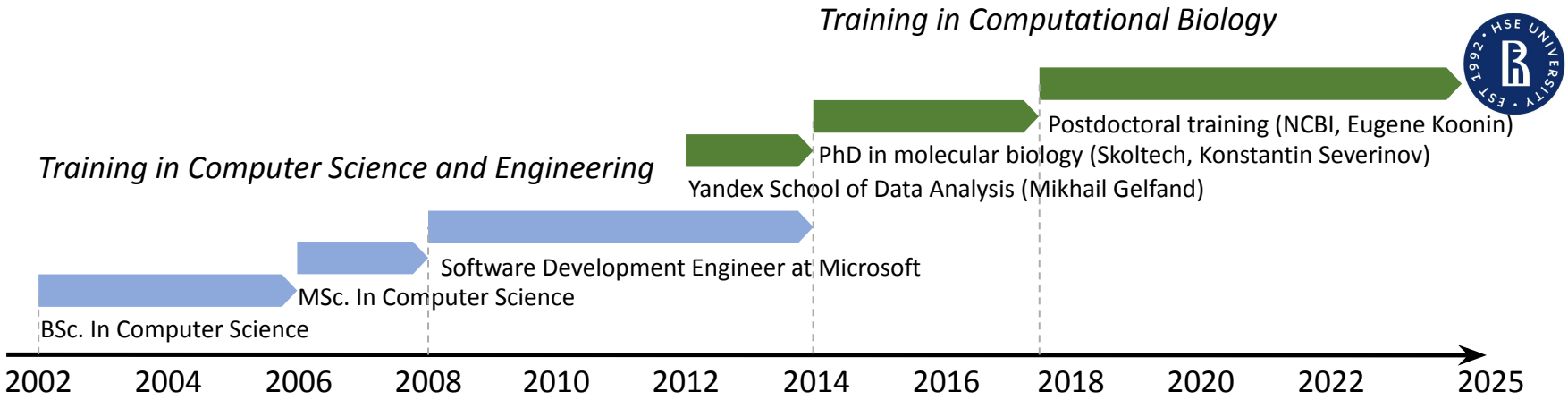
NATIONAL RESEARCH UNIVERSITY

# Машинное обучение для открытия и создания новых генетических систем

Шмаков Сергей

III научная конференция ФКН
Вороново, НИУ ВШЭ, 28.10.2025

# Training & Background



# CRISPR-Cas revolutionized biomedical research


**CRISPR-Cas applications since 2012**

**Treatments:**

- Clinical Trials for: Sickle cell disease and beta-thalassemia, Cancers, Hereditary blindness, Transthyretin amyloidosis, Cardiovascular diseases
- Personalized Medicine, treatments for unique genetic conditions

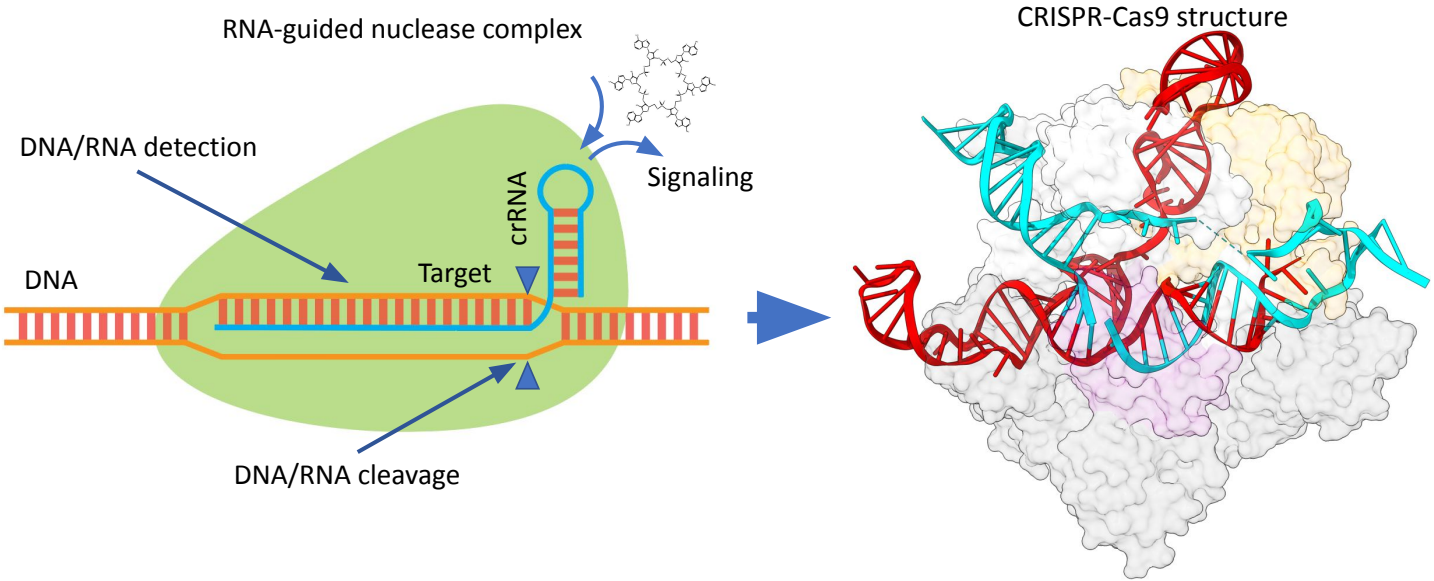
**Biotechnology:**

- Crop improvement, Livestock breeding, Biofuel production, Biomaterials production, Diagnostics

**Science:**

- Gene function, Gene expression, Disease modelling, Drug discovery

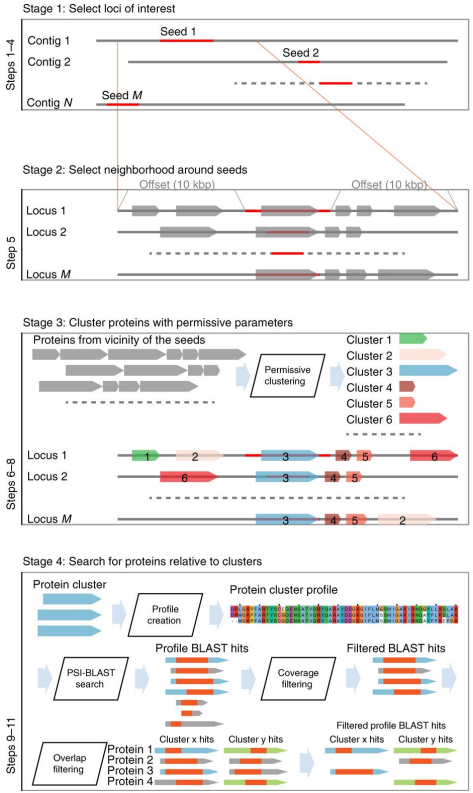
# Cas9: from defense system to genome editor

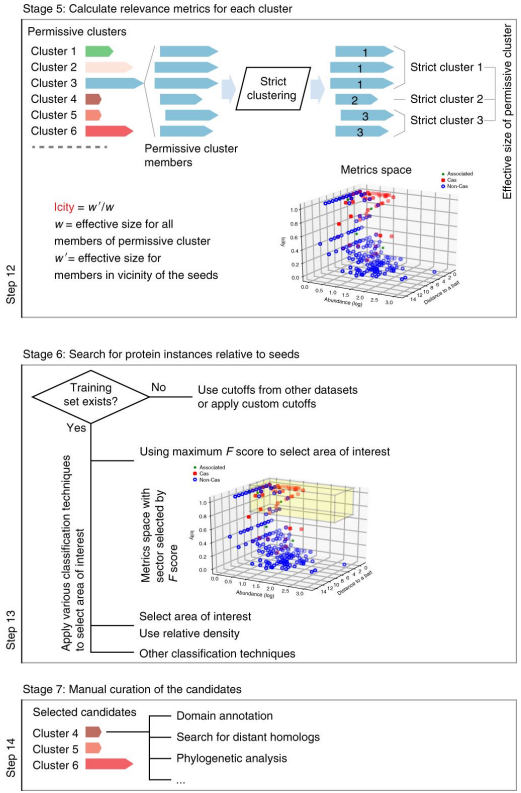





Bravo JPK, Liu MS, Hibshman GN, Dangerfield TL, Jung K, McCool RS, Johnson KA, Taylor DW. Structural basis for mismatch surveillance by CRISPR-Cas9. Nature. 2022 Mar;603(7900):343-347. doi: 10.1038/s41586-022-04470-1

# Computational approach for novel gene systems discovery




Candidate scoring using Gene Functional Linkage Measure (GFLM):

$$\text{Icity} = \frac{f(w')}{f(w)}$$

w - all cluster hits
w′ - all cluster hits close to a bait
$f$ - cluster effective size (number of different proteins)

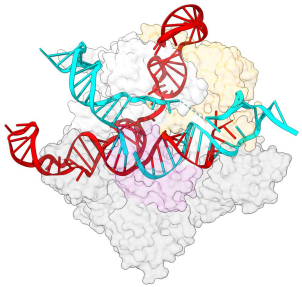
We developed the protocol for the prediction of functionally linked genes


**Systematic prediction of functionally linked genes in bacterial and archaeal genomes**

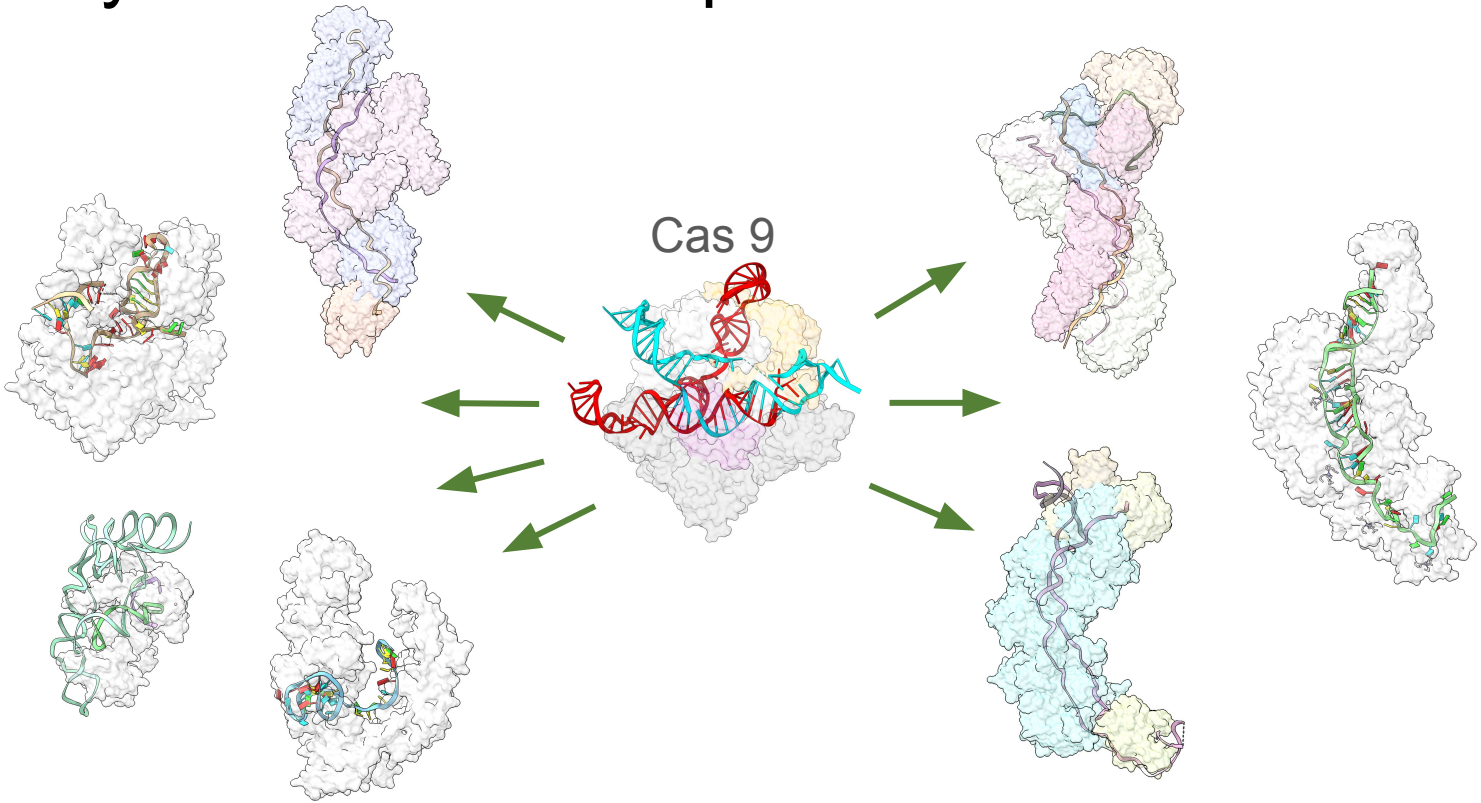
Sergey A. Shmakov[1,2], Guilhem Faure[1,3], Kira S. Makarova[1], Yuri I. Wolf[1], Konstantin V. Severinov[2,4,5] and Eugene V. Koonin[1]*

# Diversity of CRISPR-Cas protein families

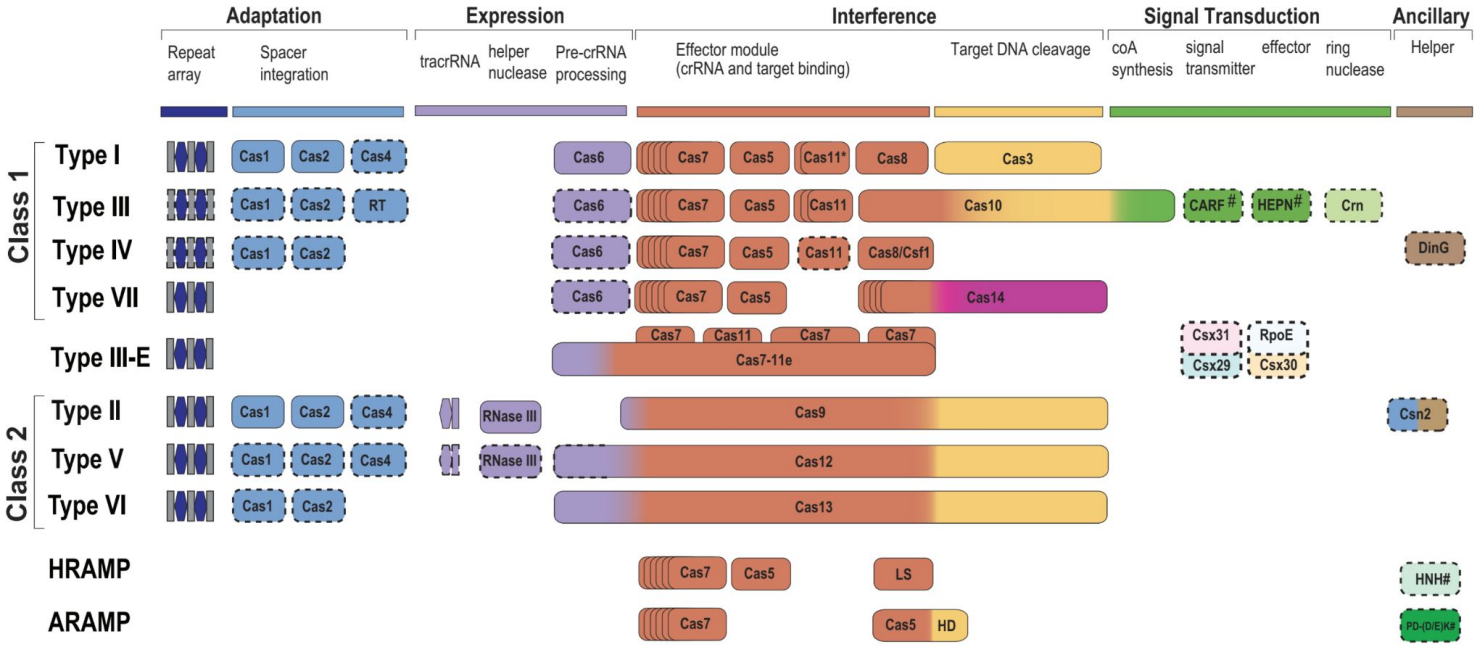
Cas 9



# Diversity of CRISPR-Cas protein families

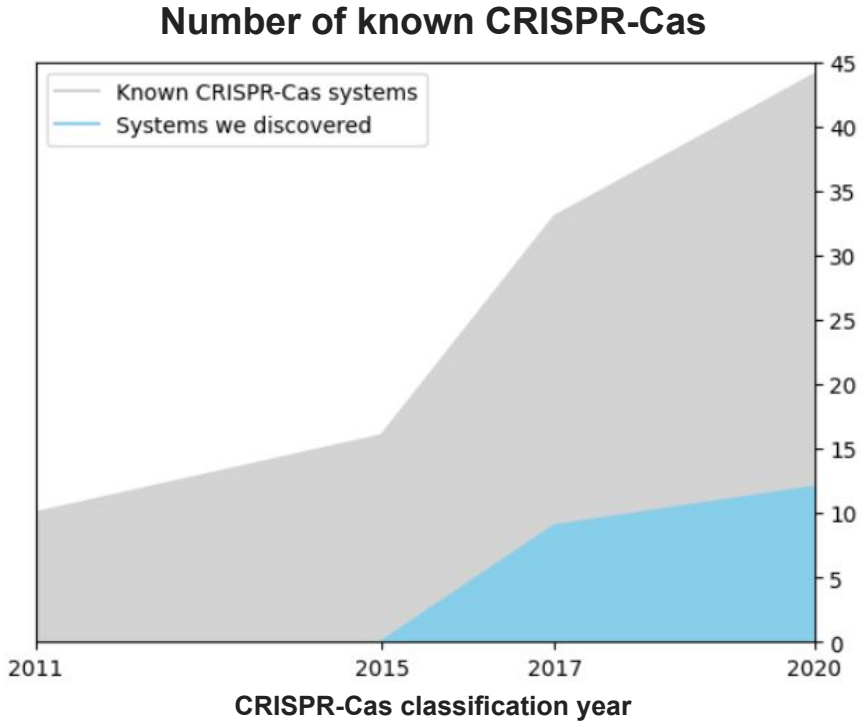


# Evolutionary classification of CRISPR-Cas systems




accepted in Nature Microbiology

# Summary for past research

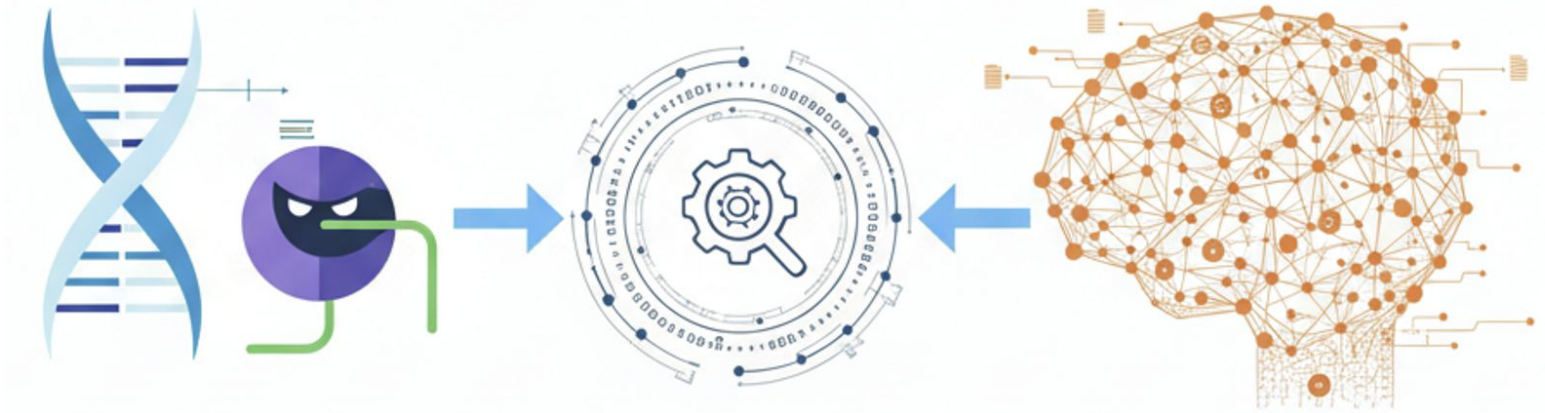



**Achievements:**

- Developed novel computational methods to systematically search for functionally associated prokaryotic genes
- Identified 12 novel CRISPR-Cas systems, plus 2 in preparation
- Characterized novel CRISPR-Cas functions, including widespread regulatory activity
- Provided classification of CRISPR-Cas systems
- Patents filed for application of Cas9, Cas12, Cas13 proteins

I have published 32 papers, which have been widely accepted by the scientific community, accumulating more than 11,000 citations

# Research plans



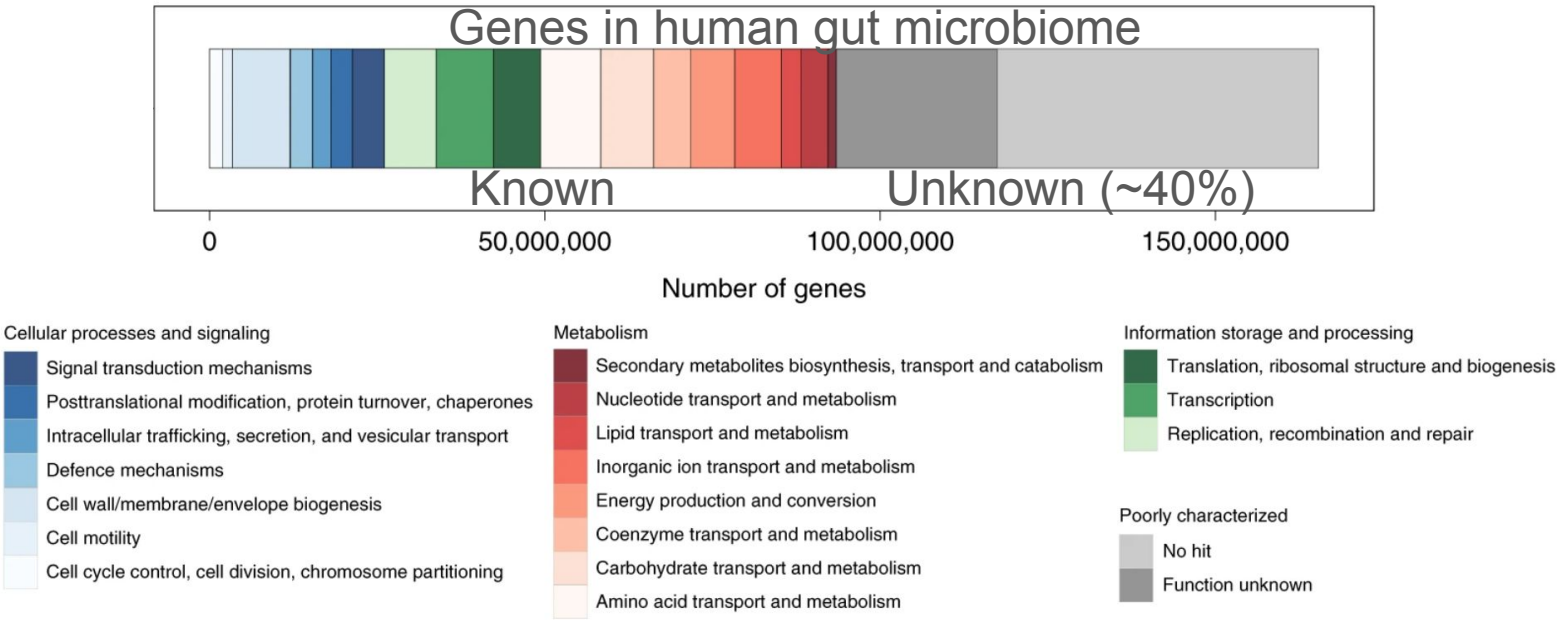
Revolution in biology

ML methods for biology and data-driven discovery

Revolution in ML

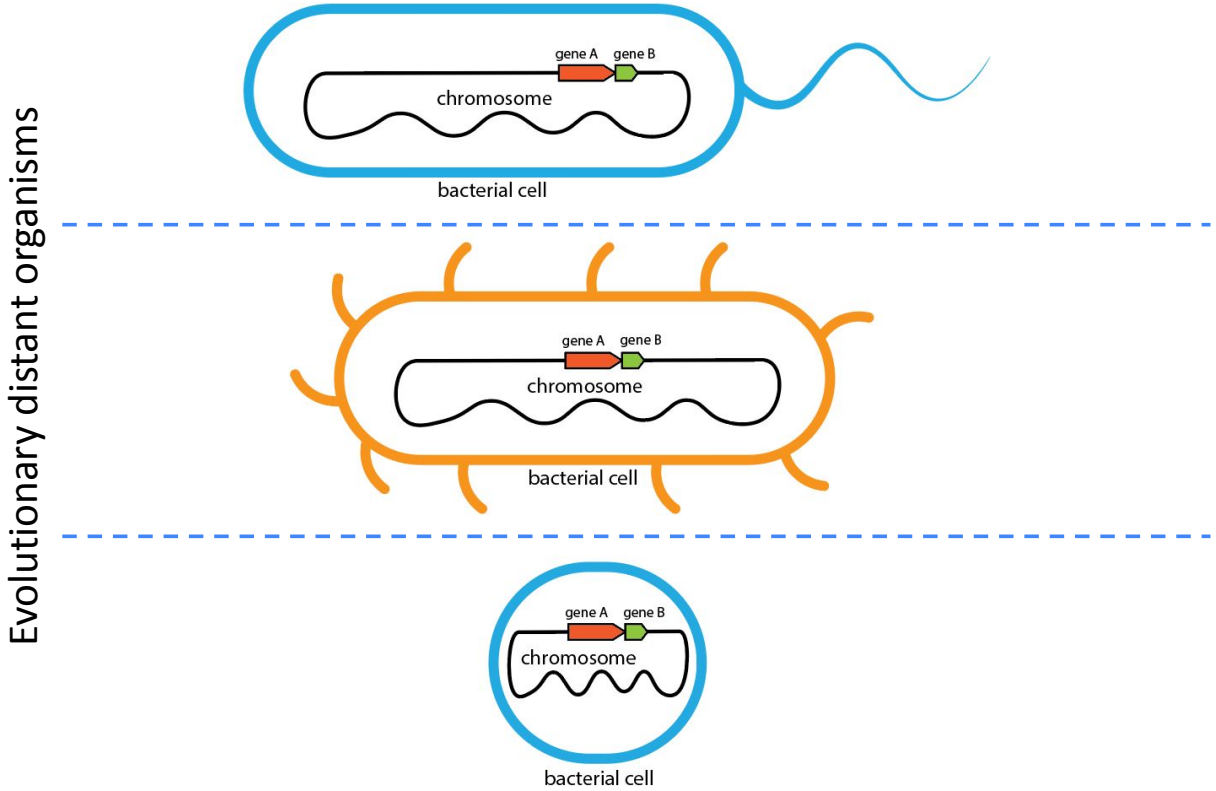
Generated with Imagen 3

# Persistent gaps in biological knowledge

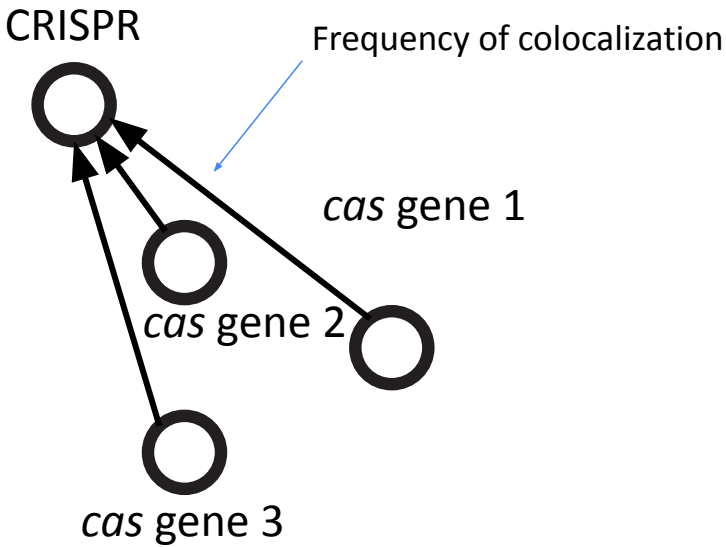



Almeida A., et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021 Jan;39(1):105-114. doi: 10.1038/s41587-020-0603-3. Epub 2020 Jul 20.

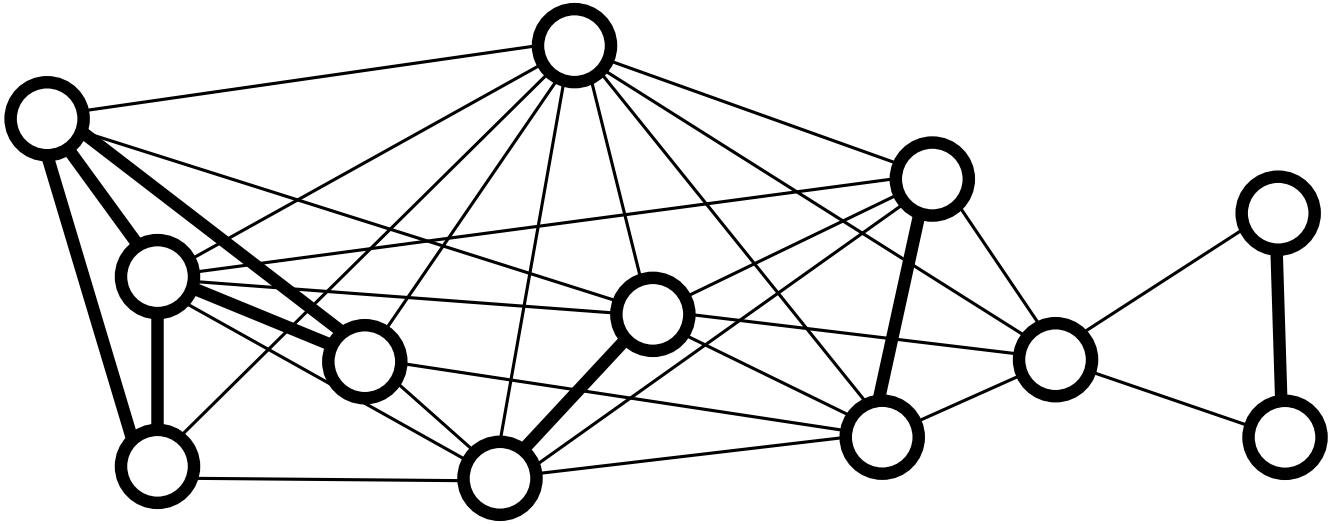
# Context is the key

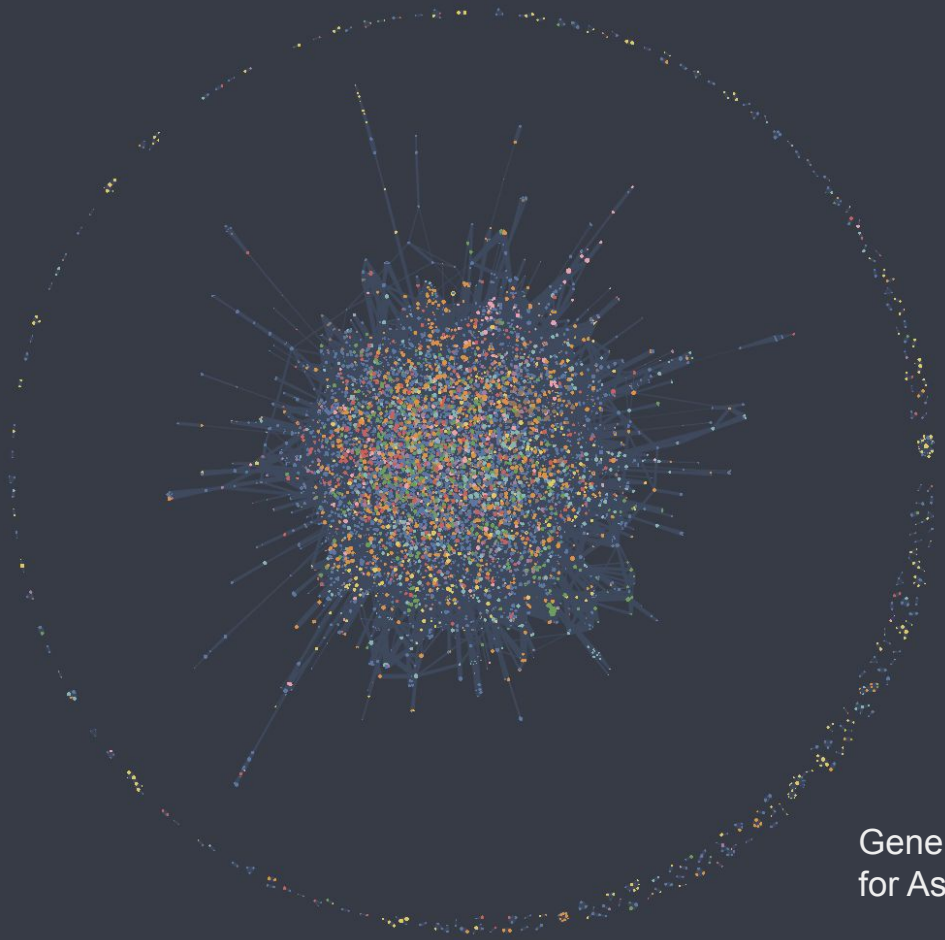


# Insights into novel gene discovery methodologies



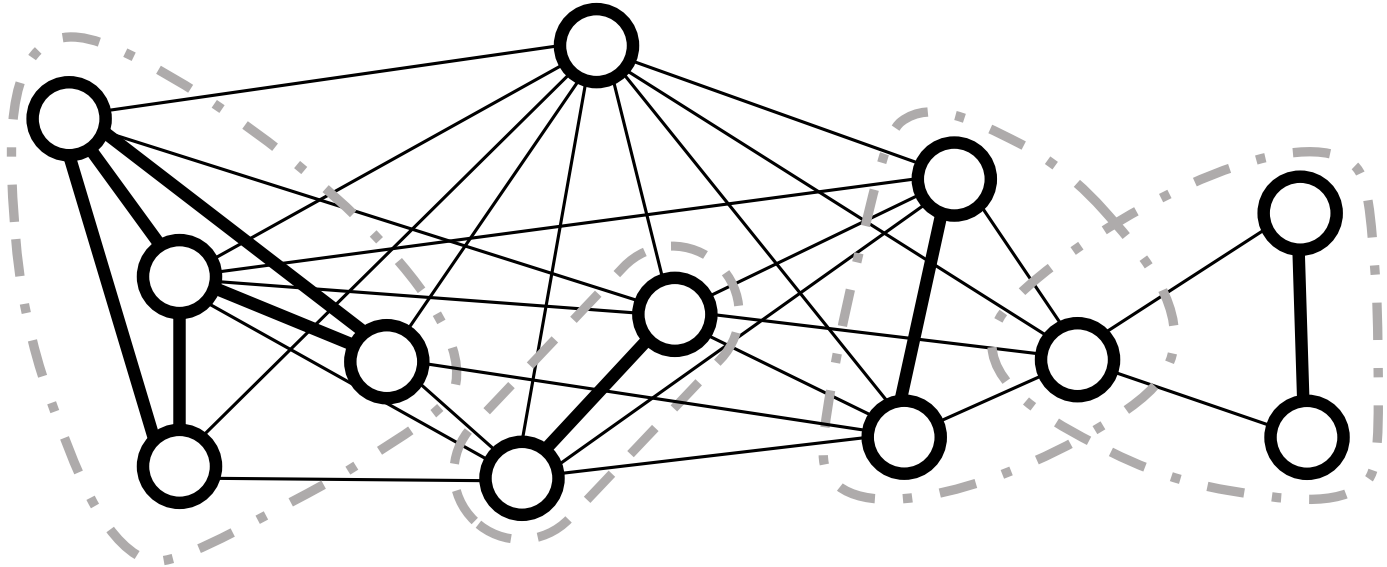
# Insights into novel gene discovery methodologies



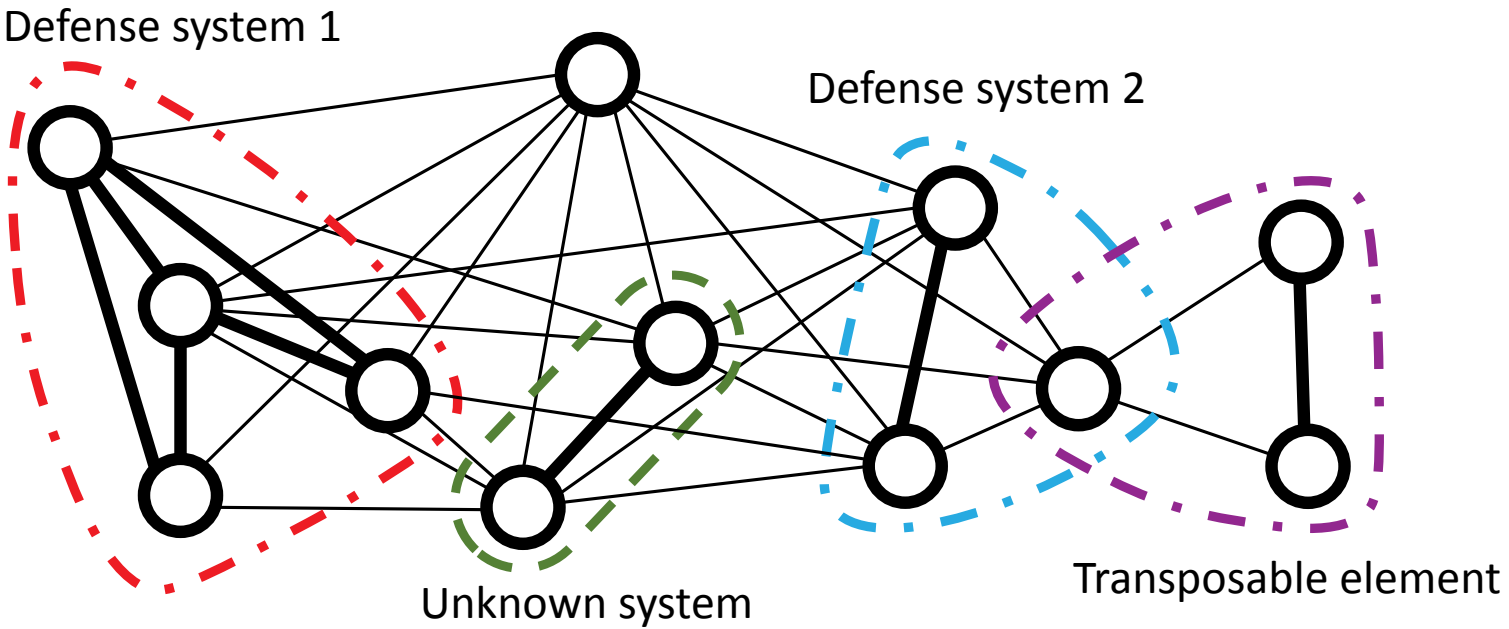


Gene association network
for Asgard archaea

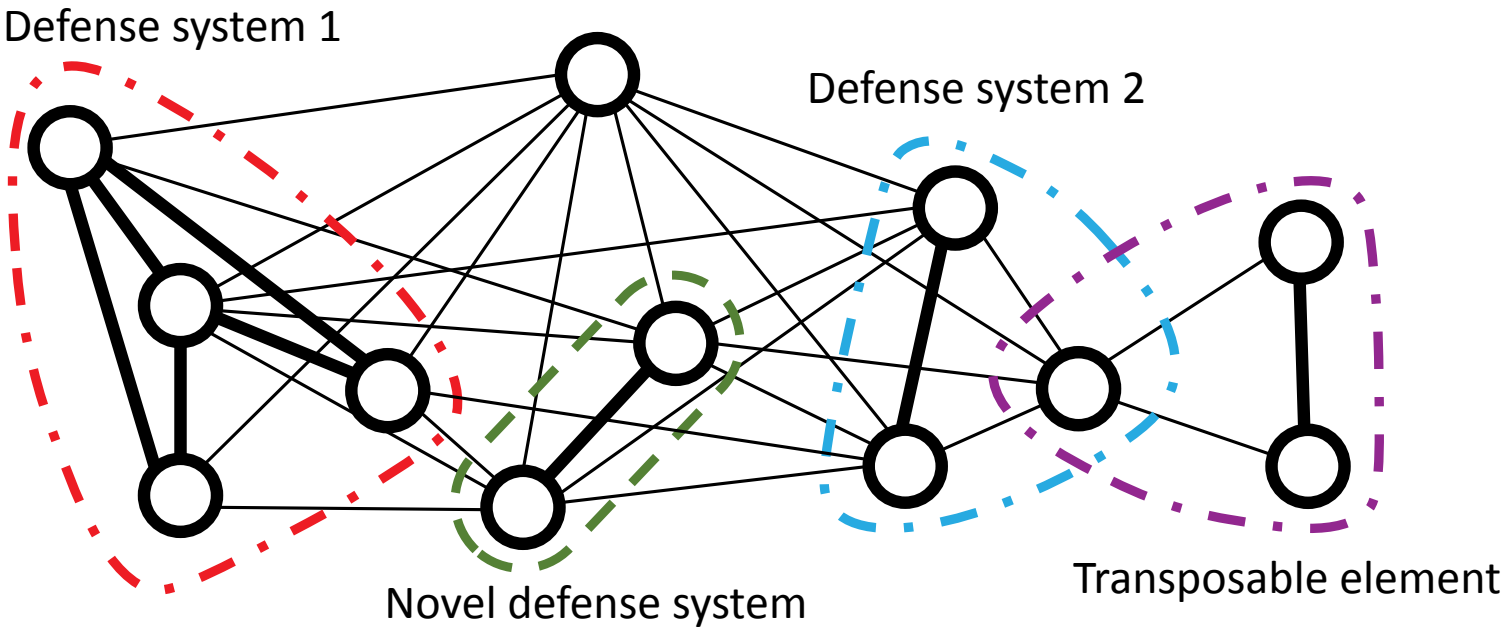
# Insights into novel gene discovery methodologies



# Insights into novel gene discovery methodologies



# Insights into novel gene discovery methodologies



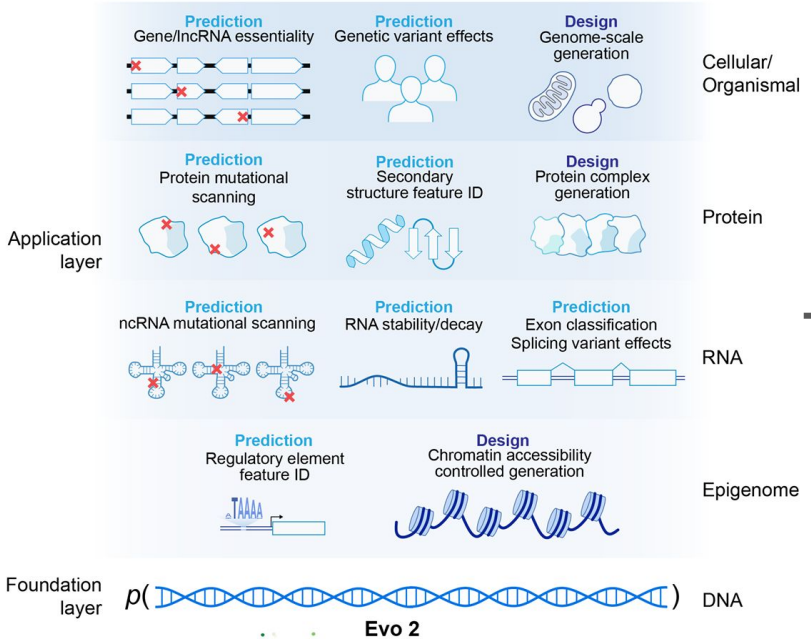
# Network approach for gene discovery

Expected results:

- Novel methodologies
- Gene systems database
- Novel genes and gene systems

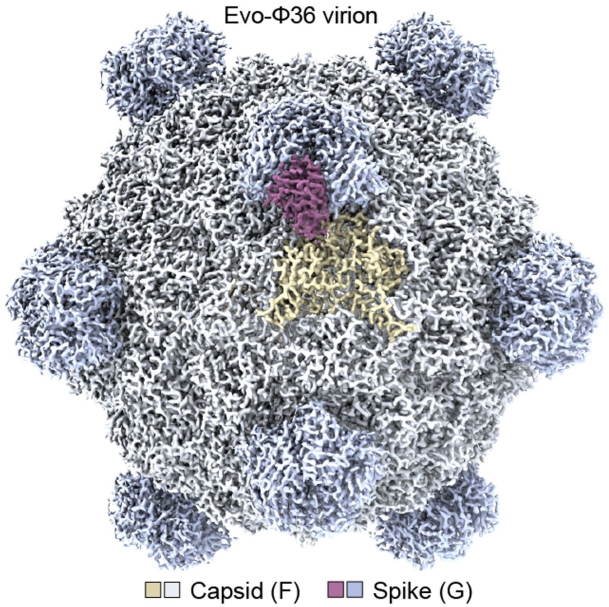
# Generative models for biological sequences

## Evo 2 - genomic foundation model




Genome modeling and design across all domains of life with Evo 2
Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, Brian L. Hie
bioRxiv 2025.02.18.638918

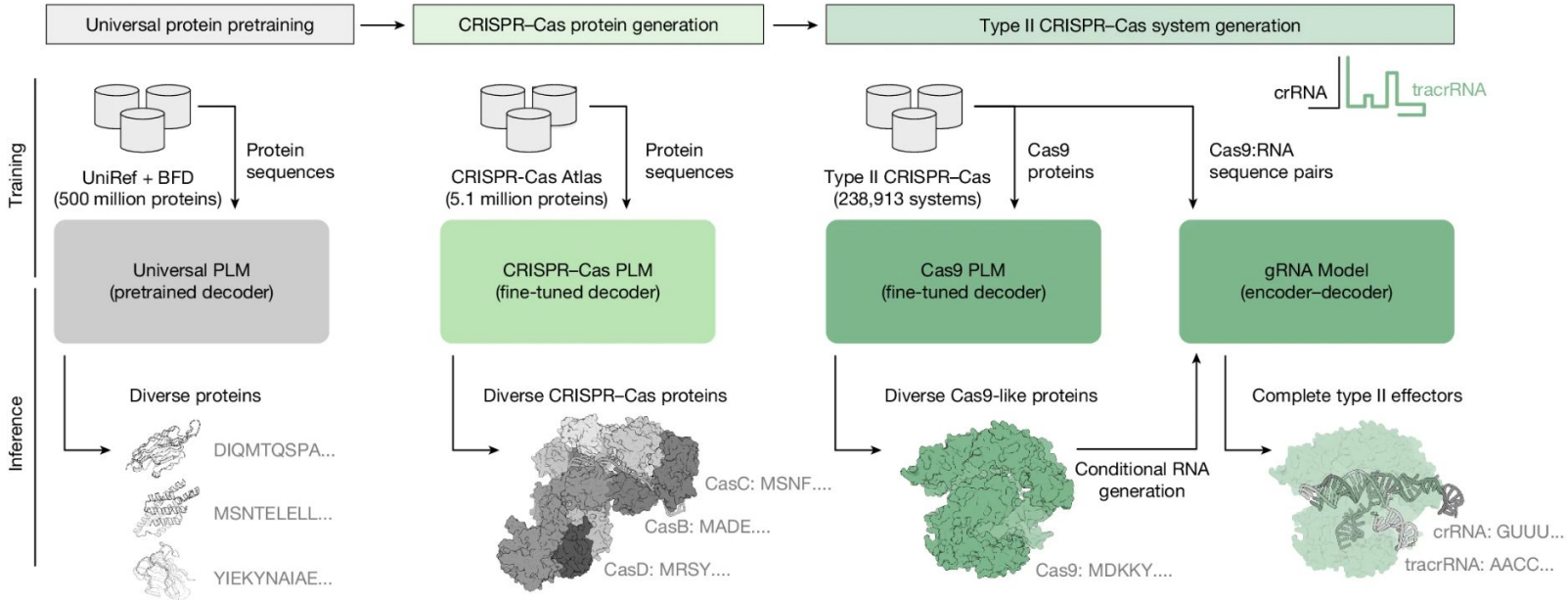
## First AI generated functional virus




Generative design of novel bacteriophages with genome language models
Samuel H. King, Claudia L. Driscoll, David B. Li, Daniel Guo, Aditi T. Merchant, Garyk Brixi, Max E. Wilkinson, Brian L. Hie
bioRxiv 2025.09.12.675911

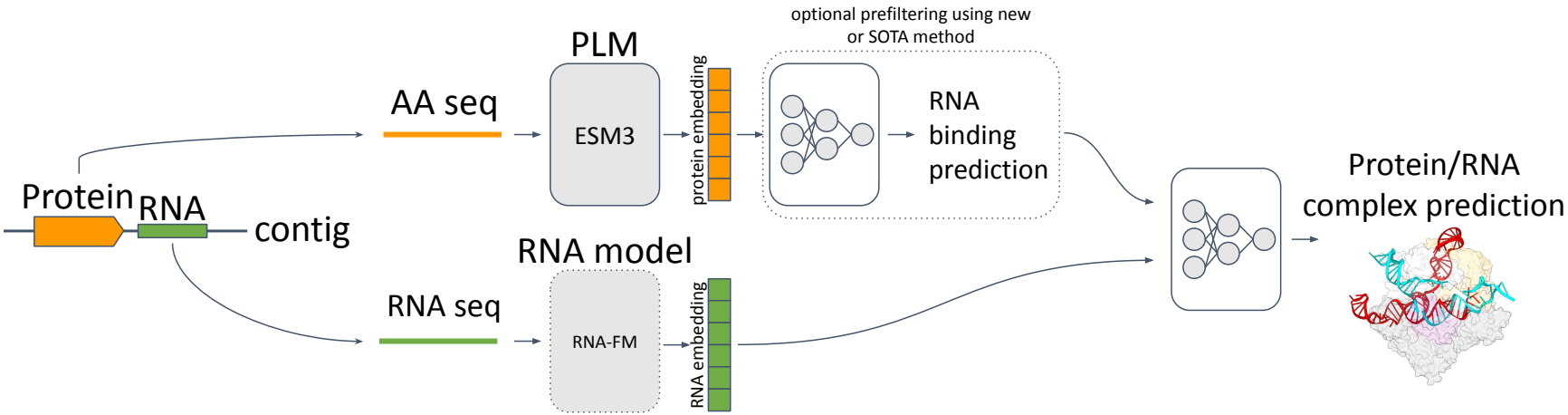
# Generative models for biological sequences

Language-modelling approach to design CRISPR–Cas proteins




Ruffolo, J.A., Nayfach, S., Gallagher, J. et al. Design of highly functional genome editors by modelling CRISPR–Cas sequences. Nature 645, 518–525 (2025).

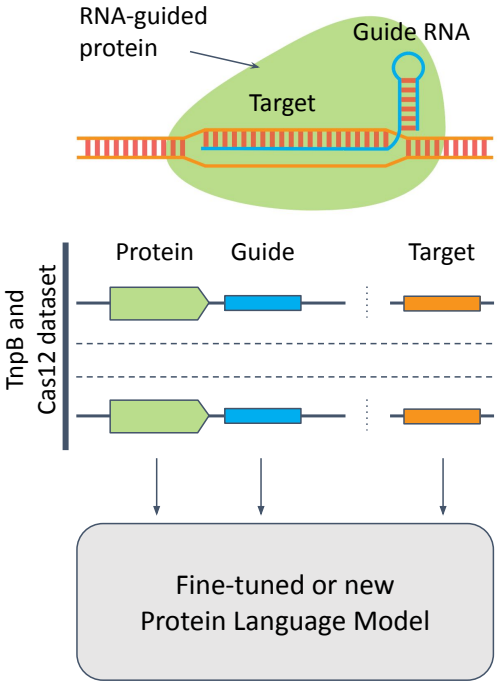
# Search for novel RNA-guided systems



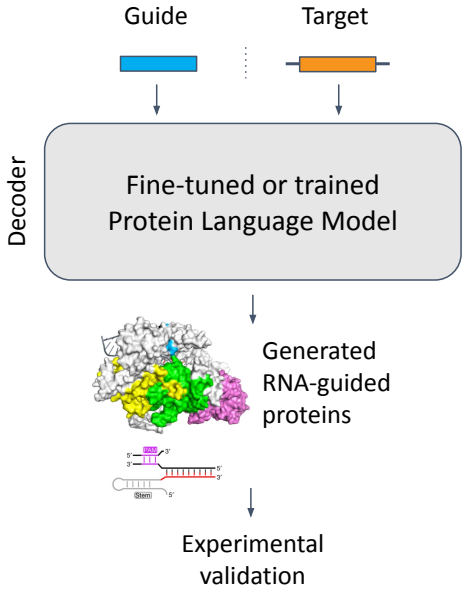

PLM - Protein Language Model

# Generative models for RNA-guided proteins

## 1: Assemble protein-guide-target dataset and fine-tune or train PLM




## 2: Generate new RNA-guided proteins



# AI for discovery and design of genome editors

Expected results:

- Comprehensive database of RNA-guided systems
- Platform for generation of RNA-guided proteins
- Novel ML models

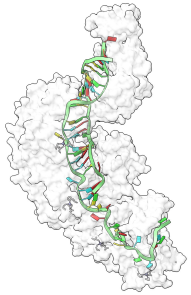
# Research plans

Project areas:

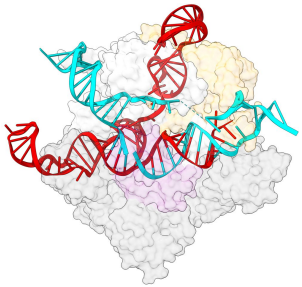
- Methodology and tools development for computational biology
- ML models to study natural genes and to generate synthetic proteins
- Biological data analyses

# Diversity of CRISPR-Cas protein families

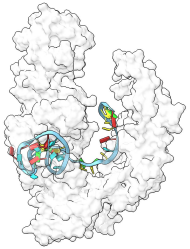
Cas7-11



Cas 9

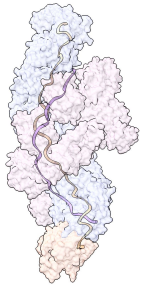


Cas12a

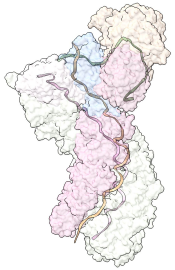


Cas12f

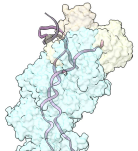
Type VII



Type I



Type IV



Cas13a