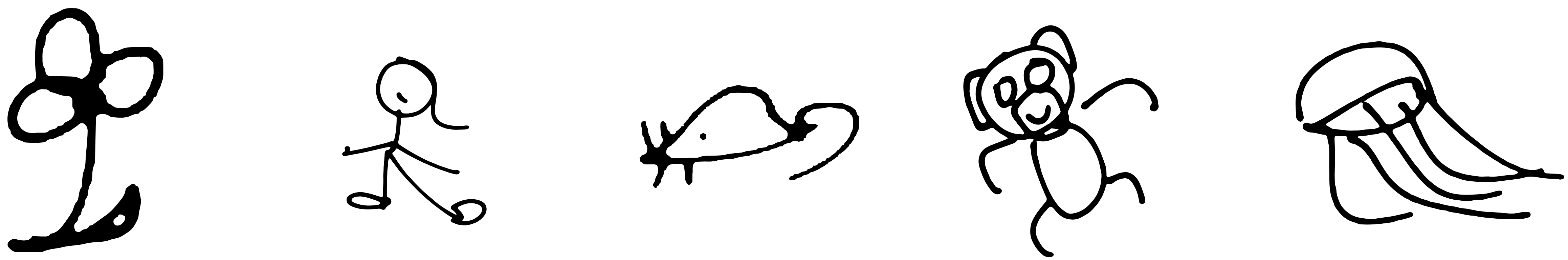


Sequence - Oxford Nanopore

https://www.uniprot.org/statistics/TrEMBL

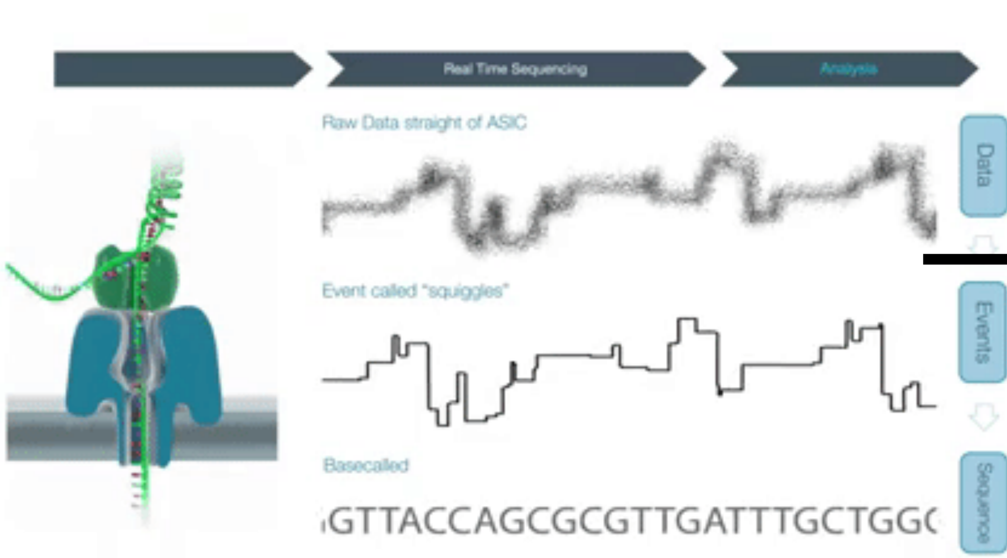
https://www.youtube.com/watch?v=GUb1TZvMWsw

DNA sequencing is becoming very **cheap** and **easy**



Isolate DNA

Deposit in public databases

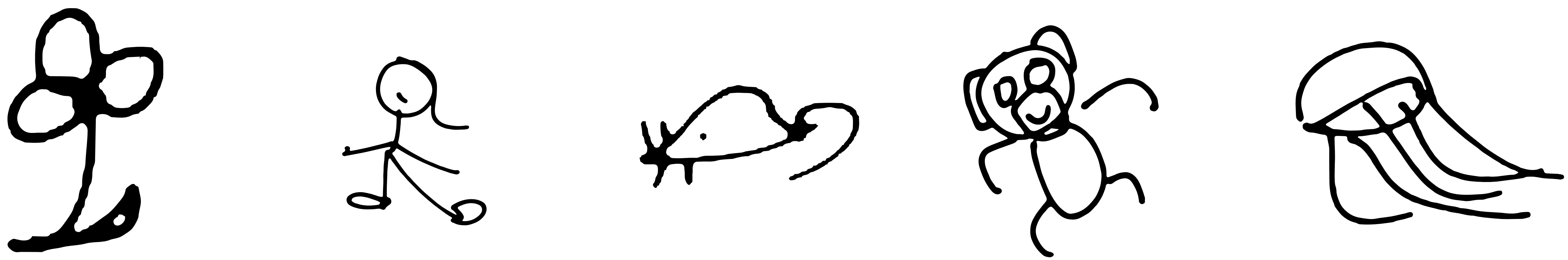


Sequence - Oxford Nanopore

https://www.uniprot.org/statistics/TrEMBL

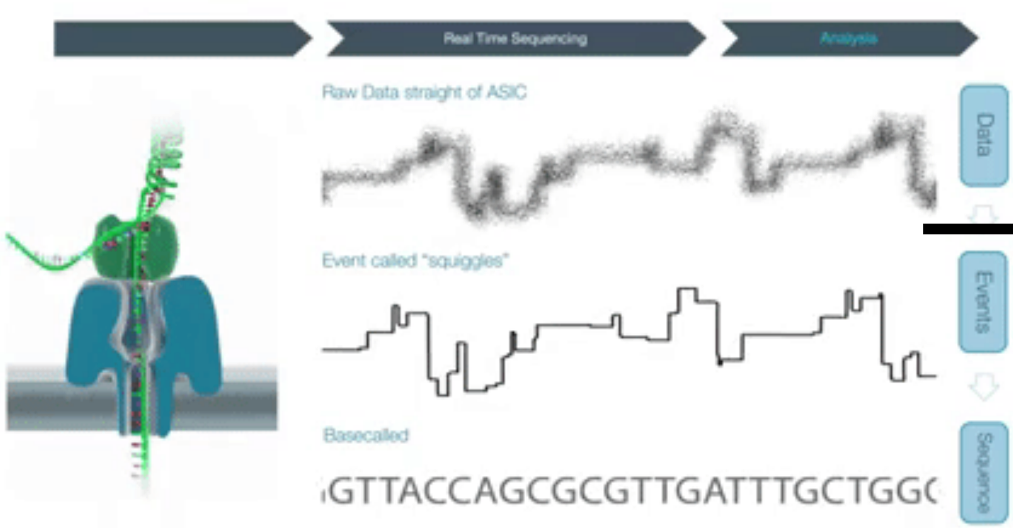
https://www.youtube.com/watch?v=GUb1TZvMWsw

DNA sequencing is becoming very **cheap** and **easy**

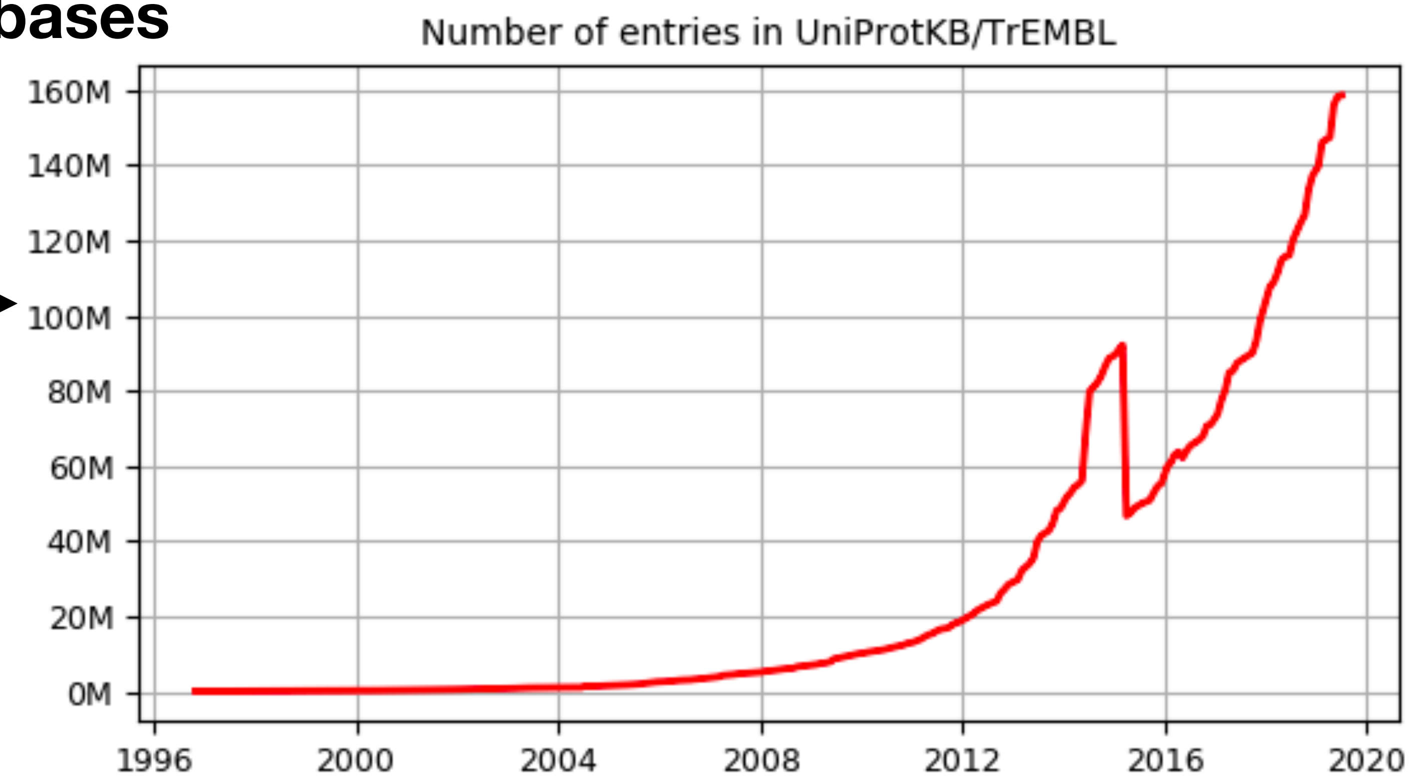


Isolate DNA

Deposit in public databases



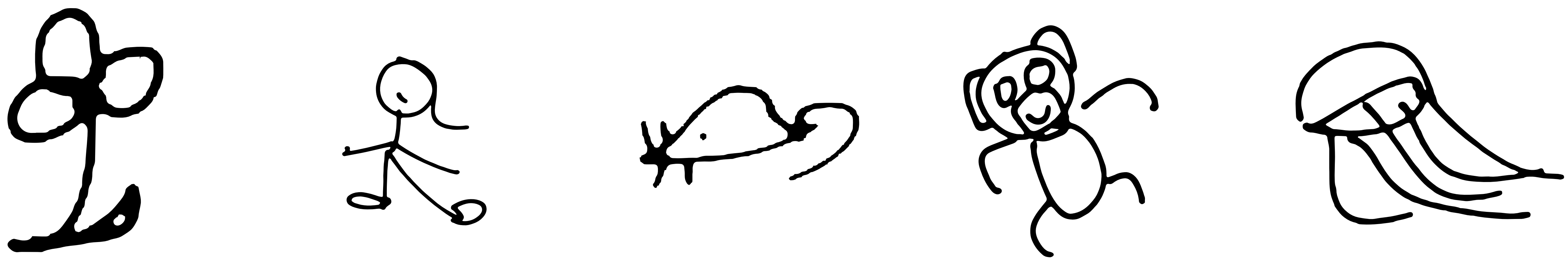
Sequence - Oxford Nanopore



https://www.uniprot.org/statistics/TrEMBL

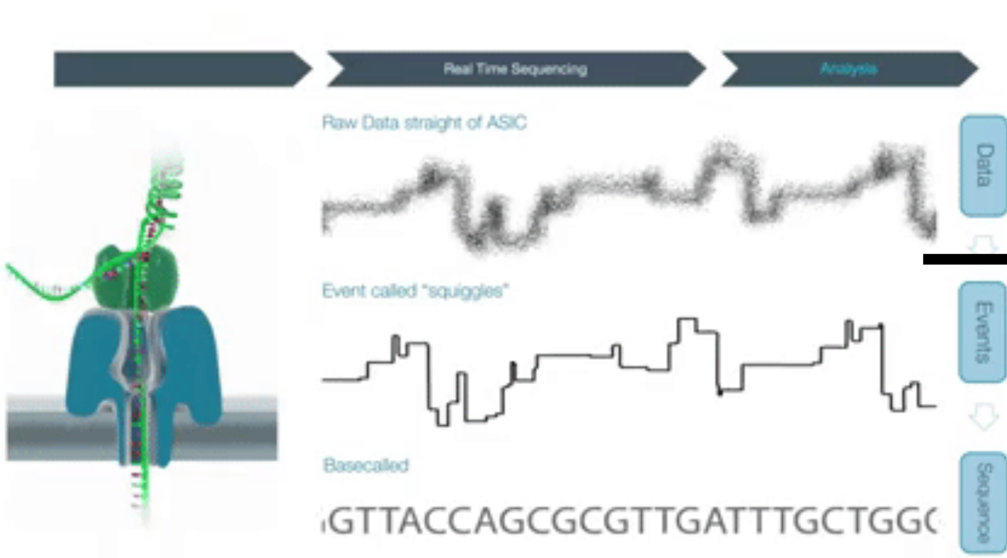
https://www.youtube.com/watch?v=GUb1TZvMWsw

DNA sequencing is becoming very **cheap** and **easy**

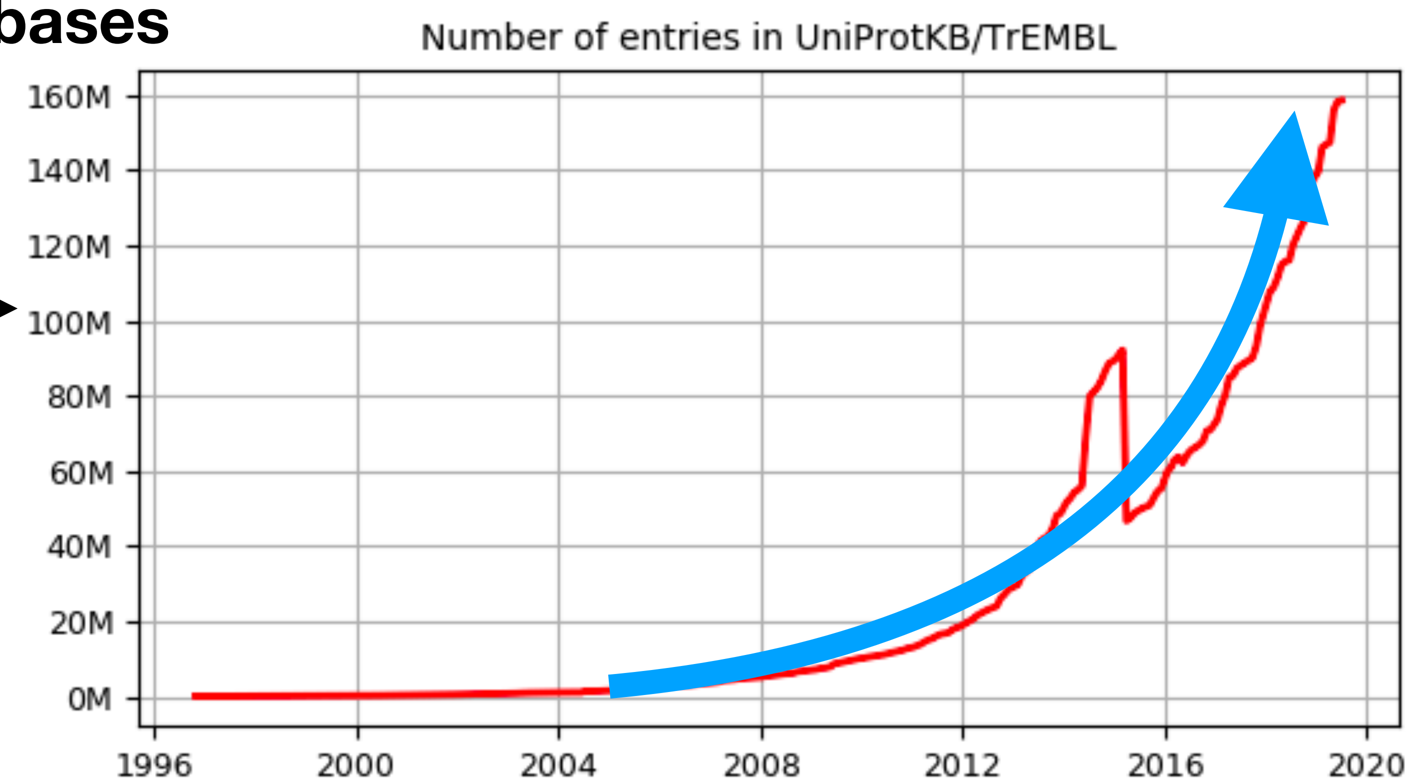


Isolate DNA

Deposit in public databases



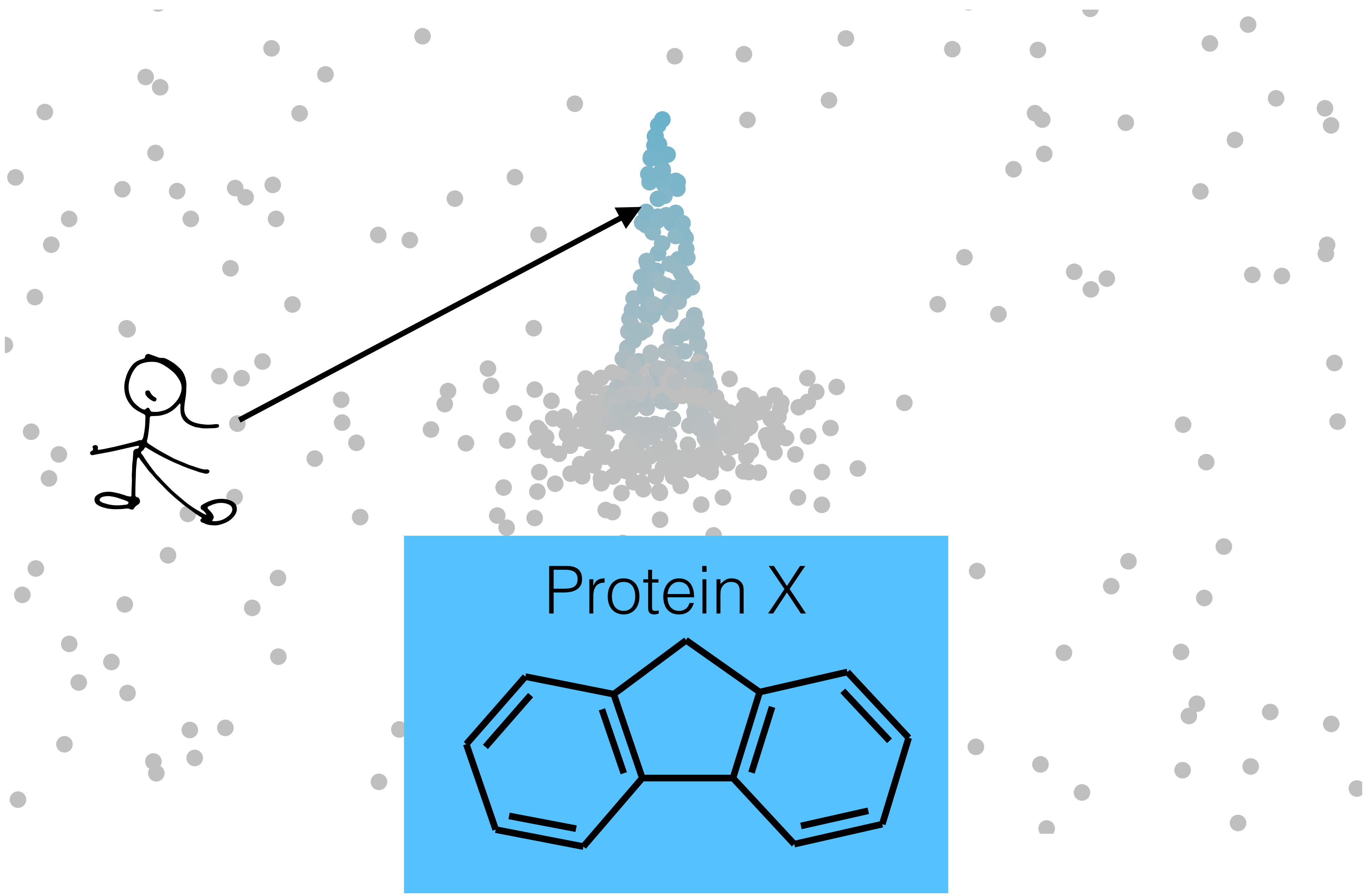
Sequence - Oxford Nanopore



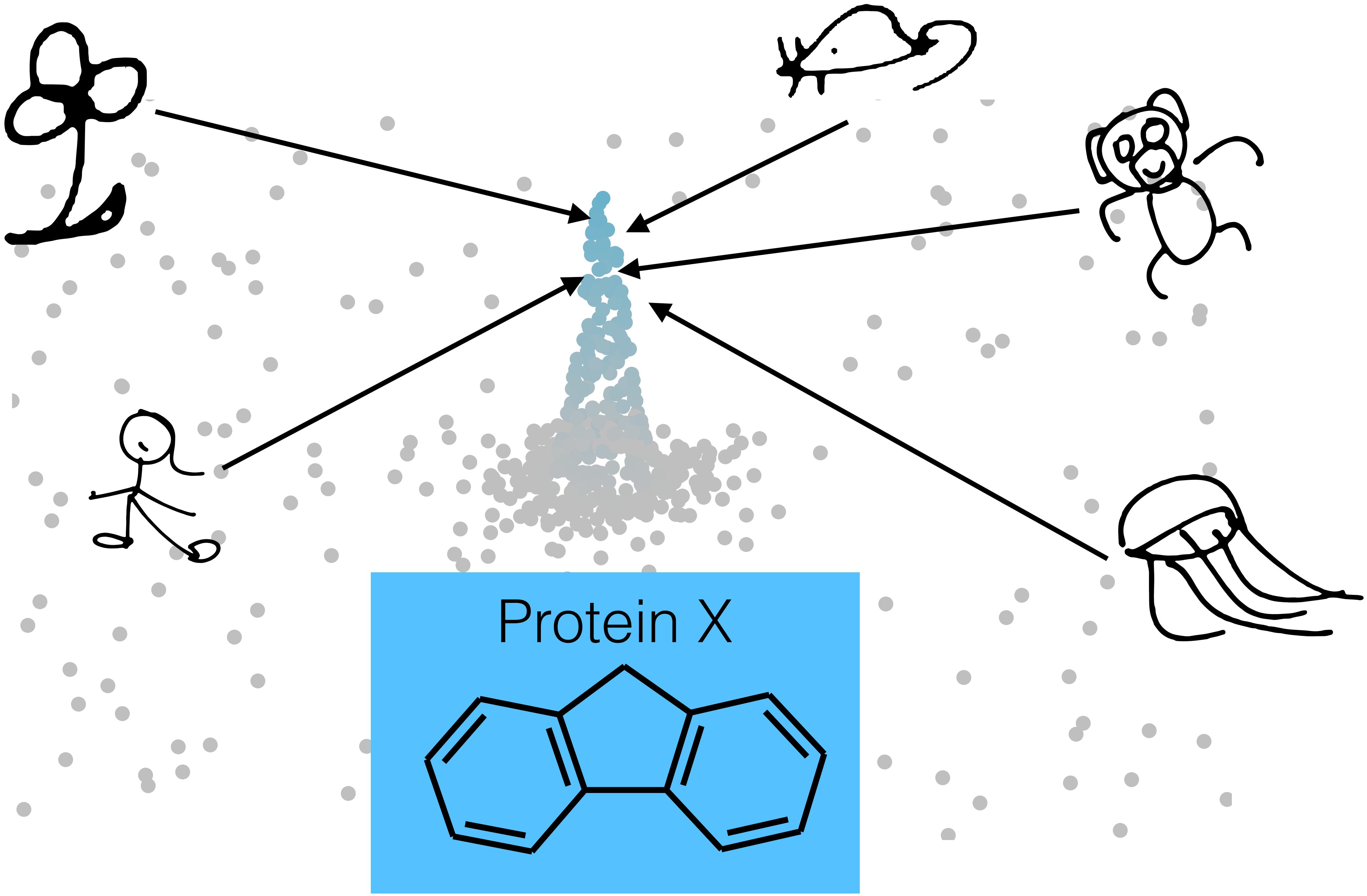
https://www.uniprot.org/statistics/TrEMBL

https://www.youtube.com/watch?v=GUb1TZvMWsw

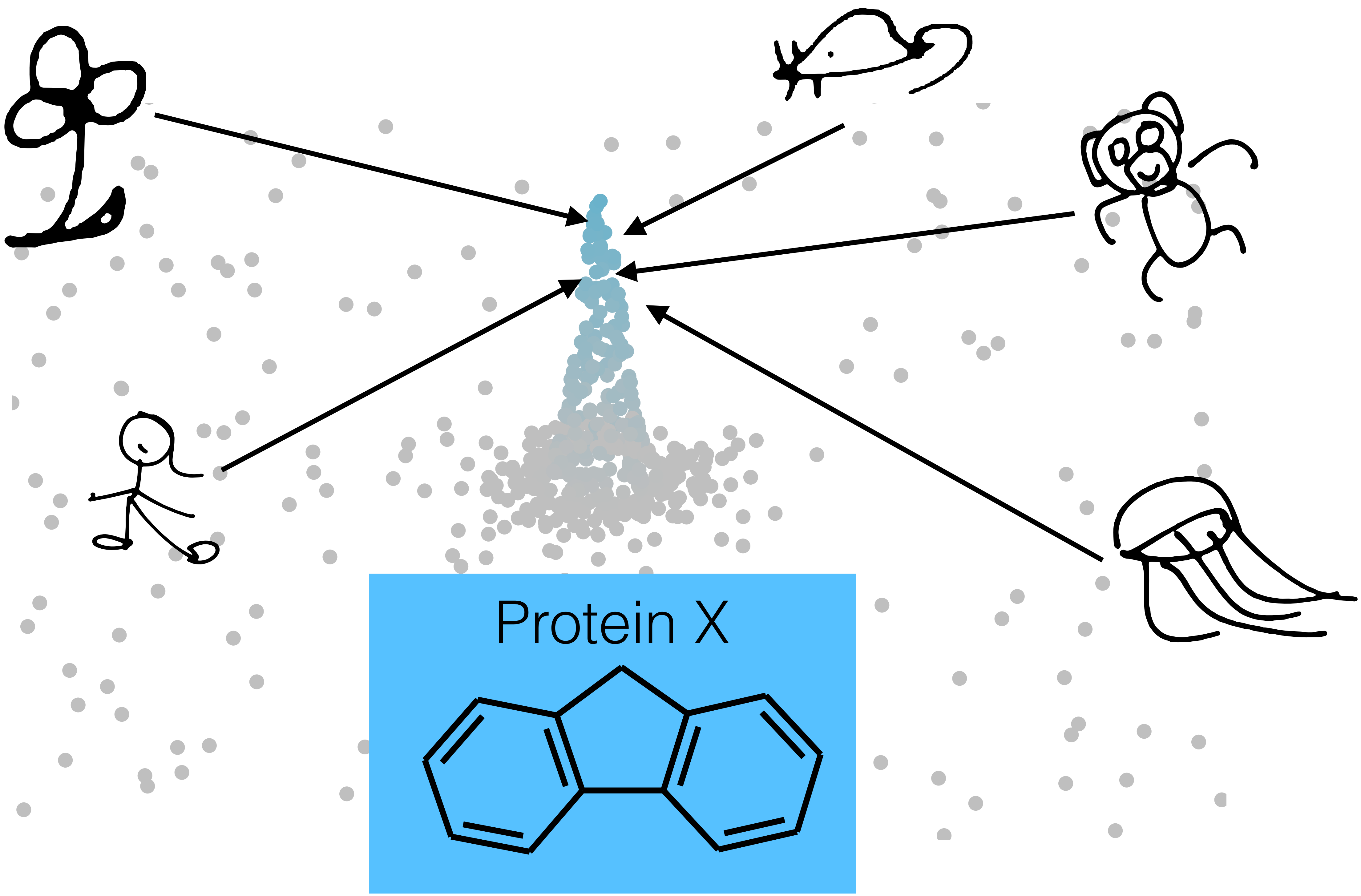
Lots of other **examples** of **Protein X** are **available**



Lots of other **examples** of **Protein X** are **available**



Lots of other **examples** of **Protein X** are **available**



All are **functional, homologous examples** of Protein X

Lots of other **examples** of **Protein X** are **available**

Sequences are found in public **genome databases**.

Lots of other **examples** of **Protein X** are **available**

Sequences are found in public **genome databases**.

JGI
JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

Natural **evolution** is an **experiment**, in **parallel**.

Lots of other **examples** of **Protein X** are **available**

Sequences are found in public **genome databases**.

Natural **evolution** is an **experiment**, in **parallel**.

<u>**Assumption:**</u>

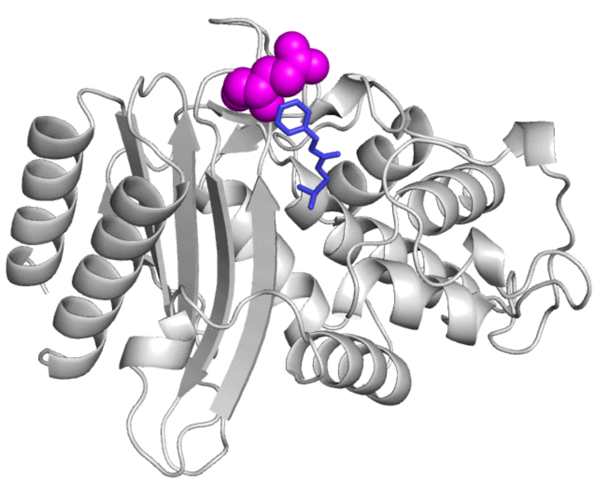
Present in database: Tolerated

Not in database: Deleterious

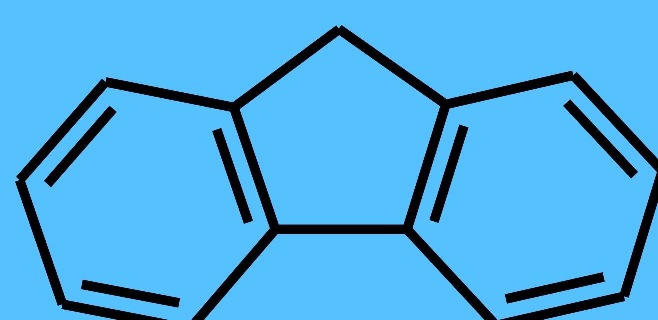
Can we **predict** which **mutation** will be **tolerable**?

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW



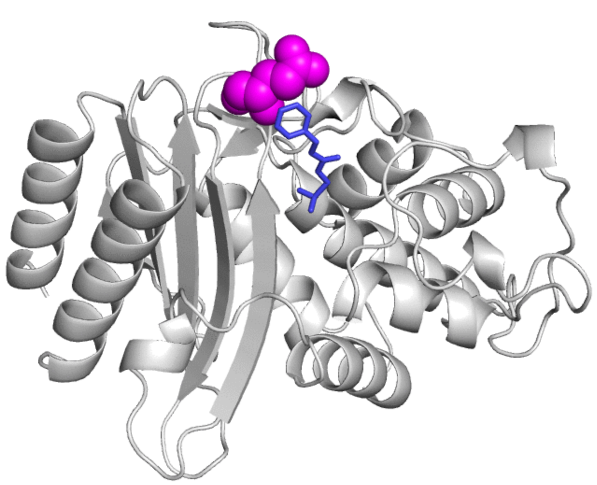
Protein X



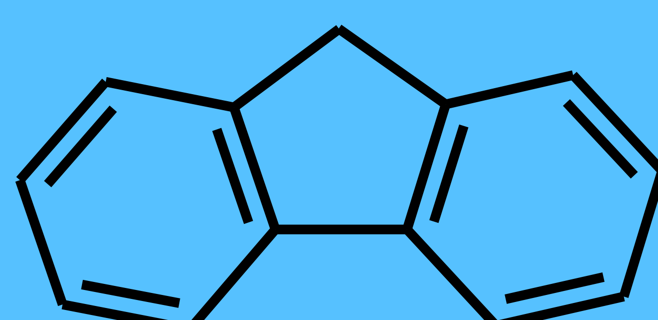
Can we **predict** which **mutation** will be **tolerable**?

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW



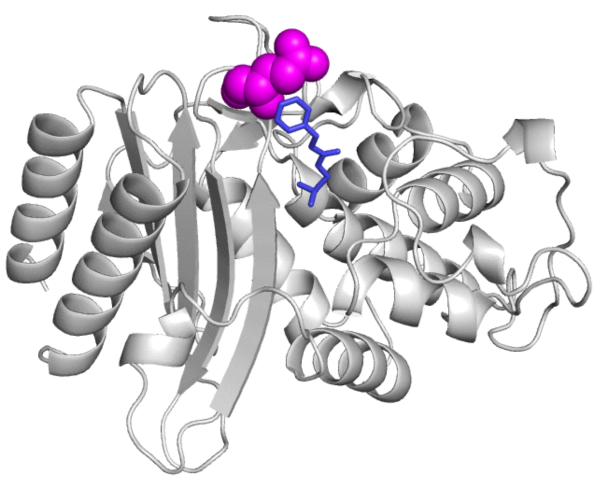
Protein X



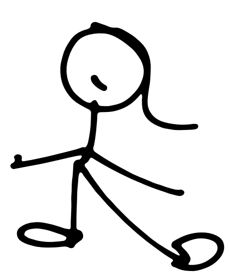
Can we **predict** which **mutation** will be **tolerable**?

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

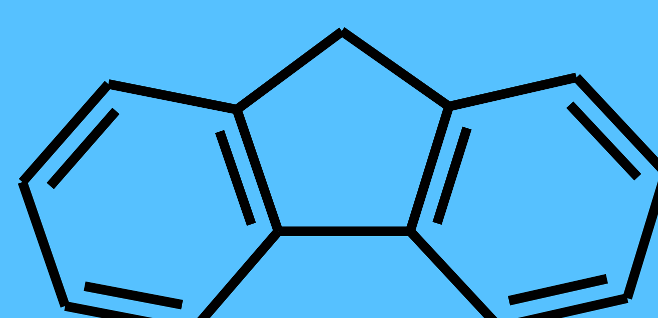
E→L




```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

Protein X



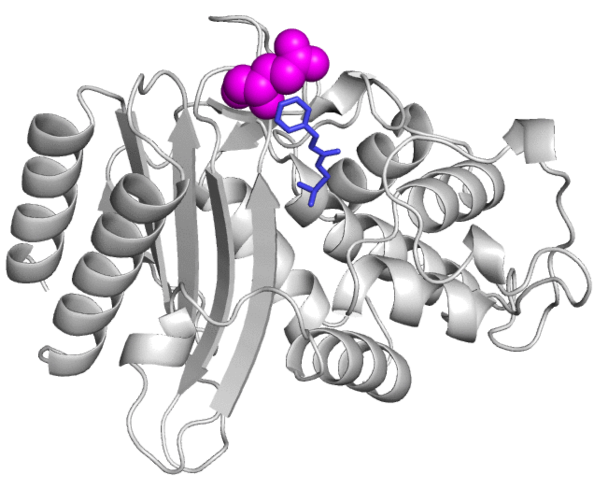
Can we **predict** which **mutation** will be **tolerable**?

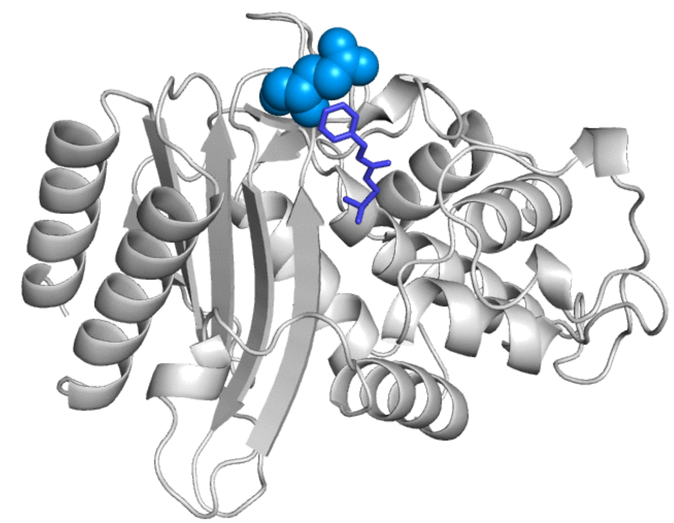
```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→L

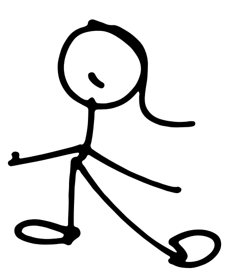
E→I

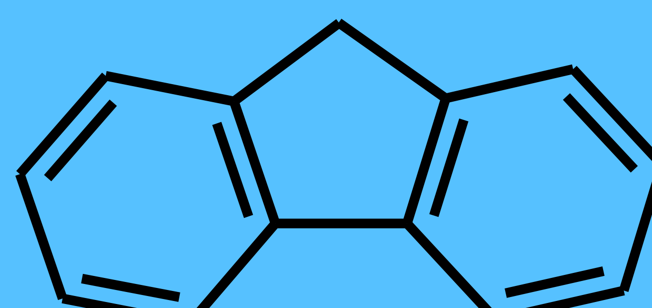

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

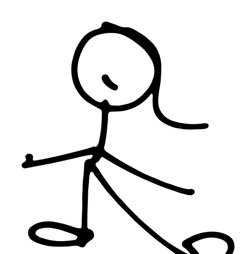





Protein X







Can we **predict** which **mutation** will be **tolerable**?

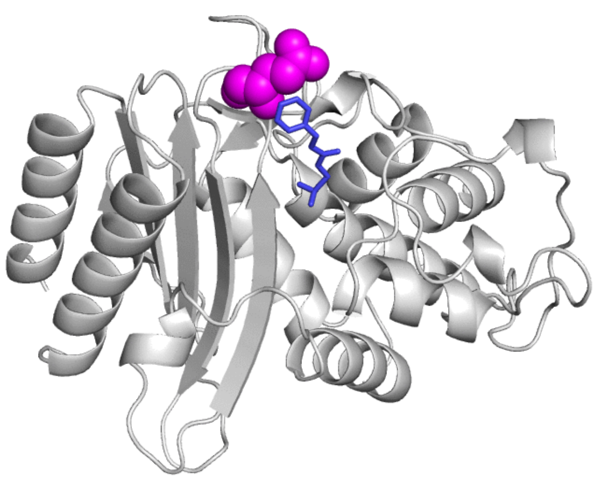
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

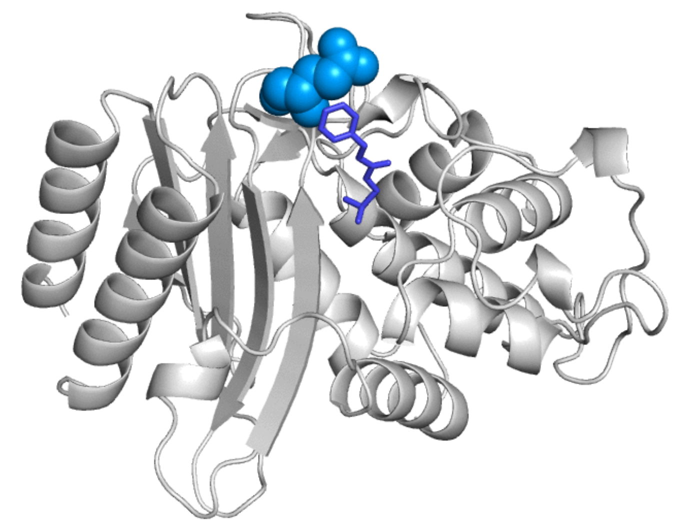
E→L

E→I

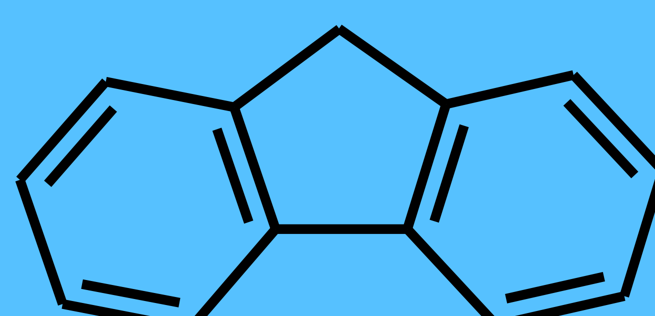
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

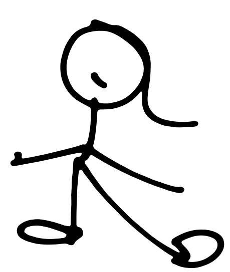
? ? ?

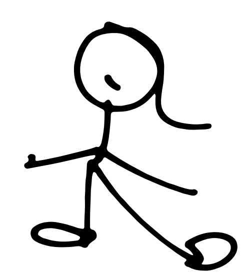




Protein X



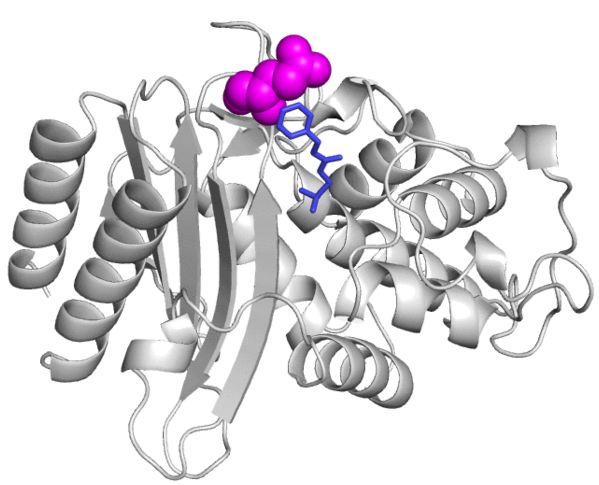




Can we **predict** which **mutation** will be **tolerable**?

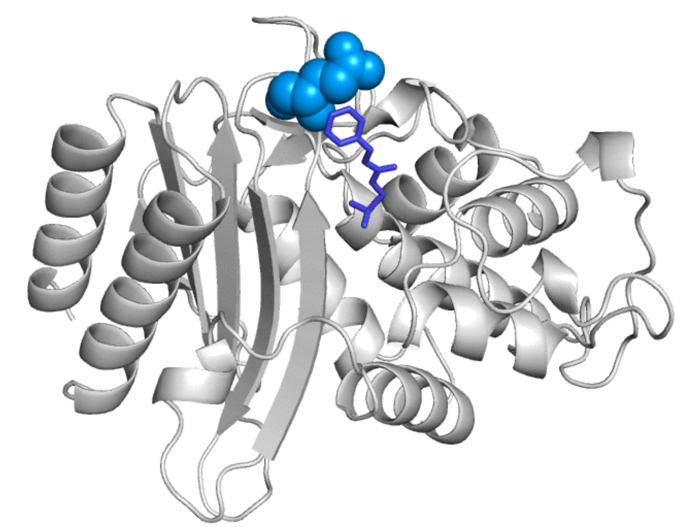
```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

E→L



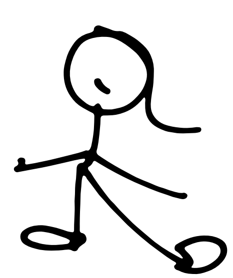
E→I

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

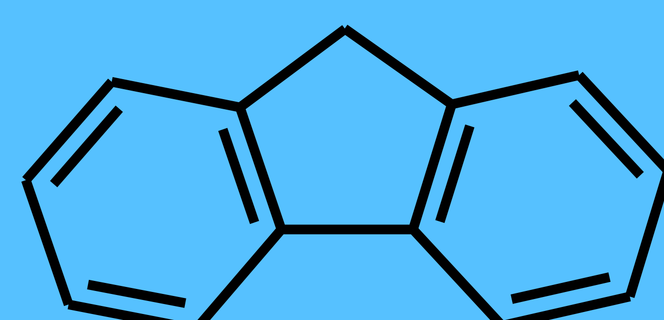


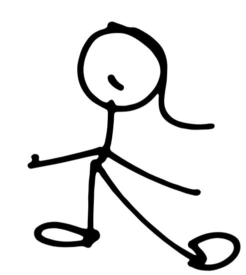
? ? ?

How can we formulate this problem?



Protein X





A **generative model** finds **probable** “words” based on **context**

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the ____.

A **generative model** finds **probable** “words” based on **context**

English sentence

banana

We are going to go to the ____.

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the ____.

banana ✘

running ✘

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the _____.

banana

running

zoo

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the zoo.

P(x) finds **probable** “words” based on **context**

A **generative model** finds **probable** "words" based on **context**

English sentence

We are going to go to the zoo.

Protein sequence

AQKLYLTHIDAEVEGADTLFITEVKQVF __

P(x) finds **probable** "words" based on **context**

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the zoo.

Protein sequence

AQKLYLTHIDAEVEGADTLFITEVKQVF __

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the zoo.

Protein sequence

AQKLYLTHIDAEVEGADTLFITEVKQVF __

H

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the zoo.

Protein sequence

AQKLYLTHIDAEVEGADTLFITEVKQVF __

H ✗

V ✗

A **generative model** finds **probable** “words” based on **context**

English sentence

We are going to go to the zoo.

Protein sequence

AQKLYLTHIDAEVEGADTLFITEVKQVF __

H

V

T

Neural networks power **generative models**

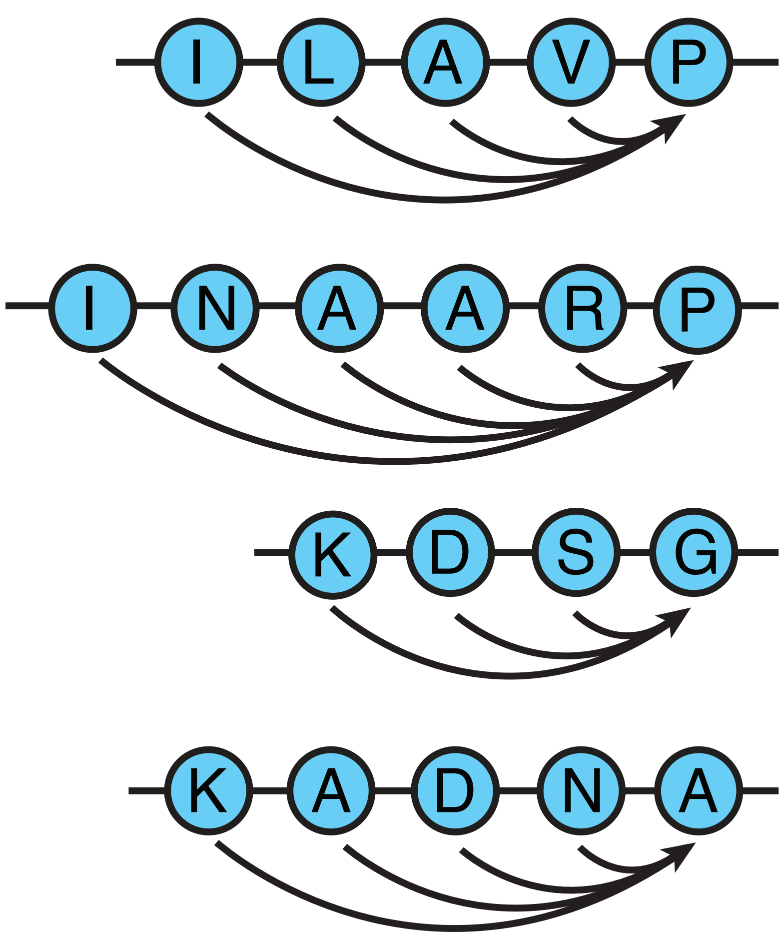
$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$

We are going to go to the zoo.

Neural networks power **generative models**

$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$

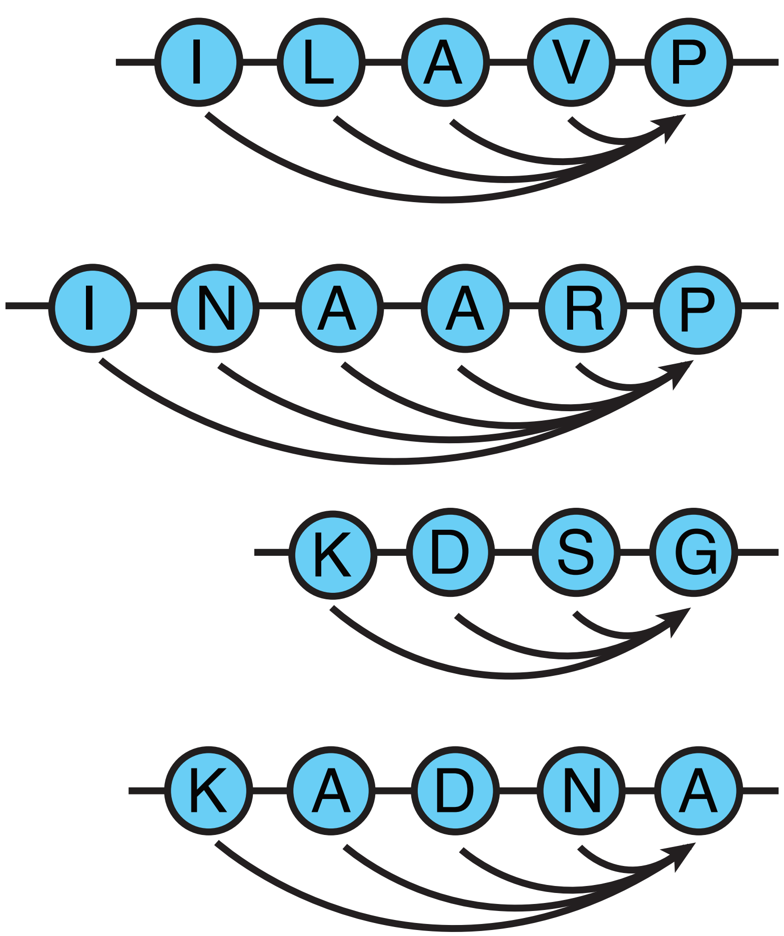
Autoregressive model



Neural networks power **generative models**

$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$

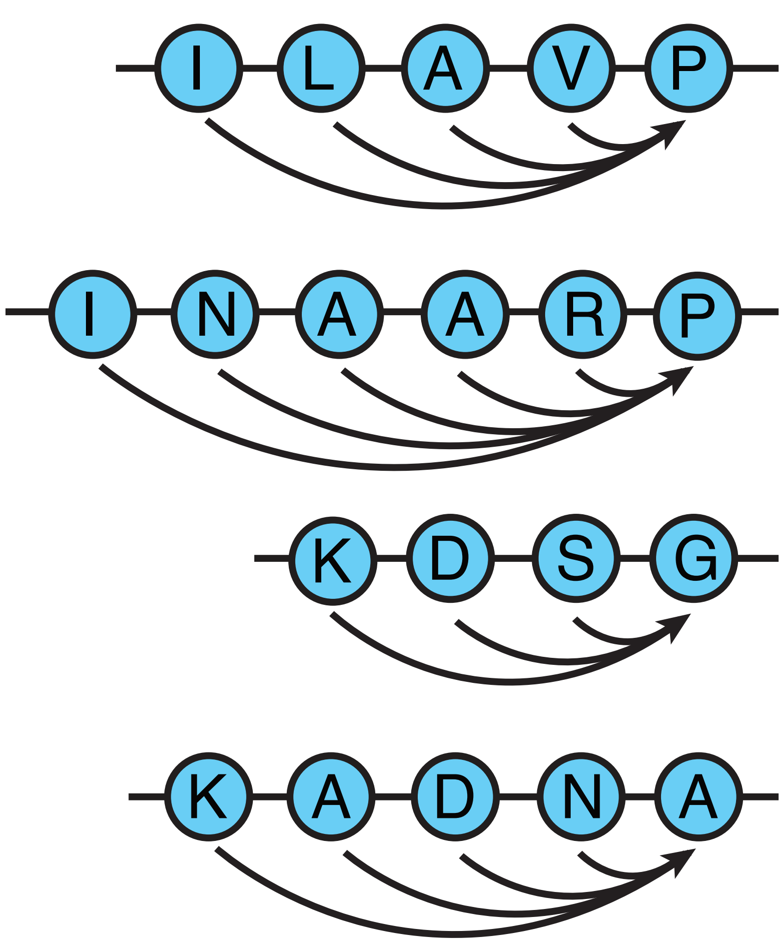
Autoregressive model

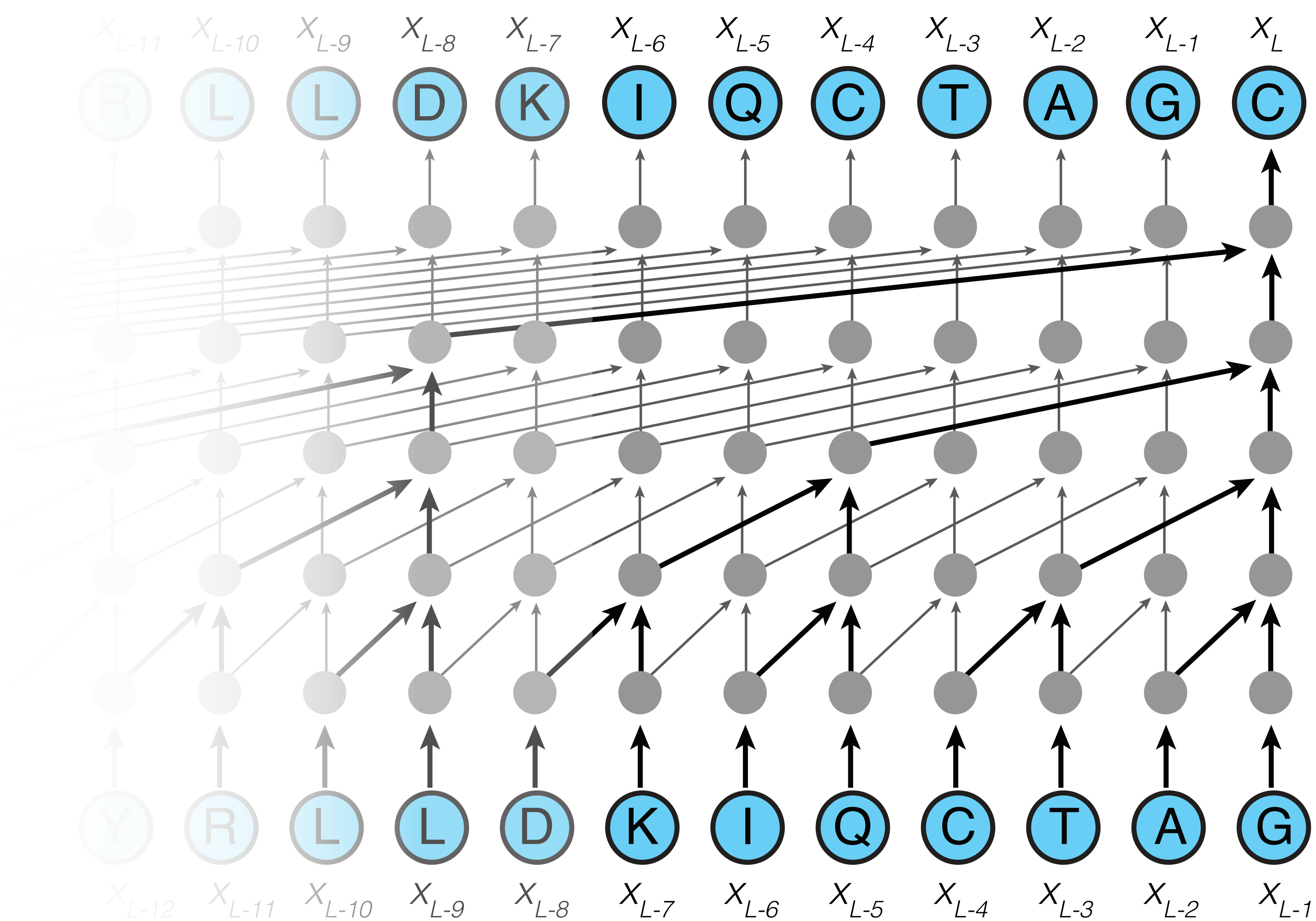


Neural networks power **generative models**

$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$

Autoregressive model



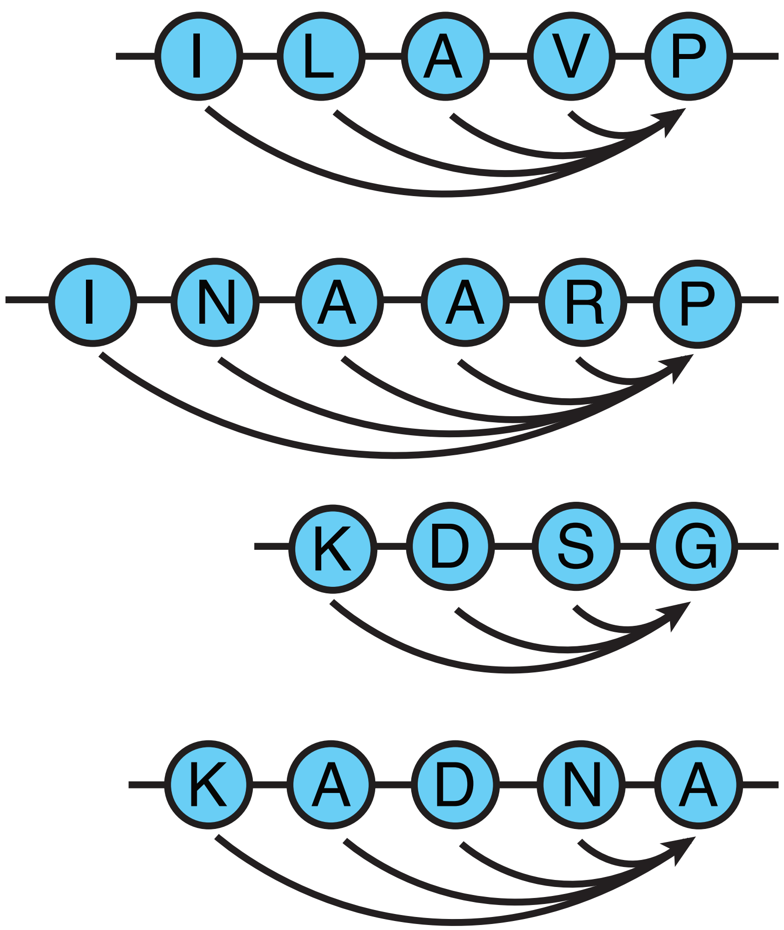


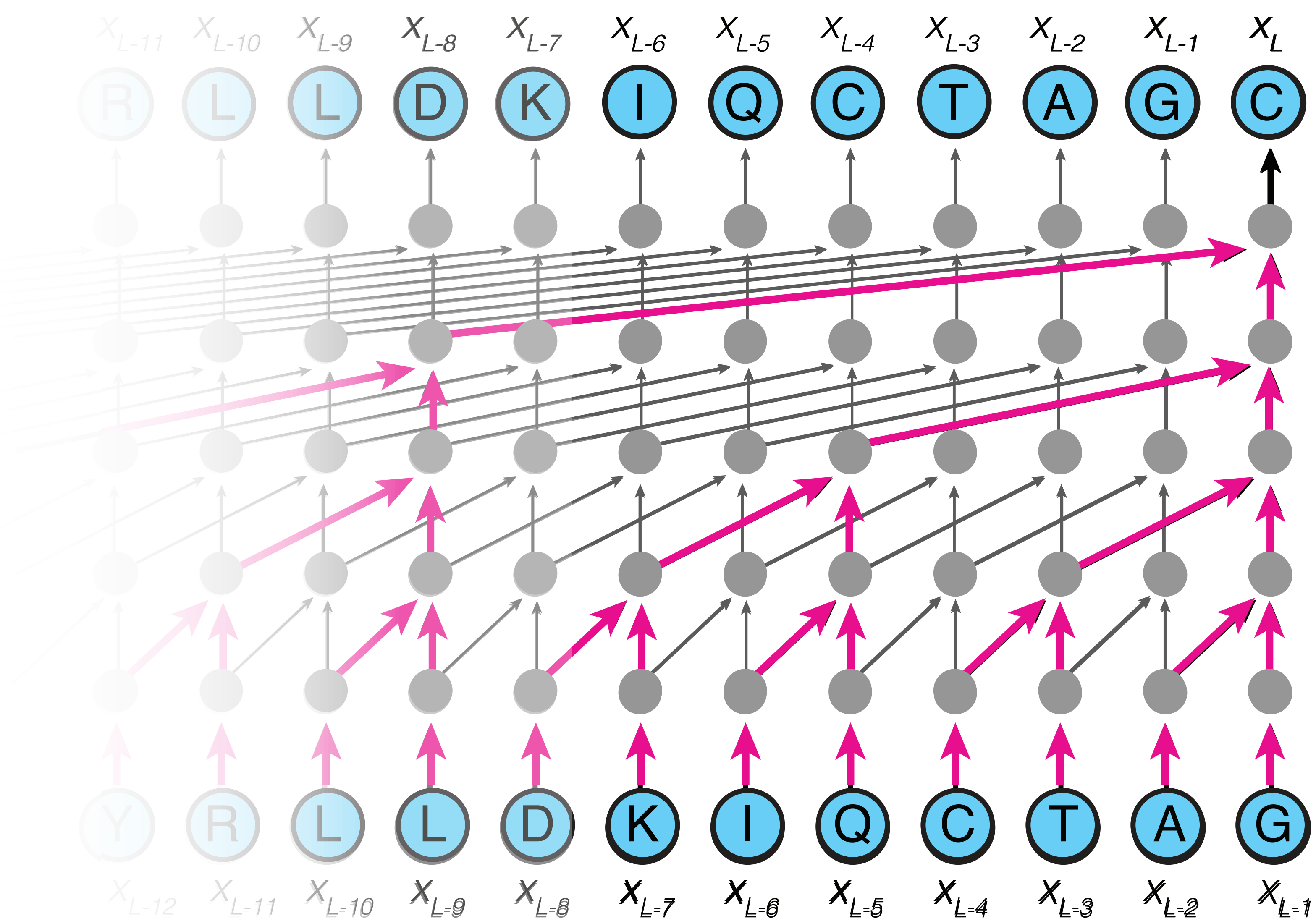
Dilated convolutional neural network

Neural networks power **generative models**

$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$

Autoregressive model

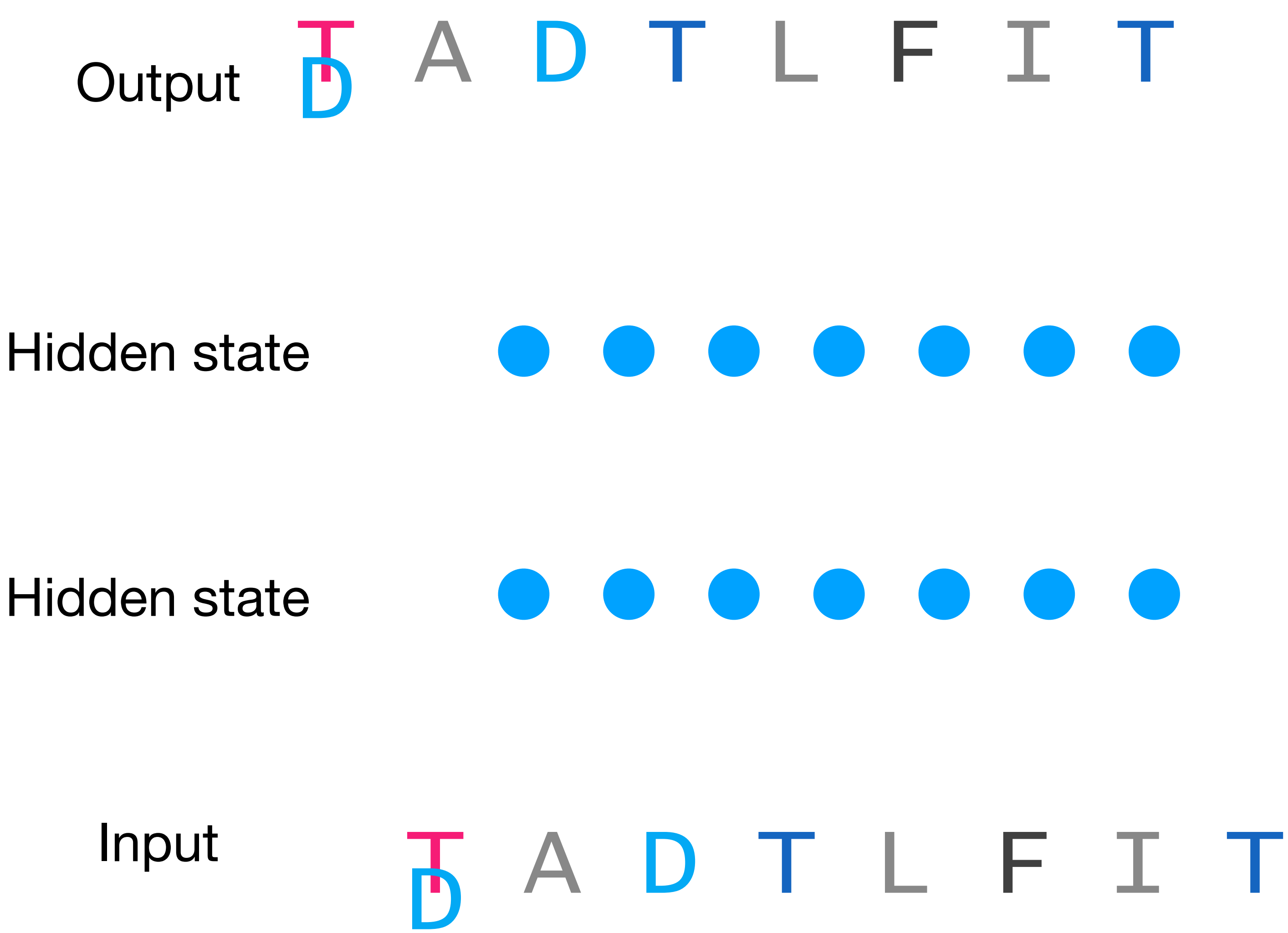




Dilated convolutional neural network

Utilizing an autoregressive likelihood

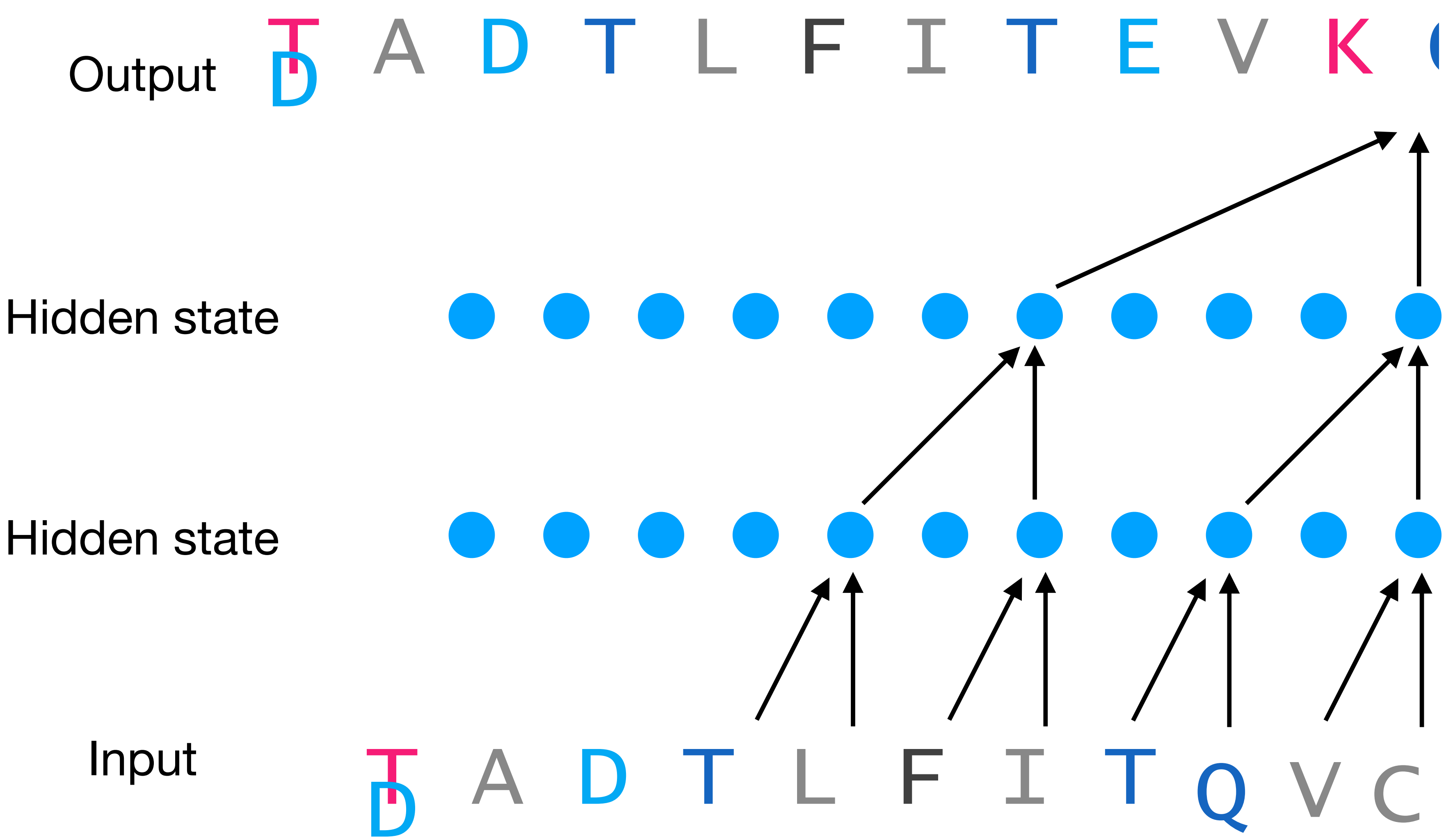
$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$



Adapted from:
https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Utilizing an autoregressive likelihood

$$p(x|\theta) = \prod_{i}^{L} p(x_i|x_{<i}, \theta)$$



Adapted from:
https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Mutation effect prediction
with an unsupervised model

Mutation effect prediction with an unsupervised model

1) Infer a **generative model** of the family

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

$$p(\mathbf{x}|\boldsymbol{\theta})$$

Mutation effect prediction with an unsupervised model

1) Infer a **generative model** of the family

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

$$\downarrow$$

$$p(\mathbf{x}|\boldsymbol{\theta})$$

2) Compute **Log Ratio** for each mutant

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

Mutation effect prediction with an unsupervised model

1) Infer a **generative model** of the family

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

$$\downarrow$$

$$p(\mathbf{x}|\boldsymbol{\theta})$$

2) Compute **Log Ratio** for each mutant

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

"How much does this mutation look like what we've seen in nature?"

Mutation effect prediction with an unsupervised model

1) Infer a **generative model** of the family

2) Compute **Log Ratio** for each mutant

Uses public data (effectively free)

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD

$$p(\mathbf{x}|\theta)$$

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\theta)}{p(\mathbf{x}_{\text{wild}}|\theta)}$$

“How much does this mutation look like what we’ve seen in nature?”

Mutation effect prediction with an unsupervised model

1) Infer a **generative model** of the family

Uses public data (effectively free)

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

$$p(\mathbf{x}|\theta)$$

2) Compute **Log Ratio** for each mutant

Fast

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\theta)}{p(\mathbf{x}_{\text{wild}}|\theta)}$$

"How much does this mutation look like what we've seen in nature?"

Mutation effect prediction
with an unsupervised model

Uses public data (effectively free)

Fast

Works on almost any protein

1) Infer a **generative model** of the family

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

$p(\mathbf{x}|\theta)$

2) Compute **Log Ratio** for each mutant

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\theta)}{p(\mathbf{x}_{\text{wild}}|\theta)}$$

“How much does this mutation look like what we’ve seen in nature?”

Mutation effect prediction
with an unsupervised model

1) Infer a **generative model** of the family

2) Compute **Log Ratio** for each mutant

Uses public data (effectively free)

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
```

Fast

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\theta)}{p(\mathbf{x}_{\text{wild}}|\theta)}$$

Works on almost any protein

$$p(\mathbf{x}|\theta)$$

Accurate

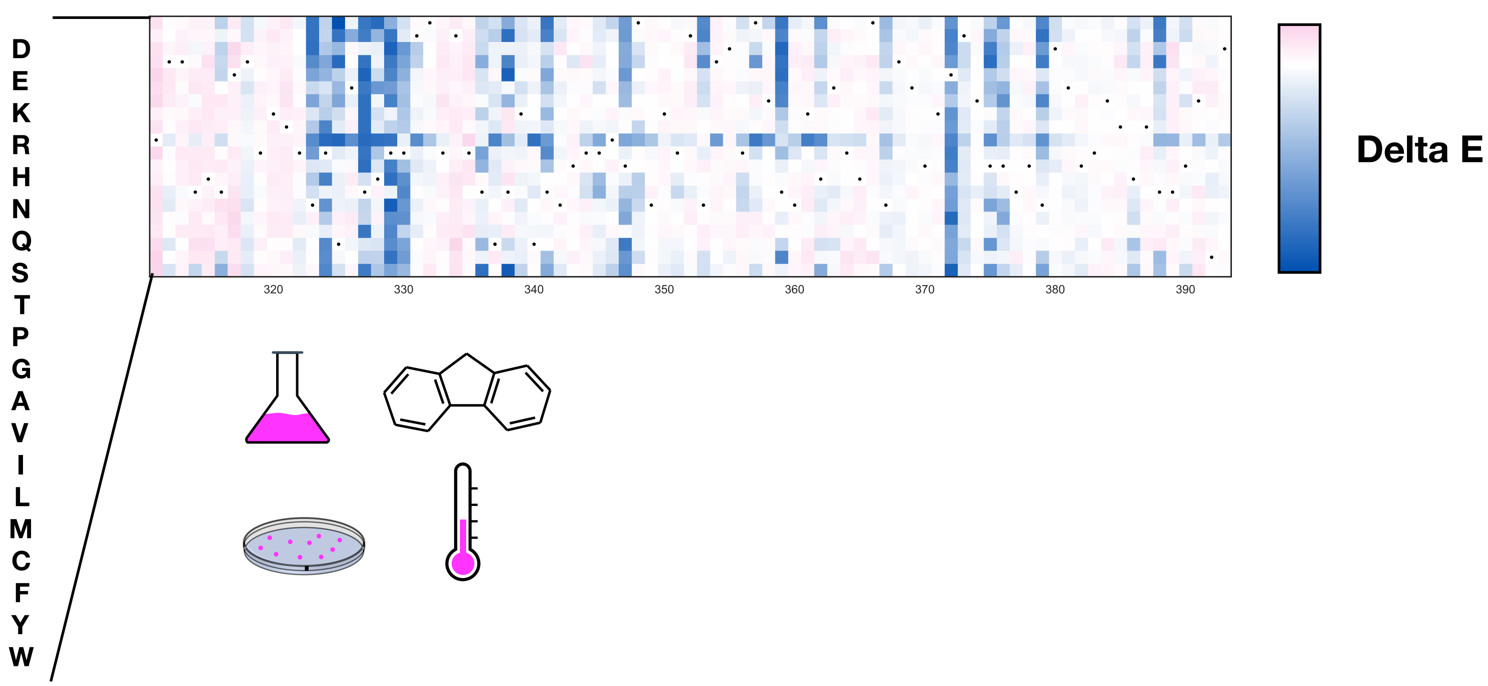
"How much does this mutation look like what we've seen in nature?"

To evaluate performance, we collected ~30 **saturation mutagenesis experiments**

McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

To evaluate performance, we collected ~30 **saturation mutagenesis experiments**

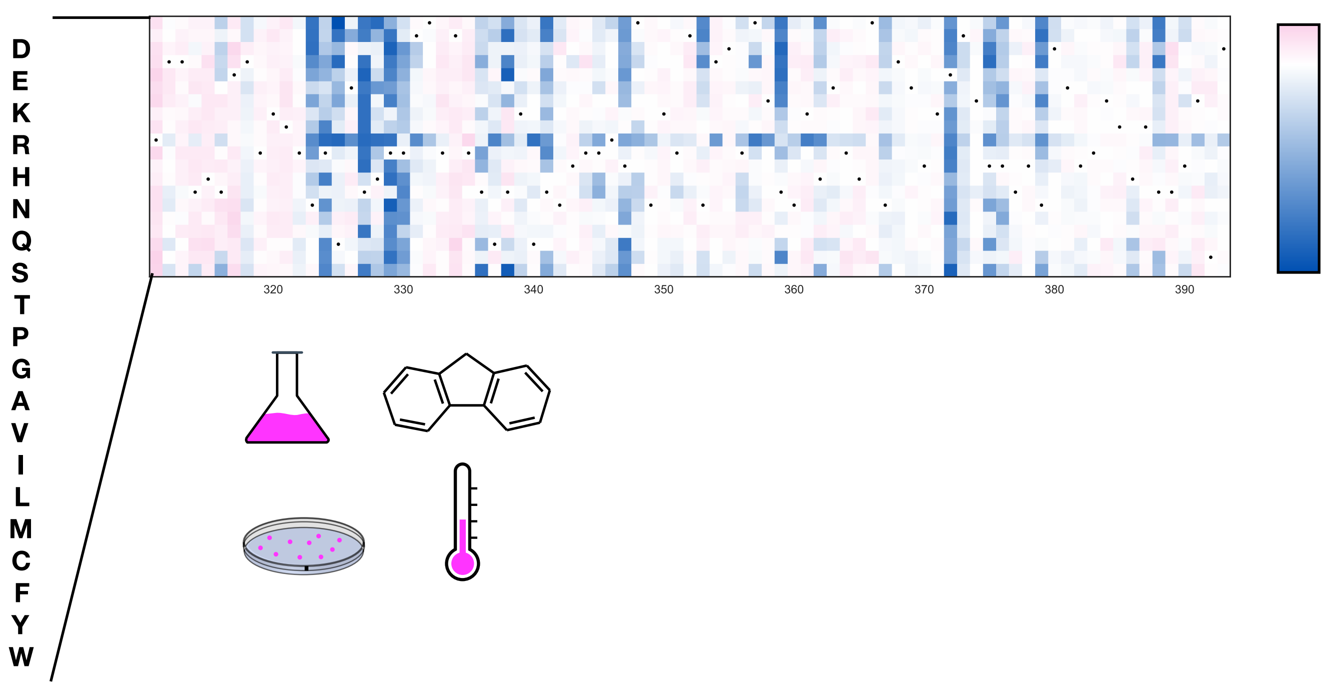
PDZ domain binding to ligand



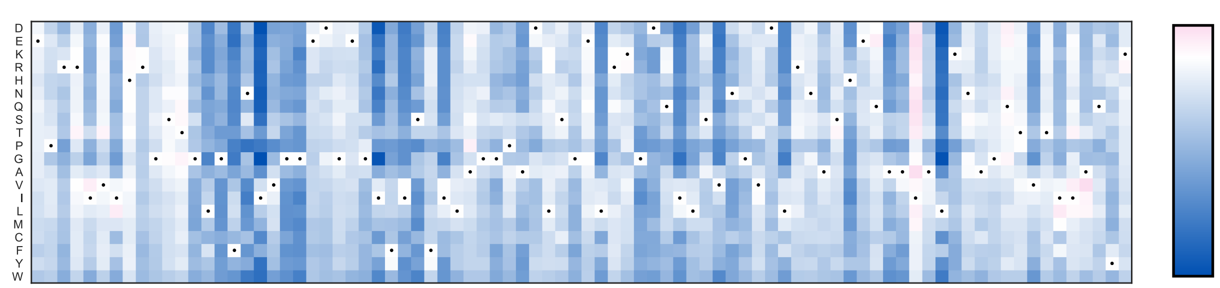
McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

To evaluate performance, we collected ~30 **saturation mutagenesis experiments**

PDZ domain binding to ligand



Log probability ratio

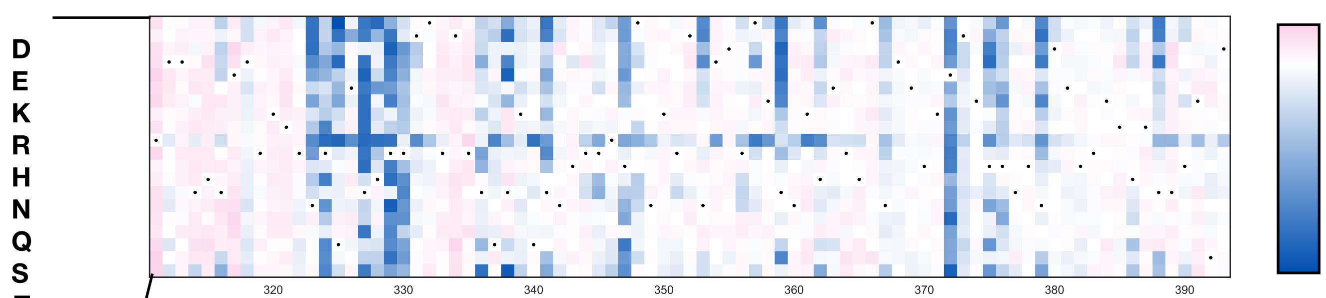


$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

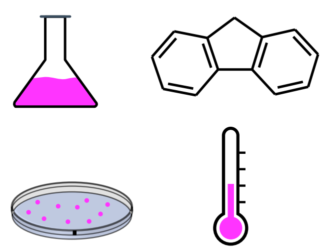
McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

To evaluate performance, we collected ~30 **saturation mutagenesis experiments**

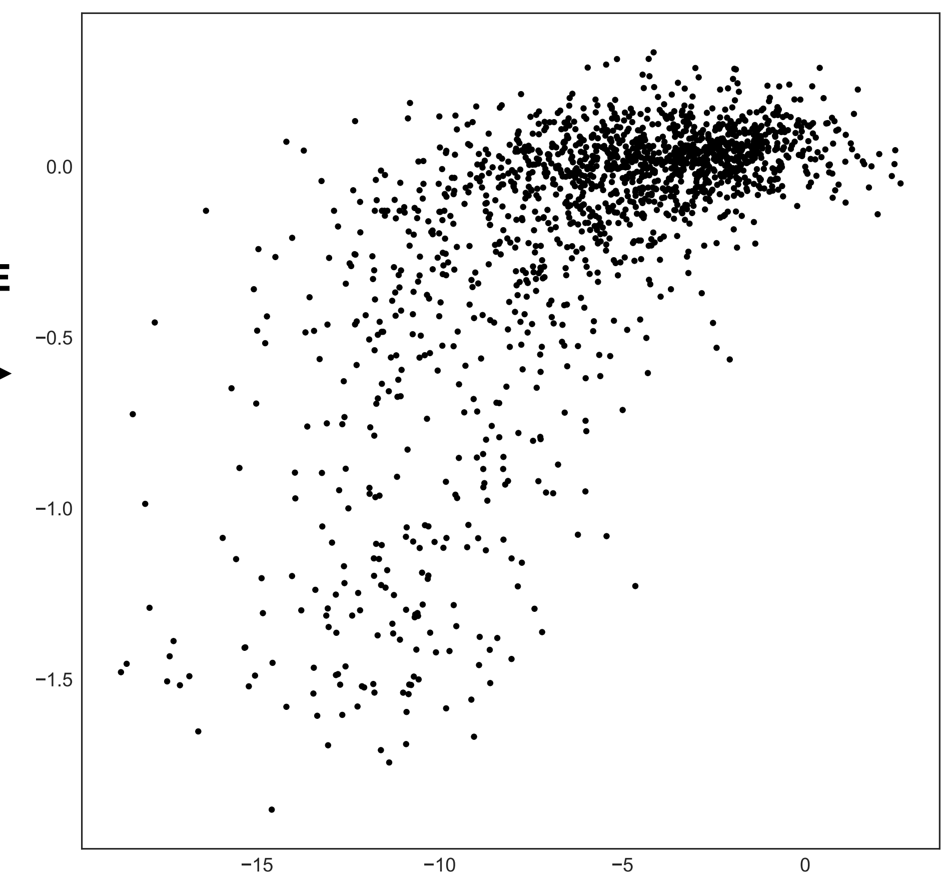
PDZ domain binding to ligand



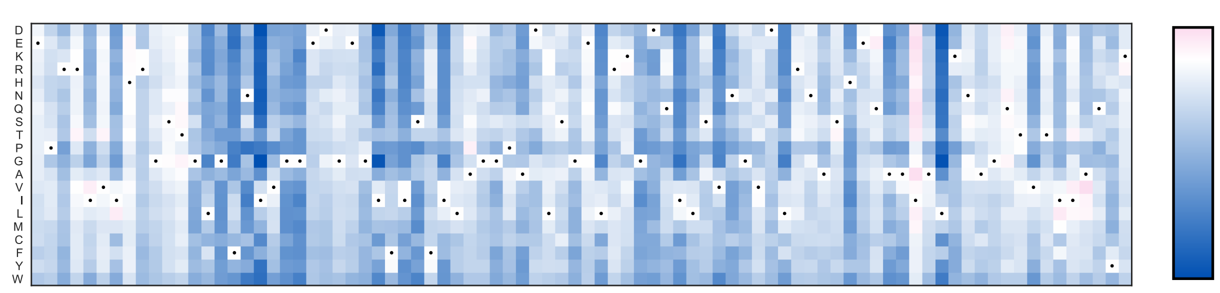
Delta E



Compare



Log probability ratio

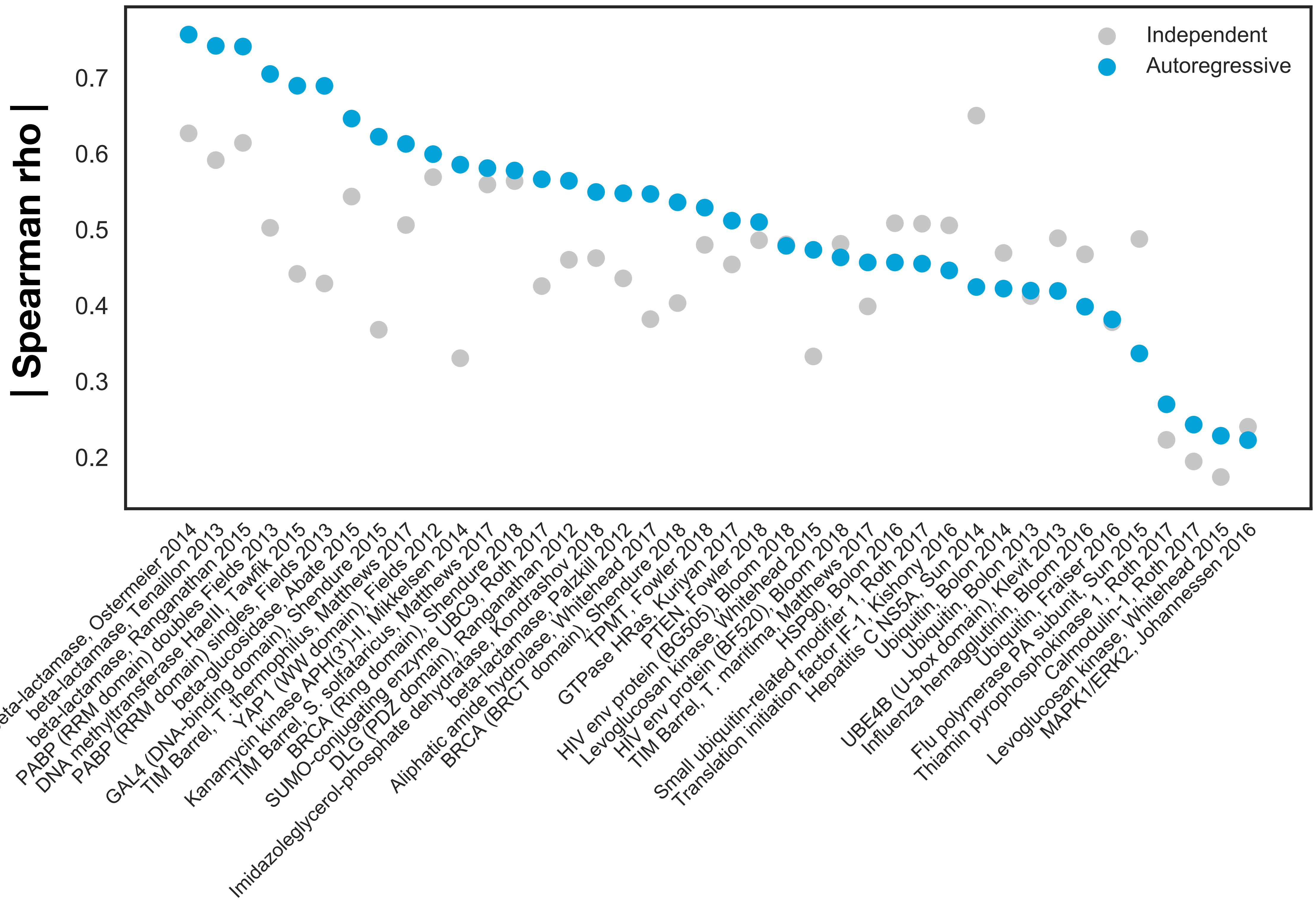


Delta Log Probability

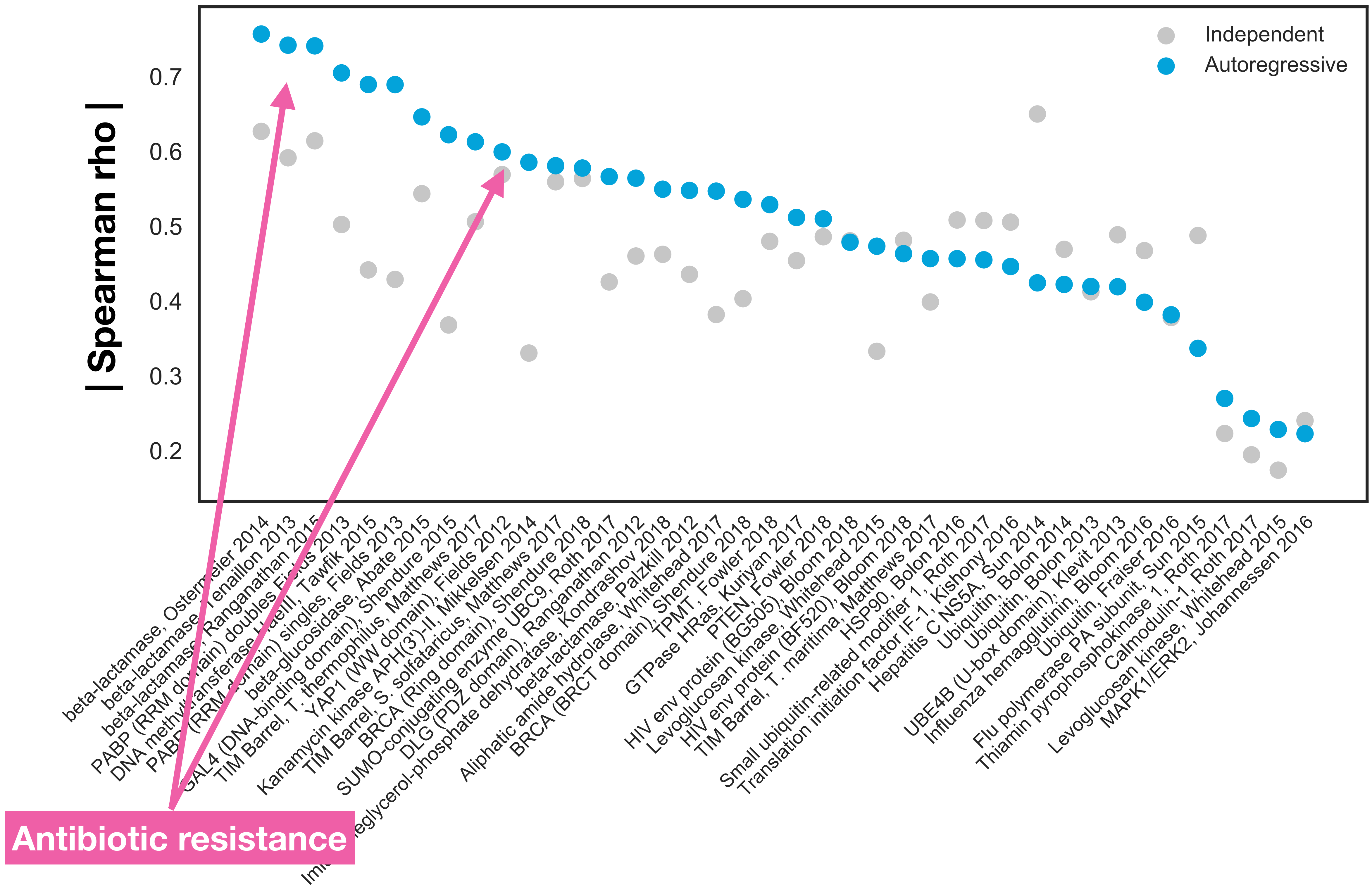
$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

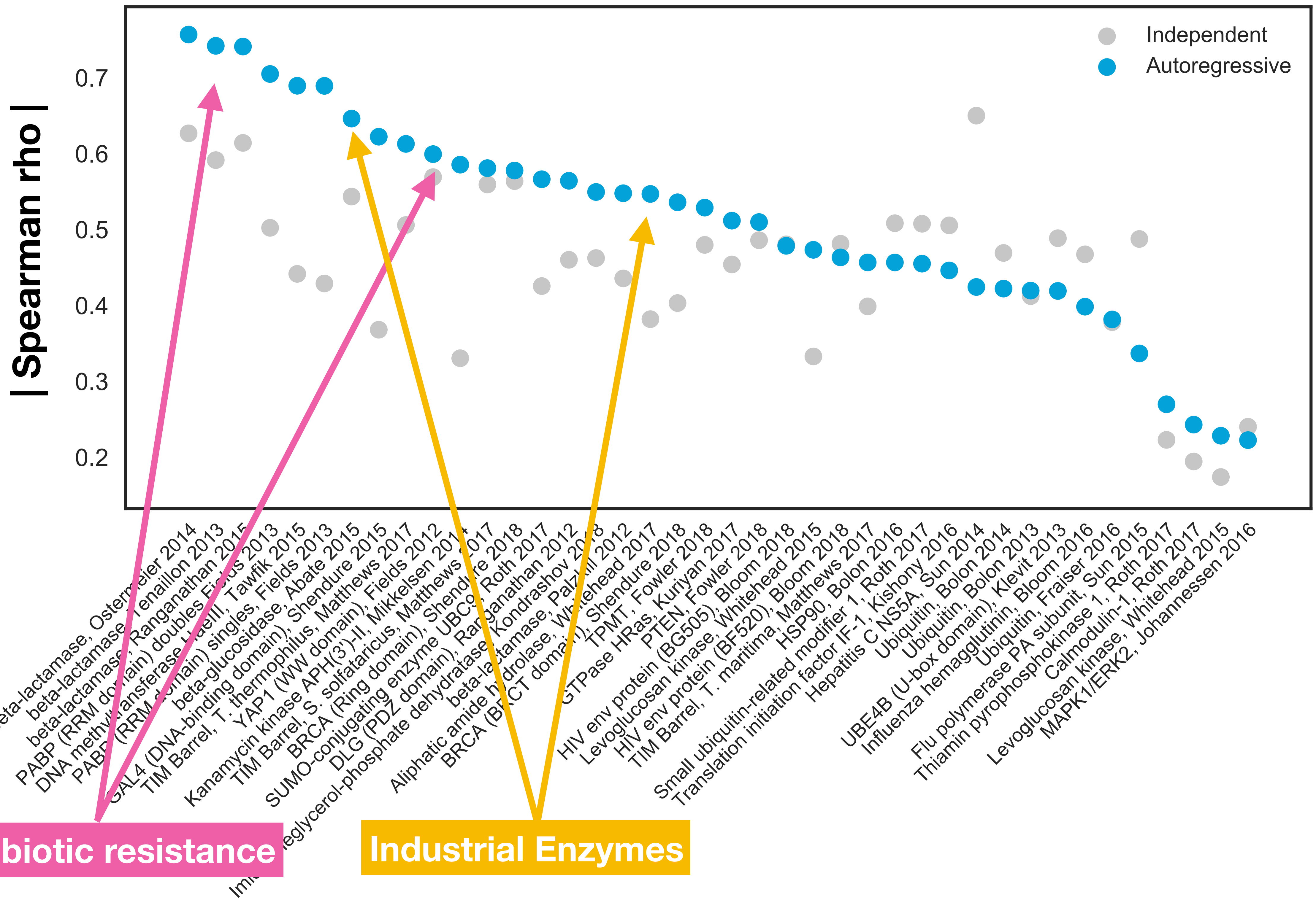
Autoregressive model is **predictive** of the effect of mutations



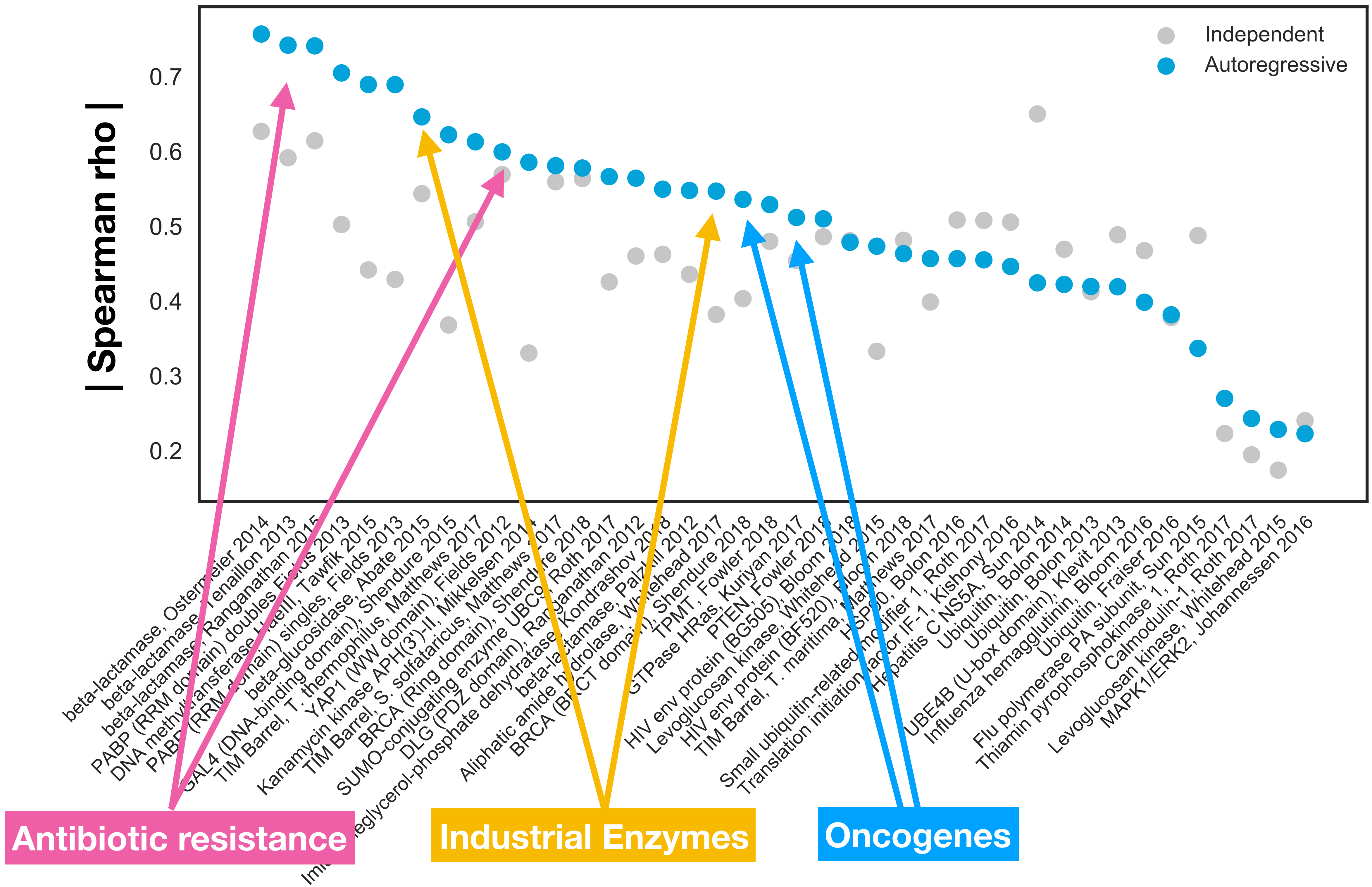
Autoregressive model is **predictive** of the effect of mutations



Autoregressive model is **predictive** of the effect of mutations



Autoregressive model is **predictive** of the effect of mutations



Autoregressive models are more **flexible**

Autoregressive models are more **flexible**

Missense mutations

…DKSGAG**V**RGSRGIIA…

…DKSGAG**I**RGSRGIIA…

Autoregressive models are more **flexible**

Missense mutations

…DKSGAG**V**RGSRGIIA…

…DKSGAG**I**RGSRGIIA…

Insertions

…DKSGAG**EGT**RGSRGIIA…

Autoregressive models are more **flexible**

Missense mutations

...DKSGAGVRGSRGIIA...

...DKSGAGIRGSRGIIA...

Insertions

...DKSGAGEGTRGSRGIIA...

Deletions

...DKSGAGRGSRGIIA...

Models **predict** the **effects** of **insertions** and **deletions**

Insertions

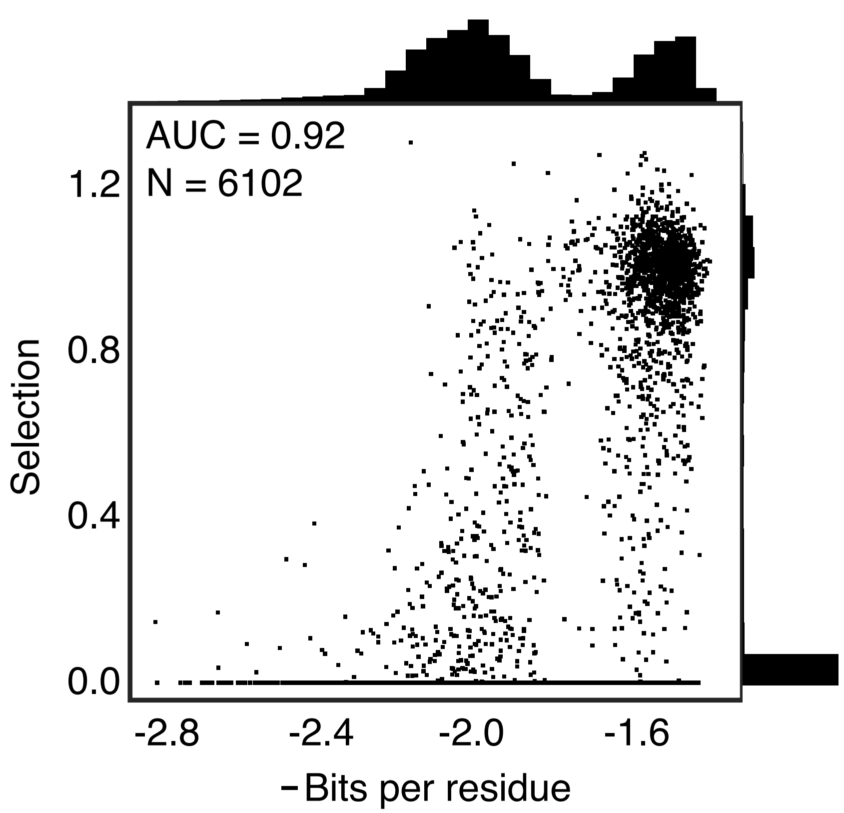
...DKSGAGEGTRGSRGIIA...

Deletions

...DKSGAGRGSRGIIA...

Imidazoleglycerol-phosphate dehydratase

Insertions & deletions



Models **predict** the **effects** of **insertions** and **deletions**

Insertions

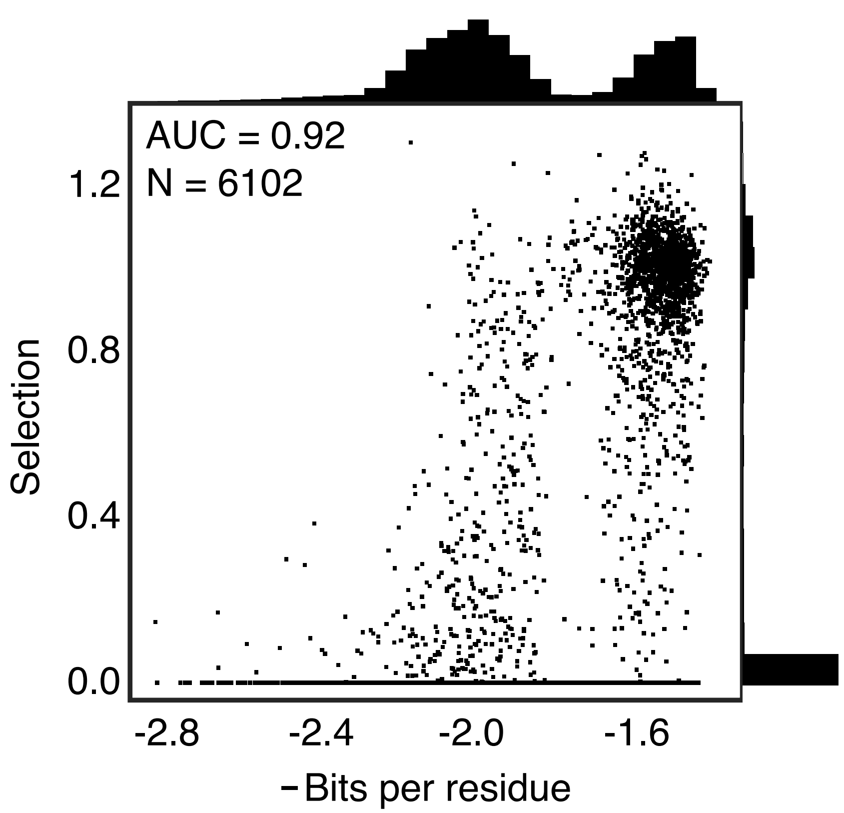
...DKSGAG**EGT**RGSRGIIA...

Deletions

...DKSGAGRGSRGIIA...

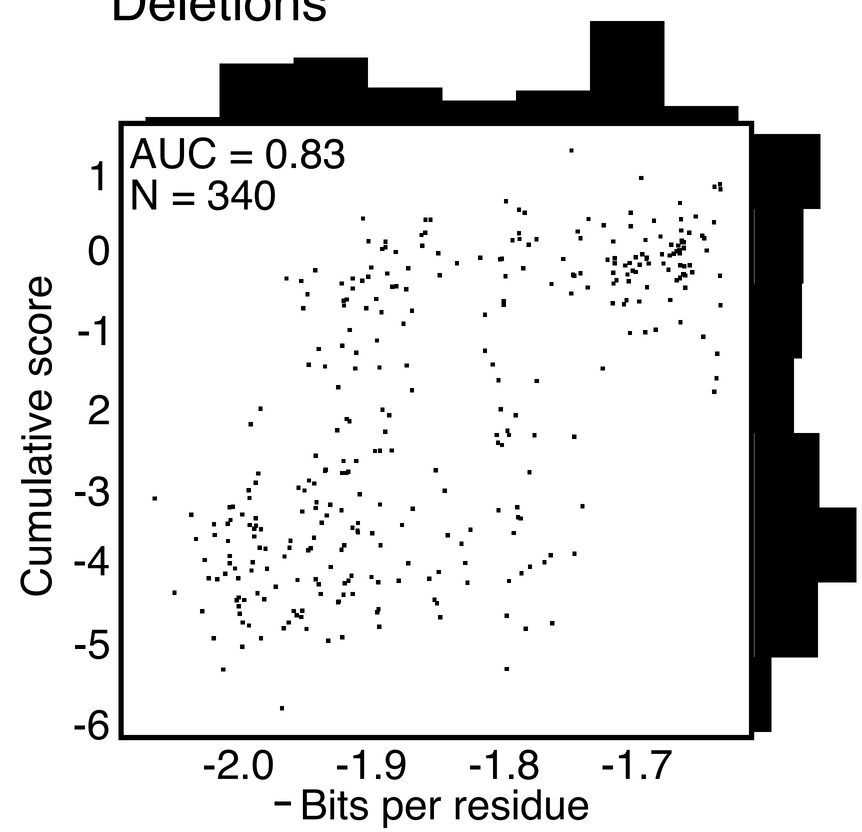
Imidazoleglycerol-phosphate dehydratase

Insertions & deletions

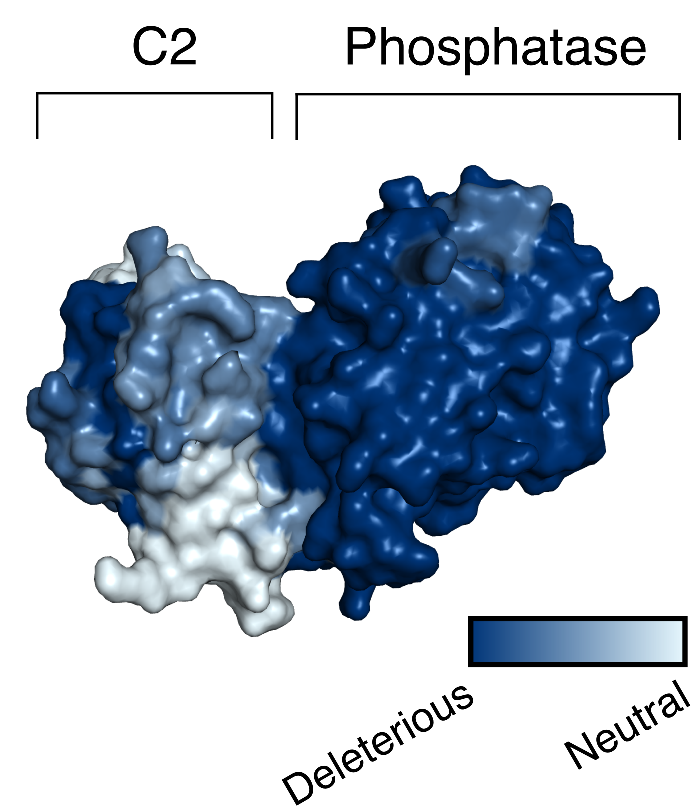


PTEN phosphatase

Deletions



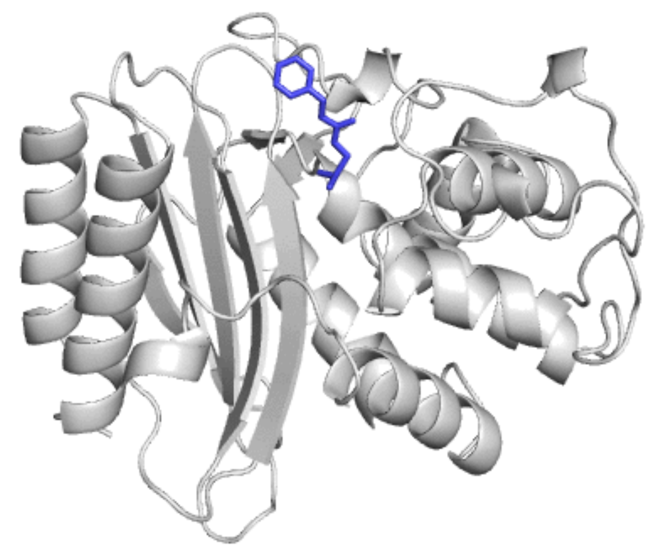
Single amino acid deletions



Recap

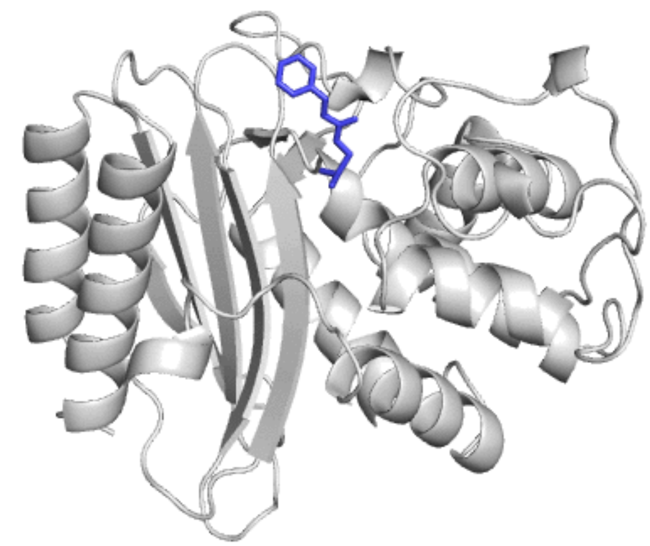
Recap

Predicting the effects of mutations is important



Recap

Predicting the effects of mutations is important

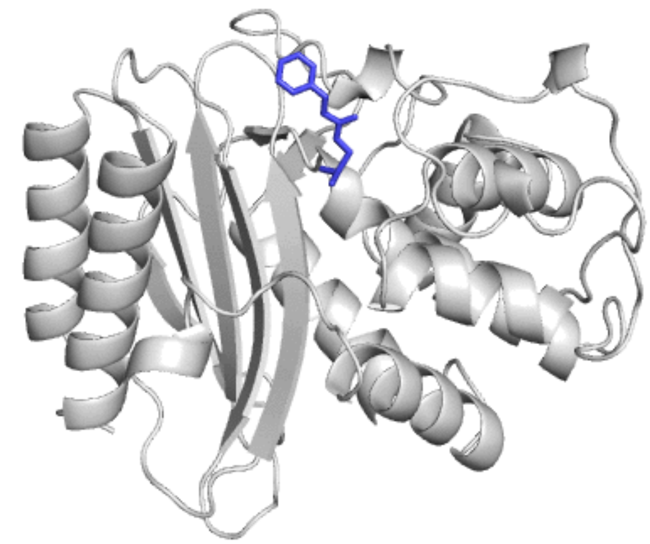


Sequencing is becoming cheaper and easier

```
GTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGG
AAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGA
TTCCAGCCAGGCAGTGGCGGATCAATATGCCGACTTCCTG
ACACCTCGTCGATGGATTACTACCATCAGTTGCGTTATGC
GTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATC
TCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGAC
AAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTC
GAAACGGGACGTGAACTGGAGCTGGCGGATATTGAAATTG
TGCCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTC
```

Recap

Predicting the effects of mutations is important



Sequencing is becoming cheaper and easier

```
GTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGG
AAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGA
TTCCAGCCAGGCAGTGGCGGATCAATATGCCGACTTCCTG
ACACCTCGTCGATGGATTACTACCATCAGTTGCGTTATGC
GTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATC
TCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGAC
AAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTC
GAAACGGGACGTGAACTGGAGCTGGCGGATATTGAAATTG
TGCCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTC
```

Generative models fit to biological sequence data can predict the effect of mutations

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

Part II: **Genotype -> phenotype for complex diseases**

insitro™

Part II: **Genotype -> phenotype** for **complex diseases**

Part II: **Genotype -> phenotype** for **complex diseases**

insitro™

Founded by Daphne Koller in 2018

Part II: **Genotype -> phenotype** for **complex diseases**

insitro™

Founded by Daphne Koller in 2018

Create a **new paradigm** for **drug development**

Part II: **Genotype -> phenotype** for **complex diseases**

Founded by Daphne Koller in 2018

Create a **new paradigm** for **drug development** that uses **high-quality data** and **data-driven models**

Part II: **Genotype -> phenotype for complex diseases**

insitro™

Founded by Daphne Koller in 2018

Create a **new paradigm** for **drug development** that uses **high-quality data** and **data-driven models** to design **novel**, **safe**, and e**ffective therapies**

Part II: **Genotype -> phenotype for complex diseases**

Founded by Daphne Koller in 2018

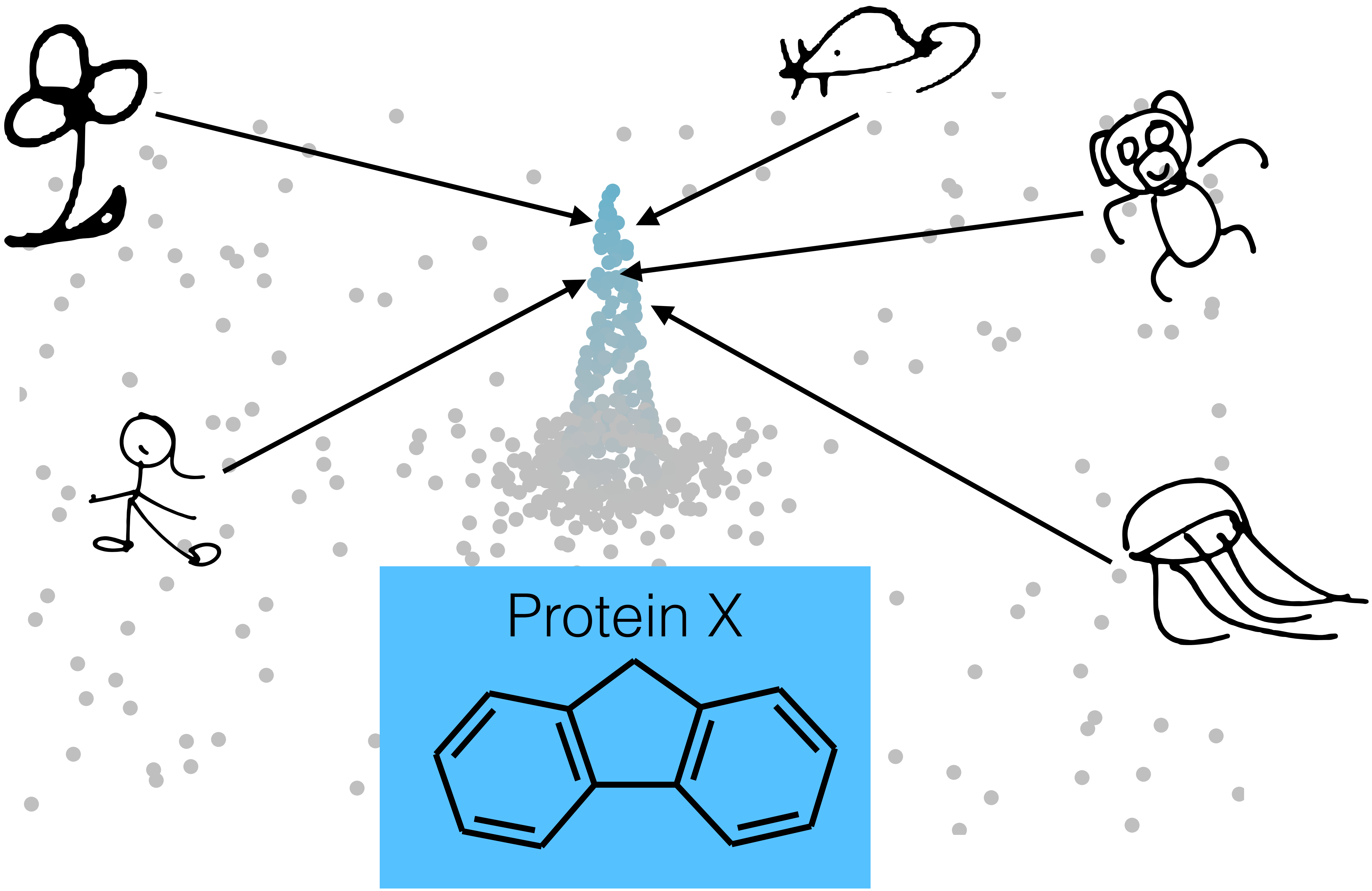
Create a **new paradigm** for **drug development**

that uses **high-quality data** and **data-driven models**

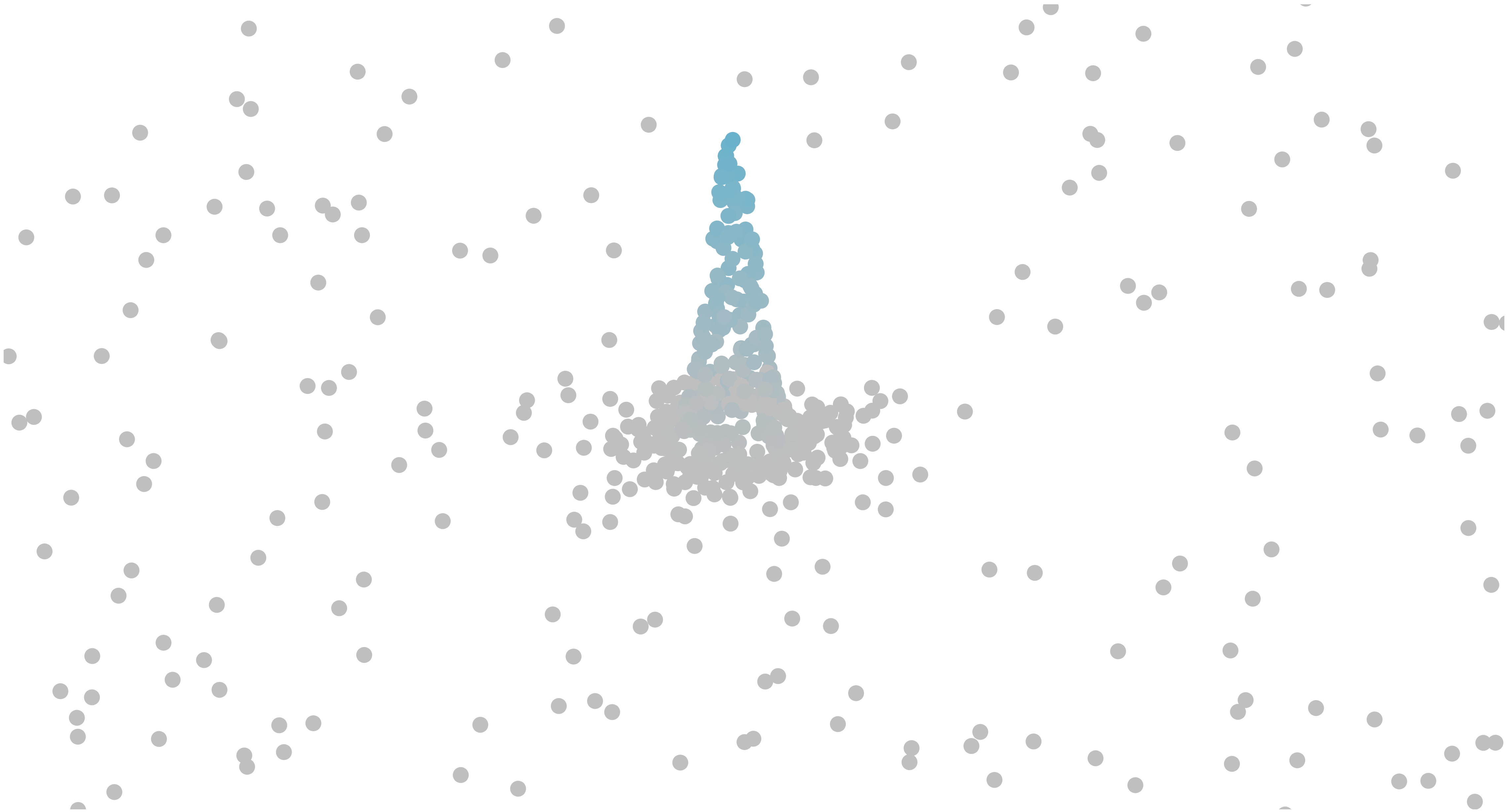
to design **novel**, **safe**, and e**ffective therapies**

that help **more people**, **faster**, and at a **lower cost.**

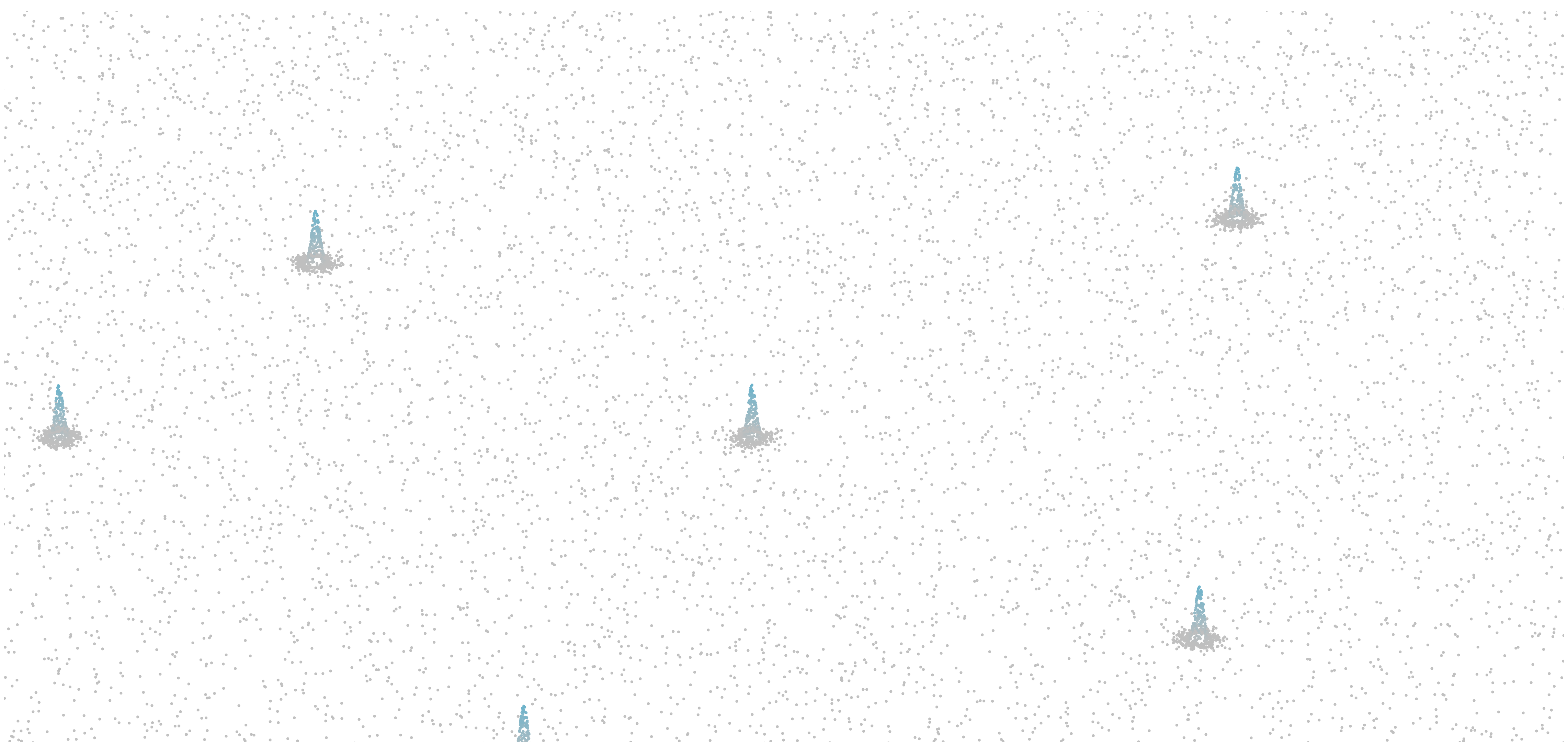
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



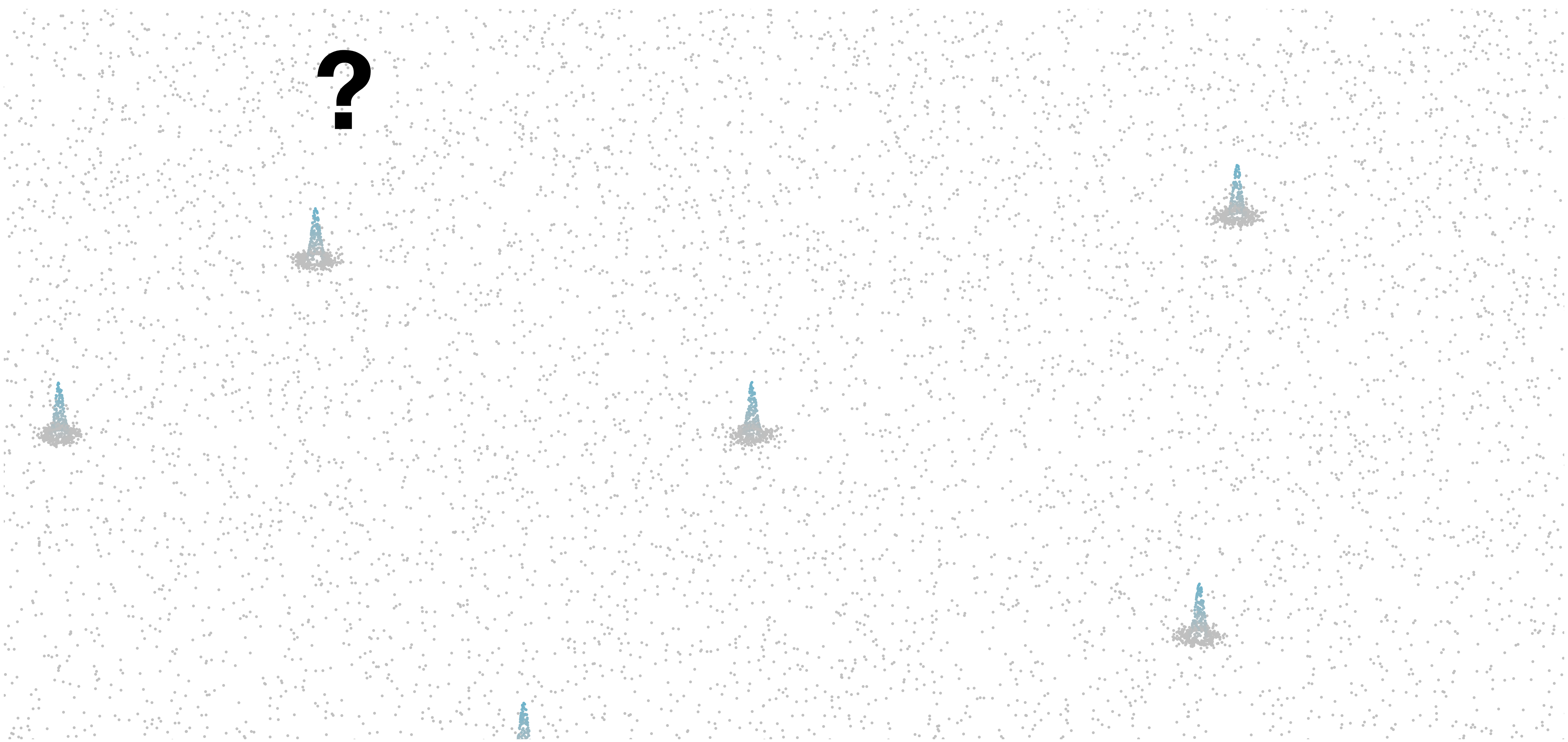
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



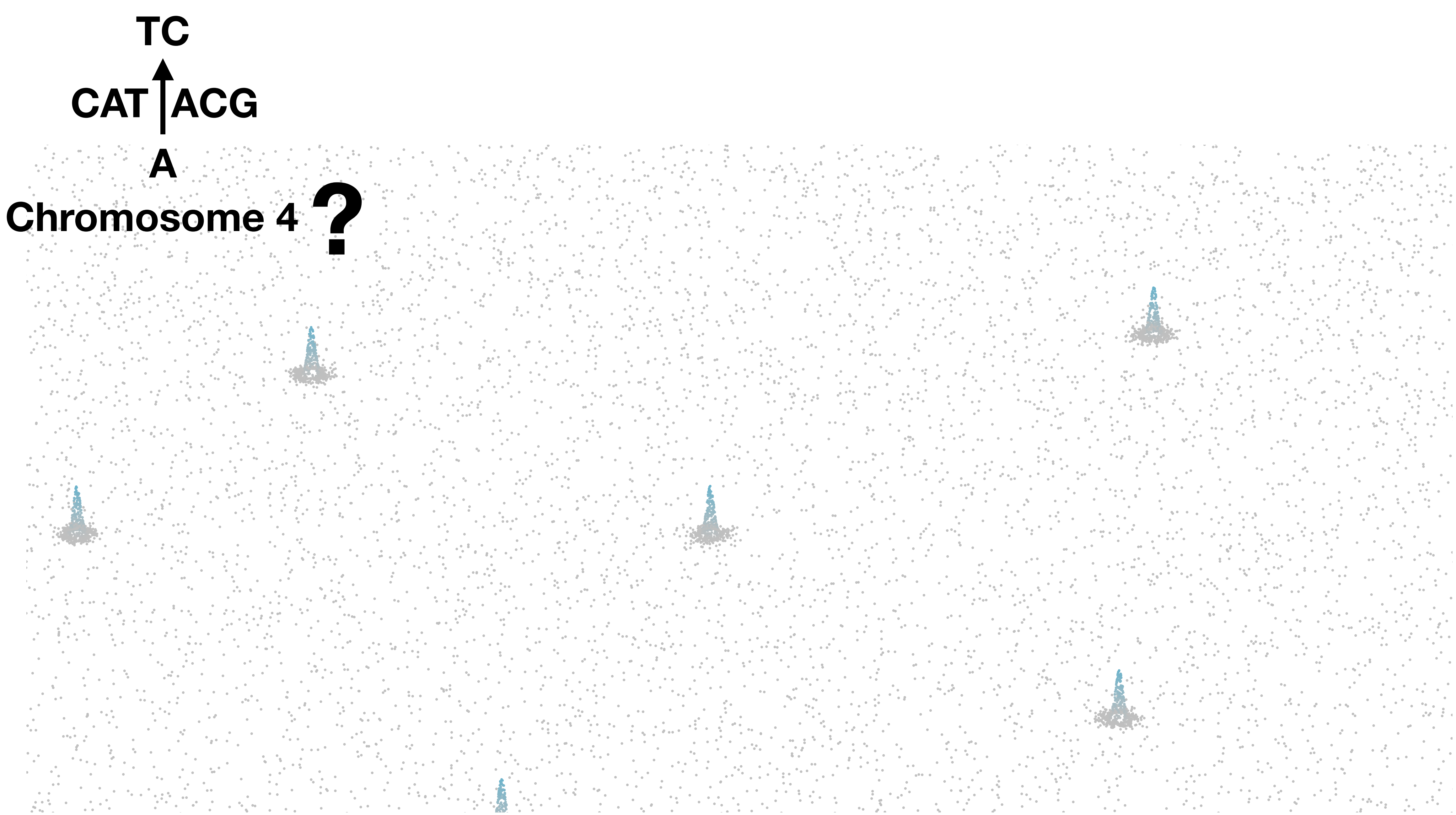
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



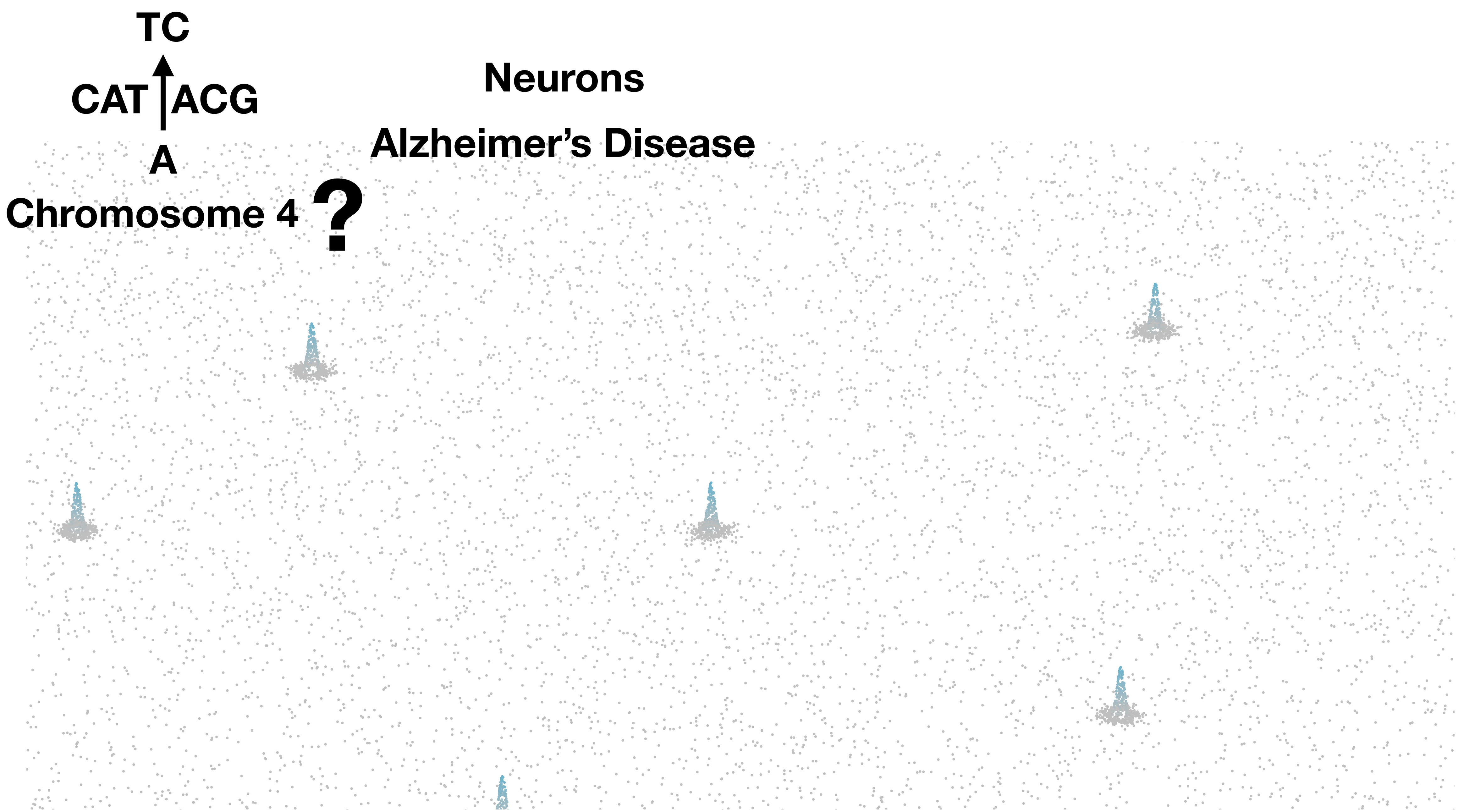
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



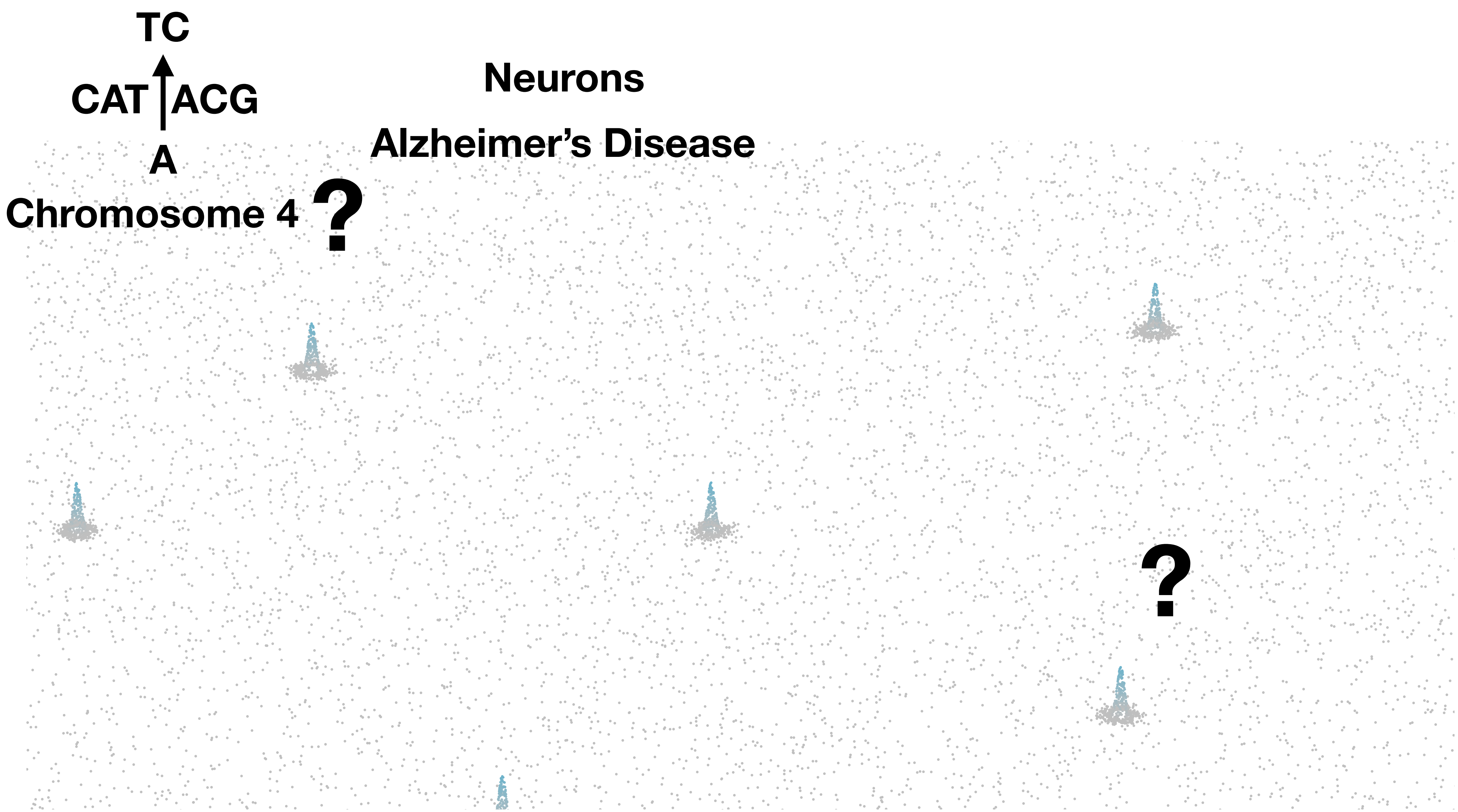
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



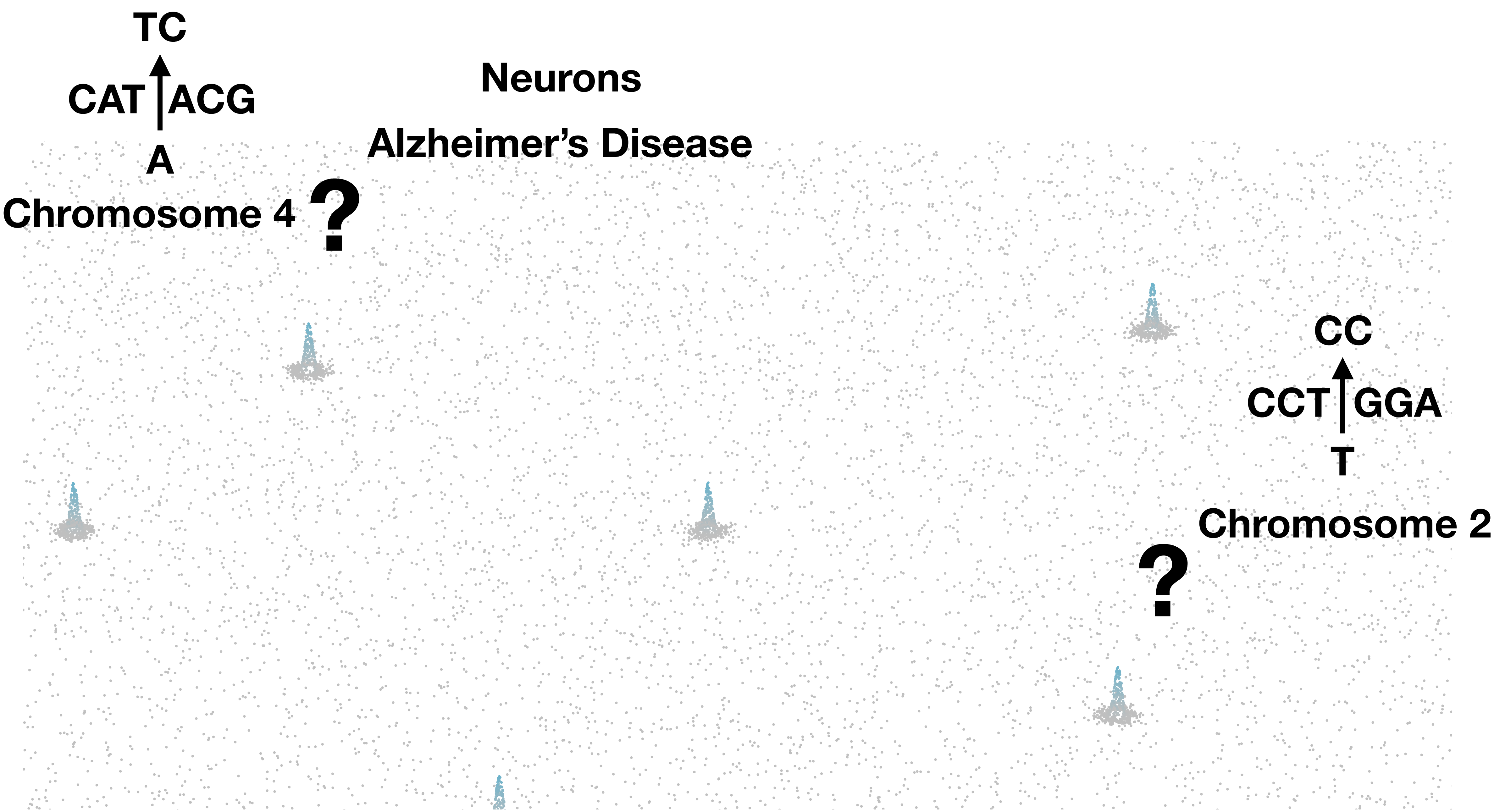
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



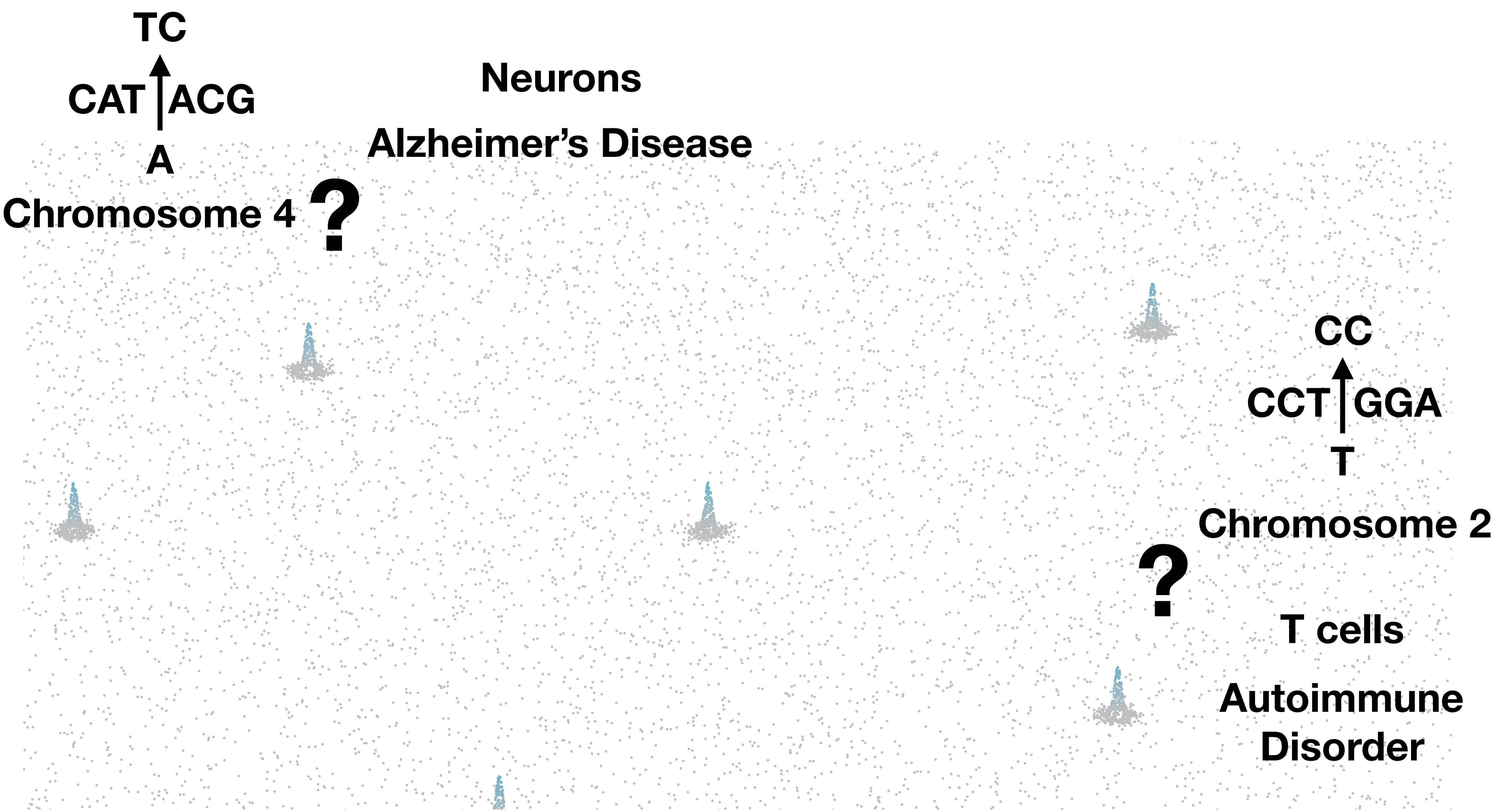
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



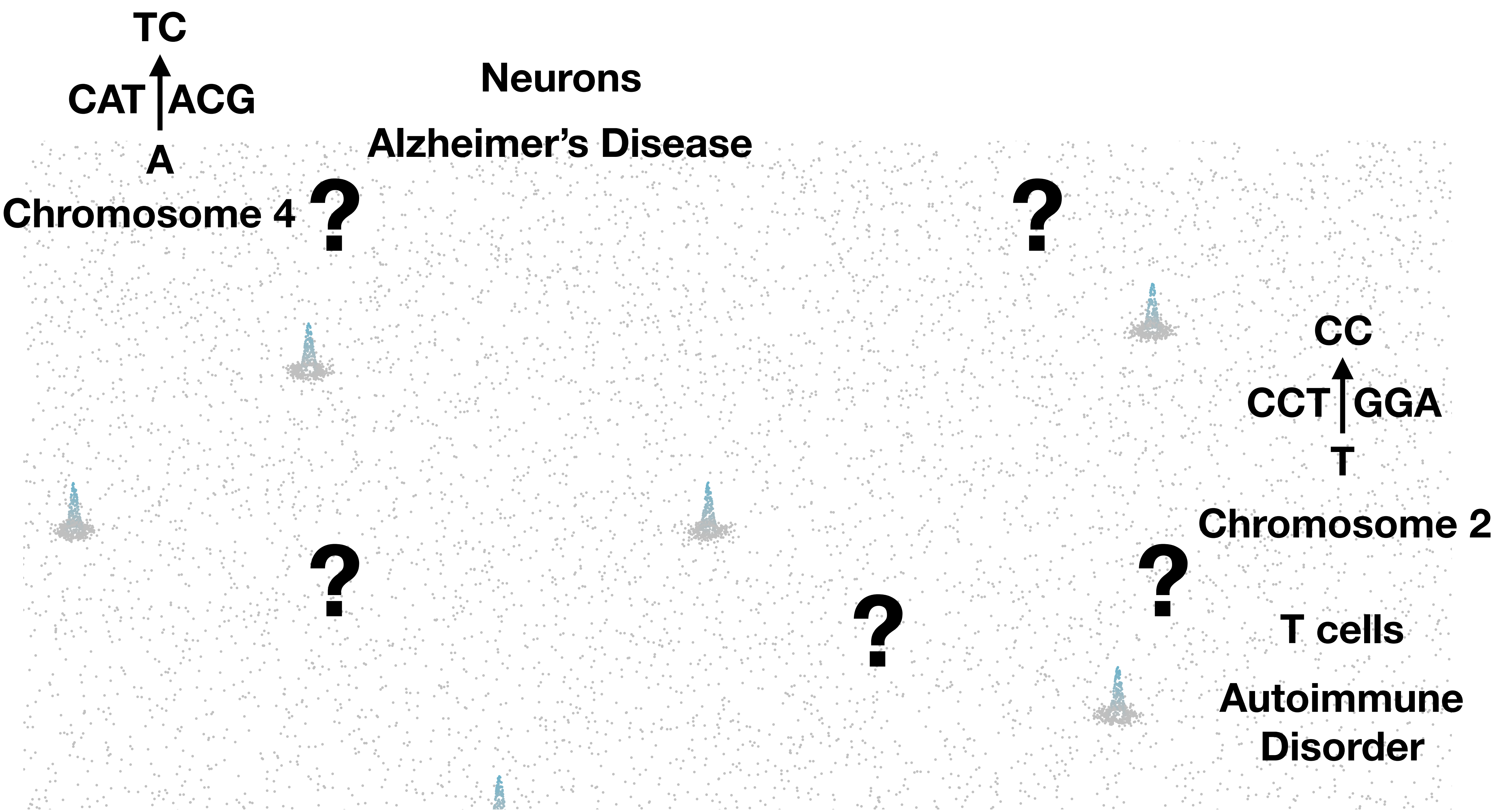
Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



Part II: **Genotype -> phenotype** for different **tissues** are **sparse**



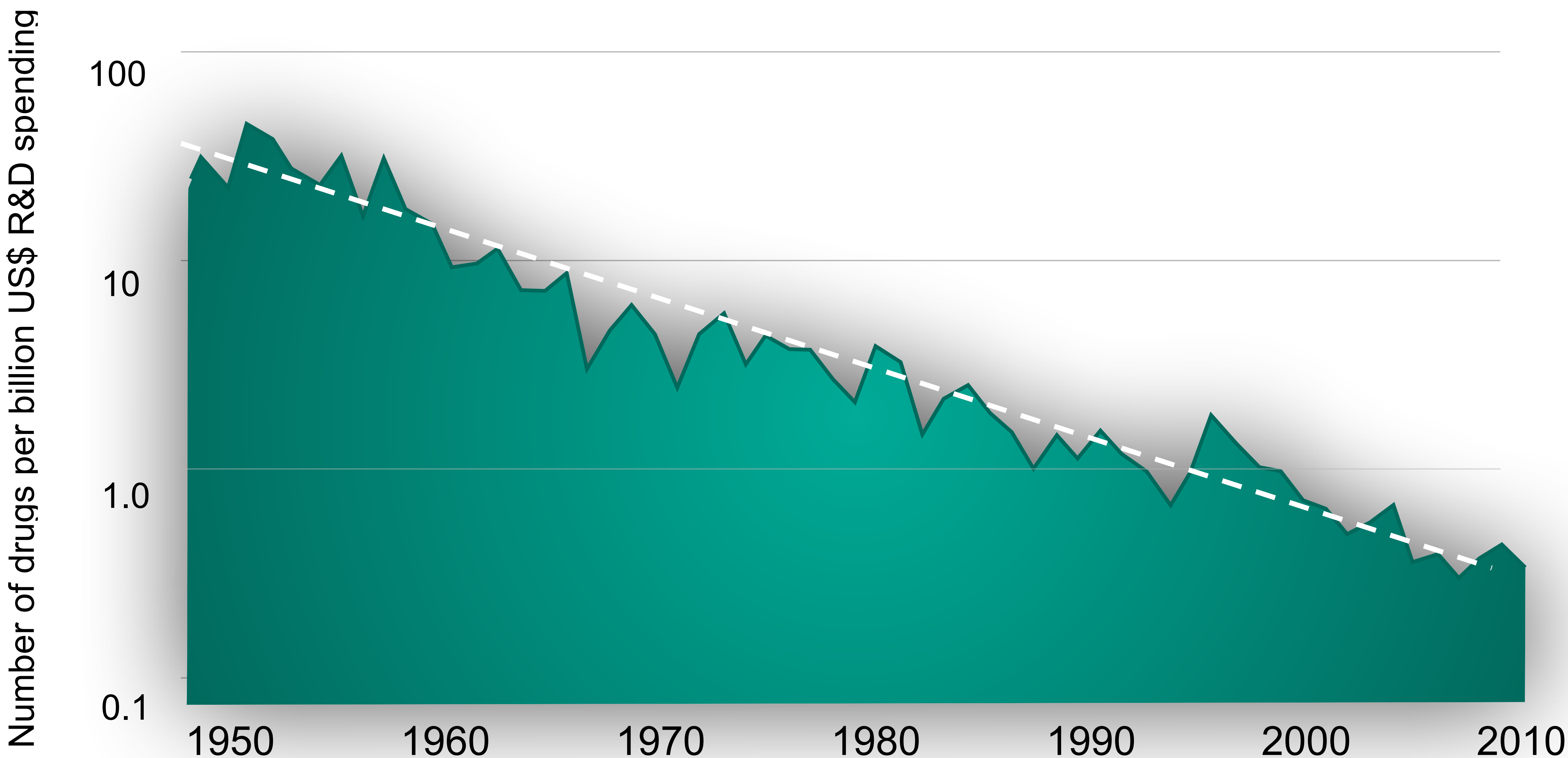
Drug discovery is becoming exponentially **less efficient**

Drug discovery is becoming exponentially **less efficient**

Moore's Law → **Eroom's Law**

Drug discovery is becoming exponentially **less efficient**

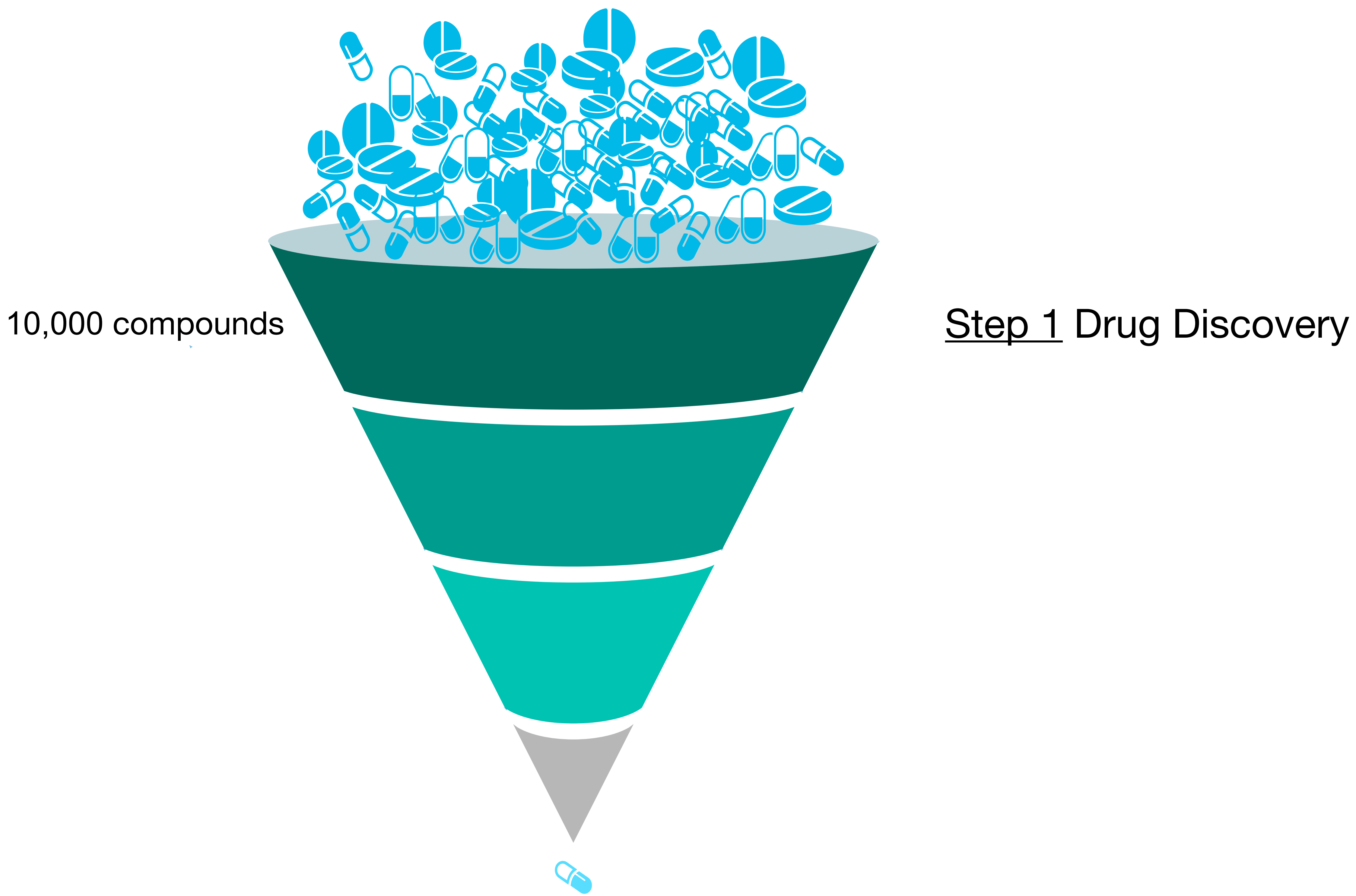
Moore's Law → **Eroom's Law**



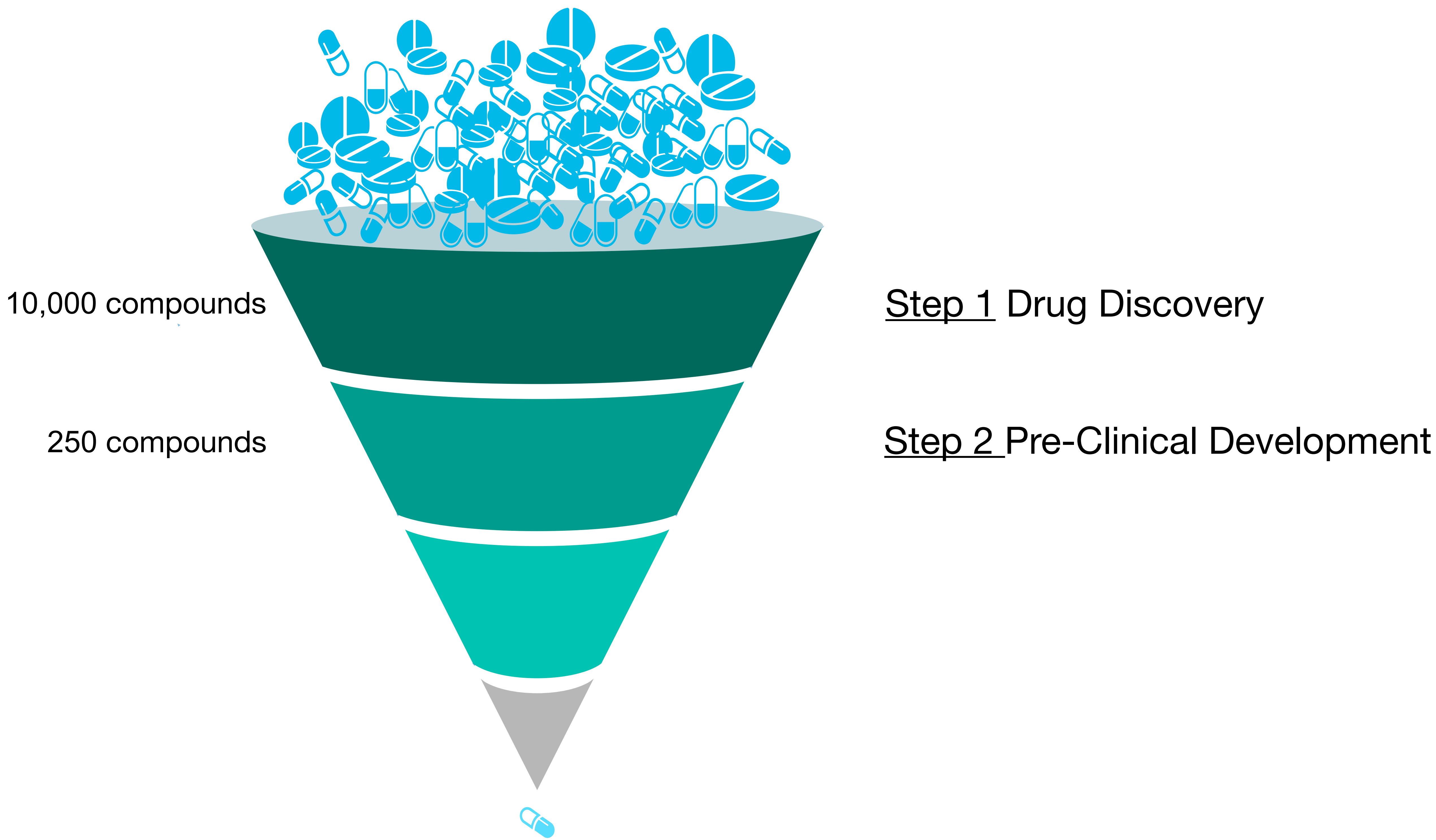
Scannell et al, Nature Reviews Drug Discovery, 2012.

Drug discovery is ultimately a **problem** of **prediction**

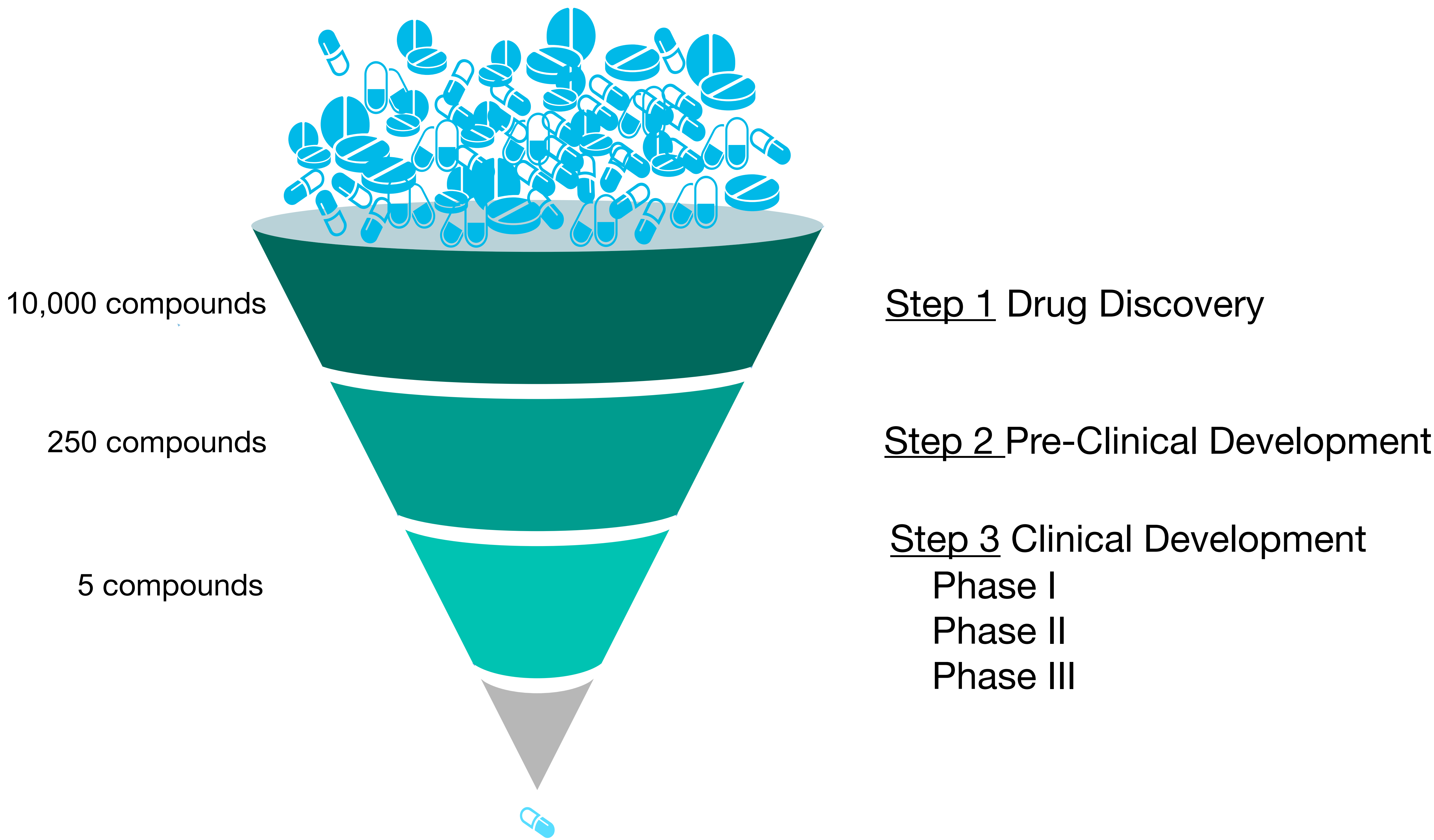
Drug discovery is ultimately a **problem** of **prediction**



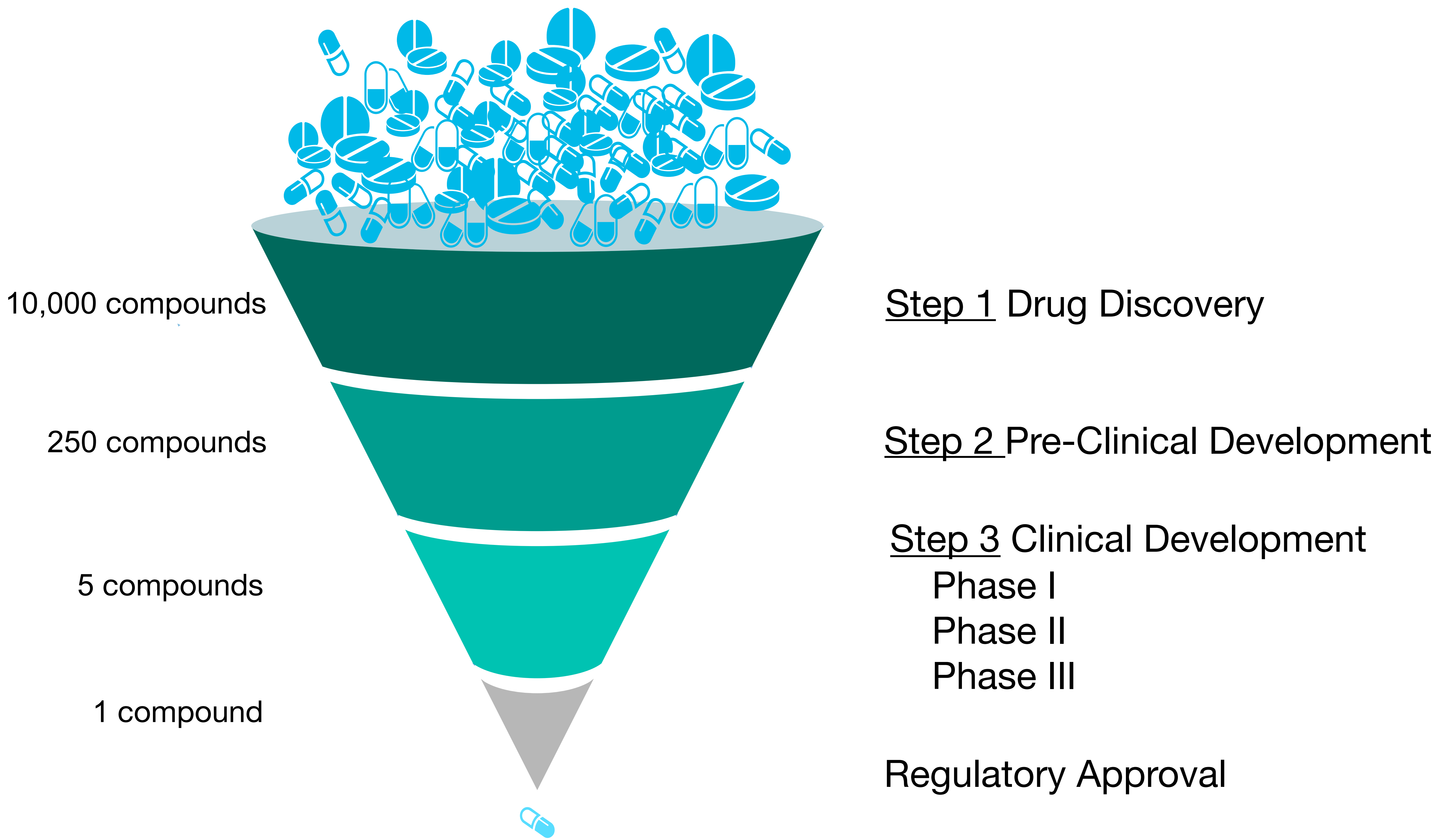
Drug discovery is ultimately a **problem** of **prediction**



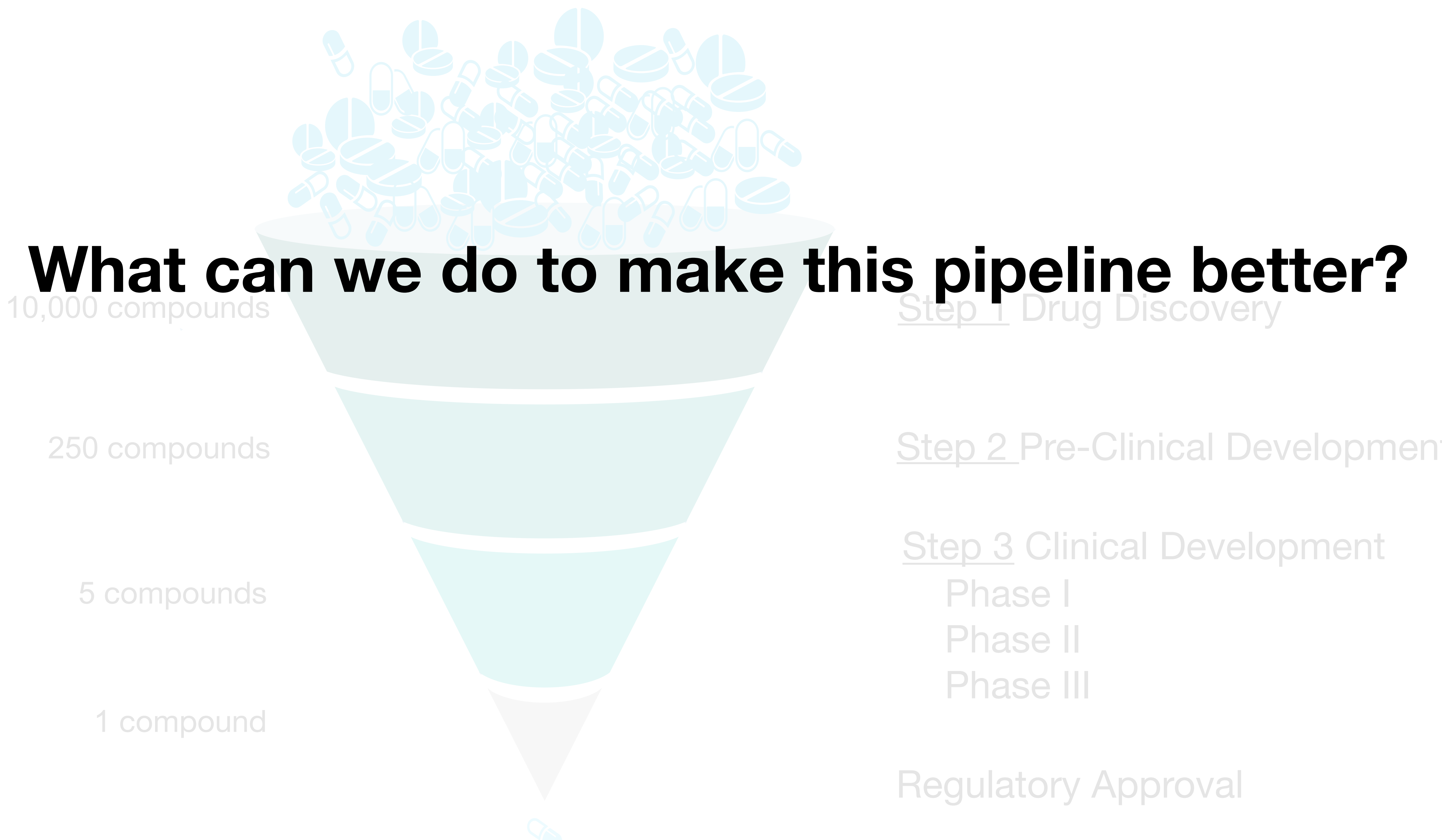
Drug discovery is ultimately a **problem** of **prediction**



Drug discovery is ultimately a **problem** of **prediction**

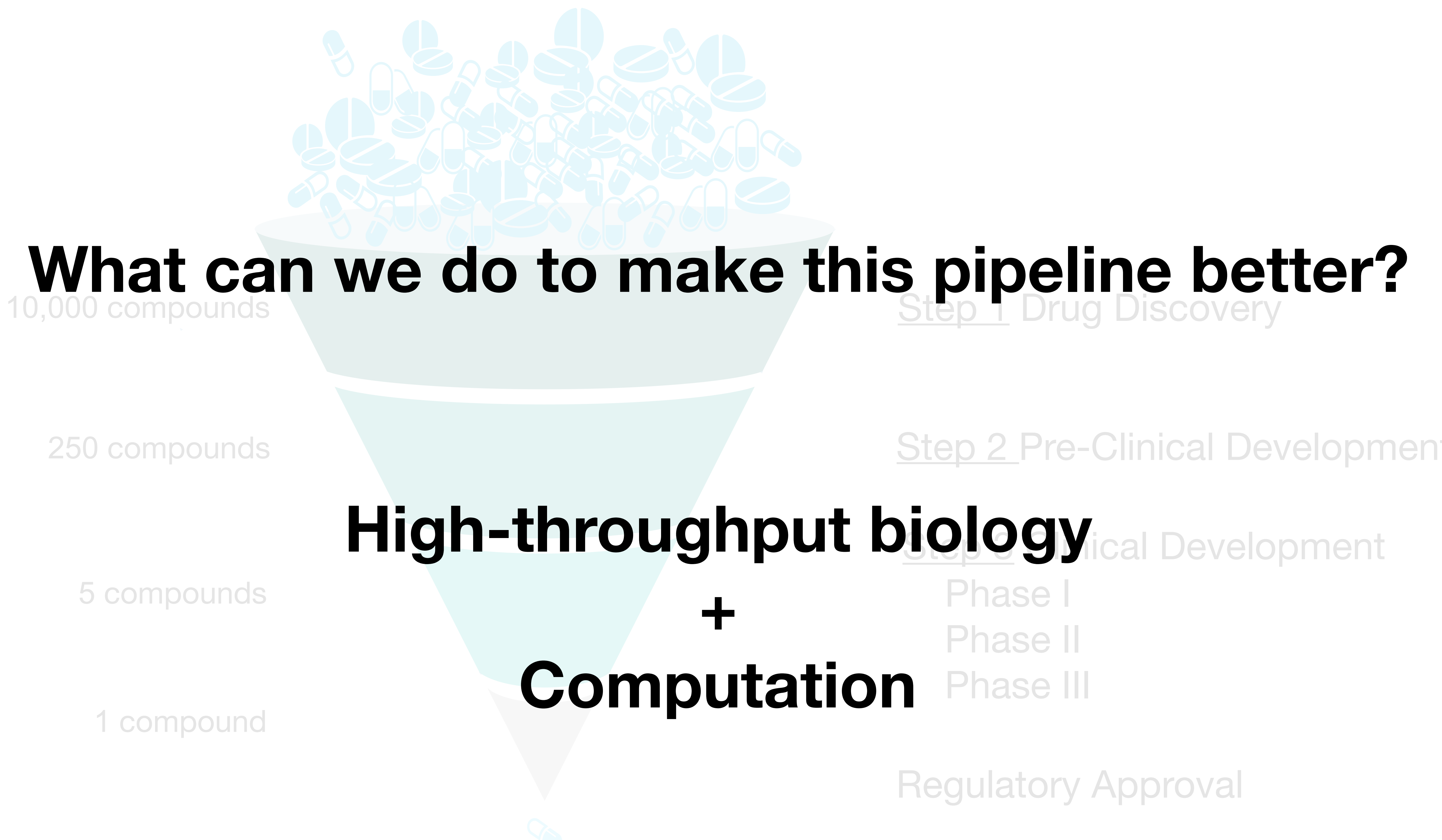


Drug discovery is ultimately a **problem** of **prediction**



What can we do to make this pipeline better?

Drug discovery is ultimately a **problem** of **prediction**



What can we do to make this pipeline better?

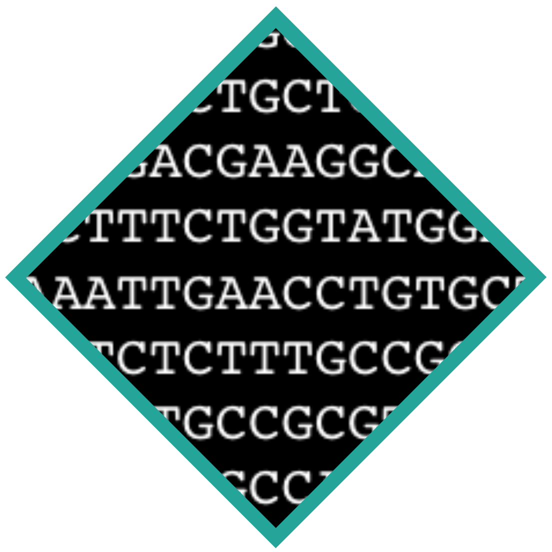
High-throughput biology

+

Computation

insitro is founded on **data-driven** technologies

Human genetics



https://www.rndsystems.com/resources/articles/differentiation-potential-induced-pluripotent-stem-cells

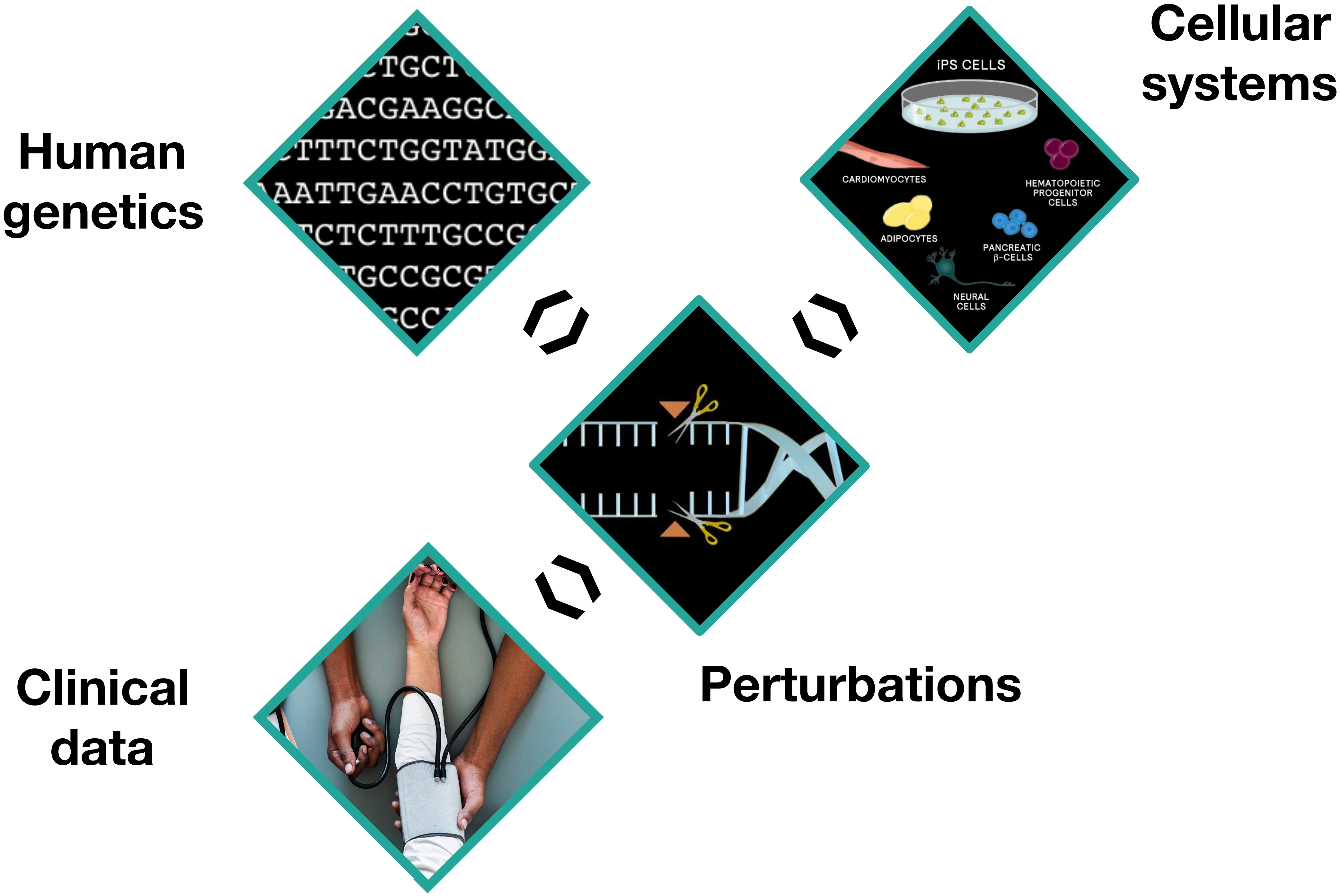
insitro is founded on **data-driven** technologies

Human genetics

Clinical data

https://www.rndsystems.com/resources/articles/differentiation-potential-induced-pluripotent-stem-cells

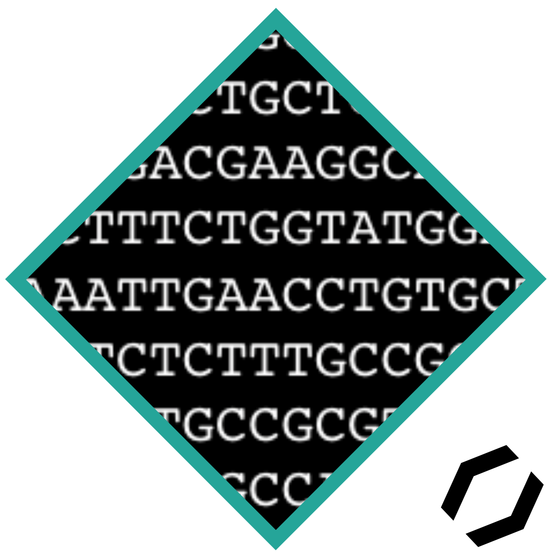
insitro is founded on **data-driven** technologies



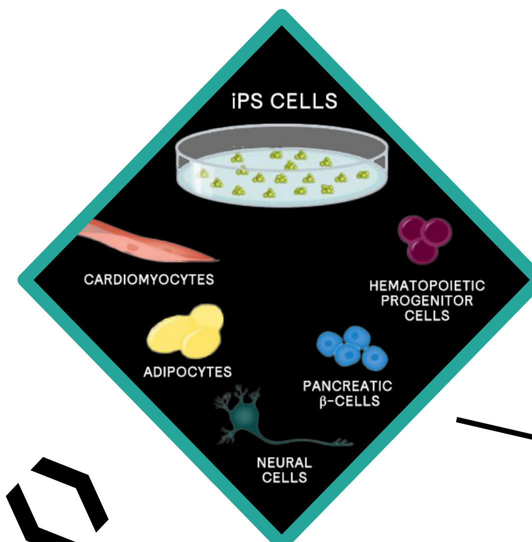
https://www.rndsystems.com/resources/articles/differentiation-potential-induced-pluripotent-stem-cells

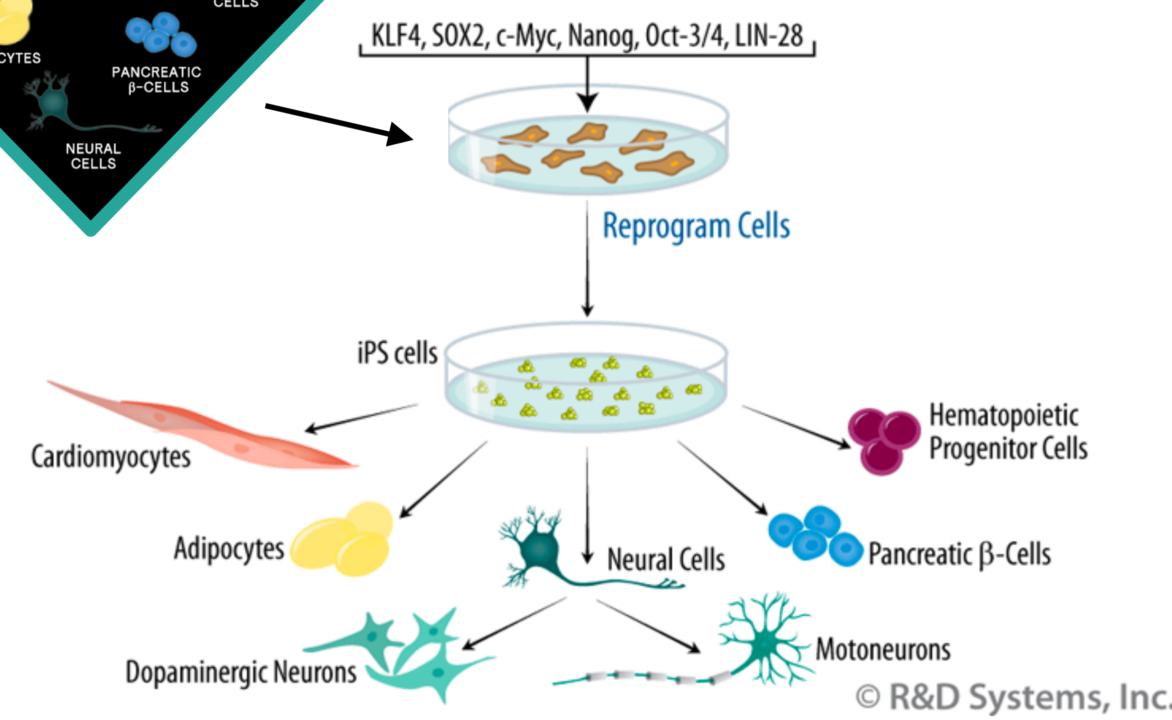
insitro is founded on **data-driven** technologies

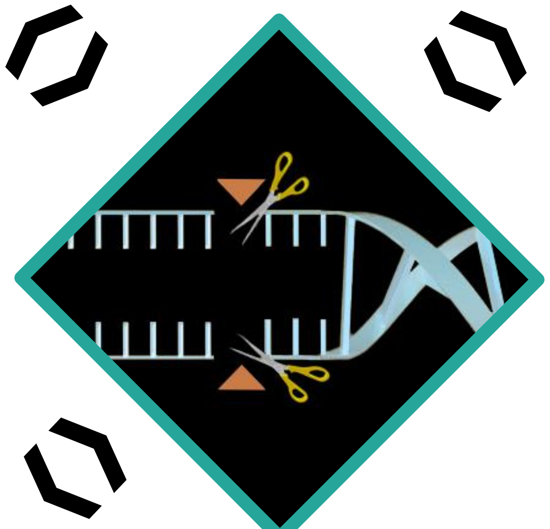
Human genetics



Cellular systems





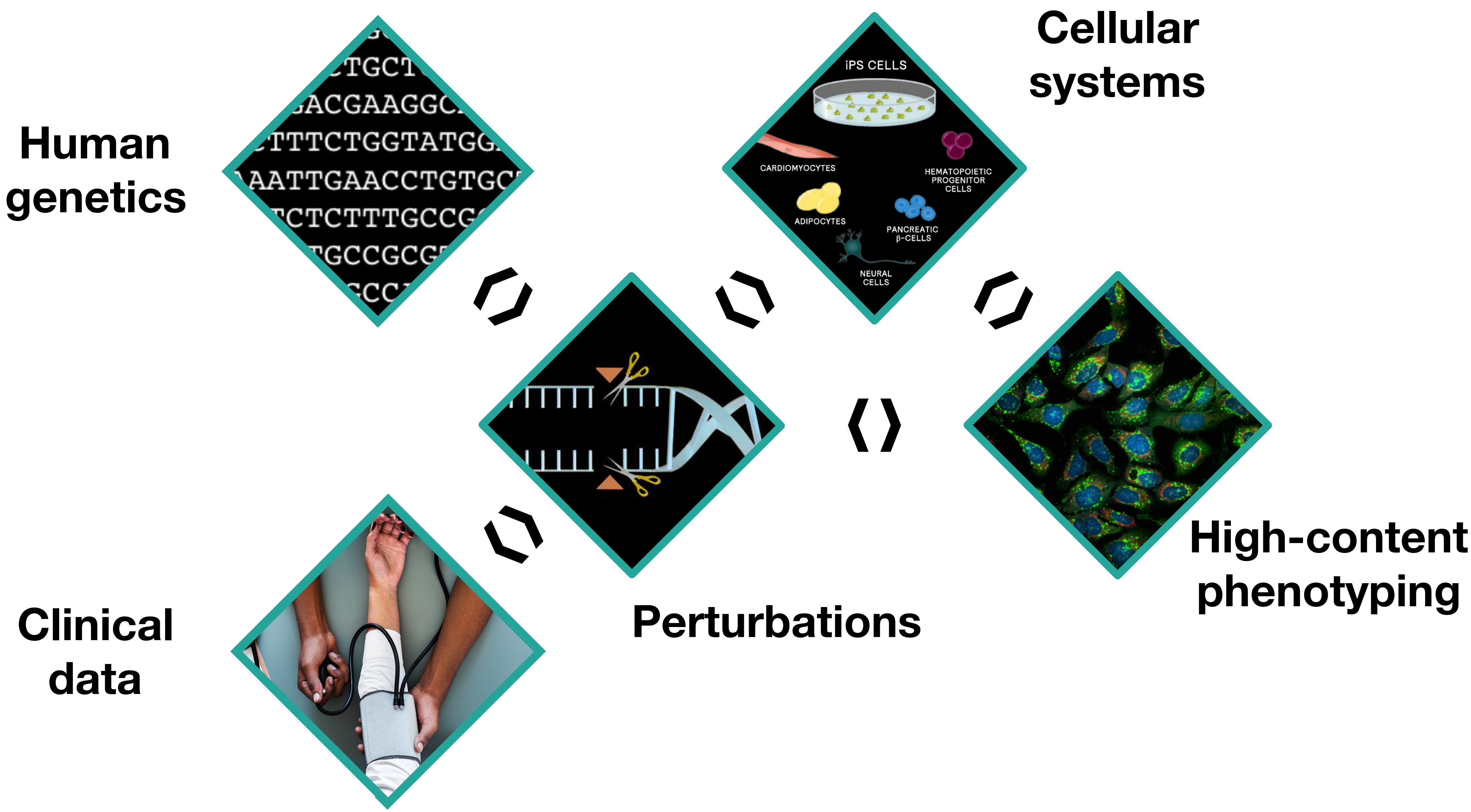


Perturbations

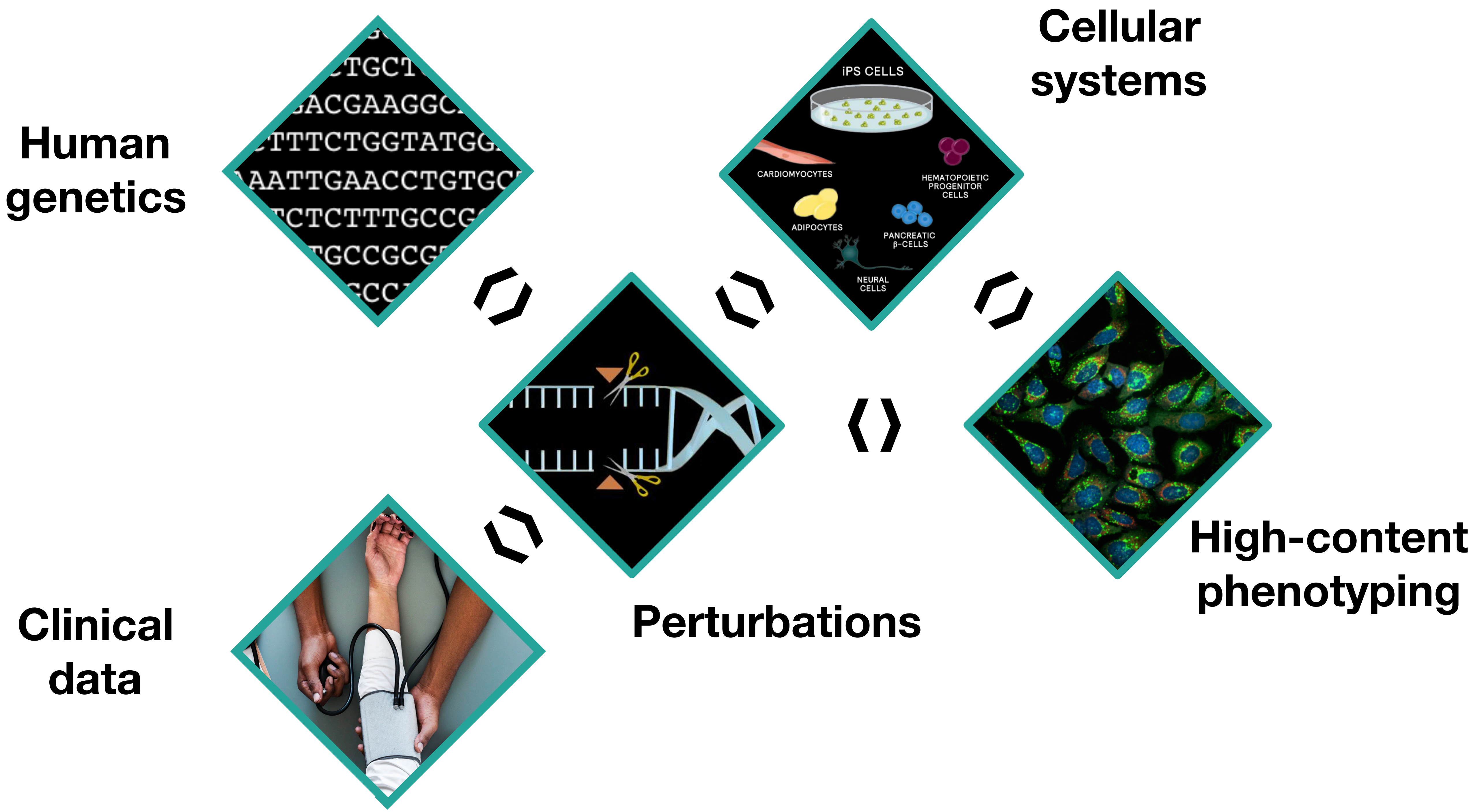
Clinical data

https://www.rndsystems.com/resources/articles/differentiation-potential-induced-pluripotent-stem-cells

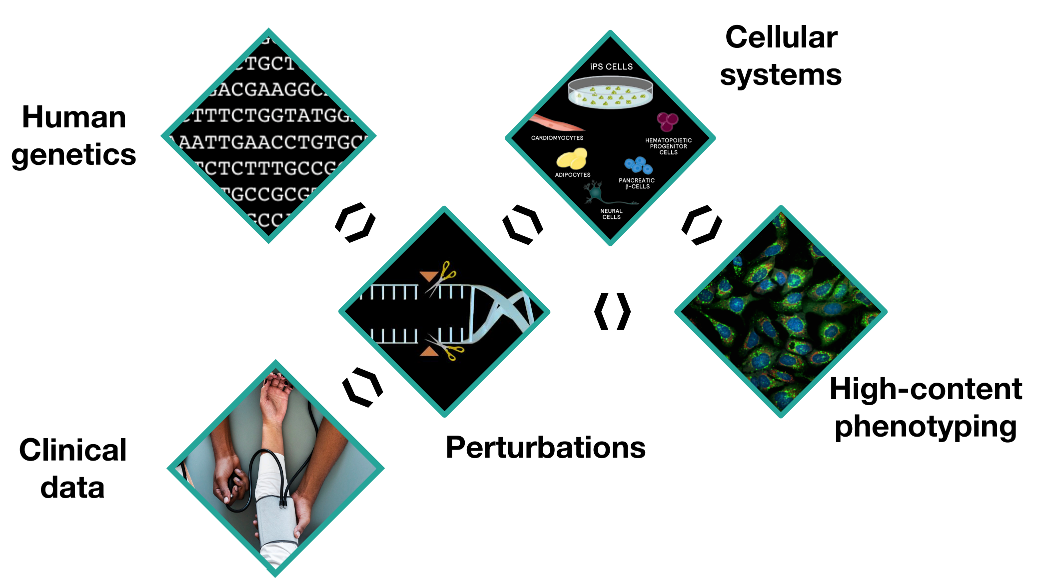
insitro is founded on **data-driven** technologies



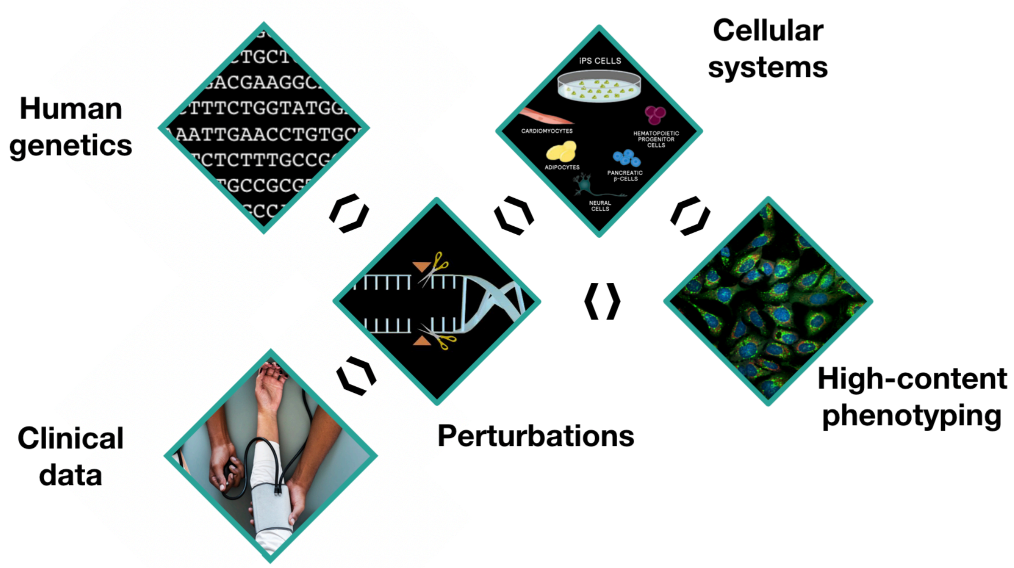
insitro is founded on **data-driven** technologies



insitro is founded on **data-driven** technologies

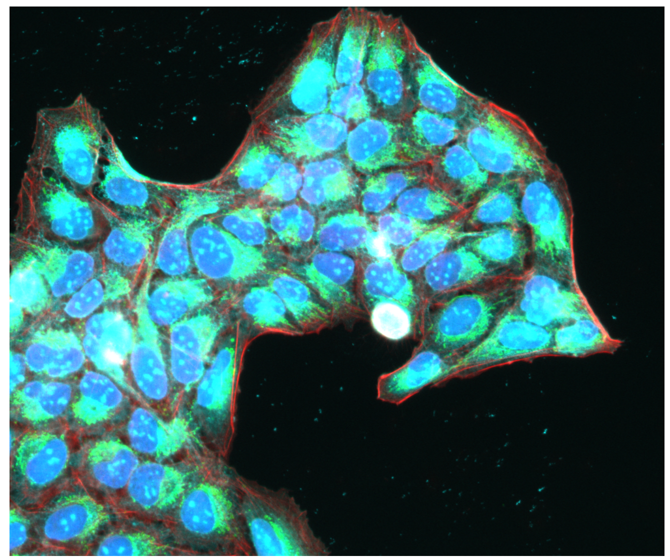


insitro is founded on **data-driven** technologies

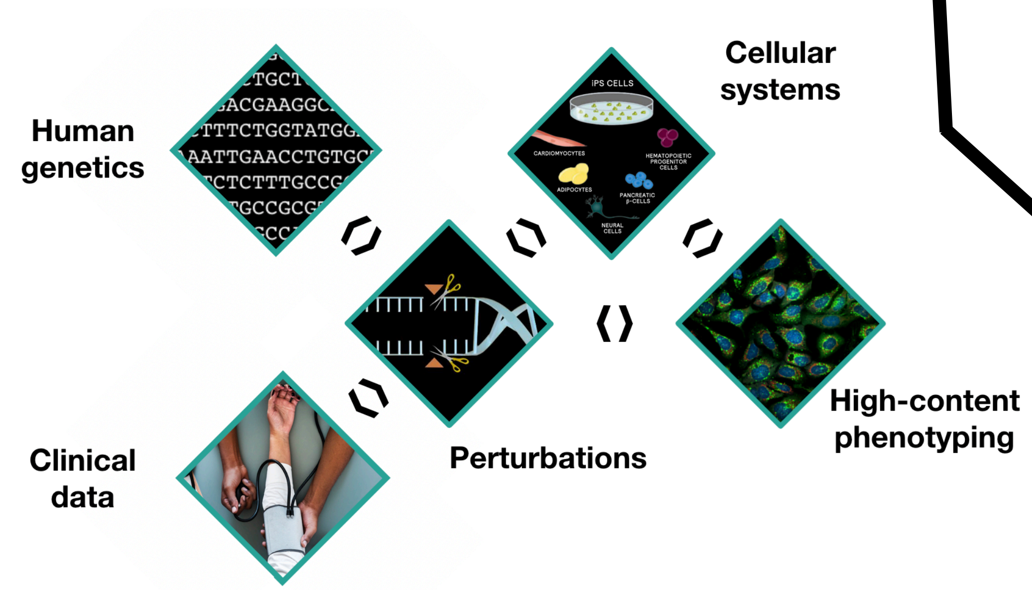


https://www.youtube.com/watch?v=vL7ptq2Dcf0

insitro is founded on **data-driven** technologies

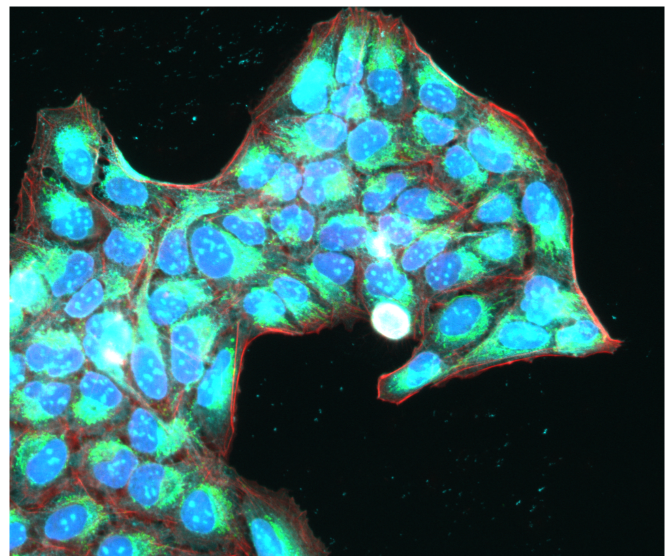


High-content microscopy

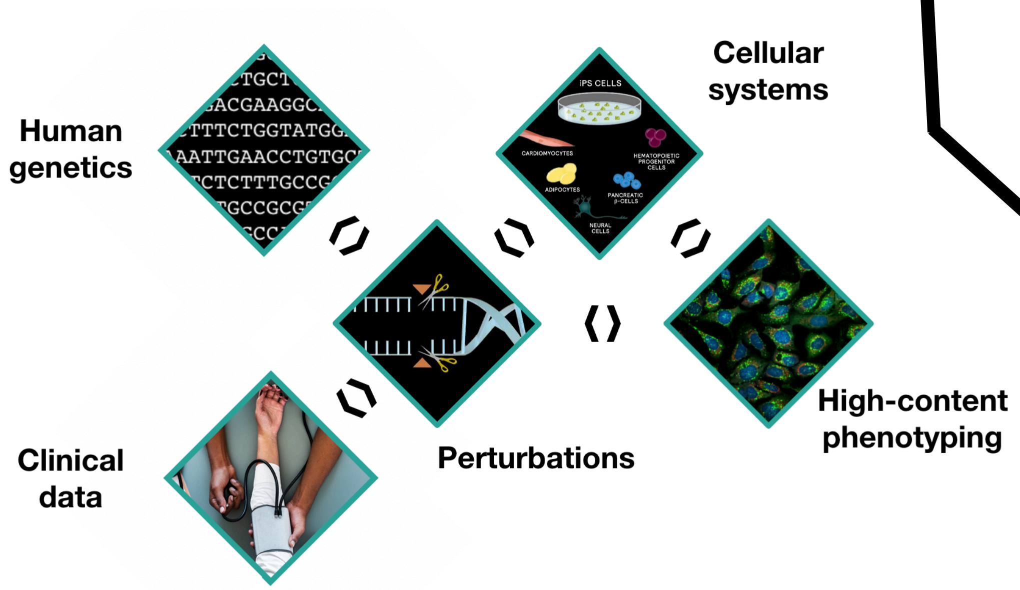


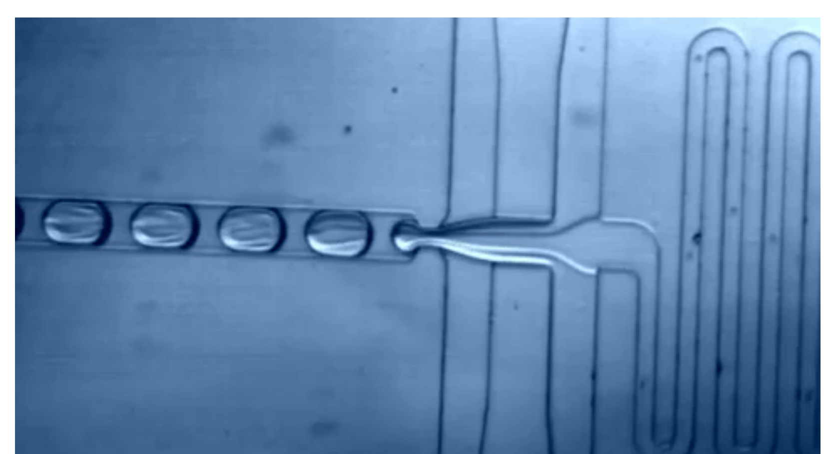
https://www.youtube.com/watch?v=vL7ptq2Dcf0

insitro is founded on **data-driven** technologies



High-content microscopy

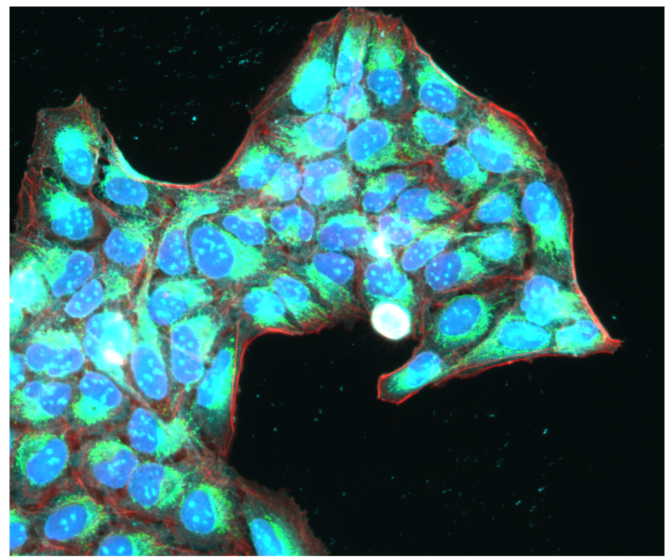




Single-cell RNA sequencing

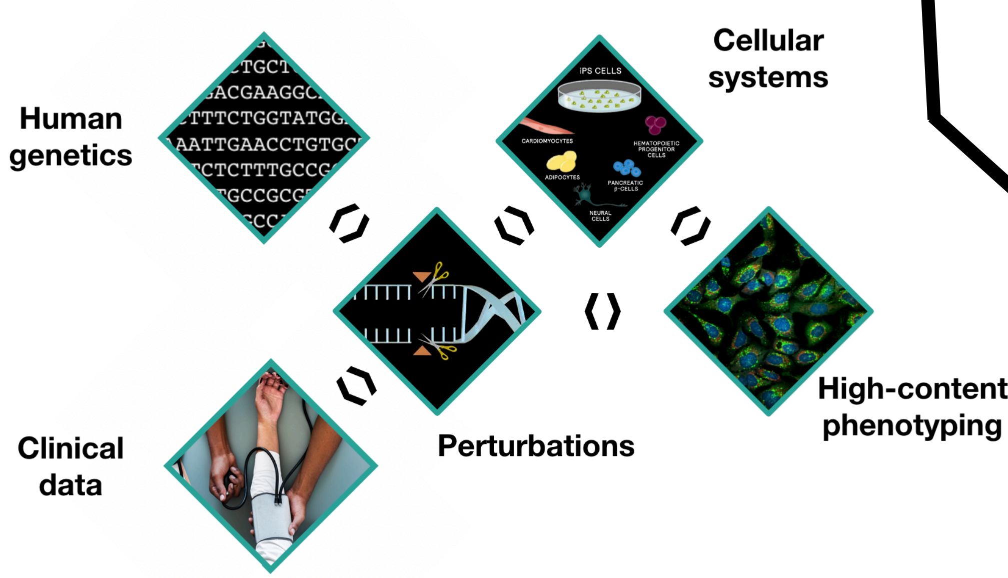
https://www.youtube.com/watch?v=vL7ptq2Dcf0

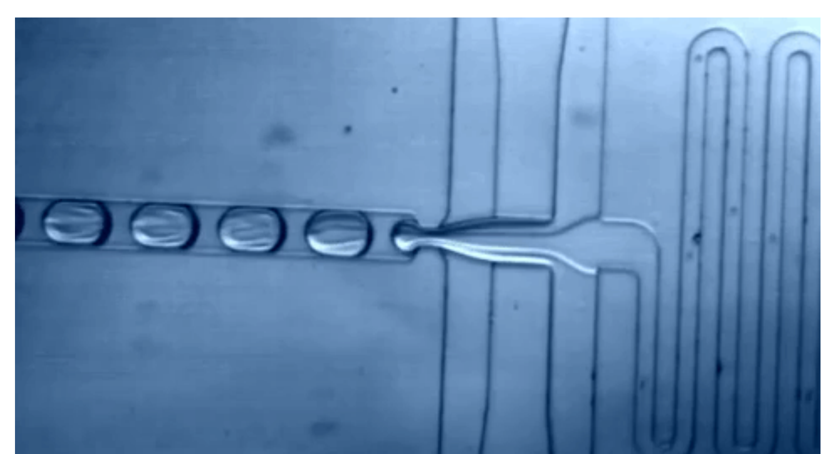
insitro is founded on **data-driven** technologies



High-content microscopy

Machine Learning

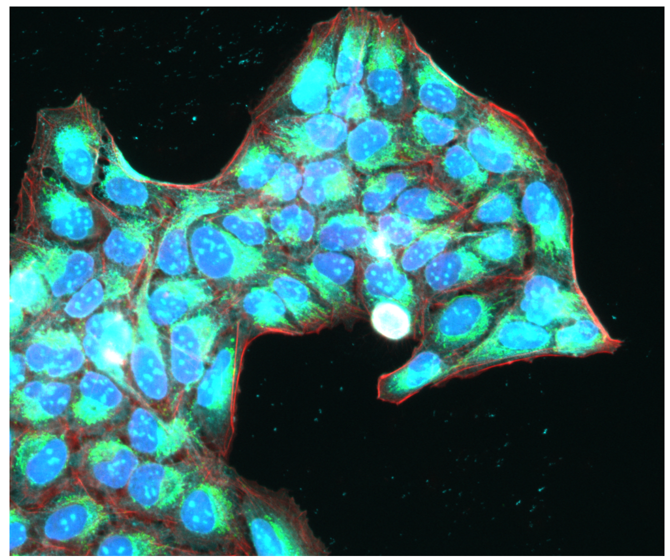




Single-cell RNA sequencing

https://www.youtube.com/watch?v=vL7ptq2Dcf0

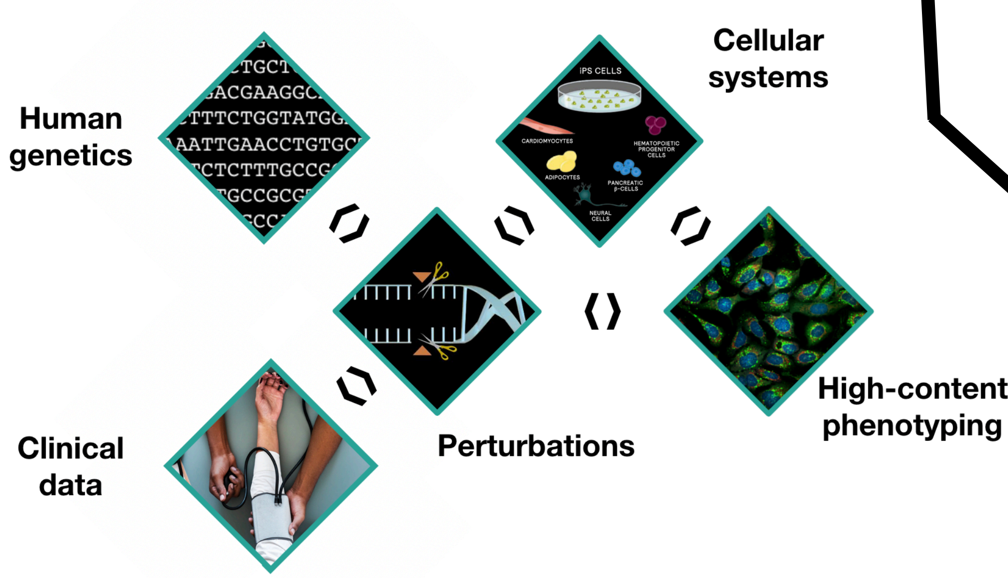
insitro is founded on **data-driven** technologies

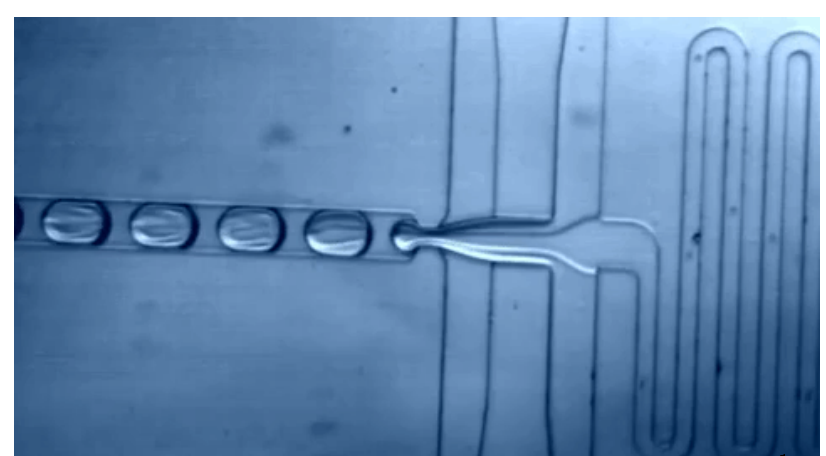


High-content microscopy

Machine Learning

Characterize phenotypes

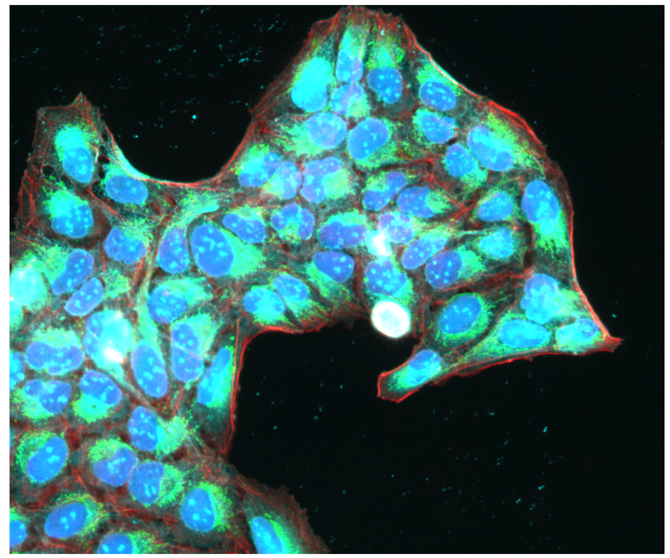




Single-cell RNA sequencing

https://www.youtube.com/watch?v=vL7ptq2Dcf0

insitro is founded on **data-driven** technologies

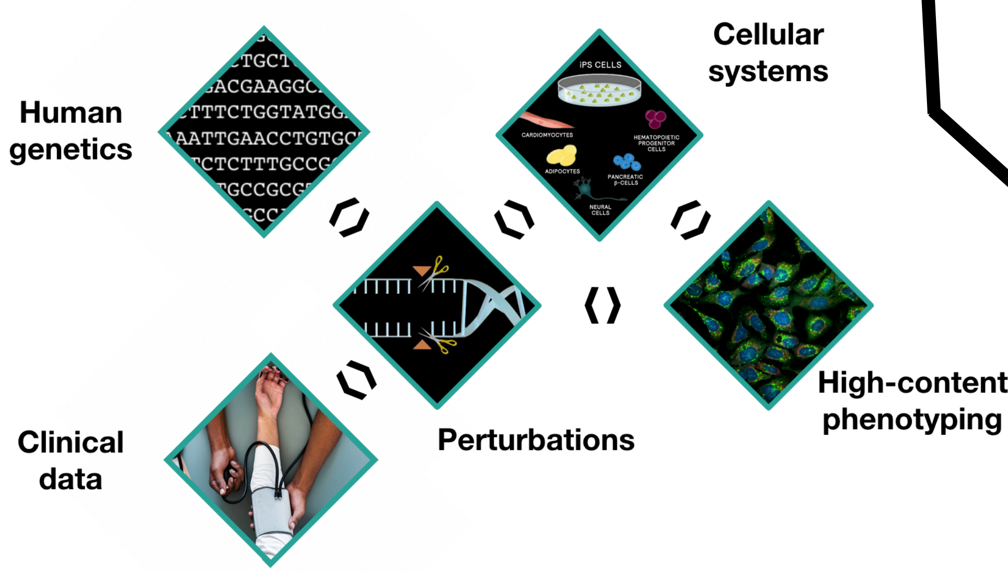


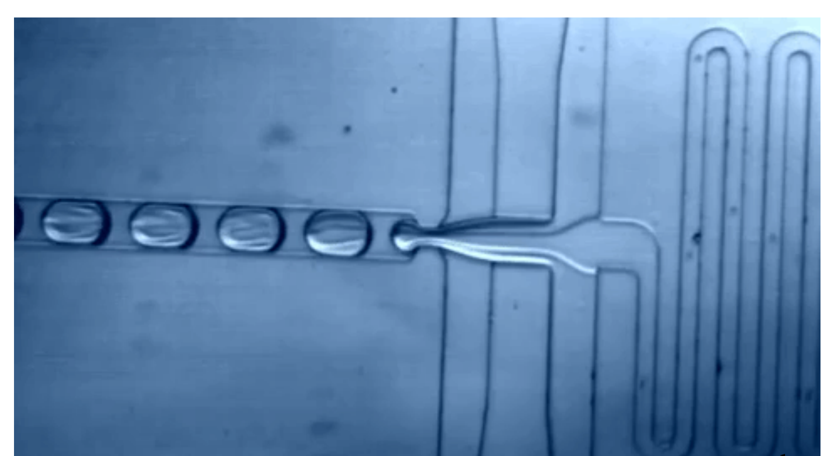
High-content microscopy

Machine Learning

Characterize phenotypes

Understand relationships between data modalities



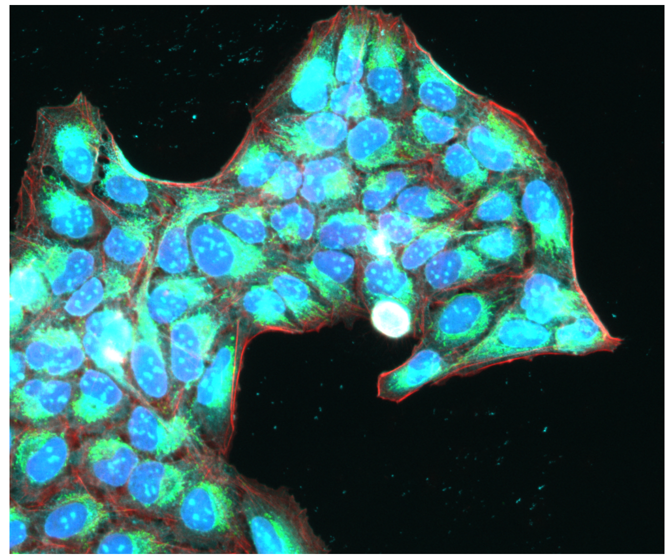


Single-cell RNA sequencing

https://www.youtube.com/watch?v=vL7ptq2Dcf0

insitro is founded on **data-driven** technologies

One experiment can test thousands of biological hypotheses

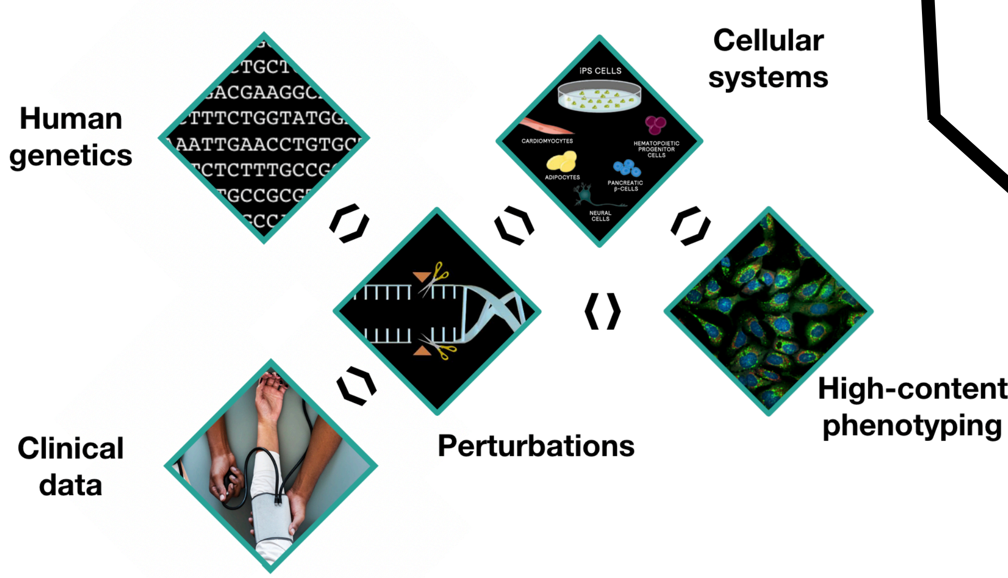


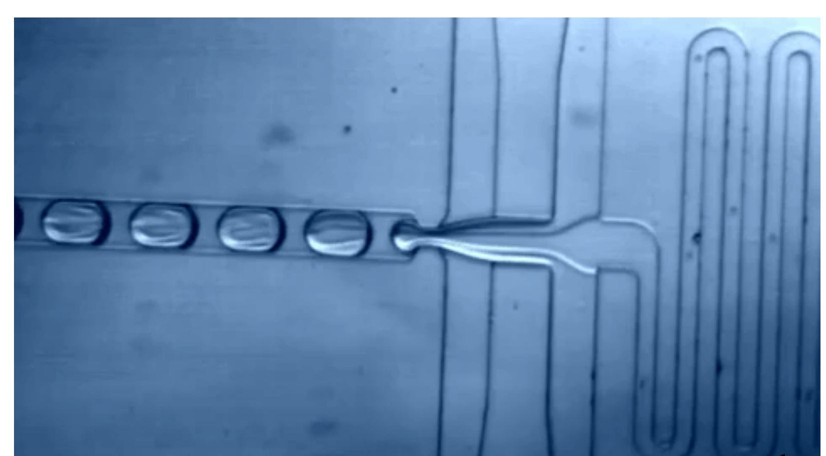
High-content microscopy

Machine Learning

Characterize phenotypes

Understand relationships between data modalities

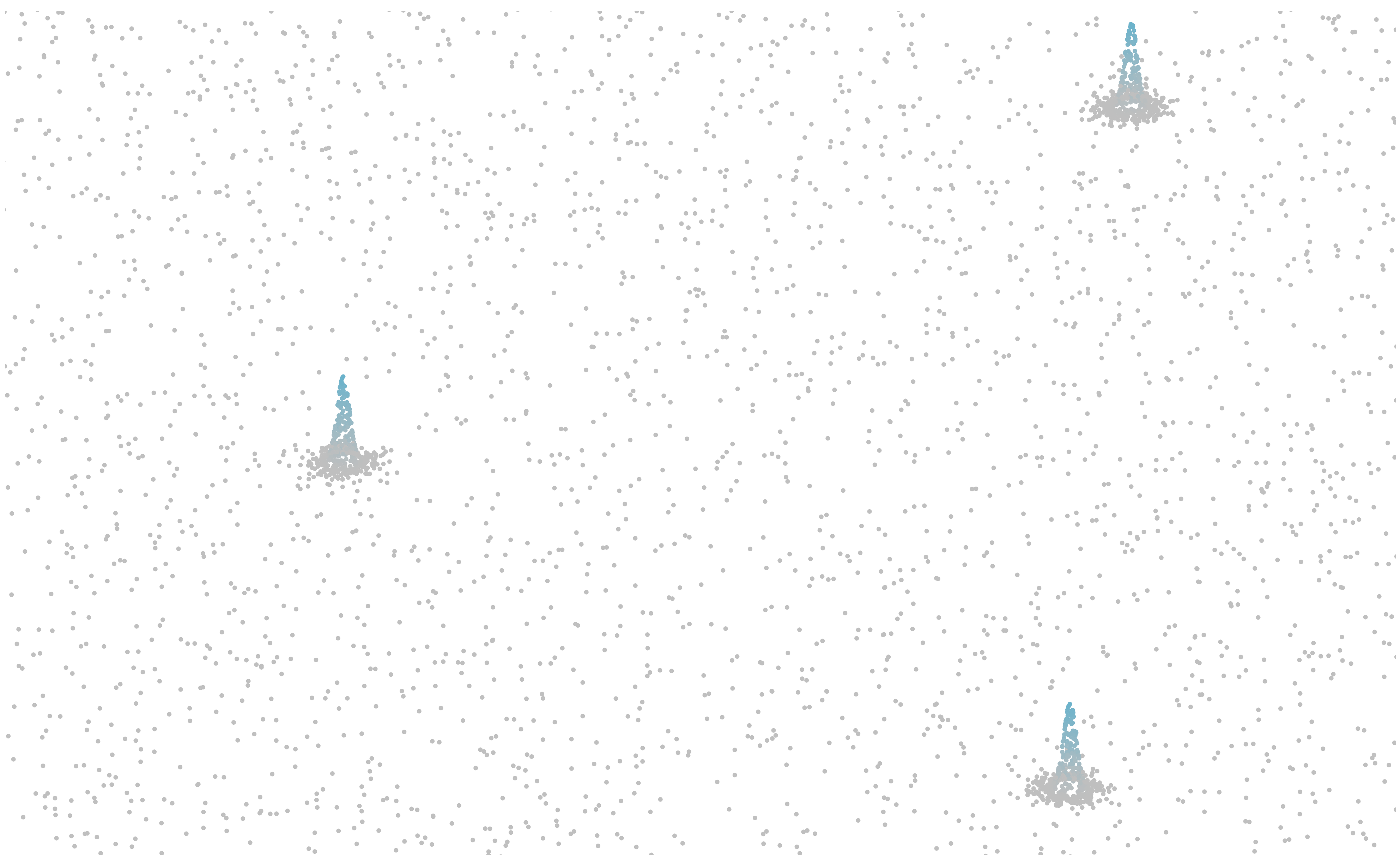




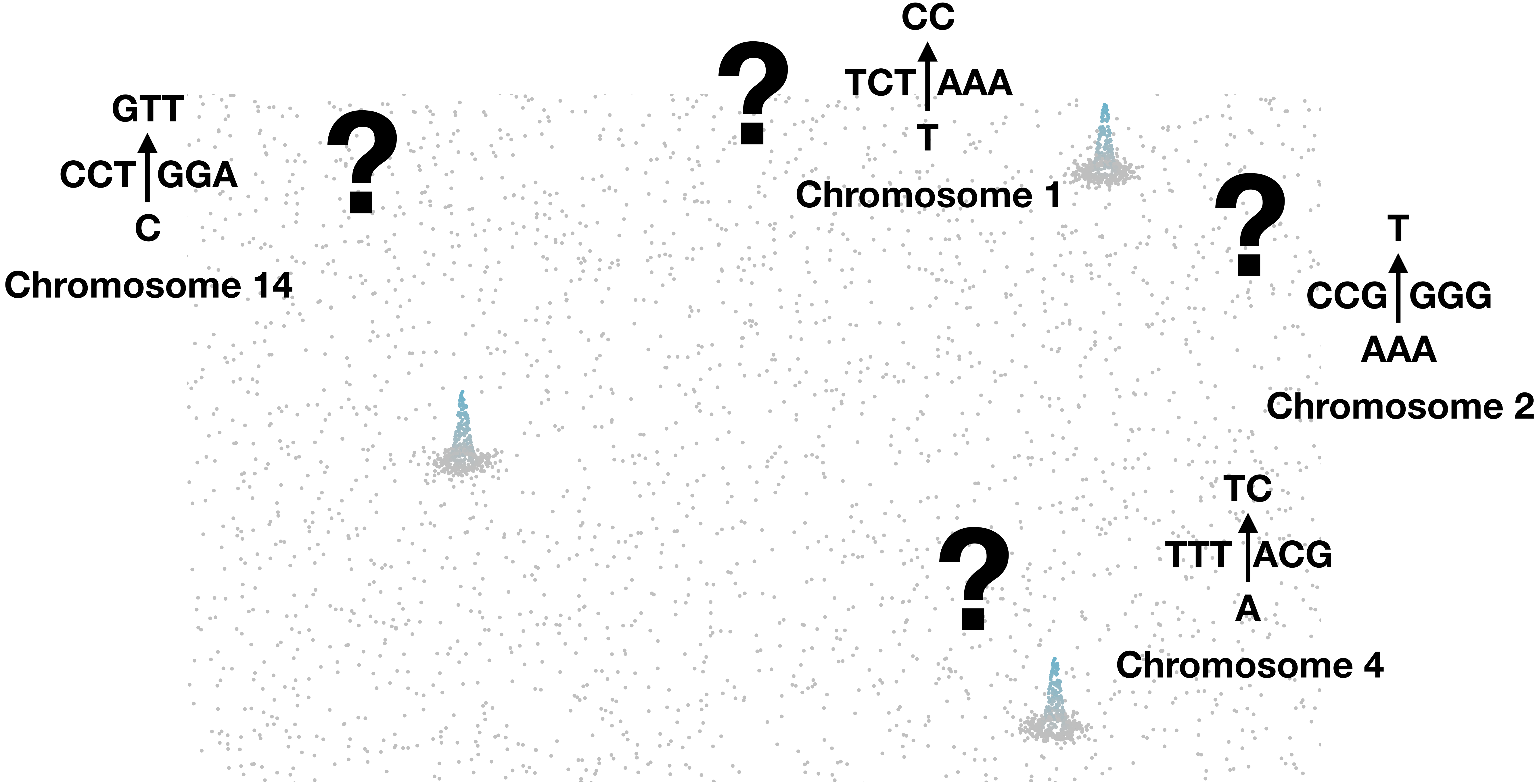
Single-cell RNA sequencing

https://www.youtube.com/watch?v=vL7ptq2Dcf0

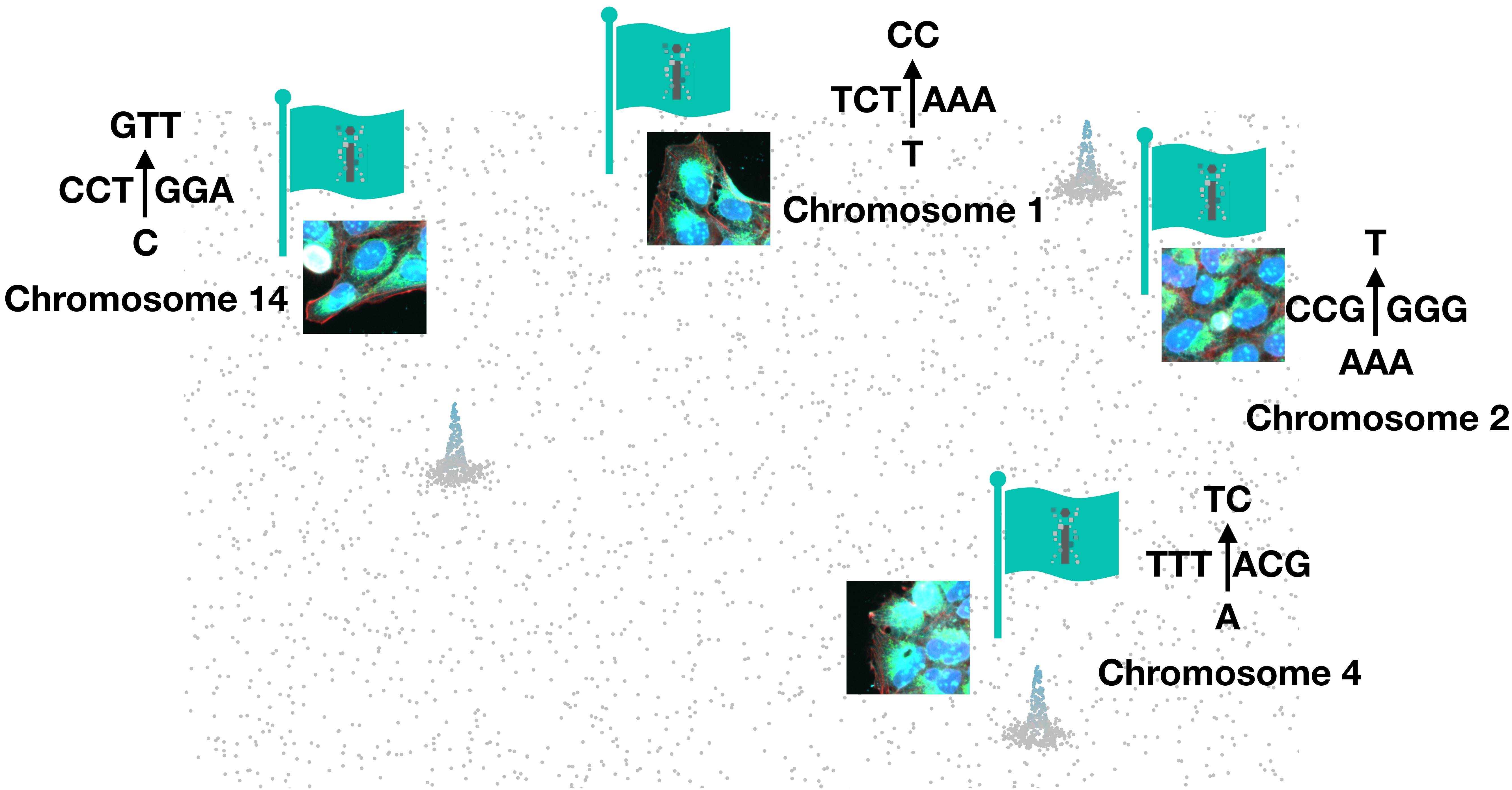
The insitro **bio-data factory** will elucidate the **landscape of disease**



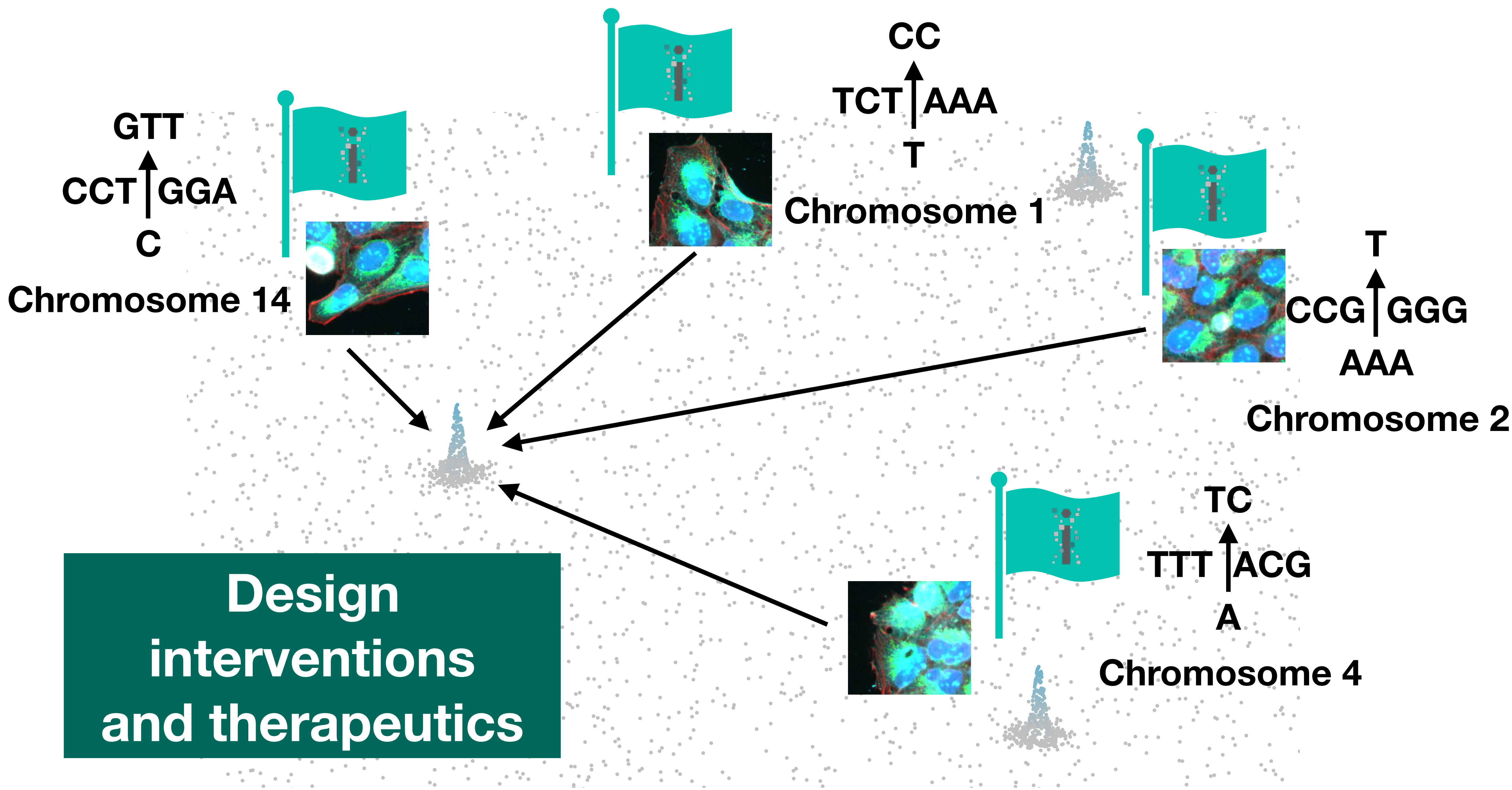
The insitro **bio-data factory** will elucidate the **landscape of disease**



The insitro **bio-data factory** will elucidate the **landscape of disease**



The insitro **bio-data factory** will elucidate the **landscape of disease**



The future is

biology + computation

Thank you!

Thank you!

DOE
CSGF

Thank you!

HARVARD + UCSF

Debora Marks
John Ingraham
Chris Sander
Andrew Kruse
Aashish Manglik
Conor McMahon
June Shin
Aaron Kollasch

Anna Green
Thomas Hopf
Charlotta Scharfe
Benni Schubert
Eli Weinstein
Kelly Brock
Rohan Maddamsetti
David Ding
Hailey Cambra
Agnes Toth-Petroczy
Perry Palmedo
Frank Poelwijk
Nick Gauthier
Jennie Epp

DOE CSGF

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Sam Deutsch

insitro™

Daphne Koller