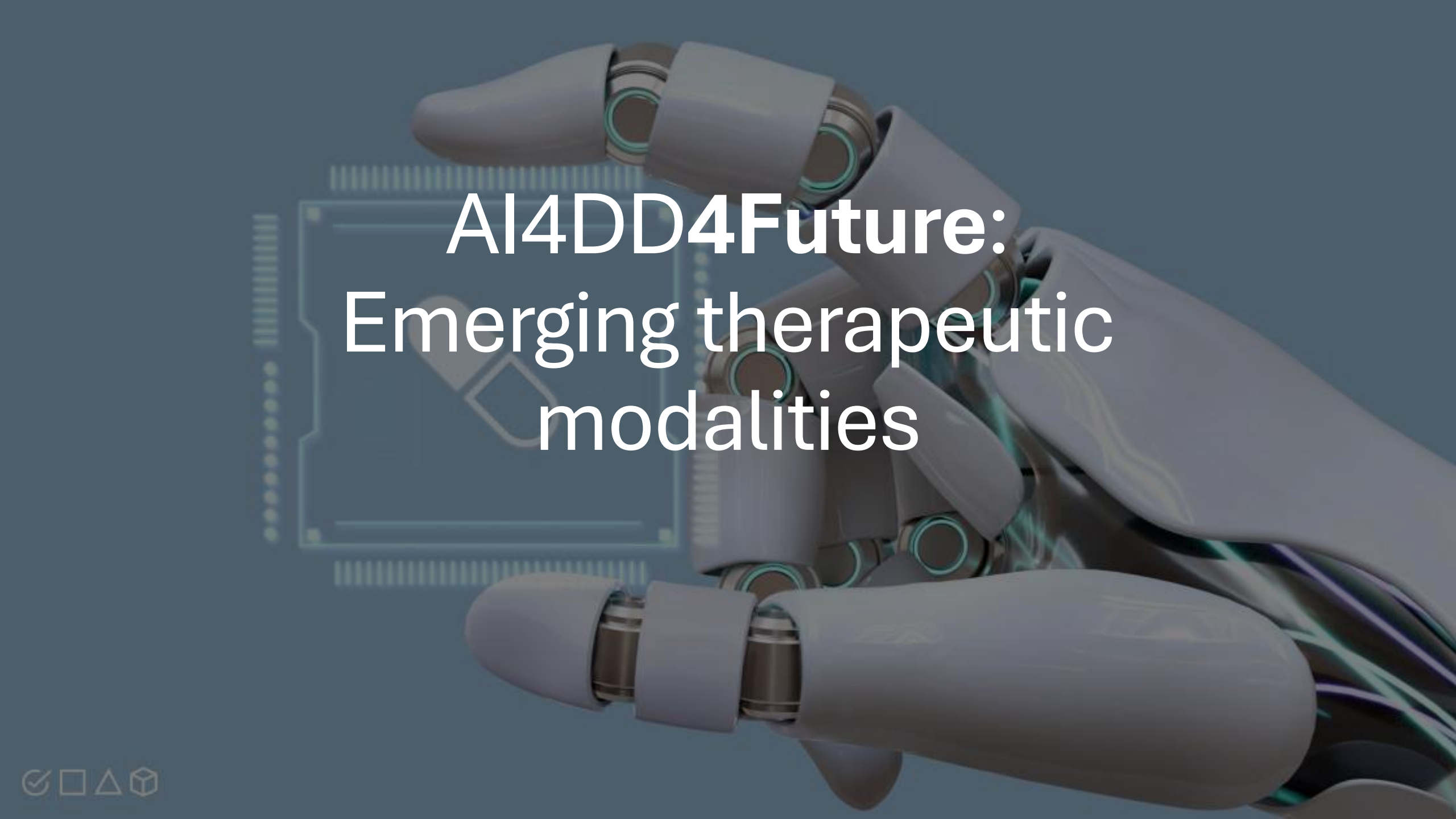


AI4DD**4Future**: Emerging therapeutic modalities

Outline

1. Targeting “undruggable” proteins
   - PROTACs
   - Molecular Glues
2. When drugs are peptides and proteins
   - AI for biologics and bioengineering
3. When drugs are genes
   - Gene therapy and CRISPR therapy

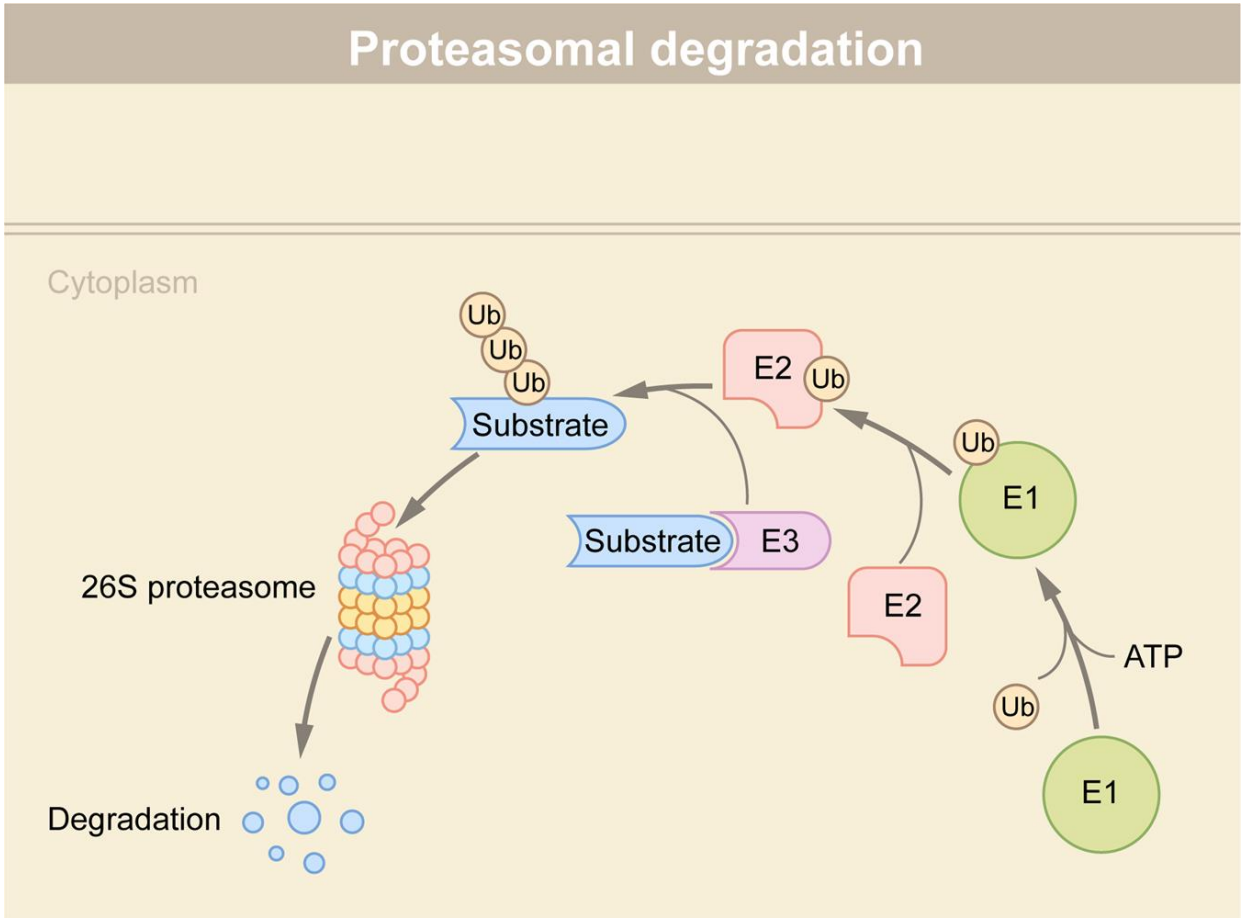
1. Targeting the “undruggable”:
Moving beyond traditionally-druggable proteins

- Recall: only 25% of proteome considered “druggable”
- Yet protein-protein interactions (often undruggable due to large protein surface interfaces) are highly disease-relevant

- Enter therapeutic modalities based on induced proximity
  - Bifunctional compounds (PROTACS, bispecific antibodies)
  - Molecular glues

- Considerations for ternary complex:

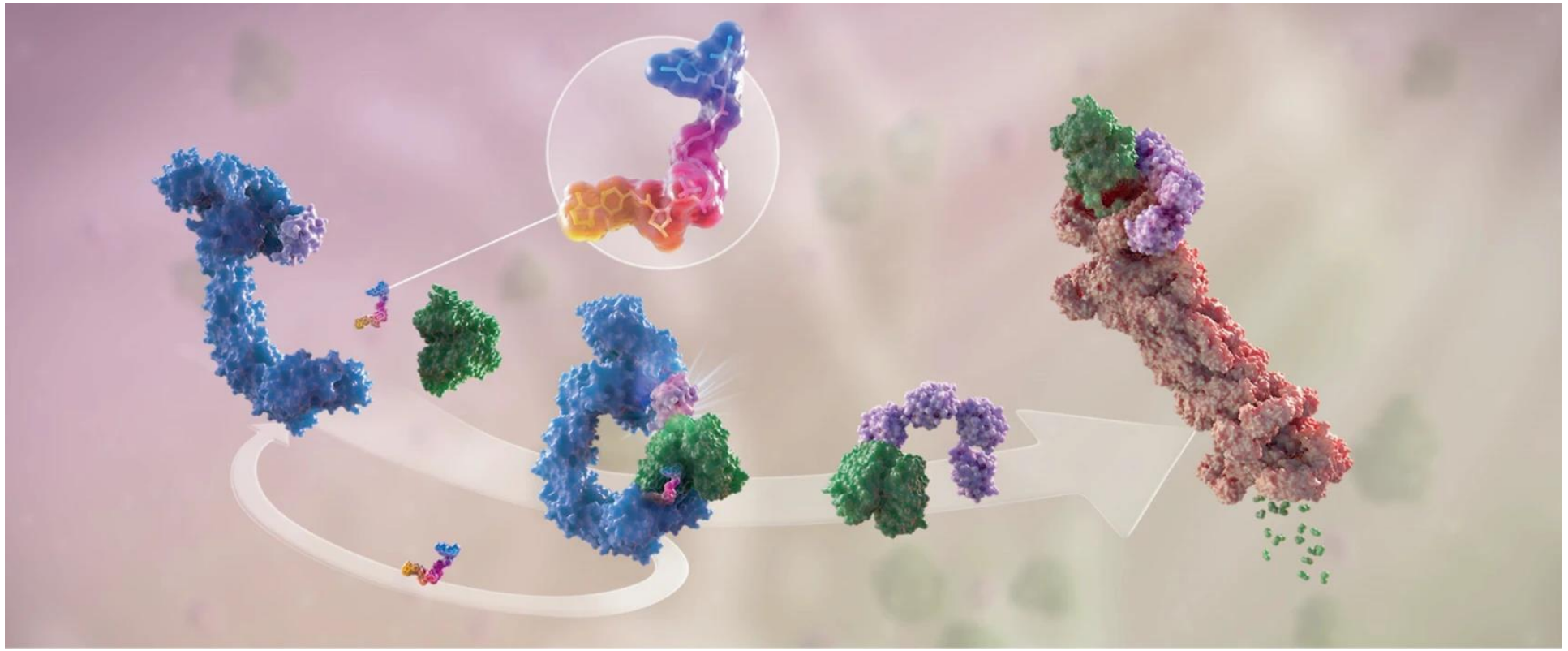
Harnessing the cell's recycling system (proteasome) to degrade proteins



For "undruggable" proteins, instead of inhibiting them, try to target them for degradation

Zhao et al., 2022, Signal Transduction and Targeted Therapy

Schematic of PROTAC action

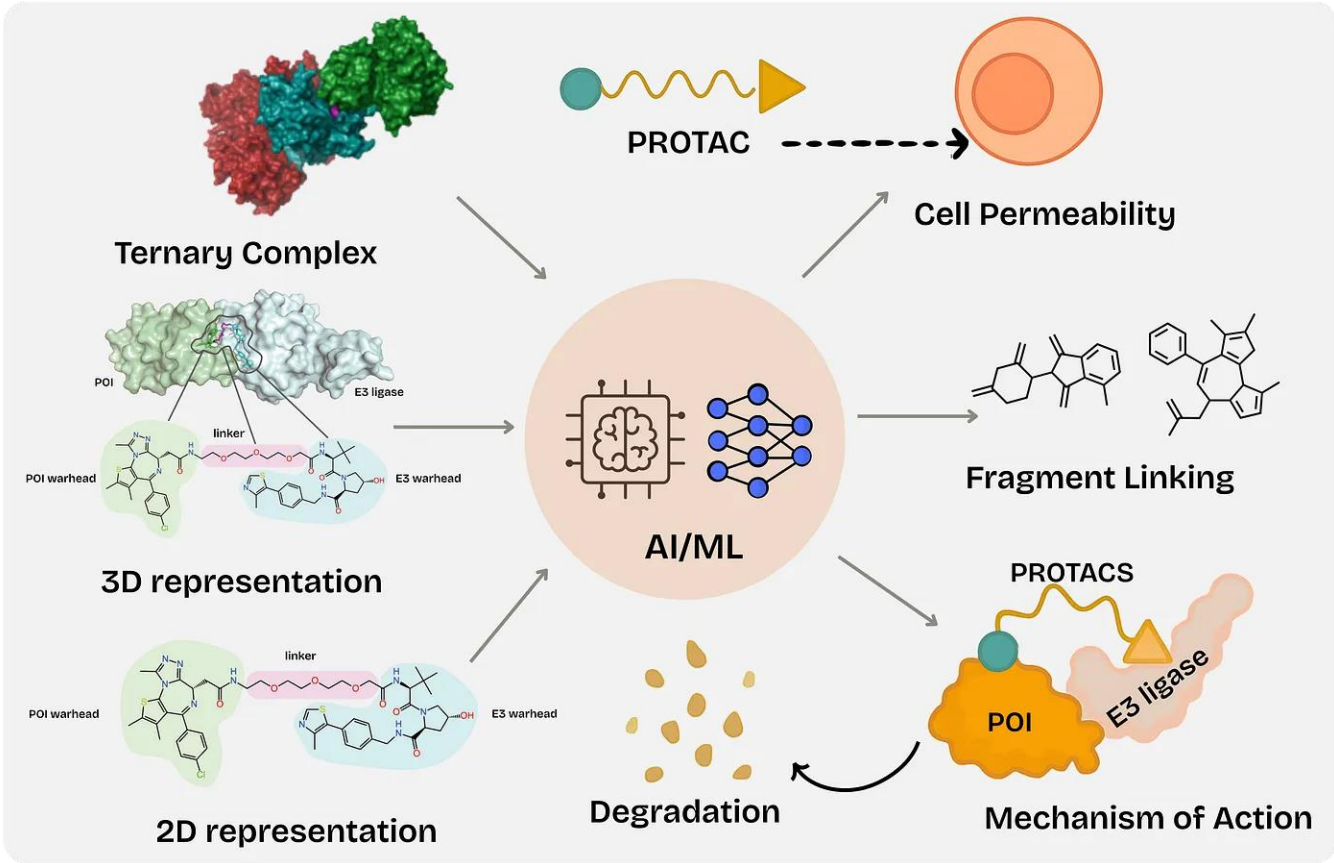


Circle = PROTAC. Blue = ubiquitin ligase, green = target protein. PROTACs bring the ligase and target close together (“induced proximity”). The ligase activity joins ubiquitin (pink/purple) to the target, which destines it for recycling.

Bekes et al., 2022. Nat Rev Drug Discovery

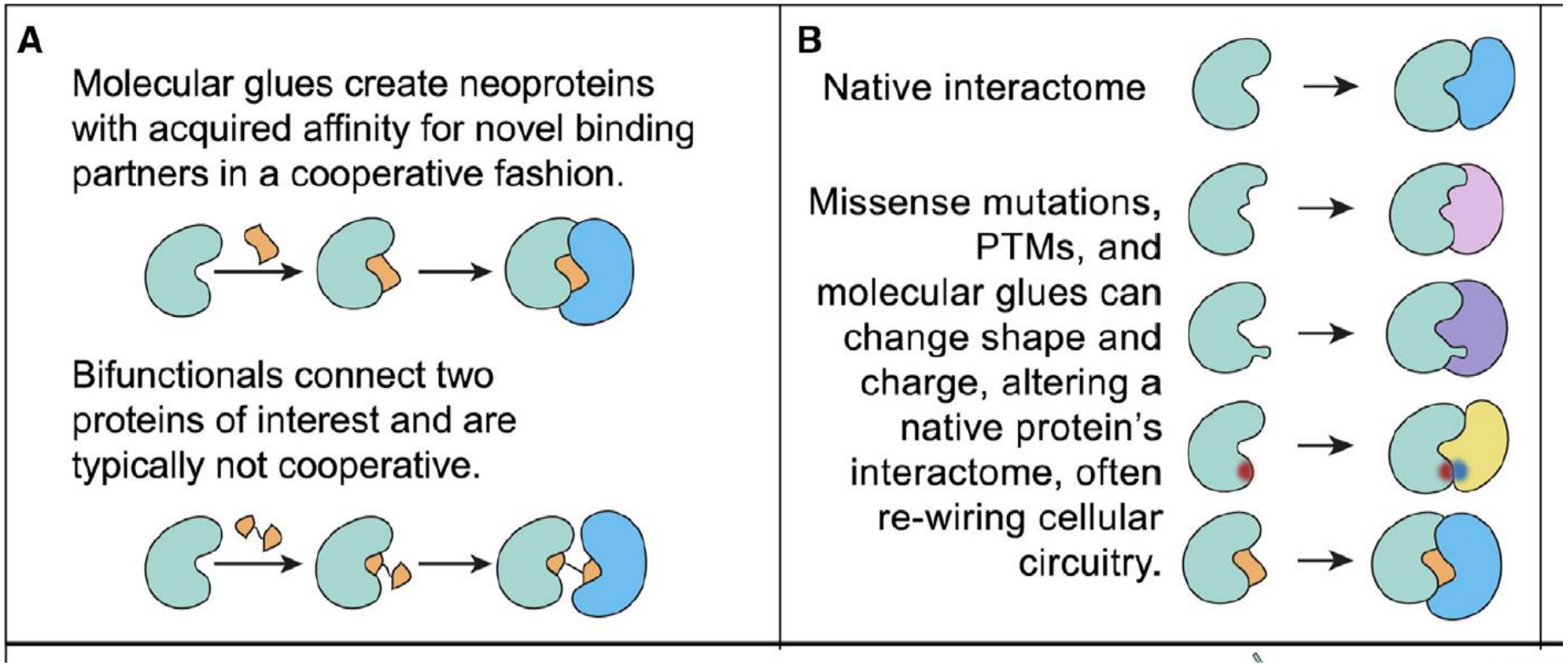
PROTACS – **Pro**teolysis **ta**rgeting **c**himera**s**

- “3-body problem” drugs
- Bulky linker
- In vivo efficacy needs in vivo degradation
  - Cell permeability
  - Enzymatic degradation in cells



Gharbi Y et al., Digital Discovery, 2024

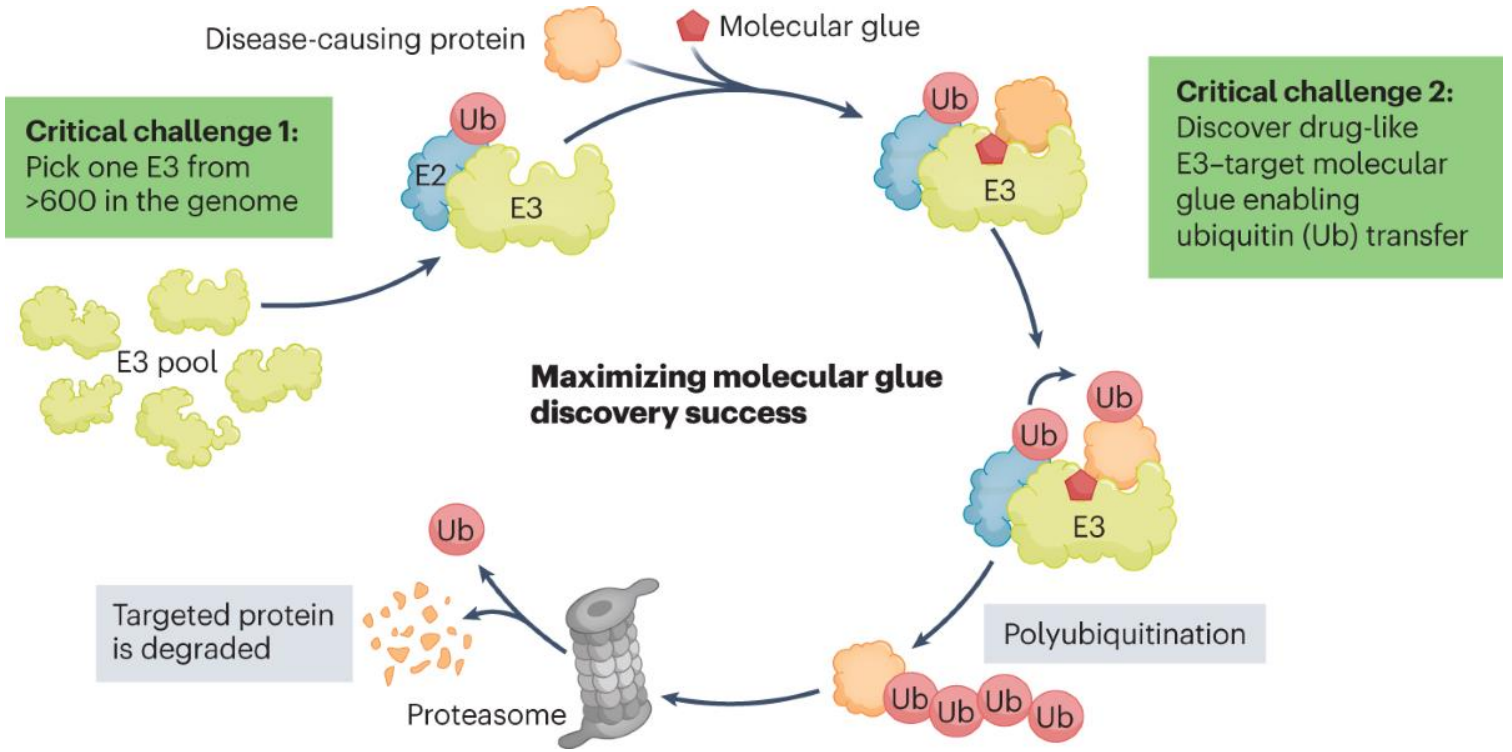
Newer therapeutic strategy: Discover small molecules to rewire cellular circuits via creation of “neoproteins”



Schreiber Cell Chemical Biology 2024

Molecular glues to degrade specific disease-relevant proteins

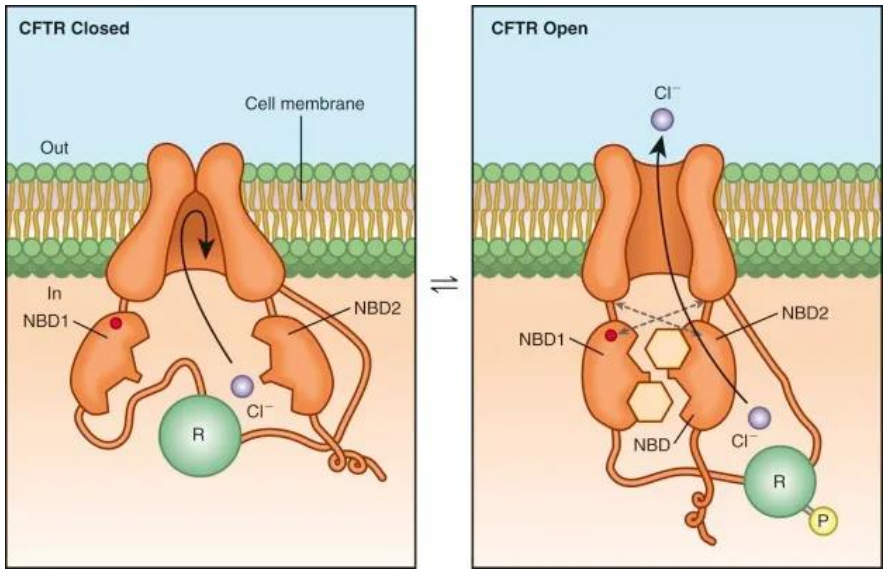
- Small molecules that promote protein degradation by:
- Stabilizing an existing protein-protein interaction or create a new PPI
- Glues are monovalent (compared to bivalent PROTACS) and much smaller than PROTACs
- Found through phenotypic screening
- Low hit rate



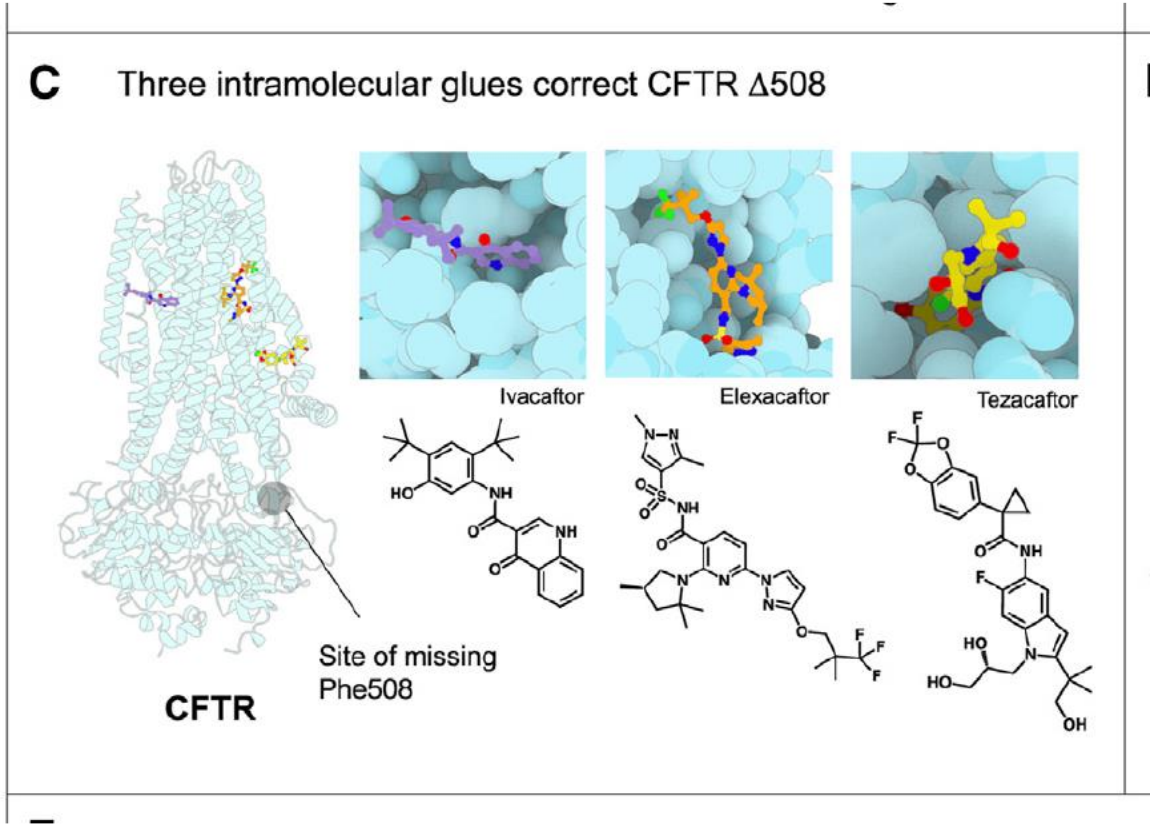
Garber, Nat Biotechnol 2024

Molecular glues: Approved clinical drugs as stabilizers of variant proteins

- Drugs to treat cystic fibrosis

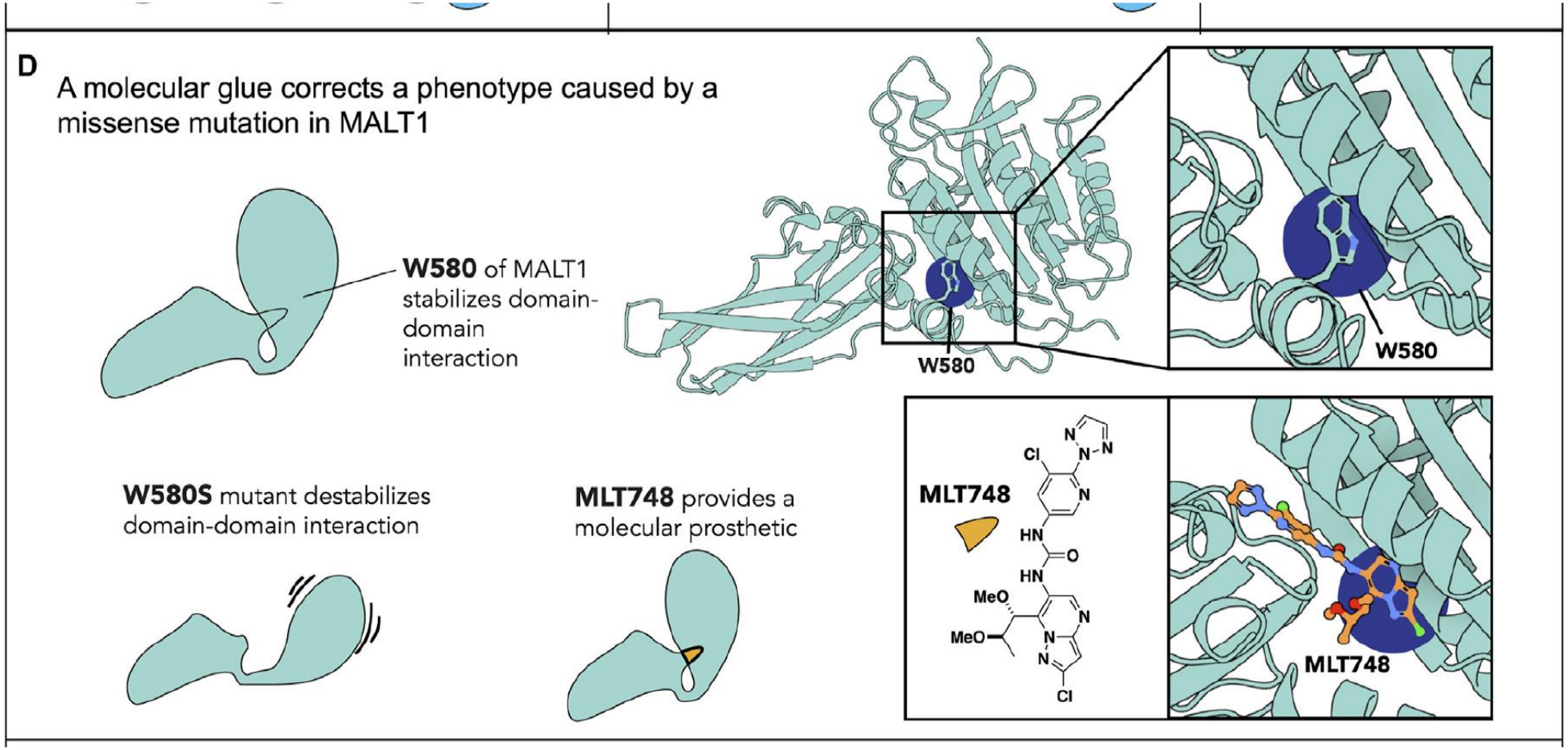


Chloride channel: Cystic Fibrosis Transmembrane Conductance Receptor



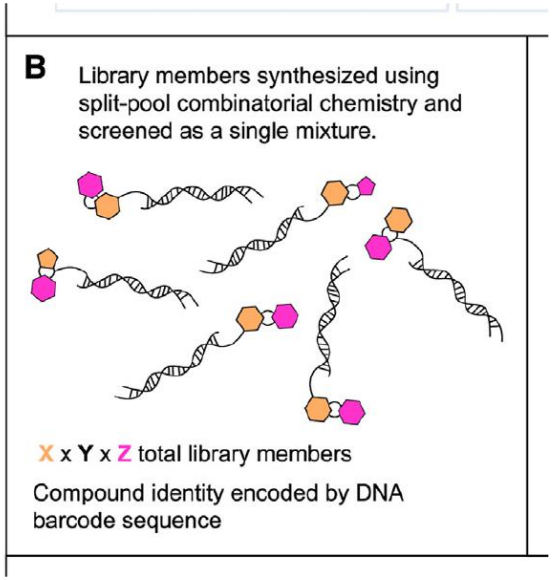
Images: Thoracic Key, Schreiber Cell Chemical Biology 2024

Example: “correcting” phenotypic consequences of proteins encoded by DNA variants (mutations)

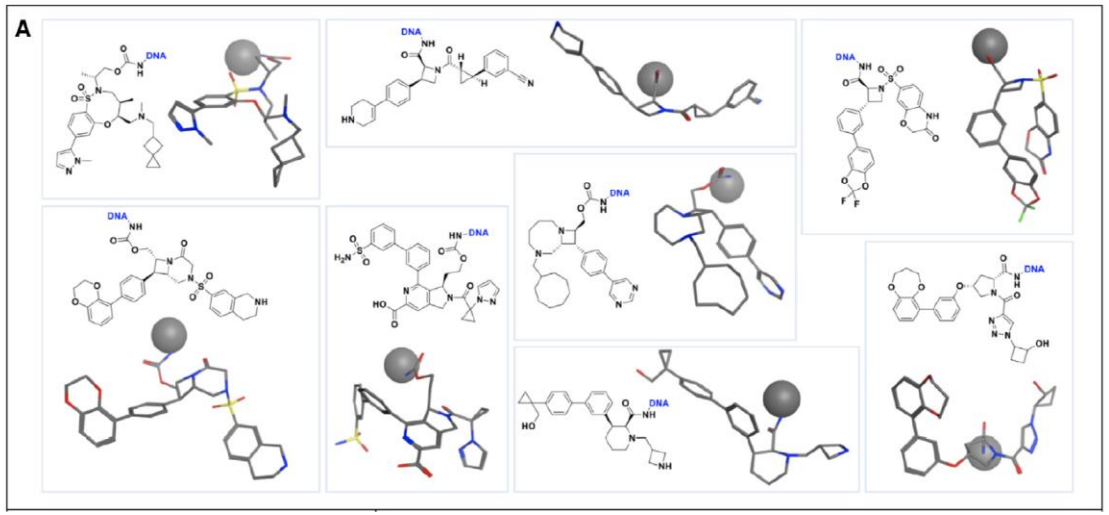


Schreiber Cell Chemical Biology 2024

Chemical space for seeking these kinds of drugs?



DNA-encoded libraries
For high-throughput screening

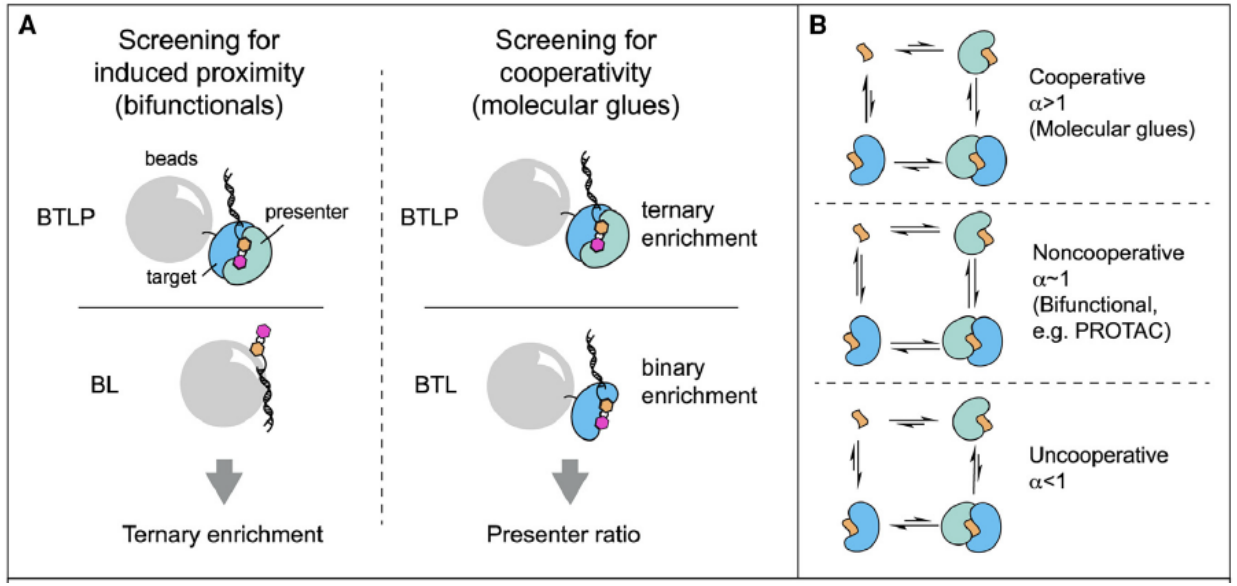


Add geometric and spatial diversity in the DEL →
In order to find drugs that induce proximity or create neoproteins

Our drugs should “do” different things than traditional drugs
→ Explore different regions of chemical space

Screening Strategies

- Analyze the enrichment of DNA barcodes
  - in DNA-Encoded Library screening
- Implied mechanisms based on co-operativity (alpha) score



Issues for **therapeutic modalities of induced proximity**

- Structural moieties involve multiple proteins (complexity)
- The drug has a complex mode of action (not just simply “inhibit”)
- The chemical space is more complex
- Screening using DNA-encoded libraries (DNA as data)
- Consequence (“success”) involves attention to interactome, rather than single readout
- Clinically – a chance for much more selectivity and conditional interactions
  - → personalized medicine
  - 15 molecular glues are FDA-approved drugs (although not originally “discovered” as molecular glues)

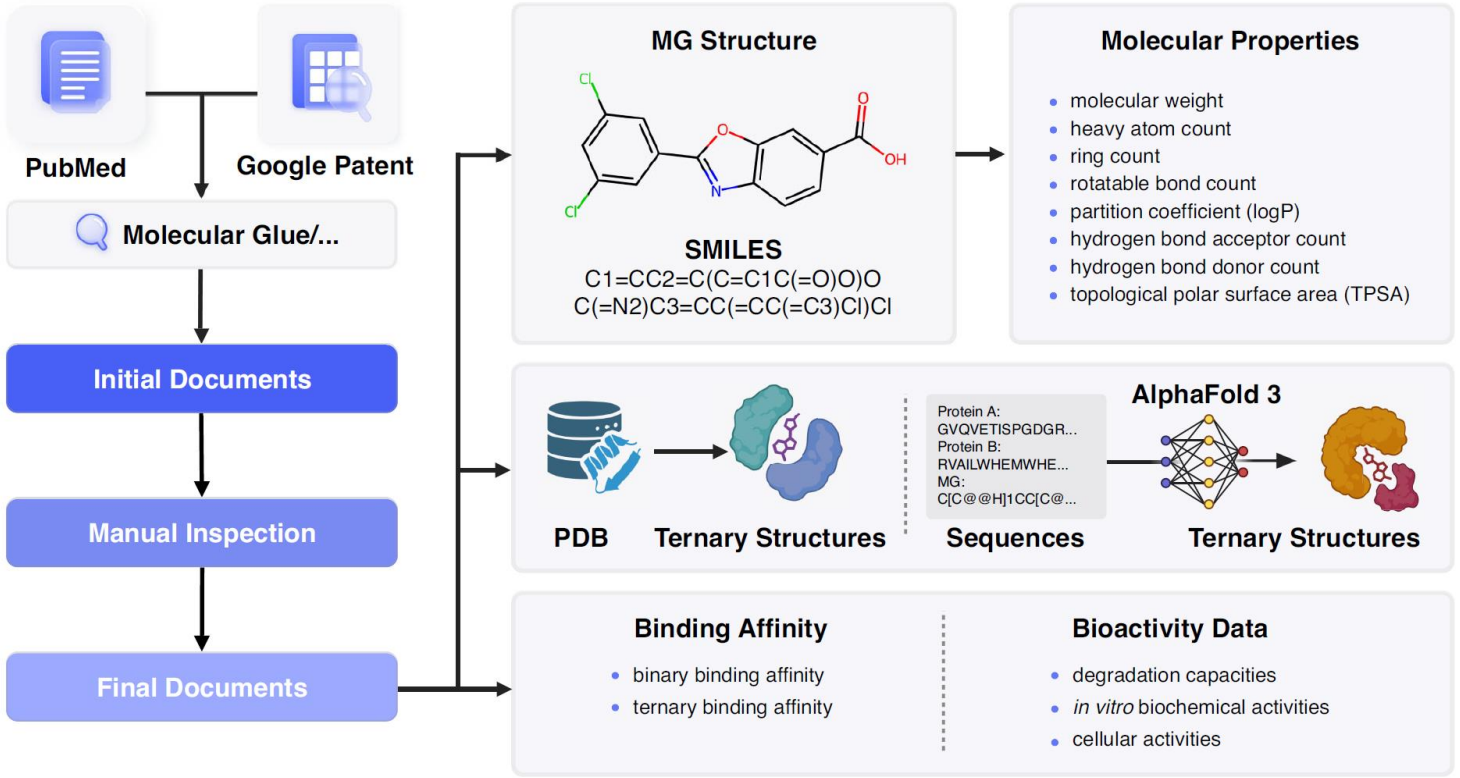
Molecular glue database to explore

Opportunities for AI co-folding models to improve molecular glue discovery

Zhu et al., NAR 2026

Molecular glue database to explore

Multiple datasets combined



https://mgtbind.pkumdl.cn/

Zhu et al., NAR 2026

Implications for AI/ML

- PROTAC = engineered bifunctional
  Glue = emergent interface stabilization

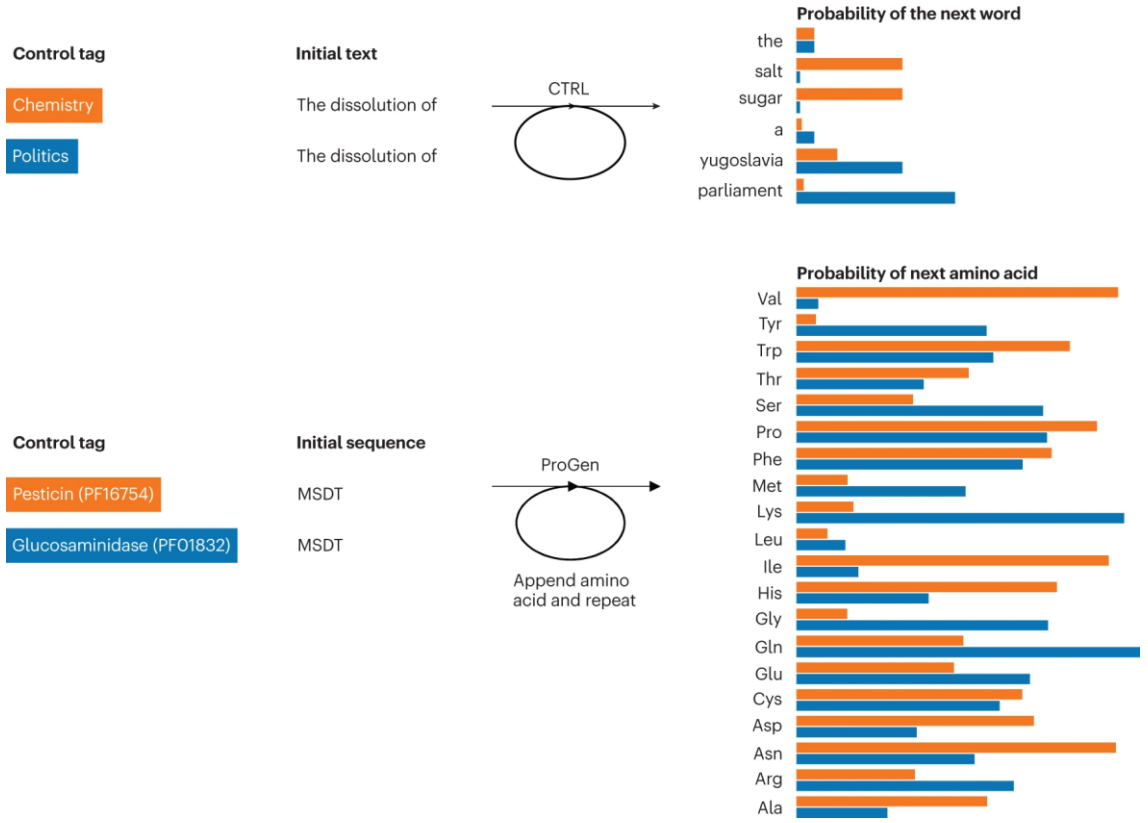
- Ternary complex modeling
  - (eg E3 Ubiquitin ligase + PROTAC + target protein)
- How do we represent three-body interactions?
- Can we predict cooperativity (α factor)?

- Data is:
  - Very sparse
  - Biased (abundance, positive result publication bias) for just a few E3 ubiquitin ligases. Serendipity for molecular glues

- Modeling tasks:
  - Survival models (half-life of protein)
  - Geometry?
  - Drugs need to enter cells to work. What do we want to optimize ? Discuss…

2. When drugs are peptides and proteins: AI for peptide and protein design

Generating artificial proteins with LLMs e.g. (ProGen models)



"ProGen samples protein sequences one amino acid at a time; the probability of each amino acid is influenced by the prefix of the sequence generated so far and a set of control tags that specify a desired protein function. Given the same prefix but different control tags, the resulting distribution may be quite different (orange vs. blue histograms)."

Belanger and Colwell, 2023 Nat Biotech (commentary)

Generating artificial proteins with LLMs (ProGen models)

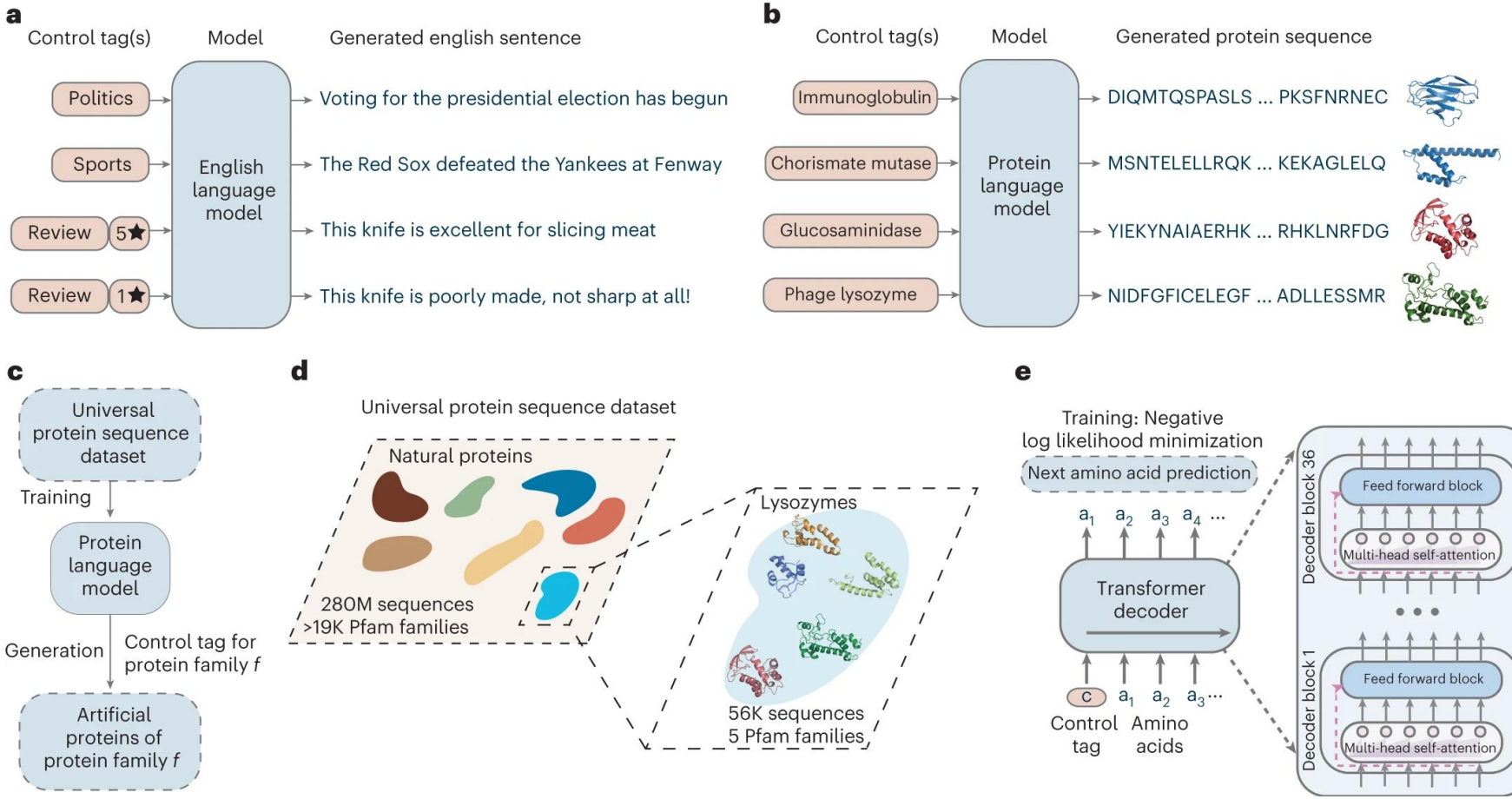
nature biotechnology

Article https://doi.org/10.1038/s41587-022-01618-2

Large language models generate functional protein sequences across diverse families

Received: 12 July 2022
Accepted: 17 November 2022
Published online: 26 January 2023

Ali Madani[1,2], Ben Krause[1,10], Eric R. Greene[3,10], Subu Subramanian[4,5], Benjamin P. Mohr[6], James M. Holton[7,8,9], Jose Luis Olmos Jr.[3], Caiming Xiong[1], Zachary Z. Sun[6], Richard Socher[1], James S. Fraser[3] & Nikhil Naik[1]



Madani et al., 2023 Nat Biotech

Novel artificial antibacterial proteins

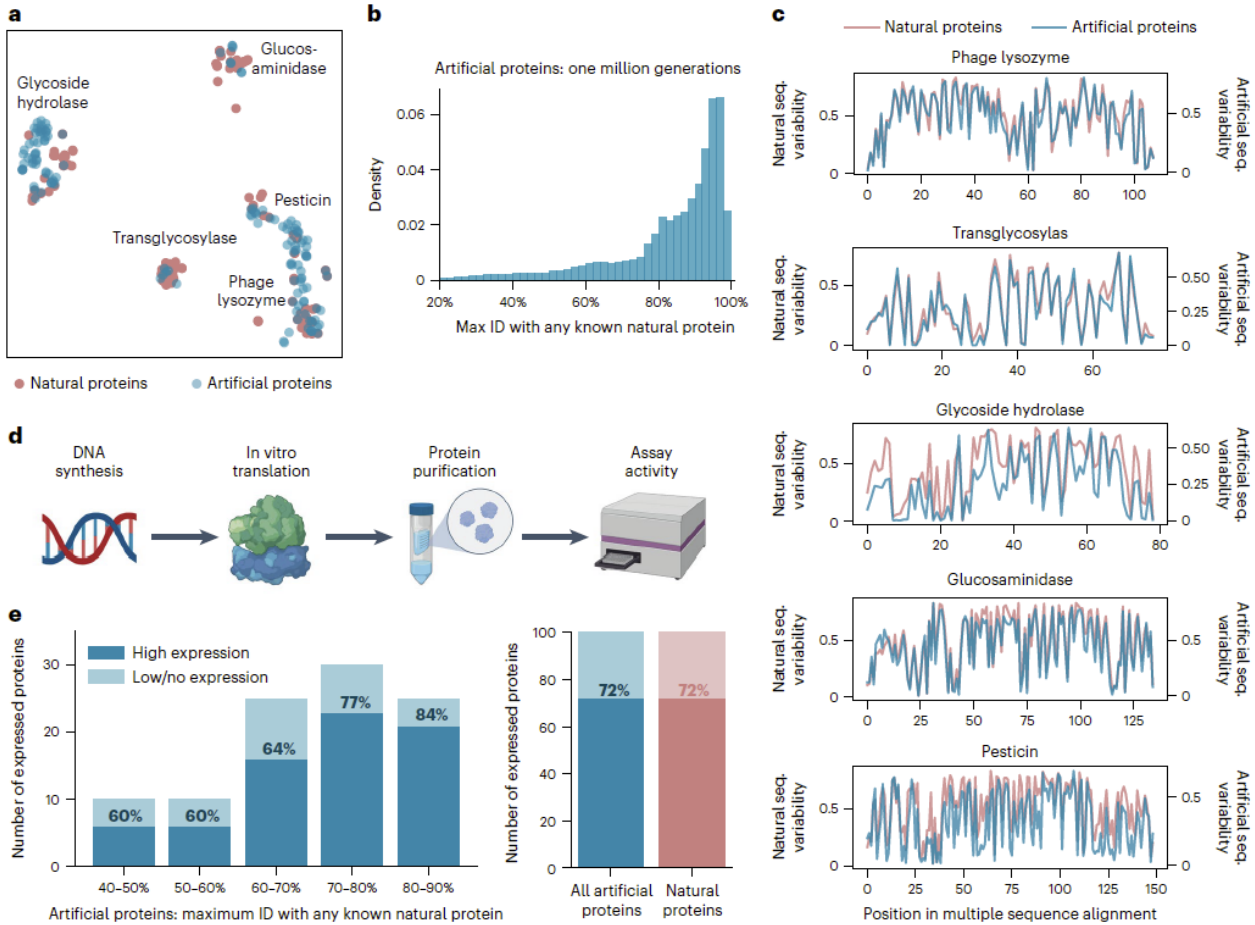
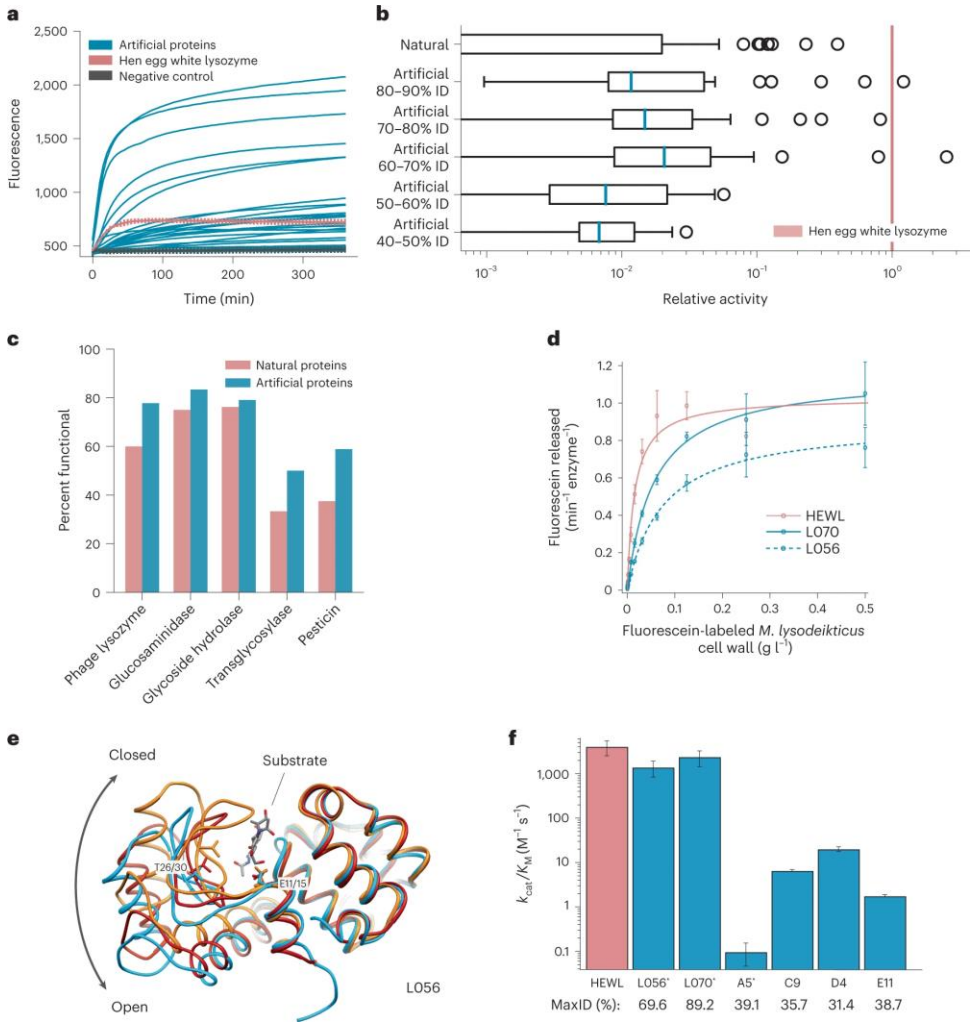


Fig. 2 | Generated artificial antibacterial proteins are diverse and express well in our experimental system. a, When analyzed using *t*-distributed stochastic neighbor embedding (*t*-SNE) as a dimensionality reduction technique for visualization purposes, artificial sequences from our model are shown to span the landscape of natural proteins from five lysozyme families. Each point represents a natural or generated sequence embedded in a two-dimensional *t*-SNE space. **b**, With sufficient sampling, ProGen can generate sequences that are highly dissimilar from natural proteins. Max ID measures the maximum identity of an artificial protein with any publicly available natural protein. **c**, Artificial proteins maintain similar evolutionary conservation patterns as natural proteins across families. Plots demonstrate the variability at each aligned position for a library of proteins. Conserved positions are represented as curve dips. seq., sequence. **d**, From our generated proteins, we select one hundred proteins for synthesis and characterization in our experimental setup. **e**, Artificial proteins express well even with increasing dissimilarity from nature (40–50% max ID) and yield comparable expression quality to one hundred representative natural proteins.

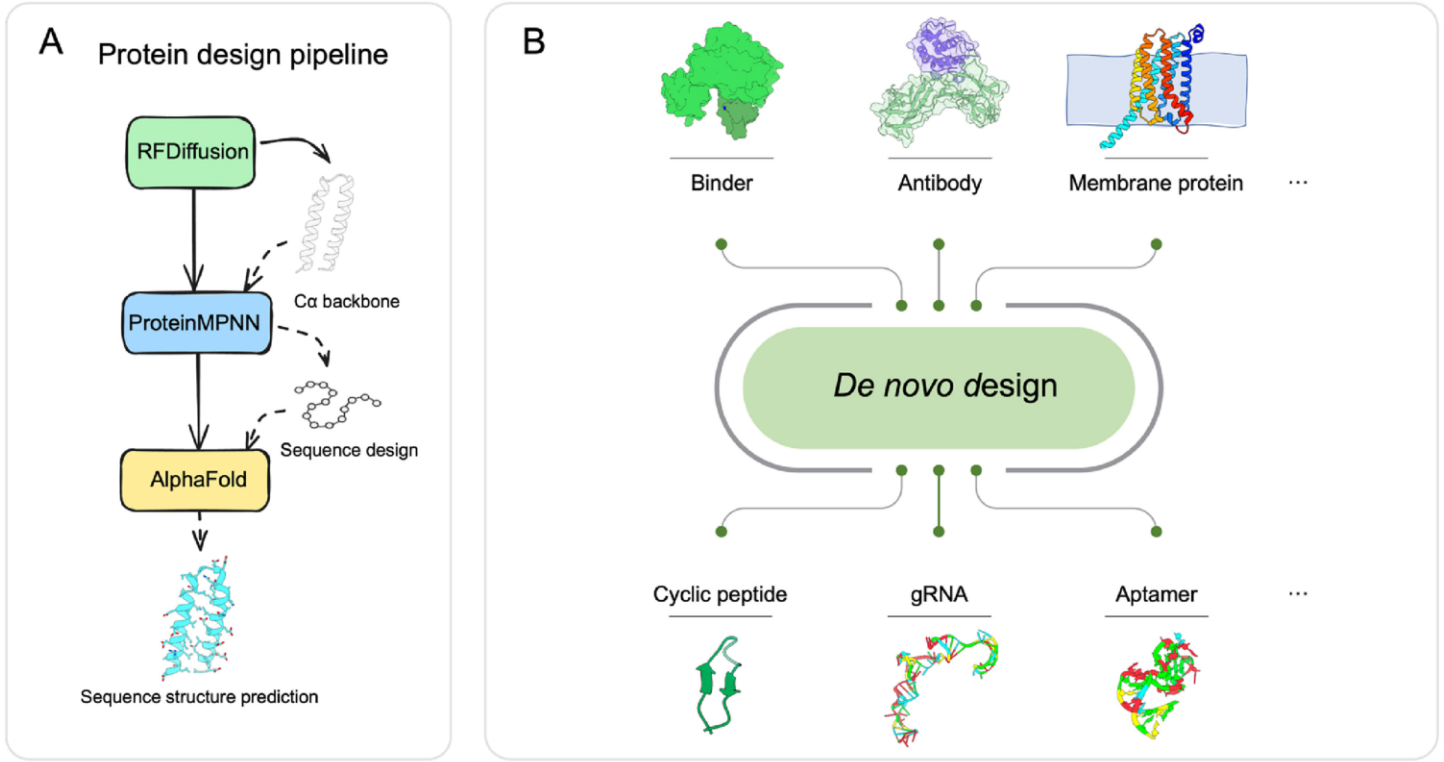
Madani et al., 2023 Nat Biotech

Artificial proteins are functional, with as low as 31% similarity to human proteins



Madani et al., 2023 Nat Biotech

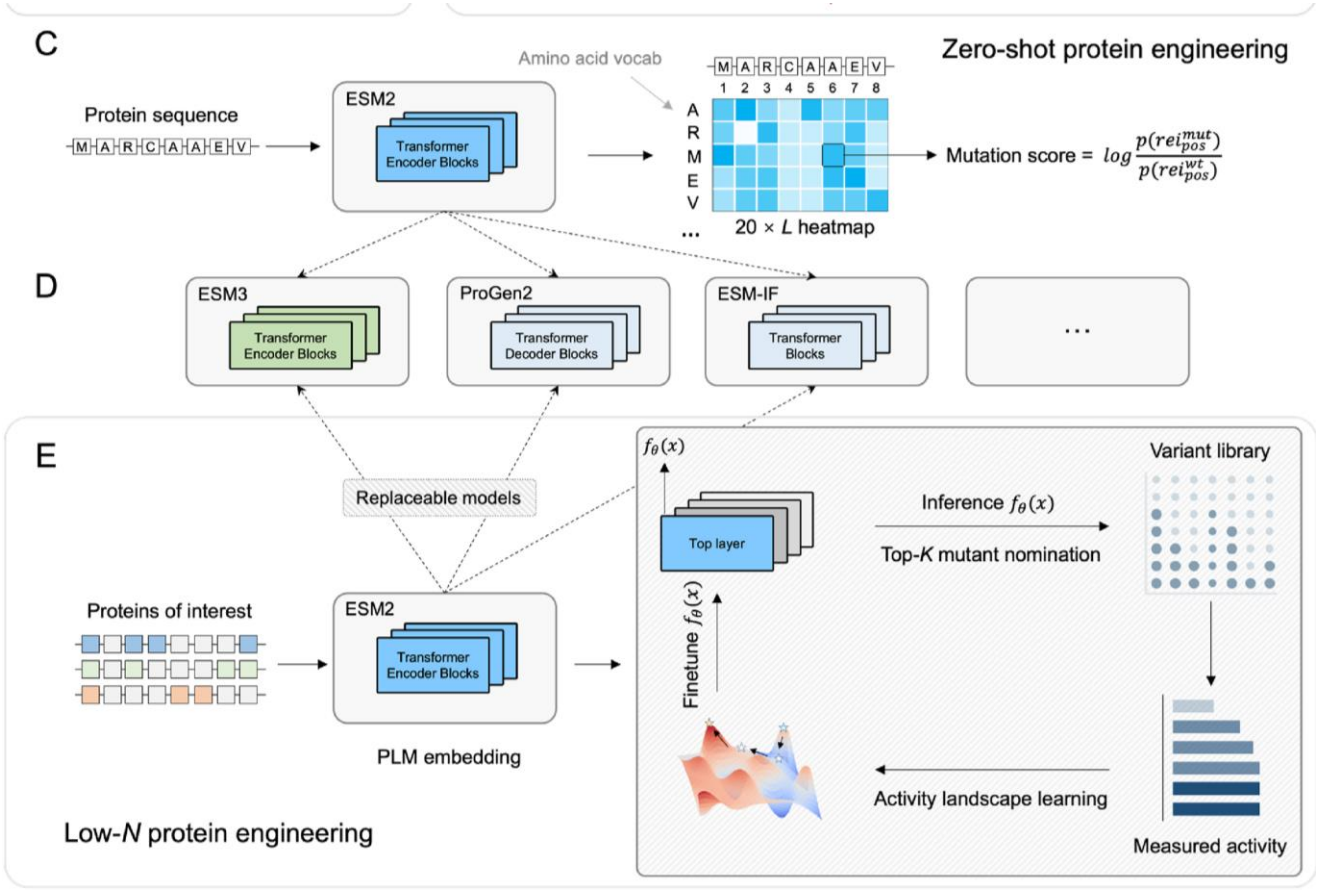
Overview: Algorithms and applications of *de novo* protein design and engineering.



(A) *De novo* biologic drug design pipeline and (B) main therapeutic application scenarios.

Tang et al., Acta Pharmaceutica Sinica 2026

Algorithms and applications of *de novo* protein design and engineering



C, D) Zero-shot protein engineering workflow and interchangeable protein base models, and (E) low-N lightweight supervised fine-tuning directed evolution methods

Tang et al., Acta Pharmaceutica Sinica 2026

nature biotechnology

Article

https://doi.org/10.1038/s41587-025-02761-2

Target sequence-conditioned design of peptide binders using masked language modeling

Received: 24 May 2024

Accepted: 2 July 2025

Published online: 13 August 2025

Check for updates

Leo Tianlai Chen[1,12], Zachary Quinn[1,12], Madeleine Dumas[2,3,12], Christina Peng[4,12], Lauren Hong[1], Moises Lopez-Gonzalez[2], Alexander Mestre[5], Rio Watson[1], Sophia Vincoff[1], Lin Zhao[1], Jianli Wu[5], Audrey Stavrand[2], Mayumi Schaepers-Cheu[2], Tian Zi Wang[1], Divya Srijay[1], Connor Monticello[6], Pranay Vure[1], Rishab Pulugurta[1], Sarah Pertsemlidis[1], Kseniia Kholina[1], Shrey Goel[1], Matthew P. DeLisa[6,7,8], Jen-Tsan Ashley Chi[5], Ray Truant[4], Hector C. Aguilar[2,3] & Pranam Chatterjee[1,9,10,11]

The computational design of protein-based binders presents unique opportunities to access 'undruggable' targets, but effective binder design often relies on stable three-dimensional structures or structure-influenced latent spaces. Here we introduce PepMLM, a target sequence-conditioned designer of de novo linear peptide binders. Using a masking strategy that positions cognate peptide sequences at the C terminus of target protein sequences, PepMLM finetunes the ESM-2 protein language model to fully reconstruct the binder region, achieving low perplexities matching or improving upon validated peptide–protein sequence pairs. After successful in silico benchmarking with AlphaFold-based docking, we experimentally validate the efficacy of PepMLM through both binding and degradation assays. PepMLM-derived peptides demonstrate sequence-specific binding to cancer and reproductive targets, including NCAM1 and AMHR2, and enable targeted degradation of proteins across diverse disease contexts, from Huntington's disease to live viral infections. Altogether, PepMLM enables the design of candidate binders to any target protein, without requiring structural input, facilitating broad applications in therapeutic development.

- Designing protein binders

Overview and evaluation of the PepMLM model (review this after the Genome Language Module)

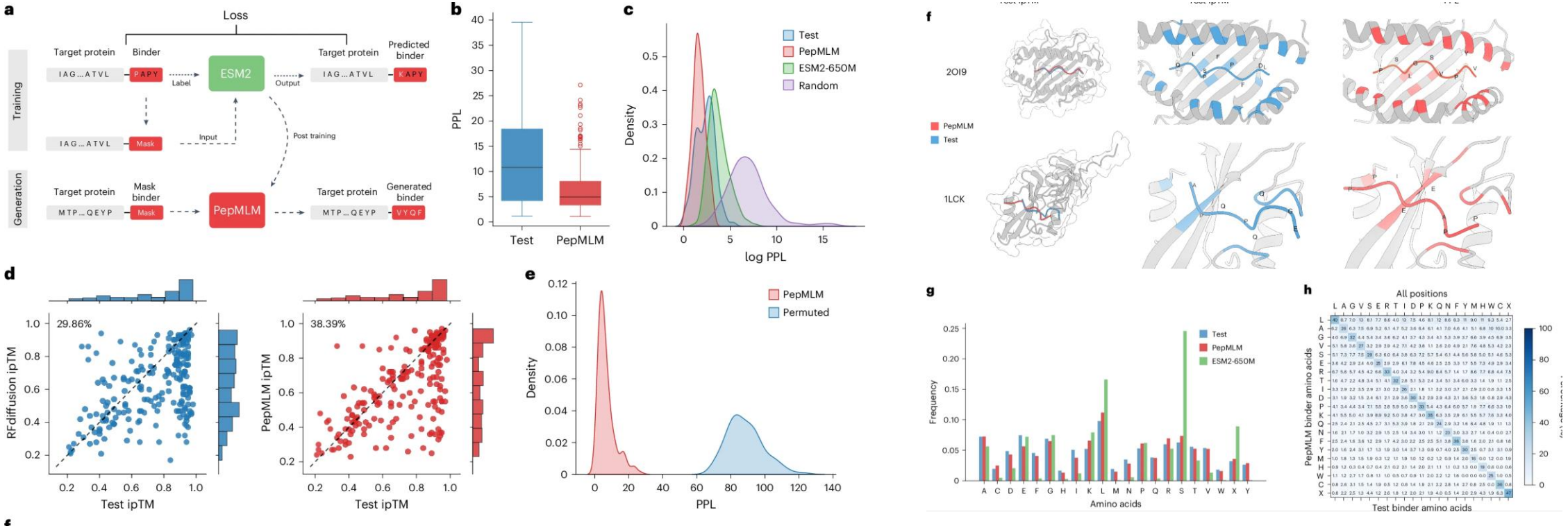
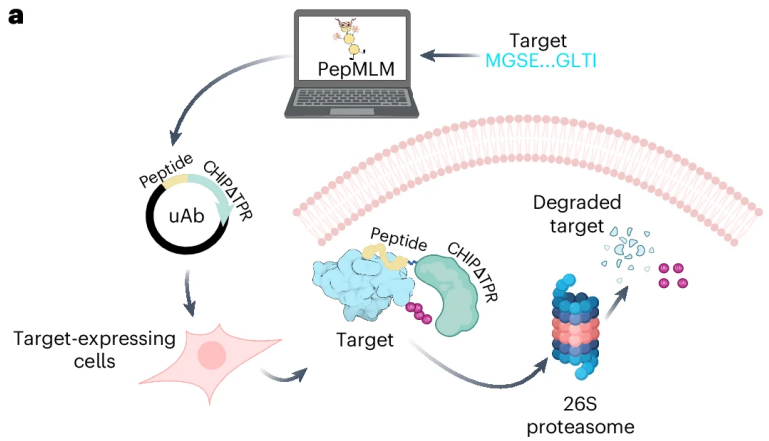
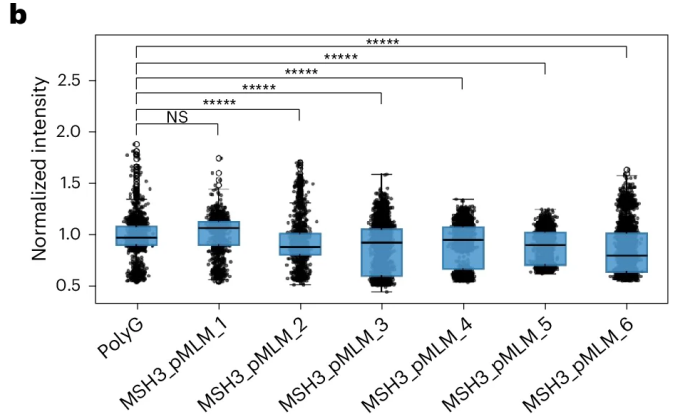


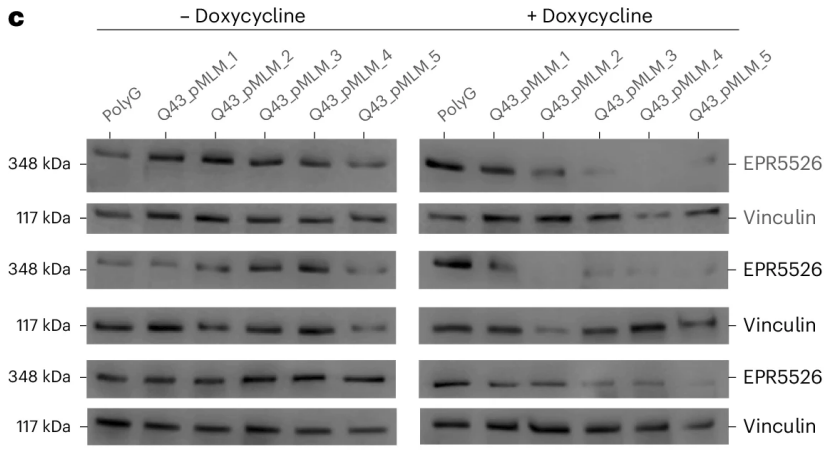
Fig. 1 | Overview and evaluation of the PepMLM model. a, The architecture of the PepMLM model. Based on the finetuning of ESM-2, the model incorporates the target protein sequence along with a masked binder region during the training phase. During the generation phase, the model can accept target protein sequences and mask tokens to facilitate the creation of peptides of specified lengths. **b**, Perplexity distribution comparison. The perplexity values were calculated for test and designed peptides, encompassing the target proteins in the test set. **c**, The density distribution visualization of the log perplexity values for target–peptide pairs, encompassing test peptides, PepMLM-650M-designed peptides, ESM-2-650M-designed peptides and random peptides. **d**, In silico hit rate assessment of RFdiffusion (left) and PepMLM (right). Using AlphaFold-Multimer, ipTM scores were computed for both the designed and test peptides in conjunction with the target protein sequence. The entries are organized in accordance with the ipTM scores attributed to the test set peptides. The hit rate is characterized by the designed peptides exhibiting ipTM scores ≥ those of the test peptides. **e**, Binding specificity analysis through permutation tests. The distribution of PPL scores for matched target–binder pairs (blue) is compared with randomly shuffled mismatched pairs (red). Each target's binder was shuffled 100 times to generate the mismatched distribution. Statistical significance was determined using t-test ($P < 0.001$). **f**, Structural comparison of computationally designed and experimental peptide binders in complex with their target proteins. Target proteins (gray) are shown in complex with PepMLM-designed binders (red) and experimental test binders (blue), with contact residues highlighted in corresponding colors. Top, mouse H-2Kb MHC complex (PDB ID: 2OI9) with designed peptide PSLGSVPYV (ipTM: 0.9) and test peptide QLSPFPFDL (ipTM: 0.9). Bottom, human tyrosine kinase complex (PDB ID: 1LCK) with designed peptide PPAEEIPP (ipTM: 0.82) and test peptide EGQQPQPA (ipTM: 0.68). **g**, Frequency distribution of individual amino acids among peptide binders ($n = 203$), comparing the test set (blue), PepMLM-designed sequences (red) and ESM2-650M-designed sequences (green). **h**, Amino-acid-specific generation distribution at contact positions (8-Å threshold). The heatmap shows the percentage of designed amino acids (y axis) given each amino acid in test binders (x axis).

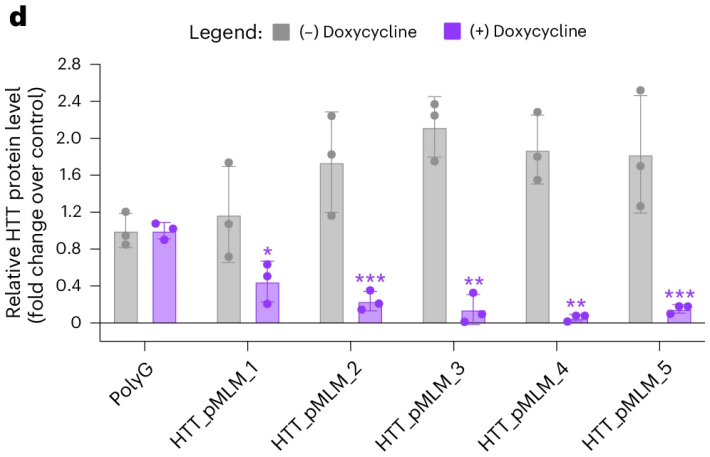
Chen et al., 2025

Degradation of disease-related Huntington proteins by PepMLM-designed binders









C,d = switchable system

Chen et al., 2025

Protein binders and artificial proteins: Implications for AI/ML

- Instead of small molecule drug discovery
  - small graphs, 30–80 heavy atoms
- We now have:
  - Sequences of 10–1500 amino acids
  - 3D structures with long-range constraints
  - Evolutionary priors
  - Functional surfaces instead of binding pockets
- How do we represent proteins in a way that captures structure and function?
  - Sequence-only models? Structural models? Combination?
- Drug discovery is shifting from predicting binding of fixed molecules to **designing dynamic biological systems**.

Discussion

- How might you adapt your current AI/ML knowledge to the context of peptide and protein design?

3. When drugs are gene therapy and genetic editing

Gene therapy in the clinic

- Decades of toxicity in clinical trials
- Delivery challenges
- A few approved gene therapy drugs, lots of phase 1 clinical trials
- Dec 2023: first CRISPR gene therapy approved
- CRISPR: gene editing
  - FDA called for long-term studies of off-target gene editing

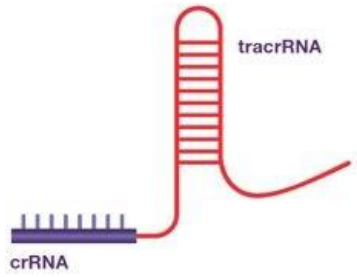
Human Gene Therapy Products Incorporating Human Genome Editing

Guidance for Industry

Gene editing therapies: CRISPR

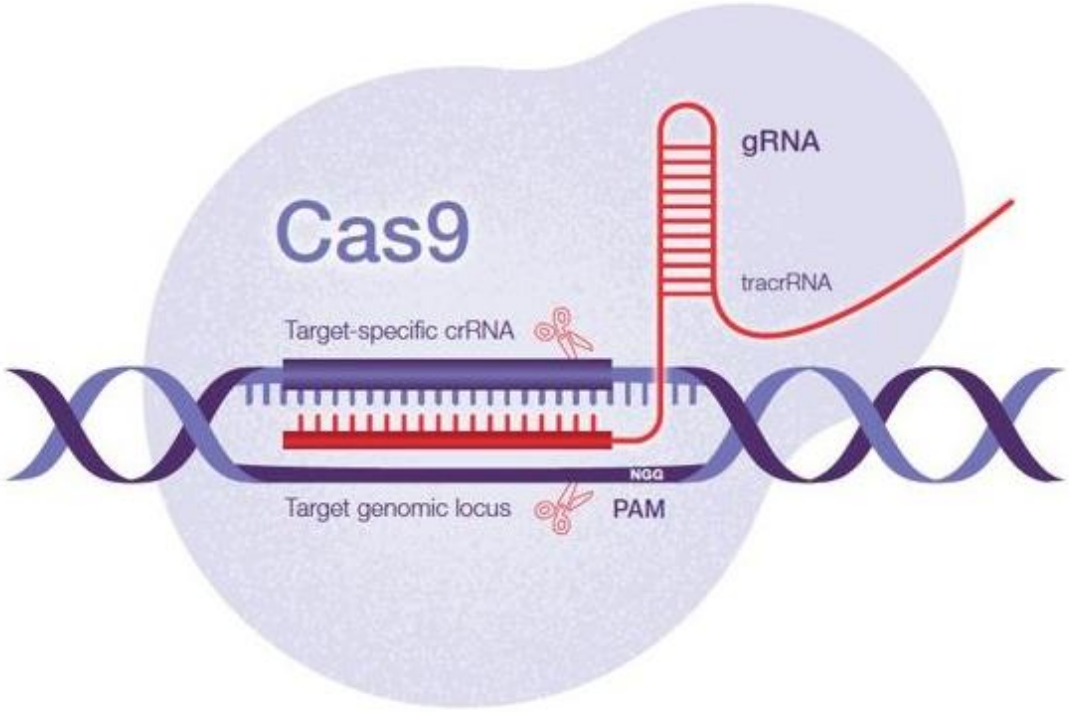
Basic CRISPR Workflow

- Guide RNA(gRNA) binds DNA
- Cas nuclease cuts



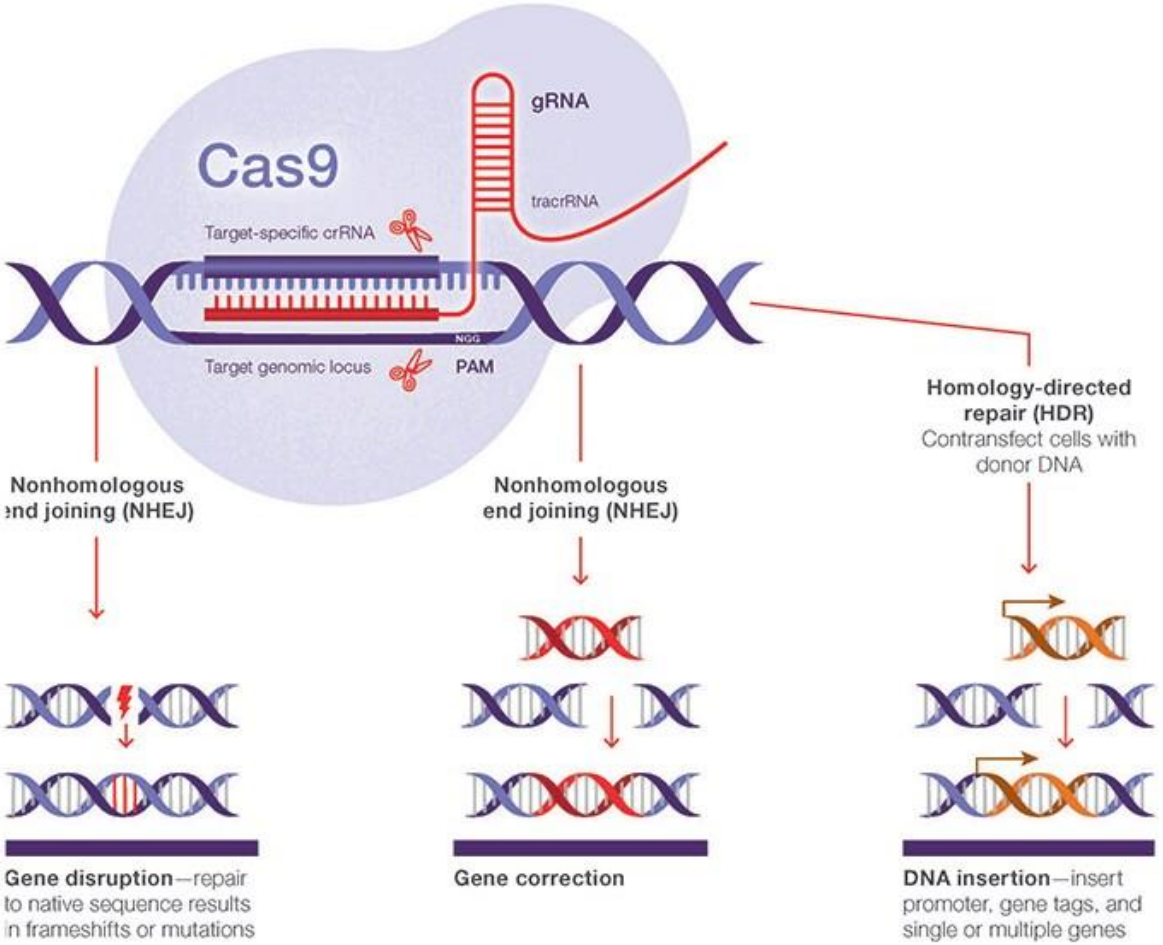
gRNA

Cas9 Protein



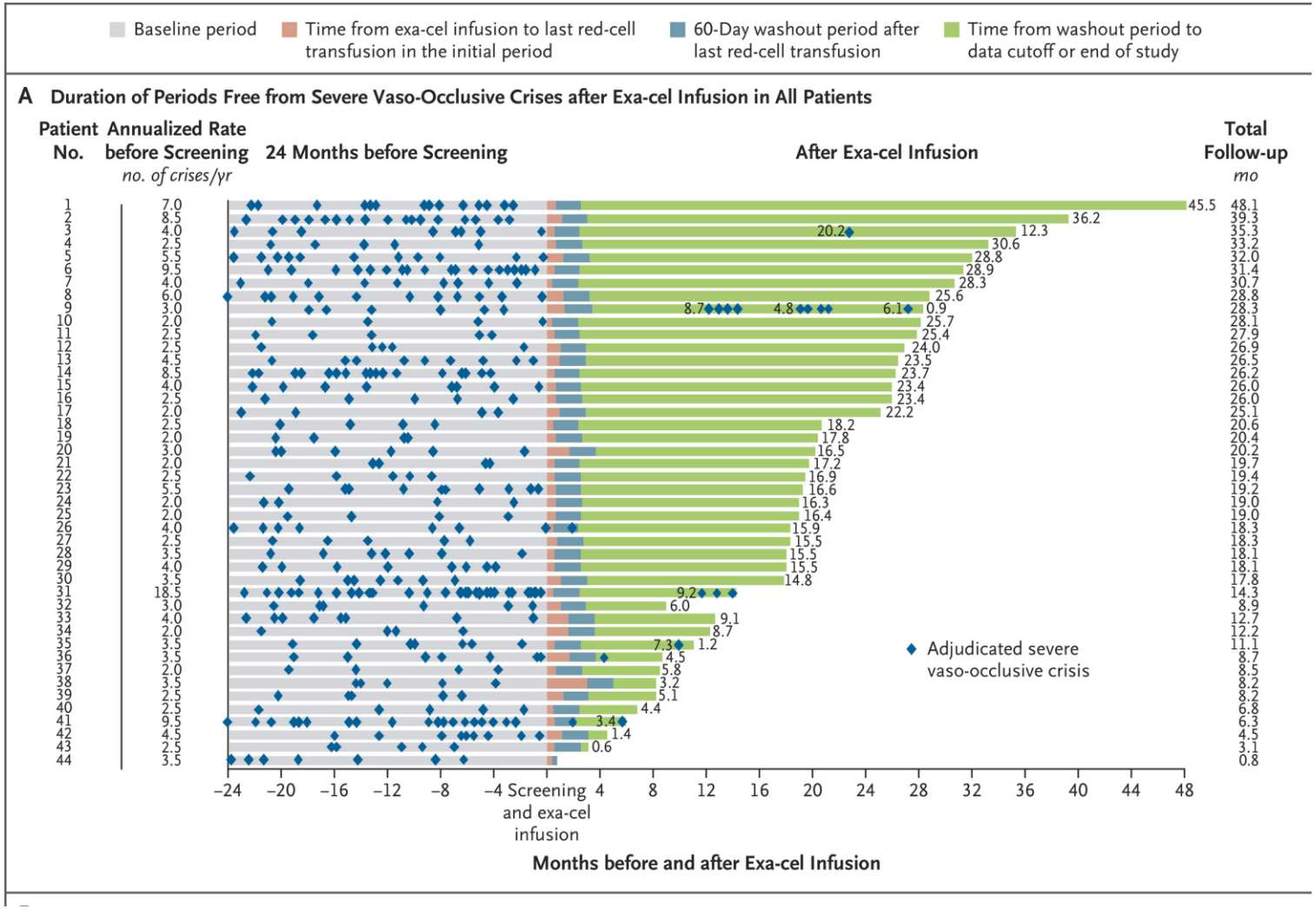
Images: ThermoFisher Scientific

Human DNA repair pathways determine outcome of gene editing



Images: ThermoFisher Scientific

Casgevy clinical trial data: Efficacy



Frangoul et al., 2024 NEJM

Gene editing therapies: Problems relevant for AI/ML

- gRNA efficiency prediction
- Off-target prediction
- Base editing outcome modeling
- Prime editing outcome modeling
- DNA repair outcome distribution modeling

} More precise and “corrective” types of gene edits

- Gene editing outcome is probabilistic, not deterministic.

Active ML contributions in this space

nature biomedical engineering

Article

https://doi.org/10.1038/s41551-025-01463-z

CRISPR-GPT for agentic automation of gene-editing experiments

Received: 25 June 2024

Accepted: 17 June 2025

Yuanhao Qu[1,9], Kaixuan Huang[2,9], Ming Yin[2], Kanghong Zhan[3], Dyllan Liu[4], Di Yin[1], Henry C. Cousins[5,6], William A. Johnson[1], Xiaotong Wang[1], Mihir Shah[5], Russ B. Altman[4,7], Denny Zhou[8], Mengdi Wang[2] & Le Cong[1]

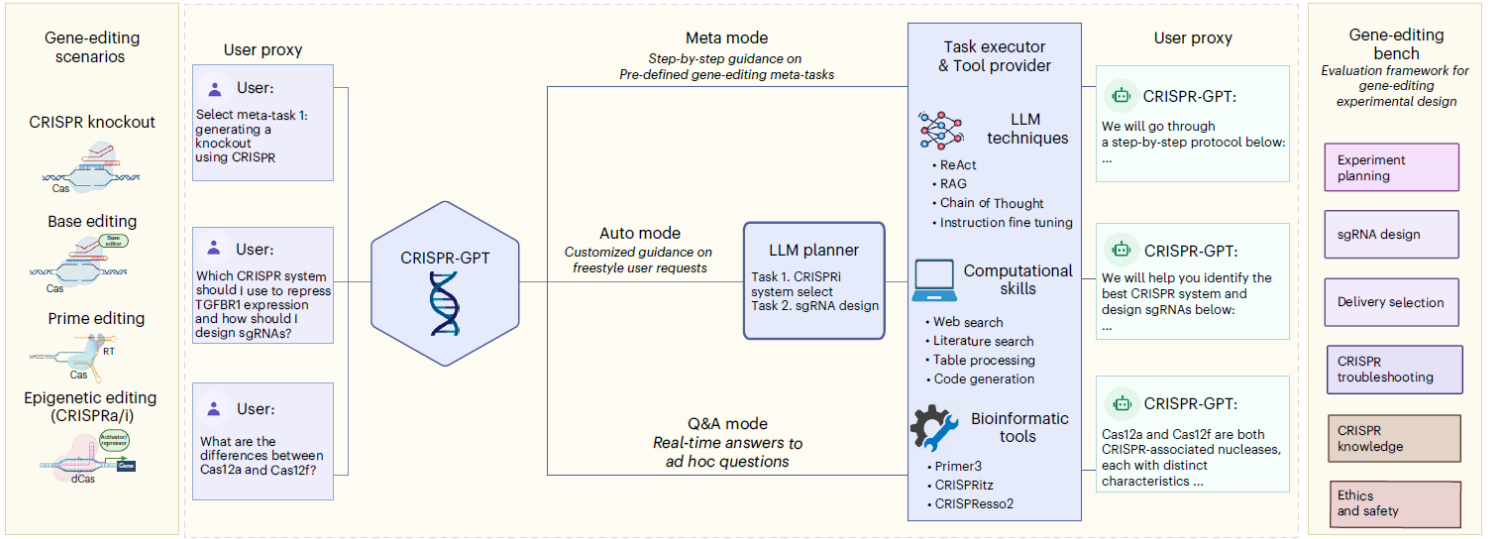


Fig. 1 | Overview of CRISPR-GPT. CRISPR-GPT is an LLM-powered multi-agent system designed to provide AI copiloting for human researchers in gene editing. It supports four primary gene-editing modalities: knockout, base editing, prime editing and epigenetic editing (CRISPRa/i). The system offers three user interaction modes: Meta mode (step-by-step guidance on predefined tasks), Auto mode (customized guidance based on user requests) and Q&A mode (real-time answers to ad hoc questions), to streamline experiment design and planning. CRISPR-GPT consists of four core components: the User proxy, LLM planner, Task executor and Tool provider. Together, these components are equipped with a comprehensive suite of tools and decision-support capabilities to facilitate the design, planning and analysis of gene-editing workflows. To evaluate CRISPR-GPT's performance, we developed the Gene-editing bench, a framework of 288 test cases covering tasks such as experimental planning, sgRNA design, delivery method selection and more. Figure was originally created with BioRender.com/tb8sq6f.

Performance of CRISPR-GPT relative to general-purpose LLMs

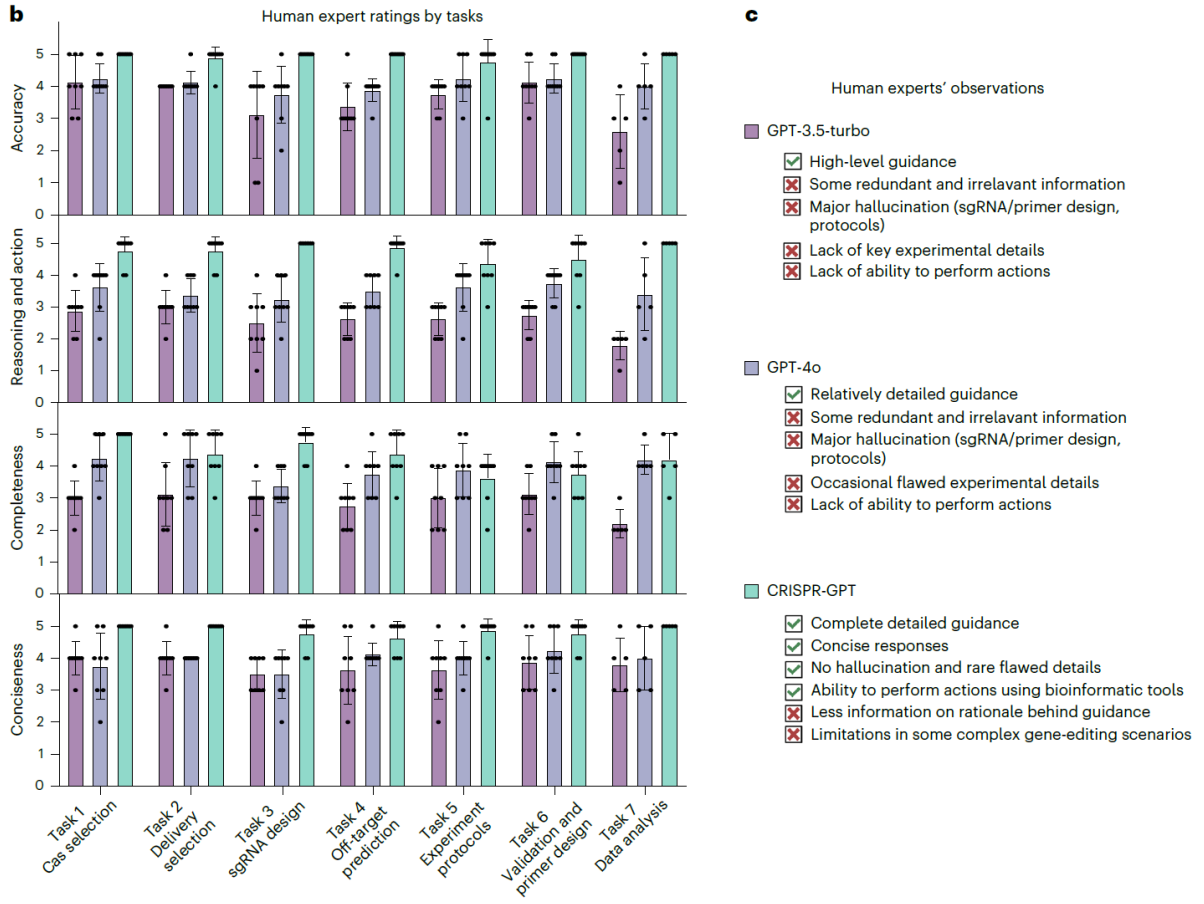


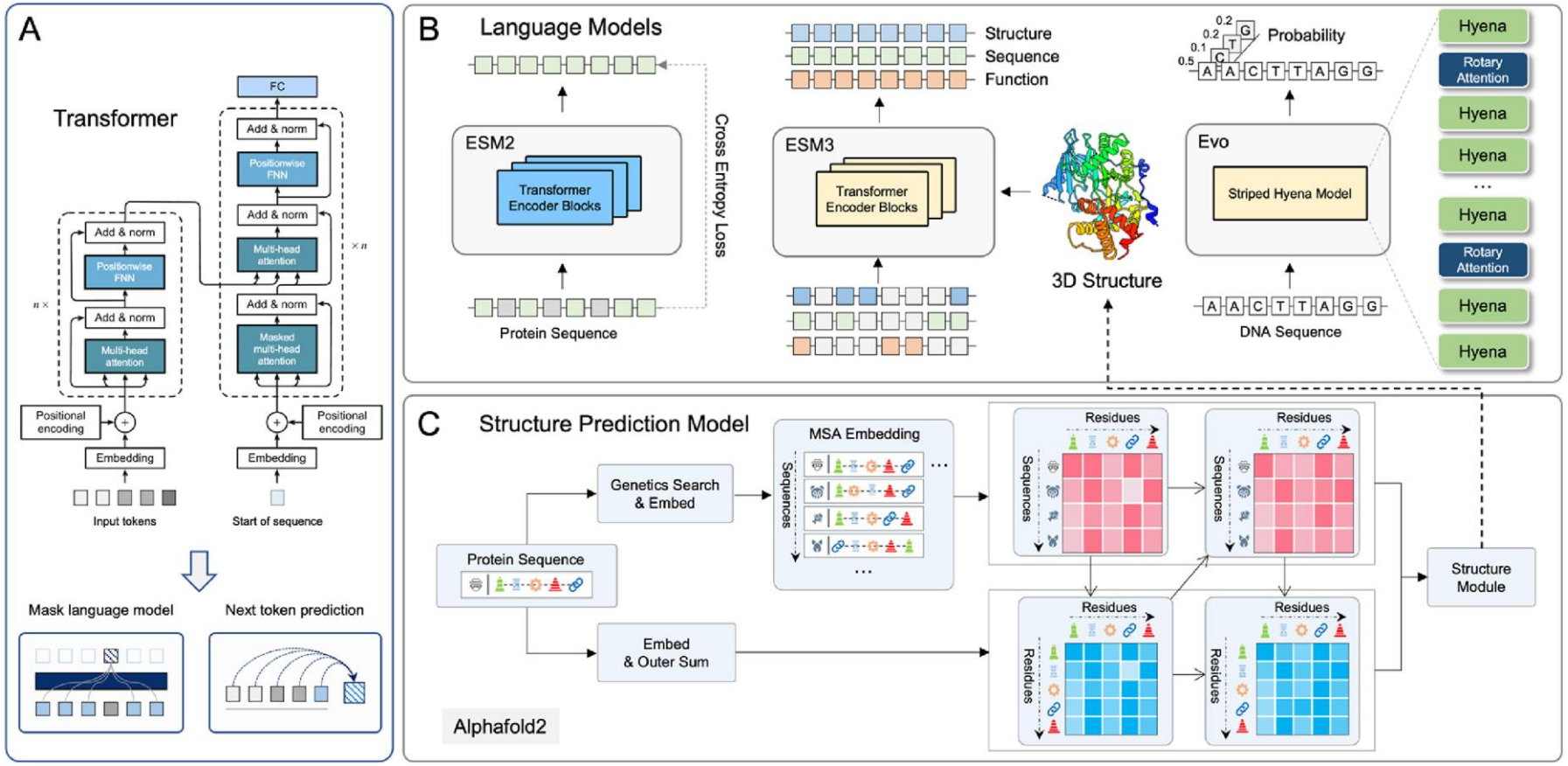
Fig. 5 | CRISPR-GPT outperforms general-purpose LLM for gene-editing research in human user experiences. a, Human user experience: evaluation in Supplementary Table 2). **b**, Human user experience: evaluation results breakdown by major gene-editing tasks. Data shown are the mean ± s.d.

Discussion:

- How do subject matter expertise and AI/ML combine in drug discovery projects involving emerging drug modalities?
- What top 3 questions do you as AI/ML students have for biologists in this space?

Supplementary Slides

Review paper: Algorithm development and application for sequence and structural biological data



(A) The Transformer architecture serves as the foundational backbone, utilizing attention mechanisms to decipher biological sequence patterns. (B) Biological language models have advanced from analyzing protein sequences (ESM2) to integrating multimodal data (ESM3) and genomic contexts (Evo). (C) Structure prediction models like AlphaFold2 leverage evolutionary information from MSAs to accurately generate 3D protein conformations.

Tang et al., Acta Pharmaceutica Sinica 2026

nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01165-w

Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer

Received: 8 May 2025

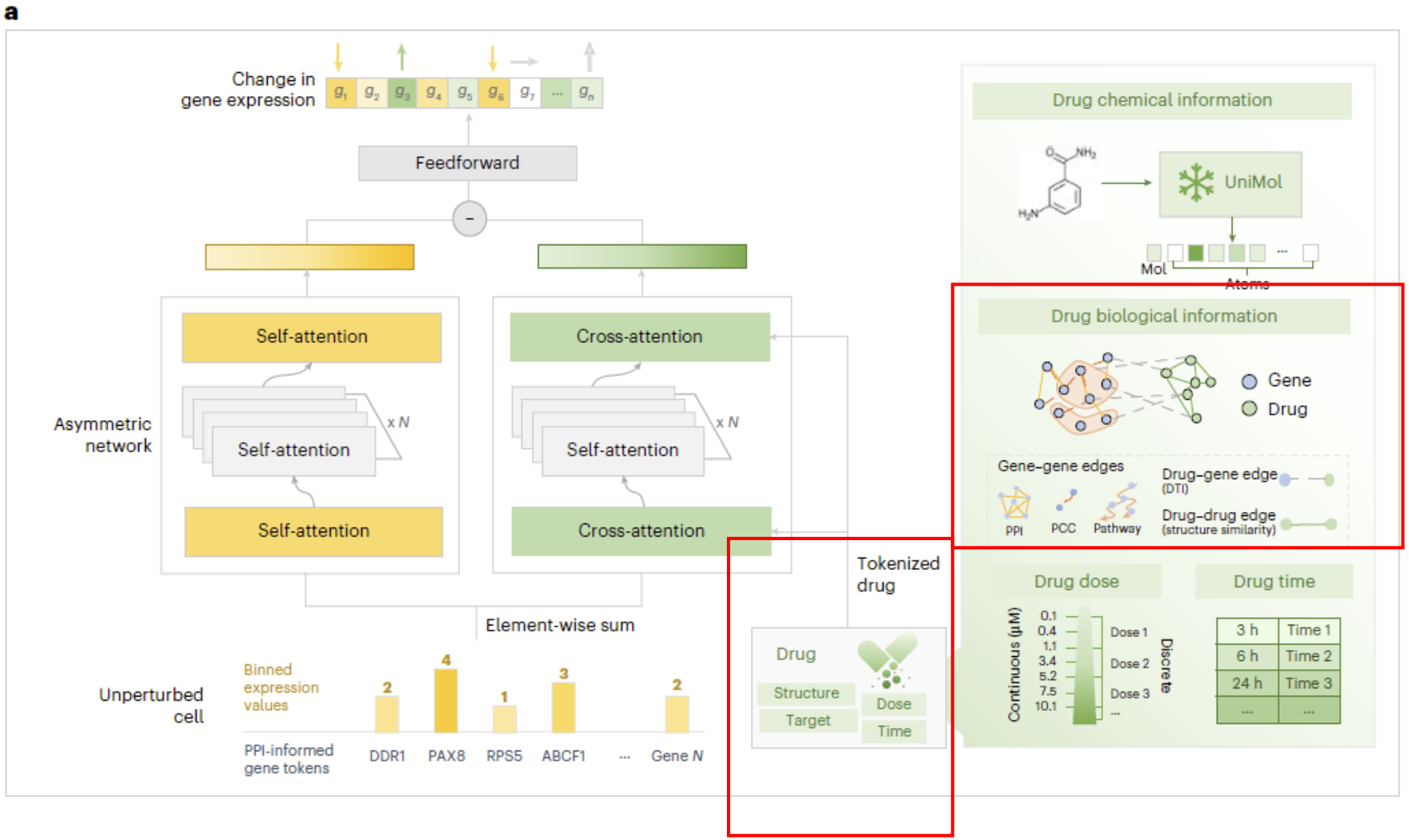
Accepted: 2 December 2025

Published online: 26 January 2026

Yue Guo[1], Hao Zhang[1,2], Haitao Hu[1,2], Jialu Wu[3], Ji Cao[1,4,5,6] ✉, Chang-Yu Hsieh[3,4] ✉ & Bo Yang[1,4,5,7] ✉

Systematic mapping of chemical perturbation responses is revolutionizing

Architecture of XPert



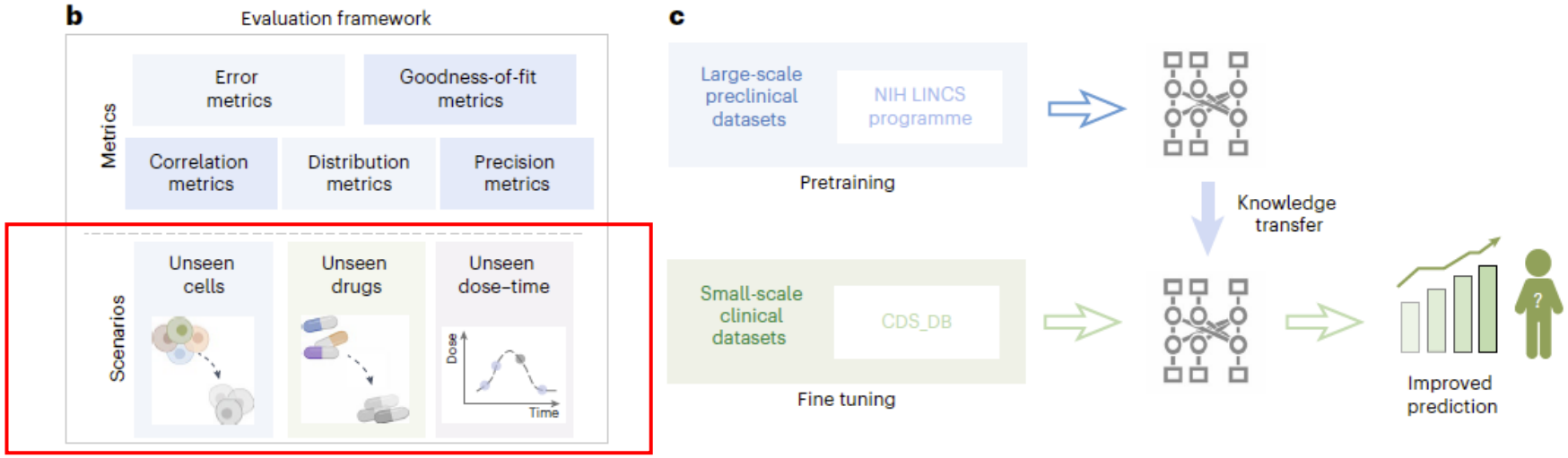


Fig. 1 | Overview of XPert. a, Architecture of XPert, featuring a dual-branch framework composed of self-attention and cross-attention modules. XPert receives inputs from both unperturbed gene expression and multiscale drug features, and predicts both change in gene expression (x_{deg}) and post-perturbation gene expression (x_{pert}). For the drug modality, chemical information is derived from the molecular representation model UniMol, whereas biological information is extracted from a pretrained heterogeneous knowledge graph, along with other tokenized variables such as drug dose and time. **b**, Evaluation framework used to assess XPert's performance, which includes five types of metric: error metrics, goodness-of-fit metrics, correlation metrics, distribution metrics and precision metrics. These are applied across three blind scenarios: novel cell lines, unseen drugs and unmeasured dose–time conditions. **c**, Pretraining and fine-tuning pipeline designed to address data scarcity in clinical applications. XPert is pretrained on large-scale preclinical perturbation datasets (for example, L1000) and then fine-tuned on smaller, clinical datasets (for example, CDS-DB), improving the prediction accuracy for clinical applications. Illustrations in **b** created with BioRender.com.

post-perturbation states.

Another fundamental challenge is how to translate chemical perturbations into biological perturbation signals. State-of-the-art (SOTA) approaches typically concatenate chemical and cellular features, capturing global-cell-state alterations but failing to resolve gene-specific responses. This limitation necessitates advanced fusion strategies that integrate prior knowledge (for example, drug–target interactions (DTIs)) to bridge chemical and biological spaces. A further gap is the inadequate modelling of the well-established dose- and time-dependent nature of drug effects[16,17]. Previous attempts have relied on simplistic encodings (for example, one-hot encoding), which is insufficient for modelling nonlinear dose–response relationships (for example, inverted U-shaped curves)[18], restricting a full understanding of transcriptional pharmacodynamics.

Drug embeddings by chemical space c) or pretrained biological / MoA space c)

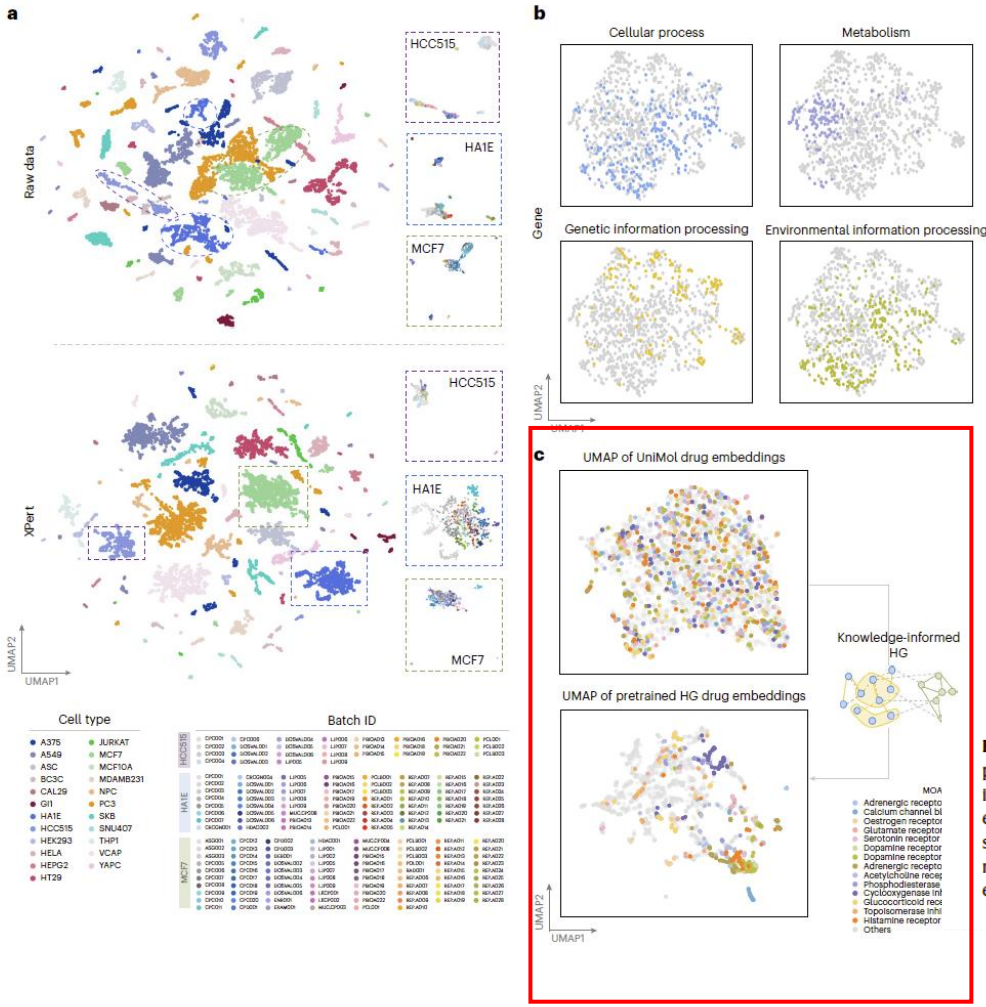


Fig. 3 | Biological knowledge interpretation in XPert. a, UMAP plots of post-treatment profiles and the < *cls* > token embeddings obtained from XPert in the test dataset, coloured by cell type and batch ID. XPert's < *cls* > embeddings effectively mitigate batch effects, leading to a more cohesive clustering of specific cell types. **b**, UMAP of gene token embeddings in XPert, coloured by four major Kyoto Encyclopedia of Genes and Genomes pathways. **c**, UMAP of drug embeddings, coloured by the drug MoA. The top plot shows the drug embeddings obtained from UniMol (representing the chemical space of drugs), whereas the bottom plot uses pretrained HG embeddings (representing the biological space). Drugs with similar MoAs cluster together in biological space rather than chemical space. **d**, SAR of EGFR inhibitors. The atom weights of two EGFR inhibitors—gefitinib and erlotinib—are displayed, highlighting key substructures and their consistency with the SAR.