Generative models of catalytic RNAs

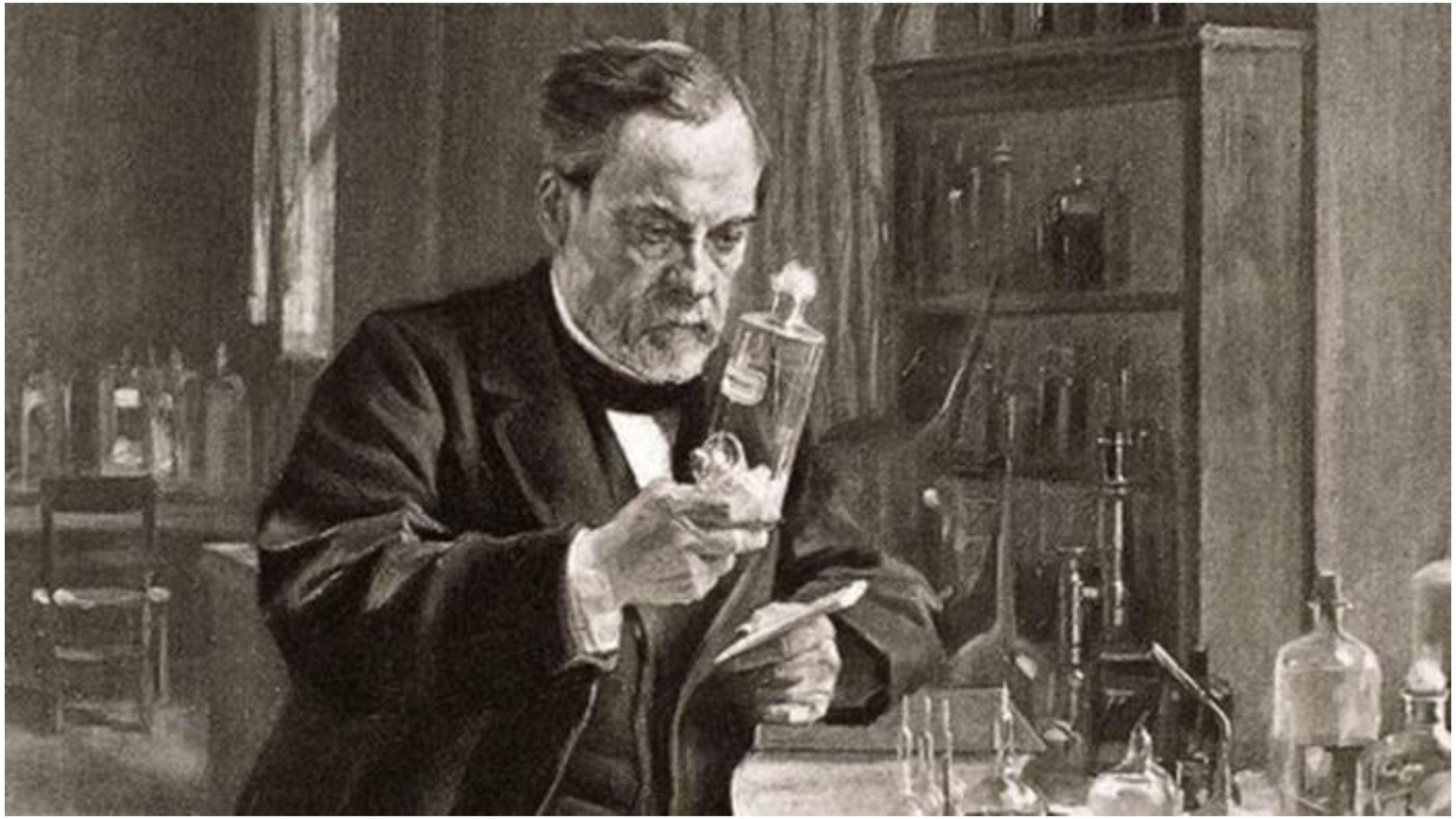
AI Chemist school
Philippe NGHE
04 03 2025

ESPCI PARIS | PSL

ÉCOLE POLYTECHNIQUE

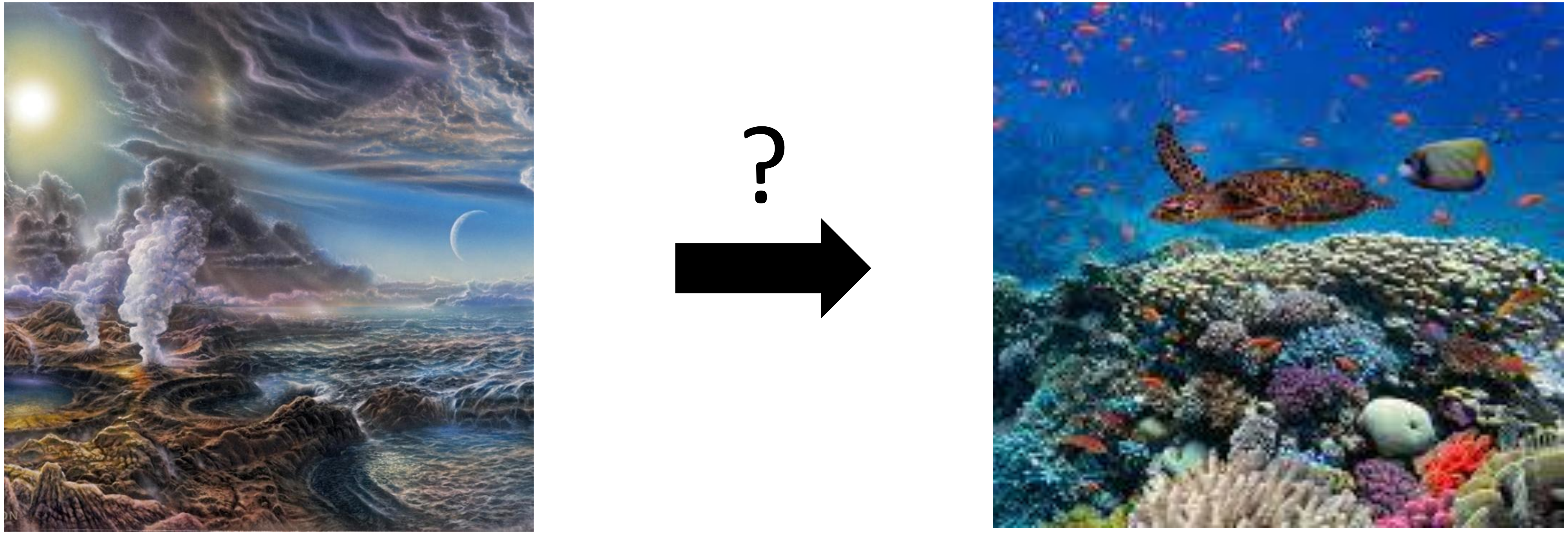
INSTITUT POLYTECHNIQUE DE PARIS

THE ORIGIN OF LIFE QUESTION



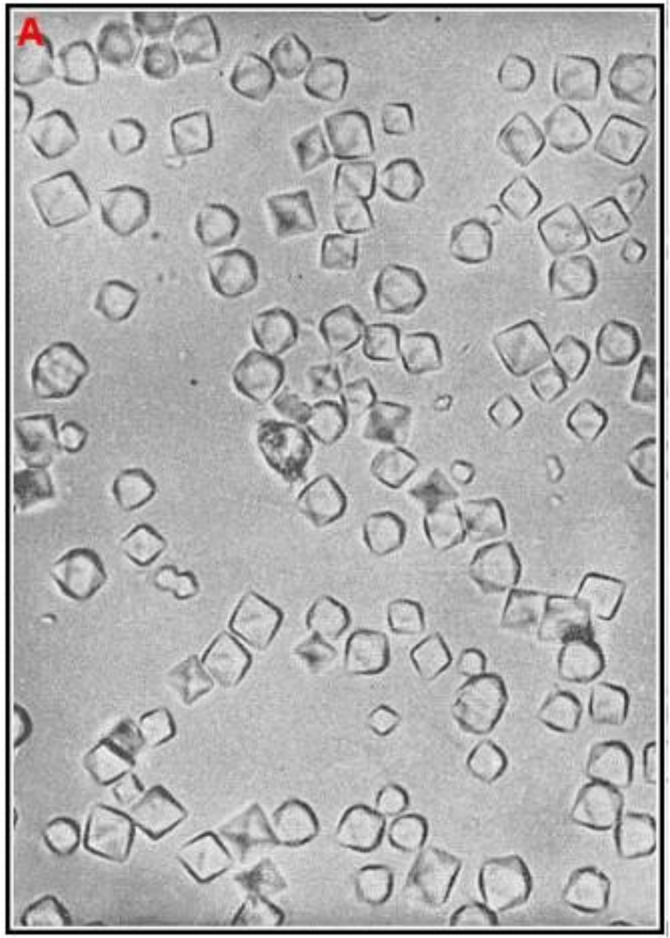
Pasteur 1864

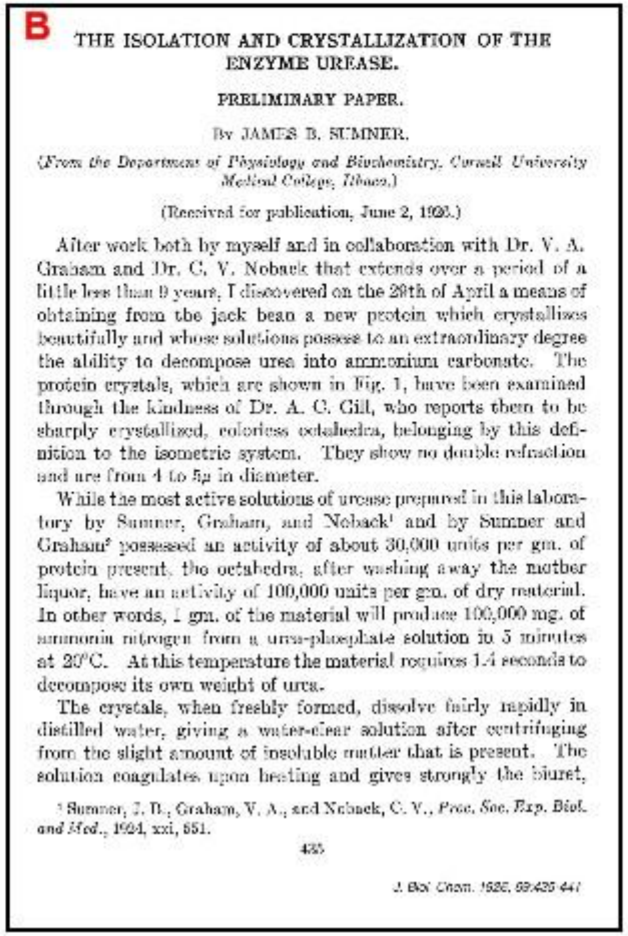
Abiogenesis



Enzymes are proteins

1926: James Sumner crystallizes ureases





THE ISOLATION AND CRYSTALLIZATION OF THE ENZYME UREASE.

PRELIMINARY PAPER.

By JAMES B. SUMNER.

(From the Department of Physiology and Biochemistry, Cornell University Medical College, Ithaca.)

(Received for publication, June 2, 1926.)

After work both by myself and in collaboration with Dr. V. A. Graham and Dr. C. V. Noback that extends over a period of a little less than 9 years, I discovered on the 28th of April a means of obtaining from the jack bean a new protein which crystallises beautifully and whose solutions possess to an extraordinary degree the ability to decompose urea into ammonium carbonate. The protein crystals, which are shown in Fig. 1, have been examined through the kindness of Dr. A. C. Gill, who reports them to be sharply crystallised, colorless octahedra, belonging by this definition to the isometric system. They show no double refraction and are from 4 to 5μ in diameter.

While the most active solutions of urease prepared in this laboratory by Sumner, Graham, and Noback[1] and by Sumner and Graham[2] possessed an activity of about 30,000 units per gm. of protein present, the octahedra, after washing away the mother liquor, have an activity of 100,000 units per gm. of dry material. In other words, 1 gm. of the material will produce 100,000 mg. of ammonia nitrogen from a urea-phosphate solution in 5 minutes at 20°C. At this temperature the material requires 1.4 seconds to decompose its own weight of urea.

The crystals, when freshly formed, dissolve fairly rapidly in distilled water, giving a water-clear solution after centrifuging from the slight amount of insoluble matter that is present. The solution coagulates upon heating and gives strongly the biuret,

---

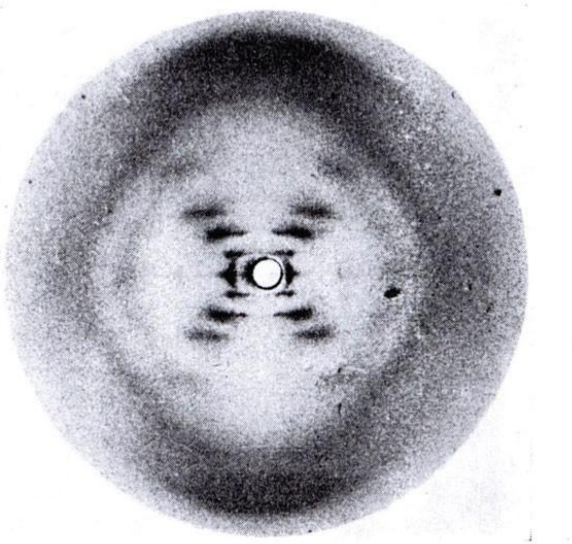
[1] Sumner, J. B., Graham, V. A., and Noback, C. V., *Proc. Soc. Exp. Biol. and Med.*, 1924, xxi, 551.

435

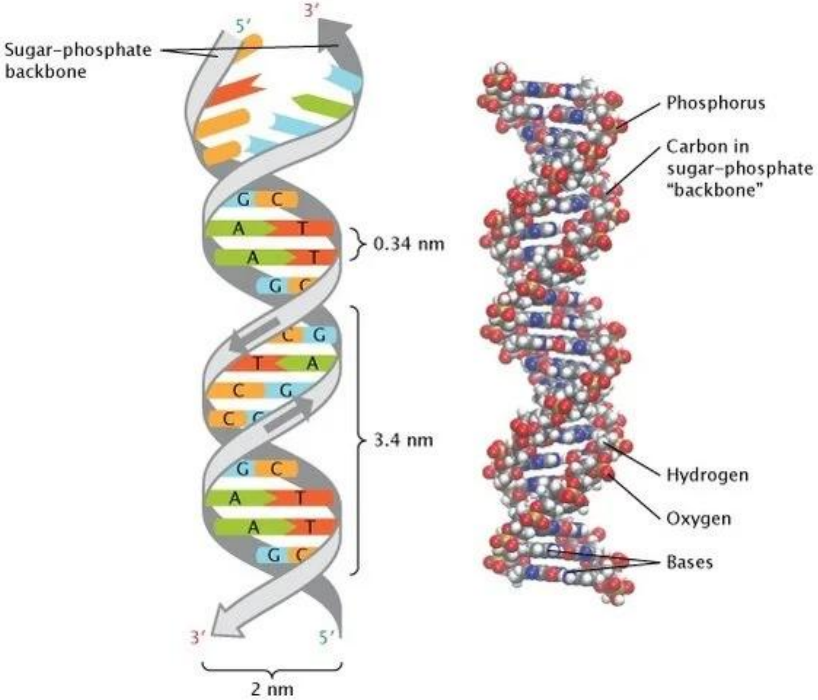
J. Biol. Chem. 1926, 69:435-441

DNA carries genetic information

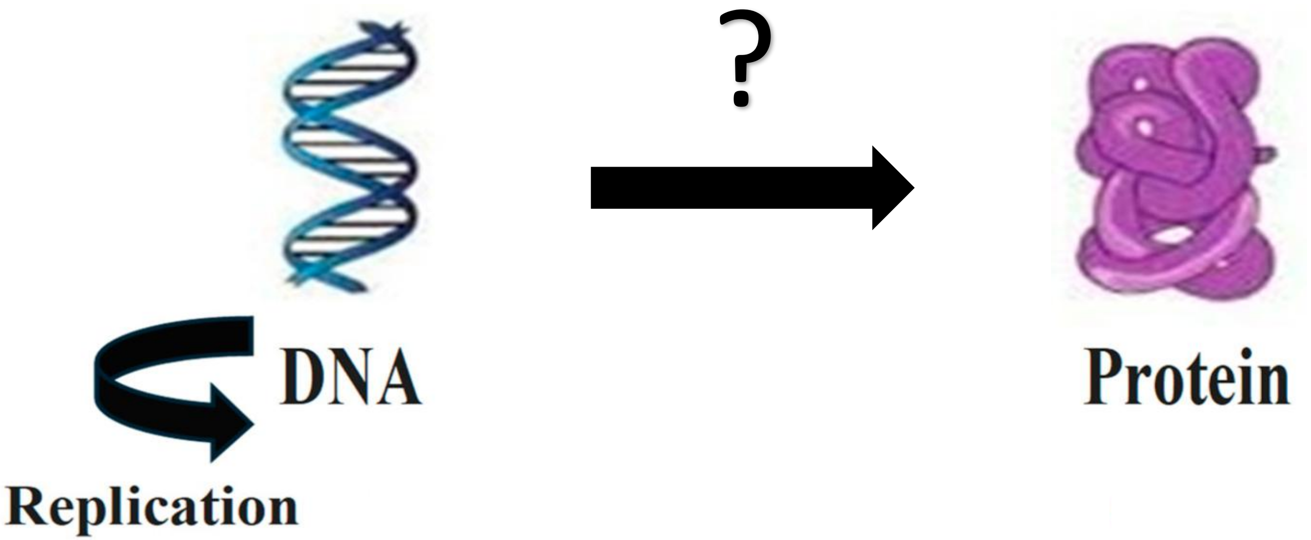
1953: elucidation of DNA structure



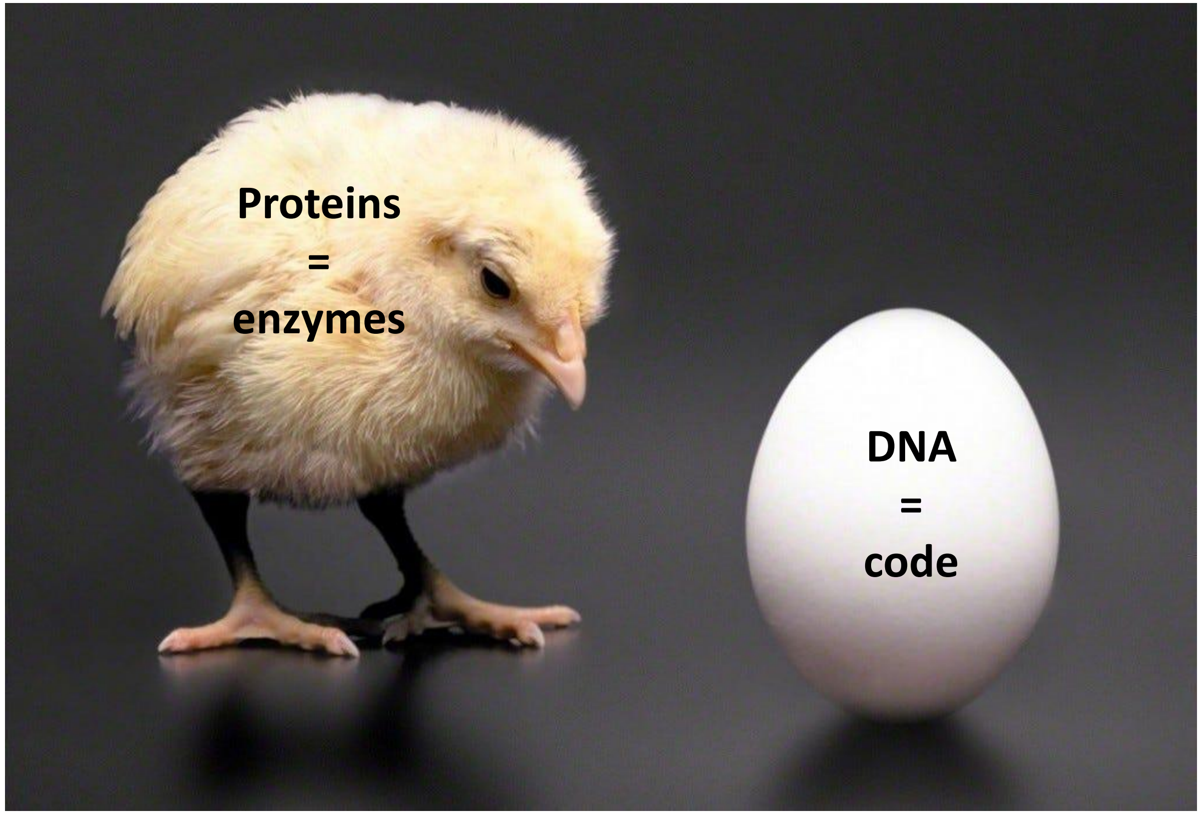
Francis Crick James Watson Maurice Wilkins Rosalind Franklin



The missing link

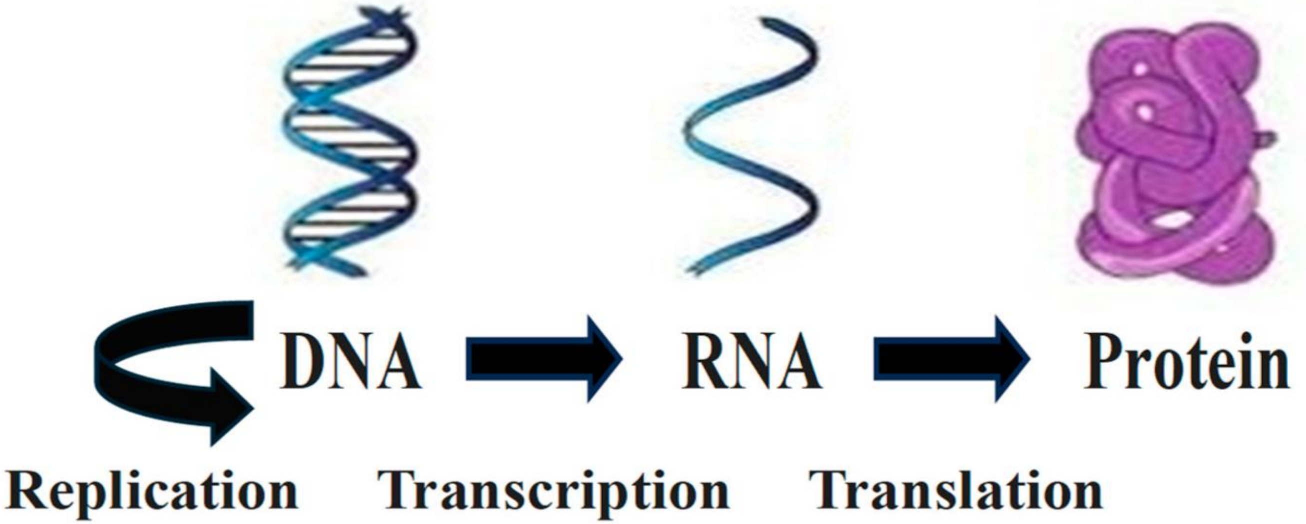


Can RNA solve the origin of life paradox?



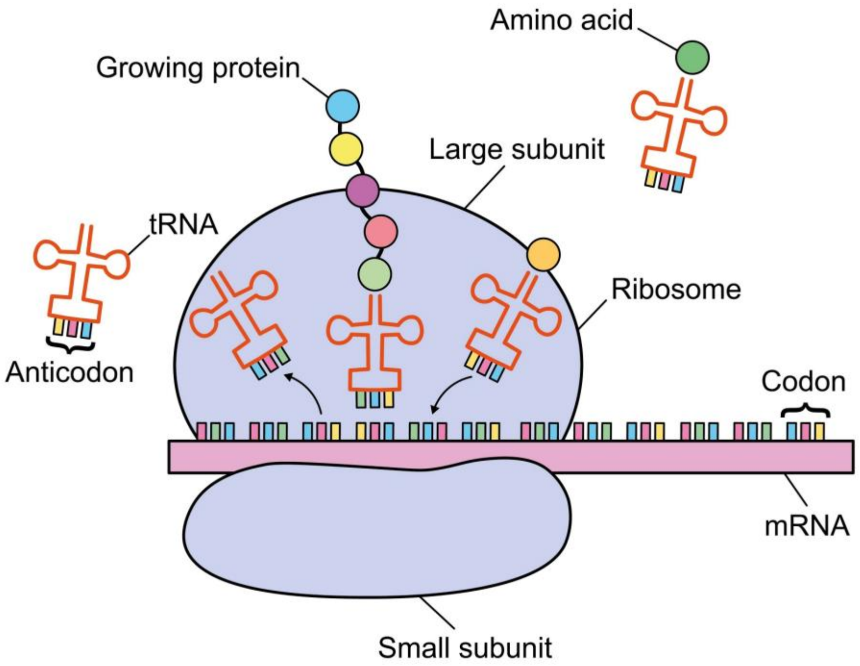
Crick, F. H. (1968). The origin of the genetic code. *Journal of molecular biology*, *38*(3), 367-379.

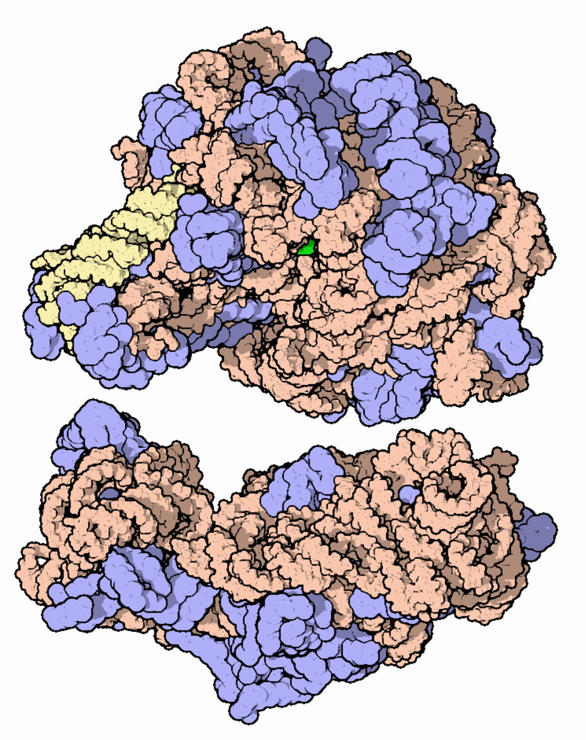
The missing link



1961: discovery of messenger RNA by Gros, Jacob, Monod (Institut Pasteur)

Translation is an RNA machinery



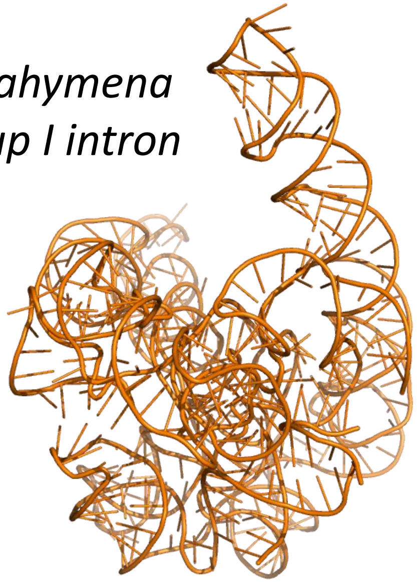


Von Böhlen, K., Makowski, I., Hansen, H. A. S., Bartels, H., Berkovitch-Yellin, Z., Zaytzev-Bashan, A., ... & Yonath, A. (1991). Characterization and preliminary attempts for derivatization of crystals of large ribosomal subunits from Haloarcula marismortui diffracting to 3 Å resolution. *Journal of molecular biology*, *222*(1), 11-15.

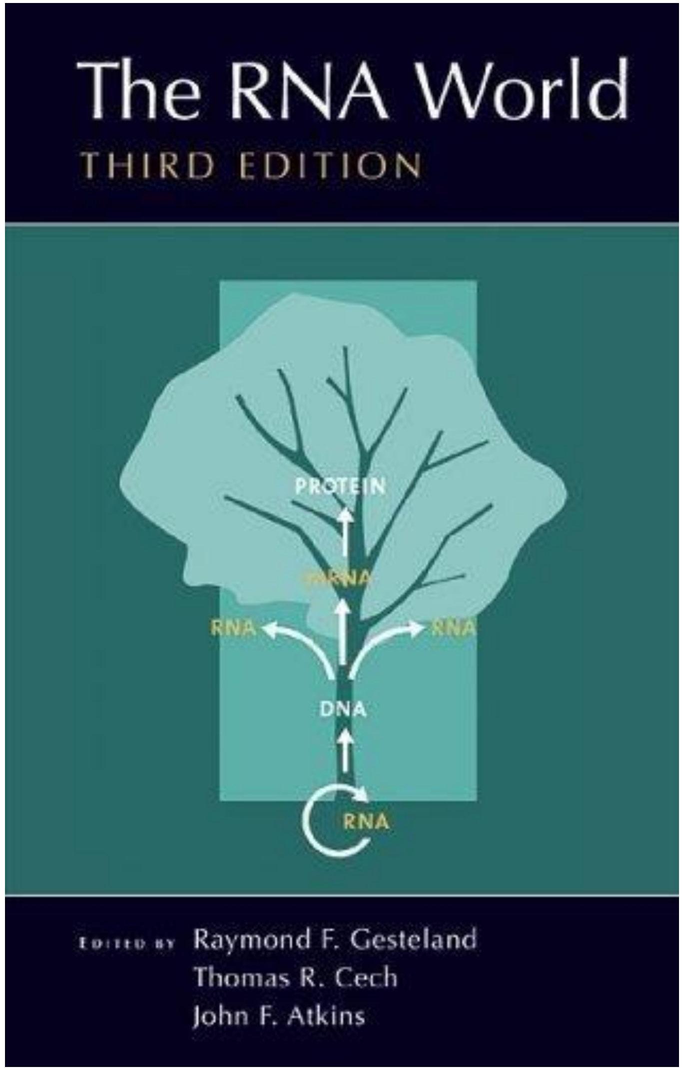
The RNA world hypothesis

1982: Like proteins, RNA is a catalyst

Tetrahymena
Group I intron



Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., & Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *cell*, *31*(1), 147-157.

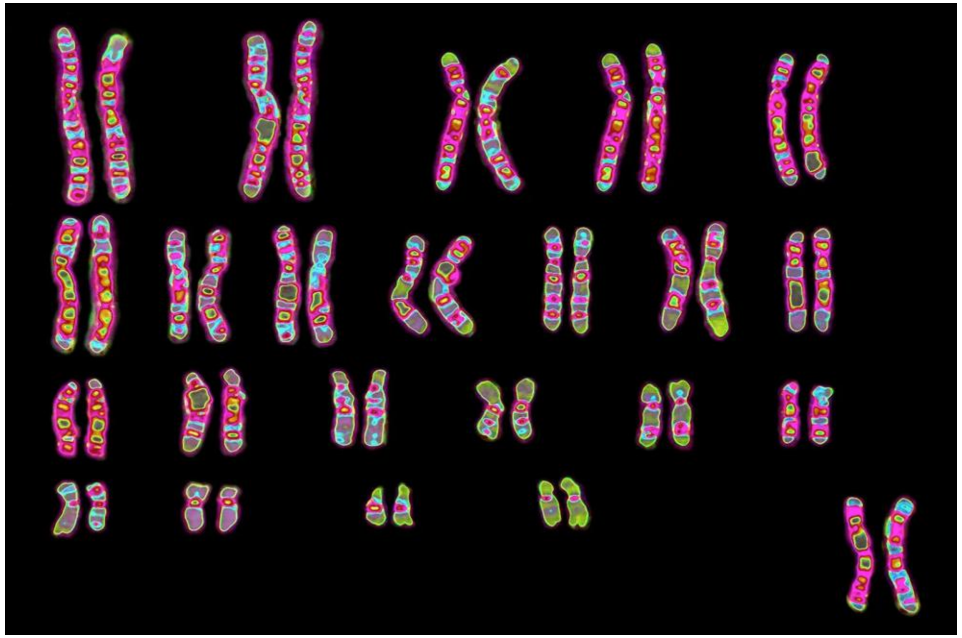


The RNA World
THIRD EDITION

Edited by Raymond F. Gesteland
Thomas R. Cech
John F. Atkins

We still live in an RNA world!

The human genome



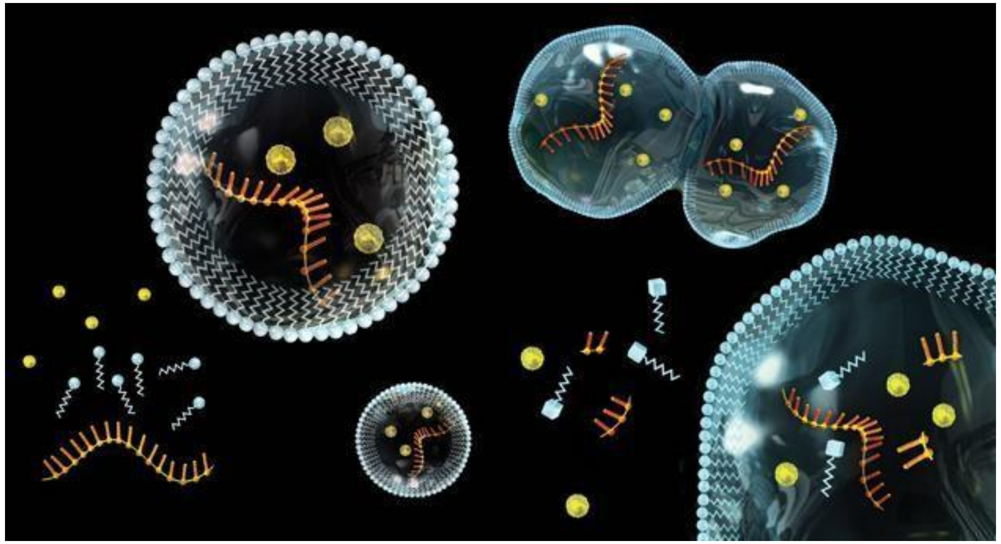
3 billion base pairs of DNA

75 - 90% is transcribed into RNA

1 - 2% is translated into proteins

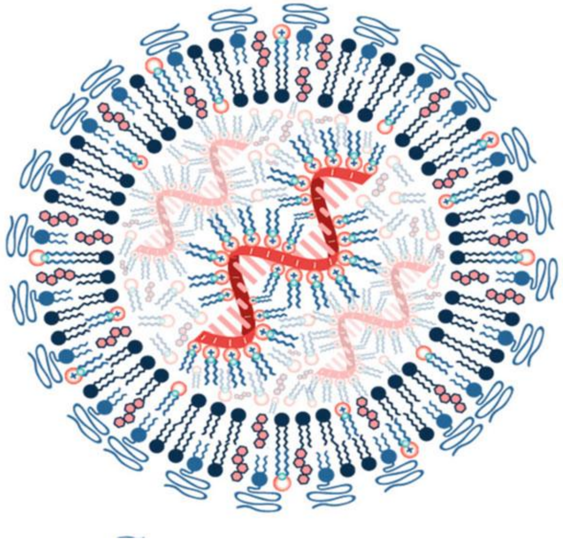
RNA is at the core of biology from the start

Understand the origin of life



Protocells

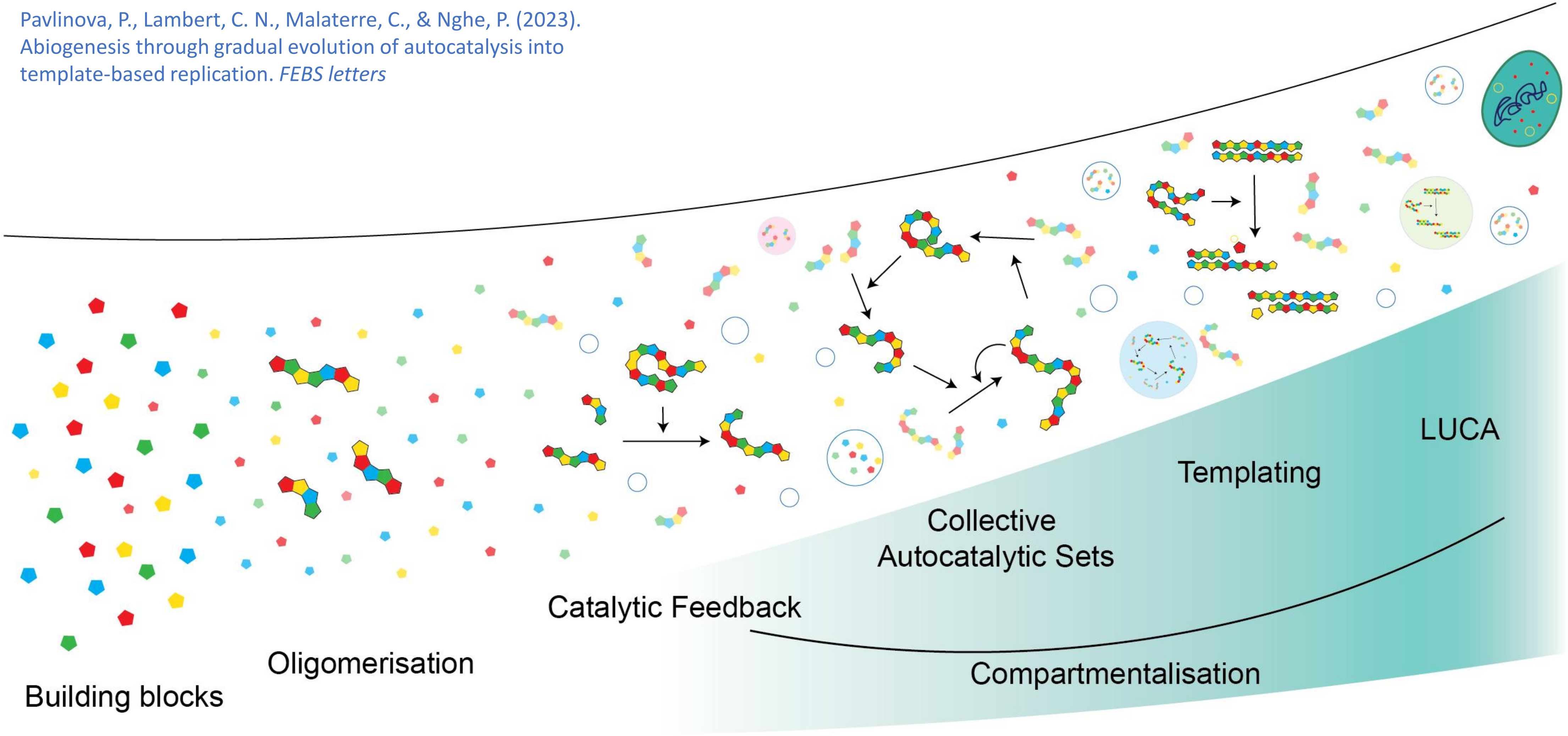
Develop novel therapies



Lipid Nanoparticle for therapeutic RNA delivery

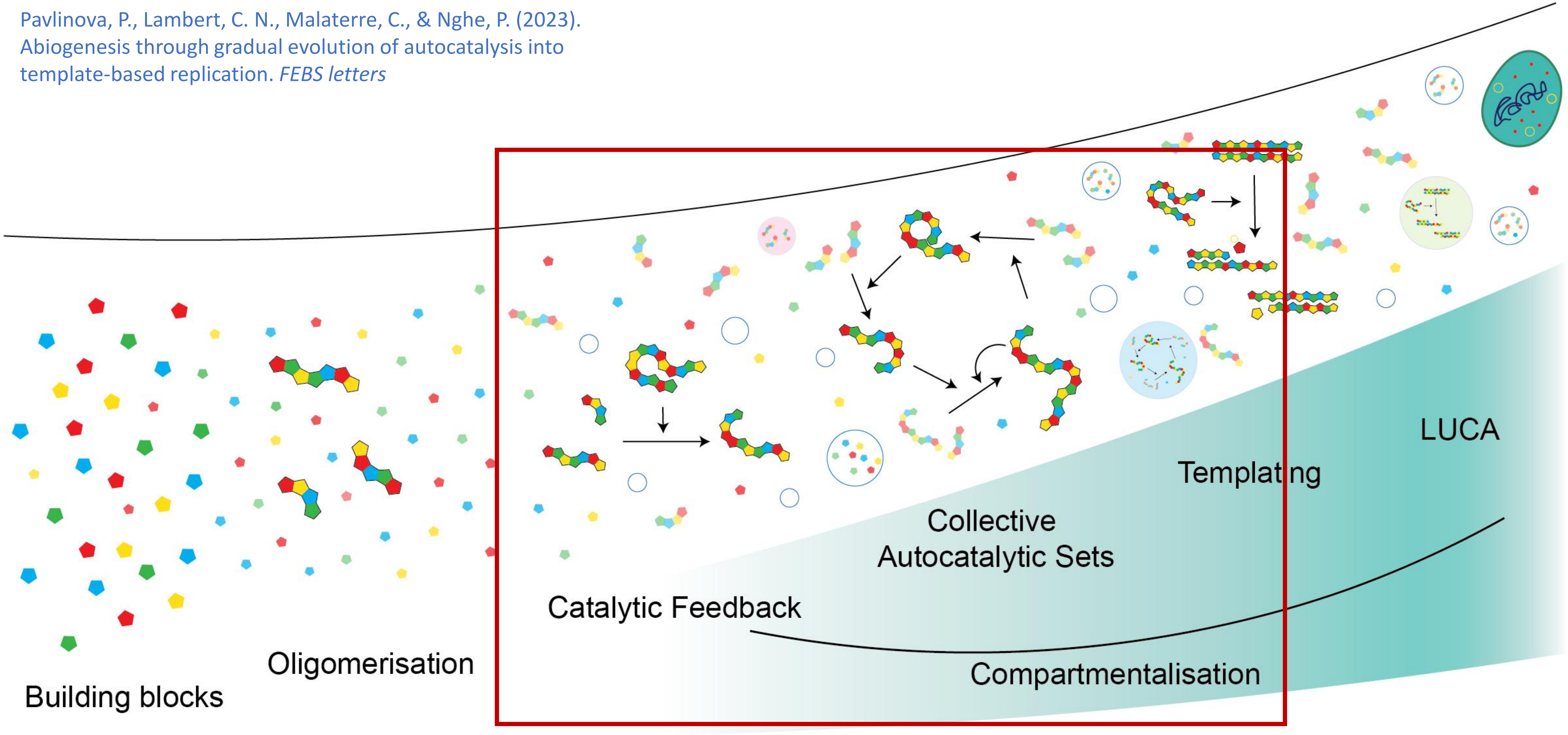
Origin of life: a motivation for RNA design

Pavlinova, P., Lambert, C. N., Malaterre, C., & Nghe, P. (2023). Abiogenesis through gradual evolution of autocatalysis into template-based replication. *FEBS letters*



Origin of life: a motivation for RNA design

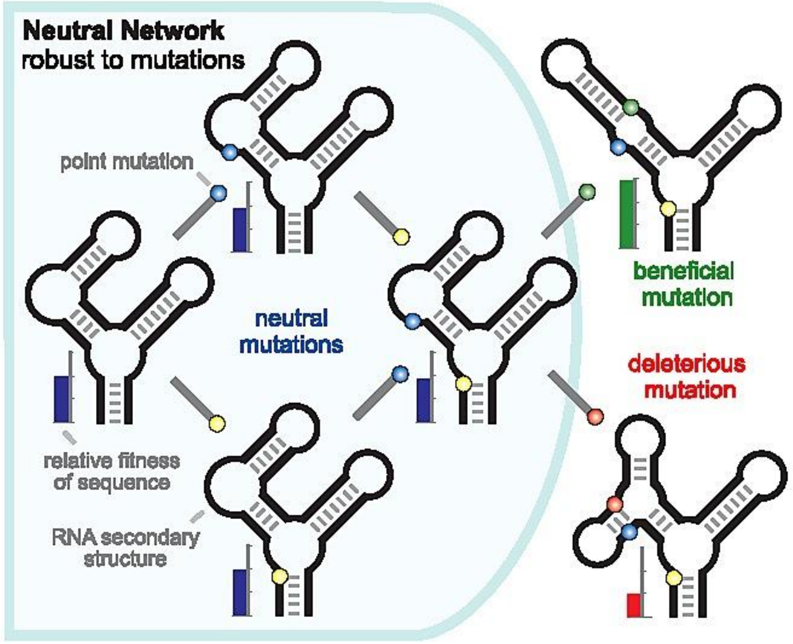
Pavlinova, P., Lambert, C. N., Malaterre, C., & Nghe, P. (2023). Abiogenesis through gradual evolution of autocatalysis into template-based replication. *FEBS letters*



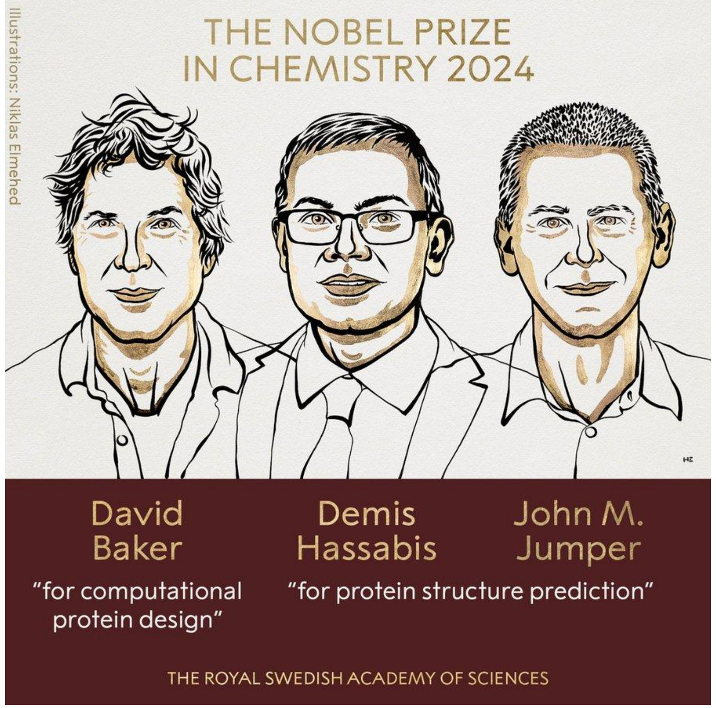
Is RNA self-reproduction exceptional or banal?

Challenge :
design RNAs with the same function but no identity

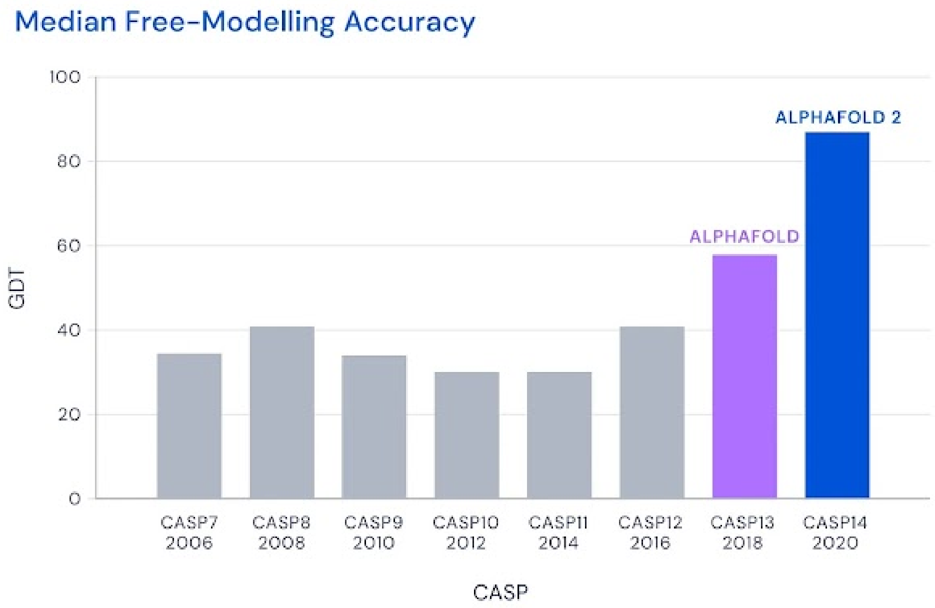
« neutral network » of RNA function

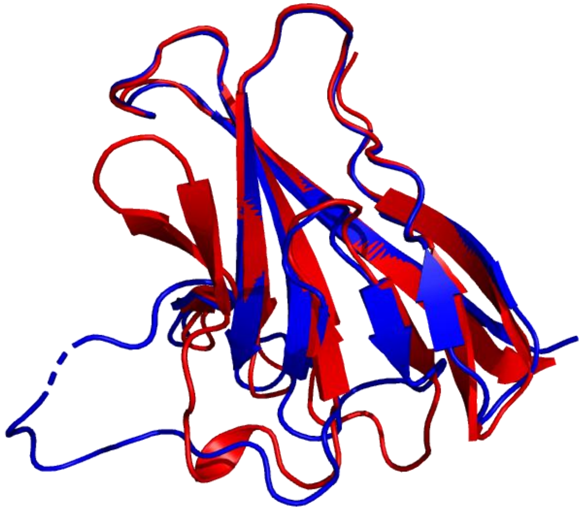


Natural protein structures can now be predicted



Alphafold2 could predict even difficult protein folds at the CASP14 competition with atomic resolution

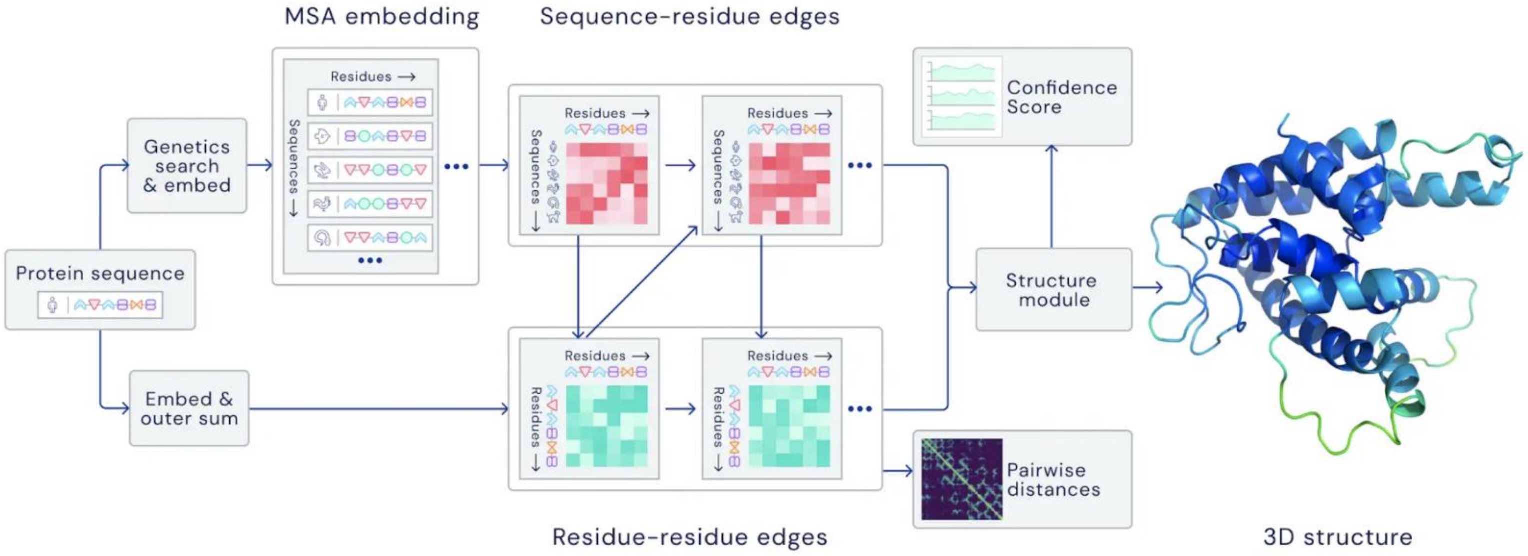




Alphafold learns context-dependencies at a large scale

Learned from:

- 170.000 structures
- $>10^7$ sequences

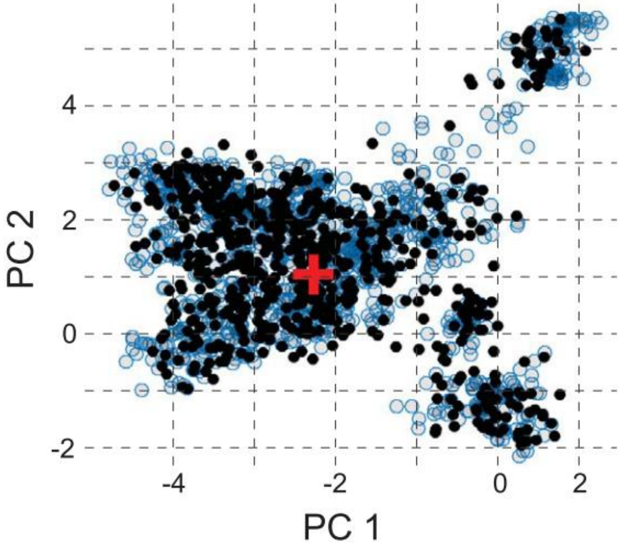


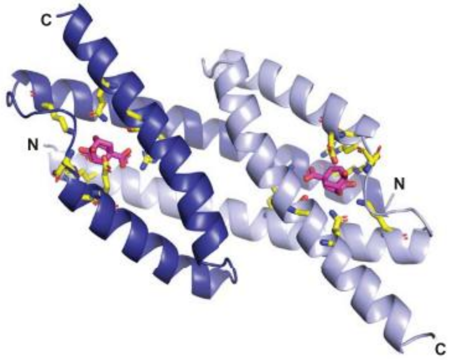
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, *596*(7873), 583-589.

Can we predict function?

Alphafold does not.

Only within protein homolog families.



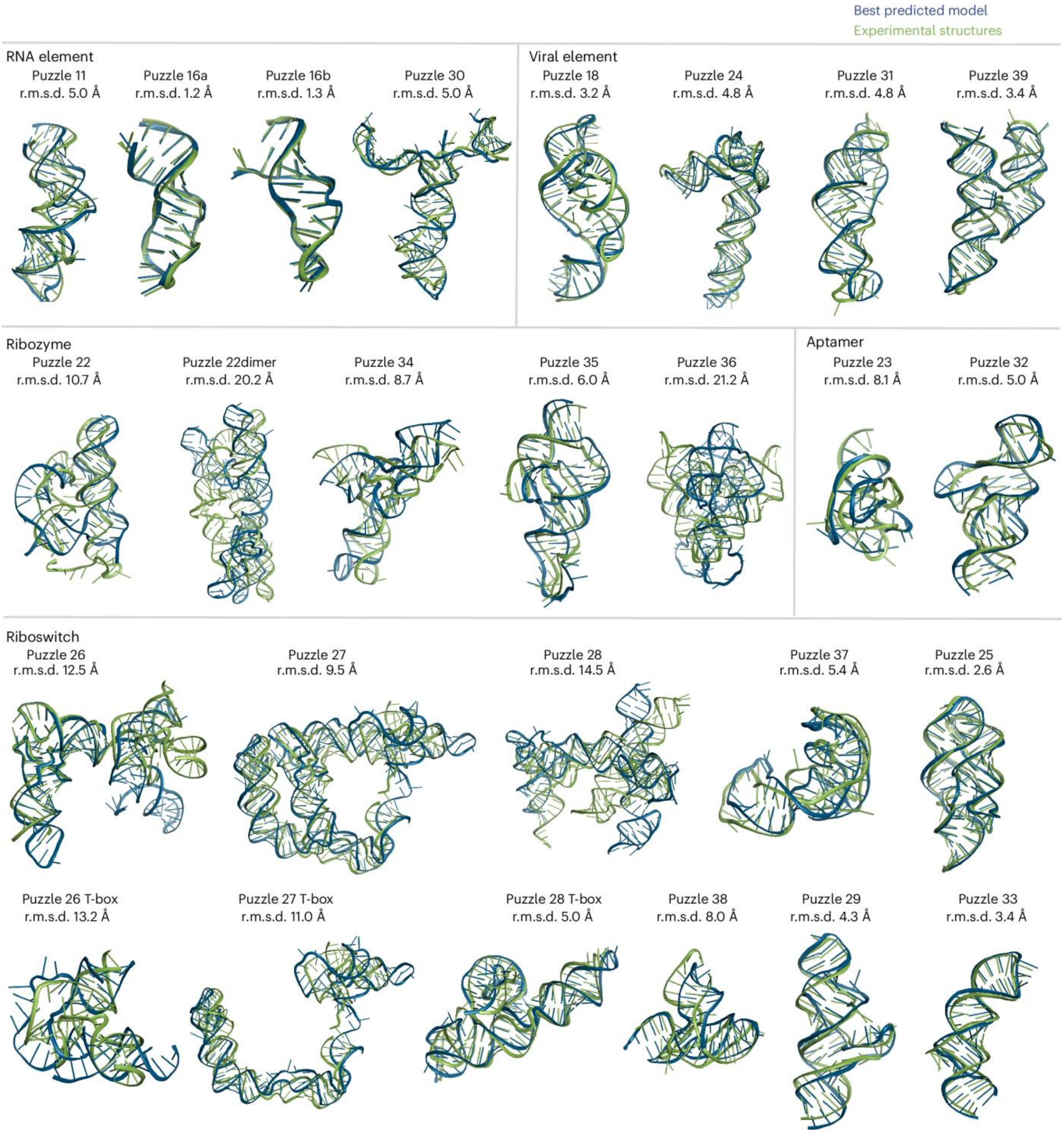


Chorismate mutase enzyme

Russ et al. "Science 369.6502 (2020): 440-445

Structure prediction fails for RNA

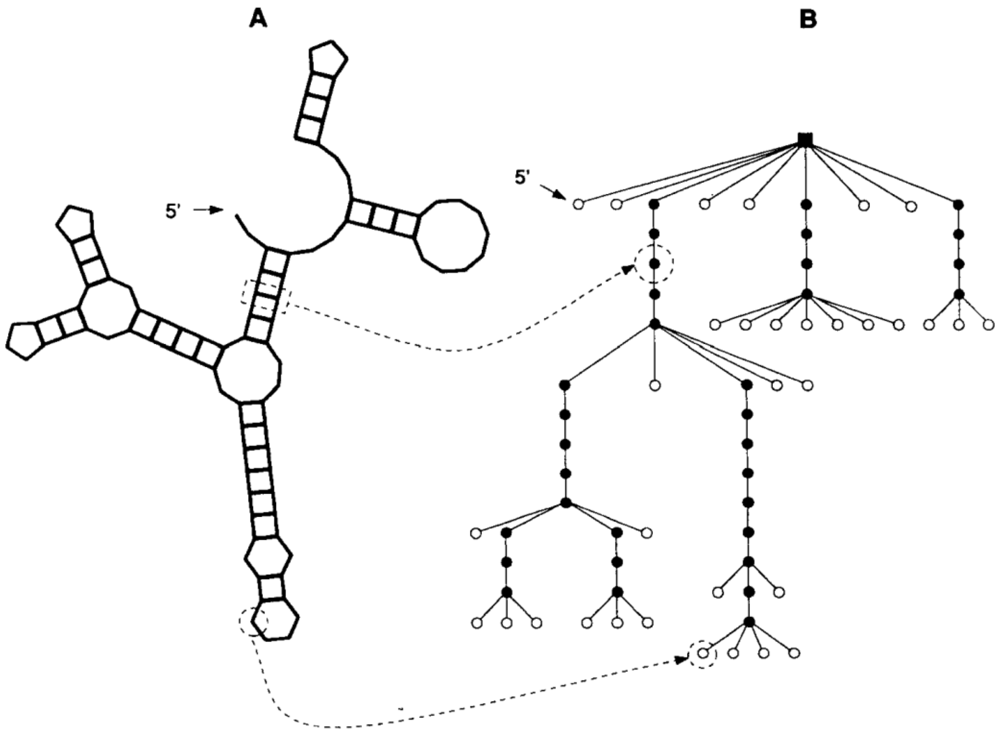
- ~10^4 structures
- ~10^7 sequences
- Heavily biased towards tRNAs and rRNAs



Bu, Fan, et al. "RNA-Puzzles Round V: blind predictions of 23 RNA structures." *Nature methods* (2024): 1-13.

Kretsch, Rachael C., et al. "Assessment of nucleic acid structure prediction in CASP16." *bioRxiv* (2025): 2025-05.

Opportunities: physics-based models

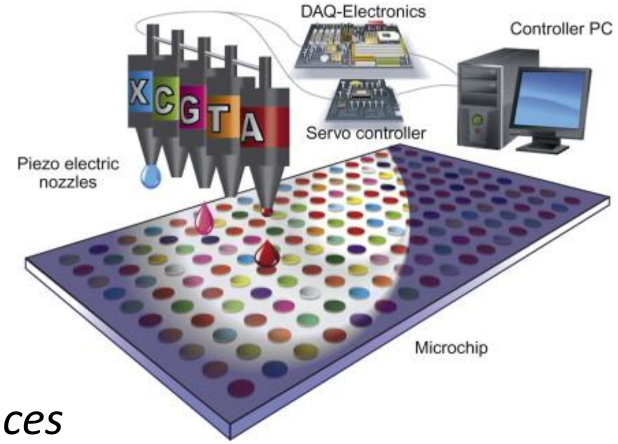


It is possible to compute secondary structure efficiently for reasonably-sized RNAs.

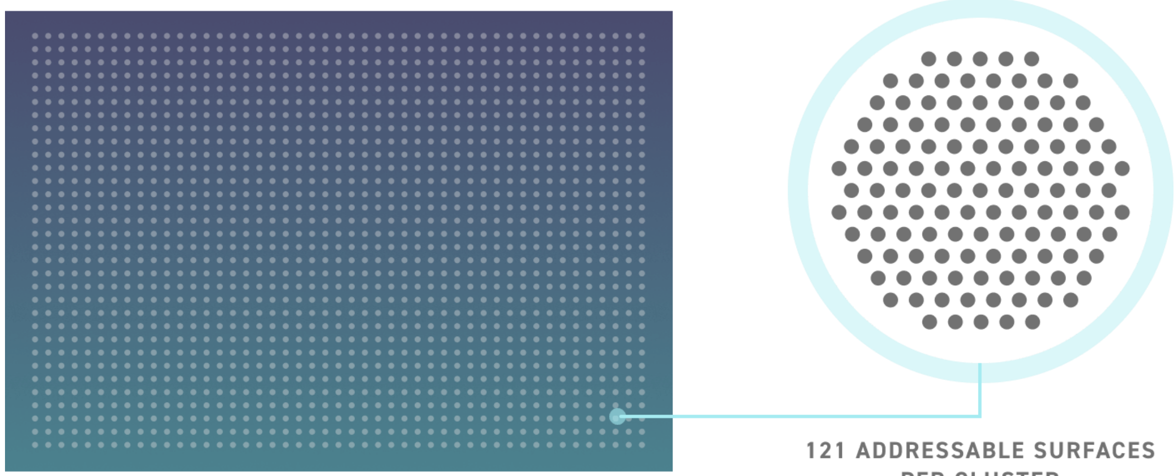
Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fur chemie*, *125*, 167-167.

Opportunities: DNA synthesis & sequencing

Custom DNA pool synthesis

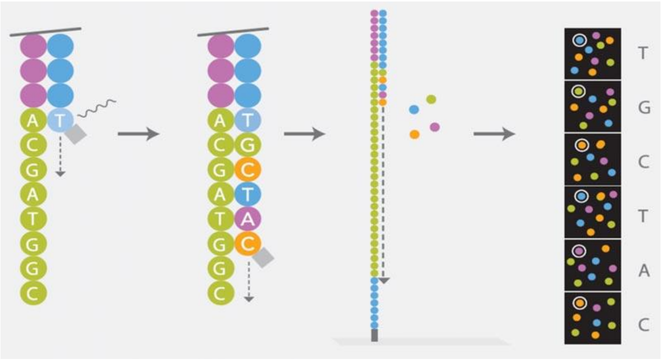


Twist Biosciences



DNA sequencing

Illumina



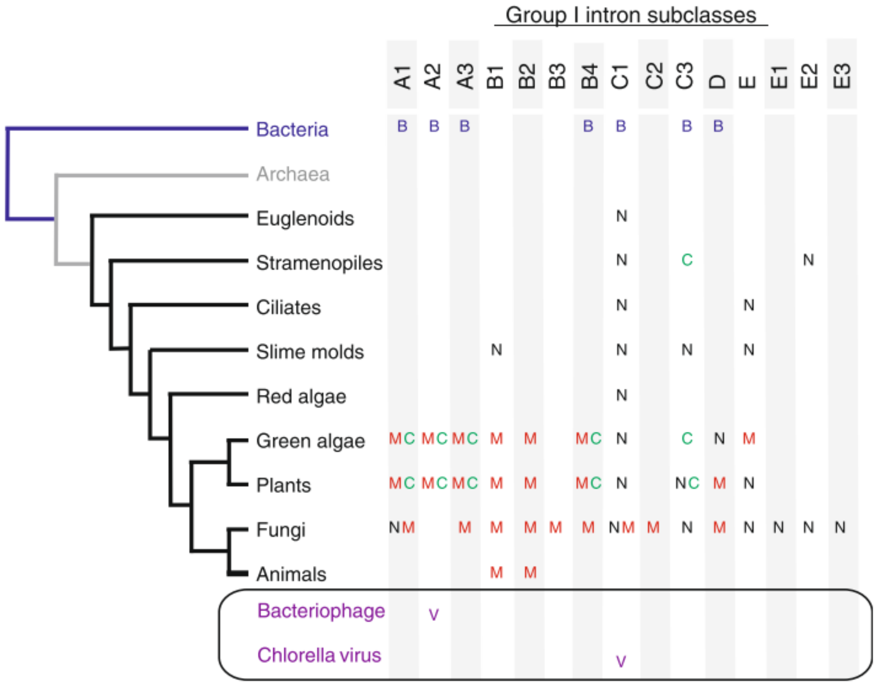
Nanopore

AI challenges for RNA

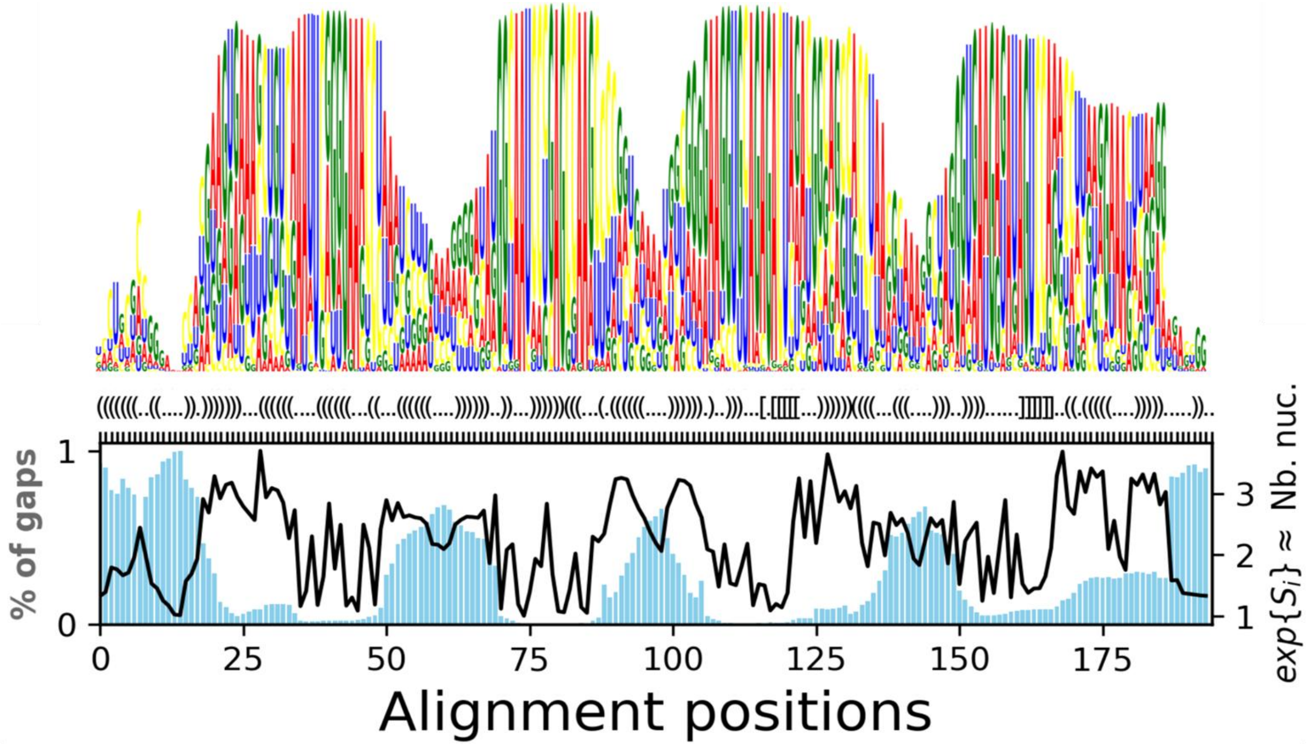
1. Can we predict RNA structure?
2. Can we predict RNA function?
3. How to integrate physics-based models?
4. How to optimally learn from modelling – experimental cycles?

- **RNA models are of high interest for biology**
- **RNA provides an ideal platform to test AI methods**

Group I intron family

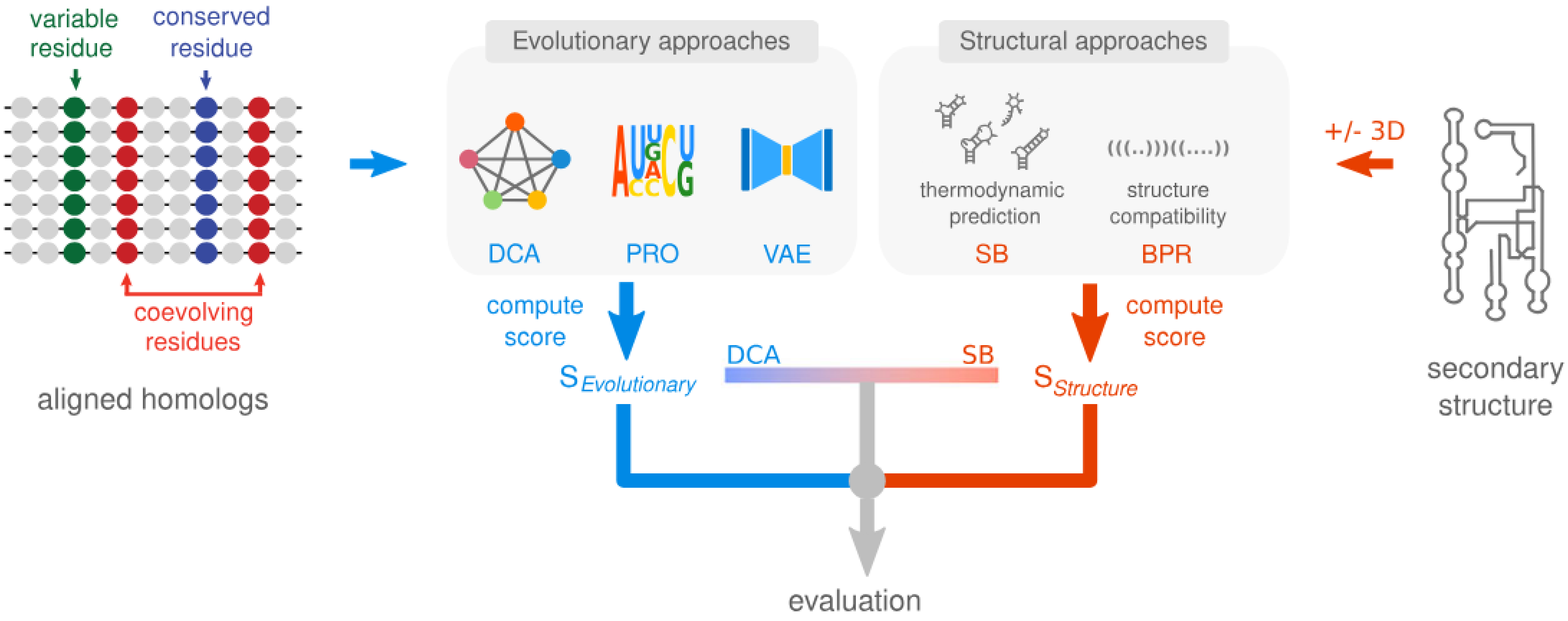


	A1	A2	A3	B1	B2	B3	B4	C1	C2	C3	D	E	E1	E2	E3
Bacteria	B	B	B				B	B		B	B				
Archaea															
Euglenoids								N							
Stramenopiles								N		C				N	
Ciliates								N				N			
Slime molds				N				N		N		N			
Red algae								N							
Green algae	MC	MC	MC	M	M		MC	N		C	N	M			
Plants	MC	MC	MC	M	M		MC	N		NC	M	N			
Fungi	NM		M	M	M	M	M	NM	M	N	M	N	N	N	N
Animals				M	M										
Bacteriophage		V													
Chlorella virus								V							

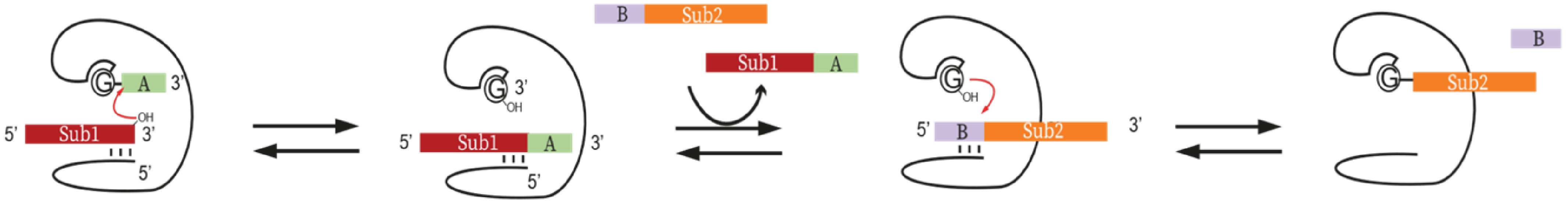


Evolutionary data of group I introns
~816 sequences from the RFAM database

Computational workflow



Experimental test of activity



Computational design

DNA synthesis

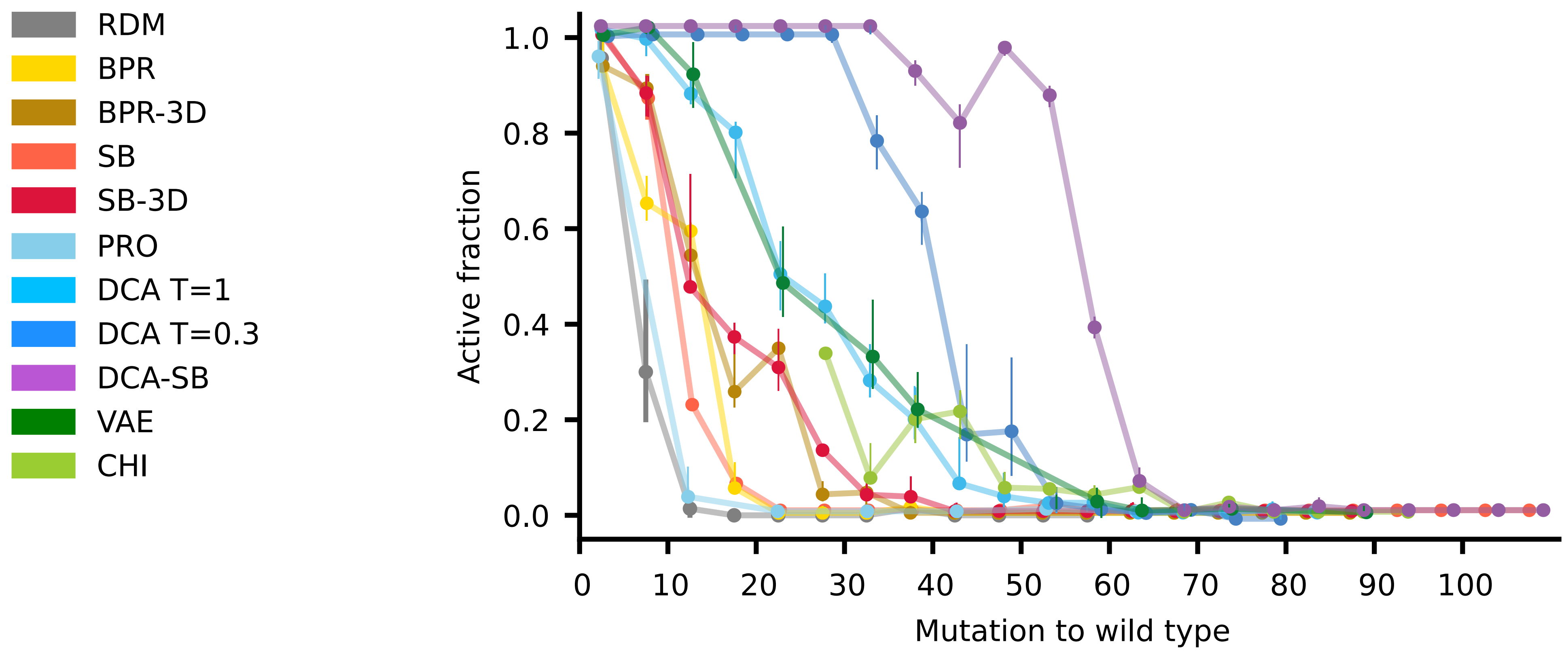
Biochemical assay

Sequencing read-out

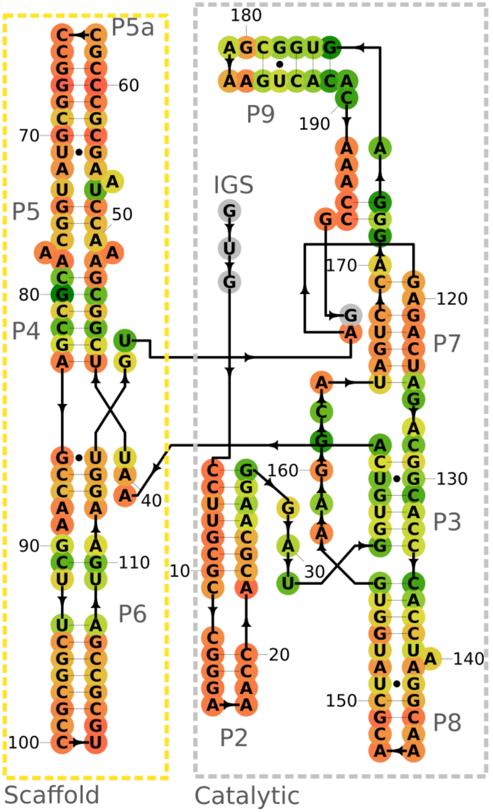
>10.000 distinct sequences can be tested at once in a single day

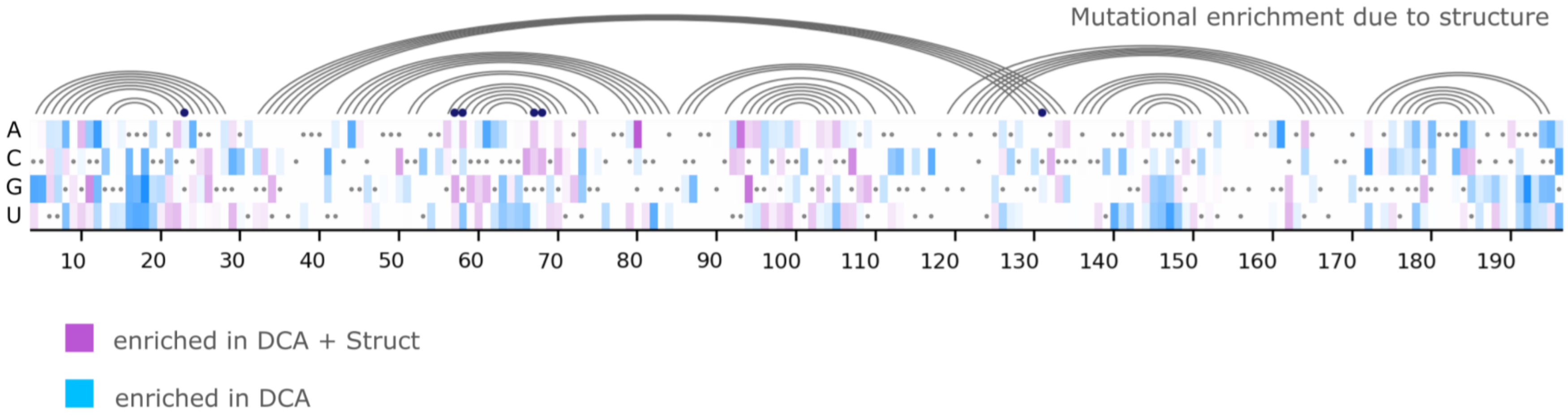
Comparison of generative models

~200 sequences per bin of 10 mutations

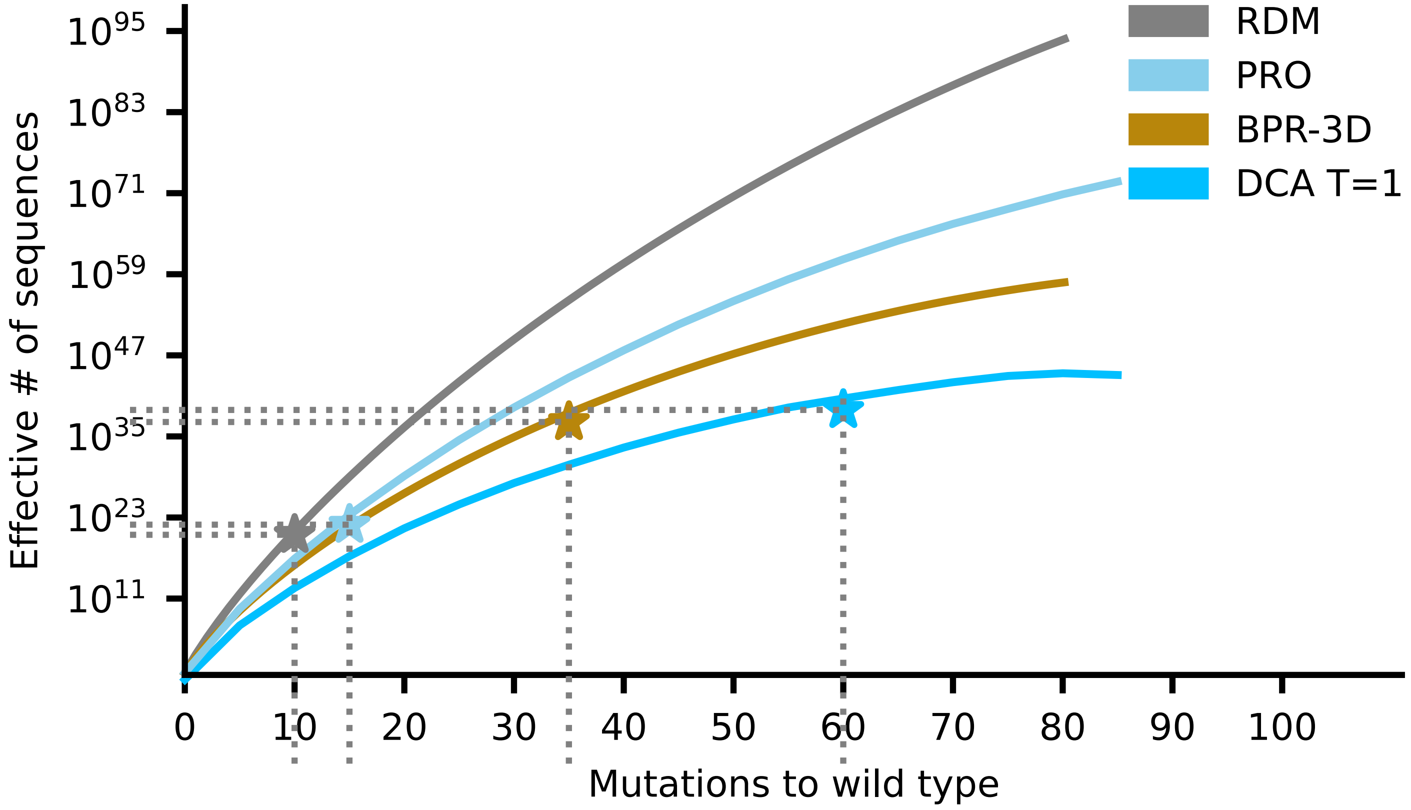


Physics generates functional molecules out of the natural distribution





Estimation of the number of autocatalytic RNAs



Estimated number of sequences of length 200 with the same function: 10^{39}

Lambert CN, Opuu V, Calvanese F, Pavlinova P, Zamponi F, Hayden EJ, Weigt M, Smerlak M, Nghe P. *Exploring the space of self-reproducing ribozymes using generative models*. Nature Communications. 2025 Aug 22;16(1):7836.

Our Impact

Technology

Players and researchers have created design tools, predictive AI models, large data sets, and DNA template libraries. All technologies are freely available for non-commercial use, and commercial licenses support Eterna's ongoing research.

Learn more

Publications

Eterna players have contributed to over 30 scientific publications, including as lead authors, in journals such as *Nature Communications*, *PLOS Computational Biology*, and *Proceedings of the National Academy of Sciences*.

Learn more

OpenVaccine

During the pandemic, the Eterna community created and published designs for a more stable mRNA vaccine for COVID-19 optimized for new variants. Eterna's open challenges address several important areas of RNA medicine.

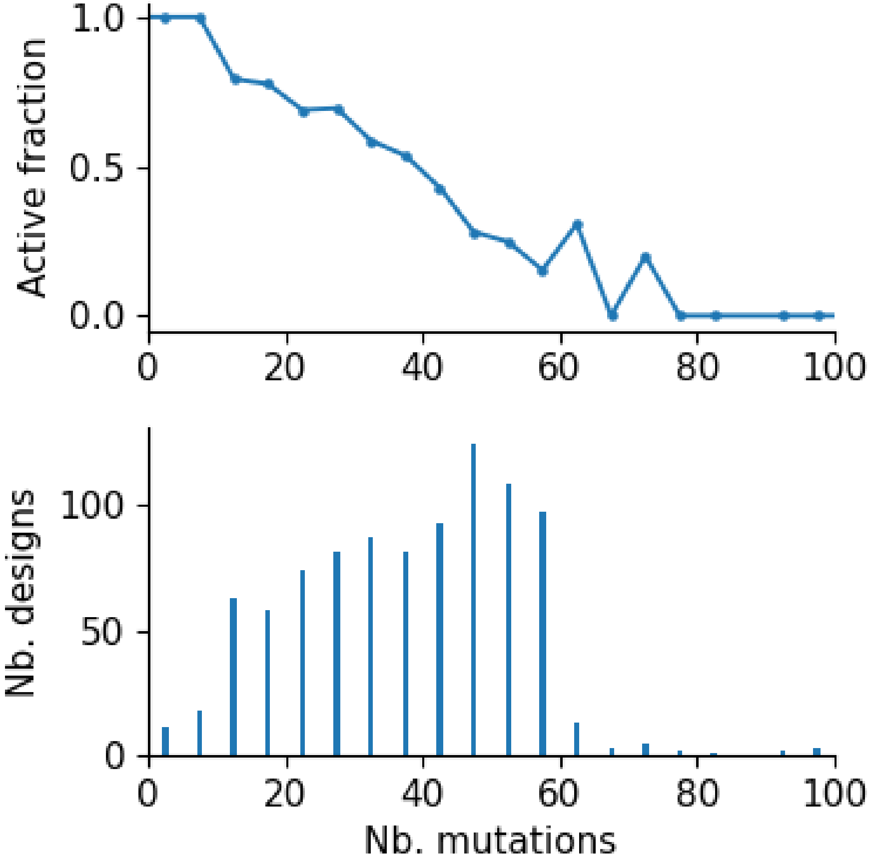
Learn more

FEATURED LAB CHALLENGE

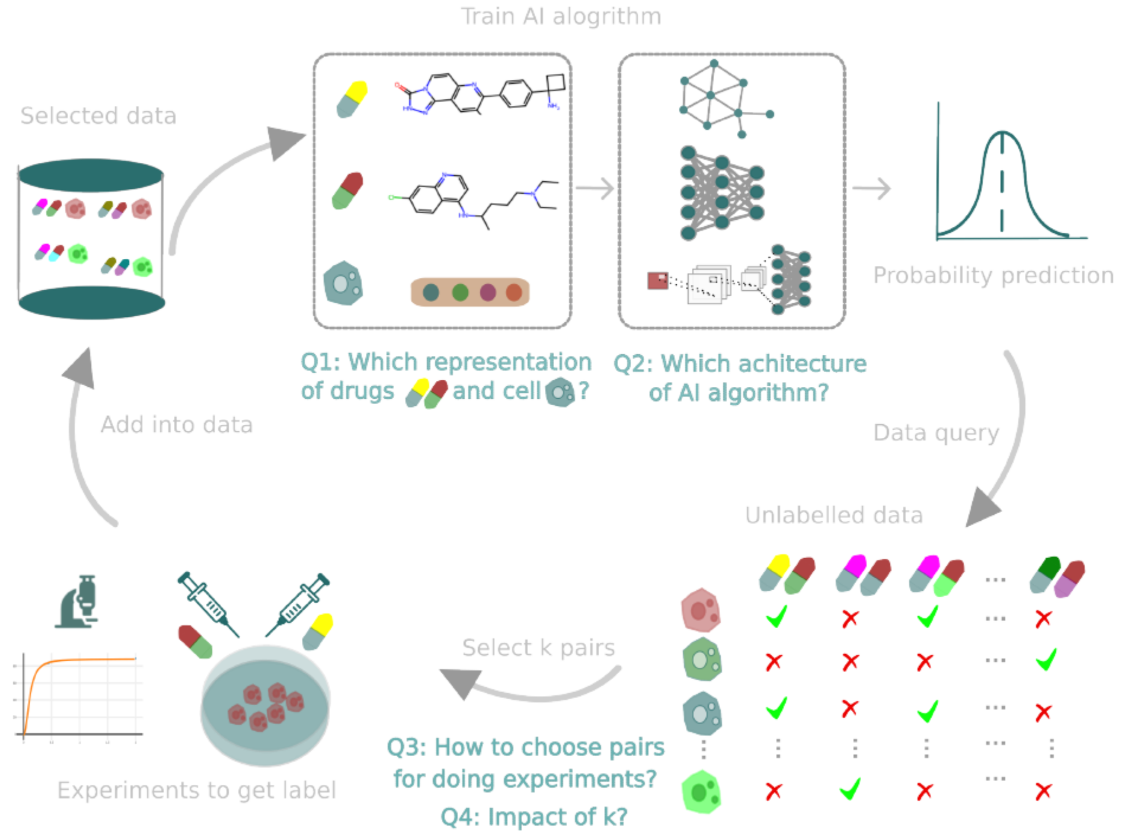
OpenKnot

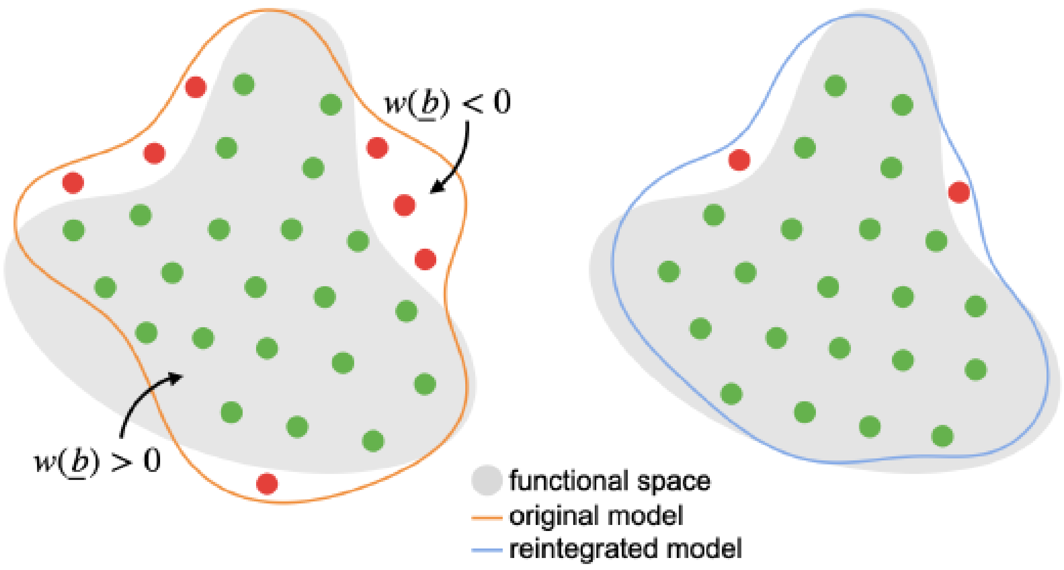
Many important biological processes depend on RNAs that form pseudoknots, and they are among the most conserved structures in evolutionary history. However, scientists still have much to learn about their properties, structure, and functions.

Humans equal machine learning



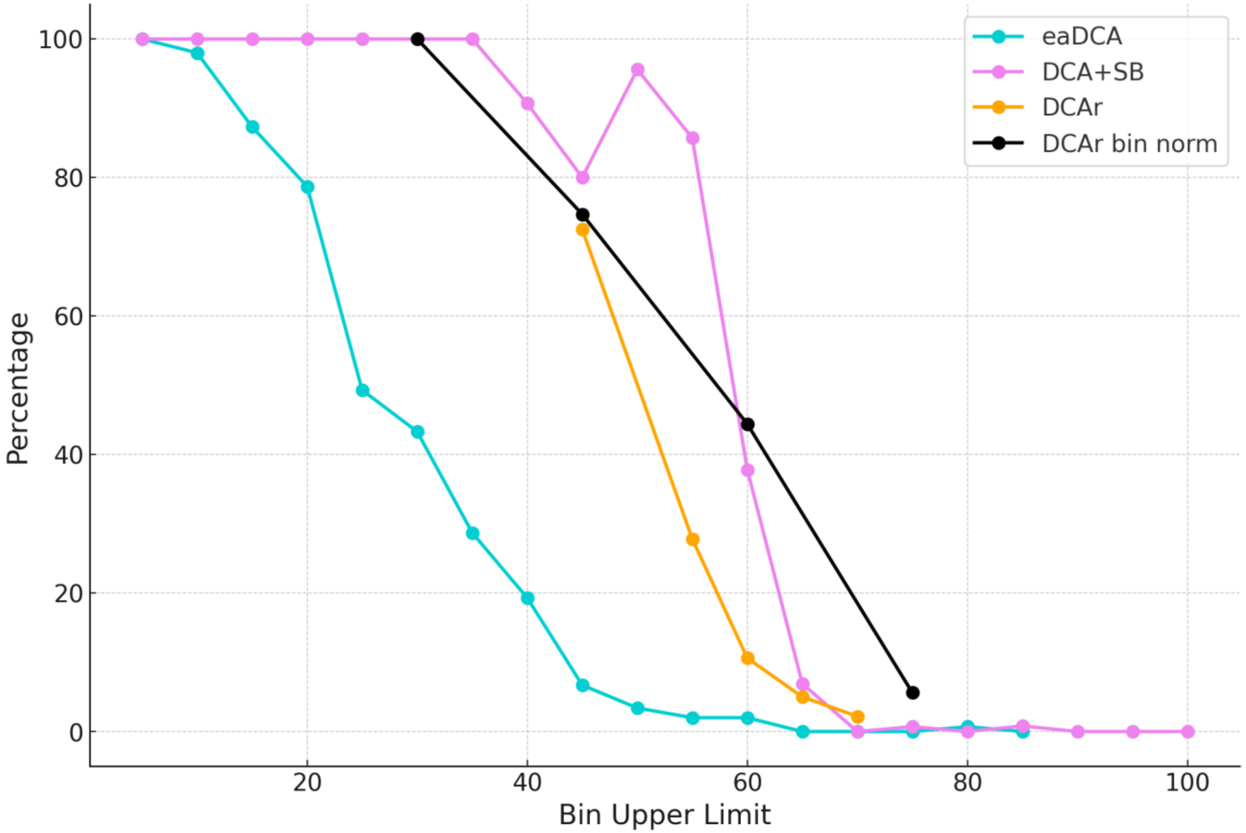
Active learning





Wang, S., Allauzen, A., Nghe, P., & Opuu, V. (2025). A guide for active learning in synergistic drug discovery. *Scientific Reports*, *15*(1), 3484

Relearning allows going out of the natural distribution



Calvanese, F., Peinetti, G., Pavlinova, P., Nghe, P., & Weigt, M. (2025). Integrating experimental feedback improves generative models for biological sequences. *Nucleic Acids Research*, *53*(16), gkaf832.

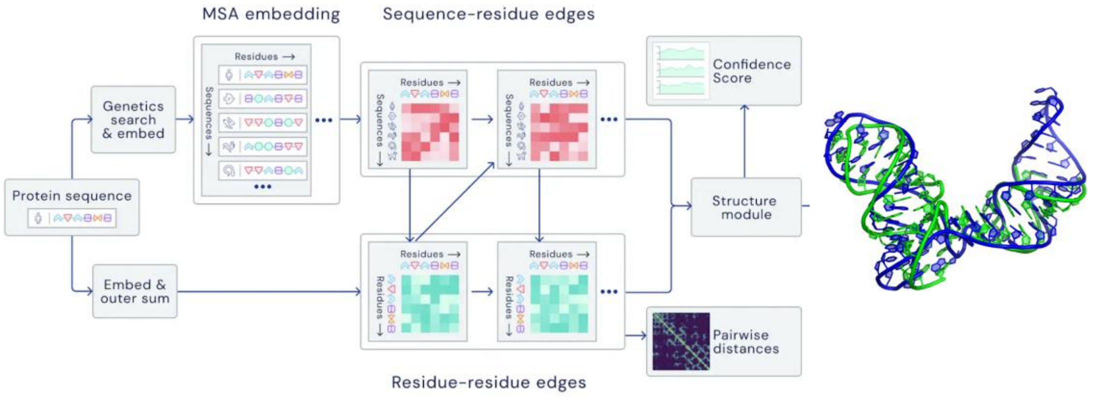
Build a LLM for RNA from our own data

Public databases
~10^4 structures
~10^7 sequences

Physics-based models of RNA
Folding
Molecular dynamics

Experiemental structures
CryoEM

Generative LLM



Active learning

High-throughput screening of RNA

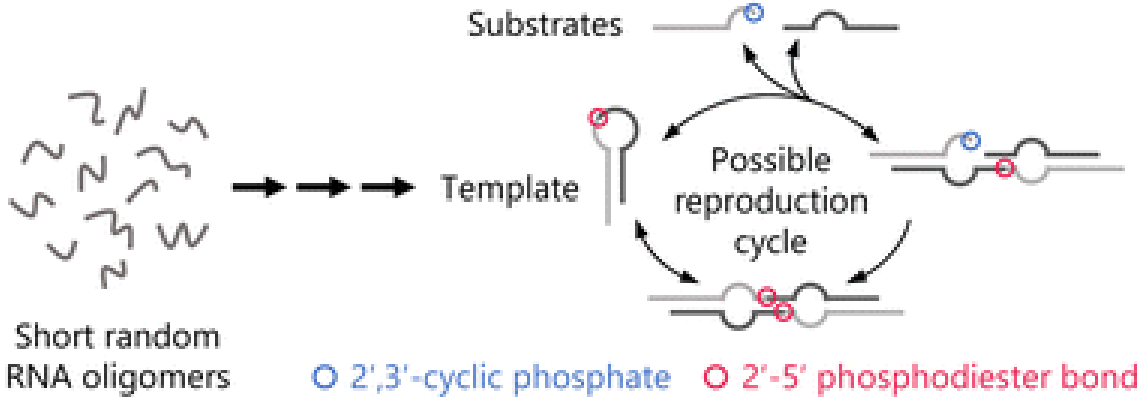
Computational design

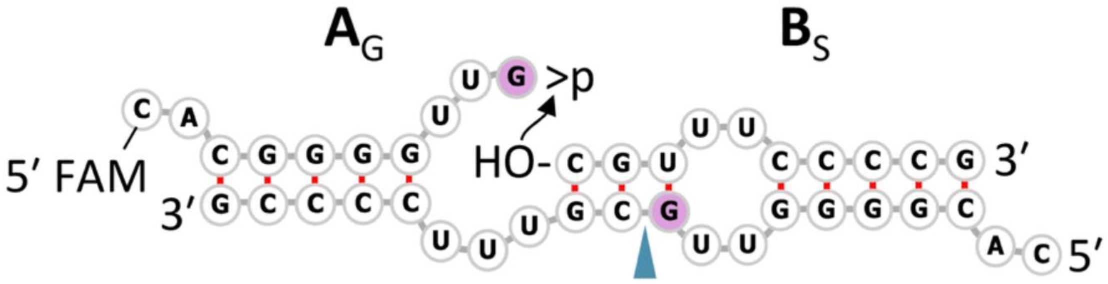
Synthesis

Pooled assay

Sequencing

Autocatalysis emerges from random mixtures

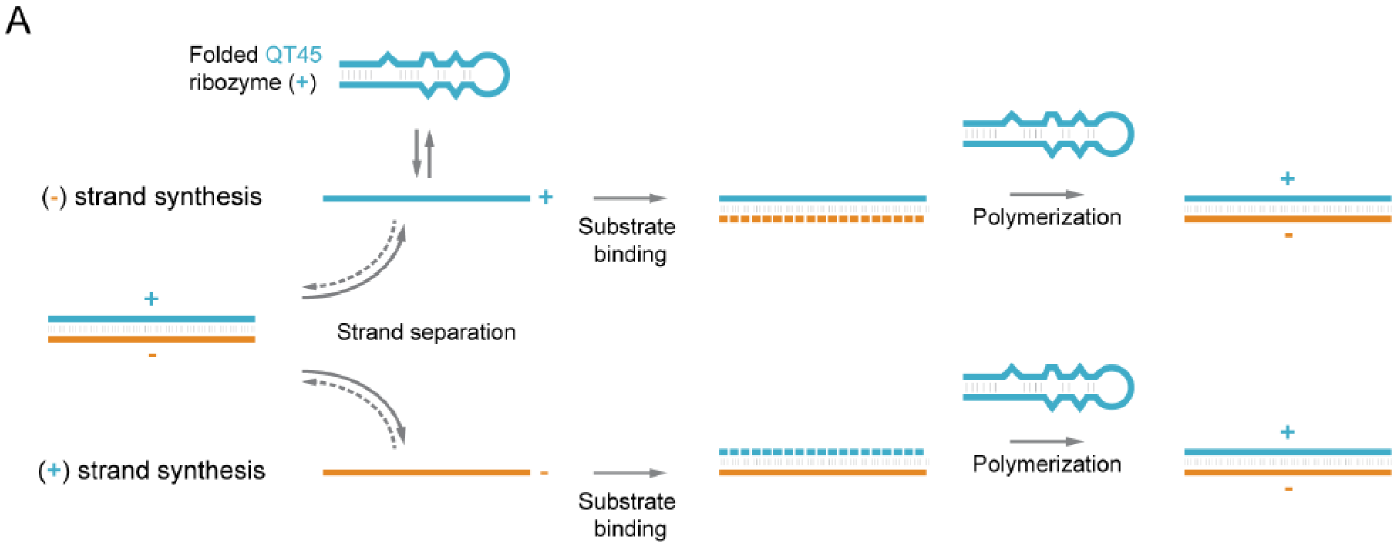




Mizuuchi R, Ichihashi N. Minimal RNA self-reproduction discovered from a random pool of oligomers. Chem Sci. 2023;14(28):7656-64.

Toward replication in the RNA world

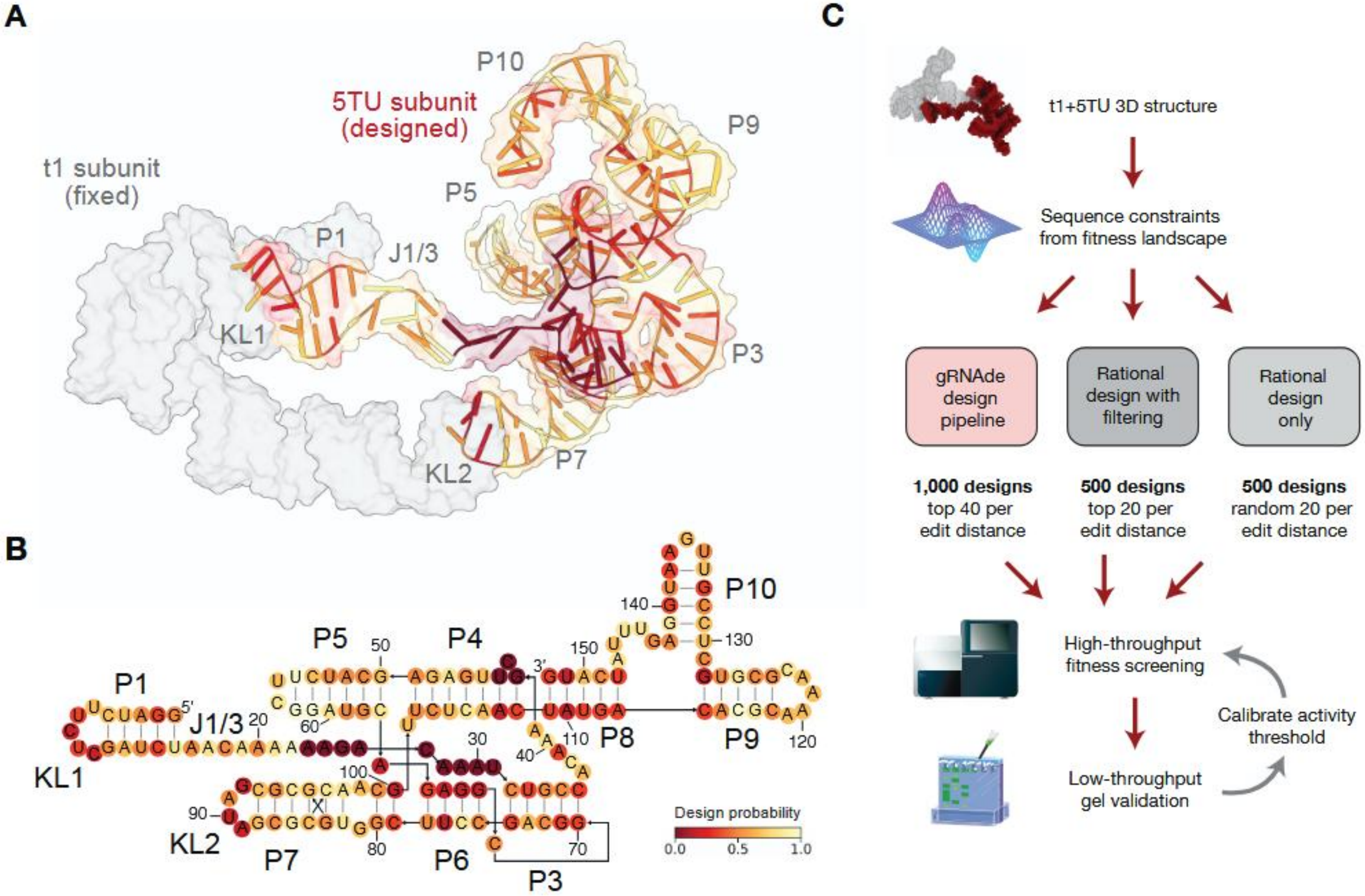
An RNA that can copy itself from trimers and hexamers



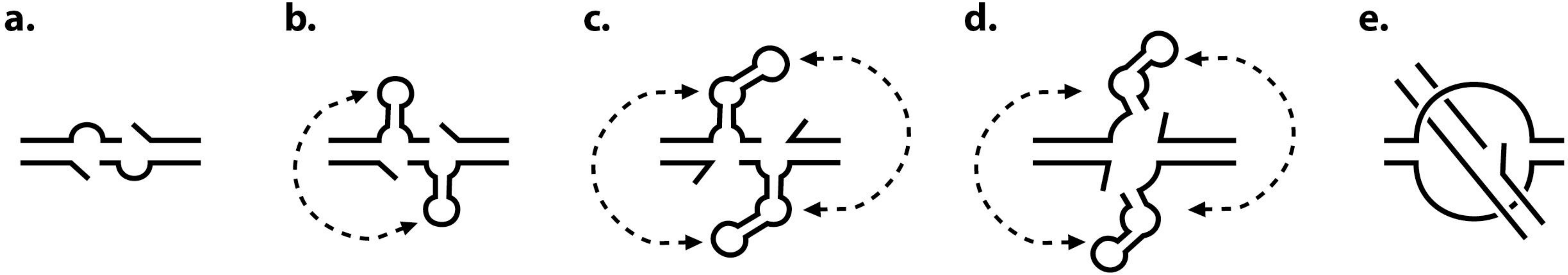
A polymerase ribozyme that can synthesize both itself and its complementary strand
E Gianni, SLY Kwok, CJK Wan, K Goeij, B Clifton, J Attwater, P Holliger, Science 2026

Diversifying replicases

Joshi, C. K., Gianni, E., Kwok, S. L., Mathis, S. V., Liò, P., & Holliger, P. (2025). Generative inverse design of RNA structure and function with gRNAde. *bioRxiv*, 2025-11.

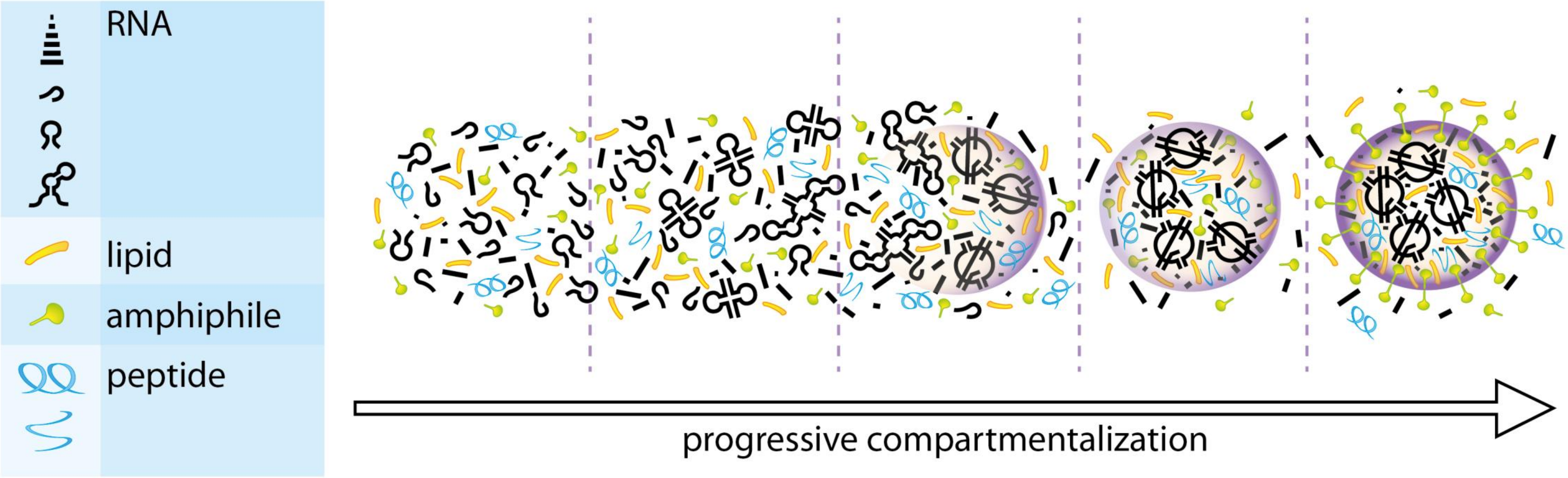


Evolution of primordial autocatalysts

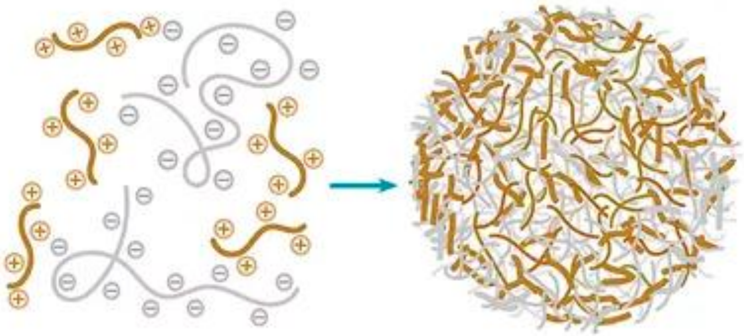


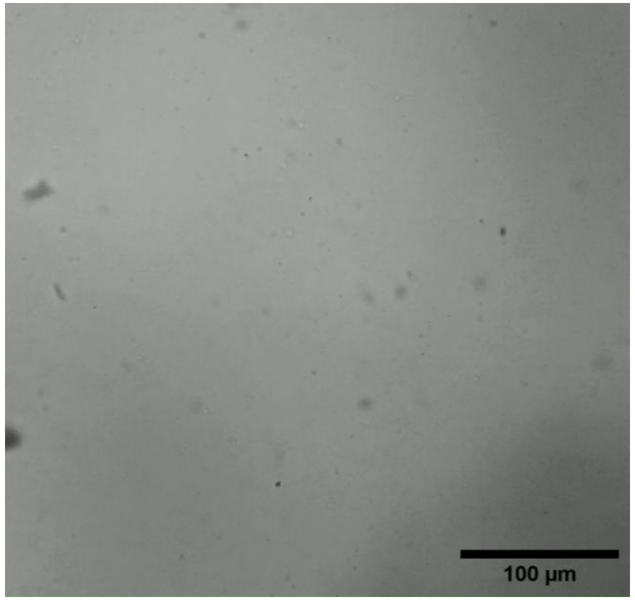
Nghe, P. A stepwise emergence of evolution in the RNA world. *FEBS letters*. (2025)

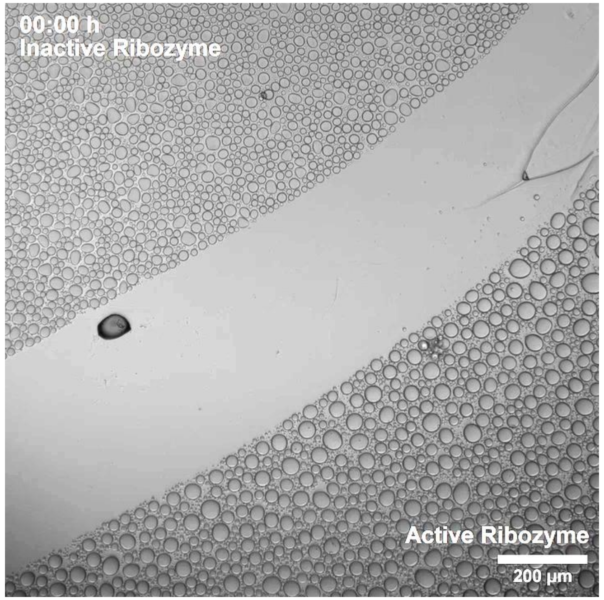
Emergence of protocells



Self-organization of RNA and peptides in compartments



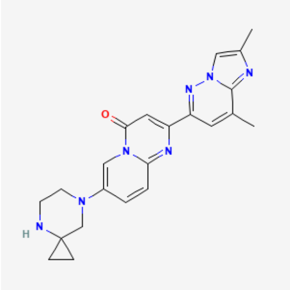




Selection in a drying pond

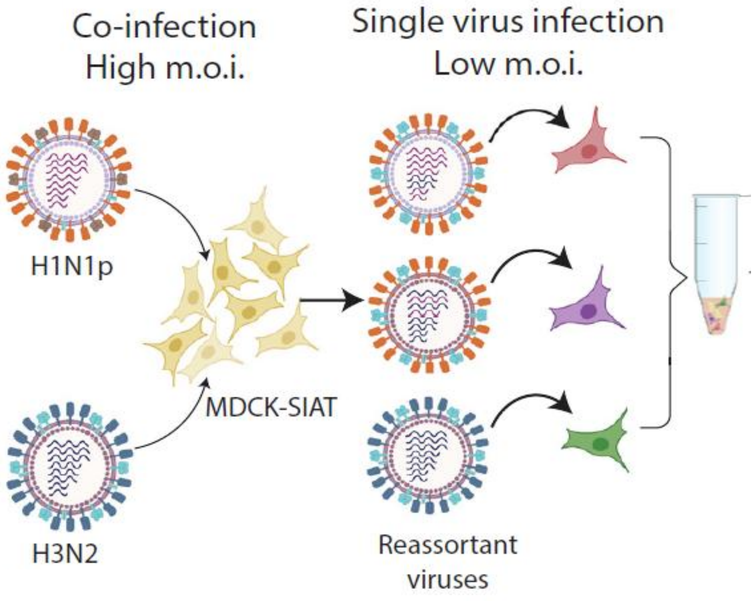
Biomedical applications

Targetting RNA with small molecules



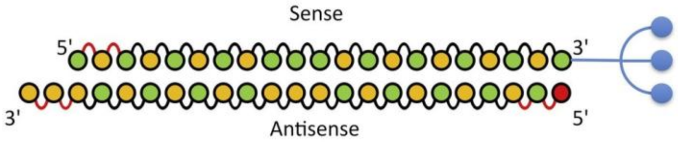
Risdiplam against SMA

Predicting the next influenza strain

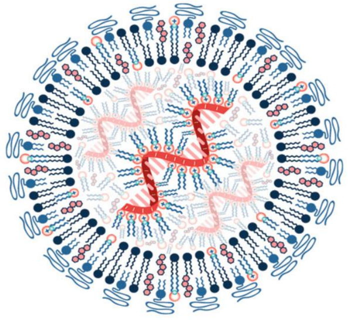


KY Chen, J Karuppusamy, MB O'Neill, V Opuu, M Bahin, S Foulon, P Ibanez, L Quintana-Murci, T Ozawa, S van der Werf, P Nghe, N Naffakh, AD Griffiths, C Isel, *High-throughput droplet-based analysis of influenza A virus genetic reassortment by single-virus RNA sequencing*, Proceedings of the National Academy of Sciences (2023),

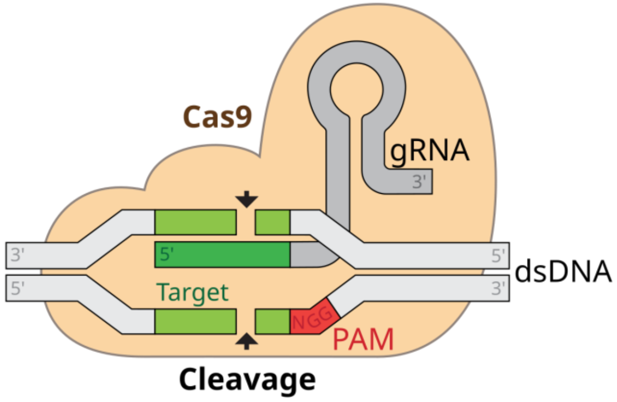
Small interfering RNAs



mRNA therapies

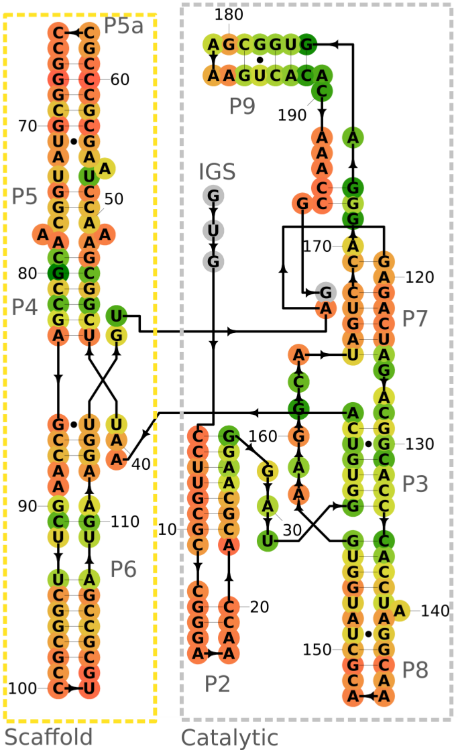


Gene-editing therapies



Conclusion

- Generative models work with as few as hundreds of sequences
- Exploration – Exploitation trade-off
- Out-of-distribution desigsn from physics and relearning
- Scaling law of the number of possible sequences



Challenges

- Toward no identity
- Diversify from a single instance
- Modify structure
- Find new activities

Opportunities

- Active learning
- Physics-informed models

Acknowledgements

ESCPI
Camille LAMBERT
Vaitea OPUU
Polina Pavlinova

Sorbonne Université
Francesco CALVANESE
Giovanni PEINETTI
Marin WEIGT

La Sapienza: Francesco ZAMPONI
MPI Dresden: Matteo SMERLAK
Boise State University: Eric Hayden

HFSP International Human Frontier Science Program Organization

erc

FRANCE 2030 PROGRAMME DE RECHERCHE ORIGINES

Horizon 2020 European Union Funding for Research & Innovation