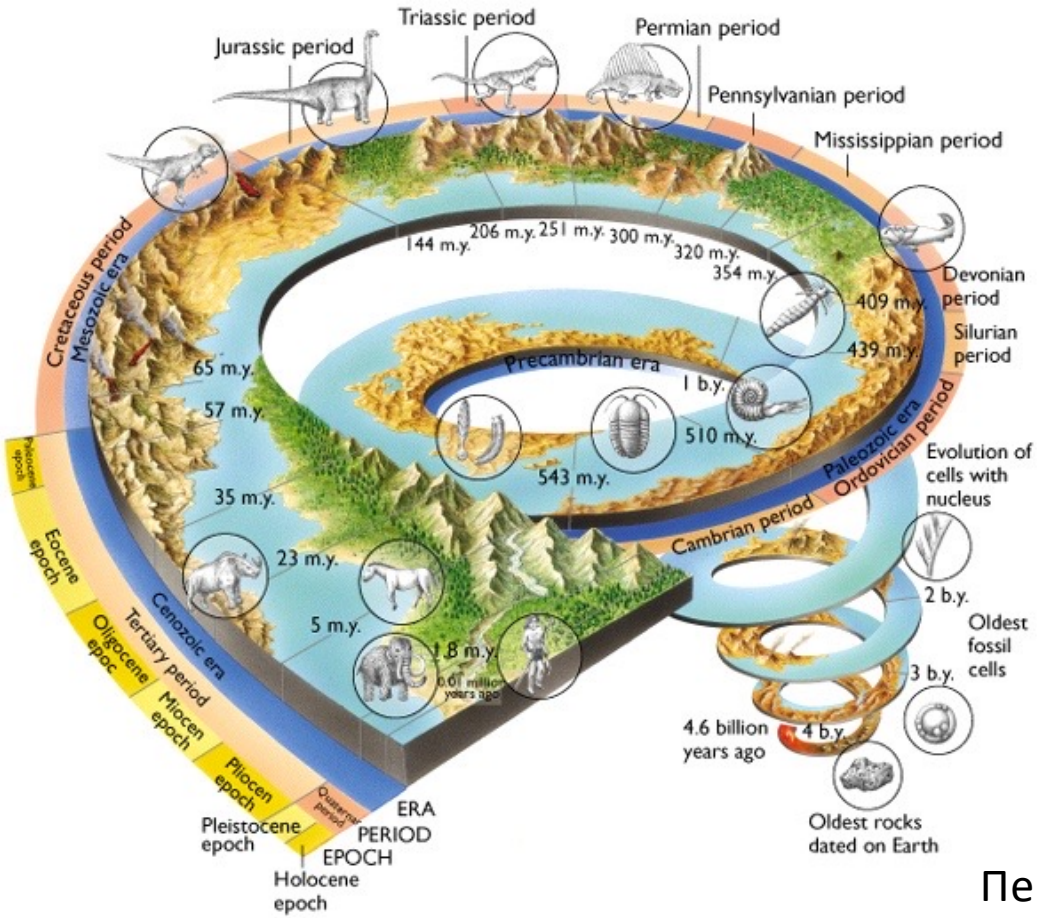
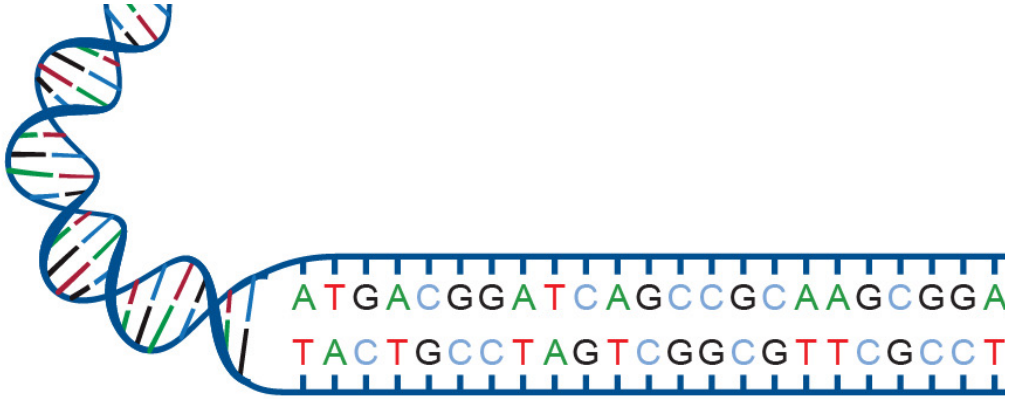
Использование суперкомпьютерных технологий в (био)молекулярных исследованиях

Алексей Константинович Шайтан
д.ф.-м.н., чл.-корр. РАН

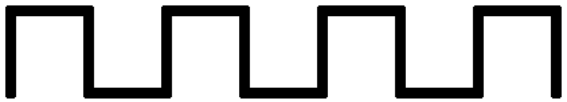
кафедра биоинженерии, биологический факультет
МГУ имени М.В. Ломоносова

Все началось с цифры …

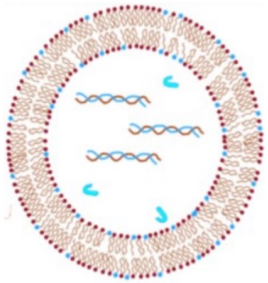




ДНК – цифровой код



1 0 1 0 1 0 1



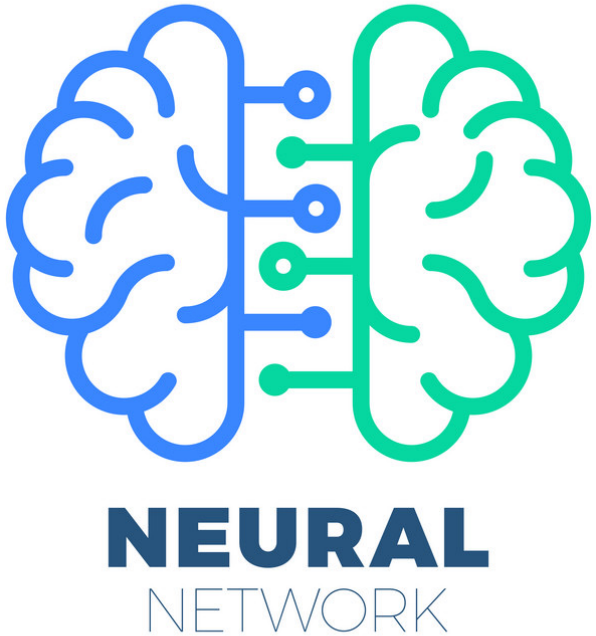
Первый репликатор

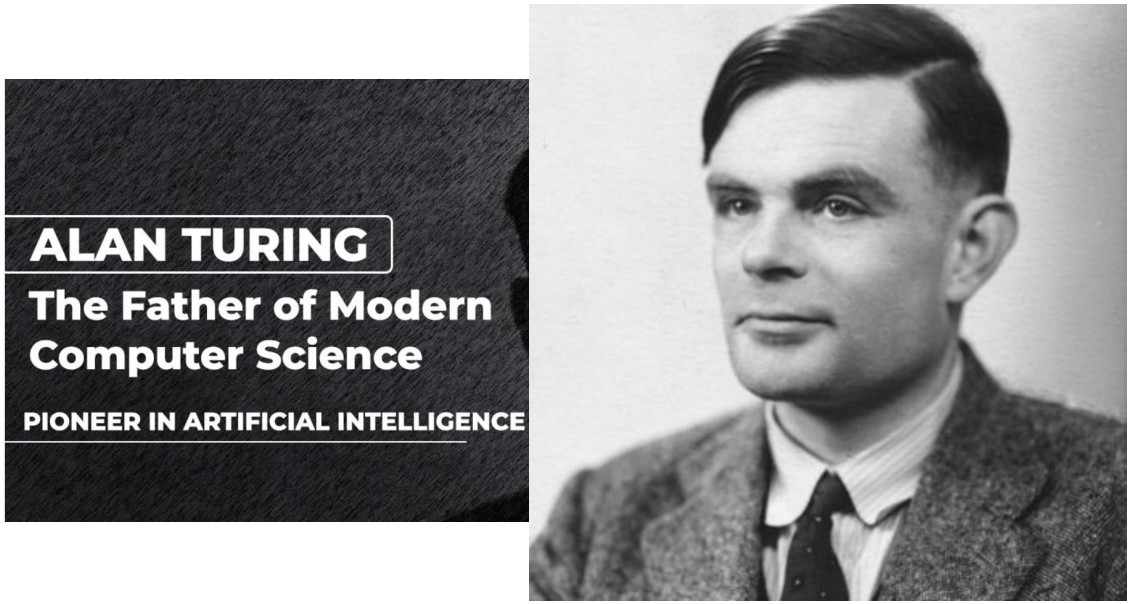
... запоминание и обработка информации …

<u>Мозг</u>

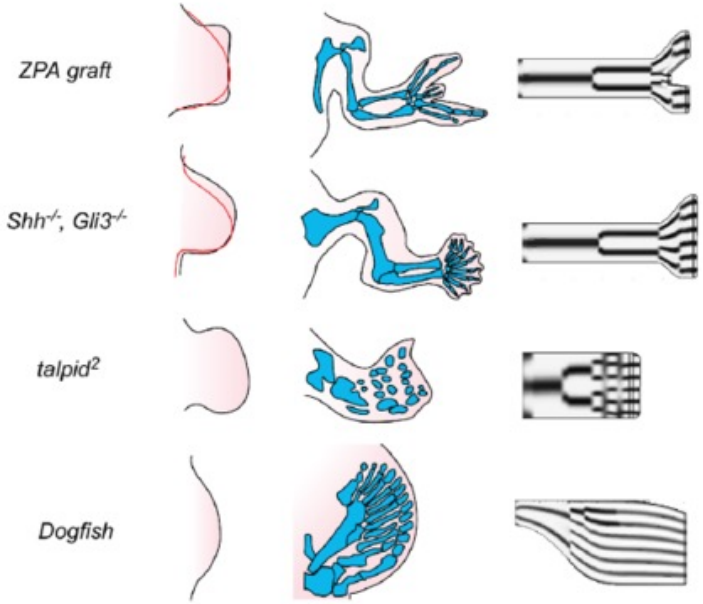
~100 млрд. нейронов

~100 триллионов синапсов





Алан Тьюринг



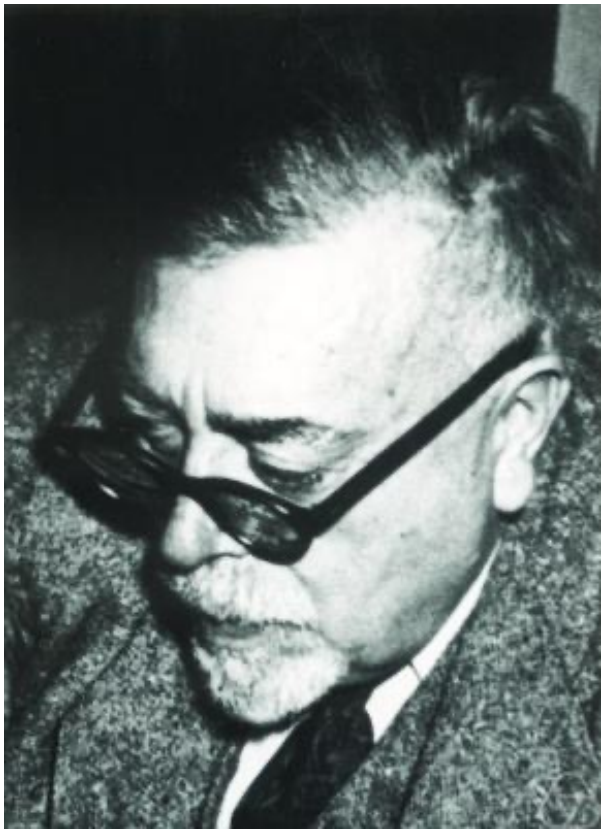
THE CHEMICAL BASIS OF MORPHOGENESIS

By A. M. TURING, F.R.S. *University of Manchester*

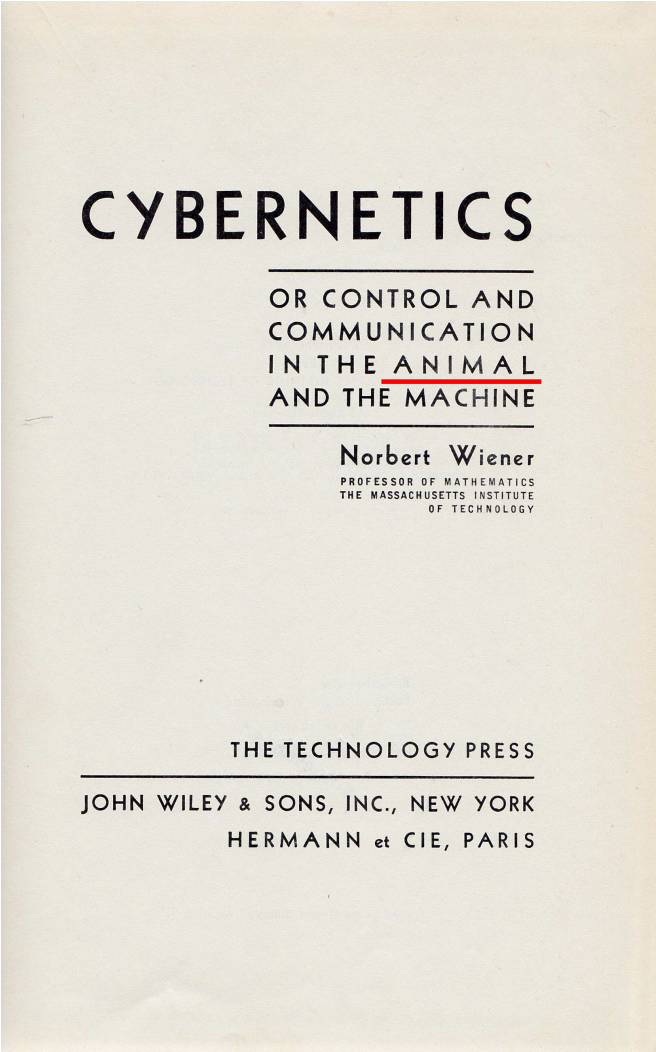
(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

теория морфогенеза

"originator of **cybernetics**, the science of communication as it relates to living things and machines”



Норберт Винер
1894 - 1964



CYBERNETICS

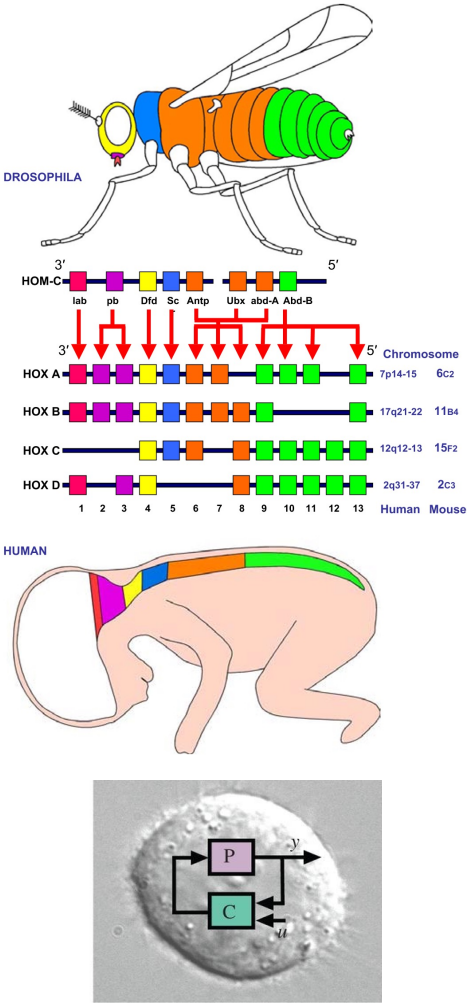
OR CONTROL AND COMMUNICATION IN THE ANIMAL AND THE MACHINE

Norbert Wiener
PROFESSOR OF MATHEMATICS
THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

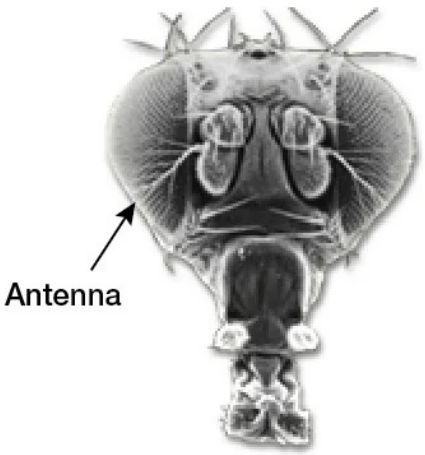
THE TECHNOLOGY PRESS

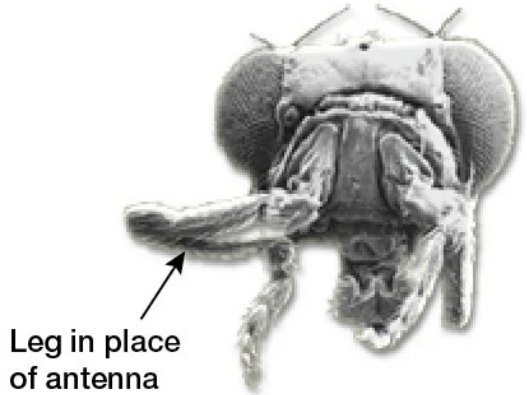
JOHN WILEY & SONS, INC., NEW YORK
HERMANN et CIE, PARIS

1948



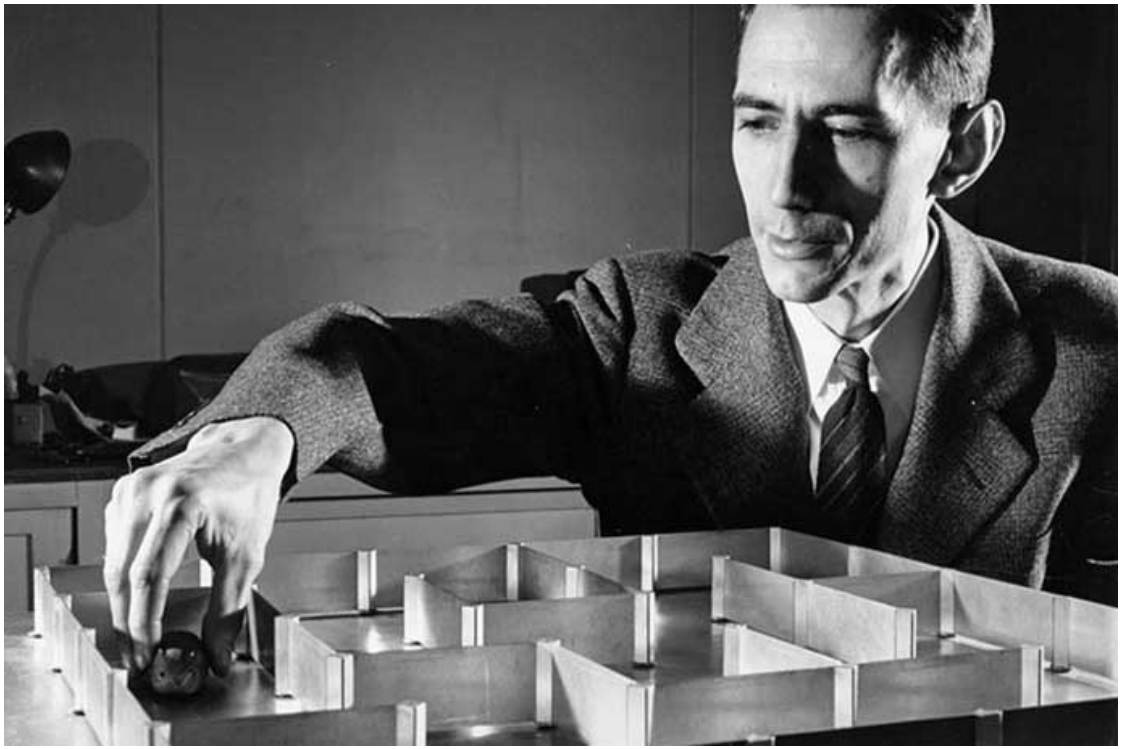
Программы развития организма закодированы в геноме



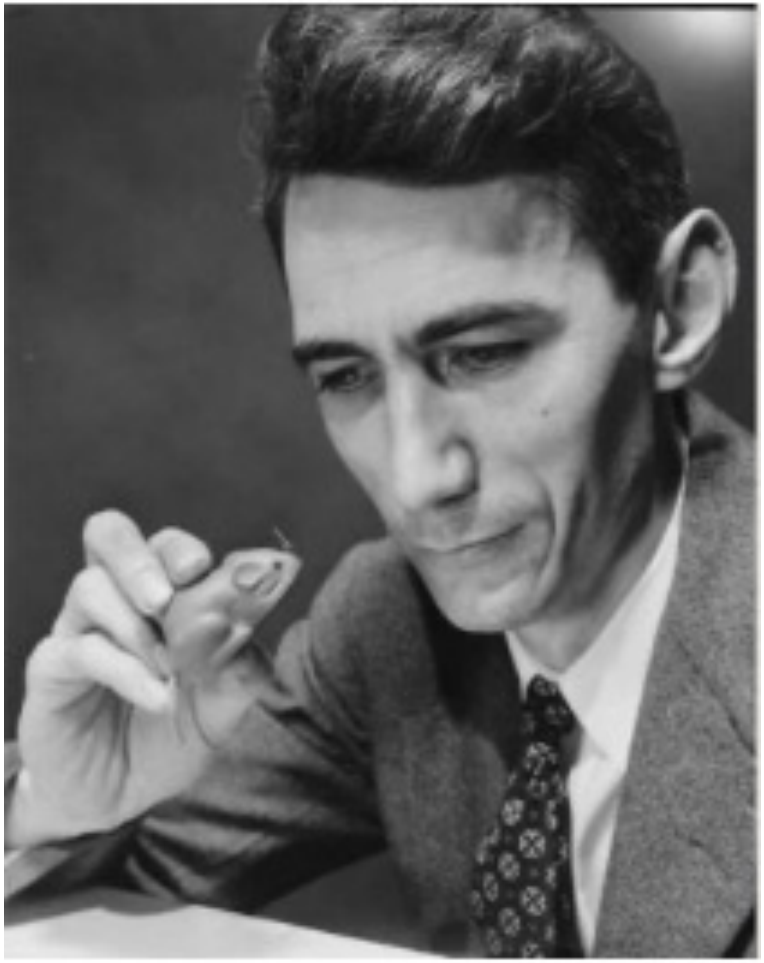


Мутации в одном гене приводят к изменению состава организма

“the father of information theory"



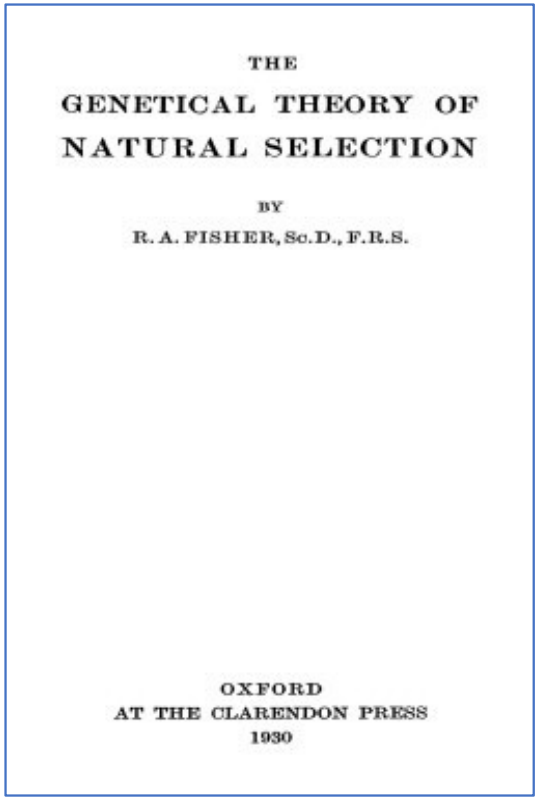
первые эксперименты по
искусственному интеллекту



Клод Шэнон
1916 –2001

"a genius who almost single-handedly created the foundations for modern **statistical science**"

Рональд Фишер
1890 - 1962



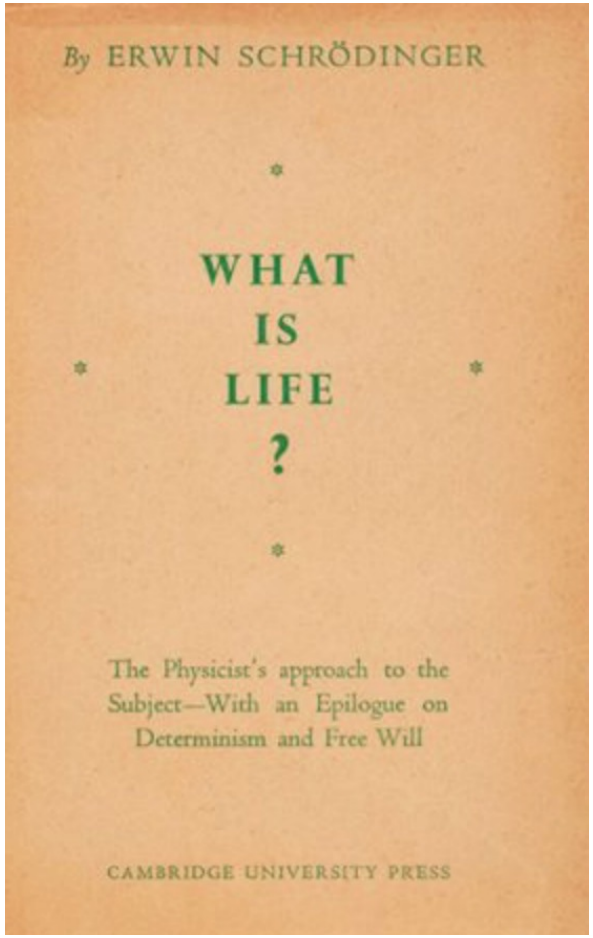
THE
GENETICAL THEORY OF
NATURAL SELECTION

BY
R. A. FISHER, Sc.D., F.R.S.

OXFORD
AT THE CLARENDON PRESS
1930

основоположник генетики

Эрвин Шредингер
1887 - 1961



предвосхитил
открытие ДНК

Клуб галстуков РНК

Георгий Гамов
1904 - 1968

RNA Tie Club

George Gamow (ALA)
F. Crick (TYR)
J. Watson (PRO)
Richard Feynman (GLY)
Melvin Calvin (HIS)
Edward Teller (LEU)
Leslie Orgel (THR)
Max Delbruck (TRY)
Sydney Brenner (VAL)
Erwin Chargaff (LYS)

Первая встреча 1954

(Супер)компьютерные технологии и решение актуальных задач биологии

<u>Большие вызовы</u> биологии XXI-века

Фундаментальные

- Можем ли мы научиться управлять материей также, как природные живые системы? Можем ли мы их превзойти?

- Можем ли мы задавать программы работы живых систем на уровне сравнимом с программами выработанными в ходе эволюции?

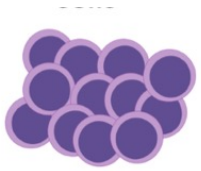
Практические

<u>Забота об окружающей среде</u>
Биоэкономика.
Переход к возобновляемым технологиям.
Устойчивое развитие.

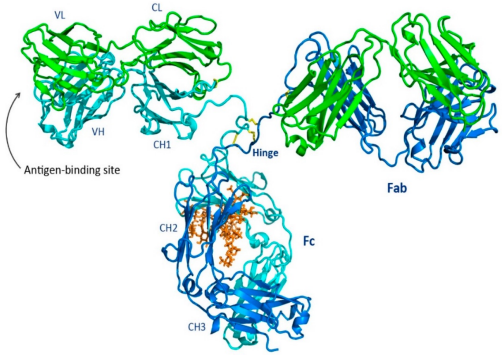
<u>Забота о человеке</u>
Питание.
Медицина.
Поддержание и расширение возможностей.

Цели и задачи

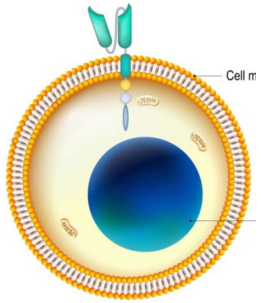
- Медицина и здоровье
  - Регенеративная медицина
  - Программирование иммунных клеток
  - Био-синтез лекарств
  - Органоиды
  - Нейротехнологии



Органоиды

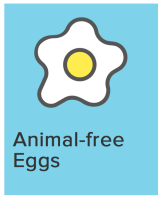


Антитело



”Умные” клетки-лекарства

- Сельское хозяйство и продукты питания
  - Производство еды, кормов
  - Живые удобрения



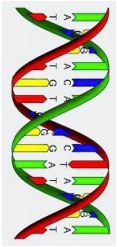
- Промышленность (синтез, переработка, энергетика)
  - Зеленая экономика
  - Биотопливо
  - Биопроизводство материалов (волокна, полимеры)

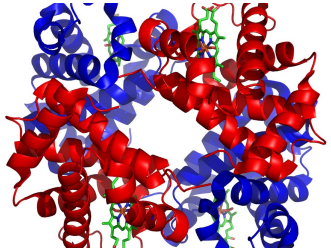
Наши дни

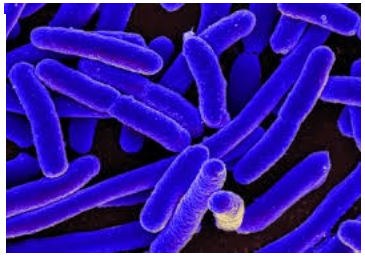
ИТ системы

- Возможности моделирования о обработки информации
- ИТ системы – пример того, что мы научились создавать сложные системы. Разработаны методы создания сложных систем

Био системы

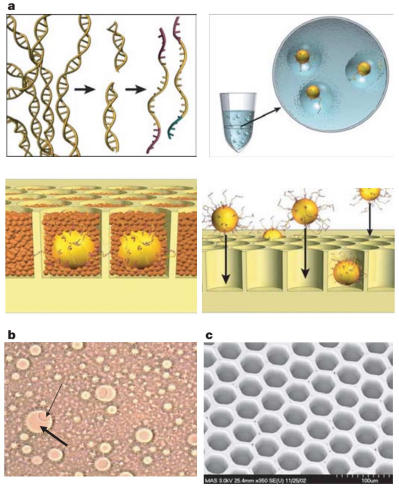






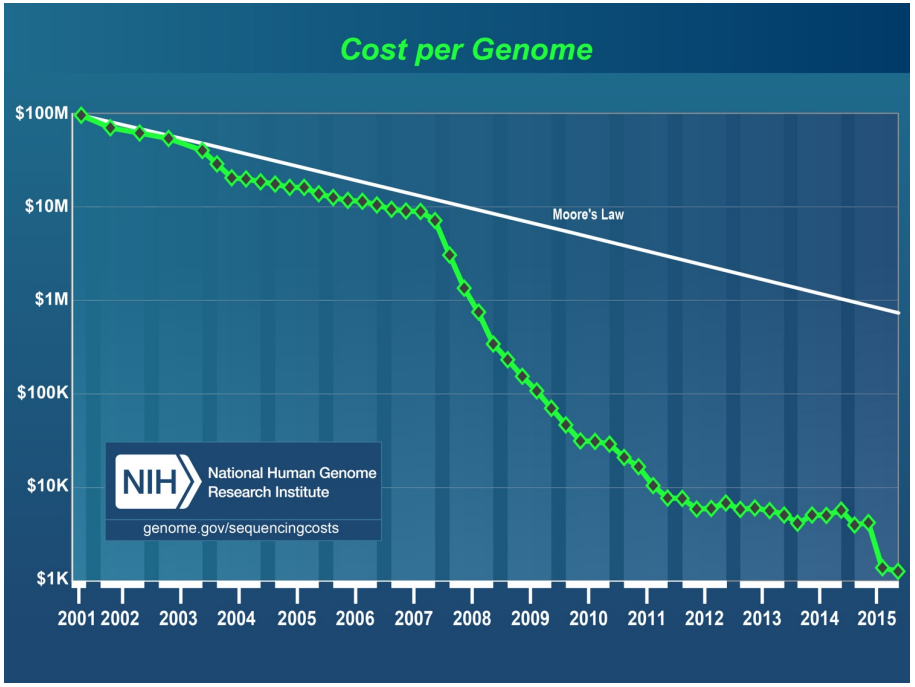
Человек
Геном: ~ 3 Гбайт
10^13 клеток

Бактерия E. coli
Геном: ~ 5 Мбайт



2003 **2022**

Технологии секвенирования нового поколения (NGS)



Удешевления секвенирования стало возможным благодаря применению микротехнологий и компьютерных технологий обработки данных

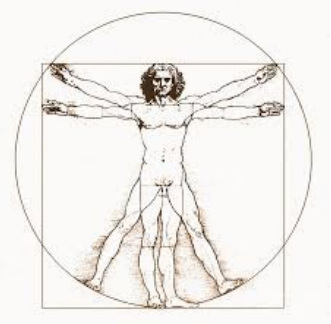
Библиотека Л. Н. Толстого в Ясной Поляне состоит из 22 тысяч книг и номеров журналов на 35 языках

<u>Геном человека (2001 год)</u>

3 млрд нуклеотидов

262 тыс страниц

3 Гбайта



"What I cannot create, I do not understand"

Р. Фейнман
(1918-1988)

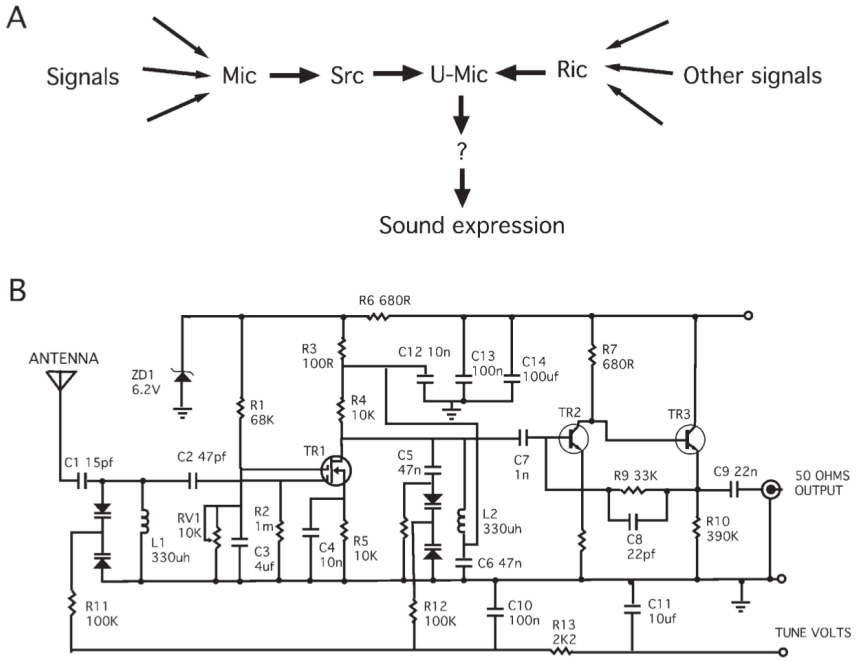
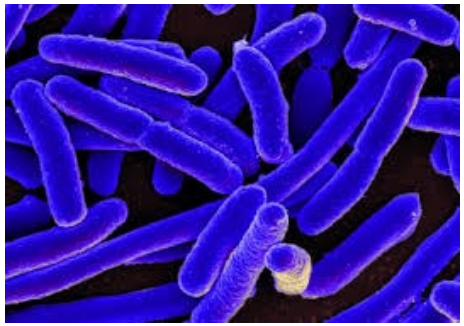


Figure 3. The tools used by biologists and engineers to describe processes of interest

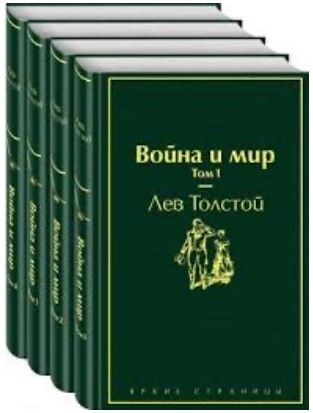
Y. Lazebnik (2002)

О сложности живых систем



Технологии придуманные человеком начали догонять живые системы по информационной сложности

Бактерия E. coli
Геном: ~ 5Мбайт
~4000 генов
3–4 миллиона молекул белков



3 млн. символов

~ 1 Тб

Самолет
~ 5 млн частей

Компьютерный процессор
~100 млрд. транзисторов

Человек
геном ~3 млрд. нуклеотидов
30 триллионов клеток

Linux

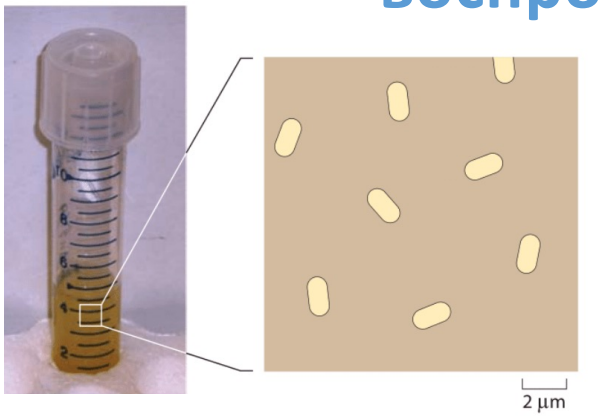
ядро ~3 млрд. букв

Мозг
~100 млрд. нейронов
~100 триллионов синапсов

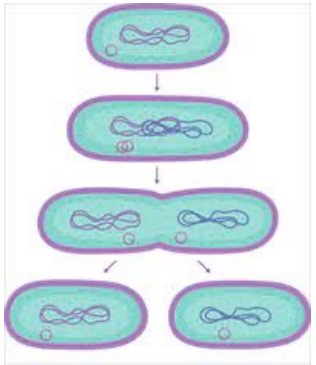
ChatGPT

200 млрд. параметров

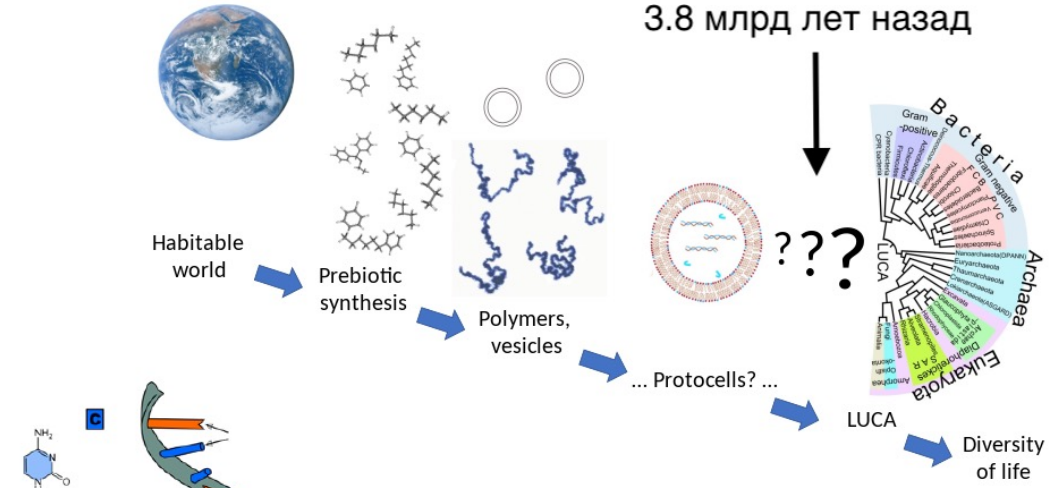
Но живые системы умеют управлять материей, воспроизводиться и работать «параллельно»!

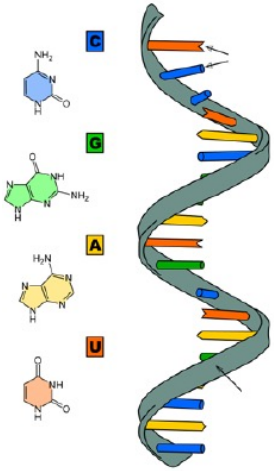


10^9 клеток на миллилитр

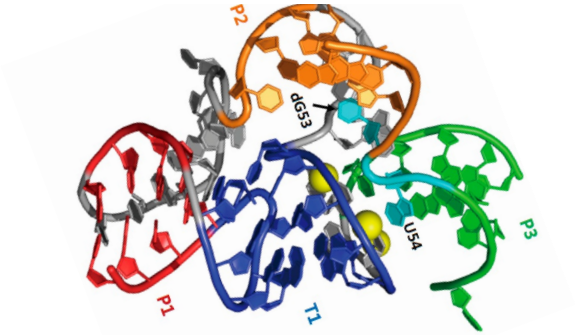


Клетки делятся





Молекулы РНК

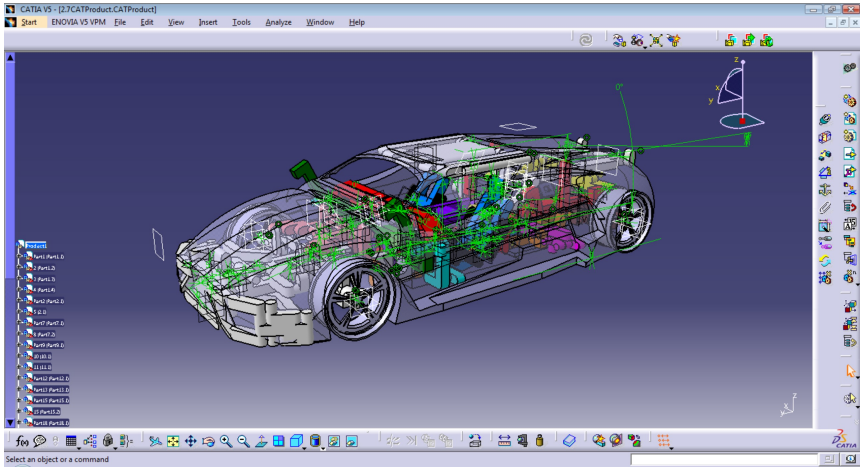


Для цепи длиной 50 нуклеотидов, $\mathbf{4^{50} \cong 10^{30}}$ (!) вариантов цепи

Как совладать со сложностью живых систем?

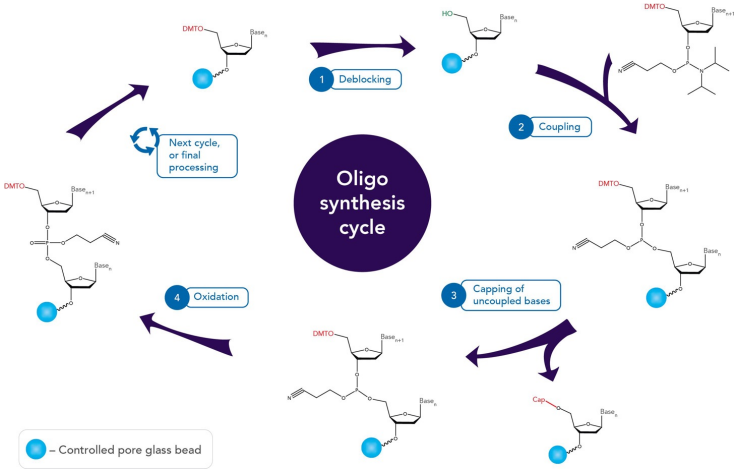
- Использование компьютерных технологий для анализа и моделирования живых систем. Включая технологии искусственного интеллекта.

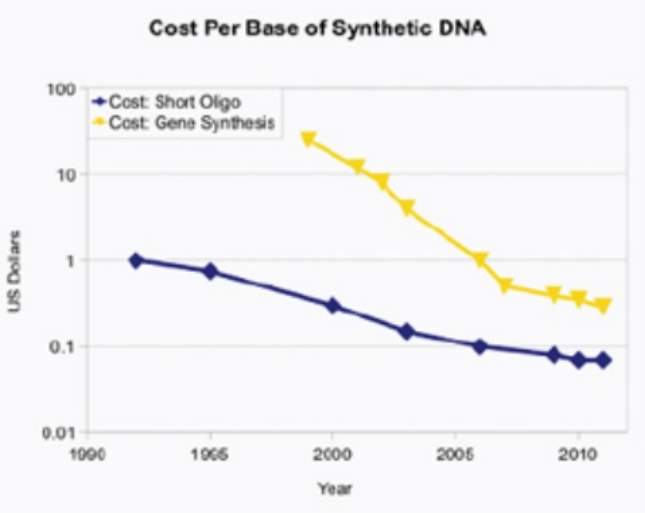
- Использование принципов, которые мы применяем для создания сложных технических систем



Пост-геномная эра: «запись» (синтез) ДНК

Традиционный синтез ДНК



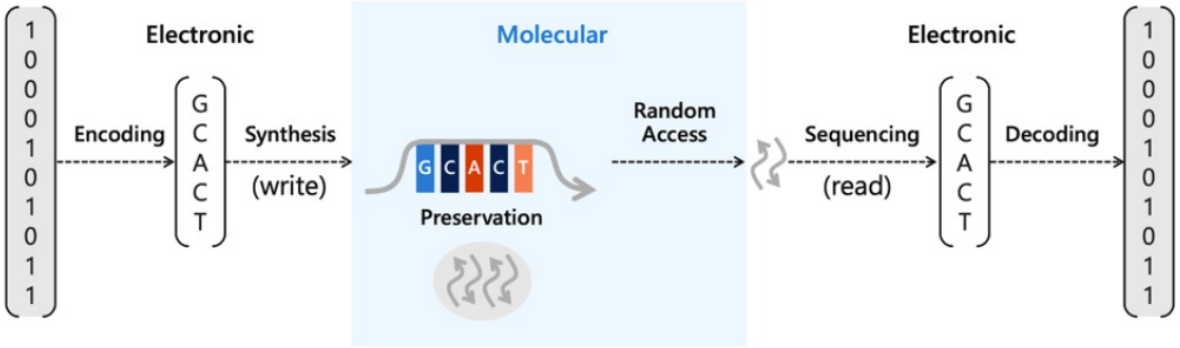


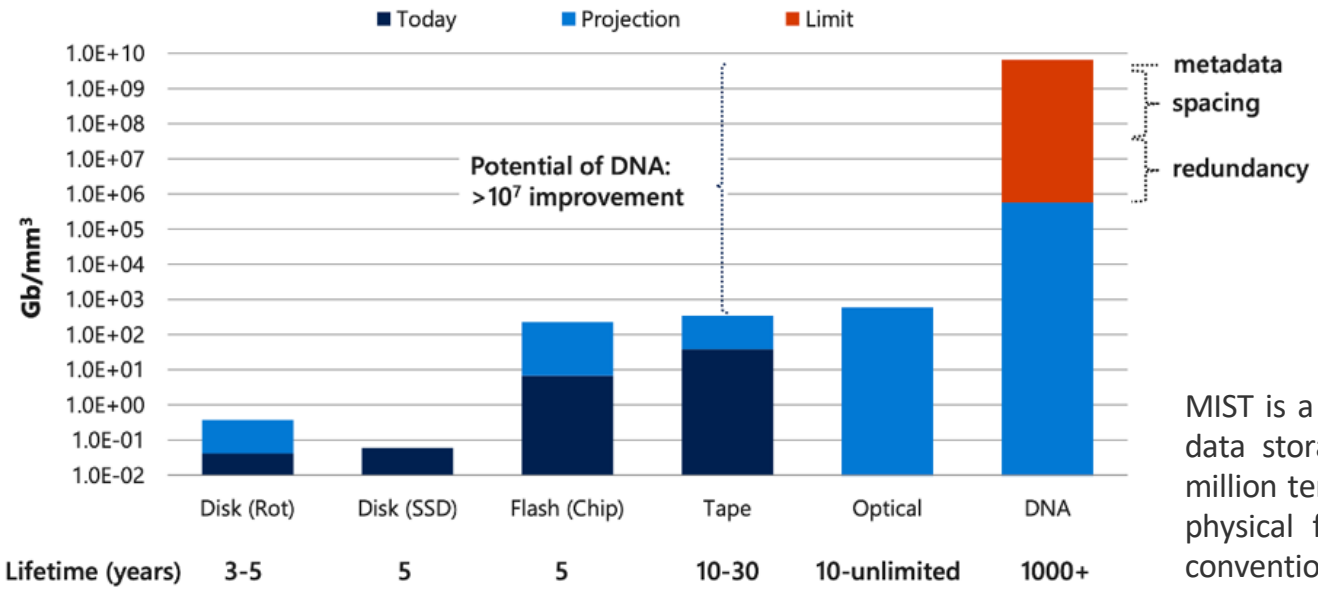
2023: цена синтеза ДНК от 7 центов за нуклеотид

Технологии микроматричного синтеза ДНК

Кремниевый чип
> 1 млн луночек для синтеза ДНК

Using DNA for long-term data storage





DNA DATA STORAGE ALLIANCE

illumina · Microsoft · TWIST BIOSCIENCE · Western Digital.

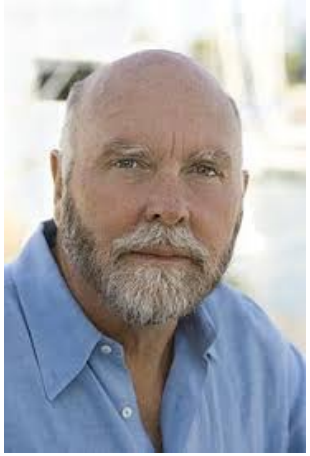
Office of the Director of National Intelligence
Intelligence Advanced Research Projects Activity
IARPA
Creating Advantage through Research and Technology

Molecular Information Storage program

MIST is a multi-year research effort to develop next-generation data storage technologies that can scale into the exabyte (1 million terabyte) regime, and beyond, with significantly reduced physical footprint, power and cost requirements, relative to conventional approaches.

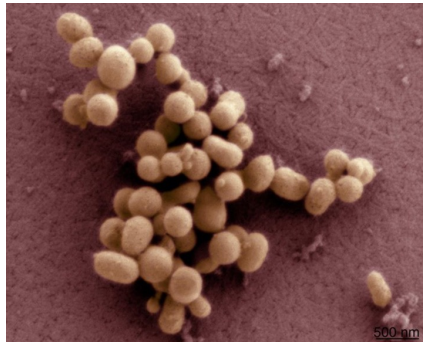
Goal $1/TB by 2030

Первая «искусственная жизнь»



Craig Venter

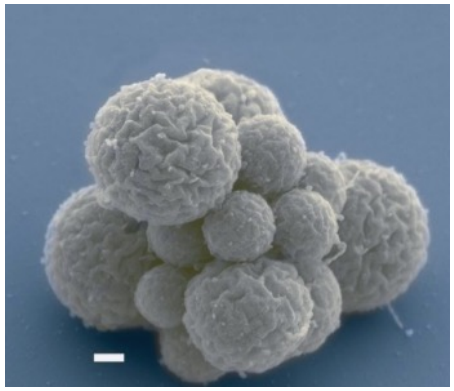
2010 *Mycoplasma mycoides* JCVI-syn1.0



Первый искусственный геном (~$40 млн)

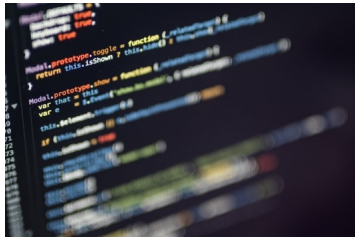
Human	~20,000–25,000
Escherichia coli (K12 strain)	~4500
Syn 1.0	901
*Mycoplasma genitalium**	525
Syn 3.0	473

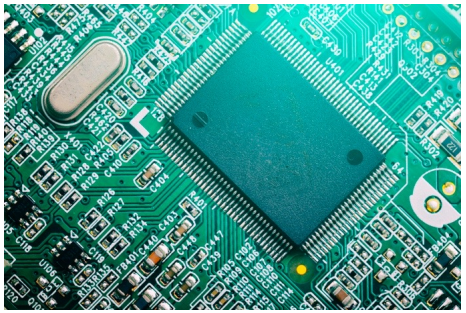
2016 *Syn3.0*



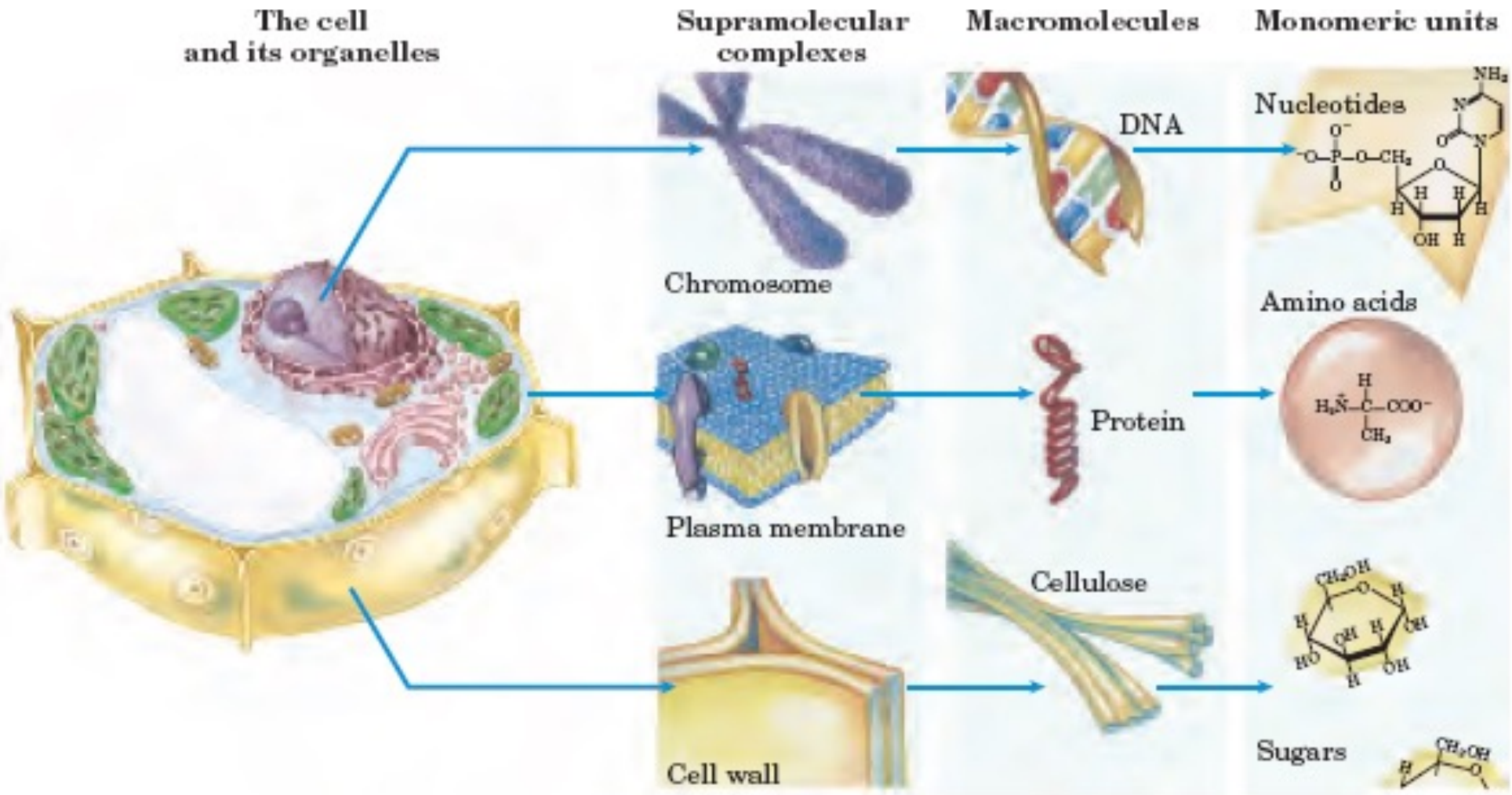
Минимальный геном

- Software boots into hardware
- Genetic code boots into cells





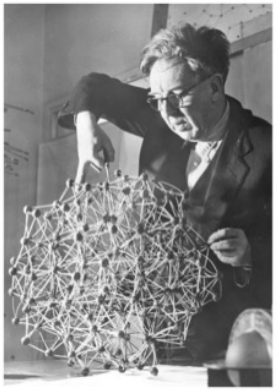
(Супер)компьютерные технологии для изучения структуры и динамики биомолекул


The cell
and its organelles
Supramolecular
complexes
Macromolecules
Monomeric units
Chromosome
DNA
Nucleotides
Plasma membrane
Protein
Amino acids
Cellulose
Cell wall
Sugars

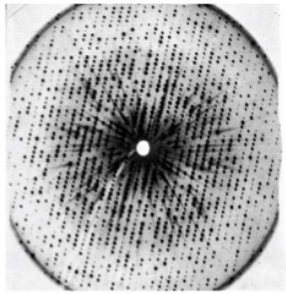
Задача моделирования биомолекул

М. Перуц, Д.К. Кендрю

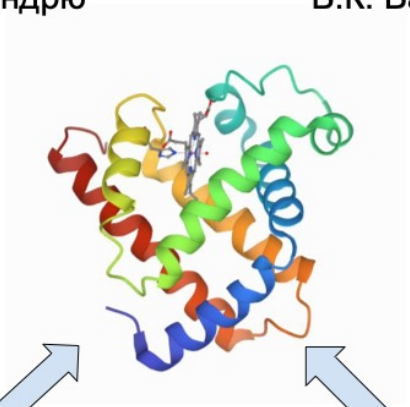
Б.К. Вайнштейн



Дж. Д. Бернал



Дифракционная картина белка



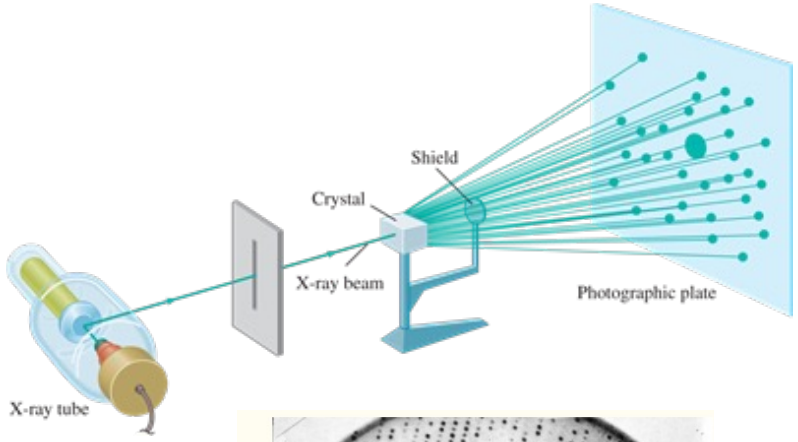
Моделирование по экспериментальным данным

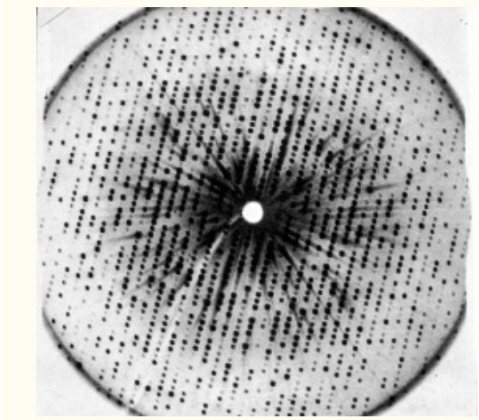
Имитационное моделирование (molecular simulations)

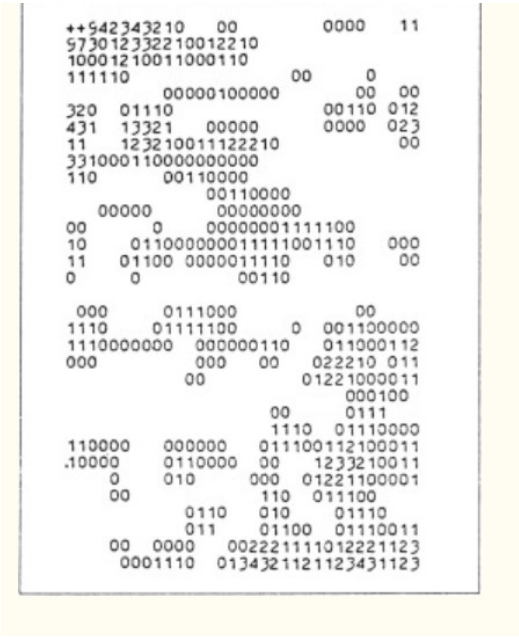
Компьютер ENIAC, 1947

Суперкомпьютер Ломоносов-2, 2015

1952: первые компьютерные программы для решения обратной задачи дифракции







$$\rho(xyz) = \frac{1}{V} \sum_{\substack{hkl \\ -\infty}}^{+\infty} |F(hkl)| \,.\, e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$$

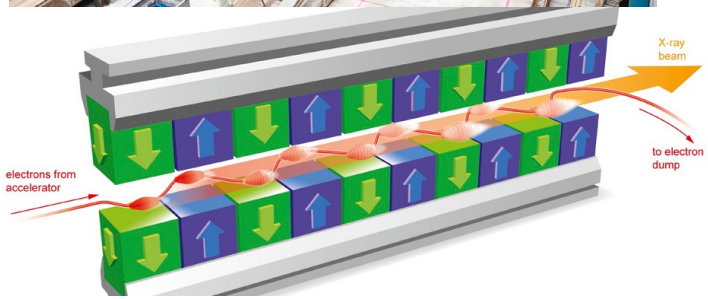
Amplitudes

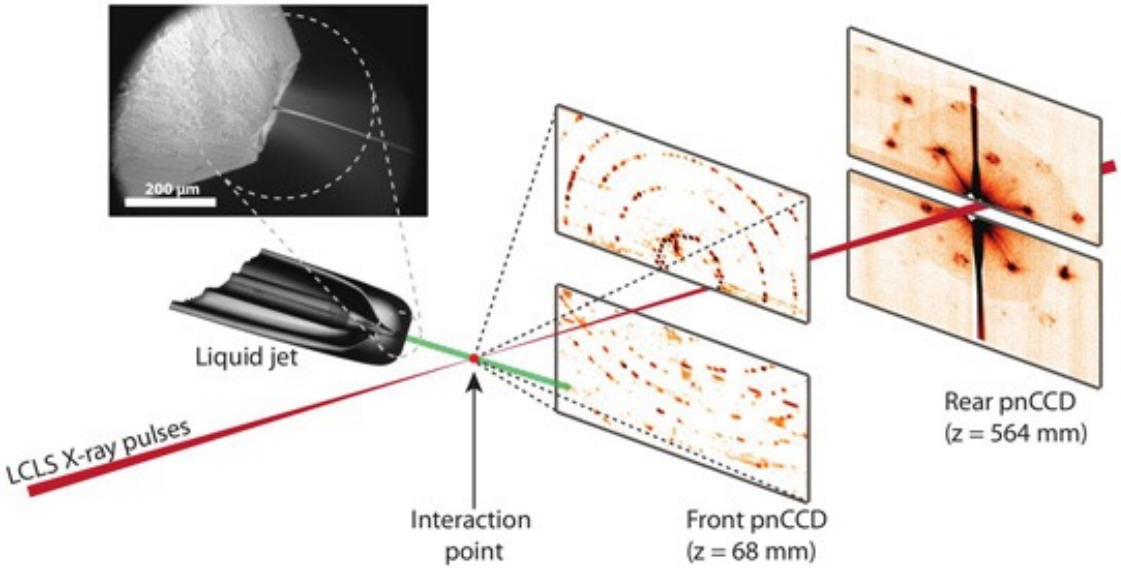
Phases?

~ 1 Гб данных

21 век: Лазеры на свободных электронах

European XFEL, Hamburg

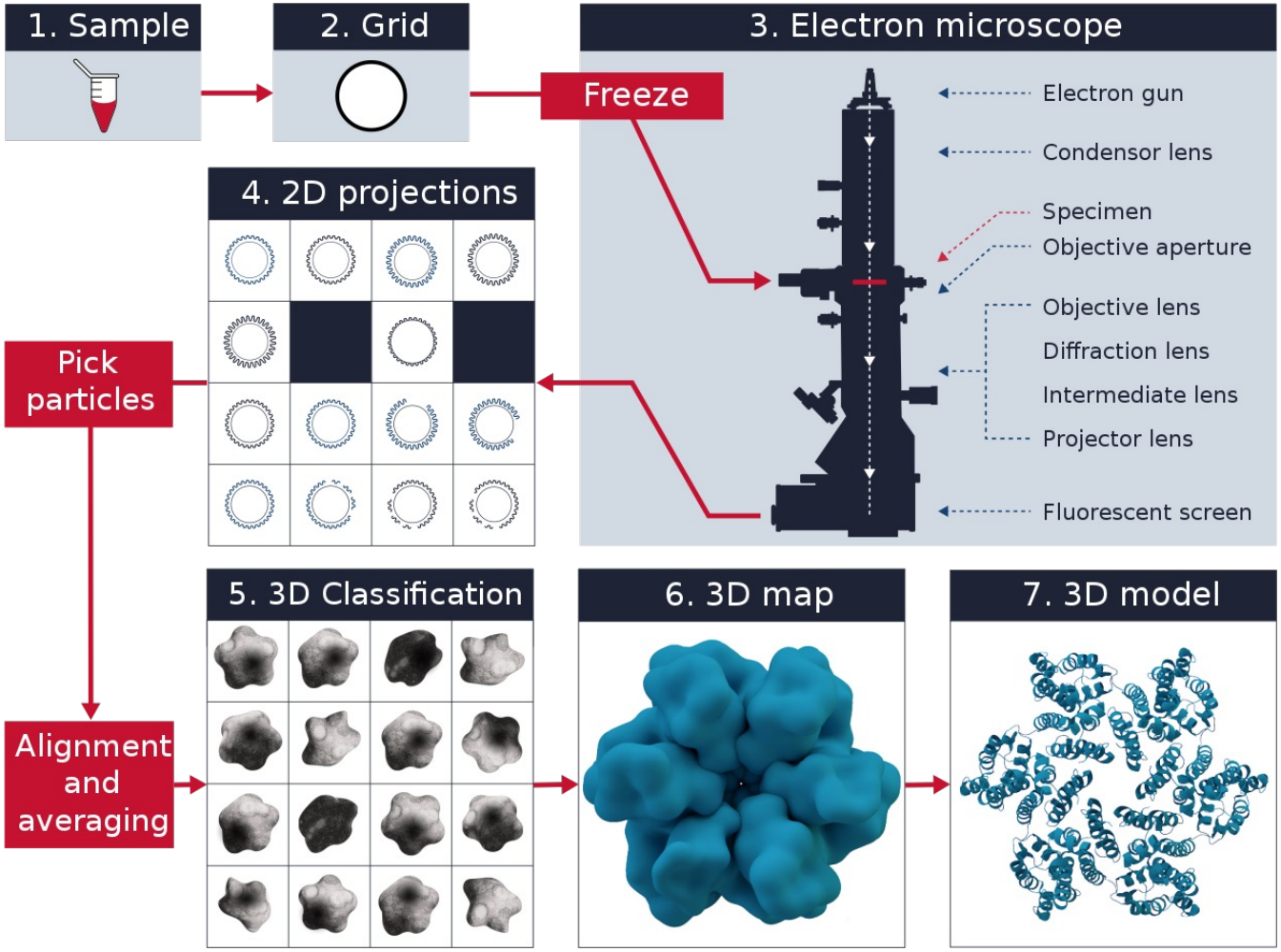




27000 импульсов в секунду

~ 100 Тб данных

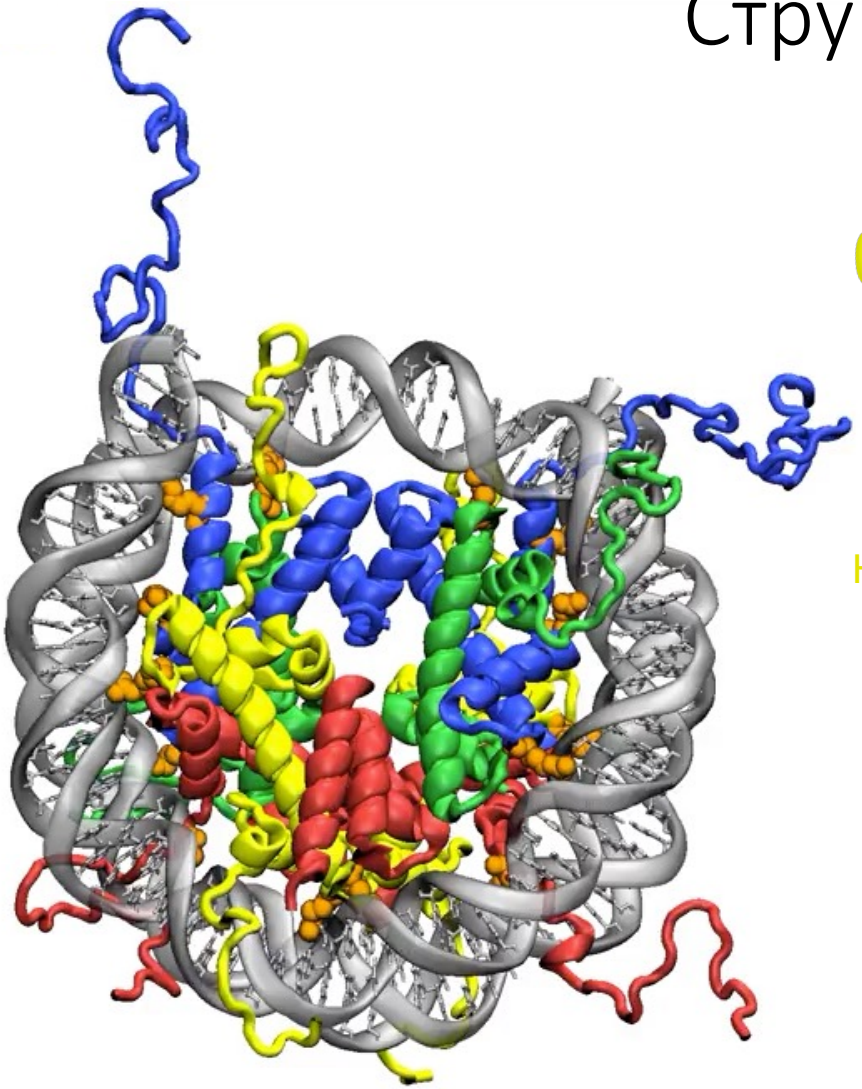
Криогенная электронная микроскопия



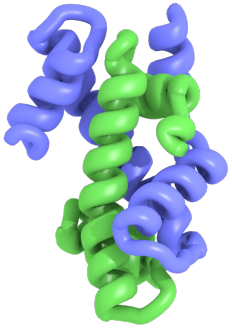
Обратная задача: восстановление структуры из ~10^6 изображений

~ 10 Тб данных

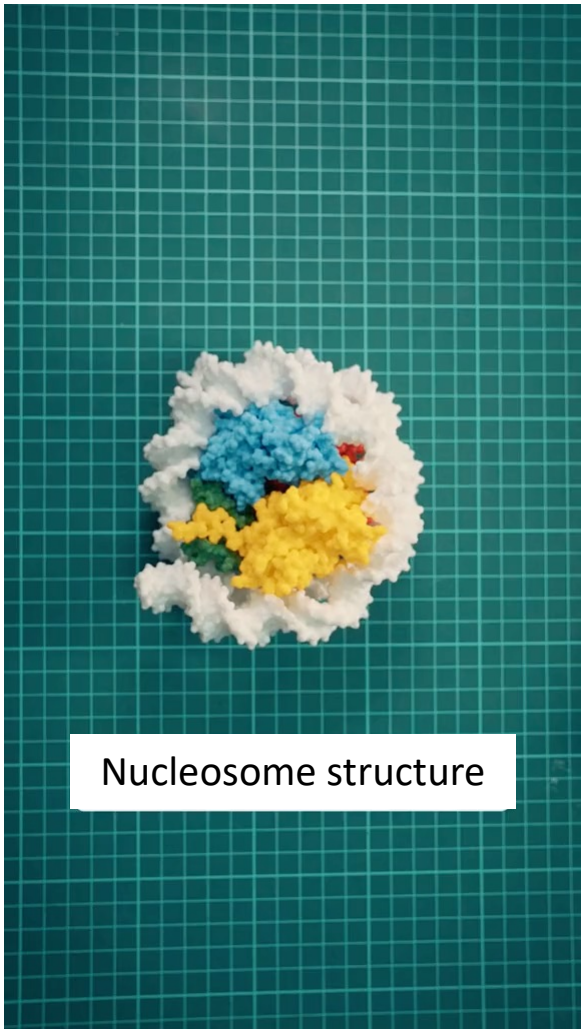
Структуры статичны



Histone
H2A-H2B dimer



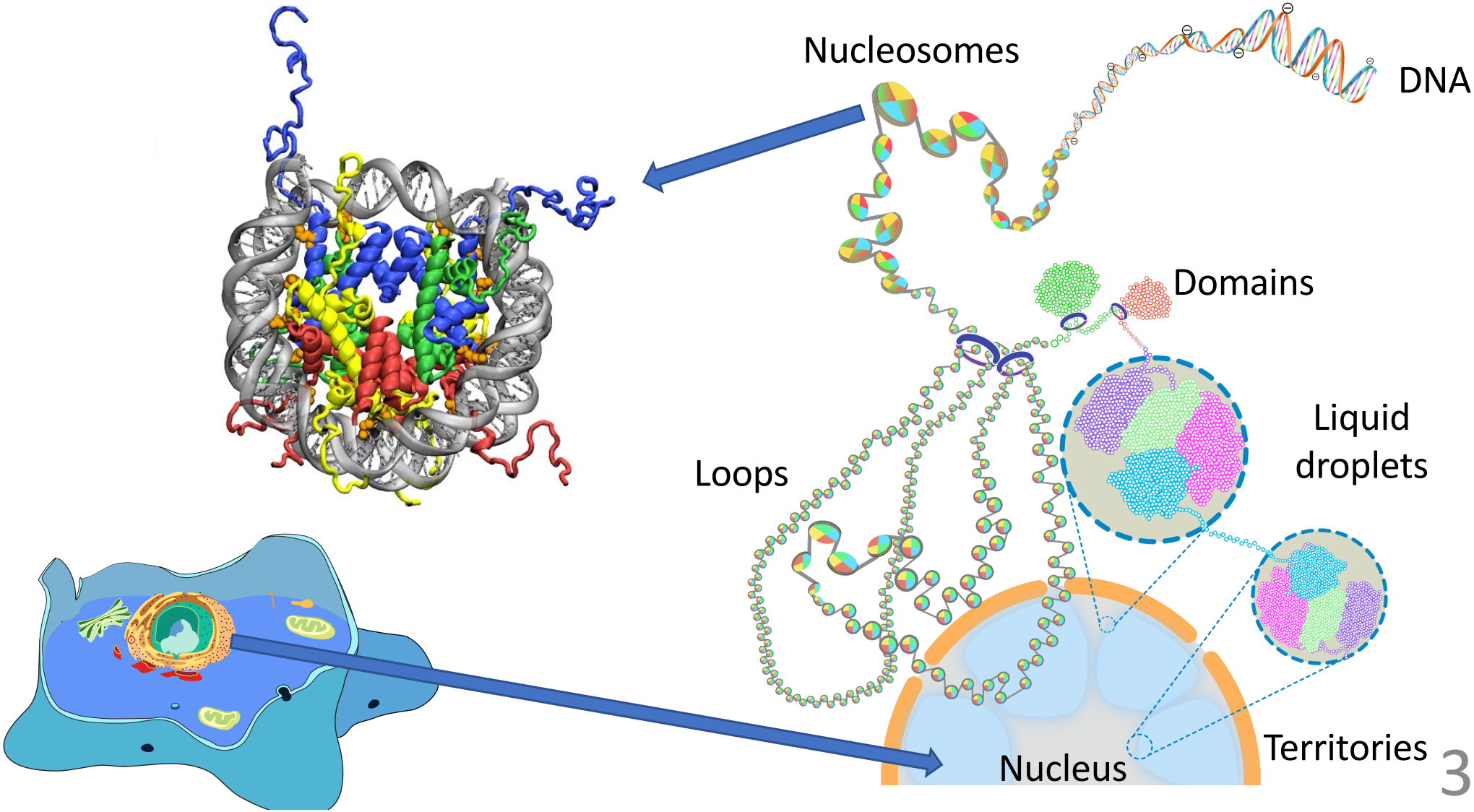
Histone
H3-H4 dimer



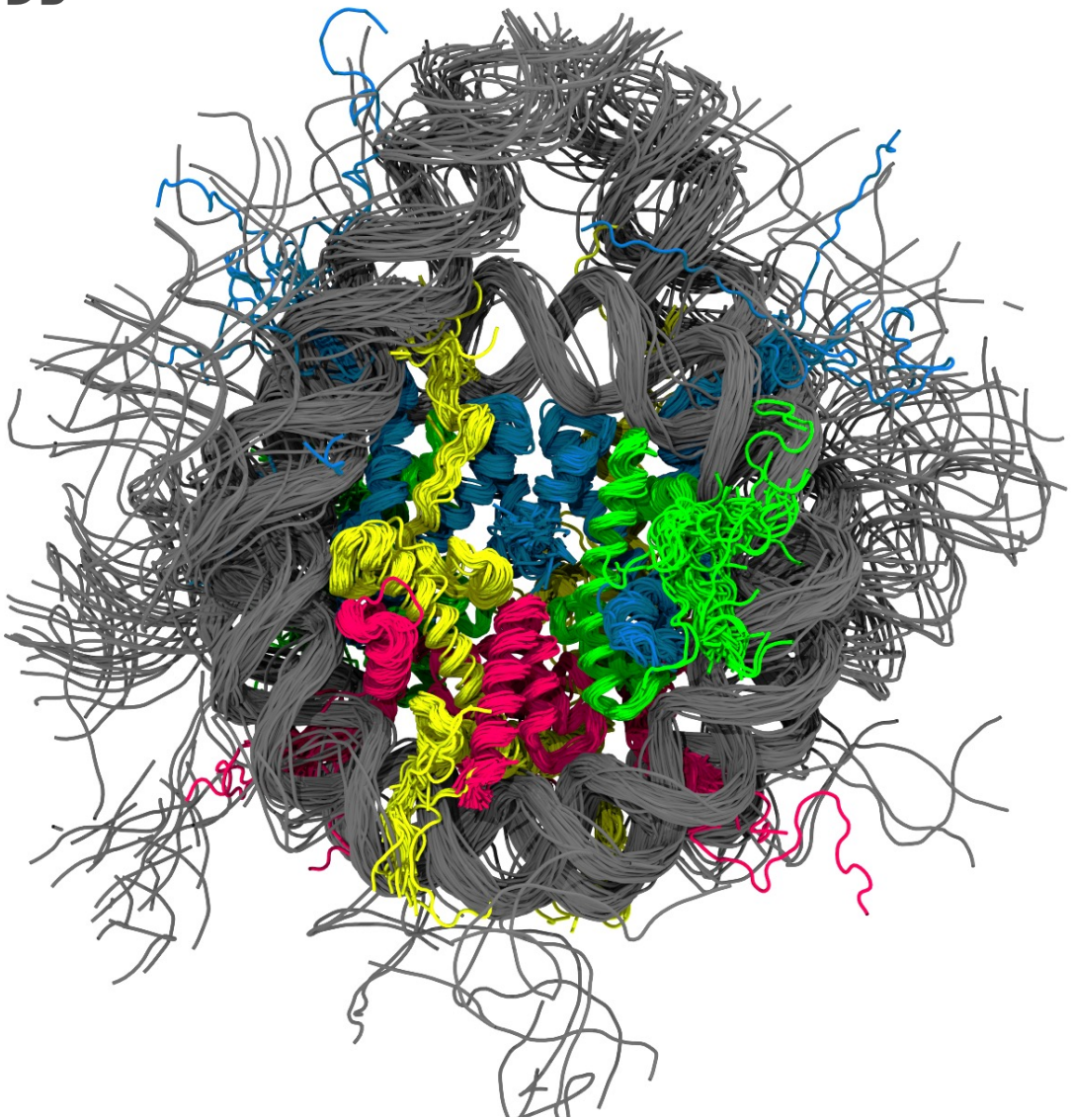
Davey, C. A.; Sargent, D. F.; Luger, K.; Maeder, A. W.; Richmond, T. J. *J. Mol. Biol.* **2002**, *319* (5), 1097

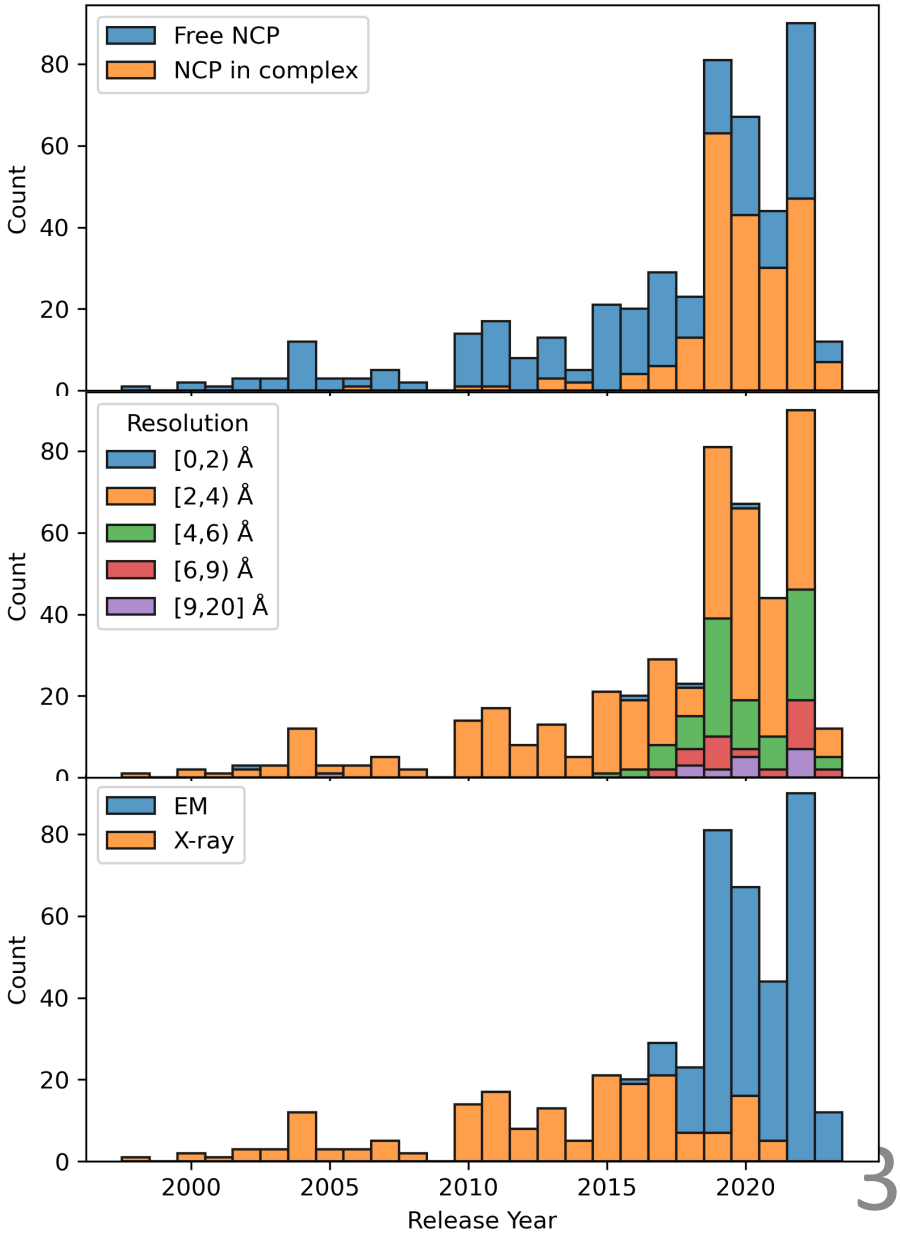
https://github.com/molsim/nuclLEGO

The complexity of chromatin structure and dynamics

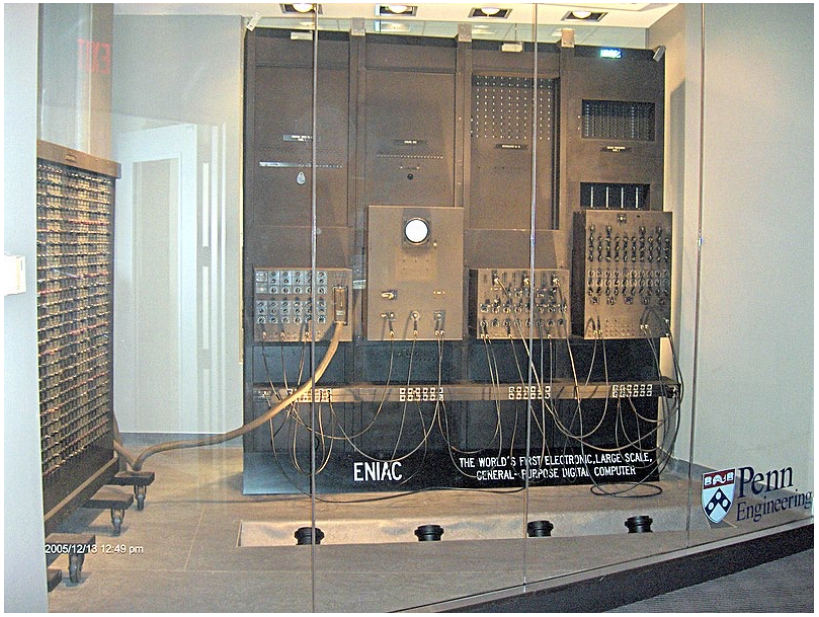


Nucleosome DB statistics

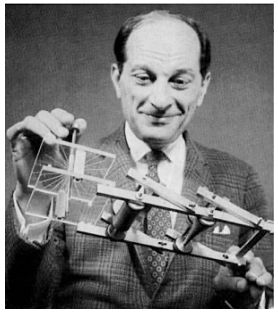


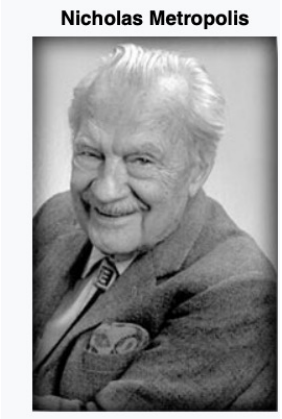


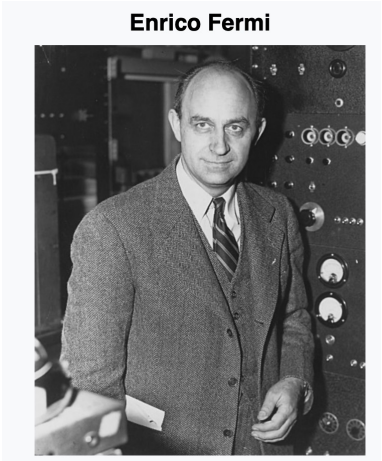
Molecular simulations



ENIAC - 1945







Метод Монте-Карло – 1949

Метод молекулярной динамики – 1955

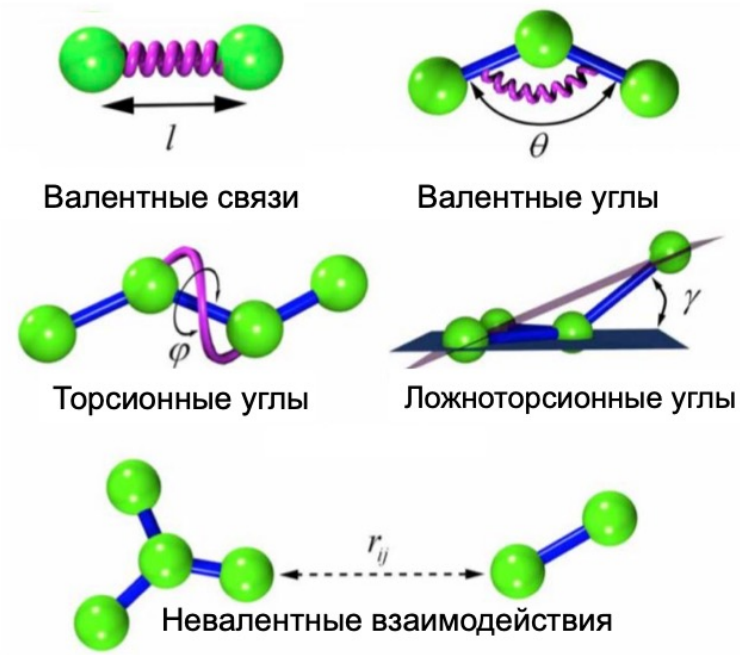
МД твердых сфер Алдер и Вейнрайт – 1957
1970-ые первые работы по МД белков – Карплюс, Варшел, Левитт

Методы молекулярной механики и динамики

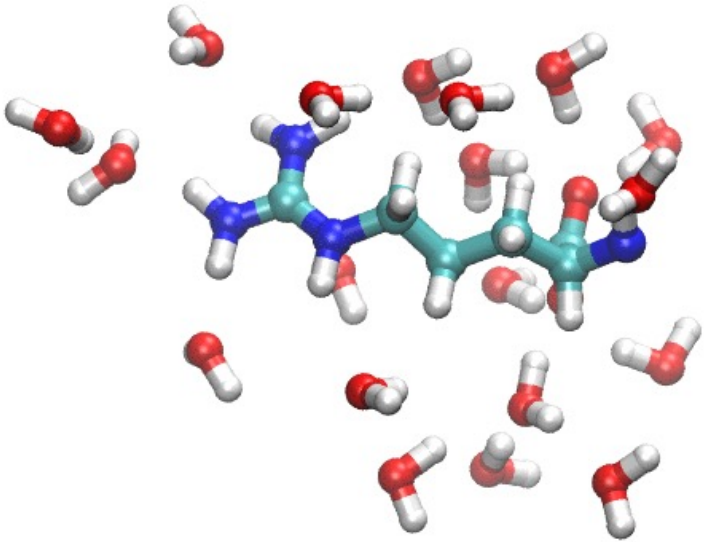
Полноатомное приближение

$$U(\{\vec{r}_i\}) = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 + \sum_{angels} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{torsions} \frac{1}{2} V_n [1 + \cos(n\varphi - \varphi_0)] +$$

$$\sum_{impropers} \frac{1}{2} k_\gamma (\gamma - \gamma_0)^2 + \sum_{j+1}^{N-1} \sum_{i=j+1}^{N} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} f_{ij}$$



Численное решение уравнений движения ($\vec{F} = m\vec{a}$)



2013 Chemistry Prize

Taking the Experiment to Cyberspace

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

- Press release
- Popular information
- Advanced information

Photo © Harvard University

Martin Karplus

Martin Karplus, U.S. and Austrian citizen. Born 1930 in Vienna, Austria. Ph.D. 1953 from California Institute of Technology, CA, USA. Professeur Conventionné, Université de Strasbourg, France and Theodore William Richards Professor of Chemistry, Emeritus, Harvard University, Cambridge, MA, USA.

- More on Martin Karplus

Photo: S. Fisch

Michael Levitt

Michael Levitt, U.S., British and Israeli citizen. Born 1947 in Pretoria, South Africa. Ph.D. 1971 from University of Cambridge, UK. Robert W. and Vivian K. Cahill Professor in Cancer Research, Stanford University School of Medicine, Stanford, CA, USA.

- Have a look at Michael Levitt's photo gallery

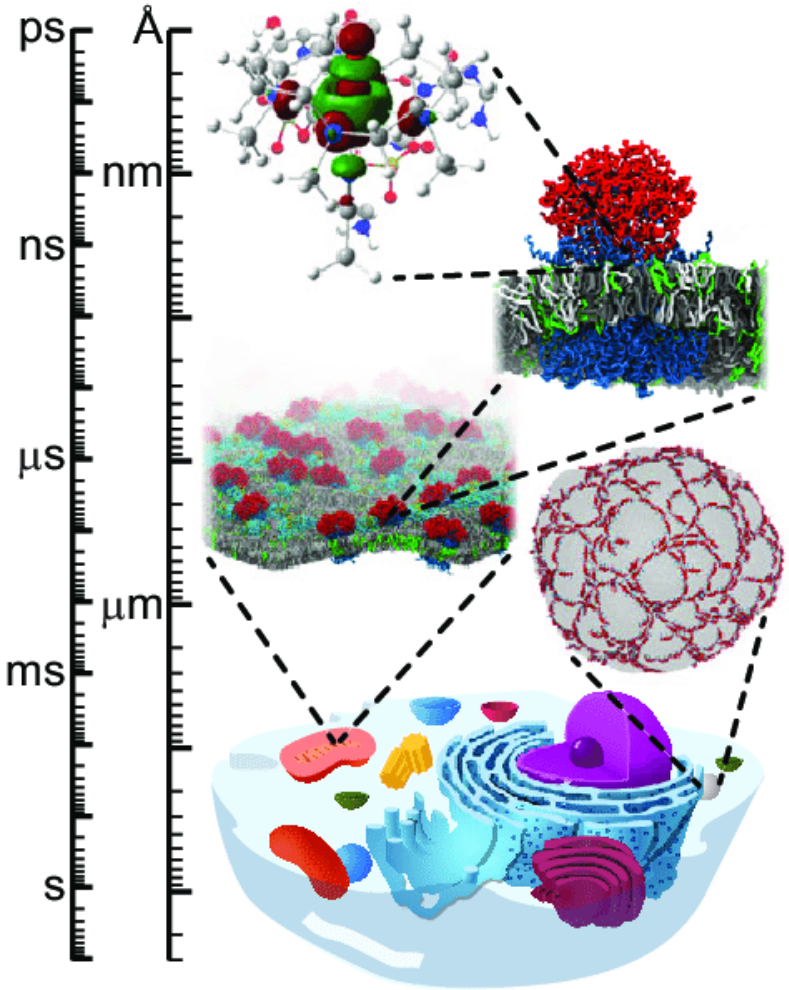
Photo: Wikimedia Commons

Arieh Warshel

Arieh Warshel, U.S. and Israeli citizen. Born 1940 in Kibbutz Sde-Nahum, Israel. Ph.D. 1969 from Weizmann Institute of Science, Rehovot, Israel. Distinguished Professor, University of Southern California, Los Angeles, CA, USA.

- Interviews with Chemistry Laureate Arieh Warshel

Методы моделирования как «вычислительный микроскоп»



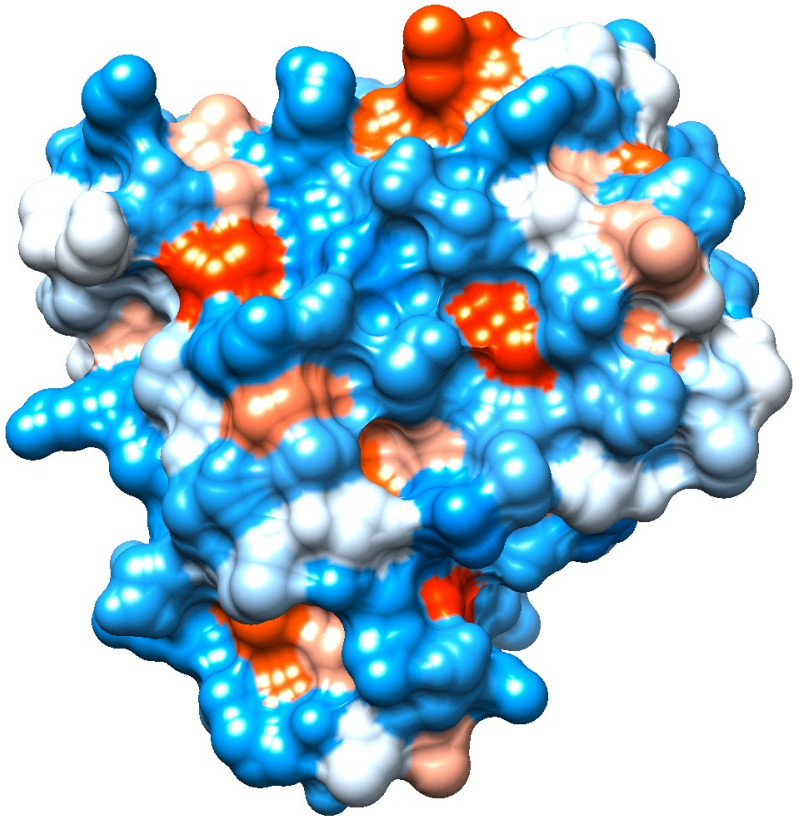
Quantum
- atoms, electrons and electron clouds included
- explicit solvent
- quantum mechanics

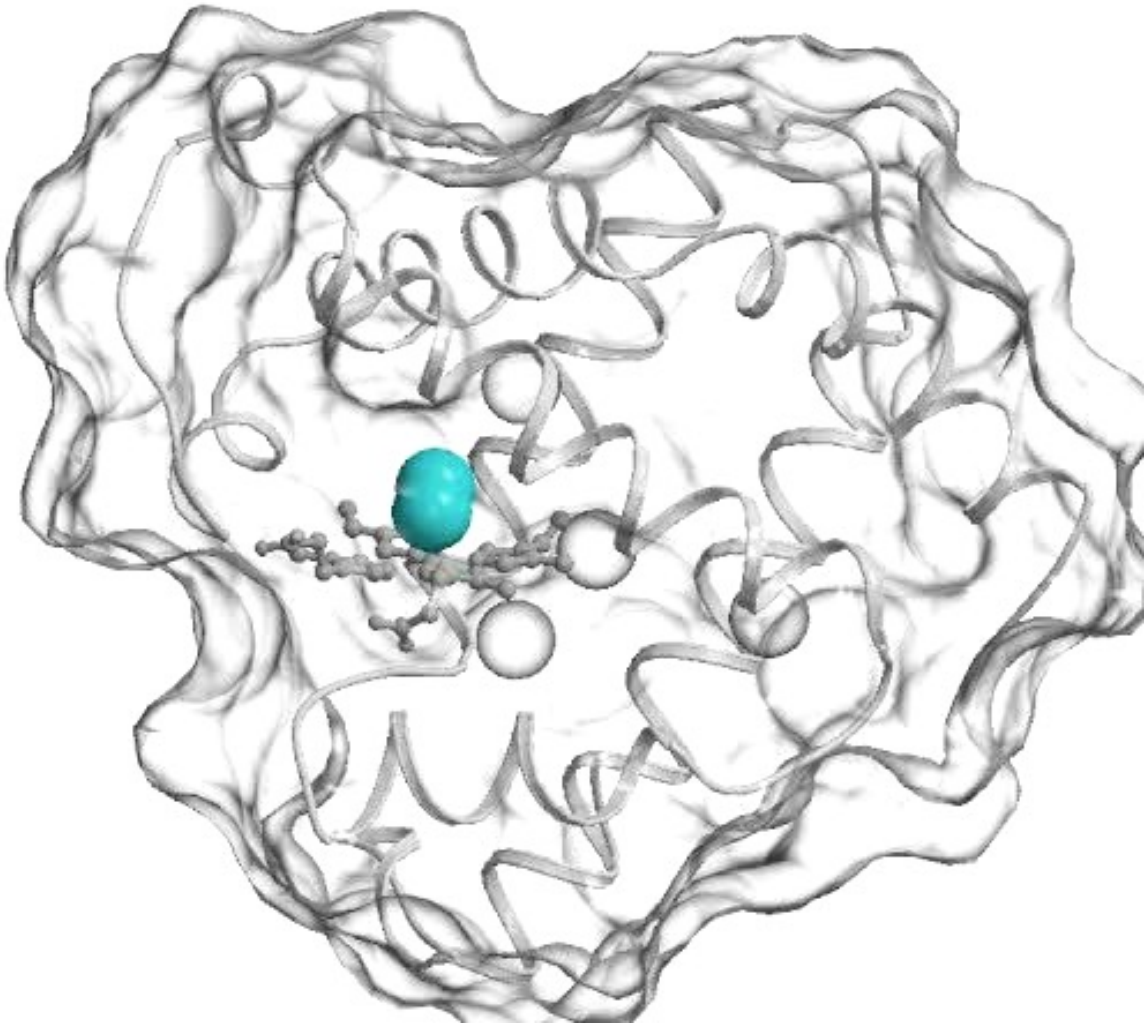
All-atom
- all or most atoms present
- explicit solvent
- molecular dynamics

Coarse-grained
- beads comprising a few atoms
- explicit or implicit solvent
- molecular dynamics

Supra-coarse-grained
- interaction sites comprising many atoms, protein parts or proteins
- implicit solvent
- stochastic dynamics

Continuum
- materials as a continuous mass
- implicit solvent
- continuum mechanics





PNAS

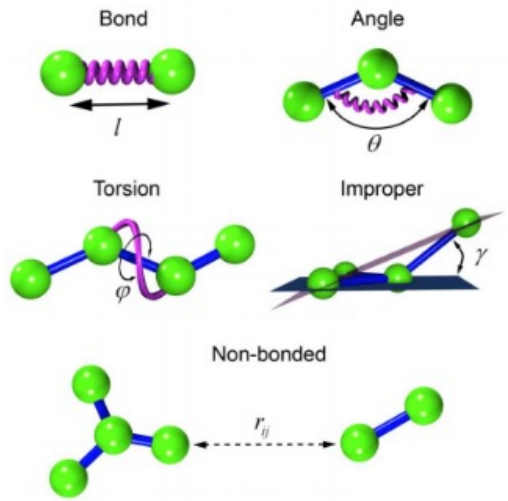
Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin

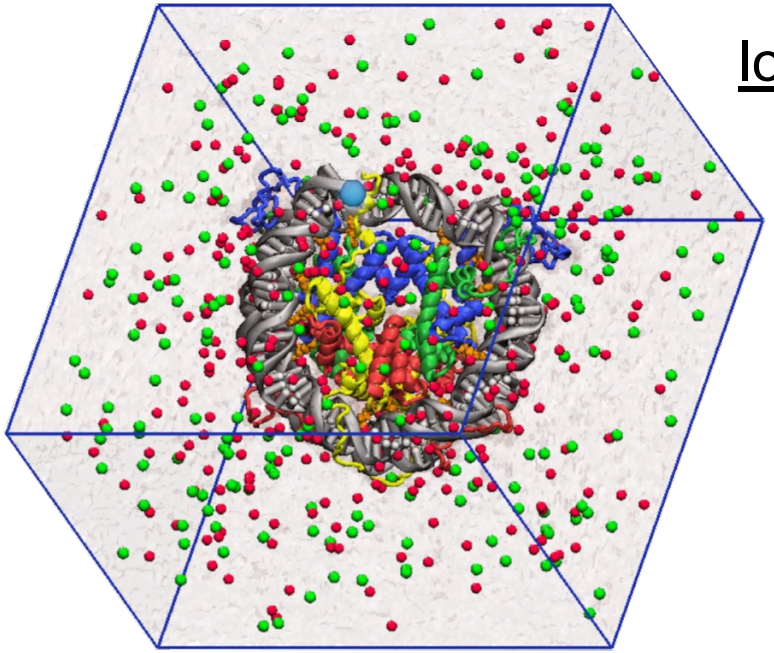
Jory Z. Ruscio*, Deept Kumar†, Maulik Shukla†, Michael G. Prisant‡, T. M. Murali†, and Alexey V. Onufriev†§¶

*Genetics, Bioinformatics, and Computational Biology Program and Departments of †Computer Science and §Physics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061; and ‡Department of Biochemistry, Duke University Medical Center, Durham, NC 27710

Edited by Robert L. Baldwin, Stanford University Medical Center, Stanford, CA, and approved March 25, 2008 (received for review November 15, 2007)

All-atom MD simulations of nucleosomes





Ions: Na^+ Cl^-

- Force field:

AMBER14SB + parmbsc1 + CUFIX + TIP3P

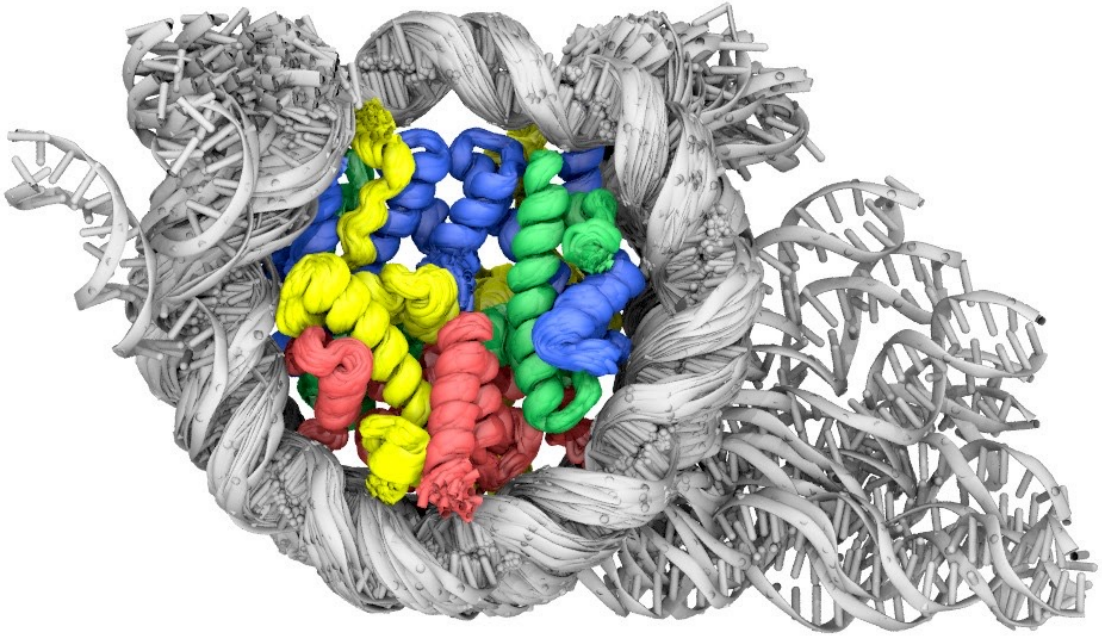
- 150 mM NaCl
- GROMACS2019 (GPU)

Lomonosov-2 supercomputer

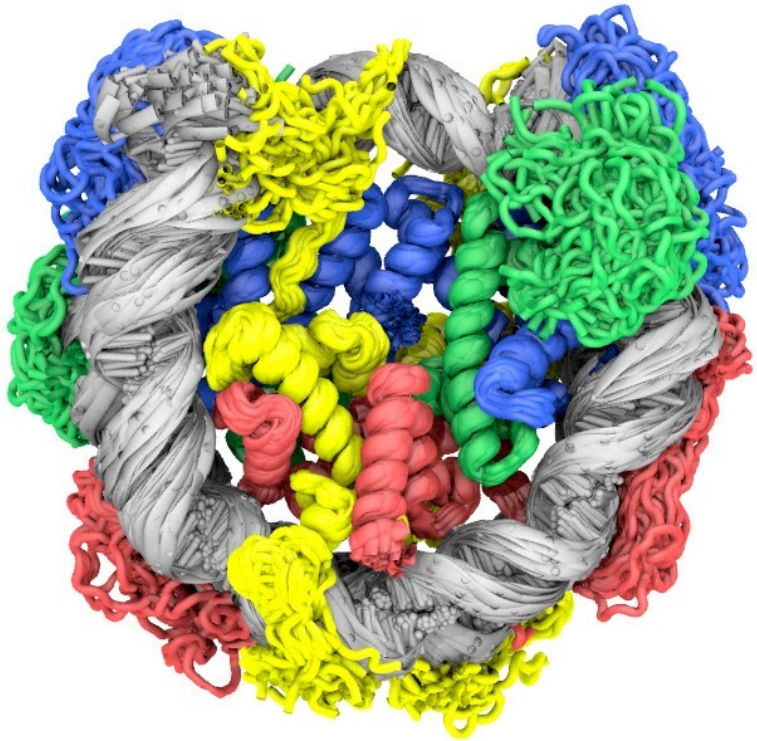
- Simulations up to 15 microseconds long.
- Several starting structures with different DNA sequences.
- With and without histone tails.

Armeev, G. A.; Kniazeva, A. S.; Komarova, G. A.; Kirpichnikov, M. P.; Shaytan, A. K. *Nature Commun.* **2021**,

Dynamics overview at 15 microsecond scale



Nucleosome core particles
without histone tails

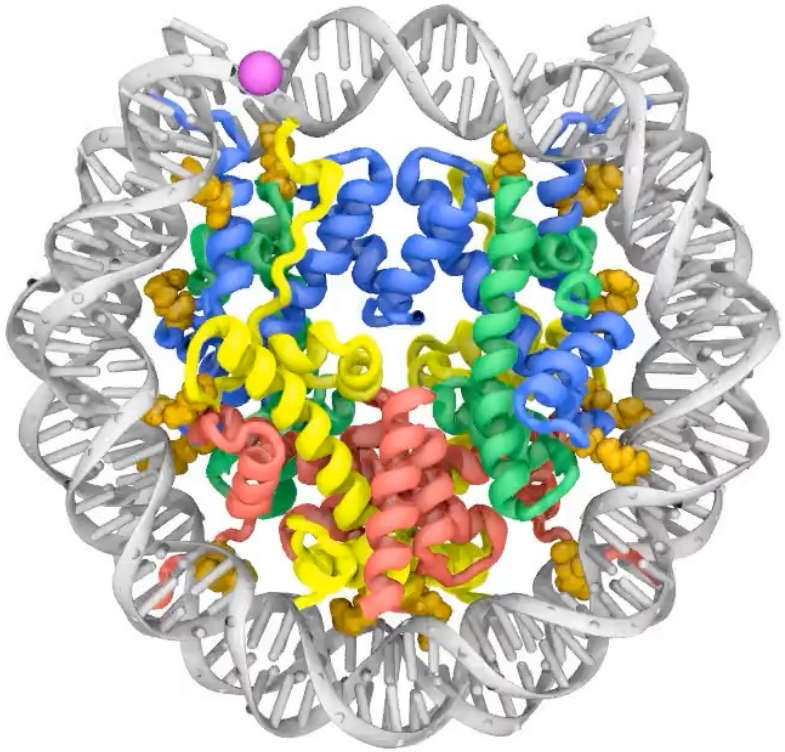


Nucleosome core particles with
full length histone tails

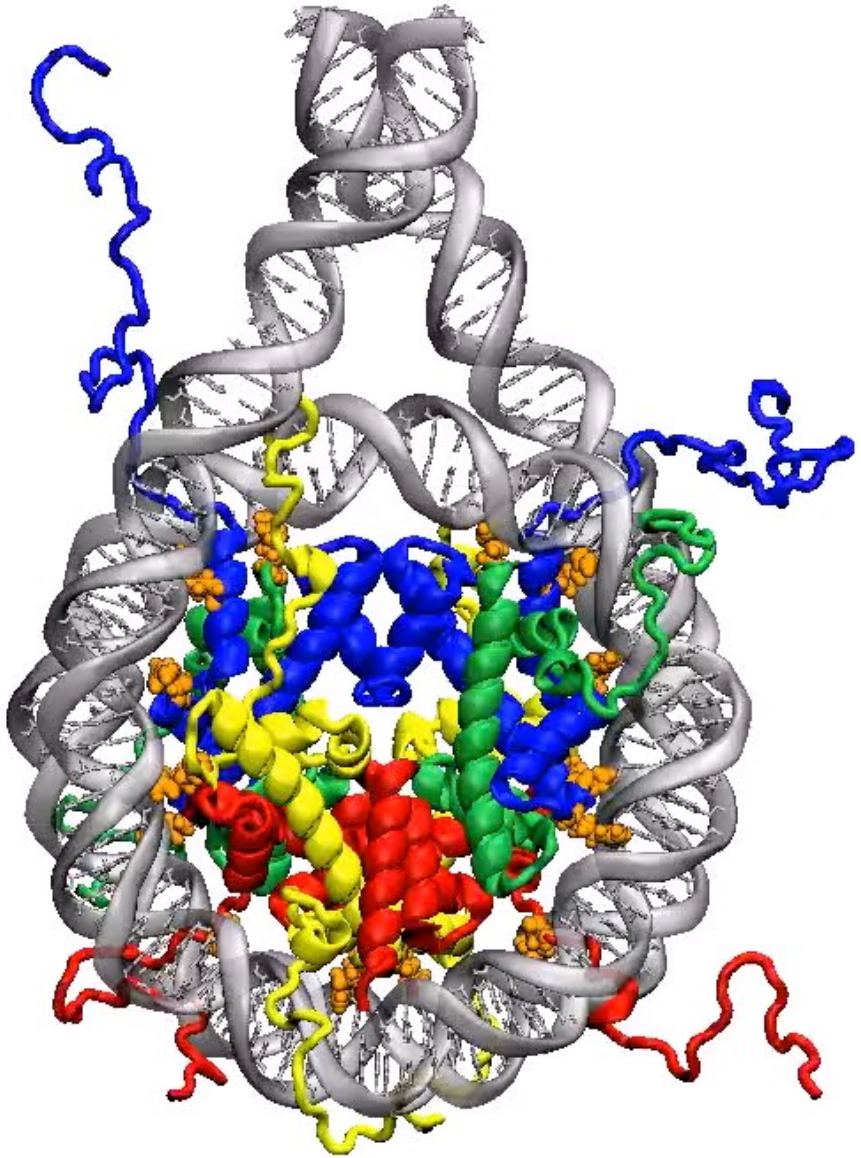
https://intbio.org/Armeev_et_al_2021/ - interactive trajecotry viewer

NCP^{tt}_{145}

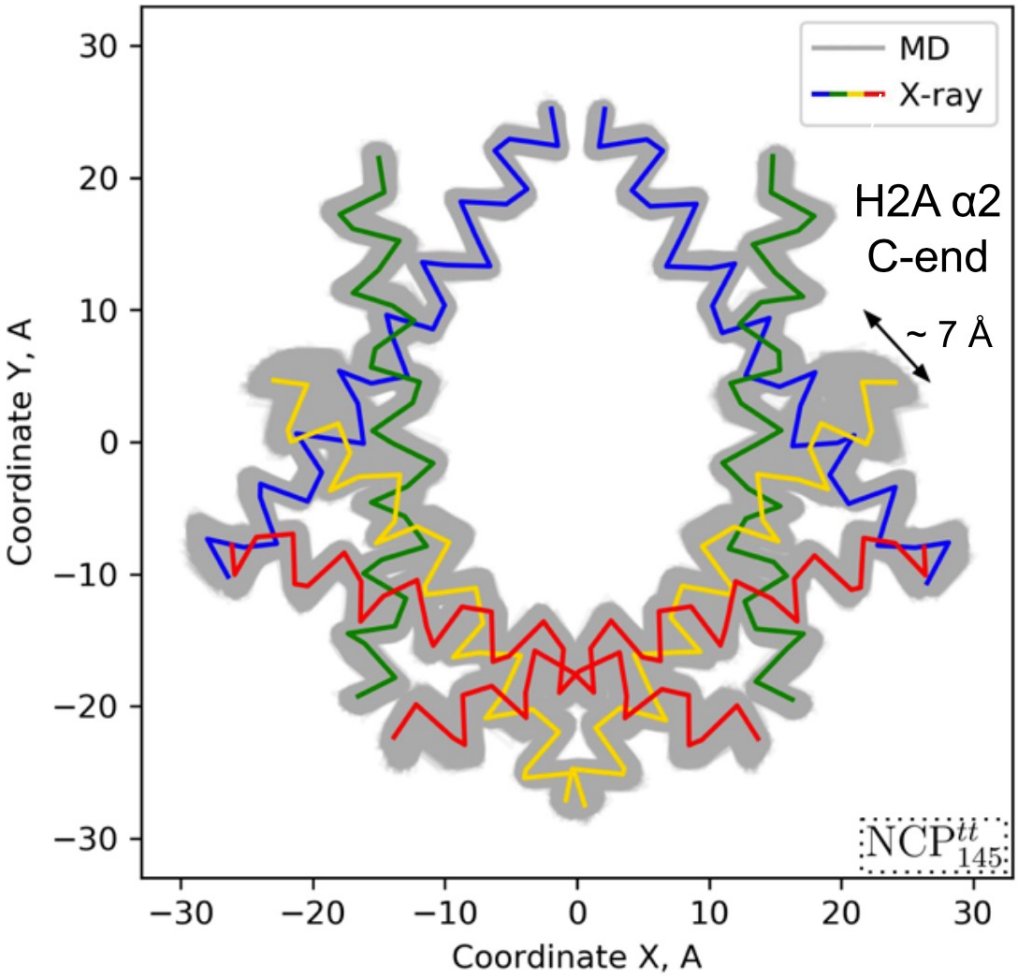
Histones H3
Histones H4
Histones H2A
Histones H2B
Min groove ARG
DNA
O5'DNA chain I

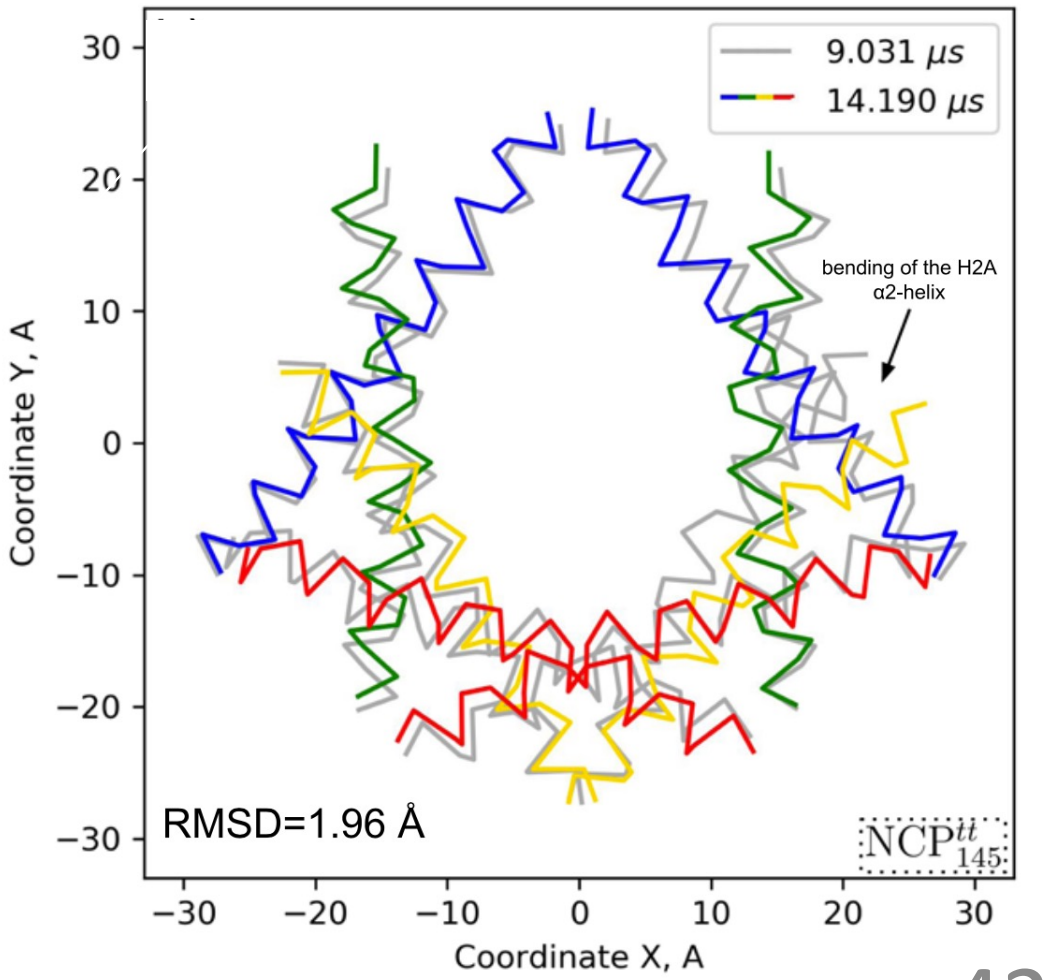


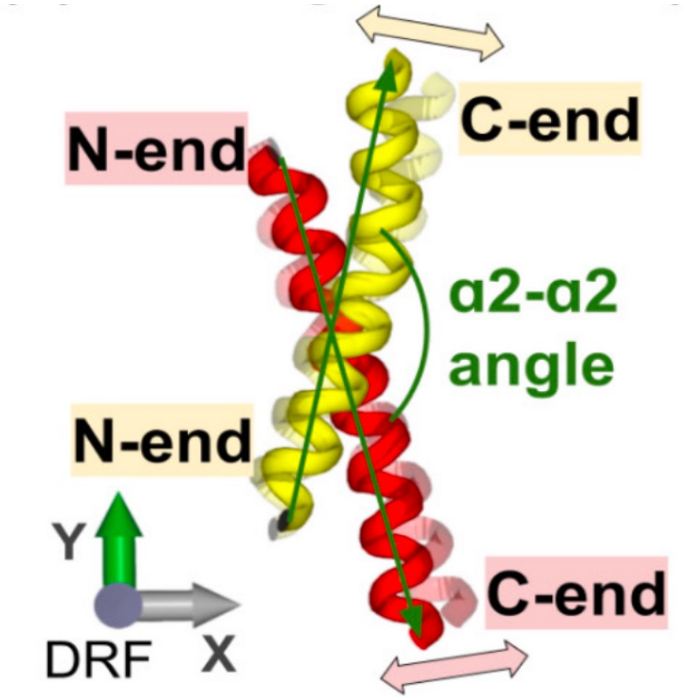
Time: 0.0 ns



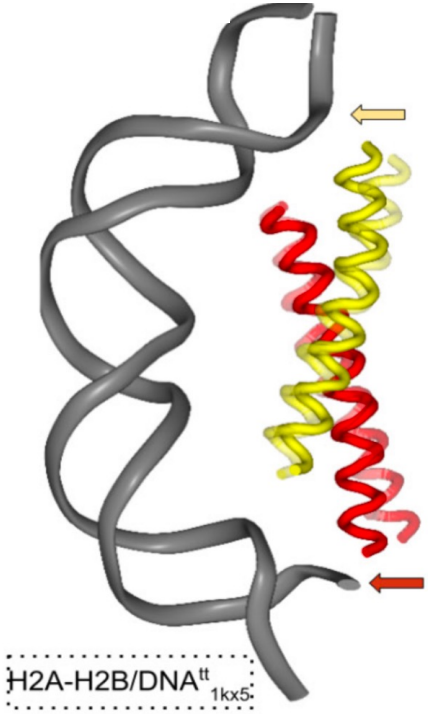
Plasticity of the globular histone core



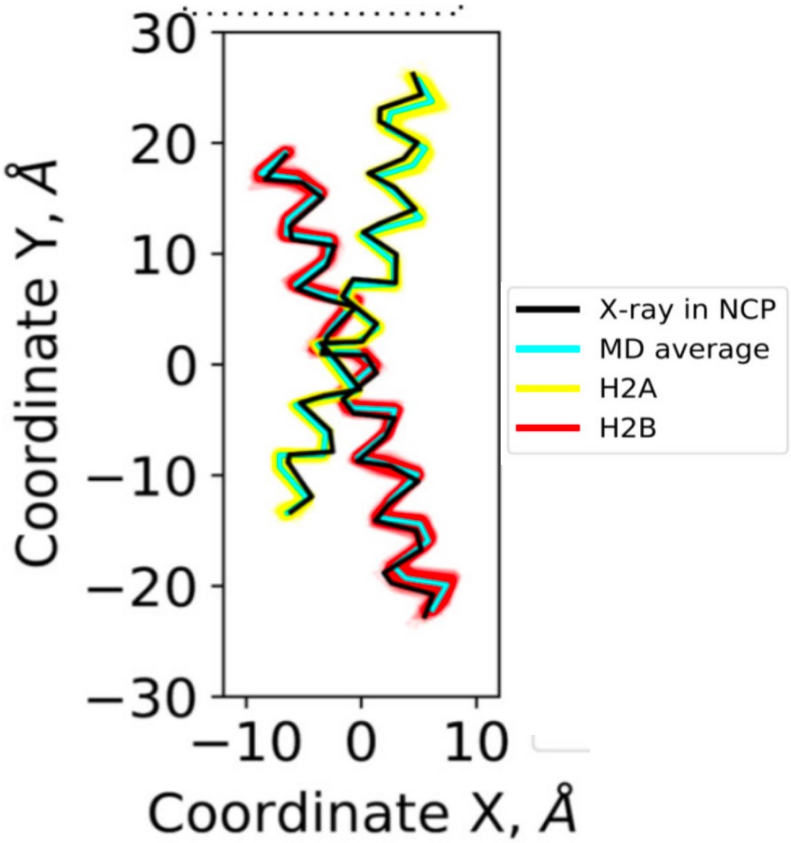




First eigenvector from PCA analysis of H2A-H2B dimer dynamics



Dimer bending upon DNA binding

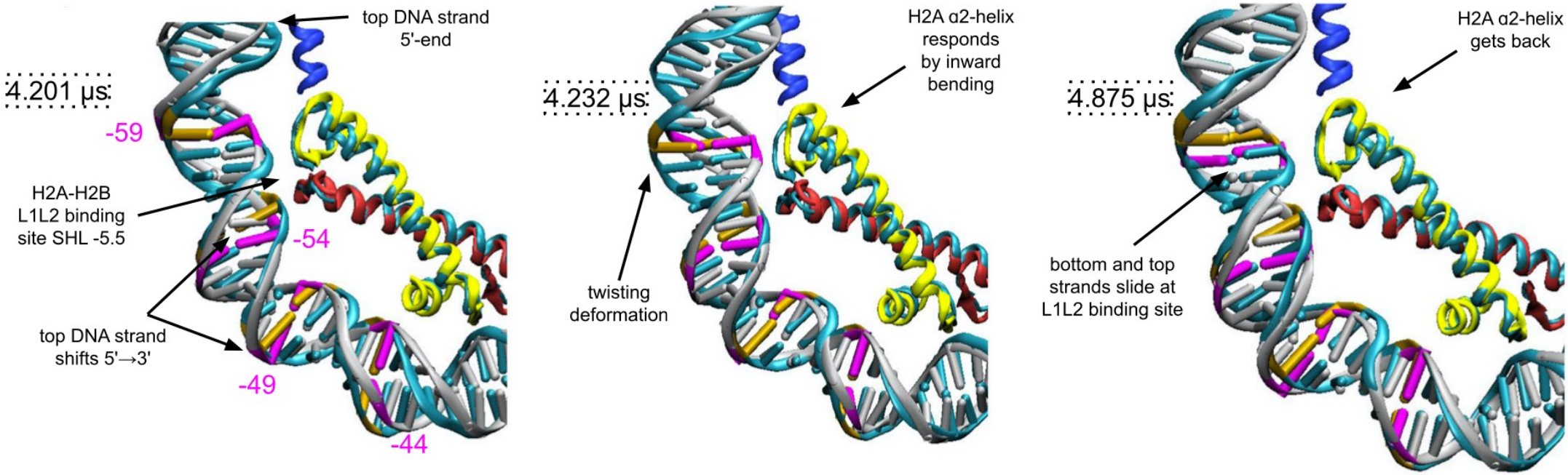


Free dimer dynamics

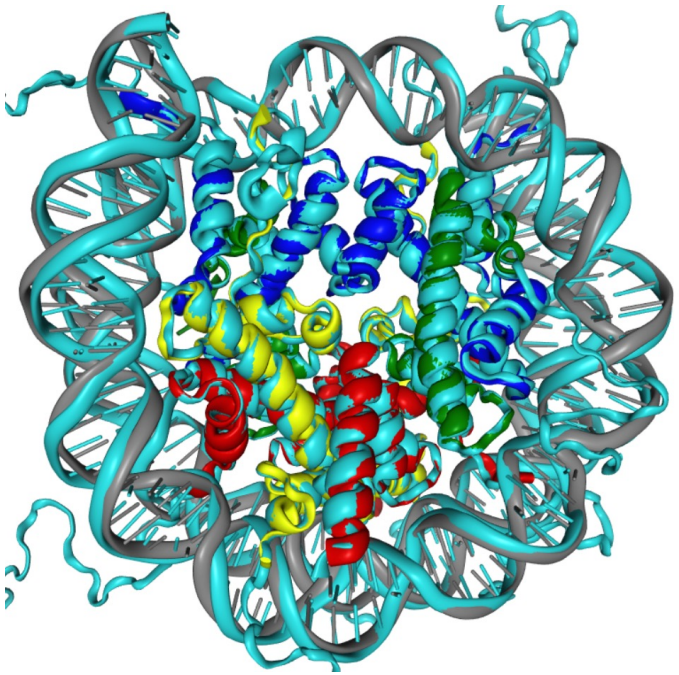
Kniazeva,A.S., Armeev,G.A. and Shaytan,A.K. (2022) H2A-H2B Histone Dimer Plasticity and Its Functional Implications. *Cells*, **11**, 2837.

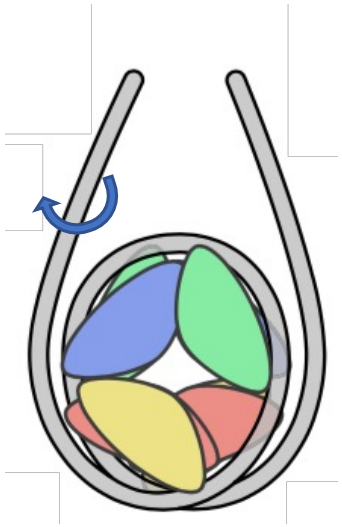
Relaxation of DNA twist-defects in nucleosome

Twist-defect distortion through top DNA strand sliding ⇒ Overtwisting at SHL -5.5 ⇒ DNA slides past SHL -5.5

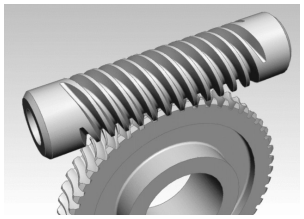


Перемещение ДНК вдоль нуклесомы

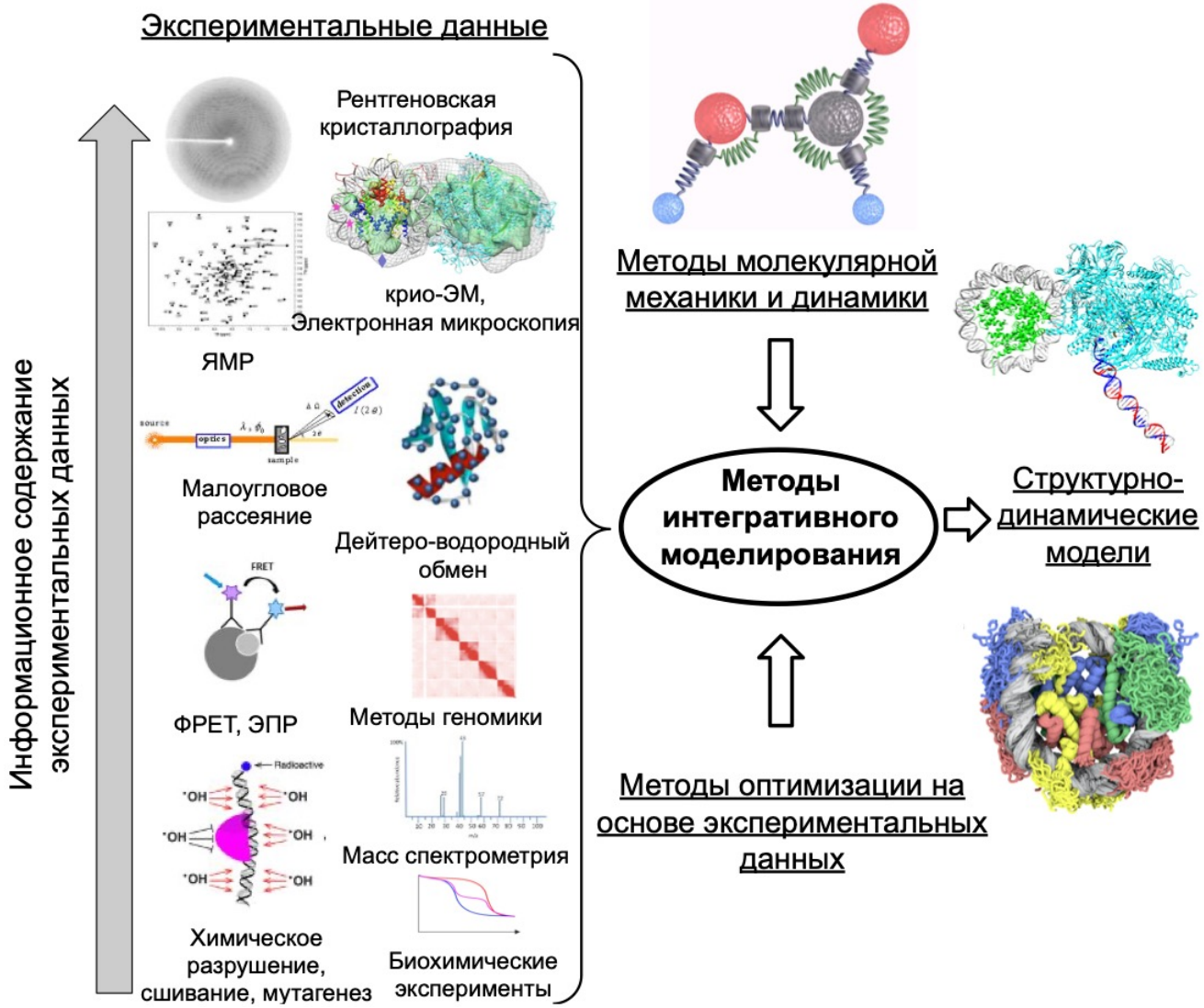




Twist-defect propagation

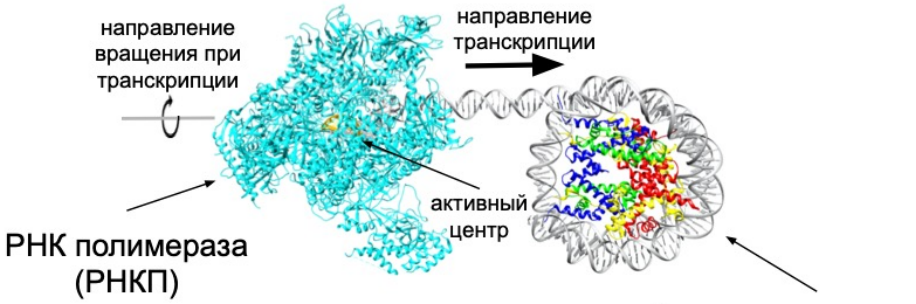


Интегративные подходы



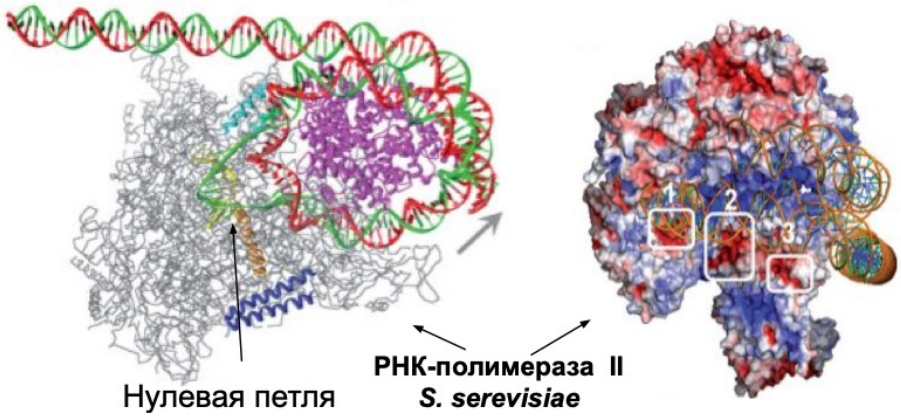
Комплексы нуклеосом и РНК-полимераз

Принципиальная схема движения РНКП по ДНК к нуклеосоме



Модель комплекса в положении +42 после входа РНКП в нуклеосому

Модель комплекса в положении +49 (нулевая петля)



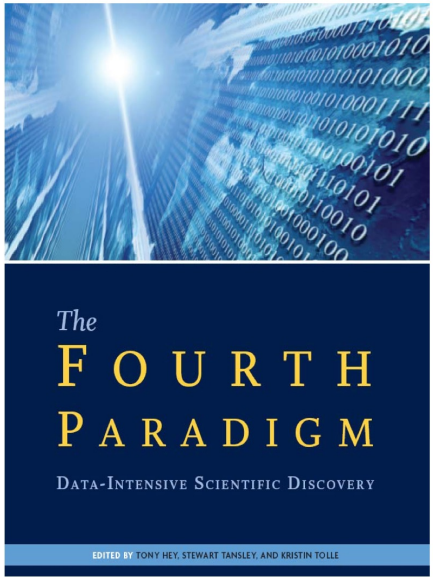
Нейросетевые подходы для изучения структуры биомолекул

Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks
   - ➢ eScience is the set of tools and technologies to support data federation and collaboration
     - For analysis and data mining
     - For data visualization and exploration
     - For scholarly communication and dissemination

Astronomy has been one of the first disciplines to embrace data-intensive science with the Virtual Observatory (VO), enabling highly efficient access to data and analysis tools at a centralized site. The image shows the Pleiades star cluster form the Digitized Sky Survey combined with an image of the moon, synthesized within the WorldWide Telescope service.

Science must move from data to information to knowledge



The Fourth Paradigm:
Data-Intensive Scientific Discovery

Tony Hey
Corporate Vice President
Microsoft External Research

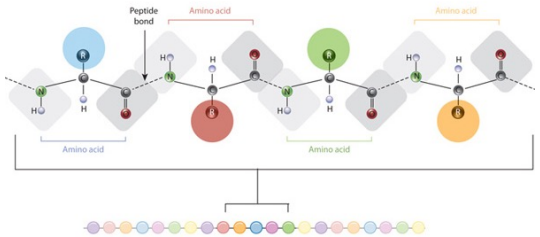
2009

This work is licensed under a Creative Commons Attribution 3.0 United States License.

With thanks to Jim Gray

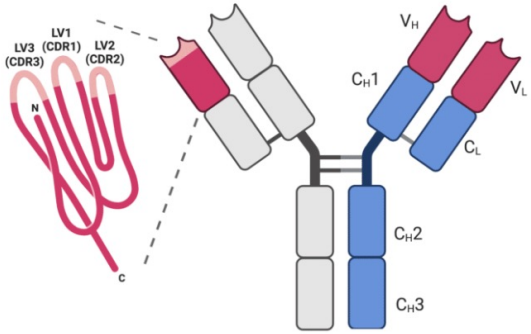
https://www.infodata-edepot.de/volltext/fhpotsdam/10445.pdf

«Пространство возможного»: белковая вселенная

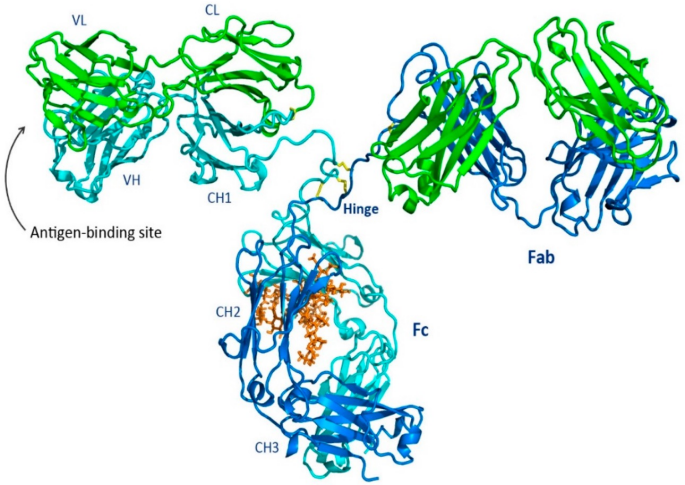


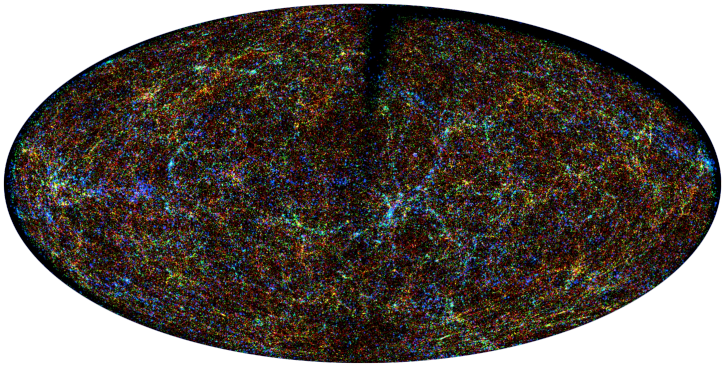
20 типов аминокислот

Для цепи длиной 100, **$20^{100} \cong 10^{130}$** (!) вариантов цепи



Антитела



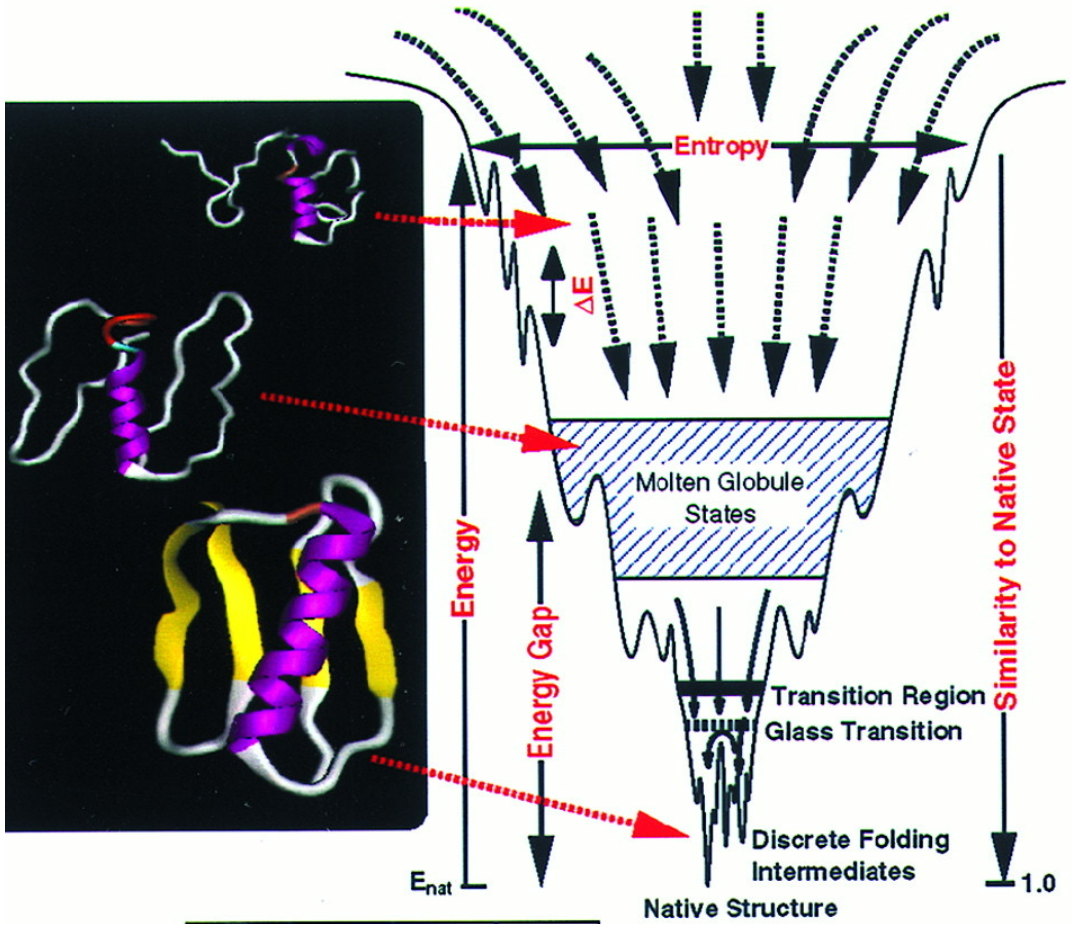


Вселенная
$4.4*10^{23}$ км
~10^{80} атомов

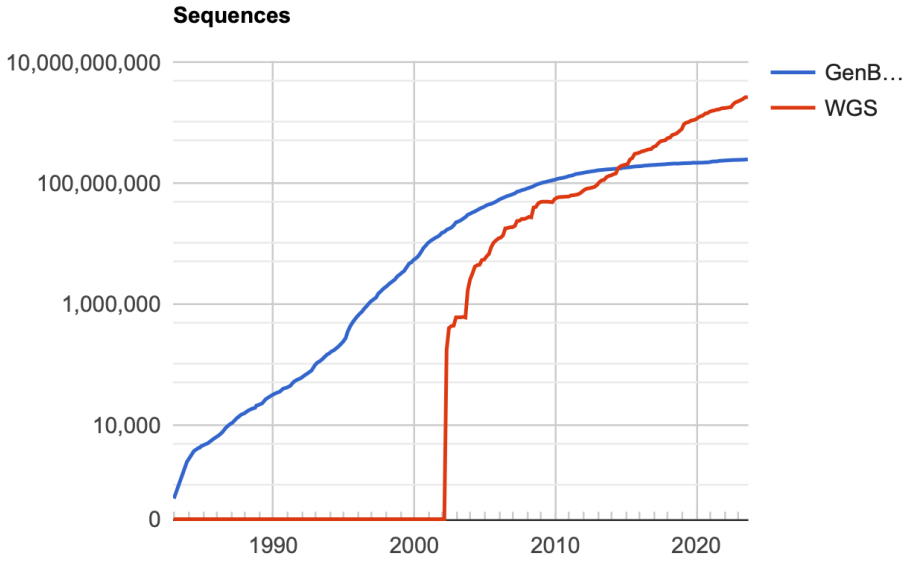
$>>10^{30}>>$ раз больше

Земля
$1.3*10^{4}$ км
~10^{50} атомов

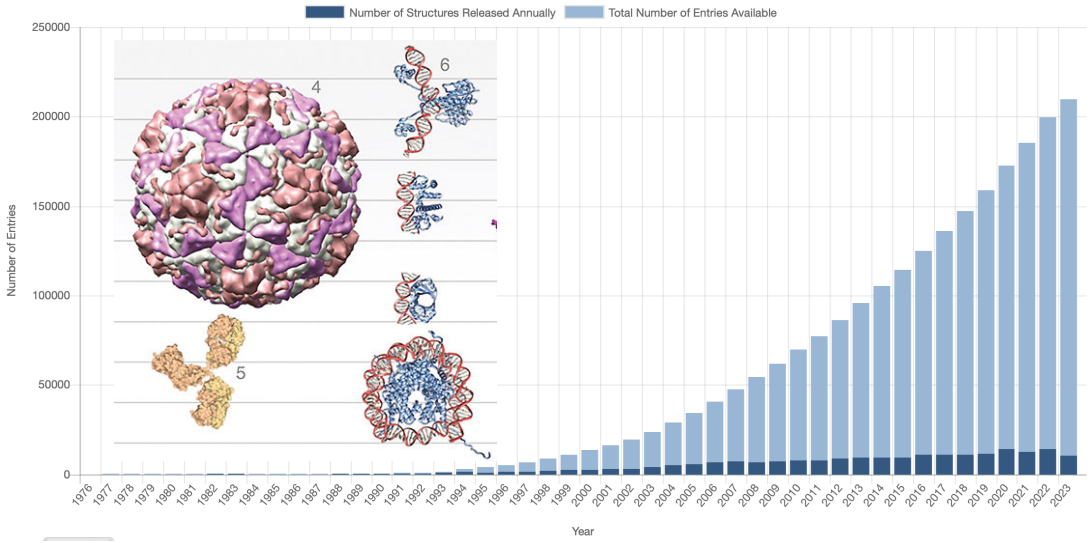
Воронка фолдинга



Информация в базах данных и структуре и последовательности биомолекул

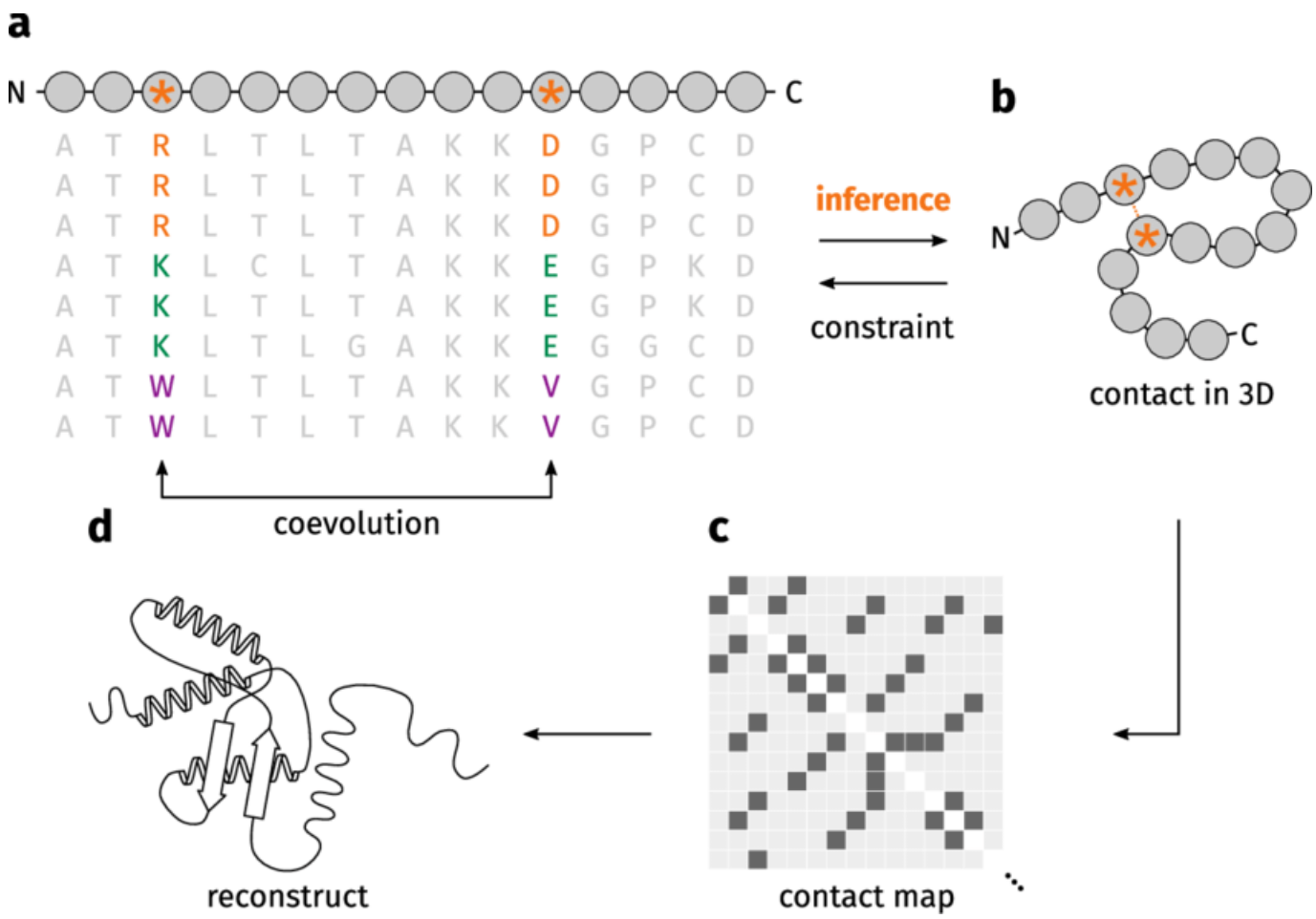


Количество белковых последовательностей в базе GenBank



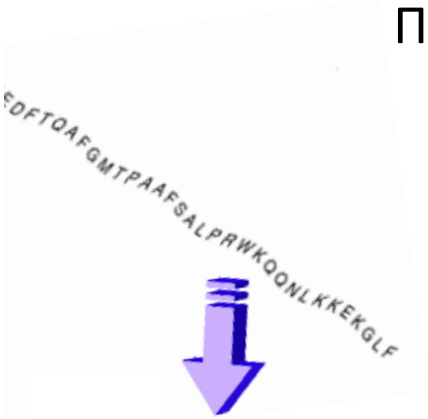
Структур в базе PDB

Коэволюция

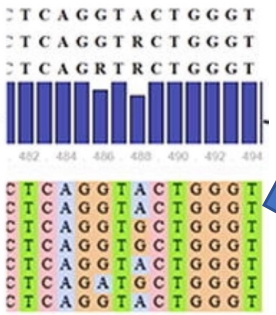

a
N
C
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
coevolution
inference
constraint
b
N
C
contact in 3D
c
contact map
d
reconstruct

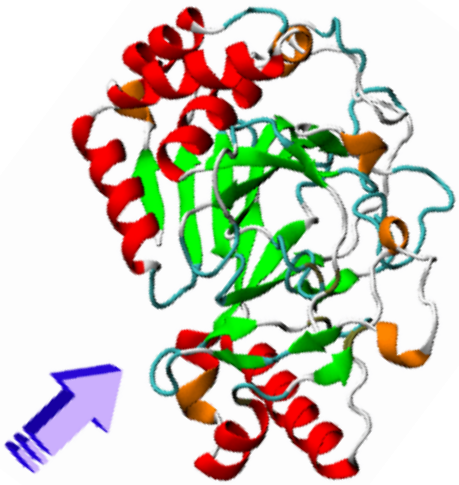
Ключевые технологи: Искусственный интеллект и большие данные

Пример: предсказание структуры белка по последовательности

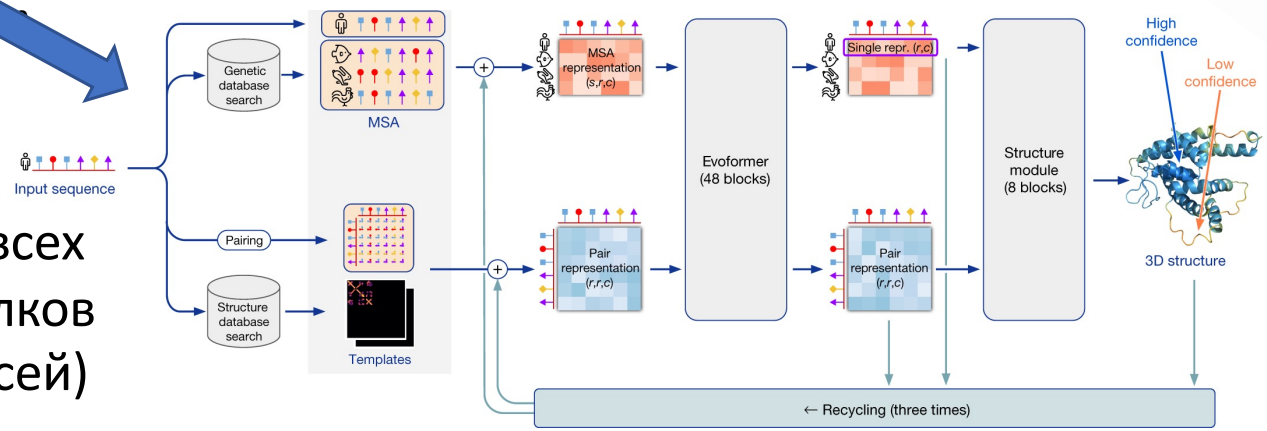


Поиск гомологов



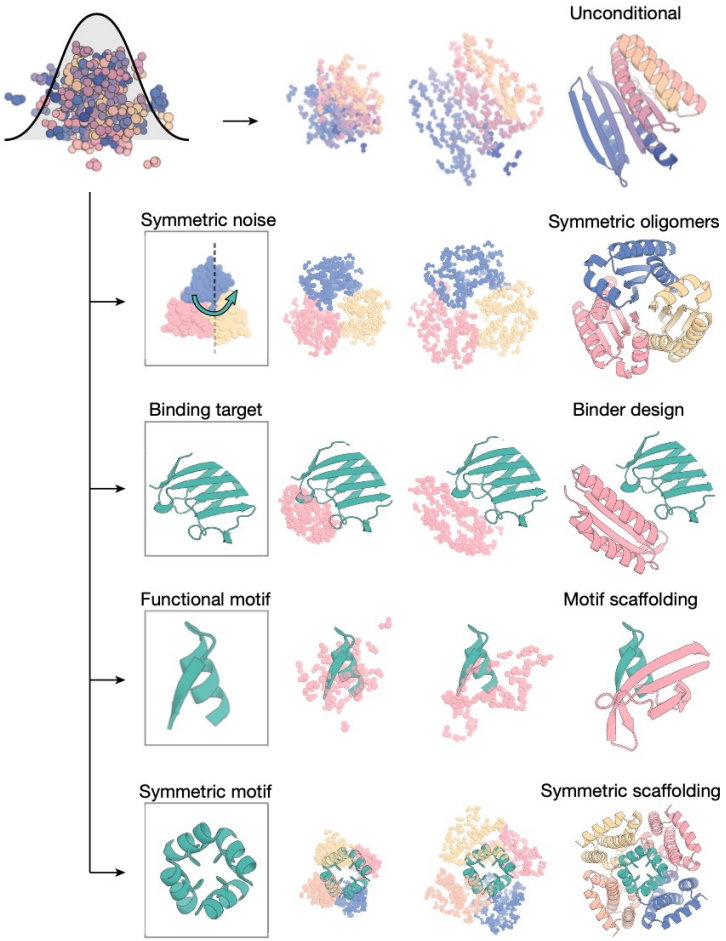


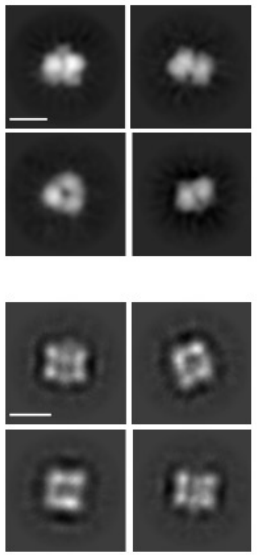
Нейросеть 21 млн. параметров

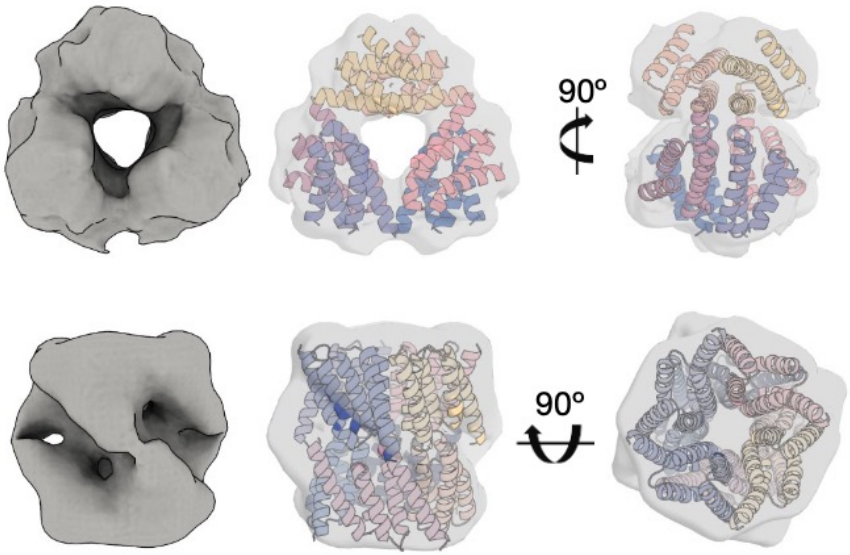


База данных всех известных белков (2.5 млн записей)

Генеративные нейросети для дизайна белков



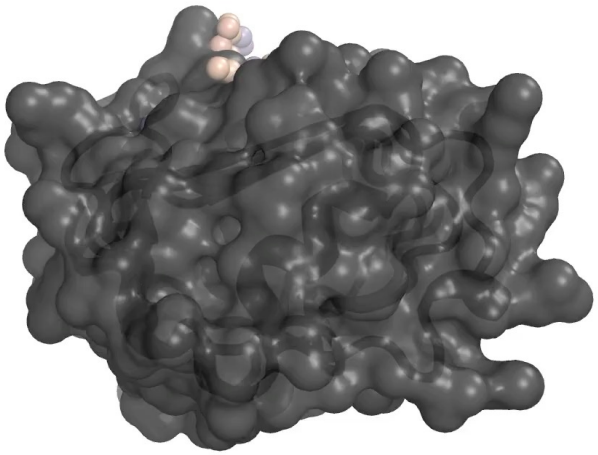


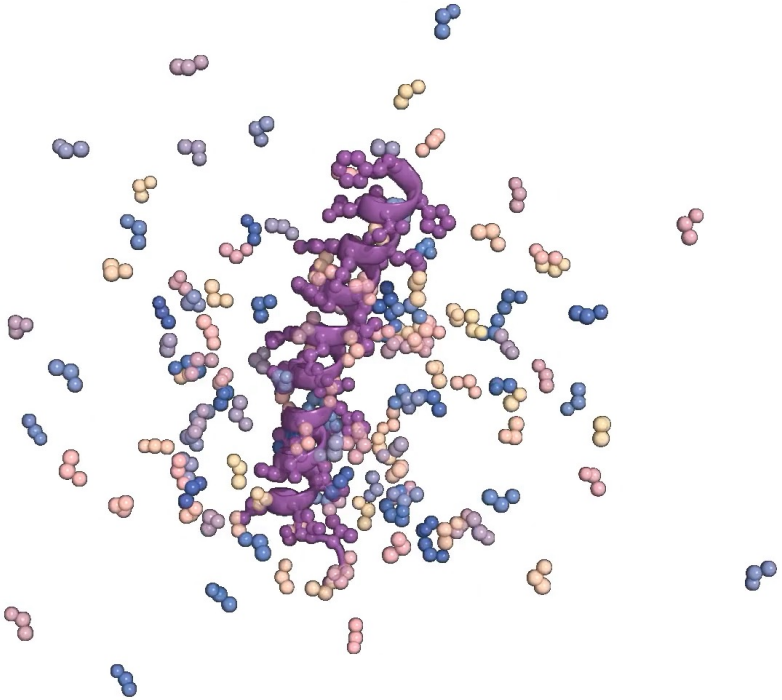


De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8
Received: 14 December 2022
Accepted: 7 July 2023
Published online: 11 July 2023
Open access

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14] ✉





https://www.bakerlab.org/2023/07/11/diffusion-model-for-protein-design/

Спасибо за внимание!