Линейный алгоритм в нелинейной биологии

Илья Белалов
ФИЦ Биотехнологии РАН

Семинар, ФКН ВШЭ
19.02.2026

Линейный алгоритм в нелинейной биологии

- Генотип определяет фенотип
- Преобразование Фурье псевдобулевых функций
- Арифметика в мире геномов
- Столбцы в выравнивании скрывают эпистаз
- Вычисление эпистаза перебором (+ смешная шутка)
- Мировые исследования эпистаза у SARS-CoV-2: пределы возможного
- Эпистаз высоких порядков у SARS-CoV-2
- Гиперграф и проекция в пространство графов
- Текущее состояние исследований
- Куда дальше?
- Альфафолд и попарные взаимодействия аминокислот
- Альфагеном и попарные связи между токенами
- Спасибо за внимание!

Исследования эпистаза в 2026

› Science. 2026 Feb 12;391(6786):eadx6931. doi: 10.1126/science.adx6931. Epub 2026 Feb 12.

Structural ontogeny of protein-protein interactions

Aerin Yang [# 1], Hanlun Jiang [# 2], Kevin M Jude [# 1 3], Deniz Akpinaroglu [4 5], Stephan Allenspach [# 2], Alex Jie Li [4 5], James Bowden [2], Carla Patricia Perez [6], Liu Liu [1], Po-Ssu Huang [7], Tanja Kortemme [4 5 8], Jennifer Listgarten [2 4 9], K Christopher Garcia [1 3]

Affiliations + expand

PMID: 41678610 PMCID: PMC12904254 DOI: 10.1126/science.adx6931

Abstract

Understanding how protein binding sites evolve interactions with other proteins could hold clues to targeting "undruggable" surfaces. We used synthetic coevolution to engineer new interactions between naïve surfaces, simulating the de novo formation of protein complexes. We isolated seven distinct structural families of protein Z-domain complexes and found that synthetic complexes explore multiple shallow energy wells through ratchet-like docking modes, whereas complexes formed by natural binding sites converged in a deep energy well with a relatively fixed geometry. Epistasis analysis of a machine learning-estimated fitness landscape revealed "seed" contacts between binding partners that anchored the earliest stages of encounter complex formation. Our results suggest that "silent" surfaces have a shallower energy landscape than natural binding sites, disfavoring tight binding, likely owing to evolutionary counterselection.

› Proc Natl Acad Sci U S A. 2026 Jan 20;123(3):e2516291123. doi: 10.1073/pnas.2516291123. Epub 2026 Jan 16.

Predicting epistasis across proteins by structural logic

Michelle Tang [1], Gareth A Cromie [1], Anowarul Kabir [2], Martin S Timour [1], Julee Ashmead [1], Russell S Lo [1], Nathaniel Corley [3], Frank DiMaio [3 4], Hiroki Morizono [5 6], Ljubica Caldovic [5 6], Nicholas Ah Mew [5 6], Andrea Gropman [7 8], Amarda Shehu [2], Aimée M Dudley [1]

Affiliations + expand

PMID: 41543897 PMCID: PMC12818424 (available on 2026-07-16)
DOI: 10.1073/pnas.2516291123

Abstract

Accurately predicting the phenotypic consequences of genetic variation is a major challenge for precision medicine. The problem is exacerbated by epistatic interactions, nonadditive effects between genetic variants that produce unexpected phenotypes. Here, we explore an understudie form of positive epistasis: intragenic complementation, in which pairs of loss-of-function variant restore near wild-type protein function. Using mutational scanning in yeast, we identify thousand such interactions in a clinically important enzyme, human argininosuccinate lyase (ASL). Restora of protein function is not due to the biochemical properties of the substituted amino acids, but ra to a structural feature of the protein, the active site assembly. We develop a machine learning algorithm that uses protein language model embeddings to predict intragenic complementation ASL with 99.6% accuracy. Additionally, the model trained on ASL generalizes to a structurally rela but sequence-divergent enzyme, fumarase, with accuracy over 90%. Our findings reveal a structu basis for this form of epistasis and provide a predictive framework that could extend to at least 4 human proteins.

Keywords: epistasis; machine learning; variant effects.

› Cell. 2026 Feb 5:S0092-8674(25)01490-4. doi: 10.1016/j.cell.2025.12.041. Online ahead of print.

Symbiotic entrenchment through ecological Catch-22

Thomas H Naragon [1], Joani W Viliunas [2], Mina Yousefelahiyeh [2], Adrian Brückner [2], Julian M Wagner [2], K Esther Okamoto [2], Hannah M Ryon [2], Danny Collinson [2], Sheila A Kitchen [2], Reto S Wijker [3], Alex L Sessions [3], Joseph Parker [4]

Affiliations + expand

PMID: 41650968 DOI: 10.1016/j.cell.2025.12.041

Abstract

Why symbiotic organisms evolve irreversible dependencies on hosts is an outstanding question. We report a biological stealth device in a beetle that permits infiltration of ant societies. Via transcriptional silencing, the beetle switches off biosynthesis of cuticular hydrocarbons (CHCs)-body surface pheromones that function pleiotropically as a waxy desiccation barrier. Silencing transforms the beetle into a chemical blank slate onto which ant CHCs are transferred via grooming behavior, leading to perfect chemical mimicry and acceptance into the colony. Silencing is irreversible, however, forcing the beetle into a chronic dependence on ants to both maintain mimicry and prevent desiccation. We show that evolutionary reversion of the silencing mechanism would render the beetle detectable to ants; conversely, reversion of the beetle's attraction to ants would render it desiccation prone. Symbiotic entrenchment can thus arise from epistasis between symbiotic traits, locking lineages into a Catch-22 that obstructs reversion to living freely.

Keywords: ants; behavior; biosynthesis; cell biology; chemical ecology; evolution; irreversibility; obligate symbiosis; rove beetles.

› Nat Commun. 2026 Feb 5. doi: 10.1038/s41467-026-69152-2. Online ahead of print.

Protein-protein interactions are a major source of epistasis in genetic interaction networks

Xavier Castellanos-Girouard [1 2 3], Adrian W R Serohijos [4 5], Stephen W Michnick [6 7]

Affiliations + expand

PMID: 41644525 DOI: 10.1038/s41467-026-69152-2

› Proc Natl Acad Sci U S A. 2026 Feb 3;123(5):e2505183123. doi: 10.1073/pnas.2505183123. Epub 2026 Jan 30.

Evolutionary pathways in epistatic mechanical networks

Samar Alqatari [1], Sidney R Nagel [1]

Affiliations + expand

PMID: 41615748 PMCID: PMC12867688 (available on 2026-07-30)
DOI: 10.1073/pnas.2505183123

Генотип определяет фенотип

Фенотип - функция *f(x)*, генотип - аргумент *x*.

Без ограничения общности считаем геном последовательностью символов двухбуквенного алфавита:

- Все локусы имеют по два аллеля {A_1,a_1,A_2,a_2, … }
- Кодируем буквы {A,C,G,T,-} в цифры {0,1}
- Изучаем инопланетные геномы {Одинин, Нулидин}

$$x \in \{0,1\}^n$$

Генотип определяет фенотип

Фенотип - функция *f(x)*, генотип - аргумент *x*.

Без ограничения общности считаем фенотип булевой функцией:

- Многомерный фенотип - набор вещественных фенотипов

  $g(x) \in \mathbb{R}^d$

  $g(x) = (h_1(x), h_2(x), \ldots, h_d(x))$

  $h_i(x) \in \mathbb{R}$

- Вещественный фенотип - линейная комбинация булевых фенотипов

  $h(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \ldots + \alpha_k f_k(x)$

  $\alpha_i \in \mathbb{R}$

  $f_i \in \{0, 1\}$

Преобразование Фурье псевдобулевых функций

Рассмотрим булеву функцию фенотипа *f(x)*, она имеет полиномиальное представление:

$$u \in \mathbb{F}_2^n \qquad \hat{u} = \{i \ : \ u_i = 1\} \qquad \tilde{u} = \prod_{i \in \hat{u}} x_i \qquad \mathrm{supp} f = \{x : f(x) \neq 0\}$$

$$f(x) = \sum_{u \in \mathbb{F}_2^n} F(u) \prod_{i \in \hat{u}} x_i = \sum_{u \in \mathbb{F}_2^n} F(u)\tilde{u} = C + \sum_{i=1}^{n} F(\delta_i)x_i + \sum_{i<j} F(\delta_{ij})x_i x_j + \sum_{i<j<k} F(\delta_{ijk})x_i x_j x_k + \ \ldots$$

$$= |\mathrm{supp} f| \sum_{u:|F(u)|=|\mathrm{supp} f|} \tilde{u}\ \mathrm{sign} F(u) \ + \ (|\mathrm{supp} f| - 1) \sum_{u:|F(u)|=(|\mathrm{supp} f|-1)} \tilde{u}\ \mathrm{sign} F(u) \ + \ \ldots$$

Коэффициенты при мономах вычисляются по формуле:

$$F(u) = \sum_{x \in \mathbb{F}_2^n} f(x)(-1)^{<x,u>} = \sum_{x \in \mathrm{supp} f} (-1)^{<x,u>}$$

$$\int_{\mathbb{R}} f(x) e^{-iux} dx$$

Арифметика в мире геномов

Возьмем два генома, x и y

Фенотип $f(x) = f(y) = 1$

x = 101001011…001001

y = 100001101…001011

Найдем отличия x и y:

a = 001000110…000010 - вектор мутаций

$a = x \text{ XOR } y = x + y \pmod 2 = x \oplus y$

$x \oplus a = y$

Возьмем третий геном z, с фенотипом $f(z) = 1$

Рассмотрим “геном” $v := z \oplus a$

$f(v) = 1$?

Арифметика в мире геномов

Возьмем два генома, x и y

Фенотип $f(x) = f(y) = 1$

x = 101001011…001001

y = 100001101…001011

Найдем отличия x и y:

a = 001000110…000010 - вектор мутаций

$a = x \text{ XOR } y = x + y \pmod 2 = x \oplus y$

$x \oplus a = y$

Возьмем третий геном z, с фенотипом $f(z) = 1$

Рассмотрим “геном” $v := z \oplus a$

$f(v) = 1$?

> Виды и роды, генетически близкие, характеризуются сходными рядами наследственной изменчивости с такой правильностью, что, зная ряд форм в пределах одного вида, можно предвидеть **нахождение параллельных форм** у других видов и родов. Чем ближе генетически расположены в общей системе роды и линнеоны, тем полнее сходство в рядах их изменчивости.
>
> Н.И. Вавилов (1922)

Арифметика в мире геномов

$x \oplus y \oplus z = v$ - геном с фенотипом $f(v) = 1$

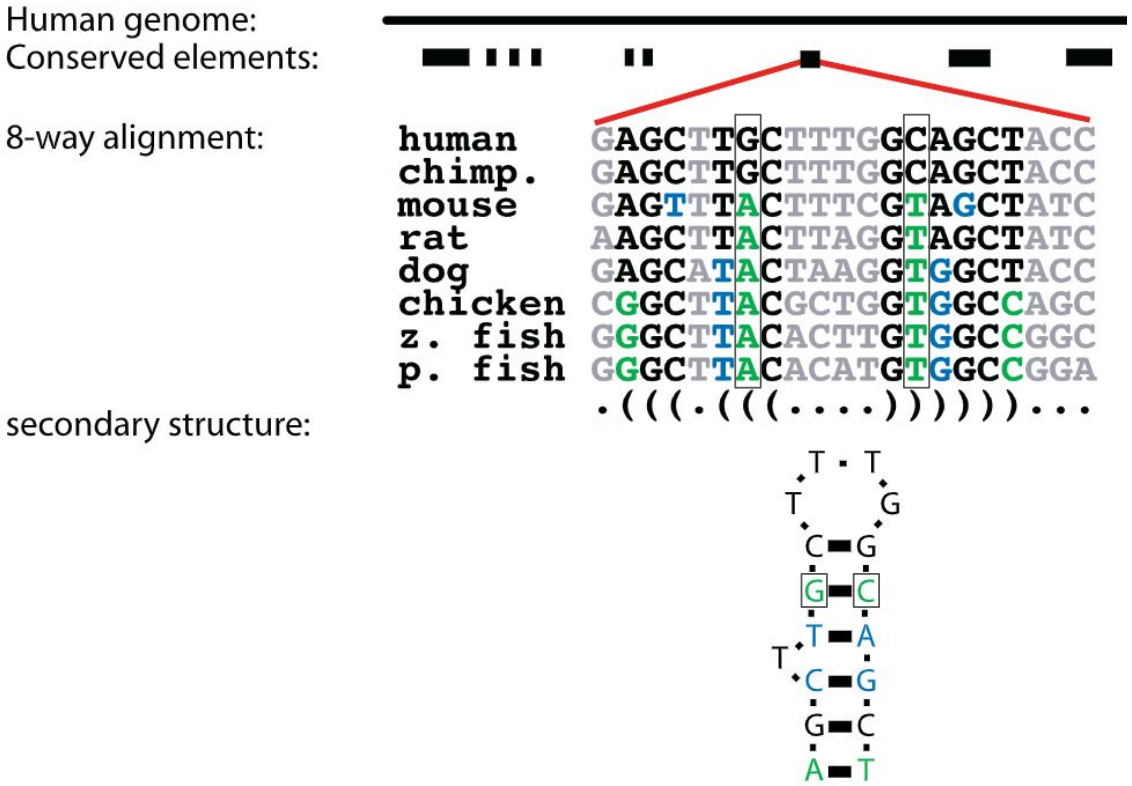
$x_1 \oplus x_2 \oplus \ldots \oplus x_p = x_0$ - геном с фенотипом $f(x_0) = 1$, p - нечетное

Геномы образуют аффинное подпространство в пространстве последовательностей

Это подпространство можно задать уравнениями (образующими соотношениями)

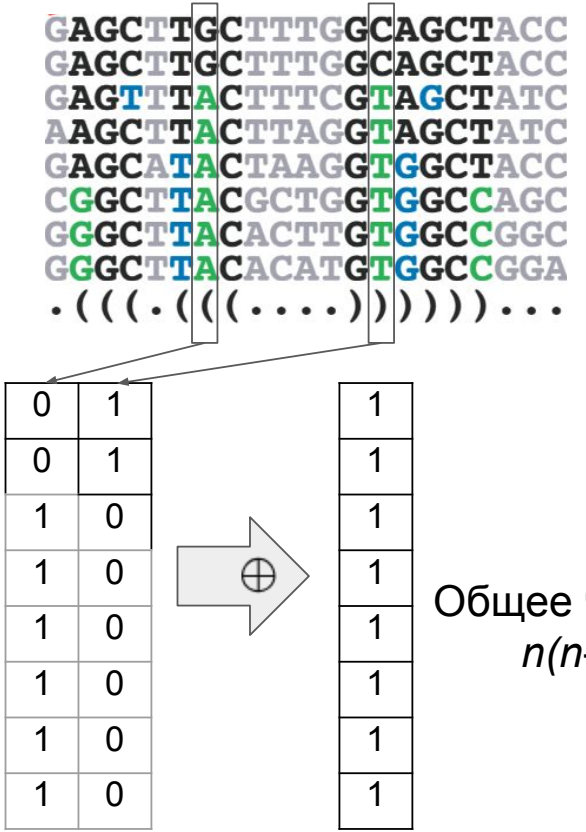
Каждое соотношение соответствует набору столбцов в выравнивании

Столбцы в выравнивании скрывают эпистаз

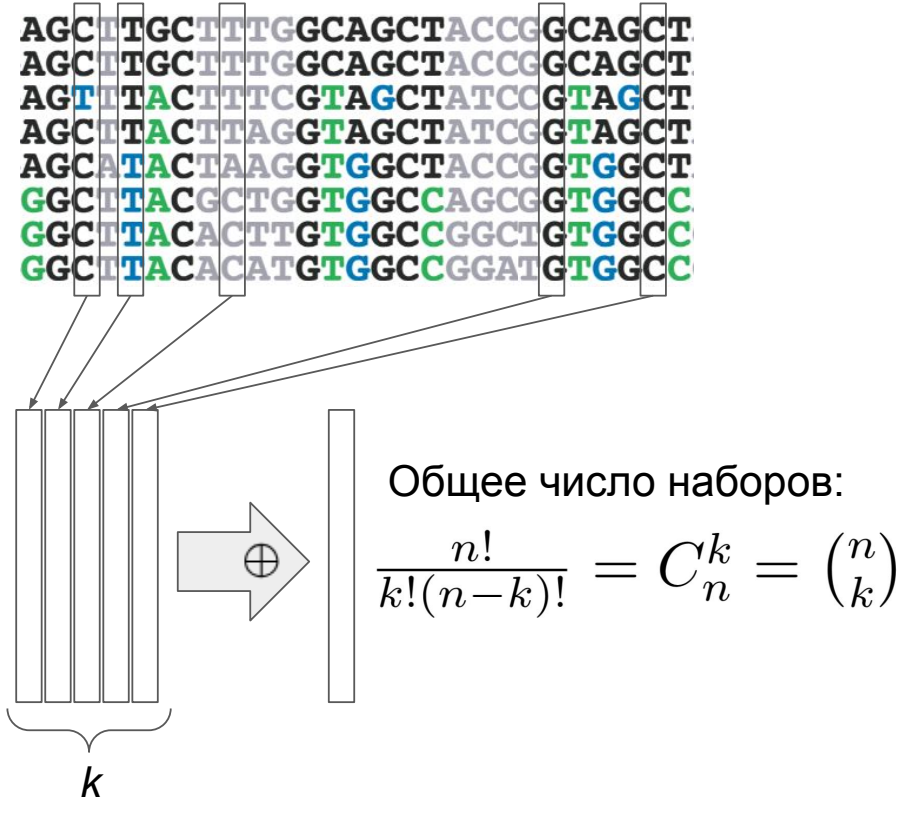


https://doi.org/10.1371/journal.pcbi.0020033

Столбцы в выравнивании скрывают эпистаз



Общее число пар:
n(*n*-1)/2



Общее число наборов:

$$\frac{n!}{k!(n-k)!} = C_n^k = \binom{n}{k}$$

Вычисление эпистаза перебором

Имеется 120 локусов (генов, нуклеотидов, …), *n* = 120. Общее количество комбинаций, $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

$2^{120} > 10^{36}$ операций
10 GHz × 100 000 CPU = 10^{15} операций в секунду
$10^{36} / 10^{15} = 10^{21}$ секунд на все комбинации
$10^{21} / 10^{18} =$ **1000**

Вычисление эпистаза перебором

Имеется 120 локусов (генов, нуклеотидов, …), n = 120. Общее количество комбинаций, $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

$2^{120} > 10^{36}$ операций
10 GHz × 100 000 CPU = 10^{15} операций в секунду
$10^{36} / 10^{15} = 10^{21}$ секунд на все комбинации
$10^{21} / 10^{18} =$ **1000**

Система	Производительность	Время на все комбинации
10 GHz × 100 000 CPU	10^{15} (1 петафлопс)	~42 триллиона лет*
Суперкомпьютер Frontier	1.2×10^{18} (1.2 экзафлопс)	~35 миллиардов лет*
YTsaurus на всех процессорах Земли	~10^{22} (10 зеттафлопс)	~4.2 миллиона лет

*Возраст Вселенной: 1.4×10^{10} лет

❌ Полный перебор непрактичен даже для скромного n = 120
✅ Нужен новый алгоритм, обходящий комбинаторный взрыв

Мировые исследования эпистаза: пределы возможного

Парные взаимодействия

Epistasis at the SARS-CoV-2 Receptor-Binding Domain Interface and the Propitiously Boring Implications for Vaccine Escape
Rochman ND, Faure G, Wolf YI, Freddolino PL, Zhang F, Koonin EV. ***mBio***. 2022;13(2):e0013522.

Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes
Zeng H-L, Dichio V, Rodríguez Horta E, Thorell K, Aurell E. ***Proc Natl Acad Sci USA***. 2020;117(50):31519–31526.

Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes
Rodriguez-Rivas J, Croce G, Muscat M, Weigt M. ***Proc Natl Acad Sci USA***. 2022;119:e2113118119.

Real-time identification of epistatic interactions in SARS-CoV-2 from large genome collections
Innocenti G, Obara M, Costa B, et al. ***Genome Biol***. 2024;25:228.

Temporal epistasis inference from more than 3,500,000 SARS-CoV-2 genomic sequences
Zeng H-L, Liu Y, Dichio V, Aurell E. ***Physical Review E***, 106(4), p.044409

Epistasis lowers the genetic barrier to SARS-CoV-2 neutralizing antibody escape
Witte L, Baharani VA, Schmidt F, et al. ***Nat Commun***. 2023;14:302.

Host heterogeneity and epistasis explain punctuated evolution of SARS-CoV-2
Nielsen BF, Li Y, Sneppen K, et al. ***PLoS Comput Biol***. 2023;19(2):e1010896.

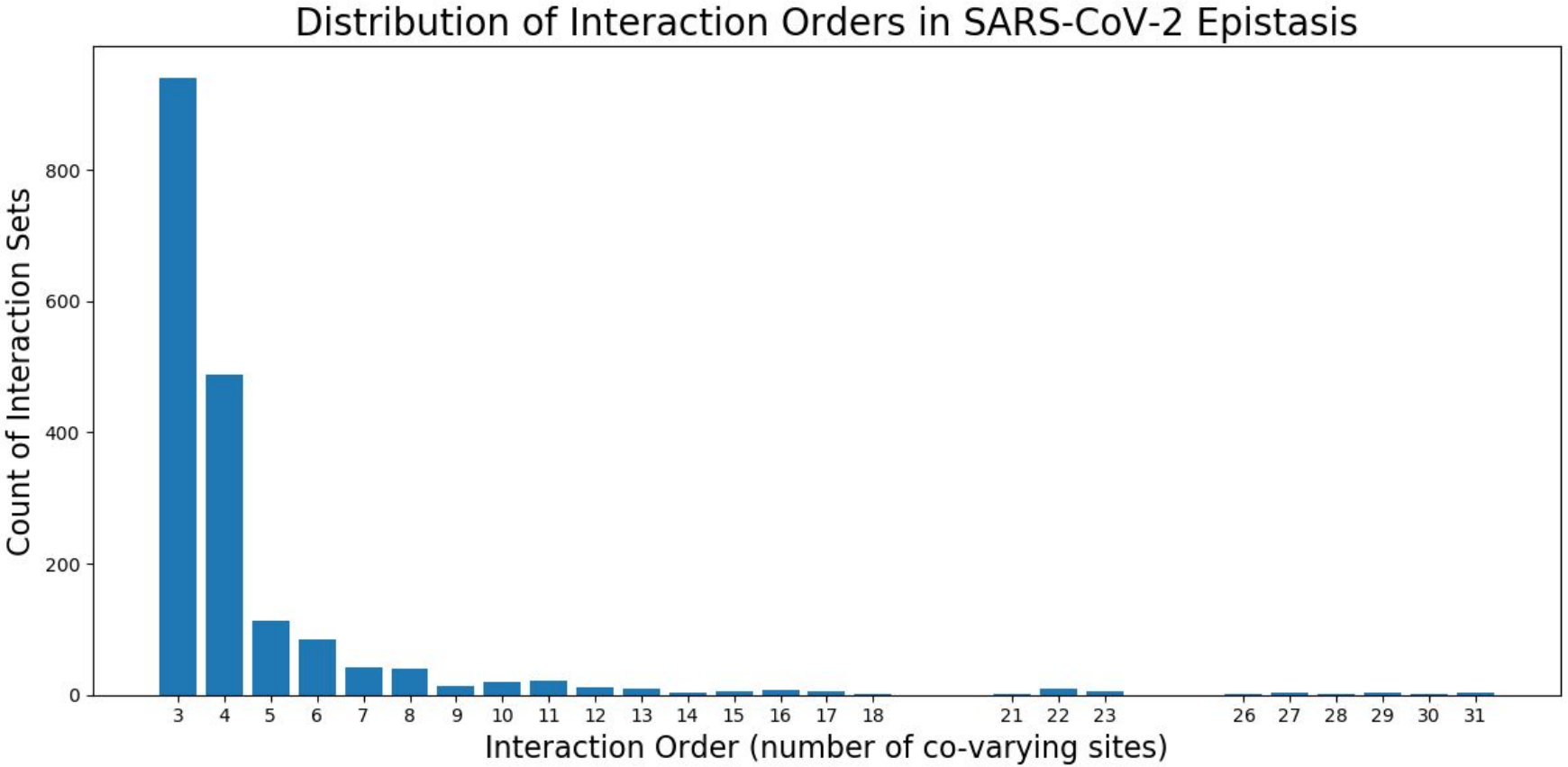
Coordinated evolution at amino acid sites of SARS-CoV-2 spike
Neverov AD, Fedonin G, Popova A, Bykova D, Bazykin G ***Elife***, 2023; p.e82516.

Взаимодействия высоких порядков

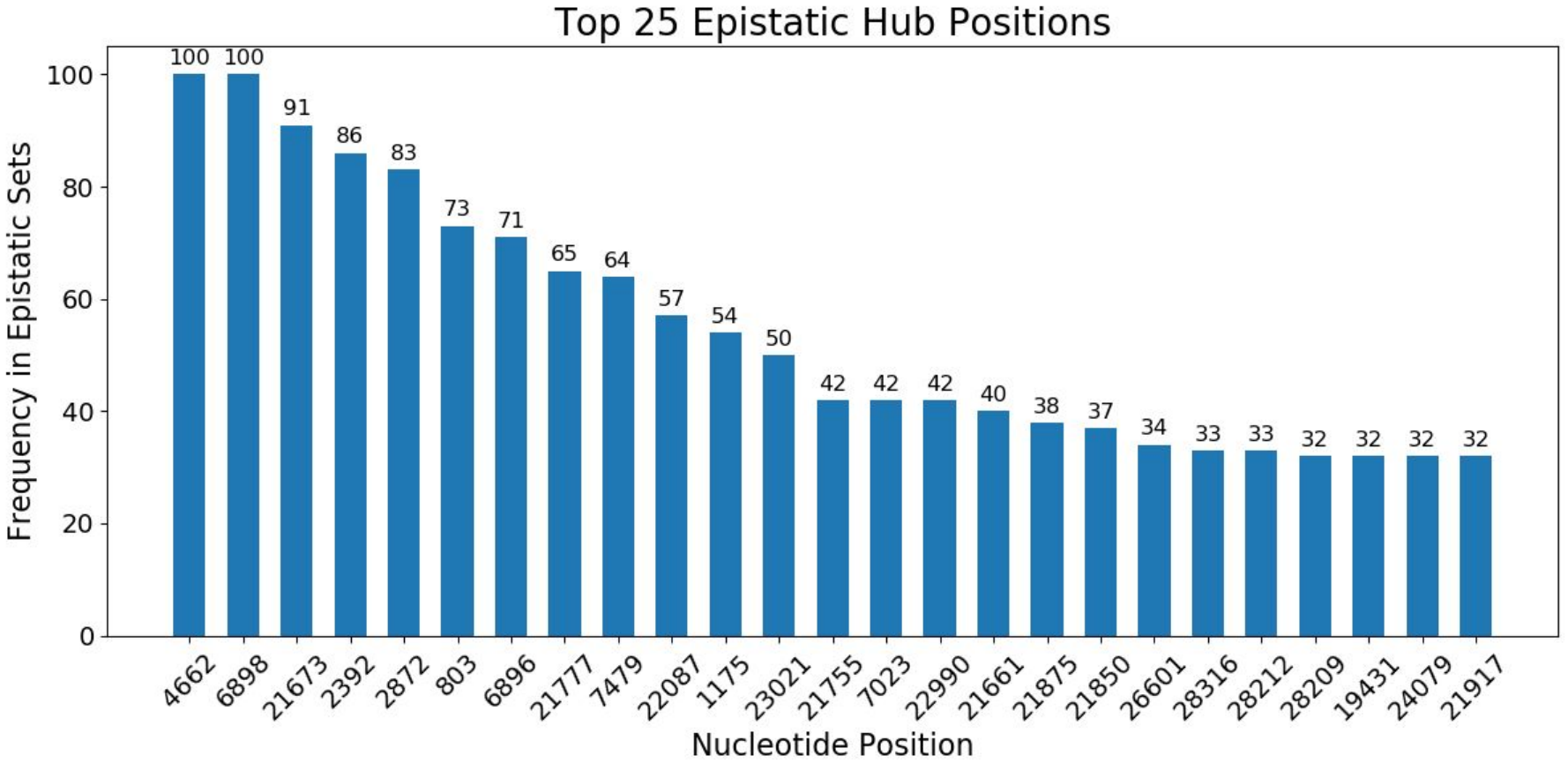
Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1
Moulana A, et al. ***Nat Commun***. 2022;13:7011.

The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution
Moulana A, et al. ***eLife*** 2023;12:e83442.

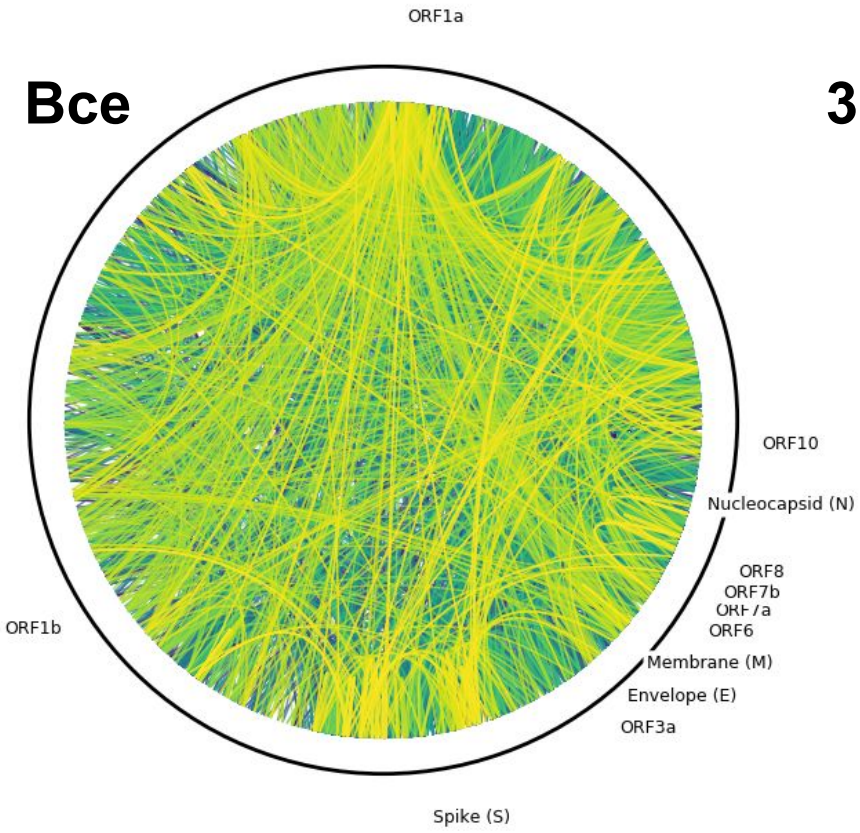
Эпистаз высоких порядков у SARS-CoV-2

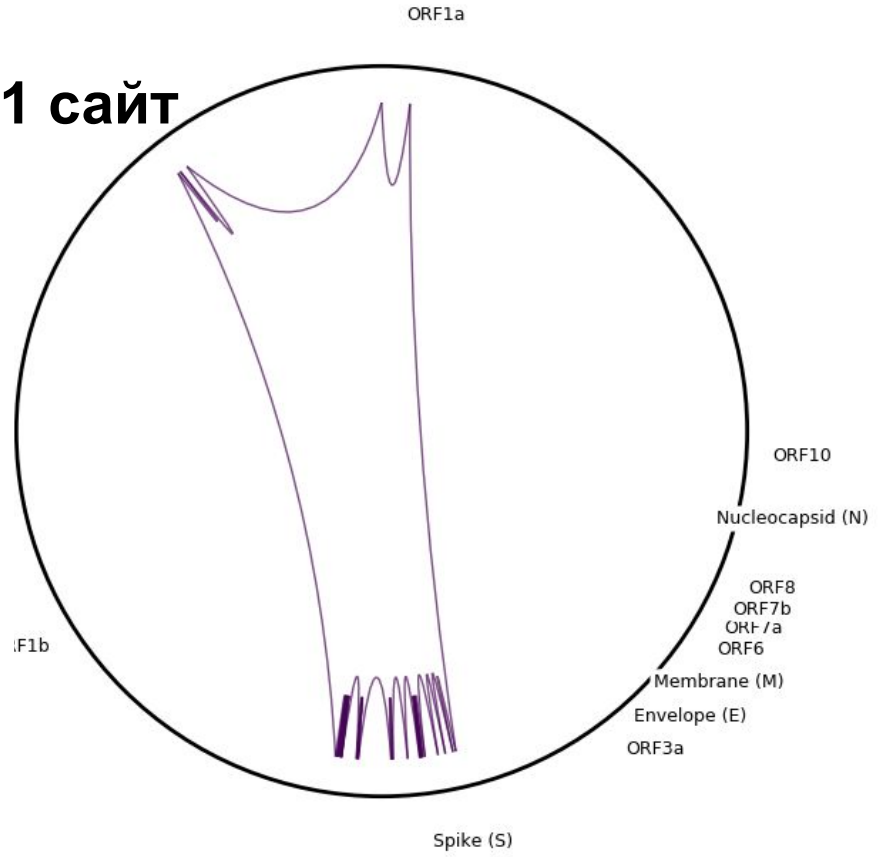


Эпистаз высоких порядков у SARS-CoV-2



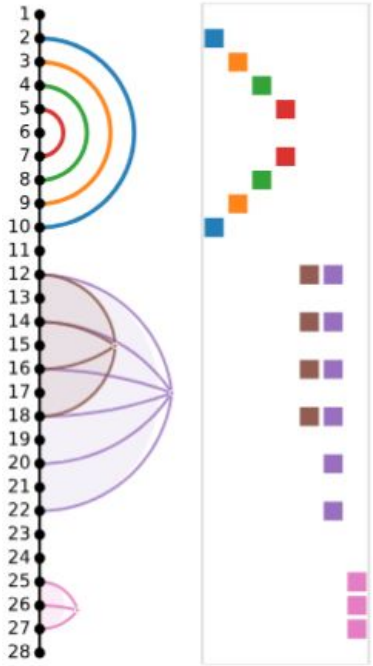
Эпистаз высоких порядков у SARS-CoV-2



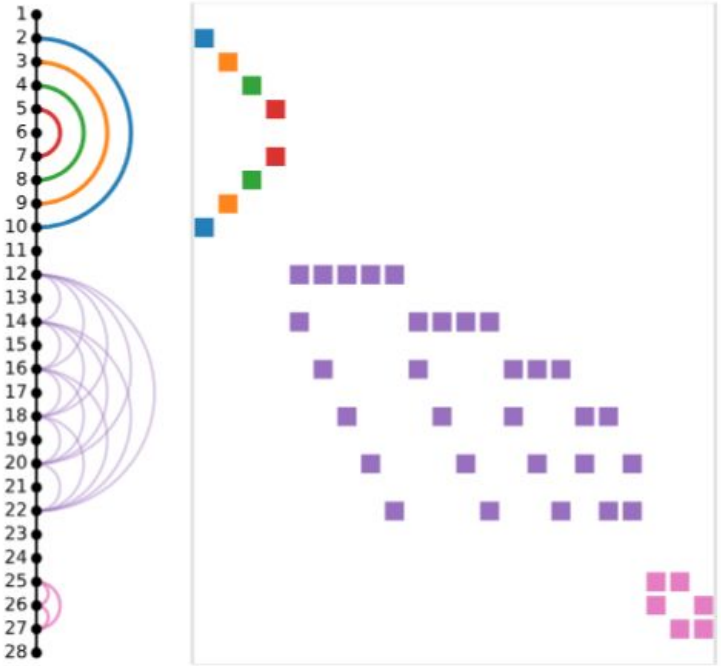


Гиперграф и проекция в пространство графов

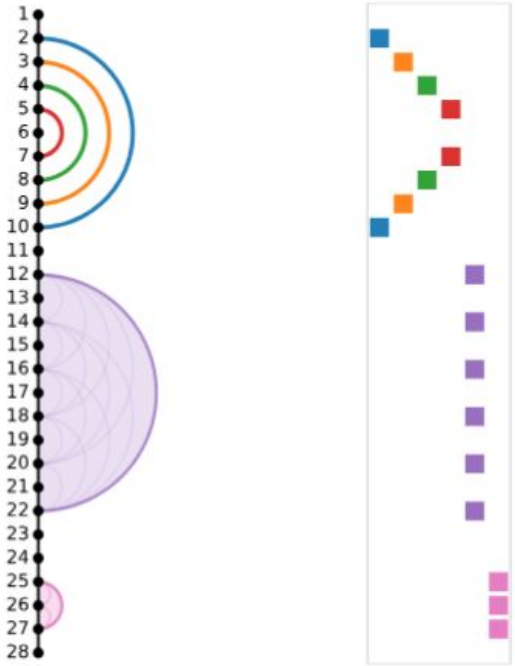
Гиперграф



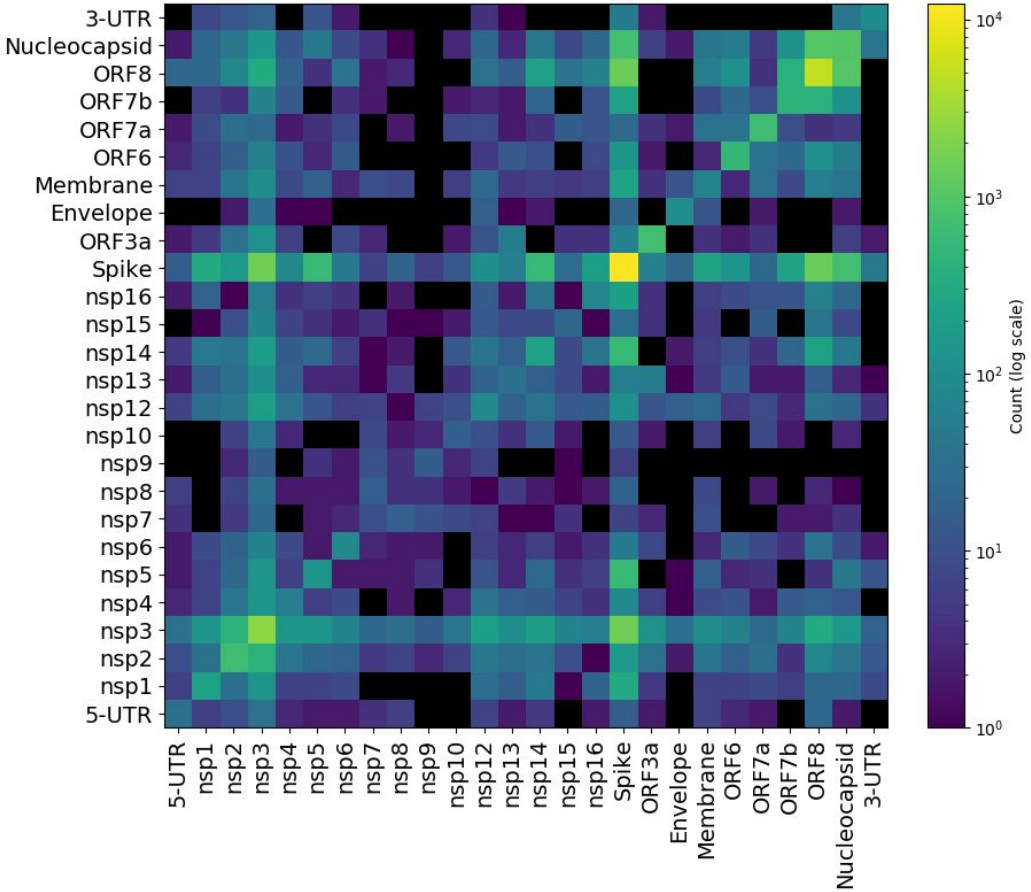
Граф



Симплициальный комплекс



Эпистаз высоких порядков у SARS-CoV-2



Текущее состояние исследований

- Завершается работа над программой (быстрая, параллельная, константная по памяти имплементация на Rust)

Следующие шаги:

- Филогенетическая обусловленность
- Интерпретация результатов на уровне внутри- и межмолекулярных взаимодействий вирусных и человеческих факторов (РНК-РНК, РНК-белок, белок-белок, ДНК-белок)
- Контекст эпидемиологического и инфекционного процесса

Куда дальше?

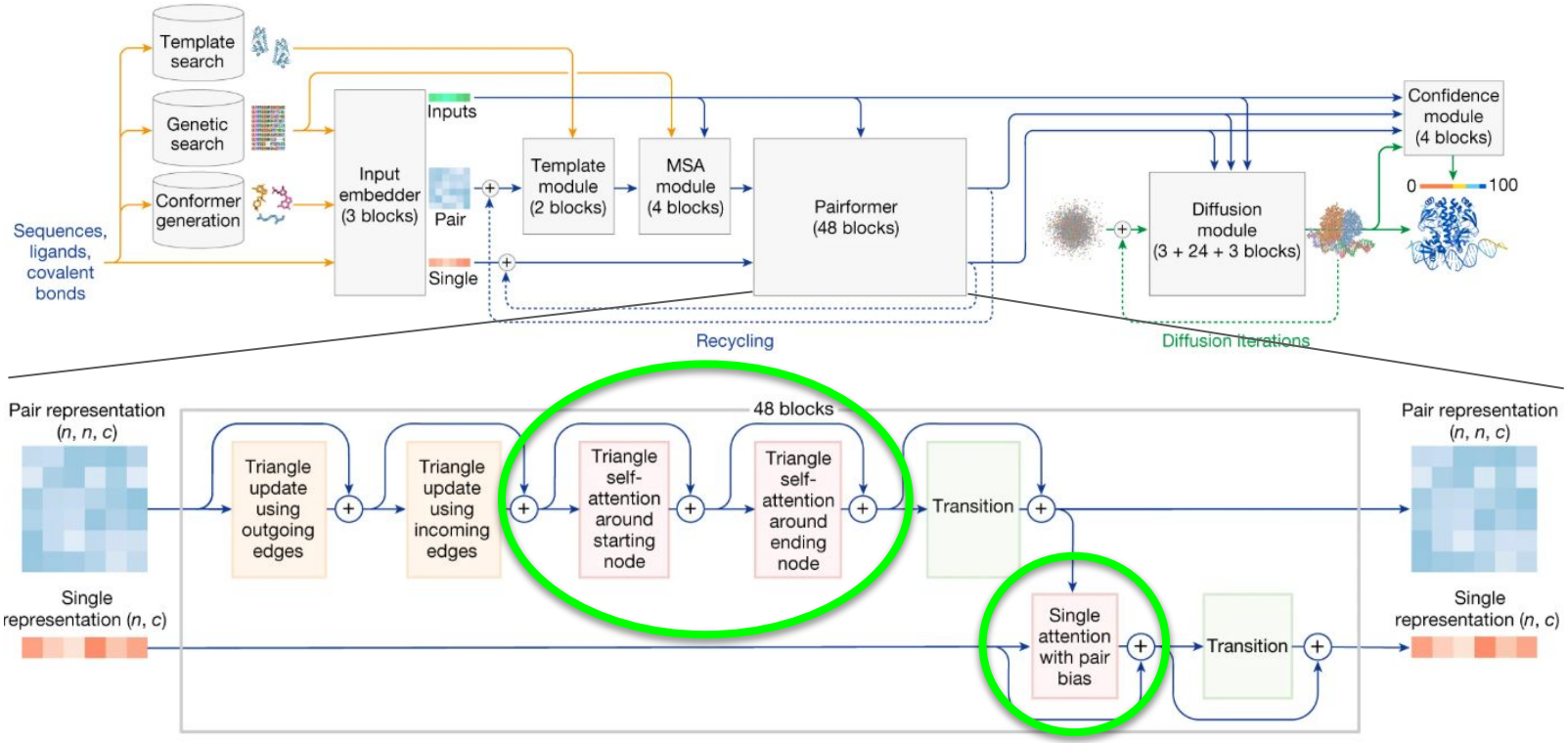
Использование по прямому назначению:
- SARS-CoV-2 и другие коронавирусы, грипп, HIV, HCV, энтеровирусы, денге, зика, западный нил
- Поиск отдаленных гомологов, обобщение PFAM
- Определение функционально-связанных генов у вирусов и прокариот

Родственные задачи в математике:
- Линейные коды с единичным кодовым расстоянием
- Динамика популяций на графах Кэли, теория квазивидов, эпидемиология
- Метрическая геометрия, расстояния Хаусдорфа и Громова-Хаусдорфа
- Не максимальные коэффициенты Фурье

Адаптация алгоритма:
- Геном человека
- Задачи CatBoost, fp-growth, TabPFN, GNN, …
- Изображения и CNN

Использование результатов:
- Прогноз эволюции вирусов
- Генеративные модели для биологических последовательностей и табличных (векторных) данных
- Альфафолд, альфагеном, механизм внимания в трансформерах

Альфафолд и попарные взаимодействия аминокислот



https://www.nature.com/articles/s41586-024-07487-w

Ограничения белковых языковых моделей

nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01176-7

A flaw in using pretrained protein language models in protein–protein interaction inference models

Received: 29 April 2025

Accepted: 18 December 2025

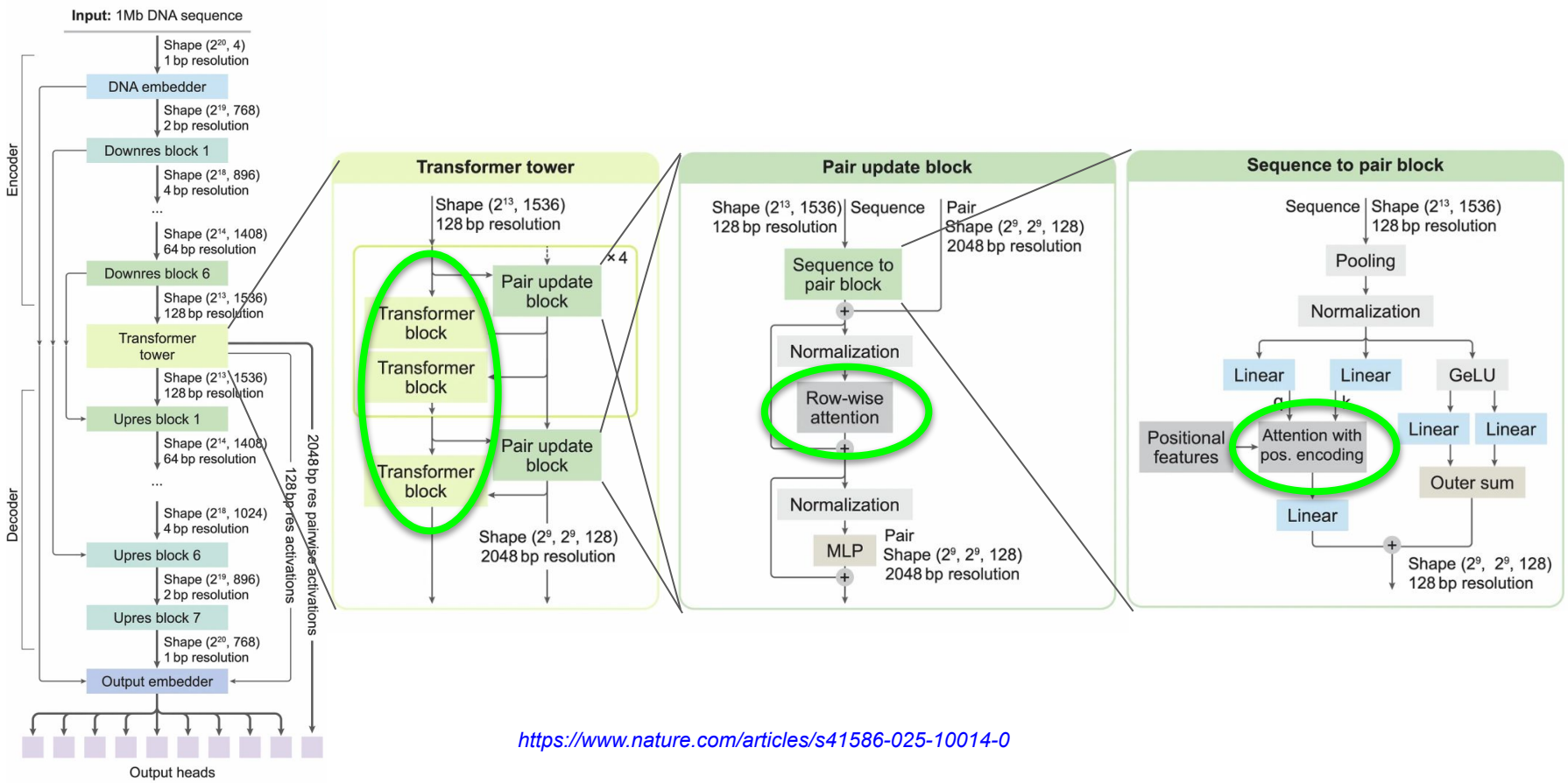
Published online: 13 February 2026

Check for updates

Joseph Szymborski [1,2] & Amin Emad [1,2,3,4]

With the growing pervasiveness of pretrained protein language models (pLMs), pLM-based methods are increasingly being put forward for the protein–protein interaction (PPI) inference task. Here we identify and confirm that existing pretrained pLMs are a source of data leakage for the downstream PPI task. We characterize the extent of the data leakage problem by training and comparing small and efficient pLMs on a dataset that controls for data leakage (strict) with one that does not (non-strict). Although data leakage from pretrained pLMs cause a measurable inflation of testing scores, we find that this does not necessarily extend to other, non-paired biological tasks such as protein keyword annotation. Further, we find no connection between the context lengths of pLMs and the performance of pLM-based PPI inference methods on proteins with sequence lengths that surpass it. Furthermore, we show that pLM-based and non-pLM-based models fail to generalize in tasks such as prediction of the human-SARS-CoV-2 PPIs or the effect of point mutations on binding affinities. This study demonstrates the importance of extending existing protocols for the evaluation of pLM-based models applied to paired biological datasets and identifies areas of weakness of current pLM models.

Альфагеном и попарные связи между токенами



https://www.nature.com/articles/s41586-025-10014-0

Спасибо за внимание!

- Вопросы?
- Замечания?
- Предложения?
- Ответы?

ilya.belalov@gmail.com